International Journal of Advanced Research in Engineering Technology & Sciences ISSN: 2394-2819
July-2016 Volume 3, Issue-7
Email:editor@ijarets.org www.ijarets.org

TOOLS AND TECHNIQUES OF DATA MINING IN EDUCATION SYSTEM FOR EFFECTIVE DECISION MAKING

ShubhanshVibhu, Sudhanshu Maurya
Research Scholar
Computer Science and Engineering
NITRA Technical Campus
Ghaziabad

Dr. B. K. Sharma
Professor
Head, Computer Science & Engg. and
Software Development Centre
NITRA Technical Campus
Ghaziabad

ABSTRACT: *Data mining is the process of extracting valid, previously unknown, and ultimately comprehensible information from large databases and using it to make crucial business decisions. The extracted information can be used to form a prediction or classification model, identify relations between database records, or provide a summary of the database(s) being mined. Data mining consists of a number of operations each of which is supported by a variety of techniques such as rule induction, neural networks, conceptual clustering, association discovery, etc.*

In many real-world domains such as marketing analysis, financial analysis, fraud detection, etc. information extraction requires the cooperative use of several data mining operations and techniques. Under this project we will be applying data mining tools and techniques in education system for quick decision making and taking necessary corrective action. We also aim in admonishing various revolutionary activities in academics by analyzing the data that we will be using. In our study we will be using various open source software. A detailed analysis in education system is required so that the student and all the stakeholder of the education system can be confronted will the simpler analysis so that they can easily fetch their desired result

KEYWORDS: *Educational Data Mining (EDM); Classification; Knowledge Discovery in Database (KDD); J48 Algorithm.*

1. INTRODUCTION

The advent of information technology in various fields has lead the large volumes of data storage in various formats like records, files, documents, images, sound, videos, scientific data and many new data formats. The data collected from different applications require proper method of extracting knowledge from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data . The main functions of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data Data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. Data mining techniques have been introduced into new fields of Statistics, Databases, Machine Learning, Pattern Reorganization, Artificial Intelligence and Computation capabilities etc.

There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational environments. Educational Data Mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor, and many others. Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for prediction regarding various attributes concerned with student's performance in class and it's result, alienation of traditional classroom teaching model, attendance and its impact on the

student, general emotional and critical thinking factors , prediction about student's performance and so on. The main objective of this paper is to use data mining methodologies to study student's performance in the courses. Data mining provides many tasks that could be used to study the student performance. In this Research, the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here. Information's like Attendance, Class test, Seminar and Assignment marks were collected from the student's management system, to predict the performance at the end of the semester. This paper investigates the accuracy of Decision tree techniques for predicting student performance.

2. KNOWLEDGE DISCOVERY IN DATABASE

The **Knowledge Discovery in Databases** (KDD) process is commonly defined with the stages:

- Selection of Data
- Pre-processing
- Data Transformation
- Data Mining
- Interpretation/Evaluation

SELECTION OF DATA

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data. Before person begins with all the happening tools that are available in WEKA or any other software for data mining purpose his goals should be clear of what he want to achieve rather on how he want to achieve. Following question one should always ask before performing data mining.

PRE-PROCESSING

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving

DATA TRANSFORMATION

Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system. The usual process involves converting documents, but data conversions sometimes involve the conversion of a program from one computer language to another to enable the program to run on a different platform. The usual reason for this data migration is the adoption of a new system that's totally different from the previous one. In real practice, data transformation involves the use of a special program that's able to read the data's original base language, determine the language into which the data that must be translated for it to be usable by the new program or system, and then proceeds to transform that data.

DATA MINING

As discussed earlier data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. It is the most important phase of data mining. One of the properties of implementation of data mining is that there is no system of input/output result.

Knowledge discovery is a constant process of analysis which requires lots of human consciousness. Prediction through final result completely depends on the understanding of user. If user prospective towards model is wrong, what so ever may the model is or irrespective to its accuracy, improper understanding may lead to wrong outcome.There are various techniques that are used in data mining, they are listed as follow

- Classification
- Clustering
- Prediction
- Association Rules
- Decision Tree

2.1 CLASSIFICATION

Classification is the most commonly applied data mining technique, which employs a set of pre classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network -based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier- training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

2.2 CLUSTERING

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem.

2.3 PREDICTION

Technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real - world problems are not simply prediction. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be

Necessary to forecast future values the same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables).

2.4 ASSOCIATION RULE

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

2.5 DECISION TREES

Decision tree is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision tree is a classifier expressed as a recursive partition of the in-stance 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.

3. USE OF DATA MINING APPLICATION IN EDUCATIONAL SYSTEM

In present day educational system, a student's performance is determined by the its previous educational qualification, the emotional andcritical thinking ability of the student, its interaction with the educational infrastructure and finally it can be judged by the end semester examination. The end semester examination is one that each student has to pass with some minimum marks to get a pass status.

