Mathematical model to estimate the probability to approve given the first grade at the Escuela Politécnica Nacional

Tarquino Sánchez, MBA.[1], Iván Sandoval, MSc.[1], Diego Salazar, Math.[1], Ximena Sánchez, Eng.[1], and Vanessa Guevara, Eng.[1]
[1]*Escuela Politécnica Nacional, Ecuador, tarquino.sanchez@epn.edu.ec, ivan.sandoval@epn.edu.ec, diego.salazar@epn.edu.ec, ximena.sanchez@epn.edu.ec, vanessa.guevarav@epn.edu.ec*

***Abstract*– Ecuadorian universities must undergo a rigorous process of internal and external evaluation regarding the quality of education by the Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad del Sistema de Educación Superior (CEAACES). The institutional evaluation model for Higher Education Institutions from Ecuador seeks to establish a standard procedure to have a continuous monitoring of the approval probability of the students in each class. These activities allow academic units to improve their quality management. A logistic regression model was developed using 5 years of historical data, a brief discussion about the pertinence about main variables was also elaborated. A simulation and its results are finally shown, interpretation of the probability results and the estimated number of future student were found.**

The interpretation of the probability results and the estimated number of future student is presented, and a simulation and its results are finally shown.

Keywords—Capacity of students, education, accreditation, passing rate, quality of education.

I. Introduction

Ecuadorian universities must undergo a rigorous process of internal and external evaluation regarding the quality of education, in order to obtain the Certificate of Social Accreditation by the Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad del Sistema de Educación Superior (CEAACES).

The institutional evaluation model for Higher Education Institutions (HEI) addresses six criteria which support for the processes of teaching, research and linkage activities. These criteria are the parameters for the evaluation, and they are: Organization, Academia, Research, Linkage with Society, Resources and Infrastructure, and Students. The organization criterion considers the processes of institutional organization that allows the institution to establish, monitor and evaluate the achievement of institutional objectives, considering principles of Quality Management. The academia criterion evaluates the quality of the teaching staff and the working conditions in which their activities are carried out. The research criterion evaluates the projects and research activities of HEI. The linkage with society criterion evaluates the transfer of knowledge to meet the needs and solve environmental problems in order to generate development. The infrastructure criterion evaluates the characteristics of the physical infrastructure and the information technologies of the HEI. The students criterion considers the policies and actions undertaken by HEI to guarantee the adequate conditions for the academic success of students, as well as the results measured in terms of academic efficiency [1].

Because of the criteria of organization, resources and infrastructure, and students, the HEIs from Ecuador need to develop and apply institutional approval procedures, periodic monitoring and management of its academic offer. These activities allow academic units to improve their quality management. With this purpose the HEIs from Ecuador seek to establish a standard procedure to have a continuous monitoring of the passing rate of students in each class. Thereby, the academic units can optimize the planning both in teachers and infrastructure.

The present investigation describes a model that allows projecting the number of students who pass for class according to the first grade corresponding to the first partial grade in different careers and the Leveling Course (LC) at the Escuela Politécnica Nacional (EPN). In this way, the institution can improve the quality management principles.

II. Methodology

A. Documentary analysis of the accreditation at Escuela Politécnica Nacional

In June 2014, after the CEAACES accreditation process, EPN obtained a 5-year Category "A" accreditation diploma. Following this, the institution generated the Institutional Improvement Plan (PMI) which was applied and concluded in December 2015. The results report showed the inexistence of associated policies or regulations with quality management. Therefore, the EPN is seeking to establish strategies that can guarantee the consolidation of the university quality management system. These strategies could contribute to make appropriate decisions and promote the continuous improvement of the processes, activities and results in all the academic units of the institution [2].

Digital Object Identifier (DOI): http://dx.doi.org/10.18687/LACCEI2018.1.1.284
ISBN: 978-0-9993443-1-6
ISSN: 2414-6390

B. Documentary analysis of the evaluation system of the EPN

At the EPN, each class has two grades corresponding to the results obtained through continuous assessment events. The first grade in the middle of the semester (first partial) and the second one at the end of it (second partial). Students who reach 14 points or more in the sum of the two grades will be exempted of the final exam and will pass the class. The approval grade will be equal to said sum multiplied by two and this is about 40 points. Students who fail to reach 14 points, but have at least 9 points in the sum of the two grades, must take a final exam on 20 points, to complete a minimum of 24 points and pass the class [3].

