Using Logistic Regression to Predict Secondary School Student Performance

Shahnoor Ali, Hasan Sarwar, Hala Yasmin, Umar Hayyat, Zain-ul-Abidin, Muhammad Rehman Shahid
shahnoorali9716@gmail.com, hasansarwarwarraich@yahoo.com, halayasmin41@gmail.com,
umar187hayyat@gmail.com, zainkhan8788@yahoo.com, mrehman0892@gmail.com
Department of Computer Science, National Textile University, Faisalabad, Pakistan

***Abstract*— Education is one of the principle establishments for the student's advancement and furthermore for public human asset improvement. Failure at college and grade retention is an essential concern among students and their parents. This study aims to assess student's failure in the core subject Mathematics and the factors that influence failures mendaciously. Logistic Regression was applied to the dataset. Dataset was trained and tested, and then the predictions were made about the accuracy of the results. The whole dataset was firstly transformed from nominal to numeric values then into binary classes of 0 or 1 (whether the student failed or not). Absences, weekly study time, age ranges from 15 to 19, going out with friends are the radix in student's failure. The study was performed on a total of 395 students from which mostly male students were classified as failed in mathematics. The accuracy of results was found through a classification report and confusion matrix. Confusion matrix gives an accuracy of 85%. This study has thriving highlighted the prevalence of multi-factorial contributors such as social and health related factors for college failure. Social related factors were found to be more prevalent. The cynosure of this paper is the application of Logistic Regression and finding out how accurate the model results are predicted. The precision of 85% is predicted, revealing a good fit.**

Keywords*— *Machine Learning, Prediction, Performance, Students, Logistic Regression.

I. INTRODUCTION

Our education system today is in a state of obvious disrepair. The failure of students that erupted during the last quarter year 2018 convulsing the country's social and economic development. Because students are the most critical asset for the educational institution, their performance takes on an essential job to become the best graduate class that will end up being a leader, a pioneer, a premier and a worker of a specific nation, responsible for social and financial progress of the nation. Educating proficient and compelling human powers is considered among the primary obligations of colleges. Consistently, colleges graduate and concede newcomer students; in this constant cycle, education quality has a critical position. Hence, expanding the nature of instructive framework is viewed as the most persuasive factor In building up the nations; this is because students accomplish a situation because of their scholarly achievement in which their greatest inward and external powers are utilized for accomplishing objectives of advanced education and acquiring fundamental conditions for fruitful public activity. Then again, the absence of achievement in training clears the ground for a few individuals, particular and social issues and deviation from accomplishing the objectives of the instructive framework. The reasons for academic failure incorporate familial, health and financial issues that lead to an absence of eagerness for establishments, inspirational and physiological issues, psychological and neurological hindrances to realizing which prompts the misuse of current consumption and time. High rates of college disappointment have been trailed by review redundancy which has turned into a particular trademark in colleges even in the creating nations. This was not the first-run in which the disappointment of students was evaluated, various examinations and research were enveloped to asses disappointment in the nearness of social, mental, wellbeing and school related elements. Smoking, drinking, drug abuse can be one of the factors of student failure with its ramification of a student losing self-confidence, becoming discouraged and decreasing their effort in work. Other factors evolve such as truancy from classes, dropping out, redoing the grade or nether education. It has been observed that a significant number of students (about 20%) are hypothetically primitive and failed to achieve good marks. To assist those students who are encountering academic failure, fall into diverse categories as an evasion, impedance, and remediation. Considering the above-mentioned points, student failure is of crucial importance at this hour as the advancements have been made in developing countries like Europe.

