STUDENTS' PERFORMANCE PREDICTION USING MACHINE LEARNING ALGORITHMS

Submitted in partial fulfillment of the requirements for the award of the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

Submitted by

SUWETHA S 224058033

Under the guidance of

Mr. K. B. EASHWAR, Assistant Professor, Department of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SASTRA DEEMED TO BE UNIVERSITY

(A University under section 3 of the UGC Act, 1956)

Srinivasa Ramanujan Centre

Kumbakonam - 612001

Tamil Nadu, India

May 2024

SHANMUGHA ARTS, SCIENCE, TECHNOLOGY & RESEARCH ACADEMY

(SASTRA DEEMED TO BE UNIVERSITY)

(A University established under section 3 of the UGC Act, 1956)

Srinivasa Ramanujan Centre

Kumbakonam —612001

Tamil Nadu, India

BONAFIDE CERTIFICATE

Certified that this project work entitled "**STUDENTS' PERFORMANCE PREDICTION USING MACHINE LEARNING ALGORITHMS**" submitted to the Srinivasa Ramanujan Centre, SASTRA Deemed to be University, Kumbakonam – 612001 by **Suwetha S (224058033)** in partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE** work carried out under the guidance of **Shri. K. B. Eashwar** during the period December 2023 to May 2024.

Signature of Project Supervisor :

Name with Affiliation :

Date :

Project Viva Voce held on :

Examiner I Examiner II

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SHANMUGHA ARTS, SCIENCE, TECHNOLOGY & RESEARCH ACADEMY

SASTRA DEEMED TO BE UNIVERSITY

(A University established under section 3 of the UGC Act, 1956)

Srinivasa Ramanujan Centre

Kumbakonam - 612001

Tamil Nadu, India

DECLARATION

I submit this project work entitled "**STUDENTS' PERFORMANCE PREDICTION USING MACHINE LEARNING ALGORITHMS**" to Srinivasa Ramanujan Centre, SASTRA Deemed to be University, Kumbakonam – 612 001, in partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE**. I declare that this is my original work carried out under the guidance of **K. B. Eashwar,** Assistant Professor, Department of Computer Science and Engineering, Srinivasa Ramanujan Centre.

Place: Kumbakonam

Date:

Name: Signature:

Reg No:

ACKNOWLEDGEMENT

I pay my sincere pranams to God ALMIGHTY for his grace and infinite mercy and for showing on me his choicest blessings.

First, I would like to express my sincere thanks to our honourable Chancellor **Prof. R. Sethuraman**, Vice-Chancellor **Dr. S. Vaidhyasubramaniam** and Registrar **Dr. R. Chandramouli** for allowing me to be a student of this esteemed institution.

I express my deepest thanks to revered Dean **Dr. V. Ramaswamy**, SRC and respected Associate Dean **Dr. A. Alli Rani**, SRC for all their moral support and suggestions when required without any reservations.

I exhibit my pleasure in expressing my thanks to **Dr. V. Kalaichelvi**, Associate Professor, Department of Computer Science and Engineering, Srinivasa Ramanujan Centre, for her encouragement during our project work.

I exhibit my pleasure in expressing my thanks **Mr. K. B. Eashwar**, Assistant Professor, Department of CSE my guide for his ever-encouraging spirit and meticulous guidance for the completion of the project.

I would like to express my deep sense of gratitude to the project coordinators, **Dr. R. Sujarani** and **Dr. K. Rajesh**, Assistant Professor, Department of Computer Science and Engineering for their cordial support and meticulous guidance which enabled me to complete this project successfully.

I would like to thank the panel members **Dr. R. Thanuja** and **Dr. A. Sumathi** Assistant Professor, Department of CSE for their support and encouragement to complete the project successfully.

I owe my sincere thanks to all faculty members in the department who have directly or indirectly helped me in completing this project.

Without the support of my parents and friends, this project would never have become a reality. I owe my sincere thanks to all of them. I dedicate this work to all my well-wishers, with love and affection.

TABLE OF CONTENT:

STUDENTS' PERFORMANCE PREDICTION USING MACHINE LEARNING ALGORITHMS

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, SVM and Gradient Boosting 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**.** This research supports students, parents, and educators in obtaining an outlook on academic success. Based on the examination of the model, it is determined that maintaining a lifestyle focused on health is associated with positive correlations to academic success. Conversely, the presence of stress is identified as having an adverse effect on academic outcomes.

Keywords: Machine Learning, Student, Education, Academic, Random Forest.

1.1 INTRODUCTION:

Due to the exposure to the advancements of the technology and the entertainment at their fingertips, z-generation students are too vulnerable to these technological environments. Apart from that various other factors are taking control over them from reaching their goals in the academic domain. Most of the students are still being pertained to the domestic issues and their altered student cult. The technology has widened the portal of the world to perceive anything. Some are good to their improvement in their smartness and most of them are creating impact on their behavior, character, action, and even in their psychological state. For example, now-a-days students doesn't expect from their teachers or faculties to learn something new. Because of the various online courses and programmes, they can learn without their faculties or teachers. The technology is in their hands. But at the same time, there exist the dark side for this technology, which spoils their career, goals and other good sides of their life. This project aims to identify the students who are performing below the threshold level and try to find the cause of their lower performance in the academic related activities.

The primary aim is to find out the causes for the students' lower performance as well as to identify the most efficient algorithm among the three most familiar machine learning algorithms. They are, Decision Tree, Random Forest, and Gradient Boosting algorithms.

1.2 PROBLEM STATEMENT:

Create a machine learning model that forecasts student achievement based on a range of behavioral, academic, and demographic characteristics. The objective is to develop a predictive system that can forecast academic performance, graduation rates, and grade levels for students. Using past student data grades, attendance records, socioeconomic status, and other pertinent details the model ought to be able to spot trends and connections that affect student performance. Helping educators, administrators, and policymakers identify students who are at-risk early, implement focused interventions, and enhance overall educational results is the ultimate goal.

