Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 3, March 2014, pg.22 – 35

RESEARCH ARTICLE

DATA MINING TECHNIQUES FOR PREDICTING SCHOOL FAILURES AND DROPOUT

Mrs.KALPANA.N[1], Dr.K.DAVID[2]

kalp8na@gmail.com [1], jdbdavid@gmail.com [2]

[1]M.E Student Department of Computer Science and Engineering,

[2]Assistant professor Department of Computer Science and Engineering,

ANNA University Chennai, India.

***ABSTRACT**: In this paper we propose to apply data mining tech-unique to How to Improve Listening skill. We use real data on 100 students from Tamil Nadu College, such as induction rules and decision trees. In experiments, attempt to improve their accuracy for predicting which students Personality Test using all the available attributes. Next selecting the best attributes; and finally, rebalancing data and using cost sensitive classification. The outcomes have been compared and the models with the best results are shown.*

Index Terms: Classification, educational data mining (EDM), Personality Test, Weka tool

I. INTRODUCTION

Recent years have shown a growing interest and concern in many countries about problem of college students How to class listen or not, and the determination of its main contributing factors. The great deal of research has been done on identifying the factors that affect the low performance of students (Listening Skill) at different

educational levels (Anna Univ, deemed Univ and Government) using the large amount of information that current computers can store in databases. All these data are a "gold mine" of valuable information about students. Identify and find useful information hidden in large databases is a difficult task. A very promising solution to achieve this goal is the use of knowledge discovery in databases techniques or data mining in education, called educational data mining, EDM. This new area of research focuses on the development of methods to better understand students and the settings in which they learn. In fact, there are good examples of how to apply EDM techniques to create models that Listening skill and Performance test specifically. These works have shown promising results with respect to those sociological, economic, or educational characteristics that may be more relevant in the prediction of low academic performance. It is also important to notice that most of the research on the application of EDM to resolve the problems. More specifically to online or distance education. However, very little information about specific research on part time and full time education has been found, and what has been found uses only statistical methods, not DM techniques.

There are several important differences and/or advantages between applying data mining with respect to only using statistical models:

1) Data mining is a broad process that consists of several stages and includes many techniques, among them the statistics. This knowledge discovery process comprises the steps of pre-processing, the application of DM techniques and the evaluation and interpretation of the results.

2) Statistical techniques (data analysis) are often used as a quality criterion of the verisimilitude of the data given the model. DM uses a more direct approach, such asto use the percentage of well classified data.

3) In statistics, the search is usually done by modeling based on a hill climbing algorithm in combination with a verisimilitude ratio test-based hypothesis. DM is often used a meta-heuristics search.

4) DM is aimed at working with very large amounts of data (millions and billions). The statistics does not usually work well in large databases with high dimensionality.

The paper is organized as follows: Section II presents our proposed method for Listening skill. Section III describes data used and the information sources from we gathered. Section IV describes the data pre-processing step.

EXISTING SYSTEM:

Starting from the previous models (rules and decision trees) generated by the DM algorithms, a system to alert the teacher and their parents about students who are potentially at risk of failing or drop out can be implemented.

Limitations:

- The problem of imbalanced data classification occurs when the number of instances in one class is much smaller than the number of instances in another class or other classes.

PROPOSED SYSTEM:

In the method we propose that once students were found at risk, they would be assigned to a tutor in order to provide them with both academic support and guidance for motivating and trying to prevent student failure.

In the method we have shown that classification algorithms can be used successfully in order to predict a student's academic performance and, in particular, to model the difference between Fail and Pass students.

