International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Oct-2022)

https://dx.doi.org/10.12785/ijcds/120196

Systematic Approach for Re-Sampling and Prediction of Low Sample Educational Datasets

ARHAM TARIQ[1], YASIR NIAZ KHAN[2], AHMAD AMIN[3], MUDASSAR NASEER[2] and BILAL HUSSAIN[2]

[1]*Department of Computer Science, University of Central Punjab, Lahore, Pakistan*
[2]*Department of Computer Science, University of Lahore, Lahore, Pakistan*
[3]*Department of Computer Science, Superior University, Lahore, Pakistan*

Received 22 Jan. 2021, Revised 15 Jul. 2022, Accepted 23 Jul. 2022, Published 31 Oct. 2022

Abstract: From last couple of decades, literacy rate is increased all over the globe so as the educational datasets. Prediction of student's performance is considered as an emerging research area under educational data mining. Previous studies have noticed that most of the available educational datasets are of small sample size. These datasets provide fewer generalization opportunities, which makes them difficult to analyze. Previous approaches use noise filtering, data balancing, GAN-based oversampling, or mostly rely on classifiers' performance. In this paper, we proposed an approach that provides an improved model that optimizes the classifier's performance and removes the adverse effects of noisy instances, and increase data balancing tendency in a better way. The proposed model is based on CTGAN (Conditional Tabular Generative Model), NCC (Nearest Centroid Classifier) combined with data balancing algorithm SMOTE-IPF (Iterative-Partitioning Filter) to increase dataset size by keeping their balanced nature intact and also to minimize the negative effect of noisy data points. Finally, for prediction six classifiers Random Forest (RF), Gradient Boosting (GB), CAT Boost (CT), Extra Tree (ET), K-Nearest Neighbor (KNN), and AdaBoost (AB) are used and their parameters are tuned. After parameter optimization stacking among different combination of classifiers is applied using Logistic Regression. The detailed analysis of results elaborates that the proposed model supersedes previous approaches by 2-2.5% in terms of Accuracy, and ROC.

Keywords: Low Sample Educational Datasets, Conditional Tabular Generative Model, Students' academic performance, Educational Data Mining, SMOTE-IPF

1. Introduction and Overview

The ability to predict a student's academic performance in an educational environment could be notable in a number of ways. Taking any level of education under consideration, many factors contribute to student success e.g. the pressure of performance and prevention from drop out introduces a new level of fear in students that is needed to be overcome.

With the enhancement in technology, one can contribute to society by effectively analyzing the key factors that can help stakeholders to get an insightful overview of the student's attributes and improve student's performance in academia. Data Mining can be used as a powerful practice to detect patterns from datasets. Notable relationships between different attributes of a dataset can be extracted by using different data mining techniques. Educational Data mining (EDM) can be referred to as the analysis of student attributes to correlate it with their academic success at any level of academia.

Due to the reason that mostly educational datasets are of small sample size [1-5]. These datasets provide fewer underlying patterns for classifiers to generalize, which makes them difficult to analyze. Taking previous approaches under consideration, they used data balancing, GAN-based oversampling, or used only Machine Learning classifiers performance and noise filtering based models. This study aims to produce a model capable enough to handle limitations in previous approaches.

The paper will investigate the classifier's performance evaluated on to the data set re-sampled with enhanced data balancing and noise filtering approach. The proposed approach utilizes systematic configuration of both data balancing algorithms, GAN-based model to oversample data, and also manages noisy data points and optimizes classifiers performance as well. To develop this mixed hybrid improved approach, the proposed model is based on CTGAN [6] for creating and NCC for extracting the nonnoisy closest neighbors of majority synthetic samples and combining these extracted majority synthetic samples

Email:Arhamtariq99@gmail.com,yasir.niaz@cs.uol.edu.pk,ahmadaminvhr@gmail.com,mudasser.naseer@cs.uol.edu.pk,bilal.hussain@cs.uol.edu.pk

with the original dataset so, that tendency of data balancing algorithms can be enhanced. After it, the data balancing algorithm SMOTE-IPF [7] is applied so that more synthetic minority samples with noise filtering can be created by considering the increased ratio of majority samples. In the last step, we performed parameter tuning of Random Forest, Gradient Boosting, Ada Boost, Cat Boost, Extra Tree Classifier, and KNN, and then a Stacked ensemble among the best of them is created using Logistic Regression as a Meta classifier for the prediction of the oversampled dataset.

The rest of the paper is organized as follows: section two elaborates problem description section three highlights the previous work, section four provides the information of the dataset, and section five gave a brief description of the proposed model. Detailed results are defined in section six, section seven compares the proposed approach with the existing techniques, and conclusion of our work is given in section eight.

2. PROBLEM DESCRIPTION

Most of the educational datasets are not created with a high number of attributes or instances generally, they have a low sample size. Analyzing these datasets can generate a lack of optimization and generalization opportunities that may lead to the poor performance of predictive models. The proposed model lead us towards the improved approach by handling limitations present in the previous approaches. Limitations of previous approaches are mentioned below:

- Most of the approaches rely only on classifiers, this can result in over fitting for small-size datasets. The presence of noisy instances can also cause a negative effect on the classifier performance.
- The filter-based approaches can handle noisy instance problems, but they cannot balance datasets or increase small dataset size. This can lead to fewer underlying patterns in the dataset to generalize..
- Data balancing algorithms can easily handle skew datasets. But they are limited to only oversampling of minority classes. If the skew between the number of minority and majority classes of a dataset is less then low number of new data points will be created.
- GAN-based oversampling methods can increase both minority and majority class data points. Previously in EDM, GAN based approach was used to increase dataset size, not to balance it, both of them were not configured at the same time. Apart from these Tabular based GAN models are specially designed to replicate privacy issues dataset, so there are possibilities of including new noisy instances to an existing dataset.

3. Related Work

Imran et al. [8] Used a dataset gathered from UCI Machine Learning Repository having 33 attributes and 1044 instances. The model predicts academic success with the approach based on discretization, filter-based feature selection, class balancing, and homogeneous ensembles machine learning method with an accuracy rate of 95.78 Two principal approaches have been followed with regard to high performance computing CPU power measurement: direct measurement and estimation.

Injadat et al [9] used two datasets of university students having 52 and 480 instances. The model predicted grades at course level by training datasets at different stages, hyper-tuning machine learning algorithms with Grid search, and finding the best combination of ensemble classifiers on different dataset stages. The model achieved an accuracy of 89.9datasets used.

