European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 08, 2020

An in Depth Analysis of Machine Learning Classifiers for Prediction of Student's Performance

Thingbaijam Lenin[1,*], N. Chandrasekaran[2]

[1]Research Scholar,
[1] Martin Luther Christian University (MLCU), Meghalaya, India
[2]ex IBM and Director, CDAC-India, Visiting Prof., MLCU

[1] lenin.th@gmail.com

Abstract

Machine learning algorithms are sensitive to the nature and the dimension of the data that are fed into the model for analysis. These algorithms tend to perform significantly different depending upon the dataset used for analysis and training. It then becomes difficult to discover the best algorithm to handle a particular dataset. In the current work, we have made an attempt to verify 24 different state of the art supervised machine learning algorithms in an effort to find the most suitable classifier for predicting the performance of students in our University. Out of the 24 algorithms that we have identified, we found Naïve Bayes (NB) and Stabilized Nearest Neighbor Classifier (SNN) to be the most suitable for deployment followed by K-Nearest Neighbors (KNN) and Cost Sensitive C5.0 (C5.0Cost). We have also determined that handling missing values using KNN improves the classification of minority classes. The classifiers have been evaluated with the sensitivity, specificity, precision, kappa and F-score metrics. It has further been established that the performance metric "Accuracy" is misleading when dealing with imbalanced dataset and balanced accuracy provides far better and reliable information for the model being developed.

Keywords: Educational Data Mining, Machine Learning, Classification, Data Imputation, R Programming, Stabilized Nearest Neighbor, Naïve Bayes.

1. Introduction

Recent times have witnessed enormous strides being made in the applications of machine learning techniques in various fields like health care, retail, education, etc. This trend has largely been driven by the tremendous acceleration of data generation, as well as the efforts being made to advance the capabilities of machine learning techniques. The launch in the year 2008 of an annual "International Conference and a Journal" related to educational data mining played an important role in enhancing research in the area of education (*International Educational Data Mining Society*, 2008)

The main objective of this study is to develop a model to identify students, who are at risk based on their demographic, prior academic performance and their intelligence as defined by Gardner's "Multiple Intelligence"(Gardner, 2011). This lends an opportunity to the academic administrators to identify students who are likely to drop out of the program due to poor academic performance. We consider that this study will provide a solution to both the students and the University to tackle issues relating to poor performance and drop out of students.

For the current research, we have used popular machine learning algorithms like Random Forest (RF), K-Nearest Neighbors (KNN), Naïve Bayes (NB), J48, Adaboost and other classification algorithms comprising a total of 24 algorithms and these have been implemented using R programming language. We have made use of the extensive capabilities of the RStudio integrated development environment for the development of the new model.

We categorized the students under "GD" and "FR", which refers to "Good" and "Fair" respectively. Students predicted to fall under the "FR" category are considered to be the ones who are at risk and are likely to fail or drop out of the University while students predicted "GD" are expected to perform well without any additional assistance from the faculty.

The rest of the paper is organized as follows. In Section II, we have provided a brief summary of various works related to the educational data mining and in other domains found in literature. Section III describes the methodology used for model development and this is followed by Section IV, in which the results and discussions are presented in detail.

2. Literature Review

The literature abounds with several studies that were conducted to predict students' performance by using different techniques of machine learning. We have explored some of the works that are directly relevant to our study.

Zulfiker *et al.* (2020) used data obtained from a private university in Bangladesh to demonstrate that 7 different machine learning classifiers, viz., SVM, KNN, Logistic Regression, DT, adapative boosting algorithm, AdaBoost, Extra Tree Classifier and Multilayer Perceptron (M.L.P) classifiers, were able to predict the student performance well. The approach involved training a student dataset of 400 instances with 8 attributes. The resulted output of the "base classifiers" were aggregated by "weighted voting" approach and had resulted in producing an accuracy of 81.73%.

The work of Krishna, et al. presented classification by CART (Classification and Regression Trees), which is a decision tree algorithm. Data collected from 352 graduating students were used for training and to make predictions relating to the identification of students at risk. The dataset was extracted from the Moodle LMS log file, which consists of the activities of the student while using the LMS. The authors

achieved a very high accuracy of 99.1%. Moodle is a free and open-source learning management system (LMS) written in PHP. It is possible to customize LMS to create websites with online courses for both educators and trainers to realize certain learning objectives (Krishna *et al.*, 2020).

