Global Journal of Computer Science and Technology
Volume 12 Issue 2 Version 1.0 January 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Improving Academic Performance of Students of Defence University Based on Data Warehousing and Data mining

By Dr. Vuda Sreenivasarao, Capt. Genetu Yohannes
Defence University College, Debrezeit, ETHIOPIA

Abstract - The student academic performance in Defence University College is of great concern to the higher technical education managements, where several factors may affect the performance. The student academic performance in engineering during their first year at university is a turning point in their educational path and usually encroaches on their general point average in a decisive manner. The students evaluation factors like class quizzes mid and final exam assignment are studied. It is recommended that all these correlated information should be conveyed to the class teacher before the conduction of final exam. This study will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students. Statistics plays an important role in assessment and evaluation of performance in academics of universities need to have extensive analysis capabilities of student achievement levels in order to make appropriate academic decisions. Academic decisions will result in academic performance changes, which need to be assessed periodically and over span of time. The performance parameters chosen can be viewed at the individual student, department, school and university levels. Data mining is used to extract meaning full information and to develop significant relationships among variables stored in large data set/ data warehouse. In this paper is an attempt to using concepts of data mining like k-Means clustering, Decision tree Techniques, to help in enhancing the quality of the higher technical educational system by evaluating student data to study the main attributes that may affect the performance of student in courses.

Keywords : *Data base, Data warehousing, Data mining, Academic Performance, Educational data mining , Student performance analysis and K-Means clustering algorithm .*

GJCST Classification: *H.2.8*

Strictly as per the compliance and regulations of:

© 2012 Dr. Vuda Sreenivasarao, Capt. Genetu Yohannes. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction inany medium, provided the original work is properly cited.

Improving Academic Performance of Students of Defence University Based on Data Warehousing and Data mining

Dr. Vuda Sreenivasarao[α], Capt. Genetu Yohannes[Ω]

Abstract - The student academic performance in Defence University College is of great concern to the higher technical education managements, where several factors may affect the performance. The student academic performance in engineering during their first year at university is a turning point in their educational path and usually encroaches on their general point average in a decisive manner. The students evaluation factors like class quizzes mid and final exam assignment are studied. It is recommended that all these correlated information should be conveyed to the class teacher before the conduction of final exam. This study will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students. Statistics plays an important role in assessment and evaluation of performance in academics of universities need to have extensive analysis capabilities of student achievement levels in order to make appropriate academic decisions. Academic decisions will result in academic performance changes, which need to be assessed periodically and over span of time. The performance parameters chosen can be viewed at the individual student, department, school and university levels. Data mining is used to extract meaning full information and to develop significant relationships among variables stored in large data set/ data warehouse. In this paper is an attempt to using concepts of data mining like k-Means clustering, Decision tree Techniques, to help in enhancing the quality of the higher technical educational system by evaluating student data to study the main attributes that may affect the performance of student in courses.

Keywords : Data base, Data warehousing, Data mining, Academic Performance, Educational data mining , Student performance analysis and K-Means clustering algorithm .

I. Introduction

Data mining techniques have been applied in many application domains such as Banking, Fraud detection, Instruction detection and Communication. Recently the data mining techniques were used to improve and evaluate the engineering education tasks. Some authors have proposed some techniques and architectures for using data warehousing and data mining for higher technical education. Data mining is a process of extracting previously unknown, valid, potentional useful and hidden patterns from large data sets. As the amount of data stored in educational data bases in increasing rapidly. In order to get required benefits from such large data and to find hidden relationships between variables using different data mining techniques developed and used. Clustering and decision tree are most widely used techniques for future prediction. The aim of clustering is to partition students in to homogeneous groups according to their characteristics and abilities. These applications can help both instructor and student to improve the quality education. Analyze different factors effect a students learning behavior and performance during academic career using K-means clustering algorithm and decision tree in an higher educational institute. Decision tree analysis is a popular data mining technique that can be used to explain different variables like attendance ratio and grade ratio. Clustering is one of the basic techniques often used in analyzing data sets. This study makes use of cluster analysis to segment students in to groups according to their characteristics. Academic decisions may require extensive analysis of student achievement levels. Statistical data can also be used to see the results of important academic decisions. It is necessary to have measurements to make appropriate academic decisions on one hand; while on the other hand, there is a need to see the results of academic decisions by taking measurements. The decision, implementation, measurement and evaluation mechanisms work like a chain one leading to the other. Their relationship is shown in Fig.1.

