SASTRA

ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION

DEEMED TO BE UNIVERSITY

(U/S 3 of the UGC Act, 1956)

THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

THANJAVUR | KUMBAKONAM | CHENNAI

Students' Performance Prediction using Machine Learning algorithms

Under the guidance of :

Mr. EASHWAR K B

Asst. Professor – II

Department of CSE

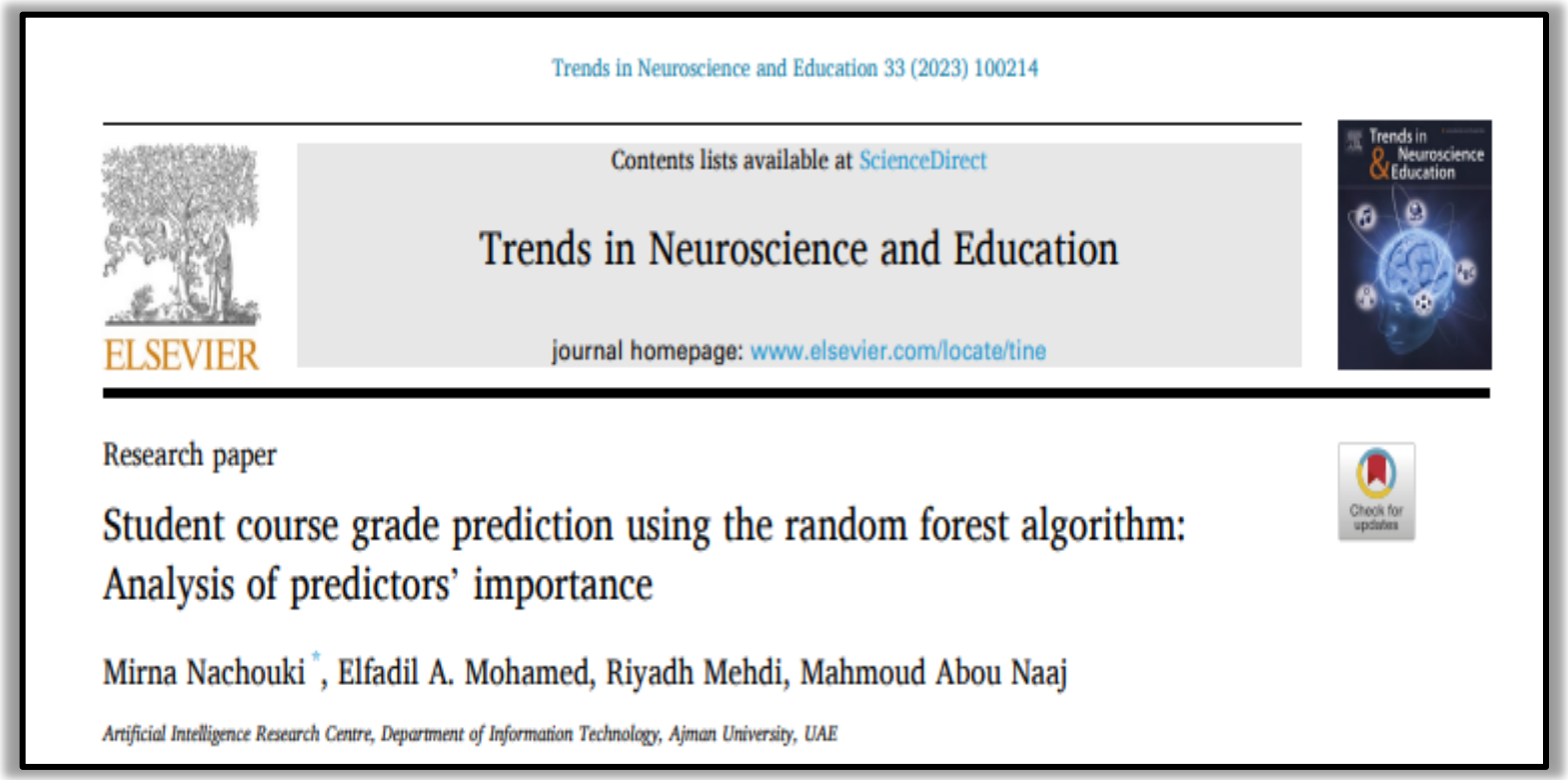
Presented by :

Suwetha S

224058033

II-M.Sc. Computer Science

BASE PAPER



Trends in Neuroscience and Education 33 (2023) 100214

Contents lists available at ScienceDirect

Trends in Neuroscience and Education

journal homepage: www.elsevier.com/locate/tine

Research paper

Student course grade prediction using the random forest algorithm: Analysis of predictors' importance

Mirna Nachouki *, Elfadil A. Mohamed, Riyadh Mehdi, Mahmoud Abou Naaj

Artificial Intelligence Research Centre, Department of Information Technology, Ajman University, UAE

OBJECTIVE

Students' performance prediction uses analytical methods, including machine learning and statistics, to estimate students' future academic performance. It predicts student performance by considering the parameters that past-grades, attendance records, socio-economic background, and study attitudes and health records. By analyzing different datasets derived from that student transcripts

ABSTRACT

This proposed project is aimed to predict the students' performance using the following machine learning algorithms: Random Forest, Support Vector Machine (SVM) and Gradient Boosting. By analyzing different dataset derived from student transcripts, it is aimed to understand how useful each algorithm is in forecasting the students' academic performance. Through a detailed validation the strengths and limitations of these predictive abilities of Random Forest, Gradient Boosting and Support Vector Machine (SVM) within the environment of student academic performance. This study increases to the conversation on improving predictive methodologies in education, providing valuable understanding for institutions looking to increase student success through knowledgeable decision making and personalized interventions.

LITERATURE REVIEW

Year	Title	Journal Name	Techniques used	Limitations
2018	Early segmentation of students according to their academic performance: A predictive modelling approach	Decision Support Systems	decision trees, support vector machines, naive Bayes, bagged trees and boosted trees	The data is from 2003 to 2015, so the study may not be very helpful for understanding the current education situation since things may have changed
2019	Prediction of academic performance associated with internet usage behaviors using machine learning algorithms	Computers in Human Behavior	decision tree, neural network and support vector machine	None of the problems or difficulties related to this approach are discussed in this paper

CONT.,

Year	Title	Journal Name	Techniques used	Limitations
2020	Multiple Features Fusion Attention Mechanism Enhanced Deep Knowledge Tracing for Student Performance Prediction	IEEE Access	Recurrent Neural Network (RNN)	What should be done next or how to improve the study is not discussed in this paper
2021	Multiclass Prediction Model for Student Grade Prediction Using Machine Learning	IEEE Access	Decision Tree, Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbor (kNN), Logistic Regression and Random Forest	This paper discusses two approaches for selecting particular characteristics, but it provides insufficient information about these approaches

CONT.,

Year	Title	Journal Name	Techniques used	Limitations
2022	A machine learning prediction of academic performance of secondary school students using radial basis function neural network	Trends in Neuroscience and Education	Bayes net, decision tree, k-nearest neighbors, logistic regression, naive Bayes, random forest and random tree	The data is only from students in Bangladesh, making it unsuitable for use in any other educational setting or location
2023	Machine Learning Algorithms based Student Performance Prediction based on Previous Records	IEEE Access	Bayesian classification	The Bayesian classifier technique is only used in this paper

