# *Predicting Students' Academic Performance Through Supervised Machine Learning*

Engr. Sana Bhutto
*Department of Software Eengineering*
*Mehran University Of Engineering And Technology*
Jamshoro, Pakistan
sanabhutto163@hotmail.com

Dr. Isma Farah Siddiqui
*Department of Software Eengineering*
*Mehran University Of Engineering And Technology*
Jamshoro, Pakistan
isma.farrah@faculty.muet.edu.pk

Dr. Qasim Ali Arain
*Department of Software Eengineering*
*Mehran University Of Engineering And Technology*
Jamshoro, Pakistan
qasim.arain@faculty.muet.edu.pk

Maleeha Anwar
*Research Assistant, Dept. of Computer Science,*
*University of Karachi*
Karachi, Pakistan
famemaleeha@gmail.com

***Abstract*— There are many supervised and unsupervised types of machine learning approaches that are used to extract hidden information and relationship between data, which will eventually, helps decision-makers in the future to take proper interventions. The variety of powerful algorithms used in different areas of daily life that includes our educational system as well. This paper introduces students' academic performance prediction model that uses supervised type of machine learning algorithms like support vector machine and logistic regression. The results supported by various experiments using different technologies are compared and it is showed that sequential minimal optimization algorithm outperforms by achieving improved accuracy as compared to logistic regression. And the knowledge found through this research can help educational institutes to predict the future behavior of students so that they can categorize their performance into good or bad. The objective is not just to predict future performance of students but also provide the best technique for finding the most impactful features to work on like teacher's performance, student's motivation that will eventually decrease the student's dropout ratio.**

***Keywords*— *Education Data mining, Learning Analytics, Prediction model, machine learning***

## I. INTRODUCTION

Learning Analytics plays major role in improving educational system by focusing the different perspective e.g student perspective, teacher perspective and administrative perspective. Student's proper assessment, a clear understanding of educational problems, selecting, and planning proper interventions at the right time are few goals of learning analytics. For this regard data analytics and different fields associated with it like knowledge discovery in databases, analysis of future prediction predictive, text recognition and mining, neural network based and artificial intelligence techniques now have seeped into the nerves of the educational system. The major goal is to achieve a technological shift from traditional learning and teaching practices towards automation. The student's data can be collected through online learning management system. We have massive collection of data but we are hardly able to extract useful information from it, the data generated through different sources contain valuable hidden information that increased the research interest towards field of knowledge discovery in databases, the extracted information can help educational practitioners and decision makers as stated by Padhy, Mishra, and Panigrahi in [1].

Educational data mining is an emerging research area composed of a big set of psychological and computational approaches for providing a roadmap of how students learn. The Latest automated interactive learning tools like creative games, simulation-based applications, and intelligent tutoring systems have given ways to analyze and discover student's data and patterns those data contains. Online learning management systems contain a variety of students related data and valuable features that affect the performance of students that can be used in the building of prediction model.

In [2] Baker, Gowda and Corbett defined few goals of educational data mining and to achieve those goals different methods are used: 1) students' future performance prediction model can be created for class prediction based on a training set and test data set. These models are especially helpful in online tutoring systems for understanding student educational outcomes where student behavior or interaction with system like participation in discussion groups, engaging in sample tests, going through provided materials and answering the questions in quizzes plays very important role in predicting which student will pass or fail a class. 2) Clustering is used to group data points having similar attributes or closet values e.g clustering applications can be used to identify the students having same type of patterns while interacting with any online learning management system so that the students can be categorized based on their learning approaches. 3) Relationship Mining is used to identify relationships between data points and considering them as a rule so that they can be used later. In 2009, Baker and Yacef [3] stated that relation mining is used in identifying relations between different parameters that affect the students' performance in any learning management system. Basic two techniques of relation mining(association rule mining and sequential rule mining) are used generally. In 2010, Mercercon and Yacef [4] defined one technique association rule mining that is used to find mistakes done by students jointly, used to create recommendation systems for course selection based on student needs and learning approach. Sequential pattern mining builds the rules that find

978-1-7281-6899-9/20//$31.00 ©2020 IEEE

links between existence of consecutive events like the types of mistakes students do sequentially and help they seek.