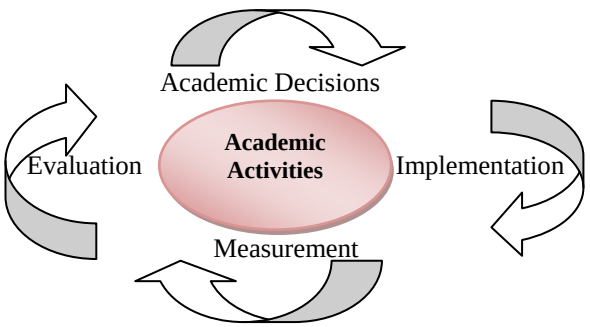


Figure 1 : Academic decision phases.

Author α : Professor, Dept of Computer & Information Technology, Defence University College, Debrezeit, ETHIOPIA.
E-mail : vudasrinivasarao@gmail.com
Author Ω : Head of Department, Dept of Computer & Information Technology, Defence University College, Debrezeit, ETHIOPIA.
E-mail : genetu77@yahoo.com

© 2012 Global Journals Inc. (US)

II. Related Work

a) Data Base

A data base is a collection of data usually associated with some organization or enterprise. Unlike a simple set, data in a data base are usually viewed to have a particular structure or schema with which it is associated. For example,(ID, Name, Address, Salary, Job No) may be the schema for a personal data base.

b) Data warehousing

Data warehouse is a data base devoted to analytical processing. Data warehouse to be a set of data that supports DSS and is subject-oriented, integrated, time-variant, and non-volatile. A complete repository of historical corporate data extracted from transaction systems that is available for ad-hoc access by knowledge workers. The processes of DW involve taking data from the legacy system together with corresponding transactions of the system's data base and transforming the data in to organized information in a user friendly format. The data warehouse market supports such diverse industries as manufacturing, retail, telecommunications and health care. It has access a warehouse includes traditional querying, OLAP, and data mining. Since the warehouse is stored as a data base, it can be accessed by traditional query languages.

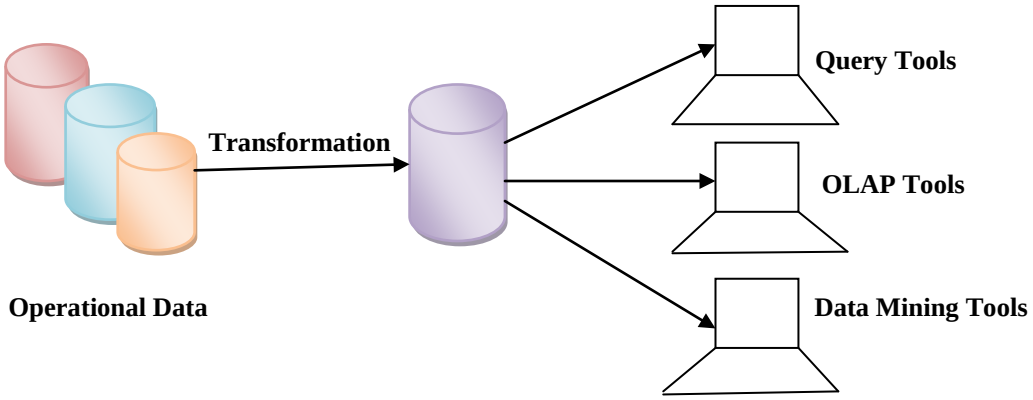


Figure 2 : Data ware house

Example of data warehousing can be defined in any of your organization. Consider the case of a Bank; a bank will typically have current accounts and saving accounts, foreign currency account etc. The bank will have an MIS system for leasing and another system for managing credit cards and another system for every different kind of business they are in . However, nowhere they have the total view of the environment from the customer's perspective. The reason being, transaction processing systems are typically designed around functional areas, within a business environment. For good decision making you should be able to integrate the data across the organization so as to cross the LoB (Line of Business) . So the idea here is to give the total view of the organization especially from a customer's perspective within the data warehouse, as shown in below figure 3.