Urkude and K. Gupta worked on 395 student records to predict the student's performance. The dataset consists of attributes such as age, occupation of the parent, health condition, and internet access and school information. Three classifiers, such as Decision Tree, Naïve Bayes and Support Vector Machine were used by taking three different samples of sizes, 100, 200 and 300. The performance of the classifiers were evaluated by using F1 and observed that The F1 score increased with sample size and the highest score of 0.7838 was achieved by the SVM at 300 sample size (Urkude and Gupta, 2019).

The work of Fatima and Mahgoub indicate that both the Decision Tree and Bayesian Network algorithms have been effective in predicting student's academic success. They used a dataset consisting of 165 students for analysis and were able to achieve an accuracy of 96.6%. Their observations include that gender, mother's education, score at high school, previous semester score and attendance and score in SE were the most informative features for the analysis of the student performance (Fatima and Mahgoub, 2019).

Imran, et al have worked on an ensemble model of J48 and Real AdaBoost to obtain a significantly improved accuracy 95.78% using the dataset at UGI machine learning repository consisting of 1044 instances with 33 attributes. The dataset was found to exhibit class imbalance, which necessitated the deployment of a re-sampling class balancing method. Artificial Neural Network (ANN) based techniques have been used by Lau, et al [8] to evaluate and predict CGPA of students. The dataset of 1000 University students used consisted of information on socio-economic background and the score obtained in the national university entrance examination. The overall accuracy of the model was 84.8 %. They observed that the role of mother played was more significant in the student's academic performance than that of the father. The authors also observed that imbalance data of gender decreased the prediction accuracy and ANN performed poorly in classifying students based on their gender(Imran *et al.*, 2019).

The machine learning techniques of ANN, Naïve Bayes, Decision Tree, and Logistic Regression were also analyzed employing the records of 161 students of the Al-Muthanna University in an effort to predict the student performance at the study conducted by Altabrawee, et al. Their work revealed that ANN provided the best performance when pitted against the other classifiers with an accuracy of 77.04%. The five most important attributes that had the most influence were found to be the student grades obtained in computer subjects, types of accommodation, interest in studying computer core subjects, satisfaction with the educational environment and type of residency(Altabrawee, Ali and Ajmi, 2019).

It is pertinent to digress at this juncture to also refer to the work of our co-workers at MLCU in the field of data mining (Dawngliani M.S, Chandrasekaran N, 2019; Dawngliani *et al.*, 2020; Dawngliani M S, Chandrasekaran N, 2020). In their study, the authors have explored various algorithmic tools to facilitate the prediction of the chances of breast cancer survivability, to assist a better understanding of the survival factors.

3. Methodology

The proposed methodology is an adaptation to the CRISP-DM (Shearer *et al.*, 2000), which stands for "Cross Industry Standard Process for Data Mining". The methodology identifies 6 steps to conceive a particular data mining project that can have cycle iterations to suit developers. The 5 primary stages involve business understanding, understanding data, data preparation, modeling and evaluation. The framework of the methodology is shown in figure 1.

Dataset and Description: To assist in the development of this model to assess the performance of students, we have gained access to the vast data collected from our university, MLCU, Meghalaya. The problems are unique as the students enroll from remote places of the region situated across the north eastern parts of India. The data consists of information collected at the time of admission. This includes data relating to demographic as well as academic performance prior to joining the University, their intelligence as defined by Gardner's Multiple Intelligence (MI) (*Multiple Intelligences Inventory*, 2017) and the students' transcripts. We have collected 497 instances each with 20 attributes. All the attributes and the values assigned to them for developing the model are shown in table 1.

The total instance collected is 497 with 20 attributes. All the attributes and the values assigned for developing the model are shown in table 1.

Understanding the Dataset: In order to meet the research objective and to gain greater accuracy, transformation of the dataset was performed, which in turn also precedes to furnishing some valuable intuition. For instance, we have found that some missing values were present in the dataset, please refer to table 2 for the percentage of missing values present in the attributes. Figure 2 also shows the missing values present in each attribute.

The nature of this value is MCAR, which stands for 'Missing Completely At Random' and it is considered that it will introduce no bias (R. J. Little and D. B. Rubin, 1987).
The computations also indicate that the dataset is imbalanced and that the positive class 'FR' consists merely of 4.8% of the dependent feature. We, however, followed appropriate techniques to handle missing values and to tackle the problems relating to imbalanced class in the study. These are explained in greater detail in the latter part of the paper.

Data Preparation: In this phase, we partition the data into 70% training data and 30% testing data thereby obtaining the following number of instances:

The computations also indicate that the dataset is imbalanced and that the positive class 'FR' consists merely of 4.6% of the dependent feature. We, however, followed appropriate techniques to handle missing values and to tackle the problems relating to imbalanced class in the study. These are explained in greater detail in the latter part of the paper.

- Removing the record with missing values : This method is found to be the simplest one as it simply deletes the record that consists of missing values in any of the attributes. We use it as there is enough instances left after the process.
- Using K-Nearest Neighbors algorithm: KNN performs data imputation by taking the majority of the K nearest values (G. E. A. P. A. Batista and M. C. Monard, 2002). It is clustering base technique and we used k=10 for performing the imputation. We can then compare the performance of the model developed using the data imputed by the above two techniques.

Modelling: Choosing the right classifier for a dataset is very much crucial for any machine learning study. Classifier which performs well in one dataset may not do so in another dataset. In order to identify the most appropriate classifier which provides the best performance, various state of the art classifiers are explored in this study using R as a programming language in the RStudio IDE (*RStudio*, 2011). The performance of the various classifiers are analyses using R Caret package (Max *et al.*, 2020). We use "FR" as positive class and performed analysis with the classifiers Naïve Bayes(nb), K-Nearest Neighbors (knn), J48, random forest (rf) , ROC-Based Classifier (rocc), Stochastic Gradient Boosting (gbm),

eXtreme Gradient Boosting(xgbTree), JRip, OneR, Classical Soft Independent Modelling of Class Analogy (CSimca), Robust Soft Independent Modelling of Class Analogy (Rsimca), C5.0Cost, PART, Conditional Inference Tree (ctree), Bagged CART (treebag), Bayesian Generalized Linear Model (bayesglm), Averaged Neural Network(avNNet), Neural Network(nnet), C5.0, C5.0Rules, C5.0Tree and Stabilized Nearest Neighbor Classifier (snn). While executing the classifiers, we use 10 fold cross validation to minimize the overfitting.

As has been cited earlier, the dataset is imbalanced and hence, while using 'accuracy' as a metric of the performance of the classifier, we have found out that it leads to deceiving, misleading and ambiguous outcome. When the computations provide high accuracy, the penalty that comes with it includes very less or no classification of the positive class. Similar mishaps occur for classification of negative classes by certain classifiers.

It was thus considered to be more appropriate to use 'balanced accuracy', as defined by the following formula to measure the performance of a classifier:

Balanced Accuracy = (Sensitivity + Specificity)/2 Other computations include metrics like Kappa, Precision, etc.

