DOI: https://doi.org/10.15276/aait.04.2020.5
UDC 004.93

APPLICATION OF MACHINE LEARNING MODELS IN ENROLLMENT AND STUDENT TRAINING AT VIETNAMESE UNIVERSITIES

Kim Thanh Tran[1)]
ORCID: https://orcid.org/0000-0002-4241-1065, tkthanh2011@gmail.com
The Vinh Tran[2)]
ORCID: https://orcid.org/0000-0002-4241-1065, tranthevinh@opu.ua
Manh Tuong Tran[1)]
ORCID: https://orcid.org/0000-0002-0495-9388, tmtuong@ufm.edu.vn
Anh Linh Duy Vu[1)]
ORCID: https://orcid.org/0000-0003-2432-0319, vuduy@ufm.edu.vn

[1)] University of Finance-Marketing, 2/4 Tran Xuan Soan St. District 7. Ho Chi Minh City, Vietnam
[2)] Odessa National Polytechnic University, 1, Shevchenko Ave. Odessa, 65044, Ukraine

ABSTRACT

In Vietnam, since 2015, the Ministry of Education and Training of Vietnam has decided to abolish university entrance exams and advocates the use of high school graduation exam results of candidates for admission to go to universities. The 2015 and 2016 exam questions for the Math exam are the essay questions. From 2017 up to now, the Ministry of Education and Training of Vietnam has applied the form of multiple-choice exams for Mathematics in the high school graduation exam. There are many mixed opinions about the impact of this form of examination and admission on the quality of university students. In particular, the switch from the form of essay examination to multiple-choice exams led the entire Vietnam Mathematical Association at that time to send recommendations on continuing to maintain the form of essay examination for mathematics. The purposes of this article are analysis and evaluation the effects of relevant factors on the academic performance of advanced math students of university students, and offer solutions to optimize university entrance exam. The data set was provided by Training Management Department and Training Quality Control and Testing Laboratory of the University of Finance – Marketing. This dataset includes information about math high school graduation test scores, learning process scores (scores assessed by direct instructors), and advanced math course end test scores of 2834 students in courses from 2015 to 2019. Linear and non-linear regression machine learning models were used to solve the tasks given in this article. An analysis of the data was conducted to reveal the advantages and disadvantages of the change in university enrollment of the Vietnamese Ministry of Education and Training. Tools from the Python libraries have been supported and used effectively in the process of solving problems. Through building and surveying the model, there are suggestions and solutions to problems in enrollment and input quality assurance. Specifically, in the preparation of entrance exams, the entrance exam questions should not exceed 61-66 % of multiple choice questions.

Keywords: Cross-Sectional Data; Essay Exam; Test Exam; Linear Regression; Non-Linear Regression; Least Squares Regression; Support Vector Regression

For citation: **Tran Kim Thanh, Tran The Vinh, Tran Manh Tuong, Vu Anh Linh Duy. Application of Machine Learning Models in Enrollment and Student Training at Vietnamese Universities. *Applied Aspects of Information Technology*. 2020; Vol.3 No.4: 276–287. DOI: https://doi.org/10.15276/aait.04.2020.5**

INTRODUCTION

From 2015 Ministry of Education and Training of Vietnam removed the university entrance exam and used the results of the high school exit exam to enter universities in Vietnam. There are many conflicting opinions about the impact of this decision on the quality of university students.

Therefore, it becomes necessary to study the problems based on survey data and data on the performance of school graduates and students. When investigating at different times on a large population. In that case, at each survey time, a random sample of survey subject's data of the survey times.

This data is called Independently Pooled cross-section data over time, simply called Pooled cross-section data [1, 2]. Thus, Pooled Cross-section data is a very common data type in surveys. For independently Pooled Cross – Section data, the subjects surveyed (cross units) at different times may differ, and the number of these objects is not fixed over time. From a statistical point of view, independently pooled cross-section data have the important feature that they are constituted by independently sampled observations. This important feature is an advantage for cross-data analysis, as it eliminates correlation in noise error. However, an independently pooled cross data differs from a cross data (a single random sample) in that sampling from the population at different points in time likely leads to observations that are not identically distributed. For example, the variable "income" or variable "education" have distribution that changes over time in most countries.

Data in this article is data set, which is provided by Training Management Department and Training Quality Control and Testing Laboratory of the University of Finance – Marketing, which includes the following information of 2834 students of the courses since 2015 to 2019: the math high school graduation test scores, the Advanced Maths learning process scores (scores assessed by direct instructors), and the Advanced Maths final exam scores. This dataset is an Independently Pooled Cross-section data set with 5 observation time units of 5 years. In Vietnam, Admission to universities in the period 2015 – 2019 is based on the results of high school graduation exam scores, in which Mathematics in this period has changed the form of exam: The 2015 and 2016 essay exam period, which we simply call the essay examination period, the period from 2017 to 2019 on multiple-choice exams, which we call the multiple-choice test period. So, the data is divided into two samples: the essay test data sample and the multiple-choice test data sample. Both of these samples are independent pooled cross-section data samples.

Regression Analysis [3, 4], [5]. Regression is a data mining technique used to predict a range of numeric values given a particular dataset. Regressions measure the relationship between a dependent variable (what you want to measure) and an independent variable (the data you use to predict the dependent variable).

Regression analysis is used to solve the following types of problems:

– Determine which independent variable is associated with the dependent.

– Understand the relationship between the dependent and independent variables.

– Predict the unknown values of the dependent variable.

Regression is used across multiple industries for business and marketing planning, financial forecasting, environmental modeling and analysis of trends. For example, in the field of education, department of education analyst examines the effectiveness of a new school feeding program. The analyst constructs a regression model for performance indicators using independent variables such as class size, household income, student funding per capita, and school feeding. The model equation is used to identify the relative contribution of each variable to school performance.

Linear and nonlinear regression analyzes are detailed in the Literature review below.

Over the years Python has several powerful and popular libraries that are designed to work with data mining: analysis, visualization, trend forecasting [29, 30], [31]. For example, the Matplotlib library is one of the most popular data visualization libraries. The Pandas library is used to analyze information. Scikit-learn library provides simple and efficient tools for predictive data analysis.

LITERATURE REVIEW

Ordinary least squares (OLS) regression [6, 7], [8, 9], [10, 11], [12] is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable; the method estimates the relationship by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line. OLS regression will be discussed in the context of a bivariate model, that is, a model in which there is only one independent variable (X) predicting a dependent variable (Y).