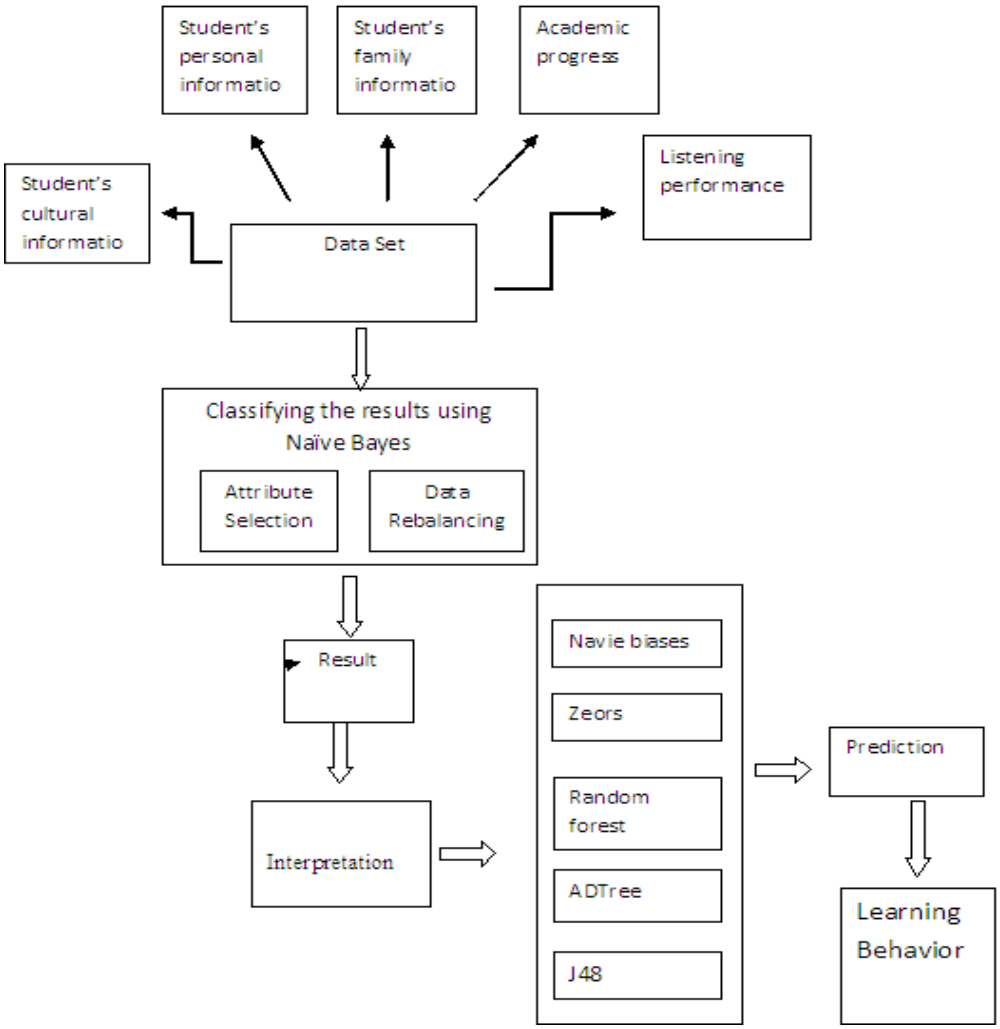
Advantages:

- Data mining is a broad process that consists of several stages and includes many techniques, among them the information.
- In knowledge discovery process comprises the steps of pre-processing, the application of DM techniques and the evaluation and reading of the results.
- DM is aimed at working with very large amounts of data (millions and billions).

At The statistics does not usually work well in large databases with high dimensionality.

II. SYSTEM ARCHITECHTURE

SYSTEM ARCHITECHTURE:

