TRAINEE PERFORMANCE PREDICTION MODEL FOR HAWASSA POLYTECHNIC COLLEGE USING KDP

TIZITA G/SILASSIE

HAWASSA UNIVERSITY, HAWASSA, ETHIOPIA

November, 2022

TRAINEE PERFORMANCE PREDICTION MODEL FOR HAWASSA POLYTECHNIC COLLEGE USING KDP

TIZITA G/SILASSIE

MAJOR ADVISOR:DEGIF TEKA(PhD)

CO-ADVISOR:MR. KIBREBEAL

A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE

HAWASSA UNIVERSITY

INSTITUTE OF TECHNOLOGY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

HAWASSA, ETHIOPIA

November, 2022

APPROVAL SHEET-I

This is to certify that the thesis entitled “TRAINEE PERFORMANCE PREDICTION MODEL FOR HAWASSA POLYTECHNIC COLLEGE USING KDP” submitted in partial fulfillment of the requirements for the degree of Master's with specialization in Computer Science, the Graduate Program of the Department/School of Informatics, and has been carried out by Tizita G/silassie. Therefore we recommend that the student has fulfilled the requirements and hence hereby can submit the thesis to the department.

_____	_____	_____
Name of major advisor	Signature	Date

_____	_____	_____
Name of co-advisor	Signature	Date

APPROVAL SHEET-II

We, the undersigned, members of the Board of Examiners of the final open defense by Tizita G/silassie have read and evaluated his/her thesis entitled "TRAINEE PERFORMANCE PREDICTION MODEL FOR HAWASSA POLYTECHNIC COLLEGE USING KDP ", and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirements for the degree.

_____	_____	_____
Name of Major Advisor	Signature	Date
_____ _	_____	_____
Name of Internal Examiner-I	Signature	Date
_____	_____	_____
Name of Internal Examiner-II	Signature	Date
_____	_____	_____
Name of External examiner	Signature	Date
_____	_____	_____

ACKNOWLEDGEMENT

First and foremost, praises and thanks to the God, the Almighty, for His blessings, courage, and patience in order to complete the research successfully. I would like to express my deep and sincere gratitude to my research Advisor, Dr. Degif Teka and my co-advisor Mr. Kibrebeal for their invaluable guidance, constructive comments and suggestions throughout the preparation of this research paper.

I am extremely grateful to my mother W/ro Tayech Bahiru, for her love, prayers, caring and sacrifices for educating and preparing me for my future. I am extending my heartfelt thanks to my beloved husband Ato Bereket Buche and my daughter Abiya for their love, understanding, prayers and continuing support to complete this research work.

Also I express my sincerest graduated and heartfelt thanks to W/ro Muluwork Derese for her encouragement, courteous support and helpful personality. Moreover, I would like to thank staff members of Hawassa Polytechnic college especially W/ro Rosa Mekonen for their cooperation and giving of required information about Hawassa Polytechnic college.

Finally, my deepest gratitude to Hawassa University for giving me this chance to pursue my MSc study. And thanks to all the people who have supported me to complete the research work directly or indirectly.

STATEMENT OF THE AUTHOUR

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: ……...……. Signature: …….……

Place: Institute of Technology, Hawassa University, Hawassa

Date of Submission: …….

LIST OF ABBREVIATIONS

KDP	Knowledge Discovery Process
CRISP-DM	Cross Industry Standard Process for Data Mining
TVET	Technical & Vocational Education Training
DM	Data Mining
VB	Visual Basic
SEMMA	Sample Explore Modify Model Assess
SAS	Statistical Analysis System
SMOTE	Synthetic Minority Over Sampling Technique
WEKA	Waikato Environment for Knowledge Analysis

LIST OF FIGURES

Figure 2.1. The six-steps of KDD model 9
Figure 2.2: Phases of the CRISP-DM .....**Error! Bookmark not defined.**
Figure.2.3. Data mining tasks. 16
Figure 2.4. Decision tree Structure 18
Figure 2.5.Four Outcomes of Classifier 27
Figure3.1: The six steps of Hybrid data mining model (KDP) 34
Figure 3.2: Ethiopian Educational Cycle 35
Figure 4.1.WEKA Snapshot Side by side view of the class attribute creditor's Original data ..45
Figure 4.2 Side by side view of the class attribute creditor's Resample (balanced) data 45
Figure 4.3. WEKA Snapshot of attribute selection using infogain ranker. 46
Figure 4.4 WEKA Snapshot of 10-Fold cross Validation 51
Figure 4.5: WEKA Snapshot of J48 pruned decision tree with 80 % split test mode 52
Figure 4.5: WEKA Snapshot J48 pruned decision tree with 66% split test mode 53
Figure 4.6: WEKA Snapshot J48 Un-Pruned Decision Tree with 10-fold cross validation 55
Figure 4.7: WEKA Snapshot of J48 Un-Pruned Decision Tree with 80% split test mode 56
Figure 4.8: WEKA Snapshot of J48 Un-Pruned Decision Tree with 66% split test mode 57
Figure 4.9: WEKA Snapshot of JRIP 10-Fold cross Validation 59
Figure 4.10: WEKA Snapshot of JRIP Pruned 80% split Test Data 60
Figure 4.11: WEKA Snapshot of JRIP Pruned 66% split Test Data 61
Figure 4.12: WEKA Snapshot of JRIP Un-Pruned 10-Fold cross Validation.....63
Figure 4.13: WEKA Snapshot of JRIP Un-Pruned 80% split Test Data.....64
Figure 4.14: WEKA Snapshot of JRIP Un-Pruned 66% split Test Data.....65
Figure 4.15: WEKA Snapshot of Naïve Bayes 10-fold cross validation Test Data 67
Figure 4.16: WEKA Snapshot of Naïve Bayes 80% split Test Data.....68
Figure 4.17: WEKA Snapshot of Naïve Bayes 66% split Test Data.....69
Figure 4.18: WEKA Snapshot of PART Pruned 10-fold cross validation Test Data.....71
Figure 4.19: WEKA Snapshot of PART Pruned 80% split Test Data 72
Figure 4.20: WEKA Snapshot of PART Pruned 80% split Test Data 73

Figure 4.21: WEKA Snapshot of PART Un-Pruned 10-fold cross validation Test Data 75
Figure 4.22: WEKA Snapshot of PART Un-Pruned 80% split Test Data 76
Figure 4.23: WEKA Snapshot of PART Un-Pruned 66% split Test Data 77
Figure 4.24: WEKA Snapshot of PART Un-Pruned 10-fold Cross validation 80
Figure 4.25: Snapshot Log In Page 85
Figure 4.26 Snapshot of Main Page for Performance Prediction 86

LIST OF TABLES

Table 2.1. Two dimensional confusion matrix 27
Table3.1: Description of the selected attributes from Hawasa Poly Technic College dataset ...38
Table 3.2: Additional attributes with their description.....39
Table 3.3: Discretized age attribute 43
Table 3.4: A discretized transcript result attribute.....41
Table 4.1. Experimentation ways of whole attributes and selected attributes.....47
Table 4.2: Description of parameters to be tuned in all classification modeling 48
Table 4.3: Summary of experimental result of J48 pruned Decision Trees' algorithm 54
Table 4.4 Summary of experimental result of J48 unpruned Decision Trees' algorithm 58
Table 4.5 Summary of experimental result of JRIP pruned algorithm.....62
Table 4.6 Summary of experimental result of JRIP Un-pruned algorithm 66
Table 4.7 Summary of experimental result of Naïve Bayes algorithm 70
Table 4.8 Summary of experimental result of PART Pruned algorithm.....78
Table 4.9 Summary of experimental result of PART Un-Pruned algorithm.....78
Table 4.10: Comparison of experiment accuracy form the selected Algorithms 79

ABSTRACT

Data mining is the key tool for discovery of knowledge from large data set. It is this technology in most of the educational organization of the world currently helping to know the organization data explicitly and pave the way to produce quality citizens. Unlike other sectors the power of data mining is not much exploited in educational sector. Although there are studies regarding academic performance of students using data mining techniques, they are all about university students. we cant find academic performance of trainee research in Technical and vocational Institute. Thus, the purpose of this study is to develop Trainee performance prediction model for Hawassa polytechnic college.

A total of 8200 records with 13 attributes were collected from Hawassa Polytechnic College registrar data set of the past 5 years ranging from 2009 to 2013 E.C. An experiment has been conducted using the Knowledge Discovery Process (KDP) Model using WEKA software version 3.8.4. Four data mining algorithms namely J48 Decision Trees, JRip rules induction, Naïve Bayes and PART with seven experiments (J48 Pruned and Un-pruned decision tree algorithm, Naive Bayes classifier, JRIP Pruned and Un-pruned and PART Pruned and Un-pruned) were used to develop trainees performance predictive model. All the experiments were carried out with the same dataset and evaluated with 10-fold cross validation, 80% and 66% split test parameters. The study shows PART Un-pruned 10-fold cross validation test has the highest accuracy with 95.4268% and attributes such as trade/occupation, EGSECE, transcript, level, sex, English, and sector can be used at a time of decision making as they have shown strong prediction power which can help to predict trainees performance. Finally the researcher develop a prototype based on the rules generated from the selected algorithm.

TABLE OF CONTENT

Declaration.....**Error! Bookmark not defined.**
Acknowledgement..... I
List Of Abbreviations.....III
List of Figures..... IV
List Of Tables..... VI
Abstract..... VII
CHAPTER ONE.....1
INTRODUCTION.....1
1.1 Background.....1
1.2 Background of the Study.....2
1.3 Statement of the problem.....3
1.4 Objectives of the Study.....4
1.5 Methodology.....4
1.6 Significance of the Study.....6
1.7 Scope of the study.....7
1.8 Organization of the Thesis.....7
CHAPTER TWO.....8
LITERATURE REVIEW.....8
2. OVERVIEW OF DATA.....8
2.1 Data Mining Process Models.....8
2.2 Data Mining Tasks.....15
2.3 Predictive Data Mining.....23
2.4 Attribute Selection Measures.....23
2.5 Implementation Tools.....25
2.6 Model Evaluation.....26
2.7 Educational Data Mining.....29
2.8 Related work.....30
CHAPTER THREE.....33
METHODOLOGY.....33

3. Research Design 33
3.1 Understanding of the Problem Domain 34
3.2 Understanding of the Data 36
3.3 Preparation of the Data 37
3.4 Data mining 42
3.5 Evaluation of the discovered knowledge 42
3.6 Use of the discovered knowledge 42
CHAPTER FOUR 44
EXPERMENTAL ANALYSIS AND RESULT 44
4.1. Introduction 44
4.2. Balancing Instances 44
4.3. Attribute Selection 46
4.4 Experimental Setup 47
4.5 Experimental Results 48
4.5.1 MODEL BUILDING USING J48 DECISION TREE 50
4.5.2 MODEL BUILDING USING JRIP ALGORITHM 59
4.5.3 MODEL BUILDING USING NAÏVE BAYES ALGORITHM 67
4.5.4 MODEL BUILDING USING PART ALGORITHM 71
4.6 Comparison of the experimented classification models 79
4.7 Selecting Classification Model 81
4.8 Generating Rules 81
4.9 Deployment of the result 83
CHAPTER FIVE 87
CONCLUSION AND RECOMMENDATION 87
5.1. Conclusion 87
5.2. Recommendations 88
References 90
Appendix I: Sample Trainee Records in CSV format 94
Appendix II: Sample rules generated from PART Un-pruned 10-fold cross validation 95
Appendix I: Sample Trainee Records in CSV format 94
Appendix II: Sample rules generated from PART Un-pruned 10-fold cross validation 95

CHAPTER ONE

1.1. INTRODUCTION

Background

One of the basic and main roles of Higher education is to support the development of a society and the world. To increase trainees success rate, it is vital to understand and define academic success[1]. According to Ramageri [2], a trainee's academic performance is a crucial part of an academic institution. This is considered one of the important measures for many universities. Some people say that the trainee's academic performance can be measured through learning assessment and co-curricular activities. The majority of people have mentioned that the trainee's past performances, achievements, and grades can play a vital role to predict the trainee's success rate. Predominantly, most of the higher-level institutions use grades as the main measure to assess trainee's performance. In addition, assignment marks, and final exam scores will affect the trainee's academic performance.

Data Mining involves analysis and improvement in the prediction methods of trainee performance. Based on prediction results, if the trainee needs are fulfilled timely then the overall result and performance will increase year by year. For performance analysis and prediction, important attributes of the trainees are gathered. Subsequently, various data mining techniques and classification algorithms are applied to get deeper insights and predictions. The learning and experimentation from the trainee performance data can help educational experts and other stakeholders make decisions in improving the teaching-learning process.

Educational data mining (EDM) is a technique used in educational and academic institutions. It is theory-oriented and aims to develop computational approaches that combine theory and data to assist with and enhance the quality of education and improve academic performance of trainees and graduates. Machine learning techniques in educational data mining aim to develop a model for discovering meaningful hidden patterns and exploring useful information from educational settings[3]. Prediction of trainee's performance became an urgent desire in most educational entities and institutes. That is essential to help at-risk trainees and assure their

retention, providing excellent learning resources and experience, and improving the university's ranking and reputation. Extensive efforts have been made to predict trainee performance for different aims, like detecting at-risk trainees, assurance of trainee retention, course and resource allocations, and many others[4]. Before designing and implementing any pedagogical and instructional interventions to improve trainee performance, it is important to develop an effective model to predict trainee academic performance so the instructor can know how well or how poorly the trainees in the class will perform[5].

1.1. Background of the Study

The study considers Hawassa Polytechnic College as a case. Hawassa Polytechnic College is one among public TVET colleges in the Sidama national regional state, Ethiopia and located at Hawassa city administration. Since its establishment in 1998 G.C, the college is offering technical and vocational education and training for the local community through both formal and non-formal programs. even though there were limited resources the college struggles to produce qualified graduates. As TVET institute becoming more and more unavoidable for development of a nation the government give remarkable attention in terms of economic and strategic direction for all TVET institution of a nation and this helps Hawassa Polytechnic College to provide quality service in almost 26 occupations, some ranges to level five, to produce graduates having supervisory to managerial skill.

Hawassa Polytechnic College aims for developing quality technical and vocationally trained person. The College provide opportunities to foster entrepreneurship and deliver standardized industry extension services in the community. Furthermore, it shall play a vital role in reducing unemployment of youth population and enabling the industries to have trained workforce which is among factor for their productivity.

As colleges' brochure published on 2013 regarding with human resource there are 318 workers of which 207 male 111 female. When we see this figure the majority goes to academic which accounts 223 (160 male and 63 female) and they provide services in organized manner as industrial, economic and Hotel, Tourism and sport sector. In addition, the academic status of

these trainers for last year were 30 at “A” level, 90 at “B” level and 103 “C” level trainers serving the college both in training, technology transfer and industry extension services.

Finally, these days the college is offering training for over 6,000 trainees in regular and extension programs and more than 10,000 trainees at short term program using its limited physical facilities when the standard to deliver quality training concerned.

1.2. Statement of the problem

One of the most important indicators of the effectiveness of teaching can be the academic achievement of learners, which can be influenced by different factors such as learning methods and individual motivations. However, some studies have shown that some trainee characteristics are significant predictors of trainees academic achievement whereas some are not[6]. The problem of accurate trainee performance prediction is still a challenging task due to various issues and many other factors are involved in it. The main issues in the performance prediction methods are inefficiency and the use of improper attributes or variables. This problem can be considered a hard problem because the performance depends on many characteristics related to the trainees. These characteristics can be categorized into a trainee's status or assessment result, demographics, physical profile, academic progress, and educational background. The trainee's assessment result is the most important attribute used to predict performance.

Appropriate career selection is one of the key decisions in trainee’s life. Wrongly chosen careers sometimes destroy trainee’s life. Right career selection for a trainee is always based on the selection of the right educational program.

Several studies have been conducted to identify the factors that determine the academic performance of university students. But insufficient studies are conducted on TVET that help to predict the trainee’s academic performance especially in Ethiopian TVET education. To this end, this study tries to answer the following research questions.

1. What are the major attributes that affect trainee academic performance at Hawassa Polytechnic College?

2. Which Data Mining techniques are more appropriate to predict TVET's trainee academic performance

1.3. Objectives of the Study

1.3.1. General Objective

The general objective of this study is developing a predictive model for predicting the Trainees academic performance the case of Hawassa Poly Technical College.

1.3.2. Specific Objectives

The specific objectives of this study are:

- To Explore literature for trainee performance prediction and factors affecting trainee performance
- To Identify factors that affect trainee performance
- To identify attributes that are associated with Trainee Academic Performance using data mining technique
- To Select the best algorithms that have been used for predicting the trainees academic performance,
- To Pre-processing the data to make it ready for experimentation
- To evaluate the developed model by using performance evaluation metrics.

1.4. Methodology

There are different types of standards and methodologies being used in data mining (DM) researches. The development of academic and industrial models has led to the development of hybrid models i.e., models that combine aspects of both. This process model is selected because it provides more general, research-oriented description of the steps, introducing a data mining step instead of the modelling step, introducing several new explicit feedback mechanisms. One such model is a six-step KDP model developed by [7]. It was developed based on the CRISP-DM model by adopting it to academic research. A description of the six steps follows.

Understanding of the problem domain: This initial step involves working closely with Hawassa Poly technic College Registrar officer, Vice dean and COC Co-ordinator to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem through discussion and furthermore literature review has been carried out on the existing study area, such as documents published by College, Regional Or National TVET Agency.

Understanding of the data: This step includes collecting sample data and deciding which data, including format and size, will be needed. The researcher used data collected from Hawassa College trainee's database that covers from 2009-20013 E.C the data extracts from Microsoft Excel 2007 and 2010 and a total of 13 attributes (columns) and 8200 records (rows) were identified. Background knowledge can be used to guide these efforts. Finally, the step includes verification of the usefulness of the data with respect to the DM goals.

Preparation of the data: This step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms by derivation of new attributes and by summarization of data. The results are data that meet the specific input requirements for the DM tools selected in Step1.

Data mining: Hybrid Data mining process model was selected and this method involves six steps and the researcher thoroughly passes through all the steps and iterated as needed. The study was conducted using WEKA software Version 3.8.4. Different experiments are conducted using J48 decision tree algorithm, PART rule induction, Naïve Bayes, and JRIP rule algorithm using 10- fold cross validation and percentage split.

Evaluation of the discovered knowledge: Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative

actions could have been taken to improve the results. A list of errors made in the process is prepared.

Use of the discovered knowledge: This final step consists of planning where and how to use the discovered knowledge. In this study, we also developed a simple prototype using VB.NET programming language. To use the hidden patterns extracted using data mining technique, a user interface is designed with the help of VB.NET programming languages. This programming language is selected since it is general purpose, object object-oriented language and its familiarity with the researcher.

1.5. Significance of the Study

This study will have significance for various bodies to show a conceptual approach to predict the trainee's performance. The study will help to understand what factors influence the performance of the trainees in TVET. Besides, it will serve as a viewpoint for various research works intended to study for a specific determinant of the trainee's performance.

Predicting the academic performance of the trainee has an essential significance for the trainer to predict trainee academic performance and then take some proactive measures to improve the trainee's performance. This study plays a significant role in supporting parents and educational leaders in predicting trainee's performance. One of the key contributions of this study will be a critical analysis of the factors influencing trainee's academic performance which play an important role in performance prediction.

1.6. Scope of the study

The scope of this study will be limited to develop a model for predicting Hawassa Poly Technical College trainee's performance. The trainee's academic records will be collected from Hawassa Poly Technical College registrar office.

1.7. Limitation

The limitation of this study was we could not get all needed factors to determine higher education trainees 'performance because of unavailability of clear data found in the database and the researcher decides such as family size, family background and health-related data not included under this study.

1.8. Organization of the Thesis

This thesis is divided into five chapters. The first chapter is an introduction part, which contains background of the study, statement of the problem, objectives, Significance of the study, scope and methodologies of the research. The second chapter is devoted to literature review on data mining technology and machine learning. And it also covers review of applicable data mining techniques including related research works. The third chapter discusses about the methodology followed to conduct the study. That comprises model building steps such as business understanding, data understanding, and data preparation. The fourth chapter is dedicated to Experimental analysis and result using different data mining algorithms, evaluation, and deployment of the result. Results of the experiment are also analysed and interpreted in chapter four. The last chapter which is Chapter five presents the conclusion and recommendation that summarize the major points of the research and recommendations forwarded for practice and further research and adjustments about organization on the ground of the research results.

CHAPTER TWO

2. LITERATURE REVIEW

OVERVIEW OF DATA MINING

As illustrated in [3] the term data mining is referred as “the process of extracting or mining Knowledge from large amount of data”. Today, the ever-growing influence of computer technology put its impact on every sector. Thus, the large amount of data must have to be extracted or mined in order to get a hidden pattern which is useful for the sector to make it more competitive and profitable.

2.1. Data Mining Process Models

Data mining process defines a sequence of steps that should be followed to discover knowledge (e.g., patterns) in data. Each step is usually realized with the help of available commercial or open-source software tools. To formalize the knowledge discovery processes (KDPs) within a common framework, there is the concept of a process model [4]. The model helps organizations to better understand the KDP and provides a roadmap to follow while planning and executing the project. This in turn results in cost and time savings, better understanding, and acceptance of the results of such projects. We need to understand that such processes are non-trivial and involve multiple steps, reviews of partial results, possibly several iterations, and interactions with the data owners [4].

There are different DM process model standards that are used in different research and business data mining projects [4].

- ✓ KDD process (Knowledge Discovery in Databases),
- ✓ SEMMA (Sample Explore Modify Model Assess)
- ✓ CRISP-DM (Cross Industry Standard Process for Data Mining), and
- ✓ KDP (Knowledge Discovery Process)

2.1.1 THE KDD PROCESS

The KDD process is the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required pre-processing, sub sampling, and transformation of the database [1]. The general KDD process is depicted.