Moreover, to meet the furtherance of these countries, new techniques, proficiency in a particular field, craftsmanship must be instigated and cramming must be halted in our education system. The factors that positively out-turn the student's academic performance include teacher-student relationship, teacher's welfare (teacher's salaries paid on time), home background (parents' ability to supervise their children about admission, knowledge), friendly principle-teacher relationship, effective teacher's supervision [1]. If these factors have contemplated by all the people who are accountable and indispensable in student's success and failure, then student's mental, psychological, social and

family problems can be obliterated. In this paper, we applied an algorithm, i.e. Logistic Regression on a dataset collected on the failure of Portuguese students collected by one of the Portuguese researchers. The dataset is cut to short incorporating 13 variables according to the feature selection of 5 research papers, integrating common features. We have reckoned students' failures on the premise of causes and rationale given by other researchers according to their dataset. This paper comprises of at least five almost same research papers with different findings, variables, techniques, and algorithms but the question was the same: What are the reasons and causes for the failure of students? Our goal was to apply the non-identical algorithm to the dataset and check the results and accuracy of the dataset through a classification report. Portuguese researchers applied Correlation, Random Forest and Decision Tree to his dataset [2]. In comparison to these algorithms, we hand-pick Logistic Regression.

The paper is organized as in Section 1 it introduces about the importance of education and student's academic performance. Section 2 reviews the literature survey of student's academic performance in different research papers and factors influencing their performance. Section 3 is about all the materials and methods that are used in this paper which includes student's variables, statistical techniques, data mining algorithm and in which computational environment model is tested in. Section 4 comprises of the results that come from the algorithm and its accuracy. Section 5 is the conclusion of the whole research paper.

II. Literature Survey

Most of the assorted studies have been conducted to find out student's academic performance [3][4][5][6]. According to studies, about 20% of students are failed no less than one time amid their education and this disappointment not just aims some psychological, mental issues for them, yet in addition puts them at the danger of educational deprivation and hardship thinking contemplating their academic breakthrough, harms optimum utilization of scientific principles for training human powers and financial resources and furthermore social disappointment [7]. Moreover, students' dropout and academic failure cause a few difficulties, issues and challenges for the students themselves alongside colossal misfortune for the nation. Researches have demonstrated that students with academic failure are increasingly foreseeable to utilize drugs at older ages; like this, dropout and academic failure might ensue liquor and drug addictions [8].

A study in one of the Arabian colleges on students[9] who perpetrate suicide, encountered coma/fainting, cardiovascular diseases, asthma, visual problems showed academic failure as the most popular reason for their disease [7]. Different examinations have proposed that different factors can prompt academic failure; some studies have considered the use of illicit drugs and various investigations have shown that personality factors, incentives, interest, fulfilment, abandonment, achievement desire, and family circumstances can influence the level of academic success in colleges [2]. In a thorough modus operandi, the variables associated with college failure can be organized into three classifications of internal organizational factors (proficient characteristics of instructors, space and appropriate facilities and equipment)[10], external organizational factors (guardians' education level and their dealing with students' academic failure, financial situation of families, misty and indeterminate job prospects [1]) and individual factors (components like having a objective, inspiration, planning, examining strategy, intelligence, consideration, anxiety, affective disorder and mental problems and lack of attendance to the course [7]). Education's quality may take a back seat to education's quantity. There are too many endeavors that make our education system hitting the rock-bottom in quality. According to studies, this issue is escalating each year with the goal that numerous students cannot manage the curriculum or finish it in due course [11].

According to the last few years[12][13], several various significant studies have been carried out to develop different models for assessing students' performance by considering different factors like family pay, direction from parents, teacher and student relationship, school distance and sex of the students, but these studies have not investigated the learning structures, communication skills and proper guidance of parents. *Table 1* comprises of several research papers that describes the causes and factors of students' academic failure, the variables used in the research paper, their author name, in which year they are published in, their sample size, statistical analysis and in which tools the algorithms are implemented in [11][14].

