Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 3s (2025)

A Streamlined Approach to Student Stream Prediction Using an Ensemble Machine Learning Model

Rajan Saluja[1] **Munishwar Rai**[2]

[1]Research Scholar, MMDU Mullana, Ambala Cantt, Haryana, India, Assistant Professor, UIC, Chandigarh University, Gharun-14030 Mohali, Punjab, India, rajansaluja@yahoo.com

[2]Professor, MMICTBM, MMDU, Mullana, Ambala Cantt, Haryana

Article History:

***Received**: 22-09-2024*

***Revised**: 03-11-2024*

***Accepted**: 18-11-2024*

Abstract:

Finding a stream or course after secondary or senior secondary education is a daunting challenge for students and parents, as numerous options are available in various engineering and non-engineering courses. This decision potentially influences a student's academic success and career. Most frequently, they take courses with the advice of relatives, neighbors, or career counsellors. Online platforms and Learning Management Systems also exist to offer guidance on stream selection. Still, these systems rely on short-term assessments such as tests, quizzes, or interviews, potentially restricting a student's options. Our research employed the Rajan and Rai (RR) student performance prediction model based on a sophisticated Ensemble Machine Learning approach. Our model incorporates a stack of four multiclass classifiers, namely Decision Tree, k-Nearest Neighbor, Naïve Bayes, and One vs. Rest Support Vector Machine classifiers, and demonstrates a remarkable accuracy rate of 80% for predicting the most suitable academic stream for a student in an Institution. To develop this model, we utilized data from five distinct branches of students. We aim to enhance students' academic success so they can complete their degrees with excellent Grades. Exploring our model in the education sector empowers students with the timely facilities they need for a successful and fulfilling educational journey.

Keywords: Academic Guidance; Decision Tree; K-Nearest Neighbor; Naïve Bayes; One vs. Rest Support Vector Machine; Ensemble Machine Learning.

1. Introduction

A student's academic success is of utmost importance as it serves as the pivot that determines their future path. The two phases of a student's educational journey are the foundational school education and the succeeding college education. After graduation, students face many duties, all of which play a part in shaping their future. Choosing a career stream is essential and forms the cornerstone of their future professional journey. This decision requires careful consideration and judgment because, once taken, it is difficult to undo. After middle school, the curriculum opens into main disciplines, including non-medical science, medical science, commerce, and the arts. Somewhere, educational policies allow students to create a personalized curriculum, which adds a deeper level of complexity to their academics. Students who prefer the arts, business, or medical sciences have limited college-level course alternatives. The available options correspond to their interests and skills in the core topics they studied at the school level. Conversely, students studying sciences in schools other than medicine choose engineering as their primary choice, with a wide range of specializations and subfields available.

A student's academic success is a crucial pillar that significantly impacts their future path. A student's life is played out in two parts: one at the school level education and the other at college education after school. Students are burdened with many duties after college, each shaping the course of a more promising life. The choice of career stream is a critical turning point that will determine their future career path. As such, once taken, this choice takes a fearsome permanence; thus, it requires careful study.

After completing intermediate school, the curriculum opens into critical courses, including, but not limited to, Arts, Commerce, Medical Science, and Non-medical Science. A new curriculum that recognizes changing paradigms allows students to customize their course schedules. Within the arts, business, and medical science fields, the variety of college possibilities corresponds well with personal preferences and skills developed in core topics. On the other hand, students studying non-medical sciences tend to go into engineering, which is their favourite career path with a wide range of specialized specializations.

There are many options available to a non-medical science student, including computer applications, business administration, science courses, and science honors, to name just a few. For parents and students alike, navigating this sea of options is daunting. Making a wise choice might help a student succeed academically, while making a wrong choice can lead to a nightmare. Erroneous stream selection by students has far-reaching consequences that affect not just the individual but all parties involved in an educational setting, including parents, instructors, administrators, managers, and society.

In our research endeavors, we have harnessed our previously proposed student performance prediction model EMLRR(Ensemble Machine Learning Rajan and Rai Model) [1] to forecast a student's most fitting academic stream. Designed initially to anticipate academic success at the earliest juncture, our RR model strives to facilitate timely interventions. Selecting an appropriate stream post-senior secondary school is pivotal to a student's academic triumph. Thus, we have adapted the same predictive model to ascertain the optimal stream for a student's educational success. Our dataset encompasses over 2000 students from four engineering branches and the Bachelor of Computer Applications (BCA) program at the Panipat Institute of Engineering and Technology (PIET). While a multitude of platforms and applications exist for stream/subject selection, often relying on quizzes or small tests, many operate on parameters such as gender, parents' income and education, student's preferences, guidance from relatives, school-level percentage or CGPA, college reputation, trending courses, and market demand. Unfortunately, these methods may lead to selecting a course misaligned with a student's interests or one that loses market relevance after a few years.

