Analysis of student performance using machine learning approaches for Outcome Based Education

Joshi Neeti Dattuprasad
M.Tech Student
Institute of Technology
Nirma University, Ahmedabad, Gujarat 382481
Email: 15mcei09@nirmauni.ac.in

Priyanka Sharma
Professor at CE Department
Institute of Technology
Nirma University,Ahmedabad, Gujarat 382481
Email: priyanka.sharma@nirmauni.ac.in

***Abstract*—Outcome based education (OBE) refers to the analysis of student performance on the basis of Program outcomes, Course Learning outcomes, Assessment Matrix and rubrics for each course. Data analysis can assist in terms of predicting and analyzing the performance based on machine learning algorithms. This helps in incorporating student performance with learning outcomes and program outcomes and classify them by quality indicators which represents how much goal is achieved by studying. However, these analysis take into consideration, only the academic performance of the students. An adaptive approach that incorporates both academic information and personal characteristics of the student can be used for a more precise prediction. By using the different data mining algorithms, the accurate prediction and more significant analysis could be procure. This paper presents a comparative study of data analysis for OBE and experimental results of OBE modules. The experimental results contain the predictive analysis, data analysis and comparative analysis of student performance for an intricate analysis of the OBE based implementation.**

***Index Terms*—Data analysis, Outcome based education, Learning outcome, Program outcome, Data mining and Machine learning Algorithms, LMS and Adaptive LMS.**

I. Introduction

Outcome Based Education aims that student should be able to determine and understand the specific course work as per the program outcomes (PO) and Learning outcomes (LO). The OBE system is largely accepted and implemented in different rural areas of the countries like Australia (25 years), Europe (4 years), Hong-Kong (11 years), Malaysia (8 years), South Africa (6 years), United States (17 years), Pakistan (just started), India (2 years).

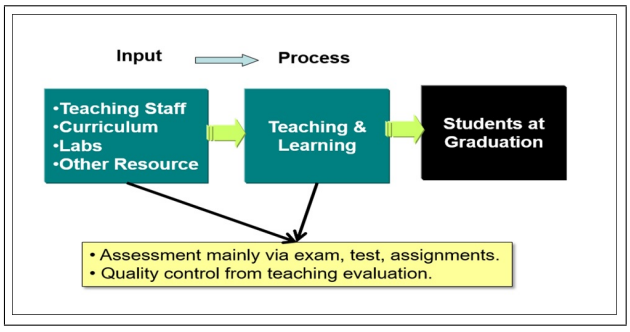

Fig. 1: Traditional education system

The figure 1 depict the process of the traditional education system. It belongs to the classroom teaching system in which teacher teaches the students in the classrooms and labs as per the designed syllabus. Students prepare different subjects from the textbook for the exams. To became a graduate, student should obtain minimum passing grade. The assessment of the student's performance is on the basis of exams, tests and assignments. The interactive assessment is not a part of the evaluation system which in fact is necessary for the assessment of a student intelligence towards the curriculum. The assessment system can improve the quality of education. Due to monotonous structure of the exam, the assessment of students become formal because the exams contain the questions from the question bank (normally direct questions), which are asked repetitively and the students get through easily.

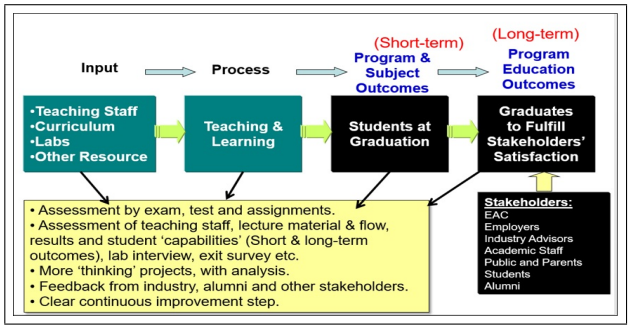

Fig. 2: Outcome based education system

Figure 2 represents the process of outcome based education system. It contains e-learning, classroom teaching, quizzes, group discussion etc. The conceptual learning is more important in the teaching process. The outcome based education contains large number of modules to grasp the entire course-work. To complete the course-work, each module should be successfully finished. Along with the classroom teaching, the students also learn from the videos, group discussions, quizzes to get in-depth knowledge of the course work. To improve on the concepts and practical knowledge of the subject, the projects are given on the respective topics. Due to indirect learning process of Outcome based education, the evaluation