OLS regression is one of the major techniques used to analysis data and forms the basis of many other techniques [7]. The usefulness of the technique can be greatly extended with the use of dummy variable coding to include grouped explanatory variable [8], for a discussion of the analysis of experimental designs using regression) and data transformation methods [9]. OLS regression is particularly powerful as it relatively easy to also check the model assumption such as linearity, constant variance and the effect of outliers using simple graphical methods [10].

Simple linear regression: general problem formulation [3], [13, 14], [15, 16].

Suppose we have k predictor variables $x_1, \dots, x_k$ and a dependent variable y.

We consider the simple linear relation (where the hat on top of vector y symbolizes that this is a vector of predicted y values:

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}.$$

The parameters $\beta_0, \beta_1, \dots, \beta_k$ of this equation are called regression coefficients. In particular, β_0 is called the regression intercept and $\beta_1, \dots, \beta_k$ are regression slope coefficients.

Based on the predictions of a parameter vector ($\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$) we consider the residual sum of squares as a measure of prediction error:

$$RSS_{(\beta_0,\beta_1,\dots,\beta_k)} = \sum_{i=1}^{k} [\, y_i - \hat{y}_i(\beta_0, \beta_1, \dots, \beta_k)]^2,$$

We would like to find the best parameter values (denoted traditionally by a hat on the parameter's variable: $\hat{\beta}_i$,) in the sense of minimizing the residual sum of squares:

$$\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\right) = \arg\min_{(\beta_0,\beta_1,\ldots,\beta_k)} RSS_{(\beta_0,\beta_1,\ldots,\beta_k)}$$

In statistics, **the mean squared error (MSE)** [17, 18] – the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss.The MSE is a measure of the quality of an estimator, it is always non-negative, and values closer to zero are better

$$MSE = \frac{1}{k}\sum_{i=1}^{k}[\, y_i - \hat{y}_i]^2.$$

Coefficient of determination R^2 [5], [19, 20] is the square of the sample correlation coefficient, written as a percent. When evaluating the goodness-of-fit of simulated (Y_{pred}) vs. measured (Y_{obs}) values, it is not appropriate to base this on the R^2 of the linear regression (i.e., $Y_{\text{obs}} = m \cdot Y_{\text{pred}} +$ b). The R^2 quantifies the degree of any linear correlation between Y_{obs} and Y_{pred}, while for the goodness-of-fit evaluation only one specific linear correlation should be taken into consideration: $Y_{\text{obs}} = 1 \cdot Y_{\text{pred}} + 0$.

So, its value is between 0 % and 100 %. A value of 0 % means that there is no linear dependence between the sample values of X and Y. while a value of 100 % means there is a perfect linear dependence. Clearly, the larger the value of R2, the more confidence we can have that there really is a linear relationship between X and Y.

Support Vector Machines (SVM) are very specific class of algorithms, characterized by usage of kernels, absence of local minima, sparseness of the solution and capacity control obtained by acting on the margin, or on number of support vectors, etc. [5], [21, 22], [23, 24], [25, 26]. Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). Still, it contains all the main features that characterize maximum margin algorithm: a non-linear function is leaned by linear learning machine mapping into high dimensional kernel induced feature space. The capacity of the system is controlled by parameters that do not depend on the dimensionality of feature space.

The SVM regression algorithm (Support Vector Regression or SVR) [23, 24], [25], [28] is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

Advantages of Support Vector Regression are as mentioned below:

– It is robust to outliers.

– Decision model can be easily updated.

– It has excellent generalization capability, with high prediction accuracy.

– Its implementation is easy.

Some hyperparameters in SVR are as below:

– Hyperplane. Hyperplanes are decision boundaries that is used to predict the continuous output. The data points on either side of the hyperplane that are closest to the hyperplane are called Support Vectors. These are used to plot the required line that shows the predicted output of the algorithm.

– Kernel [27]. A kernel is a set of mathematical functions that takes data as input and transform it into the required form. These are generally used for finding a hyperplane in the higher dimensional space. The most widely used kernels include Linear, Non-Linear, Polynomial, Radial Basis Function (RBF) and Sigmoid. By default, RBF is used as the kernel. Each of these kernels are used depending on the dataset.

– Boundary lines. These are the two lines that are drawn around the hyperplane at a distance of ε. It is used to create a margin between the data points.

Support Vector Regression performs linear regression in the high-dimension feature space using ε-insensitive loss and, at the same time, tries to reduce model complexity by minimizing $\|\omega\|^2$. This can be described by introducing (non-negative) slack variables $\xi_i, \xi_i^*, i = 1, \ldots, n$, to measure the deviation of training samples outside ε-insensitive zone.

Thus, SVR is formulated as minimization of the following functional:

$$\min\left(\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}|\xi_i|\right),$$

$$\begin{cases} y_i - f(x_i, \omega) \le \varepsilon + \xi_i^* \\ f(x_i, \omega) - y_i - \le \varepsilon + \xi_i. \\ \xi_i, \xi_i^* \ge 0, i = 1, \ldots, n \end{cases}$$

This optimization problem can transform into the dual problem and its solution is given by

$$f(x) = \sum_{i=1}^{n_{sv}}(\alpha_i - \alpha_i^*)K(x_i, x),$$

where:

$$0 \le \alpha_i^* \le C$$
$$0 \le \alpha_i \le C$$

n_{sv} is – the number of Support Vectors (SVs);

K – the kernel function

$$K(x, x_i) = \sum_{j=1}^{m} g_j(x) g_j(x_i).$$

Parameter C determines the tradeoff between the model complexity (flatness) and the degree to which deviations larger than ε are tolerated in

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

optimization formulation for example, if C is too large (infinity), then the objective is to minimize the empirical risk only, without regard to model complexity part in the optimization formulation.

Parameter ε controls the width of the

ε – insensitive zone, used to fit the training data.

The value of ε can affect the number of support vectors used to construct the regression function. The bigger ε, the fewer support vectors are selected. On the other hand, bigger ε -values results in more "flat" estimates. Hence, both C and ε – values affect model complexity (but in a different way).

Linear SVR

$$y = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b.$$

Non-linear SVR [26]. The kernel functions transform the data into a higher dimensional feature space to make it possible to perform the linear separation.

Kernel functions [27].

– Polynomial

$k(x_i, x_j) = (x_i, x_j)^d$.

– Gaussian Radial Basic function

$k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$.

Example of 1D regression [32] using linear, polynomial and RBF kernels is shown in Fig. 1.