1.3 EXISTING WORK:

The existing approach focused on the essential variables that are affecting the successful completion rate of schooling. The researchers developed a model to predict course grades based on these characteristics by employing the random forest algorithm. The study's conclusions emphasized the significance of variables like high school graduation rate and grade point average in predicting course grades. The study also stressed the importance of class attendance percentage and course category as significant variables influencing student success. Overall, the research contributes to the advancement of educational data mining by uncovering factors influencing students' academic achievement and offering recommendations for enhancing graduation rates while reducing dropout rates in schools.

1.4 PROPOSED WORK:

The goal of the proposed research is to predict students' academic performance using machine learning techniques including gradient boosting, random forest, and support vector machines (SVM). To ascertain how well each algorithm predicts academic success, a variety of datasets collected from student transcripts will be analyzed. The study aims to determine the advantages and disadvantages of these predictive models in relation to student performance using stringent validation procedures. By providing insightful information to organizations looking to enhance student outcomes through customized interventions and well-informed decision-making, the research adds to the continuing conversation about improving predictive techniques in education. By examining the predictive capabilities of Random Forest, SVM, and Gradient Boosting, the study aims to provide students, parents, and educators with insights into factors influencing academic success.

Additionally, the initiative investigates the connection between lifestyle choices and academic achievement. It implies that sustaining a healthy lifestyle has a favorable correlation with academic achievement and that stress has a negative correlation with academic performance. This area of the study adds to our knowledge of the larger context that affects academic achievement and offers practical advice for fostering student success.

2. OBJECTIVE:

The objective of this project is to assess and contrast the predictive abilities of three machine learning algorithms: Random Forest, Support Vector Machine (SVM), and Gradient Boosting. To this end, a variety of datasets derived from student transcripts will be analyzed. Through extensive validation, the study endeavors to ascertain the benefits and drawbacks of each algorithm with respect to predicting student performance. In doing so, it seeks to advance the discourse regarding the enhancement of predictive techniques in education and offer valuable data to institutions seeking to enhance student achievement through customized interventions and informed decision-making**.**

3. SYSTEM REQUIREMENTS:

The proposed technique is executed in Google Colab. Google Colab was utilized for implementing the proposed technique through a web browser. The system configurations are 8GB RAM, Windows OS-11th, 64-bit operating system, x64-based processor, intel i5 processor with version 23H2.

4. LITERATURE SURVEY:

In several studies, machine learning techniques were used to predict the performance of students based on their academic outcomes. O.A. Olabanjo et al. [1] The paper developed a predictive model using a Radial Basis Function Neural Network (RBFNN) to forecast students' academic performance based on academic records, cognitive abilities, and psychomotor ratings. The model achieved an accuracy of 86.59% with full features including psychomotor and cognitive ratings, and 82.80% without these factors, demonstrating the importance of considering holistic student attributes in predicting academic success. V.L. Miguéis et al. [2] The paper focuses on early segmentation of students' academic performance using data mining techniques like decision trees, Random Forests, and Support Vector Machines (SVM). It proposes a two-stage approach to predict students' future academic success groups with an emphasis on time to degree completion. The study achieves high accuracy in predicting academic performance levels among engineering students. Shaikh Rezwan Rahman, et al. [3] The study utilized data mining algorithms such as logistic regression, multilayer perceptrons, and random forest to analyze the impact of co-curricular activities on academic performance, achieving high accuracies ranging from 94.654% to 99.5294%. The research focused on predicting students' performance through machine learning techniques, highlighting the correlation between CGPA and participation in extracurricular activities, with a 69% correlation observed. Hina Gull, et al. [4] To predict student grades, the researchers in this study used machine learning techniques such as support vector machines, logistic regression, linear discriminant analysis, K-nearest neighbors, regression and classification trees, and Gaussian Naive Bayes. In terms of predicting student performance results, the linear discriminant analysis technique proved to be the most accurate, with a prediction accuracy of 90.74%. Ulloa-Cazarez, R. L., et al. [5] In the study, Logistic Regression was utilized for predicting student placements based on academic data collected from Kaggle. Data pre-processing, model training, and validation were conducted in PyCharm. The accuracy of the classification using Logistic Regression was calculated, and a confusion matrix was created to assess the model's performance, achieving successful placement predictions with the model. Shelly Gupta and Jyoti Agarwal [6] In this paper, the work done includes data acquisition from the UCI repository, data pre-processing, feature selection, data visualization, and classification for student performance prediction. The algorithms used are K-Nearest Neighbors (KNN) and Logistic Regression. The accuracy achieved was 90.75% with KNN and 85.71% with Logistic Regression.

S. D. Abdul Bujang, N. A. Mat Isa, and N. A. Ismail [7] The research conducted a comprehensive analysis of machine learning techniques to predict student grades, focusing on improving predictive accuracy using Decision Tree (J48), Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Random Forest (RF) algorithms. The study also implemented oversampling SMOTE and feature selection methods to address imbalanced multi-classification issues, achieving high prediction performance with J48 and RF algorithms leading in precision scores. D. Liu, Yunping Zhang, and Jun Zhang [8] The work in the paper involves enhancing deep knowledge tracing for student performance prediction using a Multiple Features Fusion Attention Mechanism. The study utilizes techniques such as LSTM networks, Adam algorithm for training, and Pytorch implementation. The MFA-DKT framework outperforms

traditional models like BKT, achieving an AUC of 0.85 and an ACC of 0.79 in student performance prediction. A. Alshanqiti and A. Namoun [9] The work includes developing a hybrid regression model and an optimized multi-label classifier to predict student academic performance and identify influential factors. The algorithms and techniques used in the study involve collaborative filtering, fuzzy set rules, Lasso linear regression, and a weighted mean scheme with Self Organizing Map. The study achieved high accuracy in predicting student performance through the integration of these techniques, as demonstrated in extensive performance evaluations against benchmark datasets. Suchithra Rajendran, et al. [10] Using machine learning algorithms like logistic regression, artificial neural networks, random forests, gradient boosting, and stacking techniques, the study performed a predictive analysis on academic performance. The stacking method yielded the most accuracy, with logistic regression, random forest, gradient boosting, artificial neural network, and random forest coming in order of preference. The study's objective was to categorize and rank the variables that affect academic performance in order to give educational institutions useful information. Xing Xu, et al. [11] The researchers in this study examined actual internet usage data from 4000 students and extracted factors including online duration, internet traffic volume, and connection frequency in order to predict academic success. Neural networks, decision trees, and support vector machines were some of the machine learning techniques used to analyze the data. As the number of features increased, so did the prediction accuracy; SVM had the highest accuracy, at 72.75%.

