e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 1, January 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.54

☎ 6381 907 438 6381 907 438 ijmrset@gmail.com www.ijmrset.com

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 1, January 2024 |

| DOI:10.15680/IJMRSET.2024.0701008 |

Data Mining Techniques to Analyze the Impact of Social Media on Academic Performance of School Students. : A Review

Tanya jain

Assistant Professor, Department of Computer Science, Madhyanchal Professional University, Bhopal, India

Abstract: The motive of purpose of educational institutionss is to provide quality education to their students. However, it is difficult to analyze data manually. Educational data mining is more effective as compared to statistics methods used to explore data in educational settings to analyze student's performance. The objective of the study is to use different data mining techniques and find their performance and impact of different features on students' academic perfrmance. The dataset was collected from the Kaggle repository. To analyze the dataset, different classification algorithms were applied like decision tree, random forest, SVM classifier, SGD classifier, Ada Boost classifier, and LR classifier. This research revealed that random forest achieved a higher score (90%). The score of decision tree, Ada Boost, logistic regresion, SVM, and SGD is 89%, 86%, 83%, 80%, and 77%, respectively. Results show that technology g.reatly influences student performance. The students who use social media throughout the week showed low performance as compared to the students who use it only at weekends. Further more, the impact of other features on the performance of students is also measured.

I. INTRODUCTION

Student's performance modeling is one of the challenging and popular research topics in educational data mining (EDM) [1]. Multiple factors influence the academic performance in nonlinear ways. The widespread availability of educational datasets further made educational data mining more attractive to the researchers. EDM is a field in which data mining algorithms are applied on educational data to improve and predict the performance of education in institute students [2].

Information Technology (IT) is an important part of learning process [3]. It greatly influences the online student performance and GPA. The study [4] believes that the use of technology such as internet is one of the most important factors that can influence the educational performance of students positively or negatively. Students are spending too much time on social media sites like Facebook and do have not enough time to study. This behavior leads students toward poor performance during high school studies, and they consider themselves difficult to survive in higher studies. EDM can detect this poor behavior pattern at right time to maximize the student grades and minimize the failure rate of weak students.

Social media greatly influence the school-age students. Social media influences students' academic and personal lives. Students use social media for academic purposes to improve their performance. Teachers and students both can use social media as a teaching and learning tool for ease and improve learning and teaching process [5].

The objective of the study is to predict student performance with the use of technology, weekday-social-media-use, and weekend-social-media-use and, furthermore, to find out the student who desires to get higher education in advance and to find how parent's education influences the student performance.

In this paper, data is collected from Callboard 360 LMS (learning management system). We used six machine learning techniques (DT, SVM classifier, SGD classifier, RF classifier, AdaBoost, and LR classifier) to determine the patterns inside the student performance data.

The results show that technology greatly influences the student performance. Six classifiers are used to identify the performance on the basis of different features like romantic status, use of technology, weekday-social-media-use, weekend-social-media-use, parent education, and living area. It helps the teachers to identify the fair, good, and poor students.

The proposed work analyzes performance and finds the desirable and undesirable student behaviors of students, which will help to combine students into classes based on different performance capabilities, furthermore predicting the student's social activities.

II. RELATE WORK

Educational institutes face different problems to identify reasons of drop-out, graduate not on time, pass to fail ratio, effect of parent's involvement on student performance, effect of attendance on student performance, predicting student's performance on the basis of previous marks, and many more. For solving such problems, several studies present machine learning and statistical solutions.

The study [4] believes that the use of technology such as internet is one of the factors that can influence educational performance of students positively or negatively. Students are spending too much time on social media websites like Facebook and do have not enough time to study which leads students towards poor performance during high school studies and ultimately consider themselves difficult to survive in higher studies. EDM can detect this poor behavior pattern at right time to maximize the student grades and minimize the failure rate of students. Another study [5] shows that social media influences the school-age students positively. Students use social media for academic purposes to improve their performance. Teachers and students both can use social media as a teaching and learning tool for ease and improve learning and teaching process.