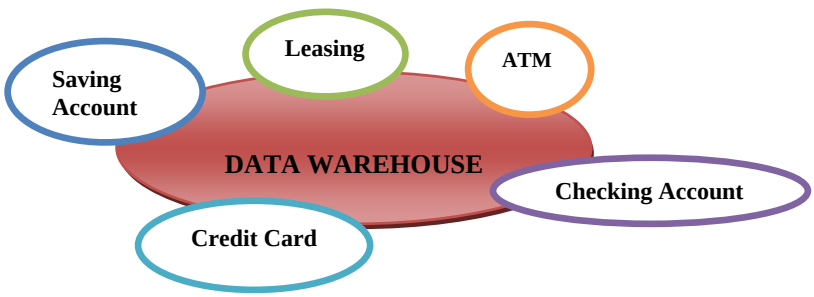


Figure 3 : A data warehouse crosses the Line of Business.

© 2012 Global Journals Inc. (US)

c) Data Mining

Data mining Techniques are used to extract useful and valid patterns from huge data bases. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. Large amount of data is accumulated in university students. Data mining software allow the users to analyze data from different dimensions categorize it and a summarized the relationships, identified during the mining process. Different data mining techniques are used in various fields of life such as medicine, statistical analysis, engineering, education, banking, marketing, sale etc. Data mining techniques can be differentiated by their different model functions and representation, preference criterion, and algorithms .The main function of the model that we are interested in is Classification, as normal, or malicious, or as a particular type of attack . We are also interested in link and sequence analysis. Additionally, data mining systems provide the means to easily perform data summarization and visualization, aiding the security analyst in identifying areas of concern .The models must be represented in some form. Common representations for data mining techniques include rules, decision trees, linear and non-linear functions (including neural nets), instance-based examples, and probability models.

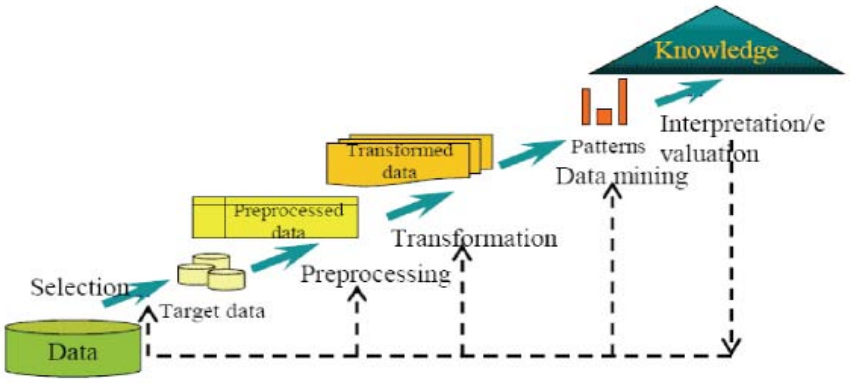


Figure 2 : The transition from raw data to valuable knowledge.

III. Clustering

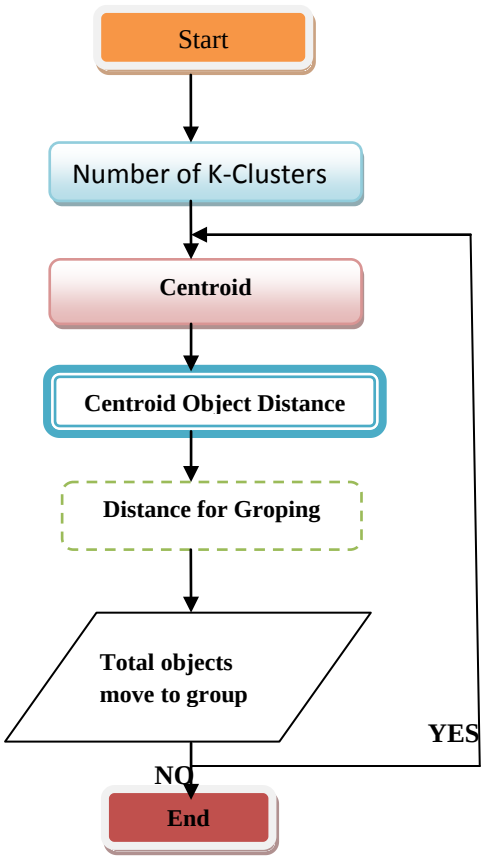
Clustering is a method to group data in to classes with identical characteristics in which the similarity of intra-class is maximized or minimized. Cluster analysis used to segment a large set of data in to subsets called clusters. Each cluster is a collection of data objects that are similar to one another are placed within the same cluster but are dissimilar to objects in other clusters. A cluster of data objects can be treated collectively as one group in many applications. Cluster analysis is an important human activity. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, and market research. Clustering is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes. Current clustering techniques can be broadly classified in to three categories; partitional, hierarchical and locality-based algorithms.

