Classification and Regression Decision Tree: A Mining Technique for Students' Insights on the University Services with Text Analysis

Dr. Maryli F. Rosas
Computer Studies Department
De La Salle University – Dasmarinas
Dasmarinas, Philippines
mfrosas@dlsud.edu.ph

Dr. Shaneth C. Ambat
Graduate School
AMA University
Project 8 Quezon City, Philippines
shaneth_ambat@yahoo.com

Dr. Alejandro D. Magnaye
Emilio Aguinaldo College
Dasmariñas, Philippines
alex16magnaye@ymail.com

John Renbert F. Rosas
De La Salle University – Dasmarinas
Dasmarinas, Philippines
johnrenbertrosas@gmail.com

***Abstract*—According to Gatpandan, P [1], "the role of the university as a service provider in education sector considers several aspects from student admission to graduate career, and the student is the primary consumer in Higher Education Institution (HEI) services and has implications for the management of service of quality in higher education organizations."

Quality education can be determined thru the quality of services that were given to the students. Satisfaction level of the students can be measured based on their experience all throughout the entire stay in the university.

In educational setting, exit interviews are conducted with students who have graduated from an educational institution. The exit interview is intended to "*gather information about students' experience while attending in the institution, what they benefited from, what was missing, and what could be improved to enhance the experience of the next generation of students. This type of interview can also point to areas in which the institution should invest more or less resources to enhance a student's learning and development experience*." [2]

A University in Dasmariñas has customized exit interview for their graduating students. This exit interview is in the form of questionnaire and used a five-point Likert scale. The strategic value of this Exit interview can be effectively achieve through applying data mining.

Data mining is a process of extracting useful information from huge data [3] and finding patterns. Data mining process can also be applied to educational environment in particular to Higher Education Institutions.

With this, the researchers is motivated to come up with a study that would help the education sector especially the management to address improvements in their institution through applying data mining technique on the Students' insight on the University's Academic and Student Services particularly on the areas of: Facilities, Student services, and Teachers.

Implementation of cross validation 10-folds logistic regression and decision tree analysis were used in this study.

***Keywords*— Data Mining, Decision Tree, Classification, University Services**

I. INTRODUCTION

Quality education requires a continuous quality improvement of every stakeholders. The quality of services rendered by the university would determine the quality of education.

Gaining insights from the students using survey tools would help the university determine the satisfaction level of the students as regards to the quality of university services.

The study would try to solve the following problems: (1) What attributes can be used to determine the satisfaction level of the students. (2) How can data mining techniques be utilized to identify the significant patterns in the feedback or opinions of the students. (3) How can the model help improve the quality of services of the university.

The general objective of this study is to develop a model that will serve as a basis for continuous quality improvement of student services. Specifically, it aims to: (1) To identify the attributes that can be used to determine the satisfaction level of the students. (2) To determine data mining techniques that can be utilized to identify the significant patterns in the feedback or opinions of the students. (3) To implement the model in developing strategies to improve the quality of services of the university.

The study used data mining tools such as SPSS and Weka to train and test data.

II. LITERATURE REVIEW

In the study of Perri, Heijden and Peter [4] they have discussed the classification trees as nonparametric data mining tool for pattern recognition expressed as tree structures. They even cited that since the work of Breiman, Friedman, Olshen, and Stone on classification and regression trees (CART), a great deal of research has been done in this area.

Decision tree according to Romero and Ventura [5], classifies specific entities into particular classes based on the features of the entities. Decision tree technique can be used to extract models to predict future trends or simply describing sequences of interrelated decisions. Some of the most well known algorithms are ID3, C4.5 and classification and regression trees.

Milanović and Stamenković[6] paper shows conceptual features of decision tree. According to them, this decision tree is remarkably suitable in detecting the data structure due to its nature.

Koyuncugil and Ozgulbas[7] presented a data mining risk model for detecting financial and operational risk indicators. They have discussed in their paper the different decision tree algorithms such as the ID3, C4.5, CART or Classification and Regression Tree and CHAID. Classification and Regression Trees or CART, according to them, a relatively new and popular non-parametric analysis technique. CART builds a decision tree and chooses its splits through entropy or Gini metrics for choosing optimal splits and CART generate binary trees.