Study [6] used decision tree, a nonlinear classifier to generate tree and rules. For analyzing results, the J48 algorithm is used as an analyzing tool. Dataset is collected through the surveys of students of master and Ph.D.

Reference [10] declared that data mining techniques increasingly merged day by day with the educational field. Data mining and education field are combined called educational data mining that help to identify the features and information of students. This study uses to predict and analyze the performance of bachelor and master students at university level students. The performance is analyzed with two algorithms: decision tree and fuzzy genetic algorithms. The dataset contains features internal-marks, sessional-marks, and admission-marks which are used to identify the results. Internal-marks contain attendance-marks, AVG-marks, sessional-marks, and assignment-marks. Weighted marks obtain from matric and interclass. In master degree, examination marks are also included. A systematic model is used to enhance the performance of students in the early stage and in time. To find the result and solution in early stage, conduct good result in final examination. Students also view their result and new updates. Many companies connect to the educational organizations to find out students according to their needs.

Reference [11] declared that large amount of data is stored in different technological spaces and makes new data quickly and easily. Data mining is also combined with these technologies. With the help of data mining techniques, important information from ordinary data can be taken out. Because of these techniques, data can be produced quickly and easily daily or each second. Using data mining methods provide meaningful knowledge. An educational database contains huge amount of data related to student data mining methods applying on this data. This study defines how to use DM algorithms such as KNN, naïve Bayes, and DT algorithm. Apply these algorithms on student raw data and find the best result.

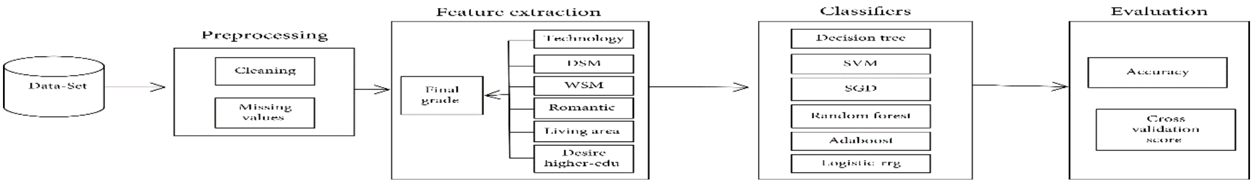
Reference [12] narrated that college students have great facility of internet. Internet educates the students in living and learning process. This study discloses the connection between internet use behavior and educational performance of students. It also analyzes students that are undergraduates by using machine learning algorithms. The dataset of 4000 students has attributes of online-duration, internet-traffic-volume, and connection-frequency, which were extracted, calculated, and normalized from the real internet usage. DT, NN, and SVM were used to find student educational performance by using these attributes. Internet-con and frequency attribute are positively linked, and internet-traffic-volume attribute is negatively linked with academic performance of students. The online-time and internet-time suffering results in surprising performance among different datasets. The number of features increases and improves accuracy. The results define that internet usage is able to distinguish and analyze student's academic performance.

Reference [13] narrated that in higher education, data mining approaches are used and createmeaningful data from large meaningless data. By using a supervised data mining method, find the results of student progress. To find the student progress is helpful for current educational organizations. The basic purpose of the study is to make a model with the help of classification methods. This model analyzes the student performance in Malaysia. This model is used to find the most important features from the large dataset. Many approaches which are KNN, naïve Bayes, DT, and logistic regression approaches are used to analyze the student academic result performance. These approaches are based on accuracy measure, precision, recall, and ROC curve. The output showing the naïve Bayes algorithm is better. NB is disclosing important attributes that are used to find excellent students whose grades are A+ and A.

Reference [14] reported that large number of students dropped out a major worry of higher education organizations. It greatly influences the fee of students and discarded public resources. It is necessary to find those students who are in danger of dropping out and find those attributes that are the cause of higher dropout rate. Educational data mining methods are used to recover this problem. In this study, the University Teknologi MARA students of computer science undergraduate after three years. DT, logistic regression, random forest, KNN, and NN algorithms are matched to analyze student performance. Several machine learning methods are combined and make an efficient model. The logistic regression method is the best algorithm to analyzing and predicting the dropout students.

