Prediction of Academic Performance of College Students with Bipolar Disorder using different Deep learning and Machine learning algorithms

S. Peerbasha[#1], M. Mohamed Surputheen[#2]
[1] bashapeer2003@gmail.com [2]msurfudeen@yahoo.com
[1]Research Scholar, PG & Research Department of Computer Science, Jamal Mohamed College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, Tamilnadu, India.
[2]Associate Professor, PG & Research Department of Computer Science, Jamal Mohamed College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, Tamilnadu, India.

Abstract
In modern years, the performance of the students is analysed with lot of difficulties, which is a very important problem in all the academic institutions. The main idea of this paper is to analyze and evaluate the academic performance of the college students with bipolar disorder by applying data mining classification algorithms using Jupiter Notebook, python tool. This tool has been generally used as a decision-making tool in terms of academic performance of the students. The various classifiers could be logistic regression, random forest classifier gini, random forest classifier entropy, decision tree classifier, K-Neighbours classifier, Ada Boost classifier, Extra Tree Classifier, GaussianNB, BernoulliNB are used. The results of such classification model deals with 13 measures like Accuracy, Precision, Recall, F1 Measure, Sensitivity, Specificity, R Squared, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, TPR, TNR, FPR and FNR. Therefore, conclusion could be reached that the Decision Tree Classifier is better than that of different algorithms.

Keywords: *Logistic Regression, Random Forest classifier, Decision tree classifier, K-Neighbours classifier, Ada Boost classifier, Extra Tree classifier, GaussianNB.*

I. INTRODUCTION

Data Mining could be a promising and prospering wilderness in the examination of information and moreover, the outcome of the investigation has numerous applications. Information Mining can likewise allude to Knowledge Discovery from Data (KDD).This framework capacities as the machine-driven or helpful extraction of examples addressing information verifiably keep or caught in immense data sets, information stockrooms, the Web, information archives, and data streams. Data Mining is a multidisciplinary field, enveloping regions like data innovation, AI, insights, design acknowledgment, information recovery, neural organizations, data-based frameworks, man-made reasoning, and information perception.

The use of information mining is generally common in training frameworks. Instructive information mining is an arising field that can be successfully applied in the field of training. The instructive information mining employments a few thoughts and ideas, for example, Association rule mining, grouping, and bunching. The information that arises can be utilized to all the more likely comprehend student's' advancement rate, student's' consistency standard, student's' progress rate, and the student's' prosperity. The information mining framework is urgent and essential to gauge the student's' execution improvement. The order calculations can be utilized to characterize and investigate the student's' information set in precise way. The attributes that may be used for student's academic performance such as Name, College, Class Tutor Name, Gender, Age, Address, Family size, Parent's Status, Mother's Education, Father's Education, Mother's Job, Father's Job, Reason to choose this college, Student's Guardian, Travel Time, Study time (Weekly), Number of Subjects failed so far, College Support (Extra curricular), Family Educational Support, Paid courses attended, Extra curricular Activities involves, Nursery schools attended, Higher education, Internet access at home, love relationships, Family relationships, Free time after college, Going out with friends, Week days – Alcohol / Smoking, Weekends – Alcohol / Smoking, Current Health Status, Number of days absent, CGPA, Number of Psychological motivation sessions attended, CGPA Grade.

The main idea of this research paper is to use data mining classification algorithms to study and analyze the academic performance of the college students with bipolar disorder. This research paper constitutes 5 sections. Section-1 deals with introduction; Section-2 enumerates a related work. Section-3 presents the idea and aspects of different classifiers. Section-4 elaborates the Data Pre-processing. Section-5 explains the implementation of model construction. Section-6 handles results and discussions. Conclusion will be given in Section-7.

Manuscript received July 5, 2021
Manuscript revised July 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.7.40