According to Dimple [8], the decision tree is a powerful classification algorithm that is popular in the information

978-1-5386-7693-6/18/$31.00 ©2018 IEEE

systems. The decision tree is performed with separate recursive observation in branches to construct a tree for prediction. The splitting algorithms – i.e. Information gain (used in ID3, C4.5, C5), Gini index (used in CART), and Chi-square test (used in CHAID) – are used to identify a variable and the corresponding threshold, and then split the input observation into two or more subgroups. In this paper it is identified and evaluated that the most commonly used DM algorithms resulting as well-performing on medical databases, based on recent studies. The following algorithms have been identified: decision trees (DT's), artificial neural networks (ANNs) and their multilayer perceptron model, bayesian networks and naïve Bayes.

Song and Lu [9] in their scholarly works entitled: "Decision tree methods: applications for classification and prediction" introduced frequently used algorithms used to develop decision trees (including CART, C4.5, CHAID, and QUEST) and described the SPSS and SAS programs that can be used to visualize tree structure. According to them, "decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure".

Educational Data Mining according to Rajadhyax and Shirwaikar[10] uses many techniques such as Decision Trees, Neural, Networks, Naïve Bayes, KNearest neighbor, and many others. Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering.

In the study of Lakshmi, Nagesh and VeeraKrishna[11], they have used three data mining techniques namely Artificial Neural Networks, Decision tree and Logical Regression. These techniques were used to elicit knowledge about the interaction between variables and patient survival. A performance comparison of three data mining techniques is employed for extracting knowledge in the form of classification rules. The concepts introduced in their research have been applied and tested using a data collected at different dialysis sites.

Decision trees according to Hong are a great tool for exploratory analysis. He noted that CART can work well with all types of variables be it numeric, ordinal or even nominal. "Variables that have little or no effect on our outcome variable would never be selected in the algorithm. Trees can also handle missing data efficiently." As he compared linear regression with trees, he cited that trees can capture non-linear relationships and interactions without necessarily requiring explicit specification in the modeling process[12].

The approach of Classification and Regression Tree analysis can easily interpreted, computationally driven and practicable method for modelling interactions between health-related variables as stated by authors Kuhn, Page and Worrall-Carter [13].

The paper of Rutkowski, Jaworski, Pietruczuk and Duda presented a new algorithm that is based on the commonly known Classification and Regression Tree algorithm. According to them to solve the problem of determining the best attribute to make a split, thye have applied Gaussian approximation. [14].

Authors Bergner, et.al applied collaborative filtering (CF) to dichotomously scored student response data finding optimal parameters for each student and item based on cross-validated prediction accuracy. [15]

In the paper of Baker, he cited that "classification methods have been used to develop detectors of student affect, including frustration, boredom, anxiety, engaged concentration, joy, and distress. Detectors of affect and emotion have been used to drive automated adaptation to differences in student affect, significantly reducing students' frustration and anxiety and increasing the incidence of positive emotion".[16]

III. RESEARCH METHODOLOGY

All Graduating students across the seven (7) colleges are required to answer interview survey. The goal of this survey is to capture the students' insights, and their satisfaction level on the three (3) different areas such as the facilities, faculty and student services of the university. Applying Decision Tree as classification technique was used in the study. Classification and Regression Tree, also known as C&RT was the main concentration of the study. This is one type of the decision tree that most authors used because of its simplicity.

After applying all the algorithms and be used as inputs for the desired improvement plan of the three areas such as Facilities, Student Services, and Faculty As we all know, educational processes entails continuous quality improvement in a lot of interrelated areas. Facilities will be measured based on the satisfaction of students in terms of the utilization of the physical infrastructures such as but not limited to the library, laboratories (computer, science, architecture, and all the specialized laboratories) that would contribute to the learning of the students during their stay. Moreover, the services experienced by the students from different servicing units like registrar, accounting, security, admission, and other related offices that in a way add up to the success of the students. Teachers on the other hand may use the result to focus on what is really needed to improve their teaching methodology.

With the use of the data mining, it will help the university identify the present and future requirements of students and their preferences to enhance their satisfaction levels.

The descriptive method was used in this study to gather information about the existing conditions of each College and Academic servicing units in the university.

A three-year historical data was acquired and used in this study. Based on the gathered data, attributes that determines the satisfaction level of the students was identified. Weka and SPSS data mining tools are used to help the researchers analyzed the gained data.

