Proceedings of the Fourth International Conference on Smart Systems and Inventive Technology (ICSSIT-2022)
IEEE Xplore Part Number: CFP22P17-ART; ISBN: 978-1-6654-0118-0

Machine Learning based Prediction of Dropout Students from the Education University using SMOTE

M.Revathy, Research Scholar,
Department of Computer Science,
Vels Institute of Science, Technology & Advanced Studies. (VISTAS)
Pallavaram, Chennai, India
revathysathya85@gmail.com

S. Kamalakkannan, Associate Professor,
Department of Information Technology,
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Pallavaram, Chennai, India
kannan.scs@velsuniv.ac.in

P.Kavitha, Research Scholar,
Department of Computer Science,
Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Pallavaram, Chennai, India
pkavikamal@gmail.com

2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT) | 978-1-6654-0118-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSSIT53264.2022.9716450

Abstract— **In past decade, there have been several students who have dropped out from the educational institutions and it is increasing rapidly. This has become one of the challenging factors for the educational institution. The students are getting into the institution and embarking their learning with several expectations and dreams. The expectations of the students have not fulfilled due to various factors such as staff, management, parents, course chosen etc., that make them drop from their registered curriculum. However, this has become the main issue for all educational institutes wherein several researchers introduced the technique of data mining for analysing as well as predicting the student's dropout. Therefore, this paper focuses on early finding of dropout variables as an advance by dimensionality reduction using feature selection and extraction methods. In feature extraction, there may be an occurrence of imbalanced data that may affect the significance of Machine Learning (ML) techniques. Thus, Synthetic Minority Oversampling Technique (SMOTE) is subsequently added with Principal Compound Analysis (PCA) whereas the oversampling of imbalanced data is managed to balanced dataset. Moreover, the 1,243 student's data have been collected and analysed using proposed PCA-SMOTE to allow for a more accurate forecast in case of dropout. Accuracy performance related to PCA-SMOTE has been 97.6% that is evaluated through confusion matrix parameter and compared with existing ML to find out the exact students who are not satisfied with their fulfilment in the environment of education institute.**

Keywords: Dropout student, Imbalanced data, Synthetic Minority Oversampling Technique (SMOTE), Principal Compound Analysis (PCA), Machine Learning (ML), Education institute, prediction, imbalanced data

1. INTRODUCTION

The phrase "dropout" refers to a student's dismissal from a college, school or any other educational institution after failing to complete the required course. The society, organization and nation-building may get affected, if students are getting dropped out of the programme. In all educational sectors, long-term student retention is a constant objective. Filling seats in a certain amount of time period is one of the existing practices in higher education. Each year, hundreds of students enroll and the amount of data collected has grown steadily. According to the university's requirements, all registered student must take an external, internal and practical examination. Many students fail to reach the standards of university because they are unable to meet the requirements of the institution [1]. Millions of students or more leave school or any other educational institution devoid of completing the course requirements every year and 8,000 students per day leave throughout an academic year. Institutional and individual factors both impact dropout rates. The number of students dropping out of school has increased dramatically in contemporary years. As a result, it connects all other educational institutions to the key thread. There is in need of component that influences student's decision to drop out while evaluating them. Numerous studies are being conducted to determine the specific variables which cause students at various educational levels such as elementary, secondary and higher education to dropout. Educational Data Mining (EDM) is utilized and deployed in this study for forecasting the specific reasons or factors to student dropout [2]. It is utilized for recognizing students in academics who have displayed inferior performance [3].

As a human resource, higher education is frequently used as a measure to determine the quality of student education. Higher education is often regarded as a respectable institution when students attain good results and are competent in their disciplines. Alternatively, failure of students may have a negative influence on institutions and students. Currently, the issue of student failure is identified as an ongoing university challenge to examine various variables that cause students to drop out like demographics, academic performance, economic assistance and student conduct among others. Students are unable to finish their education by the end of the given study time are considered to have dropped out. It reduces dropout students' skills and abilities in their professions, which has a negative impact on the quality of institution [4]. Researchers frequently utilize variables including internal assessment, high

school education, extracurricular events, student demographics, external assessment, Cumulative Grade Point Average (CGPA) and social interaction network to predict dropouts [5]. The internal assessment indicators and CGPA are the most promising factors, since their value make the most of the measurement of a student's future and current abilities [6]. Demographic factors particularly gender has persuaded learning characteristics in the first two years of study in traditional higher education and also in student's online programme. Economic limitations, student absenteeism, age, parental effect, career prospects and marital status are all factors that might lead to dropout [7] [8].

Data which contains large dimensions are beyond human interpretation as they do not fetch valuable information leading to time and space complexity along with the undue influence on cost for processing them. To solve the problem of multiple dimensions, the attributes which are most appropriate from the larger spaces is chosen. Reducing the high dimensional data into their lower forms while upholding the intrinsic information extracted from the larger spaces is what is come to be known as dimensionality reduction. Dimensionality reduction is used for various purposes such as visualizing the data in 2D or 3D, compressing the data for effective analysis and noise removal to ensure that accurate data is obtained in response to a query, which is important as the noisy data contribute a detrimental influence on the model's accuracy. The best algorithm of dimensionality reduction technique is compared and adopted in the student dropout model to analyse and predict the student's dropout. The student dropout is dependent on academic performance and can be analysed using z factor [9].