II. LITERATUR SURVEY

Academic performance of students in advanced education (HE) is investigated broadly to handle scholarly underachievement, expanded college dropout rates, graduation delays, among other persevering difficulties [1]. In straightforward terms, student performance alludes to the degree of accomplishing present moment and long haul objectives in education [2]. Nonetheless, academicians measure student's accomplishment according to alternate points of view, going from student's' last grades, grade point normal (GPA), to future occupation possibilities [3]. The writing offers an abundance of computational efforts striving to further develop student execution in schools and colleges, most remarkably those determined by data mining and learning investigation procedures [4]. The opportune expectation of student execution empowers the identification of low performing student's, along these lines, engaging instructors to mediate right on time during the learning cycle and carry out the necessary intercessions. Productive intercessions incorporate, however are not restricted to, student prompting, execution progress observing, shrewd mentoring frameworks advancement, and policymaking [5]. This undertaking is emphatically supported by computational advances in information mining and learning examination [6]. A new complete study features that around 70% of the checked on work explored student execution expectation utilizing student grades and GPAs, while just 10% of the investigations reviewed the forecast of student accomplishment utilizing learning results [3]. This hole impelled us to altogether research the work did where the learning results are utilized as an intermediary for student scholastic execution. Result based instruction is a worldview of schooling that spotlights on executing and achieving the purported learning results [7]. As a result, student learning results are objectives that action the degree which student's achieves the proposed capabilities, explicitly information, abilities, and qualities, toward the finish of a specific learning measure. In our view, the student results address a more comprehensive measurement for making a decision about student scholastic accomplishments than simple appraisal grades. This view agrees with the case that the learning results address basic components of student scholarly achievement [8]. Also, eminent HE accreditation associations, like ABET and ACBSP, utilize the learning results as the structure blocks for surveying the nature of instructive projects [9]. Such significance calls for more exploration endeavours to anticipate the fulfilment of learning results, both at the course and program levels. The absence of methodical reviews exploring the expectation of student execution utilizing student results has roused us to seek after the destinations of this exploration. In an efficient writing survey (i.e., SLR), a bit by bit convention is executed to distinguish, select, and assess the integrated investigations to respond to explicit examination questions [10, 11].

III. CLASSIFICATION

Classification is a process that is used to categorize data into predefined unconditional class labels. Classification can be a two venture process comprising of training and testing. In the initial step, a model is developed by investigating the data tuples from training data having a collection of attributes. For each tuple in the preparing data, the value of class label attribute is understood. Categorization rule is applied on preparing data to frame the model. In the second step of arrangement, test data is utilized to look at the exactness of the model. Assuming the precision of the model is proper, the model can be utilized to group the unknown information tuples.

3.1. Classification Algorithms

The fundamental techniques for classification used in this work are logistic regression, random forest classifier gini, random forest classifier entropy, decision tree classifier, K-Neighbours classifier, Ada Boost classifier, Extra Tree Classifier, GaussianNB, and BernoulliNB. We will discuss those classifiers one by one.

3.2 Logistic regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

3.3 Random Forest Classifier gini

Random Forest Classifier gini or Mean Decrease in Impurity (MDI) calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits.

3.4 Random Forest Classifier Entropy

Random Forest Classifier Entropy helps us to select the best splitter. It can be defined as the measure of the purity of the sub split. It always lies between 0 and 1.

$$H(s) = - P (+) \log_2 P (+) – P (-) \log_2 P (-)$$

3.5 Decision Tree Classifier

Decision tree develops classification or regression models as a tree structure. It's everything except a dataset into more humble and more modest subsets while at the same time, a connected decision tree is steadily evolved. The possible outcome is a tree with choice hubs and leaf hubs.

3.6 K-Neighbours Classifier

K neighbours classifier is used to store all available cases and classifies new cased based on similarity measure such as distance functions. It has been used in statistical estimation and pattern recognition.

3.7 Ada Booster Classifier

It is a boosting procedure that is utilized as a gathering technique in AI, in which the weights are re-allotted to each case, with higher loads to inaccurately classified cases.

3.8 Extra Tree Classifier

It is a kind of outfit learning procedure that totals the results of various de-associated choice trees gathered in a "forest" to yield its grouping result.

3.9 GaussianNB Classifier

It is the fast, accurate and reliable algorithm that has high accuracy and speed on large datasets. It assumes that the effect of a particular feature in a class in independent of other features.

3.10 BernoulliNB Classifier

It is designed for binary / Boolean features. Threshold for binarizing (mapping to Booleans) of sample features.

IV. DATA PRE-PROCESSING

Datasets used inside classification algorithm should be clear and can be pre-processed for taking care of absent or excess attributes. The information are to be taken care of with effectiveness to prompt the best result from the Data Mining measure.

4.1 Identification of Attributes

Dataset collected from college students database are given with the attributes.