Definition: Given a data base $D = \{ t_1, t_2, \text{--------} t_n\}$ of tuples and an integer value K, the clustering problem is to define a mapping **f:D** $\{1,2,3\text{--------} K\}$ where each t_i is assigned to one cluster K_j, $i \leq j \leq k$. A cluster K_j, contains precisely those tuples mapped to it; that is $K_j = \{t_i / f(t_i) = k_j, 1 \leq i \leq n$ and $t_i \in D\}$.

a) K-Means Clustering

K-Means is one of the simplest unsupervised learning algorithms used for clustering. K-means partitions "n" observations in to k clusters in which each observation belongs to the cluster with the nearest mean. This algorithm aims at minimizing an objective function, in this case a squared error function.

© 2012 Global Journals Inc. (US)



Flow chart: K-Means clustering.

Algorithm: K-Means Clustering.

1. **Select number of K points as the initial centroids.**
2. **Repeat.**
3. **Form K clusters by assigning all points to the nearest centroid.**
4. **Recomputed the centroid of each Cluster.**
5. **Until the centroids don't change.**

IV. Decision Tree

Decision tree induction can be integrated with data warehousing techniques for data mining. A decision tree is a predictive node ling technique used in classification, clustering, and prediction tasks. Decision tree use a "divide and conquer" technique to split the problem search space in to subsets.

A decision tree is a tree where the root and each internal node are labeled with a question. The arcs emanating from each node represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration.

Given a data base D = $\{t_1,t_2,$--------$t_n\}$ where t_i = (t_{i1}----------t_{in}) and the data base schema contains the following attributes {A1,A2,-------An}. Also given is a set of classes C = {C1, C2, --------Cm}. A decision tree or classification tree is a tree associated with D that has the following properties:

1. Each internal node is labeled with an attribute, Ai.
2. Each arc is labeled with a predicate that can be applied to the attribute associated with the parent.
3. Each leaf node is labeled with a class, Cj.

© 2012 Global Journals Inc. (US)

The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner.

Decision Tree Algorithm: generate a decision tree from the given training data

1. Create a node **N**
2. **If** samples are all of the same class, C **then**
3. Return N as a leaf node labeled with the class C;
4. **If** attribute-list is empty **then**
5. Return N as a leaf node labeled with the most common class in samples.
6. Select test-attribute, the attribute among attribute-list with the highest information gain;
7. Label node N with test-attribute;
8. **For** each known value a_i of test-attribute.
9. Grow a branch from node **N** for the condition test attribute = a_i;
10. Let Si be the set of samples for which test-attribute = a_i;
11. **If** Si is empty **then**
12. Attach a leaf labeled with the most common class in samples;
13. **Else** attach the node returned by generate-decision-tree(S_i,attribute-list-attribute);

Each internal node tests an attribute, each branch corresponds to attribute value, and each leaf node assigns a classification.