Chakrabarty et al. [12] predicted admission status using the graduate student's dataset having 8 attributes and 500 instances with R-score of 0.84. The model was based on Random Forest Regressor for feature selection, Grid search for parameter optimization, and Gradient Regressor Boosting for prediction. Chui et al. [1] proposed a model based on the improved Conditional Generative Adversarial Network- Based Deep Support Vector Machine, in which Conditional GAN (generative adversarial network) was used to increase data points and Network-Based Deep Support Vector Machine was used for prediction. The model attained 95.710.971gathered from UCI Machine Learning Repository having 33 attributes and 1044 instances. Rohani et al. [13] proposed a hybrid model by combining Simulated Annealing Algorithm and Genetic algorithm for student academic performance prediction. The model used a dataset collected from the UCI Machine Learning Repository, having average accuracy of 92.70used a dataset gathered from UCI Machine Learning Repository having 33 attributes and 1044 instances. The model was based on the Wrapper attribute selection method and comparative analysis between three different classifiers (Decision Tree, Random Forest, and Naive Bayes) to achieve an accuracy of 93.67Ashfaq [10] achieved an accuracy of 84.1by using a dataset having 16 attributes of 480 students gathered from the Kalboard 360 Learning Management System. The model was based on hyper- parameter tuning Random Forest with Random Search and ADASYN balancing technique.

Ajibade et al [11] configured a model by using discretization, an under- sampling approach to balance the dataset, they used six different classifiers Naïve Bayes, Decision Tree, K-Nearest Neighbor, Discriminant Analysis, Pairwise Coupling combined with the boosting approach for prediction. The model achieved 94.1accuracy on the dataset gathered from Kalboard 360 Learning management system.

Walia et al. [15] used a dataset collected from UCI Machine Learning Repository to predict academic success. The model was based on three algorithms Ranker, BestFit, and Greedy Stepwise for attribute selection and five machine learning classifiers NB, DT, RF, JRip, and ZeroR with a maximum accuracy of 84.81

Rimadana et al. [3] predicted the academic success of students using a time skill management dataset. The approach was configured on the dataset of 125 students based on 23 attributes. The model attained an accuracy of 84machine learning classifiers and using the most frequent value to handle missing values. Huda et al. [16] predicted academic success using a dataset having 33 attributes and 395 instances gathered from the University of Minho in Portugal. The model was based on SVM classifier compared with KNN-classifier. The model achived an accuracy of 90.25Utomo et al. [17] configured a model based on K-NN,C4.5 with SMOTE data balancing approach to predict academic performance. The model was based on a dataset having 16 attributes of 480 students gathered from the Kalboard 360 Learning Management System. The model achieved an accuracy of 74.09

Thammasiri et al. [18] configured a model for dealing with class imbalance problems in the academic dataset. The model was based SVM classifier combined with SMOTE (minority oversampling). The dataset used in the approach was based on 34 attributes gathered during the 2005-2011 academic session from a school based in Tulsa, USA. The model achieved an accuracy of 90.24

Satyanarayana et al. [19] configured an approach for noise data points elimination for educational datasets prediction. Base level classifiers (J48, RF, NB) in the model eliminates the noisy instances using majority-based voting. The model achieved an accuracy of 94.5a dataset having 1044 instances and 33 attributes gathered from UCI Machine Learning Repository. Ahmed et al. [20] proposed Fast KNN for improving the speed and accuracy of the traditional KNN algorithm. The model introduces the concept of moment descriptor in KNN. The model achieves an accuracy of 90.25academic performance dataset gathered from UCI Machine Learning Repository.

4. DATASET

The dataset which we used in our proposed work was originally used in [21]. It was gathered in the 2005-2006 academic session from two schools of Portugal. It contains 33 attributes having information related to demographics, habits, grades, and features of school students. The dataset contains 1044 instances, 649 samples in the dataset belong to the Portuguese language class and 395 samples belong to the Mathematics class. Out of 33 attributes, 29 are acquired from the questionnaire and 4 of them are gathered from school reports.

Table 1 describes attributes present in the dataset used in the proposed methodology.

5. PROPOSED METHODOLOGY

This section will provide a detailed description of the proposed model. The proposed model provides a better approach as compared to the previous approaches and generates more appropriate results. The proposed model is based on a mixed hybrid approach. It increases the size of the dataset and also balances it with the presence of noise filtering. Fig.1 illustrates the detailed architecture of the proposed model which is divided into several steps: the first step contains data pre-processing; the second step involves synthetic data generation to increase dataset size. In step number three, the nearest and non-noisy majority class samples have been extracted and then combine with the original dataset in step four. Data balancing technique is applied in step five. In step 6 7, different classifiers are firstly parameter tuned and then stacking is applied among different combinations of them. Lastly model performance is evaluated using 10-fold cross validation in step 8

A. *Data Pre-Processing*

Data pre-processing is considered an essential step in machine learning because it transforms data into a more digestible form to achieve best performance of the algorithm. To achieve this task first of all features are encoded using binary, ordinal, and label encoding. The actual data was distributed into two classes' Portuguese language and Mathematics class, both of them are combined into a single dataset by introducing a new attribute named Course (P for Portugal or M for Mathematics). The target attribute named G3 in the dataset is converted into two different classes (Pass/Fail) by transforming its continuous values into ordinal values. If the G3 attribute value is equal to or greater than 10 then Pass else labeled as Fail. The continuous attributes of G3 have been converted into nominal by considering the same grading system mentioned in the origin of the dataset [21].

B. *Synthetic Data Generation*

CTGAN is a deep learning-based model. It is specially designed for generating synthetic samples of diversified tabular numeric datasets [6]. CTGAN can create synthetic data effectively because its network structure contains a fully connected layer for capturing maximum co-relation among data points. In this step, CTGAN is used to generate synthetic samples from the original dataset. Synthetic samples are created so that more majority class samples can be added to the original dataset that will increase the tendency of data balancing algorithms to add moreminority class samples by considering an increased ratio of majority samples. The original dataset consists of814 instances of the Pass (majority) class and 230 of Fail(minority) class.

Fig. 2 shows the original dataset class distribution. Y-axis in Fig 2 represents the frequency of the number of data points present in the original dataset. X-axis represents two classes Pass (1) and Fail (0) data points present in the dataset. CTGAN default parameters are used to generate synthetic samples from the original dataset. Table 2 illustrates the parameters of CTGAN used for generating a synthetic dataset.

TABLE I. Attribute description with their value.

