An Optimal Random Forest Model for Enhancing Decision-Making in Improving Students’ Performance via Educational Data Mining

Samuel Odoom[1], Eric Opoku Osei[2]

[1]Department of Computer Science, Kwame Nkrumah University of Science and Technology
[2]Ph.D Department of Computer Science, Kwame Nkrumah University of Science and Technology

Abstract

The need for measures to enhance students’ performance, especially in Science and Math as a panacea for development in this 4th industrial revolution has become a global call. The large stream and heterogeneous nature of educational data make it difficult to use traditional statistical methods for data analysis to support decisions. However, Educational data mining has been effective and efficient in addressing this issue but much focus of researchers has been on only students' academic records to determine performance. This study sought to propose a Random Forest model to predict and enhance students’ performance. The study adopted a 5-stage mining process to mine psycho-socio-economic demographics educational data with Mutual information, Chi-squared test, Featurewiz, RandomizedSearch CV, and GridSearch CV to optimize the model’s performance. The study’s outcome revealed the key factors affecting students’ performance and that the model was enhanced by a 10.4% increment in precision and f1-score and 9.1% recall value, 7.1secs (62%) improvement in execution time and 78.7% improvement in Root Mean Square Error. This outcome remains a contribution to guiding decision-making in the educational setting and a basis for further studies on model optimization.

Keywords: Enhancing students’ performance, Decision support model, Educational data mining, Optimized Random Forest model, and Machine Learning.

1. Introduction

Science and technology as a panacea for countries’ growth and development has gained global recognition in this 4th Industrial Revolution (4IR) regime [1]. As posited by [2], technology, science, and knowledge-based information are the forces driving the world’s economy. [3] also joined the argument by stipulating that the existence of knowledge in science and technology which has mathematics as the foundation is a necessity to attain growth and development in every facet of life, hence, governments worldwide are compelled to highly prioritize investment in the education of science, technology, engineering, and mathematics (STEM). However, studies have revealed that the performance of students in math and science, especially in the second cycle has not been encouraging and remained a major challenge which persistently calls for solutions to remedy the situation [4] [5] [6] [7]. Thus, has left many countries, especially the developing ones to be trailing and deficient in digital and STEM skills required for growth and development in this era. This situation has rendered such countries meek in pursuing SDG 17 (17:6-17.8) and SDG 4 (4.1-4.C).

In the quest to mitigate this challenge, [17] postulated that the nature and volume of data generated by contemporary systems make it difficult for traditional methods and statistical tools to be used effectively for the required analysis, hence the necessitates for a new paradigm to be explored, thus ML. This has yielded an enormous prediction model deployed from machine learning (ML) to mine data from the educational setting to predict students' performance [17].

Most of these ML models are deployed in aid of facilitating and enhancing decision-making in the educational environment for students' performance improvement. However, most of the studies on educational data mining have focused on academic data/records to predict students' performance without or with little attention paid to the non-academic records [7], [2], [8], and [16].

This study was conceived on the premise of proposing a prediction model for assessing students' performance based on non-academic records (psycho-socio-economic demographics) from the educational setting to support decisions to improve students' performance, especially in Math and Science [9].

2. Related Work

2.1 The concept of Machine learning (ML)

The collection of data about instructional processes in education presents enormous opportunities for the enhancement of teaching and learning environment via the implementation of new instructional experiences [10], hence, analyzing data generated from the educational setting results in enhancement in students' behavior prediction, learning analytics, and new paradigm approach to policy implementation in education. Machine Learning technique is a term that describes the ways by which hidden but useful relationships and patterns among features in a given dataset are revealed via analysis with ML algorithm(s). In EDM, the prediction of students' academic performance, the tendency to dropout, violence, success in enrolled programs, and guidance and counseling needs, just to mention a few constitute the application of ML [11] [12].