## II. RELATED WORK

In 2009, El Halees [5] worked on Moodle case study by applying association rule mining and clustering technique where the author compares the traditional educational system with online learning system and observed that the academic performance of students is enhanced with the use of private tutoring. In 2010, Romeo and Ventura [6] stated that the educational data is in a hierarchical form and that's the unique feature of it. The data is linked and nested with one another like classroom-related data, students related data, teachers related data. And the other three most important features are based on time, series and context. Time represents how long student relates to classroom, login and logout details, time spent on practice tests or discussion groups. Series represents how tutoring system and learning process are organized. And context is an important concept in examining the result of prediction model where it works or where it is failed. In 2014, Herrmannova and Hlosta [7] worked on the model that identify at-risk students on distance learning platform and the identified students are sent to the module team for interventions that are possible. The results were represented by designing the mockups of dashboards. But the main limitation of this research work is it does not consider any important set of student's features but dealt with all feature sets. In 2015, Amerieh, Thair, and Ibrahim [8] presented a research work considering a predictive model that works on students' behavioral features and analyze the impact of those features by applying three classification algorithms like Decision Tree, Naïve Bayes, and Artificial Neural Network and compare the results. The data gathering process was done by XAPI learner activity tracker tool and use of activity tracker tool made the model complex. Failed to achieve good accuracy. In 2015, Hayan Agarwal and Mavani [9] stated that there exist some limitations to above algorithms because when input is given to classification algorithms like Bayesian in a continuous range it affects the overall accuracy of algorithm. In 2016, Amrieh, Thair, and Ibrahim [10] which is the continuing work of above-defined authors, introduce a prediction model. The model applies data mining techniques for evaluating student's behavioral features by classifying students' academic data. The model achieved improved accuracy but with the extra use of ensemble methods to solve problems. In 2018, Bendangnuksung [11] introduce a Deep Neural Network which was a linear classifier model to predict the academic performance of students. Extra support of different tools, libraries, and Ubuntu operating systems increases the model complexity and all the feature set were considered during analysis. So far there is a lot of work that has been done in education data mining field where different models and algorithms are used to predict students' academic performance but unfortunately unable to identify important features that affect the student's performance. So, this research work introduces a model that will identify the features which should be focused on analyzing students' academic performance and based on these features the students are categorized into three different classes i.e. good, average, and bad.

## III. PROBLEM STATEMENT

One of the major challenges that educational institutes are facing now-a-days is the rapid and exponential growth in educational data and the problem to apply that type of data to improve the overall quality of education system as stated by Baradwaj and Pal in [12]. Therefore, these challenges of learners and students can be addressed through proper analysis of educational data by having insights or tests hypothesis or model on a dataset. Data analytics helps organizations or educational institutes to find and understand data in enormous databases using their reporting capabilities and using that information to built models so that they can be applied to predict individual's performance and behaviors towards their studies with high accuracy. So that the recourses can be allocated properly and effectively by institutes. The System does not offer a path between course curriculum and the student who is learning course so that student can have access to knowledge sources, authority to change the contents of curriculums, and reconstructs it based on his or her own needs and interests. The teacher is behaving like knowledge transferring medium, whereas teacher is main pillar of any educational system, having the responsibility of facilitating in teaching process. The educational system shift from conventional to outcome based system can be successfully implemented if all involved actors like fulfills their responsibilities and play their roles properly as stated by Malan in [13].

## IV. METHODOLOGY

In this research work, we have proposed a model based on classification to predict the student's future performance, which can be used to categorize students based on their academic record. Fig. 1 shows the steps of proposed model, which will help educational decision-makers to analyze the factors that can affect the student's academic progress and can make proper interventions timely. The first step is collecting the data from an online educational source followed by preprocessing step, i.e., cleaning of variables. This step is followed by selecting an appropriate feature set and at last, to build the model data mining approach e.g classification technique is used and algorithms like Support Vector Machine using Sequential Minimal Optimization and Logistic Regression.

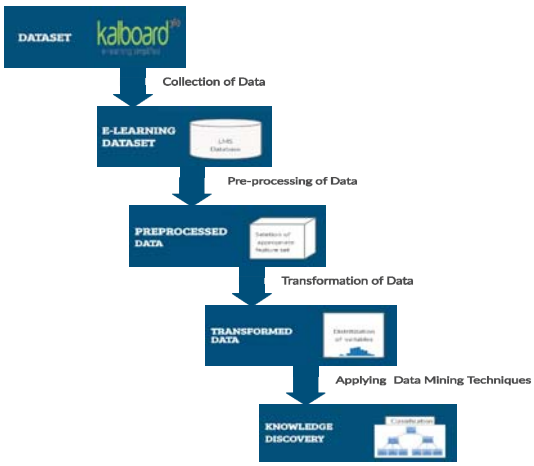


Fig. 1. The student performance model having KDD steps.