PURPOSE AND TASKS OF WORK

Since 2015 up to now, in Vietnam, the Ministry of Education and Training of Vietnam has removed the university entrance exam and used the results of the high school graduation exam to be admitted to universities. For the years 2015, 2016 the Math exam of this exam is the essay. Since 2017 until now, the Ministry of Education and Training of Vietnam has applied the form of multiple-choice exams in Mathematics in the high school graduation exam. There are many mixed opinions about the impact of this form of examination and admission on the quality of university students.

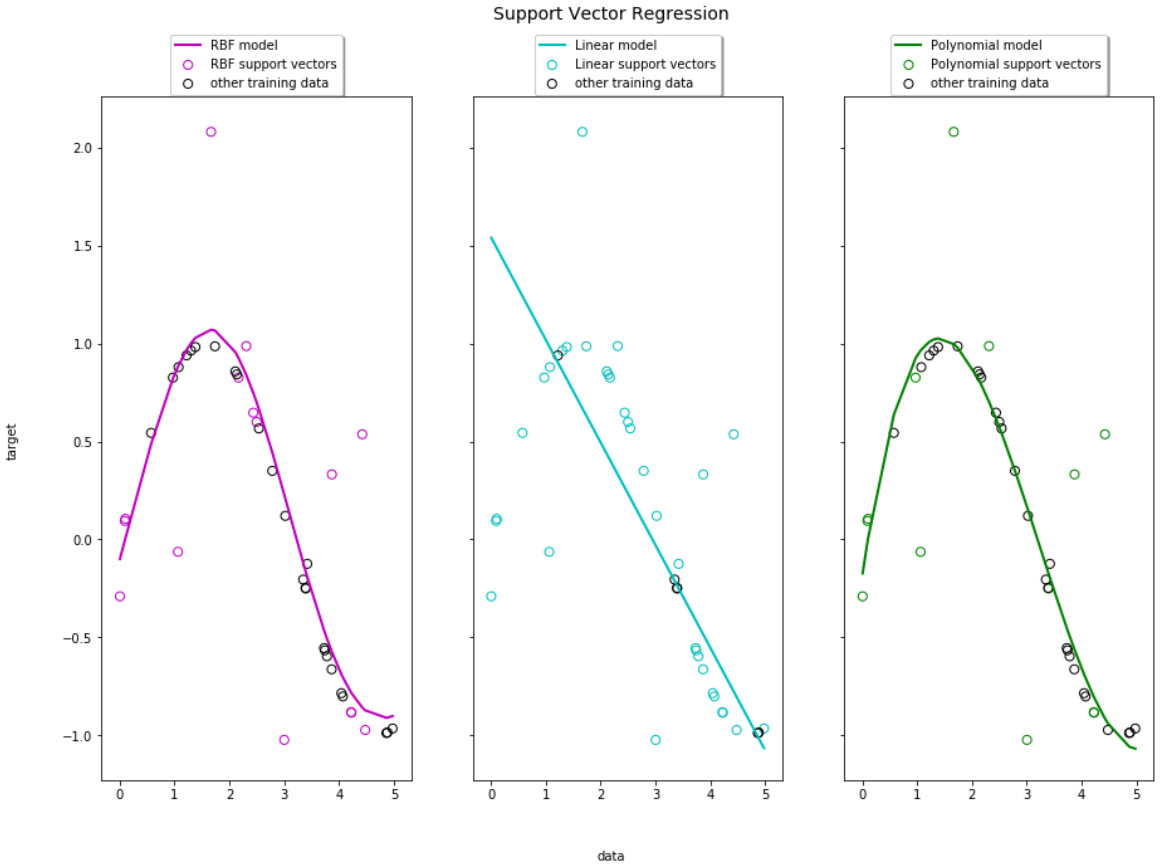


Fig. 1. **Example of 1D regression using linear, polynomial and RBF kernels [32]**
***Source*: [32]**

Based on the reviews of linear regression, SVR, and given Python library features, to survey problems from enrollment data and student training quality, as well as to propose solutions to solve problems, tasks have been introduced including:

– Data analysis. We choose this data set to apply the Independent Pooled Cross-section data analysis model to analysis and assess the impact of the basic factors such as: input score (score high school graduation exam in Math), process scores (attendance and learning attitude along course), the students' final exam scores in Advanced Math.

Apply linear regression models to analyze data in both university entrance examination periods from 2015 to 2019.

– Using Epsilon Support Vector Regression model to solve the problems, which were given during the survey in the data analysis section, and propose optimal solutions for university entrance exam.

DATA ANALYSIS

Data set is provided by Training Management Department and Training Quality Control and Testing Laboratory of the University of Finance – Marketing, which includes the following information of 2834 students of the courses since 2015 to 2019: the math high school graduation test scores, the Advanced Maths learning process scores (scores assessed by direct instructors), and the Advanced Maths final exam scores. The 2015 and 2016 essay exam. period, which we simply call the essay examination period (EE period), the period from 2017 to 2019 on multiple-choice exams, which we call the multiple-choice test period (MCT period).

Let X1, X2, Y denote respectively the entry score in Mathematics, the process scores in Advanced Math and the final exam score of a student's Advanced Math course. For the sake of simplicity, X_1 is called the input score, Y is called the output score.

Based on the available data, to serve the analysis and evaluation, the entry scores are classed as the following:

– Group A is a group with an output score from 7 to 10;

– Group B is the group with an output score from 4.5 to 6.9;

– Group C is group of output points below 4.5.

To analyze the data, we used the students' entry scores, progress scores, and math study scores for the students grouped in both exam preriods.

The process of data analysis and evaluation includes:

– general data analysis, group analysis by conventional statistical methods. The obtained results are shown in Table 1 and Table 2

– using a linear regression model [29] to evaluate the influence of input variables X1, X2 on the output variable Y. Two linear regression models are set up corresponding to 2 data pairs (X1, Y) and (X2, Y). The parameters and performance of the linear regression models are shown in Table 3.

– representation of scatter plots of data with graphs of a linear regression model. Use the chart (Fig. 2 and Fig. 3) to see trend as well as the anomaly occurring in the statistical data.