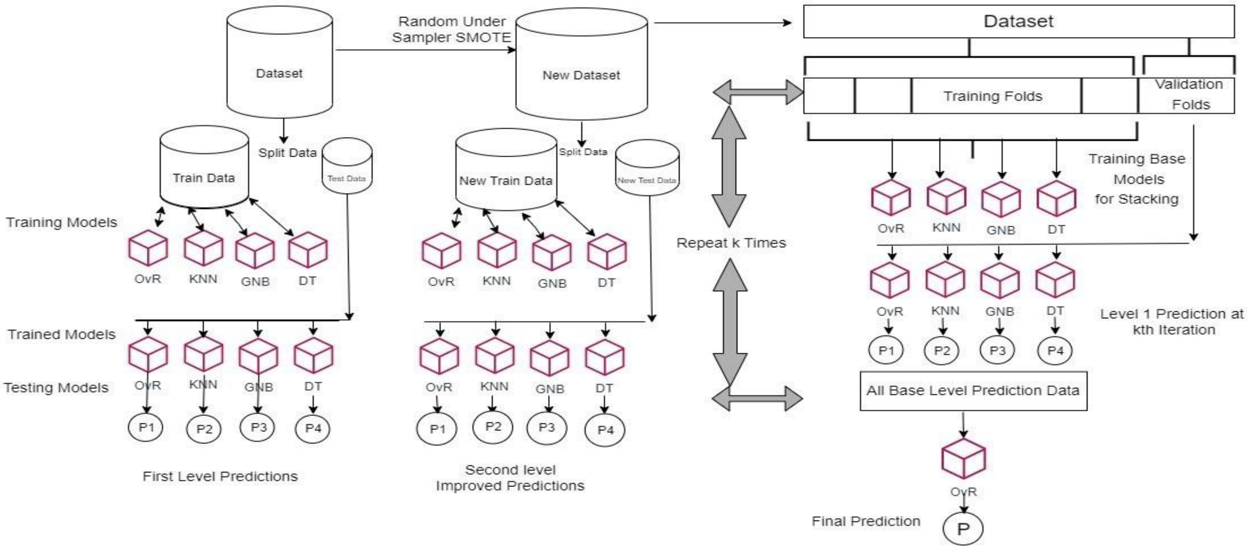


Figure 1. EMLRR Model for Prediction

Our model, honed with data from diverse branches, regions, and periods, leverages ensemble machine learning to predict the optimal stream for a student with an impressive accuracy of 80%. We are employing Decision Tree (DT) [2–5, 10], Support Vector Machine (SVM) [6-10], k-nearest neighbours (KNN) [10-12], and Gaussian Naïve Bayes (GNB) [8,10,11] as base models, with One vs. Rest Support Vector machine serving as the meta-model for Stacking, our approach incorporates various attributes such as student school-level CGPA, parents' financial status, city and area of residence, connectivity, zodiac sign, gender, willingness to reside in a hostel, and test scores. The structure of our paper unfolds across various sections: Section 2 outlines the research objectives, Section 3 delves into the literature review, Section 4 explains the methodology, Section 5 presents results and analysis, and Section 6 draws conclusions.

2. Research Objectives

Our study endeavors to enhance student academic success by applying our ensemble machine learning model, a pivotal step involving the meticulous selection of an appropriate academic stream for each student. The research is structured into two distinct phases: the initial phase consists of the collection, cleaning, and curation of data, focusing on identifying pertinent parameters influencing the choice of academic stream. In the subsequent phase, our model is trained using this refined dataset, and its predictive capabilities are tested in real-world scenarios, specifically at the point of student admission.

Our investigation addresses several research questions aimed at unraveling the intricacies of stream selection and model performance:

- What factors influence a student's choice of academic stream?

- Meta-Model Comparison: How does the prediction accuracy of our proposed model compare with alternative meta-model options?

- SMOTE Evaluation in RR Model: How do the prediction accuracy and other pertinent parameters compare when employing Synthetic Minority Over-sampling Technique (SMOTE) options within the RR model?

This structured approach allows us to delve into the complexities of stream selection, assess the comparative efficacy of our ensemble machine learning model against alternative meta-models, and evaluate the impact of different SMOTE options within the RR model. Ultimately, we aim to provide insights that contribute to an informed and effective stream selection process, thereby bolstering students' academic success.

3. Literature Survey

The pursuit of selecting the optimal academic stream for a student is a crucial factor in enhancing their educational success. In the contemporary era of digitized education systems, where vast student data are readily available, cutting-edge technologies such as machine learning and data science offer the potential to transform machines into intelligent expert systems. Leveraging historical data, these machines can guide students through informed counseling and predict the most suitable academic stream.

Our model employs stacking [13], where we have combined OvR [14], DT [15], GNB [16], and KNN classifier [17] are combined with a meta-model to predict the academic stream as a multi-class output. Additionally, our study examines the various parameters, including the accuracy and precision of these four ML Models, with and without applying SMOTE. In essence, our problem constitutes a multi-class classification challenge analogous to image classification or handwriting classification problems. The dataset comprises three target class labels: Class 0 for Bachelor in Computer Applications (BCA), Class 1 for Computer Science & Engineering (CSE)/Information Technology (IT), Class 2 for Electronics and Communication (ECE), and Mechanical Engineering (ME).

A comparative analysis with previous studies reveals opportunities for refinement. Alsayed et al. [18] demonstrated the efficacy of supervised learning techniques but suggested room for improved prediction accuracy. Kapil Sethi and Mohd Dilshad Ansari [12] achieved classification accuracy exceeding 80% but faced limitations due to a restricted dataset and focus on only two output class data. Yara Zayed et al. [19] enhanced their intelligent recommendation system with ML hyper-tuning, achieving accuracy surpassing 90%, but faced constraints in dataset variety. Samuel A. Stein et al. [20] leveraged student data for significant recommendations, obtaining a 61% success rate in predicting actual majors. Charbel Obeid et al. [21] proposed an ontology-based recommender system yet lacked transparency on dataset attributes. Inssaf El Guabassi et al. [22] developed a predictive model for students’ admission using machine learning algorithms but faced limitations with a small dataset and binary class output. Lamees Al-AlawiIn et al. [23] identified significant parameters affecting academic success but grappled with data gaps due to the COVID-19 pandemic.

Recent studies by Laszlo Bognar and Tibor Fauszt [24] and others underscore the ongoing exploration of factors influencing the predictive ability of exam-level models in predicting students’ success in Learning Management Systems. An analysis of different machine learning algorithms for predicting a student's success was done [25] to find out the strengths and weaknesses of various ML techniques.

Our choice of machine learning techniques, such as KNN, GNB, DT, and SVM, aligns with recent trends in academic performance prediction.

Bagging, boosting, and stacking, like ML techniques, are incorporated into complicated problems such as Emotional recognition, speech recognition, disease detection, and spam/fraud detection. Stacking [13] is an ensemble technique [26] that is being used to design a new model with the help of existing models so that performance can be boosted [27-30]. The stacking architecture [31] includes two models; at the low level, more than one base model is planted, and at the upper level, there is one meta-model. Ensemble means training the base models with the dataset and getting prediction data at a low level; after that, training the meta-model with prediction data collected from base models and getting a final prediction.

Our study advances the field by combining diverse machine-learning techniques and employing stacking for improved accuracy in predicting academic streams. By addressing the limitations observed in prior research, our work contributes to the ongoing discourse on leveraging technology to enhance academic student success.

4. Methodology

The methodology employed in our research delineates a systematic approach to accomplish our objectives, encompassing the following key steps:

4.1 Data Collection, Data Preprocessing, and Cleaning

4.2 Data Splitting into Train and Test Data

4.3 Training the Model with Training Data

4.4 Testing the Model to Predict the Stream

This structured methodology, as shown in Figure 2, ensures a systematic approach to solving our research objectives. Following these steps, the outcome of our research endeavors to provide reliable and meaningful insights into predicting academic streams for students.

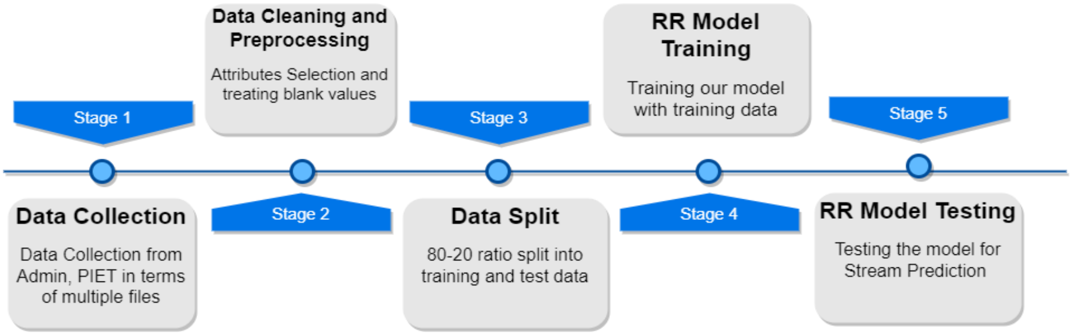
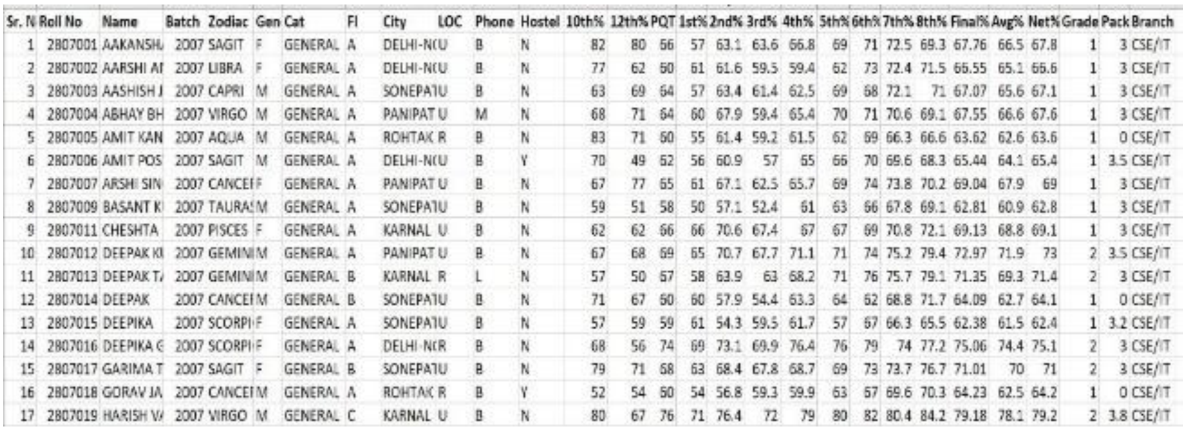

Stage 1
Data Collection
Data Collection from Admin, PIET in terms of multiple files
Stage 2
Data Cleaning and Preprocessing
Attributes Selection and treating blank values
Stage 3
Data Split
80-20 ratio split into training and test data
Stage 4
RR Model Training
Training our model with training data
Stage 5
RR Model Testing
Testing the model for Stream Prediction

Figure 2. Methodology

4.1 Data Collection, Data Preprocessing, and Cleaning

Data was collected from the Academic branch of the Panipat Institute of Engineering and Technology (PIET), Samalkha, Delhi NCR, Haryana, India. This data was in multiple Word and Excel files, each capturing distinct aspects of student information. A comprehensive dataset was curated by arranging all attributes into a single sheet, as represented in Figure 3. In total, the dataset encompasses 26 attributes, each of which is described in Table 1.



