Multiclass Prediction Model for Student Grade Prediction Using Machine Learning

Priyanka V
Scholar
Information Technology
Kongu Engineering College
Erode, Tamilnadu, India
Priyankav.19it@kongu.edu

Vani Sridhar C
Scholar
Information Technology
Kongu Engineering College
Erode, Tamilnadu, India
vanisridharc.19it@kongu.edu

Vasanth T
Scholar
nformation Technology
Kongu Engineering College
Erode, Tamilnadu, India
vansantht.19it@kongu.edu

Jeevanantham A
Assistant Professor
Information Technology
Kongu Engineering College
Erode, Tamilnadu, India
jeeva@kongu.ac.in

***Abstract* – Student grade is one of the key performance indicators that can help educators to monitor student's academic performance. Predicting this student's grade is a tedious task for the teachers. This paper presents a comprehensive analysis of machine learning techniques to predict the student grade.**

The accuracy performance of four machine learning techniques namely Decision Tree, Naïve Bayes (NB), K-Nearest Neighbor (KNN) and Random Forest (RF) is compared using student's course grade dataset. Multiclass prediction model is proposed to reduce the overfitting and misclassification results caused by imbalanced multi-classification based on oversampling Synthetic Minority Oversampling Technique (SMOTE) with forward features selection methods.

The obtained results show that the proposed model integrates with RF give significant improvement in accuracy with 98%. This model indicates the comparable and promising results for imbalanced multi-class dataset for student grade prediction.

Keywords* – *kNN Imputer, SMOTE, Forward Feature Selection, k Nearest Neighbor, Naïve Bayes, Decision Tree, Random Forest.

I. INTRODUCTION

This project is to predict the student grade using various multiclass classification algorithms like Decision Tree, Random Forest, k-NN, Naive Bayes and analyze the accuracy and performance of the algorithms, the most accurate algorithm is used to predict the student grade. This prediction helps to analyze the performance of the student in the academic year and helps to improve their performance.

A. Machine Learning

Machine Learning gained popularity in the last decade and every application whether it is scientific or consumer applications are using ML for prediction. Many applications of ML such as voice recognition, weather prediction, image detection and email filters are very popular in today's era. It predict the result by learning the patterns in the dataset. Prediction accuracy increases with continued training. There are three main types of machine learning

namely supervised learning, unsupervised learning and reinforcement learning. In this project supervised learning is used for learning patterns in the. Dataset.

B. Supervised Learning

The function of mapping input to the output was supervised by using example input and output pairs, this type of learning is called as supervised learning. There are two types of supervised learning namely regression and classification. Regression is used for the dataset with continuous value. Classification is used for the dataset with definite value. Classification algorithms like Decision Tree, Random Forest, k-NN, Naive Bayes are used in this project.

C. Objective

The objectives of this project is given below:

- To balance the imbalance dataset by using SMOTE and to choose best features using forward feature selection.
- To fill or replace the missing values by using k-NN Imputer
- To develop a multi models using various algorithms and analyze its accuracy.

To use the most accurate model to predict the grade of the students and to analyze their performance.

D. Scope

The scope of the project is given in detail below:

- To predict students grade using various data of the student.
- To help the teachers to analyze the performance of the student and reduce their work load.
- To improve the student's performance and involvement in academic activities.

II. SYSTEM ANALYSIS

A. Literature Review

According to [1], Siti Dianah Abdul Bujang, Ali Selamat, Roliana Ibrahim, Ondrej Krejcar, Enrique Herrera-Viedma(2021) propose a Model for Student Grade Prediction. This paper discuss about the performance of the student in his/her academics by analyzing both internal and external factors that affect the student's grade. This data may be imbalanced or balanced, this model should work for all type of datasets. This uses oversampling technique SMOTE to balance the imbalance dataset. This uses two feature selection technique to select only necessary feature that improves the accuracy of the model to train the sample dataset. And by doing this the overfitting will be stopped. It uses Logistic Regression, Naive Bayes, Support Vector Machine, k-Nearest Neighbor, Random Forest algorithms to build a model and analyze the accuracy of these models. Random forest produced the good accuracy based on the analysis.

According to [2**],** This paper focus mainly about the programming skills of the student. The student is categorized based on his/her performance level and trained well. It tries to predict the placement skills of the student by using various machine learning models. K-nearest Neighbor, Decision tree, Random forest, Logistic Regression and Support Vector Machine are the algorithms used to train the dataset. Random produced the accurate value compare to other algorithms.

According to [3], This paper focuses on student performance when he is in distance education. Due to COVID-19 student need to do their education in online. So that the teacher can't able to monitor the student. The model will predict the education capability of the students in online education.

Hierarchical Linear Model is used to train the data. It shows higher education capability of the students compare to offline mode.

According to [4], This paper aims to calculate the academic achievements and quotients for placement by using machine learning models. In this paper student grades and other data are used to calculate the IQ of the student. The algorithms like Naive Bayes, k-Nearest Neighbor, Support Vector Machine, Decision tree to develop the model. Decision tree produced the accurate result compare to other algorithms.

According to [5], This paper uses important factors that affect the student performance to develop a model. This paper mainly focus on way of increasing the student's engagement in the online education. It uses Random Forest a supervised learning algorithm and clustering a unsupervised learning algorithm to find the relationship in the data and to develop a model to meacure student engagement level.

According to [6], This paper focus on predicting the student's academic performance level. Both internal and external factors of the students are used to develop the model. Feature selection is used to get the related data, which increase the accuracy of the model. The classification algorithms used are Random Forest, Logistical Regression and Naive Bayes. Random Forest classification is better than other classification algorithm to predict student's performance by using both internal and external data. It compares the accuracy by using different feature selection algorithms. Random Forest with WrapperSubsetEval method gives more accuracy compare to other methods.

According to [7], In Institution all the data of the student related to the academics are stored in digital form. Data mining technique is used to find the pattern in the dataset, so that the accuracy of the will be increased. The algorithms used are Naïve Bayes, Logistic Regression, Decision Tree, Multilayer Perceptron and Random Forest. The non-related data are removed by using feature selection. Then this paper analyze the results of the model developed using different algorithms.

B. Summary

Feature Selection is used to find the best attributes in the dataset. Various classification algorithms like Random Forest, Decision Tree, k Nearest Neighbor, SVM (Support Vector Machine) and Multilayer Perceptron are used to develop a model and the efficiency of these developed models are measured using Mean Square Error (MSE) and Confusion Matrix. The most efficient model is used to predict the student's grade.

III. SYSTEM REQUIREMENTS

A. Hardware Requirements

Processor: AMD / Intel

RAM: Minimum 4 GB

Disk Space: Minimum free space of 2GB

B. Software Requirements

OS: Windows/ Linux

Tool: Python

IDE: Co-lab

C. Python

Python is the used for machine learning projects because python has large number of modules/libraries and writing code in python is simple compare to other languages. So the complexity of the

algorithms will be reduced. In addition to this, python is also a platform independent language.

D. Google Colab Notebook

The google co-lab is the best environment for the python. It allows the user to write and execute the code, it is very comfortable for developers and suitable for complex machine learning python codes. It allows user to connect their drive to it, so that the upload of dataset became easy process. And also user can upload their dataset from local file storage too. The execution time will be reduced and it provides good error checking mechanism.

IV. DATASET SPECIFICATION

A. Description of Dataset

For this project, Academic 6th semester mark is taken as dataset. This dataset can be imbalanced or balanced. The dataset contains information related to student education. The dataset has seven attributes and two hundred rows. The column attributes are id, IOT mark, ML mark, CNS mark, internal mark, attendance percentage of the student and student grade.

Student grade is used as class label. It has strong correlation with attributes internal mark and attendance percentage. This dataset has multi class labels which are categorical. The imbalanced dataset will be handled by SMOTE.

B. Attributes Information

- IOT mark – Mark scored in IOT (value: 0 to 100)
- ML mark – Mark scored in ML (value: 0 to 100)
- CNS mark – Mark scored in CNS (value: 0 to 100)
- Internal Mark – Internal mark of the student (value: 0 to 50)
- Attendance percentage – Attendance percentage of the student (value: 0 to 100)
- Grade – Final Grade of the student (value: A or B or C or D or E)

V. SYSTEM DESIGN

A. Dataset Preprocessing

The preprocessing is required for all type of datasets in order to make it fit to this model development. This preprocessing involves many steps, which is differed based on our dataset type and model going to be used.

In this project k-NN imputer is used to replace the null values or missing values, SMOTE is a oversampling technique used to make imbalance dataset to a balanced dataset so that the accuracy will be increased and the forward feature selection technique is used to choose the correct features that will increase the accuracy of our prediction.

1) *Mapping of Variables:* The algorithms will accept only numerical values for training the model. So map non-numeric data to numeric data. This is our first step.
2) *Missing Value Replacement:* The replacement of missing values is done by the k-NN imputer which uses the k-Nearest Neighbor algorithm to replace the missing values.

   The process involved in k-NN Imputer are:

   - k-NN Imputer works only for numerical value.
   - Find n nearest neighbor in the training set using Euclidian distance (root (square(x2-x1) + square(y2-y1))).

- Then it calculate the mean of the neighbors.
- Then it replace the missing value with this mean.

3) *Balancing Dataset:* Oversampling technique used to make the imbalance dataset to balanced dataset. SMOTE (Synthetic Minority Oversampling Technique) is the best method for balancing the dataset, so that the accuracy will be increased. It will create new data for minority class vector.

The process involved in SMOTE:

- It first identifies the minority class vector.
- Then it decides the k (nearest neighbors to be chosen) value. It uses k-NN to find the nearest neighbors.
- Then compute the line between the chosen data points based on k and create the new point in the link.
- The previous step is repeated for all minority data points till the dataset is balanced.

4) *Feature Selection:* Forward feature selection is used to find out the correct set of feature to be used to find the final grade of the student with higher accuracy. It comes under Wrapper Method. It stops overfitting.

Steps in forward feature selection:

- It add attribute one by one to create a model and check its accuracy.
- If the accuracy increases then that feature will be selected.
- It accuracy is not selected then that feature will not be selected.
- This process will continue until the last attribute of the dataset.

```
Selected Columns: [0, 3, 4]
RangeIndex(start=0, stop=5, step=1)
```

Fig. 1.Feature Selection

B. Algorithms

Even in many good universities the students are losing their interest in education due to lack of support, family background. So predicting students grade based on all internal and external factors of the students will be used to predict the performance of the student in the beginning itself. So that, the teacher can understand the students situation and help them provide the necessary support they needed.

The dataset chosen is of type multi class classifier so the algorithms used for this dataset are k Nearest Neighbor, Decision Tree, Random Forest and Naive Bayes. After pre-processing we developed the models using these algorithms and the results are analyzes. Based on accuracy the best model is used to predict the final grade of the student based on all factors.

1) *k-Nearest Neighbor:* The process involved in kNN are:
   - It uses Euclidian distance to find the k nearest neighbors.

(1)

Euclidian distance = (root (square(x2-x1) + square(y2-y1)))

   - The default value of k is 3. If need, the value of k can be changed.

- Then it check the class label of nearest neighbors. And find the class label which has maximum count.
- The class label with higher count is assigned to the new data point.

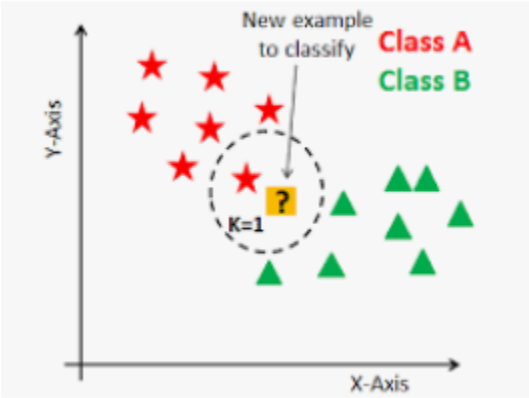


Fig. 2. k-NN Working Model

1) *Naïve Bayes:* The Multinomial Naive Bayes classification is suitable for multi class classification dataset. It works for dataset having frequency count. The categorical integer value is the input for this algorithm. The Multinomial Naive Bayes classification is works based on frequency of data.

Formula:

(2)

$$P(A/B) = (P(B/A).P(A)) / P(B)$$

- P(B) – Marginal likelihood
- P(A) – Prior probability of A
- P(B/A) – Likelihood
- P(A/B) – Posterior probability

2) *Decision Tree:* In decision tree, the data is considered as tree like structure having root, decision and terminal nodes. Root node is the starting node from which the data is starts dividing. The decision nodes are the splitting nodes and the terminal node is the last node. It provide pruning (remove unwanted nodes/data) to solve problems due to overfitting. It is like if else statement. It checks the condition and if the condition is true then it moves to next node. Entropy is the amount of uncertainty/error in the dataset. Information gain measures the decrease in uncertainty in the dataset. Entropy and Information gain is used to make decision about splitting the data.

Formula:

- Entropy for one attribute

(3)

$$E(T) = \text{Summation}(- Pi \log2 Pi)$$

- Mathematically Entropy for multiple attributes

(4)

$$E(T, X) = \text{Summation}(P(i) E(i))$$

- Information Gain

(5)

$$\text{Information Gain}(T, X) = \text{Entropy}(T) – \text{Entropy}(T, X)$$

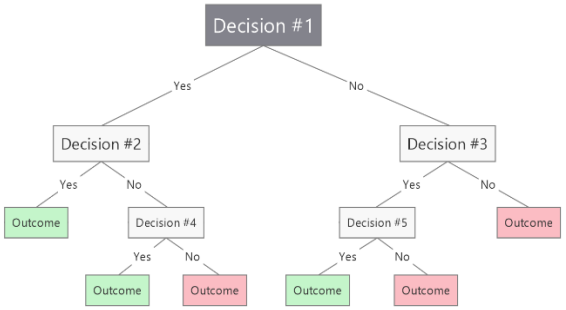


Fig. 3. Decision Tree Block Diagram

3) Random Forest: Combination of many decision trees to get more accuracy is called as random forest. The main reason for using it is, the group of decision tree will perform well when compare to the performance of single decision tree.

In Random forest each tree gives the vote, the random forest pick the classification with the majority of vote. The model developed by the Decision tree in Random forest are not related to one another, but most of them will produce accurate result.

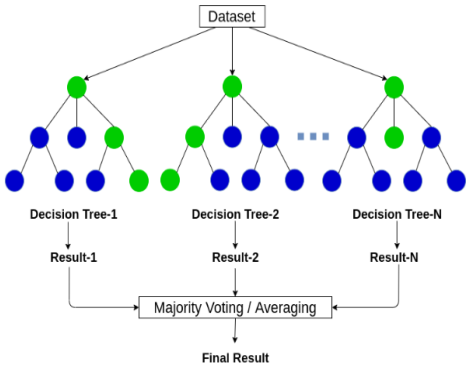
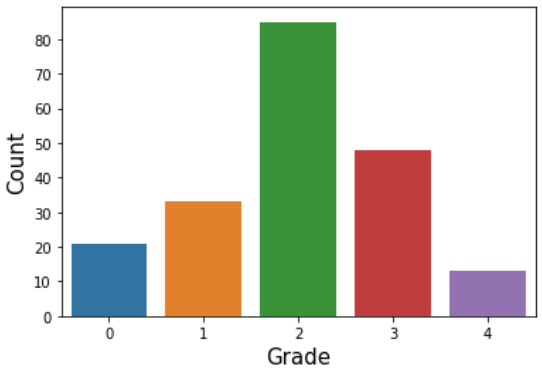


Fig. 4. Random Forest Block Diagram

VI. RESULTS AND OUTPUTS

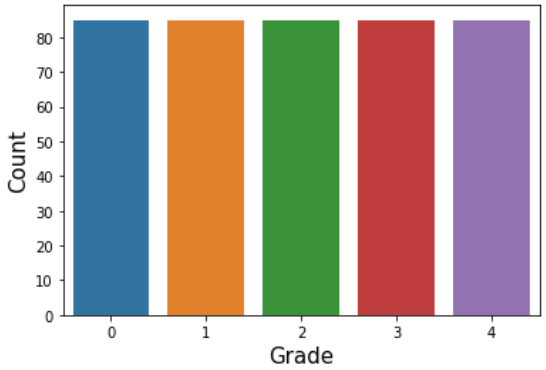
- Data visualization before oversampling



	iot	ml	cns	internal_mark	attendence
0	19	50	0	11	75
1	4	8	8	4	74
2	4	50	16	12	79
3	33	74	26	22	80
4	26	64	52	24	81
..	...	..	...	...	...
195	76	34	67	27	82
196	12	52	60	25	83
197	23	83	34	25	84
198	54	21	12	17	78
199	8	16	24	10	72

[200 rows x 5 columns]

- Data visualization after oversampling



	0	1	2	3	4
0	19.000000	50.000000	0.000000	11.000000	75.000000
1	4.000000	8.000000	8.000000	4.000000	74.000000
2	4.000000	50.000000	16.000000	12.000000	79.000000
3	33.000000	74.000000	26.000000	22.000000	80.000000
4	26.000000	64.000000	52.000000	24.000000	81.000000
..	...	...	...	...	...
420	93.373249	77.776050	88.383100	42.616900	96.000000
421	85.891676	84.000000	81.939567	44.000000	96.000000
422	83.668323	83.665838	88.002485	43.665838	95.665838
423	83.000000	82.755523	87.622239	43.688881	96.000000
424	85.439666	88.160890	84.681002	44.000000	97.040223

[425 rows x 5 columns]