The data set used in this study was obtained from a survey done in NITRA Technical Campus, Ghaziabad (Uttar Pradesh). The size of the training data was 58 and the size of the test data is 18. Incomplete or inaccurate data were identified and removed.

s.no.	**Attributes**	**Scale**
1	Tenth	$\begin{cases} \boldsymbol{if\ marks \geq 80 \rightarrow A} \\ \boldsymbol{if\ 80} > marks \geq 60 \rightarrow B \\ \boldsymbol{if\ 60} > marks \geq 33 \rightarrow C \end{cases}$
2	Twelfth	$\begin{cases} \boldsymbol{if\ marks \geq 80 \rightarrow A} \\ \boldsymbol{if\ 80} > marks \geq 60 \rightarrow B \\ \boldsymbol{if\ 60} > marks \geq 33 \rightarrow C \end{cases}$
3	BTech	$\begin{cases} \boldsymbol{if\ marks \geq 80 \rightarrow A} \\ \boldsymbol{if\ 80} > marks \geq 60 \rightarrow B \\ \boldsymbol{if\ 60} > marks \geq 30 \rightarrow C \end{cases}$
4	Attendance	$\begin{cases} \boldsymbol{if\ Attendance} > 95 \rightarrow Ve \\ \boldsymbol{if\ 95} < Attendance \geq 85 \\ \boldsymbol{if\ 85} < Attendance \geq 75 \rightarrow \\ \boldsymbol{if\ Attendance} > 75 \rightarrow \end{cases}$
5	Proficiency In English	**{Low, Medium, High}**

6	Travel Time	$\begin{cases} if\ \mathbf{Time} \leq \mathbf{15}\ \boldsymbol{min} \rightarrow \boldsymbol{Ve} \\ if\ \mathbf{15}\boldsymbol{min} < Time \geq 30\ m \\ if\ \mathbf{30}\ \boldsymbol{min} < Time \geq 1\boldsymbol{Hr} \rightarrow \\ if\ \boldsymbol{Time} > 1\boldsymbol{Hr} \rightarrow \boldsymbol{M} \end{cases}$
7	Physical_Activity	**{Low, Medium, High, Very High}**
8	T_Total	**{Numerical value}**
9	Result	**{Pass, Fail}**

3.1 ALGORITHM J48

STEP 1:- In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class**.**

STEP 2:- The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.

STEP 3:- Attributes with highest potential Information is selected as ROOT node.

STEP 4:- Repeat **STEP 2** and **STEP 3** for each attribute, until instances of parent node is impure or no attributes are lest for further classification.

3.2 DATA ANALYSIS

The Dataset of 76 students used in the study were obtained from our College are given in Table-1

Table-1 : Student Training Data

Student ID	Tenth	Twelfth	BTech	Attendance	Prof_In_English	Travel_Time	Physical_Activity	T_Total
CSE38	A	A	B	Very Good	Medium	Very Less	High	8
CSE42	B	B	C	Marginal	Medium	Very Less	Medium	7
CSE12	A	A	A	Very Good	High	Very Less	Medium	8
CSE50	A	A	B	Very Good	High	Less	High	7
CSE04	B	C	C	Good	Medium	Very Less	Medium	10
CSE01	B	B	B	Marginal	High	Very Less	Very High	6
CSE09	B	C	C	Poor	Low	Less	High	5
CSE21	B	B	C	Marginal	Medium	Very Less	Medium	4
CSE48	A	A	B	Good	Medium	Much	Medium	6
CSE24	A	A	B	Good	Medium	Very Less	Medium	7
CSE22	A	A	B	Good	Medium	Very Less	Low	5
CSE31	A	A	A	Very Good	Medium	Sufficient	High	8
CSE34	A	B	A	Very Good	Medium	Sufficient	Very High	7
CSE19	A	A	B	Marginal	High	Less	Low	6
CSE18	B	B	B	Good	Medium	Very Less	Very High	7
CSE08	A	B	B	Marginal	High	Very Less	Very High	9
CSE15	A	A	B	Marginal	High	Less	High	7
CSE25	A	A	B	Marginal	High	Very Less	High	7