2.2 Application of Machine Learning

Machine learning (ML) is a rapidly growing field of computer science that focuses on the development of algorithms and models that can learn from data and make predictions or decisions without explicit programming. ML algorithms have been applied to various tasks in a wide range of fields, including natural language processing, computer vision, robotics, finance, healthcare, and many others [13]. According to [14], in the field of computer vision, ML algorithms have been used to improve the accuracy of image recognition and object detection tasks. He further posited that; ML algorithms have been used to improve the accuracy of stock market prediction tasks in the field of finance. Also, ML algorithms have been used to improve the accuracy of portfolio optimization tasks [15]. They also articulated that, ML algorithms have been used to improve the accuracy of text classification, sentiment analysis, language translation, and other tasks in the field of natural language processing (NLP). The healthcare sector has also experienced a massive improvement in the accuracy of medical diagnosis tasks with the help of ML algorithms. In the field of education, ML's application has resulted in a new paradigm of research termed Educational data mining. In EDM, a review of studies on the application of ML in education revealed the following; prediction of students' academic performance, tendency of dropout, violence, success in enrolled programs, guidance and counselling needs, just to mention a few [11][12].

2.3 Proposed Models for Performance Prediction

[19] conducted a study to predict students' performance in final exams based on mid-semester results using Random Forest (RF), NN, Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), and K-NN. The results revealed that RF recorded the highest accuracy of 74.6%. However, the study failed to explore the impact of non-academic factors on students' performance. [1] then conducted a study to accommodate this gap by employing Multi-layer perceptron (MLP) and BLR to explore inherent factors accountable for students' performance. The findings show that students' perception of the subject, parents' level of education, and accommodation status are the most determinant factors for predicting students' performance. This was criticized by [20] for not exploring most of the spatial and behavioral features of the study. A study was then conducted on this premise using the Fuzzy Delphi method to solicit experts' views on factors relevant to predicting students' performance. The study then integrated a geo-spatial-based ML technique to determine the correlation between academic semester behavior and location features to predict academic performance. The outcome of the study revealed that academic factors, socioeconomic factors, and semester behavior were the major factors affecting students' performance. Though their study explored additional factors to previous studies but neglected the psychological and social factors that are equally relevant in determining students' performance. In addition to the above, [4] [18] also revealed in their criticism that most of the studies have placed less focus on enhancing the proposed models' performance. They proceeded to emphasize the need to use only features relevant to the prediction of the target variable as a means to attain optimization of the model's performance. Their study proposed an ensembled Chi-MI, thus merging Chi-Square with Mutual Information as a feature selection model. The outcome of their model was encouraging but the discovery of Featurewiz necessitated a study to leverage the assembling of the three (3) coupled with hyper-parameters tunning techniques to attain more better outcome in terms of performance optimization of the proposed model.

3. Methodology

The study sought to propose a decision support model with a Random Forest algorithm to facilitate decision-making according to enhancing students' performance in Math and Science at the senior high level given psycho-socio-economic factors. A five-stage process of discovering knowledge based on an integration of elements from the CRISP-DM and KDD processes model of mining data was employed for this study [26][27]. The choice of the proposed ML algorithm and data mining process model was informed by the miner's quest to leverage the optimization power of the combination of feature selection and parameters-tuning techniques to enhance the performance of the proposed model compared to other similar models for the prediction task [23][18][14][21][24]. The various processes that constitute the mining process model have been elaborated on below.

3.1 Data gathering

Data used for the study include students' demographics, personal-related factors, socio-economic factors, school-related factors, and academic records. Part of the data was retrieved from the Computerized Schools Selection and Placement System (CSSPS), an information system designated for newly admitted senior high students. [1] and was coupled with data collected via Questionnaire. A questionnaire of test items from the International Personality Item Pool (IPIP) was administered to collect primary data that was not available but required for the study in the educational setting with a justification of inclusion of items for the construction of the questionnaire given the study's objectives from the EDM literature.

