Data Mining on Student Database to Improve Future Performance

Kashish Kohli
School of Computer Science Engineering
Vellore Institute of Technology
Chennai, India

Shiivong Birla
School of Computer Science Engineering
Vellore Institute of Technology
Chennai, India

ABSTRACT
Data Mining refers to the process of extracting information from large sets of data. Its primary implication is finding relationships between different variables to extract meaningful information. In this paper, we apply Data Mining techniques to find and evaluate future results and factors which affect them. Following preprocessing of data, several data mining techniques have been applied namely association, classification and clustering. We present the result and analysis after each process.

General Terms
Pattern Recognition, Data Mining, Clustering, Predictive Analytics

Keywords
Data Mining; Association; Classification; Clustering; Education

1. INTRODUCTION
In the recent years, the interest in Data Mining and Business Intelligence has boomed [1] [2]. This has led to an exponential growth in storage capacities hence leading to an increase in databases of several organizations. These databases contain important trends and patterns that can be utilized to improve success rate. Data Mining when applied to Educational Databases is used to detect patterns and extract knowledge that can help in improving the current education system. Education is critical for a nation to develop. Whether it is financially or socially, education assumes a fundamental part in the development of these two imperative components. In a country like India, which possesses one of the largest youngest populations in the world, the need for educational reform and modernization assumes an even greater importance. The Indian education has developed rapidly in the past decades, however it still occurs at the global education's tail end. School dropout rate amongst students in India is as high as 40% as compared to a global average of 24% [3].

The principle goal of advanced educational institutes is to give quality training to their students. This improves their capability of making executive is by extracting knowledge and patterns from the students' database and study factors that will affect their education and learning capabilities. These learning patterns can help teachers understand their students better. In this manner, they can approach concepts from a point of view that most students will be comfortable with [4]. This will help in reducing failure rate and improving the quality of teaching. The database used for this work has been obtained from students in various educational institutes of India and contains approximately 400 entries.

Several techniques are available to perform mining on this data. Data Mining also yields rules which can be harnesses in several techniques like Association Rules etc. [5]. The data mining in this paper follows the step-by-step procedure. The data that can be collected and is maintained in the database is shown. Then that data is subjected to preprocessing to ready it for mining. The subsequent mining operations are then applied which provides us the results. Finally, the usage of information obtained is explained. Mining operations vary over a large scale. In this paper, we investigated Association, Classification and Clustering. The software used was RStudio since it supports a wide variety of databases (including .csv which was used for this work). It also has in-built functionalities of KMeans among others. From the obtained results, it is interpreted as to how these results can help Indian educational authorities improve students' academics.

2. DATASET AND PREPROCESSIN
The dataset used in this work has been obtained from students of various educational institutes of India by the means of Google Forms. The database contains close to 400 records.

Attribute	Description
Name	Name of the Student
Sex	Sex (binary: female or male)
Age	Age (numeric: from 15 to 22)
Home	Home Type(binary: U- Urban; R-Rural)
City	City where secondary education is being/was obtained
FSize	Size of Family(binary: 0- Less than or equal to 4; 1- Greater than 4)
FEdu	Father Education Status(numeric: 0 - none, 1 - primary education, 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
MEdu	Mother Education Status(numeric: 0 - none, 1 - primary education, 2 – 5th to 9th grade, 3 – secondary-education or 4 – higher education)
Pstatus	Parent's cohabitation status (binary: "T" - living together or "A" - apart)
FJob	Father's job (nominal: Education/Health Care; Civil: Administrative/Police; At Home; Other)
MJob	Mother's job (nominal: Education/Health Care; Civil: Administrative/Police; At Home; Other)
TravellingTime	Home to School travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
ExtraActivities	Extra-Curricular activities (binary: yes or no)
Studytime	Weekly Study Time (numeric: 1- < 2 hours, 2- 2-5 hours, 3- 5-10 hours, 4- >10 hours)

Freetime	Free time after school (numeric: from 1 - very low to 5 - very high)
Outing	Going out with friends (numeric: from 1 - very low to 5 - very high)
NoofFail	Number of past class failures (numeric: n if 1<=n<3, else 4)

