Managing Student Performance: A Predictive Analytics using Imbalanced Data

Usman Ashfaq, Booma P. M., Raheem Mafas

Abstract*: *Big data has revolutionized every field of life, which accumulates human learning as well. The field of education has progressed in past couple of decades, and addition to that, rapid growth in the number of educational institutions has created a tough competition. The massive accumulation of data in the educational sector has created a great scope of EDM (Educational Data Mining) with the support of robust predictive models. It is quite necessary to regularly examine the performance of the students to make them perform better, thus helps to maintain the reputation of the institution. This study proposed a predictive model through which the performance of the student can be forecasted depending upon various characteristics. The KDD(Knowledge Discovery in Databases) methodology was followed stepwise in this study for developing predictive models to predict student performance. The data balancing techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) were employed to handle the unbalanced effect of data which causes bias predictions. Also, for the selection of significant features techniques, FCBF (Fast Correlation Based Feature selection) and RFE (Recursive Feature Elimination) were used. The EDM algorithms Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Network (ANN) were utilized for predicting student performance with suitable hyper-parameter tuning using random search to enhance the performance of the model. The results obtained were cross-validated using Ensemble Method and benchmarked with previous studies. The random forest model achieved the highest accuracy of 86% after data balancing and careful selection of significant features.

Keywords*: *Predictive Algorithm, Data Balancing, Educational Data Mining, Feature Selection, Student Performance.

I. INTRODUCTION

These days prediction of student performance is contemplated as a standout amongst the education sector challenges because of deficiency of effective models. This issue is especially existing as of now the research of inadequate measure inclusion in the assessment of student performance is lacking and incompatibility of the present models to the institutions' frameworks.

According to the study [1], the students' performance is linked to the drop-out tendency of the students as per the education system. Severely, in a previous couple of decades, there has been a major downfall in the performance of the students irrespective of the universities' desideratum. The presence of this issue has an enormous fortification in light of the more unfortunate profession prospects.

The prior mentioned problem can be effectively watched even in the nations with the solid monetary improvement which elevates a worry around the globe's authorities of education. Higher Education Statistics Agency (HESA) stated that since 2013 the increment in drop-out for students has escalated yearly [2]. United States' condition is much severe as students who finished a six-year degree had a percentage of 56% only, while that of a two-year degree was 29% of the enrolled students [3].

For adapting up to this problem in the greatest nations of Europe, after the Bologna Process, a mentoring framework has been actualized [4]. Amongst the fundamental targets of this procedure is to provide counseling to students for scholarly advancements by assisting them in academic path selection. Albeit, the availability of similar frameworks in the greater part of American and European institutions, despite that, to identify students who require them the most is still tough for academics and management. This is the result of the express size of the accessible understudy information and of its quickly regularly expanding nature, as well. Subsequently, those who require individual counseling would be ignored because of no efficient identification method.

EDM known as Educational Data mining is the transformation of primary data, from the educational environment to instructive information for getting in-depth knowledge of the insights and assist students and educational institutions for effective path selection [5]. According to researchers [6], EDM can be considered as a convergence of three primary zones, education, statistics, and computer science. The other areasnamely, Learning Analytics (LA), Data Mining (DM) and Machine Learning (ML) and computer-based education, formed due to convergence of these primary zones are also linked to EDM as shown in Fig. 1. ML is categorized as supervised ML and unsupervised ML where the supervised ML requires a training set upon which the model is trained for precise predictions, whereas for unsupervised there is no requirement for training set [7]. Helping the struggling university students on their performance was investigated by the application of EDM on their data, for alerting them about their performance and assisting them to accomplish more [8].

Revised Manuscript Received on March 09, 2020.

Usman Ashfaq, Department of Computing, Engineering & Technology, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia.

Dr. Booma P. M., Department of Computing, Engineering & Technology, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia.

Raheem Mafas, Department of Computing, Engineering & Technology, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia.

EDM has demonstrated to be valuable for institutions to develop and plan their program better, academics for focusing up to individual needs of students and mainly for students to assist them in better learning. In addition by the EDM execution in online learning, concealed flows and prime acuity can be withdrawn [9].