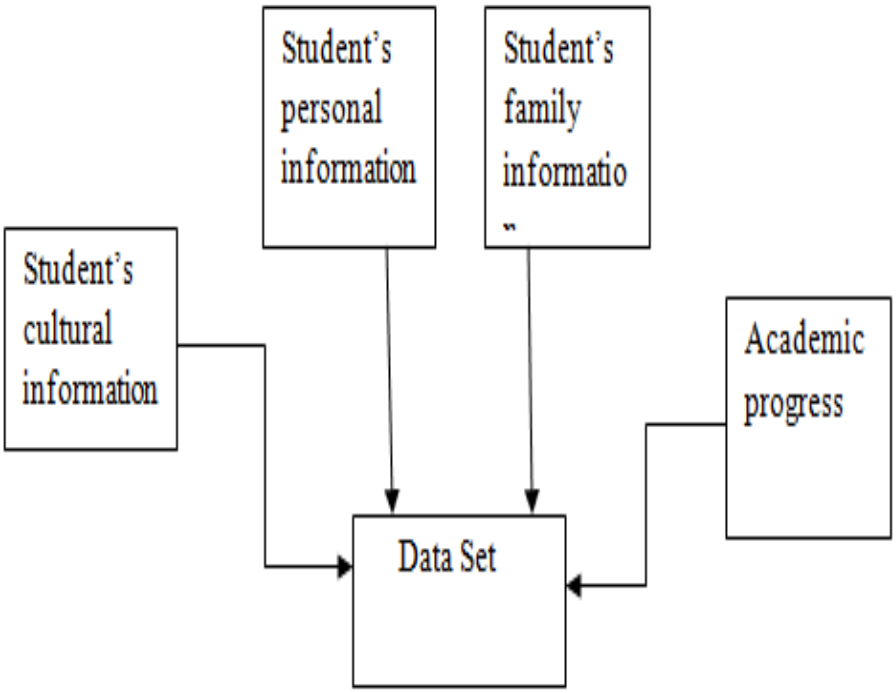


III. RELATED WORKS

A. Data gathering

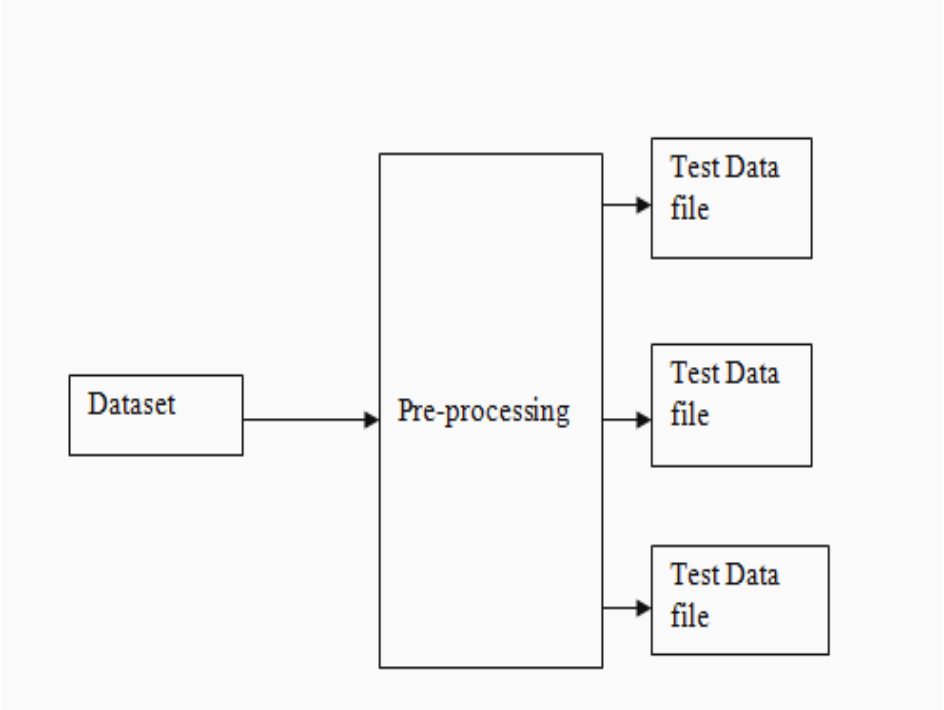
The process of data gathering is that involves in collecting all available information about students .the set of factor should be identified that can affect student's performance and collected from different available data sources .the collected characteristics or risk factors that can influence to students failure or dropped out. Risk factors contain the information about student's cultural, social, educational background, socioeconomic status, psychological profile and academic

progress .In which most of the students are aged between 15 and 16 and this is the years with the highest rate of failure. Finally the survey is to obtain personal and family information to identify important risk factors of all students and school services provides the score obtained by the students in all subjects of course. All those information are integrated into single dataset.