### A. Collection of Dataset

The rapid increase in digital technologies and internet usage has created the vision of an online learning platform. Kakasevski, Mihajlov, Arsenovski, and Chungurski in [14] defined the concept of learning management system has simplified online learning process. The major responsibility of LMS is to manage, control and track student and instructor activities and interaction with system as stated by Rapuano and Zoino in [15]. The dataset used in this paper is acquired from e-learning system known as Kalboard 360[16]. The dataset contains 500 records and having 16 distinct attributes/features. The features are further divided into four different categories: 1) Students' demographical features which include gender, country, birthplace, and parent of the student (mother or father). 2) Academic background feature includes level of education(low, medium and high), section ID(A, B or C), student grade level(G-1 to G-12), semester of student(1st or 2nd), courses offered by LMS(IT, Maths, Science, Arabic, English), punctuality of student in class 3) Parental participation category includes parents involvement and satisfaction ratio gained through surveys 4) Interaction feature includes all the parameters which shows students' engagement in the classroom or participation in class activities like participation in group discussions, viewing material uploaded, attempting quizzes on time, raising hands and submitting assignments Improvement of student's performance class especially in online learning management system partially or fully depends on the frequency of interaction with system and motivation or encouragement student have to engage with classroom activities.

### B. Preprocessing of Data

Data preprocessing is a very important task that is used before applying machine learning algorithms. Data preprocessing involves further data cleaning, data transformation, and feature selection. In data cleaning step, the dataset is checked for any null or redundant values. After data cleaning process the dataset is left with 480 records. The 20 records were removed from the dataset which contains missing values. The dataset contains 16 attributes of different categories.

*1) Data Cleaning:* Data cleaning is one of the necessary parts of transformation, data cleaning is process of eliminating irrelevant or null records from data. The dataset used in this work contains 500 records of students. After this step, the final version of dataset left with 480 records remaining 20 records contains null values so those records are separated from dataset.

*2) Feature/Attribute Extraction:* Student interaction is considered one of the most important aspects in achieving progressed outcomes in learning process. In [17] Gunuc and Kuzu defined student interaction as "the quality and quantity of students' psychological, cognitive, emotional and behavioral reactions to the learning process as well as to in-class/out-of-class academic and social activities to achieve successful learning outcomes". So in this research among 16 different attributes, the main focus would upon the most appropriate set of features that plays vital role in students' academic success or failure.

Feature extraction is among widely used preprocessing techniques that should be applied before mining the dataset so that the overall quality of mined patterns and time required in mining is improved. Data preprocessing is noteworthy process in knowledge discovery because having quality data can help to make quality decisions. The objective of feature selection or attribute reduction is to know the minimum space of attributes because mining such reduced and appropriate set of features has extra benefits. It can help to form patterns simpler to understand because of having reduced number of attributes in found patterns. It can also increase the overall accuracy of classifiers and defined by Han, Jiave, and Kamber in [18].

Methods that are used for data space reduction categorized into two types: Wrapper based method and filter-based method. The working of wrapper-based method is in such a way that it uses the technique of classification which is used to evaluate the value of feature set, so the features selection purely depends upon the classifier itself. Wrapper method provides better results and performance in comparison with filter methods because the process of feature selection is amended for the classification techniques or algorithms, which are used. Whereas, wrapper based techniques proved to be an expensive type of technique in terms of massive databases or increased dimensionality so it costs more in terms of time and computational complexity because every feature available in feature set which is considered needs to be examined along with used classification algorithm. On the other hand, filter-based techniques are used before the process of classification. This approach does not depend upon the machine learning algorithm, it is considered to be simple, computationally easier and fast. Filter method provides feature selection, which further can be used as input to different classification algorithms. So far, many feature extraction and feature ranking techniques have been used among which few famous techniques are Information Gain, Gain Ratio attribute selection, Correlation Based Feature Selection, Euclidean Distance, Principal Component Analysis and so on. In our research work, we have used gain ratio technique for ranking the features available in our dataset to know the impact of those features on students' success. Fig. 1 shows the rank value associated with different features. Sine two categories interaction feature and students' parent's participation in learning process have got highest ranks among all 16 features. This shows that both categories of features have great impact on the progress of students' academic performance since ranks are calculated while considering students' marks in semester.

