Student's Success Prediction in the Secondary Level of Education Using a Linear Regression Model

Fuad Dedić[*], Nina Bijedić[*], Elmir Babović[*] and Dražena Gašpar[**]

[*] Faculty of Information Technologies, University "Džemal Bijedić", Mostar, Bosnia and Herzegovina

[**] Faculty of Economics, University of Mostar, Mostar, Bosnia and Herzegovina

Abstract - The paper presents research about the use of linear regression model for the analysis of the interrelation of the grades by subject in 1st semester to a final grade of class (after 2nd semester) in secondary school education level. The authors used linear regression to create equations for description relation between grades as independent variables and rating (final grade) as a dependent variable. Final grade calculates as an average of all subjects' grade. The research was based on the set of data related to the success of Persian-Bosnian College students, Bosnia and Herzegovina. This research should help educational institutions to predict students' success, i.e., the final grade of semesters of 1st and 2nd class based on subjects' grades in 1st semester of 1st class. Based on this pre-diction, educational institutions can make the teaching processes more efficient and be more attractive for future students.

I. Introduction

The secondary level of the education system in Bosnia and Herzegovina aims to prepare students for entry into the labor market or higher education. The primary quality indicator is students' success and school management policy to help students to achieve better success [1].

Management of secondary schools is under a considerable challenge to prepare students for entry into a highly competitive, dynamic, high-tech, complex, and interdisciplinary global labor market and higher education environment. They can respond to that challenge by equipping their students with appropriate knowledge, skills, and competences [2]. One of the significant changes in the education market of Bosnia and Herzegovina is the continuous activity of high school managers to achieve a better position in the education market. The market struggle is caused not only by competition between public and private schools but also by the presence of a lot of organizations that offer various informal programs. Although this type of education cannot be compared with formal education, it is a fact and, as such, must be considered [4]. For example, these organizations offer a comprehensive list of short-term specialized courses to citizens [5-6]. Such learning delivery platforms put pressure on institutions providing formal education to upgrade their curriculum continually and finally define a more flexible educational process. The goal is to maintain systems' standards concerning learning outcomes.

Regression analysis is a statistical technique for investigating and modeling the relationship between variables and is used to build a prediction model, i.e., the best fitted model with minimum squared errors of the fitted values and further applied to data for continuous value predictions.

In the process of the curricula development and implementation, the decisions what is to be taught, for what reasons, and how learning should look like are crucial for the success of curricula. In order to ensure successful achievement of planned learning outcomes, the curricula have to be carefully designed and implemented.

The way parts of the curricula are designed, their complexity and volume, and their sequence within curricula are defined by the Ministry of education, which is a competent state institution.

That is why the authors raised the question: is it possible by using linear regression analysis to determine equations for describing the relation between grades by courses (subjects) of 1st semester of 1st class and final grade of another semester of 1st and 2nd class.

The educational process is divided into two parts. 1st and 2nd classes make the first part and subjects are the same for all students in both classes. 3rd and 4th classes are considered as a second part, and subjects depend on students' choice. Research is performed using data from 1st and 2nd class only.

The research aims to determine the degree of impact of successes achieved in 1st semester of 1st class to the success achieved on a later semester during the 1st and 2nd class.

II. METHODOLOGY

Within the research, the set of data related to the success of students of Persian-Bosnian College is used. The final grade of every semester is the average of subjects' grades. There is not any kind (especially not mathematical equation for calculating) relation between subjects' grades of 1st semester of 1st class and final grade of other semesters.

Aims of research are to:

a) define the relation between grades from 1st semester of 1st class and final grade of 2nd semester of 1st class and final grade of 1st and 2nd semester of 2nd class

b) using linear regression model create an equation for description relation between grades as predictors (independent variables) and final grade as a response (dependent variable)

The research consists of 4 cases. Aim of all cases is to determine the linear regression model between subjects' grades of 1st semester of 1st class and final grades of all other semesters. The analysis was performed on 1st, and 2nd class because curricula are the same and consist of the same subjects (courses). This paper presents linear regression equation for subjects' grades of 1st semester of 1st class versus 1st semester of 2nd class.

If schools' management has early information about the potential worst success of students (lower grades), it can organize the additional support for students to help them successfully pass the courses recognized as potentially challenging for most of them.

The analysis was performed in R programming language using packages: RSQLite and dplyr [7], [8].

Accuracy, precision, recall, and F-measure are parameters used within research to determine the models' efficiency. Accuracy is the necessary parameter, but for more reliable results of efficiency, it is advisable to use other parameters. In order to determine the values of the efficiency parameters, it is necessary to create a confusion matrix [9]. The general form of the confusion matrix is given in Table 1.

TABLE I. The general form of confusion matrix [9]

		Actual	
		T	N
Predicted	T	True positive (TP)	False positive (FP)
	N	False negative (FN)	True negative (TN)

Using the marks from the confusion matrix, the parameters can be represented by the following expressions.

