Student Academic Performance analysis using Classification algorithms

E. Iyappan[1], K. Rajapriya[2], A. Santhosh[3], T. Sasikumar[4], P. Vinitha[5], M.S. Sassirekha[6], Anbarasan Balakrishnan[7]

[1,2,3,4,5]*Department of Computer Application, Thiagarajar college of Engineering (Affiliated to Anna University), Madurai, India*

[6]*Assistant Professor, Department of Computer Application, Thiagarajar college of Engineering (Affiliated to Anna University), Madurai, India*

[7]*Director- Engineering, Capgemini Engineering*

Abstract **- In this paper, we performed analysis is to measure performance of the students and factors affecting their performance. The model can be used for early predicting student performance to help in improving student performance on the subject. Supervised algorithms like Decision tree, Adaboost, Random Forest, Stochastic classification and Logistic regression are used to predict the model.**

Index Terms **– Classification, Support Vector classification, Adaboost, Matplotlib, Seaborn, Logistic regression, Decision tree, Random Forest, Stochastic classification.**

I.INTRODUCTION

Education is an important element of the society. With the corona-virus outbreak that has disrupted life around the globe in 2020, the educational systems have been affected in many ways; studies show that student's performance has decreased and their attention span has been reduced due to online classes. This analytics study highlights the need to deal with this problem more seriously and try to find effective solutions, as well as the influencing factors

Primary goals of this analysis are

- Do an exploratory analysis of the student performance dataset
- Do a visualization analysis of the student performance dataset
- To find whether a student will pass or not using classification models
- Check the accuracy of each classification model

II. RELATED WORKS

EDM is used to identify students' learning patterns and process. They can be used to predict their performance to identify at-risk students at early stage. Prediction can be done based on their learning activities, coursework grades and learning outcome.

Educational big data and learning analytics approaches were applied in blended Calculus course for early prediction of students' academic performance.

Logistic regression was used to predict students' final grade performance. Seven critical factors had been identified, whereas they consisted of three traditional factors and four online factors that impacted students' academic performance.

Decision tree classifier was used to develop an early warning system to identify at-risk student. A data consisted of 300 students with 13 online attributes was used to build a prediction model. The model achieved 100% accuracy based on performance predicting whether students would pass or fail.

Adaboost classifier was applied on the data collected during freshman year to predict students' grades in their final year. Meanwhile, used regression to predict students' grades.

Stochastic classification was used to identify affective information to improve predictive accuracy for the early identification of students who are likely to fail in a subject

Table 1- Student Background

Attribute	Description	Type	Value
sex	gender of student		male | female
school	school of student		Mourinho da Silveira | Gabriel Pereira
address	type of student's home address		rural | urban

Pstatus	cohabitation status of	binary	living together | apart
famsize	size of		< 3 | > 3
schoolsup	extra educational school support		yes | no
famsup	educational support from family		yes | no
Mjob	job of mother	nominal	- at home - civil services - teacher - health care related
Fjob	job of father		
reason	reason to choose this school		- close to home - school reputation - course preference - other
guardian	guardian of student		- father - mother - other
Medu	education of mother	numeric	0 # none 1 # primary education 2 # 5th to 9th grade 3 # secondary education 4 # higher education
Fedu	education of father		
famrel	quality of family relationships		very bad (1) to excellent (5)
age	age of student		15 - 22
traveltime	travel time from home to school		1 # < 15 min 2 # 15 to 30 min 3 # 30 min. to 1 hour 4 # > 1 hour
studytime	weekly study time		1 # < 2 hours 2 # 2 to 5 hours 3 # 5 to 10 hours 4 # > 10 hours
failures	number of failures in past class		n if 1 ≤ n < 3, else 4

Table II Student Social Activities

Attribute	Description	Type	Value
activities	extra-curricular	binary	yes | no
higher	plans for higher education		
internet	home internet access		
nursery	nursery school attended		
paidclass	extra paid classes		
absences	absences from school	numeric	very low (1) to very high (5)
health	status of current health		
freetime	free time after school		
goout	outing with friends		
Dalc	consume alcohol in weekday		
Walc	consume alcohol in weekend		0 - 93

Table III Student Coursework Result

Attribute	Description	Type	Value
GI	1st grade period	numeric	0 - 20
G2	2nd grade period		

III. METHODOLOGY AND DATASET

There are 395 rows and 32 columns in the dataset. The following are the inferences from the dataset

- We have school name, gender, age and other details of each student
- We have the reason for choosing the school, parental status and support.
- We also have other factors like outing, internet, free time and health which affects the study.
- Performance of students based on the factors are main target
- Performance can be either pass or fail. students background with 18 attributes (TableI) student social activities with 12 attributes (Table II) student coursework results with 2 attributes (Table III)

These subsets attributes will be used to predict final grade (G3). G3 is a numeric datatype with range of 1 – 10 used to measure student performance on their final grade. The subset attributes will be evaluated under two models:

• 2-level classification (pass / fail)

• 4-level classification (Good/Fair/Poor/Very poor) (Table IV)

TABLE IV. 4-LEVEL CLASSIFICATION ON PARENTS EDUCATION RESULTS

Education	>=4	3	2	1
Results	Good	Fair	Poor	Very Poor

Data conversion and normalization have been applied on following attributes prior to the evaluation of prediction models.

•Age attribute (student background subset) is converted to nominal.

• Failures attribute (student social activities) is normalized to categorial value as depicted in Table V. The parents' education levels are converted to 2-level classification and 4-level classification to predict the performance at 2-level classification and 4-level classification models.

