Journal of Computer Science and Information Technology
March 2014, Vol. 2, No. 1, pp. 21-46
ISSN: 2334-2366 (Print), 2334-2374 (Online)

American Research Institute for Policy Development
42 Monticello Street, New York, NY 12701, USA.
Phone: 1.347.757.4901 Website: www.aripd.org/jcsit

A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm

Ogunde A. O[1]. and Ajibade D. A[1].

Abstract

The desire of every organization is to extract hidden but useful knowledge from their data through data mining tools. Also, the recent decline in the standard of education in most developing countries has necessitated researches that will help proffer solutions to some of the problems. From the literature, different analysis has been carried out on university data, which includes student's university entrance examination and Ordinary level results but the relationship between these entry results and students' final graduation grades has been in isolation. Therefore, in this work, a new system that will predict students' graduation grades based on entry results data using the Iterative Dichotomiser 3 (ID3) decision tree algorithm was developed. ID3 decision tree algorithm was used to train the data of the graduated sets. The knowledge represented by decision trees were extracted and presented in form of IF-THEN rules. The trained data were then used to develop a model for making future prediction of students' graduation grades. The developed system could be very useful in predicting students' final graduation grades even from the point of entry into the university. This will help management staff, academic planners to properly counsel students in order to improve their overall performance.

Keywords: Data mining, Decision trees, Prediction, ID3 algorithm, Knowledge extraction

1.0 Introduction

Over the years, data mining has been used for extraction of implicit, previously undetected and useful information from large amounts of data. It is often used for prediction from the knowledge pattern.

[1] Computer Science Programme, Department of Mathematical Sciences, Redeemer's University, Redemption Camp, Nigeria, E-mail: ogundea@run.edu.ng

Data Mining is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of statistics and Database Management Systems. It also detects hidden knowledge and patterns which were previously unknown from large databases for easy and fast retrieval of data and information. Different analysis has been done on university students' entrance examination and ordinary level result but the relationships between them and the final graduation grades have not received much attention from the research world. Extraction of hidden knowledge and subsequent management of these much important enrolment data could be of very great importance to the management and stakeholders of academic institutions, most especially in the area of decision making, which would in turn improve students' performance through proper guidance and counselling.

Data mining techniques has also been used in several occasions in solving educational problems and to perform crucial analysis in the educational sector. This is to enhance educational standards and management such as investigating the areas of learning recommendation systems, learning material arrangement, continuous student assessments and evaluation of educational websites. Data Mining discovers relationships among attributes in data set, producing conditional statements concerning attribute-values. Classification is one of the most commonly applied data mining technique, which uses a set of pre-classified examples to develop a model that can classify the population of records at large (Samrat and Vikesh, 2012). Therefore in this work, a data mining system, using the ID3 decision tree algorithm for classification of data, was deployed to mine students' enrolment data and the knowledge gained was used to predict possible student graduation grades.

2.0 Literature Review

This section examines the concepts and ideas in the literature relating to this research area.

2.1 Data Mining (DM)

DM refers to a particular step in the Knowledge discovery process. It consists of particular algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (models) over the data (Albashiri, 2013).

Data mining is also a process of extracting nontrivial, valid, novel and useful information from large databases. DM can be viewed as a kind of search for meaningful patterns or rules from a large search space, that is, the database (Srinivasa et al., 2007). As this knowledge is captured, it can become a very vital tool to gaining a competitive advantage over competitors in an industry. This then creates an important need for tools to analyze and model these data. According to Rao and Vidyavathi (2010), data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behaviour of their customers and potential customers. DM also refers to the process of extracting novel, useful, and understandable pieces of information (e.g., patterns, rules, regularities, constraints) from data in databases (Saravanan and Christopher, 2012).

Recently, researchers had also proposed that the future of data mining has a lot to benefit from other technologies such as cloud computing. Cloud computing combined with data mining can provide powerful capacities of storage and computing and an excellent resource management (Qureshi et al., 2013). Due to the explosive data growth and amount of computation involved in data mining, an efficient and high-performance computing is very necessary for a successful data mining application. Data mining in the cloud computing environment can be considered as the future of data mining because of the advantages of cloud computing paradigm. Cloud computing provides greater capabilities in data mining and data analytics. The major concern about data mining is that the space required by the operations and itemsets is very large. By combining data mining with cloud computing, a considerable amount of space can be saved and the cost of computational resources can be greatly reduced (Pareek and Gupta, 2012).

Data mining is the process of discovering interesting knowledge from large amount of data stored in database, data warehouse or other information repositories. Data Mining is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems (Surjeet & Saurabh, 2012). Data mining is data analysis methodology used to identify hidden patterns in a large data set. It has been successfully used in different areas including the educational environment.

Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process (Surjeet & Saurabh, 2012).

The main functionality of data mining techniques is applying various methods and algorithms in order to discover and extract patterns of stored data. These interesting patterns are presented to the user and may be stored as new knowledge in knowledge base. Data mining has been used in areas such as database systems, data warehousing, statistics, machine learning, data visualization, and information retrieval. Data mining techniques have been introduced to new areas including neural networks, patterns recognition, spatial data analysis, image databases and many application fields such as business, economics and bioinformatics. Some types of data mining techniques are: Clustering, Association Rule Mining, Neural Networks, Genetic Algorithms, Nearest Neighbor Method, Classification Rule Mining, Decision trees and many others.

2.2 Decision Trees

A decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labelled with distinct outcomes of the test. Each leaf node has a class label associated with it. Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome (Surjeet & Saurabh, 2012).

In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorisation and generalisation of a given set of data. The four widely used decision tree learning algorithms are: ID3, CART, CHAID and C4.5.

2.2.1 ID3

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric - information gain. To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function. (Surjeet & Saurabh, 2012). A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. They are commonly used for gaining information for the purpose of decision making. Decision tree starts with a root node on which it is for users to take appropriate actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. ID3 is a simple decision tree learning algorithm developed by Ross Quinlan. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric - information gain.

ID3 algorithm is one important method in the technology of decision tree classification and so is widely applied. ID3 algorithm searches through attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the n (where n = number of possible values of an attribute) partitioned subsets to get their "best" attribute. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices. The central principle of ID3 algorithm is based on the information theory.

2.4 Related Works

These are some of the related work that describes data mining and knowledge discovery in the context of this work.

Surjeet & Saurabh (2012) used student’s past academic performance to create a model using ID3 decision tree algorithm for prediction of student’s enrolment in MCA course. Shreenath & Madhu (2012) developed a model for the placement database, so that institutions can use it to discover some interesting patterns that could be analyzed to plan their future activities. The work used Apriori Algorithm to discover student academic behaviours. Al-Radaideh et al. (2006) applied a decision tree model to predict the final grade of students who studied the C++course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5 and the Naïve Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models.

Hijazi & Naqvi (2006) conducted a study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of colleges affiliated to Punjab University of Pakistan. The hypothesis that was stated as “Students’ attitude towards attendance in class, hours spent in study on daily basis after college, students family income, students mother's age and mother's education are significantly related with student performance” was framed. By means of simple linear regression analysis, it was found that the factors like mother’s education and student’s family income were highly correlated with the student academic performance.

Khan (2005) conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Bray (2007) in his study on private tutoring and its implications, observed that the percentage of students receiving private tutoring in India was relatively higher than in Malaysia, Singapore, Japan, China and Sri Lanka.

It was also observed that there was an enhancement of academic performance with the intensity of private tutoring and this variation of intensity of private tutoring depends on the collective factor namely socio-economic conditions. Oladipupo & Oyelade (2009) analysed a data using undergraduate students' result in the department of Computer Science from a university in Nigeria. The department offers two programmes; Computer Science and Management Information Science. A total number of 30 courses for 100 level and 200 level students were considered as a case study. The analysis revealed that there is more to students' failure than the students' ability. It also reveals some hidden patterns of students' failed courses which could serve as bedrock for academic planners in making academic decisions and an aid in the curriculum re-structuring and modification with a view to improving students' performance and reducing failure rate.