Selection: - This stage consists on creating a target Dataset, or focusing on a subset of variables or data samples, on which discovery is to be performed.

Pre-processing: - This is the steps where the target data cleaning and pre-processing performed in order to obtain consistent, complete and better quality of data.

Transformation: - This stage consists of the transformation and discretization of the data using dimensionality reduction and or sampling methods.

Data Mining:- This is the knowledge extraction steps for extracting patterns of interest in a particular representational form, depending on the data mining objective (usually, prediction).

Interpretation/Evaluation: – in this stage the interpretation and evaluation of the mined patterns will be done using different effectiveness measures, such as accuracy. In Figure 2.2. It comprises the following steps [4].

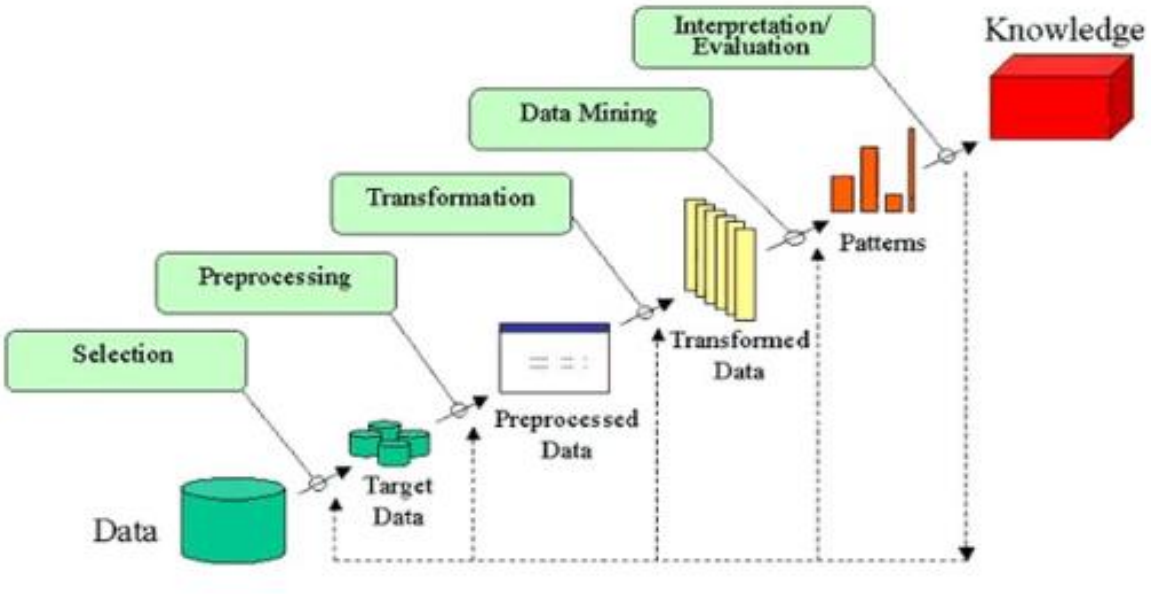


Figure 2.1. The six-steps of KDD model [6].

The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user [31]. Additionally, the KDD process must be preceded by the development of an understanding of the application domain, the relevant prior knowledge and the goals of the end-user. It also must be continued by the knowledge consolidation through incorporating this knowledge into the system [1].

2.1.2 THE SEMMA PROCESS

The SEMMA process was developed by the SAS Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project. The SAS (Statistical Analysis System) institute considers a cycle with 5 stages for the process [33].

Sample: - This stage consists of sampling the data by extracting a portion of a large data set big enough to contain the significant information, yet small enough to manipulate quickly. This stage is pointed out as being optional.

Explore:-This stage consists of the exploration of the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.

Modify: - This stage consists of the modification of the data by creating, selecting, and transforming the variables to focus the model selection process.

Model: - This stage consists of modelling the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.

Assess: - This stage consists of assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs.

Although the SEMMA process is independent from the chosen DM tool, it is linked to the SAS Enterprise Miner software and pretends to guide the user on the implementations of DM applications. SEMMA offers an easy to understand process, allowing an organized and adequate development and maintenance of DM projects. It thus confers a structure for the conception, creation and evolution, helping to present solutions to business problems as well as to find DM business goals.

2.1.3 THE CRISP-DM PROCESS

Analysing the problems of DM & KD projects, a group of prominent enterprises (Teradata, SPSS - ISL, Daimler-Chrysler and OHRA) proposed a reference guide to develop DM & KD projects. This guide is called CRISP-DM (CRoss Industry Standard Process for Data Mining) [23]. CRISP-DM is vendor-independent so it can be used with any DM tool and it can be applied to solve any DM problem. The CRISP-DM methodology is described in terms of a hierarchical process model, comprising four levels of abstraction (from general to specific): phases, generic tasks, specialized tasks, and process instances.

CRISP-DM defines the phases to be carried out in a DM project. CRISP-DM also defines for each phase the tasks and the deliverables for each task.

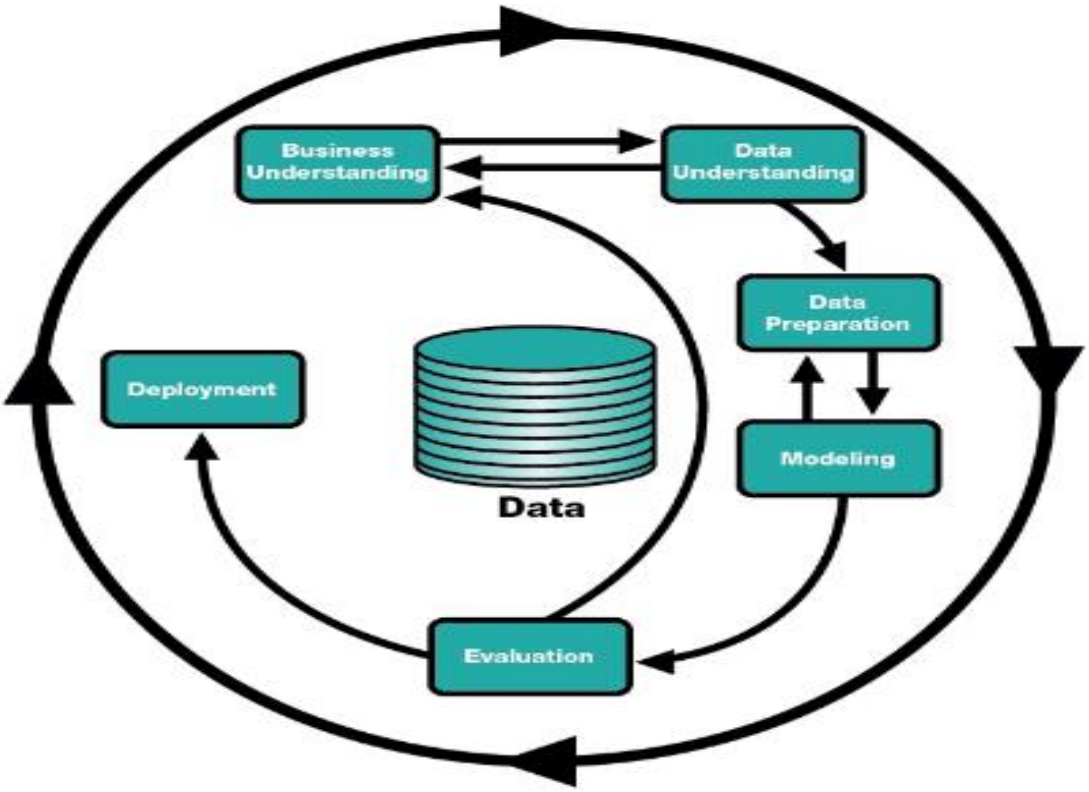


Figure 2.2: Phases of the CRISP-DM [31]

CRISP-DM divides the life cycle of a data mining project in to six phases which are shown in Figure 2.2 [31].The sequence of the phases is not strict. The arrows indicate only the most important and frequent dependencies between phases, but in a particular project, it depends on the outcome of each phase, or which particular task of a phase, has to be performed next.

The outer circle in Figure 2.2 symbolizes the cyclic nature of data mining itself. Data mining is not finished once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more focused business questions. Each phase of CRISP-DM briefly describes as follows [31]:

- **Business Understanding**

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives.

- **Data Understanding**

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. There is a close link between Business Understanding and Data Understanding. The formulation of the data mining problem and the project plan require at least some understanding of the available data.

- **Data preparation**

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.

- **Modeling**

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats. There is a close link between Data Preparation and Modeling. Often, one realizes data problems while modeling or one gets ideas for constructing new data.

- **Evaluation**

At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

- **Deployment**

Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

2.1.4 Knowledge Discovery Process(KDP)

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of both. One such model has six-steps such as understanding of the problem domain, understanding of the data, Preparation of the data, Data mining, Evaluation of the discovered knowledge and Use of the discovered knowledge process (KDP) model developed by [7]. It was developed based on the Cross Industry Standard Process for Data Mining (CRISP-DM) model by adopting/take on it to academic research.

The main differences and extensions of hybrid Model: It provide more general research oriented description of the steps, introducing a data mining step instead of the modelling step, introducing several new explicit feedback mechanisms, (the Cross –Industry Standard Process for Data Mining (CRISP-DM) model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains.

Step.1. Understanding of the Problem Domain

This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology. Description of the problem, including its restrictions/boundaries, is prepared. Finally, project goals are translated into Data mining (DM) goals, and the initial selection of Data mining (DM) tools to be used later in the process is performed.

Step.2. Understanding of the Data.

This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data are checked for completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the Data mining goals.

Step.3. Preparation of the Data

The collected data pre-processed and cleaned in a way to fulfil the requirement of data mining software. In this step of data preparation, tasks like handling missing values, handling outliers' data, transformation of data, and data reduction were taken place. Feature selection and extraction algorithms process handled to acquire cleaned data. The results are data that meet the specific input requirements for Data Mining tools.

Step.4. Data Mining Techniques

Here the data miner uses various classification Data Mining (DM) methods to derive knowledge from the pre-processed data. Among the available algorithms the analysis is performed using WEKA machine learning environment. Among the different available classification algorithms in WEKA decision tree, rule induction, Bayesian and regression algorithm are selected for experimentation. The reason for selecting the above algorithms is its popularities in recently published papers, easy to understand and interpret the result of the model for the studies

Step.5. Evaluation of the Discovered Knowledge

Evaluation includes understanding the results, checking whether the discovered knowledge is novel/new and interesting, interpretation of the results by domain experts. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared. Thus after the development of the model based on the training dataset, the accuracy of the model were tested using test datasets. A confusion matrix and the result of the model were assessed to determine the impact of the discovered knowledge. Using confusion matrix, accuracy, sensitivity, specificity, and precision were calculated to evaluate the performance of each models.

Step.6. Use of the Discovered Knowledge

This final step consists of planning where and how to use the discovered knowledge. Also even if the purpose of the study is academic purpose and the use of model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the Organization can use it. Prototype development is needed to show that the data mining model developed could be deployed. A researcher developed the prototype using VB.NET programming language for selected rules of the model

2.2. Data Mining Tasks

There are two data mining tasks which are known as primary goals of data mining (prediction and description) [34] data mining tasks. Data mining tasks like classification, regression, and

prediction and Time series analysis are categorized under predictive data mining techniques whereas clustering, association, rule discovery, summarization and sequence analysis are categorized under descriptive data mining tasks. Figure 2.4 illustrates the two category of data mining tasks. The description of each tasks presented below.

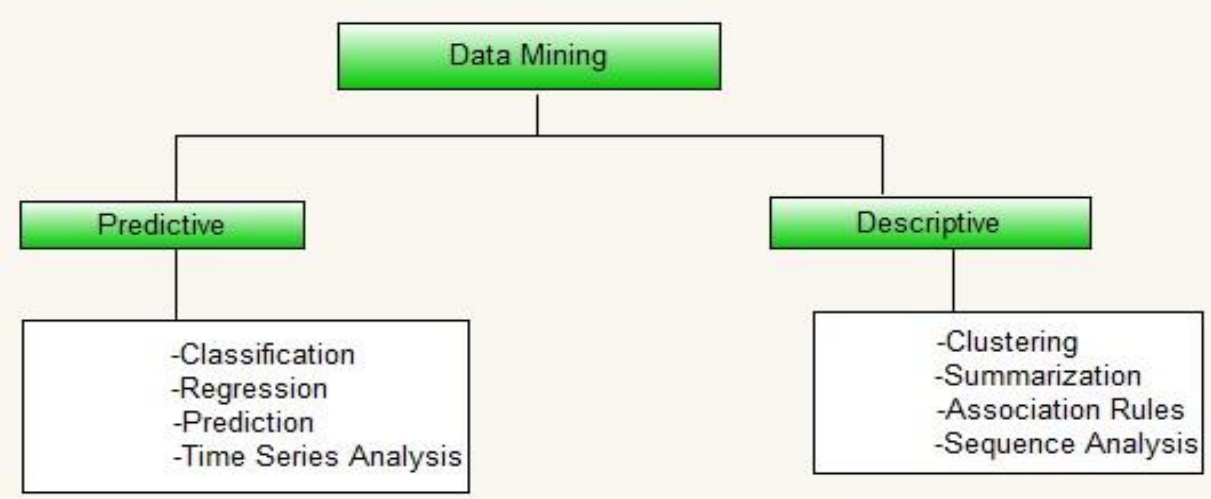


Figure.2.3. Data mining tasks. [34]

Prediction is the objective of data mining which has an aim of predicting unknown or future values of the attributes of interest using other attributes in the databases and it is referred to as supervised learning, since calculated or estimated values are compared with known results Whereas **description** a is data mining task which has an aim of describing the data in a manner understandable and interpretable to humans and its techniques are referred to as unsupervised learning which interrogate the database to identify patterns and relationship in the data. The relative importance of description and prediction depend on the use in different applications [28]. Prediction requires having labels for the output variable for a limited data set, where a label represents some trusted "ground truth" information about the output variable's value in specific cases [5].

2.2.1. Classification

Classification is to classify items into several predefined classes. It is one of the most common tasks in supervised learning, but it has not received much attention in temporal/time based data mining [35]. The classification task is characterized by a well-defined definition of the class

labels, and a training set consisting of pre classified examples. The task is to develop a classifier model of some kind that can be applied to unclassified data in order to classify it. The developed model is based on the analysis of a set of training data whose class label is known and the derived model may be represented in various forms such as IF-THEN rules, decision trees, mathematical formulae, semantic network etc. [36]. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attributes set and the class label of the input data. The techniques are lists as the following [36]. Some popular classification methods include decision tree, logistic regression .In this study, decision tree and rule induction are discussed. A Rule-based classifier extracts a set of rules that show relationships between attributes of the data set and the class label. It uses a set of IF-THEN rules for classification. Rules are easier for humans to understand.

2.2.2. Decision Tree

Decision tree technique uses tree structure to builds regression or classification models. In this technique dataset is divided into smaller subsets and at the same time an associated decision tree is incrementally developed. That result in a tree having decision node and leaf nodes. A decision node is one which has two or more branches. Leaf node represents a decision or classification. The root node known as a best predictor is the top most decision node in a tree. Decision trees handle both numerical data and categorical data [28]. The final result is a tree with leaf nodes and decision nodes where the leaf represents a decision or classification [37].The decision tree algorithm is probably the most popular data mining technique because of the fast training, performance, a high degree of accuracy, and easily understandable patterns. Splitting your data into subsets is the main idea behind the algorithm [38]. When decision tree induction is used for attributes subset selection, a tree is constructed from the given labeled data. All attributes that do not appear in the tree are assumed to be irrelevant. There is a large number of decision-tree induction algorithms described primarily in the machine-learning and applied-statistics literatures that construct decision trees from a set of input-output training samples [29]. In decision tree construction, selection of splitting attributes is necessary in order to avoid irrelevant attributes by examining the effect of each attribute for the distinct class and its likelihood for improving the overall decision performance of the tree, since the feature with minimum impact on dependent variable may distort the tree's performance and the

classification accuracy. One of the most attractive aspects of decision trees lies in their interpretability especially with respect to the construction of decision rules which is constructed from a decision tree simply by traversing any given path from the root node to any leaf [29]. Therefore, to make a decision tree model more readable, a path to each leaf can be transformed into an IF-THEN rule. Figure 2.5 illustrates the root node and leaf node as follows [39].

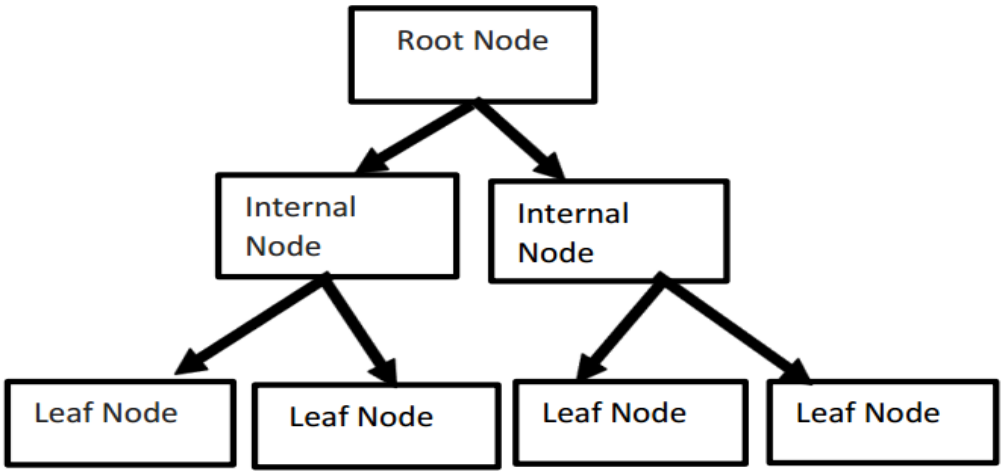


Figure 2.4. Decision tree Structure [39].

The challenge with decision tree is over fitting. As the dataset grows larger and the number of attributes grows larger, we can create trees that become increasingly complex [36]. This potentially leads to the concept of over fitting which consequently brings the notion of pruning; this implies removing of the branches of the classification tree in order to make tree as simple and compact as possible, with as few nodes and leaves as possible. This is done through pruning a tree by halting its construction by partition the subset of training tuples at a given node or removing sub trees from a fully grown tree [36].

The main advantages of decision trees over other algorithms are that they are quick to build, efficient and easy to understand as each node is labelled in terms of the input attributes. The basic algorithm for decision tree induction is greedy algorithm that constructs decision trees in a top-down recursive divide and conquer manner [40]. The algorithm is summarized as follows as stated in [40]

Create a node N;

If samples are all of the same class, C then

Return N as a leaf node labeled with the class C;

If attribute-list is empty then

Return N as a leaf node labeled with the most common class in samples; select test-attribute, the attribute among attributes-list with the highest information gain; label node N with test-attribute; for each known value AI of test-attribute grow a branch from node N for the condition test-attribute= ai;

Let'si be the set of samples for which test-attribute= ai;

If si is empty then attach a leaf labeled with the most common class in samples;

else attach the node returned by Generate_decision_tree (si, attribute-list test-attribute).

Classification Algorithm

Decision Tree J48: C4.5 is an evolution of Dichotomiser ID3, presented by Quinlan J.R [41] for generating a pruned or un-pruned C4.5 tree and all the possible tests are considered during decision making based on information gain value of each attribute [42]. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets.

REP Tree: It Reduces Error Pruning (REP) Tree where Classifier is a fast type of decision tree learner which is built for the decision tree or for the regression tree by using the information gain with entropy and as in C4.5 Algorithm, which deals with the missing values by breaking the corresponding instances into pieces [43].

Rule Based Classification Algorithms

A rule is represented by the IF-THEN form, where the IF part is called the condition and the THEN part is called the action [44].The basic unit and format of knowledge in rule-based reasoning is the rule. The IF-THEN rules are quite natural for humans and are easily understood by both programmers and domain experts. However, accurate description of the domain expert's knowledge in simple rules is often difficult.

IF Condition THEN Conclusion

The rule based classifier is constructed on the concept that IF the information supplied by the user satisfies the conditions of a rule, THEN the actions of the rule are executed [44].For example, one could have the following set of rules to classify the student academic performance. According to [45], knowledge extracted from dataset, IF performance of student in English subject at every level of Grade= Excellent or very good and Maths= Excellent or very good and related subjects, THEN performance = Excellent. The advantage of IF-THEN rule is the rules are order independent i.e. regardless of the order of rules executed, the same classification of the classes is possible to reach [44]. PART and JRIP are algorithms are an example of rule based classifiers.

PART: It is a separate-and-conquer/master rule learner. The algorithm producing sets of rules called decision lists which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in its each iteration and makes the best leaf into a rule. The algorithm is a combination of C4.5 and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) rule learning [46].

JRIP: It implements a propositional rule learner. JRip proposed a Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is an inference and rules--based learner (RIPPER) that can be used to classify elements with propositional rules. The RIPPERR algorithm is a direct method used to extracts the rules directly from the data [46]. JRIP (WEKA'S implementation of the RIPPER rule learner) is a fast algorithm for learning "IF

THEN" rules. Like decision trees rule learning algorithms are popular because the knowledge representation is very easy to interpret.

2.2.3. Regression

Regression, sometimes also called estimation, is a kind of statistical estimation technique which is used to map each data object to a real value provided prediction value [47]. In prediction the aim is to predict a value of a given continuously valued variable based on the values of other variables, assuming either a linear or nonlinear model of dependency [48]. That means the estimation approach has the great advantage that the individual records can be rank ordered according to the estimate. Uses of regression include prediction, modelling of causal relationships, and testing hypotheses about relationships between variables. Well suited techniques for regression tasks are (linear) regression models and none linear regression [49].

2.2.4. Time Series Analysis

In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behavior and the historical time series plot is used to predict future values of the variable.