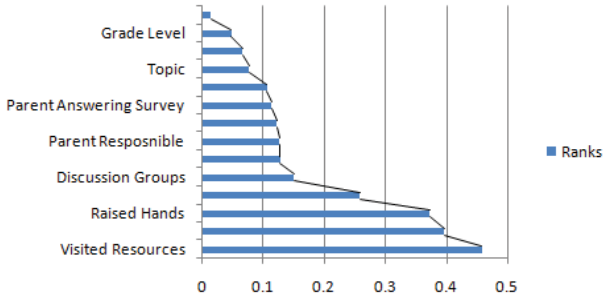


Fig. 2. Feature ranks associated with each feature.

*3) Data Transformation:* Data transformation is also considered one of the main steps in preprocessing process. So the dataset used in this research is educational dataset used for students' performance prediction where students' cgpa or marks in previous semesters should be considered to predict students' future performance that's why student marks which were numerical values are converted into nominal values to represent the class labels for classification and prediction. In Table 1 we have categorized the marks of students into three classes (Good, Average, Bad).

TABLE I. CLASSES BASED ON STUDENTS' NUMERICAL MARKS VALUES

| **Three Classes** | |
|---|---|
| ***Student Marks*** | ***Class Labels*** |
| 0-69 | Bad |
| 70-89 | Average |
| 90-100 | Good |

*4) Applying Data Mining Techniques:* Multiple supervised and unsupervised techniques used for extracting hidden patterns from enormous databases. In this research work, we have used a supervised type of machine learning approach i.e. classification with three class intervals(good, average and bad) for all variables or attributes like Romero, Ventura and Garcia in [19]. As many machine learning algorithms are used to train the student predictive models so after analyzing many of them for comparison purposes we have selected logistic regression and support vector machine using sequential minimal optimization two widely known classifiers for prediction.

*a) Logistic Regression:* Logistic regression is commonly used where categorical target variable is used and is considered most famous parametric method. Generally, logistic regression models are linear models used for the prediction of some categorical dependent value based on some independent predictor value. It can solve binary(two classes) classification problems as well as multinomial(two or more classes) values problems. Logistic regression model works on the concept of logit function for finding the linear combination or probabilities of occurrence of some targeted class value based on predictor values(continuous or discrete).

*b) Support Vector Machine(SMO):* Support vector machine also known as SVM is one of the most popular machine learning algorithms used for classification, prediction and regression. SVM algorithms have great usage in text recognition or categorization and bioinformatics because of these algorithms best work with analyzing large datasets having many predictor attributes. The SVM model is considered maximum-margin model which is discriminative and works on the concept of classifying data point into two or more classes by locating an optimistic decision boundary which is known as N-dimensional hyperplane and it should be at large distance from data point of each class. Vectors considered as support vectors that are near the hyperplane. It is considered as binary linear classifier having non-probabilistic nature. When dealing with non-boundaries kernel function is used to map data point actual space to new feature space. In this work, we are using sequential minimal optimization that because it is considered best maximum-margin hyperplane solver by breaking the problem space into N-dimensional subspaces without using any other algorithm for optimization.

## V. RESULTS AND ANALYSIS

The results used in this research work are based on the student prediction model using two different classification techniques have been calculated through two different tools.

### *A. Setting Environment*

For performing the experiment, we have used PC with 8GB RAM, core i7 processor. Further two different approaches are used to calculate the evaluation measures and accuracies so that prediction model gives better results for students' future performance prediction.

*1) Weka tool:* Weka tool provides simple, interactive and easy to understand environment as defined by Arora in [20]. It is a machine-learning tool it provides improved accuracies and better prediction results. Weka tool has many different options for preprocessing, classification, clustering, attribute selection, and visualization. We have transformed and cleaned the dataset by converting numerical values into nominal for proper results. Attribute extraction is also done using weka, we have selected gain ratio techniques and got the appropriate set of attributes. Further, we have used classification technique based on two classifiers(Logistic regression and SVM). In classification, weka provides two options for training and testing the dataset with the name of "use training set and supplied test set". Different options are available for evaluating performance measures few of them are cross validation using different number of folds, percentage split which is used to split the dataset into two partitions one is used for training the dataset and other is used for test purpose. We have used percentage split for splitting the dataset. 70% of dataset has been used for training the model and remaining 30% is used for testing purposes.