Index	Attribute	Description	Possible Values
1	Sex	Student’s Gender	Binary = F, M
2	School	Student’s School (Ga briel or Mousinho)	Binary = G, M
3	Age	Age of student	Numeric = 15 to 22
4	Address	Student’s residence	Binary = Urban or rural
5	Meducation	Mother’s education	Numeric = 0-None, 1- Primary, 2-Fifth to Ninth, 3-Secondary Education,4-Higher Education
6	Feducation	Father’s education	Numeric = 0-None, 1- Primary, 2-Fifth to Ninth, 3-Secondary Education,4-Higher Education
7	Father’s education	Guardian of student	Nominal = Father, mother or other
8	Family Size	Student’s family members	Binary = greater or less than 3
9	Family Relation	Quality of family relationship	Numeric = from 1-very bad to 5-Excellent
10	Reason	Purpose to join the school	Nominal= near to home, repute of school, course preference or other reasons
11	Travel Time	Time required to reach school	Numeric = 1-15 min, 15-30 min, 30 min -1 hour, more than 1 hour
12	Study Time	Time required to reach school	Numeric = 1-15 min, 15-30 min, 30 min -1 hour, more than 1 hour
13	Failures	Number of student’s failures in	Numeric 1 to 4
14	Schoolup	Extra educational assistance provided by school	Binary = Yes, No
15	Famup	Family Education Support	Binary = Yes, No
16	Activities	Extra –curricular activities at school	Binary = Yes, No
17	Paidclass	Extra Paid classes	Binary = Yes, No
18	Internet	Student’s internet availability status	Binary = Yes, No
19	Nursery	Nursery Student attended nursey school or not	Binary = Yes, No
20	Higher	Student interested in pursuing higher education	Binary = Yes, No
21	Romantic	Student’s romantic relationship status	Binary = Yes, No
22	Freetime	Free time after school	Numeric = from 1 very low to 5 very high
23	Goout	Going out with friends	Numeric = from 1 very low to 5 very high
24	Walc	Students’ weekly Alcohol	Numeric = from 1 very low to 5 very high
25	Dalc	Student’s Alcohol consumption at the workplace	Numeric = from 1 very low to 5 very high
26	Health	Student’s current health status	Numeric = from 1 very bad to 5 very good
27	Absences	Student’s total absences	Numeric = 0 to 93 yearly
28	Pstatus	Parent’s cohabitation status	Binary = ’1’ - living together or ’0- apart
29	Mjob	Mother’s job	nominal: ’teacher’, ’health”care related, ’civil services’ (e.g. administrative or police),’at home’ or ’other
30	Fjob	Father’s Job	nominal: ’teacher’,’health”care related, ’civil services’ (e.g. administrative or police),’at home’ or ’other
31	G1	First period grade marks	Numeric = 0 to 20
32	G2	Second period grade marks	Numeric = 0 to 20
33	G3	Third period grade marks	Numeric = 0 to 20
34	Course	Third period grade marks	Binary = Mathematics (mat) and Portuguese language (por)

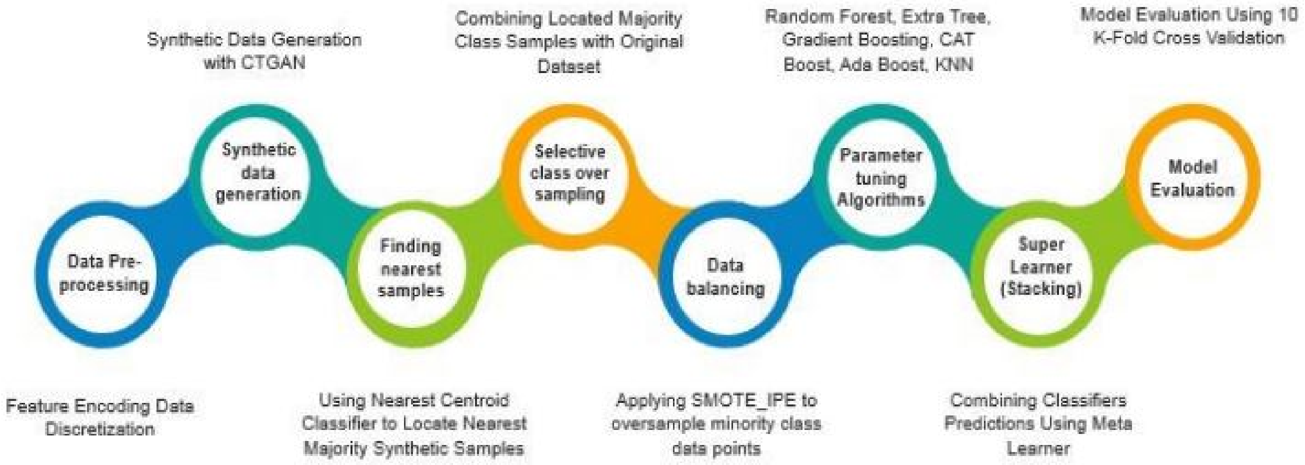
Synthetic Data Generation with CTGAN
Combining Located Majority Class Samples with Original Dataset
Random Forest, Extra Tree, Gradient Boosting, CAT Boost, Ada Boost, KNN
Model Evaluation Using 10 K-Fold Cross Validation
Synthetic data generation
Selective class over sampling
Parameter tuning Algorithms
Model Evaluation
Data Pre-processing
Finding nearest samples
Data balancing
Super Learner (Stacking)
Feature Encoding Data Discretization
Using Nearest Centroid Classifier to Locate Nearest Majority Synthetic Samples
Applying SMOTE_IPE to oversample minority class data points
Combining Classifiers Predictions Using Meta Learner

Figure. 1 Proposed Methodology

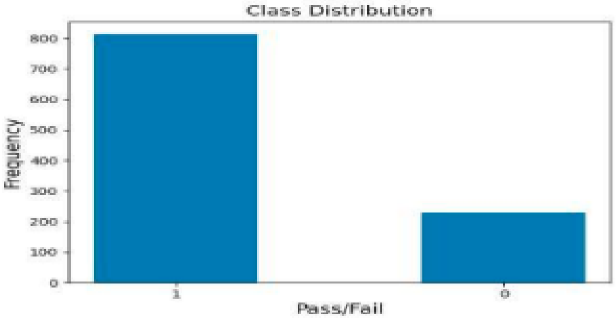

Fig. 2. Original Dataset Class Distribution

TABLE II. CTGAN parameters

Parameter Name	Value
Epochs	300
Batch size	500

C. Data Integration

In this phase nearest synthetic majority samples are extracted from synthetic data created from CTGAN by using NCC. Extracted nearest majority samples are then combined with the original dataset. This combination will increase data set skew and enhance data balancing to create more minority samples by considering the increased ratio of majority samples. There are three types of data points in a tabular dataset as described in Figure 3.These can be termed as (i) safe zone instances (data points that can be easily separated),(ii) borderline instances (data points very close to the decision boundary), and(iii) noisy instances (data points that cannot be separated easily and are located far away from decision boundary. Furthermore noisy instances are those data points that can be a cause to reduce classifier's performance.

