International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 6 Issue 1, November-December 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Identifying the Key Factors of Training Technical School and College Teachers in Bangladesh Using Data Mining

Md. Mehedi Hasan[1], Md. Imran Ali[1], Nakib Aman Turzo[2], Golam Rabbani[3]

[1]Department of Computer Science & Engineering, Varendra University, Rajshahi, Bangladesh
[2]Research Co-ordinator, National Academy for Computer Training and Research, Bogura, Bangladesh
[3]Instructor, National Academy for Computer Training and Research, Bogura, Bangladesh

ABSTRACT

According to the Bangladesh Bureau of Statistics (BBS), the literacy rate in Bangladesh is increasing day by day. But, It's not acceptable to our present day. The role model of an education system is a teacher or instructor. Proper education can improve our literacy rate and also be a huge change for our future Digital Bangladesh. This enhancement is only possible to highly trained instructors or teachers. In order to improve an organization's training process, it's important to assess how instructors are trained their students. This research has worked on identifying the key factors of training Technical School and College teachers in Bangladesh. The proposed work is conducted by Data Mining and Machine Learning. The methods of this experiment are Data Processing, Data Mining, and Analysis & Evaluation. Filtering our data is completed by using the Data Processing method. After that, the datasets are trained and tested by the Data Mining and Machine Learning tools. Finally, the experimental results are evaluated and analyzed by the different assessment tools. The accuracy of our trained models are 0.97%, 0.97%, 0.96%, 0.96%, 0.96%, 0.96%, 0.94%, 0.93%, 0.93%, 0.92%, 0.91%, 0.33%, 0.22% using the Logistic Regression, Extra Trees Classifier, Random Forest Classifier, Gradient Boosting Classifier, Light Gradient Boosting Machine, SVM - Linear Kernel, Ada Boost Classifier, K Neighbors Classifier, Linear Discriminant Analysis, Decision Tree Classifier, Ridge Classifier, Quadratic Discriminant Analysis, Naive Bayes, respectively. As a result, the Logistic Regression does accurately identify and classify the key factors of training Technical School and College teachers. The Logistic Regression model accuracy is 0.97% which gives better accuracy than other machine learning algorithms.

KEYWORDS: *Bangladesh Bureau of Statistics, Instructors, Teachers, Technical School & College, Data Processing, Data Mining, Machine Learning, Analysis & Evaluation, Logistic Regression, Extra Trees Classifier, Random Forest Classifier, K Neighbors Classifier, Linear Discriminant Analysis, Decision Tree Classifier, Naive Bayes, Key factors*

How to cite this paper: Md. Mehedi Hasan | Md. Imran Ali | Nakib Aman Turzo | Golam Rabbani "Identifying the Key Factors of Training Technical School and College Teachers in Bangladesh Using Data Mining" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-6 | Issue-1, December 2021, pp.864-869, URL: www.ijtsrd.com/papers/ijtsrd47901.pdf

IJTSRD47901

INTRODUCTION

There has been a growing interest in the use of educational data mining in research in recent years. In the current field of education and science, data mining has become a powerful tool by which the future of an organization can be changed. Data mining can easily be used to analyze the performance of an instructor or students in an organization.

Data mining completes the task of discovering new patterns from huge amounts of data. To collect this huge amount of data, data can be collected through Google Forms using social media or various digital media [1].

In analyzing the process of training instructors, it is important to measure how the performance of instructors can be analyzed to improve an organization by collecting data on the activities that take place each year. To analyze the performance compared to different algorithms of data mining, it is divided into several sections to see what kind of performance good instructors give.

This paper gives good accuracy especially for logistics regression for performance analysis using data mining algorithms. This paper also attempts to highlight the dataset complement and enables the instructors to analyze the performance.

LITERATURE REVIEW

Ms.Tismy Devasia, Ms.Vinushree T P, Mr.Vinayak Hegde's prediction performance in 2016 by using education data mining on 19 attributes of 700 students of Amrita Vishwa Vidyapeetham, Mysuru. This paper, divided the students into 4 categories to analyze their performance and showed that the percentage of good student performance is higher. Several algorithms have been used in the paper, among which the Naive Bayesian algorithm has given good accuracy [2].