Educational Data Mining Model

This study evaluates the impact of technology on student's educational performance. This study proposed the educational data mining (EDM) model that is divided into five major sections such as collection of dataset, preprocessing of dataset, feature extraction, selection of classifier, and model evaluation, see Figure 1. Each section may contain more than one subsection. In step one, dataset was cleaned and checked if there is no missing value. After cleaning the dataset, required features were extracted from the dataset like "use of technology," "weekly-social-media-use," or "weekday-social-media-use." In the next step, different learning models were used to predict student's final grade performance. After that, model's performance was compared based on accuracy score to select the best learner for the problem. The algorithms used in the study are DT, SVM classifier, SGD classifier, random forest classifier, AdaBoost, and logistic regression classifier.



Dataset Collection

The data used for the analysis is collected from an electronic-learning system called Kalboard 360.

Preprocessing

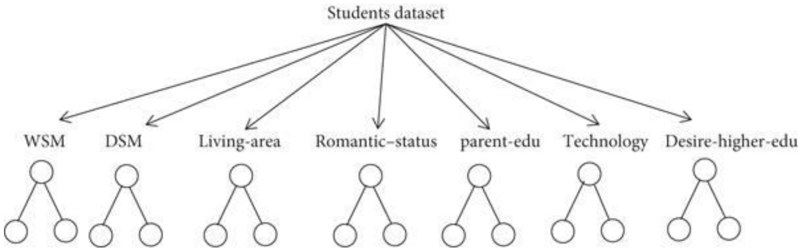
In the preprocessing phase, first, we make sure that there is no irrelevant and unacceptable value existed inside the dataset. This process is called cleaning. After cleaning process, we analyzed the data and removed unnecessary fields that are not relevant to our research objective. This process makes data more refined and relevant to research objective. In the preprocessing, we also handled the null values in the dataset.

Selection of Classifier

After obtaining the required features, different classifiers were trained on the dataset. The algorithms used in the study are DT, SVM classifier, SGD classifier, random forest classifier, AdaBoost, and LR classifier.

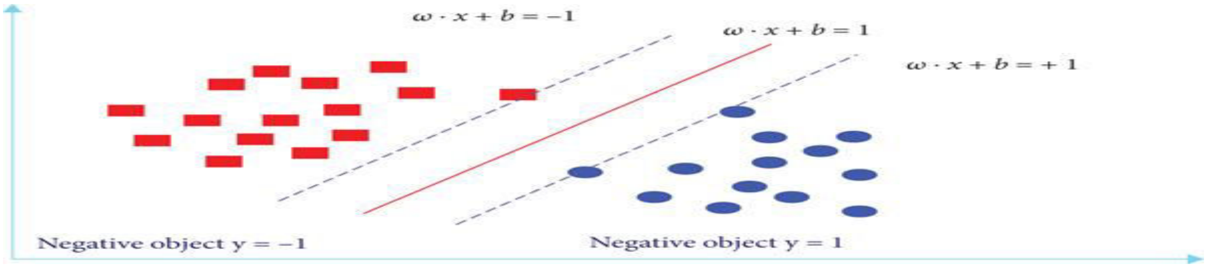
Decision Tree

The DT classifier is simple and understandable by analysts and end users. It is a tree shape model built based on the features, see Figure 2. These are WSM, DSM, living area, romantic status, parent education, technology, and desire-higher-education; these all features are called nodes and influence the student final scores. Every node is divided into subnodes. Every node makes decision on the basis of numeric value.



