International Journal for Research in Applied Sciences and Biotechnology

ISSN: 2349-8889
Volume-8, Issue-3 (May 2021)

www.ijrasb.com

https://doi.org/10.31033/ijrasb.8.3.14

A Comparative Study of School Parent Satisfaction Predictors using different Classifiers

Saronyo Lal Mukherjee[1] and Shawni Dutta[2]
[1]Third Year Student, Department of Computer Science, The Bhawanipur Education Society College, Kolkata, INDIA.
[2]Lecturer, Department of Computer Science, The Bhawanipur Education Society College, Kolkata, INDIA.

[2]Corresponding Author: shawnidutta83@gmail.com

ABSTRACT

Educational data mining (EDM) is applied on voluminous student information for obtaining some useful information. This research focuses on the parents' satisfaction based on their executed study. Instead of focusing only from the educational institutions, it is also required to put concentration to the parents' side. Depending on the factors such as how the student carries out their study, their examination result and many more, parental satisfaction is predicted. For carrying out the analysis of these parameters, machine learning methods are implemented and applied to the educational dataset. Several machine learning models such as Support Vector Machines (SVM), k-Nearest Neighbours (KNN), Decision Tree classifiers, and Multi-layer Perceptron classifier (MLP) are constructed for predicting parental satisfaction level. Comparative analysis shows the highest accuracy of 92% executed by the SVM model. Executing this predictive modeling will assist the parents to guide and motivate their children towards areas that demand improvement.

Keywords- Parent Satisfaction, Student performance, EDM, Machine learning, classification.

I. INTRODUCTION

Learning analytics (LA) is an interesting field to explore as it investigates the learning procedure of learners and their contexts. The objective of learning analytics is to understand and optimize the students' risks [1]. Education data mining (EDM) is another field which detects the uncovered patterns present in the students' database and apply computerized methods to understand the huge amount of data [2]. Use of LA and EDM is often beneficial as it inspects the educational dataset from the perspective of teacher, students, parents, and administrators. This study has focused mainly from the parents' perspective while examining the students' performance parameters. As a stakeholder of this domain, guardians are notified with an early alarming system. This will assist the guardians to encourage the students' improvements [3].

Education in the 21st Century has seen advancements in both quality and quantity of it. Educational institutions have witnessed an exponential increase of parent awareness in institutional quality, which has led to the extensive corporatization. As a result parent satisfaction in institutional quality index is of utmost importance. Involvement of parents is closely associated with children's academic performance. It is observed that the students whose guardians have more participation in their education pursue higher levels of academic performance than those students whose parents are less involved in their study [4]. Thus a good parent satisfaction rating is highly sought after, so is a fair parent satisfaction rating prediction. Institutions require knowing predicted satisfaction rating based on the performance of their students for internal administrative purposes. An idea of how satisfied a parent is in his or her child's educational institution also helps the institution to check on the fields it lags in. It also enables the institution to judge the mood of its clients, and the opportune moment to introduce new policies.

The objective of this study is to predict and assess the guardians' satisfaction regarding their children's study progress. To fulfil the aforementioned objective, it is required to explore the relationships among the interfering performance factors. This hidden relationship knowledge extraction from a large dataset can be obtained by using Machine Learning (ML) techniques. These hidden patterns can enable smart decision making processes. Construction of an automated tool enables extraction and understanding of the dataset and predicts the parents' satisfaction level. Supervised ML techniques are approached while solving this particular problem. Classification is a supervised ML technique that utilizes training data for understanding the hidden patterns in the data and later uses the knowledge for prediction purposes. The prediction is basically whether or not the parent is satisfied with the student's study; hence this problem is defined as a binary classification problem [5]. This paper uses different Machine Learning (ML) classifier algorithms and compares their respective accuracy scores and other parameters to propose a useful model which can be used to predict parent satisfaction based on various factors like student grade, nationality, place of birth, number of times a student's puts up his or her hand when a question is asked, how willing parents are to answer satisfaction surveys, etc. The classifier predictive models compared include Support Vector Machines (SVM) [6], k-Nearest Neighbours (KNN) [7], Decision Tree classifiers [8], and the Multi-layer Perceptron

This work is under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

classifier (MLP) [9]. The parameters used to compare these algorithms are Accuracy Score [10], Matthews's correlation coefficient (MCC) [10], and Mean squared error regression loss [11]. Comparative study based on these metrics enables to identify the best performing classifier model.