Sr. N	Roll No	Name	Batch	Zodiac	Gen	Cat	FI	City	LOC	Phone	Hostel	10th%	12th%	PQT	1st%	2nd%	3rd%	4th%	5th%	6th%	7th%	8th%	Final%	Avg%	Net%	Grade	Pack	Branch
1	2807001	AAKANSH/	2007	SAGIT	F	GENERAL	A	DELHI-NC	U	B	N	82	80	56	57	63.1	63.6	66.8	69	71	72.5	69.3	67.76	66.5	67.8	1	3	CSE/IT
2	2807002	AARSHI AI	2007	LIBRA	F	GENERAL	A	DELHI-NC	U	B	N	77	62	50	61	61.6	59.5	59.4	62	73	72.4	71.5	66.55	65.1	66.6	1	3	CSE/IT
3	2807003	AASHISH J	2007	CAPRI	M	GENERAL	A	SONEPAT	U	B	N	63	69	64	57	63.4	61.4	62.5	69	68	72.1	71	67.07	65.6	67.1	1	3	CSE/IT
4	2807004	ABHAY BH	2007	VIRGO	M	GENERAL	A	PANIPAT	U	M	N	68	71	64	60	67.9	59.4	65.4	70	71	70.6	69.1	67.55	66.6	67.6	1	3	CSE/IT
5	2807005	AMIT KAN	2007	AQUA	M	GENERAL	A	ROHTAK	R	B	N	83	71	60	55	61.4	59.2	61.5	62	69	66.3	66.6	63.62	62.6	63.6	1	0	CSE/IT
6	2807006	AMIT POS	2007	SAGIT	M	GENERAL	A	DELHI-NC	U	B	Y	70	49	52	56	60.9	57	65	66	70	69.6	68.3	65.44	64.1	65.4	1	3.5	CSE/IT
7	2807007	ARSHI SIN	2007	CANCEI	F	GENERAL	A	PANIPAT	U	B	N	67	77	65	61	67.1	62.5	65.7	69	74	73.8	70.2	69.04	67.9	69	1	3	CSE/IT
8	2807009	BASANT K	2007	TAURAS	M	GENERAL	A	SONEPAT	U	B	N	59	51	58	50	57.1	52.4	61	63	66	67.8	69.1	62.81	60.9	62.8	1	3	CSE/IT
9	2807011	CHESHTA	2007	PISCES	F	GENERAL	A	KARNAL	U	B	N	62	62	66	66	70.6	67.4	67	67	69	70.8	72.1	69.13	68.8	69.1	1	3	CSE/IT
10	2807012	DEEPAK KI	2007	GEMINI	M	GENERAL	A	PANIPAT	U	B	N	67	68	69	65	70.7	67.7	71.1	71	74	75.2	79.4	72.97	71.9	73	2	3.5	CSE/IT
11	2807013	DEEPAK T/	2007	GEMINI	M	GENERAL	B	KARNAL	R	L	N	57	50	57	58	63.9	63	68.2	71	76	75.7	79.1	71.35	69.3	71.4	2	3	CSE/IT
12	2807014	DEEPAK	2007	CANCEI	M	GENERAL	B	SONEPAT	U	B	N	71	67	60	60	57.9	54.4	63.3	64	62	68.8	71.7	64.09	62.7	64.1	1	0	CSE/IT
13	2807015	DEEPIKA	2007	SCORPI	F	GENERAL	A	SONEPAT	U	B	N	57	59	59	61	54.3	59.5	61.7	57	67	66.3	65.5	62.38	61.5	62.4	1	3.2	CSE/IT
14	2807016	DEEPIKA G	2007	SCORPI	F	GENERAL	A	DELHI-NC	R	B	N	68	56	74	69	73.1	69.9	76.4	76	79	74	77.2	75.06	74.4	75.1	2	3	CSE/IT
15	2807017	GARIMA T	2007	SAGIT	F	GENERAL	B	SONEPAT	U	B	N	79	71	68	63	68.4	67.8	68.7	69	73	73.7	76.7	71.01	70	71	2	3	CSE/IT
16	2807018	GORAV JA	2007	CANCEI	M	GENERAL	A	ROHTAK	R	B	Y	52	54	60	54	56.8	59.3	59.9	63	67	69.6	70.3	64.23	62.5	64.2	1	0	CSE/IT
17	2807019	HARISH V/	2007	VIRGO	M	GENERAL	C	KARNAL	U	B	N	80	67	76	71	76.4	72	79	80	82	80.4	84.2	79.18	78.1	79.2	2	3.8	CSE/IT

Figure 3. Final dataset Used for prediction of Stream

To facilitate a standardized evaluation, marks obtained in both school and engineering were converted into percentages, considering the inherent variations in performance evaluation criteria across different educational boards at the school level. Though initially included, marks from the graduation level were ultimately excluded from the dataset as they proved unnecessary for branch prediction. The grades were calculated based on the average net percentage over eight or six semesters, depending on the course duration (applicable to the BCA course). The 'Grade' field was explicitly considered, with entries restricted to either grade 1 or grade 2. Subsequently, data about these selected grades was retained, while the 'Grade' attribute was omitted.

The Pandas module in Python allowed for a smooth dataset loading during the data handling and analysis. Extraneous attributes were carefully removed to create a simplified and pertinent dataset for our further research stages. This systematic approach to gathering and preparing data is a fundamental step in our effort to use machine learning to forecast academic streams and provide insightful information to the field of education. The factors affecting the choice of stream, as determined by a thorough review of the literature, include primary student data like age, gender, category, and subject preference—furthermore, parents' information like their work, annual income, and level of social connectivity. Decision-making is further aided by residential data that spans the state, district, and area. One crucial factor is academic performance, which includes entrance, senior secondary, and school levels academic grades. It is also acknowledged that market demand is critical in influencing stream selection.

To ensure the robustness of our predictive model, we gathered student data spanning from 2007 to 2021 across disciplines such as Computer Science Engineering, Information Technology, Electronics and Communication Engineering, Mechanical Engineering, and Bachelor of Computer Applications (BCA). The inclusion criteria for data entries were stringent, focusing on students with good and excellent academic grades or those who secured job placements.

In refining the dataset for analysis, extraneous fields such as roll number, name, batch, graduation-level percentages, grade, and package were judiciously removed. This meticulous curation process, outlined in Table 1, ensures that

the dataset is tailored for predictive modelling, focusing on relevant attributes conducive to accurate stream selection predictions.

Table 1 Attributes, their description, and selection for stream prediction

SR NO.	ATTRIBUTE	DESCRIPTION
1.	Roll No	The unique identity of the student. Not needed for stream prediction.
2.	Name	Name of the student.
3.	Batch	The year when a student got admitted
4.	Zodiac	The Zodiac sign was calculated using an Excel formula with the Date of Birth.
5.	Gender	Gender of Student.
6.	Cat	Category of student: General, BC, and SC
7.	FI	Financial Income Status A, B, and C
8.	City	The part of the country
9.	Location	Rural or Urban areas
10.	Phone	Connectivity with Student/Father/Mother/Guardian
11.	Hostel	Willing to live in a Hostel or day scholar.
12.	10th %	Intermediate School Level Marks
13.	12th %	Senior Secondary School Level marks
14.	PQT	Entrance Test Marks
15.	1st %-8th %	Marks 1st sem to 8th sem.
16.	Net%	The average percentage
17.	Grade	Average, Good, and Excellent.
18.	Package	The salary package
19.	Branch/Stream	The stream of students: CSE/IT, BCA, ECE/ME.