Even if various techniques for solving classification issues have been devised, the features and quantity of the available datasets comprises constantly a considerable, if not larger effect on the predictions eminence. The idea of class imbalance, bestowing to which one class's label is signified to a smaller degree is one data characteristic that significantly confuses classification issues. When classifiers are trained on unbalanced datasets, they produce models that are heavily biased toward the majority class [10]. On top of that, the interest class is generally the class with the fewest occurrences that results in more false positive predictions [11]. Such a result can have dire consequences in detecting faults or intrusions, problems arising from medical domains and several others. The former implies that conventional classification methods are modified for improving the challenge of learning for the minority class, whilst latter means adjusting the class imbalance ratio to produce a more equal dissemination of classes. Due to their adaptability and simplicity, data-level approaches that either conduct under sampling of majority instances or oversampling of minority ones are more commonly used [12]. To avoid the elimination of important majority instances, oversampling algorithms represent the preferable choice. The Synthetic Minority Oversampling Technique (SMOTE) algorithm as introduced by Chawla et al is their most prominent example [13]. In reminiscence of the origins of SMOTE [14], Chawla highlights two classification results that have been sought to improve when developing this algorithm is

1) Performance on minority class
2) Generalization capacity

However, the novelty of SMOTE is in its synthetic instance creation mechanism that tried to cope with challenges arisen from overfitting the minority class when performing random oversampling, the state-of-the-art method at that time. Hence, this paper has proposed PCA-SMOTE algorithm to identify the exact reasoning of those students who have mentioned "Yes" in the dropout interest status. Thus, the proposed PCA-SMOTE gets implemented in various ML models for justifying the accuracy performance of the respective algorithm.

The paper is organized as follows. The prior research on several predictions modelling in the realm of education as well as educational data mining research is detailed in section 2. The method of proposed research along with SMOTE working procedure for discovering classification approaches in predicting student dropout is outlined in Section 3. The PCA-SMOTE with KNN model has performed better in confusion matrix values than other existing ML model is discussed in Section 4. Finally, the performed PCA-SMOTE algorithm has assisted in identifying the exact reasoning with better accurate prediction is concluded in Section 5.

2. Literature Review

Earlier research has developed many definitions of student dropout. The utmost general definition examines whether students have been dynamic till the end of week or if present week is the last week in which they are active. Primary detection of students on the verge of dropping out is critical for minimizing the problem and permitting for the implementation of needed conditions. EDM is an interdisciplinary field concerned with the creation of tools for examining a wide range of unique data in the field of education, with the goal of defining appropriate learning approaches and better understanding students' needs [15]. For the purpose of enhancing the performance of students as well as the teaching-learning process, EDM is used to foresee its issues [16]. Due to the vast data in educational dataset, it has worries about how to adapt data mining approaches and identify patterns that are typically very difficult to address [17]. Data mining has aided in the discovery of datasets with various methodologies like mathematical methods, statistical models and machine learning algorithms as a decision-making standard [18].

W. Yu,T.-C. Lin,Y.-C. Chen and D. Kaufman centered their examination on the application of a probability model to envisage college student retention. Amongst the years 2000 and 2008, they gathered data. Chi-square analysis is used to examine an attribute from the data set. A strong GPA in pre-university and high school decreases attrition as per the research. Students benefit from orientation in core courses, English language, on-campus jobs and a comfortable hostel stay. The researchers employed non-linear regression model built particularly for binary dependent variables [19]. As indicated by V. Arul Kumar et al, this paper suggests applying the techniques of dimensionality reduction for reducing multiple dimensions with in detail explanation [20]. Cattell, R.B. et al, has focused PCA to be the best approach and says that covariance matrix, eigen values are the best pursuits for dimensionality reduction is the most widely used dimensionality reduction technique and in the PCA for minimizing the problem we use covariance matrix, eigenvectors and also Eigen values [21]. PCA can be often used in reducing dimensionality, multiple linear regression and also several regression techniques, projection pursuit and independent component analysis. Muhammad Shakil Pervez et al., has completely reviewed the various techniques of feature selection and differentiated them [22]. Albert Bifet et al., to reduce the multi-dimensionality author have compared a genetic algorithm with existing PCA technique [23]. The genetic algorithm is used as a search engine, then the strategy that has minimum validation error is selected finally as the summation of all separate results are taken for final accuracy. Later this accuracy is compared with PCA accuracy.