5. METHODOLOGY:

5.1 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.

5.2 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.

5.3 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.

5.3.1 Random Forest:

Random Forest is a well-liked machine learning algorithm that is applied to supervised learning methods. Regression and classification-based machine learning challenges can be solved with it. Its foundation is the idea of ensemble learning, which is the process of merging several classifiers to enhance the model's performance and resolve a challenging issue. According to its name, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of depending just on one decision tree, the random forest makes forecasts based on the majority vote of projections from each one.

5.3.2 Gradient Boosting:

With gradient descent, each new model is trained to minimize the loss function, such as the mean squared error or cross-entropy of the preceding model. Gradient boosting is a potent boosting procedure that turns multiple weak learners into strong learners. The approach calculates the gradient of the loss function in relation to the current ensemble's predictions for each iteration and then trains a new weak model to minimize this gradient. The modified model's predictions are then included into the ensemble, and this process is repeated until a predefined condition is satisfied, at which point the process is considered to be finished.

5.3.3 Support Vector Machine (SVM):

Support vector machines, or SVM, are among the most popular supervised learning methods for problems involving both regression and classification. However, it's primarily used for classification problems in machine learning. To simplify the process of classifying new data points in the future, the SVM approach looks for the best line or boundary that can split n-dimensional space into classes. This ideal decision boundary is known as a hyperplane. To help create the hyperplane, SVM chooses the extreme vectors and points. Given that these extreme circumstances are referred to as support vectors, the method is thought of as a support vector machine.

5.4 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.

6. IMPLEMENTATION:

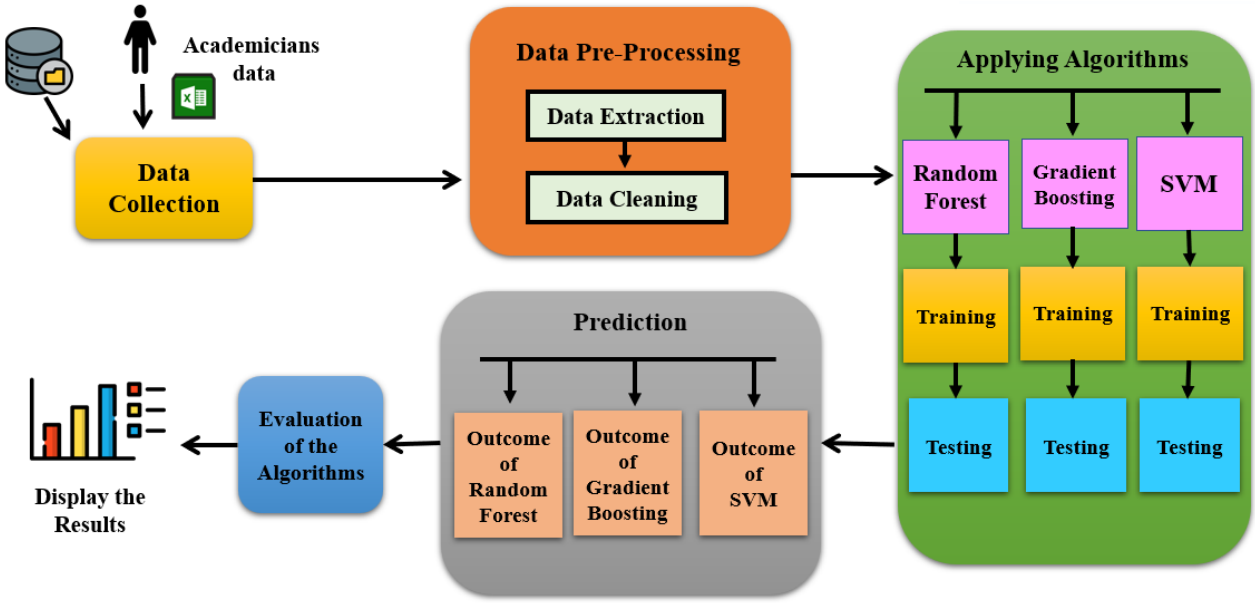

Figure 1: Block Diagram

First, data from students was collected using a carefully designed questionnaire with pertinent questions meant to elicit desirable insights. A Google Form was then created and distributed to enrolled students. Local students completed the offered Google Form to actively engage in this study.

Data preprocessing was the next step after data gathering. At this point, more features were gathered, but only those that were judged pertinent to the task were chosen, unnecessary features were left out. This procedure guaranteed a simplified dataset that was suitable for precise analysis. This data manipulation was made easy using Python programming and modules like Pandas and Numpy.

After the dataset was ready, different machine learning techniques were applied in the next stage. Among the techniques used to glean insights and patterns from the data were Random Forest, Gradient Boosting, and Support Vector Machines. The project's overall goals were enhanced by the use of these methods, which made it possible to explore the dataset's complicated linkages. Throughout the project, Python was the main programming language used, offering a stable environment for machine learning algorithm implementation. For data visualization, libraries like matplotlib.pyplot were used, especially for making bar graphs that effectively represented the data. These graphics were very helpful in clarifying important discoveries and improving the data' readability.

The final step of the prediction process was to obtain results from the algorithms that were put into practice. The performance of each algorithm was evaluated through extensive testing and evaluation, offering important insights into how effective they were for the task at hand. This phase was the project's apex, where the preprocessing, analysis, and data collection activities came together to produce suggestions and results that could be put into practice.

Looking back, the project's path required a methodical strategy that included gathering data, preprocessing, implementing algorithms, and making predictions. Every stage was carried out with great care, making use of the right instruments and techniques to guarantee that the project's goals were reached. Together with the use of cutting-edge tools and methods, the cooperation from the students' efforts enabled the project to be completed successfully and yielded insightful results.