On the descriptive end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationships in large data sets. Clustering, Summarization, and Visualization of databases are the main applications of descriptive data mining. The usefulness of this concept is that it enables one to generalize the data set from multiple levels of abstraction, which facilitates the examination of the general behaviour of the data, since it is impossible to deduce that from a large database.

2.2.5. Clustering

According to [50] explained, clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. Clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Clustering is commonly used to search for unique groupings within a data set. The distinguishing factor between clustering and classification is that in clustering there are no predefined classes and no examples [51]. The difference with the classification task is that clusters were unknown at the time the algorithm starts. In other words, there are no predefined

classes it relies on [51]. The objects are grouped together based on self-similarity and the applicability can be seen, for instance, in discovering new student behavior pattern.

2.2.6. Summarization

The aim of summarization is to find a brief description for a subset of data [52]. Tabulating the mean and standard deviations for all fields is a simple example of summarization. There are more sophisticated techniques for summarization and they are usually applied to facilitate automated report generation and interactive data analysis (Dependency modelling – to find a model that describes significant dependencies between variables. For example, probabilistic dependency networks use conditional independence to specify the structural level of the model and probabilities or correlation to specify the strengths (quantitative level) of dependencies [28].

2.2.7. Association Rule Discovery

Association Rule is a descriptive data mining task which is one of the relationship mining like as correlation mining, sequential pattern mining, and causal data mining methods. It includes determining patterns, or associations between elements in datasets. Associations are represented in the form of rules. The association technique is used for associating tasks [49]. Examples are market basket analysis, which more or less is determining what things go together in a shopping cart at the supermarket, and cross selling programs, which helps to design attractive packages or groupings of products and services [51]. One of the tasks of data mining is association rule mining. Association rule mining finds interesting association or correlation relationships among a large dataset [44].With a massive amounts of data continuously being collected and stored, many industries became interested in mining association rules from their datasets and the discovery of interesting association among huge amount of business transaction records can help in many business decision making processes [44]. Association rules are in the form of “If antecedent, then consequent,” together with a measure of the support and confidence associated with the rule.

2.2.8. Sequence Analysis

Sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend. Given is a set of objects, with each object associated with its own

timeline of events, find rules that predict strong sequential dependencies among different events.[44]

2.3. Predictive Data Mining

Data mining is the exploration of historical data (usually large in size) in search of a consistent pattern and/or a systematic relationship between variables; it is then used to validate the findings by applying the detected patterns to new subsets of data [37]. The roots of data mining originate in three areas: classical statistics, artificial intelligence (AI) and machine learning [38]. Pregibon [21], described data mining as a blend of statistics, artificial intelligence, and database research, and noted that it was not a field of interest to many until recently.

According to [35] data mining can be divided into two tasks: predictive tasks and descriptive tasks. The aim of data mining is prediction; therefore, predictive data mining is the most common type of data mining and is the one that has the most application to businesses or life concerns. Prediction task predicts the possible values of missing or future data. Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest. Predictive data mining has three stages, they are: data pre-processing, prediction and deployment.

The data mining process starts with the collection and storage of data in the data warehouse. Data collection and warehousing is a whole topic of its own, consisting of identifying relevant features in a business and setting a storage file to document them. It also involves cleaning and securing the data to avoid its corruption. According to Kimball, a data warehouse is a copy of transactional or non-transactional data specifically structured for querying, analyzing, and reporting [9]. Data exploration, which follows, may include the preliminary analysis done to data to get it prepared for mining. The next step involves feature selection and or reduction. Mining or model building for prediction is the third main stage, and finally come the data post-processing, interpretation, and/or deployment.

2.4. Attribute Selection Measures

There are different kinds of attribute selection algorithm in data mining of which the most popular attribute selections measures are Information Gain, Gain Ratio, and Gini Index.

Attribute selection sometimes called as Feature selection is one of the major tasks in decision tree because it will enhance the accuracy and it will help us for selecting a subset of relevant features for building robust learning models. The researcher decides to use the Information gain measures.

2.4.1. Information Gain

This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or "information content" of messages [3]. Let node N represents or hold the tuples of partition D. The attribute with the highest information gain is chosen as the splitting attribute for node N. This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions.

Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found. The expected information needed to classify a tuple in D is given by [3]

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

Where p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_i,D|/ |D|$. A log function to the base 2 is used, because the information is encoded in bits. Info(D) is just the average amount of information needed to identify the class label of a tuple in D. Note that, at this point, the information we have is based solely on the proportions of tuples of each class. Info(D) is also known as the entropy of D.

Now, suppose we were to partition the tuples in D on some attribute A having v distinct values, $\{a_1, a_2 \ldots a_v\}$ as observed from the training data. If A is discrete-valued, these values correspond directly to the v outcomes of a test on A. Attribute A can be used to split D into v partitions or subsets, $\{D_1, D_2, : : : , D_v\}$,where Dj contains those tuples in D that have outcome aj of A. These partitions would correspond to the branches grown from node N. However, it is

quite likely that the partitions will be impure (e.g., where a partition may contain a collection of tuples from different classes rather than from a single class). After partitioning in order to arrive to the exact classification the amount of information needed is measured by [3]

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the j^{th} partition. $Info_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information (still) required, the greater the purity of the partitions.

Information gain is defined as the difference between the original information requirements (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is [3],

$$Gain(A) = Info(D) - Info_A(D).$$

In other words, Gain(A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest information gain, (Gain(A)), is chosen as the splitting attribute at node N. This is equivalent to saying that we want to partition on the attribute A that would do the "best classification," so that the amount of information still required to finish classifying the tuples is minimal (i.e., minimum InfoA(D)).

The researcher tried to rank the attribute based on information gain, ranking the attributes to the mining task of the decision tree was implemented by *weka.attributeSelection.InfoGainAttributeEval* which ranks Attribute using search methods of Ranker.

2.5. Implementation Tools

In order to mine hidden knowledge from the pre-processed dataset and compare the performance of classifiers, WEKA 3.8.4 version is used. WEKA is chosen since it is proven to be powerful for data mining and used by many researchers for mining task and the researchers is familiar with the tool. It contains tools for data pre-processing, clustering, regression, classification, association rules and visualization. WEKA is written in the Java language and contains a Graphical User Interface (GUI) for interacting with data files and producing visual results.

2.6. Model Evaluation

The standard classifier performance measures such as Prediction Accuracy, True Positive, False Positive, Precision, Recall and F-Measure are commonly used [54]. Confusion matrix helps to see a breakdown of a classifier‘s performance by showing how frequently instances of a class let us say class X are classified as class X or misclassified as some other class, say class Y [55]. According to Weiss and Zhang [56], a model performance evaluation should answer questions such as: How accurate is the model? How well does the model describe the observed data? How much confidence can be placed in the model‘s predictions? How understandable is the model?

The application’s results help to interpret the generated models in relation to the research question posed in the first chapter. The classification algorithm predicts the class label. The final output will be patterns which are used to find out the performance of students. Some of the performance measures are given below in Table 2.1 Confusion Matrix .

2.6.1. Confusion Matrix

Confusion matrix can show the selected model prediction performance by comparing it to the actual value. In evaluating the performance of a model, a confusion matrix or correct classification matrix can be used. Confusion matrix focuses on the predictive capability of a model rather than how fast takes to classify or build models, scalability, etc. [57]. In the class case prediction, the result is often displayed as a two dimensional confusion matrix with a row and column for each class.

Table 2.1. Two dimensional confusion matrix

	Predicted class		
Actual class	Class (+)	Class (+)	Class=(-)
		True Positive(TP)	**False** Negative(FN)
	Class=(-)	**False** Positive(FP)	**True** Negative(TN)

The two-class case with classes yes and no, a single prediction has the four different possible outcomes shown in Table 2.1. The true positives (TP) and true negatives (TN) are correct classifications. A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive. The four outcomes of classifier and meanings are figuratively as follows.

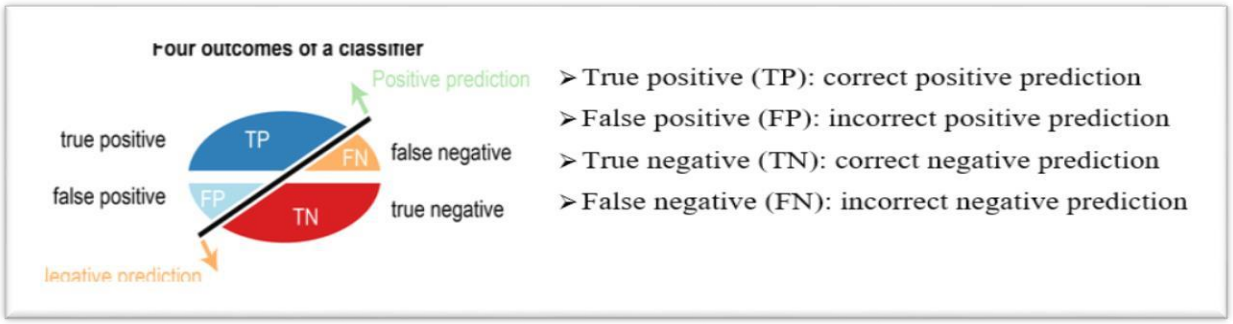


Figure 2.5.Four Outcomes of Classifier

2.6.2. True Positive Rate (TPR)

True Positive rate (TPR) is the proportion of positive or correctly classified instances as positive or correct instances. It is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true Specificity (SP). The best True Positive rate is 1.0, whereas the worst is 0.0.

$$TPR = \frac{TP}{TP+FN} \quad (2.1)$$

2.6.3. False Positive Rate (FPR)

The False Positive (FPR) rate is measures the proportion of negative instances that are erroneously classified as positive. It is calculated as the number of incorrect positive predictions divided by the total number of negatives. The best false positive rate is 0.0 whereas the worst is

1.0. It can also be calculated as 1 – specificity.

$$FPR = \frac{FP}{TN+FP} \quad \text{.....(2.2)}$$

2.6.4. Precision

Another metrics for performance evaluation of the classifier is precision which measures what percent of tuples that the classifier labelled as positive are actually positive. The best precision is 1.0, whereas the worst is 0.0.

$$Precision = \frac{TP}{TP+FP} \quad \text{.....(2.3)}$$

2.6.5. Recall

Recall is also another performance evaluation which measures what percent of positive tuples the classifier labelled as positive for both True and False classes. It is what percent of positive/negative tuples the classifier labeled as positive or negative for both True and False Classes.

They are summarized in the following formulas [55].The formula is:

$$Recall = \frac{TP}{TP+FN} \quad \text{.....(2.4)}$$

2.6.6. F-Measure

The final metric used for performance evaluation of classifiers on confusion matrix is F-measure. The F- measure is the inverse relationship between precision & recall, and calculated as the harmonic mean of recall and precision. It is the point to conclude that the precision and

recall of the model are significantly balanced [55].F-Measure: is calculated as the harmonic mean of recall and precision.

$$\textbf{F-Measure (F)} = \frac{2*recall*percision}{recall+Percision} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.5)$$

$$TNR = \frac{TN}{FP+TN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.6)$$

Specificity is calculated as the number of correct negative predictions (TN) divided by the total number of negatives (N). Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR).

The best specificity is 1.0, whereas the worst is 0.0.

2.6.7. Predictive Accuracy

$$\text{Predictive Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.7)$$

Among the standard performance metrics, accuracy is the most widely-used metric to check the performance of the model. The correctness accuracy for a data mining classifier is defined as the degree of closeness of its prediction to the actual values, either true or false [55].The accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives instance by the total number of samples. The accuracy [28] of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Prediction accuracy measures the proportion of instances that are correctly classified by the classifier. Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0.

2.7. 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 [13]. 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 [14].

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 trainee outcomes [14]. 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 [13].

2.8. Related work

Several studies addressed the analysis of educational data to get useful information that affects learning quality. Muluken [67] has developed predictive model that predicts Student Success and Failure by using two classification algorithms namely J48 Classifier and Naïve Bayes. The proposed Models Compared to discover the best model for predicting student performance SAP to identify students at risk. Only one independent parameter used in their study, which is the midyear mark in order to predict the students of Computer Science final marks. The classification rule generation process is based on the decision tree and Bayes as a classification

technique from Sample dataset of 11,873 instances. The J48 classifier performed better with 92.34% of the tuples being predicted correctly compared to accuracy for Naïve Bayes. CRISP-DM (Cross Industry Standard Process for Data mining) is a data mining methodology has used by the research. Analysis is done by using WEKA3.7 application software on attributes such as Sex, age, region, Higher Education Entrance Certificate Examination result, field of study, College, Number of courses given in a semester, Total credit hours given in a semester, Number of students in a class, semester Cumulative GPA of a student. Students' performance were categorized in five groups: Very good, with a high probability of succeeding; Good students, who are above average and with a little more effort can succeed with good Grades ; Satisfactory students, who may succeed; Below Satisfactory students, who require more efforts to succeed; and Fail, who have a high probability of dropping out. In conclusion, there are many data mining technique but the study uses only classification model and very few scenarios for experimentation is considered as the limitation of the study.

[12] Developed a model that enables the higher learning institutions (HLI) to predict the trainee's dropout and analyze the causative factors that lead to the dropout of such trainees. In this situation, the classification method of data mining will be used to predict useful information through the historical information by data mining tools. The Authors analyzed data by using three decision tree algorithms: Iterative Dichotomiser3 (ID3), J48 and classification and regression (CART) decision tree. WEKA is used by the authors of this research to analyze data and to build a trainee's dropout model. The authors in [12] were taking 797 trainees as samples from different department of Kavumu TVET School. The data attributes include personal gender, district, and sponsorship and tuition fees) and it was collected by using Kavumu TVET school reports of 2019 and real questionnaires. The result indicated that ID3 is the best algorithm with an accuracy of 80 % compared to J48 with accuracy of 79.5 % and CART with 78.5%. And finally the authors concluded that few trainees are more likely to drop out due to home sickness, illness, peer problems, high school fees, and adjustment problems.

[8] Conducted a significant data mining research on the student performance by using 300 students' data from 5 different degree colleges using the Naïve Bayes classification method, on a group of BCA (Bachelor of Computer Applications) students who appeared for the final

examination in the year 2010. A questionnaire was distributed for collecting data from each student before the final examination, which had multiple personal, social, and psychological questions that was used in the study to identify relations between these factors and the student‘s performance and grades. They found that the most influencing factor for student‘s performance is their grade in senior secondary school, which further tells us, that those students, who performed well in their secondary school, were definitely perform well in their Bachelors study. Furthermore, it was found that the living location, medium of teaching, mother‘s qualification, student other habits, family annual income, and student family status, all of which, highly contribute in the students educational performance, thus, it can predict a student‘s grade or generally their performance if basic personal and social knowledge was collected. Limitation of this study is the experiment was conducted with a less rich dataset; the total dataset used in this experiment was only 300.

In another research, authors in [26] discuss association rule discovery techniques improve the quality of education by analyzing the data and discover the factors that affect the academic results. Apriori algorithm is used which extracts the set of rules, specific to each class and analyzes the given data to classify the student based on their performance in academics. The researchers uses different reports of Term Tests, Attendance, and University Results That contains details of students with 5 listed attributes which include Rollno, Term Test1 marks, TermTest2 marks, Attendance, University Results. The data contains numeric value. The research analyzed the potential use of one of the data mining technique called association rule mining in enhancing the quality of students performances. The extracted rules helps to predict the performance of the students and it identify poor, good and excellent students. The Research helps to identify those students which need special attention to reduce fail ratio and taking appropriate action for the next semester examination.

CHAPTER THREE

MATERIALS and METHODS

Methodology is a way that deals with data collection, analysis and interpretation in order to help the investigators achieve the objective of the research. Hence, the following methods and processes followed in this research work.

3. RESEARCH DESIGN

This study follows experimental research. Experimental research designs are selected because the primary approach used to investigate causal (cause/effect) relationships and to study the relationship between one variable and another. Researchers use experimental research to compare two or more groups on one or more measures [1].

To conduct an extensive experiment, the study uses KDP (Knowledge Discovery process) model. This process model is selected because it provides more general, research-oriented description of the steps, introducing a data mining step instead of the modelling step, introducing several new explicit feedback mechanisms. The model has six steps (see figure 3.1); understanding of the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and finally use of the discovered knowledge.

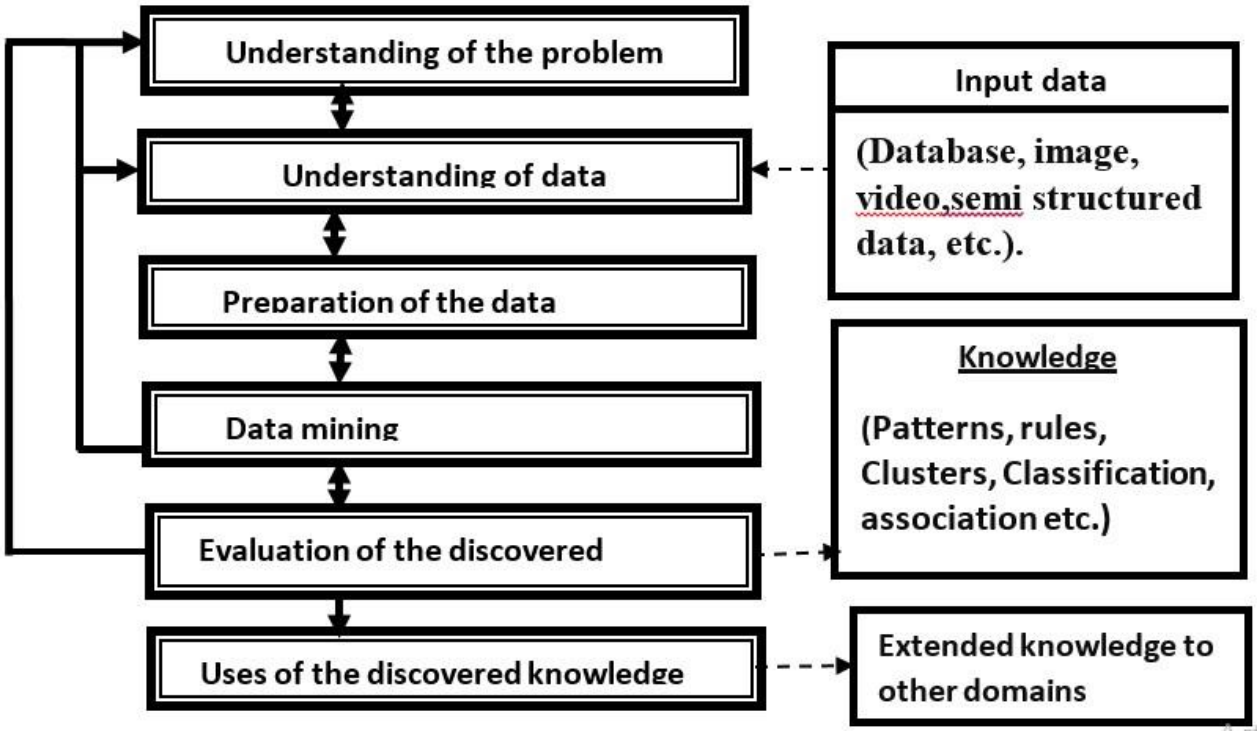


Figure 3.1: The six steps of Hybrid data mining model (KDP) [7]

Therefore, the overall research design was to build a model that can be used to predict the trainee's performance on Hawassa poly Technique College data set. One of the most important aspects of this model is iterative and interactive feature. The feedback loops are necessary because any changes and decisions made in one of the steps can result in changes in following steps. The researcher tries to discuss tasks done and methods used at each step in detail below.

3.1. Understanding of the Problem Domain

In Ethiopia 8-2-2 formal education structure has been adopted [10]. Primary school has an official entry age of around 7 and duration of 8 grades. Secondary school is divided in to two cycles: Lower secondary consists of grade 9-10 and upper secondary consists of grade 11-12. In principles, primary education school is free and compulsory. Students sit for the primary school Certificate Examination at the end of grade 8, the General secondary education certificate examination (EGSECE) at the end of grade 10. And the Higher education Entrance certificate

Examination (EHEECE) at the end of grade 12. [10] In this case Grade 10 attained mostly at the age of 17 and Grade 12 attained mostly at the age of 19.Therefore, the researcher decides to take TVET trainees at age 17 and above. The bellow graph figure 3.2 illustrate the above Ethiopian educational cycle.

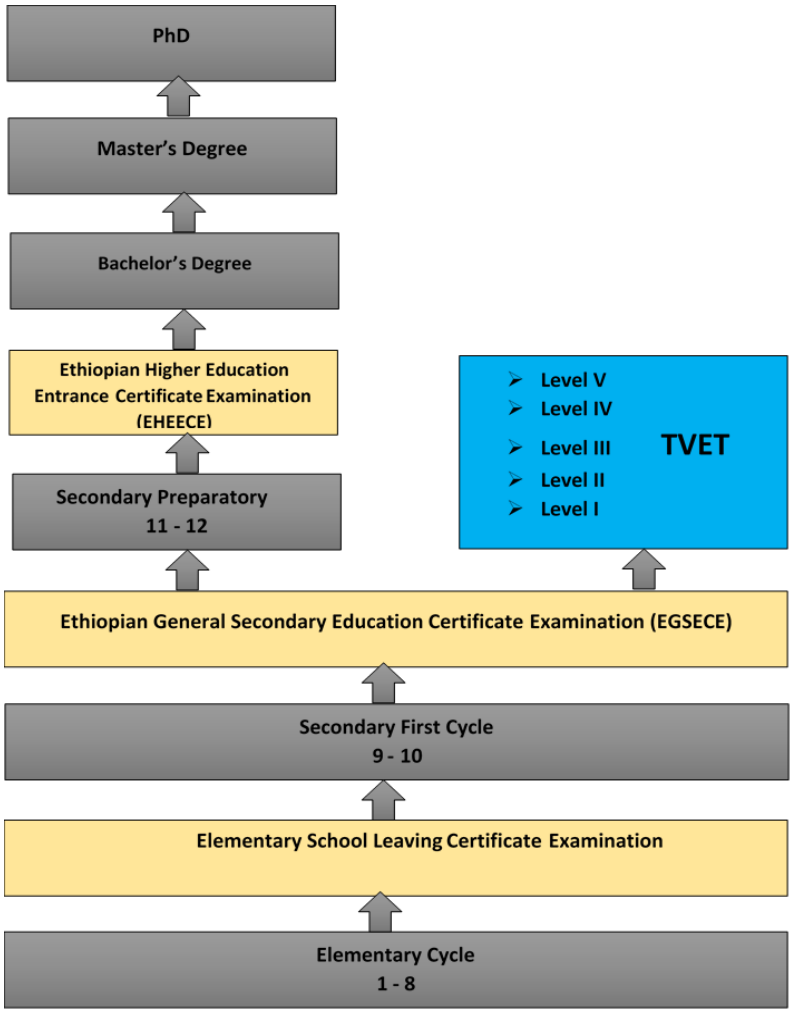


Figure 3.2: Ethiopian Educational Cycle

This initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem. It also involves learning domain-specific terminology [2].