Student ID	Tenth	Twelfth	BTech	Attendance	Prof_In_English	Travel_Time	Physical_Activity	T_Total	Result
CSE17	B	B	C	Marginal	Medium	Much	Medium	8	Fail
CSE26	B	B	C	Good	Medium	Very Less	Medium	5	Fail
CSE16	A	A	B	Marginal	Medium	Very Less	Medium	8	Fail
CSE30	A	A	B	Good	High	Very Less	Medium	6	Pass
CSE10	A	B	B	Marginal	Medium	Very Less	Medium	4	Fail
CSE08	A	A	C	Good	Low	Very Less	Medium	3	Fail
CSE22	A	B	C	Marginal	Medium	Very Less	Very High	7	Fail
TT09	B	A	B	Good	Medium	Less	Medium	2	Pass
CSE35	B	B	B	Good	Medium	Very Less	Medium	7	Fail
CSE23	B	B	C	Good	Medium	Much	Medium	6	Fail
CSE13	A	A	A	Marginal	Medium	Very Less	High	12	Pass
CSE20	A	B	C	Good	Medium	Sufficient	Medium	4	Fail
CSE21	A	A	B	Good	Medium	Sufficient	Medium	4	Fail
CSE31	A	A	C	Marginal	Medium	Much	Medium	5	Fail
CSE29	A	A	C	Good	Medium	Very Less	Medium	3	Fail
TT02	A	A	B	Good	High	Less	High	5	Pass
CSE06	B	A	C	Poor	Medium	Much	Medium	7	Fail
CSE43	B	B	C	Poor	Medium	Very Less	Medium	6	Fail
CSE44	B	A	B	Poor	Medium	Very Less	Medium	6	Pass
CSE05	B	B	C	Good	Medium	Sufficient	Medium	9	Fail
CSE42	B	B	C	Marginal	High	Much	Very High	7	Fail
TC07	B	A	B	Very Good	Medium	Very Less	Medium	2	Pass
TC12	B	B	B	Good	Medium	Very Less	Medium	6	Pass
TC02	B	B	C	Poor	High	Very Less	High	1	Fail
TC05	B	A	C	Poor	Medium	Much	High	6	Fail
TC18	A	A	A	Very Good	Medium	Less	High	6	Pass
TC14	A	B	B	Good	High	Much	Medium	3	Fail
TC17	B	B	B	Marginal	High	Very Less	High	5	Pass
TC08	B	B	B	Marginal	Medium	Less	Medium	2	Pass
TC21	B	B	C	Marginal	Medium	Very Less	Medium	7	Fail
TC06	A	A	B	Good	Medium	Very Less	High	3	Pass
TC16	B	B	C	Poor	Medium	Very Less	Medium	4	Fail
TC09	A	A	C	Marginal	Medium	Very Less	High	3	Fail
TC15	B	B	C	Very Good	Medium	Very Less	High	3	Fail
TC03	A	B	C	Poor	High	Very Less	High	9	Fail
CSE41	B	B	B	Marginal	Medium	Less	High	2	Pass
TC01	A	A	C	Marginal	Medium	Very Less	High	5	Fail
TC20	A	B	B	Good	Medium	Very Less	Medium	4	Fail
CSE48	A	B	C	Marginal	Medium	Very Less	Medium	7	Fail
TT21	A	B	C	Marginal	Low	Much	Low	5	Pass
TT08	B	A	B	Very Good	Medium	Very Less	Medium	8	Pass
TT07	A	A	D	Good	Medium	Very Less	Medium	7	Pass
CSE46	A	B	B	Good	Medium	Very Less	Medium	7	Fail
TT11	B	B	B	Poor	Medium	Very Less	Medium	5	Pass
TT06	A	B	C	Very Good	Medium	Less	Very High	5	Fail
CSE50	A	A	B	Good	Medium	Sufficient	Medium	9	Fail
CSE33	C	B	C	Good	Low	Very Less	High	2	Fail
TT18	A	B	C	Poor	Medium	Very Less	Medium	7	Fail
TT15	B	B	B	Very Good	Medium	Much	Medium	3	Pass
TT22	A	A	B	Very Good	Medium	Less	Medium	5	Pass
TT10	A	B	B	Very Good	Medium	Very Less	Medium	7	Pass
CSE40	A	A	B	Marginal	Medium	Much	Medium	4	Pass
CSE39	A	B	C	Good	Medium	Very Less	Medium	5	Fail
CSE49	A	A	B	Marginal	High	Much	Medium	7	Pass
CSE07	A	B	C	Poor	Medium	Very Less	High	7	Fail
CSE18	A	B	B	Very Good	Medium	Sufficient	Low	8	Pass
CSE02	C	B	C	Good	Medium	Very Less	Medium	3	Fail
CSE15	B	B	B	Good	Low	Less	Medium	2	Fail

3.3 ENTROPY GAIN

This process uses the "Entropy" which is a measure of the data disorder. The Entropy of $\vec{y}$ is calculated by

$$Entropy\ (\vec{y}) = -\sum_{j=1}^{n} \frac{|y_1|}{|\vec{y}|} log(\frac{|y_1|}{|\vec{y}|})$$

$$Entropy\ (j|\ \vec{y}) = \sum_{j=1}^{n} \frac{|y_1|}{|\vec{y}|} log(\frac{|y_1|}{|\vec{y}|})$$

And Gain is

$$Gain(\vec{y}, j) = Entropy(\vec{y} - Entropy(j|\vec{y})$$

The objective is to maximize the Gain, dividing by overall entropy due to split argument $\vec{y}$ by value j. Calculated Entropy value is given in the Table-2.