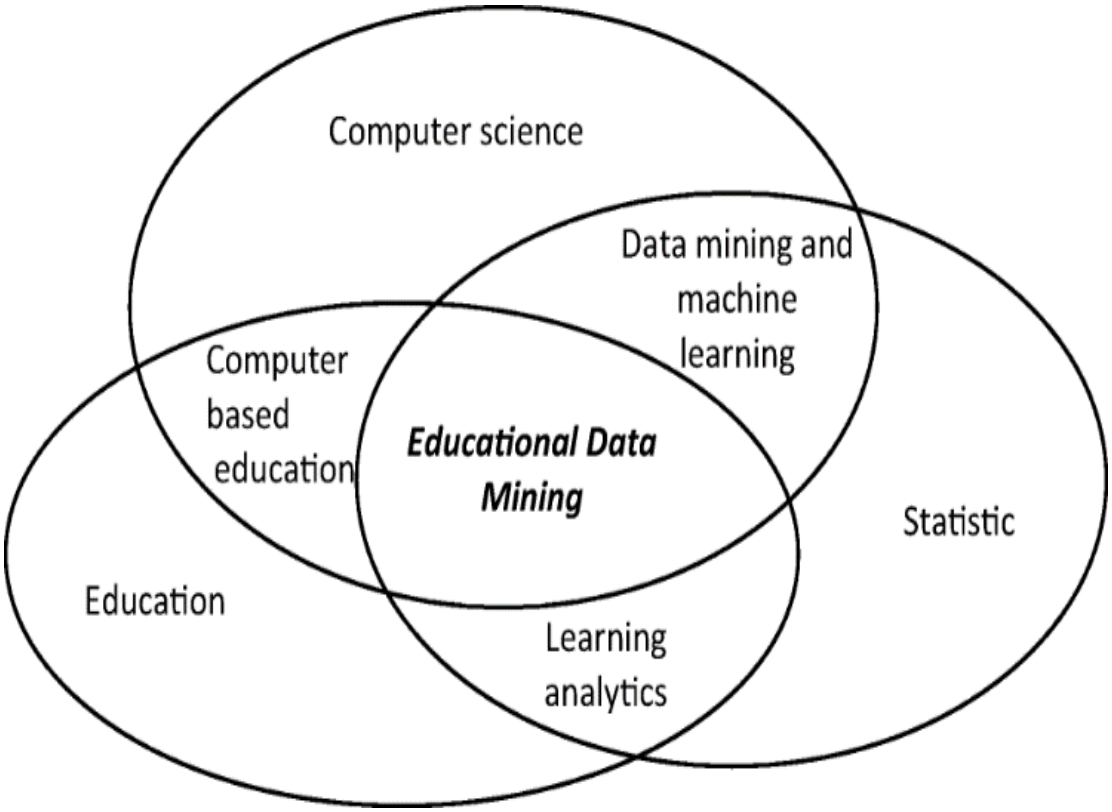

Fig.1. EDM Structure[6]

In the field of EDM, the prediction of student's performance has a ton of consideration among the researchers. Conventional techniques of EDM has been deployed to manage various errands identified with the student's characteristics [10]. The poor performance of students in the educational organization is a significant social issue, which leads to student drop-out, and it has turned out to be significant for educational institutions and academics to more likely comprehend the reasons for a large number of failures in the student community. This isn't a simple case, as lot of factors or attributes are to be assessed that affect students' performance [11]. ML empowers the forecast of student's conduct, expertise, and performance by investigating different student's ventures, as individuals or in groups.

According to the student's pursuit, their performance can be anticipated utilizing the techniques of DM, which could help teachers to focus more on the struggling students [12]. Studies have been done for predicting student performance based onthe number of features of the studentlike previous grades, financials, behavior, etc. for the precise prediction of student performance [13].Prediction of the student performance is highly dependent on the performance of his/her past or present academic records, as the subsequent performance of the student is more likely to be similar to his/her past records[14], [15].

In addition to academic features, the other which are quite feasible for student's performance prediction are student behavior characteristics and their demographic characteristics comprising age, financial status, gender, marital status, etc. These characteristics have a huge role in student performance, so these can't be foreseen during their performance prediction. Studies [16], [17] proposed a framework for struggling students to alert

them early about their performance based on these additional characteristics.Most of the past researches have only focused on student's academic-related attributes. In addition to academic factors, the behavioral characteristics related to students, like their attendance, class participation, reviewing course materials, etc., are also needed to be included in the analysis as the demographical factors also have a huge impact.