Table 1: Summary of a description of the dataset used for modeling

Category	Code	Description	Values
Student's Background data	Program	The course pursued by a student	"General Arts/General Science/Business/Home Economics
	Gender	Gender of Student	"Male/Female"
	FamTyp	The nature of the student's family	"Extended/Nuclear/Single Parenting/Reconstituted"
	PSP	Perceived Style of Parenting	"Neglectful/Authoritative/Authoritarian
	FIL	Family Income Level	"High/Moderate/Low
	Scholarship	Student's scholarship status	"Yes/No"
	LFFE	Level of Father's Formal Education	"Tertiary/Secondary/Basic/None
	LMFE	Level of Mother's Formal Education	"Tertiary/Secondary/Basic/None
	BO	Birth Order of the student	"First/Middle/Last"
	Community	Student's community of residence	"Urban/Rural"
	BECE	Student's BECE grade	"6 – 54"
Emotional Intelligence Test	EQ_SelfMotivation	Emotional Quotient (Self-Motivation)	"Good/Average/Poor"
	EQ_SelfAwareness	Emotional Quotient (Self-Awareness)	"Good/Average/Poor"
	EQ_ManagingEmotions	Emotional Quotient (Managing Emotions)	"Good/Average/Poor"
	EQ_SocialSkills	Emotional Quotient (Social Skills)	"Good/Average/Poor"
Personality Trait Test	PT_Extraversion	Personality Trait (Extraversion)	"High/Average/Low"
	PT_Agreeableness	Personality Trait (Agreeableness)	"High/Average/Low"
	PT_Conscientiousness	Personality Trait (Conscientiousness)	"High/Average/Low"
	PT_Neuroticism	Personality Trait (Neuroticism)	"High/Average/Low"
	PT_Openness	Personality Trait (Openness)	"High/Average/Low"

Category	Code	Description	Values
Teacher's Attribute	STA	Science Teacher's Attributes on Competency	"Highly Competent/Competent/Less Competent/Incompetent"
	MTA	Math Teacher's Attributes on Competency	"Highly Competent/Competent/Less Competent/Incompetent"
School's Environment	SSE	Science School Environment Conduciveness	"Highly Conducive/ Conducive /Less Conducive /Not Conducive"
	MSE	Math School Environment Conduciveness	"Highly Conducive/ Conducive /Less Conducive /Not Conducive"
Academic Resources	SAR	Availability of science Academic resources	"Adequately Available/ Available /Inadequately Available /Not Available"
	MAR	Availability of Math Academic resources	"Adequately Available/ Available /Inadequately Available /Not Available"
Student's Motivation	SSM	Student's motivation for learning science	"Highly Motivated/ Motivated /Less Motivated /Not Motivated"
	MSM	Student's motivation for learning Math	"Highly Motivated/ Motivated /Less Motivated /Not Motivated"
Target Variable	Science Remarks	Student's performance in Science	"PASS/FAIL"
	Math Remarks	Student's performance in Math	"PASS/FAIL"

3.2 Preprocessing of Data

The researcher committed a conscious effort to improve and change the nature and format of data to enhance its quality in pursuit of the model's performance optimization. The preprocessing activities entailed; importing the dataset, statistical description of the dataset, cleaning of data, exploratory data analysis, correlation analysis, label encoding (feature scaling), and definition of features and labels of the dataset (splitting dataset) [28][25].

3.3 Modelling

Modeling was done in two main stages in each experiment, thus, the training and the testing stage where the train set and test set were used respectively. The result of the optimized model was evaluated and compared to the initial outcome to determine whether the model's performance had been improved. Once optimization was attained, then the results as well as the discovered information were analyzed.

3.4 Optimization of Proposed Model

3.4.1 Feature Selection (FS)

As part of the model's performance optimization process, dominant features were selected from among all the features in the entire dataset using Chi-square, Mutual Information (MU), and FeatureWiz feature selection methods [14][1].

i) The Chi-squared test is used to determine whether two given variables (actual and expected) relate in a way. Thus, the actual can be based on anticipating the expected. Chi-Square is mathematically denoted as

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

O_i *implies the actual event* $(Observed\ value)$ *and* E_i *denotes the expected value*. When implemented, the Chi-square algorithm produced probability values (p-values) as well to aid in the determination of the correlation among the attributes.

ii) Mutual Information (MI) on the other hand, determines the dependency of two or more variables. The lower the value, the more independent the variables, and vice versa. MI is represented mathematically as

$$I(X;Y) = H(X) - H(X|Y). \quad (2)$$

iii) FeatureWiz is one of the libraries in Python for automatic feature selection. The ultimate function of featureWiz is to extract the most significant features from any given set of data given the target variable. Hence, after their implementation, the miner was able to extract 17 relevant dominant features which were sandwiched with the ones selected by the Chi-Square test and Mutual Information to get the most relevant features for the modeling.