The supervised learning approaches are used to anticipate a student's decision-making process when it comes to choosing a company [24]. This is a binary classification issue in which several classifiers are used for forecasting student acceptance rates. It is discovered that predicting a student's acceptance offer with a logistic regression classifier have an accuracy of 76.9%. Data mining approaches like Nave Bayes, K-Nearest Neighbor, Neural Network and Decision Tree are used to forecast student dropout. In order to examine the students admissions forecast rate, a genetic algorithm is utilized [25]. An effort is made to examine the answer for the strength of universities and colleges around the country [26]. They predicted that students' willingness to carry on their studies at colleges and universities would be a key element in the world's intellectual and social development. For the purpose of forecasting the admission choice of Davidson College students, machine learning is used and discovered that it is 86 percent accurate. The management faces a significant difficulty in attracting a large number of students to the institution [27].

The latest such studies have begun to place more emphasis on experimental analysis than on plain review. In analysis performed by Bajer et al. [28], six SMOTE-based algorithms are statistically related in aspects of the geometric mean achieved using three distinct classifiers and a huge number of disparate unbalanced datasets. An interesting conclusion of this analysis is that even a slight change in the synthetic instance creation method can alter the performance of the overall algorithm. More recent research is conducted by Kovacs [29], wherein they compared and evaluated 85 different minority oversampling strategies using 104 different unbalanced datasets and four different classifiers. In order to clean data noise prior to and subsequent to SMOTE procedure, the FRIPS-SMOTE-FRBPS study for example employed fuzzy as the prototype selection technique [30]. In the interim, the Rough Set Theory (RST) technique for reselecting every synthetic data sample created by SMOTE centered on similarity relations in an effort to remove noise is the research that coupled the SMOTE technique with data selection technique [31]. Additionally, SMOTE-IPF sought to combine an SMOTE-based oversampling approach with an iterative filtering method known as Iterative-Partitioning Filter (IPF) to exclude fake data samples that are deemed noise [32]. LN-SMOTE [33] that uses more specific data regarding the immediate area of analysed instances and selective oversampling in empowering SMOTE are the other research [34].

Sasikumar Iyer et.al [35] provides a complete run down of MOOCs student's dropout probability with the help of machine learning techniques. Moreover, we have highlighted a few answers being used to solve the dropout problem, provide an examination of the challenge of prediction models, and give some important overview and suggestions that may pave the way to develop useful ML solutions for overcoming the MOOCs dropout problem. Long Short-Term Memory neural network (LSTM) prediction model makes use of time-control units; the unit has the ability to model early learning behaviour with varying time intervals. Bonifro et.al [36] address this problem wherein we developed a tool that, by exploiting machine learning techniques, allows to predict the dropout of a first-year undergraduate student. The proposed tool allows estimating the risk of quitting an academic course, and it can be used either during the application phase or during the first year, since it selectively accounts for personal data, academic records from secondary school and also first year course credits.

The findings of the survey summarized about goal of defining appropriate learning approaches and better understanding students' needs. The researchers employed various methodologies like mathematical method, statistical models and machine learning algorithm as a decision making. The existing SMOTE challenges arisen from over fitting the minority class when performing random oversampling. Therefore, the proposed PCA-SMOTE algorithm to identify the exact reasoning of those students in the dropout interest status.

3. RESEARCH METHODOLOGY

3.1 Data capturing and preprocessing

Data have been captured by or0ganizing the survey for undergraduate student present in the university, wherein nearly 1,243 students have participated. This survey may specifically show the dropout status with gender specification as shown in figure 1. The "0" representation in gender is female and "1" as male as well as Dropout status also defined as "0" for not interested to dropout student and "1" for interested to dropout. According to the figure 1, the male students are high in both interests for dropout and not interest for dropout while compared to female students.

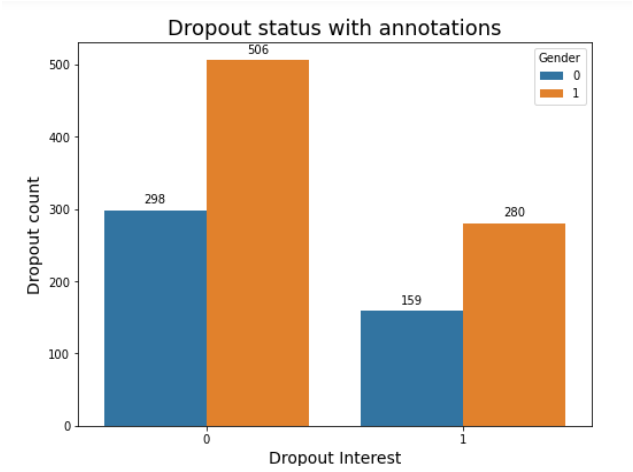

FIGURE 1 DROPOUT INTEREST STATUS COUNT FROM THE STUDENT OF EDUCATION INSTITUTE

However, the reasons for dropout are the essential key factors that affect the student's psychology classes namely higher school environment, staffs and administration of education institute, parents, acceptance, friend's relationship etc. Hence, this survey is involved with 31 reason questionnaires involved with above defined classes, whereas the questionnaires have been raised to their respective class staff in-charge and parents in order to determine an exact dropout status prediction. In addition, the 31 questions are the variables considered for dataset that is grouped into several classes like academic, socio-economic, parents background, staff and management etc. have been shown in table 1. Once the data is captured, then it need to be pre-processed by identifying the missing data and imputing an adequate value and transforms all values in term of numerical variables for improving the accuracy of prediction in accordance to the ML algorithm requirements. The 31 survey questions are named with Q1, Q2,…Q31 along with student ID, academic year, gender and dropout interest status have been shown in figure 2.