SPSS is an effective tool that was used by the researchers to visualized the data in a tree structure using its visualization designer. SPSS is so friendly yet powerful because of the features it has. Aside from the statistics program (used for statistical functions and cross tabulations in the study), it also has the modeler program which helps the researchers to build and validate models.

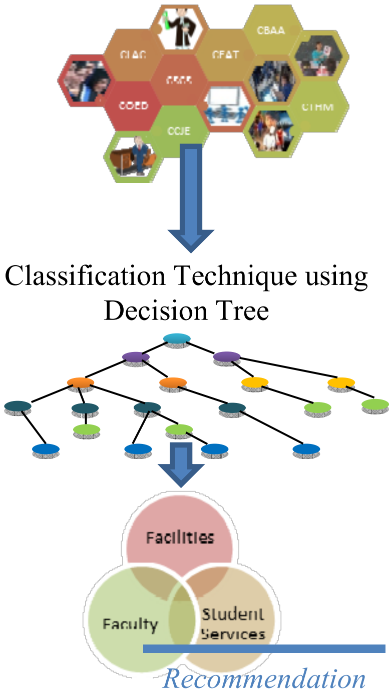

Fig. 1 Conceptual Framework

A data model was built for this study that will serve as a basis of the researchers as it progresses in her study. In the data model, all important fields are identified based on the survey tool. The comment was dichotomized in terms of context and its sentiments. The researchers create additional fields such as the section and items. Each identified area are further classified according to its section.

The researchers also applied classification technique complying the four (4) different steps such as data cleansing, data transformation, pattern discovery and interpretation of results.

Data cleansing is the most tedious part of the study since the comments of the students are in qualitative form. The researchers first handle the missing values, as well as resolve inconsistencies of data. The cleansed data are not ready for data transformation. The comments of the students are first classified according to its context. It was classified to what key area that the student comments would fall. Let us say the comment is “the parking lot is too far”, This comment falls on the Facilities area. And the comment is identified to section : parking lot. All data are transformed into a standard format. The next step is to find pattern. In discovering patterns, different data mining algorithms may be applied.

To perform data analysis, data mining algorithms are to be used. The researchers used regression and decision tree particularly the C&R Tree.

The researchers have used survey questionnaires, interview (structured and unstructured) to help the researchers evaluate the current business processes and systems used within the organization to create and apply valuable information.

The researchers also gathered students' feedback data from school year 2011 to 2016. The students' feedback data would come from different programs under each of the seven colleges. However, since not all the seven colleges has a complete repository of the five year data, the researchers just choose the three year historical data from 2013 to 2016 where all the seven colleges has complete data.

The researchers conducted interviews to the Associate deans, deans, and college secretaries regarding the administering and consolidating the exit interviews.

In developing the system, the researchers used the prototyping methodology. Prototyping is a concept that bases its development by creating prototypes after prototypes until the “perfect” software is created. It is an iterative process in which requirements are converted to a working system that is continually revised through close work between the researchers and the target users. The developed online application was also presented to the respondents for evaluation purposes.

IV. RESULTS AND DISCUSSION

With the use of the data model, the researchers was able to come up with seven (7) attributes or variables. Out of these attributes, the researchers was able to identify the target variable Sentiment and the predictor variables Area, College and Comment.

The researchers used the WEKA Data mining tool. Applying Logistic regression analysis, the researchers was able to derive relationships between variables.

Using 10-fold cross validation logistic regression, below shows the result.

TABLE 1. Summary

Correctly Classified Instances	71670	98.0975%
Incorrectly Classified Instances	139	1.9025%
Kappa statistic	0.9619	
Mean absolute error	0.0203	
Root mean squared error	0.1144	
Relative absolute error	4.0521%	
Root relative squared error	22.8724%	
Total Number of Instances	7306	

With the above results, there were 98.0975 % correctly classified instances and only 1.9025 % were incorrectly classified.

The result of Kappa statistics above is 0.9619 which is greater than 0.81, just like in the first test it is almost perfect agreement. The mean absolute error is only 0.0203 which is forecast with almost minimal error just like the previous test option. The Root mean squared error of this model is 0.1144, hence it also indicates better fit. The result of Relative Absolute Error of 4.0521% and Root Relative Squared Error (RRSE) is only 22.8724% indicates that the model is good.