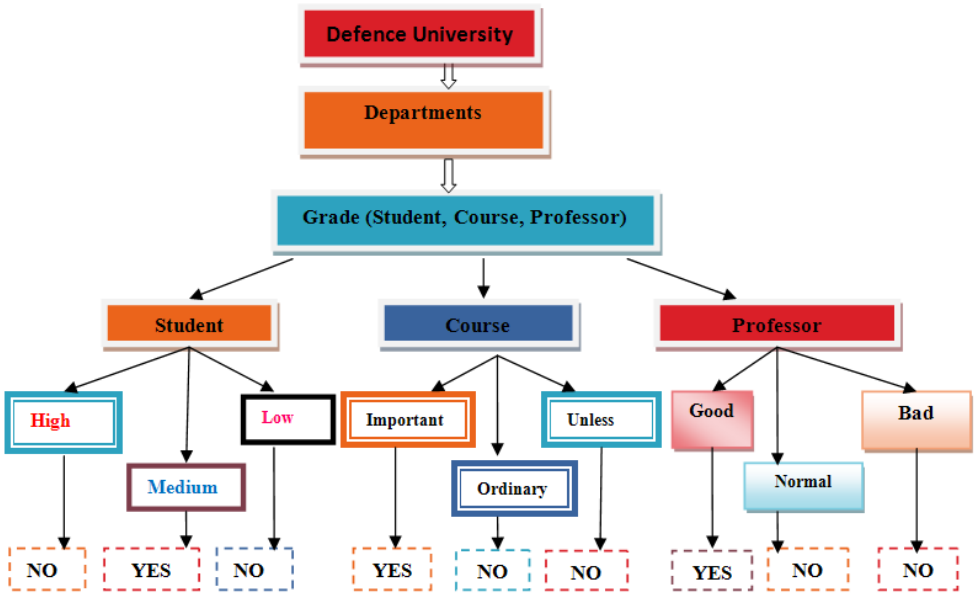


Figure 4 : Decision Tree

© 2012 Global Journals Inc. (US)

Table 1 shows the form of training data.1500 student score records are used for training.

Table 1 : Form of training examples.

S NO	Student	Course	Professor	Marks	Grade	Results
1	Ferede Adugna	Cryptography	Dr. Rao	86	A	YES
2	Rwibasira M	Interfacing Tech	Miss Maria	75	B	NO
3	Makuei Nyok	Operating Systems	Genetu Yohannes	85	A	YES
4	Daniel Tekif	Microprocessor	Michael	55	D	NO
5	Mesfin Dadi	Distributed Systems	Lea	95	A	YES
6	Debebe Shibeshi	Computer Networks	Genetu Yohannes	90	A	YES
7	Gidey Abrha	Network security	Dr. Srinivas	98	A	YES
8	Samuel Hagos	Web technology	Oliver	45	F	NO
9	Desta Desisa	Compiler Design	Melissa	73	B	NO
10	Tibabu Beza	Cloud computing	Praveen	50	D	NO
11	Desta Hagos	Mobile Communication	Dr.K.A.Lathiaf	70	B	NO
12	Ferede Adugna	Cryptography	Dr. Rao	86	A	YES
13	Rwibasira M	Interfacing Tech	Miss Maria	75	B	NO
14	Makuei Nyok	Operating Systems	Genetu Yohannes	85	A	YES
15	Daniel Tekif	Microprocessor	Michael	55	D	NO

Table 2 : shows Data base for Previous Semester student with effort taken from Marks taken.

Student Roll number	Marks	Effort
DEC-01	10-50	More Attention, Conducting Special Classes, Assigments, Conducting more practical classes, Daily tests and Parents and Faculty meeting.
DEC-02	51-60	Conducting Special Classes, Assignments and Conducting more practical classes.
DEC-03	61-75	Assignments and Conducting more practical classes.
DEC-04	76-85	Assignments and Conducting classes for Interviews.
DEC-05	86-100	Conducting classes for Interviews and Giving exposure for career important.

Formation of Decision tree from students marks table and applying effort depending on marks.

© 2012 Global Journals Inc. (US)