On the other hand, the prior studies have looked upon diverse student features, while overlooking the imbalance effect of student data. The imbalanced data can lead to biased predictions [18], as this has a huge impact on prediction model building. This issue has become more vibrant to be tackled to prevent significant losses for the institution and students, as well. Therefore, this forms the gap for improvement to bear out the prior mentioned issues. In this line, the following are set as the objectives of this study:

1. To cater imbalance factor by balancing techniques for better model performance.
2. To analyze features influencing student performance by the application of feature selection methods.
3. To compare the data mining algorithm's performance for effective model development.

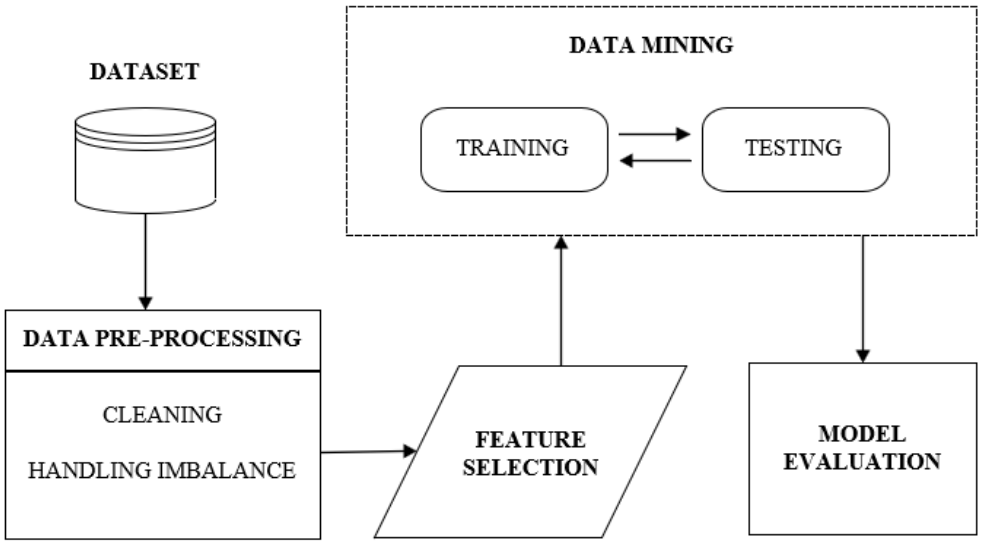

Fig.2. Methodology Outline as per KDD

II. METHODOLOGY AND DATA

The study was conducted utilizing the secondary data, while the methodology outline of this study is shown in Fig. 2. The outline of the proposed methodology for this study is structured upon KDD (Knowledge Discovery in Databases). The first step of the outline was to obtain the data from the data source. Then after that, data was processed in the second step, where the data cleaning and imbalance issue were handled, as it is vital for the machine to understand data and for the efficient running of the algorithms as well. The next step was feature selection in which features that are significant were selected through feature selection techniques, as the features which are not significant for the model should be removed to get better results and effective model running.

DATA

The dataset of students, for this study, have been taken from the online data portal [19], which was originally collected from The University of Jordan, Amman, Jordan, through Kalboard 360 which is an LMS (Learning Management System). Kalboard 360 has been intended to encourage learning using leading-edge innovation. The data was collected using a tracker tool for learner activity, xAPI (Experience API). xAPi empowers to screen learning advancement and student activities like watching a preparation video or perusing an article, as it is part of TLA (Training and Learning Architecture). It helps the learning action suppliers to decide the student, movement, and items that portray a learning knowledge. The dataset comprises of 480 student records and 16 attributes, in which 11 are of type character and 5 of numeric, related to students. The features in the dataset are of three categories, academic, behavioral and demographic.

EXPLORATARY DATA ANALYSIS (EDA)

The EDA was performed using Tableau software, for descriptive analysis. Various aspects of data were analyzed, where data was complete and had no outliers. While figure 3 shows the distribution of classes in the dependent attribute, i.e. 'Class'. 'H' is for the high performing (High Level) students for which count was 142, 'M' for medium performing (Medium Level) students for which count was 211 and 'L' for low performing (Low Level) students for which count was 127. It can be perceived that there is an issue of the imbalance of classes in the data. The model developed on such data performs bias predictions towards the majority class.