WORKFLOW


Data Collection
Pre-Processing
Applying Algorithms
Prediction

DATASET DESCRIPTION

Sample Dataset

Source: Real-time Data from Students'

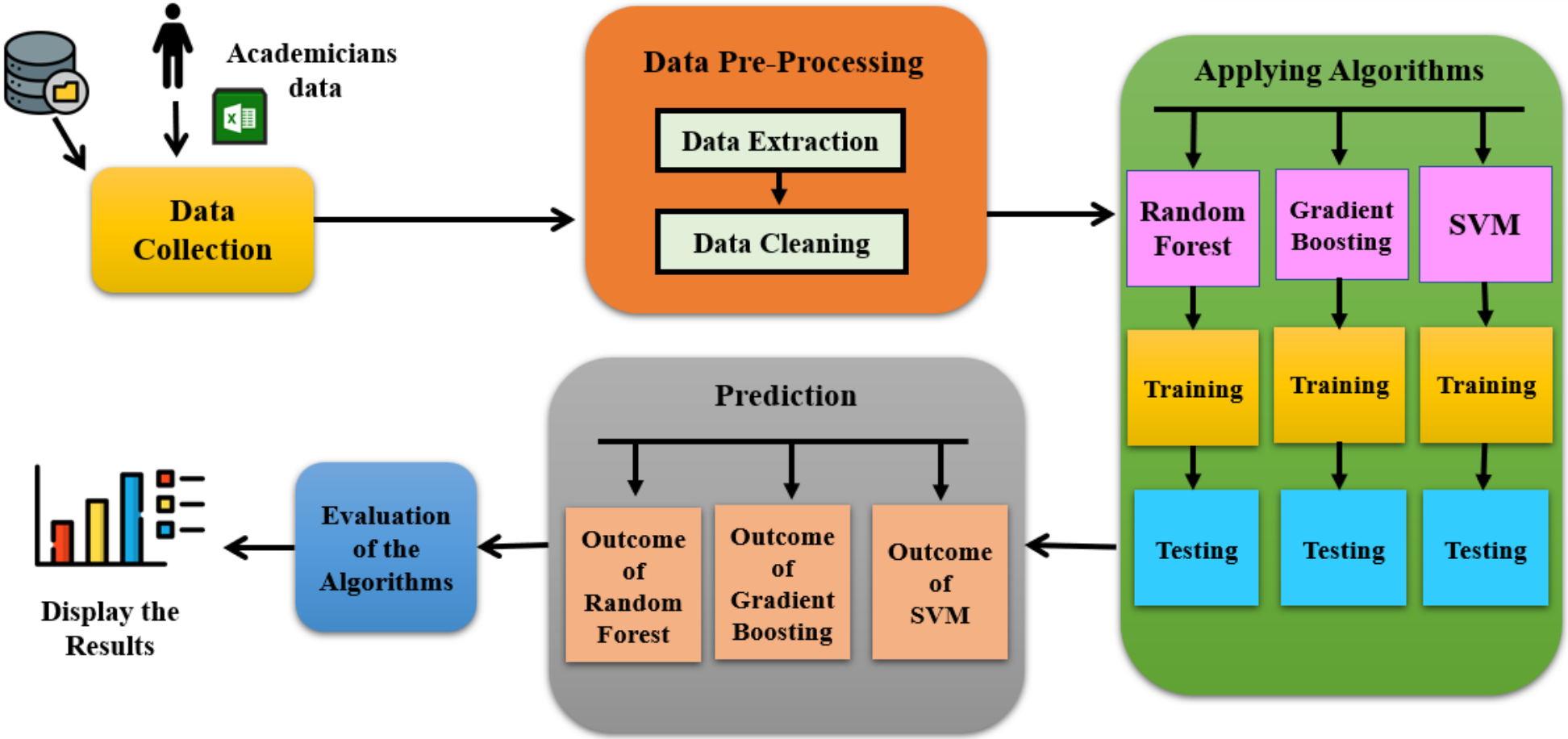
A	B	C	D	E	F	G	H	I	J	K
Name	Register No	Age	Gender	Your Current Course of Study	10th Board of Education	12th Board of Education	12th Specialization	Mode of Study	If you are a Day Scholar and traveling some km (kilometres) on every day, What are the problems you face ?	What is your current attendance percentage(%) ?
Aarthika S	224026002	19	Female	3rd B.Com	State Board	State Board	Commerce with Computer application	Day Scholer	Headache	80 - 90
Abinaya K	224026004	19	Female	3rd B.Com	State Board	State Board	Commerce with Accountancy	Day Scholer	Weight loss because of traveling	90 above
Abinaya.M	225026005	20	Female	3rd B.Com	State Board	State Board	Commerce with Accountancy	Day Scholer	Headache	80 - 90
Abinaya R	224026009	20	Female	3rd B.Com	State Board	State Board	Commerce with Accountancy	Day Scholer	Weight loss because of traveling	70 - 80

CONT.,

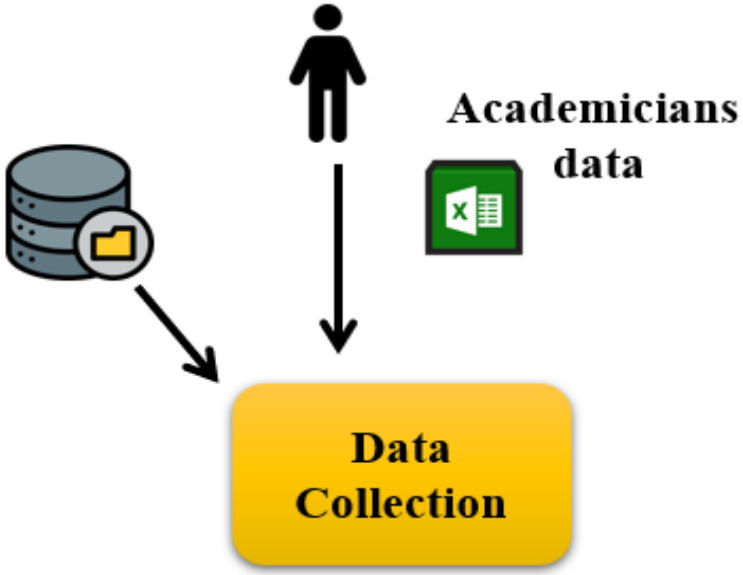
L	M	N	O	P	Q	R	S
How is your Grade Point Average calculated ?	Does family situations affect your studies?	If you said "Yes", which factor influences you ?	What is your family's annual income?	If you get a low grade in a course, what do you think could be the reason ?	If you are less focused on your studies and your friends score higher than you, does that discourage you?	What kind of course do you struggle with the most?	Why are you struggling with that course?
Test in classes	Maybe	Education Investment	10000 upto 50000	Hopelessness	No	Managing Business Process	Lack of interest in course
Internal Marks	No	Nothing	10000 upto 50000	Anxiety	No	Business Law	Difficulty Level
Internal Marks	No	Nothing	10000 upto 50000	Loneliness	No	Insurance	Difficulty Level
Assignments	No	Education Investment	50000 upto 1,00,000	Fear about Future	No	Business Law	Difficulty Level