Table 1. **General data analysis in both exam preriods**

Medium scores of	In EE period	In MCT period
Input scores	6.71	7.19
Process scores	7.63	7.7
Output score	5.85	5.78

***Source*: compiled by the author**

Table 2. **Data analysis in both exam preriods by divided groups**

Statistics / Group	In EE period				In MCT period			
	Input	Process	Output	%	Input	Process	Output	%
Group A	6.9	8.93	7.9	23.5	7.46	8.68	8.11	25.9
Group B	6.7	7.65	5.64	63	7.16	7.87	5.63	53.5
Group C	6.45	5.3	3.27	13.5	6.92	6.01	3.21	20.6

***Source*: compiled by the author**

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

Table 3. **The parameters and performance of the linear regression models [29]**

Periods \ Parameters	Model 1 for Input-ouput scores				Model 2 for process-ouput scores			
	a	b	ε (MSE)	R^2	a	b	ε (MSE)	R^2
On EE period	0.57	2.04	2.43	0.05	0.52	1.89	1.49	0.42
On MCT periods	0.59	1.57	3.34	0.06	0.6	1.19	2.24	0.37

***Source*: [29]**

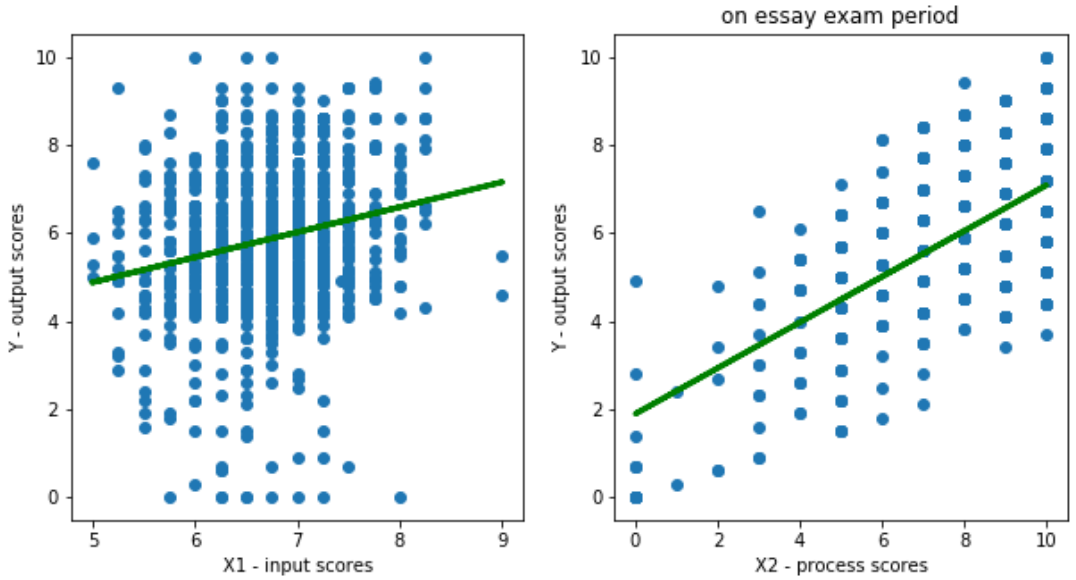


Fig. 2. **Representation of graphs of a linear regression models (green lines) with scatter plots (blue points) of data in EE period**
***Source*: compiled by the author**

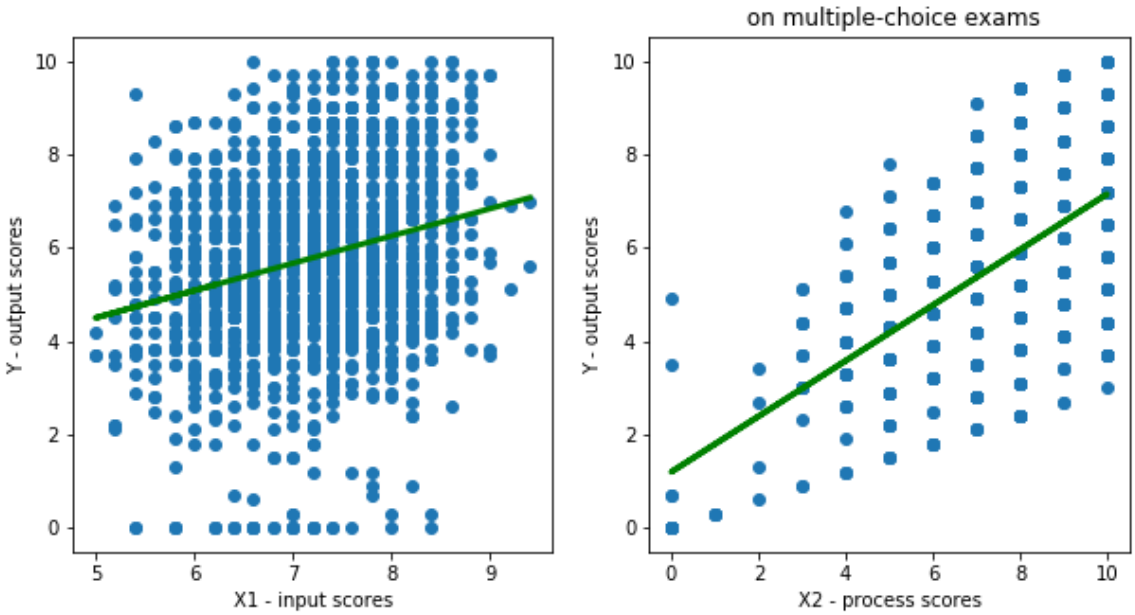


Fig. 3. **Representation of graphs of a linear regression models (green lines) with scatter plots (blue points) of data in MCT period**
***Source*: compiled by the author**

From the obtained results (Table 1, Table 2 and Table 3; Fig. 1 and Fig. 2) we have clearly seen that:

– The average of the math input scores for the EE period (6.71) was lower than the MCT period (7.19), average of their math output scores higher (5.85), and less difference with average of input score than on MCT period (Table 1).

– In group C (Table 2), the percentage of students with a failing semester math score in EE period (13.5 %) was much lower than in MCT test (20.6 %), although in the MCT period the math input score is higher (6.92), than in the EE period (6.45).

– The lines obtained from the linear regression models in the both periods (Fig. 1 and Fig. 2) showed the increasing trend of Y (outpput scores) according to X1 (input scores) and according to X2 (process scores).This means that in general, students with higher entry math scores or higher progress scores will most likely have higher Advanced Math scores. The uptrend of Y according to X1 and the uptrend of Y according to X2 is approximately the same.

– From the scatter chart of the distribution from the original data (Fig. 1 and Fig. 2) showed that there are quite a few students who have an entry score of more than 7 points but only achieve the final exam results of Advanced Math below 4.5 points. Specifically, in the MCT period the number of students with such status accounts for 42.1 % of group C and in EE period is 12.32 % of group C). This showed that there are still risks in the admissions problem, especially the multiple choice test.