process is rather made complicated to achieve the best accurate results. The evaluation is based on tests, assignments, project work, seminars, group discussion, laboratory work and quizzes. Then the mapping of the students performance is done on the basis of program outcomes and learning outcomes defined by the course curriculum. Stakeholders also give their feedback, whether the students are desirable and compatible for corporate world. At each and every step, student is getting assessed and mapped with program outcomes and learning outcomes. All these evaluation steps are required to clear the course successfully. Though the assessment process of outcome based education is clumsy, it turns out to get the quality result.

Outcome based education is implemented by learning management system (LMS) which provide the services to teachers, students and admin for on-line exams, analysis of student's performance, e-tutorials and so on.

In this paper, the theoretical perspective of OBE using machine learning will be discussed with comparative studies and the experimental results outcome will be obtained using data-set of students. The present study deals with the analysis of student's performance using different machine learning algorithms like k-means, polynomial regression, neural network, apriori algorithm, decision tree and standard deviation.

II. Literature Review

Midhun, et al [1] have identified the academically at risk student and predictive analysis of 9th class of CBSE students. They have found out the result using k-means algorithm and multiple regression algorithm. They have defined six clusters to classify the students who contain similar marks. From that, one cluster contains the students who have low marks. So, that cluster would be the result of at-risk student. To find out the GPA of 4th test, they used the training dataset which contains the student id, name and marks of four tests. Another dataset is a test dataset which contains the student id, name, and the marks of three tests. So using the multiple regression, they have predicted the 4th test marks. The challenge for this method is less accuracy of multiple regression algorithm for prediction. The accuracy can be improved by using Neural Network algorithm.

Dimokas, et al [2] proposed the statistics, data warehousing and mining methodology to find out the students who fail to complete their study in stipulated time and they compared the male and female who complete their study more faster using Pearson correlation and cross tabulation method. They conclude that 58 percent students complete their study after 4 years (because of strong negative correlation between duration and students) and female completed their study before male. The another algorithm to find the students who would fail to complete their study in stipulated times is using k-means clustering algorithm which would give the clusters similar performance of the students so from that we can definitively detain students and the students who can be detained.

Camilo, et al [3] have found out the students who have low academic performance and the subjects in which students frequently fail using decision tree and Naive base classifier algorithms. The result is maths and social science are the subjects in which student frequently fail and students who are younger (23-25 years) have low academic performance. The Apriori algorithm can also be used to get the frequent subjects in which students frequently failed. Apriori is less complex than decision tree and nave based classifier to find frequent pattern mining.Nguyen, et al [4] suggested the analysis of four algorithms when the dataset is inconsistent. They compared four algorithms K-means, SOM, OCFSCM and NPFSCM to find out which algorithm is more consistent for incomplete dataset. They proved that SOM and K-Means is more consistent if the dataset is incomplete more than 50 percent.

Rover, et al [5] proposed survey paper which contains the assessment by ABET at Iowa State University. They have defined different objectives and quality indicators to classify the students. They compare the assessment of the student and teacher and they conclude that the teacher has more knowledge than student the result of the survey is hierarchical assessment gives a more accurate result. But they didnt classify the objectives into LO and PO. So there is no structural analysis of students. Romero, et al [6] presented the survey paper of in journal of the IEEE. They explained the data mining techniques, distant learning, e-learning, text mining, web mining and a traditional education system. They also describe the different data mining tools. They did a comparative analysis of which data mining task is appropriate to which educational system.

Guleria, etal [10] proposed the feedback system from the student. Here feedback (for different parameters like teaching , intrastate, department, institute) of the student is calculated by using the standard deviation to check the significance of the feedback.Almeida, et al[7](2010) suggested This paper contains the survey of AEHS system and they have d that Category Theory is very useful in AEHS system. But the limitation is that, for hierarchical assessment the automated tool is needed to implement category theory. To overcome this challenge computer aided engineering tool is needed for adaptive navigation.