3.4.2 Parameters-Tuning

The study utilized both RandomizedSearch Cross-Validation (RS CV) and GridSearch Cross-Validation (GS CV) to enhance the performance of the proposed model via tuning and producing the best parameters used by the RF algorithm to optimize the mining outcome. After the initial modeling, which was done with basic (default) parameters of the RF algorithm, RS CV and GS CV were deployed to identify the best parameters and used for subsequent modeling. The researcher then compared the results emanated from both modeling processes to demonstrate the optimization of the proposed model's performance given the feature selection techniques and parameter-tuning methods applied.

3.5 Model Performance Evaluation (MPE)

Confusion Matrix, F-Score, Accuracy, Precision, Recall, and Area under the curve (AUC) – Receiver Operating Characteristic curve (ROC) were used to evaluate the model's performance.

Accuracy: (A) $A = \frac{CP}{TP}$ (3)

where CP is the number of correctly predicted instances and TP is the total number of predictions.

Precision: is determined as: $P = \frac{TP}{TP+FP}$ (4)

TP implies True positive and FP is for False positive. F1-Score: computed as:

$$F - score = 2 * \frac{P*R}{P+R} \quad (5)$$

P is precision and R is Recall.

ROC Area: To plot the curve, True Positive Rate (TPR) which connotes Recall, and True Negative Rate (TNR) which represents Specificity are used as parameters. Mathematically, $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{TP + FP}$ (6)

TP is True Positive, FP is False Positive, TN is True Negative, and FN is False Negative.

Root Mean Square Error (RMSE): is represented as = $\sqrt{\sum_{i=1}^{S}\left(\frac{Predicted\ \ Values_i - Actual\ Values_i}{S}\right)^2}$ (7)

where S = the number of data points in the entire iteration.

Macro Average Accuracies: is represented as = $\left[\frac{F1-score1+f1-score2+\cdots+f1-scoreN}{n}\right]$ (8)

where 'n' implies the number of classes.

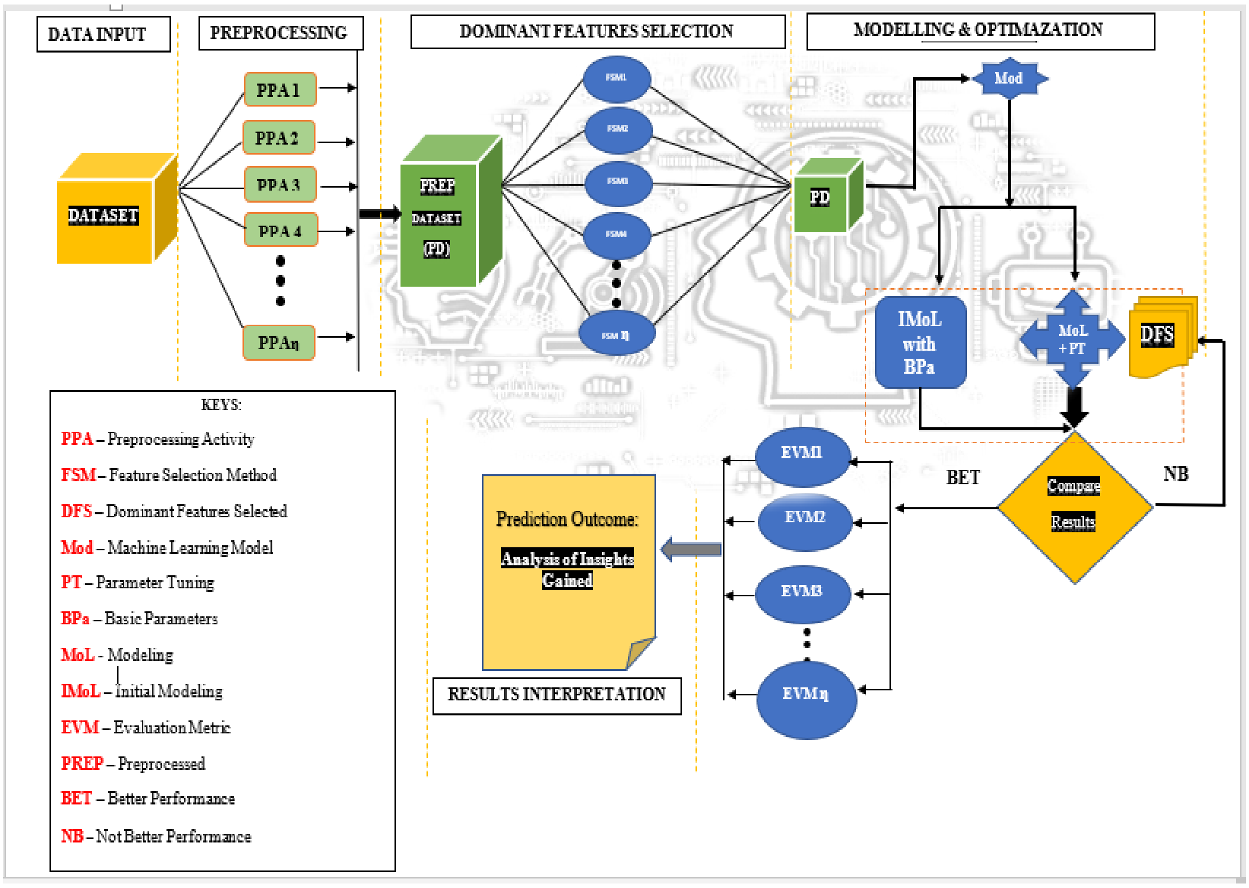
Weighted Average Accuracies: is = $\left[\frac{(F1-score1 * t(s))+(f1-score2 * t(s))+\cdots+(f1-scoreN* t(s))}{n}\right]$ (9)