II. RELATED WORKS

Assessing the students' performance is one of the most important concerns of an academic institution. Many researchers mentioned that students' success rates can be predicted by examining their past grades, co-curricular activities, achievements and many other factors. This section briefly describes numerous research studies those were carried out for students' performance prediction.

With the aid of AI, students' performance can be predicted by using machine learning based models such as back propagation (BP), Support Vector Regressor (SVR), Long-short term memory (LSTM) and Gradient Boosting Classification (GB). Experimental results concluded highest prediction accuracy of 87.78% as shown by BP [12]. Another study [13] presented a strategy of identifying students who have a poor academic performance of the computer science subject offered by Al-Muthanna University, College of Humanities. Four classification algorithms like feed forward Artificial Neural Network (ANN), Naïve Bayes, Decision Tree (DT), and Logistic Regression, were implemented to identify the poor students. ANN model has shown that ROC index of 0.80 and accuracy 77.04%. Additionally, DT model identified the most influential factors such as Computer Grades-Course1, Accommodation, Interest in studying computer, Educational Environment Satisfaction, and the Residency which can instigate the students' performance [13].

Kaur et al. [14] performed predictive modelling for identifying the slow learners using classification algorithms. This study collected real time data and implemented classification algorithms such as Naïve Bayes, J48, REP Tree, SMO, and Multilayer Perceptron (MLP) using the WEKA tool. Among these specified classifiers, this study has identified slow learners with the highest accuracy of 75% as exhibited by Multi-Layer Perceptron. Edin Osmanbegović et al. [15] performed academic success prediction by a means of MLP, J48 and Naive Bayes (NB) classification techniques. To fulfil this objective, a dataset from the University of Tuzla is collected. Analytical results confirmed highest efficiency of NB technique with the highest efficiency of 76.65%.Another study [16] has constructed a classification model based on deep learning, NB, MLP and SVM. Comparative results have identified that SPPN (Students Performance Prediction Network) outperformed the other models [16]. Students' performance prediction system is also built by using Deep Neural Network (DNN). The performance of this model is compared to Decision Tree (C5.0), Naïve Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbor. Finally, DNN has exhibited the highest accuracy of 84% [17].

Considerable amount of studies have also focused on assessing the students' engagement in web-based learning platforms. Students engage themselves in watching videos and shorter video length attracts students to some greater extend [18]. Another study [19] investigated the students' engagement in higher education blended-learning classrooms using cross-lagged modelling technique. The investigation revealed that proper course design can enhance students' engagement. The investigation on the relationship between gamification and student engagement in online discussion forums has been conducted in [20]. Another study [21] investigated the relationship between course materials and students' scores.

As mentioned in this section, considerable research has been conducted from the students', instructors, and/or academic institutions' point of view. However, this study has tried to investigate the relationship between students' performance and parent's satisfaction. For carrying out this investigation, numerous existing machine learning methods are applied. Finally, the best classification technique is identified for parental satisfaction level prediction system.

III. DATASET COLLECTION AND PRE-PROCESSING

The dataset used was obtained from Kaggle [22] and consists of total 17 parameters. Description of these parameters is summarized in Table 1. The dataset consists of 480 instances. The Parent School Satisfaction index was selected as the target attribute or the attribute to be predicted. It has two options as its answers, ie. Yes and No. The rest of the 16 columns were treated as the input attributes. The entire dataset was then divided into two parts with 80% of the data in one and 20% in the other. The 80% portion was treated as the Training Data and the rest was treated as the Testing Data. Meanwhile The Parent School Satisfaction column was used as the Target Data. The existence of this target variable will distinguish the training and testing dataset.

Before splitting the dataset into training and testing data, some pre-processing techniques are applied. The string values present in the dataset are encoded into labels using the Label Encoder function. Numeric data present in the dataset were scaled using the Min Max Scaler tool. Both of these tools are found under the pre-processing sub-module of the Scikit-learn [23] package.

This work is under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Table1: Summary of the collected dataset