SVM Classifier

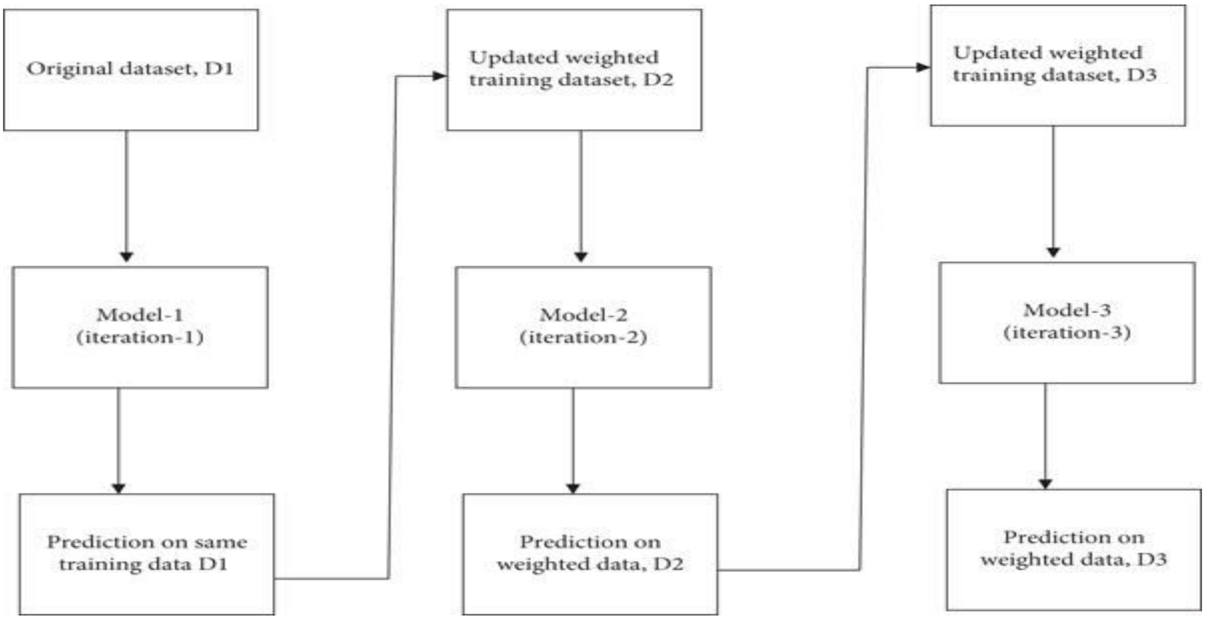
This classifier is a linear algorithm that is suitable for small datasets. Support vector machines are not suitable for large datasets because it takes small memory and needs more training time. I used this classifier because my dataset is small; it correctly classifies features that are influencing the student performance. It divides features into two classes, for example, living area divides into a rural and urban area and the use of technology divides the yes or no class. The study divides the use of internet into two classes such as “low use” and “high use”. If a person uses 1-2 days, then it is considered in “low use” and -1 weight is given to the user. If a person uses 4-5 days, then it is considered in “high use” and 1 weight is given to the user. In case of 3 days, 0 weight is given to it. All categories are shown in Figure 3.



AdaBoost

The AdaBoost classifier is a meta-algorithm of machine learning. Meta-algorithms mean different low accuracy classifiers merged into a single highly predictive model to increase performance. This classifier is sensitive to error porn data and outliers. This algorithm is less risky in overfitting problems as compared to other algorithms. The AdaBoost classifier is used to build a high-performance classifier whose accuracy is high. This classifier combines weak and poor classifiers and makes a strong and highly performing classifier.

As shown in Figure 5, the AdaBoost classifier works in the following steps:(1)Firstly, AdaBoost selects training samples randomly(2)It trains the AdaBoost machine learning algorithm by selecting the samples based on the correct analysis of the last training(3)It allocates the higher weight to wrong classified samples so the next repetition of classification gets the high probability for classification(4)It allocates the weight to the trained classifier in each repetition according to the accuracy of the classifier. It generates a high-performance classifier(5)This process repeats until the complete training samples fits without any error(6)To perform voting process on all the learning algorithms you generate


Original dataset, D1
Updated weighted training dataset, D2
Updated weighted training dataset, D3
Model-1 (iteration-1)
Model-2 (iteration-2)
Model-3 (iteration-3)
Prediction on same training data D1
Prediction on weighted data, D2
Prediction on weighted data, D3