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

Accuracy represents the percentage of correctly classified instances in the data set.

$$precision = P = \frac{TP}{TP + FP} \quad (2)$$

Precision (also called positive predictive value) is the number of positive predictions divided by a total number of positive class values predicted.

$$recall = R = \frac{TP}{TP + FN} \quad (3)$$

Recall (also called the true positive rate, the sensitivity) is the number of positive predictions divided by the number of positive class values in the test data.

$$F - measure = \frac{2 \cdot R \cdot P}{R + P} \quad (4)$$

F-measure (also F-score or F1 score) is a measure of a test's accuracy. That is, the harmonic mean (average) of the precision and recall, and F-measure is the best when precision (P) and recall (R) are balanced [18].

III. RESULTS

Figure 1 shows the process flow. The process is starting with data preprocessing. After the pivot table is created, and the missing value is imputed, the appropriate table from the database is connected, and data is attached to the data frame. The Data frame is divided to train and test data. Using a training dataset linear regression model is created. The created model is evaluated, and if the model is good enough, it is used for prediction on test data. Prediction results are evaluated by creating confusion matrix and calculating accuracy, precision, recall, and F-measure.

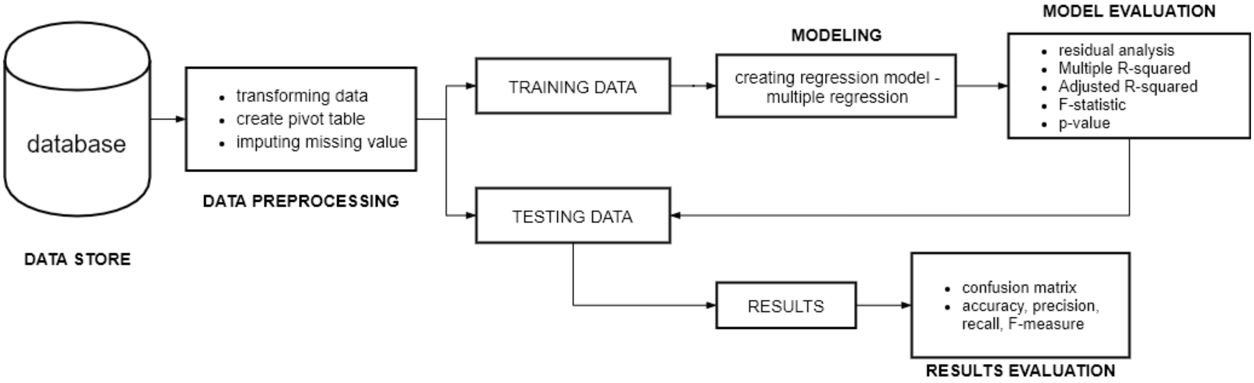

Figure 1: Diagram process flow

IV. CASE STUDY

All subjects of 1st semester of 1 class as predictors

Aim of case study is to define the linear regression model between the final grade of 1st semester of 2nd class and grades by subjects of 1st semester of 1st class.

New data frame c2s1 (class_2_semester_1) is created using data frame class_2_rating. The new data frame is consisted of grades by the subject of 1st semester of 1st class and 1st semester of 2nd class final grade.

There is not subject with p-values smaller than 0.1, and they will not be performed next iteration with a new linear regression model consist of these subjects' grades as predictors.

It was created diagnostic plot of model. The red line on the bottom-left plot is relatively flat in range 1-4 of fitted values. Generally, the red line in the bottom right plot is partially flat regarding the dashed line. On the plot exists two Cook's distance values (0.5, 1), but there is no point which lies outside of Cook's distance, so this is very good fit regression.

Regression equation - all subjects' grades of 1st semester of 1 classes predictor

class_2_sem_1 ≈ -0.009362 + 0.090047s3_a + 0.126428s5_a + 0.393115s6_a-0.226720s11_a + 0.273442s12_a - 0.191545s13_a - 0.063349s14_a - 0.119358s15_a-0.102640s16_a + 0.110209s19_a + 0.075932s623_a + 0.226710s773_a + 0.344334s5021_a

Prediction is performed on test dataset, i.e. data frame prediction_c2s1_a. Quality of prediction is checked by confusion matrix.

TABLE II: Accuracy, precision, recall, and F-measure for case study

Accuracy	Precision	Recall	F-measure
0.7	0.7777778	0.875	0.8235294

All parameters (accuracy, precision, recall and F-measure) have high values, so the model is rated as very good.

V. EXPERIMENTAL RESULTS DISCUSSION - PROOF OF HYPOTHESIS

This research is based on Statistical Learning. Experimental results show success in proof of concept and proof of the hypothesis. Aim of all cases is to define linear regression model final grade of semesters of 1st and 2nd class against grades by subjects of 1st semester of 1st class.

All models' residuals are small. Medians of all models are close to 0. Distinctions 3Q and 1Q are low (≈ 0.01), and the absolute value of distinction 1Q and 3Q to the median are low (lower than 0.01). F-statistics are predominantly excellent. P-values of all models are lower than 0.1. The degrees of freedom are relatively small. Multiple R-squared and adjusted R-squared are close to 0.