In this research discussion was made with Registrar officer, Vice Dean and COC coordinator from now on wards we call them as domain experts and additionally tries to refer relevant documents such as journal articles, conference papers and the internet were also reviewed to understand the

current problem domain.

The business objective of Technical and Vocational Education and Training (TVET) in Ethiopia seeks to create competent and self-reliant citizens to contribute to the economic and social development of the country, thus improving the livelihoods of all Ethiopians and sustainably reducing p

overty. Prediction helps to identify trainees with weak performance and help them to score better marks for improving their performance.

Domain experts were consulted; direct observation of the College existing system was done and reviewing of the different literature were performed to have brief understanding on the factor affecting trainees academic performance. Investigation on issues was studied to see a gap where the data mining can be used to fill with the use of Data Mining techniques.

3.2. Understanding of the Data

This step includes collecting sample data and deciding which data, including format and size, will be needed. Background knowledge can be used to guide these efforts. Data was visualized using Microsoft excel format to check completeness, redundancy, missing values, plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the DM goals. This step need for additional or more specific information about the data in order to guide the choice of specific data pre-processing algorithms [2].

Data understanding is next to domain understanding in hybrid data model. This step includes collecting sample data and deciding which data will be needed including data format and size. Background knowledge can be used to guide these efforts and data are checked for completeness, redundancy, missing values, acceptability of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the Data mining (DM) goals. Thus, five years from 2009 to 2013 initial dataset is collected in excel 2007 and 2010 format from Hawassa Poly Technique College registrar office.

3.3. Preparation of the Data

This step is concerned with deciding which data are used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data may be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (say, by discretization), and by summarization of data (data granularization) [2]. The end results are data that meet the specific input requirements for the DM tools selected in Step 3.4.

The major tasks undertaken in this phase include: description of data sources, carrying out statistical summary measure, filling missing with mode values, and data transformation/reduction activities. Additionally, converting data from one format to another format was also made.

3.3.1. Collect Initial Data

The initial data for this study were collected after getting permission from the administrations of Hawassa Poly Technic College. The data collected was about TVET trainees. The table was saved to MS Excel file.

The study used data collected from Hawassa College trainee's database that covers from 2009-20013 E.C. To understand the data, discussions were made with domain experts from Hawassa Poly Technic College registrar office.

3.3.2. Description of the Collected Data

The decision on the data that was used for the analysis is based on several criteria, including its relevance to the data mining goals. The attributes are selected based on understanding the problem domain with the help of domain expert and literature reviews. Table3.1 presents the description of the collected attributes with their data type. The final selected data was prepared and pre-processed before developing the model.

Table3.1: Description of the selected attributes from Hawassa Poly Technic College dataset

No.	Attribute Name	Data Type	Description	Value
1	Gender	Nominal	Sex of the trainees	Male= M, Female= F
2	Division	Nominal	Trainees study program	Extension and Regular
3	High school	Nominal	Previously attended High School	Private or Government
4	Sector	Nominal	Trainees sector	Economy, Industry ...etc
5	Occupation	Nominal	Trainees department	ICT, ACCOUNTING, ETC
6	Level	Nominal	Contains Occupational trainee's Level	Level I, II, III and level IV
7	Woreda	Nominal	Name of WOREDA	
8	Entry year	Date	Trainees entry year	
9	EGSECE Result	Numeric	Result in EGSECE	
10	Transcript	Numeric	Average result of grade 9 and 10	(50-100)
11	English	Nominal	EGSECE Result trainees Get in English	A,B,C,D and F

12	Mathematics	Nominal	EGSECE Result trainees Get in Mathematics	A,B,C,D and F
13	Status *(class attribute)*	Nomonal	Describes the final Assessment performance	Competent (C) or Not Yet Competent (NYC)

This data set was organized in rows and columns where each column represents an attribute and each row stands for a single record of an individual. The data set that was taken from Hawassa Poly Technic College registrar has a total of 13 attributes (columns) and 8200 records (rows) identified;

This study derived one attributes from the existing original data set as shown in table 3.2

- **Age**: Age attribute was derived from date of birth found on Trainee profile. We derived this attribute because it helps us to determine trainee's status with respect to their age and convert into age category to simplify analysis.

Table 3.2: Additional attributes with their description

No	Attribute Name	Data Type	Description	Data Values
1	Age(derived)	Numeric	Age of the trainees	[15– 50]

3.3.3. Data Cleaning

Witten and Frank[4],described Data cleaning as a time consuming and labour intensive procedure but one that is absolutely necessary for successful data mining. Data cleaning routines work to cleaning the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Quality of data plays an important role in information oriented organizations, where the knowledge is extracted from data. Consistency, completeness, accuracy, validity and timeliness are the important characteristics of quality data. So, it's important to obtain quality data for knowledge extraction. Data cleaning is an important step in KDP process in order to recognize any inconsistency and incompleteness in the data set and to improve the quality [5]. Under data cleaning there are many activities are done.

In this study, the data set of Hawassa Poly Technique college Contains 164 missed values of age Attribute and it comprises (2%) of instances from total data set. To handle the problem of missing values for nominal variables, replacing with modal value and for the numeric type attribute the missing values were replaced by the Mean value which is recommended by many scholars and it is also cited in [15].

As it has been described in table 3.1 the attributes description table, the Gender Attribute is nominal type. This attribute has 2 valid attribute values which mean the distinct value is 2. These are M (Male) and F (Female). The sex attribute which has a missed value of 246 (3%) of instances from total data set. The frequency of the attribute M (Male) is 4868 and F (Female) is 3332. Then as you can see modal value for this attribute is M (Male). Different literatures recommend that the missing vales for the nominal type attribute shall be replaced by the modal value [15]. But here it is difficult to insert 246 3% of the record in the data set it may change the true nature of the data set. Then the researcher decided to handle the missing values manually by looking trainees name even though it takes time and tedious.

3.3.4. Data Transformation

Data transformation is the process of discretizing continuous value. In this page, the continuous

valued age, Ethiopian higher education entrance examination English/Mathematics result and transcript result were discretized as shown in table 3.3, 3.4 and table 3.5 respectively.

In consultation with domain experts the age ranges of trainees joined the TVET College are divided into 4 intervals. After completing the discretization process distinct values of the age attribute were reduced to 4 from 28 distinct values.

Table 3.3: Discretized age attribute

Age	**Represented value**
17 –26	Early Young
27 –34	Young
35 –42	Early Adult
43 –50	Adult

The same was done with status score to convert into the stated success, average and weak.

Table 3.4: A discretized **transcript result attribute**

Grade9&10 Transcript Result	**Represented value**
[90 –100]	Excellent
[80 - 90]	Very good
[60 –80]	Good
[50-60]	Satisfactory

3.3.5. Data Preparation for WEKA Software

The raw data initially stored in Microsoft Excel 2007 & 2010 format which is not understandable by the WEKA tool. To make such data format understandable by WEKA saved the exported Microsoft Excel file is saved as a CSV (Comma Separate Values) file format. The last task done to make the data format suitable for WEKA tool was converting the CSV file into ARFF (Attribute Relation File Format) file format.

3.4. Data mining

One of the major tasks of data mining research is modelling; here the data miner uses various DM methods to derive knowledge from pre-processed data.

In this phase, various data mining techniques were selected and applied and their parameters calibrated to optimal values. Typically, there are several techniques and tool for the same DM problem type. In this study **WEKA version 3.8.4**(Waikato Environment for Knowledge Analysis) is selected for DM. It is free software available under the GNU (General Public License). The WEKA work bench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. It is an open source; powerful tool for data mining algorithm will be used for this study. We experimented different DM classification algorithms such as J48, Naïve Bayes, JRiP and PART. These algorithms are selected because they are usually used for educational data mining and understandable rules can be extracted.

3.5. Evaluation of the discovered knowledge

Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models are retained, and the entire process is revisited to identify which alternative actions could have been taken to improve the results. At this stage, the models that were generated after the application of DM tools and techniques on data are evaluated in terms of achieving the initially stated aims and objectives. Models generated are measured against each other to select the better performing one. The evaluation is done using confusion matrix by comparing accuracy, recall, precision, ROC area, and f-measure.

3.6. Use of the discovered knowledge

This final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. A plan to monitor the implementation of the discovered knowledge is created and the entire project documented.

Finally, the discovered knowledge is deployed.

In this research the discovered knowledge is used by integrating the user interface which is designed by VB.NET programming language with a WEKA system in order to show the prediction of TVET trainee's performance. VB.NET is selected because it is easy to build applications and its familiarity.VB.NET supports object-oriented programming approach which suits to develop a desktop application.

CHAPTER FOUR

EXPERMENTAL ANALYSIS AND RESULT

4.1. Introduction

This chapter discuss about the experimentation phase applied on the pre-processed dataset using with the selected classification techniques. The chapter has consisting of experimental setup, experimental analysis, experimental result, and comparison of the result. The experimental part describes how the pre-processed dataset is partitioned for training and testing purpose. The researcher decides to use 10-fold cross validation, 80% and 66% split test parameters for all selected algorithm and finally the model with the best performance is selected.

4.2. Balancing Instances

According to[68], a dataset is imbalanced if the classification categories are not approximately equally represented. Performance of DM algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the cost difference of error is large.

Under-sampling of the majority (normal) class has been proposed as a good means of increasing the sensitivity of a classifier to the minority class. However, Chawla[68], showed a method of oversampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance (in ROC space) receiving operation characteristics, than only under sampling the majority class. The Trainees performance Data of Hawassa Poly Technique College has a higher imbalance for the class variable (not yet competent). Therefore, the researcher used re-sampling of Weka (Weka. filters. supervised. Instance. Resample) to over sample the minority classes (not yet competent) and under sample the majority (competent). As a result, the class distribution in the dataset changes and probability of correctly classifying the instances of the class increases.

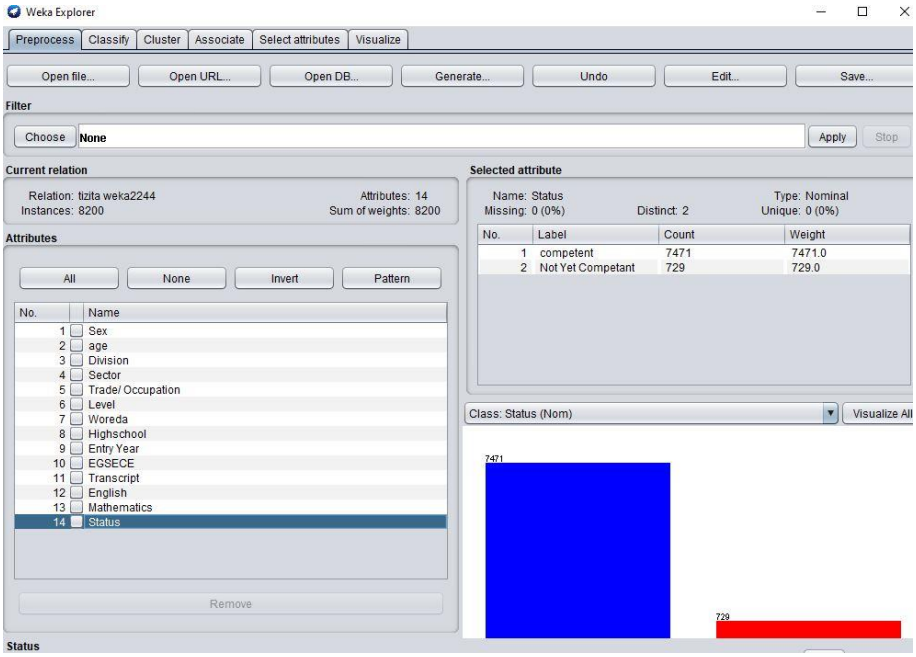


Figure 4.1.WEKA Snapshot Side by side view of the class attribute creditor's Original data

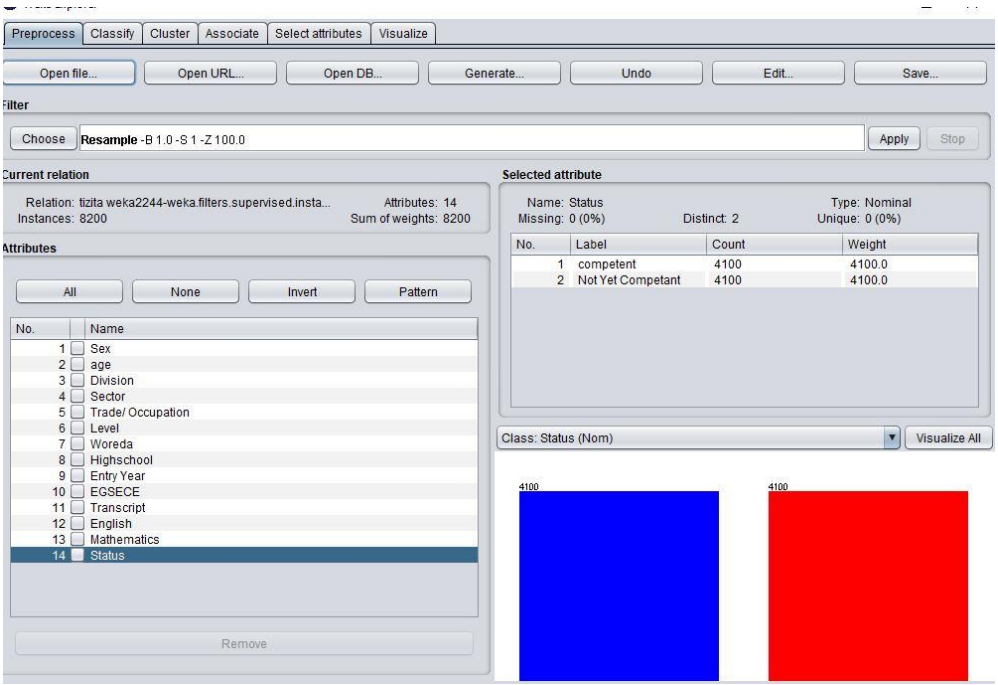


Figure 4.2 Side by side view of the class attribute creditor's Resample (balanced) data

Figure 4.1 shows a side by side view of the class attribute is status of Competent the majority class. Originally there were 7471 records in the majority class and only 729 records the minority class. Figure 4.2 shows after resampling, which is over sampling the minority which has moved the instances from 7471 to 4100 and under sampling the majority diminishes from 729 to 4100.

After passing the above pre-processing stage the researcher has got 8200 records. Then the pre-processed dataset in excel is converted to Comma Separated Values (.csv) to make it compatible with WEKA software.

4.3. Attribute Selection

The attribute selection process is performed using the ranker selection attribute values by using info gain ratio attribute selection measure to select the most relevant attributes, whole attributes Out of 14 ranked attributes using info gain ratio, and all 13 attributes were selected based on relevance value by assuming 1attribute as class attribute. The following Figure 4.3 depicts the 13 attribute selection ranker using info gain ratio from largest to lowest.

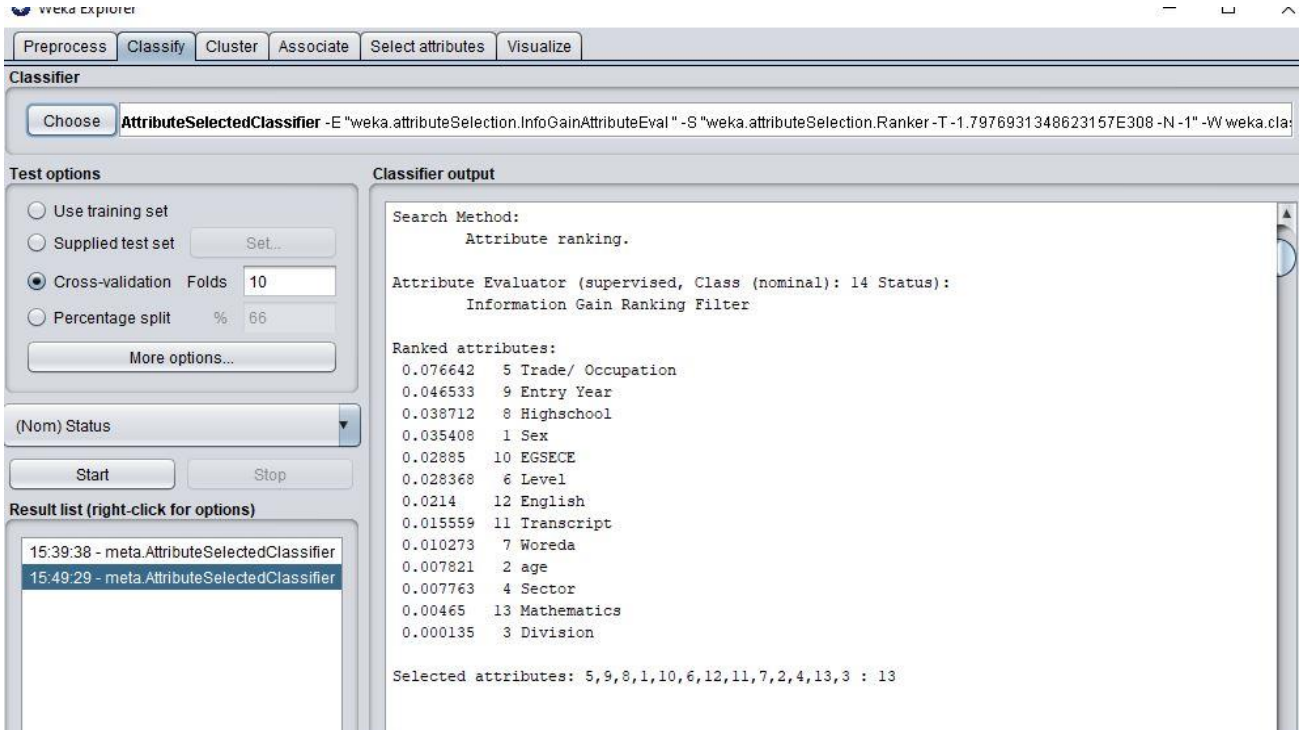


Figure 4.3. WEKA Snapshot of attribute selection using infogain ranker.

Next, the researcher decides to perform all experiments based on the above selected attributes

4.4 Experimental Setup

In any data mining research before developing a model, we should generate a mechanism to test the model performance. For instance, in the supervised data mining task, such as classification, it is common to use classification accuracy measure, True Positive rate (TP), precision, recall and F-measure of the experts are used as to measure the performance of the developed data mining model. Each algorithm was conducted in different ways using attributes (whole and selected attributes) based on infogain ratio values in WEKA tool with balanced dataset. Thus, the whole attributes and selected attributes have experimented with the four classifies (J48, Naïve Bayes, JRiP and PART) in the manner shown in Table 4.1.

Table 4.1. Experimentation ways of whole attributes and selected attributes

Algorithm	**Test Options**	**Parameter**	**Name of experiment**
J48 Pruned	10 Cross Validation Test	Whole Features	Experiment #1
	80/20 Split Test Mode	Whole Features	Experiment #2
	66/34 Split Test Mode	Whole Features	Experiment #3
J48 Un-Pruned	10 Cross Validation Test	Whole Features	Experiment #4
	80/20 Split Test Mode	Whole Features	Experiment #5
	66/34 Split Test Mode	Whole Features	Experiment #6
JRIP Pruned	10 Cross Validation Test	Whole Features	Experiment #7
	80/20 Split Test Mode	Whole Features	Experiment #8
	66/34 Split Test Mode	Whole Features	Experiment #9
JRIP	10 Cross Validation Test	Whole Features	Experiment #10

Algorithm	Test Options	Parameter	Name of experiment
Un-Pruned	80/20 Split Test Mode	Whole Features	Experiment #11
	66/34 Split Test Mode	Whole Features	Experiment #12
Naïvebais	10 Cross Validation Test	Whole Features	Experiment #13
	80/20 Split Test Mode	Whole Features	Experiment #14
	66/34 Split Test Mode	Whole Features	Experiment #15
PART Pruned	10 Cross Validation Test	Whole Features	Experiment #16
	80/20 Split Test Mode	Whole Features	Experiment #17
	66/34 Split Test Mode	Whole Features	Experiment #18
PART Un-Pruned	10 Cross Validation Test	Whole Features	Experiment #19
	80/20 Split Test Mode	Whole Features	Experiment #20
	66/34 Split Test Mode	Whole Features	Experiment #21

4.5 Experimental Results

In order to come up with the best classification model, both Pruned and Unpruned methods are implemented. These pruning mechanisms are labeled in Weka as *ConfidenceFactor* (CF) and *MinNumObj* (MNO) respectively. In this study these two parameters shown on Table 4.2 are tuned for different values including the default to optimize the classification model.