7. PERFORMANCE ANALYSIS:

7.1 Accuracy:

Accuracy in machine learning is a measure of the model's ability to correctly classify instances. It is calculated as the ratio of correctly predicted instances to the total number of instances. A higher accuracy indicates better performance, but it may not be suitable for imbalanced datasets. It is a fundamental evaluation metric used to assess the overall effectiveness of a machine learning model.

Accuracy = Number of Correct Predictions / Total Number of Predictions

7.2 Precision:

Precision in machine learning is the proportion of true positive predictions among all positive predictions made by the model. It reflects the model's ability to avoid false positives, ensuring that when it predicts a certain outcome, it is highly likely to be correct. Precision is a crucial metric in tasks where false positives are costly or undesirable, such as medical diagnosis or fraud detection. Higher precision indicates better performance in correctly identifying positive instances.

Precision = True Positives / True Positives + False Positives

7.3 Recall:

Recall in machine learning measures the proportion of true positive predictions identified by the model out of all actual positive instances in the dataset. It gauges the model's ability to capture all relevant positive cases, minimizing false negatives. A higher recall indicates that the model can effectively identify most of the positive instances, even at the cost of more false positives. It is essential in tasks where missing positive instances are critical, such as medical diagnosis or anomaly detection.

Recall = True Positives / True Positives + False Negatives

7.4 F1- Score:

In machine learning, the F1-score is a harmonic mean of recall and precision. It is helpful for assessing model performance in binary classification problems because it offers a single score that strikes a balance between precision and recall. Better overall performance, taking into account both false positives and false negatives, is indicated by a higher F1-score. When there is an imbalance in the dataset between positive and negative cases, it is especially helpful.

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

7.5 Results:

7.5.1 For 1st Year Students':

Table 1: 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 Machine	0.37	0.61	0.46	0.61

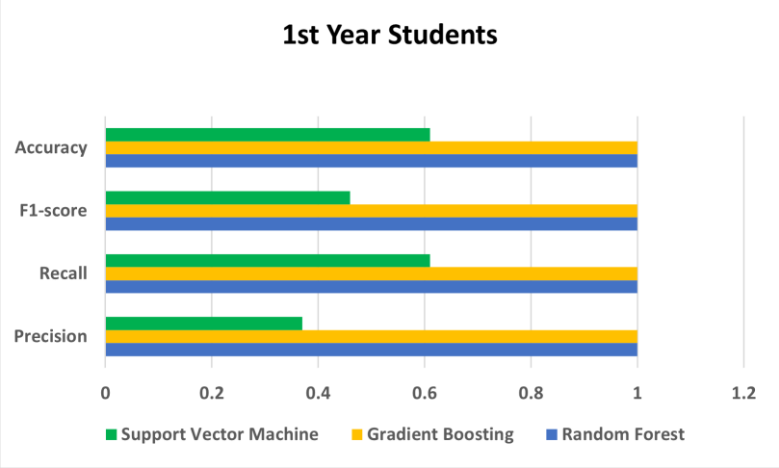

Figure 2: 1st Year Students' Results

7.5.1.1 Query Results for students' with CGPA below 6.0 and their corresponding parameters threshold values:

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	1	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

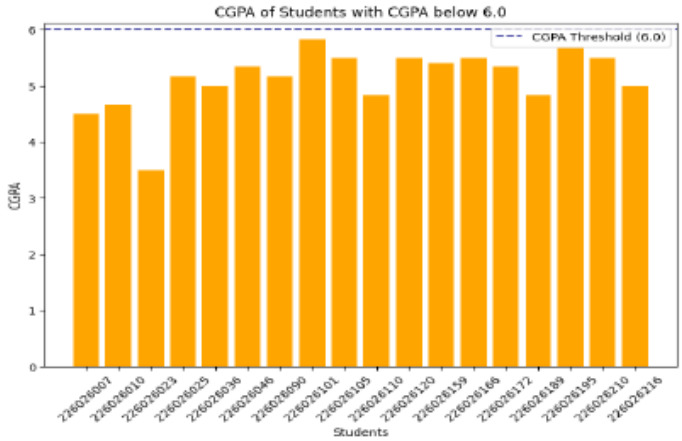

Figure 3: 1st year students' with CGPA below 6.0

7.5.1.2 The parameter “income” and its corresponding threshold value:

4 – 2,00,000 above
3 – 1,00,000 upto 2,00,000
2 – 50000 upto 1,00,000
1 – 10000 upto 50000

```
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
```

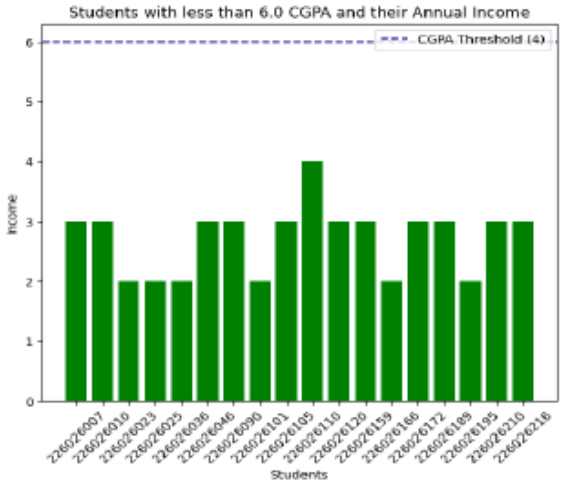

Figure 4: 1st year students’ with CGPA below 6.0 and their annual income

7.5.1.3 The parameter “psychological_issue” and its corresponding threshold value:

5 – Hopelessness
4 – Fear about Future
3 – Depression
2 – Anxiety
1 – Loneliness
0 – Nothing

```
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
```

