International Journal of Advanced Technology in Engineering and Science
Vol. No. 11, Issue No. 04, April 2023
www.ijates.com
ijates
ISSN 2348 - 7550

Multiclass Prediction Model for Student Grade Prediction Through Machine Learning

G. Sitharamani, CH. Priyanka Neha, B. Yoga Deepthi, A.Srinivas

(Under the guidance of Mr. S. Ramesh Babu, Assosiate Professor,
Department of Computer Science and Engineering, Tirumala Engineering College)

ABSTRACT

Today, predictive analysis applications became an urgent desire in higher education institutions. Student grade is one of the key performance indicators that can help educators monitor their academic performance. However, there are severe challenges in handling imbalanced datasets for enhancing the performance of predicting student grades. For analysis the student grade, machine learning techniques such as Logistic Regression, KNN, SVM, Random Forest, Decision Tree, Naive Bayes, SMOTE are used. This proposed model indicates the comparable and promising results that can enhance the prediction performance model for imbalanced multi-classification for student grade prediction.

Keywords— Machine learning, Predictive model, imbalanced problem Student grade prediction, Multi Class Classification

I. INTRODUCTION

Predictive analytics used advanced analytics that encompasses machine learning implementation to derive high-quality performance and meaningful information for all education levels. In higher education institutions, every institution has its student academic management system to record all the data of student containing information about the student academic results and grades in different courses and programs. Predictive analytics can be used to predict grade which is one of the key performance indicators that can help educational institutions monitor their academic performance.

However, the related works on mechanism to improve imbalanced multi- classification problem in predicting students' grade prediction are difficult to found. To address this, we will be using oversampling techniques and feature selection techniques along with the machine learning models to predict the student's grade in this project.

II. LITERATURE SERVEY

H. Kanegae, K. Suzuki, K. Fukatani, T. Ito, N. Harada et al. [18] proposed a prediction model for the onset of new hypertension. The model was tested on clinical data of hypertensive patients. Any missing values in the data were imputed using last observation carried forward, mean and mode substitutions. The model was a combination of logistic regression, random forest and XGBoost technique combined with the help of bagging technique. The model achieved 0.992 AUC.

M. Ambika, G. Raghuraman, and L. Sai Ramesh et al. [16] developed a personalized decision support system based on a support vector machine (SVM) and fuzzy association rule mining (ARM) to predict the

probability of acquiring hypertension. The missing values in the data are substituted using mean and mode value substitution and the interquartile range(IQR) technique is used to remove the outliers. The model also took into account personal behavioral factors along with medical history for prediction. The model reported a prediction accuracy of 91.8%.

J. Chorowski, J. Wang, and J. M. Zurada, [26] developed a mode based on Support VectorMachine (SVM), the Lea st-Squares SVM (LSSVM), the Extreme Learning Machine (ELM), and the Margin Loss ELM (MLELM) are discussed to demonstrate how specific parameterizations of a general problem statement affect the classifier design and performance, and how ideas from the four different classifiers can be mixed and used together. Comparison of classification accuracies under a nested cross- validation evaluation shows that with an exception all four models perform similarly on the evaluated datasets.

Y. Isler, A. Narin, M. Ozer[10] proposed a model based on three detrending methods, the smoothness prior approach, the wavelet and the empirical mode decomposition, were compared on artificial R-R interval series with four types of simulated trends. Results indicated that the wavelet method showed a better overall performance than the other two methods, and more time-saving, too. Therefore it was selected for spectral analysis of real R-R interval series of thirty-seven healthy subjects.

Xiaohan Li, S. Wu [14] developed a model based on Recurrent Neural Networks (RNNs), especially those using Long Short-Term Memory (LSTM) units, can capture long range dependencies,so they are effective in modeling variable-length sequences. They conduct experiments on the BP data set collected from a type of wireless home BP monitors, and their experimental results show that theproposed models outperform several competitive compared methods. The model reported the predicted accuracy of 82.4%.

III. PROPOSED SYSTEM

Students marks of other subjects are taken as input for evaluation students' performance. Data set is pre-processed and features and labels are extracted from dataset then dataset is split in to test and train sets then linear regression is applied to dataset for prediction. Before final marks of all subjects are evaluated prediction can be performed. Using machine learning process automation of marks prediction can be done.

