RESUMEN

El proceso educativo está condicionado por varios factores, sin duda, los más preponderantes están en el ámbito social, económico y familiar del entorno estudiantil; la preocupación de los estudiantes y sus familiares, una vez que terminan la educación media, es el acceso a la educación superior. El presente estudio de investigación, empleó técnicas de Big Data. El mismo que permitió que, a partir del examen de ingreso Ser Bachiller 2019 y de la encuesta de Factores Asociados 2019, del INEVAL, pronosticar el acceso a la educación superior de bachilleres en todo el país. De las datas proporcionadas se construyó una sola matriz, sobre ésta se depuró la información; eliminando así variables similares y otras que no presentan información en por lo menos el 50 % de observaciones.

Así mismo, se eliminaron observaciones que no presentaron resultados de las evaluaciones parciales. Posteriormente, se redujo la matriz, seleccionando las variables utilizadas en estudios similares realizados en Colombia, Argentina y Costa Rica. Además, se revisaron las variables que emplean los proyectos Pisa y Terce, y así tomar los indicadores más recurrentes como las variables significativas para este estudio. Se logró una matriz de 265 915 observaciones por 48 variables; de estas 48 variables, 45 correspondieron a factores asociados. Se aplicó el Análisis de Componentes Principales para representar los 45 factores asociados en un conjunto condensado de menor dimensión; sobre esta base se trabajaron los modelos predictivos que determinan el ingreso, o no, a la educación superior de un determinado individuo. Utilizando las técnicas de Regresión Logística, Análisis Discriminante y Máquina de Soporte Vectorial. Se estableció que, de estos modelos, el más adecuado es el de Regresión Logística el mismo que produce sus pronósticos con un 70 % de eficiencia. Adicionalmente, se determinaron los factores asociados que más influyen en forma positiva y en forma negativa, para que el estudiante ingrese a la universidad.

Palabras clave: Big Data, Análisis de Componentes Principales, