Application of Data Mining for Predicting Student Performance at Haramaya University Health Science College
Demmelash Abate Garbado
Lecturer
Department of Information Systems
Wolaita Sodo University, Ethiopia

Abstract:
Nowadays the amount of data maintained in an electronic format is increasing at alarming rate. Data Mining is used to pull out meaningful information and to develop significant relationship among variables stored in large data warehouse. Mining in educational environment is called educational data mining. It is concerned with developing new methods to discover knowledge from educational database. Educational data mining is an emerging discipline with developing methods for extracting unique type of data that come from educational setting and using those methods to better understand students and the setting which they learn in as defined by the educational data mining community. Education is an essential element for the progress of country. Educational data mining provides a set of techniques, which can help the educational system to overcome these issues. Because there are large data bases in higher learning institution the researcher need to use this advantage to applying data mining on Haramaya University Health Science College to see whether it is possible to predict failure and success on the first year. So that necessary decisions would be made accordingly. This paper briefly reviewed the various data mining applications in university databases.

Key words: Data Mining, Dataset, Higher education.

I. INTRODUCTION
The amount of data collected and stored in an electronic format is increasing at alarming rate in recent years. Each and every day the human beings are using the vast data and again these data are in the different fields and different formats. It may be in the form of documents, may be graphical formats, may be the video, and may be records. There is huge volume of data but The researcher hardly able to turn them in to useful information and knowledge for managerial decision making in organizations. The database management systems efficiently manage large corpus of data and effective and efficient retrieval of particular information from a large collection whenever needed and also contributes to recent massive gathering of all sorts of information. This retrieval of data as and when needed contributes the technology of data mining. To generate information from massive collection of data, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to have powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining is the best solution for such problems. Data mining is the extraction of hidden information from large databases and it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses (Larose, D. T, 2005). Data mining tools can answer the questions that traditionally (manually) were time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. The researcher begin with an overview of data mining capabilities.

1.1. OBJECTIVE OF THE PAPER
Due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge data mining is currently in use in different disciplines. Because there are large data bases in higher learning institution the researcher need to use this advantage to applying data mining on Haramaya University Health Science College to see whether it is possible to predict failure and success on the first year. So that necessary decisions will be made accordingly. So the main objective of the paper is to show the advantage of data mining in predicting student’s success and failure of students to complete their study.

II. LITERATURE REVIEW
Data mining is the extraction of implicit, previously unknown, and potentially useful information from data (Berry, M. & Linoff, S. 2000). It is about solving problems by analyzing data already present in large quantities of data in order to discover meaningful patterns and rules. Data mining is the process of discovering interesting knowledge from large amount of data stored in database, database warehouse or other information repository (Erdogan and Timor, 2005). It is a process of searching for interesting knowledge which mean new and useful from large data or it is the process to uncover hidden trends and patterns. These trends and patterns form the basis of predictive models that enable analysts to produce new observations from existing data. Data mining can be further divided into supervised and unsupervised knowledge discovery. Unsupervised knowledge discovery is to recognize relationships in the data and supervised knowledge discovery is to explain those relationships once they have been found. Unsupervised knowledge discovery as a
2.1. THE DATA MINING TASKS
Data mining tasks can be classified as follows (Monika Goyal and RajanVohra, 2012)
• Anomaly detection (Outlier/change/deviation detection): The identification of unusual data records, that might be interesting or data errors which require further investigation.
• Association rule learning (Dependency modeling): Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes.
• Clustering: It is a task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
• Classification: It is the task of generalizing known structure to apply for new data. For example, an email program might attempt to classify an email as legitimate or spam.
• Regression: It attempts to find a function which models the data with the least error.
• Summarization: It providing a more compact representation of the data set, including visualization and report generation.

2.2. DATA MINING LIFE CYCLE
In order to systematically conduct data mining analysis, a general process is usually followed. There are some standard processes, two of which are described as follows. One (CRISP) is an industry standard process consisting of a sequence of steps that are usually involved in a data mining study. The other (SEMMA) stands for sample, explore, modify, model, assess. While each step of either approach isn’t needed in every analysis, this process provides a good coverage of the steps needed, starting with data exploration, data collection, data processing, analysis, inferences drawn, and implementation. The researcher discuss the first one for now. The life cycle of a data mining project consists of six phases (Larose, D. T., 2005). The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase.

The main phases are:
i. Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
ii. Data Understanding: It starts with an initial data collection, 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.
iii. Data Preparation: In this stage, it collects all the different data sets and construct the varieties of the activities basing on the initial raw data
iv. Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
v. Evaluation: In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.
vi. Deployment: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. There are many types of graphs used by different researchers to represent these data mining life cycles, sometimes called stages. The researcher prefer the two most expressive models or graphs as seen below.