The Seaborn library's barplot function generated a bar graph to display the count of different categorical attributes, as represented in Figure 4. Categories within the dataset are assigned three distinct values: General, BC (Backward Class), and SC (Scheduled Caste). The city attribute is the residential place of the student; geographically, the country is segmented into six parts, including four regions within the Haryana State, denoted as majority locations, along with the Delhi-NCR and Other States (OS) category for locations outside Haryana. In addition to the Residential area, the Location attribute was taken for two places, urban and rural, along with five variations in parents' phone/mobile connectivity. The categories were encoded into integers using the LabelEncoder method for practical analysis and model compatibility. The resulting dataset, post-label encoding, is illustrated in Figure 5. This process ensures that categorical data is appropriately transformed for subsequent modeling and analysis, facilitating a more comprehensive understanding of the dataset's structure and patterns.

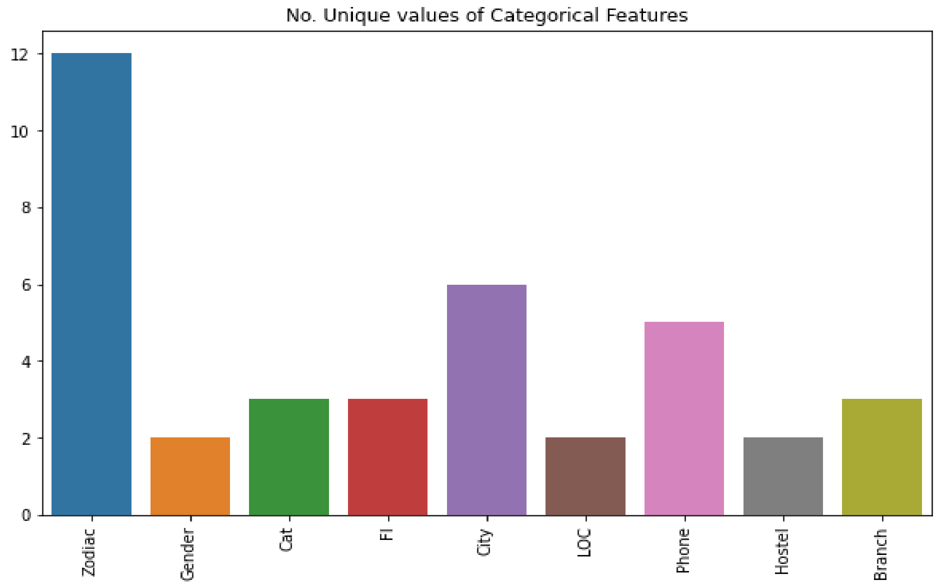
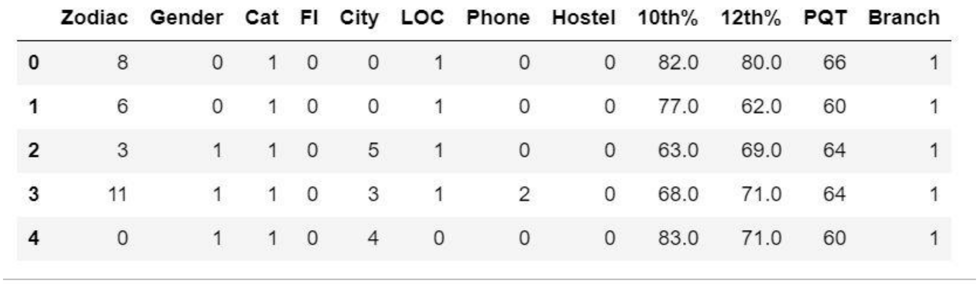


Figure 4. Categorical attributes value count

4.2 Data Splitting into Train and Test Data

The dataset is split into separate sets for testing and training. This task is essential for evaluating the model's performance by conducting a thorough assessment to predict the results. The dataset is being partitioned into independent and dependent attributes. The independent features X in our dataset constitute the attributes 'Hostel,' '10th%,' '12th%,' 'Zodiac,' 'Gender,' 'Cat,' 'FI,' 'City,' 'LOC,' 'Phone,' and 'PQT.' The dependent attribute in our dataset is "branch": the target variable denoted by y, which requires prediction. The dataset was then separated into X_train, X_test, y_train, and y_test sets in an 80:20 proportion; the ratio was chosen after evaluating different proportions.



	Zodiac	Gender	Cat	FI	City	LOC	Phone	Hostel	10th%	12th%	PQT	Branch
0	8	0	1	0	0	1	0	0	82.0	80.0	66	1
1	6	0	1	0	0	1	0	0	77.0	62.0	60	1
2	3	1	1	0	5	1	0	0	63.0	69.0	64	1
3	11	1	1	0	3	1	2	0	68.0	71.0	64	1
4	0	1	1	0	4	0	0	0	83.0	71.0	60	1

4.3Training of RR Model

Each ML model needs a thorough training process using the training dataset to create accurate predictions, as the model must first understand the intricate relationships, patterns, and other features included in the data. The objective attribute to be predicted ('Branch'), data transformation instructions, training parameters to regulate the learning process, and the input training data source (X_train and y_train) were all specified throughout the model's training. Different independent characteristics and train-test split ratios were used to train the RR model. Final training utilized the following independent attributes: 'Zodiac,' 'Gender,' 'Cat,' 'FI,' 'City,' 'LOC,' 'Phone,' 'Hostel,' '10th%,' '12th%,' and 'PQT.'

4.4 Testing RR Model

The trained model is to be tested with a separate testing dataset. Its efficacy is gauged by its ability to predict students' academic stream accurately. This step serves as a crucial validation of the model's practical utility. The testing of the RR model was completed in three different phases. In phase 1, predictions have been computed using a provided dataset for all four multiclass classification models. In phase 2, predictions were calculated with the help of oversampled data, and in the final phase, stream prediction was computed using stacking [31].

Table 2 Predictions using the Given Dataset

Model	Accuracy	Recall	Precision	F1-Score
DT	69	67	67	68
OvR	55	54	55	50
KNN	56	55	56	56
GNB	68	68	67	68

Figure 5. Final dataset used for training the model

The performance of the model was measured on four key metrics: accuracy, F1-Score, Recall, and Precision [1]—the accuracy of predicting stream for a student after the final prediction was found to be 80%. Analysis of first phase predictions for all classification models is outlined in Table 2, in Table 3, second level predictions are described 3, and final level prediction analysis is mentioned in Table 4.

Table 3 Predictions with SMOTE

Models testing after SMOTE	Accuracy	Recall	Precision	F1-Score
DT	74	77	75	75
OvR	62	64	60	57
KNN	58	57	57	57
GNB	71	71	71	71

5. Results And Discussion

In the first phase, the observation revealed that DT, OvR, KNN, and GNB accuracy levels were 69%, 55%, 56%, and 68%. Recall, precision, and F1-Score values across all models remained consistently robust under various sampling criteria. Skewness [32] was visible because of the different counts in the BCA, CSE/IT, and ME/ECE branches. SMOTE [33-34] was utilized at second-level prediction to tackle this imbalance in the dataset, which is well known for managing data that is not in balance. After implementing SMOTE, new data belonging to the minority class was introduced. The model underwent thorough testing on the same dataset with various oversampling techniques [35–41], like

Random Under Sampler [35], Borderline SMOTE [36], SVMSMOTE [37] and ADASYN [38]. They used the newly SMOTE-enhanced data to get second-level predictions (Table 3).

Following the application of SMOTE, noticeable enhancements were observed in the performance of each technique compared to scenarios without SMOTE. DT and GNB achieved a branch prediction accuracy of more than 70%. We utilized the ensemble machine learning technique in the third phase to further boost overall prediction accuracy. We selected OvR as the meta-model for the ensemble model, incorporating all four ML classification models as base models. Rigorous testing of various environmental parameters ensued, significantly elevating prediction accuracy within our design. The detailed result analysis has been briefed in Table 4.

Table 4 Results analysis of the stream prediction using the RR model

Ensemble Model (Stacking)	**Meta-model**	**Base models**	**SMOTE**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
RR model	**SVM One vs. Rest**	**DT, OvR, KNN, GNB**	**Borderline Smote**	**80**	**80**	**80**	**80**
			Random Under Sampler	**76**	**75**	**76**	**76**
			SVMSMOTE	**79**	**79**	**79**	**79**
			SMOTE	**76**	**75**	**75**	**75**
M2	Decision tree	DT, OvR, KNN, GNB	Borderline Smote	76	78	76	75
			Random Under Sampler	71	71	71	71
			SVMSMOTE	77	78	77	77
			SMOTE	76	75	75	75
M3	K-nearest neighbors	DT, OvR, KNN, GNB	Borderline Smote	78	78	78	78
			Random Under Sampler	72	72	72	72
			SVMSMOTE	76	77	77	77
			SMOTE	77	79	79	79
M4	Gaussian Naïve Bayes	DT, OvR, KNN, GNB	Borderline Smote	77	78	78	78
			Random Under Sampler	76	76	76	76
			SVMSMOTE	77	78	77	77
			SMOTE	79	80	79	79

The highest prediction accuracy for all three classes reached 80%. At the first level of stacking, KNN, NB, OvR, and DT were selected as base models. In the optimal performance scenario, the Borderline SMOTE emerged as the most effective oversampling technique for our RR Model. Additionally, other SMOTE techniques demonstrated superior performance in our model, achieving a remarkable accuracy of up to 80% in predicting the branch of a student. The model showcased excellent values for precision, recall, and F1-Score.

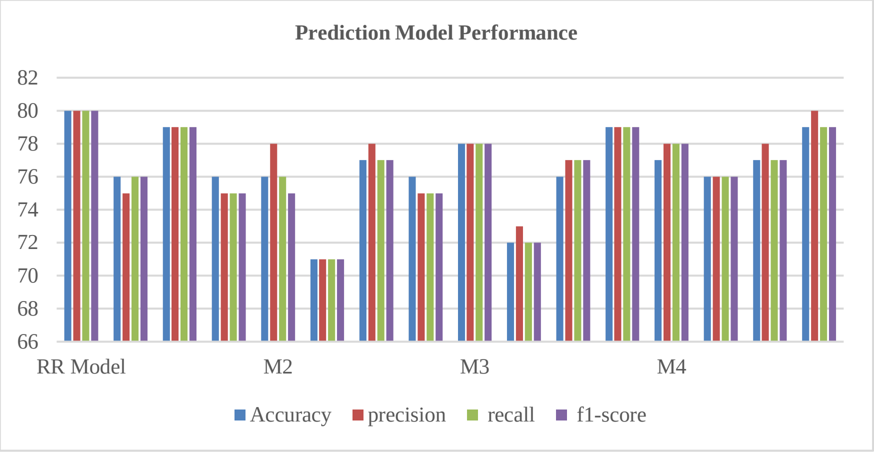


Figure 6. The Prediction Performance Analysis of the RR model.

Evaluations were conducted with KNN, DT, and GNB as meta-models. The graphical representation in Figure 6 illustrates that across all SMOTE methods (Random Under Sampler, Borderline SMOTE, SVMSMOTE, SMOTE, or ADASYN), the RR model consistently outperformed other ensemble models (M2, M3, and M4).

6. Conclusions

In this paper, we have used our predictive model to predict the stream/branch of a student seeking admission to a particular Institution. The model works on ensemble machine learning techniques expressly, the stacking of OvR as a meta-model, DT, KNN, GNB, and OvR as base models, and the Borderline SMOTE oversampling technique has been incorporated in the model. Our model demonstrates an impressive accuracy of 80% in predicting students' stream at the time of admission to an Institution. Precision, recall, and F1-Score metrics have been accurately calculated and exhibit a commendable performance. Compared to other SMOTE variants, our model stands out as the top performer, surpassing existing models described in the literature. Our Model will serve the students, ensuring their academic success and successful tests conducted in the institution. Our model showcases a substantial improvement in the institution's academic performance. Our model's applicability significantly extends beyond educational contexts, making it suitable for other multiclass classification problems such as fraud detection, cancer detection, or image classification.