– The linear regression model built on data pair (X2, Y) model 2 gives better performance with MSE value for the both periods of 1.49 and 2.24 respectively; coefficient of determination – 0.42 and 0.37. This means that data pair (X2; Y) more accurately assess student quality. The model 2 can be used to predict student's output score.

Thus, the analysis of the above data has clearly shown that this form of essay examination has ensured better student input, and less quality problem risks. However, in reality, the organization of examination in the form of multiple choice test brings great advantages in terms of organization such as budget, time, and manpower. So a test that includes both multiple choice and essay format will guarantee the admission advantages of both.

In this paper, we have created both-stage scrambled datasets according to different ratios. Build machine learning models and apply their to survey and evaluate the results according to the criteria: giving the highest test rate possible, but the quality of students is still relative to the built model. From there, giving a reasonable rate for the test and the essay in a university entrance exam.

USING EPSILON-SUPPORT VECTOR REGRESSION (ESVR) MODEL TO SOLVE THE PROBLEM

Building dataset for learning process

To solve the problem given above. The first step to take is to set up the data set for machine learning. A data set will be generated from statistical data of EE period and MCT period, including input scores, progress scores, output-scores of students. The multiple-choice ratio on the entrance exam will determine the data mixing rate for MCT period in the new data. New data sets were generated from data mixed at a ratio of 1 % to 99 % of MCT period's data. New data are categorized into groups A, B, and C based on semester exam scores in math.

This new data set will be surveyed using the developed ESVR model (will be discussed in the next section) to find the data with the appropriate mixing ratio that yields the highest performance.

Establishment of ESVR models and conducting surveys

The implemented ESVR model, which used scikit-learn Python [30], is based on below parameters:

–Specifies the kernel type to be used in the algorithm kernel='rbf'.

– Degree (=3) of the polynomial kernel function.

– Kernel coefficient for 'rbf' gamma=0.5.

– Regularization parameter C=10.

– Epsilon in the epsilon-SVR model $\varepsilon = 0.5$.

This ESVR model is trained from sample data, which have been classified by groups A, B, C, whose input is process score X2 and output Y-math semester exam score. Examples are shown in Fig. 4.

The performance survey of this ESVR model was performed with using the input data (the entrance test scores X1) from the new data set, which were according to subgroups A, B, C (Fig. 5). From there we find out the highest possible rate that can be mixed into the college entrance exam questions, while ensuring the highest model performance. The obtained results are shown in Table 4.

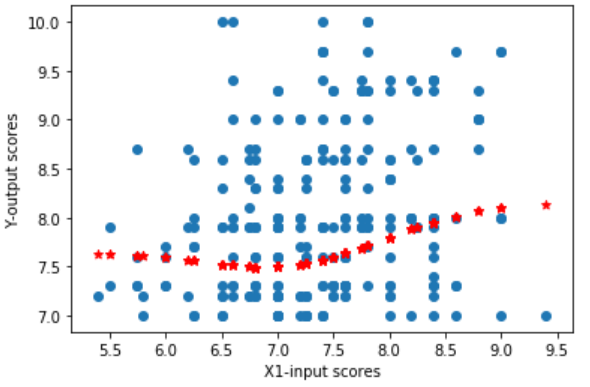
From the obtained results in the Table 3, we see that the performance of the generated ESVR model is highest with data with a mixing ratio of 61%-66 % of the multiple choice test. From there we can conclude that according to the built model, to ensure the quality of students, the highest test rate can be put on the university entrance exam from 61 % to 66%.

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

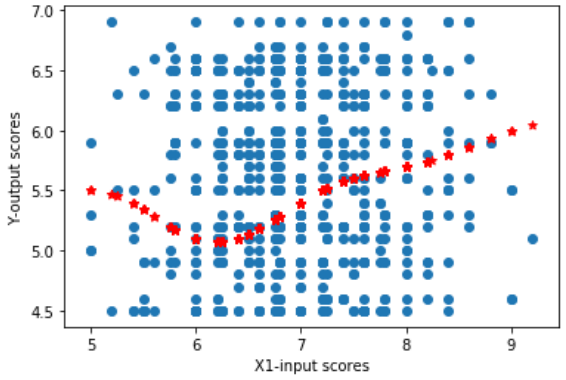
Table 4. **Result of the survey performance of developed ESVR model with new mixed datasets**

	Group A	Group B	Group C
Mean (MSE)	0.5772	0.5354	1.7486
Max (MSE)	0.5935	0.5418	1.7938
Min (MSE)	0.5633	0.5308	1.6956
Obtained ratio of MCT (%)	66	61	65