Table -2 : Entropy Value

Name of Attribute	Entropy Value
Tenth	0.90713
Twelfth	0.91251
B.Tech.	0.58412
Attendance	0.86835
Prof_In_English	0.94759
Travel Time	0.88586
Physical Activity	0.87111

As entropy for each individual variable has been calculated next step will be calculation of Gain. For each attribute Entropy Gain is calculated and listed below in the Table-3.

Table -3 : Gain Value

Name of Attribute	Gain
Tenth	0.05042
Twelfth	0.04504
B.Tech.	0.37343
Attendance	0.08920
Prof_In_English	0.00996
Travel Time	0.07169
Physical Activity	0.08644

The knowledge represented by the decision tree can be extracted and represented in the form of if than rules.

IF B.Tech= 'C' AND Physical Activity = 'Low' THEN Result= 'Pass'
IF B.Tech= 'C' AND Physical Activity = 'Very High' THEN Result= 'Fail'
IF B.Tech= 'C' AND Physical Activity = 'Medium' THEN Result= 'Fail'
IF B.Tech= 'C' AND Physical Activity = 'High' THEN Result= 'Fail'
IF B.Tech= 'B' AND Physical Activity = 'High' THEN Result= 'PASS'
IF B.Tech= 'B' AND Physical Activity = 'Low' THEN Result= 'PASS'

IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Very Good' THEN Result= 'PASS'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Poor' THEN Result= 'Pass'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Marginal' AND Travel Time='Very Less' THEN Result= 'Fail'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Marginal' AND Travel Time='Less' THEN Result= 'Pass'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Marginal' AND Travel Time='Much' THEN Result= 'Pass'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Good' AND Twelfth= 'C' THEN Result= 'Fail'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Good' AND Twelfth= 'A' AND Travel Time= 'Very Less' THEN Result= 'Pass'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Good' AND Twelfth= 'A' AND Travel Time= 'Less' THEN Result= 'Pass'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Good' AND Twelfth= 'A' AND Travel Time= 'Sufficient' THEN Result= 'Fail'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Good' AND Twelfth= 'B' AND Tenth= 'A' THEN Result= 'Fail'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Good' AND Twelfth= 'B' AND Tenth= 'B' AND Prof_In_English= 'Low' THEN Result= 'Fail'
IF B.Tech= 'B' AND Physical Activity = 'Medium' AND Attendance= 'Good' AND Twelfth= 'B' AND Tenth= 'B' AND Prof_In_English= 'Medium' AND T_Total= '>6' THEN Result= 'Fail'
IF B.Tech= 'B' AND Physical Activity =

'Medium' AND Attendance= 'Good' AND Twelfth= 'B' AND Tenth= 'B' AND Prof_In_English= 'Medium' AND T_Total= '<=6' THEN Result= 'Pass'

4. CONCLUSION

In this paper, the classification technique is used on student database to predict the student result on the basis of previously known entities and the various tests through which they have undergone through. As there are many approaches that can be used for data classification, but here we have used the decision tree method. Information like Basic Educational Details of Secondary and Senior Secondary Schooling, Attendance, Travel time, Physical Activity, were collected from the students through the survey and they were asked to sit in for Emotional, General English and Critical Thinking Tests. These were done to predict the performance in the end of the semester. This study will help to the students and the teachers to identify the division of the students. This study will also work to identify those students which need special attention to reduce fail ratio and thus will help in taking appropriate action for the next semester examination.

5. REFERENCES

1. Dr. B.K. Sharma & Dr. S.P Singh, 2015. *A naive Bayes approach for converging learning objects with open educational resources, DOI: 10.1007/s10639-015-9416-2*
2. Ian H Witten &Aibe Frank, *Data Mining Practical Machine Learning Tools and Techniques*
3. S.K. AlthafHussainBasha, Y.R. Ramesh Kumar, A. Govardhan and Mohd.Zaheer Ahmed, *Predicting Student Academic Performance Using Temporal Association Mining*
4. Nitya Upadhyay and VinodiniKatiyar, *A Survey on the Classification Techniques In Educational Data Mining*
5. Paulo Cortez and Alicia Silva, *Using data mining to predict secondary school student performance*
6. Brijesh Kumar Bhardwaj and Saurabh Pal,(2011),*Mining Educational Data to Analyze Students' Performance*
7. Ian H. Witten, *Video lectures on WEKA*