ISSN 2278-3091

Volume 12, No.1, January - February 2023

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse011212023.pdf

https://doi.org/10.30534/ijatcse/2023/011212023

Predicting the Assessment Course Performance of Criminology Students Using Data Mining

Eugene P. Iglesias[1], Rogelio Badiang[2]

[1]Graduate Student, University of the Immaculate Conception, Philippines, eiglesias_210000000556@uic.edu.ph

[2]Graduate School Professor, Philippines, University of the Immaculate Conception, rbadiang@uic.edu.ph

Received Date December 15, 2022 Accepted Date: January 18, 2023 Published Date: February 06, 2023

ABSTRACT

Living in this changing era where things constantly change overtime, transitioning means not a big thing. Obviously, nowadays even learning often takes place outside of traditional educational settings. This study aims to determine the performance of criminology students in the assessment course. The research covers the analysis of the performance of Criminology students of Legacy College of Compostela in the Assessment Course with six (6) subject areas. The data were taken from the College of Criminal Justice Education (CCJE) students' evaluation, Multiple Linear Regression was employed to predict the students' performance in the assessment course. The data mining study results were acquired using IBM SPSS as the Modeler to transform the data and extract relevant information, which was then used for the conclusion. Based on the results, it can be concluded that the subjects of Crime Detection and Investigation and Law Enforcement Administration significantly influence the outcome of the students' assessments. Therefore, to concentrate on reviewing other areas with students, such as correlational administration, criminalistics, criminal law and procedure, and criminal sociology, might be useful in improving students' performance.

Key words: Data Mining, assessment of course performance, criminology students, Multiple Linear Regression

1. INTRODUCTION

The world is constantly changing, which means that nowadays, learning occurs outside of formal school settings. Students must choose what and when to study, whether they have mastered the content sufficiently to quit practicing it, and how to divide their time between other subject [10]. This possibly results in achieving the basis of having quality education and capabilities of an institution to produce professionals. Before a criminology student will graduate, an assessment course or other call it pre-review course needs to be passed by the students. This course includes six subject areas with different courses under it.

The application of data mining methods in the field of education has attracted great attention in recent years. Data Mining (DM) is the discovery of data. It is the field of discovering new and potentially useful information or meaningful results from big data [12]. It also aims to obtain new trends and new patterns from large datasets by using different classification algorithms [2].

1.1 Purpose of the Study

This section describes the purpose of conducting the study.

1. To determine the performance of criminology students in the assessment course.

2. To assist educators in determining what subject/s could affect the performance of the students in the assessment.

1.2 Scope and Delimitation of the Study

This research covers the analysis of the performance of Criminology students of Legacy College of Compostela in the Assessment Course with six (6) subjects such as Crime Detection and Investigation, Correctional Administration, Criminalistics, Law Enforcement and Administration, Criminal Jurisprudence and Procedure, and Criminal Sociology for the academic year 2021-2022.

1.3 Related Literature and Studies

A few studies have been made in education data mining for discovering different patterns to improve the students' performance. [4] studied the use of data mining techniques using the Apriori algorithm on a set of students of Istanbul Eyup I.M.K.B.Vocational Commerce High School. They have taken the dataset of 28 students for 74 courses for minimum support rate 9 and as minimum confidence rate 85%. In their study they have revealed that if a student failed a particular subject in class 9th then he/she will fail next year as well. It discovered the rate of successful students by finding the rate of unsuccessful students which will help the student in choosing the right subject.

There are many approaches to prediction student performance, data mining techniques are one of the most well-recognized and the most well-used techniques in data mining are classification and regression. The researchers applied them to forecast student performance, Strecht et.al. [6] conducted to predict students' grades in their work and their results (pass/fail). To predict the student's results, a classification model was applied, while a regression model was used to forecast the grades. For classification, decision trees and SVM were used, and for regression analysis, SVM Random Forest and AdaBoost.R2 were used. Based on their study, the classification model was shown to be capable of obtaining useful patterns, while regression methods were unsuccessful to prevail over a simple baseline.