Table 4.2: Description of parameters to be tuned in all classification modeling

Parameter	Description	Default Value
ConfidenceFactor (CF)	The confidence factor used for pruning (smaller values incur more pruning)	0.25
MinNumObj (MNO)	The minimum number of instances per leaf	2

The following parameters are used throughout the experiments [2],

Accuracy is the percentage of correct predictions. According to confusion matrix, it can be calculated as

$$AC = \frac{TN+TP}{TP+FP+TN+FN} \quad \text{..... (4.1)}$$

Where

TN: is the true negative, i.e., instances that are correctly classified as negative.

TP: is the true positive, i.e., instances that are correctly classified as positive

FP: is the false positive, i.e., instances that are predicted to be positive but should have been classified as negative.

FN: false negative, i.e., instances that are predicted to be negative but should have been classified as positive.

Precision is the ratio of correct prediction to the sum of true and wrong prediction. It can be calculated as,

$$Precision = \frac{TP}{TP+FP} \quad \text{..... (4.2)}$$

Recall is the ratio of correct prediction to the sum of true prediction and false negative prediction.

It is calculated can be calculated as,

$$Recall = \frac{TP}{TP+FN} \quad \text{..... (4.3)}$$

F-Measure can be calculated as follows,

$$F - Measure = \frac{2(Precision * Recall)}{Precision + Recall} \quad \text{..... (4.4)}$$

Totally, 21 Experimentation have been performed on the pre-processed dataset. Each of this experiments and the result explained below.

4.5.1 MODEL BUILDING USING J48 DECISION TREE

Decision tree is known as state of the art in data mining. The reason this classifier is chosen is that it has a very good performance for large data size analysis [3], the result can be easily interpreted and they can be applied to categorical data and also most researches have chosen it as the best classifier when it is compared with other algorithms [4].

This experiment is conducted in the J48 decision tree using the 10-fold cross validation and the percentage split classification test option in Weka, with default Confidence Factor (CF) and MinNumObj (MNO) values including the default parameters shown in Table 4.5. The default value of percentage split, which is 66% for training and 34% for testing is applied. A data set of fourteen selected attributes and 8200 instances has been used in this experiment.

EXPERIMENT #1 J48 pruned decision tree with 10-fold cross validation test mode. By setting the value of CF=0.25 default and MNO=2 (Default), and using 10-fold cross validation test option in WEKA, the snapshot of this experiment is shown in **Figure 4.4.**

```
Time taken to build model: 0.33 seconds

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

Correctly Classified Instances        7654               93.3415 %
Incorrectly Classified Instances       546                6.6585 %
Kappa statistic                          0.8668
Mean absolute error                      0.0904
Root mean squared error                  0.2432
Relative absolute error                 18.0748 %
Root relative squared error             48.6405 %
Total Number of Instances            8200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.891    0.024    0.974      0.891    0.930      0.870    0.960     0.968     competent
                 0.976    0.109    0.900      0.976    0.936      0.870    0.960     0.936     Not Yet Comp
Weighted Avg.    0.933    0.067    0.937      0.933    0.933      0.870    0.960     0.952

=== Confusion Matrix ===

   a    b   <-- classified as
3653  447 |    a = competent
  99 4001 |    b = Not Yet Competant
```

Figure 4.4 WEKA Snapshot of 10-Fold cross Validation

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (3653+4001)/ (3653+99+4001+447) = 0.933414634146=**93.3415%**

Therefore, the overall accuracy of the model was measured at 93.3415 %.

This result shows that the J48 pruned decision tree with 10-fold cross validation algorithm as per this set up scored an accuracy of 93.3415% from the total instance 7654 instances were correctly classified, while 546 instances were incorrectly classified.

EXPERIMENT #2. J48 pruned decision tree with 80 % split test mode

This experiment conducts under J48 pruned decision tree with 80 % split test mode option by setting the value of CF=0.25and MNO=2 to their default values and using J48 pruned decision tree with 80% split test option in WEKA. The snapshot of this experiment is shown in Figure 4.5. Here and for all the next subsequent tests the class attribute is set to Class (having values of C=Competent, NYC=Not Yet Competent).

The Experiment Result is as shown below in figure 4.5

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.09 seconds

=== Summary ===

Correctly Classified Instances        1527               93.1098 %
Incorrectly Classified Instances       113                6.8902 %
Kappa statistic                          0.8616
Mean absolute error                      0.0967
Root mean squared error                  0.2455
Relative absolute error                 19.3316 %
Root relative squared error             49.0787 %
Total Number of Instances             1640

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.891    0.032    0.963      0.891    0.926      0.864    0.960     0.963     competent
                 0.968    0.109    0.905      0.968    0.936      0.864    0.960     0.943     Not Yet Comp
Weighted Avg.    0.931    0.072    0.933      0.931    0.931      0.864    0.960     0.952

=== Confusion Matrix ===

   a   b   <-- classified as
 704  86 |   a = competent
  27 823 |   b = Not Yet Competant
```

Figure 4.5: WEKA Snapshot of J48 pruned decision tree with 80 % split test mode

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (704+823)/ (704+27+823+86) = 0.931097560975=**93.10975%**

Therefore, the overall accuracy of the model was measured at 93.1098 %. This result shows that the J48 pruned decision tree with 80% split test learning algorithm as per this set up scored an accuracy of 93.1098% From the total training set 1527 instances were correctly classified, while 113 instances were incorrectly classified.

EXPERIMENT #3. J48 pruned decision tree with 66% split test mode

By setting the value of CF=0.25 default and MNO=2 (Default), and using 66% split test option in WEKA, the snapshot of this experiment is shown in Figure 4.6

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances        2581               92.5753 %
Incorrectly Classified Instances       207                7.4247 %
Kappa statistic                          0.8514
Mean absolute error                      0.102
Root mean squared error                  0.2526
Relative absolute error                 20.4015 %
Root relative squared error             50.519  %
Total Number of Instances             2788

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.879    0.028    0.969      0.879    0.922      0.855    0.955     0.961     competent
                 0.972    0.121    0.890      0.972    0.930      0.855    0.955     0.931     Not Yet Comp
Weighted Avg.    0.926    0.075    0.929      0.926    0.926      0.855    0.955     0.946

=== Confusion Matrix ===

    a    b   <-- classified as
 1216  168 |    a = competent
   39 1365 |    b = Not Yet Competant
```

Figure 4.5: WEKA Snapshot J48 pruned decision tree with 66% split test mode

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (1216+1365)/ (1216+39+1365+168) = 0.9257532281=**92.5753%**

Therefore, the overall accuracy of the model was measured at 92.5753 %.

This result shows that the J48 pruned decision tree with 66% split test learning algorithm as per this set up scored an accuracy of 92.5753% from the total training set 2581 Instances were correctly classified, while 207 instances were incorrectly classified.

Summary of performance result is presented in table 4.3 using J48 pruned algorithm by changing test modes.

Table 4.3: Summary of experimental result of J48 pruned Decision Trees' algorithm

S. No	**Comparing parameters**	**Experiments' No**		
		#1	**#2**	**#3**
1	Testing Mode	10-Fold Cross Validation	80%	66%
2	Confidence Factor	2.5	2.5	2.5
3	TP Rate	0.933	0.931	0.926
4	FP Rate	0.067	0.072	0.075
5	Time Taken (sec.)	0.33	0.09	0.02
6	Precision	0.937	0.933	0.929
7	Recall	0.933	0.931	0.926
8	F-Measure	0.933	0.931	0.926
9	MCC	0.870	0.864	0.855
10	ROC area	0.960	0.960	0.955
11	PRC Area	0.952	0.952	0.946
12	**Accuracy (%)**	**93.3415%**	**93.1098%**	**92.5753%**

The above J48 pruned decision tree experiments were basically done by using 80% and 66% split test modes and Cross-validation fold of 10. As shown in Table 4.4, the overall accuracy of the

selected classifier is **93.3415%** is selected among the above three experiment. This means J48 pruned Decision tree algorithm with 10-fold cross validation test option has a better classification performance for identifying “Competent“ and “Not Yet Competent” class.

EXPERIMENT #4. J48 unpruned decision tree with 10-fold cross validation test mode

By setting the value of CF=0.25 default and MNO=2 (Default), and using 10-fold cross validation test option in WEKA, the snapshot of this experiment is shown in **Figure 4.6.**

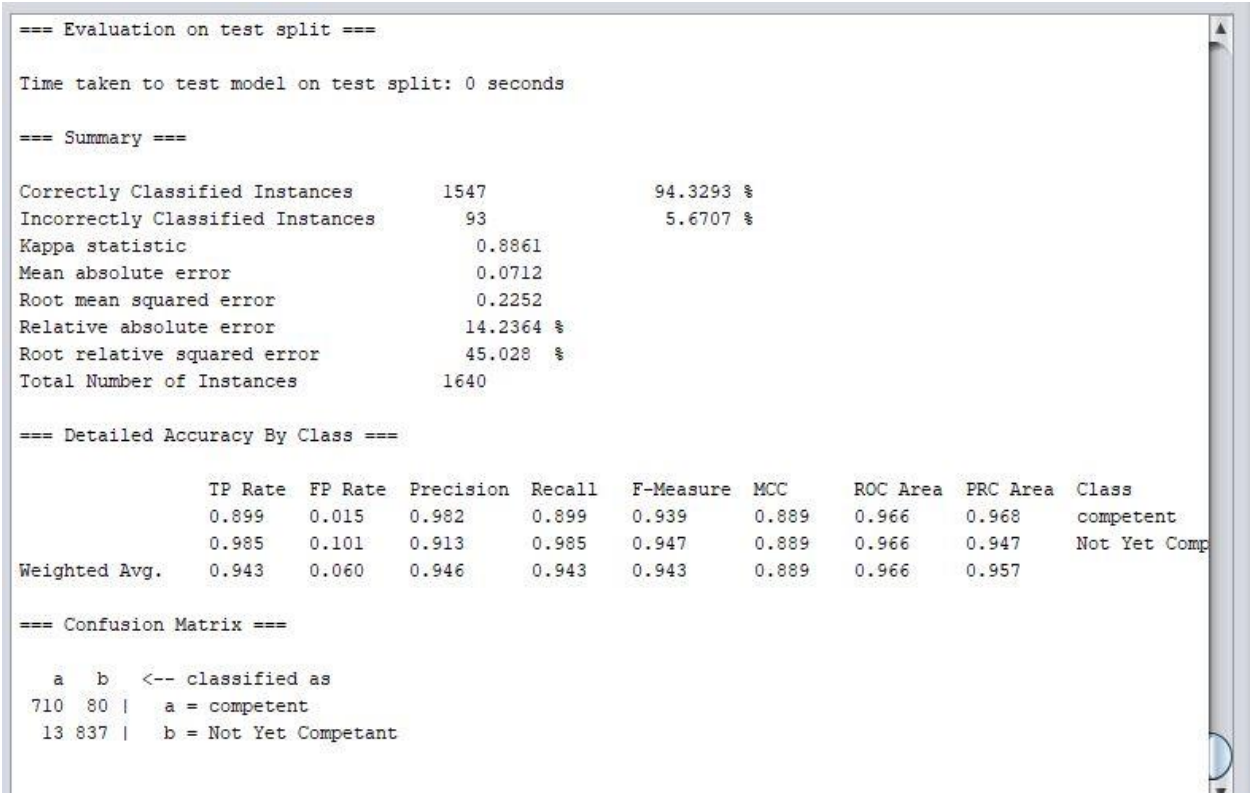
```
Time taken to build model: 0.19 seconds

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

Correctly Classified Instances        7754               94.561  %
Incorrectly Classified Instances       446                5.439  %
Kappa statistic                          0.8912
Mean absolute error                      0.0649
Root mean squared error                  0.2191
Relative absolute error                 12.9835 %
Root relative squared error             43.8256 %
Total Number of Instances            8200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.903     0.012     0.987      0.903    0.943      0.894    0.968     0.973     competent
               0.988     0.097     0.911      0.988    0.948      0.894    0.968     0.944     Not Yet C
Weighted Avg.  0.946     0.054     0.949      0.946    0.946      0.894    0.968     0.959

=== Confusion Matrix ===

    a    b   <-- classified as
 3704  396 |    a = competent
   50 4050 |    b = Not Yet Competant
```

Figure 4.6: WEKA Snapshot J48 Un-Pruned Decision Tree with 10-fold cross validation

Therefore, the overall accuracy of the model was measured at 94.561%. This result shows that the J48 unpruned decision tree with10-fold cross validation test mode as per this set up scored an accuracy of 94.561% From the total training set 7754 instances were correctly classified, while 446 instances were incorrectly classified.

EXPERIMENT #5: J48 Un-Pruned Decision Tree with 80% split test mode

By setting the value of CF=0.25and MNO=2 to their default values and using J48 unpruned decision tree with 80% split test option in WEKA, the snapshot of this experiment is shown in Figure 4.7.



```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances        1547               94.3293 %
Incorrectly Classified Instances        93                5.6707 %
Kappa statistic                          0.8861
Mean absolute error                      0.0712
Root mean squared error                  0.2252
Relative absolute error                 14.2364 %
Root relative squared error             45.028  %
Total Number of Instances             1640

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.899    0.015    0.982      0.899    0.939      0.889    0.966     0.968     competent
                 0.985    0.101    0.913      0.985    0.947      0.889    0.966     0.947     Not Yet Comp
Weighted Avg.    0.943    0.060    0.946      0.943    0.943      0.889    0.966     0.957

=== Confusion Matrix ===

   a   b   <-- classified as
 710  80 |   a = competent
  13 837 |   b = Not Yet Competant
```

Figure 4.7: WEKA Snapshot of J48 Un-Pruned Decision Tree with 80% split test mode

According to equation 4.1.,

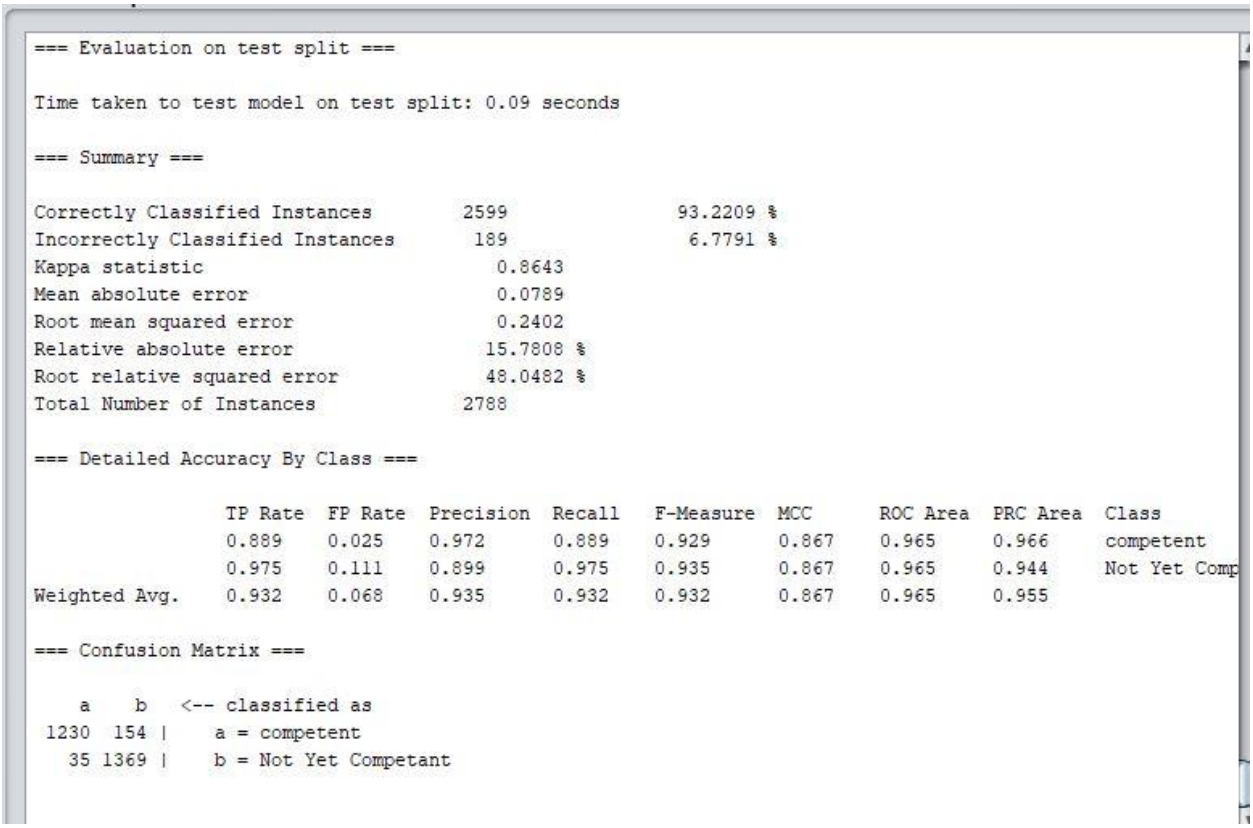
Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (710+837)/ (710+13+837+80) = 0.94329=**94.32926%**

Therefore, the overall accuracy of the model was measured at 94.3293 %. This result shows that the J48 unpruned decision tree with 80% split test learning algorithm as per this set up scored an accuracy of 94.3293% From the total training set 1547 instances were correctly classified, while 93 instances were incorrectly classified.

EXPERIMENT #6. J48 Un-Pruned Decision tree with 66% split test mode

By setting the value of CF=0.25 default and MNO=2 (Default), and using 64% split test option in WEKA, the snapshot of this experiment is shown in **Figure 4.8.**



```
=== Evaluation on test split ===

Time taken to test model on test split: 0.09 seconds

=== Summary ===

Correctly Classified Instances        2599               93.2209 %
Incorrectly Classified Instances       189                6.7791 %
Kappa statistic                          0.8643
Mean absolute error                      0.0789
Root mean squared error                  0.2402
Relative absolute error                 15.7808 %
Root relative squared error             48.0482 %
Total Number of Instances             2788

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.889     0.025     0.972      0.889    0.929      0.867    0.965     0.966     competent
               0.975     0.111     0.899      0.975    0.935      0.867    0.965     0.944     Not Yet Comp
Weighted Avg.  0.932     0.068     0.935      0.932    0.932      0.867    0.965     0.955

=== Confusion Matrix ===

    a    b   <-- classified as
 1230  154 |    a = competent
   35 1369 |    b = Not Yet Competant
```

Figure 4.8: WEKA Snapshot of J48 Un-Pruned Decision Tree with 66% split test mode

Therefore, the overall accuracy of the model was measured at 93.2209 %.

This result shows that the J48 unpruned decision tree with 66% split test learning algorithm as per this set up scored an accuracy of 93.2209% From the total training set 2599 instances were correctly classified, while 189 instances were incorrectly classified.

Summary of experimental results of J48 unpruned Decision Tree is summarized in table 4.3.

Table 4.4 Summary of experimental result of J48 unpruned Decision Trees' algorithm

S. No	**Comparing parameters**	**Experiments' No**		
		#4	**#5**	**#6**
1	Testing Mode	10-Fold Cross Validation	80%	66%
2	Confidence Factor	2.5	2.5	2.5
3	TP Rate	0.946	0.943	0.932
4	FP Rate	0.054	0.060	0.068
5	Time Taken (sec.)	0.19	0.0	0.09
6	Precision	0.949	0.946	0.935
7	Recall	0.946	0.943	0.932
8	F-Measure	0.946	0.943	0.932
9	MCC	0.894	0.889	0.867
10	ROC area	0.968	0.966	0.965
11	PRC Area	0.959	0.957	0.955
12	**Accuracy (%)**	**94.561%**	**94.3293%**	**93.2209%**

As shown in Table 4.3, the overall accuracy among the above three experiment 94.561% is selected. This means J48 Un-pruned Decision tree algorithm with 10-fold cross validation test

option has a better classification performance for identifying “Competent” and “Not Yet Competent” class.

4.5.2 MODEL BUILDING USING JRIP ALGORITHM

This experiment was performed on the JRiP pruned and Unpruned algorithm; it is an alternative representative of a classification Rules follows the same procedure applied on the previous experiment which are presented above. The experiments were run on the training dataset to build the model and its quality was estimated on the test dataset. The result six experiments conducted with 10-fold cross validation, 80% and 66% split test is presented in table 4.6 and 4.7 respectively.