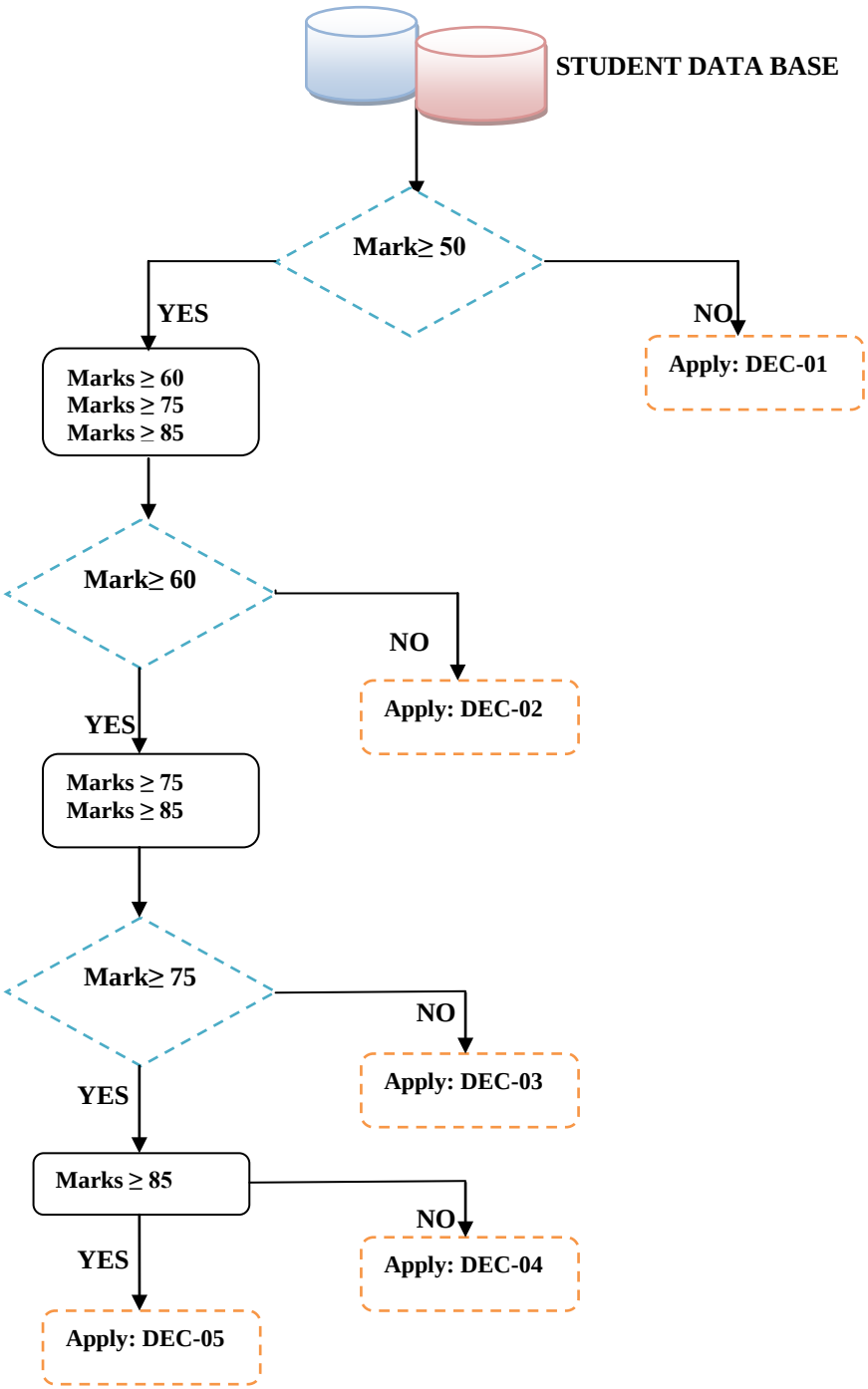


Figure 5 : Flowchart of students marks table and applying effort depending on marks.

After the pattern is classified from the decision tree we can obtain the specify knowledge discovery to form the knowledge base system. Similarly the same data mining process can be done to the professors for classifying their performance which help in improve Technical education system.

© 2012 Global Journals Inc. (US)

V. Results

Both K-Means clustering, Decision tree algorithm were applied on the data set.