III. Material and Methods

STUDENT DATA:

Table 3 comprises of 13 variables that includes sex, age, cohabitation status of parents, mother's education, father's education, weekly study time, number of past class failures, extra-curricular activities, desire to obtain higher education, access to the internet at home, in a romantic relationship, quality of family relationships, meeting with friends, current state of health and number of school absences. Some variables were numeric, and others were nominal. In order to apply Logistic Regression, Data wrangling was done on the dataset. Some of the nominal variables of dataset were transmuted into numeric and also into dichotomous, i.e. 1 or 0. Above mentioned variables were further censored in significance with failure. Only those variables were transformed that were correlated closely to failure, i.e. sex, weekly study time, internet access at home, extra-curricular activities, number of past class failures, cohabitation status of parents, desire to obtain higher education, in a romantic relationship (
Table 2). Incorporating with the data which is only in numerical form because for evaluating final results it would be difficult to convert the string into a float. For the description of the variables used in the dataset, *Table 3* elucidates the variables, i.e. their category, when in 0 or 1 form and what they interpret.

Table 1: Literature Survey

Author	Context	Variable	Sample Size	Statistical Analysis	Analysis Tools
P. Cortez (2008)	Failure of students	32	650	Classification, Regression	RMiner
Madeeha (2009)	Child's failure in school and grade retention	64	699	Simple random sampling method	Excel
C. Gbollie (2017)	Cause and Reasons for the failure of students	13	323	Correlation, Mean, Standard Deviation	Statistical packages for the social sciences (SPSS 17.0.)
Irfan Mushtaq (2012)	Factors contributing in failures of students	5	155	Mean, Standard deviation, correlation,	Appropriate statistical package
L. Kalagbor (2012)	Factors positively influencing on student's academic performance	10	650	Frequency, percentage, Mean	Excel

DATA MINING MODEL:

This study is based on the information gathered during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal. P. Cortez and A. Silva [2] have estimated failure of students using data mining techniques which includes Classification and Regression, Decision Trees, Random Forests and by integrating Business Intelligence techniques in Education. Predictive Modeling Technique, Regression Analysis, which always implies prediction. It estimates the relationship between an independent variable(predictor) and a dependent variable(target). In this paper, estimation is done on predictive analysis, not on prescriptive analysis as it is used to solve classification problems, not regression problems [15].

In this paper, Logistic Regression is done on the dataset of 396 students after picking out the 13 variables out of 32, and dataset only belongs to the students who failed in Mathematics. Failure of students in the Portuguese language is not estimated. Those variables have extracted that ascendance on failure the most. This algorithm produces results in a binary format that is used to predict the outcome of a categorical dependent variable. A statistical technique, Logistic Regression, used in research projects that require the analysis of the relationship of dependent variable or of a result with one or more independent variables or predictors when the dependent variable is either (a) Dichotomous, with only two classifications, for instance, if one has failed (yes or no); (b) unordered polytomous, which is a nominal scale variable with three or more categories, for example, the quality of family relationships (from 1 - very bad to 5 - excellent); or (c) ordered polytomous, which is an ordinal scale variable with three or more categories, for example, the completed level of education (e.g., less than primary school, primary school, secondary school, an undergraduate degree, or a post-graduate degree). The logistic regression was employed to study the relationship between the failure status of the student, their Age, Gender, study time (intensity of course) and the extra-curricular activities are other than studying Mathematics. The failure status of a student was categorized as never failing in any subject (1) and failing in at least one subject (0). The histogram in *Figure 1* implements that some failures in mathematics (one or more than one) are less than the number of no failures and failure rate is high in males than in females *(Figure 3).* The fact that math is arduous for males because activities including going out with friends' which consequences in a short period of study time. As this subject requires the most attention and hard work, they fail to do so that fails the respective subject.

Figure 2 explains how badly these two factors (going out with friends and failure) influence each other. Different colors indicate several class failures: Red (3 class failures), Green (2 class failures), Yellow (1 class failure) and blue (no failure) in *Figure 2*. The Logistic Regression is derived from the *Straight-Line Equation(1)* and then reducing the *equation(1)* ranging only from 0 to 1 resulting *equation(2)*. In this way, the Logistic Regressions' predictions are in the form of probabilities of an occasion happening, i.e., the likelihood of y=1, given specific estimations of input variables x. Hence, the results of LogR range between 0-1. LogR models the information using the standard logistic function, which is an S-shaped curve given by the equations *(3)* and *(4)*.