For future work, we plan to leverage this model to test at other Institutions where other courses are available and different classes of students belong to other regions and cultures. We aim to explore more sophisticated ensemble machine-learning models to enhance prediction accuracy at the next level. We conclude that the RR model offers an excellent service to students and institutions to improve a student's academic success.

Acknowledgment

I am highly thankful to the academic and administrative staff of the Panipat Institute of Engineering and Technology for providing each kind of support. I thank the faculty of MM Institute of Computer Technology & Business Management (MMDU) for their support and guidance.

Conflict of Interest

I declare that there is no conflict of interest regarding the publication of this article.

References

[1] Saluja R., Rai M., Saluja R. Designing new student performance prediction model using ensemble machine learning (2023) *Journal of Autonomous Intelligence*, 6 (1), pp. 1 - 12. DOI: https://doi.org/10.32629/jai.v6i1.583

[2] Ünal F. Data mining for student performance prediction in education. In: *Birant D (editor). Data mining Methods, applications, and systems*. London: IntechOpen; 2021. doi: 10.5772/intechopen.91449

[3] Palacios CA, Reyes-Suárez JA, Bearzotti LA, et al. Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. Entropy 2021; 23(4): 485. doi: https://doi.org/10.3390/e23040485

[4] Kaunang FJ, Rotikan R. Students' academic performance prediction using data mining. In: 2018 *Third International Conference on Informatics and Computing (ICIC)*; 2018 Oct 17–18; Palembang, Indonesia. New York: IEEE; 2019. doi: 10.1109/IAC.2018.8780547

[5] Ruiz S, Urretavizcaya M, Rodríguez C, Fernández-Castro I. Predicting students' outcomes from the emotional response in the classroom and attendance. *Interactive Learning Environments* 2020; 28(1): 107–129. doi: 10.1080/10494820.2018.1528282

[6] Begum S., Padmannavar S.S. Prediction of Student Performance using Genetically Optimized Feature Selection with Multiclass Classification (2022) *International Journal of Engineering Trends and Technology*, 70 (4), pp. 223 - 235. DOI: 10.14445/22315381/IJETT-V70I4P219

[7] Bujang SDA, Selamat A, Ibrahim R, et al. Multiclass prediction model for student grade prediction using machine learning. *IEEE Access 2021*; 9: 95608–95621. doi: 10.1109/ACCESS.2021.3093563

[8] Pang Y, Judd N, O'Brien J, Ben-Avie M. Predicting students' graduation outcomes through support vector machines. In: *2017 Frontiers in Education Conference (FIE)*; 2017 Oct 18–21; Indianapolis, IN, USA. New York: IEEE; 2017. doi: 10.1109/FIE.2017.8190666

[9] Ma X., Zhou Z. Student pass rates prediction using optimized support vector machine and decision tree (2018) *2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018*, 2018-January, pp. 209 - 215. DOI: 10.1109/CCWC.2018.8301756

[10] Latif G., Abdelhamid S.E., Fawagreh K.S., Brahim G.B., Alghazo R. Machine Learning in Higher Education: Students' Performance Assessment Considering Online Activity Logs (2023*) IEEE Access*, 11, pp. 69586 – 69600 DOI: 10.1109/ACCESS.2023.3287972

[11] Cervera DEM, Parra OJS, Prado MAA. Forecasting model with machine learning in higher education ICFES exams. *International Journal of Electrical and Computer Engineering 2021*; 11(6): pp. 5402–5410. doi: http://doi.org/10.11591/ijece.v11i6.pp5402-5410

[12] Sethi K, Jaiswal V, Ansari MD. Machine learning-based support system allows students to select stream (subject). *Recent Advances in Computer Science and Communications 2020*; 13(3): 336–344. doi: 10.2174/2213275912666181128120527

[13] Nti IK, Adekoya AF, Weyori BA. A comprehensive evaluation of ensemble learning for stock-market prediction. *Journal of Big Data 2020*; 7: 20. doi: 10.1186/s40537-020-00299-5

[14] Xu J. An extended one-versus-rest support vector machine for multi-label classification. *Neurocom-puting 2011*; 74(17): 3114–3124. doi: 10.1016/j.neucom.2011.04.024

[15] Trabelsi A, Elouedi Z, Lefevre E. Decision tree classifiers for evidential attribute values and class labels. *Fuzzy Sets and Systems 2019*; 366: 46–62. doi: 10.1016/j.fss.2018.11.006

[16] Wang S, Jiang L, Li C. Adapting naive Bayes tree for text classification. *Knowledge and Information Systems 2015*; 44: 77–89. doi: 10.1007/s10115-014-0746-y.

[17] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747

[18] Alsayed, A.O.; Rahim, M.S.M.; AlBidewi, I.; Hussain, M.; Jabeen, S.H.; Alromema, N.; Hussain, S.; Jibril, M.L. Selection of the Right Undergraduate Major by Students Using Supervised Learning Techniques. Appl. Sci. 2021, 11, 10639. https://doi.org/10.3390/app112210639

[19] Zayed, Y.; Salman, Y.Hasasneh, A. A Recommendation System for Selecting the Appropriate Undergraduate Program at Higher Education Institutions Using Graduate Student Data. Appl. Sci. 2022, 12, 12525. https://doi.org/10.3390/app122412525

[20] Samuel A. Stein, Gary M. Weiss, Yiwen Chen, and Daniel D. Leeds. 2020. A College Major Recommendation System. In Proceedings of the *14th ACM Conference on Recommender Systems (RecSys '20).* Association for Computing Machinery, New York, NY, USA, 640–644. https://doi.org/10.1145/3383313.3418488