Karampiperis et al [8] presented a description the concept competence description ontology and learners competence records using Competence Description Ontology. They conclude that Less success rate in the complex competence level and highest success rate in simple competence level.Ioannis, et al [10] compared the Learning management system and Adaptive educational hypermedia system features. They conclude that LMS uses only academic information while AEHS uses academic and personal information too. So AEHS gives more accurate results than LMS. But the limitation is to use the AEHS is, very difficult to implement and it is not coasting most effective. So to overcome the challenges of LMS and AEHS is to use adaptive LMS system.

Table 1 demonstrate the comparative analysis of the different literature survey.

Author	Title	Methodology	Remarks
Midhun[1]	A Big Data Approach for classification and prediction of student result using Map Reduce	• Algorithm: K- means Map Reduce Algorithm, Multi leaner Regression Algorithm • Characteristics: (1) Identifying academically at risk student(2) Predictive Analysis of student result. • Input: Dataset of CBSC Student of 10th class • Result: (1)There are 6 clusters of students which have similar marks by which we can defined at risk student (2) From GP1, GP2, GP3 the GP4 has predicted	Here the Learning analysis and Predictive Analysis is used to identify academically at risk students which helps who need special attention to enhance student learning outcomes. The challenge is the accuracy of the multiple regression algorithm.
Nikolaos[2]	A Prototype System for Educational Data Warehousing and Mining	• Algorithm: Cross Tabulation Method, Pearson Correlation • Characteristic: Goal: (1) find the students who have failed to complete their studies in stipulated time (2) find out who have completed their study more faster. OLAP and statistical operations are used to find the useful knowledge • Input: Dataset of 1184 Greek students • Result: (1) As per the Duration curve analysis the Female will complete their studies earlier than male. (2) 58 percentage students have complete their studies after 4 years	By statistical analysis we can find out the relationship between 2 variables
Ioannis[3]	Comparing LMS and AEHS-Challenges for improvement with exploitation of Data Mining	• Characteristics: Compression between two system. That is LMS Adaptive Educational Hyper (Learning Management System) and AEHS (Adaptive Educational Hypermedia System). In LMS the Student's characteristics(knowledge level, goals, Learning style) is totally ignored.	AEHS gives more accurate result as compare to LMS
Camilo[4]	A Model to Predict Low Academic Performance at a Specific Enrolment Using Data Mining	• Algorithm: Nave Based Classifier, Decision Tree • Characteristics : Goal :The main aim is to find out the students who have blocks and have low academic performance. • Input: (1) Initial Academic Information (2) Demographic and Scio-Economic Information (3) Academic Potential . • Result: (1) Poor Maths and Social science leads to loose in academics (2) Younger students (age: 23-28) have poor academic record	The Naive based classifier Algorithm gives more accuracy. The accuracy of classifier improved when the academic data was added
Vo[5]	A robust and effective algorithmic framework for incomplete educational data clustering	• Characteristics: Goal: Main aim is grouping of the students on the basis of learning behaviour, skills, performance, Preference. The robust algorithm is developed for incomplete dataset • Algorithm : (1) K means (2) SOM (3) OCFSCM (4) NPFSCM • Input: (1) D: Incomplete Dataset (2) K: number of clusters (3) Threshold. Output: (1) C: Number of clusters (2) D: Complete Dataset. • Result : K-means and SOM algorithm are used as robust algorithm for incomplete dataset.	K-means and SOM algorithm are more robust for dataset which are more than 50 percentage incomplete. K-means is hard version of OCFSCM and SOM is hard version on NPFSCM
Rover[6]	Implementation and Results of a Revised ABET Assessment Process	• Characteristics: This paper includes ABET assessment process at different levels. The objectives and Quality Indicators/ Rubrics of the institute are provided. • Input: The student Performance and Teacher performance are mapped with the objectives and they are classified into the different classes as per the Rubrics. • Result: Hierarchical assessment reduce the effort and it is effective	Having a small committee of faculty knowledgeable about the accreditation process adds significantly to the quality of assessment results.
Guleria[8]	Increasing Quality of Education Using Educational Data Mining	• Algorithm: Standard Deviation • Characteristics: In this paper the quality of education is rated by the students. • Input: There are 3 Quality parameters. (1) Teaching skills (practical knowledge, Responsiveness, Cooperative) (2) Infrastructure (Library Facilities, Classrooms, Labs) (3) Contents (course content, Course Material) • Output: Student rated this parameter out of 10	Using Feedback approach the quality of the education can be measured and improved.
Pythagoras[10]	Adaptive Learning Objects sequencing for competence based Learning	• Algorithm: Competence Description Ontology • Characteristics: It is the concept • competence description ontology and learners competence records • Input: 50 simulated instances • Output: Less success rate in complex competence level and highest success rate in simple competence level	Learning objects sequencing is used to recognise lifelong learning