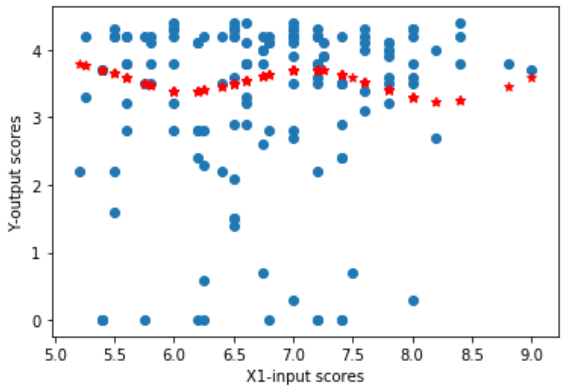
***Source*: compiled by the author**



a

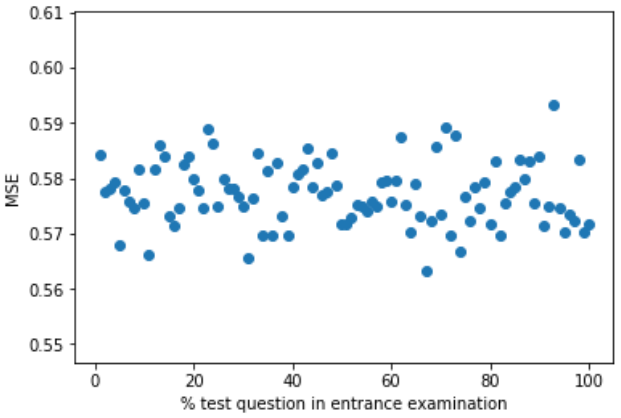


b

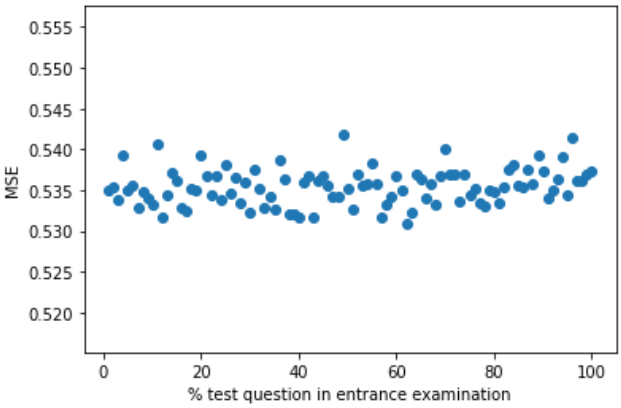


c

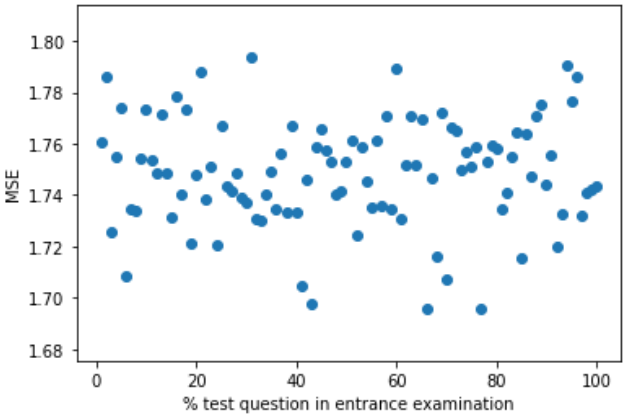
Fig. 4. **Representation of graphs of ESVR models (red star points) with scatter plots (blue points) of data of:**
a – group A; b – group B; c – group C
***Source*: compiled by the author**



a



b



c

Fig. 5. **Performance surveying of developed ESVR model with new mixed datasets:**
a – group A; b – group B; c – group C
***Source*: compiled by the author**

CONCLUSION

In this article, the data set was provided by Training Management Department and Training Quality Control and Testing Laboratory of the University of Finance - Marketing. This dataset includes information about math high school graduation test scores, learning process scores

(scores assessed by direct instructors), and advanced math course end test scores of 2834 students in courses from 2015 to 2019. By mathematical statistical method with 4using NumPy, Pandas library Python, initial data analysis was performed (Table 2 and Table 3). The article discusses the features of building linear (OLS) and non-linear regression (ESVR) machine learning models were used to solve the tasks. For this, the capabilities of matplotlib, scikit-learn libraries Python are used (Fig. 1; Fig. 2; Fig. 3 and Fig. 4). Performance surveying of developed OLS models and ESVR models with processed data were tested (Table 4 and Table 5). The paper has analyzed the advantages and disadvantages of both forms of university enrollment (multiple-choice test and essay) from the obtained results of the data analysis. From the built-in ESVR model, the model performance was investigated on the new data set, which generated from the original data set. The ratio of mixing data from two periods (EE period and MCT period) to get the highest modeling performance was found. From there we have a solution for the entrance exam questions. From this result, in the article, it is possible to propose enrollment options to ensure the learning quality of students while ensuring the factors of saving budget, time and human resources. That is, in the university entrance examination can be used both multiple-choice test and an essay forms with the highest rate of multiple-choice test questions from 61 % to 66 %.

REFERENCES

1. Wooldridge, M. Jeffrey. “Introductory Econometrics: A Modern Approach”. *South-Western Centage Learning.* [5th Edition]. Mason Ohio: USA. 2013. 878 p.

2. Wooldridge, M. Jeffrey. “Econometric Analysis of Cross Section and Panel Data”. *The MIT Press.* [2nd Edition]. Cambridge Massachusetts. USA. 2010. 1096 p.

3. Ryan, T. P. “Modern Regression Methods”. *Wiley-Interscience.* [2nd Edition]. Hoboken New Jersey: USA. 2018. 672 p.

4. Cherkassky, V. & Mulier, F. “Learning from Data: Concepts, Theory, and Methods”. *Wiley-IEEE Press.* [2nd Edition]. Hoboken New Jersey: USA. 2007. 560 p.

5. Draper, N. R. & Smith, H. “Applied Regression Analysis”. *Wiley-Interscience.* [3rd Edition]. Hoboken New Jersey: USA. 1998. 736p.

6. Hutcheson, G. D. “Ordinary Least-Squares Regression”. *The SAGE Dictionary of Quantitative Management Research. SAGE Publications.* Thousand Oaks California: USA. 2011. p.224–228. DOI: https://doi.org/10.4135/9781446251119.n67.

7. Rutherford, A. “Introducing ANOVA and ANCOVA: a GLM Approach”. *John Wiley & Sons, Inc.* [2st Edition]. Chichester West Sussex. England. 2011. 360 p. DOI: https://doi.org/10.1002/9781118491683.

8. Hutcheson, G. D. & Moutinho, L. “Statistical Modeling for Management”. *Sage Publications. Online Publication*. December 27. 2012. DOI: https://doi.org/10.4135/9781446220566.

9. Fox, J. “An R Companion to Applied Regression”. *Sage Publications.* Inc. [2st Edition]. Thousand Oaks California: USA. 2011. 449 p.

10. Hutcheson, G. D. “The Multivariate Social Scientist”. *Sage Publications*. Thousand Oaks California: USA. 1999. 288p. DOI: https://doi.org/10.4135/9780857028075.

11. Agresti, A. “An Introduction to Categorical Data Analysis”. *Wiley Series in Probability and Statistics. Wiley-Interscience.* [3rd Edition]. *Hoboken*. New Jersey: USA. 2018. 400 p.

12. Koteswara, R. K. “Testing for the Independence of Regression Disturbances*”. Journal Econometrica.* 1970; Vol. 38 Issue 1: 97–117. DOI: https://doi.org/ 10.2307/1909244.

13. Bremer, M. “Multiple Linear Regression”. MATH 261a, San Jose State University. USA. 2012. Available from: http://mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement%205%20%20 multiple%20regression.pdf.

14. Gonzalez, P. & Orbe, S. “The Multiple Regression Model: Estimation”. *Dpt. Applied Economics III (Econometrics and Statistics).* University of the Basque Country. Spain. 2014. Available from: https://www.coursehero.com/file/46666912/multiple-regression.pdf.

15. Kirchner, James W. “Data Analysis Toolkit 10: Simple Linear Regression Derivation of Linear Regression Equations”. University of California. Berkeley: USA. September 2001. Available from: http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf.