TABLE 2. Detailed Accuracy By Class

Class	0	1	Weighted Avg.
TP Rate	0.983	0.978	0.981
FP Rate	0.022	0.017	0.019
Precision	0.979	0.983	0.981
Recall	0.983	0.978	0.981

F-Measure	0.981	0.981	0.981
MCC	0.962	0.962	0.962
ROC Area	0.997	0.998	0.998
PRC Area	0.996	0.998	0.997

```
=== Confusion Matrix ===
    a       b   <-- classified as
 3631    61 |    a = 0
   78  3536 |    b = 1
```

The result of TP rate, Recall, and Precision is a good indication that the class was captured and predicted correctly. ROC or Receiver Operating Characteristic measures discrimination, that is the ability of the test to correctly classify those negative and positive sentiments. Since the result of ROC area is nearly 1 (0.997) it represents an almost perfect test. In the confusion matrix, 3,631 instances were correctly classified as negative sentiments, and only 61 were misclassified as negative sentiments. Moreover, there are 3536 instances that were correctly classified as positive sentiments and 78 were incorrectly classified.

To answer problem number 2, the researchers used one of the major components of data mining which is Classification. By using classification, the researchers analyzed a set of data and generate a set of grouping rules which can be used to classify future data. The target attribute Sentiments are classified according to its predictors Area, Comments and Section.

Data Classification method using classification-rule learning involves finding rules or decision trees that partition the training data into predefined classes. This decision tree predicts an outcome and describes how different criteria affects the outcome.

The researchers had used the Classification and Regression Tree. In this case, the researchers used the IBM SPSS to derive the tree. Fig. 2 shows the resulting decision tree.

Node 0 has predicted Area split into two (2) nodes: Node 1 and Node 2. Node 1 has 63.673% positive sentiments and 36.327% negative sentiments. So, if the Area is equal to Administrator or Faculty or Quality of Education then it predicted to have positive sentiment. And if the Area is Facility or Student Services then it is predicted to have negative sentiment.

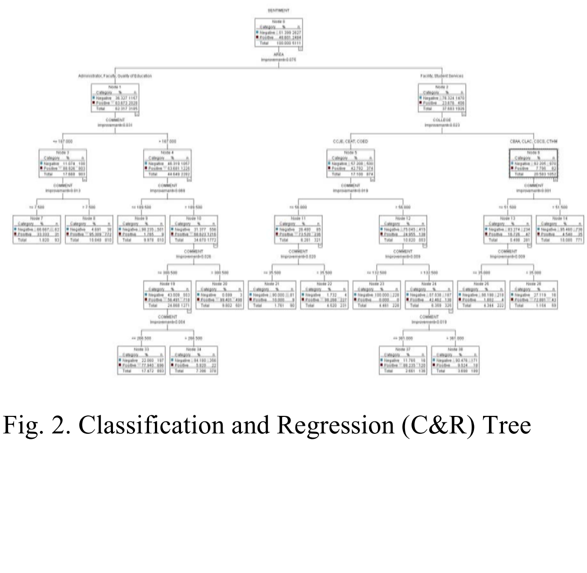

Fig. 2. Classification and Regression (C&R) Tree

From Node 1, it is again partitioned into two splits. The first split covering comments that are less than or equal to 167 (Node 3) while the second split are comments greater than 167 (Node 4). Comments from 1 to 167 are pertaining to administrator, facilities and some faculty. While comments greater than 167 are those that pertain to some faculty, quality of education and student services.

Node 3 is partitioned again into two splits: Node 7 and Node 8. Comments from 1 to 7.5 predicted Negative sentiments (66.667%) while comments greater than 7.5 but less than or equal to 167 predicted positive sentiments with 95.309%.

Same with Node 3, the Node 4 is further partitioned into two splits: Node 9 and Node 10. Node 9 gained 98.235% predicted negative sentiments while node 10 predicted 68.623% positive sentiments. Among these positive sentiments, the highest predicted positive sentiments (99.401%) are those comments greater than 300.50 (Node 20). Therefore, comments that are greater than 7.5 (some comment on administrator), 168 up to 266.5 (comments on faculty up to some quality of education) and greater than 300.5 (comments on quality of education and student services) predicted to have positive sentiments. The Administrator, Faculty, and Quality of Education are three areas where it can be considered that the students are satisfied.