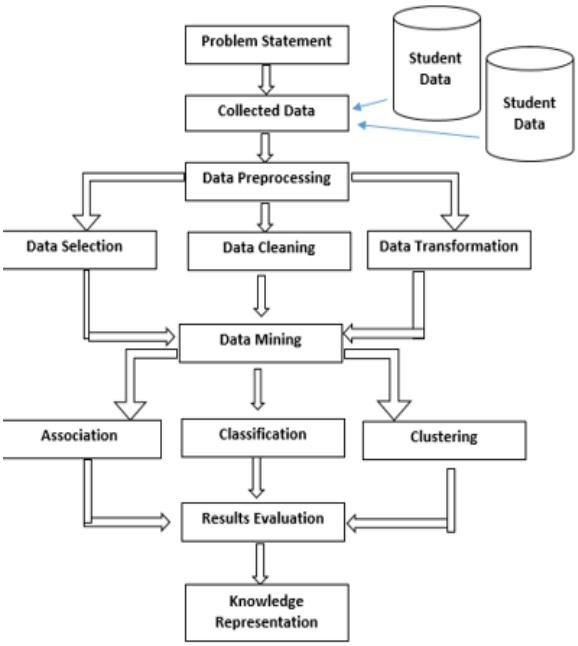

Figure 1: Methodology of Data Mining

Alco	Weekly alcohol consumption (numeric: from 1-very low to 5- very high)
Health	Current Health Status (numeric: from 1 - very bad to 5 - very good)
Gr1*	Mathematics grade (numeric: from 0 to 10)
Gr2*	Science grade (numeric: from 0 to 10)
Gr3*	Final grade (numeric: from 0 to 10)

*Grades are discretized after Preprocessing.

In India, the secondary education system consists of 9 years of primary education, following which 3 years of higher education is imparted (Classes 10th, 11th and 12th). Distribution into respective streams is done following 10th standard. There are several streams (e.g. Engineering Sciences, Medical Sciences, Commerce, Arts etc.). Subjects like Mathematics and Science (Medical/Social/Engineering) still form the backbone of most of these streams. The data collected is for the 2015-2016 period. Grades Gr1 and Gr2 refer to academic grades in Mathematics and Science receptively. Grade Gr3 represents the last evaluation of year/semester.

Several of the data fields in the collected database contain irrelevant information or information that cannot be processed with other values. This information is removed or categorized and this constitutes preprocessing.

- Some fields are discarded as part of Preprocessing and Data Preparation. Deletion of some fields promotes better input for Data Mining. The attributes are then subjected to preprocessing softwares e.g. RapidMiner etc. Now attributes that do not present any relevant information and only provide personal details of students are removed. These include Name, City etc. Variables like these provide high variance and little information and hence are of no consequence.
- The dataset consisted of some records with missing values. These records are deleted.
- Most values of Grade 3 are continuous in nature. These are discretized into grades. Hence five categories are now created: A, B, C, D and E.

Following Preprocessing, the remaining data is visualized especially for G3 (Figure 2):

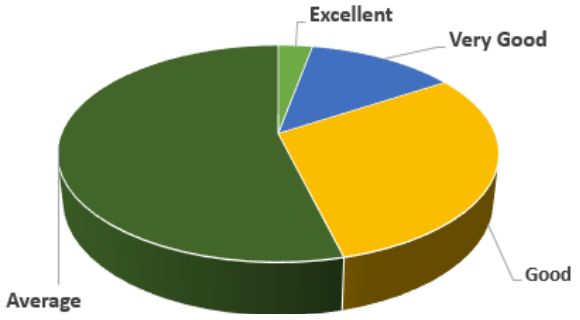

Figure 2: Visualized Data for G3

3. APPLICATIONS OF DATA MINING TECHNIQUES

Data Mining is governed by a fixed set of steps or methodology. All the operations are applied consequently and the results are subsequently obtained [6]. The basic framework used in this paper has been shown in Figure 2.Figure 2 describes the work methodology that is employed here. It begins from obtaining the problem. Then the data regarding the problem is collected. The obtained data is preprocessed to make it compatible to given methodologies. This has already been done in the previous section. Following up by employing pre-decided methodologies of Association, Classification and Clustering, we obtain patterns which are then evaluated. Finally the results are presented as Knowledge Representation.