Table-1 Symbolic Attribute Description

Attributes	Description	Possible Values
Name	Name of the Student	Text
College	Name of the College	Jamal Mohamed College, Holy Cross College, Bishop Heber College, St. Joseph's College, EVR College, National College, Cauvery College for Women, Indira Gandhi College, SRC, MIET, Others
CTN	Name of the Class teacher	Text
Gender	Gender of the student	Male, Female, Transgender
Age	Age of the students	18 – 30
Address	Whether the students is from city or village	Urban, City
FS	Family size	Less than 3, Greater than 3
PS	To identify whether they are joined or not	Joined, Apart
ME	Educational qualification of mother	0 (None), 1 (Primary Education, Upto 4th Standard), 2 (5th std to 9th std), 3 (10th Std to 12th std), 4 (UG / PG)
FE	Educational qualification of father	0 (None), 1 (Primary Education, Upto 4th Standard), 2 (5th std to 9th std), 3 (10th Std to 12th std), 4 (UG / PG)
MJ	Job of the mother	Teacher, Health care related, Job less, Others
FJ	Job of the father	Teacher, Health care related, Job less, Others
RTCC	Exact reason for joining the college	Close to home, College Reputation, Course Preference, Others
SG	Guardian of the student	Father, Mother, Others
TT	Time of the travelling	Less than 15 minutes, 15 minutes – 30 minutes, 30 minutes – 1hour, more than 1 hour
ST	Time duration of the week	Less than 2 hours, 2 – 5 hrs, 5 – 10 hrs, more than 10 hrs
NSFS	Arrears so far in the subjects studied	Nil, 1,2, 3 and more than 3
CSIEC	Importance given by the college for extra curricular activities	Yes, No
FES	Educational Support given by the family	Yes, No
PCA	Any additional course for cost attended	Yes, No
ECA	Involved in any extra-curricular activities or not	Yes, No
Nursery	Attended in Nursery school or no	Yes, No
HE	Want to join Higher Education	Yes, No
IAH	To identify whether the internet is accessed or not	Yes, No
AYL	Love relationships	Yes, No
FR	About relationships of the family	Very bad, bad, good, very good, excellent
FTAC	Free time after the class	Very bad, bad, good, very good, excellent
GOWF	Time spent with friends	Very bad, bad, good, very good, excellent

WDAC	Alcohol consumption during Monday to Friday	Very bad, bad, good, very good, excellent
WEAC	Alcohol consumption during Saturday and Sunday	Very bad, bad, good, very good, excellent
CHS	Status of the health	Very bad, bad, good, very good, excellent
NDA	Total number of days absent for the classes	0 to 90
CGPA	Cumulative Grade Point Average	0.0 to 10.0
NPMSA	Number of Psychological motivation sessions attended	0 to 10
CGPAG	Grade for the CGPA	Distinction, First Class, Second Class, Pass, Fail

V. IMPLEMENTATION USING JUPITER NOTEBOOK, PYTHON

Jupiter Notebook is an open-source web application that permits us to create and share documents that contain real code, equations, and visualizations, and summarized text. It is used for data cleaning, transformation, numerical simulation, statistical modeling, data visualization, machine learning, etc. From the above attributes given, the student analysis.arff file was created and uploaded into Jupiter Notebook. The academic performance of the students is influenced by the above-specified attributes. The various classifiers such as logistic regression, random forest classifier Gini, random forest classifier entropy, decision tree classifier, K-Neighbours classifier, Ada Boost classifier, Extra Tree Classifier, GaussianNB, and BernoulliNB are used. The results of such classification model deal with 13 measures like Accuracy, Precision, Recall, F1 Measure, Sensitivity, Specificity, R Squared, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, TPR, TNR, FPR, and FNR.

VI. RESULTS AND DISCUSSION

Table 2 Extraction of required features from the dataset

Address	Study_hours	Extra_curriculum	Internet_access	Family_interaction	Drug_use_weekdays	Drug_use_weekends	Leave_log	CGPA	Psychological motivation sessions attended	CGPA_Grade
Rural(Village)	2	0	1	5	1	1	5.0	8.0	7	First Class
Rural (Village)	10	0	0	5	1	1	0.0	10.0	10	Outstanding
Rural (Village)	2	0	1	5	1	1	5.0	9.9	10	Outstanding
Rural (Village)	2	1	1	5	2	2	6.0	7.0	5	First class
Urban (City)	10	1	1	3	1	1	4.0	8.0	7	First class

We have given the results of CGPA Grade, which is based on psychological motivation sessions demonstrated that the rural students give outstanding performance than urban students.

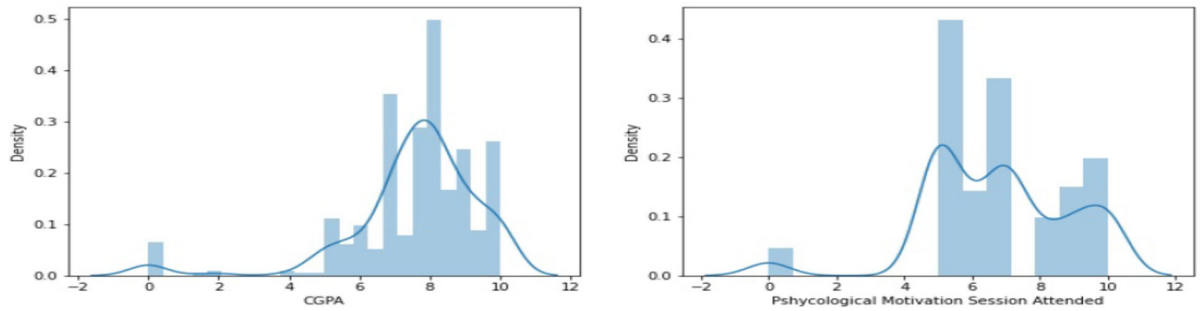


Fig 1: Analysis of CGPA with respect to Psychological motivation session attended