C. Leveling Course of EPN

The EPN has two types of Leveling Course: Engineering, Sciences and Administrative Sciences (CNIC, in its Spanish acronym), and Higher Technological Level (CNTS, in its Spanish acronym). These courses aim to level the basic knowledge of high school graduates in order to continue with their chosen careers. The EPN Leveling Course offers five classes:

- Physics
- Geometry
- Math
- Chemistry
- Literature

In this model, the EPN Leveling Course has been considered as a study case. The LC students do not have an academic record and their admission process does not depend on the institution. For this reason, it is critical to determine the number of students that pass to have a better planning, management and monitoring of academic processes focused on quality. Existing data from the 2012-B period to 2017-A of LC student academic performance were analyzed.

D. Correlation Analysis and Selection of variables

The procedure of the proposed method began with a statistical analysis. With this, the probability distributions of final grades in the different classes were obtained. We can see that in Physics, Math, Geometry, and Chemistry the distribution densities are concentrated between 9 and 12 and around 24 (Fig.1 to Fig.4). The number of scores between 9 and 12 is larger than 24, showing that most of the students had to take the final exam. In literature class the situation is reversed; the scores on the two first exams are higher, and most students are not required to take the final exam (Fig.5).

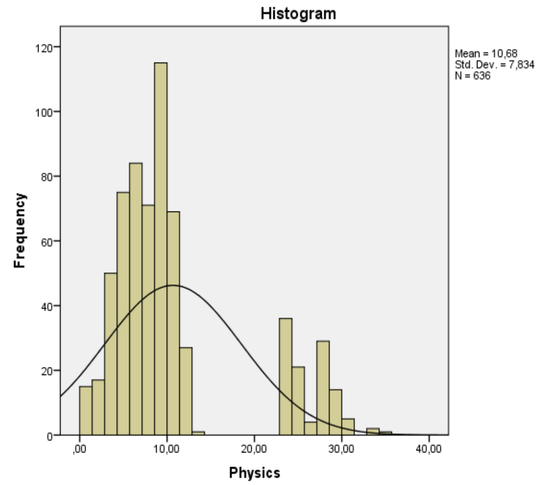

Fig 1. Physics Distribution Densities

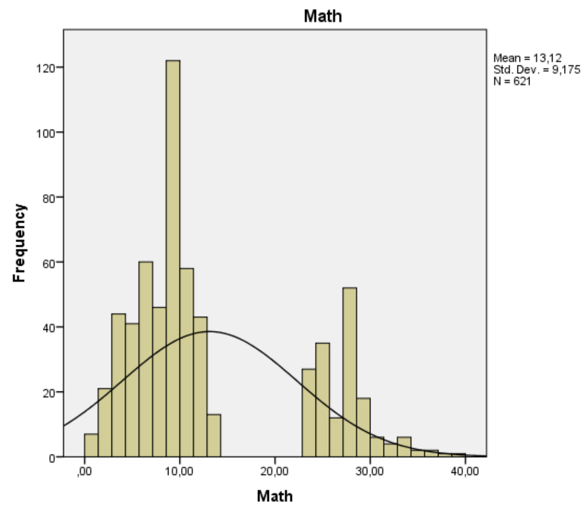

Fig 2. Math Distribution Densities

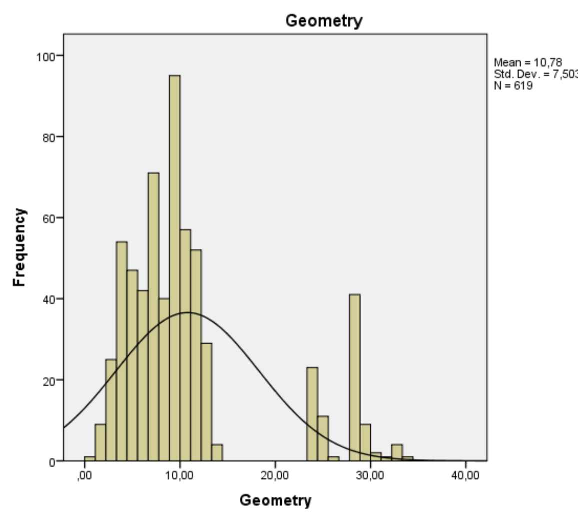

Fig 3. Geometry Distribution Densities

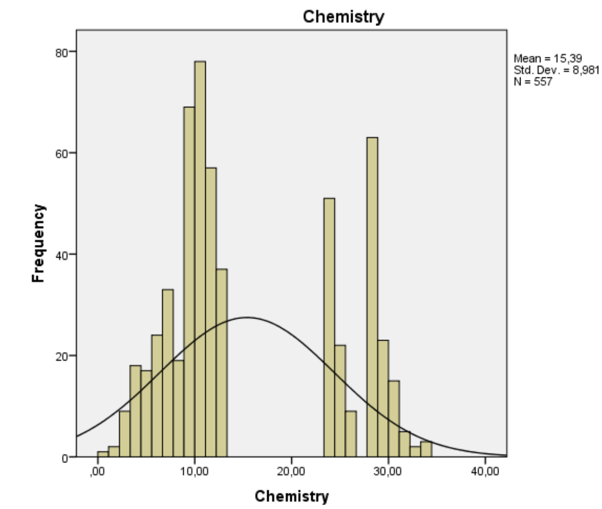

Fig 4. Chemistry Distribution Densities

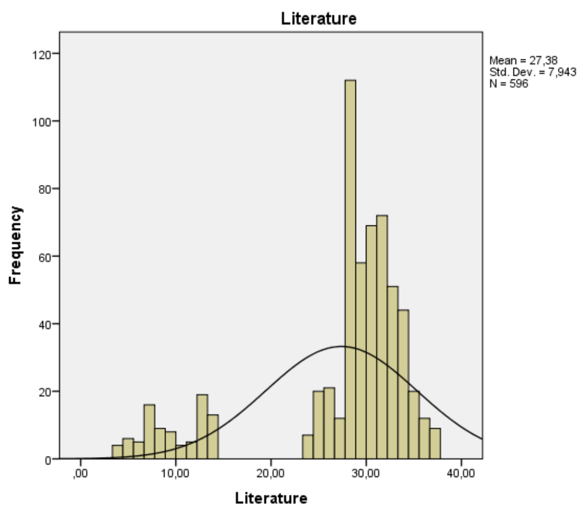

Fig 5. Literature Distribution Densities

When the ENES and class grade averages were analyzed, we can see a certain relationship between them. With the exception of chemistry, usually in each group a high ENES grade average involves a high class grade, as it indicates (Fig. 6). So, in Table I all these values are shown.