EXPERIMENT #7. JRIP Pruned 10-fold cross validation

By setting WEKA default classification test option to 10-fold cross validation, the following result is obtained from WEKA the snapshot for this experiment is shown in Figure 4.9

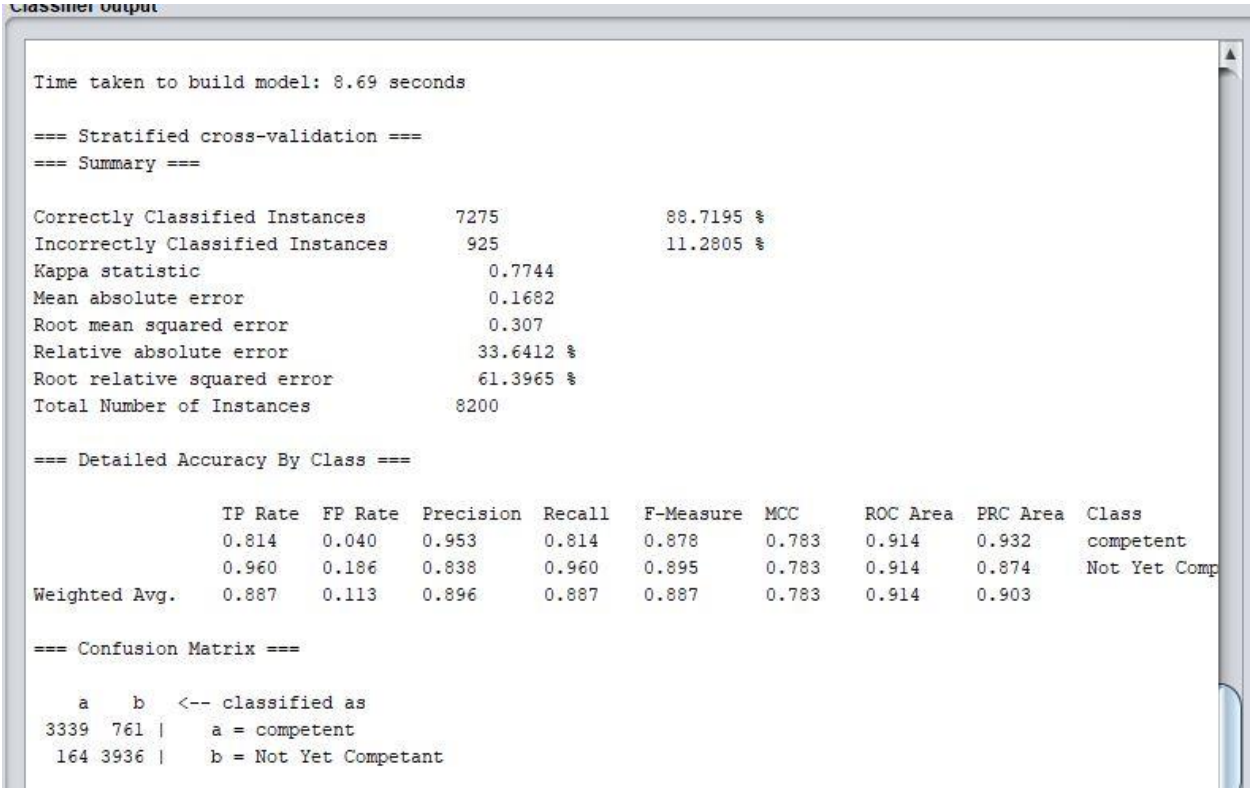


Figure 4.9: WEKA Snapshot of JRIP 10-Fold cross Validation

Therefore, the overall accuracy of the model was measured at 88.7195%. This result shows that the JRIP with 10-fold cross validation algorithm as per this set up scored an accuracy of 88.7195% From the total training set 7275 instances were correctly classified, while 925 instances were incorrectly classified

EXPERIMENT #8. JRIP Pruned with 80% split test mode

By setting WEKA default classification test option to 80% split, the following result is obtained, the snapshot for this experiment is shown in Figure 4.10

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.19 seconds

=== Summary ===

Correctly Classified Instances        1431               87.2561 %
Incorrectly Classified Instances       209               12.7439 %
Kappa statistic                          0.743
Mean absolute error                      0.202
Root mean squared error                  0.3229
Relative absolute error                 40.3929 %
Root relative squared error             64.5587 %
Total Number of Instances             1640

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.770    0.032    0.957      0.770    0.853      0.757    0.893     0.881     competent
                 0.968    0.230    0.819      0.968    0.887      0.757    0.893     0.861     Not Yet Competant
Weighted Avg.    0.873    0.135    0.886      0.873    0.871      0.757    0.893     0.871

=== Confusion Matrix ===

   a   b   <-- classified as
 608 182 |   a = competent
  27 823 |   b = Not Yet Competant
```

Figure 4.10: WEKA Snapshot of JRIP Pruned 80% split Test Data

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (608+823)/ (608+27+823+182) = 0.87256=**87.256%**

Therefore, the overall accuracy of the model was measured at 87.256%. This result shows that the JRIP Pruned with 80% split test learning algorithm as per this set up scored an accuracy of

87.256% From the total training set 1431 instances were correctly classified, while 209 instances were incorrectly classified.

EXPERMENT #9: JRIP Pruned with 66% split test mode

By setting WEKA default classification test option to 66% split, the following result is obtained, the snapshot for this experiment is shown in Figure 4.11

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances        2452               87.9484 %
Incorrectly Classified Instances       336               12.0516 %
Kappa statistic                          0.7587
Mean absolute error                      0.1816
Root mean squared error                  0.3134
Relative absolute error                 36.3191 %
Root relative squared error             62.6819 %
Total Number of Instances             2788

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.796     0.038     0.953      0.796     0.868      0.769     0.913     0.908     competent
               0.962     0.204     0.827      0.962     0.889      0.769     0.913     0.882     Not Yet Comp
Weighted Avg.  0.879     0.122     0.890      0.879     0.879      0.769     0.913     0.895

=== Confusion Matrix ===

    a    b   <-- classified as
 1102  282 |    a = competent
   54 1350 |    b = Not Yet Competant
```

Figure 4.11: WEKA Snapshot of JRIP Pruned 66% split Test Data

Therefore, the overall accuracy of the model was measured at 87.9484%.

This result shows that the JRIP with 66% split test learning algorithm as per this set up scored an accuracy of 87.9484% From the total training set 2452 instances were correctly classified, while 336 instances were incorrectly classified

summary of experimental results of JRIP algorithms is summarized in table 4.5.

Table 4.5 Summary of experimental result of JRIP pruned algorithm

S. No	Comparing parameters	Experiments' No		
		#7	**#8**	**#9**
1	Testing Mode	10-Fold Cross Validation	80%	66%
2	Confidence Factor	2.5	2.5	2.5
3	TP Rate	0.887	0.873	0.879
4	FP Rate	0.113	0.135	0.122
5	Time Taken (sec.)	8.69	0.19	0.03
6	Precision	0.896	0.886	0.890
7	Recall	0.887	0.873	0.879
8	F-Measure	0.887	0.871	0.879
9	MCC	0.783	0.757	0.769
10	ROC area	0.914	0.893	0.913
11	PRC Area	0.903	0.871	0.895
12	**Accuracy (%)**	**88.7195%**	**87.2561%**	**87.9484%**

When we compare the performance of the models produced by JRip Pruned algorithm the use of 10-fold Cross Validation test learning algorithm registered the highest accuracy of 88.7195%.

EXPERIMENT #10. JRIP Un-Pruned 10-fold cross validation

By setting WEKA default classification test option to 10-fold cross validation, the following result is obtained from WEKA the snapshot for this experiment is shown in Figure 4.12

```
Time taken to build model: 5.2 seconds

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

Correctly Classified Instances        7076               86.2927 %
Incorrectly Classified Instances      1124               13.7073 %
Kappa statistic                          0.7259
Mean absolute error                      0.1891
Root mean squared error                  0.3231
Relative absolute error                 37.8242 %
Root relative squared error             64.6132 %
Total Number of Instances             8200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.874    0.148    0.855      0.874    0.864      0.726    0.915     0.924     competent
                 0.852    0.126    0.871      0.852    0.861      0.726    0.915     0.877     Not Yet Comp
Weighted Avg.    0.863    0.137    0.863      0.863    0.863      0.726    0.915     0.900

=== Confusion Matrix ===

    a    b   <-- classified as
 3584  516 |    a = competent
  608 3492 |    b = Not Yet Competant
```

Figure 4.12: WEKA Snapshot of JRIP Un-Pruned 10-Fold cross Validation

Therefore, the overall accuracy of the model was measured at 86.2927%. This result shows that the JRIP Un-Pruned with 10-fold cross validation algorithm as per this set up scored an accuracy of 86.2927% From the total training set 7076 instances were correctly classified, while 1124 instances were incorrectly classified

EXPERIMENT #11. JRIP Un-Pruned with 80% split test mode

By setting WEKA default classification test option to 80% split, the following result is obtained, the snapshot for this experiment is shown in Figure 4.13

```
Time taken to test model on test split: 0.05 seconds

=== Summary ===

Correctly Classified Instances        1394               85      %
Incorrectly Classified Instances       246               15      %
Kappa statistic                          0.7022
Mean absolute error                      0.2099
Root mean squared error                  0.3483
Relative absolute error                 41.9708 %
Root relative squared error             69.6396 %
Total Number of Instances             1640

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.966    0.258    0.777      0.966    0.861      0.722    0.854     0.767     competent
                 0.742    0.034    0.959      0.742    0.837      0.722    0.854     0.845     Not Yet Competant
Weighted Avg.    0.850    0.142    0.871      0.850    0.849      0.722    0.854     0.808

=== Confusion Matrix ===

   a   b   <-- classified as
 763  27 |   a = competent
 219 631 |   b = Not Yet Competant
```

Figure 4.13: WEKA Snapshot of JRIP Un-Pruned 80% split Test Data

According to equation 4.1.,

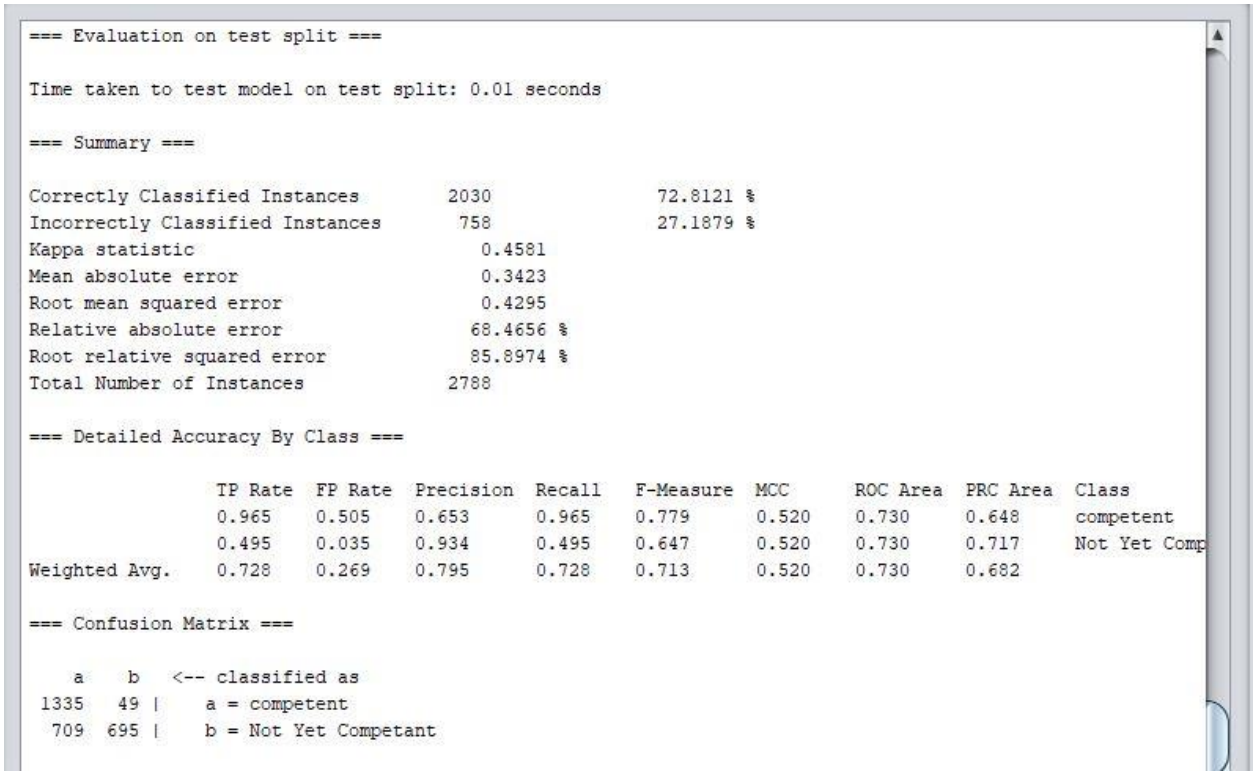
$$\text{Accuracy} = (TP+TN)/ (TP+FP+TN+FN)$$

$$= (763+631)/ (763+219+631+27) = 0.85=\mathbf{85\%}$$

Therefore, the overall accuracy of the model was measured at 85%. This result shows that the JRIP Pruned with 80% split test learning algorithm as per this set up scored an accuracy of 85% From the total training set 1394 instances were correctly classified, while 246 instances were incorrectly classified.

EXPERMENT #12: JRIP Un-Pruned with 66% split test mode

By setting WEKA default classification test option to 66% split, the following result is obtained, the snapshot for this experiment is shown in Figure 4.14



```
=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances        2030               72.8121 %
Incorrectly Classified Instances       758               27.1879 %
Kappa statistic                          0.4581
Mean absolute error                      0.3423
Root mean squared error                  0.4295
Relative absolute error                 68.4656 %
Root relative squared error             85.8974 %
Total Number of Instances             2788

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.965    0.505    0.653      0.965    0.779      0.520    0.730     0.648     competent
                 0.495    0.035    0.934      0.495    0.647      0.520    0.730     0.717     Not Yet Comp
Weighted Avg.    0.728    0.269    0.795      0.728    0.713      0.520    0.730     0.682

=== Confusion Matrix ===

    a    b   <-- classified as
 1335   49 |    a = competent
  709  695 |    b = Not Yet Competant
```

Figure 4.14: WEKA Snapshot of JRIP Un-Pruned 66% split Test Data

Therefore, the overall accuracy of the model was measured at 72.8121%.

This result shows that the JRIP with 66% split test learning algorithm as per this set up scored an accuracy of 72.8121% From the total training set 2030 instances were correctly classified, while 758 instances were incorrectly classified

Summary of experimental results of JRIP Un-Pruned algorithms is summarized in table 4.6

Table 4.6 Summary of experimental result of JRIP Un-pruned algorithm

S. No	Comparing parameters	Experiments' No		
		#10	**#11**	**#12**
1	Testing Mode	10-Fold Cross Validation	80%	66%
2	Confidence Factor	2.5	2.5	2.5
3	TP Rate	0.863	0.850	0.728
4	FP Rate	0.137	0.142	0.269
5	Time Taken (sec.)	5.2	0.06	0.01
6	Precision	0.863	0.871	0.795
7	Recall	0.863	0.850	0.728
8	F-Measure	0.863	0.849	0.713
9	MCC	0.726	0.722	0.520
10	ROC area	0.915	0.854	0.730
11	PRC Area	0.900	0.808	0.682
12	**Accuracy (%)**	**86.2927%**	**85%**	**72.8121%**

When we compare the performance of the models produced by JRiP Un-Pruned algorithm the use of 10-fold Cross Validation test learning algorithm registered the highest accuracy of 86.2927%.

4.5.3 MODEL BUILDING USING NAÏVE BAYES ALGORITHM

The Third classification model experiment is conducted using the Naïve Bayes statistical classifier. Also, this algorithm is mostly used to define classes and predict future behavior of existing instances. The WEKA default classification test option, which is 10-fold cross validation and the percentage split with the default distribution of instances, 66% for training and 34% for testing and 80% for training and 20% for testing are used.

EXPERIMENT #13. Naive Bayes 10-fold cross validation

By setting WEKA default classification test option to 10-fold cross validation, the following result is obtained, the snapshot for this experiment is shown in Figure 4.15

```
Time taken to build model: 0.09 seconds

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

Correctly Classified Instances        5933               72.3537 %
Incorrectly Classified Instances      2267               27.6463 %
Kappa statistic                          0.4471
Mean absolute error                      0.3639
Root mean squared error                  0.4366
Relative absolute error                 72.7775 %
Root relative squared error             87.314  %
Total Number of Instances             8200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.737    0.290    0.718      0.737    0.727      0.447    0.784     0.759     competent
                 0.710    0.263    0.730      0.710    0.720      0.447    0.784     0.774     Not Yet Comp
Weighted Avg.    0.724    0.276    0.724      0.724    0.723      0.447    0.784     0.767

=== Confusion Matrix ===

    a    b   <-- classified as
 3022 1078 |    a = competent
 1189 2911 |    b = Not Yet Competant
```

Figure 4.15: WEKA Snapshot of Naïve Bayes 10-fold cross validation Test Data

Therefore, the overall accuracy of the model was measured at 72.3537 %.

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (3022+2911)/ (3022+1189+2911+1078) = 0.7235365=**72.3537%**

This result shows that the Naive Bayes with 10-fold cross validation algorithm as per this set up scored an accuracy of 72.3537.5% From the total training set 5933 instances were correctly classified, while 2267 instances were incorrectly classified.

EXPERIMENT #14. Naive Bayes with 80% split test mode

By setting WEKA default classification test option to 80% split, the following result is obtained, the snapshot for this experiment is shown in Figure 4.16

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances        1180               71.9512 %
Incorrectly Classified Instances       460               28.0488 %
Kappa statistic                          0.4401
Mean absolute error                      0.364
Root mean squared error                  0.4354
Relative absolute error                 72.7755 %
Root relative squared error             87.0397 %
Total Number of Instances             1640

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.754    0.313    0.691      0.754    0.722      0.442    0.787     0.749     competent
                 0.687    0.246    0.751      0.687    0.717      0.442    0.787     0.789     Not Yet Comp
Weighted Avg.    0.720    0.278    0.722      0.720    0.719      0.442    0.787     0.770

=== Confusion Matrix ===

   a   b   <-- classified as
 596 194 |   a = competent
 266 584 |   b = Not Yet Competant
```

Figure 4.16: WEKA Snapshot of Naïve Bayes 80% split Test Data

Therefore, the overall accuracy of the model was measured at 71.9512 %.

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (596+584)/ (596+266+584+194) = 0.71951219=**71.9512%** This result shows that the Naive Bayes with 80% split test learning algorithm as per this set up scored an accuracy of 71.9512% From the total training set 1180 instances were correctly classified, while 460 instances were incorrectly classified.

EXPERIMENT #15. Naive Bayes with 66% split test mode

By setting WEKA default classification test option to 66% split, the following result is obtained, the snapshot for this experiment is shown in Figure 4.17

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances        1975               70.8393 %
Incorrectly Classified Instances       813               29.1607 %
Kappa statistic                          0.417
Mean absolute error                      0.3639
Root mean squared error                  0.4366
Relative absolute error                 72.7821 %
Root relative squared error             87.3158 %
Total Number of Instances             2788

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.730    0.313    0.697      0.730    0.713      0.417    0.783     0.756     competent
                 0.687    0.270    0.721      0.687    0.704      0.417    0.783     0.780     Not Yet Comp
Weighted Avg.    0.708    0.291    0.709      0.708    0.708      0.417    0.783     0.768

=== Confusion Matrix ===

    a    b   <-- classified as
 1010  374 |    a = competent
  439  965 |    b = Not Yet Competant
```

Figure 4.17: WEKA Snapshot of Naïve Bayes 66% split Test Data

Therefore, the overall accuracy of the model was measured at 70.8393 %.

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (1010+965)/ (1010+439+965+374) = 0.708393=**70.8393%** This result shows that the Naive Bayes decision tree with 66% split test learning algorithm as per this set up scored an accuracy of 70.8393% From the total training set 1975 instances were correctly classified, while 813 instances were incorrectly classified.

The summary result for three experiments conducted with 10-fold cross validation, and 80 and 66% split test is presented in table 4.7 below.

Table 4.7 Summary of experimental result of Naïve Bayes algorithm

S. No	**Comparing parameters**	**Experiments' No**		
		#13	**#14**	**#15**
1	Testing Mode	10-Fold Cross Validation	80%	66%
2	Confidence Factor	2.5	2.5	2.5
3	TP Rate	0.724	0.720	0.708
4	FP Rate	0.276	0.278	0.291
5	Time Taken (sec.)	0.09	0.02	0.03
6	Precision	0.724	0.722	0.709
7	Recall	0.724	0.720	0.708
8	F-Measure	0.723	0.719	0.708
9	MCC	0.447	0.442	0.417
10	ROC area	0.784	0.787	0.783
11	PRC Area	0.767	0.770	0.768
12	**Accuracy (%)**	**72.3537%**	**71.9512%**	**70.8393%**

When we compare the performance of the models produced by Naive Bayes the use of 10-fold validation registered the highest accuracy of 72.3537%

4.5.4 MODEL BUILDING USING PART ALGORITHM

EXPERIMENT #16. PART Pruned 10-fold cross validation

By setting WEKA default classification test option to 10-fold cross validation, the following result is obtained, the snapshot for this experiment is shown in Figure 4.18

```
Time taken to build model: 2.91 seconds

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

Correctly Classified Instances        7763               94.6707 %
Incorrectly Classified Instances       437                5.3293 %
Kappa statistic                          0.8934
Mean absolute error                      0.0675
Root mean squared error                  0.2211
Relative absolute error                 13.506  %
Root relative squared error             44.2281 %
Total Number of Instances             8200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.907    0.014    0.985      0.907    0.945      0.896    0.964     0.972     competent
                 0.986    0.093    0.914      0.986    0.949      0.896    0.964     0.938     Not Yet Competant
Weighted Avg.    0.947    0.053    0.950      0.947    0.947      0.896    0.964     0.955

=== Confusion Matrix ===

    a    b   <-- classified as
 3719  381 |    a = competent
   56 4044 |    b = Not Yet Competant
```

Figure 4.18: WEKA Snapshot of PART Pruned 10-fold cross validation Test Data

Therefore, the overall accuracy of the model was measured at 94.6707 %.

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (3719+4044)/ (3719+56+4044+381) = 0.946707=**94.6707%**

This result shows that the PART Pruned with 10-fold cross validation algorithm as per this set up scored an accuracy of 94.6707% From the total training set 7763 instances were correctly classified, while 437 instances were incorrectly classified.

EXPERIMENT #17. PART Pruned with 80% split test mode

By setting WEKA default classification test option to 80% split, the following result is obtained, the snapshot for this experiment is shown in Figure 4.19

```
Time taken to test model on test split: 0.17 seconds

=== Summary ===

Correctly Classified Instances        1549               94.4512 %
Incorrectly Classified Instances        91                5.5488 %
Kappa statistic                          0.8886
Mean absolute error                      0.0755
Root mean squared error                  0.227
Relative absolute error                 15.0952 %
Root relative squared error             45.3731 %
Total Number of Instances            1640

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.903    0.016    0.981      0.903    0.940      0.891    0.960     0.961     competent
                 0.984    0.097    0.916      0.984    0.948      0.891    0.960     0.938     Not Yet Competant
Weighted Avg.    0.945    0.058    0.947      0.945    0.944      0.891    0.960     0.949

=== Confusion Matrix ===

   a   b   <-- classified as
 713  77 |   a = competent
  14 836 |   b = Not Yet Competant
```

Figure 4.19: WEKA Snapshot of PART Pruned 80% split Test Data

Therefore, the overall accuracy of the model was measured at 94.4512 %.