*2)* Jupyter using python: Jupyter Notebook is web-based software having an open source. It is used for creating and sharing live codes on Github, the code can contain equations, text, and graphs. It supports Julia, Python and R programming language. Jupyter works with IPython kernel to write codes in python language whereas it supports more than about 100 kernels. Among different distributions of python language CPython also known as Python 3 and Anaconda are few widely used. We have processed and run the results using Python 3. We have used pandas which is a famous python package it provides data analysis manipulation with flexibility having powerful data structures. Pandas library provides data frames and NumPy to represent a different record. Also imported Matplotlib which is 2-dimentional plotting library of python to produce good quality figures and graphs, sklearn python library is used for supervised machine learning algorithms, this library is based on pandas, numpy, and matplotlib technologies. Using sklearn preprocessing and model selection is done. classification_report, confusion _matrix, and accuracy_scores libraries are also imported from

sklearn.metrics library. Further fit_transform function is used to transformation as this function calculates means of all columns and replace the missing values with those means, hence cleaning transforming the dataset. To split dataset into training and test or predict set we have converted data frames into an array using df.values. At last, we have evaluated the model using logistic regression and support vector machine algorithms and print the classification report, accuracy score and confusion matrix.

### *B. Evaluation Measures*

Results and quality of applied algorithms on proposed model are evaluated using four different measures like recall, precision, f1-score, and accuracy. Accuracy is considered as most effective performance measure and defined as ratio of correctly classified outcomes to the total number of inputs. Precision is defined as ratio of accurately predicted positive outcomes to the total positive predicted outcomes. Recall is the ratio of accurately positive predicted outcomes to all outcomes in a respected class. Weighted average of precision and recall is called f1-score in case of uneven class distribution f1-score is considered more useful.

### *C. Results*

In this section, different results have been evaluated based on two machine learning algorithms using two different technologies. As we have compared these algorithms based on three evaluation measures recall, precision and f1-score, these measures are calculated using true positive, false positive, true negative and false values of confusion matrix. Fig. 3 shows the comparison of these three measures between two classification algorithms conducted using Weka tool. Fig. 4 shows the comparison of results between logistic regression and support vector machine algorithm using python 3 run on Jupyter notebook.

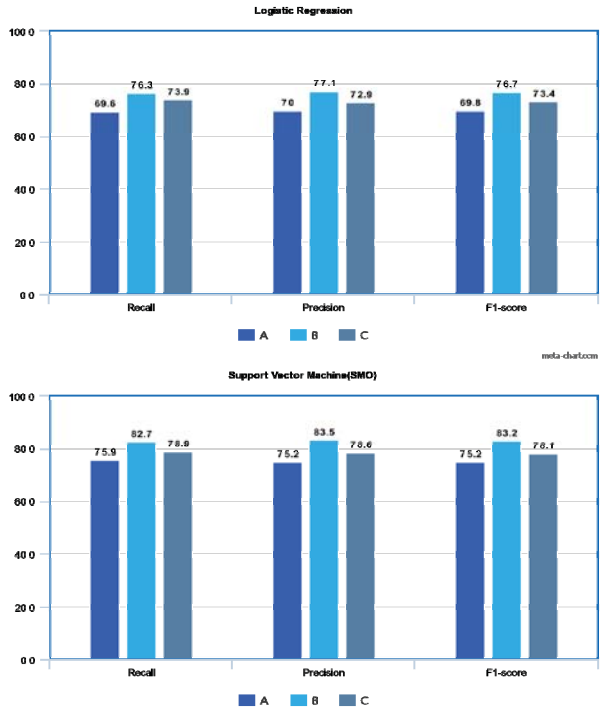


Fig. 3. Comparison of three measures using logistic regression and svm algorithm on weka tool.