TABLE 1 FRAMING OF SURVEY QUESTIONS TO THE FOLLOWING VARIABLE CLASSES

Distribution of survey questions		
S.No	**Class of variable**	**Number of questions**
1	Academic	7
2	Sports related	1
3	Social integration	3
4	Parent's background	5
5	Travelling distance	1
6	Attitude	5
7	Higher school environment	2
8	Socio-economic	2
9	Harassment	2
10	Staffs and Management	2
11	Prior studies	1

Student ID	Academic Year	Gender	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Dropout I
20171001	2017-2018	Male	No	No	Yes	No	Yes	No	No	Yes	No	No	No	No	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	No	No	No	No	Yes	Yes	No	Yes	Yes	No
20171002	2017-2018	Male	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	No	No	No	Yes	No	No	Yes	Yes	Yes	No	No	No	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes
20171003	2017-2018	Male	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes
20171004	2017-2018	Female	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
20171005	2017-2018	Male	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	No	No	Yes	No	No	No	No	No	Yes	No	No	No	No	Yes	Yes	No	No	Yes	Yes	No
20171006	2017-2018	Male	Yes	Yes	Yes	No	Yes	No	No	No	No	Yes	No	No	No	No	No	No	No	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	No	No	Yes	No
20171007	2017-2018	Male	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	Yes	No	Yes	No	No	No	Yes	No
20171008	2017-2018	Female	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes
20171009	2017-2018	Male	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	No	No	No	Yes	No	Yes	Yes	No
20171010	2017-2018	Male	Yes	Yes	Yes	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No	Yes	No	No	No	No	Yes	No	Yes	No	Yes	Yes	No
20171011	2017-2018	Male	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No
20171012	2017-2018	Female	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No	No	No	Yes	No	Yes	No	No	No	Yes	Yes	No	No	No	No	No	No	Yes	Yes	No	Yes	Yes	No
20171013	2017-2018	Female	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	No	No	No	No	Yes	No	No	No	No	No	No	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No
20171014	2017-2018	Male	No	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No	No	Yes	No	No	No	No	No	Yes	Yes	No	No	Yes	No	No	No	Yes	Yes	Yes	No	Yes	No
20171015	2017-2018	Female	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	No	Yes	No	No	No	No	No	Yes	No	No	No	No	Yes	Yes	No	No	Yes	Yes	No
20171016	2017-2018	Female	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	Yes	Yes	No	No	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	No	No	Yes	No
20171017	2017-2018	Female	Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No
20171018	2017-2018	Female	Yes	Yes	No	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	No	No	No	No	No
20171019	2017-2018	Male	No	No	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes
20171020	2017-2018	Male	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes
20171021	2017-2018	Male	Yes	No	Yes	No	Yes	No	No	Yes	No	No	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes
20171022	2017-2018	Male	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	Yes	No	No	No	No	No
20171023	2017-2018	Male	No	No	Yes	No	Yes	No	No	Yes	No	No	No	No	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	No	No	No	No	Yes	Yes	No	Yes	Yes	No
20171024	2017-2018	Male	Yes	Yes	No	Yes	Yes	No	No	No	No	No	Yes	No	No	No	Yes	No	No	Yes	Yes	Yes	No	No	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes

FIGURE 2 ORIGINAL DATASET OF SURVEY ABOUT DROPOUT INTEREST

3.2 Workflow of PCA-SMOTE

In general, answering for all the questions are mandatory and all set of questions needs to be filled from students, staff in-charge as well as parents. However, all reasons for student dropout are essential to understand the respective student difficulties in deviating from the concentration. Therefore, the data mining algorithms play an important role in dimensionality reduction and even balancing the oversampling data from the survey dataset. Thus, the paper focus on PCA-SMOTE algorithm to do dimension reduction of feature extraction and signify the obtained PCA factors for determining an accurate prediction of dropout planning students earlier. The working flow of PCA-SMOTE is shown in figure 3. The PCA has permitted to reduce the redundant attribute of questions and the respective individual question attribute has different dimensions. The goal of PCA is to scale the attribute and transform the core questionnaires attribute into new attribute sets that explain the data variance. As a result, the new characteristics are linked to a linear combination known as principle components. As a result of PCA, the attribute dimensionality has been decreased and visualization has been provided.