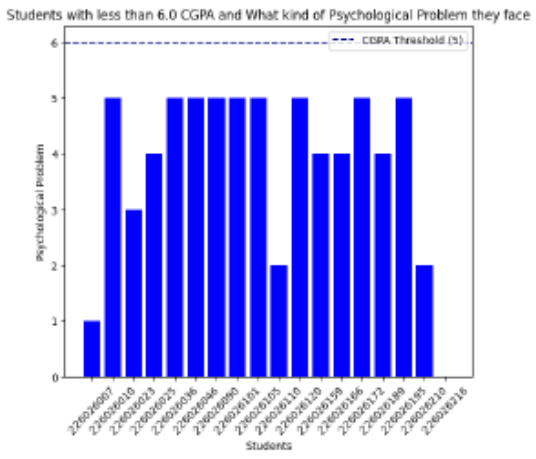

Figure 5: 1st year students with CGPA below 6.0 and what kind of Psychological problems they may face

7.5.1.4 The parameter “financial_issue” and its corresponding threshold value:

5 – Education Investment
4 – Family Income
3 – Father or Mother Health issue
2 – Misunderstanding between relation
1 – Assets related problems
0 – Nothing

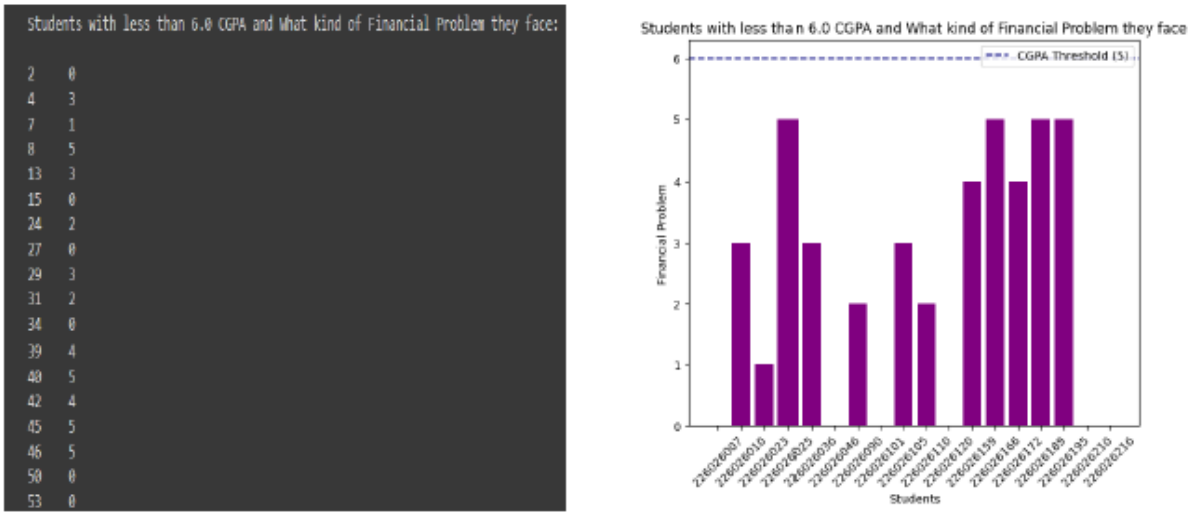

Figure 6: 1st year students with CGPA below 6.0 and what kind of Financial problems they may face

7.5.2 For 2nd Year Students’:

Table 2: 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 Machine	0.41	0.64	0.5	0.64

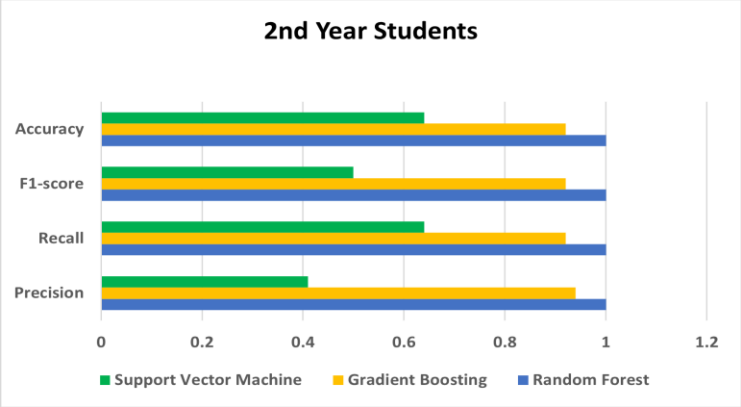

Figure 7: 2nd Year Students' Results

7.5.2.1 Query Results for students' with CGPA below 6.0 and their corresponding parameters threshold values:

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	6.2000	6.2000	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	5.3049	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]

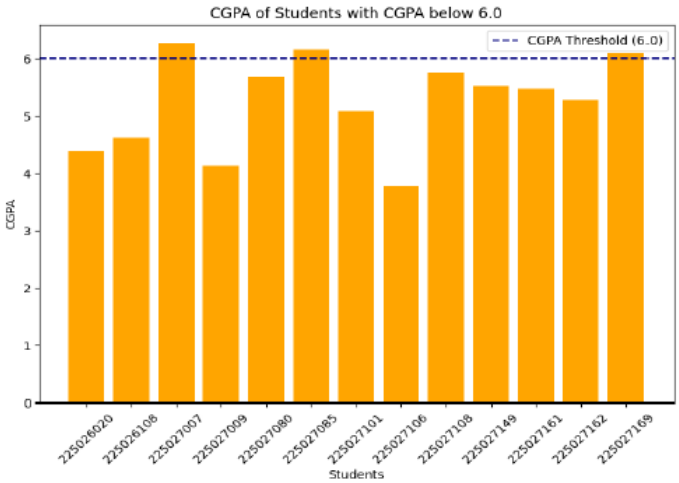

Figure 8: 2nd year students' with CGPA below 6.0

7.5.2.2 The parameter “income” and its corresponding threshold value:

4 – 2,00,000 above
3 – 1,00,000 upto 2,00,000
2 – 50000 upto 1,00,000
1 – 10000 upto 50000

```
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
```