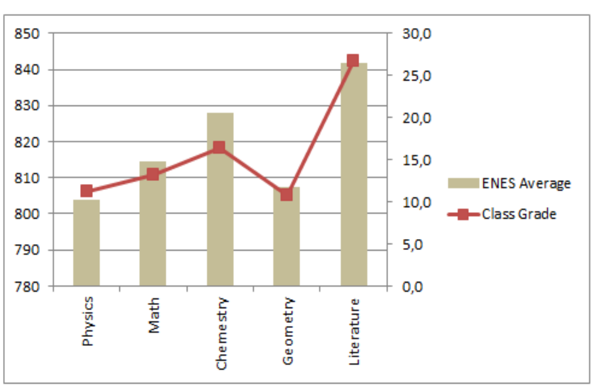

Fig 6. ENES and class grade averages

TABLE I
ENES AND LEVELING COURSE GRADES BY CLASS

Class	ENES Average	Class Grade
Physics	804	11,2
Math	815	13,2
Chemistry	828	16,3
Geometry	807	10,8
Literature	842	26,8

It should be noted that, the population taken in account for the compute of such averages are the total populations, i.e. students who had approved or not each class. All the population were considered because the ENES grade takes into account all and every single student who is registered at EPN and it is relevant for our purposes to determine the effectiveness level of such an exam. After this, the correlation between both averages was determined, like Table II shows.

TABLE II
CORRELATIONS ENES-GRADES BY CLASS

		ENES	Class
ENES	Pearson Correlation Sig. (2-tailed) N	1 5	,955* ,011 5
Class	Pearson Correlation Sig. (2-tailed) N	,955* ,011 5	1 5

* Correlation is significant at the 0.05 level (2-tailed).

In principle, we identified a strong, linear Pearson correlation between the ENES grade average and the average grade of each class. Hence, we can also make a quite simple linear relationship, given by:

$$CG = 0.4022 \cdot EG - 313.82 \qquad (1)$$

Where CG is the Class Grade and EG is the ENES Grade. Also, the distinct correlations class to class were analyzed. There is a positive and strong correlation between these two variables.

Now, we can see how the different strength correlations are related between them, as shown in Table III to VII. the 2-tailed correlation was chosen because it was not known a priori the direction in which such a correlation would be.

TABLE III
CORRELATIONS ENES-PHYSICS GRADE

		Physics	ENES_Physics
Physics	Pearson Correlation Sig. (2-tailed) N	1 636	,310** ,000 636
ENES_Physics	Pearson Correlation Sig. (2-tailed) N	,310** ,000 636	1 636

** Correlation is significant at the 0.01 level (2-tailed).

TABLE IV
CORRELATIONS ENES-MATH GRADE

		Math	ENES_Math
Math	Pearson Correlation Sig. (2-tailed) N	1 621	,323** ,000 621
ENES_Math	Pearson Correlation Sig. (2-tailed) N	,323** ,000 621	1 621