T	U	V	W	X	Y	Z	AA	AB	AC
1st Sem SGPA	1st Sem CGPA	2nd Sem SGPA	2nd Sem CGPA	3rd Sem SGPA	3rd Sem CGPA	4th Sem SGPA	4th Sem CGPA	5th Sem SGPA	5th Sem CGPA
6.1539	6.1539	8	7.1273	8	7.4146	8.3929	7.6636	8.7037	7.8686
4.6539	4.6539	6.931	6.4	7.7037	6.8293	7.4643	6.9909	7.7037	7.1314
7.6539	7.6539	8.2414	7.9636	8.0741	8	7.8571	7.9636	8.6667	8.1022
7.2692	7.2692	7.8966	7.6	7.8519	7.6829	7.8929	7.7364	8	7.7883

PROPOSED ARCHITECTURE



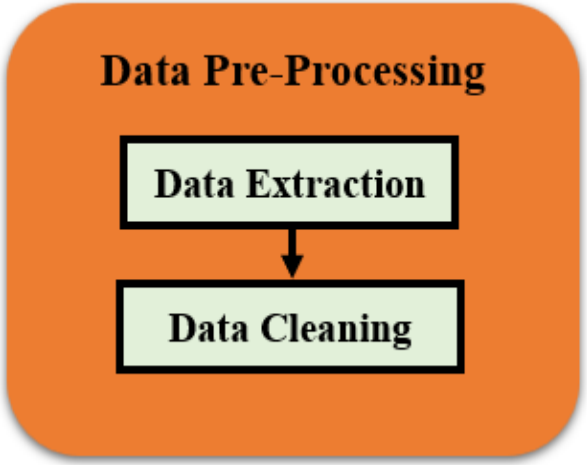
MODULE 1 : DATA COLLECTION



Data Collection

- The information was gathered from students using a questionnaire with some questions
- Additionally, the information was collected from the student academic repository. This was the way the data were collected

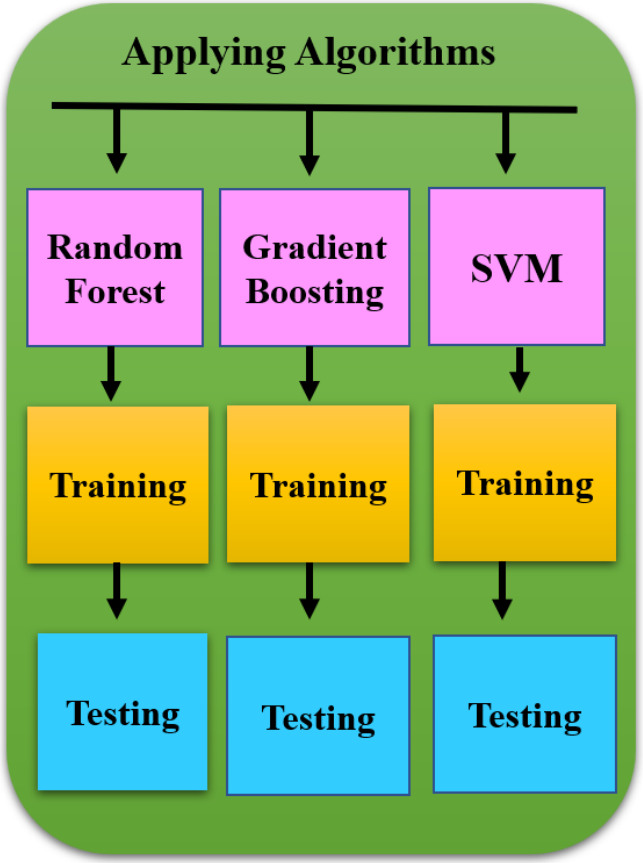
MODULE 2 : DATA PRE-PROCESSING



In this stage, using the following two methods, Data Extraction and Data Cleaning, the Data Pre-Processing was performed.

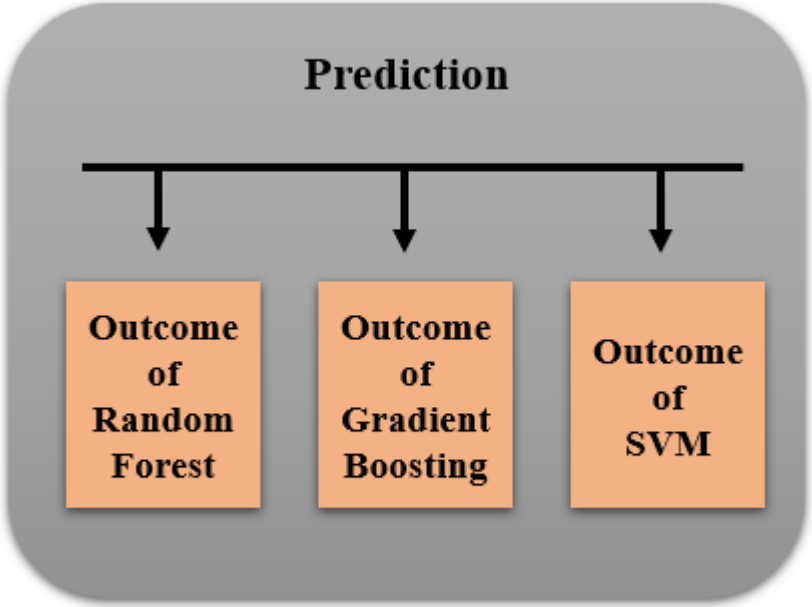
- **Data Extraction:** It was the process of collecting data samples from different sources. For example, Student academic repository and Spreadsheets documents
- **Data Cleaning:** A dataset sometimes contained entries with incomplete information. Those datasets or entries were not regarded for further processing. There existed specific techniques to impute fill in the missing data for these incomplete entries

MODULE 3 : APPLYING ALGORITHMS



- In this stage, various machine learning algorithms were applied: Random Forest, Gradient Boosting, and Support Vector Machine (SVM)
- Additionally, the dataset was divided into two subsets: one for training the models and another for testing the models, both derived from the same dataset