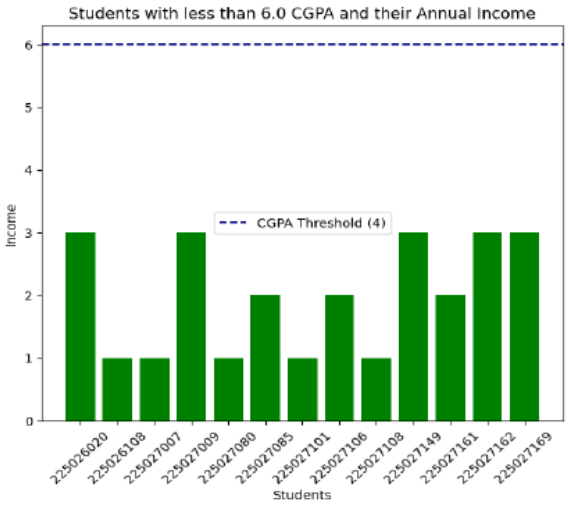

Figure 9: 2nd year students’ with CGPA below 6.0 and their annual income

7.5.2.3 The parameter “psychological_issue” and its corresponding threshold value:

5 – Hopelessness
4 – Fear about Future
3 – Depression
2 – Anxiety
1 – Loneliness
0 – Nothing

```
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
```

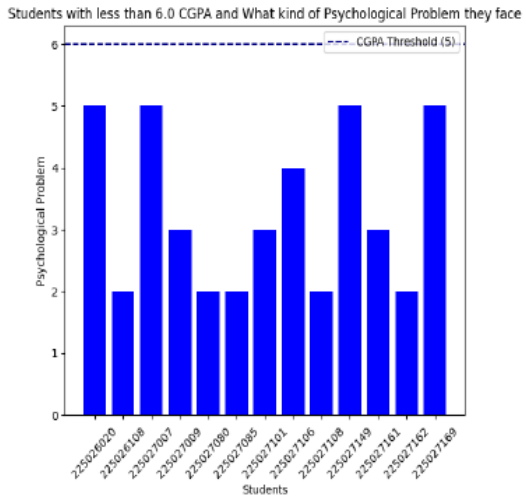

Figure 10: 2nd year students’ with CGPA below 6.0 and what kind of Psychological problems they may face

7.5.2.4 The parameter “financial_issue” and its corresponding threshold value:

5 – Education Investment
4 – Family Income
3 – Father or Mother Health issue
2 – Misunderstanding between relation
1 – Assets related problems
0 – Nothing

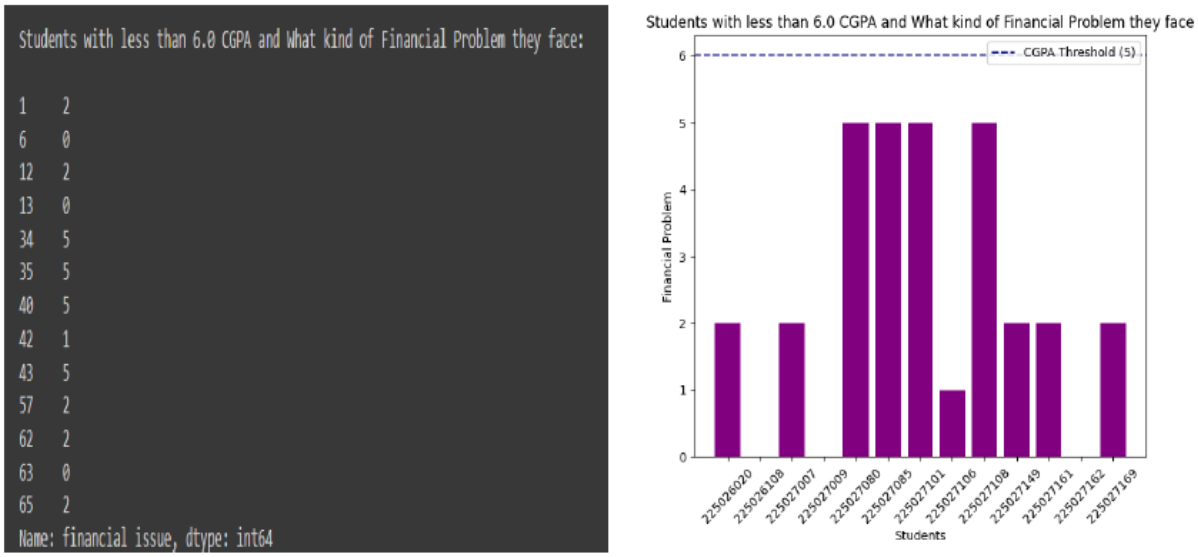

Figure 11: 2nd year students’ with CGPA below 6.0 and what kind of Financial problems they may face

7.5.3 For 3rd Year Students’:

Table 3: 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 Machine	0.69	0.83	0.75	0.83

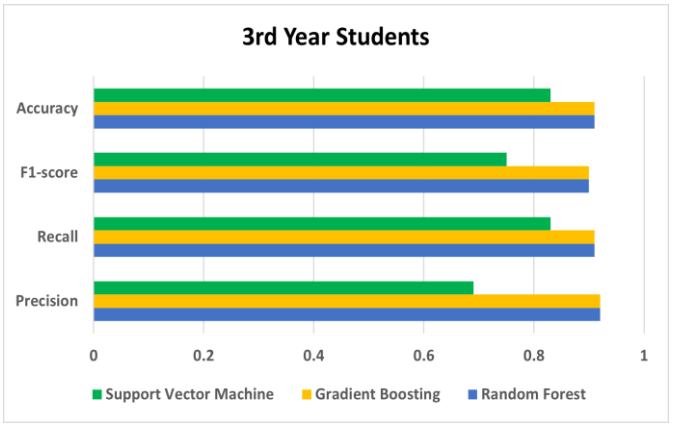

Figure 12: 3rd Year Students' Results

7.5.3.1 Query Results for students' with CGPA below 6.0 and their corresponding parameters threshold values:

Students who are below 6.0 CGPA:

	Register_no	Age	Gender	Your Current Course of Study	10th BoE	\
12	224026057	19	0	1	1	
20	224026088	19	1	1	1	
33	224026143	19	0	1	1	
38	224026181	20	1	1	1	
40	224026184	20	0	1	1	
55	224026246	19	1	1	1	
59	224026259	19	0	1	1	

	12th BoE	12th Specialization	MS	Problems_faced	attendance	...	\
12	1	1	1	0	2	...	
20	1	2	1	3	3	...	
33	1	2	1	4	1	...	
38	1	1	1	1	1	...	
40	1	2	1	3	2	...	
55	1	2	0	0	2	...	
59	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]