** Correlation is significant at the 0.01 level (2-tailed).

TABLE V
CORRELATIONS ENES-GEOMETRY GRADE

		Geometry	ENES_Geometry
Geometry	Pearson Correlation Sig. (2-tailed) N	1 619	,007 ,867 618
ENES	Pearson Correlation Sig. (2-tailed) N	,007 ,867 618	1 618

TABLE VI
CORRELATIONS ENES-CHEMISTRY GRADE

		Chemistry	ENES_Chemistry
Chemistry	Pearson Correlation Sig. (2-tailed) N	1 557	,267** ,000 557
ENES	Pearson Correlation Sig. (2-tailed) N	,367** ,000 557	1 557

** Correlation is significant at the 0.01 level (2-tailed).

TABLE VII
CORRELATIONS ENES-LITERATURE GRADE

		Literature	ENES_Literature
Literature	Pearson Correlation Sig. (2-tailed) N	1 596	,092** ,025 596
ENES_Literature	Pearson Correlation Sig. (2-tailed) N	,092** ,025 596	1 596

**Correlation is significant at the 0.05 level (2-tailed).

This time, correlations are regardless of the average, and there is a weaker strength correlation. So, even in average we can detect a correlation between both grades. If it is analyzed in more detail, students individually in each student group, the correlation is not as significant as it seems it was. This paper considers this kind of analysis very significant due to many times average tends to hide some relations in the data.

Considering the last point, we decided to avoid using the ENES grade in our analysis. However, this does not imply that ENES grade is useless, rather we will consider it irrelevant for the probability estimation.

Finally, ENES grade apart, a correlation matrix between all the classes grades was developed. For this occasion, we also considered all the student's universe, independently if they approved or not. The result for this analysis is shown in (Fig.7) and the unique statistical significant correlation that we can point out it is between Geometry and Math (marked in the figure).

Correlations

		Physics	Math	Geometry	Chemistry	Literature
Physics	Pearson Correlation	1	,029	,068	-,054	,073
	Sig. (2-tailed)		,466	,089	,202	,077
Math	Pearson Correlation	,029	1	0,415	-,025	,043
	Sig. (2-tailed)	,466		0,000**	,549	,293
Geometry	Pearson Correlation	,068	0,415	1	,027	,027
	Sig. (2-tailed)	,089	0,000**		,532	,517
Chemistry	Pearson Correlation	-,054	-,025	,027	1	-,012
	Sig. (2-tailed)	,202	,549	,532		,770
Literature	Pearson Correlation	,073	,043	,027	-,012	1
	Sig. (2-tailed)	,077	,293	,517	,770	

**. Correlation is significant at the 0.01 level (2-tailed).

Fig 7. Correlation Analysis

E. Probability estimation to approve

Like we pointed out before, to pass the course means that each student ought to get at least 24/40 in the total grade to pass. This grade is really the final one, and it's obtained to multiply the raw grade by 2. We summarize the variables as follows:

- GRADE 1B: (G1/10) is the first grade above 10 corresponding to the first partial.

- Grade 2B: (G2/10) is the second grade above 10 corresponding to the second partial.
- Final exam: (FE/20) if G1+G2 < 14, FE is mandatory.
  FE is above 20 and the minimal grade is 12/20. The needed grade (NG) is computed with: NG=24-G1-G2
- Final Grade: (FG) is above 40.
  FG=2(G1+G2) if you do not need FE and if you need it FG=G1+G2+FE
- Status: (S)
  E= Passed without FE.
  A= Passed with FE.
  R= Withdrawn.
  F=Failed.

One of the main purposes of this study is to try to catch the uncertainty, which occurs along each semester with the EPN students, through its historical data, in order to achieve some probabilities.

One among lots of techniques to achieve such probabilities, like we can see in [5], that is mainly an approximation of the probability distribution.

The main probability we want to model, is if a student with a certain first grade (G1) will approve or not, i.e. what is the probability to pass the class, in terms of formulas:

$$\boldsymbol{P}\left(\boldsymbol{x}=\frac{\boldsymbol{P}}{\boldsymbol{x}}=\boldsymbol{G1}\right)=\frac{\boldsymbol{P}(\boldsymbol{x}=\boldsymbol{P}\cap\boldsymbol{x}=\boldsymbol{G1})}{\boldsymbol{P}(\boldsymbol{x}=\boldsymbol{G1})} \qquad (2)$$

It was considered, in order to modelize the Bayes formula, the probability of intersection, instead of estimating the different distributions of these two probabilities. It was decided to estimate a regression with our historical data. It was used [2] and [3]. A logistic regression was the most adequate way to model it, as as seen below.

First, the model is developed describing the distribution in each class. We can verify that a Binomial distribution arises due to there are two possible states: Pass (E, A) or Fail (F) in each class. Then, the first step to the regression is done.