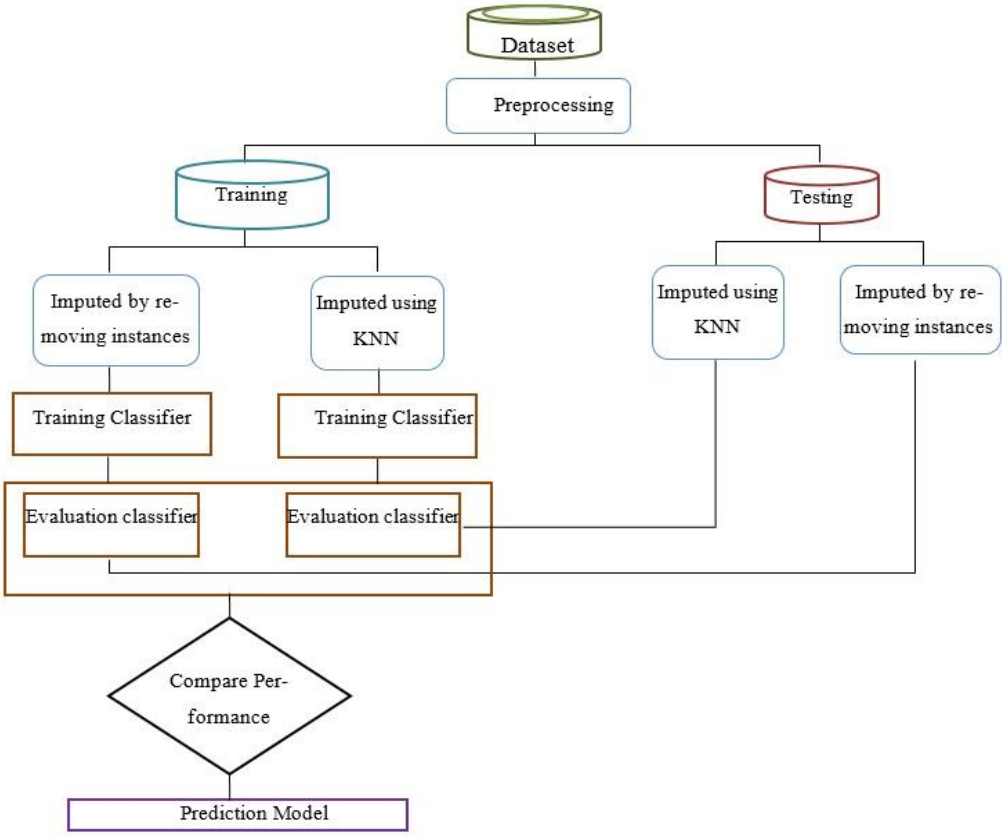

Figure 1. Framework of the Methodology

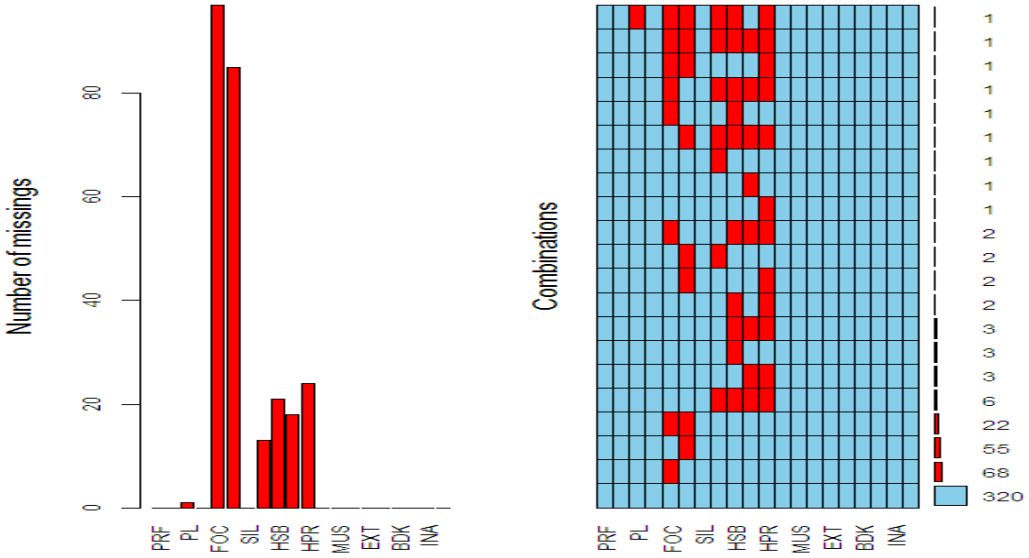

Figure 2. Attributes with missing values

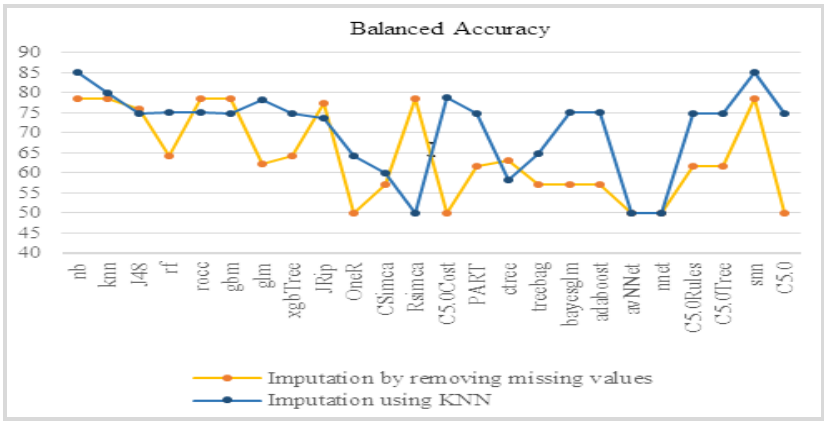

Figure 3. Comparison of Balanced Accuracy

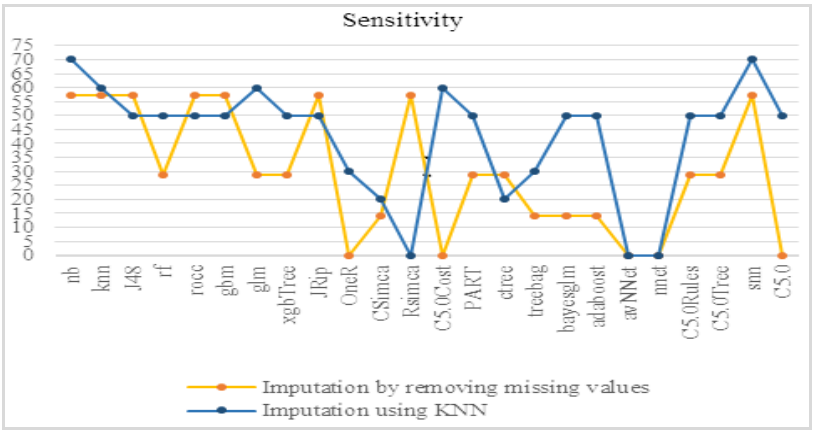

Figure 4. Comparison of Sensitivity

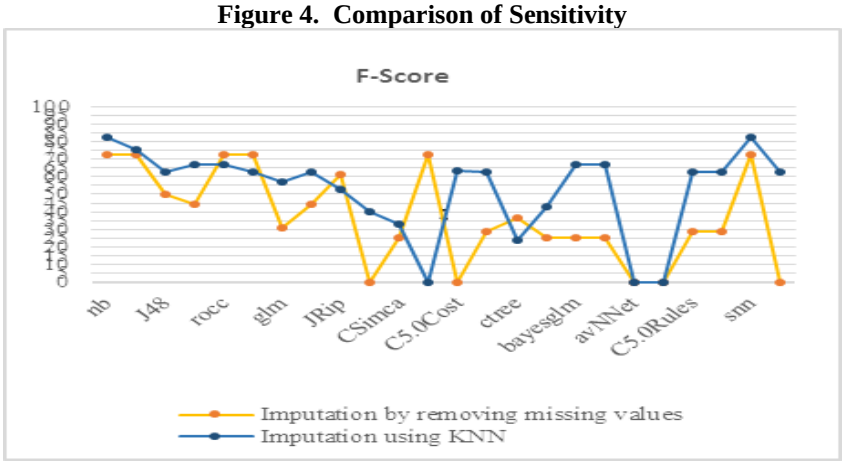

Figure 5. Comparison of F-Score

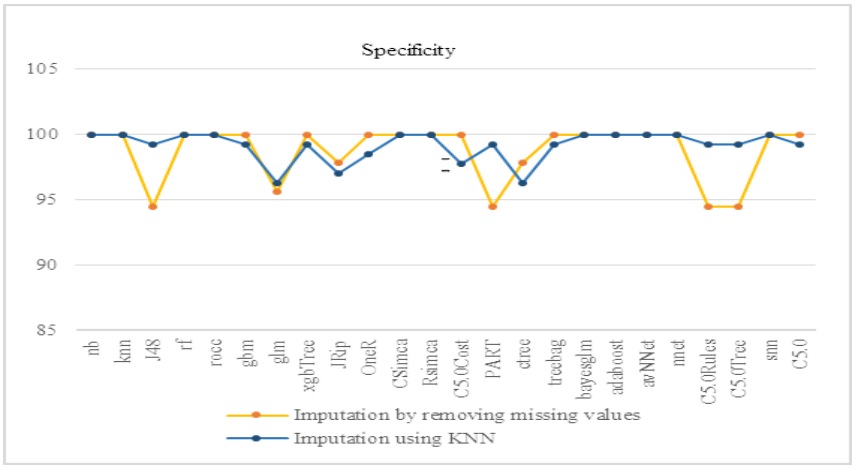

Figure 6. Comparison of Specificity

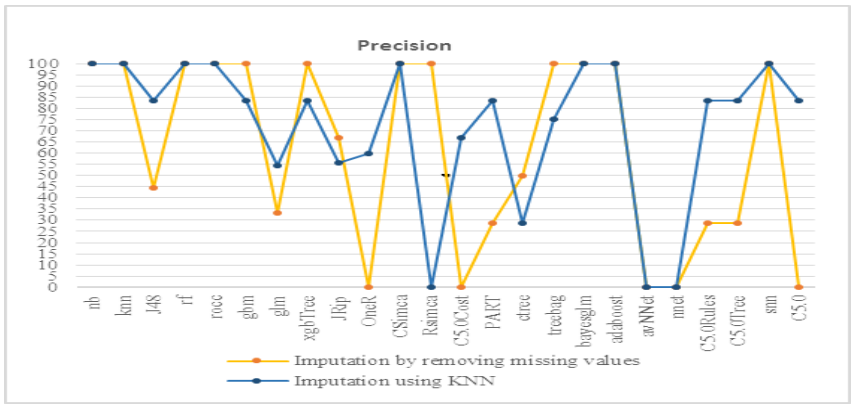

Figure 7. Comparison of Precision

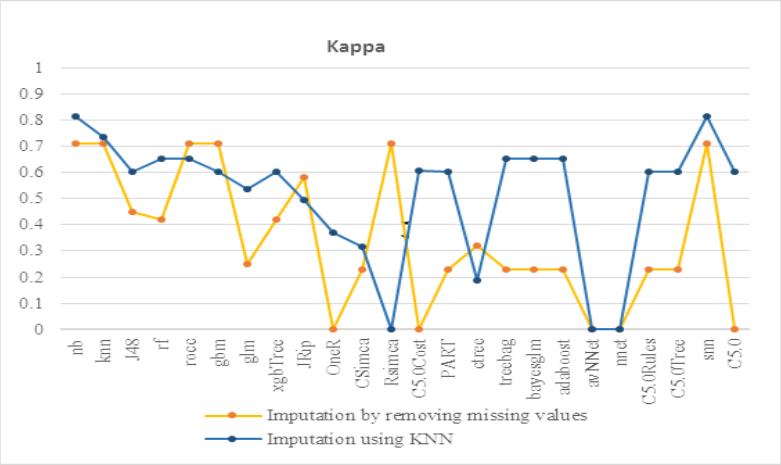

Figure 8. Comparison of Kappa Value

Table I. Data Description

SL. No.	Attribute	Description	Values
1	PRF	Performance of the student as provided by CGPA	If CGPA>=1.6, GD denoting Good else FR for Fair
2	GND	Gender of the student	"M" denoting Male and "F" denoting Female
3	PL	Permanent Location where the student was brought up	RR denoting Rural, UR denoting Urban
4	CAT	Whether the student belongs to General, Other backward class, schedule caste or schedule tribe as categories at India	"GEN" denoting General, "ST" for Schedule Tribe, "SC" for schedule caste and "OBC" for other backward class
5	FOC	Father's Occupation	"TR" denoting Teacher, "GS" denoting Govt. Servant , "BS" denoting Businessman , "FR" denoting Farmer/cultivator/Laborer /Disability, " OTH" denoting Pastor/Social Worker/Driver /Hospital worker /Ministry Admin etc.
6	MOC	Mother's Occupation	" HW" denoting house wife and " NHW" denoting working woman

7	SIL	Location of the institute where the student studied or appeared for the Matriculation	” RR” denoting Rural, ” UR” denoting Urban and ”PT” for Private candidate
8	MPR	Performance of the student at the Matriculation as provided by the Matriculation Exam	“VG” for >=60%; “GD” >=45% else ” FR“
9	HSB	Subjects studied at the Higher Secondary/12thStd	"A" for Arts, "B" for Science, "C" for Commerce, "V" for Vocational
10	HSL	Location of the institute where the student studied or appeared for the Higher Secondary Examination/12th Std.	” RR” denoting Rural, ” UR” denoting Urban and ”PT” for Private candidate
11	HPR	Performance of the student at the higher secondary as provided by the Board Examination at 12thStd	“VG” for >=60%; “GD” >=45% else ” FR“
12	NLT	Naturalistic Intelligence	Score denoted by integer value between 1 to 10
13	MUS	Musical Intelligence	Score denoted by integer value between 1 to 10
14	LOM	Logical Mathematical Intelligence	Score denoted by integer value between 1 to 10
15	EXT	Existential Intelligence	Score denoted by integer value between 1 to 10
16	INE	Interpersonal Intelligence	Score denoted by integer value between 1 to 10
17	BDK	Bodily Kinesthetic Intelligence	Score denoted by integer value between 1 to 10
18	VLI	Verbal Linguistic Intelligence	Score denoted by integer value between 1 to 10
19	INA	Intrapersonal Intelligence	Score denoted by integer value between 1 to 10
20	VSP	Visual Spatial Intelligence	Score denoted by integer value between 1 to 10