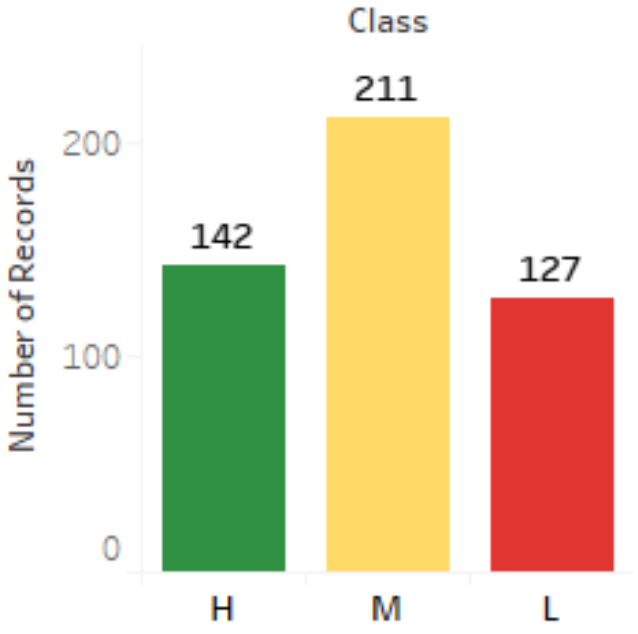

Fig.3. Class Distribution

DATA BALANCING

The techniques employed for minimizing the aspect of data imbalance were SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling).

Researchers presented an approach for oversampling, where the synthetic samples are created for a class at minority, SMOTE, instead of replacing with the samples generated form oversampling[20]. The synthetic samples are created in SMOTE by analyzing features instead of data as a whole. The synthetic samples are generated by random choice for k nearest neighbors, as per the requirement of the scenario. Researchers presented method, ADASYN, for tackling imbalanced data[21]. ADASYN utilizes the distribution of weights for various samples of the minority class, in accordance with their learning difficulty level. So, the generation of synthetic samples is upon the level in the minority class. The ADASYN helps to enhance the distribution of data by minimizing the bias aspect in data due to imbalanced data and making classification easy by shifting sample weights. The generated samples are made realistic by the addition of some random value in ADASYN. As compared to SMOTE, ADASYN only alters the minority class count, while the majority class remains intact.

FEATURE SELECTION

One of the main sections in EDM modeling is feature selection. The significance of the features is determined by feature importance and minimizing the effect of variance among the features. The two methods used for selecting features in student performance prediction were FCBF and RFE. The FCBF (Fast Correlation Based Feature selection), it comes under Filter Methods of feature selection. In the FCBF process, the features are selected upon the significance level of features by the correlation of features and the selected significant features are returned in their particular position format in the form of an array. The FCBF begins with all features and selects the significant features by the variance between the features, depending upon the prediction capability [22]. The RFE (Recursive Feature Elimination), fundamentally it is a process of looping. It comes under Wrapper Methods of feature selection. According to the significance of the features, they are ranked in a particular order in this process. While, at each iteration of the loop structure, the features which are less significant are eradicated. The application of looping structure is due to the variation in the significance of features, at each iteration, by the removal of less significant features [23].

PREDICTIVE ANALYTICS

The prediction of student performance was done using EDM (Educational Data Mining) algorithms. The algorithms to be used in this study for classification are SVM (Support Vector Machine), RF (Random Forest) and ANN (Artificial Neural Network). The concept of SVM is based upon decision planes. The decision boundaries are defined by decision planes in the prediction process. The objects from different classes are separated in decision planes, by which classification is done between different classes. So, hyper-planes are constructed in multidimensional space for differentiating between the classes and multiple categorical and continuous variables can be handled in SVM [24]. Among the combination of multiple DT (Decision Tree) in RF, the best model for predicting classes precisely in selected as the final model. The data is fed into each tree in RF and for classifying between the classes, where vote from each class is obtained. While the selection of interested class depends upon the highest count of the vote for that class [25]. ANN is a combination of neurons and hidden layers. It consists of input and output layers. It is employed for deep learning in any scenario. A neuron is the product of input with their respective weights. The activation function is utilized to scale the output for each neuron. The hidden layers count depends upon that how much model needed to be complex, more the number of hidden layers, more the model is complex, where there is an increase in model efficiency too [26].

OPTIMIZATION AND CROSS VALIDATION

After the creation of models for prediction of student performance, enhancing the model's prediction performance optimization of hyper-parameters was utilized. Hyper-parameters are those characteristics of the model which can't be determined from data, as compared to the parameter which can be calculated thorough data. The main goal of using a random search was to build a model using different values of the hyper-parameters and evaluate all the models build individually [27]. After the development process, predictive models were validated using the 'K-Fold Cross Validation' approach in the Ensemble Method process. The process of cross-validation comprises testing predictive models on a chunk of data in a repeated manner [28], [29].