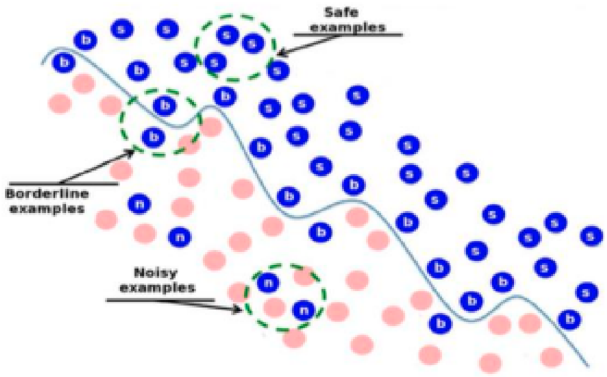

Figure 3 data points distribution types

To ensure that only useful and non-noisy samples are extracted from synthetic data a filtering process is required. For this purpose, NCC is trained on the original dataset and tested on a synthetic dataset. NCC predicts those data instances whose centroid's are close to the trained dataset instances. The majority of samples that are classified are separated and then combined with the original dataset. Data points generated from CTGAN are filtered with an NCC model and only the majority of samples are extracted because of two reasons. Firstly, the NCC classifier will ensure that only non-noisy samples are extracted from synthetic data. Secondly, because traditional data balancing algorithms oversamples minority instances by considering the ratio of majority samples [22-25] as described in Eq. (1). Adding synthetic majority class samples will increase the tendency of oversampling algorithms to generate more new data points, which will eventually increase generalization and optimization opportunities better than the traditional oversampling approach. The improved data balancing formula configured in the proposed methodology can be illustrated in Eq. (2) below. Eq. (1) represents a traditional approach for data balancing, previous approaches create minority samples by finding a difference between a total numbers of majority samples with a total number of minority samples present in the dataset. Eq. (2) represents oversampling approach introduced in the proposed approach. It increases the tendency of oversampling algorithms to create more balanced points by firstly adding synthetic majority data points, then making oversampling algorithms to create more minority samples by considering an increased ratio of majority class samples. mathtools

$$\begin{gathered}\textbf{Traditional Oversampling}\\ \textbf{=original majority-original minority instances}\end{gathered}\tag{1}$$

$$\begin{gathered}\textbf{Improved Oversampling =}\\ \textbf{(original majority+synthetic majority)–original}\\ \textbf{minority instances}\end{gathered}\tag{2}$$

Total 560 synthetic majority samples have been extracted in this phase and combined with the original dataset. Fig. 4 displays the class distribution after the combination of an original dataset with the nearest synthetic majority samples. Yaxis in Fig 4 represents the frequency of the number of data points present in the dataset oversampled with the CTGAN and NCC combined approach. The X-axis represents two classes Pass (1) and Fail (0) data points present in the dataset.

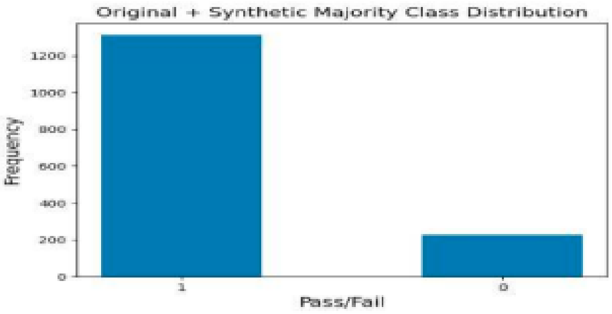

Figure. 4 Original + Synthetic Majority Class Distribution

D. Data Balancing

In this step skewed nature of data is managed. Dataset remained imbalanced if data instances of some particular class are not equal to other class instances. Managing the skew/imbalance dataset is important otherwise it may result in a negative impact on the performance of a machine learning classifier [26]. Tabular GAN's cannot remove skew nature completely but increase dataset size at the same time. Because Tabular GAN is originally designed for replicating privacy issues datasets or missing values imputation [27-29]. That is why a SMOTE IPF has been used instead of CTGAN for class balancing. SMOTE IPF overcomes noisy instances problems in the dataset by using Iterative-Partitioning-based Filter and then oversamples minority instances with SMOTE data balancing algorithm [7]. Because the tendency of data oversampling was improved before applying SMOTE IPF as mentioned in Eq. (2). So, that's why it results in a greater number of both minority and majority samples. It eventually creates more useful data points, resulting in improving classifiers performance. Comparison of traditional and improved oversampling results can be seen in Table 6 and Table 7. Fig.5 illustrates the class distribution after application of SMOTE IPF on the dataset increased with synthetic majority instances.

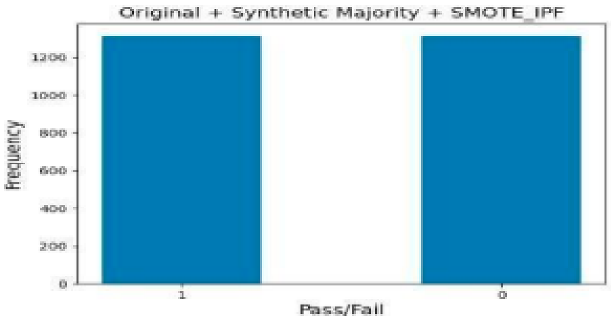

Figure. 5 Original Synthetic Majority+SMOTE_IPF

E. Model Construction

In this step of proposed approach, classifiers are applied to predict oversampled datasets. Initially, six classifiers RandomForest, GradientBoosting, CATBoost, ExtraTree, AdaBoost with Decision Tree as a base classifier and KNN are parameter tuned using Grid search. Combining multiple classifiers can perform better than a single classifier [30].After parameter optimization of these classifiers, the best among them are selected, and then their predictions are combined using Stacking. Stacking is an ensemble-based approach that involves two levels of predictions, termed Base and Meta level. Final predictions in Stacking are made by Meta level classifier after obtaining predictions from Base level classifiers in the form of vectors. In the proposed model Logistic Regression is used as a Meta classifier. Parameters list of each classifier used during hyperparameter tuning is given in Table 3 below

TABLE III. Machine Learning Classifiers

No	Classifier	Parameter Space
1	Random Forest	Estimators = 10,50,100,200,500
2	Gradient Boosting	Estimators = 10,50,100,200,500 , Learning rate = 0.05,0.5,1.0
3	Ada Boost	Estimators = 10,50,100,200,500 , Learning rate = 0.05,0.5,1.0 Decision Tree Depth=1,5,10
4	Cat Boost	Estimators = 10,50,100,200,500 , Learning rate = 0.05,0.5,1.0
5	Extra Tree	Estimators = 10,50,100,200,500
6	KNN	K = 1,2,3,4,5,6,7,8

F. Evaluation Metrics

During our experimentation, we used five commonly available measures for the evaluation of the machine learning classifiers. Details of these evaluation metrics is given below:

- Precision: It represents the accuracy of minority class samples present in the dataset. The formula for calculating the precision of a classifier is given in Eq. (3), below

$$\textbf{\textit{Precision}} = \textbf{\textit{TP/(TP+FP)}} \quad (3)$$

- Recall:It represents the fraction of the total number of instances correctly predicted by the classifier over the total number of positive instances present in the dataset. The formula for calculating recall is given in Eq. (4), below

$$\textbf{\textit{Recall}} = \textbf{\textit{TP/(TP+FN)}} \quad (4)$$

- Accuracy:It is the most commonly used evaluation metric used to measure the classification quality of a classifier. Accuracy can be termed as the total number of correctly classified or predicted data points over the total number of data points. The formula for accuracy of a classifier is given in Eq. (5), below

$$\textbf{\textit{Accuracy}} = \textbf{\textit{TP+TN/(TP+TN+FN+FP)}} \quad (5)$$

- ROC (receiver operating characteristic curve):It can be termed as the capability of a classifier to distinguish between different class instances present within the dataset in terms of predicted probabilities. It can be computed by plotting the true positive rate along Y-axis and the false positive rate along X-axis in the graph.