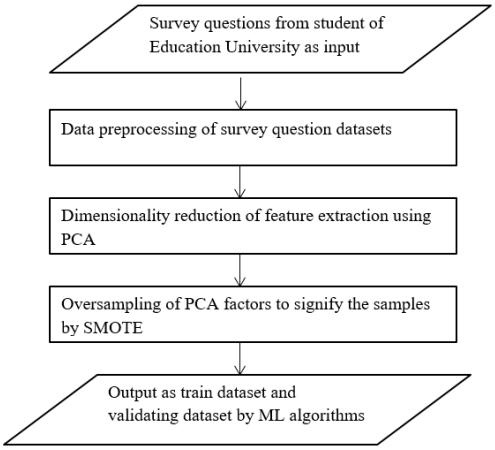

FIGURE 3 WORK FLOW FOR PROPOSED PCA-SMOTE FOR DROPOUT STUDENT PREDICTION

When the survey is conducted to the student, staffs and parents with all probable reasonable questions are explained, the value of maximum behavioural has been captured. Moreover, a dataset have several attributes based on the various subsequent conditions is mentioned as follow.

- Identifying the correlation of various attributes that may lead to multicollinearity.
- Possibility of confusion in decision making of feature selection that may be adequate for classifying technique.

- Selection of an adequate strategy in identifying the suitable attribute for predicting the undergraduate dropout students.

PCA has detected the correlation among variables of classes from the questionnaires. When the correlation is analysed to be strong among variables exists, the algorithm has endeavour in reducing the dimensionality in order to make the dimensionality meaningful for better accuracy.

Figure 4 illustrate that 75% - 80% of dropout reasoning is identified by required principle components. According to the elbow curve defining, it is observed that 82.52% of dropout information is identified with the available 1st sixteen components. Therefore, 17 components have been selected as principle components. Moreover, the minority in sampling of PCA features may endure in less accuracy in predicting student's dropout due to various correlations among variable of classes. Thus, the procedure of SMOTE is introduced for refining of oversampling PCA factors sample to signify the samples as train dataset for better earlier prediction of students dropout.

In addition, the classifier in minority class performance has been improved by SMOTE that seek in developing its capability of generalization, eliminating the probable over fitting occurrence. These are the various reasons involved in the mechanism of synthetic instance creation that assist in generating new minority classes that are not getting duplicated with existing classes, but it may be located in their neighbourhood.

Let the each minority class is mentioned as x in M and the k-neighbourhood is decided initially, whereas the several metric functions have been implemented using Euclidean distance. The subsequent generation of q

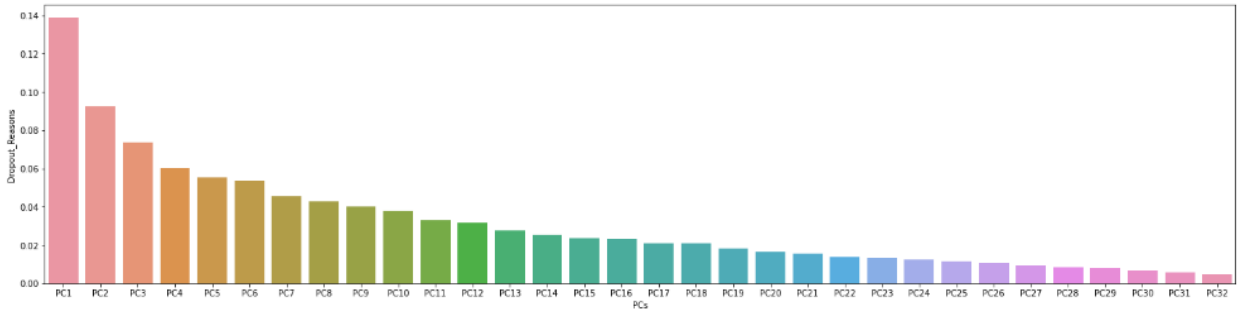

FIGURE 4 SELECTIONS OF PCA COMPONENTS BY ELBOW CURVE DIAGRAM

synthetic instances are created in term of convex combination that is expressed in equation (1).

$$y^i := x + U_i(0,1) * \left(\hat{x}^{r(i)} - x\right), \qquad i = 1,2,\ldots n \quad (1)$$

Where,
x = line among the minority class
$\hat{x}^{r(i)}$ = Random neighbor oversampling from the k-neighbourhood
$U_i(0,1)$ = Uniform distribution

The generated class is located on the line among the minority class (x) along with its k-neighbourhood ($\hat{x}^{r(i)}$) and the accurate position is defined by generation of random number using uniform distribution. Thus, the working of SMOTE to predict the student dropout is outlined in the given algorithm.

SMOTE algorithm

Let P_c be positive class set and N_c be negative class set, whereas "n" represents number of variables, k represent k-neighbourhood for SMOTE oversampling and f represent the oversampling factors.

Step 1: Initialize and portioning of N_c and computing $n_c := (f+1)|P_c|$ and $\{N_{c1}, N_{c2}, \ldots N_{cn}\} := Do.partition\ (N, n)$.

Step 2: The iteration $i := 1$, while $i \leq n$, then SMOTE oversampling is generated by computing $S := SMOTE(P, k, f)$, negative class for under sampling variables is defined as $N'_c := undersample(N_{ci}, n_c)$.