It is shown in this figure that Facility and Student Services are predicted with negative sentiments of 78.324%. Comments that are less than or equal to 35 (comments mostly on facilities) gained the highest negative sentiments on both Node 25 (98.198%) and Node 5 (90.00%) and it was predicted from all of the seven colleges. However, the students from CBAA, CLAC, CSCS and CTHM have greater negative sentiments in terms of facilities which means that they are not that satisfied. Moreover, comments that are greater than 56 less but than 132.5 (still referring to facilities), have predicted a 100% (Node 23) negative sentiment from predicted colleges CCJE, CEAT and COED (Node 5). In addition, comments that are greater than 361 (mostly Student Services) have predicted 90.476% (Node 38) negative sentiments from predicted colleges CCJE, CEAT, and COED.

To answer problem number 3, the researchers used the produced significant patterns of the algorithm as a basis of improvement on the quality of overall services of the university. In addition, the pattern discovered among the dataset was used in developing the recommendation plan of the system.

In order to have a knowledgebase of recommendation, the comments from the historical data plays a vital role. These past comments were analyzed and extracted to generated key phrases. The key phrases were stored in the knowledge base so that when a new comment of the student arises and have the same sentiment score and key phrases, the system could output a learned recommendation. Sentiment score is used to determine the negativity or positivity of the comments. Since the goal of the study is to come up with an automated recommendation, only the negative comments are subject for text analytics. Microsoft Azure Text Analytics API Version 2.0 was used and also integrated in the system. This API uses a machine learning classification algorithm to generate a sentiment score. The sentiment score is between 0 and 1 for each comment, where scores close to 1 is a positive sentiment and scores close to 0 as negative sentiments.

To validate the quality of the developed online system, the researchers asked the respondents to evaluate the system. A 5-point likert scale was used having 5 as strongly agree and 1 as strongly disagree. As a result of the software evaluation, the system gained a 4.72 weighted mean which means that the system was highly functionally suitable, highly efficient, highly compatible, highly usable, highly reliable, highly maintainable and highly secured.

IV. CONCLUSIONS

The researchers used the following technology: SPSS, MS SQL Server 2016 to generate the decision tree. Entity Framework Version 6 for the backend. For the frontend, the researchers used Microsoft Web Application Framework: ASP.NET MVC version 5 Framework and Angular js1.6 Framework for the client side. Bootstrap 3 was used for the report styles, layouts, etc. For the sentiment analysis, the researchers used Quartz.Net which is a scheduling library used in .net This was used to create a job scheduler that fetches the remarks of the students. Text Analytics Service of Microsoft Azure was also used in the study.

The developed online application was evaluated by respondents. The average computed standard deviation results to 0.49 and the weighted mean for all the criteria is 4.72 which can be interpreted that the system gains an overall very high quality rating.

The purpose of this study was to come up with a recommendation plan that would help the education sector especially the management to address improvements in their institution through applying data mining technique on the students' response on the University's Academic and Student Services particularly on the areas of: Quality education, Facilities, Student services, Teachers and Administrators.

Recurring comments from the students all over the university often exist. These comments usually show their sentiments on the Quality of Education they received from teachers, facilities, and services contributing to their satisfaction level.

Weka and SPSS were data mining tools used to mine the students' comments and produce significant attributes and patterns. The generated Classification and Regression Tree corroborates with the result of the developed online system. This simply concludes that the comments of the students three years ago are most likely the same the comments of today. Though the university continuously address those past comments of the students however, the comments and sentiments may still may evolve. This also implies that the quality of education is really a continuous process, that even though the university has already address issues, still there will arise new issues for higher level of improvement. Change is inevitable, yet the good thing with this study, the university can now look forward on the factors that truly affects the satisfaction level of the students.

The researchers highly recommends the following: 1) The developed Online Exit Interview Application should be implemented in the university. And 2) The instrument used in the Online Exit interview may be revised to lessen the number questions given to students.

ACKNOWLEDGMENT

The authors of this research wish to express their utmost gratitude and appreciation to the following individuals: Paulino Gatpandan, Dr. Mengvi Gatpandan, Dr. Johnny A. Ching, Ms. Marivic Mitschek for their utmost support. To our parents especially mommy Lita in heaven, parents-in-law, husband Romel and kids for their unending support towards the end of this research. Their patience, love and understanding which encouraged the researchers to finish this research. Thank you! And most especially, to the heavenly Father for His guidance. Without Him all of this will never be possible.