MODULE 4 : PREDICTION

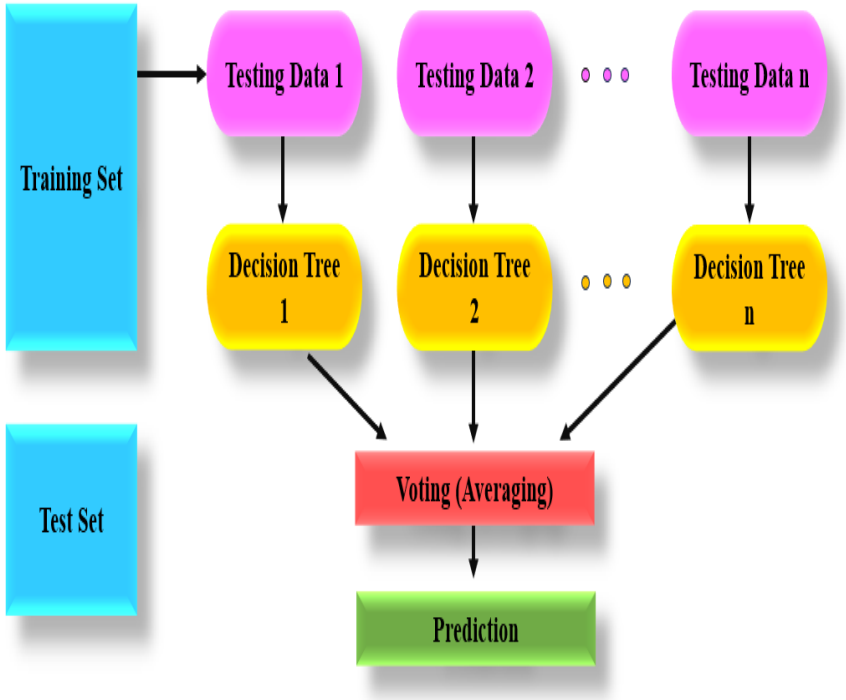


- In this stage, the outcomes of the Random Forest algorithm, Gradient Boosting algorithm, and Support Vector Machine (SVM) algorithm were obtained separately
- Predictions of results were obtained through various evaluation techniques

ALGORITHM USED

1) Random Forest

- Random Forest is an ensemble learning algorithm used for both classification and regression tasks
- It builds multiple decision trees during training and combines their predictions to make more accurate and robust predictions
- Each tree in the forest is trained on a random subset of the data and makes independent predictions and the final outputs are determined by a majority vote or averaging depending on the task



CONT.,

2) Gradient Boosting

- Gradient Boosting is a popular boosting algorithm in machine learning used for classification and regression tasks
- Gradient Boosting is also a powerful algorithm that utilizes boosting, where it constructs an ensemble of decision trees is acceptable for handling large datasets

3) Support Vector Machine

- A Support Vector Machine(SVM) is a supervised machine learning algorithm used for classification and regression tasks
- It works by finding the hyperplane that best separates data points into different classes, with a maximum margin between the classes

OUTPUT FOR 3rd YEAR

For 3rd Year Students':

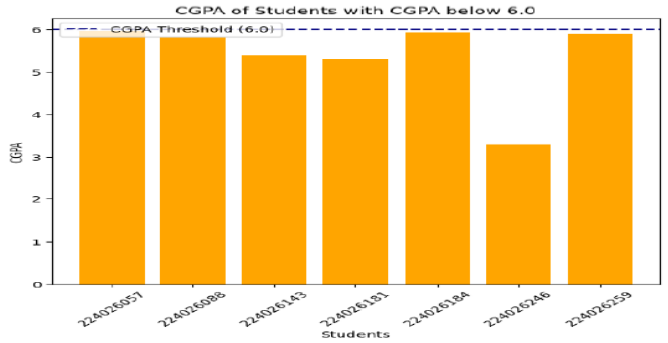
Students who are below 6.0 CGPA:

	Register_no	Age	Gender	Your Current Course of Study	10th BoE	12th BoE	12th Specialization	MS	Problems_faced	attendance	...
12	224026057	19	0	1	1	1	1	1	0	2	...
20	224026088	19	1	1	1	1	2	1	3	3	...
33	224026143	19	0	1	1	1	2	1	4	1	...
38	224026181	20	1	1	1	1	1	1	1	1	...
40	224026184	20	0	1	1	1	2	1	3	2	...
55	224026246	19	1	1	1	1	2	0	0	2	...
59	224026259	19	0	1	1	1	2	1	0	2	...

	CGPA_1	SGPA_2	CGPA_2	SGPA_3	CGPA_3	SGPA_4	CGPA_4	SGPA_5	CGPA_5
12	4.6154	6.0000	4.6727	6.0000	5.3049	6.3571	6.0364	5.6296	5.9562
20	3.5000	5.1724	5.6182	5.9630	5.7120	5.1786	5.7727	5.3704	5.8102
33	5.2195	5.4615	4.6071	4.5862	5.3909	5.0000	4.7407	4.9259	5.3796
38	4.9678	5.5862	5.3010	5.3704	5.2000	4.7857	5.1270	5.9000	5.3212
40	5.7683	3.4231	6.0714	3.7931	5.8455	4.7818	6.2593	5.8148	5.9270
55	5.2000	3.5000	3.5000	4.3000	4.6000	3.5000	3.2000	3.8000	3.3000
59	5.3462	5.3448	5.3455	4.4074	5.0366	5.4286	5.8546	5.4444	5.8905

	target
12	0
20	0
33	0
38	0
40	0
55	0
59	0

[7 rows x 28 columns]

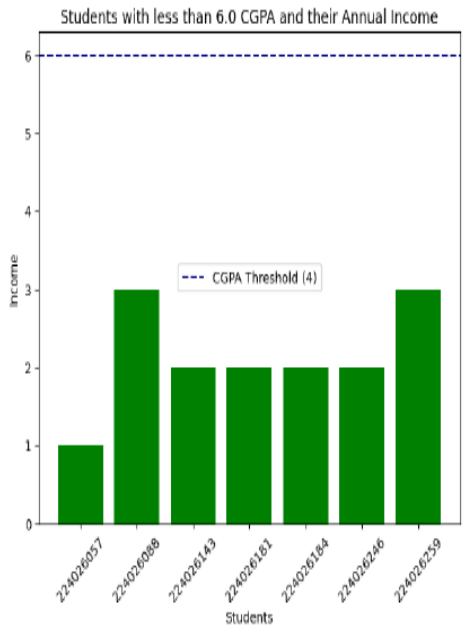


3rd year students' with CGPA below 6.0

Students with less than 6.0 CGPA and their Annual Income :

12	1
20	3
33	2
38	2
40	2
55	2
59	3

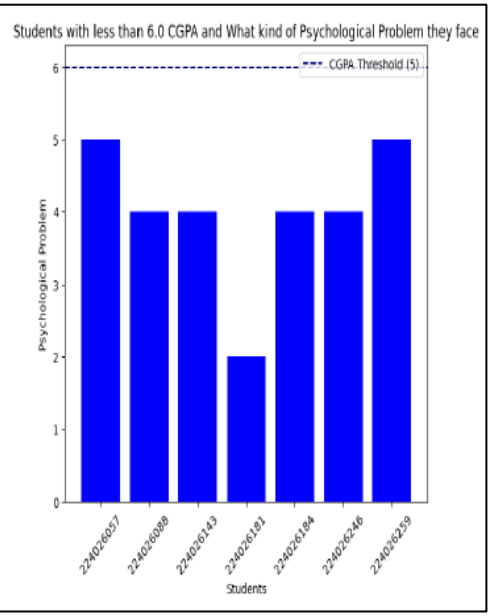
Name: income, dtype: int64



3rd year students' with CGPA below 6.0 and their annual income

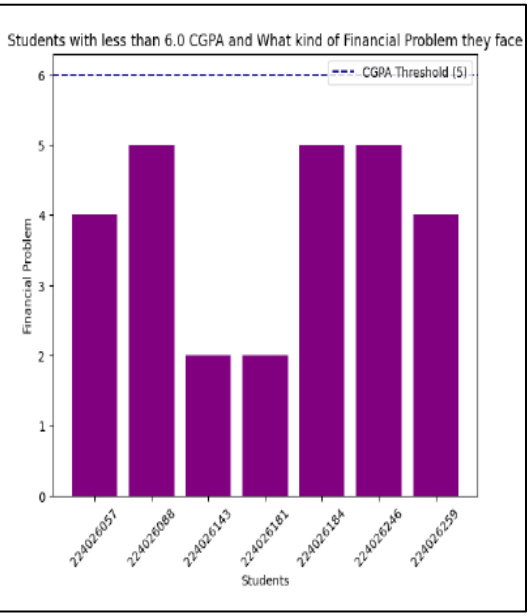
CONT.,

```
Students with less than 6.0 CGPA and What kind of Psychological Problem they face:

12    5
20    4
33    4
38    2
40    4
55    4
59    5
Name: psychological_issue, dtype: int64
```