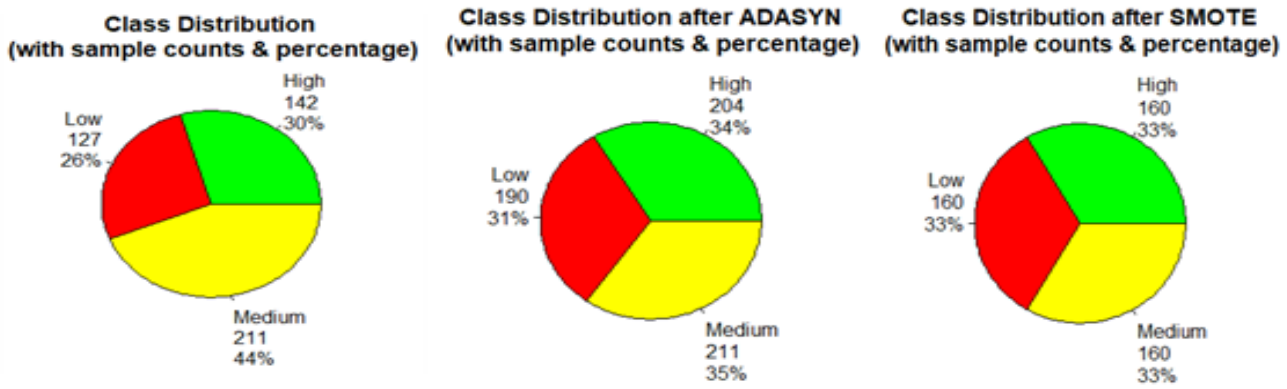

Fig.4. Data Balancing by SMOTE and ADASYN

III. RESULTS

Data under consideration was complete and had no outliers, so just data transformation was performed from qualitative to quantitative. Balancing techniques, ADASYN and SMOTE were employed, for which results can be observed in figure 4. The total frequency for SMOTE was the same as that of original data, while the frequency for each of the class was altered. The synthetic samples were created for high and low classes, while the samples for majority class, i.e. medium, were removed. For ADASYN, the total frequency was increased, as the frequency count for majority class, i.e. medium, wasn't altered. Whereas, synthetic samples were generated for other, high and low, classes. The resulted frequencies of classes for SMOTE data were equal, while that in case of ADASYN weren't equal but close to the majority class.

The feature selection techniques, FCBF and RFE, were implemented. The total independent variables present in ADASYN and SMOTE data, which were '16', excluding the target variable. Upon the implementation of RFE upon the balanced datasets, ADASYN and SMOTE, three variables, i.e. 'StageID', 'SectionID' and 'Semester' which were non-significant, were removed. While the remaining selected variables used. Where, in the case of FCBF deployment upon balanced datasets only one variable, i.e. 'Semester' was removed.

The predictive models developed in this study were random forest (RF), support vector machine (SVM) and artificial neural network (ANN). The data was split into '70%' for training and remaining '30%' for testing. The models after development were optimized by hyper-parameter tuning, which was performed using a random search. Table-I shows the results obtained for predictive models developed for original, balanced and feature selected data. The accuracies for the models were increased after hyper-parameter optimization. The parameters tuned were 'mtry' for RF, 'sigma' and 'tau' for SVM and, 'size' and 'decay' for ANN.

Table-I: Predictive Model Results

DATA	ACCURACY	OPTIMIZED ACCURACY
RANDOM FOREST (RF)		
Original	0.7483	0.8398
ADASYN & RFE	0.7845	**0.8674**
SMOTE & RFE	0.8056	0.8403
ADASYN & FCBF	0.8232	0.8453
SMOTE & FCBF	**0.8333**	0.8542
SUPPORT VECTOR MACHINE (SVM)		
Original	0.7014	0.7153
ADASYN & RFE	0.7624	0.7790
SMOTE & RFE	0.7917	**0.7986**
ADASYN & FCBF	0.7680	0.7956
SMOTE & FCBF	**0.7917**	0.7986
ARTIFICIAL NEURAL NETWORK (ANN)		
Original	0.6084	**0.7902**
ADASYN & RFE	0.6304	0.7403
SMOTE & RFE	0.5826	0.7639
ADASYN & FCBF	**0.6467**	0.7514
SMOTE & FCBF	0.6190	0.7639

The cross-validation of the models developed was performed using a 10 fold cross-validation process under the ensemble method, results listed in table-II. The parameters for developing the model were chosen by various values upon the repetition of model development upon them and the best model was decided upon the accuracy of the model. With the models developed in this study, RF, SVM, and ANN, three other models were also incorporated in this ensemble method which includes decision tree (DT), K-Nearest Neighbors (K-NN) and Naïve Bayes (NB).