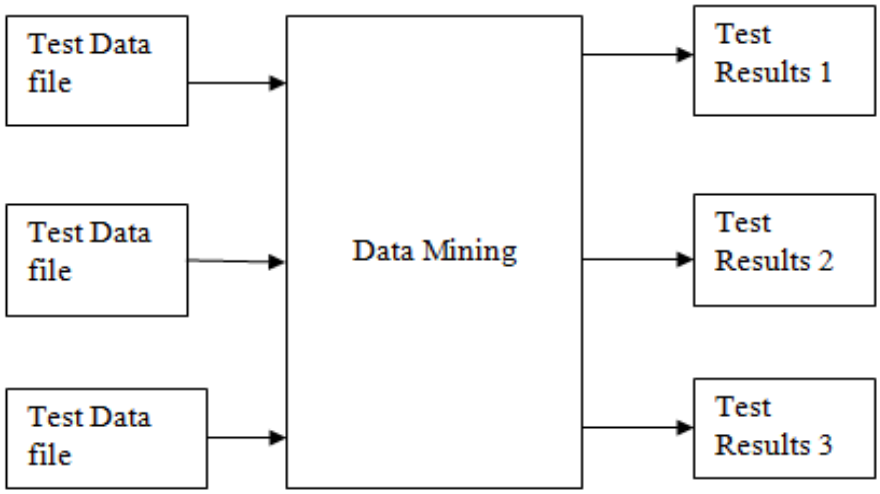
B. Pre-processing

In this stage dataset is prepared for applying data mining technique. Before applying data mining technique, pre-processing methods like cleaning, variable transformation and data partitioning and other technique attribute selection is must be applied. Here new attribute of age is created using date of birth of each students. The continues variables are transformed into discreet variable that is scores obtained by each student is changed into categorical values (i.e) Excellent score between 9.5 and 10,Very good the score between 8.5 and 9.4.all information's are integrated in single dataset that is stored in .arff format of Weka tool. Finally entire dataset is divided randomly into 10 pairs of training and test data files. After pre-processing we have attributes or variables for each student. Each test file will contain best attributes and rebalanced.



C. Data mining

In this stage Data mining technique is going to be applied. Here the data mining technique is mainly used for classification. The classification is based on best attribute selection from data set. In which the naive bays algorithm is implemented for classification of data. Traditionally the Weka Software tool is used for data mining. It contains verity of data mining algorithms.