Anoushka Jain, Tanupriya Choudhury, Parveen Mor, and A.Sai Sabitha wrote a paper in 2016 comparing different data mining techniques to analyze the intellectual performance of students. This paper collected data using social media and various digital media through Google Forms. In the paper, students are marked in 5 categories which are classified by different discussion trees. Out of 5 algorithms, 3 algorithms have predicted mode CGPA as "good" and 2 algorithms say "very good" [2].

J.K. Jothi Kalpana, K. Venkatalakshmi wrote a paper in 2014 to analyze the performance of graduate students. In the paper, data sets of 5-years students of the College of Engineering and Technology, Villupuram were collected. K-Means divided students into groups using clustering algorithms and methods such as centroid-based, distribution-based, and density-based clustering were used. The paper divided the student's CGPA into 5 categories namely excellent, very good, good, average, and poor which showed the average CGPA of 54% of students using data mining techniques [3].

Chitra Jalota, Rashmi Agrawal wrote a paper on education data mining in 2019 by using classification. In this paper, the data set of 480 students contains 163 instances and 16 attributes. Five classifications are used under Weka and comparisons are made based on the accuracy between these classifications and different error measurements are used to diagnose different classifications. Multilayer Perceptron has been shown to give the best performance among other classifiers [4].

Vinayak Hegde and Sushma Rao H S wrote a paper in 2016 on the performance analysis of student programming languages using educational data mining. In this paper, a survey of students on 15 questions for C, 24 for Math, 23 for C ++, and 20 for Java divided their marks into 5 categories namely outstanding, excellent, good, average, and poor respectively. Analyze student performance based on that test provides a big factor in identifying poorly performing students [5].

Masna Wati, Wahyu Indrawan, Joan Angelina Widians, Novianti Puspitasari wrote a paper in 2017 that uses data from students within the academic database or from outside the academic database using the Naïve Bayes Classifier and Tree C4.5 Predicts educational outcomes. This paper shows that the mental level acquired by age, the cultural habits of the students, and the level of participation of the students have some effect on the graduation time of the students when they study them [6].

METHOLODOGY

The proposed educational data mining process in this study is illustrated in Fig.1 below. Data mining is the process of discovering knowledge that deals with huge amounts of information. In this proposed method, a large dataset is created where the performance of instructors from the National Academy of Computer Training and Research is evaluated.

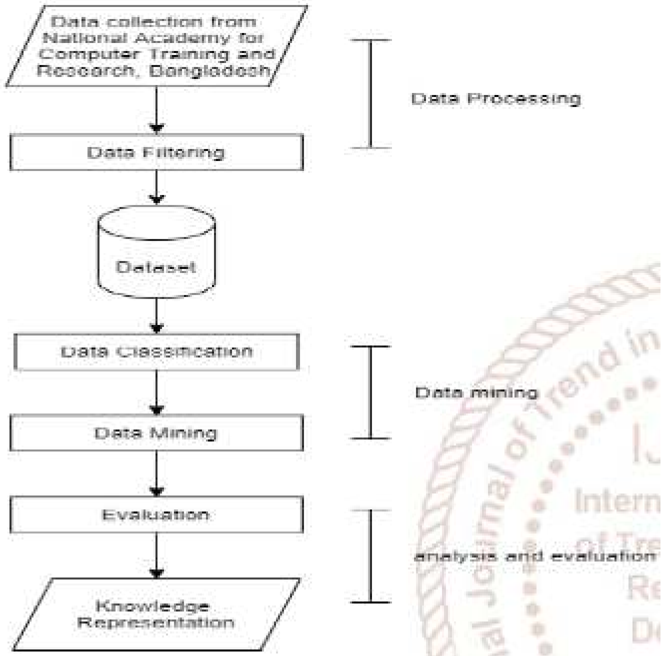

Fig.1: Proposed Methodology Diagram

A. Data Processing

Information about the trainers was collected from 80 students of 4 batches from the National Academy of Computer Training and Research through Google Forms. In this step, tasks like selecting, cleaning, filtering were done in the data.

B. Dataset

A dataset with 1772 data was selected from 80 students of 4 batches. There were 32 instructors on the dataset and 10 questions were taken in each of the 16 topics.