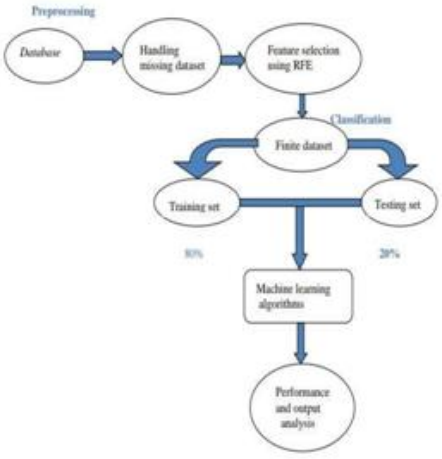


Fig: Proposed system

IV. TRAINING AND TESTING THE DATASET

The dataset we will be using contains 23 attributes and 1044 instances in which the class distribution of dataset is non-uniform indicating an Imbalanced Dataset that can lead to overfit results.

This is an imbalanced dataset because it consists of 55,112,470,337 and 70 instances of grade 1,2,3,4 and 5 respectively. SMOTE technique can be used to balance this Dataset. The dataset is partitioned into 80% and 20% by using 5 fold Stratified cross validation.

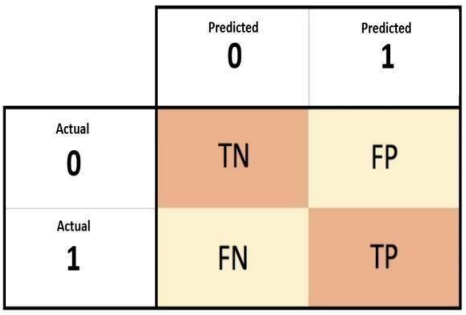


Fig: Confusion matrix

Metrics are used to evaluate the efficiency of various machine learning algorithms. The training process is followed by a validation process that is used to measure the performance of the model and in particularits ability to generalize to (input, output) features that were not used in training the model. We will calculate the metrics from the confusion matrix.

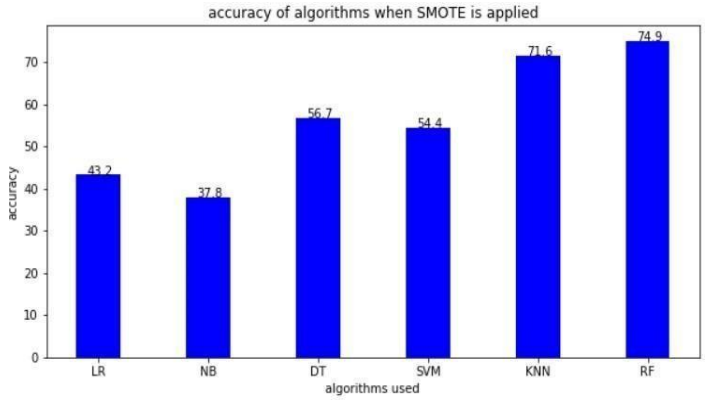
Attribute	Type	Values	Description
School	Nominal	SI-SI641	Student Identification
Sex	Nominal	[M,F]	Gender of Student
Age	Numeric	[1..100]	Age of student
Address	Nominal	DEC_JUNE	Address of student
FamSize	Nominal	[GT3,LE3]	Size of Student Family
Guardian	Nominal	[Mother,Father]	Guardian of Student
Medu	Numeric	[1,4]	Education of Student's Mother
Fedu	Numeric	[1,4]	Education of student's Father
Travel Time	Numeric	[1,2]	Travel time of Student
Study Time	Numeric	[1,4]	Student studying time
Failures	Numeric	[1,4]	Failures of a Student
Schools up	Nominal	[Yes,No]	Extra Education if any
Famsup	Nominal	[Yes, No]	Support of family
Paid	Nominal	[Yes,No]	Paid for extra courses or not
Activities	Nominal	[Yes,No]	Extra curricular activities
Higher	Nominal	[Yes,No]	Willing to take Higher Education
Internet	Nominal	[Yes,No]	Internet Access at home
Subject			
Famrel	Numeric	[1,5]	Status of family relationships
Freetime	Numeric	[1,5]	Free time after School
Goout	Numeric	[1,5]	Going with Friends
Health	Numeric	[1,5]	Status of Health
Absences	Numeric	[0,93]	No.of absences
G	Numeric	[0,20]	Grade of student
Grade	Nominal	[A ,B,C,D,E]	Final Grade Predicted

Table : Dataset Description

V. RESULTS

In this project, six machine learning algorithms namely Logistic Regression, Naïve Bayes, Decision Tree, Support Vector Machine, K Nearest Neighbors and Random Forest Classifiers are applied on an imbalanced student dataset to predict the grade of the student.