3.1 Association

Association Rules refer to a particular format of If/Then type statements. These rules can extract relationships between unassociated data. They are of the form If antecedent Then (probable) Consequent. An antecedent is generally found in the data while the consequent follows it [5]. For a student database, we deal with variables that affect Grade 3. Hence the consequent here is Grade = "X".

X can vary between the 5 established Grades: A, B, C, D and E.

As part of the Association method, we now apply Frequent Pattern Growth Algorithm. FP Growth Algorithm is scalable and more efficient at detecting patterns when compared to other pattern detecting algorithms. It uses a 2 step procedure which includes:

1. Building the FP Growth Tree.

2. Extracting frequent item sets from the tree. [7]

```
[G1=D, Home=U, FEdu=4, sex=M] --> [G3=D]
(support: 0.195, confidence:0.757. lift=1.396)

[Home=U, FEdu=4, Alco=4, sex=M] --> [G3=D]
(support: 0.101, confidence:0.754. lift=1.391)

[G1=D, Home=U, Outing=5, sex=F] --> [G3=D]
(support: 0.155, confidence:0.745. lift=1.375)

[G1=A, Home=U, FEdu=4, MEdu=4] --> [G3=B]
(support: 0.105, confidence:0.831. lift=1.449)

[G1=D, Home=U, MEdu=4, sex=M] --> [G3=D]
(support: 0.301, confidence:0.731. lift=1.348)

[G1=D, Home=R, FEdu=0, sex=F] --> [G3=E]
(support: 0.146, confidence:0.716, lift=1.321)

[G1=D, Home=R, FEdu=4, sex=M] --> [G3=D]
(support: 0.103, confidence:0.712. lift=1.314)

[G1=D, Home=U, , FEdu=0, sex=M] --> [G3=D]
(support: 0.133, confidence:0.706. lift=1.303)

[G1=D, Alco=4, Outing=5, sex=M] --> [G3=D]
(support: 0.176, confidence:0.699. lift=1.298)

[G1=D, Home=R, FEdu=4, sex=F] --> [G3=D]
(support: 0.232, confidence:0.677. lift=1.250)
```

Figure 1: Association Rules

The rules which are obtained are then sorted by Lift Metric [8]. Lift Metric is the ratio of the percentage of the correct classification by the model to the percentage of the actual correct classification in database. If the value obtained is greater than or equal to 1, it indicates a strong relationship between the antecedent and the consequent. A sample of Association rules is depicted in Figure 3:

Several observations can be made here for the first sample rule:

- The lift metric is 1.396. This indicates a strong correlation between the antecedent (G1=D, Home=U, FEdu=4, sex=M) and the consequent (G3=D).
- The rule can be interpreted as: For the situation: G1=D, Home=U, FEdu=4, sex=M, which is supported by 19.5% students (support=0.195), will have about 75.7% students (confidence= 0.757) getting a Grade D in G3.

The other association rules can be interpreted in a similar manner.

3.2 Classification

Classification is used for assigning labels to discrete data. It helps by predicting the target class for given records. Classification can only be applied for datasets in which classes are known [7]. In our case, we divided G3 into discrete values of Grades (A, B, C, D and E).

```
If G1<C and G2<C then G3<=C
If (G1<C or G2<C) and Sex=M then G3<=C
If (G1<C or G2<C) and Home=R then G3<=C
If FEdu=4 and Home=U and Alco=0 then G3<=A
and G3>=E
If (G1>=B or G2>=B) and StudyTime>=2 then
G3>=B
If Sex=F and (G1>=B or G2>=B) then G3>=B
If Sex=M and (G1>=B or G2>=B) then G3>=C
If Sex=F and Home=U and FEdu=4 then G3>=B
If FEdu=4 and MEdu=4 then G3>=D
If StudyTime>=3 then G3>=B
If Sex=F and Home=R then G3<=C
If Sex=F and Home=U then G3<=A
If Sex=M and Home=R then G3<=B
If Sex=M and Home=U then G3<=A
If FEdu<=3 and MEdu=4 then G3<=B
If MEdu<=3 and FEdu=4 then G3<=A
If G1<=B and G2<=B and StudyTime>=3
then G3<=A
If G1<=A and G1>=B and G2<=A and G2>=B
then G3<=A and G3>=B
If StudyTime>=3 and FEdu=4 then G3<=A
If Home=U and G1<=C and FEdu=4 then
G3<=B
If FEdu<=3 and MEdu<=3 then G3<=C
If Alco<=5 and Alco>=3 and Home=R then G3<=D
If Alco<=5 and Alco>=3 and Home=U then G3<=C
If FEdu<=2 and Alco>=2 then G3<=D
If StudyTime>=3 and Alco=0 and FEdu=4 then G3<=A
If Home=U and Sex=M and G2<=A then G3<=A
If FEdu=<=3 and MEdu<=3 and Home=R then G3<=C
If Home=U and StudyTime>=3 then G3>=C
```

