iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54210

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Educational Data Mining using ML

Mudrakola Bhavani[1], Podila Mounika[2]
[1]*Assistant Professor of the Department of Information Technology*
[2]*Assistant Professor of the Department of Computer Science and Technology*

Abstract: ***The ability to forecast students' performance is one of the most useful and important academic issues in the world today because of the development of technology. In the subject of education, data mining is incredibly useful, particularly for analysing student performance. The imbalanced datasets in this subject have made it extremely difficult to estimate students' performance, and there is no comparison of the various resampling techniques. This study compares multiple resampling procedures to manage the unbalanced information problem when projecting student performance of two distinctive datasets, including Borderline SMOTE, SMOTE-ENN, SMOTE, Random Over Sampler, SVM-SMOTE, and SMOTE-Tomek. Additionally, the dissimilarity between binary classification and multiclass, as well as the features' structures, are looked at. must be able to evaluate the effectiveness.***

I. INTRODUCTION

A lot of information has been gathered recently due to developments in many disciplines. Mining procedures are generally used to cite valued information from the data because analysing large amounts of data to extract required information is a difficult task for humans. It is common knowledge that universities operate in a complicated and competitive environment. The major task for universities is to thoroughly assess their performance, pinpoint their distinctiveness, and devise strategies for future growth and success. The educational system is aware of the potential for data mining to significantly enhance its performance. The use of data mining techniques for the analysis of available data in educational institutions is identified as educational data mining. Despite the fact that knowledge discovery is facilitated by data mining.

II. DRAWBACKS

A. SMOTE-ENN

This technique is one of the recognized ones which enhances the results by combining the SMOTE as an over-sampling model and the ENN as an under sampling model.

B. SMOTE-Tomek

Another popular hybrid technique, SMOTE is connected to Tomek connections as an under-sampling model in order to improve the findings.

C. Shuffle 5-fold cross-Validation

The retention method was repeated five times in this study using mixed 5-fold cross validation to divide data into 5 groups. Each time, out of five subdivisions one is used as the test set and the remaining additional four as the training set. Table 10 displays the variance and average accuracy of the machine learning models used to implement this validation method. Because this method works, the consequences of random 5-fold cross-validation are reliable and adequate. The findings showed that the accuracy of some models slightly improved when inconsistent data were resolved. The results of data matching using SVM-SMOTE are better than other data.

D. Imbalanced Data Problem

Several real-world data with unequal class distribution suffer from data inconsistency issues. It's significant to the memo that utmost machine learning models effort finest while there are approximately equal numbers of trials from each class. The mainstream class dominates the marginal class as a result of the unbalanced data problem; as a result, the classifiers are further likely to favour the majority class, and their performance isn’t trusted. The grade point average of the four group students is not equal, as shown by an analysis of the introduced datasets. In reality, the dataset of Iran has additional trials from the Intermediate and Decent (each 40%) classifications than from the further dual classes (has only the Poor class)

A lot of research has been done in the field of education. A neural network (ANN) model was developed in 2008 to envisage the performance of 1,407 students. The hold-out method is the most widely utilized cross-validation method, and is used to train and test the preparation process. It has to be noted that the ANN approach has been used as a predictive model in previous studies. In 2015, two models of the ANN algorithm were developed

III. LITERATURE

A. *Educational data mining: study from 1995 to 2005*

Expectedly to the progress of the data science education community, now there is a growing significance in data mining and education. This study explores how data mining is used in traditional schools, specialized online courses, effective content management practices, and flexible and intelligent methods. Each of these systems has specific information and learning objectives. After preprocessing the data available in each case, the following data can be used: statistics, visualization, distribution, distribution and analysis of differences; association principle mining, model mining and text mining. For data mining to be successful, more effort is required before it reaches maturity..

B. *Predictive Data Mining Approaches in Medical Diagnosis: An Assessment of Some Diseases Prediction*

A huge amount of information has been gathered about the many things that need to be done due to the continuous development of technology in all sectors. Data mining is the method of finding and analysing confidential data from multiple views to attain information. One of the many applications of data is just diagnostics. Many diseases are now considered deadly and dangerous. The most common causes of death are diabetes, heart disease and breast cancer. This study reviewed 168 articles discussing the use of data mining to diagnose these diseases. This study focuses on 85 documents selected for further attention between 1997 and 2018. Each algorithm, model, and evaluation process for the document is only reviewed.

C. *Student Performance Prediction By Using Data Mining Classification Algorithms*

The quality of education a school provides to its students can be used to measure the success of the school. The highest level of education is achieved through an accurate study of student performance. The lack of an appropriate framework for assessing student performance and development is still a problem today. This usually happens for two reasons. First, it is difficult to predict student achievement using existing methods. Secondly, due to the neglect of many important factors that affect students' grades. Predicting student performance is a more difficult task due to the large amount of information in the curriculum. These teaching methods can help better predict a child's performance. These fit well.

D. *Data Mining Approach to Expecting the Performance of first-year Student in a University by the Admission Requirements,"*

A student's academic success in school is affected by many non-academic and academic factors. While students who have previously failed at family involvement may succeed in college by staying away from home, students who have previously succeeded in secondary education at home will be disappointed for social and lifestyle reasons. In Nigeria, university admissions are usually determined by the acceptance rates of students, which are often academic in nature and may not translate into quality once the student has started school. This study used six variables to analyze the relationship between perception of acceptance and students' first-year academic presentation using GPA and grade level.

E. *Educational Data Mining: Predictive Analysis of the Academic Performance of Public school Students in the Capital of Brazil,*

There is a lot of educational information available, but it is necessary to do well and analyze the information that can be used to improve the educational process. Data mining classification and clustering algorithms have the potential to help businesses. In this study, we use various classification and integration methods to learn data using WEKA. The findings show that this algorithm can accurately predict what students will do after training; however, it is important to use the right algorithm. We work with academics and administrators to use a variety of data analytics to monitor and predict student performance and use this data as a starting point. In education, we create rules and regulations to improve student learning. Education Information, K-Nearest and Education Content.

F. *Proposed Work*

The retention method, one of the most widely used cross validation methods, is used to train and test input algorithms. It should be distinguished as the ANN approach has been used as a predictive model in previous studies. In 2015, two prototypes were developed based on the neural network algorithm.

The outcomes of the study showed that the neural network algorithm was able to predict 95% of student performance and showed the predictive power of the model. In addition, the total correct score for the ANN model test is 84.6% indicates the model's ability to predict student performance. No doubt further machine learning models have been advanced as well. A pure Bayesian model was developed.

G. Educational Data Mining

The application of data mining procedures to analyse data presented to schools is called educational data mining (EDM). Data mining facilitates data sighting, while algorithms make available the necessary tools. The ability to predict student performance with a high degree of accuracy is useful as it allows early detection of failing students. Learning data mining enables learning organizations to gain a deeper understanding of their learning processes by analysing relevant learning data. In fact, predicting a student's academic performance is important for ensuring student success, but this can be difficult as there are many factors that influence student performance. Researchers have recently developed new methods to study data mining.

H. Advantages

1) In this way, samples and samples close to the borderline will be separated more than samples far away from the borderline.
2) The techniques use a number of nearest neighbors for each subsample to split the subsample into clusters: safe, dangerous, and noisy.
3) Note danger group is applied to create new events.

IV. RESULTS AND DISCUSSION

A. Student Dataset

We require a student dataset with several variables in order to predict student success. The students' information is shown in the figure below with various features. Every piece of information about a pupil can be used to forecast performance. Therefore, the ability to forecast student performance with a high degree of accuracy is helpful since it allows for the early identification of individuals who perform poorly academically. It speeds up kids' academic advancement.

					Matriculation Scores					Junior High School Grades				
Number	Student ID	Gender	IQ	Interest	Math	Physic	English	Eco	AVG Matriculation	Math	Science	Social	English	AVG Grades
001	1310A01	M	116	MNS	40	47	52	40	44.75	94.8	89.0	92.4	93.2	92.35
002	1310A02	M	131	MNS	80	68	76	70	73.50	93.8	90.2	94.0	88.8	91.70
003	1310A03	F	109	MNS	84	75	52	70	70.25	90.8	86.2	91.4	90.2	89.65
004	1310A07	F	124	MNS	76	63	56	50	61.25	84.0	82.0	81.2	81.2	82.10
005	1310A08	M	131	MNS	60	60	68	70	64.50	93.4	92.8	89.6	91.6	91.85
006	1310S41	M	109	SS	80	67	60	60	66.75	82.6	80.4	85.0	83.0	82.75
007	1310S42	M	101	SS	20	51	28	50	37.25	80.2	82.6	81.6	86.8	82.80
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
875	1310S62	P	97	SS	28	30	45	43	36.50	73.5	77.8	78.2	74.4	75.97

Fig 1: Student dataset

B. Pre-processing

Real-world data characteristically contains noise, and missing values, may be in an unwanted format, making it incredible to build machine learning models on it directly. To clean the data and make a machine learning model Data preprocessing is essential, which also advances the model's accuracy and efficiency. Pre-processing the student dataset and setting it up in a structured style is shown in figure 5 below. Here, a score of 0 indicates a graduate and a score of 1 indicates a non-graduate who made less academic progress. Figure 6 displays the informational graphs for each student.