Historically, not all the classes have the same difficulty, as we can see in each class distribution figure (Fig.8 to Fig.12).

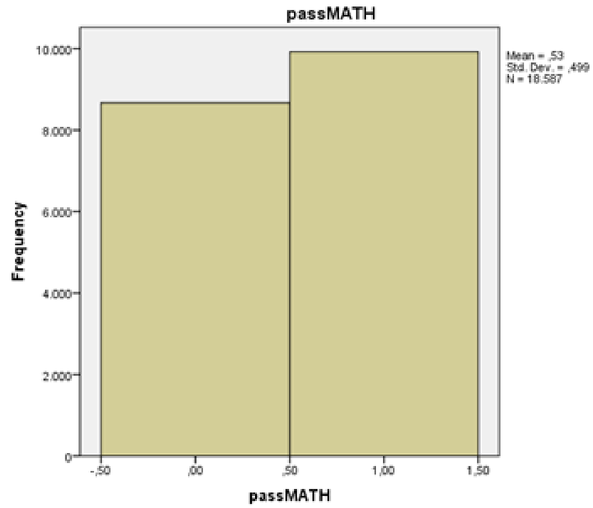

Fig 8. Math binomial distribution

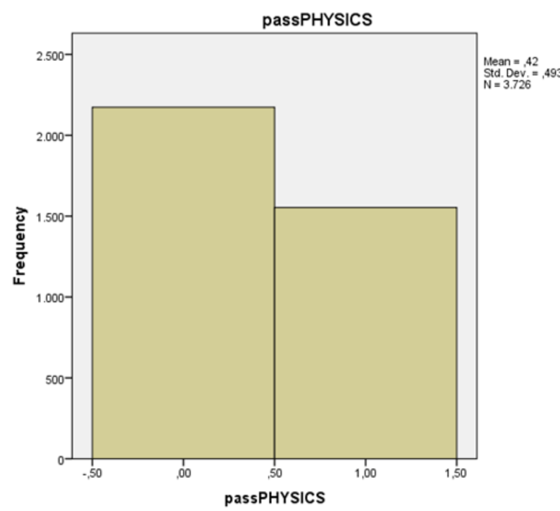

Fig 9. Physics binomial distribution

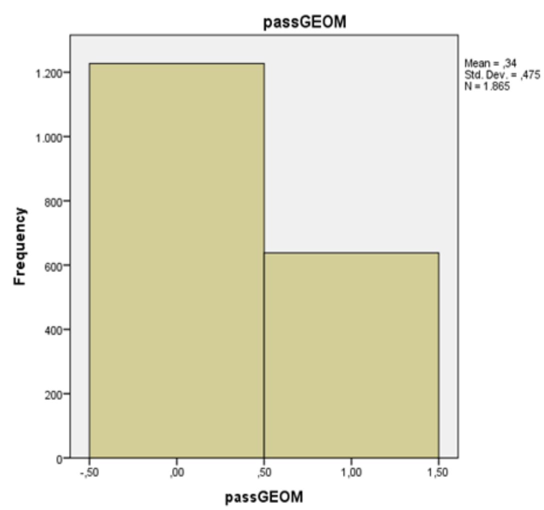

Fig 10. Geometry binomial distribution

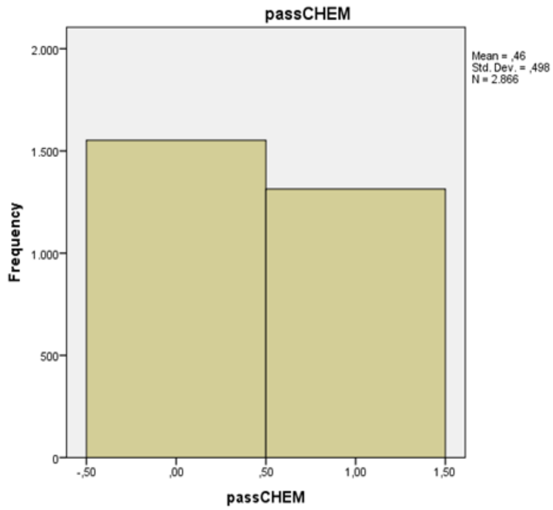

Fig 11. Chemestry binomial distribution

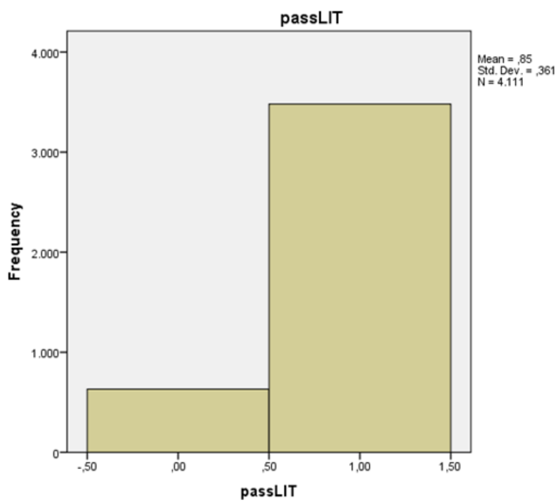

Fig 12. Literature's binomial distribution

Again, it is really interesting the student's behavior in Literature: it is considerable the amount of success cases in regard to the failed cases. Also, in Math classes the amount of approved students is higher than the failed students, but in all the cases we can assume a Binomial distribution.

Thus, we can assume that:

$$\boldsymbol{G1 \sim B(n,p)} \quad (3)$$

with n∈N and p∈[0,1].

After, one way to develop a probability function to estimate such probability is using in each class a Binomial logistic regression, i.e.:

$$\boldsymbol{P(G_1) = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot G_1)}}} \quad (4)$$

In the Fig.13, the shape of each of these functions is shown. The criteria used to have choose such functions are the need of smooth functions with a range between [0,1] and at the same time with an appropriate slope in the middle of its domain.

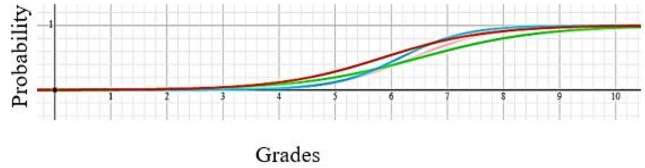

Fig 13. Probability functions for the logistic regression

As we can see in the probability functions were developed with the aim to have the enough slope between 5-8, which are the most critical grades to approve or not each class. Logistic regressions were used because it treats the same set of problems as the probit regression, with the latter using a cumulative normal distribution instead. We also are assuming a standard logistic distribution of errors like we can see in [8]. It was decided to use such technique instead of another generalized linear model because the conditional distribution is a Bernoulli distribution due to the dependent variable is binary.