Attributes	Description
Gender	Gender of the students ('Male' or 'Female')
Nationality	The nationality of the student (' Kuwait',' Lebanon',' Egypt',' SaudiArabia','USA','Jordan','Venezuela','Iran','Tunis','Morocco','Syria','Palestine',' Iraq',' Lybia')
Place of Birth	Where the student was born (' Kuwait',' Lebanon',' Egypt',' Saudi Arabia',' USA','Jordan','Venezuela','Iran','Tunis','Morocco','Syria','Palestine','Iraq',' Lybia'
Stage ID	ID of the Stage of the student ('lower level', 'Middle School', 'High School')
Grade ID	ID of the grade the student is in ('G-01', 'G-02', 'G-03', 'G-04', 'G-05', 'G-06', 'G-07', 'G-08', 'G-09', 'G-10', 'G-11', 'G-12 ')
Section ID	ID of the section the student is in ('A','B','C')
Topic	The topic the student is studying (' English',' Spanish', 'French',' Arabic',' IT',' Math',' Chemistry', 'Biology', 'Science',' History',' Quran',' Geology')
Semester	The semester the student is in (' First',' Second')
Relation	The relation between the parent and the student ('mom','father')
Raised Hands	The number of times the student has raised his/her hands (0-100)
Visited Resources	The number of times student has visited the course materials (0-100)
Announcements Viewed	The number of announcements viewed (0-100)
Discussion	The number of discussions attended (0-100)
Parent Answering Survey	Whether the parent has answered the survey or not ('Yes','No')
Parent School Satisfaction (to be predicted)	Whether the parent is satisfied with school or not ('Yes','No')
Student Absence Days	The number of days the student was absent (above-7, under-7)
Class	The class of the student (Low-Level where interval ranges from 0 to 69, Middle-Level where interval ranges from from 70 to 89, High-Levelwhere interval ranges from from 90-100.)

IV. BACKGROUND

In Supervised Machine Learning, externally supplied examples are used to search for an appropriate algorithm, which will produce general hypotheses and make predictions on future examples. In any dataset that is being used by a machine learning algorithm, each data is represented by the same set of features. These features may be binary, continuous or even categorical. If the supplied or training examples are labelled with previously known labels, then the type of learning is called Supervised Learning [5]. Otherwise, if the labels are unknown, the learning is known as Unsupervised Learning. This paper will be dealing with Supervised Learning. The goal of a Supervised Classification Model is to train an appropriate model with the predictor features and their corresponding distribution of class labels. The thus trained classifier is then used to classify testing examples for which the predictor features are known, but the respective classes are unknown and thus predicted.

This section briefly elaborates the classifier models used in this study. The employed classifiers are Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Decision Tree (DT), Neural Network.

Support Vector Machine (SVM)

A Support-Vector Machine or SVM Algorithm uses a pre-chosen non-linear mapping technique to map input predictor vectors into a feature space Z which has a high dimension. In the aforementioned feature space, a linear decision surface with special properties is constructed so that the network is able to maintain high generalization ability. The linear decision function with the maximal margin between vectors of any two classes in SVM is called the Optimal Hyper plane. This margin is determined by a small subset of the training set, called the support vectors, is used to construct an Optimal Hyper plane [6].

SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of the kernel is to take data as input and transform it into the required form. There are possible types of kernels

This work is under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

such as Gaussian radial basis function (RBF), Sigmoid Kernel, Polynomial Kernel [23].

K- Nearest Neighbours (KNN)

The Nearest Neighbour Decision Rule is used to classify, unclassified sample points to the nearest set of previously classified points. The x_is in a set of n pairs, (x_1,θ_1), (x_2,θ_2), ..., (x_n,θ_n), take up values in a metric space X., while the θ_is take up the values {1, 2, ...,M}. A metric d is defined upon these pairs. The category of each it individual is indexed by θ_i, and the corresponding x_is are the outcomes obtained from the respective set of measurements made. In the testing cases, a new pair, (x,θ) is introduced. 'x' is the only parameter observable. The corresponding θ is obtained using the information contained in the previously correctly points.

The Nearest Neighbour Rule is used to decide which x belongs to which category of θ_n' it belongs to with nearest neighbour being x_n'. The k-Nearest Neighbours rule maintains that x converges at x with probability unity, as sample size n increases and *k* remains fixed [7].

Decision Tree

Decision Trees are algorithms which states rules to represent underlying data with hierarchical, and sequential structures by recursively partitioning the data. The Decision Tree structure contains some or no internal nodes and one or more leaf nodes. Each internal node has two or more child nodes. To test the value of an expression of the attributes, all internal nodes contain splits. The internal nodes, t, have arcs to its children which are labelled with distinct outcomes of the test at t. Thus a class label is given to each node. Thus tree induction, tree growing, and tree building is different names given to the job of constructing a tree from the training set. A greedy top-down process is followed by most tree induction systems [8].

Neural Networks

Neural Networks are a computational learning system that uses a network of functions to take input, process, and render a desired output. The concept of an Artificial Neural Network (ANN) was taken from the human brain, and how its network of neurons work together to take input, make decisions, and then ultimately take actions. There are two main layers, called the input and output layer of neurons, the purpose of which are as their names suggest. There may be a number of neuron layers in between these two layers which actually make the decisions and are used to decide the output. Neural Networks are nowadays used for pattern recognition, face recognition, AI and other such fields [9].