TABLE V. 4-LEVEL CLASSIFICATION ON STUDENTS ARREARS RESULTS

Failures	>=4	3	2	1
Results	Good	Fair	Poor	Very Poor

The unwanted data, outliers and the null values in the dataset can produce the biased output while doing prediction. So, we have to clean them by removing them or replacing them.

The following data are removed due to their non importance

- Daily alcohol consumption
- Weekly alcohol consumption
- First period grade (G1)
- Second period grade (G2)
- Third period grade (G3)

III. EXPERIMENTAL RESULTS

For analyzing the data, we used classification algorithms to perform analysis on the dataset with the help of python. Supervised algorithm techniques, namely Decision tree, Random Forest, Logistic regression, supported vector Machine, Adaboost, Stochastic classification The evaluation has been performed on the three subset attributes on 2-level classification and 4-level classification models as shown in Figure 2. The experimental analysis also performed on all attributes which is referred as all subsets dataset.

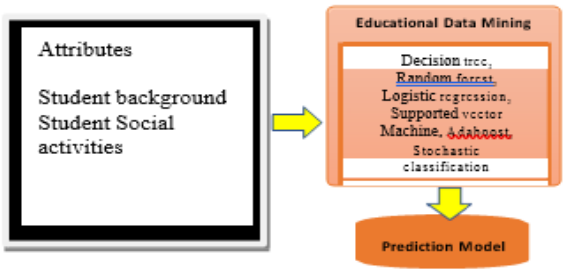

Fig. 2. Experimental Models

As depicted in Table VI (Figure 3) decision tree outperformed other educational data mining algorithms in 2- level classification and 4-level classification on all subsets dataset except student social activities subsets in 2-level classification. Meanwhile, Logistic regression outperformed other in evaluating student social activities subsets in 2-level classification and 4-level classification models.

As shown in Table VI (Figure 3), the models accuracy indicated that student background and student social activities are viable to be used to perform early analysis and prediction of at-risk student to determine whether it pass or fail the subject.

An explanatory analysis was performed on 2-level classification model. Decision trees are generated to identify the relevant attributes that might direct impact students' performance. As depicted in Figure 3,4,5, following attributes are significant to predict student performance:

- Parental status, desire to pursue higher studies, study time, living area influence the performance of students in positive manner
- Outing, health, failures in previous exams affects the performance in negative manner.

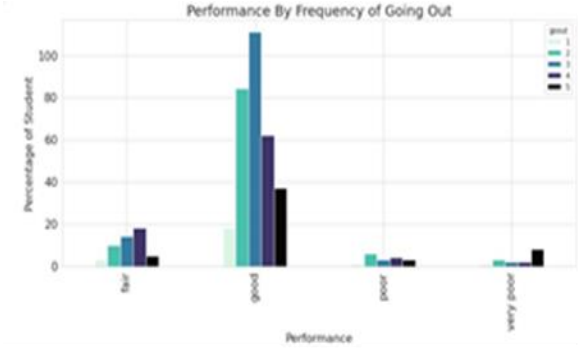

Fig. 3. GRAPH OF STUDENTS – GOING OUT [1]

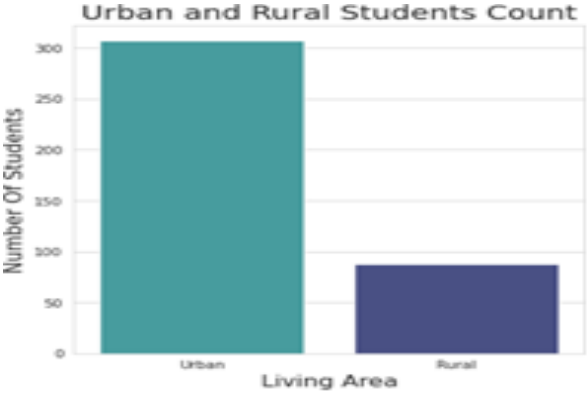

Fig. 4. GRAPH OF URBAN STUDENTS - RURAL STUDENTS

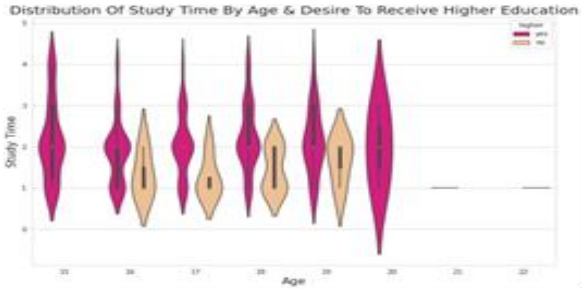

Fig. 5. GRAPH OF DESIRE OF STUDENTS INFLUENCES THE PERFORMANCE IN HIGHER LEVEL

IV. CONCLUSION AND FUTURE WORK

In this paper, we have analyzed the factors that affect the students positively and negatively. We've also analyzed which factor affects the most and whether students would pass or fail.

For this project entitled "students' performance Dataset", we used several classification methods such

as logistic regression, decision tree, random tree, adaboost, stochastic and SVM and finally we got a higher value with Decision tree of 100% compared to other algorithms.

We have also generated comparison of analytics with actual analytics.

TABLE VI. ACCURACY SCORE OF CLASSIFICATION ALGORITHMS

Accuracy score						
Model	Decision Tree	Random Forest	Logistic Algorithm	SVC	Adaboost	Stochastic Gradient Descent
Score	100%	87.34%	96.20%	97.46%	98.73%	87.34%

REFERENCES

[1] https://github.com/Shwetago/Student-grades-prediction
[2] https://github.com/AbhishekMali21/student-grade-analysis- prediction
[3] https://github.com/sachanganesh/student-performance-prediction
[4] https://github.com/Caellwyn/ou_student_predictions
[5] https://github.com/mohammedAljadd/Students-performance-and- difficulties-prediction