About the coefficients estimation, the usual techniques of maximum probability were used. The model reach the convergence, such an hypothesis is possible according to [8]. Also, the logistic model showed an accuracy near to the 85%, i.e. 85 of each 100 students were correctly classified as approved or failed. Next, the coefficients for each class in the Table VIII.

TABLE VIII
REGRESSION'S COEFFICIENTS

Class	Beta_0	Beta_1
Math	-7,796	1,296
Physics	-9,025	1,414
Geometry	-6,071	0,933
Chemistry	-10,454	1,698
Literature	-6,356	1,088

III. RESULTS

We simulated different scenarios for each one of the classes, using different first grades, as seen in the Fig.14 to Fig.16:

Simulated Grade
1,5

CLASS	Probability to approve
Math	0,29%
Physics	0,10%
Geometry	0,93%
Chemistry	0,04%
Literature	0,88%

Fig 14. Simulation with grade 1.5

Simulated Grade
3

CLASS	Probability to approve
Math	1,97%
Physics	0,83%
Geometry	3,65%
Chemistry	0,47%
Literature	4,34%

Fig 15. Simulation with grade 3

Simulated Grade
4,5

CLASS	Probability to approve
Math	12,30%
Physics	6,53%
Geometry	13,32%
Chemistry	5,66%
Literature	18,84%

Fig 16. Simulation with grade 4.5

In these three scenarios, the probability is low for grades below to 5, like the function shows us (Fig.17 to Fig.18):

Simulated Grade
5,5

CLASS	Probability to approve
Math	33,88%
Physics	22,32%
Geometry	28,10%
Chemistry	24,68%
Literature	40,80%

Fig 17. Simulation with grade 5.5

Simulated Grade
7

CLASS	Probability to approve
Math	78,16%
Physics	70,56%
Geometry	61,30%
Chemistry	80,71%
Literature	77,90%

Fig 18. Simulation with grade 7

With the next two grades, it is seen a fast improvement of the probability to approve, with the grade 7. Next, with a 9.5 grade the probability is near to the maximum (Fig. 19). Obviously, even with a grade of 10 the probability will not necessarily be 1, which means that even with the maximum grade there are people who did not pass.

Simulated Grade
9,5

CLASS	Probabilidad de aprobación
Math	98,92%
Physics	98,80%
Geometry	94,23%
Chemistry	99,66%
Literature	98,17%

Fig 19. Simulation with grade 9.5

A. *Simulation*

We tested our mathematical model with the students of the 2017-B academic period (Fig 20 to Fig. 24). With different number of population, there is a different distribution; almost all classes are the same with the exception of Literature.