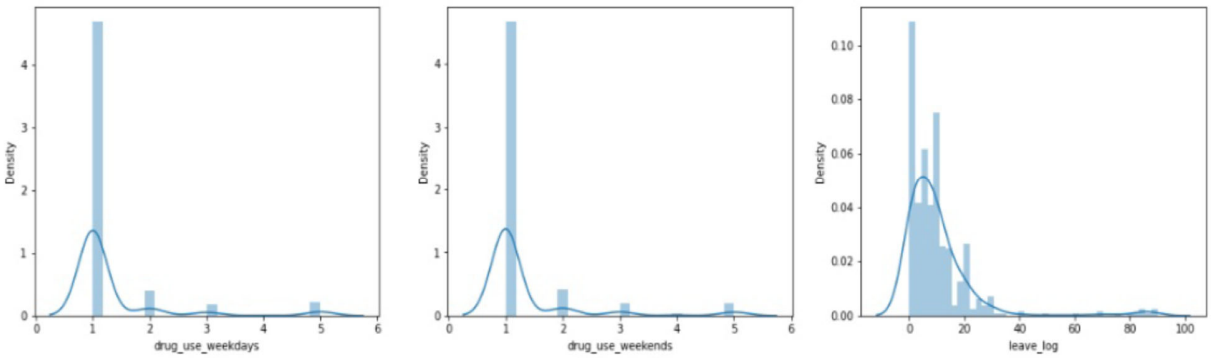


Fig 2: Analysis of Drug usage in Weekdays, Weekends and the Leave Logs

From the above figure, we showed that most of the students do not use the drugs. Few students use the drug weekly one or two times that also makes them to take leave.

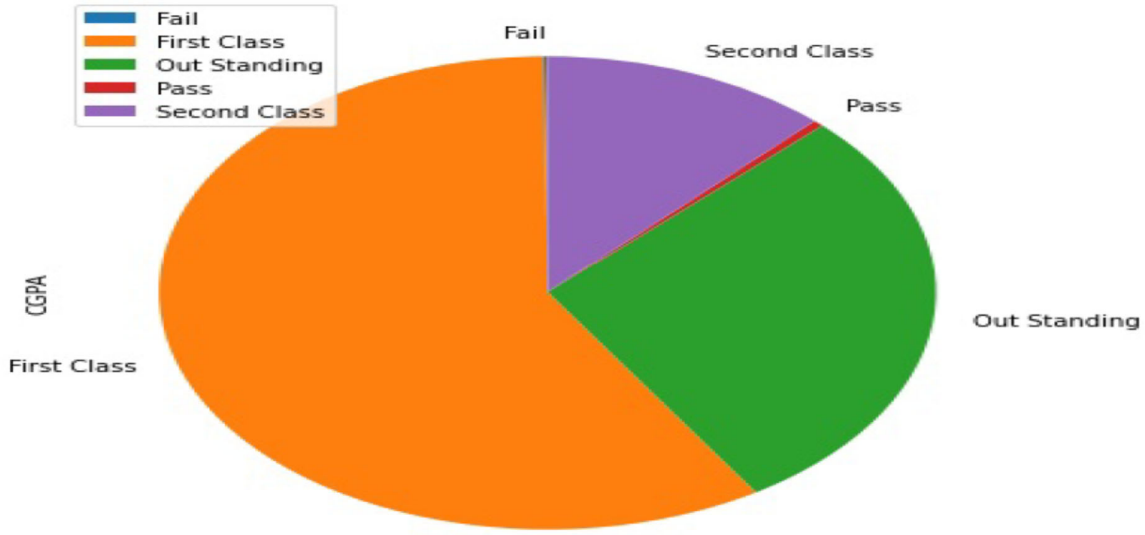


Fig 3: Analysis of the Grades obtained by the students

From the results obtained, it is cleared that the students obtained first class than the other classes of grades.