SGD Classifier

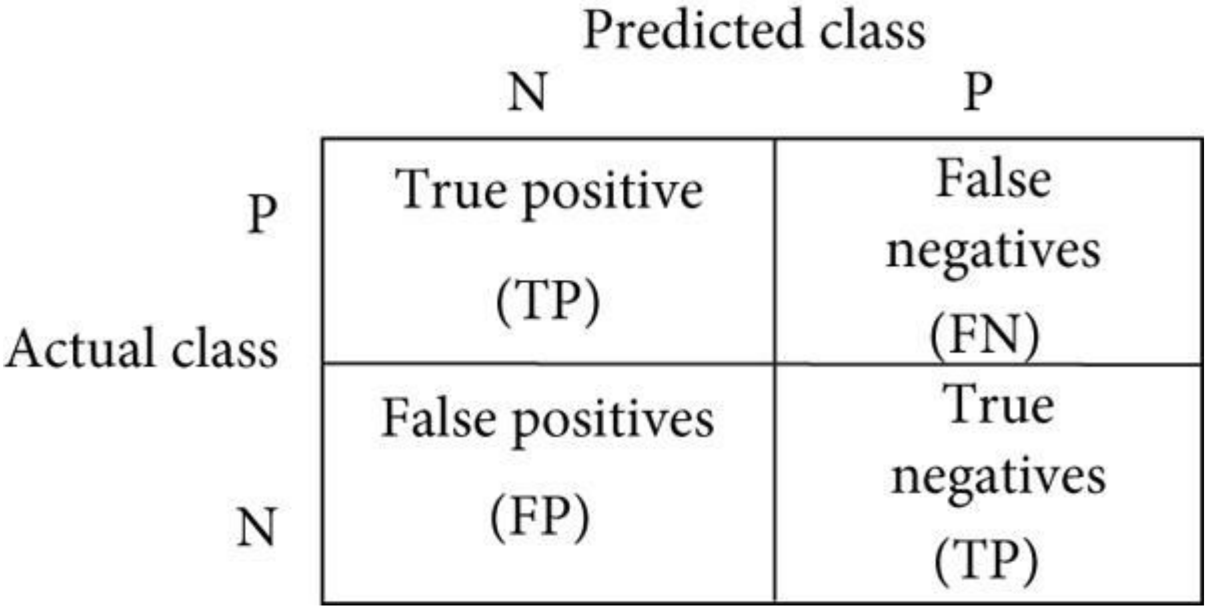
Stochastic Gradient Descent (SGD) is an optimization technique. In the Stochastic Gradient Descent approach, complete dataset is not selected; some samples are selected randomly. Total samples are called a batch. The batch is created from the complete dataset. The complexity is high when the dataset is big. It uses a single sample of data. The next time the sample is exchanged with the next sample randomly and then performs repetition. It increases the efficiency of the classifier and is easy to implement.

Results

The results of the educational data mining model are presented as follows.

Model Evaluation

The evaluation phase of our model analyzes the outcomes of every classifier on the basis of the following factors. Confusion or error matrix is used to evaluate the performance of a classifier, see Figure 6.



		Predicted class	
		N	P
Actual class	P	True positive (TP)	False negatives (FN)
	N	False positives (FP)	True negatives (TP)