Figure 2: Classification Rules

Classifications are discrete. Hence they do not imply order. Prediction is used for continuous values as opposed to discrete ones. Predictive models use a regression algorithm.

Rule-based induction is used. Here a set of rules are extracted which show the relationship between class label

and attributes. It makes use of If-Then rules for classification. It is of the following format:

If condition (Antecedent) - Then consequence (Consequent). Given below is the set of Rule-Induction based classification generated rules which show the G3 Grade as the consequent. As seen, the factors that affect the consequent are: Grades G1, G2 and G3, Sex, Home, FEdu, MEdu, StudyTime and Alco. This model possess an accuracy of 71.25% which is acceptable accuracy.

Also Classification rules differ from Association in the regard that Classification rules are used for predicting the state while Association rules are used for assessing the current state.

The above rule can be understood as:

(For the first rule) If Grades for Mathematics and Science are both less than equal to C, then Final Grade can be predicted as less than equal to C.

The advantage of using Classification is that low grades of students can be predicted on time, and teachers can take extra initiative to help these students.

3.3 Clustering

In data mining, clustering is a technique that finds groups of objects in a way that objects belonging to one cluster are more similar to each other than the objects belonging to other clusters. Clustering is used to find high quality of clusters in a way such that Inter-cluster distances are maximized and Intra-

cluster distances are minimized. We will be applying the K-means clustering algorithm which would choose the best cluster center as the centroid. The K-means clustering method used produced a model with 3 clusters of sizes 68, 200 and

```
K-means clustering with 3 clusters of sizes 68, 200, 127

Cluster means:
         G1        G2        G3
1  7.088235  5.294118  2.632353
2  9.810000 10.020000 10.150000
3 14.685039 14.708661 15.000000

Clustering vector:
  [1] 1 1 2 3 2 3 2 1 3 3 2 2 3 2 3 3 3 2 1 2 3 3 3 3 2 2
 [54] 2 2 2 3 3 2 3 2 2 2 2 2 3 3 1 2 3 3 2 1 3 2 2 2 2 2
[107] 2 3 2 3 3 2 2 3 2 3 3 3 2 3 3 3 3 3 2 3 2 2 1 3 1 1
[160] 2 1 2 1 2 1 2 2 3 1 3 1 3 2 1 2 2 2 1 2 2 2 2 3 2 2
[213] 3 2 2 3 3 1 1 2 2 1 1 3 3 3 2 3 2 2 2 3 2 2 3 1 2 3 2
[266] 3 2 2 2 1 2 3 2 3 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 2
[319] 2 2 3 2 2 3 3 2 3 2 2 3 2 3 1 1 1 3 3 1 3 2 2 1 3 1
[372] 3 2 1 3 2 3 2 3 2 3 1 2 1 1 2 1 1 2 1 2 3 2 2 2

within cluster sum of squares by cluster:
[1] 1365.397 1596.200 1255.622
 (between_SS / total_SS =  76.8 %)
```

Figure 4: K-Means clustering with 3 clusters

127.

Clustering should basically have the property of cohesion and separation which occur internally and externally respectively i.e. the between_SS/total_SS ratio should approach 1. We obtain a ratio of 0.768 i.e. 76.8% which is acceptable.