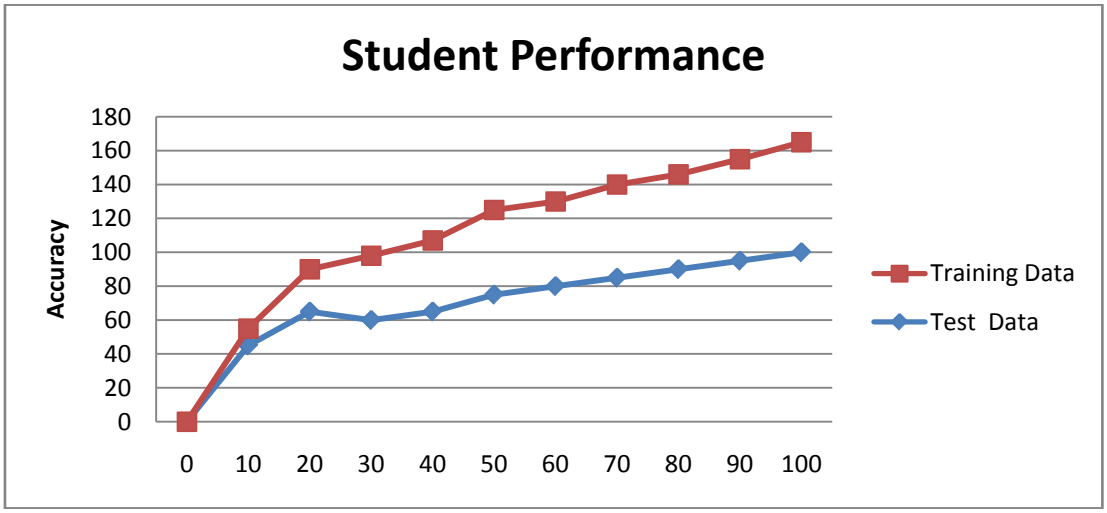


Figure 6 : Accuracy of Decision tree.

VI. Conclusion

In this study of research paper idea is a starting attempt to use data warehousing and data mining techniques to analyze and find out student academic performance and to improve the quality of the engineering system. The managements can use some techniques to improve the course outcomes according to the improve knowledge. Such knowledge can be used to give a good understanding of student's enrollment pattern in the course under study, the faculty and managerial decision maker in order to utilize the necessary steps needed to provide extra classes. Other hand, such type of knowledge the management system can be enhance their policies , improve their strategies and improve the quality of the system.

References Références Referencias

1. Vuda sreenivasarao, Dr.S.Vidyavathi, G.Ramaswamy,and Sk.Shabber, " A Research on result oriented learning process from university students based on distributed data mining and decision tree algorithm", Journal of Advance research in Computer Engineering, (Vol.4 No.2 2010), pp- 223-226 .
2. P.V.Subbareddy and Vuda Sreenivasarao, "The result oriented process for process for students based on distributed data mining", International journal of advanced computer science and applications, Vol. 1, No.5.Nov-2010, pp-22-25.
3. N.V Anand Kumar, G.V.Uma "Improving Academic Performance of Students by Applying Data Mining Technique" European Journal of Scientific Research, Volume 34, Issue 4, pp-526-534.
4. Chady EL moucary,Maric khair and Walid Zakhem " Improving student's performance using data clustering and neural networks in Foreign Language based higher education", The research bulletin of Jordan ACM, Vol 11(III),PP-27-34.
5. Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza sattar and M.Inayat khan, " Data mining model for higher education system", European Journal of Scientific Research,Vol.43 No.1,pp-24-29.
6. Dervis.Z.Deniz and I brahim ersam, "An academic decision-support system based on academic performance evaluation for student and program assessment", Int Journal Ed, Vol 18, No 2, pp.236-244.
7. Bindiya M varghese, Jose Tome J, Unnikrishna A and Poulose Jacob K, "Clustering student data to characterize performance patterns", International journal of Advanced computer and applications, Special Issue,pp-138-140.
8. Data Mining: Concepts and Techniques, Han J Kamber M, Morgan Kaufmann Publishers, 2001.
9. Introduction to DATA MINING, Tan P., Steinbach M., Kumar V, Pearson Education, 2006.
10. A Machine Learning Algorithms in Java, Witten I. Frank E. WEKA, Morgan Kaufmann Publishers, 2000.
11. Hellenic Open University, Patras, Greece, Analyzing student performance in distance learning with genetic algorithms and decision trees,. Kalles D., Pierrakeas C, 2004.
12. Mining student's data to analyze E-learning behavior: A case study-Alaa el-Halees.
13. Data mining with SQL server 2005 by Zhao Hui.Maclennan.J, Wihely publishing.Inc-2005.

© 2012 Global Journals Inc. (US)