where 'n' implies the number of classes and t(s) connotes the proportion of supported instances for each class.

3.6 Interpretation of Modeling Outcome

The relevant but hidden information were unveiled from the dataset according to dominant features impacting the performance of students in science and Mathematics at the senior high. Hence, discussions were made on how to prioritize decision-makers and authorities in the educational setting especially, at the senior high, in terms of planning and implementing remedial actions given limited resources to maximize outcomes. Visualization in terms of patterns and evaluation metrics of modeling outcomes were made to facilitate understanding that constituted the study's results. Optimization of the RF model was also analyzed in line with how the discoveries from the model can support decision-making on students' academic performance was also presented. Figure 1 below depicts the mining process.

Figure 1: Conceptual framework for the mining process



4. Results and Discussion

The modeling was executed in two (2) major experiments viz; initial and optimized modeling. The results emanated from the experiments have been presented and discussed below.

The initial modeling was done with the default basic parameters that come with the RF algorithm as well as initial features obtained after feature extraction at the preprocessing stage and the results have been depicted in Tables 2 and 3 below.

Table 2: Initial model's performance accuracies

	MODEL'S PERFORMANCE ACCURACIES					
TASK	MACRO AVERAGE			WEIGHTED AVERAGE		
	Precision(%)	Recall(%)	F1-Score(%)	Precision(%)	Recall(%)	F1-Score(%)
Math	87	89	87	89	87	87
Science	88	88	88	88	88	88

Table 3: Initial model's performance evaluation

MODEL EVALUATED BY:							
TASK	CV(%)	RMSE	TP(%)	FP(%)	TN(%)	FN(%)	ROC	TIME
Math	87	0.3559	95	5	82	18	0.9186	1.35secs
Science	87	0.3464	91	9	85	15	0.9215	1.24secs

From Tables 2 and 3 above, the model recorded a cross-validated (CV) accuracy of 87% for the prediction of both Math and Science with an error rate (RMSE) of 0.3559 and 0.3464 within an execution time of 1.35sec and 1.24sec respectively. The AUC-ROC was also recorded as 92% approximately. Given the classification rates, the TP was 95% and 5% FP for Math with 91% TP and 9% FP for Science whilst 82% and 85% TN with 18% and 15% FN were recorded for both classes respectively.