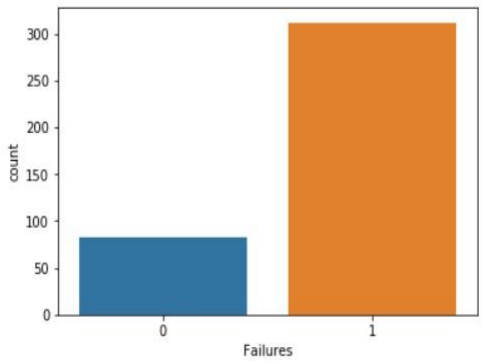

Figure 1: Histogram of Respective Failure

Table 2: Student Data Demographics

6]:

	age	famrel	goout	health	absences	Sex	Study Time b/w 2 to 5hours	Study time b/w 5 to 10hours	Internet	Activities	Failures	PStatus	Higher	Romantic
0	18	4	4	3	6	1	1	0	0	1	1	0	1	0
1	17	5	3	3	4	1	1	0	1	1	1	1	1	0
2	15	4	2	3	10	1	1	0	1	1	0	1	1	0
3	15	3	2	5	2	1	0	1	1	0	1	1	1	1
4	16	4	2	5	4	1	1	0	0	1	1	1	1	0

Table 3: Description of Student Data Variables

Attribute	Description	For Logistic Regression
Age	Age of student (numeric: from 15 to 22)	_____
Famrel	Quality of family relationships (numeric: from 1 – very bad to 5 – excellent)	_____
Gout	Going out with friends (numeric: from 1 – very low to 5 – very high)	_____
Health	Current health status (numeric: from 1 – very bad to 5 – very good)	_____
Absences	Number of school absences (numeric: from 0 to 93)	_____
Sex	Gender of Student (binary: female or male)	1 = female 0 = male
StudyTime	Weekly study time (numeric: 1 – < 2 hour, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)	1 = 2 to 5 hours 0 = 5 to 10 hours
Internet	Internet access at home (binary: yes or no)	1 = no 0 = yes
Activities	extra-curricular activities (binary: yes or no)	1 = no extra-curricular activities 0 = yes to extra-curricular activities
Failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)	1 = no failure 0 = one or more than one failure
PStatus	Cohabitation status of parents (binary: apart or living together)	1 = parents are apart 0 = parents are living together
Higher	wants to take higher education (binary: yes or no)	1 = yes to higher education 0 = no to higher education
Romantic	in a romantic relationship (binary: yes or no)	1 = in a romantic relationship 0 = not in a romantic relationship

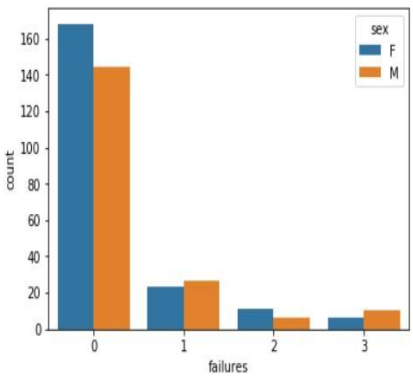

Figure 3: Histogram of Failures against gender

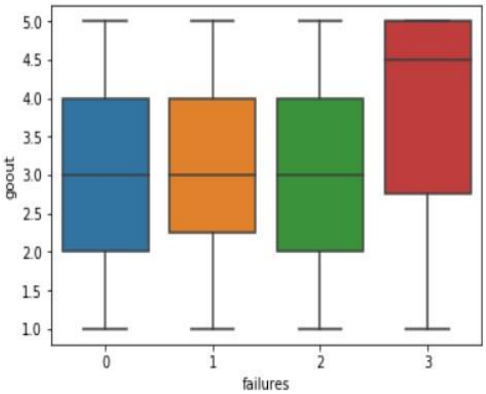

Figure 2: Box-Whisker Plot of Failures against Going Out with Friends

$$Y = C + B_1X_1 + B_2X_2 + \cdots \rightarrow Range\ is\ from\ -(infinity)\ to\ (infinity) \quad (1)$$