- Specificity: it can be termed as the measure of how well a machine learning classifier identifies negative cases present in the dataset. The formula for specificity is given in Eq. (6), below

$$\textbf{\textit{Specificity}} = \textbf{\textit{TN/(TN+FP)}} \quad (6)$$

6. Results Simulation

This section will illustrate the detailed results with and without the application of the proposed methodology. Fig. 6-9 represent scatter plots of the data points present in dataset versions of original, oversampled with CTGAN and NCC, balanced with SMOTE IPF, and the proposed methodology. Orange samples in the scatter plot represent majority (Pass/1) class data points, while Blue samples represent minority (Fail/0) class data points. The independent variable is on the x-axis, and the dependent variable is on the y-axis. These figures are generated after dimensionality reduction using Principal Component Analysis (PCA). PCA is an unsupervised approach to reduce large sets of data into smaller ones [31]. Original Dataset with 33 dimensions /attributes are reduced into 2 dimensions to visualize its distribution. Figure 6 describes a scatter plot of data points of the original dataset. It can be viewed that classes are not balanced. Orange color instances are greater in number than Blue color instances, and noisy instances of a class (1) are present between classes (0). Figure 7 represents a scatter plot of data points of dataset oversampled with majority instances extracted from CTGAN and NCC classifier. It can be viewed that only the majority (Orange) instances density is increased as compared to the original dataset. Fig. 8 represents the scatter pot of the dataset balanced with SMOTE-IPF. By comparing it with Fig. 6, we observed that more minority samples (Blue) are created due to class balancing. Fig 9 represents the scatter plot of the dataset oversampled with the proposed approach. Because the data set generated in the proposed methodology is firstly oversampled with filtered non–noisy majority (Orange) data points and then noise filtering based SMOTE IPF is applied to oversample minority (Blue) data points. That’s why it can be viewed in Fig 9 that both majority (Orange) and minority (Blue) data points presence is quite dense, fewer empty regions are left and noisy instances are not grown in size. Furthermore in Fig. 9 by its comparison with Fig 6-8 states that not only dataset size is increased, but both the number of minority (Blue) and majority (Orange) instances are oversampled with no new noisy instances generation, which will eventually lead us to better result.