Kovacic (2010) presented a case study on educational data mining to identify up to what extent the enrolment data can be used to predict students' success. The algorithms CHAID and CART were applied on student enrolment data of information system students of Open Polytechnic of New Zealand to get two decision trees classifying successful and unsuccessful students. The accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively. Bharadwaj & Pal (2011) conducted study on the student performance based by selecting 300 students from 5 different degree college conducting BCA (Bachelor of Computer Application) course of Dr. R.M.L. Awadh University, Faizabad, India. By means of Bayesian classification method on 17 attributes, it was found that the factors like students' grade in senior secondary exam, living location, medium of teaching, mothers' qualification, students other habit, family annual income and students' family status were highly correlated with the student academic performance.

Yadav et al. (2011) obtained the university students data like attendance, class test, seminar and assignment marks from the students' database, to predict the performance at the end of the semester using three algorithms ID3, C4.5 and CART and showed that CART is the best algorithm for classification of data. Ramasubbareddy et al. (2011) proposed associative classification which is a classification of a new tuple using association rules.

Associative classification is a combination of association rule mining and classification. They searched for strong associations between frequent patterns and class labels. The main aim of the paper was to improve accuracy of the classifier. The accuracy could be achieved by producing all types of negative class association rules. Al´ıpio et al. (2001) studied a new technique called post-bagging, which consists in resampling parts of a classification model rather than the data. They did this with a particular kind of model: large sets of classification association rules, and in combination with ordinary best rule and weighted voting approaches. They empirically evaluated the effects of the technique in terms of classification accuracy.

3.0 Design Methodology

This section describes the design of the system, tools, software, data flow, algorithm and other research methodology.

3.1 Data Collection and Description

The data used in this paper was collected from the Academic Department of Redeemer's University, Nigeria. The data consists of sex of the students, student entry grades in secondary school which is the O'level results, entrance examination scores and the grade obtained at graduation (B.Sc) for all graduates between 2008/2009 and 2011/2012 sessions. The data and the attributes that possibly influenced their results were selected and analysed. The grades spread between 0 – 100% taking values from A1 – F9 as detailed in Table 1.

Table 1: Student Related Variables

Variable	Description	Possible Values
GMSS	Mathematics Grade in Secondary School	{A1- 75% above, B2 – 70% - 74%, B3 – 65% - 69%, C4 - 60% - 64%, C5 – 55% - 59%, C6 – 50% - 54%, D7 – 44% - 49%, E8 – 40% - 44%, F9 - < 40%}
GESS	English Grade in Secondary School	{A1- 75% above, B2 – 70% - 74%, B3 – 65% - 69%, C4 - 60% - 64%, C5 – 55% - 59%, C6 – 50% - 54%, D7 – 44% - 49%, E8 – 40% - 44%, F9 - < 40%}
GPSS	Physics Grade in Secondary School	{A1- 75% above, B2 – 70% - 74%, B3 – 65% - 69%, C4 - 60% - 64%, C5 – 55% - 59%, C6 – 50% - 54%, D7 – 44% - 49%, E8 – 40% - 44%, F9 - < 40%}
GCSS	Chemistry Grade in Secondary School	{A1- 75% above, B2 – 70% - 74%, B3 – 65% - 69%, C4 - 60% - 64%, C5 – 55% - 59%, C6 – 50% - 54%, D7 – 44% - 49%, E8 – 40% - 44%, F9 - < 40%}
GBSS	Biology Grade in Secondary School	{A1- 75% above, B2 – 70% - 74%, B3 – 65% - 69%, C4 - 60% - 64%, C5 – 55% - 59%, C6 – 50% - 54%,

		D7 – 44% - 49%, E8 – 40% - 44%, F9 - < 40%}
EES	Entrance Examination Score	{Excellent, Very Good, Good, Fair}
GOG	Grade Obtained in Graduation (B.Sc)	{First Class, Second Class Upper, Second Class lower, Third Class, Pass}

The domain values for some of the variables were defined for the present investigation as follows:

- GMSS – Mathematics Grade in Senior Secondary education. Students register for eight subjects in O'level examination each carry 100 marks. Grades are assigned to all students using following mapping: A1- 75% above, B2 – 70% - 74%, B3 – 65% - 69%, C4 - 60% - 64%, C5 – 55% - 59%, C6 – 50% - 54%, D7 – 44% - 49%, E8 – 40% - 44% and F9 - < 40%.
- GESS – English Grade in Senior Secondary education. Students register for eight subjects in O'level examination each carry 100 marks. Grades are assigned to all students using following mapping: A1- 75% above, B2 – 70% - 74%, B3 – 65% - 69%, C4 - 60% - 64%, C5 – 55% - 59%, C6 – 50% - 54%, D7 – 44% - 49%, E8 – 40% - 44% and F9 - < 40%.
- GPSS – Physics Grade in Senior Secondary education. Students register for eight subjects in O'level examination each carry 100 marks. Grades are assigned to all students using following mapping: A1- 75% above, B2 – 70% - 74%, B3 – 65% - 69%, C4 - 60% - 64%, C5 – 55% - 59%, C6 – 50% - 54%, D7 – 44% - 49%, E8 – 40% - 44% and F9 - < 40%.
- GCSS – Chemistry Grade in Senior Secondary education. Students register for eight subjects in O'level examination each carry 100 marks. Grades are assigned to all students using following mapping: A1- 75% above, B2 – 70% - 74%, B3 – 65% - 69%, C4 - 60% - 64%, C5 – 55% - 59%, C6 – 50% - 54%, D7 – 44% - 49%, E8 – 40% - 44% and F9 - < 40%.
- GBSS – Biology Grade in Senior Secondary education. Students register for eight subjects in O'level examination each carry 100 marks. Grades are assigned to all students using following mapping: A1- 75% above, B2 – 70% - 74%, B3 – 65% -

69%, C4 - 60% - 64%, C5 – 55% - 59%, C6 – 50% - 54%, D7 – 44% - 49%, E8 – 40% - 44% and F9 - < 40%.

- EES – Entrance Examination Score. This examination is conducted by the university at the entry level. Grades are assigned as follows: Excellent – 70% above, Very Good – 60% - 69%, Good – 50% - 59% and Fair - 45% - 49%.
- GOG – Grades obtained in Graduation. This refers to the final graduation grades. It also consists of five classes which include: First Class, Second Class Upper, Second Class lower, Third Class and Pass.

3.2 The ID3 Algorithm

The ID3 algorithm is presented as follows:

```
(Examples, Target Attribute, Attributes)

• Create a root node for the tree
• If all examples are positive, Return the single-node tree Root, with label = +.
• If all examples are negative, Return the single-node tree Root, with label = -.
• If number of predicting attributes is empty, then Return the single node tree
  Root, with label = most common value of the target attribute in the examples.
• Otherwise Begin

  o A = The Attribute that best classifies examples.
  o Decision Tree attribute for Root = A.
  o For each possible value, vi, of A,

      ▪ Add a new tree branch below Root, corresponding to the test A = vi.
      ▪ Let Examples(vi) be the subset of examples that have the value vi for A
      ▪ If Examples(vi) is empty
          ➢ Then below this new branch add a leaf node with label = most
            common target value in the examples
          ➢ Else below this new branch add the subtree ID3 (Examples(vi),
            Target_Attribute, Attributes – {A})
  • End
  • Return Root
```

3.3 Design of the Prediction System

The prediction system contains a knowledge base that has accumulated experience and a set of rules for applying the knowledge base to each particular situation. It is the knowledge-base that incorporates the entrance examination score with the grades in secondary school (O'level result) and their relationships with the grades obtained in graduation. Moreover, it provides advice and guidance regarding the selection of programs. The architecture for the prediction system is provided in figure 1.

