Mining higher educational students data to analyze student’s admission in various discipline
Neeraj Bhargava¹, Anil Rajput², Pooja Shrivastava³

Abstract
In this paper our main goal to analyze Higher Education students’ data for taking admission in various disciplines. For this purpose we have used data from urban, semi-urban and rural area institutions. In the analysis we applied data mining techniques. It is the key which unlocks the road of modernization. Programs of education at all levels and especially at the higher level constitute the base of the effort to forget the bounds common citizenship to harness the energies of the people. In this research paper we have studied and taking idea to develop it [1 – 7].

Keywords: Higher Educational Data Mining, Classification; Knowledge Discovery in Database (KDD), Decision Tree

Related Work
We have described how different data mining techniques can be used in order to improve the course and the students’ learning. All these techniques can be applied separately in a same system or together in a hybrid system.

Have a survey on educational data mining between 1995 and 2005. They have compared the Traditional Classroom teaching with the Web based Educational System. Also they have discussed the use of Web Mining techniques in Education systems.

Have a described the use of k-means clustering algorithm to predict student’s learning activities. The information generated after the implementation of data mining technique may be helpful for instructor as well as for students.

The current education system does not involve any prediction about fail or pass percentage based on the performance. The system doesn’t deal with dropouts. There is no efficient method to caution the student about the deficiency in attendance. It doesn’t identify the weak student and inform the teacher. Another common problem in larger colleges and universities, some students may feel lost in the crowd. Whether they’re struggling to find help with coursework, or having difficulty choosing (or getting into) the courses they need, many students are daunted by the task of working through the collegiate bureaucracy. Since the proposed model identifies the weak students, the teachers can provide academic help for them. It also helps the teacher to act before a student drops or plan for recourse allocation with confidence gained from knowing how many students are likely to pass or fail.

Motivation Data Mining In Higher Education
Data mining techniques involves predictive modeling, cluster analysis and association and sequence (or market-basket analysis).

A Decision tree
Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node.

Classification trees are used for the kind of Data Mining problem which are concerned with prediction. Using examples of cases it is possible to construct a model that is able to predict the class of new examples using the attributes of those examples. An experiment has been set up to test the performance of two pruning methods which are used in classification tree modeling. The pruning methods will be applied on different types of data sets.

It is a machine learning technique that allows us to estimate a quantitative target variable (for example, profit, loss or loan amount) or classify observation into one category of a categorical target variable (for example, good/bad credit customer; churn or do not churn) by repeatedly dividing observations into mutually exclusive groups. The algorithm commonly used to construct decision tree is known as recursive partitioning and the common algorithms are CHAID (Chi-square Automatic Interaction Detection), CART (Classification & Regression Tree) and C5.0. This paper will focus on using CART in building the decision tree.
Decision trees represent a supervised approach to classification. Weka uses the J48 algorithm, which is Weka’s implementation of C4.5 Decision tree algorithm. J48 is actually a slight improved to and the latest version of C4.5. It was the last public version of this family of algorithms before the commercial implementation C5.0 was released.

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called “root” that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute’s value. In the case of numeric attributes, the condition refers to a range.

Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Figure 1 describes a decision tree that reasons whether or not a potential customer will respond to a direct mailing. In case of numeric attributes, decision trees can be geometrically interpreted as a collection of hyperplanes, each orthogonal to one of the axes. Naturally, decision-makers prefer less complex decision trees, since they may be considered more comprehensible.

The tree complexity has a crucial effect on its accuracy. The tree complexity is explicitly controlled by the stopping criteria used and the pruning method employed. Usually the tree complexity is measured by one of the following metrics: the total number of nodes, total number of leaves, tree depth and number of attributes used. Decision tree induction is closely related to rule induction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part, and taking the leaf’s class prediction as the class value.

### Classification Rule in Student Data Base

We are using database of Alpha College i.e. rural area, Sehore which is semi-urban area and Bhopal which is urban area from Madhya Pradesh state to apply data mining techniques and calculate results. Higher Education System of Madhya Pradesh has a great amount of data which can be analyzed and extracted for the data mining system. Faculties and departments have also important data regarding courses and modules which will be collected in document form (excel sheets).

The dataset file look like as follows:

First require attributes from these excel sheets are selected which are as follows:

- Course
- Branch
- Gender
- Category
- Class
- Income
- Date of Admission
- Minority/Non-Minority

We are using decision tree, cluster analysis and association rule data mining techniques for calculating results and for that purpose we are using Weka as data mining tool. First we open the GUI of Weka which looks like as follows:

Then we open Explorer application to preprocess the dataset. We apply the unsupervised filter to convert the attribute values form numeric to nominal as the processing is done on nominal in decision tree.
Results for rural area

Data from all excel sheets are collected in one excel sheet and excel sheet is saved in “rural.csv” file for processing in WEKA. In this dataset we are having total 134 records. Out of this 34% of data is used for training and 66% data is used as testing purpose. Ten fold cross-validations are done on data.

Decision Tree when attribute Course is selected

From this decision tree we can easily identify number of students admitted for particular course. From this data we can recognize the most selected and less selected courses by the students.

Number of Leaves: 6
Size of the tree: 7
Time taken to build model: 0 seconds

--- Stratified cross-validation ---
--- Summary ---
Correctly Classified Instances 134 100 %
Incorrectly Classified Instances 0 0 %
Total Number of Instances 134

Confusion matrix is calculated which shows the classification courses whether they are correctly classified or misclassified.
Figure 3: Preprocessing Weka Explorer

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>0</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>31</td>
</tr>
</tbody>
</table>

a = B. Com
b = B. Sc.
c = B. A.
d = B. C. A.

Figure 4: Decision Tree Visualization Attribute Course is selected
Conclusion

It showed using graph how useful data mining can be in education in particularly to improve student performance of higher education. We used students' data from database course of Sehore Distinct M.P. We collected all available data including their usage of Model learning facility of education. The data mining techniques that used are association rules and decision tree help to improve the result. They help to uncover the hidden patterns from the large data.

References