[21] Charbel Obeid, Christine Lahoud, Hicham El Khoury, Pierre-Antoine Champin: A novel hybrid recommender system approach for student academic advising named COHRS, supported by case-based reasoning and ontology. *Comput. Sci. Inf. Syst.* 19(2): 979-1005 (2022) doi: 10.2298/CSIS220215011O

[22] El Guabassi, Inssaf & Bousalem, Zakaria & Rim, Marah & Qazdar, Aimad. (2021). RSEPUA: A Recommender System for Early Predicting University Admission. 209-219. Doi: 10.1007/978-3-030-73882-2_20

[23] Al-Alawi, Lamees & Tarhini, Ali & Al-Busaidi, Adil. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Education and Information Technologies*. 28. 1-26. Doi: 10.1007/s10639-023-11700-0

[24] Bognar, Laszlo & Fauszt, Tibor & Váraljai, Mariann. (2021). The Impact of Online Quizzes on Student Success. *International Journal of Emerging Technologies in Learning (iJET)*. 16. 225. doi: https://doi.org/10.3991/ijet.v16i11.21679

[25] R. Saluja and M. Rai, "Analysis of Existing ML Techniques for Students Success Prediction," *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Solan, Himachal Pradesh, India, 2022, pp. 507-512, doi: 10.1109/PDGC56933.2022.10053236

[26] Wibawa AS, Purwarianti A. Indonesian Named-entity Recognition for 15 classes using ensemble supervised learning. *Procedia Computer Science* 2016; 81: 221–228. doi: 10.1016/j.procs.2016.04.053

[27] Hu X, Zhang H, Mei H, et al. Landslide susceptibility mapping using the stacking ensemble machine learning method in Lushui, Southwest China. Applied Sciences 2020; 10(11): 4016. doi: 10.3390/app10114016

[28] Rahman M, Chen N, Elbeltagi A, et al. Application of stacking hybrid machine learning algorithms in delineating multi-type flooding in Bangladesh. Journal of Environmental Management 2021; 295: 113086. doi: 10.1016/j.jenvman.2021.113086

[29] Chung J, Teo J. Single classifier vs. ensemble machine learning approaches for mental health prediction. Brain Informatics 2023; 10: 1. doi: 10.1186/s40708-022-00180-6

[30] Smirani LK, Yamani HA, Menzli LJ, Boulahia JA. Using ensemble learning algorithms to predict student failure and enabling customized educational paths. *Scientific Programming 2022*; 2022: 3805235. doi: 10.1155/2022/3805235

[31] Jiang W, Chen Z, Xiang Y, et al. SSEM: A novel self-adaptive stacking ensemble model for classification. *IEEE Access 2019*; 7: 120337–120349. doi: 10.1109/ACCESS.2019.2933262.2

[32] Barella VJ, Garcia LPF, de Souto MCP, et al. Assessing the data complexity of imbalanced datasets. *Information Sciences 2021*; 553: 83–109. doi: 10.1016/j.ins.2020.12.006

[33] Bej S, Davtyan N, Wolfien M, et al. LoRAS: An oversampling approach for imbalanced datasets. Machine Learning 2021; 110: 279–301. doi: 10.1007/s10994-020-05913-4

[34] Joloudari JH, Marefat A, Nematollahi MA, Oyelere SS, Hussain S. Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences. 2023*; 13(6):4006. https://doi.org/10.3390/app13064006

[35] Chaipanha, Wuttikrai & Kaewwichian, Patiphan. (2022). Smote vs. Random Under Sampling for Imbalanced Data-Car Ownership Demand Model. Communications - Scientific letters of the University of Zilina. 24. Doi: 10.26552/com.C.2022.3.D105-D115

[36] Sun Y, Que H, Cai Q, Zhao J, Li J, Kong Z, Wang S. Borderline SMOTE Algorithm and Feature Selection-Based Network *Anomalies Detection Strategy. Energies. 2022*; 15(13):4751. https://doi.org/10.3390/en15134751

[37] Bagui SS, Mink D, Bagui SC, Subramaniam S. Determining Resampling Ratios Using BSMOTE and SVM-SMOTE for Identifying Rare *Attacks in Imbalanced Cybersecurity Data. Computers. 2023*; 12(10):204. https://doi.org/10.3390/computers12100204

[38] Rahul Mitra, Anurag Bajpai, Krishanu Biswas, ADASYN-assisted machine learning for phase prediction of high entropy carbides, *Computational Materials Science*, Volume 223, 2023, 112142, ISSN 0927-0256, https://doi.org/10.1016/j.commatsci.2023.112142

[39] Davagdorj K, Lee JS, Pham VH, Ryu KH. A comparative analysis of machine learning methods for class imbalance in a smoking cessation intervention. *Applied Sciences 2020*; 10(9): 3307. doi: 10.3390/app10093307

[40] Seo JH, Kim YH. A machine-learning approach to optimize smote ratio in class imbalance dataset for intrusion detection. *Computational Intelligence and Neuroscience 2018*; 2018: 9704672. doi: 10.1155/2018/9704672

[41] Ijaz MF, Alfian G, Syafrudin M, Rhee J. Hybrid Prediction Model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, Synthetic Minority Over Sampling Technique (SMOTE), and random forest. *Applied Sciences 2018*; 8(8): 1325. doi: 10.3390/app8081325

Rajan Saluja

I am Rajan Saluja, a research scholar at MMICTBM, MMDU, Mullana, Ambala, India. I am also working as an assistant professor at UIC, Chandigarh University. The research area I am interested in, is Ensemble Machine Learning Models. I am having ACM professional membership and 5 publications related to the field. My emailid is rajansaluja@yahoo.com

Dr. Munishwar Rai

He is professor at MMICTBM, MMDU, Mullana, Ambala. He is having more 25 years experience in teaching and more than 10 years experience in research field related to machine learning and its applications. He has more 40 publication in Scopus indexed, WOS, and UGC approved Journals and conference proceedings.