An Oversampling technique i.e, SMOTE is used to overcome the problem of imbalanced dataset and two feature selection techniques.chi square and feature_importance are used to select the optimal features. When



the machine learning algorithms are applied alone, SVM performs betterthan other algorithms with an accuracy of 49%.

When SMOTE is applied alone, Random Forestperforms better with an accuracy of 74%. When each of the feature techniques are used, SVM performs better with an accuracy of 47% and 48% approximately. When SMOTE is applied together with feature selection algorithms, RF shows highest accuracy of 63% and 71.5% respectively.

From this results, we can analyze that RF performs best with SMOTE alone but it considers many features compared to SMOTE with feature_importance whose accuracy is slightly low. Hence, it is optimal and better to consider applying RF algorithm along with SMOTE and feature_importance technique.

The distribution of number of grades in the imbalanced dataset is:

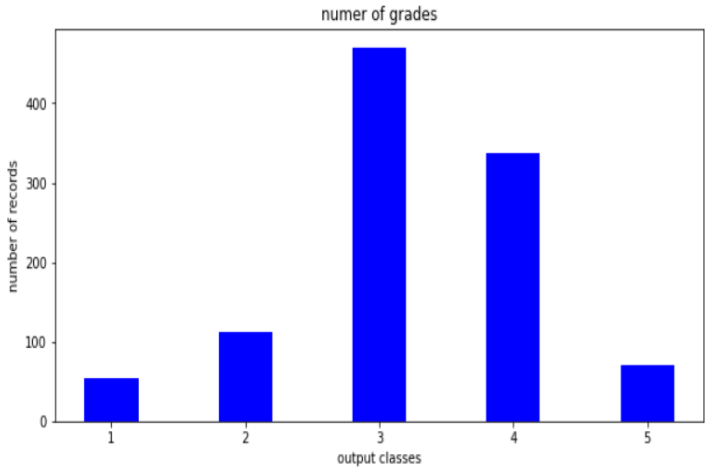


Fig: Distribution of classes

The accuracy of machine learning algorithms alone is:

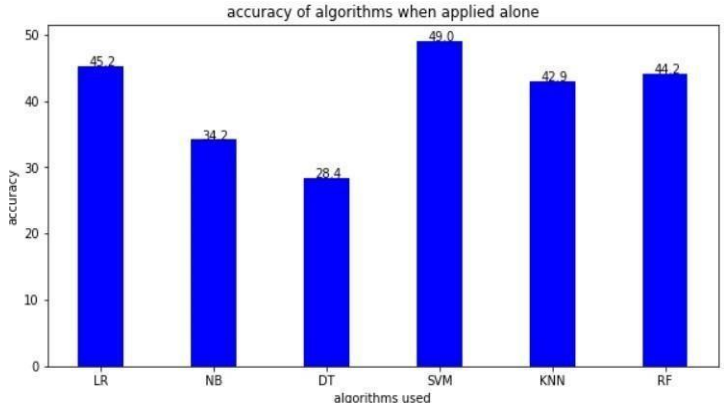


Fig: Accuracy of algorithms alone

The Accuracy of machine learning algorithms when SMOTE/ features selection methods applied invidually