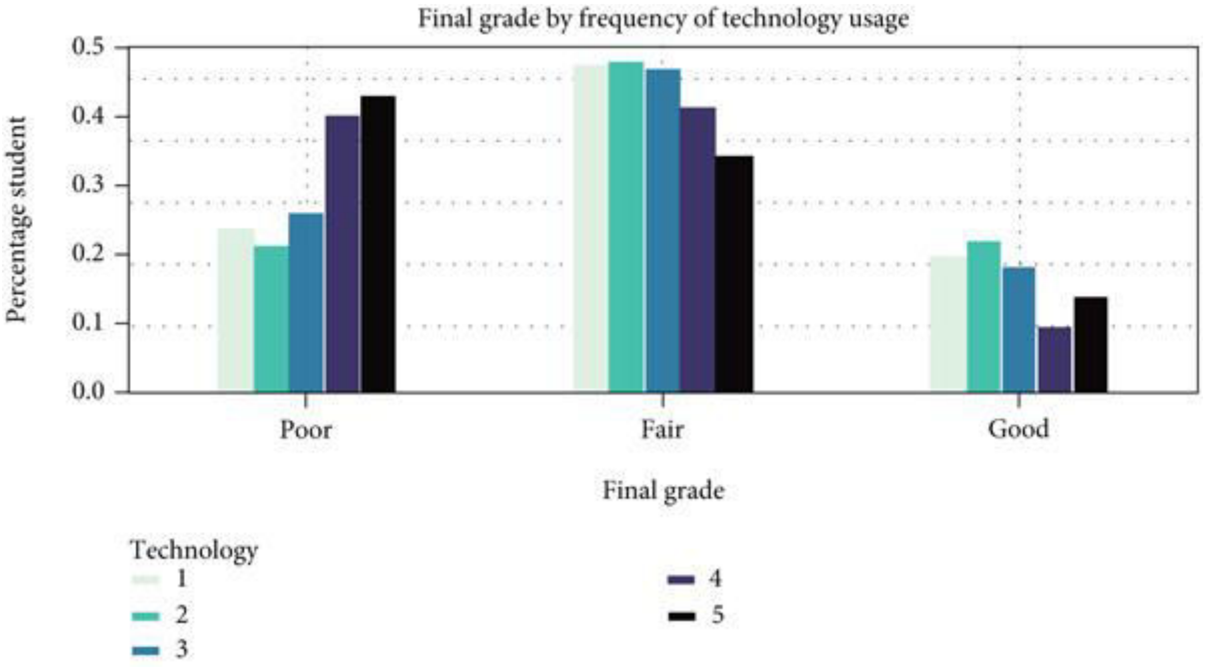
Accuracy is the basic evaluation metric to analyze the rate of correctness of the prediction. The accuracy is measured with a formula, see the following:

Correlation Heat Map

A heat map is a simple and useful tool to find out useful attributes in a dataset. Diagram represents correlation between different features. Value 1 shows two feathers are positively correlated, and when the correlation is closer to or similar to -1 increase or decrease, one variable value will decrease or increase the other variable. The main advantage to use a heat map is how a feature is useful according to my problem and cleans my dataset before its use and execution. The correlation heat map is shown in Figure 7.

Final Grade by Frequency of Technology Usage

The use of technology depends on the understanding of devices that are used by students. Good students understand the technology and use it for their studies. The performance is high because they understand the technology and cannot waste their time. The poor students are not capable of using technology. When they use devices, they cannot understand what they are working, so they waste their time and energy. So, that type of student has low performance. The fair students are that some students understand, and some cannot understand the technology; some students use technology but cannot improve their performance because they have no proper guidelines and training to use technology.