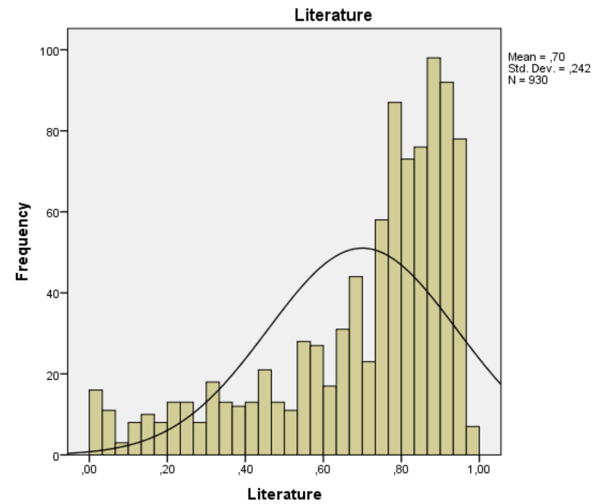

Fig 20. Literature Simulation (2017B)

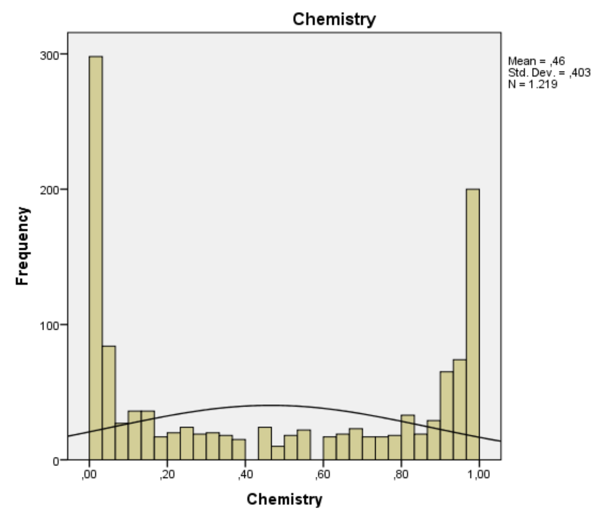

Fig 21. Chemistry Simulation (2017B)

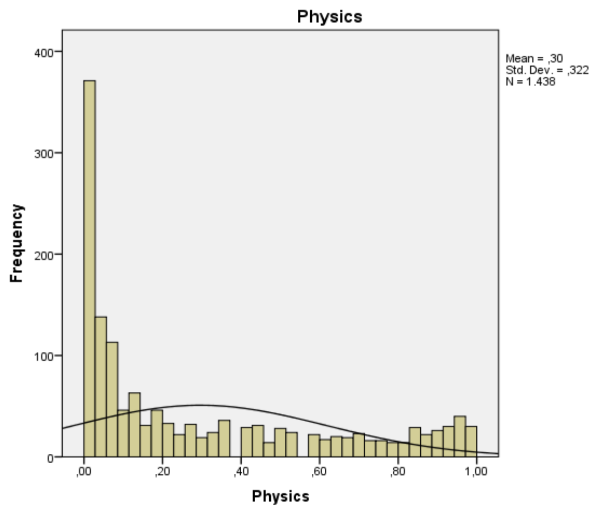

Fig 22. Physics Simulation (2017B)

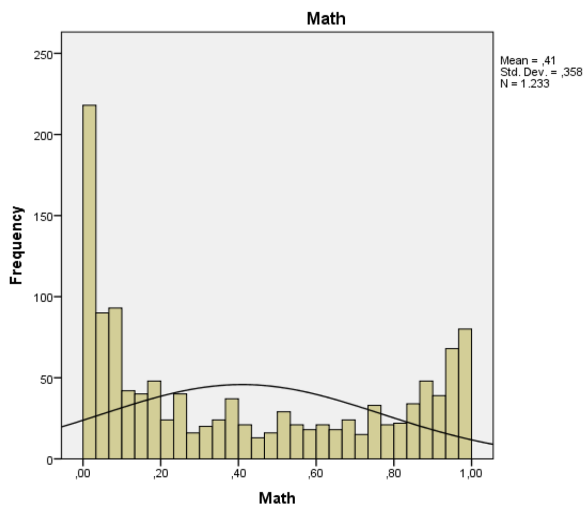

Fig 23. Math Simulation (2017B)