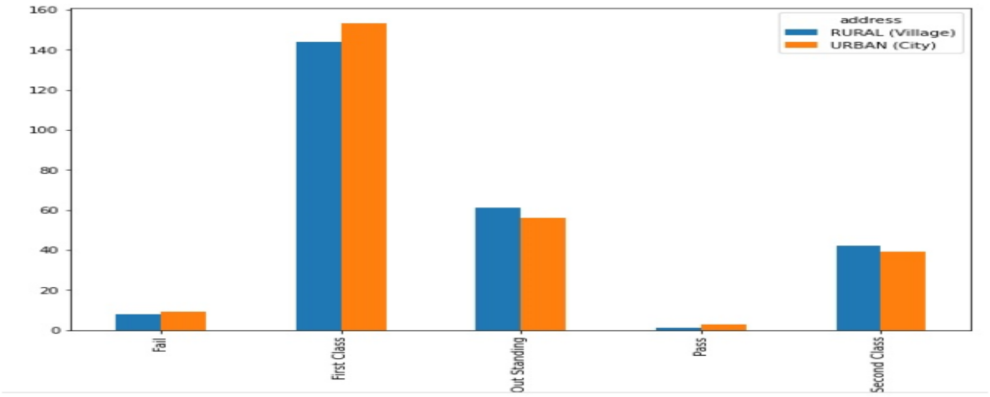


Fig 4: Performance Analysis of Rural and Urban Students

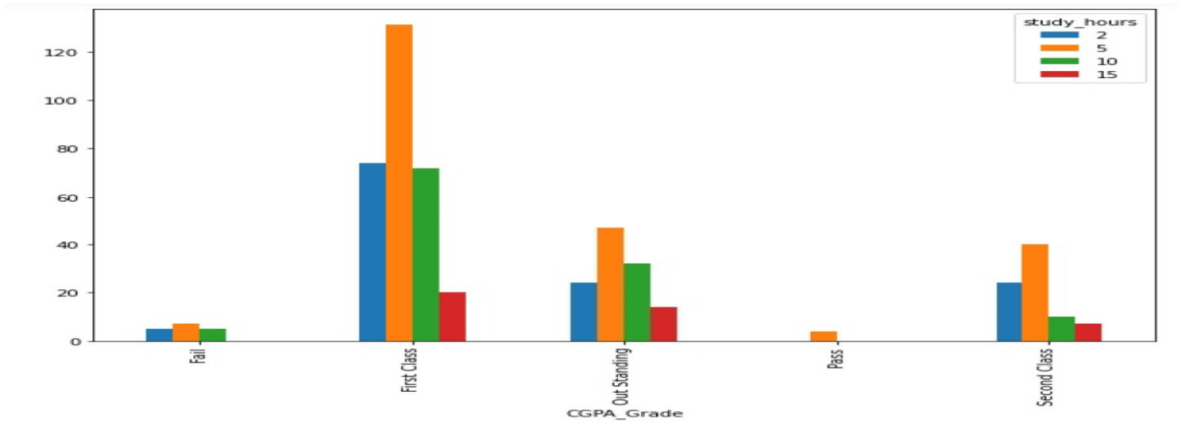


Fig 5: Analysis of Study Hours of the Students

The present study confirms the finding that on an average of five hours of study will make the students to score good grades. Here, the study also reveals that the study hour does not give the improvement in grades. Full focus of study is needed than the hours.

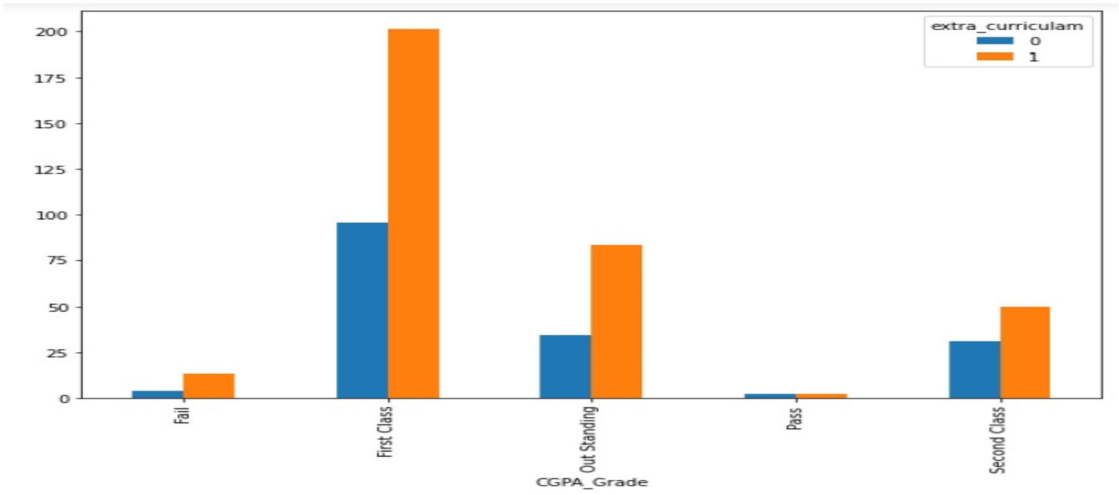


Fig 6: Analysis of the Extra Curricular activities of the students

Another promising finding was that the students with extra curricular activities are performing better than the students without extra curricular activities.