3rd year students' with CGPA below 6.0 and what kind of Psychological problems they may face

```
Students with less than 6.0 CGPA and what kind of Financial Problem they face:

12    4
20    5
33    2
38    2
40    5
55    5
59    4
Name: financial_issue, dtype: int64
```

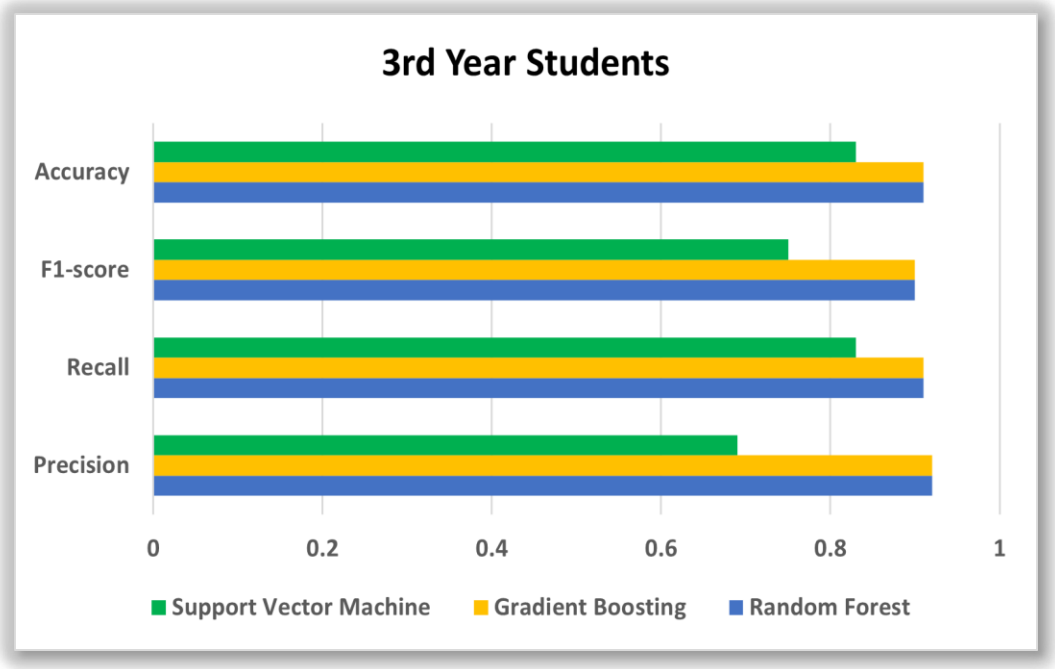


3rd year students' with CGPA below 6.0 and what kind of Financial problems they may face

CONT.,

3rd Year Students' Results:

3rd Year Students'				
Algorithms	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Random Forest	0.92	0.91	0.9	0.91
Gradient Boosting	0.92	0.91	0.9	0.91
Support Vector Machines	0.69	0.83	0.75	0.83



OUTPUT FOR 2nd YEAR

For 2nd Year Students':

Students who are below 6.0 CGPA:

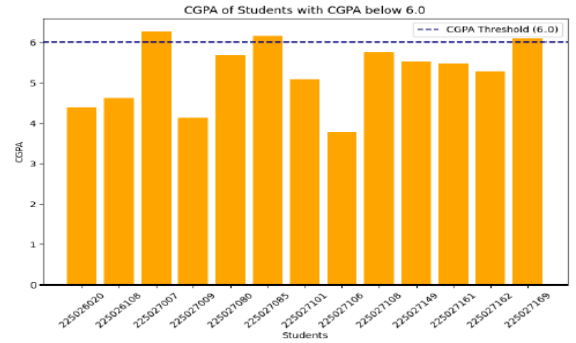
	Register_no	Age	Gender	Your Current Course of Study	10th BoE
1	225026020	19	0	1	1
6	225026108	19	1	1	1
12	225027007	20	0	1	2
13	225027009	18	1	1	1
34	225027080	19	0	1	1
35	225027085	20	1	1	1
40	225027101	19	1	1	1
42	225027106	19	0	1	1
43	225027108	19	1	1	1
57	225027149	19	1	1	1
62	225027161	19	0	1	1
63	225027162	18	0	1	2
65	225027169	19	1	1	1

	12th BoE	12th Specialization	MS	Problems_faced	attendance	...
1	1	1	1	4	1	...
6	1	1	1	3	3	...
12	2	5	1	3	3	...
13	1	4	1	4	2	...
34	1	4	0	0	1	...
35	1	4	1	4	2	...
40	1	5	1	2	2	...
42	1	4	1	2	2	...
43	1	4	0	0	1	...
57	1	4	1	2	1	...
62	1	4	1	2	1	...
63	1	5	1	3	3	...
65	1	4	1	4	1	...

	friend_discourage	course_struggle	why_struggle	SGPA_1	CGPA_1	SGPA_2
1	0	4	5	4.3846	4.3846	5.5862
6	0	5	4	5.7692	4.6154	6.0000
12	1	8	5	6.2800	6.2800	5.9600
13	0	9	4	4.1200	4.1200	5.1200
34	1	9	2	5.6800	5.6800	5.6400
35	0	9	4	6.1600	6.1600	6.2800
40	0	9	4	5.0800	5.0800	6.2000
42	0	9	1	3.7600	3.7600	3.2400
43	1	9	2	5.7600	5.7600	4.1000
57	0	9	5	5.5200	5.5200	6.0400
62	0	9	1	5.4800	5.4800	6.2000
63	1	9	4	5.2800	5.2800	5.0800
65	0	4	5	6.1000	6.1000	5.4000

	CGPA_2	SGPA_3	CGPA_3	target
1	5.5818	6.0741	5.7439	0
6	4.6727	6.0000	5.3849	0
12	6.1200	5.4000	5.8800	0
13	5.2600	3.6400	4.9867	0
34	5.6600	3.6000	4.9733	0
35	6.2200	3.8800	5.4400	0
40	5.9600	5.3200	5.7400	0
42	3.5000	2.4000	3.1333	0
43	4.6600	3.6600	4.4600	0
57	5.7000	5.8400	5.8000	0
62	5.8400	4.7600	5.4800	0
63	5.1800	4.6800	5.0133	0
65	5.7000	5.0000	5.5000	0