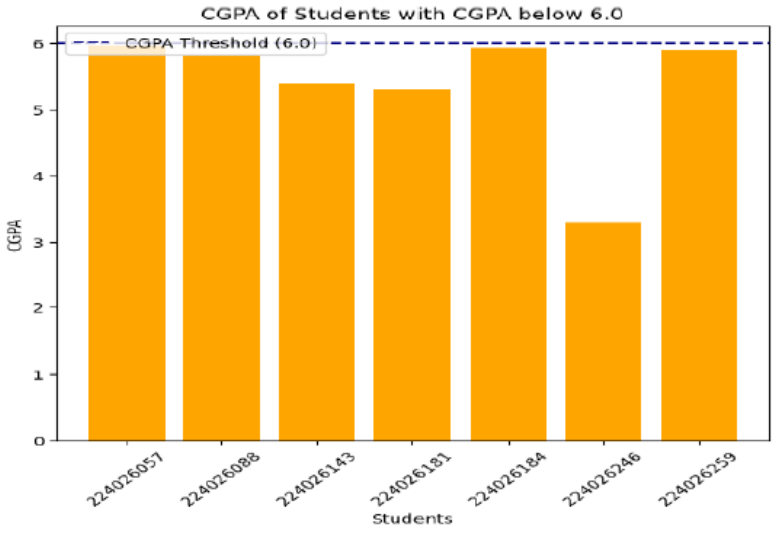

Figure 13: 3rd year students' with CGPA below 6.0

7.5.3.2 The parameter “income” and its corresponding threshold value:

4 – 2,00,000 above
3 – 1,00,000 upto 2,00,000
2 – 50000 upto 1,00,000
1 – 10000 upto 50000

```
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
```

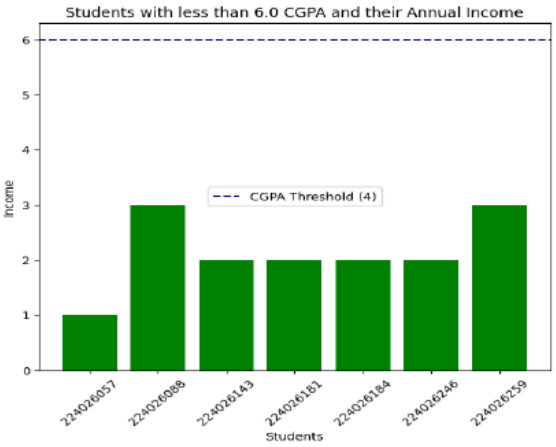

Figure 14: 3rd year students’ with CGPA below 6.0 and their annual income

7.5.3.3 The parameter “psychological_issue” and its corresponding threshold value:

5 – Hopelessness
4 – Fear about Future
3 – Depression
2 – Anxiety
1 – Loneliness
0 - Nothing

```
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
```

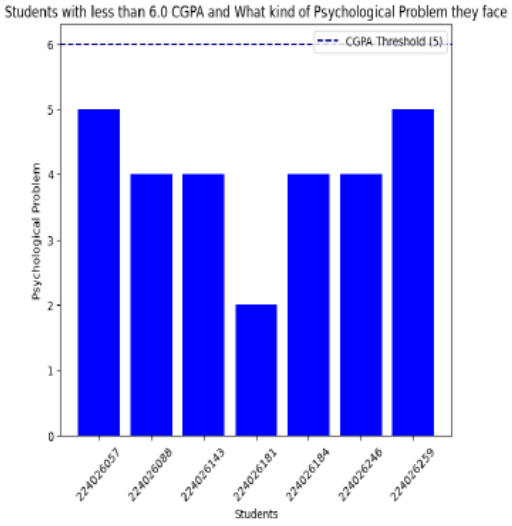

Figure 15: 3rd year students’ with CGPA below 6.0 and what kind of Psychological problems they may face

7.5.3.4 The parameter “financial_issue” and its corresponding threshold value:

5 – Education Investment
4 – Family Income
3 – Father or Mother Health issue
2 – Misunderstanding between relation
1 – Assets related problems
0 - Nothing

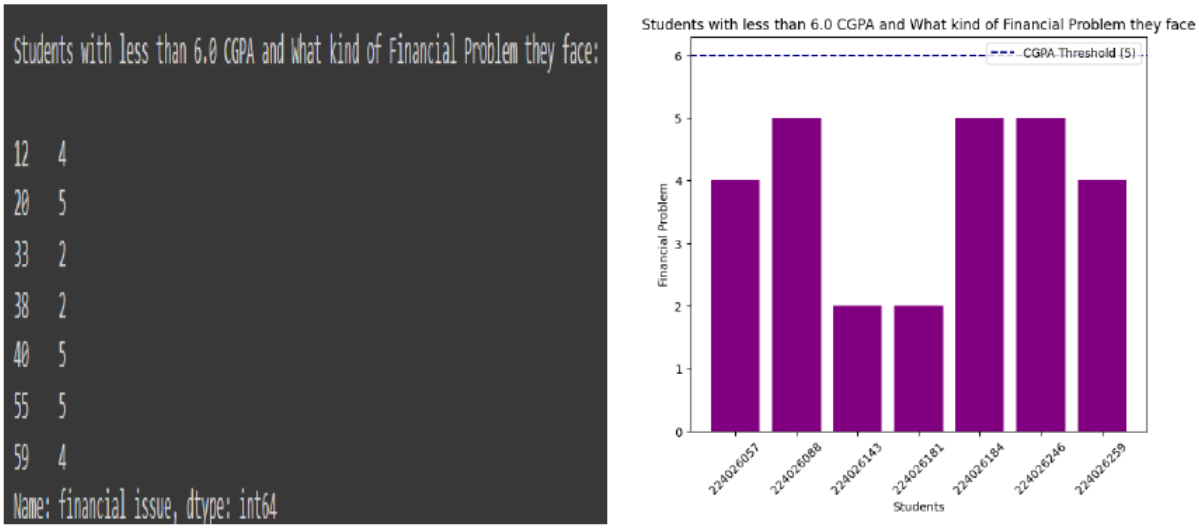

Figure 16: 3rd year students’ with CGPA below 6.0 and what kind of Financial problems they may face

8. 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.