Table II. Percentage of missing value in the attribute

PL	FOC	MOC	MPR	HSB	HSL	HPR
0.2	19.52	17.1	2.62	4.23	3.62	4.83

Table III. Training and Testing Data

Training Data		Testing Data	
PRF	Frequency	PRF	Frequency
FR	13	FR	10
GD	340	GD	134

Table IV. Result of the analysis on data imputed by removing instance with any missing value

Classifier	Accuracy	Sensitivi-ty	Specificity	Preci-sion	Kap-pa	Balanced Accuracy
nb	96.94	57.14	100	100	0.71	78.57
knn	96.94	57.14	100	100	0.71	78.57
J48	91.84	57.14	94.50	44.44	0.45	75.82
rf	94.90	28.57	100	100	0.42	64.28
rocc	96.94	57.14	100	100	0.71	78.57
gbm	96.94	57.14	100	100	0.71	78.57
glm	90.82	28.57	95.60	33.33	0.25	62.08
xgbTree	94.90	28.57	100	100	0.42	64.28
JRip	94.90	57.14	97.80	66.66	0.58	77.47
OneR	92.86	0000	100	0000	000	50.00
CSimca	93.88	14.29	100	100	0.23	57.14
Rsimca	96.94	57.14	100	100	0.71	78.57
C5.0Cost	92.86	0000	100	0000	0000	50.00
PART	89.80	28.57	94.50	28.57	0.23	61.53
ctree	92.86	28.57	97.80	50.00	0.32	63.18
treebag	93.88	14.29	100	100	0.23	57.14
bayesglm	93.88	14.29	100	100	0.23	57.14
adaboost	93.88	14.29	100	100	0.23	57.14
avNNet	92.86	0000	100	0000	000	50.00
nnet	92.86	0000	100	0000	000	50.00
C5.0Rules	89.80	28.57	94.50	28.57	0.23	61.53
C5.0Tree	89.80	28.57	94.50	28.57	0.23	61.53
snn	96.94	57.14	100	100	0.71	78.57
C5.0	92.86	0000	100	0000	000	50.00

Table V. Result of the analysis on data imputed using KNN

Classifier	Accuracy	Sensitivity	Specificity	Precision	Kappa	Balanced Accuracy
nb	97.92	70	100	100	0.812	85
knn	97.22	60	100	100	0.736	80
J48	95.83	50	99.25	83.33	0.604	74.63
rf	96.53	50	100	100	0.650	75
rocc	96.53	50	100	100	0.650	75
gbm	95.83	50	99.25	83.33	0.604	74.63
glm	93.75	60	96.27	54.54	0.537	78.13
xgbTree	95.83	50	99.25	83.33	0.604	74.63
JRip	93.75	50	97.02	55.55	0.493	73.51
OneR	93.75	30	98.51	60	0.370	64.25
CSimca	94.44	20	100	100	0.317	60
Rsimca	93.06	0	100	0	0	50
C5.0Cost	95.14	60	97.76	66.66	0.605	78.88
PART	95.83	50	99.25	83.33	0.604	74.63
ctree	90.97	20	96.27	28.57	0.188	58.14
treebag	94.44	30	99.25	75	0.650	64.63
bayesglm	96.53	50	100	100	0.650	75
adaboost	96.53	50	100	100	0.650	75
avNNet	93.06	0	100	0	0	50
nnet	93.06	0	100	0	0	50
C5.0Rules	95.83	50	99.25	83.33	0.604	74.63
C5.0Tree	95.83	50	99.25	83.33	0.604	74.63
snn	97.92	70	100	100	0.812	85
C5.0	95.83	50	99.25	83.33	0.604	74.63

4. Result And Conclusion

The results of the above modeling phase are shown in table 4 and table 5. The performance of the classifiers has found to be better for the dataset imputed using KNN algorithm rather than the one imputed by merely removing records with some missing values in any of the attributes. We have also observed that Stabilized Nearest Neighbor Classifier (SNN) and Naïve Bayes (NB) classifiers provide the highest balanced accuracy (85%), please see figure 3.

Since we are more concerned with the correct classification of FR and cannot afford to have more false negatives, the sensitivity metrics has also been taken

into consideration. Under this scenario, both "SNN" and "NB" has been found to provide 70% sensitivity, see figure 4.