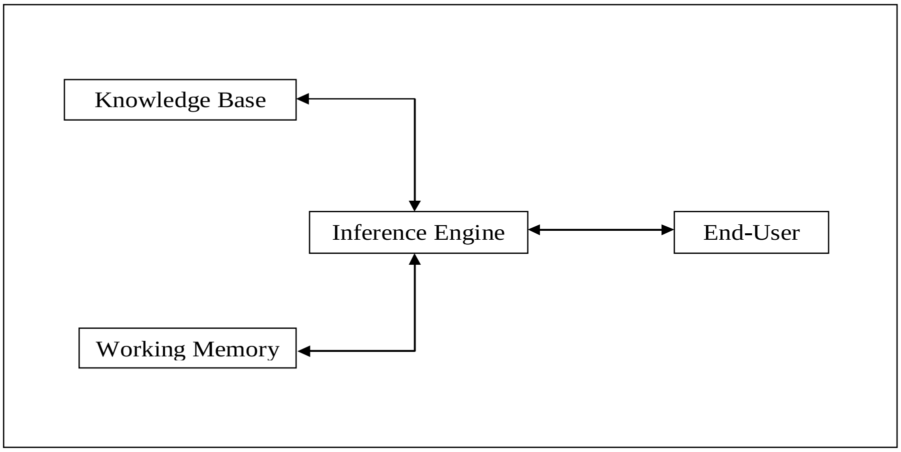


Figure 1: Architecture of the prediction system

Knowledge Base: this contains the domain knowledge and they are declarative representation often in IF-THEN rules. These rules will be generated by Waikato Environment for Knowledge Analysis (WEKA) implementation tool which represents the knowledge base of the system.

Working Memory: the user enters information on a current problem into the working memory. The system matches this information with knowledge contained in the knowledge base to infer new facts. The system then enters these new facts into the working memory and the matching process continues. Eventually the system reaches some conclusion that it also enters into the working process.

Inference Engine: it works with the information in the knowledge base and the working memory. It searches the rules for a match between their premises and information contained in the working memory. When it finds a match, it adds the conclusion to the working memory and also looks for new rules.

End-User: the individual who will be consulting with the system to get advice that would have been provided by the prediction system.

3.3.1 Data Training

The system flowchart for training the data is shown in figure 2 while the overall system flowchart is displayed in figure 3. In figure 3, the data is selected and converted into Attribute Relation File Format (arff) format using the arff converter and then classified using WEKA and the decision tree is produced (Sunita & Lobo, 2011). In figure 3, the IF-THEN rules from the decision tree is inputted into the knowledge-base and the user then interacts through the user interface and then asks related question which are used to match the rules in the knowledge base.

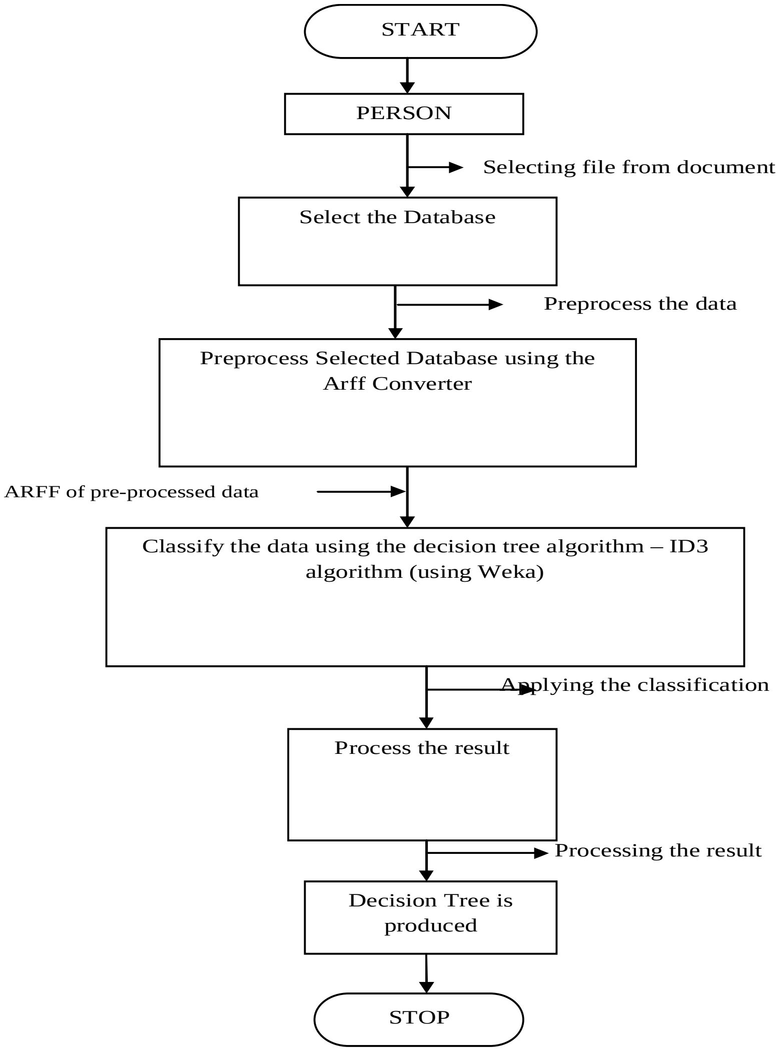

START
PERSON
Selecting file from document
Select the Database
Preprocess the data
Preprocess Selected Database using the Arff Converter
ARFF of pre-processed data
Classify the data using the decision tree algorithm – ID3 algorithm (using Weka)
Applying the classification
Process the result
Processing the result
Decision Tree is produced
STOP

Figure 2: System flowchart for training of the data

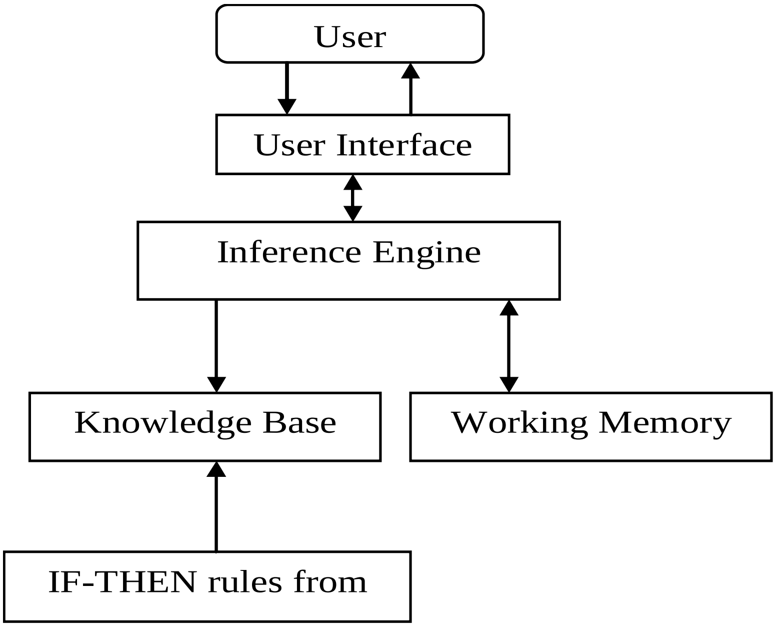
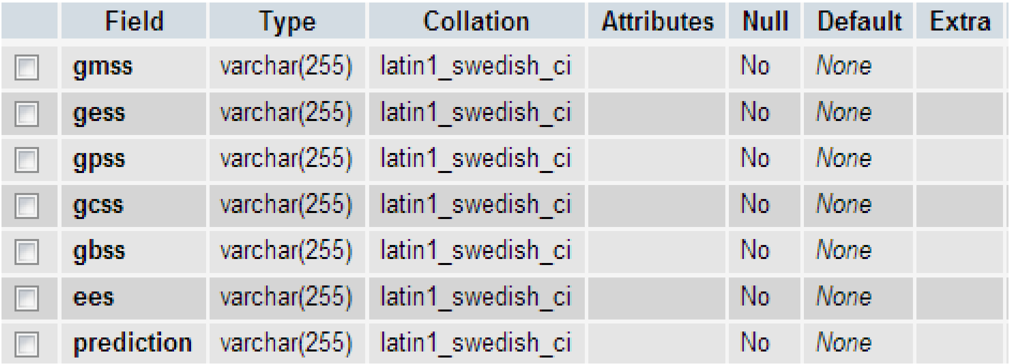


Figure 3: System flowchart

3.4 Database Design

This work made use of a single table called rules. The table contains the rules that serve as the knowledge base. The grades of a student are matched with the rules in the knowledge base to predict the student result. Table 2 shows the various field names, the data type with which they are stored and the size of each data type. The gmss is a field that holds the Grade in Mathematics in Secondary School, gess is a field that holds the Grade in English in Secondary School, gpss is a field that holds the Grade in Physics in Secondary School, gcss is a field that holds the Grade in Chemistry in Secondary School, gbss is a field that holds the Grade in Biology in Secondary School. The ees field also hold the information about Entrance Examination Score and the prediction field hold information about the final result.

Table 2: Rules Table for the Knowledge Base



	Field	Type	Collation	Attributes	Null	Default	Extra
□	gmss	varchar(255)	latin1_swedish_ci		No	*None*	
□	gess	varchar(255)	latin1_swedish_ci		No	*None*	
□	gpss	varchar(255)	latin1_swedish_ci		No	*None*	
□	gcss	varchar(255)	latin1_swedish_ci		No	*None*	
□	gbss	varchar(255)	latin1_swedish_ci		No	*None*	
□	ees	varchar(255)	latin1_swedish_ci		No	*None*	
□	prediction	varchar(255)	latin1_swedish_ci		No	*None*	

4.0 Implementation and Results

This chapter discusses the implementation of the ID3 decision tree algorithm in a special software developed for predicting students' graduation grades.

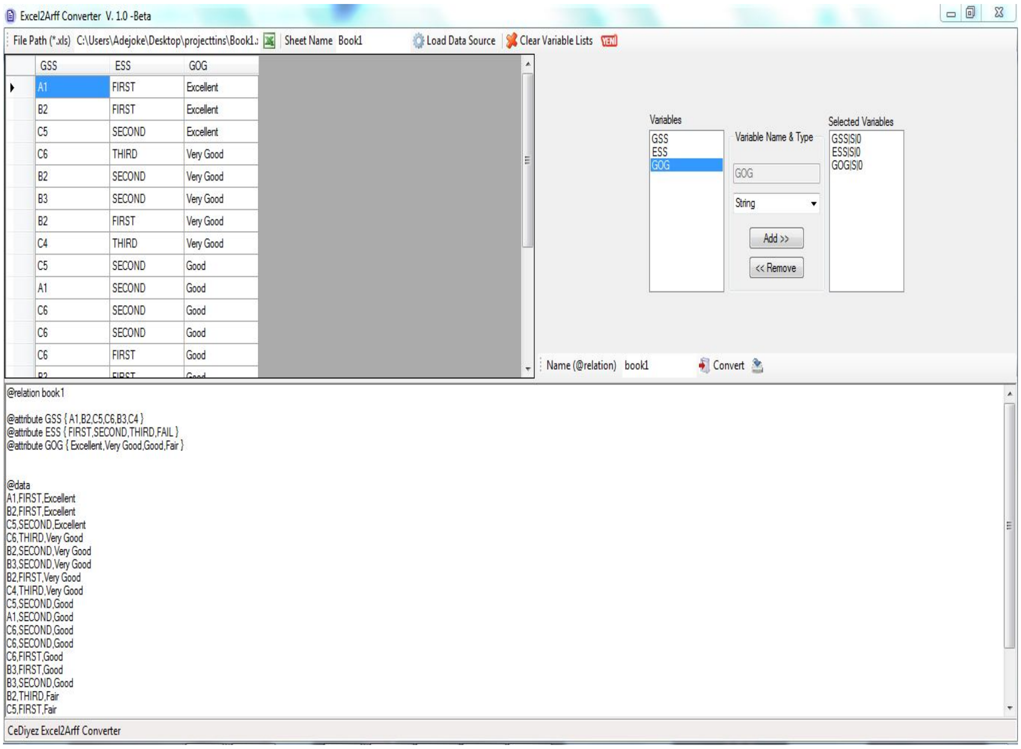
4.1 System Implementation

Implementation of the system was done with the following tools: JDK (Java Development Kit), NetBeans IDE (Integrated Development Environment) 7.0, WEKA Explorer, which 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. The main strength is that it is freely available under the GNU General Public License; it is portable & platform independent because it is fully implemented in the Java programming language and thus runs on almost any modern computing platform. Weka has several standard data mining tasks, data preprocessing, clustering, classification, association, visualization, and feature selection. There are 16 decision tree algorithms like ID3, J48, Simple CART etc. that are implemented in WEKA. The algorithm used for classification is ID3. The WEKA GUI chooser launches the WEKA's graphical environment which has six buttons: Simple CLI, Explorer, Experimenter, Knowledge Flow, ARFFViewer, & Log. The version used for the implementation was Weka 3.6.9. Others are MySQL for the system backend, ARFF Converter, which was used to transform Excel file into ARFF file. It is the external representation of an instances class.

An ARFF file consists of two distinct sections: the Header section defines attribute name, type and relations, start with a keyword @Relation <data-name> @attribute <attribute-name> <type> or {range} and the Data section lists the data records, starts with @Data list of data instances. Any line start with % is the comments. ARFF Converter 1.0 was used for the implementation.

4.2 Model Construction for ID3 Decision Tree

The raw data was selected and pre-processed by the ARFF converter (Figure 4). The ID3 decision tree model for the system was then generated from the main.arff.



Sheet Name Book1

Name (@relation) book1

GSS	ESS	GOG
A1	FIRST	Excellent
B2	FIRST	Excellent
C5	SECOND	Excellent
C6	THIRD	Very Good
B2	SECOND	Very Good
B3	SECOND	Very Good
B2	FIRST	Very Good
C4	THIRD	Very Good
C5	SECOND	Good
A1	SECOND	Good
C6	SECOND	Good
C6	SECOND	Good
C6	FIRST	Good
B3	FIRST	Good

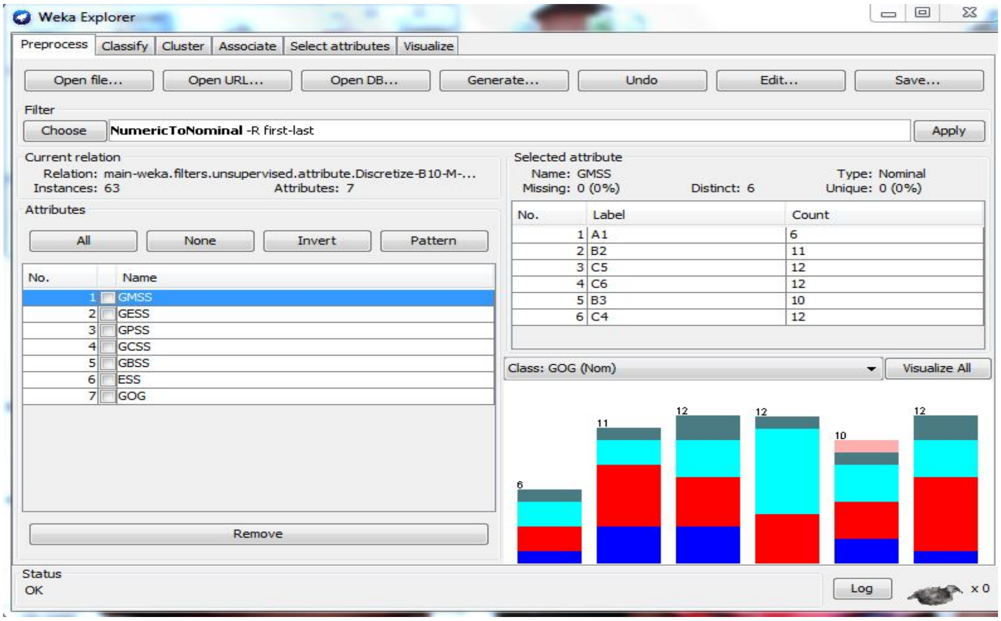
```
@relation book1

@attribute GSS { A1,B2,C5,C6,B3,C4 }
@attribute ESS { FIRST,SECOND,THIRD,FAIL }
@attribute GOG { Excellent,Very Good,Good,Fair }


@data
A1,FIRST,Excellent
B2,FIRST,Excellent
C5,SECOND,Excellent
C6,THIRD,Very Good
B2,SECOND,Very Good
B3,SECOND,Very Good
B2,FIRST,Very Good
C4,THIRD,Very Good
C5,SECOND,Good
A1,SECOND,Good
C6,SECOND,Good
C6,SECOND,Good
C6,FIRST,Good
B3,FIRST,Good
B3,SECOND,Good
B2,THIRD,Fair
C5,FIRST,Fair
```

Figure 4: ARFF Converter

The ARFF pre-processed data was then trained by the WEKA implementation tool (figure 5).



Filter

NumericToNominal -R first-last

Current relation

Relation: main-weka.filters.unsupervised.attribute.Discretize-B10-M-...

Instances: 63 Attributes: 7

Selected attribute

Name: GMSS Type: Nominal

Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

No.	Label	Count
1	A1	6
2	B2	11
3	C5	12
4	C6	12
5	B3	10
6	C4	12

Attributes

No.	Name
1	□ GMSS
2	□ GESS
3	□ GPSS
4	□ GCSS
5	□ GBSS
6	□ ESS
7	□ GOG

Class: GOG (Nom)

Status

OK

Figure 5: Weka Interface for the pre-processed data

The data is classified using the ID3 algorithm under the classify panel in WEKA and the model visualization is shown in figure 6.

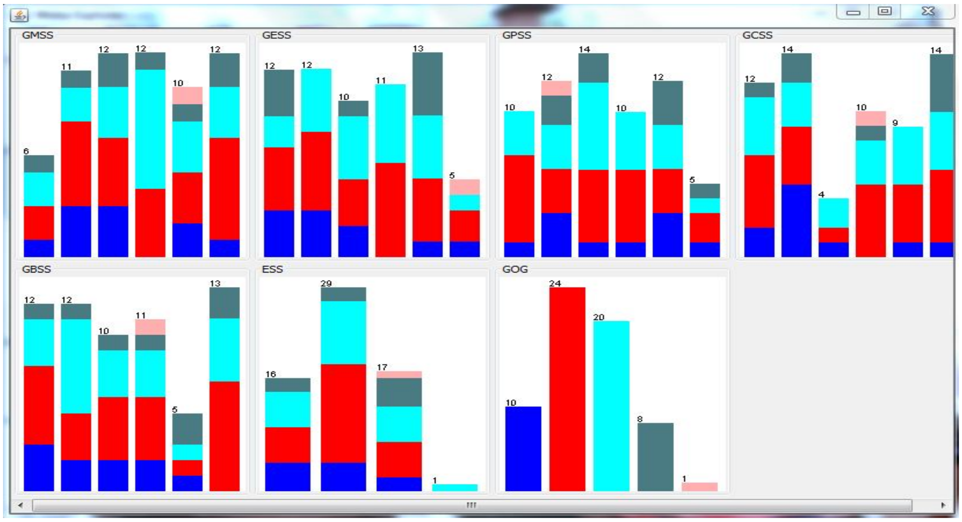
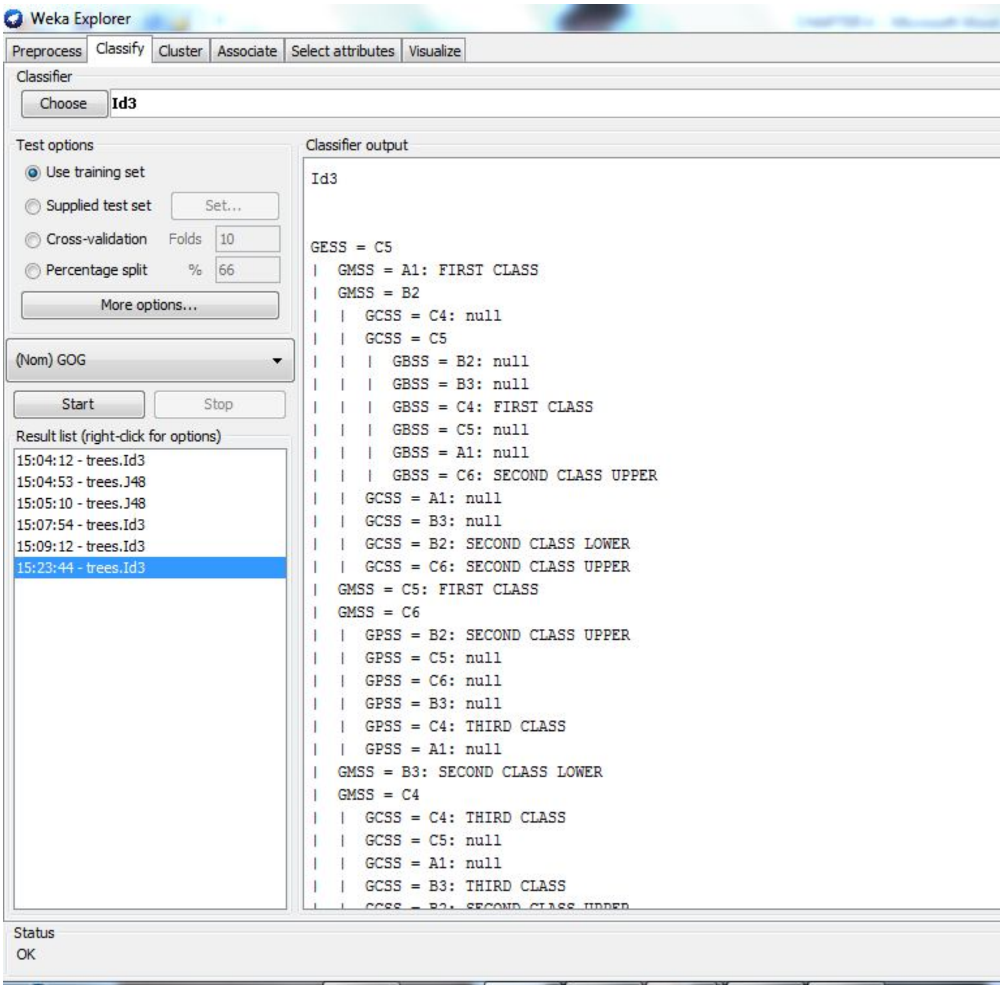


Figure 6: Model Visualization

The set of rules were generated and the results of the classifier using WEKA are shown in figures 7 and 8 respectively.



Classifier

Choose **Id3**

Test options

● Use training set

○ Supplied test set

○ Cross-validation Folds 10

○ Percentage split % 66

(Nom) GOG

Result list (right-click for options)

15:04:12 - trees.Id3
15:04:53 - trees.J48
15:05:10 - trees.J48
15:07:54 - trees.Id3
15:09:12 - trees.Id3
15:23:44 - trees.Id3

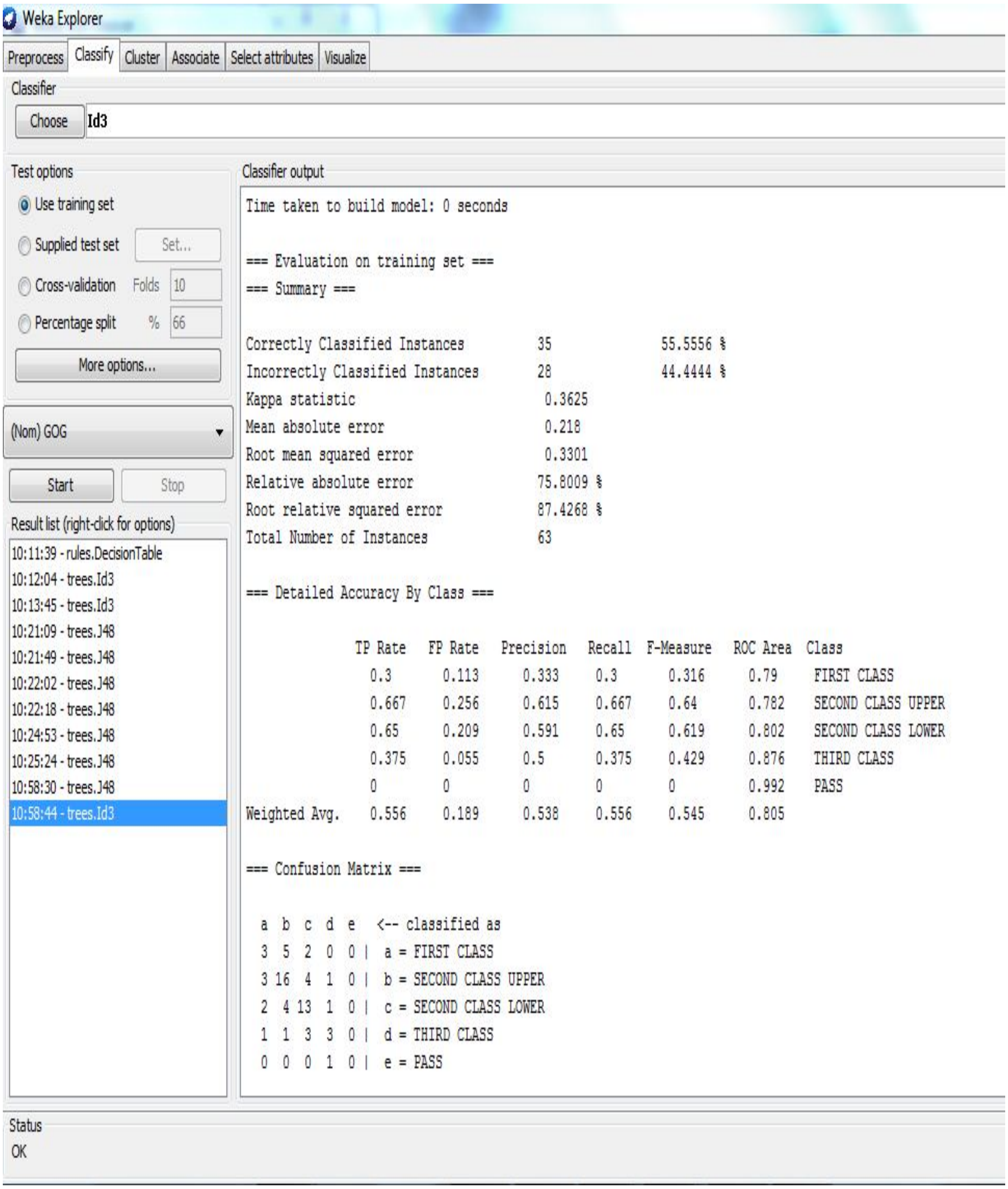
Classifier output

```
Id3


GESS = C5
|  GMSS = A1: FIRST CLASS
|  GMSS = B2
|  |  GCSS = C4: null
|  |  GCSS = C5
|  |  |  GBSS = B2: null
|  |  |  GBSS = B3: null
|  |  |  GBSS = C4: FIRST CLASS
|  |  |  GBSS = C5: null
|  |  |  GBSS = A1: null
|  |  |  GBSS = C6: SECOND CLASS UPPER
|  |  GCSS = A1: null
|  |  GCSS = B3: null
|  |  GCSS = B2: SECOND CLASS LOWER
|  |  GCSS = C6: SECOND CLASS UPPER
|  GMSS = C5: FIRST CLASS
|  GMSS = C6
|  |  GPSS = B2: SECOND CLASS UPPER
|  |  GPSS = C5: null
|  |  GPSS = C6: null
|  |  GPSS = B3: null
|  |  GPSS = C4: THIRD CLASS
|  |  GPSS = A1: null
|  GMSS = B3: SECOND CLASS LOWER
|  GMSS = C4
|  |  GCSS = C4: THIRD CLASS
|  |  GCSS = C5: null
|  |  GCSS = A1: null
|  |  GCSS = B3: THIRD CLASS
|  |  GCSS = B2: SECOND CLASS UPPER
```

Status
OK

Figure 7: ID3 rules generated using WEKA



Classifier: Id3

Test options
- ● Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) GOG

Result list (right-click for options)
- 10:11:39 - rules.DecisionTable
- 10:12:04 - trees.Id3
- 10:13:45 - trees.Id3
- 10:21:09 - trees.J48
- 10:21:49 - trees.J48
- 10:22:02 - trees.J48
- 10:22:18 - trees.J48
- 10:24:53 - trees.J48
- 10:25:24 - trees.J48
- 10:58:30 - trees.J48
- 10:58:44 - trees.Id3

Classifier output

```
Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          35               55.5556 %
Incorrectly Classified Instances        28               44.4444 %
Kappa statistic                          0.3625
Mean absolute error                      0.218
Root mean squared error                  0.3301
Relative absolute error                 75.8009 %
Root relative squared error             87.4268 %
Total Number of Instances               63

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.3       0.113     0.333      0.3       0.316      0.79     FIRST CLASS
                 0.667     0.256     0.615      0.667     0.64       0.782    SECOND CLASS UPPER
                 0.65      0.209     0.591      0.65      0.619      0.802    SECOND CLASS LOWER
                 0.375     0.055     0.5        0.375     0.429      0.876    THIRD CLASS
                 0         0         0          0         0          0.992    PASS
Weighted Avg.    0.556     0.189     0.538      0.556     0.545      0.805

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  3  5  2  0  0 |  a = FIRST CLASS
  3 16  4  1  0 |  b = SECOND CLASS UPPER
  2  4 13  1  0 |  c = SECOND CLASS LOWER
  1  1  3  3  0 |  d = THIRD CLASS
  0  0  0  1  0 |  e = PASS
```

Status
OK

Figure 8: Result of the Classifier

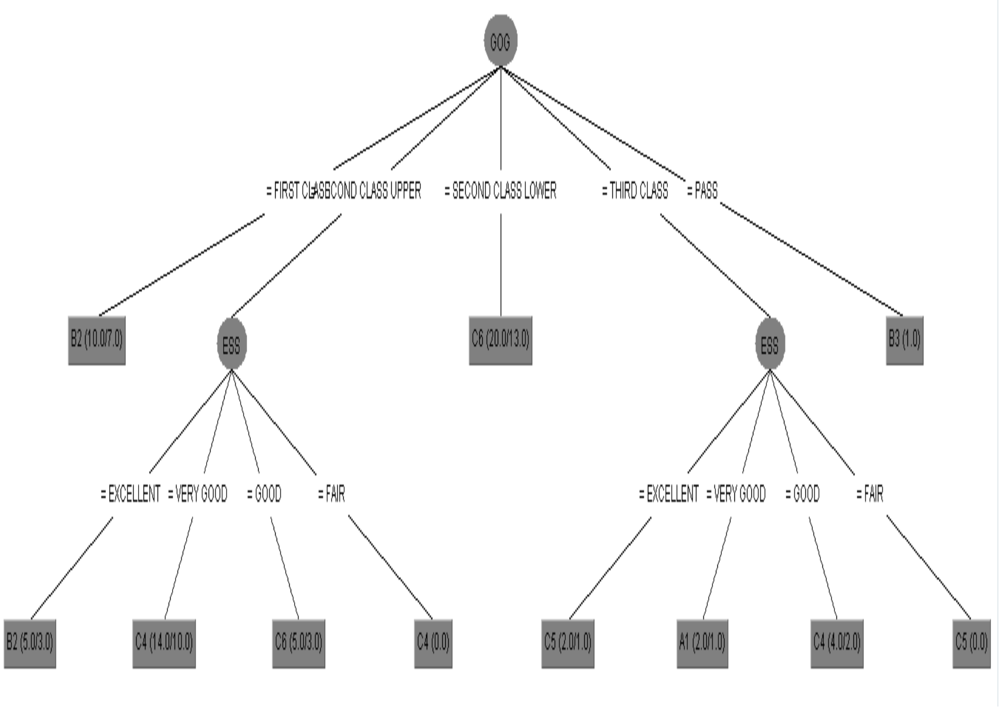


Figure 9: Decision Tree produced using WEKA

4.3 Results Analysis

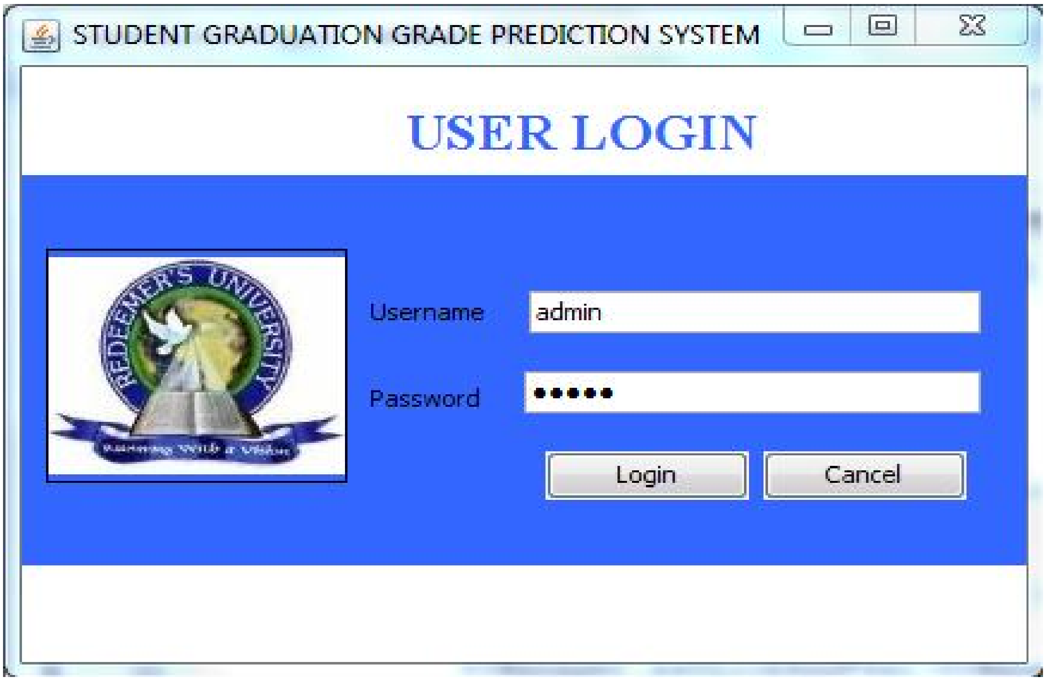
The knowledge represented by the decision tree was extracted and represented in the form of IF-THEN rules as displayed in Table 3.

Table 3: Rule Set generated by ID3

IF GMSS = ‘A1’ and GESS = ‘C5’ and GPSS = ‘A1’ and GCSS =’B3’ and GBSS = ‘C4’ and EES = ‘Excellent’, THEN GOG = ‘FIRST CLASS’
IF GMSS = ‘A1’ and GESS = ‘B3’ and GPSS = ‘A1’ and GCSS =’A1’ and GBSS = ‘C4’ and EES = ‘Very Good’, THEN GOG = ‘FIRST CLASS’
IF GMSS = ‘A1’ and GESS = ‘A1’ and GPSS = ‘A1’ and GCSS =’B3’ and GBSS = ‘C4’ and EES = ‘Good’, THEN GOG = ‘FIRST CLASS’
IF GMSS = ‘B2’ and GESS = ‘B2’ and GPSS = ‘B2’ and GCSS =’B3’ and GBSS = ‘B2’ and EES = ‘Very Good’, THEN GOG = ‘SECOND CLASS UPPER’
IF GMSS = ‘B2’ and GESS = ‘C4’ and GPSS = ‘A1’ and GCSS =’B3’ and GBSS = ‘C4’ and EES = ‘Excellent’, THEN GOG = ‘SECOND CLASS UPPER’
IF GMSS = ‘B2’ and GESS = ‘C5’ and GPSS = ‘C4’ and GCSS =’C6’ and GBSS = ‘C4’ and EES = ‘Good’, THEN GOG = ‘SECOND CLASS LOWER’
IF GMSS = ‘B3’ and GESS = ‘B2’ and GPSS = ‘A1’ and GCSS =’C4’ and GBSS =‘C4’ and EES = ‘Excellent’, THEN GOG = ‘SECOND CLASS UPPER’
IF GMSS = ‘B3’ and GESS = ‘C5’ and GPSS = ‘B2’ and GCSS =’C4’ and GBSS = ‘C4’ and EES = ‘Very Good’, THEN GOG = ‘SECOND CLASS UPPER’
IF GMSS = ‘B3’ and GESS = ‘C5’ and GPSS = ‘C6’ and GCSS =’C4’ and GBSS = ‘C6’ and EES = ‘Good’, THEN GOG = ‘SECOND CLASS LOWER’
IF GMSS = ‘C4’ and GESS = ‘C5’ and GPSS = ‘B3’ and GCSS =’C4’ and GBSS = ‘C4’ and EES = ‘Excellent’, THEN GOG = ‘SECOND CLASS UPPER’
IF GMSS = ‘C4’ and GESS = ‘B3’ and GPSS = ‘C5’ and GCSS =’B2’ and GBSS = ‘C4’ and EES = ‘Very Good’, THEN GOG = ‘SECOND CLASS UPPER’
IF GMSS = ‘C4’ and GESS = ‘C5’ and GPSS = ‘B2’ and GCSS =’C4’ and GBSS = ‘C4’ and EES = ‘Good’, THEN GOG = ‘SECOND CLASS LOWER’
IF GMSS = ‘C4’ and GESS = ‘C5’ and GPSS = ‘C5’ and GCSS =’C4’ and GBSS = ‘C4’ and EES = ‘Fair’, THEN GOG = ‘SECOND CLASS LOWER’
IF GMSS = ‘C5’ and GESS = ‘A1’ and GPSS = ‘B2’ and GCSS =’B3’ and GBSS = ‘C4’ and EES = ‘Excellent’, THEN GOG = ‘FIRST CLASS’
IF GMSS = ‘C5’ and GESS = ‘B2’ and GPSS = ‘B2’ and GCSS =’C4’ and GBSS = ‘C4’ and EES = ‘Very Good’, THEN GOG = ‘SECOND CLASS UPPER’
IF GMSS = ‘C5’ and GESS = ‘C6’ and GPSS = ‘C6’ and GCSS =’C6’ and GBSS = ‘C5’ and EES = ‘Good’, THEN GOG = ‘THIRD CLASS’
IF GMSS = ‘C6’ and GESS = ‘C5’ and GPSS = ‘B3’ and GCSS =’C4’ and GBSS = ‘C4’ and EES = ‘Excellent’, THEN GOG = ‘SECOND CLASS UPPER’
IF GMSS = ‘C6’ and GESS = ‘C5’ and GPSS = ‘B3’ and GCSS =’C4’ and GBSS = ‘C4’ and EES = ‘Very Good’, THEN GOG = ‘SECOND CLASS LOWER’
IF GMSS = ‘C6’ and GESS = ‘C6’ and GPSS = ‘C6’ and GCSS =’C6’ and GBSS = ‘C5’ and EES = ‘Fair’, THEN GOG = ‘PASS’

4.3.1 Login Page

The login page serves as an introductory page to the user. It gives access to the prediction system. The username and the password are security check to grant access to the system (Figure 10).