9. FUTURE ENHANCEMENT:

Future enhancements for student performance prediction using machine learning algorithms may involve integrating more advanced models, such as deep learning architectures like recurrent neural networks (RNNs) or transformers. This could enable capturing more intricate patterns in student data, leading to more accurate predictions. Additionally, incorporating additional features beyond traditional academic data, such as socio-economic factors or behavioral attributes, could enrich the predictive models. Furthermore, exploring ensemble methods to combine predictions from multiple models or techniques like active learning to continually refine the models with new data could enhance prediction performance. Finally, deploying personalized recommendation systems based on predicted performance could offer tailored interventions to support individual student needs, fostering better academic outcomes.

10. REFERENCES:

1. **Olabanjo, O., Wusu, A. S., & Manuel, M.** (2022, December 1). *A machine learning prediction of academic performance of secondary school students using radial basis function neural network*. Trends in Neuroscience and Education. https://doi.org/10.1016/j.tine.2022.100190

2. **Miguéis, V. L., Freitas, A., Garcia, P. J. V., & Silva, A.** (2018, November 1). *Early segmentation of students according to their academic performance: A predictive modelling approach*. Decision Support Systems. https://doi.org/10.1016/j.dss.2018.09.001

3. **Rahman, S. R., Islam, M. A., Akash, P. P., Parvin, M., Moon, N. N., & Nur, F. N**. (2021, November 1). *Effects of co-curricular activities on student's academic performance by machine learning*. Current Research in Behavioral Sciences. https://doi.org/10.1016/j.crbeha.2021.100057

4. **H. Gull, M. Saqib, S. Z. Iqbal and S. Saeed**, "Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning," *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Bangluru, India, 2020, pp. 1-4, doi: 10.1109/INOCON50539.2020.9298266.

5. **S. Chandrahasa, S. M. S. Ganeshan, R. C. Maddineni, M. S. Divya, P. Tumuluru and B. Suneetha**, "Machine Learning Algorithms based Student Performance Prediction based on Previous Records," *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2023, pp. 181-186, doi: 10.1109/ICCMC56507.2023.10084099.

6. **S. Gupta and J. Agarwal**, "Machine Learning Approaches for Student Performance Prediction," *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2022, pp. 1-6, doi: 10.1109/ICRITO56286.2022.9964821.

7. **S. D. A. Bujang *et al.***, "Multiclass Prediction Model for Student Grade Prediction Using Machine Learning," in *IEEE Access*, vol. 9, pp. 95608-95621, 2021, doi: 10.1109/ACCESS.2021.3093563

8. **D. Liu, Y. Zhang, J. Zhang, Q. Li, C. Zhang and Y. Yin**, "Multiple Features Fusion Attention Mechanism Enhanced Deep Knowledge Tracing for Student Performance Prediction," in *IEEE Access*, vol. 8, pp. 194894-194903, 2020, doi: 10.1109/ACCESS.2020.3033200

9. **A. Alshanqiti and A. Namoun**, "Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification," in *IEEE Access*, vol. 8, pp. 203827-203844, 2020, doi: 10.1109/ACCESS.2020.3036572

10. **Rajendran, S., Chamundeswari, S., & Sinha, A. A.** (2022). Predicting the academic performance of middle- and high-school students using machine learning algorithms. *Social Sciences & Humanities Open*, *6*(1), 100357. https://doi.org/10.1016/j.ssaho.2022.100357

11. **Xu, X., Wang, J., Peng, H., & Wu, R.** (2019, September 1). *Prediction of academic performance associated with internet usage behaviors using machine learning algorithms*. Computers in Human Behavior. https://doi.org/10.1016/j.chb.2019.04.015

11.APPENDIX:

```
# Import necessary libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
import matplotlib.pyplot as plt
# Load your dataset
data = pd.read_csv('/content/3rd Year Data - Sheet1.csv')
# Correct the column names in the conditions
if data['CGPA_1'].mean() < 6:
    if (data['Gender'].isin([0, 1])).all() and \
      (data['MS'].isin([0, 1])).all() and \
      (data['Problems_faced'].isin([0, 1, 2, 3, 4])).all() and \
      (data['attendance'].isin([1, 2, 3])).all() and \
      (data['fam_situations'].isin([0, 1, 2])).all() and \
      (data['financial_issue'].isin([0, 1, 2, 3, 4, 5])).all() and \
      (data['income'].isin([1, 2, 3, 4])).all() and \
      (data['psychological_issue'].isin([0, 1, 2, 3, 4, 5])).all() and \
      (data['friend_discourage'].isin([0, 1])).all() and \
      (data['course_struggle'].isin([1, 2, 3, 4, 5])).all() and \
      (data['why_struggle'].isin([1, 2, 3, 4, 5])).all():
       target = 1
    else:
       target = 0
# Split data into features and target variables
features = data[['Register_no','Gender','MS','Problems_faced', 'attendance', 'fam_situations',
'financial_issue', 'income', 'psychological_issue', 'friend_discourage', 'course_struggle',
'why_struggle', 'SGPA_1', 'CGPA_1','SGPA_2', 'CGPA_2', 'SGPA_3', 'CGPA_3', 'SGPA_4',
'CGPA_4', 'SGPA_5', 'CGPA_5']]
targets = data['target']
```

```
# Assuming 'CGPA_1' is the target variable


# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(features, targets, test_size=0.2,
random_state=42)
# Initialize Gradient Boosting Classifier with default parameters
gb_classifier = GradientBoostingClassifier(random_state=42)
# Train the model
gb_classifier.fit(X_train, y_train)
# Predict on the testing set
y_pred = gb_classifier.predict(X_test)
# Calculate evaluation metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='weighted')
recall = recall_score(y_test, y_pred, average='weighted')
f1 = f1_score(y_test, y_pred, average='weighted')
# Print the evaluation metrics
print("\n-----Gradient Boosting Algorithm-----\n")
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```