[13 rows x 24 columns]

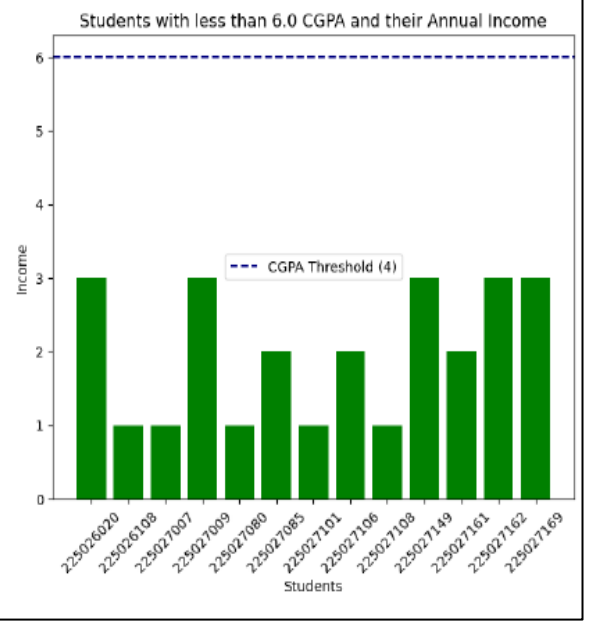


2nd year students' with CGPA below 6.0

Students with less than 6.0 CGPA and their Annual Income :

1	3
6	1
12	1
13	3
34	1
35	2
40	1
42	2
43	1
57	3
62	2
63	3
65	3

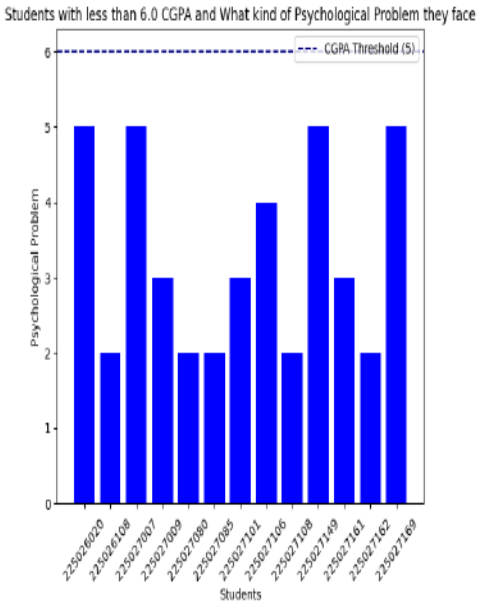
Name: income, dtype: int64



2nd year students' with CGPA below 6.0 and their annual income

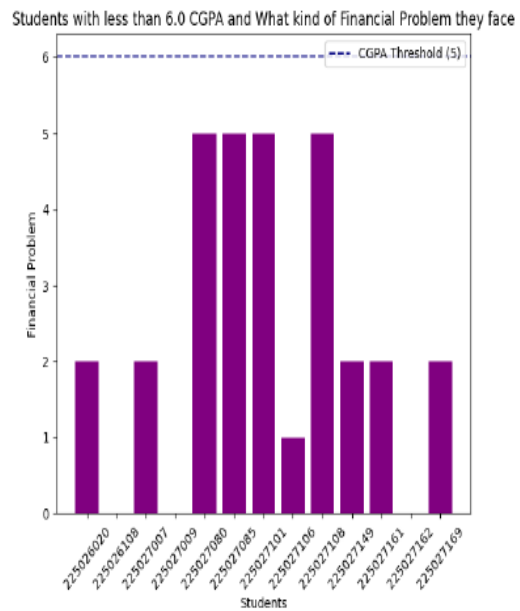
CONT.,

```
Students with less than 6.0 CGPA and What kind of Psychological Problem they face:

1     5
6     2
12    5
13    3
34    2
35    2
40    3
42    4
43    2
57    5
62    3
63    2
65    5
Name: psychological_issue, dtype: int64
```



2nd year students' with CGPA below 6.0 and what kind of Psychological problems they may face

```
Students with less than 6.0 CGPA and What kind of Financial Problem they face:

1     2
6     0
12    2
13    0
34    5
35    5
40    5
42    1
43    5
57    2
62    2
63    0
65    2
Name: financial issue, dtype: int64
```

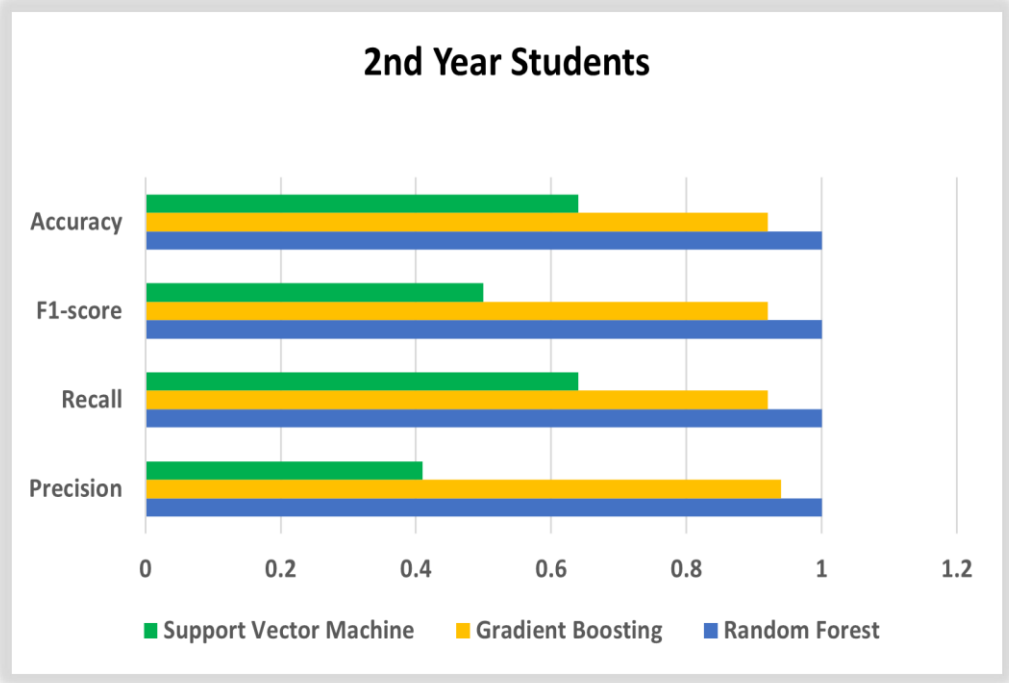


2nd year students' with CGPA below 6.0 and what kind of Financial problems they may face

CONT.,

2nd Year Students' Results:

2nd Year Students'				
Algorithms	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Random Forest	1.0	1.0	1.0	1.0
Gradient Boosting	0.94	0.92	0.92	0.92
Support Vector Machines	0.41	0.64	0.5	0.64



OUTPUT FOR 1st YEAR

For 1st Year Students':

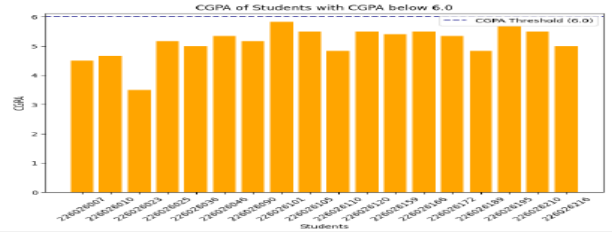
Students who are below 6.0 CGPA:

	Register no	Age	Gender	Your Current Course of Study	10th BoE
2	226026007	17	1	1	1
4	226026010	18	1	1	2
7	226026023	17	0	1	1
8	226026025	17	0	1	1
13	226026036	17	1	1	1
15	226026046	17	1	1	1
24	226026090	17	0	1	1
27	226026101	17	1	1	1
29	226026105	18	1	1	1
31	226026110	17	1	1	1
34	226026120	17	0	1	1
39	226026159	17	0	1	1
40	226026166	17	0	1	1
42	226026172	17	1	1	1
45	226026189	17	0	1	1
46	226026195	18	0	1	1
50	226026210	17	1	1	1
53	226026216	17	1	1	1

	12th BoE	12th Specialization	MS	Problems_faced	attendance
2	1	1	0	3	2
4	1	3	1	3	2
7	1	1	0	3	2
8	1	3	0	4	3
13	1	2	0	2	3
15	1	2	0	1	3
24	1	1	0	4	1
27	1	1	0	1	2
29	1	1	0	2	2
31	1	1	0	1	2
34	1	1	0	0	3
39	1	1	0	0	3
40	1	2	0	3	2
42	1	1	0	2	2
45	1	3	0	2	2
46	1	1	0	3	2
50	1	1	0	2	3
53	1	1	0	4	1

	fam_situations	financial_issue	income	psychological_issue
2	0	0	3	1
4	2	3	3	5
7	2	1	2	3
8	0	5	2	4
13	0	3	2	5
15	2	0	3	5
24	2	2	3	5
27	2	0	2	5
29	0	3	3	5
31	2	2	4	2
34	0	0	3	5
39	1	4	3	4
40	1	5	2	4
42	2	4	3	5
45	1	5	3	4
46	1	5	2	5
50	0	0	3	2
53	0	0	3	0

	friend_discourage	course_struggle	why_struggle	SGPA_1	CGPA_1	target
2	0	1	5	4.5000	4.5000	0
4	0	5	4	4.6667	4.6667	0
7	0	2	1	3.5000	3.5000	0
8	0	1	4	5.1667	5.1667	0
13	0	4	1	5.0000	5.0000	0
15	0	5	5	5.3333	5.3333	0
24	0	4	5	5.1667	5.1667	0
27	0	4	4	5.8333	5.8333	0
29	0	3	1	5.5000	5.5000	0
31	0	2	1	4.8333	4.8333	0
34	0	5	4	5.5000	5.5000	0
39	0	4	5	5.4000	5.4000	0
40	0	3	3	5.5000	5.5000	0
42	0	3	5	5.3333	5.3333	0
45	0	3	4	4.8333	4.8333	0
46	0	5	5	5.6667	5.6667	0
50	0	5	4	5.5000	5.5000	0
53	0	3	1	5.0000	5.0000	0

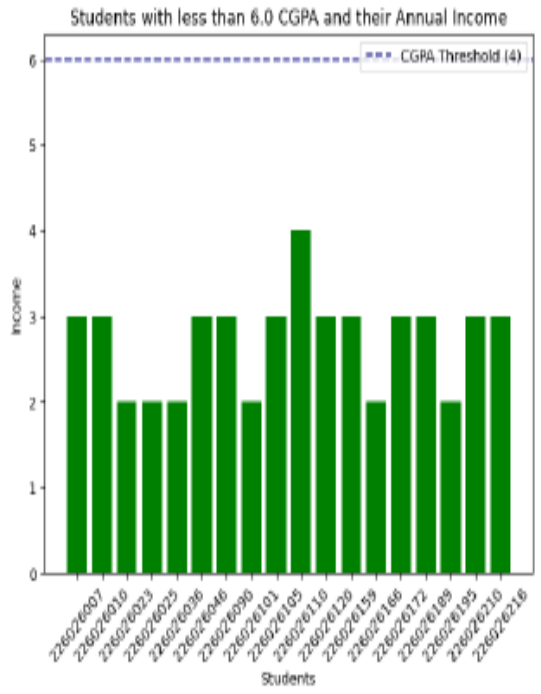


1st year students' with CGPA below 6.0

Students with less than 6.0 CGPA and their Annual Income :

2	3
4	3
7	2
8	2
13	2
15	3
24	3
27	2
29	3
31	4
34	3
39	3
40	2
42	3
45	3
46	2
50	3
53	3

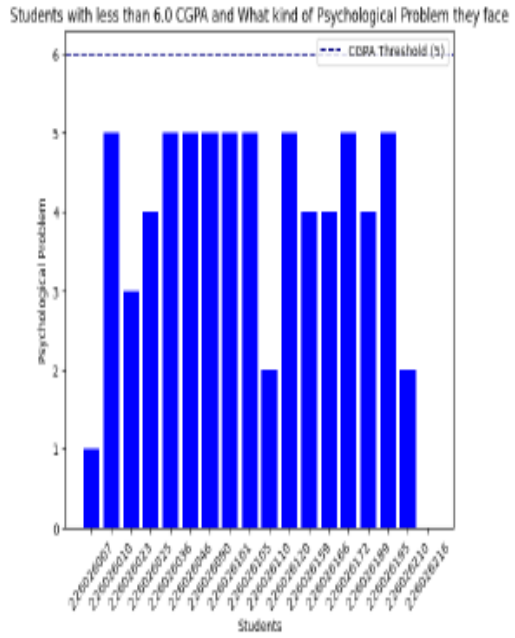
Name: income, dtype: int64



1st year students' with CGPA below 6.0 and their annual income

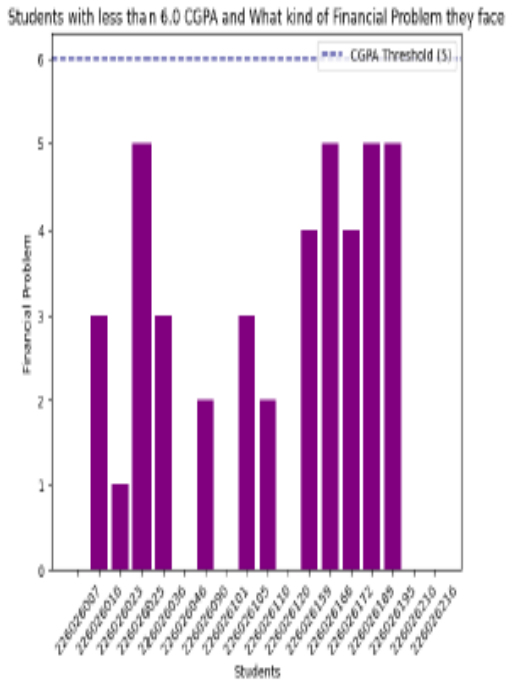
CONT.,

```
Students with less than 6.0 CGPA and What kind of Psychological Problem they face:

2    1
4    5
7    3
8    4
13   5
15   5
24   5
27   5
29   5
31   2
34   5
39   4
40   4
42   5
45   4
46   5
50   2
53   0
Name: psychological_issue, dtype: int64
```