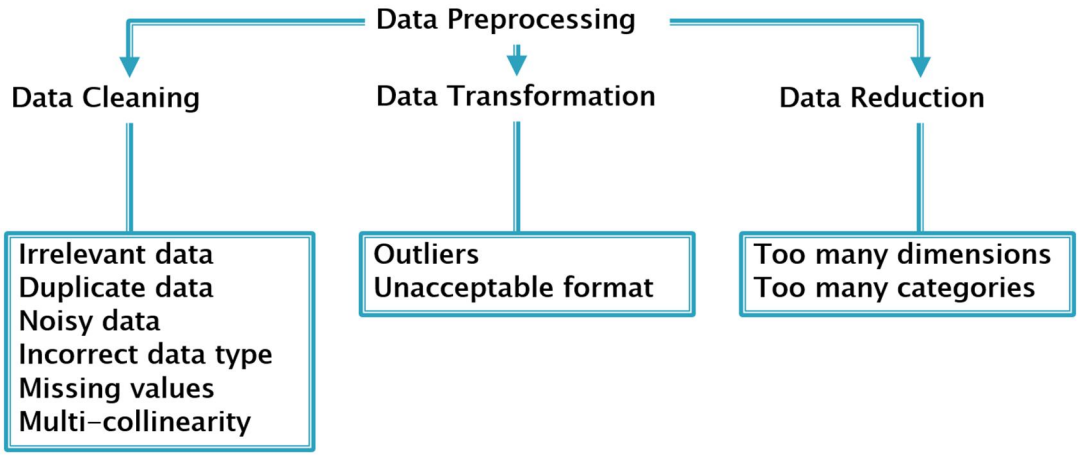


Fig 2: Pre-processing the dataset

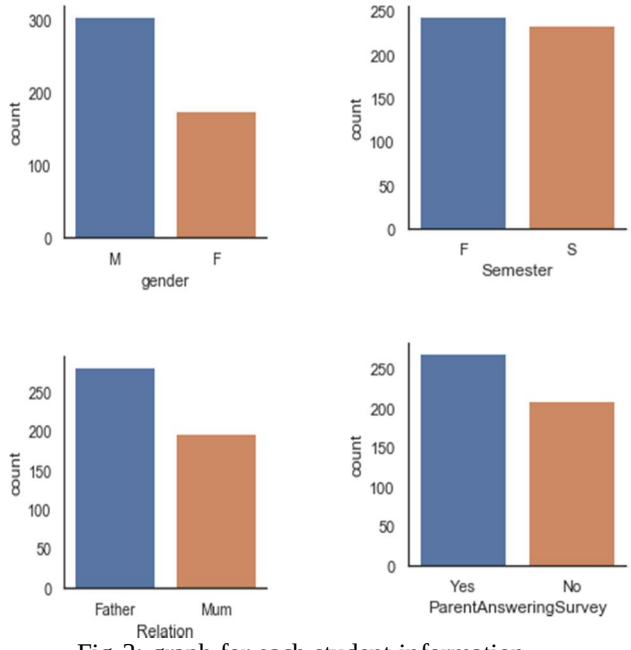


Fig 3: graph for each student information

C. Train and Test

The train/test method is a mode to scale by what means exact your model is. Since the split data set is made into two sets—a training set and a testing set—this technique is known as train/test.

Utilizing the training set, you train the model.

Using the testing set, you test the model.

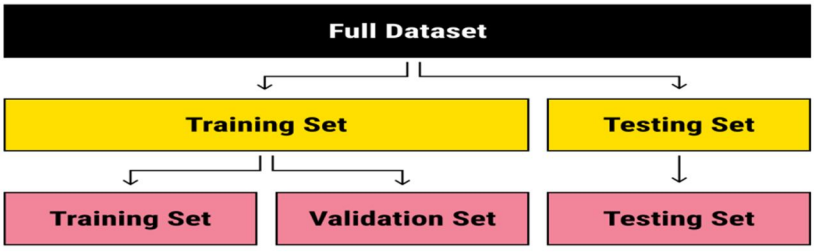


Fig 4: Train and test the dataset

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

D. Cross-validation

Training and testing the model on a subset of the input data that hasn't been used before, This confirm the model's efficiency. The subject is also a technique for defining by what means well a numerical model generalises to a changed dataset.

The stability of model is essential constantly be verified in machine learning. This specifies that we cannot fit our model to the training dataset unaccompanied. We set separately a specific example of the dataset one that was not involved in the training dataset for this use. After that, before distribution, we need to assess our model on that model, and the entire process is stated to as cross-validation. It varies after the distinctive train-test divide in this way.

The elementary stages of cross-validations are:

1) Depute a subgroup of the dataset as a validation set.
2) Offer the training to the model with the training dataset.
3) Now, evaluate model performance using the validation set. If the model performs well withthe validation set, perform the further step, else check for the issues.

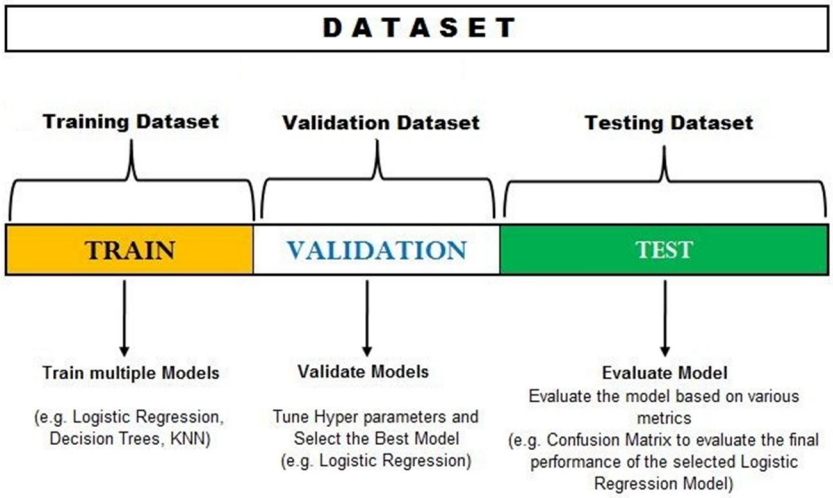


Fig 5: Applying Cross-validation on dataset

E. Linear Regression

Supervised learning in machine learning algorithms is linear regression. It performs a regression process. To model a goal estimate value Regression utilizes independent variables. It is generally used to regulate in what manner variables and forecasting relay to each other. The kind of association they take into interpretation between the independent and dependent variables, and various factors are the Regression models which differ according to the numeral of independent variables they custom.

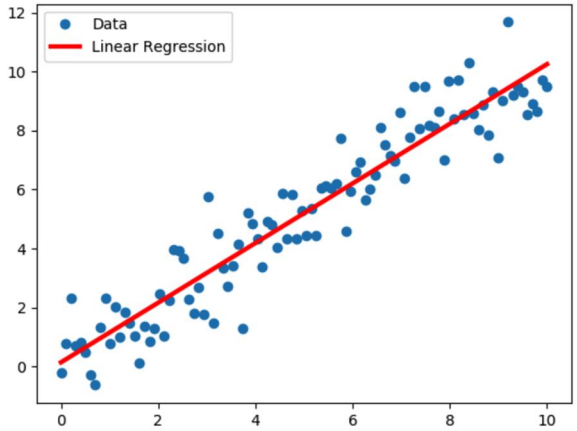


Fig 6: Applying Linear regression

F. Kappa Squared Error

Several of the simulations, the kappa coefficient, which extravagances missing ratings as a regular category, looks like somewhat biased and has a significant mean squared error. If it can be presumed that missingness is totally at random or not at random, we advise using the kappa coefficient that is created on the listwise deletion of misplaced ratings because it works well and is simple to compute.

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Fig 7: Calculating Kappa squared error

G. Mean Squared Error

The average square variance amongst the true values and the estimated value is evaluated by an mean squared deviation (MSD) or estimator mean squared error (MSE). The value that is evaluated is not a negative value for any time. So preferably the values will be zero or positive. The evaluated values of variance and the bias are included in the MSE,it is the second moment of the error (about the origin).