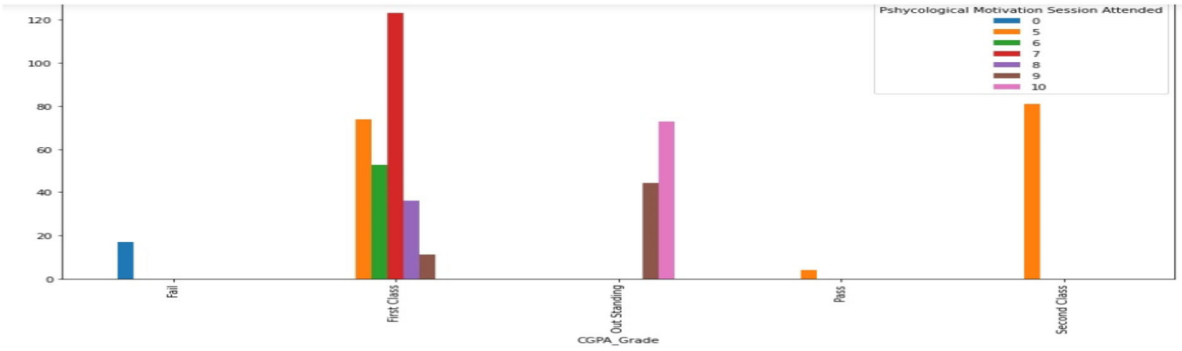
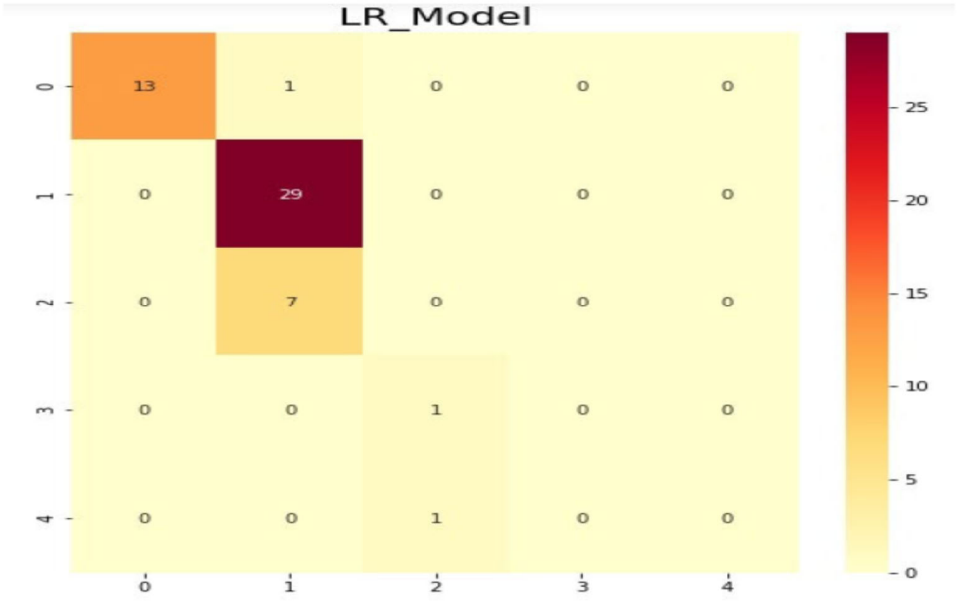


Fig 7: Analysis of Psychological Motivation Sessions attended

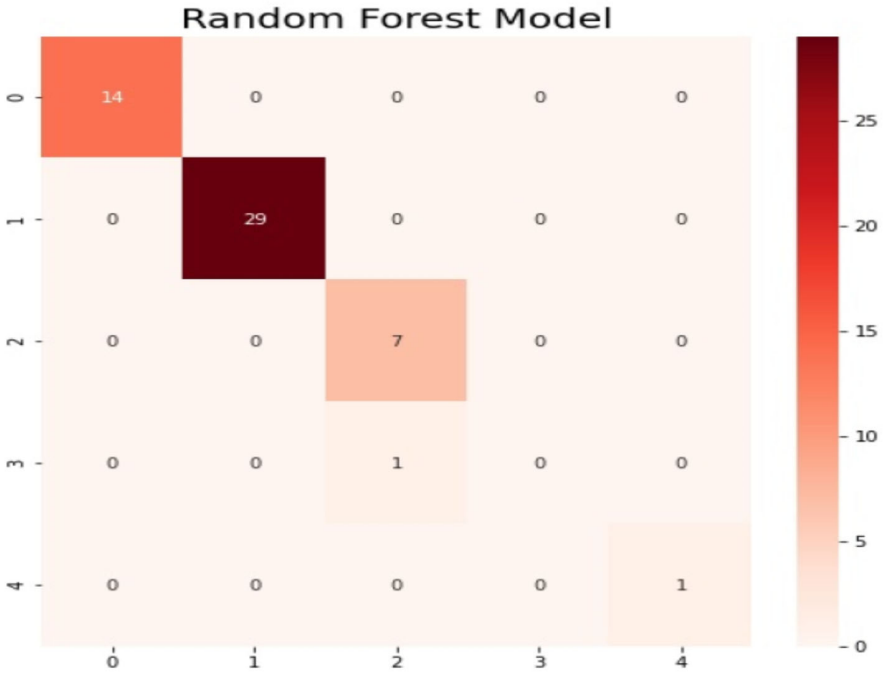
It is worth discussing the psychological motivation sessions improve the performance of students in terms of CGPA revealed by the above results.