1st year students with CGPA below 6.0 and what kind of Psychological problems they may face

```
Students with less than 6.0 CGPA and What kind of Financial Problem they face:

2    0
4    3
7    1
8    5
13   3
15   0
24   2
27   0
29   3
31   2
34   0
39   4
40   5
42   4
45   5
46   5
50   0
53   0
```

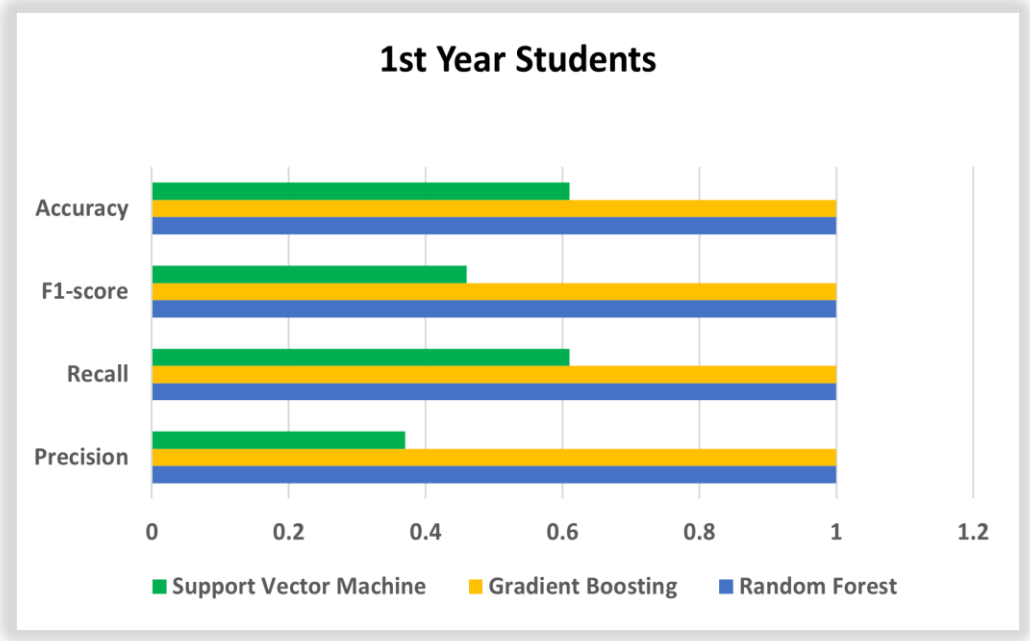


1st year students with CGPA below 6.0 and what kind of Financial problems they may face

CONT.,

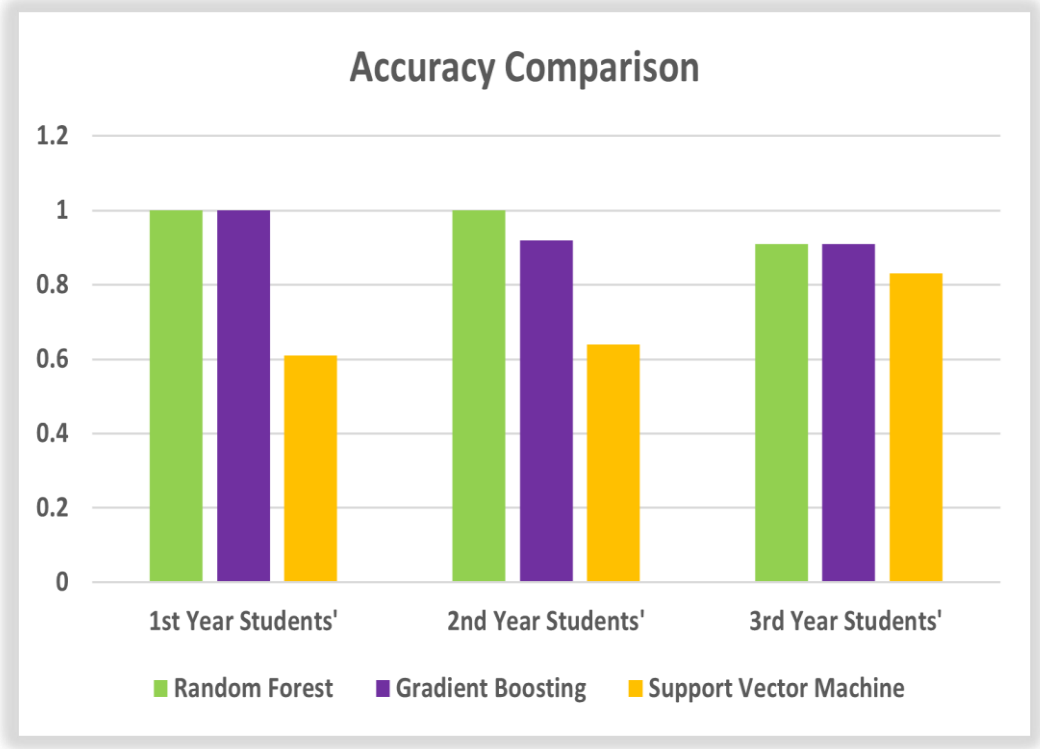
1st Year Students' Results:

1st Year Students'				
Algorithms	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Random Forest	1.0	1.0	1.0	1.0
Gradient Boosting	1.0	1.0	1.0	1.0
Support Vector Machines	0.37	0.61	0.46	0.61



ACCURACY COMPARISON

Model	1st Year Students'	2nd Year Students'	3rd Year Students'
Random Forest	1.0	1.0	0.91
Gradient Boosting	1.0	0.92	0.91
Support Vector Machine	0.61	0.64	0.83



CONCLUSION

In conclusion, for the prediction of student performance using machine learning algorithms, we gathered information from students to construct our dataset. We employed three machine learning algorithms: Random Forest, Gradient Boosting, and Support Vector Machine. Among these algorithms, Random Forest exhibited the highest accuracy. Our dataset was divided into three partitions: 1st year, 2nd year, and 3rd year. For 1st-year students, we achieved comparable levels of accuracy with both Random Forest and Gradient Boosting. In the case of 2nd-year students, Random Forest yielded the highest accuracy. Similarly, for 3rd-year students, Random Forest and Gradient Boosting demonstrated equal levels of accuracy. Overall, Random Forest consistently outperformed the other algorithms in terms of accuracy. Furthermore, this analysis can assist education management in identifying student performance based on various factors such as family situation, psychological aspects, and financial issues.

Requirements

Hardware Specifications :

Processor	:	12th Gen Intel(R) Core(TM) i51235U 1.30 GHz
Hard Disk	:	512 GB
RAM	:	8.00 GB

Software Specification :

OS	:	Windows 11
Programming Language	:	Python
Dataset	:	Real Time Data