Accuracy Comparision of Developed Model

	Classifiers	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score
0	Naive Bayes	0.694118	0.764706	0.750109	0.750304	0.749076
1	kNN	0.967647	0.952941	0.952500	0.950000	0.945611
2	Decision Tree	1.000000	0.964706	0.967619	0.959615	0.962645
3	Random Forest	1.000000	0.988235	0.990000	0.987500	0.988420

Confusion Matrix and Mean Square Error

- Naive Bayes

```
Confusion Matrix :
[[15  1  0  0  0]
 [ 1  9  2  1  0]
 [ 0  2  7  6  1]
 [ 0  0  5 13  1]
 [ 0  0  0  0 21]]
Mean Square Error
0.3058823529411765
```

- K Nearest Neighbor

```
Confusion Matrix :
[[16  0  0  0  0]
 [ 0 13  0  0  0]
 [ 0  2 12  2  0]
 [ 0  0  0 19  0]
 [ 0  0  0  0 21]]
Mean Square Error
0.047058823529411764
```

- Decision Tree

```
Confusion Matrix :
[[16  0  0  0  0]
 [ 0 12  1  0  0]
 [ 0  0 14  2  0]
 [ 0  0  0 19  0]
 [ 0  0  0  0 21]]
Mean Square Error
0.03529411764705882
```

- Random Forest

```
Confusion Matrix :
[[16  0  0  0  0]
 [ 0 13  0  0  0]
 [ 0  0 15  1  0]
 [ 0  0  0 19  0]
 [ 0  0  0  0 21]]
Mean Square Error
0.011764705882352941
```

VII. CONCLUSION

The models are developed after preprocessing using k Nearest Neighbor, Naive Bayes, Decision Tree and Random Forest algorithms.

The accuracy of k-NN model is 95%, the accuracy of Naive Bayes model is 76%, the accuracy of Decision tree model is 96% and the accuracy of Random forest model is 98%. By analyzing the above result of these models, Random forest algorithm gives more accuracy compare to other methods.

This Random Forest algorithm which gives higher accuracy, is used to predict the final grade of the student. So that the performance of the student in the academics can be measured effectively. This prediction helps the teacher to solve the problem of the student and to give him/her a proper support to increase their academic performance.

VIII. FUTURE WORK

This project can be completed at low cost, but only a few attributes are used in this project for model development. And the models are developed using four different algorithms and the performance of the models are compared.

In Future new attributes related to student like student mother's job, student father's job, father's income, number of hours student spend for studying will be added in the dataset and new algorithms like Gradient Boosting, XGBoost, AdaBoost will be used for model development.

REFERENCES

[1] Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. M. (2021). "Multiclass prediction model for student grade prediction using machine learning". IEEE Access, 9, 95608-95621.

[2] Ishizue, Ryosuke, et al. "Student placement and skill ranking predictors for programming classes using class attitude, psychological scales, and code metrics." Research and Practice in Technology Enhanced Learning 13.1 (2018): 1-20.

[3] Karadag, Engin, Ahmet Su, and Hatice Ergin-Kocaturk. "Multi-level analyses of distance education capacity, faculty members' adaptation, and indicators of student satisfaction in higher educations during the COVID-19 pandemic." International Journal of Educational Technology 18.1 (2021): 1-20.

[4] Y. Zhang, Y. Yun, H. Dai, J. Cui and X. Shang, "Graphs regularized robust matrix factorization and It's application on student grade prediction", Appl. Sci., vol. 10, pp. 1755, Jan. 2020.

[5] Jain, A., & Solanki, S. (2019, July). An efficient approach for multiclass student performance prediction based upon machine learning. In 2019 International Conference on Communication and Electronics Systems (ICCES) (pp. 1457-1462). IEEE

[6] Begum, Safira, and S. Padmannavar. "Genetically optimized ensemble classifiers for multiclass student performance prediction." Int. J. Intell. Eng. Syst 15.2 (2022): 316-328.

[7] Alharbi, Basma. "Back to Basics: An Interpretable Multi-Class Grade Prediction Framework." Arabian Journal for Science and *Engineering* 47.2 (2022): 2171-2186.

[8] Ramaswami, M. "Validating predictive performance of classifier models for multiclass problem in educational data mining." International Journal of Computer Science Issues (IJCSI) 11.5 (2014): 86.

[9] Athani, Suhas S., et al. "Student performance predictor using multiclass support vector classification algorithm." 2017 International Conference on Signal Processing and Communication (ICSPC). IEEE, 2017.

[10] Ninrutsirikun, Unhawa, et al. "Effect of the Multiple Intelligences in multiclass predictive model of computer programming course achievement." 2016 IEEE region 10 conference (TENCON*)*. IEEE, 2016.