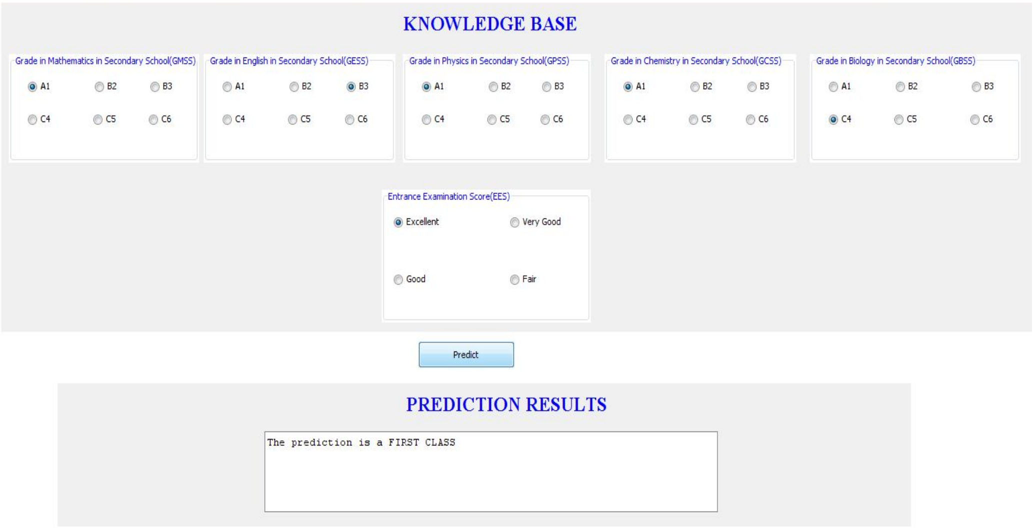
Username admin

Password •••••

Figure 10: Login Page

4.5 Sample Output

The system asks the user questions related to his grade in secondary school and the entrance examination score, and based on answers the system able to predict the grades for user. This enables the system to classify the categories of student's performance in their academic qualifications. The sample output for the prediction system is shown below in Figure 11.



KNOWLEDGE BASE

Grade in Mathematics in Secondary School(GMSS): ● A1 ○ B2 ○ B3 ○ C4 ○ C5 ○ C6

Grade in English in Secondary School(GESS): ○ A1 ○ B2 ● B3 ○ C4 ○ C5 ○ C6

Grade in Physics in Secondary School(GPSS): ● A1 ○ B2 ○ B3 ○ C4 ○ C5 ○ C6

Grade in Chemistry in Secondary School(GCSS): ● A1 ○ B2 ○ B3 ○ C4 ○ C5 ○ C6

Grade in Biology in Secondary School(GBSS): ○ A1 ○ B2 ○ B3 ● C4 ○ C5 ○ C6

Entrance Examination Score(EES): ● Excellent ○ Very Good ○ Good ○ Fair

PREDICTION RESULTS

```
The prediction is a FIRST CLASS
```

Figure 11: Sample Output for the Prediction System

4.5.1 Class-wise Accuracy and Accuracy model for Class Prediction

The class-wise accuracy is being presented in Table 4, which compared the true positive and false positive classifications. In addition, the correct precision for the five-class outcome categories was presented.

Table 4: Class-wise accuracy for five class prediction

GOG CLASS	True Positive (TP)	False Positive (FP)	Correct Precision (%)
First Class	0.3	0.113	30%
Second Class Upper	0.667	0.256	66.7%
Second Class Lower	0.65	0.209	65%
Third Class	0.375	0.055	37.5%
Pass	0	0	0%

The accuracy of the model is 79.556 %. The Table 5 shows the accuracy percentage for ID3 decision tree algorithm for classification of instances applied on the above data sets and is observed as follows:

Table 5: Accuracy Percentages

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances
ID3	79.556%	20.444%

5.0 Conclusion and Future Works

A study was conducted on the prediction of students' graduation grades using the ID3 decision tree algorithm with data such as the grade in secondary school, entrance examination score and the grade at graduation, which were trained to generate a model for predicting students' graduation grades. From the class wise accuracy, it is clear that the true positive rate for obtaining the First Class, Second Class Upper, Second Class Lower, Third Class and Pass is 30%, 66.7%, 65%, 37.5% and 0% respectively. It could be concluded from the study that the developed system will be very useful in any academic institution for the prediction of students' final graduation grades. This will also help management staff, academic planners and level/course advisers and even parents to properly counsel the students, most especially the academically weak ones, in order to improve their performance. Academically sound students could also be advised based on the results of this prediction most especially when they begin to fall below their expected grades at any point in time. The system will generally help students to benchmark their graduation grades from their entry point into the university, thereby helping them to work harder in order to achieve this. Finally, the developed system would help to significantly reduce the overall failure rate in most academic institutions as students can be well guided and counseled. The future work would include applying data mining techniques on an expanded data set with more distinctive attributes to get more accurate results. Also, a comparative analysis of these results would be carried out based on other experiments results gotten from using other types of decision tree algorithms such as C4.5, CHAID and CART.

References

Albashiri K. A. (2013). Data Partitioning and Association Rule Mining Using a Multi-Agent System. International Journal of Engineering Science and Innovative Technology (IJESIT), Volume 2, Issue 5, Pg 161-169.

Al-Radaideh, Q. A., Al-Shawakfa, E. W., & Al-Najjar M. I. (2006). Mining student data using decision trees. *International Arab Conference on Information Technology (ACIT'2006),* Yarmouk University, Jordan.

Al´ıpio, M., Paulo, J., Azevedo. (2001). An experiment with association rules and classification: post-bagging and conviction. Supported by the POSI/SRI/39630/2001/ Class Project.

Bharadwaj, B. & Pal, S. (2011). Data Mining: A prediction for performance improvement using classification. International *Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4,* pp. 136-140.

Bray, M. (2007). The shadow education system: private tutoring and its implications for planners. *(2nd ed.),* UNESCO, PARIS, France.

Hijazi, S., & Naqvi, R. (2006). Factors affecting students' performance: A Case of Private Colleges. *Bangladesh e-Journal of Sociology, Vol. 3, No. 1*.

Khan, Z. (2005). Scholastic achievement of higher secondary students in science stream. *Journal of Social Sciences, Vol. 1, No. 2*, pp. 84-87.

Kovacic, Z. (2010). Early prediction of student success: Mining student enrollment data. *Proceedings of Informing Science & IT Education Conference*.

Oladipupo, O. O., & Oyelade, O. J. (2008). Knowledge Discovery from Student's Repository: Association rule mining Approach. *International Journal of Computer Science & Security (IJCSS) Vol.4 Issue 2*, pp 199-206.

Pareek, A. & Gupta, M. (2012). Review of Data Mining Techniques in Cloud Computing Database. International Journal of Advanced Computer Research, Volume 2, Number 2.

Qureshi, Z., Bansal, J., & Bansal, S. (2013). A Survey on Association Rule Mining in Cloud Computing. International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 4, Pg 318-321.

Ramasubbareddy, B., Govardhan, A., & Ramamohanreddy, A. (2011). Classification Based on Positive and Negative Association Rules. *International Journal of Data Engineering (IJDE) Volume (2) Issue (2)*, pp 84.

Rao, V. S. & Vidyavathi, S. (2010). Distributed Data Mining and Mining Multi-Agent Data. International Journal on Computer Science and Engineering Vol. 02, No. 04, pp 1237-1244.

Samrat, S., & Vikesh, K. (2012). Classification of Student's data Using Data Mining Techniques for Training & Placement Department in Technical Education. *International Journal of Computer Science and Network (IJCSN), Volume 1, Issue 4*.

Saravanan, S. & Christopher, T. (2012). A Study on Milestones of Association Rule Mining Algorithms in Large Databases. International Journal of Computer Applications, Volume 47, No.3, Pg 12-19.

Shreenath, A., & Madhu, N. (2012). Discovery of students' academic patterns using data mining techniques. *International Journal on Computer Science and Engineering (IJCSE) Vol. 4 No. 06.*

Srinivasa K.G., Venugopal K.R., Patnaik L.M.. (2007). A self-adaptive migration model genetic algorithm for data mining applications. Information Sciences Volume 177 Pg 4295–4313.

Sunita, B., & Lobo, I. (2011). Data Mining in Educational System using WEKA. *International Conference on Emerging Technology Trends (ICETT)*, pp 20-25.

Surjeet, K., & Saurabh, P. (2012). Data Mining Application in Enrolment Management: A Case Study. *International Journal Of Computer Application (IJCA) Vol.41 No.5*, pp 1-6.

Yadav, S., Bharadwaj, B., & Pal, S. (2011). Data Mining Applications: A comparative study for Predicting Students' Performance. *International Journal of Innovative Technology and Creative Engineering (IJITCE), Vol. 1, No. 12,* pp. 13-19.