The diagnostic plot checks models of quality for all cases for training and confusion matrix for the testing phase.

The diagnostic plot consists of 4 plots: residuals vs. fitted values, standardized residuals, square root of standardized residuals, and standardized residuals vs. leverage. Fitted values are grade in the range of 1 to 5.

The first plot is residuals versus fitted values plot. The red line on the plot is a smooth curve regarding residuals, and the dotted gray line is the regression line. The analyzed case red line has a low deviation in the range of 1 to 4. In the range of 4 to 5 red line has a high deviation.

The second plot or Normal Q-Q plot shows the normal of residuals. If all residuals lie close to the gray dash line, residuals are normally distributed. The analyzed case residuals lie close to the gray dash line.

The third plot is used to measure the square root of the standardized residuals against the fitted value. If the red line is relatively flat, the assumption is correct. The analyzed case red lines are not so flat, but values are in the range of size 1 or lower, so it is concluded that assumptions are correct.

The fourth plot shows standardized residuals versus leverage. The leverage is a measurement of how each data point influences the regression. The last plot contains the Cook distance line, too. Cook's distance is a measure of how influential an instance is to the computation of a regression. The case analyzed mostly have not points which are outside Cook's distance.

Accuracy, precision, recall, and F-measure values of all cases, except case 4, are high.

The important factor of research is dataset size. Dataset has 126 rows, so training dataset in all cases have 88 rows. Even the number of rows is low, linear regression models in most cases have relatively high to very high prediction. The assumption is that greater dataset size can lead to a better model, i.e., better pre-diction.

VI. CONCLUSION

In this paper, results prove that linear regression can be trained to provide a platform for predicting the success of students based on their grades achieved on 1st semester of 1st class, i.e., at the beginning of secondary schooling. If we consider that each learning outcome builds over several different subjects, linear regression can also be used to check the definition of learning outcomes in a given curriculum.

Research is limited to predicting student success in first and second class in high school. The generalization to other classes and types of schools is not possible. According to the high school curriculum, all students in the first and second classes of high school attend the same subjects. In the third and fourth classes, students study in separate programs. Other types of schools educate students for different kinds of professions. For example, one school educates students for electricians, civil and mechanical technicians, and education of each profession types realize by the separate curriculum. In this case, the research plan for predicting students' success of each profession type should be set up differently.

Future research can be related to predicting students' success of third or fourth class in high school based on success in the first semester of the first class. Also, future research will expand the current model with other prediction methods as well as additional datasets from the different educational institutions on the current level and on other educational levels.

REFERENCES

[1] Gašpar, D., Rezić,S. (2014): Information Technology and Strategic Management of Universities, Journal of Business and Economics, ISSN 2155-7950, USA, November 2014, Volume 5, No.11, pp 249-261.

[2] M.Mabić, D.Gašpar (2018). Facebook as a Learning Tool - Students' Perspective. Proceedings of the Cen¬tral European Conference on Information and Intelligent Systems 29th CECIIS, September 19-21, Vara¬ždin, Croatia

[3] Gašpar, D., Mabić, M., (2015): Student engagement in fostering quality teaching in higher education, Jour¬nal of Educational and Research, MSCER Publishing, Rome, Italy, Vol.5. No 1 SI, April 2015, ISSN 2239-978X, ISSN 2240-0524.

[4] ElearnSA (2016). eLearning Market Analysis. Retrieved June 10, 2019, from http://elearnsa.co.za/wp-con¬tent/uploads/2015/04/Elearn-Value-Proposition.pdf

[5] Docebo (2014). E-Learning Market Trends & Forecast 2014-2016 Report, Retrieved June 10, 2019, from https://www.iconcept.nl/publicfiles/136/bestanden/elearning-market-trends-and-forecast-2014-2016-do¬cebo-report.pdf

[6] Docebo (2016). Elearning Market Trends and Forecast 2017-2021. Retrieved June 10, 2019, from https://eclass.teicrete.gr/modules/document/file.php/TP271/Additional%20material/docebo-elearn¬ing-trends-report-2017.pdf

[7] K. Müller, H. Wickham, D.A. James, S. Falcon, L. Healy (2020), Retrieved from April 15, 2020, from htt¬ps://cran.r-project.org/package=RSQLite

[8] H. Wickham, R. François, L. Henry, K. Müller, (2020) Retrieved from April 15, 2020, https://cran.r-projec¬t.org/package=dplyr

[9] Novaković, J.Dj., Veljović A., Ilić, S.S., Papić, Ž., Tomović, M. (2017). Evaluation of Classification Models in Machine Learning. Theory and Applications of Mathematics & Computer Science, Volume 7, Issue 1, pp. 39 – 46. Retrieved July 17, 2019, from https://www.uav.ro/applications/se/journal/index.php/TAMCS/article/download/158/126.