By Reducing

$$Y = C + B_1X_1 + B_2X_2 + \cdots \rightarrow In\ Logistic\ Equation, Y\ can\ be\ only\ from\ 0\ to\ 1 \quad (2)$$

$$\frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x} \quad (3)$$

$$\log\left[\frac{Y}{1-Y}\right] \rightarrow Y = C + B_1X_1 + B_2X_2 \quad (4)$$

Equation (4) is used to predict values of y. Predominantly it classifies and gives the result whether students failed or not keeping in view the factors that influence failure the most. Results will be predicted in the form of Sigmoid-curve which is used to predict the values of y. For training the data in order to apply logistic regression to the dataset, several code lines are required. The model is built on the train data, and the output is predicted on the test data. The output or predicted value is taken "Failure," and all the other values are taken as input. Failure is predicted against each input. Using the split function, dependent and independent variables are passed to the function. Split size is set as 0.3 (70:30 ratio). First, the model is created and fitted then predictions are made, and the evaluation is done on the question: how the model is performing? Performance can be checked though classification report and accuracy. The classification report is given in *Figure 4*. The f1-score describes the harmonic mean of recall and precision. The scores relating to each class clarify the accuracy of the classifier in characterizing the data points in that particular class compared to all other classes. The support is the number of samples of the true responses that lie in that class.

Accuracy is estimated with the help of a confusion matrix. Confusion matrix has four outcomes i.e. predicted no (PN), predicted yes (PY), actual no (AN), actual yes (AY). In *Table 4*, at the intersection of PN and AN, PY and AY, these can predict the accuracy of the model by adding both the values then dividing by the number of values. Accuracy score is 86%. This tells us that our results are 86% accurate.

COMPUTATIONAL ENVIRONMENT:

All experiments revealed in this study were directed utilizing Python Anaconda Jupyter Notebook, an open and free source distribution of the Python and R programming languages for scientific computing (machine learning applications, data science, predictive analysis, data processing for large-scale, etc.), which aims to simplify deployment and package management. Anaconda Navigator is a desktop GUI (Graphical User Interface) integrated into Anaconda distribution that enables users to launch applications and manage anaconda packages, environments, and channels without using command-line commands. Navigator can search for packages in Anaconda Cloud or a local Anaconda Repository, install them in an environment, run packages and update them. It is available for Windows, macOS, and Linux.

	precision	recall	f1-score	support
0	0.83	0.24	0.37	21
1	0.86	0.99	0.92	98
micro avg	0.86	0.86	0.86	119
macro avg	0.85	0.61	0.64	119
weighted avg	0.85	0.86	0.82	119

Figure 4: Classification Report of y-test and predictions

Table 4: Confusion Matrix for the Predictions

	PN	**PY**
AN	5	16
AY	1	97

IV. RESULTS

In this research, a total of 395 students who had failed in academics were identified. The study group consists of 208 females and 187 males. Before fitting the models, some preprocessing was required by the Logistic Regression Model. The nominal variables (e.g., sex, Pstatus, activities, higher, internet, romantic) were transformed into numeric, and all attributes were in 0 or 1 form. Next, the model was fitted by splitting the dataset in 70:30 ratio (Logistic Regression (C=1.0, class-weight=None, dual=False, fitintercept=True, intercept-scaling=1, max-iter=100, multiclass='warn', n-jobs=None, penalty='l2', randomstate=None, solver='warn', tol=0.0001, verbose=0, warmstart=False). The accuracy of the predictions was estimated through classification report comprises of precision, recall, f1-score, support *(Figure 4).* As an example of the quality of predictions, *Table 4* demonstrates the confusion matrices for the Logistic Regression Algorithm. 86% of the values are predicted accurately.