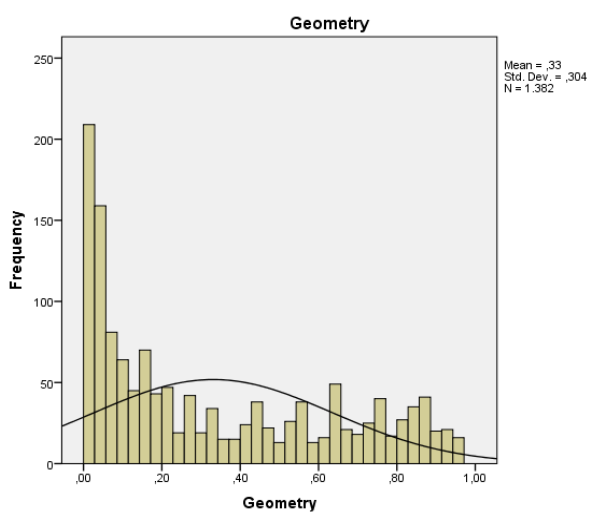

Fig 24. Geometry Simulation (2017B)

It was noted that the probability to get a low grade is always higher than to get better grades. It is important to identify that we have different N, i.e. different numbers of population because the desertion rate is different in each class (Table IX). Then, if it was simulated all the probabilities for total amount of 1793 students, a projection to pass in each class is described below.

TABLE IX
PROJECTION OF APPROVED STUDENTS IN EACH CLASS

Math*	Chemistry	Physics	Geometry	Literature
254	354	299	346	511
14,17%	19,74%	16,68%	19,30%	28,5%

*The Math class shows a strange low approval portion, because we estimated that 7.33/10 is the grade needed to pass. This grade is relatively high because the Math program changed and becomes easier over the years. We took data since 2012.

The chart above shows an important element, the estimate number of students who will pass in each class, in each semester and in each period. This is a key element in the academic planning and it ties very well with the last work. [10]

IV. CONCLUSIONS AND RECOMMENDATIONS

- Even we could have a strong relationship between the ENES grade and classes grades averages, such ENES grade was not needed to estimate the probability to approve, and it is due to a close study of the correlations showed.
- A logistic regression was the adequate technique to estimate the probability to pass, mainly because its function is smooth enough and it has a good slope growing in the middle of the grades.
- One way to improve a better classification for the regression is to consider a bigger amount of historical data.
- Literature shows a higher probability to pass, in the probability and in the simulated distribution.
- Geometry is the most difficult class to pass according to our estimation.
- One way to improve the model is to consider the periodicity of each year, it means there are two periods A and B, each one with its own academic features.
- This forecast procedure, using the mathematical model, will be periodically taken for the EPN like a standard process in order to plan and improve the academic management as well as satisfy the Organization criterion for the accreditation process.
- Monte Carlo simulations or the use of neural networks could be very useful in order to improve considerably the model accuracy, but these techniques need big amounts of data, especially with the development of neurons and its training.

REFERENCES

[1] Modelo de Evaluación Institucional de Universidades y Escuelas Politécnicas, CEAACES. http://www.ceaaces.gob.ec/sitio/wp-content/uploads/2016/06/Modelo-de-evaluacio%CC%81n-institucional-2016.pdf

[2] Listado de universidades o escuelas politécnicas públicas o privadas categoría A del país. http://programasbecas.educacionsuperior.gob.ec/wp-content/uploads/downloads/2017/04/Adjunto-2.-Listado-de-IES-del-pa%C3%ADs-Categor%C3%ADa-A.pdf

[3] Reglamento del sistema de estudios de las carreras de formación profesional y de postgrado de la escuela politécnica nacional. http://www.epn.edu.ec/wp-content/multiverso-files/18_54fd9a3cd557d/REGLAMENTO-SISTEMA-DE-ESTUDIO-DE-LAS-CARRERAS-DE-FORMACION-PROFESIONAL-Y-POSTGRADO.pdf

[4] INFORME DE AUTOEVALUACIÓN DE LA ESCUELA POLITÉCNICA NACIONAL 2015. http://www.adeponecuador.org/index.php/11-otras/51-informe-de-autoevaluacion-de-la-escuela-politecnica-nacional.

[5] Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. Statistics in medicine, 16(9), 965-980.

[6] Peng, C. Y. J., So, T. S. H., Stage, F. K., & John, E. P. S. (2002). The use and interpretation of logistic regression in higher education journals: 1988–1999. Research in higher education, 43(3), 259-293.

[7] Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. Behavior research methods, 41(3), 924-936.

[8] Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. Higher education: Handbook of theory and research, 10, 225-256.

[9] Morgan, S. P., & Teachman, J. D. (1988). Logistic regression: Description, examples, and comparisons. Journal of Marriage and Family, 50(4), 929-936.

[10] Sánchez, T. Sandoval, I. Salazar, D. Guevara, V. (2017). Modelo de Proyección para la Oferta de Cupos en el Primer Año de las Carreras de Ingeniería de la Escuela Politécnica Nacional. The fifteen LACCEI International Multi-Conference.