V. METHODOLOGY

All programming and coding was carried out using the Python programming language. The huge set of predefined modules provided by the Python Library was like SK Learn, Pandas, Numpy, etc. made to use for respective purposes. This research uses the different classification methods mentioned in this paper to classify whether the parent of a child going to school is satisfied with the services the school provides and the hollistic result shown by their wards. The classification is done by training the different classification models using 16 columns which include; gender, nationality, place of birth, the standard in which the child is studying, number of times he or she has put up their hands, the number of days they have been absent, etc. And the system is trained using whether or not the parents of the respective students are satisfied with the school (two discrete values; good or bad are used) as the target set. The test dataset is applied to predict parent satisfaction, and the obtained result is compared with the actual survey values to determine the prediction accuracy score. The prediction accuracies obtained from the different models are compared to determine the best model that can be used for parent satisfaction prediction and is proposed. Application of dataset pre-processing techniques will transform the collected dataset into a balanced dataset. Now, training and testing dataset is retrieved by partitioning the transformed dataset with the ratio of 8:2.

Implementation of each machine learning model requires a thorough parameter optimization process. While implementing the SVM model, the regularization parameter (C) was initialized the best parameter. This model has considered the gamma parameter as 0.01. These parameters have shown the best possible outcome as classification result. These parameters are also tested along with proper selection of kernel. After evaluation, Radial basis function (RBF) kernel has shown the best predictive efficiency. The KNN model was prepared. The value of k was varied within the range from 5 to 12. As a result, k=11 has shown the best testing accuracy. Next, the DT model was prepared using the 'gini' criterion and 'best' splitter. Lastly, the Multi-layer Perceptron classifier from the Neural Networks module was prepared. After picking up the best possible hyper-parameters, the classifier models start learning the uncovered patterns from the training dataset during the training phase. Later, the learning was evaluated using testing data during the testing phase. The testing outcomes are compared with the original results in the dataset using some metrics such as Accuracy score, Matthews's correlation coefficient (MCC), and Mean squared error (MSE). All the employed models are compared using this metrics. The best model is identified that exhibits the highest accuracy and MCC score and optimized MSE value.

VI. RESULTS AND DISCUSSION

The employed classifier models are dedicated to predict the parent satisfaction variable for the corresponding student based on their performance. Four classification techniques such as SVM, k-NN, DT and NN are utilized for this objective completion. The

This work is under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

classifier tools are summarized in table 2 along with their chosen parameters and accuracy, MCC and MSE. As shown in the table2, highest Accuracy Score is obtained for the Support Vector Machine algorithm with 'rbf' kernel; 0.92, while KNN has an Accuracy Score of 0.90, and Decision Trees and Neural Networks have even lesser values. Highest Michael's Correlation Coefficient is obtained for the Support Vector Machine algorithm with 'rbf' kernel; 0.8, while KNN has a Michael's Correlation Coefficient of 0.77, and Decision Trees and Neural Networks have even lesser values. Optimized Mean Squared Error is obtained for the Support Vector Machine algorithm with 'rbf' kernel; 0.08, while KNN has a Michael's Correlation Coefficient of 0.77, and Decision Trees and Neural Networks have even lesser values. After this comparison, it is quite clear that the SVM model has shown the best possible predictive efficiency with respect to all the metrics as compared to other models. Hence, the SVM model can be utilized as a tool for an intelligent model construction for carrying out the EDM process.

Table 2: Experimental Results of the ML model with optimal parameters used

Classifier Models	Parameters	Accuracy	Michael's (MCC)	MSE
SVM	Kernel = rbf	0.92	0.8	0.08
KNN	K = 11	0.9065	0.775709215281	0.09375
DT	criterion = 'gini splitter = 'best''	0.822916666667	0.640873612364	0.177083333333
NN	Alpha = 0.001	0.864583333333	0.697863157799	0.135416666667

VII. CONCLUSION

To facilitate the EDM process, construction of an intelligent model often plays an important role in assessing the performance of students. However, this study analyses the student performance and predicts the parents' satisfaction tendency by using an intelligent model. In the EDM process, parents are also one stakeholder whose existence and intervention is necessary. Several ML techniques are implemented and applied on a cleaned dataset. The collected dataset has gone through multi-step pre-processing techniques for retrieving cleaned dataset. After an exhaustive comparison of numerous machine learning models this study has identified SVM as the best predictor. Highest prediction efficiency having 92% accuracy score, MCC score of 0.8 and MSE of 0.08 is reached by the intelligent model. Early prediction will assist the parents to put more effort and concentrations for encouraging their children towards their success rates.