Step 3: Similarly, the PCA factors are rearranged based on computed train dataset sample is computed as $T \coloneqq P_c \cup S \cup N_c'$.
Step 4: The train dataset sample is then implemented in the ML algorithm with train dataset as well as validating dataset. The train dataset is computed as $M_i = ML(T)$ and $i = i + 1$.
Step 5: When i is greater than n, the while loop stop and return.

Therefore, the output of the ML set models is defined as $M \coloneqq \{M_1, M_2, \dots, M_n\}$. Thus, the PCA-SMOTE algorithm assists in improving the accuracy of student dropout prediction in the education university. The PCA-SMOTE algorithm is implemented with various ML algorithms and the performances of those algorithms are evaluated through confusion matrix parameters.

4. Result and Discussion

The evaluation of dropout student prediction using PCA-SMOTE algorithm in various ML models such as Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Logistic Regression (LR) and also with existing LR and KNN without PCA-SMOTE algorithm. This experiment is carried out using Python, which has numerous essential libraries such as scipy, numpy, scikit-learn and tensorflow, wherein the programmers use to construct data mining techniques. Built-in categorization and decision-making features are provided in these libraries. The sample count taken from the dataset describes 804 for training and 359 for testing. The confusion matrix parameters such as accuracy, precision, recall, sensitivity and specificity by the respective ML models are deal with the values of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The PCA-SMOTE algorithm with KNN has generated better accuracy with the actual and prediction confusion matrix table as shown in figure 5 and the several SVM, LDA, LR with PCA-SMOTE and without PCA-SMOTE values are listed in table 2.

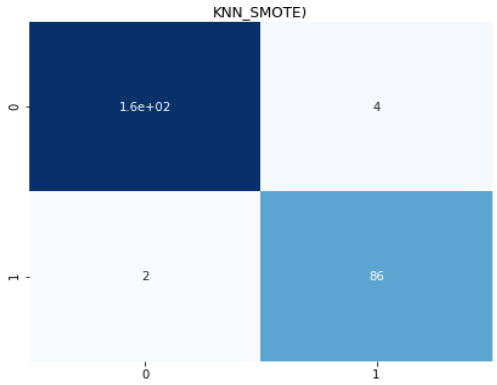

Figure 5 Confusion matrix of KNN_SMOTE

Table 2 Confusion matrix value for experimented ML models

ML models	TP	TN	FP	FN
LR	159	78	10	78
LDA	147	154	1	20
KNN	147	166	5	4
LR_SMOTE	148	161	4	9
LDA_SMOTE	151	154	1	16
KNN_SMOTE	157	86	4	2
SVM_SMOTE	146	157	6	13

According to this paper, the prediction accuracy of student dropout interest status is analysed based upon the reasons available from the 31 questions that are weighted and signified by PCA-SMOTE. This assist in providing both feature selection and extraction of exact PCA factor use of SMOTE from the over fitting sample. The evaluation of the PCA-SMOTE from various ML model is done by considering both with PCA-SMOTE and without PCA-SMOTE, whereas the PCA-SMOTE algorithm considered ML model is defined as ML_SMOTE (Eg. LR_SMOTE) and without is indicated as usual ML model name as shown in table 2. The confusion metric parameters are listed as follows in equation (2), (3) and (4).

Accuracy

The measure used to define the amount of exact prediction of student dropout from the complete dataset sample.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \dots\dots\dots(2)$$

Precision

This is the measure for correctness present in the positive prediction of student dropout that illustrates positive

prediction of student dropout as resulted in the actually positive.

$$Precision = \frac{TP}{TP+FP} \quad \text{…….(3)}$$

Recall

This is the measure of true positive that get predicted from all positive present in the student dropout dataset. This is also named as true positive rate.

$$Recall = \frac{TP}{TP+FN} \quad \text{…….(4)}$$

Similarly, the true positive rate is signified by the sensitivity and the false positive rate is signified by specificity, whereas student dropout prediction related to PCA-SMOTE algorithm perform better accuracy is shown in the table 3.

TABLE 3 CONFUSION MATRIC PARAMETER VALUES FOR VARIOUS ML MODELS

ML models	Accuracy	Precision	Recall	Sensitivity	Specificity
LR	0.9518	0.9876	0.9408	0.9408	0.9750
LDA	0.9348	0.9932	0.8802	0.8802	0.9935
KNN	0.9720	0.9671	0.9735	0.9735	0.9708
LR_SMOTE	0.9596	0.9737	0.9427	0.9427	0.9758
LDA_SMOTE	0.9472	0.9934	0.9042	0.9042	0.9935
KNN_SMOTE	0.9759	0.9752	0.9874	0.9874	0.9556
SVM_SMOTE	0.9410	0.9605	0.9182	0.9182	0.9632

Table 3 has illustrated the performance of PCA-SMOTE algorithm in ML model to perform the better prediction in student's dropout. In general, the accuracy of PCA-SMOTE algorithm related ML models like LR_SMOTE, LDA_SMOTE and KNN_SMOTE is comparatively better than respective traditional LR, LDA and KNN model. Moreover, the PCA-SMOTE algorithm models generate better accuracy in predicting dropout interest of students in the educational institution. Hence, each model with PCA-SMOTE algorithm and without PCA-SMOTE with ML models is analysed and is shown in figure 6.