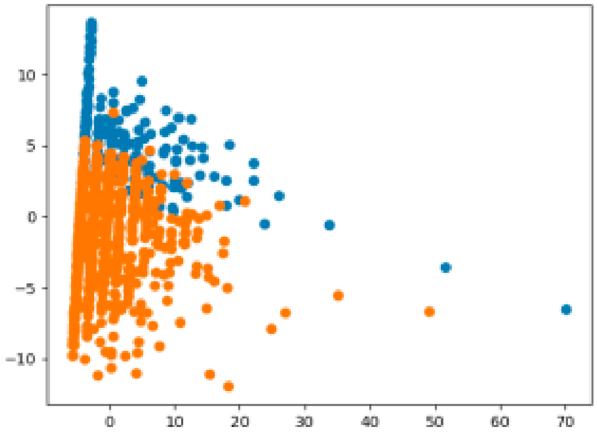

Figure. 6 Original Data Distribution

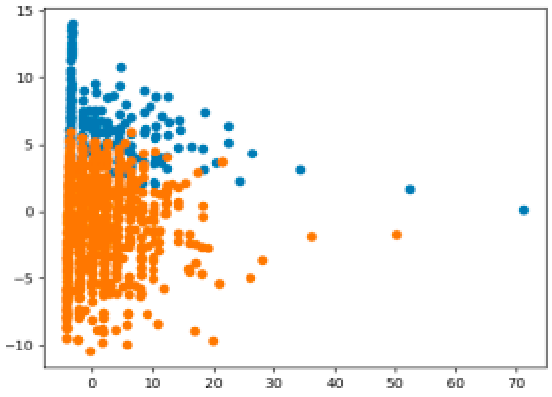

Figure.7 Majority Oversampled with CTGAN and NCC Distribution

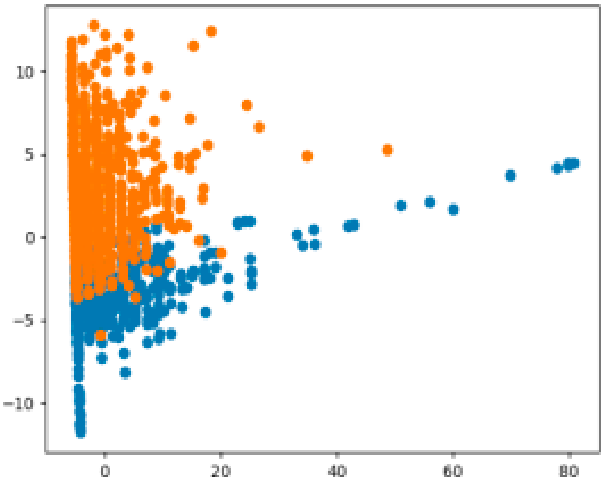

Figure. 8 Oversampled With SMOTE-IPF

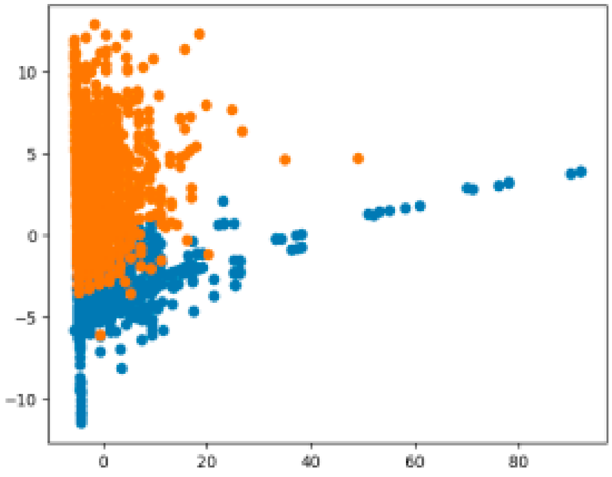

Figure. 9 Oversampled With Proposed Model

Figure from 10 to 13 represent the accuracy of classifiers applied to the original, oversampled with CTGAN and NCC, balanced with SMOTE IPF and proposed methodology dataset versions. Figure 10 illustrates the accuracy of classifiers applied to the original dataset. Because noisy instances are present, data size is small and classes are not balanced, that is why results of the original dataset are limited. Figure 11 describes the accuracy of classifiers applied to the datasets oversampled with the majority samples extracted from CTGAN and NCC classifier. Although results are improved from the original dataset, but still classes are not balanced, that's why results can further be improved. Figure 12 represents the accuracy attained by applying classifiers on the dataset balanced with SMOTE-IPF. Classifiers results are improved because classes are balanced and noise filtering is also applied. Furthermore, results can be further be improved if more balanced samples can be created. Figure 13 represents the accuracy of classifiers applied to the dataset oversampled with the proposed approach. As the data balancing tendency is improved and noisy instances are also filtered,that's why classifiers accuracy scored the most as compared to previous dataset versions.

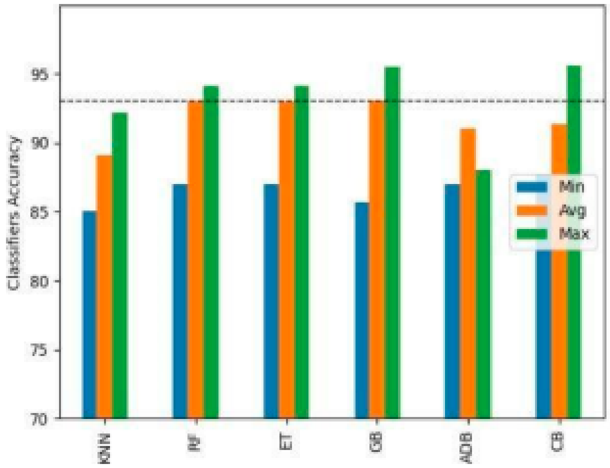

Figure 10. Results of the original dataset

Tables 4-7 represent detailed results in terms of five evaluation metrics of machine learning classifiers with 10 fold cross-validations. All the classifiers parameters are tuned by using the same parameters as mentioned in Table 3. Table 4 represents the evaluation of classifiers applied to the original dataset having 814 majority and 203 minority instances. Table 5 shows the evaluation of classifiers applied to the original dataset combined with CTGAN and NCC extracted majority class data points. Dataset version used in Table 5 contains 203 minority and 1374 majority instances. Table 6 shows evaluation of classifiers applied to the balanced dataset (SMOTE IPF) having 814 majority and 814 minority instances. Table 7 shows the evaluation of classifiers applied to the dataset oversampled with the proposed methodology having 1374 majority and 1374 minority instances. The following factors in the proposed methodology contribute

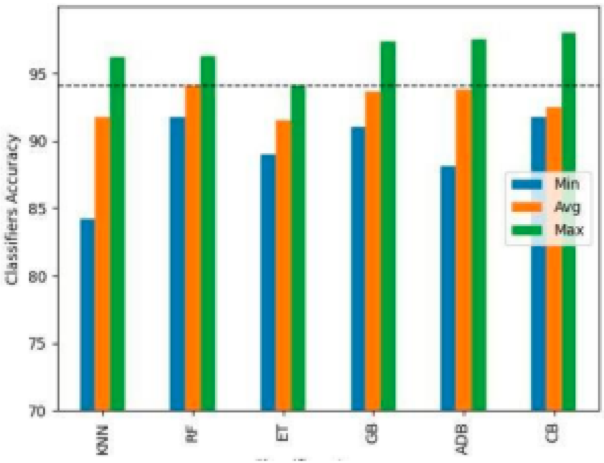

Figure1 1. Result of CTGAN + NCC

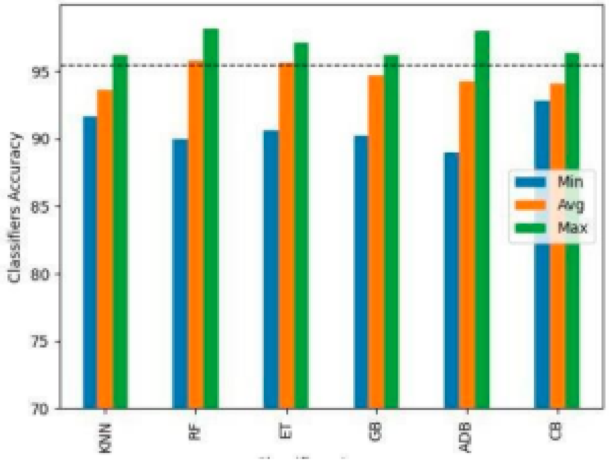

Figure 12. Results of Balance Dataset (SMOTE IPF)

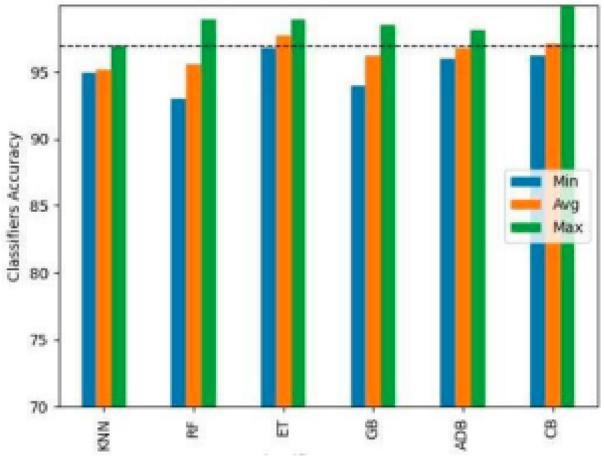

Figure 13. Results of Proposed Methodology

in producing better results:

- The number of samples generated is greater than other datasets versions that generate more underlying patterns.
- Classes are balanced and noise is filtered both during majority and minority oversampling.
- The tendency of oversampling algorithms is enhanced by adding synthetic majority samples extracted from CTGAN and NCC combined approaches.