In addition to Jayakumar [8] in educational data of college students, the result classifies the student. Based on previous student results, they predict the future student result using J48, NB, MLP, and random forest; however, classification accuracy is not very high. However, Naïve Bayes has a correct classification accuracy compared to rest three. The algorithm was trained on the same data of the 2007 batch of 51 students with the use of the software tool Weka used. The algorithms were tested on 2 A given table is showing the comparison of classification accuracy of these algorithms on the training as well as test data. The classification accuracy of these algorithms is very high for training, however, it gets reduced for cross-validation. Among these algorithms, Naïve Bayes is giving good classification accuracy.

Furthermore, Sen and Ucar [9] compared the achievements of Computer Engineering Department scholars in Karabük University according to some factors similar as age, gender, type of high academy scale and the scholars studying in distance education or regular education through data mining ways. They've taken the dataset of 3047 records. In their study they have used NN architecture called multilayer perceptron (MLP) with back propagation type supervised-learning algorithm to produce both classification and regression type prediction models and decision tree for achieving the highest possible prediction accuracy.

Lastly, in the study of Shahiria et.al. [11] they studied on predicting students' performance is mostly useful to help educators and learners improve their learning and teaching process. It reviewed previous studies on predicting students' performance with various data mining analytical methods. Most of the researchers have used cumulative grade point average (CGPA) and internal assessment as data sets. While prediction techniques are commonly used in the educational data mining area. Neural Network and Decision Tree are the two methods that were used by the researchers for predicting students' performance.

2. METHODOLOGY

This section describes the details of the dataset, pre-processing techniques, and machine learning algorithms employed in this study.

2.1 Dataset

Educational institutions regularly store all data that are available about students in the electronic medium using the Enrollment System. Data is stored in databases for processing. These data can be of many types and volumes, from students' demographics to their academic performance. In this study, the data were taken from the College of Criminal Justice Education (CCJE) students' evaluation, where all student records are stored. In these records, the final grades of 330 students who have taken all the professional courses were selected as the dataset.

2.2 Data Identification and Collection

At this phase, it is determined from which source the data will be stored, which features of the data will be used, and whether the collected data is suitable for the purpose. Feature selection involves decreasing the number of variables used to predict a particular outcome. The goals are to facilitate the interpretability of the model, reduce complexity, increase the computational efficiency of algorithms, and avoid overfitting.

The data is composed of criminology students who are able to have a pre-review exam. It has been collected and processed based on the results of their pre-review subject areas. It is composed of 6 subjects. The data is stored in excel sheets, composed of 33 students. As for this reason, the data of the students has been multiplied by 10 in order to come up with

the targeted number of data. No backlog student data is taken as the researcher wants to concentrate on students who had taken the pre-rev exam. In this study the researcher mainly concentrated on the different subjects and exams conducted by the college.

2.3 Establishing DM model and implementation of the algorithm

Multiple Linear Regression was employed to predict the relationship of the students' performance and the assessment course. Regression is a supervised machine learning technique that uses a training dataset to predict outcomes, which is similar to how classification works [5]. The output variable in classification is categorical, whereas the output variable in regression is numerical.

Multiple linear regression computes the t statistic of the overall model, the associated p value (how likely it is that the t statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true), and the regression coefficients that result in the smallest overall model error in order to determine the best-fit line for each independent variable [3].

The DM process serves two main purposes. The first purpose is to make predictions by analyzing the data in the database (predictive model). The second one is to describe behaviors (descriptive model). In predictive models, a model is created by using data with known results. Then, using this model, the result values are predicted for datasets whose results are unknown. In descriptive models, the patterns in the existing data are defined to make decisions.

3. EXPERIMENTS AND RESULTS

After determining the important variables and gathering the needed data, the researcher ran the experiment and the data mining study results were obtained using IBM SPSS as the Modeler to change the data and extract pertinent information that was then used for the conclusion.