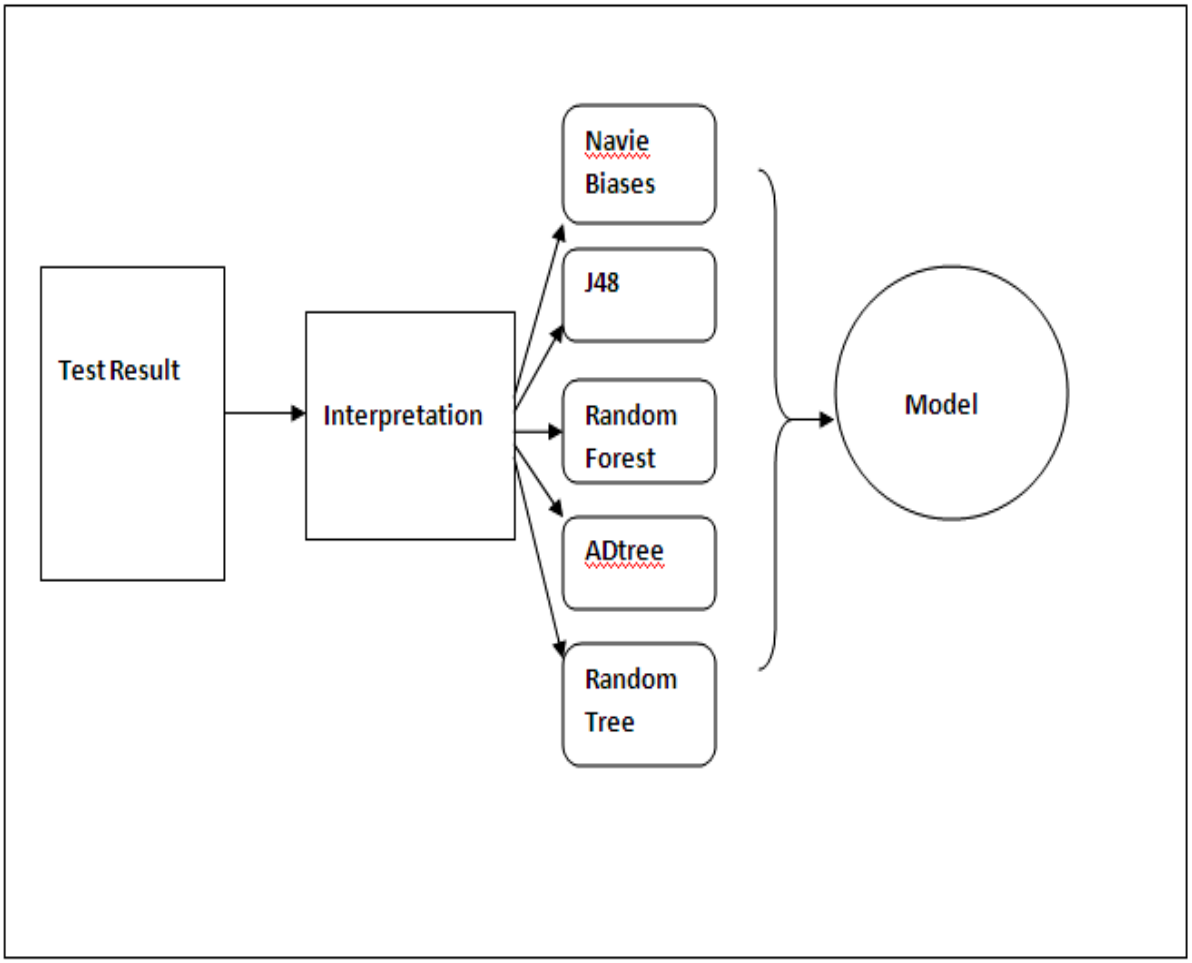


Weka implements decision tree, it is a set of condition organized in hierarchical structure. Decision tree algorithms are like J48, AD Tree, C4.5, Random

Tree etc. Here the classification algorithms were executed using cross- validation and all available information. Finally the result with the test file of classification is shown.

D. Interpretation

In which, the obtained results are analysed to predict student failure or dropped out. To achieve this previous test results are taken for comparison. At this stage classification rules are applied for predict relevant factors and relationships that lead to student pass or fail. There are attribute that indicate that student who failed are older than 15 year and some of the attribute are shows marks of poor, not presented and regular students. Finally the risk factors are analyzed from previous results of classification algorithms.



IV. Data Mining and Experiments

Test

I....	I.....
1. Am the life of the party.	26. Have little to say.
2. Feel little concern for others.	27. Have a soft heart.
3. Am always prepared.	28. Often forget to put things back in their proper place.
4. Get stressed out easily.	29. Get upset easily.
5. Have a rich vocabulary.	30. Do not have a good imagination.
6. Don't talk a lot.	31. Talk to a lot of different people at parties.
7. Am interested in people.	32. Am not really interested in others.
8. Leave my belongings around.	33. Like order.
9. Am relaxed most of the time.	34. Change my mood a lot.
10. Have difficulty understanding abstract ideas.	35. Am quick to understand things.
11. Feel comfortable around people.	36. Don't like to draw attention to myself.
12. Insult people.	37. Take time out for others.
13. Pay attention to details.	38. Shirk my duties.
14. Worry about things.	39. Have frequent mood swings.
15. Have a vivid imagination.	40. Use difficult words.
16. Keep in the background.	41. Don't mind being the center of attention.
17. Sympathize with others' feelings.	42. Feel others' emotions.
18. Make a mess of things.	43. Follow a schedule.
19. Seldom feel blue.	44. Get irritated easily.
20. Am not interested in abstract ideas.	45. Spend time reflecting on things.

Table 1—Performance test

A decision tree is a set of conditions organized in a hierarchical structure. An instance is classified by following the path of satisfied conditions from the root of the tree until a leaf is reached, which will correspond with a class label. Rule induction algorithms usually employ a specific-to-general approach, in which obtained rules are generalized (or specialized) until a satisfactory description of each class is obtained. 10 commonly used classical classification algorithms that are available in the well-known Weka DM software have been used:

5) Five Rule Navie Biases: which is a propositional rule learner; One , which uses the minimum-error attribute for class prediction; Prism, which is an algorithm for inducing modular rules; and Ridor, which is an implementation of the Ripple-Down Rule learner.

2) Five Decision tree algorithms: J48, which is an algorithm for generating apruned or unpruned C4.5 decision tree; SimpleCart , which implements minimal cost-complexity pruning; ADTree, which is an alternating decision tree; RandomTree, which considers K randomly chosen attributes at each node of the tree; and REPTree, which is a fast decision tree learner.

Finally, we have mentioned that our dataset is imbalanced. The problem of imbalanced data classification occurs when the number of instances in one class is much smaller than the number of instances in another class or other classes. Traditional classification algorithms have been developed to maximize the overall accuracy rate, which is independent of class distribution; this means that the majority

of class classifiers are in the training stage, which leads to low sensitivity classification of minority class elements at the test stage. One way to solve this problem is to act during the pre-processing of data by carrying out a sampling or balancing of class distribution. There are several data balancing or rebalancing algorithms; one that is widely used and that is available in Weka as a supervised data filter is SMOTE (Synthetic Minority Oversampling Technique). In the SMOTE algorithm, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any or all of the *k* minority class nearest neighbours. Depending on the amount of over-sampling required, neighbors from the *k* nearest neighbours are randomly chosen.