TABLE I: Comparative Survey

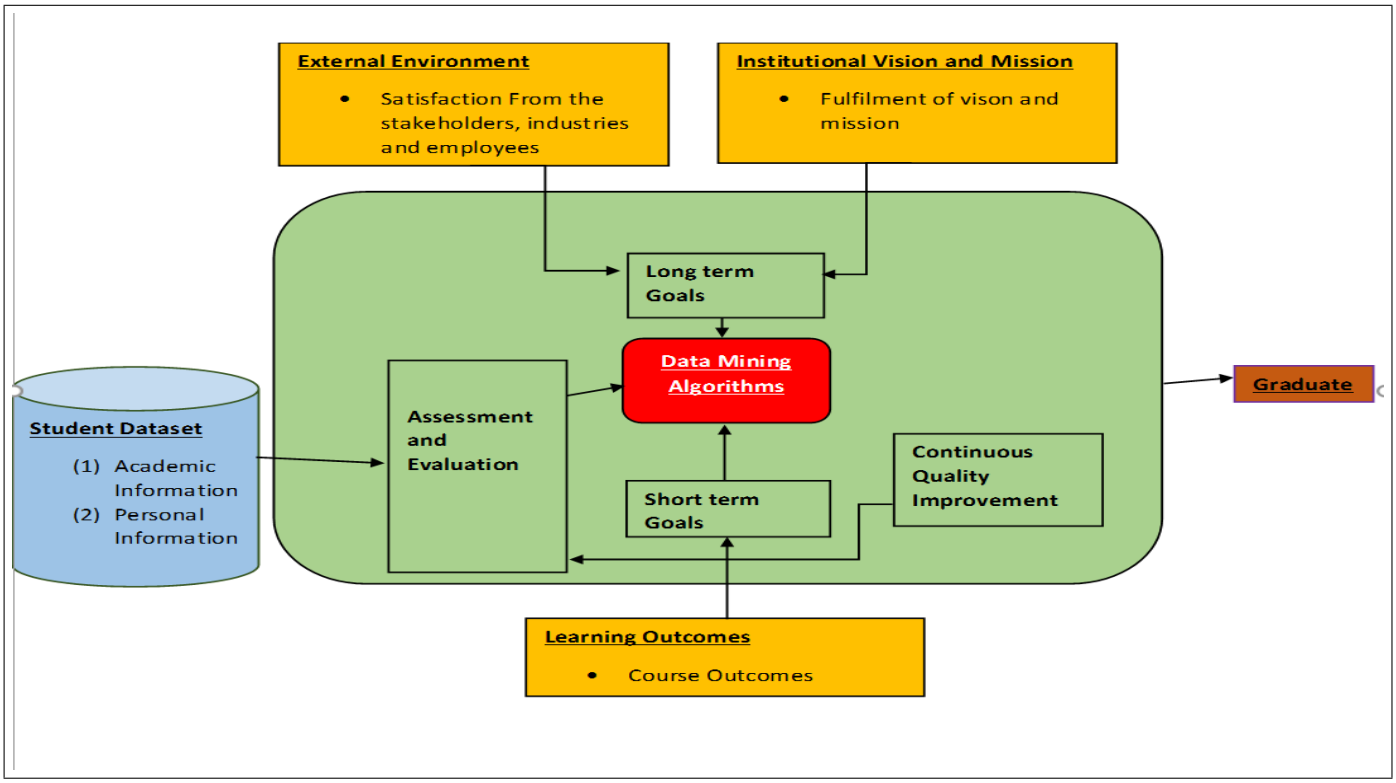

Fig. 3: Architecture of Proposed Work

III. Proposed Frame work

Prediction and analysis of the achievement of student performance based on program specific outcome and course learning objective.

The Figure 3 illustrates the proposed architecture. The academic and personal information is taken as student data. The student performance would be assessed and mapped with learning outcomes [LO] and program outcomes [PO]. LO would be course outcomes and PO would be institutional vision and mission and goals from the external environment. This mapping would be executed using the data mining algorithms. The main intent is to check continuous improvement of the student. At the end, the result analysis would be carried as the output.

IV. Methodology

This section carries the experimental approach of the comparative study. To get the results of student's performance, some machine learning algorithms are used in this experimental study. For the experimental purpose of data analysis, following five objectives are selected. 1) Find the frequent subjects which are the causes of low academic performance, 2) Discriminate the marks of at risk students, 3) Prediction of the result and comparative analysis of algorithms for accurate prediction, 4) Comparison between learning management system with Adaptive approach, 5) Analysis of Quality of Education using student's feedback.

A. *Data Collection and Preparation*

For experimental purpose, the dataset of the students is used. The first and the fourth objectives use the categorical dataset while second, third and fifth objectives use the numeric dataset, assuming no missing values in each dataset.

B. *Data Selection and Transformation*