3.1 Descriptive Statistics

The researcher chose the Assessment Result as the dependent variable and six (6) subjects such as Crime Detection and Investigation, Correctional Administration, Criminalistics, Law Enforcement and Administration, Criminal Jurisprudence and Procedure, and Criminal Sociology are the independent variables. The mean and standard deviation of each variable are illustrated in Figure 1. Moreover, there are 330 total number of students to be observed in the study.

	Mean	Std. Deviation	N
AssessmentResult	3.076	.5101	330
CrimeDetectionandInvestigationAverage	2.5379	.20889	330
CorrectionalAdministrationAverage	2.6339	.33134	330
CriminalisticsAverage	2.7097	.31794	330
LawEnforcementAdministrationAverage	2.3782	.27636	330
CriminalJurisprudenceandProcedureAverage	2.7494	.23556	330
CriminalSociologyAverage	2.5661	.27767	330

Figure 1. Descriptive Statistics

3.2 Model Summary

The information about the model's properties is provided in the Model Summary. In Figure 2, R-value was .712 which clearly shows that there is a correlation between the dependent and independent variables.

The total variation for the dependent variable that could be explained by the independent variables is displayed using the R-square value [7]. In this case, it shows .507 which means that the model is effective enough to determine the relationship.

Adjusted R-square resulted in .498 which is considered good and shows the generalization of the result.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.712[a]	.507	.498	.3615	.507	55.357	6	323	<.001

a. Predictors: (Constant), CriminalSociologyAverage, CorrectionalAdministrationAverage, LawEnforcementAdministrationAverage, CrimeDetectionandInvestigationAverage, CriminalisticsAverage, CriminalJurisprudenceandProcedureAverage

b. Dependent Variable: AssessmentResult

Figure 2. Model Summary[b]

3.3 ANOVA

Jain & Chetty [7] also mentioned that ANOVA table is used to assess whether the model is significant enough to predict the outcome. As stipulated in the Figure 3, P-value or Sig value resulted in 0.000, which is less than the normal probability of 0.05, therefore resulting in the result being statistically significant.

F-ratio, on the other hand, resulted in 55.357 which is considered a representation of the improvement in the prediction of the variable by fitting the model after considering the inaccuracy present in the model.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43.400	6	7.233	55.357	<.001[b]
	Residual	42.206	323	.131		
	Total	85.606	329			

a. Dependent Variable: AssessmentResult

b. Predictors: (Constant), CriminalSociologyAverage, CorrectionalAdministrationAverage, LawEnforcementAdministrationAverage, CrimeDetectionandInvestigationAverage, CriminalisticsAverage, CriminalJurisprudenceandProcedureAverage

Figure 3. ANOVA[a]

3. 4 Coefficient Table

The Coefficient Table shows the strength of the relationship i.e the significance of the variable in the model and the magnitude with which it impacts the dependent variable [1]. In this experiment, the significant value should be below the tolerable level of significance for the study below 0.05 for a 95% confidence interval in the study.

Figures 4 and 5 illustrate the coefficients of the dependent variable (Assessment Result) and independent variables (Crime Detection and Investigation, Correctional Administration, Criminalistics, Law Enforcement and Administration, Criminal Jurisprudence and Procedure, and Criminal Sociology).

As stipulated in Figure 5, it shows the significant relationship of each independent variable to the dependent variable. Correctional Administration, Criminalistics, Criminal Jurisprudence and Procedure, and Criminal Sociology has a significant value greater than 0.05 which simply means that it has no significant relationship to the dependent variable. Only the two (2) independent variables namely; Crime Detection and Investigation and Law Enforcement Administration got the <0.05 level of significance with the dependent variable. With this result, it can be interpreted that whenever there is an increase in these subject areas, there is also an increase to the Assessment Result. This proves that only the said variables have a significant relationship with the Assessment Result.