16. Olive, David. “Linear Regression”. *Springer International Publishing.* Cham: Switzerland. 2017. 494 p. DOI: https://doi.org/10.1007/978-3-319-55252-1.

17. Math Vault. " List of Probability and Statistics Symbols". Montreal. Canada: – Available from: https://mathvault.ca/hub/higher-math/math-symbols/probability-statistics-symbols/. – [Accessed: June, 2020].

18. Pishro-Nik, H. "Mean Squared Error (MSE)". The Department of Electrical and Computer Engineering University of Massachusetts Amherst, USA: – Available from: https://www.probabilitycourse.com/chapter9/9_1_5_mean_squared_error_MSE.php. – [Accessed: June, 2020].

19. Devore, Jay L. "Probability and Statistics for Engineering and the Sciences". *Cengage Learning.* [8th Edition]. Massachusetts United States. Boston. 2011. 768 p.

20. Barten, Anton P. "The Coeffecient of Determination for Regression without a Constant Term". In book – The Practice of Econometrics. Dordrecht. *Martinus Nijhoff Publishers.* Leiden: Belgium. 1987. p.181–189. DOI: https://doi.org/10.1007/978-94-009-3591-4_12.

21. Jingjing, Zhang. "Model Selection in SVMs using Differential Evolution". *Journal IFAC Proceedings.* 2011; Vol.44 Issue 1: 14717–14722. DOI: https://doi.org/10.3182/20110828-6-IT-1002.00584.

22. Prince Grover. "5 Regression Loss Functions All Machine Learners Should Know". – Available from: https://heartbeat.fritz.ai/5-regression-loss-functions-all-machine-learners-should-know-4fb140e9d4b0. *Tittle from the screen*. – [Accessed: June, 2018].

23. Salcedo, Sanz S., et. al. "Support Vector Machines in Engineering: an Overview". *Wires Data Mining and Knowledge Discovery.* 2014; Vol. 4 Issue 3: 161–267. DOI: https://doi.org/10.1002/widm.1125.

24. Pai, P. F. & Hsu, M. F. "An Enhanced Support Vector Machines Model for Classification and Rule Generation". *Journal Computational Optimization, Methods and Algorithms.* 2011; Vol. 356: 241–258. DOI: https://doi.org/10.1007/978-3-642-20859-1_11.

25. Smola, A. & Schölkopf, B. "A Tutorial on Support Vector Regression". *Journal Statistics and Computing.* 2004; Vol.14: 199–222. DOI: https://doi.org/10.1023/B:STCO.0000035301.49549.88.

26. Yoshioka, T. & Ishii, S. "Fast Gaussian Process Regression Using Representative Data". *International Joint Conference on Neural Networks*. 2001; Vol.1: 132–137. DOI: https://doi.org/10.1109/IJCNN.2001.939005.

27. Chiroma, H., Abdulkareem, S., Abubakar, A. I., Herawan, T., et. al. "Kernel Functions for the Support Vector Machine: Comparing Performances on Crude Oil Price Data". *Recent Advances on Soft Computing and Data Mining.* Advances in Intelligent Systems and Computing book series. *Publ. Springer. Cham.* 2014; Vol. 287: 271–281. DOI: https://doi.org/10.1007/978-3-319-07692-8_26.

28. The MathWorks, Inc. "Understanding Support Vector Machine Regression". – Available from: https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html. – [Accessed: August, 2020].

29. Custer, Charlie. "15 Python Libraries for Data Science You Should Know". Available from: https://www.dataquest.io/blog/15-python-libraries-for-data-science/. Date February 5, 2020.

30. Scikit-learn developers. "Linear Models". – Available from: https://scikit-learn.org/stable/modules/linear_model.html. – [Accessed: May, 2020].

31. Scikit-learn developers. "Epsilon-Support Vector Regression". – Available from: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html. – [Accessed: May, 2020].

32. Scikit-learn developers. "Support Vector Regression (SVR) Using Linear and non-linear kernels". – Available from: https://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression. – [Accessed: May, 2020].

Conflicts of Interest: the authors declare no conflict of interest

Received 02.10.2020
Received after revision 15.11.2020
Accepted 20.11.2020

DOI: https://doi.org/10.15276/aait.04.2020.5
УДК 004.93

ЗАСТОСУВАННЯ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ В РЕЄСТРАЦІЇ ТА НАВЧАННІ СТУДЕНТІВ В'ЄТНАМСЬКИХ УНІВЕРСИТЕТІВ

Кім Тхань Чан[1)]
ORCID: https://orcid.org/0000-0002-4241-1065, tkthanh2011@gmail.com
Тхе Вінь Чан[2)]
ORCID: https://orcid.org/0000-0002-4241-1065, tranthevinh@opu.ua
Чан Мань Тионг[1)]
ORCID: https://orcid.org/0000-0002-0495-9388, tmtuong@ufm.edu.vn
Ву Ань Лінь Діу[1)]
ORCID: https://orcid.org/0000-0003-2432-0319, vuduy@ufm.edu.vn
[1)] Університет фінансів-маркетингу, Хошимін, вулиця Чанхуа Шоан 2/4, Район 7. Хошимін, В'єтнам
[2)] Одеський національний політехнічний університет, пр-т Шевченка,1. Одеса, 65044, Україна

АНОТАЦІЯ

У В'єтнамі з 2015 року Міністерство освіти і професійної підготовки В'єтнаму вирішило скасувати вступні іспити до університетів і виступає за використання результатів випускних іспитів у середній школі кандидатів для вступу до університетів. Форма іспиту з математики 2015 і 2016 років – письмовий іспит. З 2017 року і по теперішній час Міністерство освіти і професійної підготовки В'єтнаму застосовує форму тестових іспитів з математики на випускних іспитах у середній школі. Існує багато суперечливих думок про вплив цієї форми іспитів і прийому на якість студентів університетів. Зокрема, перехід від форми письменнвого іспиту до тестового іспиту спонукав всю В'єтнамську математичну асоціацію в той час направити рекомендації щодо збереження форми письменнвого іспиту з математики. Метою даної статті є аналіз та оцінка впливу відповідних факторів на академічну успішність студентів-математиків і студентів університетів, а також пропозиції рішень для оптимізації вступних іспитів в університет. Набір даних надано Департаментом управління навчанням і Лабораторією контролю якості навчання і тестування Фінансового університету - Маркетинг. Цей набір даних включає в себе інформацію про результати випускних іспитів середньої школи з математики, балах процесу навчання (бали, які оцінюються безпосередніми викладачами) і оцінках 2834 учнів в кінці курсу математики на курсах з 2015 по 2019 рік. Моделі машинного навчання з лінійної і нелінійної регресією використані для вирішення поставлених в статті завдань. Був проведений аналіз даних, щоб виявити переваги та недоліки змін в прийомі студентів до університетів Міністерства освіти і навчання В'єтнаму. Інструменти з бібліотек Python були підтримані і ефективно використовувалися в процесі вирішення проблем. Побудова і вивчення моделі дозволяють отримати пропозиції і рішення проблем при зарахуванні і забезпеченні якості вхідних даних. Зокрема, при підготовці до вступних іспитів кількість питань вступного іспиту не повинно перевищувати 61-66 % тестових запитань.