For 1st objective, the dataset incorporates the four attributes i.e., Name, Maths, science, English. This dataset holds the information about pass or fail (i.e. Categorical attributes) in maths, science and English. So here the intention is to find out which two subjects together are the cause of failure out of three subjects. For 2nd objective, the student dataset comprises the five attributes which are id, maths, science, English and total. All the records contain numeric values. Maths, science and English contain the scores out of 100 and total contains the marks out of 300 which is total of three subjects. For 3rd objective, the dataset contains the training and test data. Training data contains five attributes id and 1st, 2nd, 3rd and 4th semester SPI and test data contains the four attribute id, 1st, 2nd and 3rd semester SPI. Here 4th semester SPI in training data is actual value, and we will predict 4th semester SPI from test data to determine the accuracy of the algorithm. In the 4th objective there are two datasets i.e. LMS dataset and Adaptive LMS dataset.

Table 2 illustrate the dataset of LMS. LMS contains the four attributes: Name (Neeti, Vidhi, Ruchi), Subjects (Maths, Science, English), Level (advance, average, poor) Result (pass,

Name	Subjects	Level	Pass
Neeti	Maths	Advance	Pass
Neeti	Science	Average	Fail
Neeti	English	Poor	Fail
Ruchi	Maths	Average	Fail
Ruchi	Science	Advance	Pass
Ruchi	English	Poor	Fail
Vidhi	Maths	Poor	Fail
Vidhi	Science	Average	Fail
Vidhi	English	Advance	Pass

TABLE II: dataset of LMS

fail). For example, If Neeti knows the advance level of maths then she always passes. As per the dataset if any of the student knows advance level of subjects then they will pass the exams and if they have average or poor knowledge then they will fail.