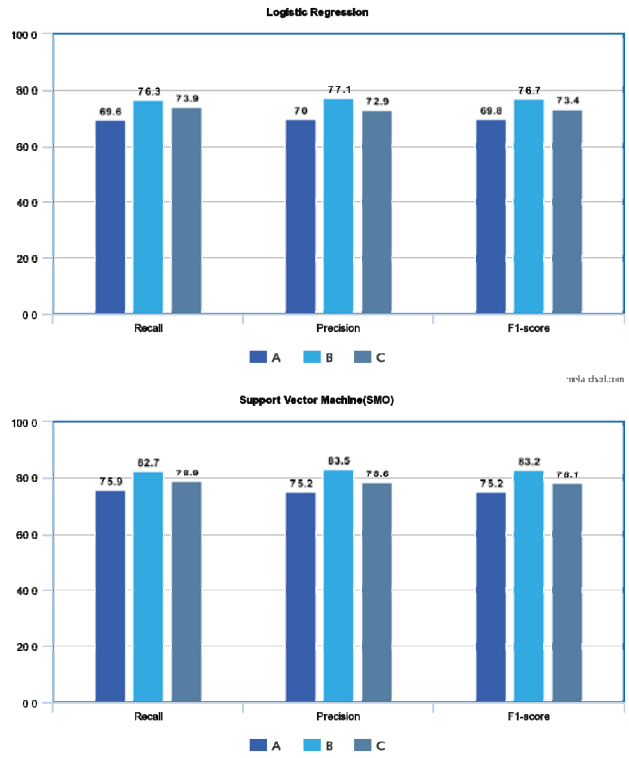


Fig. 4. Comparison of three measures using logistic regression and svm algorithm on jupyter notebook.

Fig. 5 shows the whisker plot to analyze accuracies of logistic regression and support vector machine algorithm. This plot is based on the results achieved through jupyter notebook using python, plot shows that logistic regression has accuracy of 71% whereas support vector machine has 78% accuracy. Fig. 6 shows the comparison of accuracies between these two algorithms achieved through weka tool. 73% accuracy is achieved by logistic regression algorithm and 79% accuracy is achieved by support vector machine using sequential minimal optimization algorithm.

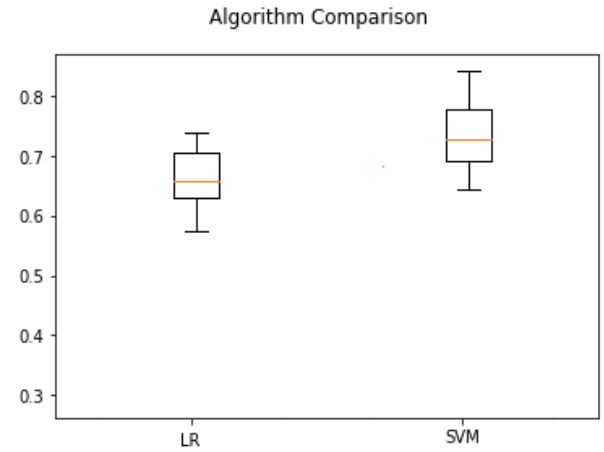


Fig. 5. Comparison of logistic regression and svm accuracies using jupyter notebook.

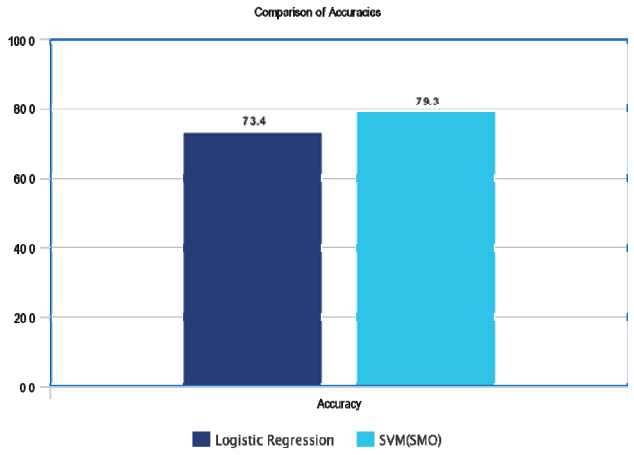


Fig. 6. Comparison of logistic regression and svm accuracies on weka tool.