Ключові слова: перехресні дані; письмовий іспит; тестовий іспит; лінійна регресія; нелінійна регресія; метод найменших квадратів; метод опорних векторів

DOI: https://doi.org/10.15276/aait.04.2020.5
УДК 004.93

ПРИМЕНЕНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ В ЗАЧИСЛЕНИИ И ОБУЧЕНИИ СТУДЕНТОВ В ВЬЕТНАМСКИХ УНИВЕРСИТЕТАХ

Ким Тхань Чан[1)]
ORCID: https://orcid.org/ 0000-0002-4241-1065, tkthanh2011@gmail.com
Тхе Винь Чан[2)]
ORCID: https://orcid.org/0000-0002-4241-1065, tranthevinh@opu.ua
Мань Тыонг Чан[1)]
ORCID: https://orcid.org/0000-0002-0495-9388, tmtuong@ufm.edu.vn
Ань Линь Диу Ву[1)]
ORCID: https://orcid.org/0000-0003-2432-0319, vuduy@ufm.edu.vn
[1)] Университет финансов и маркетинга, улица Чанхуан Шоан 2/4, Район 7. Хошимин, Вьетнам
[2)] Одесский национальный политехнический университет, проспект Шевченко 1. Одесса, 65044, Украина

АННОТАЦИЯ

Во Вьетнаме с 2015 года Министерство образования и профессиональной подготовки Вьетнама решило отменить вступительные экзамены в университеты и выступает за использование результатов выпускных экзаменов в средней школе

кандидатов для поступления в университеты. Форма экзамена по математике 2015 и 2016 годов – письменный экзамен. С 2017 года и по настоящее время Министерство образования и профессиональной подготовки Вьетнама применяет форму тестовых экзаменов по математике на выпускных экзаменах в средней школе. Существует много противоречивых мнений о влиянии этой формы экзаменов и приема на качество студентов университетов. В частности, переход от формы письменнвого экзамена к тестовому экзамену побудил всю Вьетнамскую математическую ассоциацию в то время направить рекомендации относительно сохранения формы письменнвого экзамена по математике. Целью данной статьи является анализ и оценка влияния соответствующих факторов на академическую успеваемость студентов-математиков и студентов университетов, а также предложения решений для оптимизации вступительных экзаменов в университет. Набор данных предоставлен Департаментом управления обучением и Лабораторией контроля качества обучения и тестирования Финансового университета – Маркетинг. Этот набор данных включает в себя информацию о результатах выпускных экзаменов средней школы по математике, баллах процесса обучения (баллы, оцениваемые непосредственными преподавателями) и оценках 2834 учащихся в конце курса математики на курсах с 2015 по 2019 год. Модели машинного обучения с линейной и нелинейной регрессией использованы для решения поставленных в статье задач. Был проведен анализ данных, чтобы выявить преимущества и недостатки изменений в приеме студентов в университеты Министерства образования и обучения Вьетнама. Инструменты из библиотек Python были поддержаны и эффективно использовались в процессе решения проблем. Построение и изучение модели позволяют получить предложения и решения проблем при зачислении и обеспечении качества входных данных. В частности, при подготовке к вступительным экзаменам количество вопросов вступительного экзамена не должно превышать 61-66 % тестовых вопросов..

Ключевые слова: перекрестные данные; письменный экзамен; тестовый экзамен; линейная регрессия; нелинейная регрессия; метод наименьших квадратов; метод опорных векторов

ABOUT THE AUTHORS

Tran Kim Thanh – Doctor of Philosophy (2002), Senior Lecturer University of Finance-Marketing. University of Finance-Marketing, 2/4 Tran Xuan Soan St. District 7. Ho Chi Minh City, Vietnam
ORCID: https://orcid.org/ 0000-0002-4241-1065, tkthanh2011@gmail.com
Research field: Statistical Probability; Data Analysis

Чан Кім Тхань – доктор філософії, старший викладач. Університет фінансів-маркетингу, вулиця Чанхуа Шоан 2/4, Район 7. Хошимін, В'єтнам

The Vinh Tran – Doctor of Philosophy (2016), Senior Lecturer of Department of Information Systems. Center of Ukrainian-Vietnamese Cooperation, Odessa National Polytechnic University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
tranthevinh@opu.ua.ORCID: https://orcid.org/0000-0002-4241-1065
***Research field*:** Data Analysis; Artificial Intelligence Methods and Systems

Тхе Вінь Чан – доктор філософії, старший викладач кафедри Інформаційних систем.
Центр українсько-в'єтнамського співроебітництва. Одеський національний політехнічний університет, пр-т Шевченка,1. Одеса, 65044, Україна

Tran Manh Tuong – Master of Math, Senior Lecturer University of Finance-Marketing, University of Finance-Marketing, 2/4 Tran Xuan Soan St. District 7. Ho Chi Minh City, Vietnam
tmtuong@ufm.edu.vn. ORCID: https://orcid.org/0000-0002-0495-9388
Research field: Statistical Probability; Data Analysis

Чан Мань Тионг – магістр математики, старший викладач Університет фінансів і маркетингу, вулиця Чанхуа Шоан 2/4, Район 7. Хошимін, В'єтнам

Vu Anh Linh Duy – Master of Math, Senior Lecturer University of Finance-Marketing, Odessa National Polytechnic University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
vuduy@ufm.edu.vn. ORCID: https://orcid.org/0000-0003-2432-0319
Research field: Statistical Probability; Data Analysis

Ву Ань Лінь Діу – магістр, старший викладач Університету фінансів і маркетингу. Одеський національний політехнічний університет, пр-т Шевченка,1. Одеса, 65044, Україна