Table-II: Cross Validation Results

DATA	MODEL	ACCURACY
Best performing models through Ensemble Method		
Original	Decision Tree	0.7483
ADASYN & RFE	Random Forest	0.8097
SMOTE & RFE	Decision Tree	0.8042
ADASYN & FCBF	Random Forest	**0.8260**
SMOTE & FCBF	Random Forest	0.7902

The best performing model in the case of models developed and optimized in this study is 'Random Forest' for all the forms of data under consideration, as shown in table-I. While, in the case of models developed through ensemble method, 'Decision Tree' model is performing better of original data and, SMOTE and RFE data. Whereas, for remaining forms of data, the 'Random Forest' model is performing better. The accuracies of the models developed upon on original and processed datasets are more as compared to that for the ensemble method, which validates the process of model development through optimal parameters done in this study.

IV. DISCUSSION

This research aimed to study and examine the factors affecting student performance, which could be used for better model development, to predict student performance. For this, a comprehensive amount of past studies were analyzed, which were related to the problem under consideration. The factors used by the researchers for effecting the student performance and the EDM techniques employed upon them were compared to choose the technique to be implemented in this study. The first objective of the study was:

- To cater imbalance factor by balancing techniques for better model performance.

The data balancing techniques, ADASYN and SMOTE were employed to deal with the imbalance issue of classes in the target variable.

The data resulted after the application of SMOTE, altered all the classes of data and equated the frequency for all, while the overall frequency was the same as before. In contrast to SMOTE, the data resulted by ADASYN altered the classes except for the majority class and by doing that the overall frequency was also increased. The data resulted from ADASYN didn't have an equal frequency as that of SMOTE data, but the frequency count for all the classes was quite close. Before proceeding to the modeling stage of the KDD process, after the pre-processing step, the next objective of the study accomplishedwas:

- To analyze features influencing student performance by the application of feature selection methods.

The feature selection techniques, FCBF and RFE, were deployed in this study, to identify the significant features influencing the student performance. Both techniques, FCBF and RFE, were implemented on balanced data, ADASYN and SMOTE, by resulting out four datasets in total. The application of FCBF upon ADASYN and SMOTE data, removed only one non-significant feature from the data, while in the case of RFE, three non-significant features were eradicated from the data. The remaining features were the significant features affecting the performance of the student and were utilized in model development. The next phase, as per KDD, was modeling, for which the objective of this study attained was:

- To compare the data mining algorithm's performance for effective model development.

The EDM algorithms, random forest (RF), support vector machine (SVM) and artificial neural network (ANN), was developed in this study. These algorithms were not only developed upon balanced and feature selected data, but also original data, with no techniques applied on it, for comparing the model results. In addition to that, the performance of the models developed was enhanced by optimizing the models, through the random search by selecting the optimal hyper-parameters for the models. The performance of the predictive algorithms was evaluated using confusion matrix statistics and was also cross-validated using the ensemble method. The algorithms developed in the modeling stage were developed through the ensemble method and in addition to those already developed algorithms, three additional algorithms, Decision Tree (DT), K-Nearest Neighbors (K-NN) and Naïve Bayes (NB), were developed to validate model performances. Finally, the best performing model developed in this study was benchmarked with the early studies, among which this study model performed the best, shown in table-III. The 'Random Forest' model upon ADASYN and RFE data, developed in this study, resulted in the highest accuracy, i.e. '86.74%'.