In the optimization modeling, there were also two (2) major experiments carried out with each having two (2) sessions for each class prediction. Combining the outcome of the feature selection techniques with Hyper-parameters tuning. The optimum outcomes obtained have been presented in the Tables below.

Table 4: Post-optimization accuracies of the model's performance

	ACCURACIES					
TARGET VARIABLE	MACRO AVERAGE			WEIGHTED AVERAGE		
	Precision(%)	Recall(%)	F1-Score(%)	Precision(%)	Recall(%)	F1-Score(%)
MATH	95	95	95	95	95	95
SCIENCE	96	96	96	96	96	96

Table 4 above presents results obtained from optimizing the proposed model with features selected after applying the Chi-Squared test, Mutual Information, and FeatureWiz methods coupled with GridSearchedCv (GSCV) and RandomisedSearchCv (RSCV) for the prediction of students' performance in both subjects. The outcome as shown in Table 3 above revealed that the model attained 95% and 96% accuracies for both Math and Science respectively, when macro and weighted averages were used as assessment metrics given precision, recall, and f1-scores.

In addition to the accuracies attained in the post-optimization modeling, the model was further evaluated by classification report, AUC-ROC, Cross-validated accuracy, and time for execution. The results obtained are presented in Table 5 below.

Table 5: Post-optimization model's performance evaluation

MODEL EVALUATED BY:								
TASK	CV(%)	RMSE	TP(%)	FP(%)	TN(%)	FN(%)	ROC	TIME
Math	89	0.2309	99	1	91	9	0.9680	235ms
Science	89	0.2000	99	1	95	5	0.9743	189ms

The data in Table 5 above reveals that the model attained 89% CV accuracy for the prediction of students' performance in both subjects with error rates of 0.2309 and 0.2000 for Math and Science respectively. In the classification report, the model recorded 99% True Positive (TP) with 1% False Positive (FP) as classification rates for students' performance in both Math and Science. The scores for the Negative rates were; 91% True Negative (TN) with 9% False Negative (FN) for Math and 95% TN with 5% FN for Science given the time for completing execution as 0.235secs and 0.189secs for both Math and Science respectively at AUC-ROC value of approximately 0.97 (97%) for the data used.

To validate the performance of the model, the miner performs a comparative analysis of the model's performance given accuracy, execution time, RMSE, Macro, and Weighted Average (Avg) in both the initial and post-optimization experiments using the results from the initial modeling as a baseline for the assessment. The results of the analysis are presented in Table 6 below.

Table 6: Comparative analysis of model's performance

ANALYSIS OF MODEL'S PERFORMANCE					
Modelling	Accuracy (%)	Ex. Time	RMSE	Macro Avg (%)	Weighted Avg (%)
Initial	87	1240ms	0.2000	88	88
Post-Optimization	96	189ms	0.3464	96	96
% Increment	10	85	42	9	9

The results from the comparative analysis presented in Table 6 revealed that the model's performance was significantly optimized as it was evident by a 10% increase in accuracy with an error rate reduced by 42%. The execution time was also reduced by 85% and 9% increment in both Macro and Weighted Average accuracies after the implementation of the optimization techniques. The validation was further done by comparing the model's performance to other proposed models for students' performance prediction tasks and the outcome has been shown in Table 7 below

Table 7: Model's performance comparison with other related models

Reference	Sample	ML Algorithm	Evaluation Metrics		
			Accuracy(%)	RMSE	F-score
Wei (2020)	365	RF	54.0	n/a	51.0
Adjei-Pokuaa & Adekoya (2022)	1520	RF	89.4	0.392	n/a
Ajay et al. (2020)	480	RF	81.3	n/a	n/a

Batool et al. (2021)	649	RF	75.4	n/a	81-95
Ragab et al. (2021)	480	RF	90.4	n/a	n/a
Abu Saa et al. (2019)	56544	RF	71.8	n/a	49.3
Gusnina et al. (2022)	300	RF	84.3	n/a	n/a
Khan et al. (2022)	623	RF	93.7	n/a	92.2
Jawad et al. (2022)	32593	RF	76.3-84.2	n/a	81-89
Proposed model	**598**	**RF**	**92.0-96.0**	**0.2000**	**96**

Table 6 above reveals that the proposed model outperformed all the models given accuracy, Root Mean Square Error (RMSE), and F1-score as metrics for evaluation and comparison. It was identified that the model attained an accuracy ranging from 92% to 96% with the lowest error rate of 0.20 which was seen to be higher than all the models used for the comparison. When the F-score, thus the harmonic mean of both precision and recall was used, the proposed model recorded the highest value of 96%. The proposed model effectively executed the prediction task as well as revealed the influence of non-academic data on students' performance and the relationship that exists among them which was set as the principal focus of the study.

5. Conclusion

The results revealed that the proposed model's performance was significantly improved compared to the baseline performance as well as the outcomes of other models proposed by other researchers for students' performance prediction and analysis tasks.

Input features found to be key/dominant, relevant, and responsible for the prediction of students' academic performance included; school environment conduciveness, teacher attributes (competence), father and mother's formal education level, family income status as well as nature, student's community of residence, student's motivation, scholarship, agreeableness, student's emotions management capability, student's self-awareness, openness, conscientiousness, neuroticism, extraversion, social skills as emotion quotient and personality trait features, gender, birth order, and career interest of student [19][22][1][14]. Hence, the proposed model has significantly contributed to existing knowledge of educational data mining (EDM) and the relevance and need for non-academic data (psycho-socioeconomic) to be factored in when making decisions about the improvement of students' performance. Also, as a decision support model, it can be integrated into operational systems in the educational setting to facilitate decision-making and implementation with given data on students and education in general.

Acknowledgment

Much appreciation and gratitude go to Eric Opoku Osei, Ph.D. For the guidance and support offered to me, Mr. Amiru Bakariwie, Mr. Emmanuel Baah, and all staff of Islamic Senior High School, Wa for the support rendered to me.

Author's Biography

References

1. Sokkhey, P., & Okazaki, T. (2020). Study on dominant factor for academic performance prediction using feature selection methods. *International Journal of Advanced Computer Science and Applications*, *11*(8), 492–502. https://doi.org/10.14569/IJA CSA.202 0.0110862

2. Roslan, M. H. Bin, & Chen, C. J. (2022). Predicting students' performance in English and Mathematics using data mining techniques. *Education and Information Technologies, 0123456789*. https://doi.org/10.1007/s10639-022-11259-2
3. Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H., & Wang, C. (2018). Machine Learning and Deep Learning Methods for Cybersecurity. *IEEE Access*, *6*(c), 35365–35381. https://doi.org/10.1109/ACCESS.2018.2836950
4. Ghosh, S. K., & Janan, F. (2021). Prediction of student's performance using random forest classifier. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 7089–7100.
5. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, 175–199. https://doi.org/10.1145 /3293881.3295783
6. Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences (Switzerland)*, *11*(1), 1–28. https://doi.org/10.3390/app11010237
7. Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z
8. Mantey, G. K., Edusei-tawiah, F. K., Yeboah, D. O., & Omari-sasu, A. Y. (2022). *Modeling the Determinants of Students ' Performance in Mathematics*. *18*(1), 103–115.
9. Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. *Education and Information Technologies*, *24*(6), 3577–3589. https://doi.org/10 .1007/s10639-019-09946-8
10. Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student academic performance prediction using supervised learning techniques. *International Journal of Emerging Technologies in Learning*, *14*(14), 92–104. https://doi.org/10.3991/ijet.v14i14.10310
11. Khan, M. I., Khan, Z. A., Imran, A., Khan, A. H., & Ahmed, S. (2022). Student Performance Prediction in Secondary School Education Using Machine Learning. *8th International Conference on Information Technology Trends: Industry 4.0: Technology Trends and Solutions, ITT 2022, September*, 94–101. https://doi.org/10.1109/ITT5612 3.2022.9863971
12. Beaulac, C., & Rosenthal, J. S. (2019). Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education*, *60*(7), 1048–1064. https://doi.org/10.1007/s11162-019-09546-y
13. Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques. In *Technology, Knowledge and Learning* (Vol. 24, Issue 4). Springer Netherlands. https://doi.org/10.1007/s10758-019-09408-7
14. Adjei-Pokuaa, H., & F. Adekoya, A. (2022). Predictive Analytics of Academic Performance of Senior High School (Shs) Students: a Case Study of Sunyani Shs. *International Journal of Engineering Technologies and Management Research*, *9*(2), 64–81. https://doi.org/ 10.29121/ijetmr.v9.i2.2022.1088
15. Abubakar, Y., Bahiah, N., & Ahmad, H. (2017). Prediction of Studentsâ€™ Performance in E-

Learning Environment Using Random Forest. *International Journal of Innovative Computing*, *7*(2), 1–5. https://ijic.utm.my/index.php/ijic/article/view/143

16. Mandefro Messele, A., & Addisu, M. (2020). A Model to Determine Factors Affecting Students Academic Performance: The Case of Amhara Region Agency of Competency, Ethiopia. *International Research Journal of Science and Technology*, *1*, 75–87. https://doi.org/10.46378/irjst.2020.010202
17. Lokpo, C. (2020). Mining Educational Data to Analyze Students' Performance: A Case Study of Mawuli School, Ho. *International Journal of Innovative Science and Research Technology*, *5*(5), 1920–1948. https://doi.org/10.38124/ijisrt20may635
18. Jawad, K., Shah, M. A., & Tahir, M. (2022). Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing. *Sustainability (Switzerland)*, *14*(22). https://doi.org/10.3390/su142214795
19. Batool, S., Rashid, J., Nisar, M. W., Kim, J., Mahmood, T., & Hussain, A. (2021). A Random Forest Students' Performance Prediction (RFSPP) Model Based on Students' Demographic Features. *Proceedings of the 2021 Mohammad Ali Jinnah University International Conference on Computing, MAJICC 2021*, *September*. https://doi.org/10.1109/MAJICC53071.2021.9526239
20. Ajay, P., Pranati, M., Ajay, M., Reena, P., Balakrishna, T., & Scholar, U. G. (2020). Prediction of Student Performance Using Random Forest Classification Technique. *International Research Journal of Engineering and Technology*, 405–408.
21. Gusnina, M., Wiharto, & Salamah, U. (2022). Student Performance Prediction in Sebelas Maret University Based on the Random Forest Algorithm. *Ingenierie Des Systemes d'Information*, *27*(3), 495–501. https://doi.org/10.18280/isi.270317
22. Ünal, F. (2019). We are IntechOpen, the world's leading publisher of Open Access books Built by scientists , for scientists' TOP 1 %. *IntechOpen*, *3*, 11.
23. Ghosh, S. K., & Janan, F. (2021). Prediction of student's performance using random forest classifier. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 7089–7100.
24. Abubakar, Y., Bahiah, N., & Ahmad, H. (2017). Prediction of Studentsâ€™ Performance in E-Learning Environment Using Random Forest. *International Journal of Innovative Computing*, *7*(2), 1–5. https://ijic.utm.my/index.php/ijic/article/view/143
25. Pelaez, K., Levine, R., Fan, J., Guarcello, M., & Laumakis, M. (2019). Using a latent class forest to identify at-risk students in higher education. *Journal of Educational Data Mining*, *11*(1), 18–46.
26. Kantardzic, M. (2020). *Data Mining: Concepts, Models, Methods, and Algorithms* (E. Hossain, D. A. Grier, D. Heirman, E. B. Joffe, & X. Li (eds.); 3rd ed.). John Wiley & Sons, Inc., http://www.wiley.com/go/permissions.
27. Brooks, C., & Thompson, C. (2017). Predictive Modelling in Teaching and Learning. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of Learning Analytics* (First, pp. 61–68). https://doi.org/10.18608/hla17.005
28. Gwak, Y., Jeong, C., Roh, J. H., Cho, S., & Kim, W. (2020). Multi-models of Educational Data Mining for Predicting Student Performance in Mathematics: A Case Study on High Schools in Cambodia. *IEIE Transactions on Smart Processing and Computing*, *9*(3), 203–211. https://doi.org/10.5573/IEIESPC.2020.9.3.217