TABLE IV. Results of Original Dataset

Classifier	Precision	Recall	ROC	Specificity	Accuracy
GB	77.2%	83.2%	86.5%	93.9%	91.8%
AB	76.0%	73.0%	84.5%	93.5%	89.1%
RF	78.7%	84.9%	87.3%	94.0%	92.9%
ET	64.0%	87.2%	80.7%	90.5%	92.8%
KNN	64.0%	73.0%	79.0%	90.0%	89.1%
CB	76.5%	83.7%	87.3%	94.0%	91.7%
Stacking (RF+CB +GB)	80.0%	83.5%	88.8%	94.8%	93.1%

TABLE V. Results of CTGAN,NCC(Majority) with Original Dataset

Classifier	Precision	Recall	ROC	Specificity	Accuracy
GB	76.0%	83.5%	86.8%	95.9%	93.8%
AB	73.8%	83.7%	85.7%	95.7%	94.1%
RF	67.5%	89.4%	83.1%	94.9%	94.3%
ET	56.0%	90.1%	77.8%	93.5%	91.7%
KNN	66.3%	74.1%	81.1%	94.3%	91.8%
CB	76.2%	83.5%	86.8%	95.9%	93.9%
Stacking (AB+CB +GB)	76.5%	83.6%	87.0%	96.0%	94.4%

TABLE VI. Results of Balanced Dataset (SMOTE IPF)

Classifier	Precision	Recall	ROC	Specificity	Accuracy
GB	94.9%	94.3%	94.6%	94.9%	94.6%
AB	95.7%	93.4%	94.3%	95.4%	94.7%
RF	95.8%	94.7%	95.1%	95.6%	95.8%
ET	97.2%	95.6%	96.4%	97.2%	95.9%
KNN	99.6%	89.3%	93.9%	99.5%	93.6%
CB	95.4%	94.8%	95.1%	95.4%	94.1%
Stacking (ET+RF +GB)	96.3%	95.8%	95.9%	96.2%	96.0%

TABLE VII. Results of Proposed Methodology

Classifier	Precision	Recall	ROC	Specificity	Accuracy
GB	97.3%	96.2%	96.8%	97.4%	96.2%
RF	9.74%	9.73%	9.73%	97.4%	96.2%
ET	98.2%	97.3%	97.7%	98.2%	9.77%
KNN	98.9%	93.5%	95.9%	98.9%	95.2%
CB	97.4%	97.0%	97.1%	97.4%	9.72%
Stacking (ET+RF+ CB)	98.0%	97.7%	97.8%	98.0%	97.9%

7. Model Comparison

In this section, we have compared results of our proposed methodology with other approaches configured for the prediction of low sample educational datasets. All the approaches mentioned below are configured on the same dataset used in the proposed methodology. Comparison of results is based on the presence or absence of Class balancing, Synthetic data oversampling, Noise filtering, Parameter optimization of classifiers, and Ensemble method configuration. Table 8 below provides results comparison based on accuracy and ROC using dataset of UCI student academic performance. caption

TABLE VIII. Model Comparison

Sr.No	Methodology	Class Balancing	Noise Filtering	Synthetic Data Oversampling	Ensemble Method	Parameter Tuning	Metrics	Result
1	SVM [16]	Absent	Absent	Absent	Absent	Absent	Accuracy	90.5%
2	Simulated Annealing+Genetic Algorithm[13]	Absent	Absent	Absent	Absent	Present	Accuracy	92.70%
3	Modified K Nearest Neighbor [20]	Absent	Absent	Absent	Absent	Absent	Accuracy	90.25%
4	Random Forest + Wrapper Feature Selection [14]	Absent	Absent	Absent	Present	Absent	Accuracy	93.07%
5	Ensemble Noise Filtering[19]	Absent	Present	Absent	Present	Absent	Accuracy	94.5%
6	ADASYN +Real AdaBoost[8]	Present	Absent	Absent	Present	Absent	Accuracy	95.78%
7	ICGAN DSVM [1]	Absent	Absent	Present	Absent	Present	ROC	95.1%
8	**Proposed Methodology**	**Present**	**Present**	**Present**	**Present**	**Present**	**ROC**	**97.8%**
9	**Proposed Methodology**	**Present**	**Present**	**Present**	**Present**	**Present**	**Accuracy**	**97.9%**

8. Conclusion Future Work

Accurate prediction of academic performance can prove to be beneficial for many stakeholders including teachers, students, parents,and of the educational institute. In this paper, the authors have focused on the effective oversampling of the small-sized educational dataset for increasing generalization opportunities. For this purpose systematic combination of Tabular Conditional GAN, Nearest Centroid Classifier (NCC) with hybrid Data Balancing algorithm (SMOTE-IPF) is configured for data re-sampling. For prediction six classifier's six classifiers Random Forest (RF), Gradient Boosting (GB), CAT Boost (CT), Extra Tree (ET), KNN, and AdaBoost (AB) are hyper parameter tuned and Stacked ensemble among the best of them is created. It has been found if small datasets are oversampled in such a way that class balancing algorithms capability is enhanced and noisy instances are filtered as well then machine learning classifiers performance can be improved as compared to the previous approaches. Experimental results gathered from implementing the proposed methodology show that our model outperforms previous approaches by around 2% by achieving 97.9% accuracy. The current approach is configured only for the binary problem-based dataset. In the future it can be extended for the multi-label classification problem. Replacement of the traditional ensemble method with deep learning in the proposed model can produce more predictive power.

References

[1] K. T. Chui, R. W. Liu, M. Zhao, and P. O. D. Pablos, "Predicting Students' Performance With School and Family Tutoring Using Generative Adversarial Network-Based Deep Support Vector Machine," IEEE Access, vol. 8, pp. 86745–86752, 2020, doi: 10.1109/ACCESS.2020.2992869

[2] M. R. Rimadana, S. S. Kusumawardani, P. I. Santosa, and M. S. F. Erwianda, "Predicting Student Academic Performance using Machine Learning and Time Management Skill Data," in 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Dec. 2019, pp. 511–515, doi: 10.1109/ISRITI48646.2019.9034585

[3] L. M. Abu Zohair, "Prediction of Student's performance by modelling small dataset size," Int. J. Educ. Technol. High. Educ., vol. 16, no. 1, p. 27, Aug. 2019, doi: 10.1186/s41239-019-0160-3.

[4] K. Hussain, N. Talpur, M. U. Aftab, and Zakria, "A Novel Metaheuristic Approach to Optimization of Neuro-Fuzzy System for Students' Performance Prediction," J. Soft Comput. Data Min., vol. 1, no. 1, Art. no. 1, Mar. 2020.

[5] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," Decis. Anal., vol. 2, no. 1, p. 1, Mar. 2015, doi: 10.1186/s40165-014-0010-2.

[6] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular data using Conditional GAN," 3rd Conf. Neural Inf. Process. Syst. NeurIPS 2019 Vanc. Can., Oct. 2019, Accessed: Jan. 03, 2021. [Online]. Available: http://arxiv.org/abs/1907.00503.

[7] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," Inf. Sci., vol. 291, pp. 184–203, Jan. 2015, doi: 10.1016/j.ins.2014.08.051.

[8] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student Academic Performance Prediction using Supervised Learning Techniques," Int. J. Emerg. Technol. Learn. IJET, vol. 14, no. 14, p. 92, Jul. 2019, doi: 10.3991/ijet.v14i14.10310

[9] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," Knowl.- Based Syst., vol. 200, p. 105992, Jul. 2020, doi: 10.1016/j.knosys.2020.105992.