Table-III: Model Benchmarking

STUDIES	MODELS DEVELOPED	HIGHEST ACCURACY
(Aljarah, Amrieh and Hamtini, 2015) [30]	DT, ANN, NB	ANN 73.8%
(Aljarah, Amrieh and Hamtini, 2016) [31]	DT, ANN, NB with Ensemble Method	ANN 80%
(Deepika and Sathyanarayana, 2018) [32]	DT, ANN, NB with Ensemble Method	DT 82.2%
(Bhutto, Siddiqui and Ali Arain, 2019) [33]	DT, ANN, NB	ANN 78.1%
(Li *et al.*, 2019) [34]	DT, ANN, NB, SVM	SVM 73.91%
THIS STUDY	RF, SVM, ANN with SMOTE/ADASYN and FCBF/RFE	RF 86.74% (ADASYN and RFE)

V. CONCLUSION AND FUTURE WORKS

The research comprises the implementation of various techniques and algorithms related to data mining. At first, the data balancing techniques, SMOTE and ADASYN, were applied upon unbalanced data. The results for the SMOTE technique showed that the frequency for all the classes in the target variable was altered, and due to this alteration the majority class samples were reduced to equate all the classes. This process reduced the variation in data as the dataset size originally was also small. Whereas for ADASYN, there was no reduction of samples for any class, and the resulted dataset size was increased. Continuing to the next step which was feature selection techniques, FCBF and RFE, deployment on balanced datasets, SMOTE and ADASYN. The results of FCBF on ADASYN and SMOTE were the same and were also the same in case of RFE. This was due to the less dimensionality and size of data. So, for tackling these situations more factors should be incorporated and sample size also needed to be enlarged, to get better results for the data mining techniques.

For future studies, the research can be taken up to different educational levels of student belonging and not only for physical institutions can also be useful for online learning platforms. For increasing the dimensionality and enhance real-world applications for the predictive models, factors like students' schedules and hobbies, can be incorporated, as these factors and many others also have an impact on student performance. The deployment of these systems in educational institutions will not only help to forecast student performance but will also be quite useful in terms of student drop-out, as student performance is directly linked to it.