H. Classifiers

- Decision-Tree

```
(839, 32)
R-Squared Error  is 0.35714285714285715
Full R2 score is 0.9761904761904762
```

- Linear regression

```
R-Squared Error is 0.5479510761938694
Full R2 score is 0.9755517011910761
```

- Random Forest

```
R-Squared Error is 0.5502344729937918
Full R2 score is 0.9748286358873268
```

- SVM

```
(839, 32)
R-Squared Error  is 0.4583333333333333
Full R2 score is 0.9702380952380952
```

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Predicting class: with a finalized machine learning model in the scikit-learn Python library, we control accurately by what means we can make regression estimates and the classification

Later finalizing this discussion, you resolve identify:

- To decide on a model in method to make it equipped for making predictions.
- To create probability and class estimates in sci-kit-learn.
- To create regression estimates in sci-kit-learn.

I. Multivariate Regression

A multivariate regression uses several data variables for analysis It is made up of many independent variables and a single dependent variable, which is an enlargement of multiple regressions. We effort to forecast the outcome based on the numeral of independent variables. Multivariate regression aspects of a method designate how numerous variables respond simultaneously to deviations in further variables.

J. Support Vector Regression

Support Vector Regression, a regression algorithm justifying its name, allows non-linear and also linear regressions. This tactic operates as the basis of the Support Vector Machine. SVR is a regressor used to estimate continuous well-ordered variables, In contrast to SVM, which is operated to foresee discrete categorical labels. This is a manner in which SVR and SVM diverge from each other.