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t
1	(Constant)	-1.009	.286		-3.532
	CrimeDetectionandInvestigationAverage	1.089	.177	.446	6.156
	CorrectionalAdministrationAverage	-.187	.098	-.122	-1.911
	CriminalisticsAverage	-.142	.145	-.089	-.980
	LawEnforcementAdministrationAverage	1.004	.150	.544	6.700
	CriminalJurisprudenceandProcedureAverage	.181	.197	.084	.917
	CriminalSociologyAverage	-.267	.204	-.146	-1.313

a. Dependent Variable: AssessmentResult

Figure 4. Coefficients[a] (a)

Model		Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound	Correlations Zero-order	Correlations Partial	Correlations Part
1	(Constant)	<.001	-1.570	-.447			
	CrimeDetectionandInvestigationAverage	<.001	.741	1.437	.563	.324	.240
	CorrectionalAdministrationAverage	.057	-.380	.006	.278	-.106	-.075
	CriminalisticsAverage	.328	-.428	.143	.432	-.054	-.038
	LawEnforcementAdministrationAverage	<.001	.709	1.298	.655	.349	.262
	CriminalJurisprudenceandProcedureAverage	.360	-.207	.569	.590	.051	.036
	CriminalSociologyAverage	.190	-.668	.133	.535	-.073	-.051

Figure 5. Coefficients[a] (b)

3.5 Normal P-Plot of Regression Standardized Residual

The cumulative distribution function (CDF) of the standardized residual is observed and compared to the expected CDF of the normal distribution using a probability plot. Additionally, the researcher is testing the normality of the residuals and not the predictors.

Figure 6 shows the Normal Probability Plot of the Regression Standardized Residual. As can be seen, there are some deviations that happened, meaning there are some points that fall less to the trendline, but generally the points seem to follow the line and with that, it could assume that there is a normal distribution and that the observed standardized residuals are normally distributed.

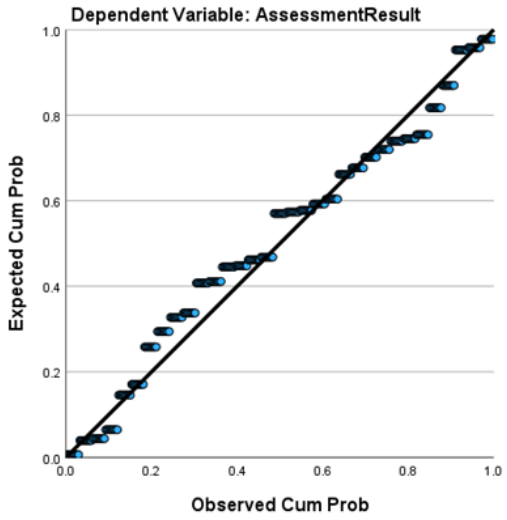

Figure 6. Normal P-Plot of Regression Standardized Residual

3.6 Scatter Plot

When using multiple linear regression, the researcher presumes that the correlation between the predictors and the response variable is linear. If this presumption is broken, the

linear regression will attempt to fit non-linear data with a straight line. This can be determined whether the relationships between the predictors and the outcome are linear using the bivariate plot of the predicted value against residuals (Tharu, 2019). Figure 6 displays the plot of the standardized residuals against the standardized projected value of assessment result and six subjects in a scatter plot.

In addition, as observed in the graph some of them are really bonds and others are far which we called outliers. A scatter plot with dots going from lower left to upper right indicates a positive correlation (as variable x goes up; variable y also goes up). A scatterplot of z scores also reveals the strength of the relationship between variables. If the dots in the scatterplot form a narrow band so that when a straight line is drawn through the band the dots will be near the line, there is a strong linear relationship between the variables.