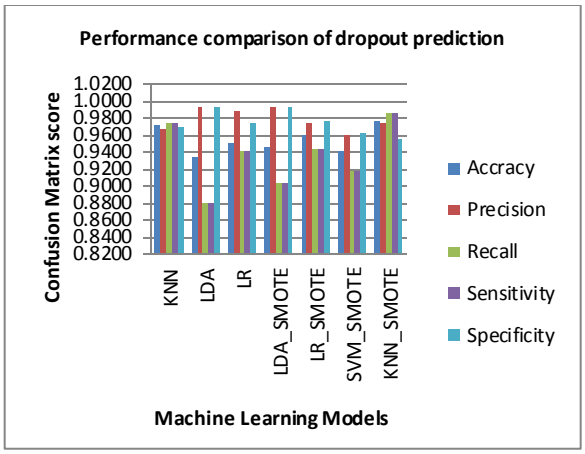

FIGURE 6 PERFORMANCE COMPARISON OF DROPOUT STUDENT PREDICTION

Figure 6 has illustrated that KNN_SMOTE has high accuracy as 97.6% while compared to other ML model in predicting that the student dropout from conducted survey. Similarly, the sensitivity and recall of KNN_SMOTE is high as 0.987 when compared with other PCA-SMOTE ML model as well as without PCA-SMOTE ML models. In addition, the specificity of KNN_SMOTE is 0.956 which is lowest than other ML models considered in this experimental explains that false positive prediction of student's dropout using PCA-SMOTE with KNN model is better while compared to other PCA-SMOTE ML model as well as without PCA-SMOTE ML models.

5. CONCLUSION

This paper is mainly focused on defining the probabilities in utilizing the survey conducted with framed 31 questions and it gets framed based on classes of variable. The variables are initially feature selected and extracted using PCA and the respective PCA factors with over fitting samples have been refined and signifying the class of variable based on minority class using SMOTE algorithm is the key for proposed PCA-SMOTE algorithm. The proposed algorithm with ML model accuracy is comparatively higher than traditional ML models. However, the SMOTE associated ML models are analysed individually has shown that KNN_SMOTE have performed at higher accuracy as 97.6% in predicting student dropout. Hence, the PCA-SMOTE algorithm with KNN model has performed better in predicting the exact reasoning of the student for the dropout. Thus, it assist for the education university to give counselling to the respective students and also concentrate as well as motivate the students to bring him out from the thought of dropout scenario. In future, the research work has

aimed to adopt collaborative learning to the students who are planned for dropping the education institute. This research may successfully minimize the count of student dropout from the educational institution or university.

REFERENCES

[1] J. Bayer, H. Bydzovská, J. Géryk, T. Obsivac, and L. Popelínský, "Predicting drop-out from social behaviour of students," Proc. 5th Int. Conf. Educ. Data Min., no. Dm, pp. 103–109, 2012.

[2] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting students performance in educational data mining," Int. Symp. Educ. Technol. ISET 2015, pp. 125–128, 2016.

[3] A. Pradeep, S. Das, and J. J. Kizhekkethottam, "Students dropout factor prediction using EDM techniques," Proc. IEEE Int. Conf. Soft-Computing Netw. Secur. ICSNS 2015, 2015.

[4] Z.J.Kovacic, "Early prediction of student success:Mining student enrollment data" pp.647-665,2010.

[5] A. M. Shahiri, W. Husain and N. A. Rashid, "A review on predicting student's performance using Data mining techniques," procedia computer science, vol.72, pp.414-422, 2015.

[6] G. S. Abu-Oda and A. M. El-Halees, "Data mining in higher education: University student dropout case study," International Journal of Data Mining & Knowledge Management Process (IJDKP), vol.5, no.1 pp. 97-106, 2015.

[7] S. Sultana, S.Khan and M. A. Abbas, "Predicting performance of Electrical Engineering students using cognitive and non-cognitive features for identification of potential dropouts," International Journal of Electrical Engineering Education, vol. 54, no.2, pp. 105-118, 2017.

[8] L. Bonaldo, and L. N. Pereira, "Dropout: Demographic profile of Brazilian university students," Procedia-Social and Behavioral Sciences, vol. 228, pp. 138-143, 2016.

[9] V. Hegde and M. S. Pallavi, "Descriptive analytical approach to analyze the student performance by comparative study using Z score factor through R language," 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, 2015, pp. 1-4.

[10] N. Japkowicz, M. Shah, "Evaluating Learning Algorithms: A Classification Perspective", 1st Ed., Cambridge University Press, 2011.

[11] S. Fotouhi, S. Asadi, M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data", Journal of Biomedical Information, Vol. 90, 2019, pp. 103089.