C. Data Mining

Data mining is the process of finding inconsistencies, patterns, and interrelationships between large data sets and predicting outcomes. Trainers were identified as having access to data, predicting achievement levels, and the need for additional efficiency. In this process, it helps to measure the performance of the trainer on a large dataset. The data is classified and compared by different models. Classification data mining techniques are much simpler and more used.

D. Analysis and Evaluation

Each question is numbered by collecting information through Google form to evaluate the performance of the trainers [2]. Marks are divided into 5 categories [5] and we can compare data with different algorithms to help trainers evaluate their performance.

Table I: Hyperparameter values set before classification.

	Parameters
C	1.043
class_weight	{}
dual	FALSE
fit_intercept	TRUE
intercept_scaling	1
l1_ratio	None
max_iter	1000
multi_class	auto
n_jobs	None
penalty	l2
random_state	5290
solver	lbfgs
tol	0.0001
verbose	0
warm_start	FALSE

Table I demonstrate sklearn model library function that makes up the proposed machine learning models. The model was built using the sci-kit-learn in the python library. All of the models were used by some parameters. These were C, class_weight, dual, fit_intercept, intercept_scaling, l1_ratio, max_iter, multi_class, n_jobs, penalty, random_state, solver, tol, verbose, and warm_start. After that, all of the models had been trained using these parameters.

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

In the present work, the overall key factors identification and classification performances of the datasets are evaluated through the 13 machine learning models. These are Logistic Regression, Extra Trees Classifier, Random Forest Classifier, Gradient Boosting Classifier, Light Gradient Boosting Machine, SVM - Linear Kernel, Ada Boost Classifier, K Neighbors Classifier, Linear Discriminant Analysis, Decision Tree Classifier, Ridge Classifier, Quadratic Discriminant Analysis, and Naive Bayes. Above all these machine learning classifiers are mostly used for classification problems.

Table II: Respective Accuracy, AUC, Recall and Precision values of different classifiers.

Model	Accuracy	AUC	Recall	Prec.
Logistic Regression (Proposed)	0.97	0.99	0.94	0.97
Extra Trees Classifier	0.97	0.99	0.93	0.97
Random Forest Classifier	0.96	0.99	0.92	0.96
Gradient Boosting Classifier	0.96	0.99	0.92	0.96
Light Gradient Boosting Machine	0.96	0.99	0.93	0.96
SVM -Linear Kernel	0.96	0	0.92	0.96
Ada Boost Classifier	0.94	0.99	0.92	0.96
K Neighbors Classifier	0.93	0.98	0.86	0.93
Linear Discriminant Analysis	0.93	0.99	0.88	0.94
Decision Tree Classifier	0.92	0.93	0.85	0.92
Ridge Classifier	0.91	0	0.82	0.91
Quadratic Discriminant Analysis	0.33	0.59	0.49	0.76
Naive Bayes	0.22	0.54	0.46	0.51

Model	F1	Kappa	MCC	TT (Sec)
Logistic Regression (proposed)	0.97	0.94	0.94	0.32
Extra Trees Classifier	0.97	0.92	0.92	0.59
Random Forest Classifier	0.96	0.91	0.91	0.65
Gradient Boosting Classifier	0.96	0.91	0.91	2.65
Light Gradient Boosting Machine	0.96	0.91	0.91	0.24
SVM -Linear Kernel	0.96	0.91	0.91	0.15
Ada Boost Classifier	0.94	0.88	0.88	0.35
K Neighbors Classifier	0.93	0.85	0.85	0.46
Linear Discriminant Analysis	0.93	0.86	0.86	0.54
Decision Tree Classifier	0.92	0.82	0.82	0.06
Ridge Classifier	0.91	0.81	0.81	0.07
Quadratic Discriminant Analysis	0.33	0.13	0.21	0.34
Naive Bayes	0.14	0.07	0.14	0.06

In this research, evaluating the classification model is measured by some criteria. When analyzing the model's performance, it's important to consider how accurate the building model's predictions are. In the discipline of Machine Learning or Deep Learning, the process of building models is evaluated in a variety of ways. Some of the methods are Confusion Matrix, Accuracy, AUC, Precision, Recall, Kappa, MCC, TT(Sec), and F1 score.