According to equation 4.1.,

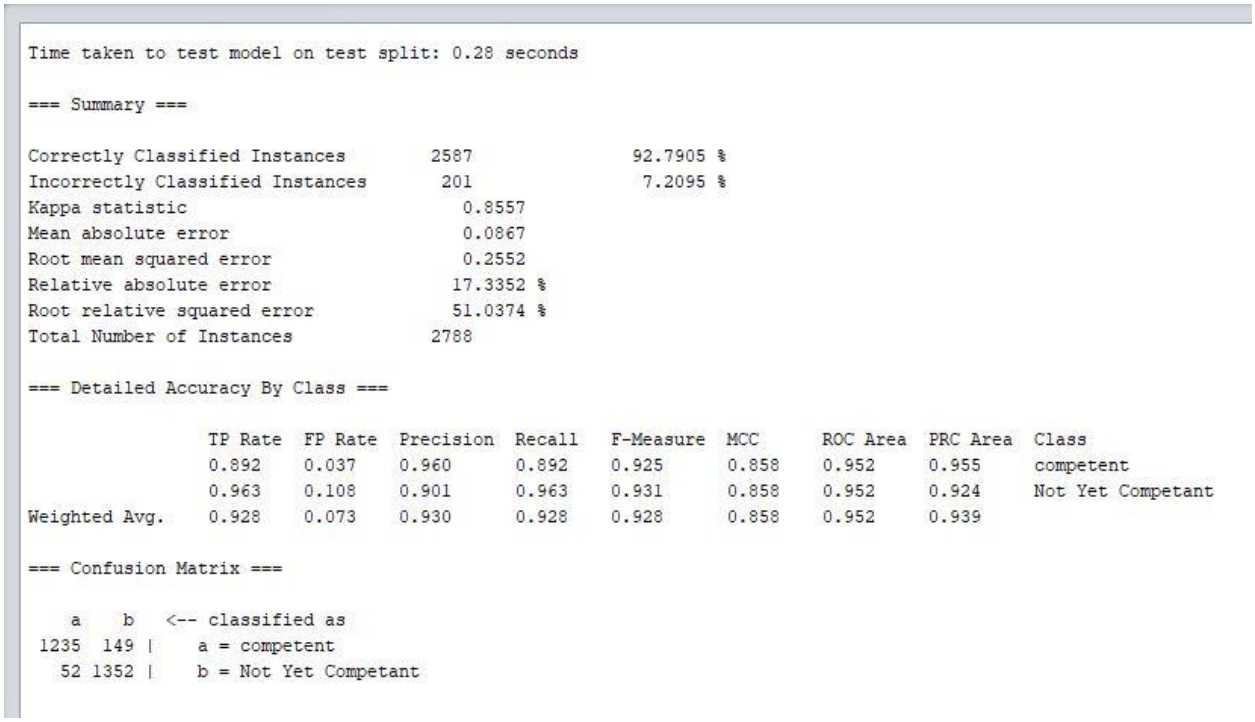
Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (713+836)/ (713+14+836+77) = 0.94451219=**94.4512%** This result shows that the PART Pruned with 80% split test learning algorithm as per this set up scored an

accuracy of 94.4512% From the total training set 1549 instances were correctly classified, while 91 instances were incorrectly classified.

EXPERIMENT #18. PART Pruned with 66% split test mode

By setting WEKA default classification test option to 66% split, the following result is obtained, the snapshot for this experiment is shown in Figure 4.20



```
Time taken to test model on test split: 0.28 seconds

=== Summary ===

Correctly Classified Instances        2587               92.7905 %
Incorrectly Classified Instances       201                7.2095 %
Kappa statistic                          0.8557
Mean absolute error                      0.0867
Root mean squared error                  0.2552
Relative absolute error                 17.3352 %
Root relative squared error             51.0374 %
Total Number of Instances             2788

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.892    0.037    0.960      0.892    0.925      0.858    0.952     0.955     competent
                 0.963    0.108    0.901      0.963    0.931      0.858    0.952     0.924     Not Yet Competant
Weighted Avg.    0.928    0.073    0.930      0.928    0.928      0.858    0.952     0.939

=== Confusion Matrix ===

    a    b   <-- classified as
 1235  149 |    a = competent
   52 1352 |    b = Not Yet Competant
```

Figure 4.20: WEKA Snapshot of PART Pruned 80% split Test Data

Therefore, the overall accuracy of the model was measured at 70.8393 %.

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (1235+1352)/ (1235+52+1352+149) = 0.927905=**92.7905%** This result shows that the PART Pruned with 66% split test learning algorithm as per this set up

scored an accuracy of 92.7905% From the total training set 2587 instances were correctly classified, while 201 instances were incorrectly classified.

The summary result for three experiments conducted with PART Pruned namely 10-fold cross validation, 80% and 66% split test is presented in table 4.8 below.

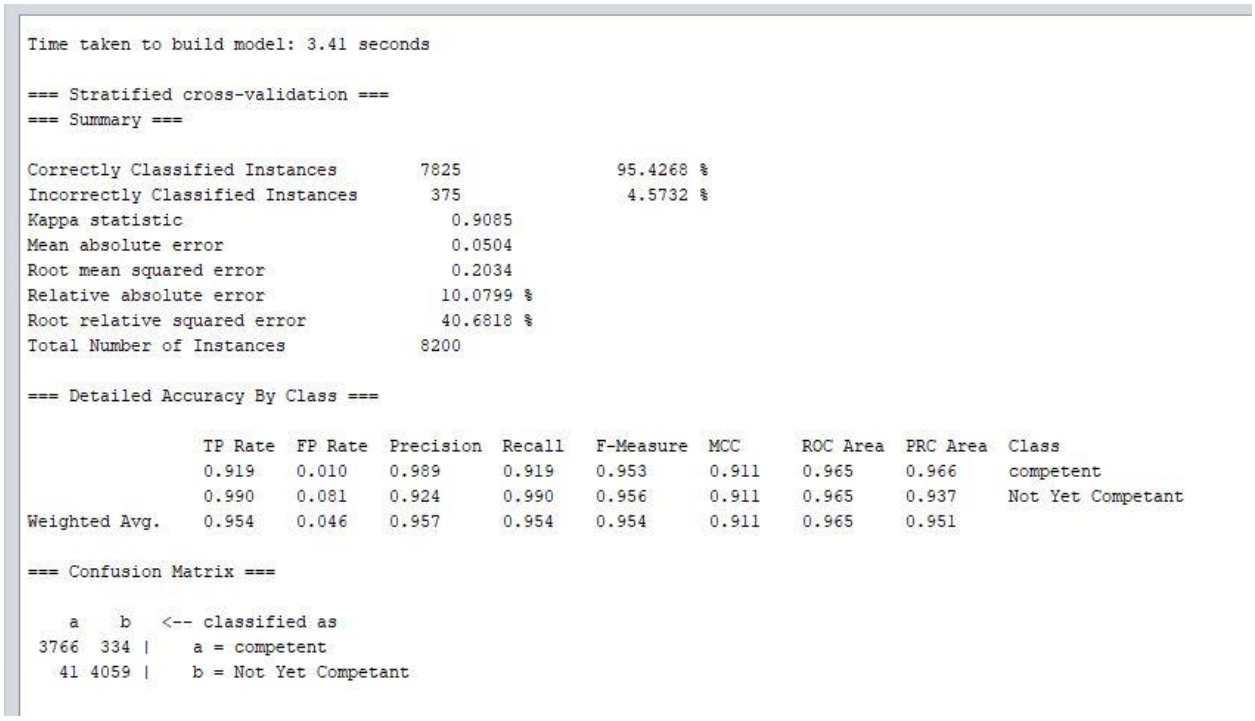
Table 4.8 Summary of experimental result of PART Pruned algorithm

S. No	**Comparing parameters**	**Experiments' No**		
		#16	**#17**	**#18**
1	Testing Mode	10-Fold Cross Validation	80%	66%
2	Confidence Factor	2.5	2.5	2.5
3	TP Rate	0.947	0.945	0.928
4	FP Rate	0.053	0.058	0.073
5	Time Taken (sec.)	2.91	0.17	0.28
6	Precision	0.950	0.947	0.930
7	Recall	0.947	0.945	0.928
8	F-Measure	0.947	0.944	0.928
9	MCC	0.896	0.891	0.858
10	ROC area	0.964	0.960	0.952
11	PRC Area	0.964	0.949	0.939
12	**Accuracy (%)**	**94.6707%**	**94.4512%**	**92.7905%**

When we compare the performance of the models produced by PART Pruned the use of 10-fold cross Validation registered the highest accuracy of **94.6707%.**

EXPERIMENT #19. PART Un-Pruned 10-fold cross validation

By setting WEKA default classification test option to 10-fold cross validation, the following result is obtained, the snapshot for this experiment is shown in Figure 4.21



```
Time taken to build model: 3.41 seconds

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

Correctly Classified Instances        7825               95.4268 %
Incorrectly Classified Instances       375                4.5732 %
Kappa statistic                          0.9085
Mean absolute error                      0.0504
Root mean squared error                  0.2034
Relative absolute error                 10.0799 %
Root relative squared error             40.6818 %
Total Number of Instances             8200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.919     0.010     0.989      0.919     0.953      0.911     0.965     0.966     competent
               0.990     0.081     0.924      0.990     0.956      0.911     0.965     0.937     Not Yet Competant
Weighted Avg.  0.954     0.046     0.957      0.954     0.954      0.911     0.965     0.951

=== Confusion Matrix ===

    a    b   <-- classified as
 3766  334 |    a = competent
   41 4059 |    b = Not Yet Competant
```

Figure 4.21: WEKA Snapshot of PART Un-Pruned 10-fold cross validation Test Data

Therefore, the overall accuracy of the model was measured at 95.4268%.

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (3766+4059)/ (3766+41+4059+334) = 0.954268=**95.4268%**

This result shows that the PART Un-Pruned with 10-fold cross validation algorithm as per this set up scored an accuracy of 95.4268% From the total training set 7825 instances were correctly classified, while 375 instances were incorrectly classified.

EXPERIMENT #20. PART Un-Pruned with 80% split test mode

By setting WEKA default classification test option to 80% split, the following result is obtained, the snapshot for this experiment is shown in Figure 4.22

```
Time taken to test model on test split: 0.39 seconds

=== Summary ===

Correctly Classified Instances        1563               95.3049 %
Incorrectly Classified Instances        77                4.6951 %
Kappa statistic                          0.9057
Mean absolute error                      0.0542
Root mean squared error                  0.2108
Relative absolute error                 10.8278 %
Root relative squared error             42.1443 %
Total Number of Instances             1640

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.910    0.007    0.992      0.910    0.949      0.909    0.961     0.959     competent
               0.993    0.090    0.922      0.993    0.956      0.909    0.961     0.934     Not Yet Competant
Weighted Avg.  0.953    0.050    0.956      0.953    0.953      0.909    0.961     0.946

=== Confusion Matrix ===

   a   b   <-- classified as
 719  71 |   a = competent
   6 844 |   b = Not Yet Competant
```

Figure 4.22: WEKA Snapshot of PART Un-Pruned 80% split Test Data

Therefore, the overall accuracy of the model was measured at 95.3049%.

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (719+844)/ (719+6+844+71) = 0.953048=**95.304878%** This result shows that the PART Un-Pruned with 80% split test learning algorithm as per this set up scored an accuracy of 95.3049% From the total training set 1563 instances were correctly classified, while 77 instances were incorrectly classified.

EXPERIMENT #21. PART Un-Pruned with 66% split test mode

By setting WEKA default classification test option to 66% split, the following result is obtained, the snapshot for this experiment is shown in Figure 4.23

```
Time taken to test model on test split: 0.55 seconds

=== Summary ===

Correctly Classified Instances        2595               93.0775 %
Incorrectly Classified Instances       193                6.9225 %
Kappa statistic                          0.8615
Mean absolute error                      0.0752
Root mean squared error                  0.2522
Relative absolute error                 15.0352 %
Root relative squared error             50.4436 %
Total Number of Instances             2788

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.884    0.024    0.974      0.884    0.927      0.865    0.942     0.941     competent
                 0.976    0.116    0.895      0.976    0.934      0.865    0.942     0.900     Not Yet Competant
Weighted Avg.    0.931    0.070    0.934      0.931    0.931      0.865    0.942     0.921

=== Confusion Matrix ===

    a    b   <-- classified as
 1224  160 |    a = competent
   33 1371 |    b = Not Yet Competant
```

Figure 4.23: WEKA Snapshot of PART Un-Pruned 66% split Test Data

Therefore, the overall accuracy of the model was measured at 93.0775%.

According to equation 4.1.,

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

= (1224+1371)/ (1224+33+1371+160) = 0.9307747=**93.07747%**

This result shows that the PART Un-Pruned with 66% split test learning algorithm as per this set up scored an accuracy of 93.0775% From the total training set 2595 instances were correctly classified, while 193 instances were incorrectly classified.

The summary result for three experiments conducted with PART Un-Pruned namely 10-fold cross validation, 80% and 66% split test is presented in table 4.9 below.

Table 4.9 Summary of experimental result of PART Un-Pruned algorithm

S. No	Comparing parameters	Experiments' No		
		#19	**#20**	**#21**
1	Testing Mode	10-Fold Cross Validation	80%	66%
2	Confidence Factor	2.5	2.5	2.5
3	TP Rate	0.954	0.953	0.931
4	FP Rate	0.046	0.050	0.070
5	Time Taken (sec.)	3.41	0.39	0.55
6	Precision	0.957	0.956	0.934
7	Recall	0.954	0.953	0.931
8	F-Measure	0.954	0.953	0.931
9	MCC	0.911	0.909	0.865
10	ROC area	0.965	0.961	0.942
11	PRC Area	0.951	0.946	0.921
12	**Accuracy (%)**	**95.4268%**	**95.3049%**	**93.0775%**

When we compare the performance of the models produced by PART Un-Pruned the use of 10-fold cross Validation registered the highest accuracy of **95.4268%.**

4.6 Comparison of the experimented classification models

In order to select a data mining model for classification tasks in the context of this study, it is necessary to evaluate the selected best model from J48 Pruned, J48 Un-Pruned, JRip Pruned, JRip Un-Pruned, PART Pruned, PART Un-Pruned and Naïve Bayes algorithms and it is summarized as follows in table 4.10.

Table 4.10: Comparison of experiment accuracy form the selected Algorithms

Comparison of **J48 Pruned, J48 Un-Pruned, JRip Pruned, JRip Un-Pruned, PART Pruned, PART Un-Pruned and Naïve Bayes** Classifier models								
S. No	**Comparing parameters**	**Selected Algorithm's**						
		J48 Pruned	**J48 Unpruned**	**Naive Bayes**	**JRiP Pruned**	**JRiP Unpruned**	**PART Pruned**	**PART Unpruned**
1	TP Rate	0.933	0.946	0.724	0.887	0.863	0.947	0.954
2	FP Rate	0.067	0.054	0.276	0.113	0.137	0.053	0.046
3	Time Taken (sec.)	0.33	0.19	0.09	8.69	5.2	2.91	3.41
4	Precision	0.937	0.949	0.724	0.896	0.863	0.950	0.957
5	Recall	0.933	0.946	0.724	0.887	0.863	0.947	0.954
6	F-Measure	0.933	0.946	0.723	0.887	0.863	0.947	0.954
7	MCC	0.870	0.894	0.447	0.783	0.726	0.896	0.911
8	ROC area	0.960	0.968	0.784	0.914	0.915	0.964	0.965
9	PRC Area	0.952	0.959	0.767	0.903	0.900	0.964	0.951
10	Accuracy (%)	**93.3415 %**	**94.561%**	**72.3537 %**	**88.7195 %**	**86.2927%**	**94.6707 %**	**95.4268%**

The best results of the above seven algorithms are compared with each other by their overall classification accuracy (performance). And, as it is shown in Table 4.10, the overall performance of the PART Un-pruned was 95.4268% with 8200 datasets and 10-fold cross validation test option. The precision recall and F-measure for this classifier is 0.957 and 0.954 respectively

while the ROC Area (AUC) is 0.965. The classification accuracy of the Naive Bayes model with the same data size and 10-fold cross validation test was 72.3537%. On the other hand, the classification accuracy of J48 pruned with 10-fold cross validation test option is 93.341%. The Fifth classification algorithm tested is JRIP Un-pruned, and it has shown overall performance of 96.0161% with 10-fold cross validation rule.

The PART unpruned Algorithm has shown better classification performance with 10-fold cross validation technique. Hence, it is reasonable to conclude that the PART Un-Pruned 10-fold cross validation model is the best classifier model for implementing Trainees academic performance prediction for Hawasa Poly Technic College. The Confusion matrix of the selected algorithm is presented as follows PART Un-pruned model with the highest accuracy has the following Confusion Matrix and related information in the bellow figure 4.24

```
Time taken to build model: 3.41 seconds

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

Correctly Classified Instances        7825               95.4268 %
Incorrectly Classified Instances       375                4.5732 %
Kappa statistic                          0.9085
Mean absolute error                      0.0504
Root mean squared error                  0.2034
Relative absolute error                 10.0799 %
Root relative squared error             40.6818 %
Total Number of Instances             8200

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.919     0.010     0.989      0.919     0.953      0.911     0.965     0.966     competent
               0.990     0.081     0.924      0.990     0.956      0.911     0.965     0.937     Not Yet Competant
Weighted Avg.  0.954     0.046     0.957      0.954     0.954      0.911     0.965     0.951

=== Confusion Matrix ===

    a    b   <-- classified as
 3766  334 |    a = competent
   41 4059 |    b = Not Yet Competant
```

Figure 4.24: WEKA Snapshot of PART Un-Pruned 10-fold Cross validation

The entries in the confusion matrix have the following meaning:

- **3766 is the number of correct predictions that an instance is Competent**
- **334 is the number of incorrect predictions that an instance is Not Yet Competent**

- **41 is the number of incorrect predictions that an instance is Competent**
- **4059 is the number of incorrect predictions that an instance is Not Yet Competent**

The above confusion matrix of PART Unpruned algorithm depicts that of 3766 are classified as **Competent** (91.854%) and the actually good 334 were classified as **Not Yet Competent**, (8.146%).

On the other hand, out of the resampled or balanced 4100 **Not Yet Competent**, 4059 were classified as **Not Yet Competent** (99%) and 41 of them were wrongly classified as **Competent** (1%). This entails that the records with the **competent** class are classified with higher error.

4.7 Selecting Classification Model

After the experimentation of this study with the J48 Pruned and Unpruned decision tree algorithm, Naive Bayes classifier, JRIP Pruned and Unpruned and PART Pruned and Unpruned Data Mining model comparing and selecting the one which performs the best is one of the deliverables of this phase.

All the experiments were carried out with the same dataset. From the output these experiments the highest accuracy is found by the PART Unpruned method. From Table 4.11 it can be seen that all the PART algorithm experiments have performed better than the J48, JRIP rule and Naïve Bayes classifier method.

4.8 Generating Rules

The experiments of Trainees Performance prediction using the PART Un-pruned 10-fold cross validation test showed better performance as compared to J48, JRip and Naïve Bayes with the same parameters. Hence this PART Un-pruned 10-fold cross validation test is selected for generating rules. The set of rules are extracted simply by traversing through the output of WEKA.

From the total generated rules for predicting trainee's performance; the following 10 rules were found interesting. Therefore, those rules which are believed to be unambiguous and relevant is selected based on the discussion with the domain expert. The numeric values which

appeared in the bracket next to the class label indicate the number of correctly and incorrectly classified records, respectively and are interpreted as follows.