Name	Subjects	Level	Habits	Rubrics	Pass
Neeti	Maths	Advance	Gadgets	High	Fail
Neeti	Maths	Advance	Emotional	Medium	Pass
Neeti	Science	Average	Gadgets	Low	Pass
Neeti	Science	Average	Emotional	High	Fail
Neeti	English	Poor	Gadgets	Medium	Fail
Neeti	English	Poor	Emotional	High	Fail
Ruchi	Maths	Average	Gadgets	Medium	Pass
Ruchi	Maths	Average	Emotional	Low	Pass
Ruchi	Science	Advance	Gadgets	High	Pass
Ruchi	Science	Advance	Emotional	Medium	Pass
Ruchi	English	Poor	Gadgets	High	Fail
Ruchi	English	Poor	Emotional	Medium	Pass
Vidhi	Maths	Poor	Gadgets	Low	Pass
Vidhi	Maths	Poor	Emotional	High	Fail
Vidhi	Science	Average	Gadgets	Medium	Fail
Vidhi	Science	Average	Emotional	High	Fail
Vidhi	English	Advance	Gadgets	Medium	Pass
Vidhi	English	Advance	Emotional	Low	Pass

TABLE III: dataset of Adaptive LMS

Table 3 demonstrate the dataset of adaptive LMS. Adaptive LMS dataset comprises the same attribute as LMS and two more attributes: Habits (gadgets, emotional) and Rubrics (high, medium, low). For example, If Neeti knows the advance level of maths and her usage of gadgets is very high then she will fail. For 5th objective, using feedback for the teachers from the students. There are five attributes: Id, Knowledge, Cooperation, Delivery, Responsiveness which contain the student's rating out of 10.

C. *Implementation of Model*

In first objective, the apriori algorithm is used for frequent pattern mining of subjects which are the causes of failure. In second objective k-means algorithm is utilized. The number of clusters is 5. In third objective there are two algorithms used for comparative analysis of machine learning algorithms i.e., Polynomial regression and neural network. First, the training dataset would apply to the model of a polynomial algorithm to train the model, then the test data would be apply. On the basis of training dataset, the prediction of test dataset would be obtained. In fourth objective, decision tree algorithm would be used and in fifth objective the standard deviation will be counted, to find the significance of the feedback.

V. Performance Analysis

The result of 1st objective would be maths and science subjects. We have taken 0.1 support-count.The 18 iteration would be performed for frequent pattern mining. There are 10 students who failed in both maths and science. In 2nd objective, there are 5 clusters taken by default. Each cluster would contain the students having similar marks in the specific subjects. For example, by using k-means algorithm, the centroid of each attribute can be found. So the student who obtains the marks near to the centroid value, that will occupy the cluster. The cluster 0 will contain the at risk students. The Id of the at-risk students are 1001(32 marks), 1003(40 marks), and 1019(106 marks)

4rth sem(Real Value)	Predicted Value	Accuracy(+/-0.5)
7.9	6.284	0
8.2	8.245	1
7.9	10.447	0
8.4	7.699	0
6.8	5.578	0
7.4	11.038	0
8.7	5.378	0
8.2	8.483	1
7.8	8.208	1
7.9	7.502	1

TABLE IV: Accuracy-Polynomial Regression

In third objective as shown in table 4, the accuracy of polynomial regression algorithm would be 40 percentage and root mean squared value would be $2.00 + / - 0.00$. As shown in table 5, the accuracy of ANN algorithm would be 80 percentage and root mean squared value would be $0.39 + / - 0.00$. So neural network is more accurate for prediction of SPI.

4rth sem(Real Value)	Predicted Value	Accuracy(+/-0.5)
7.9	7.890	1
8.2	8.212	1
7.9	7.890	1
8.4	7.699	0
6.8	7.504	0
7.4	7.288	1
8.7	8.743	1
8.2	8.159	1
7.8	8.146	1
7.9	7.431	1

TABLE V: Accuracy using Neural Network

As shown in figure 4, the decision tree of LMS represents the level of subjects, which are having more impact on the result. If the level of the subject is higher, then all the students will pass and if the level will average or poor,then students will fail. Below is the literature report of decision tree of learning management system. So level of knowledge has more impact in the result.