V. CONCLUSION

Education is even more in a crucial position today. Today's students can face future hindrances if their school activities and informal learning prepare them for adult roles such as natives, representatives, administrators, guardians, volunteers, and business visionaries. In this paper, we have addressed the prediction of the grades of secondary school students in a core class i.e., Mathematics by using previous school grades, demographic, social and other school related

data. First, the data was analyzed then trained and tested. The algorithm, i.e. Logistic Regression was applied to the dataset. It concluded that past academic performances, extra-curricular activities, going out with friends, these social factors cause academic failure. Social factors do count when it comes to academic failure. The applied algorithm predicted the model results with accuracy count of 86%.

ACKNOWLEGMENT

We are thankful to our instructor, Dr. Sohail Jabbar for his insight guidance, and suggestions during this research work that is part of our undergraduate subject. He motivated and encouraged us, helped us in gathering dataset to complete this work.

REFERENCES

[1] Ayimah J . and Agbotse G . 2012 An analysis of factors influencing students' academic performance in Ho Polytechnic *Journal of Polytechnics in Ghana(Jopog)* vol 5 pp 113–32

[2] Cortez P and Silva A 2008 Using Data Mining To Predict Secondary School Student Performance *Proc. 5 th Annu. Futur. Bus. Technol. Conf.* **2003** 5–12

[3] Gbollie C and Keamu H P 2017 Student Academic Performance: The Role of Motivation, Strategies, and Perceived Factors Hindering Liberian Junior and Senior High School Students Learning *Educ. Res. Int.* **2017** 1–11

[4] Wibawa A P, Mushtaq I and Khan S N, 2014 The Relationship Between Background Education, Socio-Demographic And Lifestyle Factors And Academic Performance *J. Holist. Nurs. Midwifery* **27** 65–73

[5] Farhan M, Jabbar S, Aslam M, Hammoudeh M, Ahmad M, Khalid S, Khan M and Han K 2018 IoT-based students interaction framework using attention-scoring assessment in eLearning *Futur. Gener. Comput. Syst.*

[6] Gómez-Aguilar D A, Hernández-García Á, García-Peñalvo F J and Therón R 2015 Tap into the visual analysis of customization of a grouping of activities in eLearning *Comput. Human Behav.* **47** 60–7

[7] Alsaimary I, Al-Sadoon M, Jassim A and Hamadi S 2009 Clinical findings and prevalence of helicobacter pylori in patients with gastritis B in Al-basrah governorate. *Oman Med. J.* **24** 208–11

[8] Farhan M, Jabbar S, Aslam M, Ahmad A, Iqbal M M, Khan M, and Maria M-E A 2017 A Real-Time Data Mining Approach for Interaction Analytics Assessment: IoT Based Student Interaction Framework *Int. J. Parallel Program.*

[9] Semerci A and Aydın M K 2018 Examining High School Teachers' Attitudes towards ICT Use in Education *Int. J. Progress. Educ.* **14** 93–105

[10] Farhan M 2011 An Interactive Assessment Methodology to Enhance Teaching Performance and Student Experience in e-Learning Environment Submitted by An Interactive Assessment Methodology to Enhance Teaching Performance and Student Experience in e-Learning Environment

[11] Farhan M, Aslam M, Jabbar S and Khalid S 2016 Multimedia based qualitative assessment methodology in eLearning: student teacher engagement analysis *Multimed. Tools Appl.* 1–15

[12] Turkish T, Journal O and Technology E 2010 TOJET: The Turkish Online Journal of Educational Technology – January 2010, volume 9 Issue 1 **9** 176–84

[13] Dickinson S J 2013 *Shape Perception in Human and Computer Vision*

[14] Iqbal M M, Farhan M, Saleem Y and Aslam M 2014 Automated Web-Bot Implementation using Machine Learning Techniques in eLearning Paradigm **4** 90–8

[15] Paul A, Ahmad A, Rathore M M and Jabbar S 2016 Smartbuddy: Defining human behaviors using big data analytics in social internet of things *IEEE Wirel. Commun.* **23** 68–74