[10] U. Ashfaq, "Managing Student Performance: A Predictive Analytics using Imbalanced Data," Int. J. Recent Technol. Eng., vol. 8, no. 6, pp. 2277–2283, Mar. 2020, doi: 10.35940/ijrte.E7008.038620.

[11] S.-S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, "A Data Mining Approach to Predict Academic Performance of Students Using Ensemble Techniques," in Intelligent Systems Design and Applications, vol. 940, A. Abraham, A. K. Cherukuri, P. Melin, and N. Gandhi, Eds. Cham: Springer International Publishing, 2020, pp. 749– 760.

[12] N. Chakrabarty, S. Chowdhury, and S. Rana, "A Statistical Approach to Graduate Admissions' Chance Prediction," Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, vol 103. Springer., 2020, pp. 333–340.

[13] Y. Rohani, Z. Torabi, and S. Kianian, "A Novel Hybrid Genetic Algorithm to Predict Students' Academic Performance," J Electr Comput Eng Innov., vol. 8, no. 2, pp. 219–232, 2020.

[14] F. Ünal, "Data Mining for Student Performance Prediction in Education," Data Min. - Methods Appl. Syst., Mar. 2020, doi: 10.5772/intechopen.91449.

[15] N. Walia, M. Kumar, N. Nayar, and G. Mehta, "Student's Academic Performance Prediction in Academic using Data Mining Techniques," presented at the 1 st International Conference on Intelligent Communication and Computational Research, 2020, doi: 10.2139/ssrn.3565874.

[16] H. Al-Shehri et al., "Student performance prediction using Support Vector Machine and KNearest Neighbor," in 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), Apr. 2017, pp. 1–4, doi: 10.1109/CCECE.2017.7946847.

[17] U. Pujianto, W. A. Prasetyo, and A. R. Taufani, "Students Academic Performance Prediction with k-Nearest Neighbor and C4.5 on SMOTE-balanced data," in 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Dec. 2020, pp. 348– 353, doi: 10.1109/ISRITI51436.2020.9315439.

[18] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," Expert Syst. Appl., vol. 41, no. 2, pp. 321–330, Feb. 2014, doi: 10.1016/j.eswa.2013.07.046.

[19] A. Satyanarayana and M. Nuckowski, "Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance," Publ. Res., Apr. 2016, [Online]. Available: https://academicworks.cuny.edu/ny pubs/79.

[20] A.-H. Rafah, S. T. Ahmad, and M. S.Croock, "Enhancement of

student performance prediction using modified K-nearest neighbor,"TELKOMNIKA Telecommun. Comput. Electron.Control, vol. 18, no. 4,Art.no.4,Aug.2020,doi:10.12928/telkomnika.v18i4.13849.

[21] P. Cortez and A. Silva, "USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE," pp. 5–12, 2008.

[22] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline oversampling for imbalanced data classification," Int. J. Knowl. Eng. Soft Data Paradig., vol. 3, no. 1, pp. 4–21, Apr. 2011, doi: 10.1504/IJKESDP.2011.039875.

[23] J. Wang, M. Xu, H. Wang, and J. Zhang, "Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding," in 2006 8th international Conference on Signal Processing, Nov. 2006, vol. 3, doi: 10.1109/ICOSP.2006.345752.

[24] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Jun. 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.

[25] A. Ghazikhani, H. S. Yazdi, and R. Monsefi, "Class imbalance handling using wrapper- based random oversampling," in 20th Iranian Conference on Electrical Engineering (ICEE2012), May 2012, pp. 611–616, doi:10.1109/IranianCEE.2012.6292428.

[26] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," Inf. Sci., vol. 513, pp. 429–441, Mar. 2020, doi:10.1016/j.ins.2019.11.004.

[27] K. Kuo, "Generative Synthesis of Insurance Datasets," ArXiv191202423 Cs Q-Fin Stat, Aug. 2020, Accessed: Apr. 07, 2021. [Online]. Available: http://arxiv.org/abs/1912.02423.

[28] Z. Zhao, A. Kunar, H. Scheer, R. Birke, and L. Chen, "CTAB-GAN: Effective Table Data Synthesizing. 2021".arXiv:2102.08369

[29] J. Kim, D. Tae, and J. Seok, "A Survey of Missing Data Imputation Using Generative Adversarial Networks," Feb. 2020, pp. 454–456, doi: 10.1109/ICAIIC48513.2020.9065044.

[30] S. Džeroski and B. Ženko, "Is Combining Classifiers with Stacking Better than Selecting the Best One?," Mach. Learn., vol. 54, no. 3, pp. 255– 273, Mar. 2004, doi: 10.1023/B:MACH.0000015881.36452.6e

[31] H. Abdi and L. J. Williams, "Principal component analysis," WIREs Comput. Stat., vol. 2, no. 4, pp. 433–459, 2010, doi: https://doi.org/10.1002/wics.101.

Mr. Muhammad Arham Tariq Muhammad Arham Tariq is working as Lecturer in the University of Central Punjab's Computer Science Department. He received the MSCS degree from The University of Lahore. He was completed Master of Computer science from the University of Lahore as well. His main area of interest is Machine learning, Computer vision, Data-mining, Mobile application and Web application development.

Dr. Yasir Niaz Khan Dr. Yasir Niaz Khan obtained his Ph.D. at the University of Tubingen in 2013. During his Ph.D., he conducted research on detection of terrain (ground surfaces) using a camera mounted on a flying and a ground robot. Upon completing his graduate studies, Dr. Khan started teaching robotics at FAST-NU, Lahore, Pakistan.

Mr. Ahmed Amin Mr. Ahmed Amin received the MSCS degree from The University of Lahore. Now he is working in the teaching faculty of the Department of CS and IT, Superior University, Lahore. His research interest includes NLP, Data Mining, Deep and Machine learning.

Dr .Mudassar Naseer Dr .Mudassar Naseer earned his Ph.D. degree in April 2010 from Beijing University of Aeronautics and Astronautics, China. He received his M.Phil. (Statistics) degree from Government College University Lahore, Pakistan in 2001 and MS(CS) degree from LUMS Lahore Pakistan, in 2004. He has a teaching experience of more than 17 years in various educational institutions including 6 years of teaching in Govt. College University Lahore, Pakistan. Currently he is holding the position of Associate Professor at University of Lahore, Lahore Campus

Mr. Bilal Hussain Mr. Bilal Hussain is currently a Lecturer in the faculty of Computer Sciences at University of Lahore (UOL). Prior to his recent appointment at the UOL, he was a lecturer in Computer Science department at University of South Asia. Bilal Hussain received his undergraduate degree from COMSATS Lahore and his Master's degree from University of Lahore. Bilal Hussain published a number of papers in preferred Journals and chapters in books, and participated in a range of forums on Computer Sciences.