REFERENCES

[1] Avella, John T., et al. "Learning analytics methods, benefits, and challenges in higher education: A systematic literature review." *Online Learning* 20.2 (2016): 13-29.

[2] Romero, Cristóbal, and Sebastián Ventura. "Educational data mining: a review of the state of the art." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40.6 (2010): 601-618.

[3] Papamitsiou, Zacharoula K., and Anastasios A. Economides. "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence." *Educational Technology & Society* 17.4 (2014): 49-64.

[4] Driessen, Geert, Frederik Smit, and Peter Sleegers. "Parental involvement and educational achievement." *British educational research journal* 31.4 (2005): 509-532.

[5] Singh, Amanpreet, Narina Thakur, and Aakanksha Sharma. "A review of supervised machine learning algorithms." *2016 3rd International Conference on Computing for Sustainable Global Development (INDIA Com)*. IEEE, 2016.

[6] Ma, Yunqian, and GuodongGuo, eds. *Support vector machines applications*. Vol. 649. New York, NY, USA:: Springer, 2014.

[7] Guo, Gongde, et al. "KNN model-based approach in classification." *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, Berlin, Heidelberg, 2003.

[8] Priyam, Anuja, et al. "Comparative analysis of decision tree classification algorithms." *International Journal of current engineering and technology* 3.2 (2013): 334-337.

[9] Windeatt, Terry. "Accuracy/diversity and ensemble MLP classifier design." *IEEE Transactions on Neural Networks* 17.5 (2006): 1194-1211.

[10] Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." *BMC genomics* 21.1 (2020): 1-13.

[11] Harville, David A., and Daniel R. Jeske. "Mean squared error of estimation or prediction under a general linear model." *Journal of the American Statistical Association* 87.419 (1992): 724-731.

[12] Sekeroglu, Boran, KamilDimililer, and Kubra Tuncal. "Student performance prediction and

 This work is under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

classification using machine learning algorithms." *Proceedings of the 2019 8th International Conference on Educational and Information Technology*. 2019.

[13] Altabrawee, Hussein, Osama Abdul Jaleel Ali, and Samir Qaisar Ajmi. "Predicting Students' Performance Using Machine Learning Techniques." *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences* 27.1 (2019): 194-205.

[14] Kaur, Parneet, Manpreet Singh, and Gurpreet Singh Josan. "Classification and prediction based data mining algorithms to predict slow learners in education sector." *Procedia Computer Science* 57 (2015): 500-508.

[15] Osmanbegovic, Edin, and Mirza Suljic. "Data mining approach for predicting student performance." *Economic Review: Journal of Economics and Business* 10.1 (2012): 3-12.

[16] Guo, Bo, et al. "Predicting students performance in educational data mining." *2015 International Symposium on Educational Technology (ISET)*. IEEE, 2015.

[17] Vijayalakshmi, V., and K. Venkatachalapathy. "Comparison of Predicting Student's Performance using Machine Learning Algorithms." *International Journal of Intelligent Systems and Applications* 11.12 (2019): 34.

[18] Guo, Philip J., Juho Kim, and Rob Rubin. "How video production affects student engagement: An empirical study of MOOC videos." *Proceedings of the first ACM conference on Learning@ scale conference*. 2014.

[19] Manwaring, Kristine C., et al. "Investigating student engagement in blended learning settings using experience sampling and structural equation modeling." *The Internet and Higher Education* 35 (2017): 21-33.

[20] Ding, Lu, Erkan Er, and Michael Orey. "An exploratory study of student engagement in gamified online discussions." *Computers & Education* 120 (2018): 213-226.

[21] Atherton, Mirella, et al. "Using learning analytics to assess student engagement and academic outcomes in open access enabling programmes." *Open Learning: The Journal of Open, Distance and e-Learning* 32.2 (2017): 119-136.

[22] Ibrahim Aljarah (November, 2016),Students' Academic Performance Dataset.Retrieved on June 06, 2020 from https://www.kaggle.com/aljarah/xAPI-Edu-Data

[23] Bisong, Ekaba. "Introduction to Scikit-learn." *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress, Berkeley, CA, 2019. 215-229.

[24] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003): 1396-1400.

This work is under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.