In Table II, we showed all the measurement tools to easily examine our machine learning models. That is the important measurement of the proposed building model.

Table III: Tuned result of Proposed Logistic Regression Model

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.99	0.99	0.97	0.99	0.99	0.97	0.97
1	0.97	0.99	0.92	0.96	0.96	0.92	0.92
2	0.95	0.99	0.90	0.95	0.95	0.89	0.89
3	0.97	0.99	0.92	0.97	0.96	0.92	0.92
4	0.97	0.99	0.94	0.97	0.97	0.94	0.94
5	0.97	0.99	0.95	0.98	0.97	0.94	0.94
6	0.98	0.99	0.96	0.98	0.98	0.95	0.96
7	0.97	0.99	0.93	0.98	0.98	0.92	0.92
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9	0.96	0.99	0.92	0.97	0.96	0.90	0.90
Mean	0.97	0.99	0.94	0.97	0.97	0.94	0.94
SD	0.02	0.01	0.04	0.02	0.02	0.04	0.04

Table III demonstrate our best performance model named Logistic Regression. During the training, we have used 10 epochs for each distinct machine learning model. In this table, we showed Accuracy, AUC, Precision, Recall, Kappa, MCC, and F1 scores for every epoch. And also, we find out Mean and Standard Deviation values. We compared all of the models result with our measurement tools (which is mentioned before). As a result, the Logistic Regression model is better accuracy than other machine learning models.

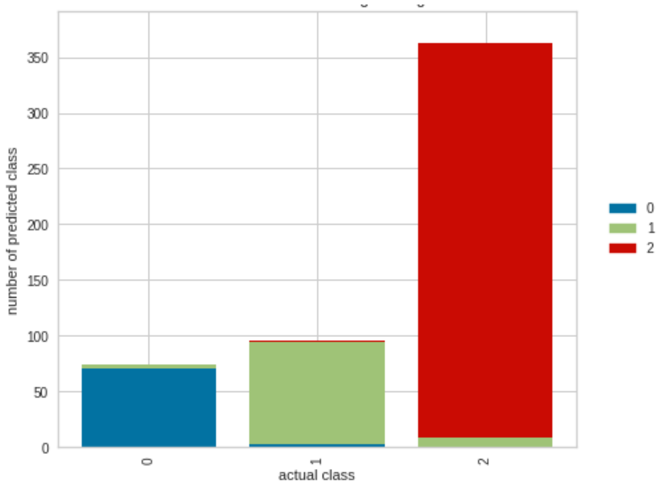

Fig.2: Class Prediction Error for Logistic Regression

In Figure 2, we exhibited our prediction loss or error in logistic regression. We compare several commonly-used prediction rules and loss functions, and we used one that reduced our prediction loss. Furthermore, when the number of the predicted classes are increasing, then the class prediction loss or error of Logistic Regression is respectfully changed.