[12] P. Skryjomski, B. Krawczyk, "Influence of minority class instance types on SMOTE imbalanced data oversampling", Proceedings of the 1st International Workshop on Learning with Imbalanced Domains: Theory and Applications, Skopje, Macedonia, 22 September 2017, Vol. 74, pp. 7-21.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Research, Vol. 16, 2002, pp. 321-357.

[14] A. Fernandez, S. Garcia, F. Herrera, N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary", Journal of Artificial Intelligence Research Vol. 61, 2018, pp. 863-905.

[15] A. K. Jain and C. K. Jha, "Dropout classification through discriminant function Analysis: A statistical approach," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 2, no. 4, pp. 572-577, 2017.

[16] A. Katare and S. Dubey, "A comparative study of Classification Algorithm in EDM using 2 Level Classification for predicting students' performance," International Journal of Computer Applications, vol.165, no.9, pp. 35-40, 2017.

[17] C. Marquez-Vera, A. Cano, C. Romero, A. Y.M. Noaman, H. M. Fardoun and S.Ventura, "Early Dropout Prediction using Data Mining: A case Study with High School Students, "Expert System Journal, vol.33, no. 1, pp. 107-124, 2016.

[18] A. Cano, A. Zafra, S. Ventura, "An Interpretable Classification rule mining algorithm," Information Sciences, vol. 240, pp. 1-20, 2013.

[19] W. Yu, T. C. Lin, Y. C. Chen, and D. Kaufman, Determinants and Probability Prediction of College Student Retention: New Evidence from the Probit Model, vol. 3, no. 3. 2012.

[20] Tian, C., Wang, Y., Lin, X., Lin, J., & Hong, J. (2016), Research on High-Dimensional Data Reduction, International Journal of Database Theory and Application, 9(1), 87-96.

[21] Pervez, M. S., & Farid, D. M. (2015), Literature Review of Feature Selection for Mining Tasks, International Journal of Computer Applications, 116(21).

[22] Porkodi, R. (2014), comparison of filter based feature selection algorithms: An overview, international journal of innovative research in technology& science, 2(2), 108-113.

[23] Sorzano, C. O. S., Vargas, J., & Montano, A. P. (2014), A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877.

[24] Kanadpriya Basu, Treena Basu, Ron Buckmire, and Nishu Lal, "Predictive Models of Student College Commitment Decisions Using Machine Learning", Data., vol. 4, no. 2, 2019, pp. 65.

[25] E. Yukselturk, S. Ozekes, and Y. K. Türel, "Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program", European Journal of Open Distance E-Learning., vol. 17, 2014, pp.118–133.

[26] Joseph Janison, "Applying Machine Learning to Predict Davidson College's Admissions Yield", proceedings of the ACM SIGCSE Technical Symposium., 2017.

[27] S. Maldonado, G. Armelini, and A. Guevara, "Assessing university enrollment and admission efforts via hierarchical classification and feature selection", Intelligent Data Analysis., vol. 21, no. 4, 2017.

[28] D. Bajer, B. Zorić, M. Dudjak, G. Martinović, "Performance Analysis of SMOTE-based Oversampling Techniques When Dealing with Data Imbalance", Proceedings of the 26th International Conference on Systems, Signals and Image Processing, Osijek, Croatia, 5-7 June 2019, pp. 265-271.

[29] G. Kovacs, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets", Applied Soft Computing, Vol. 83, 2019, pp. 105662.

[30] Verbiest, N., Ramentol, E., Cornelis, C., Herrera, F., 2014. Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. Appl. Soft Comput. J. 22, 511–517.

[31] Ramentol, E., Caballero, Y., Bello, R., Herrera, F., 2012. SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and under sampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowl. Inf. Syst. 33, 245–265.

[32] Sáez, J.A.A., Luengo, J., Stefanowski, J., Herrera, F., 2015. SMOTE-IPF: addressing the noisy and borderline examples problem in imbalanced classification by a resampling method with filtering. Inf. Sci. 291, 184–203.

[33] MacIejewski, T., Stefanowski, J., 2011. Local neighbourhood extension of SMOTE for mining imbalanced data. IEEE SSCI 2011: Symposium Series on Computational Intelligence – CIDM 2011: 2011 IEEE Symposium on Computational Intelligence and Data Mining 104–111.

[34] Nnamoko, N., Korkontzelos, I., 2020, Efficient treatment of outliers and class imbalance for diabetes prediction, Artif. Intell. Med. 104.

[35] Vivek Sasikumar Iyer, Siddharth Chaudhury and Dr.M.Shobana , "Mooc Student Dropout Prediction Using Machine Learning Algorithms", International Journal of Engineering Research in Computer Science and Engineering, Vol 8, Issue 5, May 2021.

[36] Francesca del Bonifro, Maurizio Gabbrielli, Giuseppe Lisanti and Stefano Zingaro, "Student Dropout Prediction", Artificial Intelligence in Education, pp.129-140, 2020.