Table-3 Confusion matrix of Linear Regression model



This suggests that the confusion matrix is a summary of prediction results on a classification problem. Here, the confusion matrix represents counts from predicted and actual values. We have created a confusion matrix for linear regression classifier.

Table 4 Confusion matrix of Random Forest Model



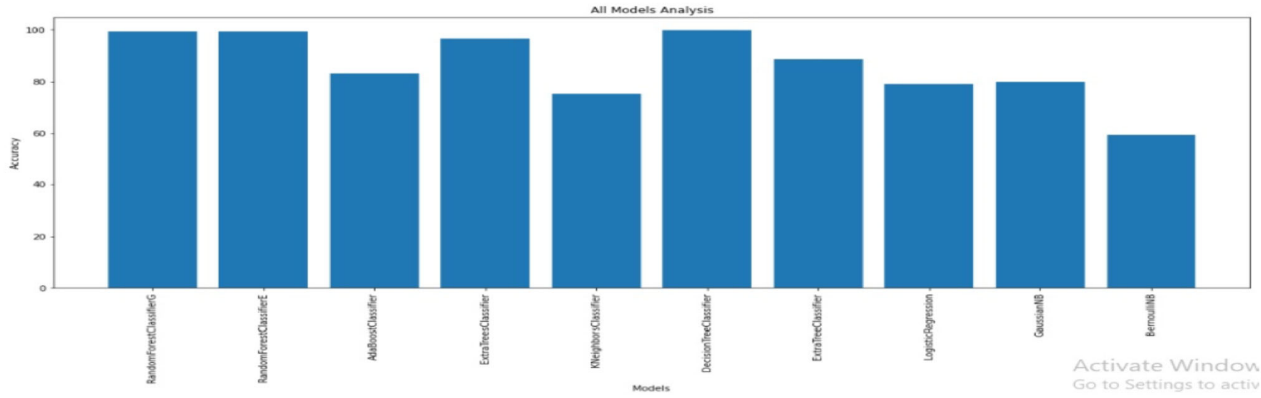


Fig 11: Performance Analysis of different machine learning algorithms

By analyzing the performance in terms of accuracy by different machine learning algorithms, we hope that decision tree classifier performs very well in the prediction of academic performance of college students with bipolar disorder

Table-5 Performance result for Classifiers

Measures	Logistic Regression	Random Forest Classifier Gini	Random Forest Classifier Entropy	Decision Tree	KNeighbors Classifier	Ada Boost Classifier	Extra Tree Classifier	Extra Tree Classifier N	GaussianNB	BernoulliNB
Precision	0.705341	0.963942	0.963942	1.0	0.701735	0.725572	0.906050	0.906050	0.931818	0.443794
Recall	0.807692	0.980769	0.980769	1.0	0.730769	0.846154	0.923077	0.961538	0.920435	0.457892
F1Measure	0.749352	0.971795	0.971795	1.0	0.711844	0.778555	0.912305	0.952086	0.920435	0.457892
Sensitivity	0.807692	0.980769	0.980769	1.0	0.730769	0.846154	0.923077	0.961538	0.923077	0.538462
Specificity	0.807692	0.980769	0.980769	1.0	0.730769	0.846154	0.923077	0.961538	0.923077	0.538462
R Squared	0.616780	0.970522	0.970522	1.0	0.410431	0.675737	0.882086	0.941043	0.882086	0.115646
MAE	0.211538	0.019231	0.019231	0.0	0.307692	0.173077	0.076923	0.038462	0.076923	0.500000
MSE	0.250000	0.019231	0.019231	0.0	0.384615	0.211538	0.076923	0.038462	0.076923	0.576923
RSME	0.500000	0.138675	0.138675	0.0	0.620174	0.459933	0.277350	0.196116	0.277350	0.759555
TPR	29.00000	29.00000	29.00000	29.0	25.000000	29.00000	29.00000	29.00000	26.000000	25.000000
TNR	13.00000	14.00000	14.00000	14.0	9.000000	14.00000	13.00000	14.00000	14.000000	2.000000
FPR	1.000000	0.000000	0.000000	0.0	5.000000	0.000000	1.000000	0.000000	0.000000	11.000000
FNR	0.000000	0.000000	0.000000	0.0	3.000000	0.000000	0.000000	0.000000	0.000000	3.000000
Accuracy	0.807692	0.980769	0.980769	1.0	0.730769	0.846154	0.923077	0.961538	0.923077	0.538462