![Figure 1&2. The data mining life cycle](image)

2.3. VISUALIZING DATA MINING MODEL
The overall idea about the data mining model is data visualization. In data mining most of the times The researcher are retrieving the data from the repositories which are very large, various in type or in the hidden form. This is the difficult task for a user. So this visualization of the data mining model helps us to provide highest levels of understanding and trust. Visualization uses graphic techniques to help people to understand and analyze data. There are several studies oriented toward visualizing different educational data such as patterns of annual, seasonal, daily and hourly user behavior on online forums. Some of such investigations are statistical graphs to analyze assignments complement, questions admitted, exam score, student tracking data to analyze student’s attendance, results on assignments and quizzes, weekly information regarding students and group’s activities. Because the user does not know beforehand what the
data mining process has discovered, it is a much important to take the output of the system and translate it into an actionable solution to a business problem. The data mining models are of two types: Predictive and Descriptive. (Shaeela Ayesha, 2010, and, Sunita B. Aher, 2011)

Predictive mining tasks perform induction on the current data in order to make predictions. The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc.

Descriptive mining tasks characterize properties of the data in a target data set. The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc.

3. DATA MINING APPLICATION IN HIGHER EDUCATIONS

Education is an essential element for the development and progress of a country. Education enables the people of a country civilized and productive. Higher learning institutions are serving the education system as an engine by producing human power.

3.1. RELATED WORKS REVIEW

Data mining is usually seen as it was only to business world but now is used almost in all disciplines. Let us see in education, As Jing Luan (2006) many of the data mining techniques used in the corporate world are transferable to higher education databases. The table, below, shows the higher education equivalents of critical business questions answered by data mining.

<table>
<thead>
<tr>
<th>Private sector questions</th>
<th>Higher education equivalents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who are my most profitable customers?</td>
<td>Which students are taking the most credit hours?</td>
</tr>
<tr>
<td>Who are my repeat Web site visitors?</td>
<td>Which students are most likely to return for more classes?</td>
</tr>
<tr>
<td>Who are my loyal customers?</td>
<td>Who are the “persistence” at my Higher education institutions/college?</td>
</tr>
<tr>
<td>Who is likely to increase his/her purchases?</td>
<td>Which alumni are likely to make larger donations?</td>
</tr>
<tr>
<td>Which customers are likely to defect to competitors?</td>
<td>What types of courses will attract more students?</td>
</tr>
</tbody>
</table>

Table.1. the higher education equivalents of critical business questions answered by data mining

One of the significant facts in higher learning institution is the explosive growth of educational data. Higher learning institutions in their lifelong teaching process have a large amount of student data stored in a database and employee’s data too. But the problem is not storing the data, rather data handling, extraction of meaningful patterns, and discovery of knowledge buried in the huge database. Deploying data mining tools is a mechanism to analyze and manage large volume of data so as to discover new patterns that are helpful for problem solving and decision making. Mining in educational environment is called Educational data mining. EDM is application of data mining which extract useful information for management decision making process, teachers as well as students (Connolly T., C. Begg et al, 1999). EDM provides a set of techniques which can help educational system to improve learning experience of students as well as increase their profits (Berry, M., &Linoff, S., 2000). EDM concerned with developing new method to discover knowledge from educational databases (Alaa el- Halees, 2009). Lack of deep and enough knowledge in higher learning institution system may prevent system management to achieve quality objectives. So that data mining methodology can help bridging this knowledge gaps in higher learning institutions. Data mining enables higher learning institutions to use their current reporting capabilities to uncover and understand hidden patterns in vast databases. These patterns are then built into data mining models and used to predict individual behavior with high accuracy. As a result of this insight, institutions are able to allocate resources and staff more effectively. Data mining is a powerful tool for academic intervention (Muluk en AlemuYehuala, 2015). The hidden patterns, associations, and anomalies that are discovered by data mining techniques from educational data can improve decision making processes in higher educational systems. This improvement can bring many advantages for the higher learning institutions which has direct consequence on the country. Some of such advantages are such as maximizing educational system efficiency, decreasing student’s drop-out rate, and increasing student’s promotion rate, increasing student’s retention rate in, increasing student’s transition rate, increasing educational improvement ratio, increasing student’s success, increasing student’s learning outcome, and reducing the employee turnover etc. In this current era The researcher are using the data mining tools for extracting the knowledge, this knowledge can be used for improving the quality of education. (NeelamadhabPadhy, 2012). EDM uses many techniques such as decision trees, neural networks, k-nearest neighbor, naïve bayes, support vector machines and many others (Shaeela Ayesha, 2010). Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for organization of syllabus, prediction regarding enrolment of students in a particular programme, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students and so on. A comprehensive literature review of various significant researches in the area of Educational Data Mining ranging from Year 2002 to 2014 can be broadly classified into five areas (PoojaThakar, Anil Mehta &Manisha, 2015):

1. Survey of papers published in Educational Data Mining.
2. Predicting Academic Performance with Pre/Post Enrollment Factors
3. Comparison of Data Mining Techniques in predicting academic performance
4. Correlation among Pre/Post Enrollment Factors and Employability
5. Other areas of Education.
4. RESEARCH METHODS AND TECHNIQUES

4.1. RESEARCH METHODS
The researcher followed the life cycle of a data mining project which consists of six phases (Larose, D. T., 2005), OR the CRoss-Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM consists of six phases intended as a cyclical process. Being cyclical model provides a chance for researchers to revise their work iteratively as discussed previously. The CRISP-DM offers a uniform framework for experience documentation and guidelines so it can be applied in different industry with different type of data.