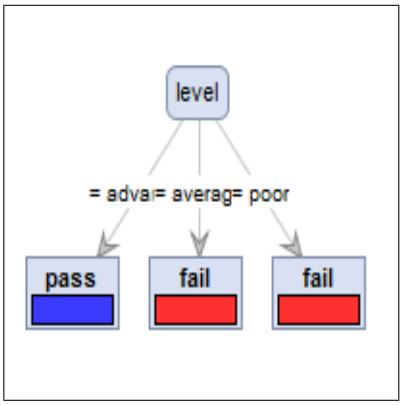

Fig. 4: Decision Tree of Learning Management System

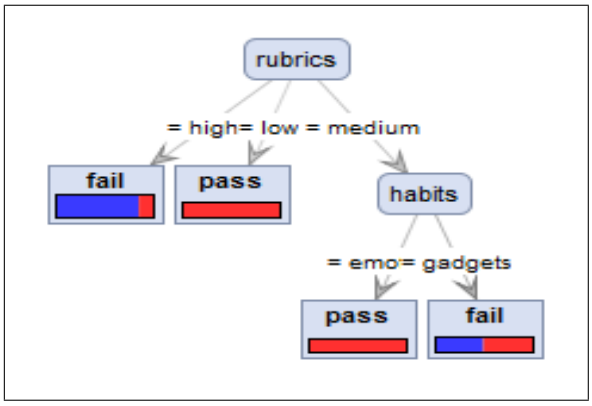

Fig. 5: Decision Tree of Adaptive Learning Management System

As shown in figure 5, tree of adaptive LMS represents the rubrics of habits, which would have the most impact. If the student's habit is severe (high) then in the 6 cases, students will fail and in 1 case student will pass. If the student's habit is low, then all will pass. If the student's habit is moderate (medium) then the result will depend on the specific case of the student. Here in one case, Neeti and Vidhi will pass while fail in another case. But Ruchi will always pass. In 5th objective, we got 1.6 standard deviation value, which is very small. So the result is significant feedback from the students. The literature result is as shown as below.

VI. Conclusion

By using the machine learning algorithms, the data analysis of student's performance gave the following results. We found the frequent subjects which are causes of failure using apriori algorithm. Also we found at-risk students using K-Means clustering algorithm. We obtained more accurate algorithm for predictive analysis of SPI using polynomial regression and neural network algorithm. After analysing the both LMS and Adaptive LMS dataset, the adaptive(behavioural) approach is more significant in the student performance and significance of the student's feedback can be measured by standard deviation using statistical analysis. So, different machine earning algorithms provide more accuracy in evaluation system for outcome based education.

VII. Future Scope

The future work can be carried out on the dataset of any learning management system for data analysis of student's performance. This can be achieved by increasing the number of machine learning algorithms, which may give more accurate results that can be checked on different data set. These algorithms will also provide the comparative accuracy between adaptive (behavioral) data and academic data.

References

[1] M. M. Mohan, S. K. Augustin, and V. K. Roshni, A bigdata approach for classifi-cation and prediction of student result using mapreduce, pp. 145150, 20.

[2] N. Dimokas, N. Mittas, A. Nanopoulos, and L. Angelis, "A prototype system for educational data warehousing and mining," pp. 199-203, 2008.

[3] C. E. L. Guarffn, E. L. Guzmfan, and F. A. Gonzfalez, "A model to predict low academic performance at a specifc enrollment using data mining," IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 10, no. 3, pp. 119-125, 2015.

[4] V. T. N. Chau, N. H. Phung, and V. T. N. Tran, "A robust and efective algorithmic framework for incomplete educational data clustering," pp. 65-70, 2015.

[5] D. T. Rover, D. Jacobson, A. Kamal, and A. Tyagi, "Implementation and results of a revised abet assessment process," 2013.

[6] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005,"Expert systems with applications, vol. 33, no. 1, pp. 135-146, 2007.

[7] P. Guleria, M. Arora, and M. Sood, "Increasing quality of education using educational data mining," pp. 118-122, 2013.

[8] M. A. F. Almeida and F. M. de Azevedo, theoretical model of the adaptive navigation support," pp. 195-200, 2010.

[9] K. Phythagoras and D. Samson, "Adaptive learning objects sequencing for competence-based learning," 2006.

[10] I. Karagiannis and M. Satratzemi, "Comparing lms and aehs: Challenges for improvement with exploitation of data mining," pp. 65-66, 2014.