REFERENCES

[1] Gatpandan, P. and Ambat, S. (2017). "*Mining Disciplinary Records of Student Welfare and Formation Office: An Exploratory Study to Enhance University Services Portfolio.*" International Journal of Information Technology, Control and Automation (IJITCA) Vol. 7, No.2, April 2017 DOI:10.5121/ijitca.2017.7202 23

[2] Definition of exit interview. https://en.m.wikipedia.org/wiki/Exit_interview

[3] Gangurde, R. A. ,Sonar, M. R. (2014). "*Knowledge Extraction using Data Mining.* " Department of MCA, K K Wagh Institute of Engineering Education and Research, NashikMaharashtra, India. Spvryan's International Journal of Engineering Sciences Technology (SEST) SEST Issue 1 , Volume 1 - Oct. 2014 http://spvryan.org/Issue1Volume1/12.pdf

[4] Perri, P.F. and van der Heijden,P. G.M. (2012). *"A Property of the CHAID Partitioning Method for Dichotomous Randomized Response Data and Categorical Predictors University of Calabria, Italy*", The Netherlands Journal of Classification 29 (2012). DOI: 10.1007/s00357-01 - 9094-8

[5] Romero and Ventura (2013), *"Data mining in education.",* WIREs Data Mining Knowl Discov 2013, 3: 12–27 DOI: 10.1002/widm.1075

[6] Milanović, M. And Stamenković, S. (2016), "*Chaid Decision Tree: Methodological Frame And Application*", Economic Themes 54(4): 563-586. DOI 10.1515/ethemes-2016-0029.

[7] Koyuncugil, A.S., Ozgulbas, N. (2009), *"Risk modeling by CHAID decision tree algorithm.*" ICCES, vol.11, no.2, pp.39-46

[8] Dimple (2014), *"A Review on Data Mining Techniques Used in Healthcare Industry SHIV SHAKTI",* International Journal in Multidisciplinary and Academic Research (SSIJMAR). Vol. 3, No. 1, February- March -2014 (ISSN 2278 – 5973)

[9] Song,Y. and Lu, Y. (2015), *"Decision tree methods: applications for classification and prediction",* Shanghai Archives of Psychiatry, 2015, Vol. 27, No. 2 • 130 Biostatistics in psychiatry (26) [Shanghai Arch Psychiatry. 2015; 27(2): pp. 130-135.

[10] Rajadhyax and Shirwaikar (2012) Data Mining on Educational Domain. Cornell University. Page 1.

Library. https://arxiv.org/ftp/arxiv/papers/1207/1207.1535.pdf.

[11] Lakshmi K.R., Nagesh, Y. and VeeraKrishna, M. (2014). *Performance Comparison Of Three Data Mining Techniques For Predicting Kidney Dialysis Survivability,* International Journal of Advances in Engineering & Technology, Mar. 2014. ©IJAET ISSN: 22311963. 242 Vol. 7, Issue 1, pp. 242-254

[12] Hong, M. (2018), *"Exploratory data mining with Classification and Regression Trees (CART)",* Psychological Science Agenda, American Psychological Association. http://www.apa.org/science/about/psa/2018/04/classification-regression-trees.aspx

[13] Kuhn, L., Ward, J and Worrall-Carter, L. (2013), "The process and utility of classification and regression tree methodology in nursing research", Wiley Journal of Advanced Nursing, Nov. 17, 2013 doi. 10.1111/jan.12288

[14] Rutkwoski, L., Jaworski, M., Pietruczuk, L. and Duda, P. (2013). *"The CART decision tree for mining data streams",* Information Sciences Volume 266, 10 May 2014, Pages 1-15, Elsevier, ScienceDirect .https://doi.org/10.1016/j.ins.2013.12.060

[15] Bergner, Y., Droschler, S. Kortemery, G., Rayyan, S., Seaton, D., Pritchard D., *"Model-Based Collaborative Filtering Analysis of Student Response Data: Machine-Learning Item Response Theory"*. Proceedings of the 5th International Conference on Educational Data Mining,

[16] Baker, Ryan., *"Mining Data for Student Models"*.http://www.upenn.edu/learninganalytics/ryanbaker/Baker-for-AITS-v8.pdf