Table 5 clearly shows the performance of every classifier based on the precision, recall, F1 measure, sensitivity, specificity, R squared, mean absolute error, mean squared error, root mean squared error, true positive rate, true negative rate, false positive rate, and false negative rate. These measures are useful for comparing the classifiers based on accuracy. From the above table, we found that the Decision tree classifier outperforms in all the aspects than the other classifier within the collected student's dataset. BernoulliNB classifier produces low values

VII. CONCLUSIONS

The paper presented different prediction models to predict the academic performance of bipolar disorder students based on Address, study_hours, extra_curriculum, internet_access, family_interaction, drug_use_weekdays, drug_use_weeends, leave_log, CGPA, Psychological motivation session attended, CGPA_Grade. This study is however limited to the college students of Jamal Mohamed College and other few college students. From the observation, it is found that the performance of a decision tree classifier is best than that of different algorithms applied in this research. The performance of every classifier compared is based on the precision, recall, F1 measure, sensitivity, specificity, R squared, mean absolute error, mean squared error, root mean squared error, true positive rate, true negative rate, false positive rate and false negative rate. This study is helpful for the educational

institutions for the prediction of academic performance of college students affected with bipolar disorder.

References

[1] M.S. Mythili, A.R. Mohamed Shanavas, "An Analysis of Student's Performance using Classification Algorithms", IOSR Journal of Computer Engineering, Volume 16, Issue I, Ver. III, pp. 63-69.

[2] Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. (2006) „Mining Student Data Using Decision Trees", The 2006 International Arab Conference on Information Technology (ACIT'2006) – Conference Proceedings.

[3] Ayesha, S. , Mustafa, T. , Sattar, A. and Khan, I. (2010) „Data Mining Model for Higher Education System", European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29.

[4] Baradwaj, B. and Pal, S. (2011) „Mining Educational Data to Analyze Student s" Performance", International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.

[5] Chandra, E. and Nandhini, K. (2010) „Knowledge Mining from Student Data", European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163.

[6] El-Halees, A. (2008) „Mining Students Data to Analyze Learning Behavior: A Case Study", The 2008 international Arab Conference of Information Technology (ACIT2008) – Conference Proceedings, University of Sfax, Tunisia, Dec 15-18.

[7] Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.

[8] Kumar, V. and Chadha, A. (2011) „An Empirical Study of the Applications of Data Mining Techniques in Higher Education", International Journal of Advanced Computer Science and Applications, vol. 2, no. 3, pp. 80-84.

[9] Mansur, M. O. , Sap, M. and Noor , M. (2005) „Outlier Detection Technique in Data Mining: A Research Perspective", In Postgraduate Annual Research Seminar.

[10] Romero, C. and Ventura, S. (2007) „Educational data Mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146

[11] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, "Mining student data using decision trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[12] U. K. Pandey, and S. Pal, "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[13] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, "Data mining model for higher education system", European Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.

[14] M. Bray, The shadow education system: private tutoring and its implications for planners, (2nd ed.), UNESCO, PARIS, France, 2007.

[15] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[16] J. R. Quinlan, "Introduction of decision tree: Machine learn", 1: pp. 86-106, 1986.

[17] Vashishta, S. (2011). Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm. IJACSA - International Journal of Advanced Computer Science and Applications, 2(4), 77-80.

Mr. S. Peerbasha, a Research Scholar at Jamal Mohamed College (Autonomous) has cleared his National Level Eligibility Test conducted by UGC NTA NET in the year 2020 and has cleared his State Level Test (TNSET-2018) conducted by Mother Teresa University, Kodaikanal. He received his M.Tech at School of Computer Science & Engineering, Bharathidasan University in 2012. He received his M.B.A at Alagappa University in 2011. He has completed his Master of Philosophy (C.S) in 2008 and Master of Computer Applications in 2007. He is currently working as an Assistant Professor of Computer Science at Jamal Mohamed College for 10 Years. He was also working as a Senior Lecturer for Southern Cross University, Australia in 2016. He is acting as a Member in Guidance and Counseling Centre, Jamal Mohamed College (Autonomous). His areas of research include Data mining, Machine learning and Deep Learning.

Communication maid id: bashapeer2003@gmail.com

Dr. M. Mohamed Surputheen, Working as an Associate Professor in Jamal Mohamed College with more than 30yrs experience. Education Qualification is M.Sc.. M.Phil and PhD. He is Guiding M.Phil Scholars and around eight PhD Scholars. Published and presented more than twenty five research papers. Currently acting as a Controller of Examinations at Jamal Mohamed College (Autonomous), Trichy. His areas of research include Wireless sensor networks, Data mining, Machine Learning, Deep Learning and Image Processing.

Communication mail id: msurfudeen@yahoo.com