Experiment and Result:

Type	Tp Rate	FP Rate	Precision	Recall	F-Measure	Roc Area
Zeors	1	1	0.84	1	0.913	0.411
Naïve Bayses	0.833	0.72	0.753	0.833	0.843	0.433
J48	1	1	0.84	1	0.913	0.412
Random forest	0.964	1	0.964	0.895	0.837	0.347
ADTree	0.929	0.875	0.848	0.929	0.886	0.514

Table – 2

Table 3- Weighted Avg.

Type	Tp Rate	FP Rate	Precision	Recall	F-Measure	Roc Area
Zeors	0.84	0.84	0.706	0.84	0.767	0.411
Naïve Bayses	0.74	0.657	0.753	0.74	0.746	0.432
J48	0.84	0.84	0.706	0.84	0.767	0.411
Random forest	0.81	0.846	0.701	0.81	0.752	0.347
ADTree	0.8	0.746	0.752	0.8	0.771	0.514

Table-4 : Confusion Matrix

TYPE	A	B
Zeros	84	0
	16	0
Naïve Bayses	70	14
	12	4

J48	**86**	**0**
	16	**0**
Random forest	**81**	**3**
	16	**0**
ADTree	**78**	**6**
	14	**2**

Table 5- Associate

HaveDifficultUnderstandingabstract											
0	9	5	11.9999	6.9992	2.0008	5	5	8.9999	5	5.0002	1
1	1.0002	1	2.0001	5.0013	6.9983	9.0001	1.0002	5.0001	1	1	8.0012
[total]	10.0002	6	14	12.0005	8.999	14.0001	6.0002	13.9999	6	6.0002	9.0012
Feelcomfortablearoundpeople											
0	1.0002	1	1	6.9988	2.0009	9.0001	5	12.9999	1	1.0001	5
1	9	5	13	5.0017	6.9982	5	1.0002	1	5	5.0001	4.0011
[total]	10.0002	6	14	12.0005	8.999	14.0001	6.0002	13.9999	6	6.0002	9.0012
FnsultPeople											
0	1.0002	5	1	9.9996	6.0001	13.0001	5	1.9999	1	1.0001	1
1	9	1	13	2.0009	2.999	1	1.0002	12	5	5.0001	8.0011
[total]	10.0002	6	14	12.0005	8.999	14.0001	6.0002	13.9999	6	6.0002	9.0012
payattentationTodetails											
0	1.0002	1	1	10.9988	6.0009	6.0001	1	1.9999	5	5.0001	4
1	9	5	13	1.0017	2.9982	8	5.0002	12	1	1.0001	5.0011
[total]	10.0002	6	14	12.0005	8.999	14.0001	6.0002	13.9999	6	6.0002	9.0012
worryAboutThinks											
0	5	5	1	5.9992	6.0008	3.0001	2	1.9999	5	5.0001	4
1	5.0002	1	13	6.0013	2.9983	11	4.0002	12	1	1.0001	5.0012
[total]	10.0002	6	14	12.0005	8.999	14.0001	6.0002	13.9999	6	6.0002	9.0012
HaveAVividImagination											
0	9	1	5	1.9994	3.0004	6	1	8.9999	5	5.0002	2
1	1.0002	5	9	10.0011	5.9986	8.0001	5.0002	5.0001	1	1	7.0012
[total]	10.0002	6	14	12.0005	8.999	14.0001	6.0002	13.9999	6	6.0002	9.0012

Associate: Best rules found

Associate:
Best rules found:

1.dropped=0 84 ==> ImQuckilyUnderstandThings=1 84 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
2.SpendMoneyUseful=1 74 ==> ImQuckilyUnderstandThings=1 74 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
3.RoamwithFriend=1 73 ==> ImQuckilyUnderstandThings=1 73 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
4.FeelOthersEmotion=1 69 ==> ImQuckilyUnderstandThings=1 69 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
5.BadHabitfromFriend=1 68 ==> ImQuckilyUnderstandThings=1 68 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
6.StressedOutEasily=1 67 ==> ImQuckilyUnderstandThings=1 67 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
7.ShrinkMyDuties=0 67 ==> ImQuckilyUnderstandThings=1 67 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
8.interestedInPeople=1 65 ==> ImQuckilyUnderstandThings=1 65 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