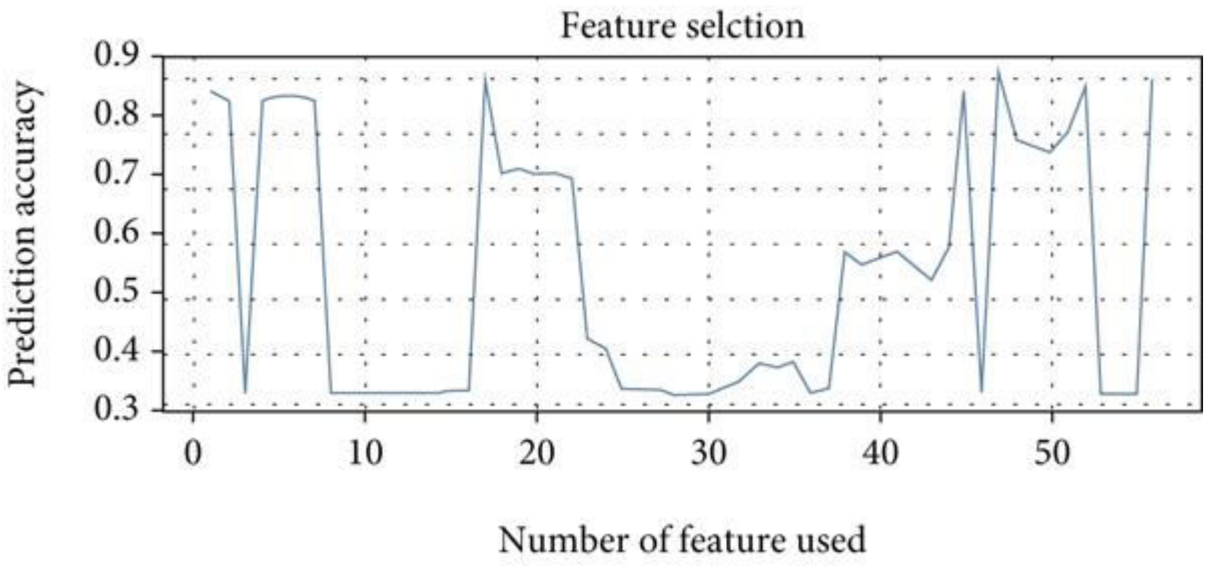


Final Grade by Social Media Consumption

Figure 10 shows good or intelligent, fair or normal, and poor or weak students use weekly social media. The use of social media divides into five levels (1, 2, 3, 4, and 5). Low social media usage on the weekend is represented by 1 and 2. The highest use of social media is represented by 4 and 5 levels. The medium use of social media on the weekend is represented by 3. Poor students that are weak in their studies use social media which results in their performance becoming slower. Fair and good students use social media highly; then, their performance is also decreasing. So, the high use of social media also influences the student performance.

Feature Effect

Figure 13 shows that the number of features increases the prediction accuracy of a classifier. Multiple features help the classifier to train and get accurate results. But condition is that features are correlated to the problem. Relevant features greatly influence the accuracy, but irrelevant features decrease the accuracy. On the other hand, multiple features can complex the classifiers and some classifier like SVM cannot work on large number of features because it has a limited memory. Deep learning correctly classifies the large number of features. So, we can say that relevant large number of features improves the accuracy, but the multiple features also increase the complexity of a classifier. Sometimes, using multiple features cannot increase the prediction accuracy because the features are irrelevant to the problem, and sometimes, a small number of features greatly influence the prediction accuracy.



Final Results of Classifiers

Supervised learning algorithms used to predict the student academic performance with the use of technology. These algorithms are DT, random forest, SVM, L-regression, AdaBoost, and SGD. The score of decision tree is 0.89% random forest score is 0.97%, support vector classifier score is 0.86%, logistic regression is 0.88%, AdaBoost is 0.87%, and SGD classifier score is 0.82%. The scores prove that the random forest classifier has the best results as compared to other classifiers. The DT classifier is the second one. The decision tree classifier score is lower than random forest because decision tree has problem of overfitting. The Stochastic Gradient Descent classifier gains the lowest scores. The comparison of classifier

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Acknowledgments

We appreciate the partial collaboration with Universiti Malaysia Sabah, Malaysia, in this work.

REFERENCES

1. Saman Amjad,[1]Muhammad Younas,[1]Muhammad Anwar,[2]Qaisar Shaheen,[3]Muhammad Shiraz, Gani “Data Mining Techniques to Analyze the Impact of Social Media on Academic Performance of High School Students Volume 2022
2. M. S. Chen, J. Han, and P. S. Yu, “Data mining: an overview from a database perspective,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866–883, 1996.

3. V. Ramesh, P. Parkaviand, and K. Ramar, "Predicting student performance: a statistical and data mining R. Sivakumar, "Effects of Social Media on Academic Performance of the Students"," *The Online Journal of Distance Education and e-Learning*, vol. 8, no. 2, p. 90, 2020.
4. S. Naseem, A. Alhudhaif, M. Anwar, K. N. Qureshi, and G. Jeon, "Artificial general intelligence based rational behavior detection using cognitive correlates for tracking online harms," Tech. Rep., Personal and Ubiquitous Computing, 2022.
5. J. Cradler, M. McNabb, M. Freeman, and R. Burchett, "How does technology influence student learning," *Learning and Leading with Technology*, vol. 29, no. 8, pp. 46–49, 2002.

INTERNATIONAL JOURNAL OF
MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com