4.2. RESEARCH TOOLS
The researcher have used Waikato Environment for Knowledge Analysis (WEKA) which is a widely used tool (Software) for data mining research. It is open source software which supports several standard data mining tasks. The researcher used ARFF (Attribute Relation File Format) format for input data sources. WEKA data mining tool has J48 implementation with tree pruning method. When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods use statistical measures to remove the least reliable branches. Here decision tree implementation with J48 available in WEKA is selected to be used for classification purpose. The nature of data used for this research can best manipulated with this classifier.

4.3. Data selection and transformation
In this step only those fields were selected which were required for data mining. A few derived variables were selected. While some of the information for the variables was extracted from the files in the record office. All the predictor and response variables which were derived from the file are given in table below for reference.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Sex of the student</td>
<td>{M, F}</td>
</tr>
<tr>
<td>Age</td>
<td>Student age</td>
<td>{Y&lt;23, G&gt;23}</td>
</tr>
<tr>
<td>Eresu</td>
<td>Entrance exam result</td>
<td>{Poor&lt;480, Average&gt;480&amp;&lt;495, Good&gt;495}</td>
</tr>
<tr>
<td>Place of birth</td>
<td>Place of birth</td>
<td>{rural, urban}</td>
</tr>
<tr>
<td>High school type</td>
<td>High school type</td>
<td>{private, governmental}</td>
</tr>
<tr>
<td>GPA</td>
<td>First semester result</td>
<td>{Poor&lt;2.75, Average&gt;2.75&amp;&lt;3.3, Good&gt;3.3}</td>
</tr>
<tr>
<td>Dpt</td>
<td>Department</td>
<td>{pub, nurs, mwif, psyn, envt, medi, labo, phar}</td>
</tr>
<tr>
<td>Class</td>
<td>Success in studying or failed</td>
<td>{succ, fail}</td>
</tr>
</tbody>
</table>

5. RESEARCH FINDINGS
The result of attribute selection in WEKA indicates the top seven determining attributes of the dataset for predicting success/failure status of a student to finish his/her study, which are: first semester result (GPA), department, entrance result, sex, high school type, place of birth (rural/urban) & age. The researcher have set testing options with two categories (using 10 folds cross validation, and percentage 80% split) to build all classification models with reduced attributes. Experiment was done on the Dataset and the following tables show the summary of results of the experiment that is done on the unbalanced and balanced dataset.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>TP Rate</th>
<th>Precision</th>
<th>ROC area</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48 without pruning</td>
<td>0.478</td>
<td>0.575</td>
<td>0.583</td>
</tr>
<tr>
<td>J48 with pruning</td>
<td>0.413</td>
<td>0.582</td>
<td>0.557</td>
</tr>
</tbody>
</table>

Table 1. Results for the accuracy of the applied classifiers on the unbalanced data set

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>TP Rate</th>
<th>Precision</th>
<th>ROC area</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48 without pruning</td>
<td>0.532</td>
<td>0.522</td>
<td>0.553</td>
</tr>
<tr>
<td>J48 with pruning</td>
<td>0.593</td>
<td>0.592</td>
<td>0.587</td>
</tr>
</tbody>
</table>

Table 2. Results for the accuracy of the applied classifiers on the balanced data set

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Number of leaves</th>
<th>Size of the tree</th>
<th>Correctly classified instances</th>
<th>Incorrectly classified instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48 without pruning</td>
<td>86</td>
<td>167</td>
<td>60.3103%</td>
<td>39.6897%</td>
</tr>
<tr>
<td>J48 with pruning</td>
<td>71</td>
<td>130</td>
<td>58.222%</td>
<td>41.4778%</td>
</tr>
</tbody>
</table>

Table 3. Results for the accuracy of the applied classifiers on the balanced data set

6. CONCLUSIONS AND RECOMMENDATION
The project has shown that data mining techniques can be applied on student’s database to predict student failure/success rate so that managing students’ performance at the beginning of the year (after first semester), assist students before they reached risk of failure, helping and guiding administrative officers to be successful in management and decision making. The objective of this project is developing a predictive model that could help Haramaya University Health Science college to identify students at risk of failure to finish their study so that they can be treated before the condition escalate into students academic dismissal and wastage of resources. The researcher use data set of 1016 records of students to build and test both decision tree and NaiveBayes models. The balanced datasets were used to minimize the impact of majority class data sets on degree of accuracy. Hence the performance of decision tree classifier on the re sampling data set was increased than the unbalanced dataset. The generalized decision tree with pruning by some parameter tuning found to be the best relevant technique on the dataset to get meaningful patterns from the decision tree experiments. Based on the findings of this project The researcher would like to make the following recommend that data mining should be applied on student’s database for many advantages that
can be obtained for the institute as well as for the country. Therefore for future the researcher should work research on our Universities for the improvement of management decision making.

7. REFERENCE

[1]. Alaa el- Halees,(2009), Mining Students Data to Analyze e-learning behavior. A case study.


[5]. Jing Luan (2006), Data Mining Applications in Higher Education.