Select Attributes:

=== Attribute Selection on all input data ===
Search Method:
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 240
Merit of best subset found: 0.174
Attribute Subset Evaluator (supervised, Class (nominal): 48 dropped):
CFS Subset Evaluator
Locally predictive attributes
Selected attributes: 1: 1

In LFA improved the original model by splitting skill Addresses model improvements even further will some skills be better merged than if they are separate skills. It Can LFA recover some elements of truth if we search from a merged model given difficulty factors, we merged some skills in the original model to remove some of the distinctions, which are represented as the difficulty factors. Circle-area and Circle-radius are merged into one skill Circle Circle-circumference and Circle-diameter into Circle CD Parallelogram-area and Parallelogram-side into Parallelogram, Pentagon-area and Pentagon-side into Pentagon Trapezoid-area Trapezoid-base Trapezoid-height into Trapezoid , In the new merged model has 8 skills Circle Crcle-CD Compose-byaddition and Parallelogram and Pentagon and Trapezoid and Triangle. Then we substituted the original skill names with the new skill name in the data ran LFA including the factors, and had the A* algorithm search through the model space.

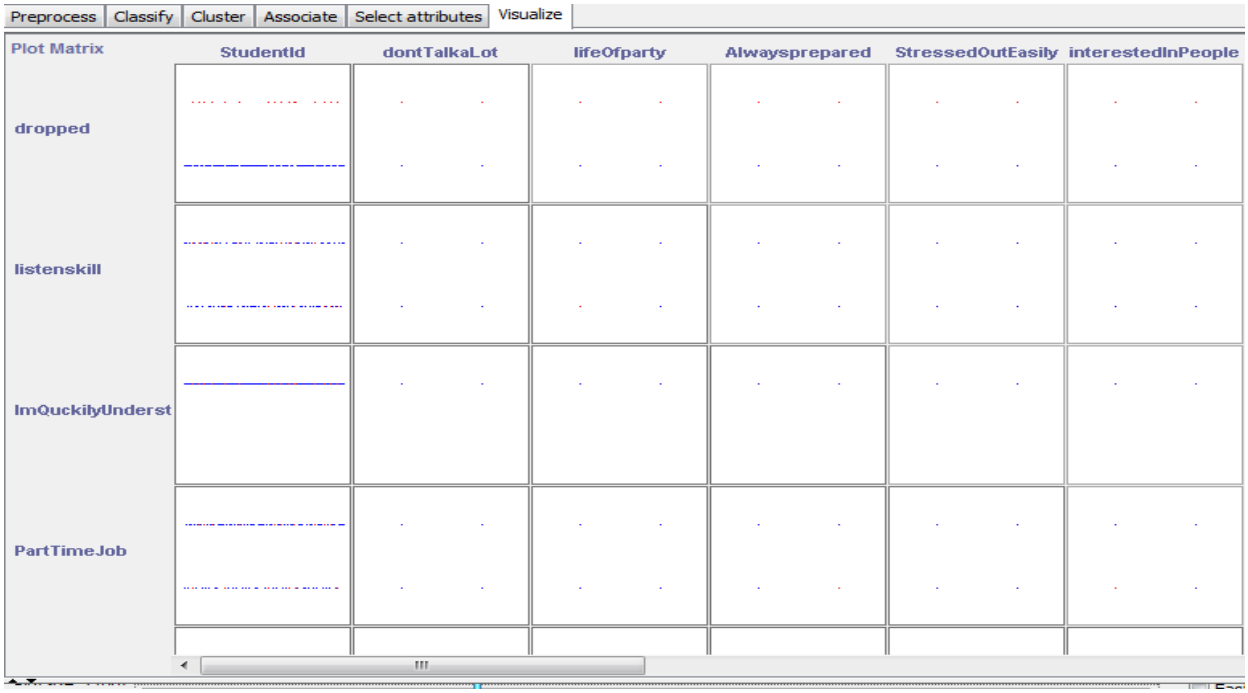


Fig: StudentId

V. FUTURE WORK

Finally, as the next step in our research we aim to carry out more experiments using more data and also from different educational levels (primary, secondary, and higher) to test whether the same performance results are obtained with different DM approaches as future work we can mention the following:

1) To develop our own algorithm for classification/prediction based on grammar using genetic programming that can be compared versus classic algorithms.