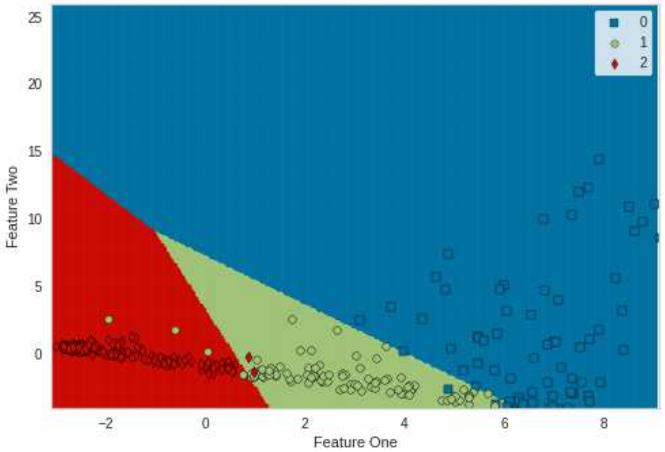

Fig.3: Class Prediction Space for Logistic Regression

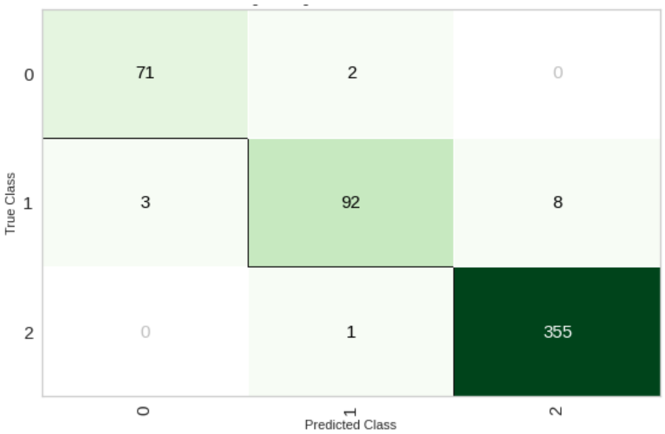

Fig.4: Confusion Matrix of Proposed Model.

In the logistic regression model, we showed our Confusion Matrix in Figure 4. The Confusion Matrix is a tabular form that is used to show the projected model performance in most circumstances. Confusion Matrix quickly determines whether a multiclass is right or incorrect. The confusion matrix is a machine learning statistic for determining a model's predictability. Precision, recall, and f1 score are all evaluated in the confusion matrix.

In the above figure, our test data's logistic regression confusion matrix summarizes a classifier's classification performance with regard to particular trained data. When compared to the actual class, the projected class is accurate. In addition, the anticipated class's error is lower than the actual class's error. So, the overall classification performance of our research is highly acceptable.

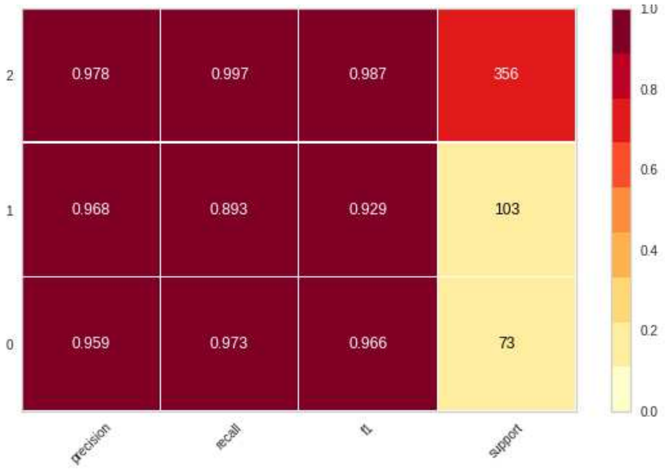

Fig.5: Heat map of the proposed model.

In Figure 5, we demonstrated our classification report in logistic regression. The classification report is used to evaluate the predictive accuracy of a classification algorithm. The report presents the major classification metrics precision, recall, and f1-score on a per-class basis. The metrics are generated using true and false positives, as well as true and false negatives.

CONCLUSION

It concludes that the data mining algorithm for evaluating instructors' performance, especially logistic regression, gives 96.69% accuracy. This study shows that out of the data obtained from 80 students in 4 batches, most of the instructors performed well. In future work, more dataset examples will be collected and compared, and analyzed with other data mining techniques.

REFERENCES

[1] T. Devasia, Vinushree T P and V. Hegde, "Prediction of students performance using Educational Data Mining," 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), 2016, pp. 91-95, doi: 10.1109/SAPIENCE.2016.7684167.

[2] Jain, T. Choudhury, P. Mor and A. S. Sabitha, "Intellectual performance analysis of students by comparing various data mining techniques,"

2017 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2017, pp. 57-63, doi: 10.1109/ICATCCT.2017.8389106.

[3] J.K.Jothi and K.Venkatalakshmi, “Intellectual performance analysis of students by using data mining techniques”, International Journal of Innovative Research in Science, Engineering and Technology, vol 3, Special iss 3, March 2014.

[4] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 243-247, doi: 10.1109/COMITCon.2019.8862214.

[5] V. Hegde and S. Rao H.S., "A Framework to Analyze Performance of Student's in Programming Language Using Educational Data Mining," 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2017, pp. 1-4, doi:10.1109/ICCIC.2017.8524244.

[6] M. Wati, W. Indrawan, J. A. Widians and N. Puspitasari, "Data mining for predicting students' learning result," 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), 2017, pp. 1-4, doi:10.1109/CAIPT.2017.8320666.