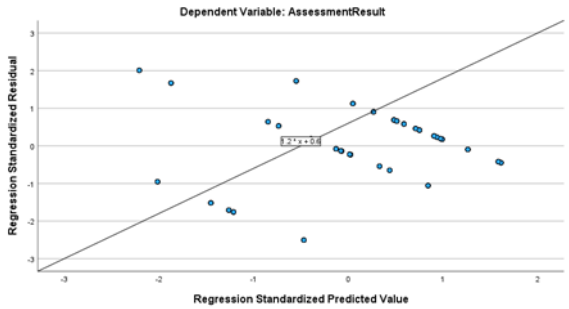

Figure 7. Scatter Plot

4. DISCUSSION AND CONCLUSION

The researcher's goal is to predict criminology students' assessment performance based on the average scores of their reviews in various subject areas. This is also to assist every educator in determining what subject/s could affect the performance of the students; by doing so, it will improve their learning and teaching technique, and it will give them an idea of what subject they need to focus on.

Based on the findings, the subjects Crime Detection and Investigation and Law Enforcement Administration have a significant relationship with the dependent variable, which is the students' assessment result. Therefore, other subject areas such as Correctional Administration, Criminalistics, Criminal Jurisprudence and Procedure, and Criminal Sociology are the subjects that educators should focus on reviewing with their students in order to improve their students' performance in assessments.

REFERENCES

[1] Abu Shanab, E., & Hammouri, Q. (2017). Exploring the factors influencing employees' satisfaction toward e-tax systems. *International Journal of Public Sector Performance Management*, *3*(2), 169. https://doi.org/10.1504/ijpspm.2017.10005371

[2] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17.

[3] Bevans, R. (2022, November 15). *Multiple Linear Regression | A Quick Guide (Examples)*. Scribbr. https://www.scribbr.com/statistics/multiple-linear-regression/

[4] Buldua, A. Üçgün,. K, (2010). Data mining application on students' data. Procedia Social and Behavioral Sciences 2 5251–5259. Retrieved from: https://doi.org/10.1016/j.sbspro.2010.03.855

[5] El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., Dakkak, A., & El Allioui, Y. (2020). A Multiple Linear Regression-Based Approach to Predict Student Performance. *Advances in Intelligent Systems and Computing*, 9–23. https://doi.org/10.1007/978-3-030-36653-7_2

[6] Strecht, Pedro, et al. (2015). "A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance." , International Educational Data Mining Society. Retrieved from: 2015EDM_FalakmasirYRK_final

[7] Jain, R., & Chetty, P. (2019). How to interpret the results of the linear regression test in SPSS?. From Project Guru: https://www.projectguru.in/interpret-results-linear-regression-test-spss/?fbclid=IwAR3SOGHCnCHwqiD1QHyCvdSmXglk_z6d_eNeW-lICitGb992Fx8aoh8ATvA

[8] Jayakumar, N. & Namdeo, N. (2014). Predicting Students' Performance Using Data Mining Technique with Rough Set Theory Concepts. Retrieved from: http://dx.doi.org/10.19044/esj.2021.v17n7p

[9] Sen, B. & Ucar, E. (2012). Evaluating the achievements of computer engineering department of distance education students with data mining methods. Procedia Technology 1 262 – 267. Retrieved from: https://doi.org/10.1016/j.protcy.2012.02.053

[10] Sense, F., Velde, M. V. D., & Rijn, H. V. (2021). *Predicting University Students' Exam Performance Using a Model-Based Adaptive Fact-Learning System* (pp. 155-169). Journal of Learning Analytics. https://doi.org/10.18608/jla.2021.6590

[11] Shahiria, A.M, Husaina, W. & Rashida, N.A.,, (2015). A Review on Predicting Student's Performance using Data Mining Techniques. Retrieved from: https://doi.org/10.1016/j.procs.2015.12.157

Tharu, R. P. (2019). Multiple regression model fitted for job satisfaction of employees working in saving and cooperative organization. *International Journal of Statistics and Applied Mathematics*, *4*(4), 43–49. https://www.mathsjournal.com/pdf/2019/vol4issue4/PartA/4-2-16-993.pdf

[12] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining practical machine learning tools and techniques (3rd ed.). Morgan Kaufmann. Retrieved from: https://doi.org/10.1016/C2009-0-19715-5