It may also be noted that sensitivity of SNN is higher than that of KNN. It has thus been established that SSN achieves greater improvement in the classification instability in comparison with KNN. The performance of the classifiers in terms of F-Score, specificity, precision and kappa in the dataset imputed with KNN and imputed by discarding the instances with the missing values are presented in Figure 5, Figure 6, Figure 7 and Figure 8 respectively. The performance of the model developed in this study has thus been proved to provide satisfactory results and hence suitable enough for deployment. Our future strategy includes exploring some hybrid approaches to improve the performance, vis-a-vis, the sensitivity.

Acknowledgment: We are thankful and grateful to the administration of Martin Luther Christian University, Meghalaya for providing the necessary data without which this study would have been a distant dream.

References

[1] Altabrawee, H., Ali, O. A. J. and Ajmi, S. Q. (2019) 'Predicting Students' Performance Using Machine Learning Techniques', JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences, 27(1), pp. 194–205. doi: 10.29196/jubpas.v27i1.2108.

[2] Dawngliani, M. S. et al. (2020) 'Comparison of Decision Tree-Based Learning Algorithms Using Breast Cancer Data', in Lecture Notes on Data Engineering and Communications Technologies. Madurai, Tamil Nadu: Lecture Notes on Data Engineering and Communications Technologies, Springer Publication., pp. 885–896. doi: 10.1007/978-3-030-43192-1_96.

[3] Dawngliani M.S, Chandrasekaran N, S. L. (2019) 'A Comparative Study between Data Mining Classification and Ensemble Techniques for Predicting Survivability of Breast Cancer Patients', International Journal of Computer Science and Mobile Computing, 8(9), pp. 1–10.

[4] Dawngliani M S, Chandrasekaran N, L. R. and T. H. (2020) 'Breast Cancer Recurrence Prediction Model Using Voting Technique', in Conference: International Conference on Mobile Computing and Sustainable Informatics (ICMCSI 2020). Nepal: EAI/Springer Innovations in Communication and Computing,.

[5] Dawngliani M S, Chandrasekaran N, S. L. (2019) 'Development of a Model to Predict Breast Cancer Recurrence Using Decision Tree based Learning Algorithms', THINK INDIA JOURNAL, 22(10).

[6] Fatima, S. and Mahgoub, S. (2019) 'Predicting Student's Performance in Education using Data Mining Techniques', International Journal of Computer Applications, 177(19), pp. 14–20. doi: 10.5120/ijca2019919607.

[7] G. E. A. P. A. Batista and M. C. Monard (no date) 'K-Nearest Neighbour as Imputation Method: Experimental Results (in print)', in Technical Report,ICMC-USP,2002. 0103-2569.

[8] Gardner, H. (2011) A Beginner's Guide to the theory of Multiple Intelligence. Available at: https://www.multipleintelligencesoasis.org/ (Accessed: 10 June 2020).
[9] Imran, M. et al. (2019) 'Student academic performance prediction using supervised learning techniques', International Journal of Emerging Technologies in Learning, 14(14), pp. 92–104. doi: 10.3991/ijet.v14i14.10310.
[10] International Educational Data Mining Society (2008). Available at:
[11] http://educationaldatamining.org/EDM2008/.
[12] Krishna, M. et al. (2020) 'Predicting Student Performance using Classification and Regression Trees Algorithm', International Journal of Innovative Technology and Exploring Engineering, 9(3), pp. 3349–3356. doi: 10.35940/ijitee.c8964.019320.
[13] Max, A. et al. (2020) 'Package " caret " R topics documented :'
[14] Multiple Intelligences Inventory (2017). Available at: http://surfaquarium.com/MI/inventory.htm (Accessed: 10 June 2019).
[15] R. J. Little and D. B. Rubin (1987) Statistical Analysis with Missing Data. New York: John Wiley and Sons.
[16] RStudio (2011). Available at: https://rstudio.com/products/rstudio/ (Accessed: 5 April 2018).
[17] Shearer, C. et al. (2000) 'The CRISP-DM model: The New Blueprint for Data Mining', Journal of Data Warehousing, 5(4), pp. 13–22. Available at: www.spss.com%5Cnwww.dw-institute.com.
[18] Urkude, S. and Gupta, K. (2019) 'Student intervention system using machine learning techniques', International Journal of Engineering and Advanced Technology, 8(6 Special Issue 3), pp. 2061–2065. doi: 10.35940/ijeat.F1392.0986S319.
[19] Zulfiker, M. S. et al. (2020) 'Predicting students' performance of the private universities of Bangladesh using machine learning approaches', International Journal of Advanced Computer Science and Applications, 11(3), pp. 672–679. doi: 10.14569/ijacsa.2020.0110383.