## VI. CONCLUSION

Future success of students depends purely upon their academic grades, now-a-days as many countries have shifted their educational system from conventional to e-learning systems. So, at distance learning, it is a very difficult task to analyze student's behavior towards their studies, their interaction with classroom activities and instructors, their participation in quizzes or discussion groups. So academic institutions can get a very good benefit by analyzing or extracting the hidden knowledge generated through the learning management system and use this type of knowledge to work for the betterment of student's academic success by predicting their future performance and taking proper interventions. In this research, we have worked on student model and evaluated the model performance using two classifiers logistic regression and support vector machine using two different technologies based on the features selected through feature selection method.

The results obtained with gain ratio feature selection and support vector machine further reveal that satisfaction level of student, interaction with system and punctuality in classroom are three major categories of factors that influence their academic grades.

The results show that support vector machine algorithm using sequential minimal optimization will work better for student's future performance prediction.

## CONTRIBUTION

This research identifies those factors that contributed most to students' academic performance and create a model that can predict the future behavior and performance of student and identify student will pass the course or not.

Overall, the implementation of this model based on EDM technique will allow the educational institutes to make better decision-making actions. This research also contributes in exploratory data analysis phase to identify most impactful features.

## REFERENCES

[1] N.Padhy , D.Mishra ,R. Panigrahi , "The survey of data mining applications and feature scope," arXiv preprint arXiv:1211.5723, November, 2012.

[2] R.S.Baker, S.Gowda, and A.Corbett, "Automatically Detecting a Student's Preparation for Future Learning: Help Use Is Key," In Proceedings of the 4th International Conference on Educational Data Mining, pp. 179–188,2011.

[3] R.S.Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," Journal of Educational Data Mining, vol.1(1) , pp. 3–17,2009.

[4] A.Merceron,K. Yacef, C. Romero, S.Ventura and M. Pechenizkiy, "Measuring correlation of strong symmetric association rules in educational data," Handbook of educational data mining, pp. 245-256, 2010.

[5] A.El-Halees, "Mining Students Data to Analyze Learning Behavior: A Case Study," The 2008 international Arab Conference of Information Technology (ACIT2008) Conference Proceedings, University of Sfax, Tunisia, December , 2008.

[6] C.R.Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol.40 (6), pp.601–618, 2010.

[7] A.Wolff, Z.Zdrahal, D.Herrmannova, J. Kuzilek and M. Hlosta, "Developing predictive models for early detection of at-risk students on distance learning modules," 2014.

[8] E.A.Amrieh, H.Thair, A.Ibrahim, "Preprocessing and analyzing educational data set using X-API for improving student's performance," 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), December, 2015.

[9] H.Agrawal and H.Mavani, " In Student Performance Prediction using Machine Learning," International Journal of Engineering Research and Technology,2015.

[10] E.AAmrieh, H.Thair, A.Ibrahim, "Mining Educational Data to predict student's academic performance using ensemble methods," International Journal of Database Theory And Application , Vol. 9 , 2016.

[11] P.P.Bendangnuksung, "Students' performance prediction using deep neural network," International Journal of Applied Engineering Research. 2018,vol. 13, pp. 1171-6.

[12] B.K Baradwa, S. Pal, " Mining educational data to analyze students' performance, " arXiv preprint arXiv:1201.3417, January, 2017.

[13] S.P.Malan, "The'new paradigm'of outcomes-based education in perspective, " Journal of Consumer Sciences, vol.28(1) ,2000.

[14] G.Kakasevski, M.Mihajlov, S.Arsenovski, and S. Chungurski, "Evaluating usability in learning management system Moodle," In ITI 2008-30th International Conference on Information Technology Interfaces, pp. 613-618,2008.

[15] S.Rapuano and F.Zoino. "A learning management system including laboratory experiments on measurement instrumentation." IEEE Transactions on instrumentation and measurement ," vol.55(5) , pp.1757-1766, 2006.

[16] Kalboard 360–E–learning system. Retrieved from http://cloud. kalboard360.com/User/Login#home/index,2000.

[17] S.Gunuc and A. Kuzu, "Student engagement scale: development, reliability and validity," Assessment & Evaluation in Higher Education, vol. 40 (4) , pp.587-610,2015.

[18] J.Han and M. Kamber, "Data mining concepts and techniques San Francisco Moraga Kaufman," 2001.

[19] C. Romero, S. Ventura and E. García, "Data mining in course management systems: Moodle case," Comput Educ, vol. 51, pp. 368–384, 2008.

[20] R.Arora, "Comparative analysis of classification algorithms on different datasets using WEKA," International Journal of Computer Applications, vol. 54(13),2012.