Prediction Model Based On Machine Learning or Softcomputing Technique for Predicting the Number of Candidates to Online Higher Education Institutions

WIT13

Ph.D. Student: Rosa Leonor Ulloa Cazarez
Universidad de Guadalajara CUCEA - DTI
Content

1. Problem
   a) Context
   b) Problem Statement
2. Research questions
3. Justification
4. Objectives
5. Hypothesis
6. Variables
7. Procedure
8. Methodology
9. Results
10. Conclusions
1. PROBLEM

a. Context
General Context

Information & knowledge Society

- Information management: capture, process & access,
- Transforming information into knowledge

National Education System

- Higher Education System
- Distance Education System

Prediction Model Based on a Machine Learning or Softcomputing Technique for Predicting the Number of Candidates to Online Higher Education Institutions

Ph. D. Student Rosa Leonor Ulloa Cazarez

Ph. D. Researcher Cuauhtémoc López Martín
b. PROBLEM STATEMENT

An overestimation of the demand of access to Online Higher Education Institutions (HEIs), could result in the investment of more resources than needed, whereas an underestimation could result in the assigning of not enough resources.
2. Research questions

<table>
<thead>
<tr>
<th></th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Does a machine learning or softcomputing technique spend less processing time than the time required for the expert judgment technologies?</td>
</tr>
<tr>
<td>2</td>
<td>Does the accuracy of a machine learning or softcomputing model outperform that of a statistical regression model, when both models are applied for predicting the dependant variable?</td>
</tr>
<tr>
<td>3</td>
<td>Could the Machine Learning or Softcomputing model be an alternative for supporting the making decision process in HEIs?</td>
</tr>
</tbody>
</table>
3. JUSTIFICATION

The prediction of the dependant variable is an impact factor for the planning process because to know the number of demandants and admitted for each season aloud the institution to take previous actions, as courses programming, teacher hiring and infrastructure disposition.
4. OBJECTIVES

1. To get a machine learning model spending less processing time than the time required for the expert judgment methodologies

2. To obtain from the machine learning model, an accuracy which outperforms that of a statistical regression model when both models are applied for predicting the dependant variable

3. To validate the machine learning model as an alternative for supporting through prediction if dependant variable, the making decision process.

Ph. D Student Rosa Leonor Ulloa Cazarez

Ph. D Researcher Cuauhtémoc López Martín
5. HYPOTHESIS

The accuracy of a machine learning or softcomputing model is statistically better than the accuracy obtained from a statistical regression model, when both are applied for predicting the dependant variable.
6. Variables

Total number of demandants

Number of demandants

Number of graduates from the next inferior level

From Jalisco State

From others States

From Jalisco State

From other States

Prediction Model Based on a Machine Learning or Softcomputing Technique for Predicting the Number of Candidates to Online Higher Education Institutions

Ph. D Student Rosa Leonor Ulloa Cazarez

Ph. D Researcher Cuauhtémoc López Martín
7. Procedure

1. To identify prediction techniques
2. To evaluate and compare the accuracy level of those techniques
3. To identify prediction models
4. To evaluate and compare the accuracy of the Machine Learning models
5. To select the machine learning model with higher accuracy level
6. To compare the accuracy of the machine learning model with that of the statistical regression
7. To propose a new Machine Learning prediction model
8. METHODOLOGY

Systematic literature review

To apply several models for the dependant variable prediction

To generate the prediction model

PH. D STUDENT ROSA LEONOR ULLOA CAZAREZ

PH. D RESEARCHER CUAUHTÉMOC LÓPEZ MARTÍN
PREDICTION MODEL BASED ON A MACHINE LEARNING OR SOFTCOMPUTING TECHNIQUE FOR PREDICTING THE NUMBER OF CANDIDATES TO ONLINE HIGHER EDUCATION INSTITUTIONS

Ph. D Student Rosa Leonor Ulloa Cazarez

Ph. D Researcher Cuauhtémoc López Martín
9. RESULTS

1. Systematic Literature Review

1.1 Main Questions

a) What was the purpose of the study?

b) What model or technique were used?

c) What variables were involved?

d) How model/technique were validated?

e) What was the result?
1.2 Results

- 622 articles found
- 17 articles revised

a) What was the purpose of the study?
- Prediction of different variables

b) What model or technique were used?
- Statistical, Neural network, Genetic fuzzy model
c) What variables were involved?

- Dropout
- Expenditure
- Course selection
- Student performance
- Academic performance
- Student participation
- Enrollment (1)
e) Results

There are some efforts to predict educative variables using statistical, machine learning or softcomputing techniques. We have found only one article with the predictable variable of enrollment.

The accuracy of the model or technique depends on the predictable variable.
2. Testing models

2.1 We applied the Statistical regression technique to the variables: ‘Aspirantes’ and ‘Admitidos’

The data shows normal distribution; there is a significative statistical correlation between the dependant & the independent variable
Dependant Variable: Admitidos
Independent Variable: Aspirantes
Lineal Model: Y = a + b * X
Equation: Admitidos = 5.88191 + 0.940013 * Aspirantes
Correlation Coefficient $r$: 0.988819 [This value shows a strong relation between both variables, to be close to +1]

$r^2$: 97.7764 [This value shows that the variable ‘Aspirantes’ explains in 97.7764% the variable ‘Admitidos’]

Magnitud Rate Error: 9.6512 [This value can be used to establish prediction limits for new predictions. Measures the dispersion of data around the regression line]

Absolut Error: 6.33692
2.2 We applied the multiple regression technique to the dependent variable ‘Admitidos’ and the independent variables ‘Aspirantes Mujeres’ and ‘Aspirantes Hombres’

There is a significative statistical relation between the variables. We have identified an atypic value that is correspondent to an extraordinary event in the institution during 2008.
Dependent Variable: Admitidos
Independent Variables: Aspirantes Hombres, Aspirantes Mujeres

Regression Equation:
Admitidos = 6.01255 + 0.944595*Aspirantes hombres + 0.938125*Aspirantes mujeres

The highest value of ‘P’ correspond to the variable ‘Aspirantes hombres’, but is less than 0.05 [0.0000]; therefore, we wouldn’t eliminate any variable from this model, including the atypic value
r2 = 97.777 [Explains the variability in the variable ‘Admitidos’ in a 97.777%]

r2 ajustada = 97.283% [This is a more appropriate statistical to compare models with different number of independent variables and the result of 97.283% is an indication of variability]

Error estándar de Est. = 10.1718

Error medio absoluto= 6.35243
Prediction model based on a machine learning or softcomputing technique for predicting the number of candidates to online higher education institutions
10. CONCLUSIONS

• Through the systematical review, we have observed the need of this type of research
  • The study field isn’t enough explored
  • Most common techniques are neural networks, genetic fuzzy, and statisticals
  • Most common variables are: dropout, performance, participation and expenditure
• The first approach to the dependant variable has proved:
  • That the statistical regression model is a good technique to do predictions, because the data shows a normal distribution and there was established a relation between both variables.
  • There was also established a relation between variables using a multiple regression statistical model.
• We are now studying a Fuzzy Logic model because in literature has proven to be as better as accurate than the lineal regression statistical model.
Thank you for your attention!