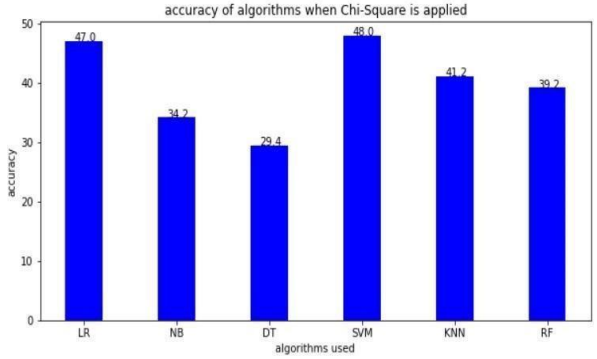


Fig: Accuracy of algorithms with SMOTE

Fig: Accuracy of algorithms with Chi-square

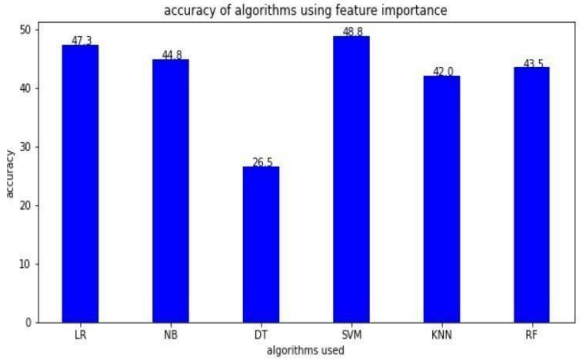


Fig: Accuracy of algorithms with feature importance

The Accuracy of machine learning algorithms when SMOTE and feature selection applied together is:

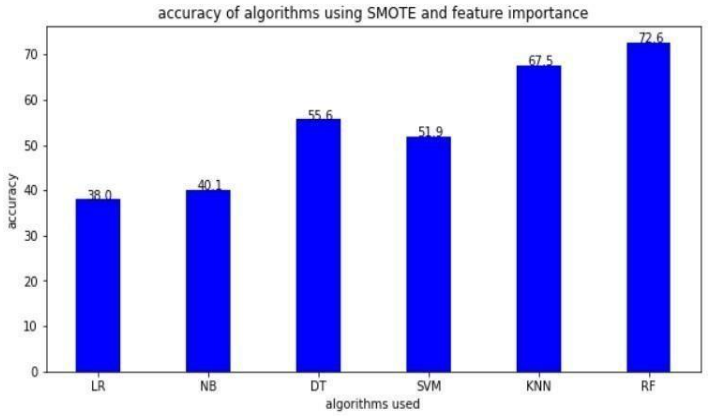


Fig: Accuracy of algorithms with SMOTE and feature importance

VI. CONCLUSION

Predicting student grades is one of the key performance indicators that can help educators monitor their academic performance. Therefore, it is important to have a predictive model that can reduce the level of uncertainty in the outcome for an imbalanced dataset. This project shows a multiclass prediction model with six predictive models to predict student's grades based on their characteristics and information. Specifically, we have done a comparative analysis of combining oversampling SMOTE with different FS methods to evaluate the performance accuracy of student grade prediction. We also have shown that the explored oversampling SMOTE is overall improved consistently than using FS alone with all predictive models. However, our proposed multiclass prediction model performed more effectively than using oversampling SMOTE and FS alone with some parameter settings that can influence the performance accuracy of all predictive models. Here, our findings contribute to be a practical approach for addressing the imbalanced multi-classification based on the data-level solution for student grade prediction.

REFERENCES

[1] A. E. Tatar and D. Düştegör, ''Prediction of academic performance at undergraduate graduation: Course grades or grade point average?'' Appl. Sci., vol. 10, no. 14, pp. 1–15, 2020.

[2] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas,J. Leinonen, C. Messom, and S. N. Liao, ''Predicting academic performance: A systematic literature review,'' in Proc. 23rd Annu. Conf. Innov. Technol. Comput. Sci. Educ., Jul. 2018,pp. 175–199.

[3] A. Polyzou and G. Karypis, ''Grade prediction with models specific to students and courses,'' Int. J. Data Sci. Anal., vol. 2, nos. 3–4, pp. 159–171, Dec. 2016.

[4] A.Verma, ''Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA,'' Int. Res. J. Eng. Technol., vol. 6, no. 3, pp.54–60, 2019.

[5] C. Jalota and R. Agrawal, Feature Selection Algorithms and Student Academic Performance: A Study, vol. 1165. Singapore: Springer, 2021.

[6] D. Berrar, ‘‘Cross-validation,’’ Comput. Biol., vols. 1–3, pp. 542–545, Jan. 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

[7] E. Alyahyan and D. Düştegör, ‘‘Predicting academic success in higher education: Literature review and best practices,’’ Int. J. Educ. Technol. Higher Educ., vol. 17, no. 1, Dec.2020

[8] L. E. O. Breiman, ‘‘Random Forests,’’ Mach. Learn., vol. 45, pp. 5–32, Oct. 2001.

[9] L. M. Abu Zohair, ‘‘Prediction of student’s performance by modelling small datasetsize,’’ Int.J. Educ. Technol. Higher Educ., vol. 16, no. 1, pp. 1–8, Dec. 2019.