REFERENCES

1. Glennie, E., Bonneau, K., Vandellen, M. and Dodge, K.A., 2012. Addition by Subtraction: The Relation between Dropout Rates and School-Level Academic Achievement. *Teachers College record (1970)*, 114(8), pp.1–26.
2. Baker, S., 2017. *Dropout rate for young UK students rises again*. [online] Times Higher Education (THE). Available at: <https://www.timeshighereducation.com/news/dropout-rate-young-uk-students-rises-again>
3. Link, C., 2015. *Pathways to Prosperity: Meeting the Challenge of Preparing Young Americans for the 21st Century*. [online] Massachusetts. Available at: <https://careertech.org/resource/pathways-prosperity>
4. Altbach, P.G. and Forest, J.J.F., 2007. *International Handbook of Higher Education*. Dordrecht: Springer Netherlands.
5. Aghabozrgi, S., Mahroeian, H., Dutt, A. and Ismail, M.A., 2015. Clustering Algorithms Applied in Educational Data Mining. *International Journal of Information and Electronics Engineering*, 5(2).
6. Mimis, M., El Hajji, M., Es-saady, Y., OueldGuejdi, A., Douzi, H. and Mammass, D., 2019. A framework for smart academic guidance using educational data mining. *Education and Information Technologies*, 24(2), pp.1379–1393.
7. Kamber, M., Pei, J. and Han, J., 2012. The Morgan Kaufmann Series in Data Management Systems. In: *Data mining: concepts and techniques*, 3rd ed. Elsevier Science, p.744.
8. Ali, S.A., Asif, R., Haider, N.G. and Merceron, A., 2017. Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113, pp.177–194.
9. Mikropoulos, T.A., Pintelas, P. and Livieris, I.E., 2016. A Decision Support System for Predicting Students' Performance. *Themes in Science and Technology Education*, 9(1), pp.43–57.
10. Daud, A., Aljohani, N.R., Abbasi, R.A., Lytras, M.D., Abbas, F. and Alowibdi, J.S., 2017. Predicting Student Performance using Advanced Learning Analytics. In: *26th International Conference on World Wide Web Companion*. Perth, Australia: International World Wide Web Conferences Steering Committee, pp.415–421.
11. Shaleena, K..and Paul, S., 2015. Data Mining Techniques for Predicting Student Performance. In: *2015 IEEE International Conference on Engineering and Technology (ICETECH)*. Coimbatore, TN, India: IEEE, pp.1–3.
12. Sin, K. and Muthu, L., 2015. Application of big data in education data mining and learning analytics – a literature review. *ICTACT journal on soft computing*, 5(4), pp.1035–1049.
13. Richardson, M., Abraham, C. and Bond, R., 2012. Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, 138(2), pp.353–387.
14. Shimada, A., Ogata, H., Okubo, F. and Yamashita, T., 2017. A neural network approach for students' performance prediction. *LAK*, pp.598–599.
15. Zhou, Z. and Ma, X., 2018. Student pass rates prediction using optimized support vector machine and decision tree. In: *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. Las Vegas, NV, USA: IEEE, pp.209–215.
16. Hayes, D., Bernacki, M., Hong, W., Markle, J. and Voorhees, N., 2017. Using LMS Data to Provide Early Alerts to Struggling Students. In: *First Year Engineering Experience (FYEE) Conference*. Daytona Beach, FL: American Society for Engineering Education.
17. Hiremath, P.G.S., Banavasi, M.N., Athani, S.S. and Kodli, S.A., 2017. Student performance predictor using multiclass support vector classification algorithm. In: *2017 International Conference on Signal Processing and Communication (ICSPC)*. Coimbatore, India: IEEE, pp.341–346.
18. Kuncheva, L.I., Arnaiz-González, Á.,Díez-Pastor, J.F. and Gunn, I.A.D., 2019. Instance selection improves geometric mean accuracy: a study on imbalanced data classification. *Progress in Artificial Intelligence*, 8(2), pp.215–228.
19. KAGGLE, 2016. *Students' Academic Performance Dataset | Kaggle*. [online] KAGGLE. Available at: <https://www.kaggle.com/aljarah/xAPI-Edu-Data>
20. Bowyer, K.W., Chawla, N. V., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, pp.321–357.
21. Yang, B., Garcia, E.A., Haibo, H. and Shutao, L., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. Hong Kong, China: IEEE, pp.1322–1328.
22. Liu, H. and Yu, L., 2003. Feature selection: a fast correlation-based filter solution. In: *Twentieth International Conference on Machine Learning (ICML-2003)*. Washington DC, USA, pp.856–863.
23. Biasioli, F., Furlanello, C., Gasperi, F. and Granitto, P.M., 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2), pp.83–90.
24. Chau, A.L., Li, X. and Yu, W., 2014. Support vector machine classification for large datasets using decision tree and Fisher linear discriminant. *Future Generation Computer Systems*, 36, pp.57–65.
25. Pal, M., 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1), pp.217–222.
26. Yegnanarayana, B., 2009. *Artificial neural networks*. New Delhi: PHI Learning Pvt. Ltd.
27. Bergstra, J.S., Bardenet, R., Bengio, Y. and Kégl, B., 2011. Algorithms for Hyper-Parameter Optimization. In: *Advances in Neural Information Processing Systems 24 (NIPS 2011)*. Neural Information Processing Systems Foundation, Inc., pp.2546–2554.
28. Krstajic, D., Buturovic, L.J., Leahy, D.E. and Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6(1), p.10.
29. Thiruchelvam V. and Wahid S., 2018. Development of a Advanced Computerised Biometric Attendance Logging System for Institutions of Higher Learning, *International Journal of Electrical & Computer Sciences IJECS / IJEN*, ISSN: 2077-1231 (Online) 2227-2739 (Print), 17(6), pp.8-15.
30. Aljarah, I., Amrieh, E.A. and Hamtini, T., 2015. Preprocessing and analyzing educational data set using X-API for improving student's performance. In: *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*.
31. Aljarah, I., Amrieh, E.A. and Hamtini, T., 2016. Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, 9(8), pp.119–136.
32. Deepika, K. and Sathyanarayana, N., 2018. Comparison Of Student Academic Performance On Different Educational Datasets Using Different Data Mining Techniques. *International Journal of Computational Engineering Research (IJCER)*, 8(9), pp.28–38.
33. Bhutto, S., Siddiqui, I.F. and Ali Arain, Q., 2019. Analyzing Students' Academic Performance through Educational Data Mining. *3C Tecnología_Glosas de innovaciónaplicadas a la pyme, Special Issue*, pp.402–421.
34. Li, F., Zhang, Y., Chen, M. and Gao, K., 2019. Which Factors Have the Greatest Impact on Student's Performance.*Journal of Physics: Conference Series*, 1288(1).

AUTHORS PROFILE

Usman Ashfaq, Post Graduate Student, School of Computing, Engineering & Technology, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia.
Email: usman.ash93@gmail.com

Dr. Booma P. M., Lecturer, Faculty of Computing, Engineering & Technology, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia.
Email: dr.booma@staffemail.apu.edu.my

Raheem Mafas, Lecturer, Faculty of Computing, Engineering & Technology, Asia Pacific University of Technology & Innovation (APU), Kuala Lumpur, Malaysia.
Email: raheem@staffemail.apu.edu.my