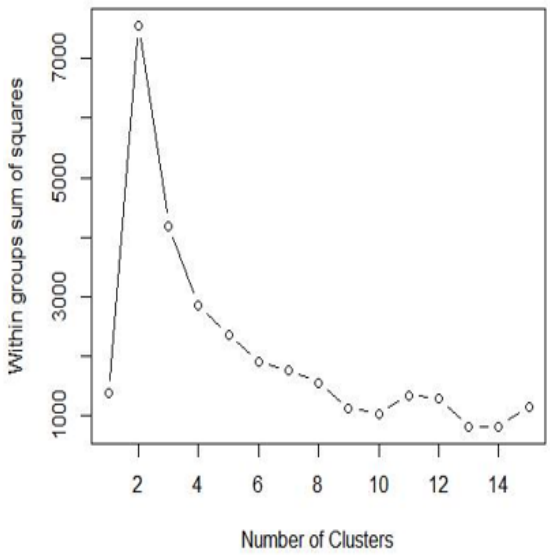

Figure 3: Elbow Criterion

A fundamental question is how to determine the value of the parameter k, which is the optimal number of clusters. Here the function of number of clusters obtained as percentage of variance, we can choose a particular number of clusters in a manner that adding more clusters does not improve modeling of data.

To be more specific, if the variance percentage is plotted against the number of clusters, a lot of information will be added by the first cluster however the marginal gain is bound to drop at some point hence giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". We see a gradual drop from 3 onwards, hence the number of clusters we should be looking for is 3.

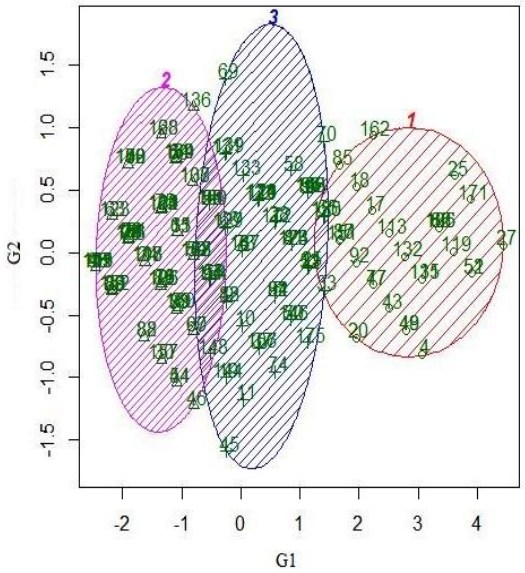

Figure 5: 2-D Representation of the Cluster Solution

After running the K-means algorithm, we see that it shows 3 different clusters of our data points, clustered with other data points that are more similar to them.

4. CONCLUSION AND FUTURE WORK

By the means of this research paper, we've shown how data mining can be used to predict and improve performance of students. This prediction can help students prepare in advance. The data used was collected from a general survey among several educational institutes of India. The analysis was performed by discovering the Association rules for the same using FP Growth Algorithm which were sorted by Lift Metric. This was followed up by Classification through Rule Based Induction Method. Then we performed clustering by the means of K-Means Algorithm. All of these operations are part of Data Mining operations which were used to predict and help students' performance.

Future Work includes using advanced techniques like Neural Nets, Outlier Detection etc. algorithms on a bigger and more generalized dataset. Finally the outcomes could be utilized by educational institutes globally.

5. REFERENCES

[1] Minaei-Bidgoli B.; Kashy D.; Kortemeyer G.; and Punch W., 2003. Predicting student performance: an application of data mining methods with an educational web-based system. In Proc. of IEEE Frontiers in Education. Colorado, USA, 13–18.

[2] Turban E.; Sharda R.; Aronson J.; and King D.,2007. Business Intelligence, A Managerial Approach. Prentice-Hall.

[3] (Ministry of Statistics and programme Implementation - MoSPI, 2012): http://mospi.nic.in/Mospi_New/upload/India_in_figures-2015.pdf

[4] Jing Luan: Data Mining Applications in Higher Education; SPSS Report

[5] Stefanos Ougiaroglou, Giorgos Paschalis; Springer; 'Association Rules Mining from the Educational Data of ESOG Web-Based Application'.

[6] Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques, 2nd edition. Morgan Kaufmann Series, Jim Gray, Series Editor.

[7] Romero, C. and Ventura, S. (2007) 'Educational data Mining: A Survey from 1995 to 2005', Expert Systems with Applications (33), pp. 135-146.

[8] Mining students data to analyze Learning Behacior: A Case Study ALAA EL-HALEES; ResearchatGate

[9] The State of Educational Data Mining in 2009: A Review and Future Visions; Ryan S.J.d. Baker, Kalina Yacef.