1. ***Sector = Economy AND English = C AND EGSECE <= 2.286 AND age <= 19 AND Trade/ Occupation = IT service support AND Sex = F AND: Not Yet Competant (6.0/1.0)***

This rule show, if a trainee sector is economy and IT service support Occupation and age is less than or equal to 19 and Sex is Female and English result is C and EGSECE is less than 2.286 then the trainee has a probability to be Not yet Competent

2. ***Highschool = Private AND* Entry Year <= 2009 AND Sex = M AND *Mathematics = B: competent (162.0)***

This rule show, if High school is Private and Entry year is 2009 and Sex is Male and Mathematics is B then the trainee has a probability to be Competent

3. Sex = F AND **Entry Year <= 2011 AND Highschool = Government AND Level = IV AND EGSECE <= 2.286 AND English = D: Not Yet Competent (86.0)**

This rule shows, if trainee's entry year is 2011 and if trainees High school is Government and Level IV and female student with EGSECE less than 2.2 and scored English D grade is going to be Not Yet Competent

4. ***Entry Year <= 2009 AND Highschool = Private AND Sex = M AND Transcript > 70.5: Competent (124.0)***

This rule show, if trainee's entry year is 2009 and if trainees High school is private and male student with a transcript greater than 70.5 is going to be Competent

5. ***Entry Year <= 2009 AND Highschool = Government AND Level = IV AND EGSECE <= 2.733243 AND Trade/ Occupation = DBA AND Division = Regular: Not Yet Competant (95.0)***

This rule shows, if trainee's entry year is 2009 and if trainees High school is Government and Level IV in DBA Occupation with EGSECE less than 2.7 and division is regular is going to be Not Yet Competent

6. ***Sex = M AND Entry Year > 2009 AND Division = Regular AND Trade/ Occupation = Tourism service AND Transcript > 55: competent (132.0)***

This rule shows, if trainee's entry year 2009 and if trainees is male & Tourism service Occupation with a transcript score greater than 55 & division is Regular is going to be Competent

7. ***Sex = M AND Entry Year > 2009 AND Division = Regular AND Level = II AND Entry Year <= 2012 AND Highschool = Government AND Trade/ Occupation = Textile technology: competent (146.0)***

This rule shows, if trainee's entry year is in between 2009 and 2012 and if trainees is male and previous high school is Government and Textile technology Occupation in level II Regular division is going to be Competent

8. ***Sex = M AND Entry Year > 2009 AND Level = II AND Highschool = Government AND Entry Year <= 2012 AND Sector = Economy: competent (242.0)***

This rule shows, if trainee's entry year is in between 2009 and 2012 and if trainees is male and previous high school is Government and sector Economy with any Occupation in level II Regular division is going to be Competent

9. ***Sex = F AND Entry Year = 2011 AND Highschool = Government AND Trade/ Occupation = Furniture making AND English = C AND Mathematics = C: Not Yet Competant (37.0)***

This rule shows, if trainee's entry year is in 2012 and if trainees is Female and previous high school is Government and Furniture making Occupation with both English and Mathematics Score C is going to be Not Yet Competent

10. ***Sex = M AND Entry Year > 2009 AND Division = Regular AND Level = IV AND Woreda = Awasa AND EGSECE > 2.4 AND Highschool = Government AND Entry Year <= 2012 AND Trade/ Occupation = Electromechanical Equipment Maintenance Supervision : competent (92.0)***

This rule shows, if trainee's entry year is in between 2009 and 2012 and if trainees is male and previous high school is Government with EGSECE score greater than 2.4 in Electromechanical Equipment Maintenance Supervision Occupation level IV Regular division is going to be Competent.

4.9 Deployment of the result

The purpose of the data mining process is to increase the knowledge gained from the data stored. And, deployment is the last step of Knowledge Discovery process. The knowledge gained from data need to be organized and presented in a way that the organization can understand and use it for successful academic performance prediction. To make this result applicable, integration of resources like people, business processes, and technology, are required. Moreover, the integration of resource is based on the information or result obtained from the classification model. In this research work different models are developed using three classification algorithms and the result of the model are evaluated using performance evaluations. Finally the best model is selected. The rules are generated using PART unpruned model and the result of the research is discussed with domain experts. Based on discussion with domain experts this model is useful for Hawassa Polytechnic College to predict the Trainees Academic performance.

To evaluate the use of the model in addition to the response of domain experts, we compare the previous systems of the organization: there is no system that supports the organization to predict the trainee's academic performance. Besides domain experts, based on the result obtained the researcher discussed with the organizations stakeholders on the advantages of the result and ways of its implementation. Also even if the purpose of the study is academic purpose and the use of model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the Organization can use it. Prototype development is needed to show that the data mining model developed could be deployed. The researcher developed the prototype using VB.NET programming language for 10 selected rules of the model as shown in figure 4.25 and sample of the code is given in the appendix part.

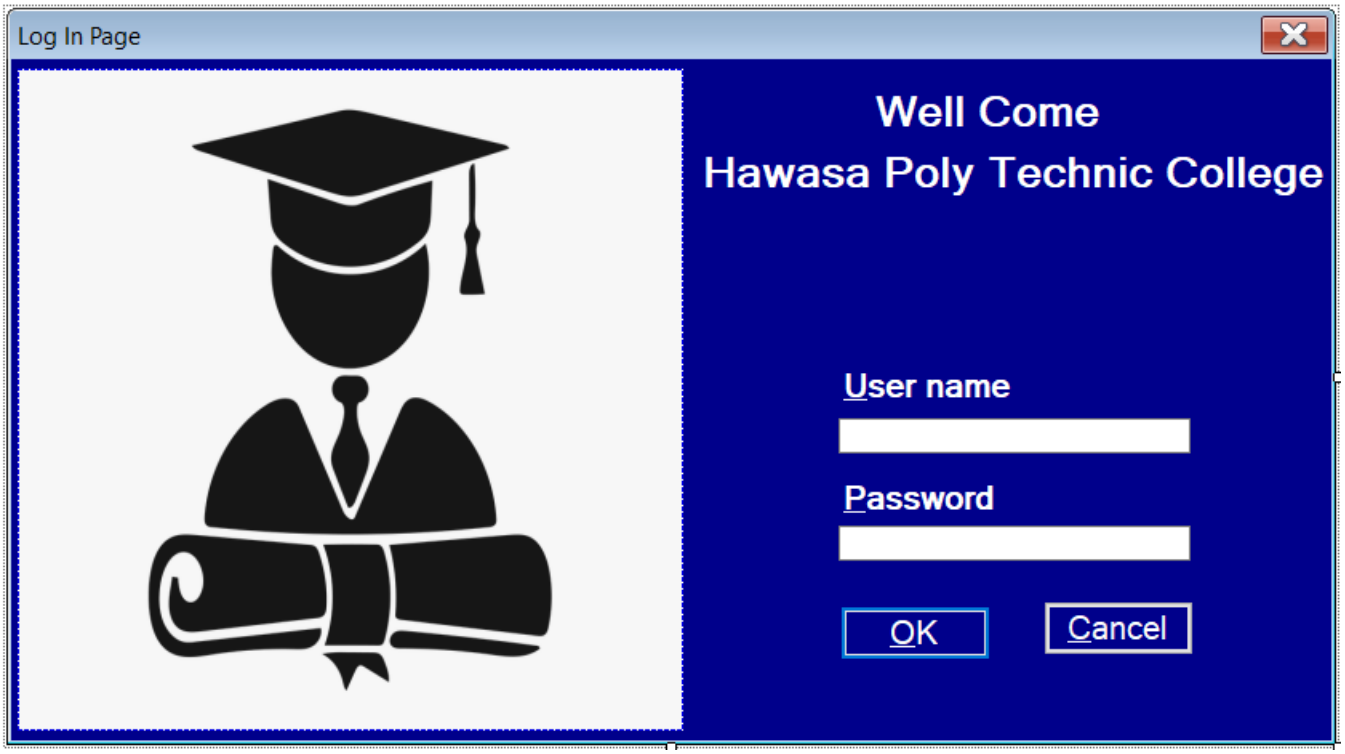


Figure 4.25: Snapshot Log In Page

Figure 4.26 Snapshot of Main Page for Performance Prediction

Finally we recommend the organization after integrating the necessary adjustments by group of domain experts the result of this study can be deployed for Trainees academic performance prediction and identification decision making and successful trainee's performance prediction in the organization and the organization will be beneficiary from the research result.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

In TVET institutions learning curriculum or process doesn‘t have a special concern to renovate student performance by showing students back history. The main focus of teaching learning is just given learning to acquire skill, attitude and knowledge. This type of process has been done for a long period of time and not gives the necessary information in resolving students learning problem.

Educational data mining (EDM) is a well-known field for high quality research that mines large data sets in order to answer educational research questions that shed light on the learning process. EDM has been concerned with developing methods for exploring unique and increasingly large scale data that come from educational settings and using those methods to better understand students and the settings which they learn in. The main objective of this research is to identify important and interesting patterns from the academic history of the trainees ‘records that can enable trainees ‘performance prediction for Hawasa polytechnic college.

In this research, the methodology employed was Hybrid Data mining process model; it involves six steps and the researcher thoroughly passes through all the steps and iterated as needed. The study was conducted using WEKA software version 3.8.4 and six data mining algorithms for classification techniques. A total of 8200datasets, 13 attributes and 1 outcome variable were used to build the model. Several experiments were conducted in order to build models that can predict student‘s performance in higher education.

Different experiments are conducted using J48 decision tree algorithm, PART rule induction, Naïve Bayes, and JRIP rule algorithm using 10- fold cross validation and percentage split. The experimental result shows that PART Unpruned algorithm outperforms with an accuracy of 95.4268%.

Based on, the extracted hidden pattern using PART Unpruned algorithm, trade/occupation, EGSECE, level, transcript ,English, mathematics , division, High school attended, sector and sex are identified as the major finding factors of student status.

The strength of this study is an achievement of all the stated goals, Data mining goal was to identify major attributes that contributes to trainees 'performance the study has identified and selected 13 attributes that are significant in predicting student performance. To design a predictive model Data mining techniques are more appropriate to predict trainees 'performance. The selected model built with PART Unpruned was able to answer this question by predicting 95.4268% of the cases correctly and develop a prototype of the trainee performance prediction interface. Understanding of the problem some attributes not considered due to data completeness and missing values found in the dataset are the challenge of this study.

5.2. Recommendations

This research work is conducted mainly for academic achievement. However, the researcher strongly believes that the findings of the study can be used by the concerned organizations to further investigate. Based on the findings obtained from the research, the researcher makes the following recommendation.

- This study used different classification algorithms in which unpruned Part rule induction performed better than others such as pruned J48 Decision Tree, unpruned J48 Decision Tree, Naïve Bayes, Unpruned JRip and pruned JRip Classification Rule algorithm. However, other classification algorithms might reveal a better accuracy. Therefore, further research must be conducted using other algorithms.
- Based on the findings of this study, attributes such as trade/occupation ,EGSECE, transcript, level ,sex, English, and sector can be used at a time of decision making as they had shown strong prediction power which can help to predict trainees performance
- In this research, the researcher use only Hawasa Poly Technique College; however further investigation is needed by including the other TVET College data.
- The study uses J48 classification, JRIP rule mining method, Naivebayes and PART algorithm to identify the determinant factor for the trainees' performance in Hawassa

Poly Technique College; both male and female trainees of Hawasa Poly Technique College most students failed to knowledge assessment in level III and IV assessment. So, the college need to give an attention to knowledge assessment in level III and IV assessment so as to improve the trainees' performance in Hawassa Poly Technique College.

- Education planners together with other interested parties should use the proposed potential set of attributes to design good and suitable plans to solve Trainees academic weakness.

References

[1] O. Negassa, "Ethiopian students' achievement challenges in science education: implications to Policy formulation," International Journal of Computer Applications, vol. 131, no. 5, 2014.

[2] Mrs.Bharati M. Ramageri, “DATA MINING TECHNIQUES AND APPLICATIONS” Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-30

[3] Han, J. (2006). *Data Mining: Concepts and Techniques* (Second edition ed.). Elsevier Incorporation.

[4] M. Anwar, "knowledge mining in supervised and unsupervised assessment," 2nd International conference on Networking and Information, vol. 17, 2015.

[5] N. S. Shah, "predicting factors that affect students’ academic performance by using data mining techniques," Pakistan business review, 2012.

[6] M. A. Yehuala, "Application of data mining technique for student success and failure prediction (the case of DebreMarkos university," International Journal of scientific &technology research, vol. 4, no. 04, 2015.

[7] K. J. Cios, P. Witold, S. Roman W. and K. Lukasz A., Data Mining: A Knowledge Discovery Approach, USA: Springer Science & Business Media, 2007

[8] Brijesh Kumar Bhardwaj and Saurabh Pal, "Data Mining: A prediction for performance improvement using classification," *(IJCSIS) International Journal of Computer Science and Information Security,*, vol. 9, no. 4, 2011.

[9] Z. Alebachew, Analysis of Urban Growth And Sprawl Mapping Using Remote Sensing And Geographic Information System, DebreBirhan: DebreBirhan Town., 2011.

[10] UNESCO IBE, world data on education, Revised 10/2010

[11] D.Delen,”predicting student attrition with data mining method”, J. college student retention,vol.13(1),pp.17-35,2011-2012.

[12] Maherezo Joseph, Tuyishimire Jean Damascène, and Niyigena Papias, “Modeling Trainee’s Dropout Predictor In KavumuTvet School”, GSJ: Volume 7, Issue 12, December 2019.

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

[14] 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.

[15] Two Crows Corporation.1999. Introduction To Data Mining And Knowledge Discovery.

[16] K. J. Sathick, "Extraction of Actionable knowledge to predict students' academic performance using data mining technique, an experimental study," vol. 1, no. 1, 2013.

[18] F. T., "Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining," International Conference on Information Acquisition, vol. 1, no. 1, 2006.

[21] A. Azwa, "First Semester Computer Science Academic Performances Analysis by using Data Mining Classification algorithms," AICS, vol. 1, no. 1, pp. 15-16, 2014.

[23] R. Naqvi, "Data Mining in Educational Settings," Pakistan Journal of Engineering, vol. 1, no. 4(2), p. 201.4615, 2015.

[26] C. Bambrah, M. Bhandari, N. Maniar, and V. Munde, "Mining Association Rules in Trainee Assessment Data," Ijarcce.Com, vol. 3, no. 3, pp. 5340–5342, 2014.

[27] J. Han, Kamber, M. and Pei, J., "Data Mining: Concepts and Techniques, third Edition ed". 225 Wyman Street, Waltham, USA: Morgan Kaufmann Publishers is an imprint of Elsevier, 2011.

[28] H. Tesfahun, "Application of Data Mining For Predicting Adult Mortality," Master Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2012.

[29] S. Institute, SAS Enterprise Minor SEMMA, 2016.

[31] K. Charly, "Data Mining for the Enterprise," 31st Annual Hawaii Int. Conf. on System Sciences, vol. 7, pp. 295-304, 1998.

[34] S. V.Sathiya and N. Dr.SaiSatya, "Data Mining Tasks Performed By Temporal Sequential Pattern," International Journal of Research and Computational Technology, vol. 2, no. 3, pp. 1-6, 2012.

[35] H. Jiawei, K. Micheline and P. Jian,, "Data Mining Concepts and Techniques," Morgan Kaufmann Publishers is an imprint of Elsevier, New York, 2012.

[36] S.O., Danso "An Exploration of Classification prediction techniques in data mining the insurance domain," Bournemouth university. Bournemouth, 2006.

[37] M. Vincent, "Introducing a data mining process framework to enable consultants to determine effective data analytics tasks," Master Thesis, University of Technology, Delft, 2012.

[38] C. d. R. Bruno and T. d. S. J. Rafael, "Identifying Bank Frauds using CRISP-DM and Decision Tree," International journal of computer science & information Technology (IJCSIT), vol. 2, no. 5, pp. 162-169, 2010.

[40] Q. J., "C4.5 Programs for Machine Learning", Los Altos: Morgan Kaufmann, 1993.

[41] M. K. J.B., Data Mining-Concepts, Models, Methods, and Algorithms, John Wiley, USA: Sons Publication Inc, 2003.

[42] T. T. R. a. F. J. Hastie, "the Elements of Statistical Learning" (2nd ed.). pp., 1- 764, 2008, Second Edition ed., Springer. USA, 2008.

[43] G., J. D. J. Z. Calivn, “Mastering Data Mining and Scince of Customer Relationship Management," A computer- Based herbsicide Injury Diagnostic Expert System. Weed Technology, vol, vol. 19, no. 7, pp. 486-491, 2005.

[44] S. Singh, "classification of students’ data using data mining techniques for training and placement department in technical education," International Journal of Computer Science and Network (IJCSN), vol. 1, no. 4, 2012.

[45] A. A. G. Aditi Mahajan, "Performance Evaluation of Rule Based Classification Algorithms," International Journal of Advanced Research in Computer Engineering & Technology , vol. 3, no. 10, pp. 12-19, 2014.

[46] C. G. Carrier and O. Povel, "Characterising data mining software," Intelligent Data Analysis, vol. 7, pp. 181 - 192., 2003.

[47] J. Luan, "Data mining and knowledge management in higher education potential applications." the association, for institutional research, vol. 1, no. 1, 2002.

[48] A. B. Michael J. and S. L. Gordon, "Data Mining Techniques for Marketing, Sales, and Customer Relationship Management", Second ed., Indianapolis, Indiana: Wiley Publishing, Inc., 2004.

[49] A.Meseret ,"A Combined Reasoning System For Knowledge Based Network Intrusion Detection," Master thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2016.

[50] E. W. T. Ngai, L. Xiu and D. C. K. Chau, "Application of data mining techniques in customer relationship management," A literature review and classification. Expert Systems with Applications, vol. 36, pp. 2592-2602, 2009.

[51] V. Kumar, "Mining Association Rules in Student’s Assessment Data," International Journal of Computer Science Issues, vol. 9, no. 5, 2012.

[52] A. Mohammed, "Towards Integrating Data Mining with Knowledge Based System: The Case of Network Intrusion detection," M.Sc. Thesis, Addis Ababa University, 2013.

[53] T .Verbraken, Bravo, C., Weber, R., &Baesens, B. , "Development and application of consumer credit scoring models using profit-based classification measures," European Journal of Operational Research, vol. 238(2), pp.505-513, (2014).

[55] M .Weiss, Sholom, Zhang, Tong. "Performance analysis and evaluation". InYeNong, editor. The Hand book of data mining. New Jeresy, USA: Lawerence Erlbaum Associates Inc., 2003.

[56] Witten, I. H. and Frank, E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco, CA, Morgan Kaufman, 2005.

[66] S. Sembiring, M. Zarlis, D. Hartama, & E. Wani, "Prediction of student academic performance by an application of data mining techniques", 2011 International Conference on Management and Artificial Intelligence, 6 (2011). 110–114.

[67] A. Muluken, "Application of Data Mining Techniques for Student Success and Failure Prediction," International Journal of Scientific & Technology Research, vol. 4, no. 4, 2015

[68] Chawla, N.,Bowyer,K., Hall, L.,Kegelmeyer, P. (2002). SMOTE: Synthetic Minority Over- Sampling Technique. Journal of Artificial Intelligence Research 16 (2002), 321-357

Appendix I: Sample Trainee Records in CSV format

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Sex	age	Division	Sector	'Trade/ Oc	Level	Woreda	Highschoc	'Entry Yea	EGSECE	Transcript	English	'Mathema	Status	
2	M	20	Regular	'Hotel and	'hotel ope	II	Awasa	Governme	2011	3.28	58	A	B	'competent '	
3	M	20	Regular	'Hotel and	'hotel ope	II	Awasa	Governme	2010	2.57	65	B	C	'competent '	
4	M	19	Regular	'Industry '	'Fruit &Ve	II	Awasa	Governme	2010	1.6	56	C	C	'competent '	
5	M	19	Regular	Economy	'Electro M	IV	Awasa	Governme	2011	2.8	67.4	C	A	'competent '	
6	F	19	Regular	Economy	'Road Con	IV	Awasa	'Private '	2009	3.28	58	A	B	'competent '	
7	M	20	Regular	Economy	Electronic	IV	Awasa	Governme	2012	2.57	50	C	C	'competent '	
8	M	19	Regular	Economy	HNS	II	Awasa	'Private '	2009	2.57	54.4	C	B	'competent '	
9	M	19	Regular	'Industry '	'Metal ma	IV	Awasa	Governme	2011	2.71	66	B	C	'competent '	
10	F	23	Regular	Economy	'Electrom	IV	Awasa	Governme	2009	2.71	66	B	C	'competent '	
11	M	25	Regular	Economy	'Construct	IV	Awasa	Governme	2012	2.14	71.5	C	C	'competent '	
12	M	18	Regular	'Industry '	'Textile te	II	Awasa	Governme	2010	2.4	59.2	A	C	'competent '	
13	M	24	Extension	Economy	Automoti	IV	Awasa	Governme	2009	1.6	56	D	C	'competent '	
14	M	21	Regular	Economy	'Road Con	IV	Awasa	Governme	2012	2.751282	67.41617	B	C	'competent '	
15	M	21	Regular	'Industry '	'Metal ma	II	Awasa	Governme	2011	1.6	54	D	C	'competent '	
16	F	18	Regular	'Industry '	'Metal ma	II	Awasa	Governme	2010	1.8	56	C	D	'competent '	
17	M	18	Regular	Economy	DBA	III	Wendo	Governme	2013	2.57	74.83	A	C	'competent '	
18	M	21	Regular	'Industry '	'Metal ma	II	Awasa	'Private '	2009	1.8	50	C	D	'competent '	
19	M	19	Regular	'Hotel and	'hotel ope	II	Awasa	Governme	2013	2.57	82.06	C	D	'competent '	
20	M	20	Regular	'Industry '	'Textile te	II	Awasa	Governme	2010	2.8	67.4	C	A	'competent '	
21	M	19	Regular	Economy	'Electro M	IV	Awasa	Governme	2012	2.733243	67.74546	B	C	'competent '	

Appendix II: Sample rules generated from PART Un-pruned 10-fold cross validation

```
Classifier Model
PART decision list
------------------

Highschool = Private  AND
Entry Year <= 2009 AND
Sex = M AND
Mathematics  = B: competent  (162.0)

Highschool = Private  AND
Entry Year <= 2009 AND
Sex = M AND
Trade/ Occupation = Textile technology: competent  (66.0)

Highschool = Private  AND
Entry Year <= 2009 AND
Sex = M AND
Trade/ Occupation = Road Construction : competent  (54.0)

Entry Year <= 2009 AND
Highschool = Government AND
Level = V AND
English = B: Not Yet Competant (26.0)

Entry Year <= 2009 AND
Highschool = Government AND
Level = V AND
age <= 21: Not Yet Competant (9.0)

Entry Year <= 2009 AND
Highschool = Government AND
Level = III AND
Division = Regular AND
Sector = Economy AND
Trade/ Occupation = DBA: Not Yet Competant (277.0)

Sex = F AND
Entry Year <= 2011 AND
Highschool = Government AND
Woreda = Aleta Wendo: competent  (11.0)

Sex = F AND
Entry Year <= 2011 AND
Highschool = Government AND
Entry Year > 2010.090898 AND
EGSECE <= 2 AND
English = B: Not Yet Competant (31.0)

Sex = F AND
Entry Year <= 2011 AND
Highschool = Government AND
Entry Year > 2010.090898 AND
Mathematics  = D: Not Yet Competant (28.0)
```