2) Predict the student failure as soon as possible in earlier the better, in order to detect students at risk in time before it is too late.

3) To propose actions for helping students identified within the risk group. Then to check the rate of the times it is possible to prevent the fail or dropout of that student previously detected.

VI. CONCLUSION

As we have seen predicting student failure at school can be a difficult task not only because it is a multifactor problem (in which there are a lot of personals and family and social and economic factors that can be influential) but also because the available data are normally imbalanced. The resolve these problems we have shown the use of different DM algorithms and approaches for predicting student failures. We have carried out several experiments using real data from high school students in Mexico and UK. We have applied different classification approaches for predicting the academic status or final student performance at the end of the courses.

Furthermore we have shown that some approaches such as selecting the best attributes cost-sensitive classification and data balance can also be very useful for improving accuracy.

References

1. L. A. A. Aldaco "Comportamiento de la deserción y reprobación en el colegio de bachilleres del estado de baja california: Caso plantel ensenada", *Proc. 10th Congr. Nat. Invest. Educ.*, pp.1 -12 2009
2. F. Araque , C. Roldán and A. Salguero "Factors influencing university drop out rates", *Comput. Educ.*, vol. 53, no. 3, pp.563 -574 2009 [CrossRef]
3 M. N. Quadril and N. V. Kalyankar "Drop out feature of student data for academic performance using decision tree techniques", *Global J. Comput. Sci. Technol.*, vol. 10, pp.2 -5 2010
4. C. Romero and S. Ventura "Educational data mining: A survey from 1995 to 2005", *Expert Syst. Appl.*, vol. 33, no. 1, pp.135 -146 2007 [CrossRef]
5. C. Romero and S. Ventura "Educational data mining: A review of the state of the art", *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp.601 -618 2010 Abstract | Full Text: PDF (410KB) | Full Text: HTML
6. S. Kotsiantis , K. Patriarcheas and M. Xenos "A combinational incremental ensemble of classifiers as a technique for predicting students\' performance in distance education", *Knowl. Based Syst.*, vol. 23, no. 6, pp.529 -535 2010
7. J. Más-Estellés , R. Alcover-Arándiga , A. Dapena-Janeiro , A. Valderruten-Vidal , R. Satorre-Cuerda , F. Llopis-Pascual , T. Rojo-Guillén , R. Mayo-Gual , M. Bermejo-Llopis , J. Gutiérrez-Serrano , J. García-Almiñana , E. Tovar-Caro and E. Menasalvas-Ruiz "Rendimiento académico de los estudios de informática en algunos centros españoles", *Proc. 15th Jornadas Enseñanza Univ. Inf., Barcelona, Rep. Conf.*, pp.5 -12 2009
8. S. Kotsiantis "Educational data mining: A case study for predicting dropout—prone students", *Int. J. Know. Eng. Soft Data Paradigms*, vol. 1, no. 2, pp.101 -111 2009 [CrossRef]
9. I. Lykourentzou , I. Giannoukos , V. Nikolopoulos , G. Mpardis and V. Loumos "Dropout prediction in e-learning courses through the combination of machine learning techniques", *Comput. Educ.*, vol. 53, no. 3, pp.950 -965 2009 [CrossRef]
10. A. Parker "A study of variables that predict dropout from distance education", *Int. J. Educ. Technol.*, vol. 1, no. 2, pp.1 -11 1999
11. T. Aluja "La minería de datos, entre la estadística y la inteligencia artificial", *Quaderns d\'Estadística Invest. Operat.*, vol. 25, no. 3, pp.479 -498 2001
12. M. M. Hernández "Causas del fracaso escolar", *Proc. 13th Congr. Soc. Española Med. Adolescente*, pp.1 -5 2002
13. E. Espíndola and A. León "La deserción escolar en américa latina: Un Tema prioritario para la agenda regional", *Revista Iberoamer. Educ.*, vol. 1, no. 30, pp.39 -62 2002

14. I. H. Witten and F. Eibe *Data Mining, Practical Machine Learning Tools and Techniques*, 2005 :Morgan Kaufman
15. M. A. Hall and G. Holmes *Benchmarking attribute selection techniques for data mining*, 2002 :Dept. Comput. Sci., Univ. Waikato