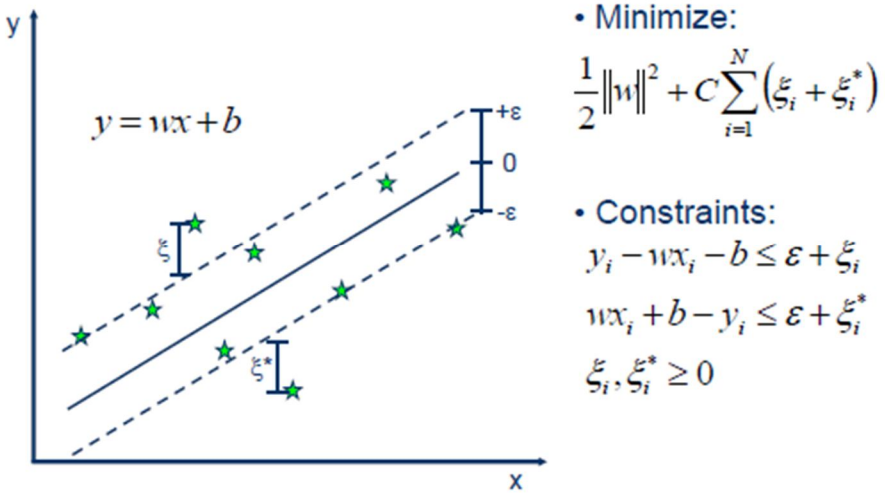


Fig 11: Support vector regression

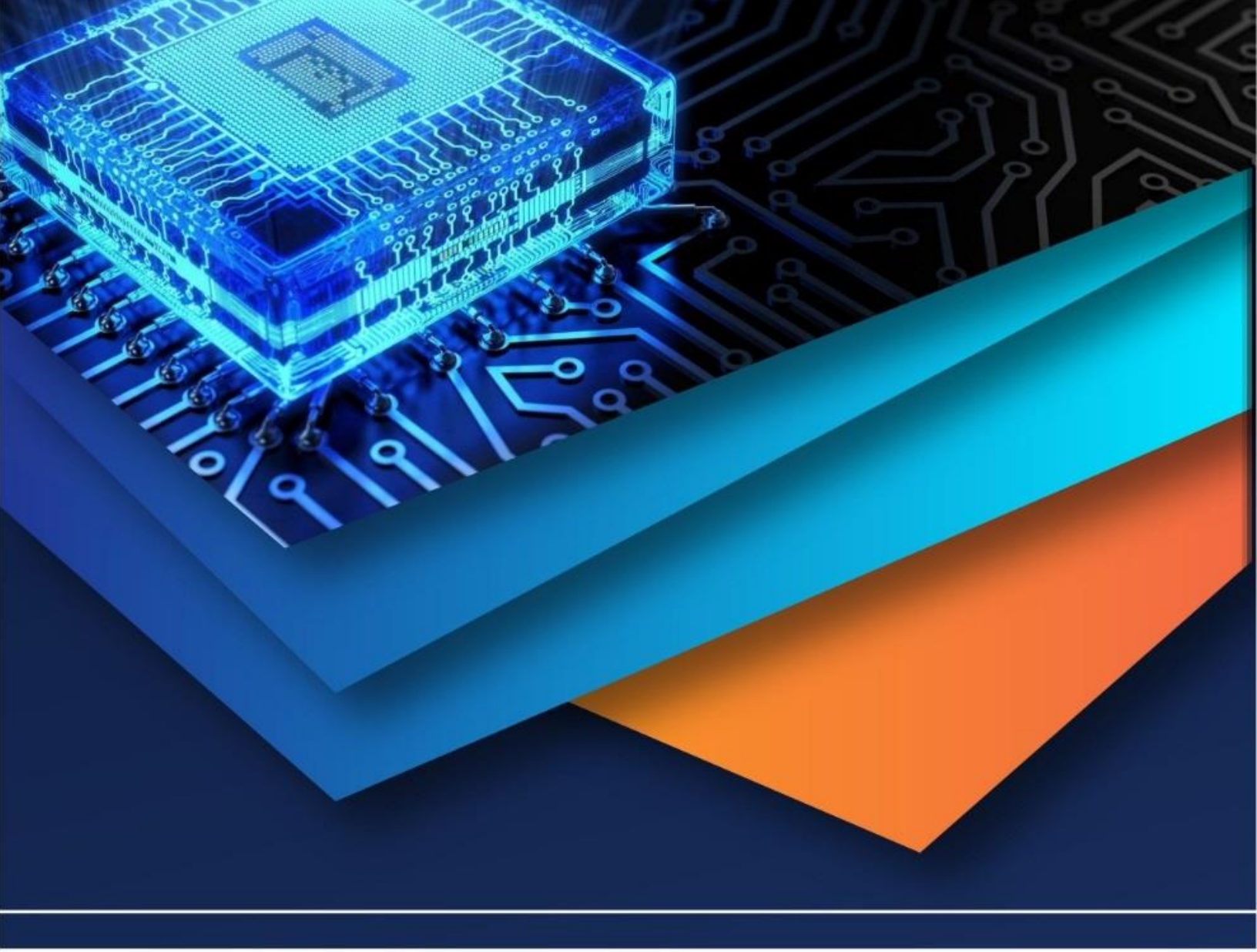
V. CONCLUSION

Present developments in recent trends and technologies are the reasons to collect the huge volumes of data. The existing methodology followed in many organizations mainly in the schools is to collect various related information about the students , from the information or related data that is collected is used for analysing various individual performance in detail. Research data mining is a powerful and logical method which is used to evaluate and analyse the necessary information or data which is relevant, The same method of analysing is necessary in the organization to evaluate student performance from academic data. Sometimes while considering the required information may also be not included, here the insufficient data may encounter a risk in identifying students performance from academic data.

The study of gives an impact of the inconsistent data and the utilization of the known resampling method between the different methods which are available such as BorderLine SMOTE. Two different datasets are used in the students performance, the including the variation between multiclass and binary distribution , as well as the structure of the features. This study can be developed in many ways and furthermore studies can be carried out in line with the following recommendations. Innovative groups and hybrid classifiers can also be integrated for better comparison and improved performance. In addition, the feature selection process can be based on the development of models to better understand important features.

REFERENCES

[1] C. Romero and S. Ventura, ``Educational data mining: A survey from 1995 to 2005," Expert Syst. Appl., vol. 33, no. 1, pp. 135_146, Jul. 2007.
[2] R. Ghorbani and R. Ghousi, ``Predictive data mining approaches in medical diagnosis: A reviewof some diseases prediction," Int. J. Data Netw. Sci., vol. 3, no. 2, pp. 47_70, 2019.
[3] D. Kabakchieva, ``Student performance prediction by using data mining classification algorithms,"Int. J. Comput. Sci. Manage. Res., vol. 1, no. 4, pp. 686_690, 2012.
[4] A. I. Adekitan and E. Noma-Osaghae, ``Data mining approach to predicting the performance of first year student in a University using the admission requirements," Educ. Inf. Technol., vol. 24, no. 2, pp. 1527_1543, Mar. 2019.
[5] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. V. Erven, ``Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," J. Bus. Res., vol. 94, pp. 335_343, Jan. 2019.
[6] R. S. Baker and K. Yacef, ``The state of educational data mining in 2009: A review and future visions," J. Educ. Data Mining, vol. 1, no. 1, pp. 3_17, 2009.
[7] E. Chandra and K. Nandhini, ``Knowledge mining from student data," Eur. J. Sci. Res., vol. 47,no. 1, pp. 156_163, 2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY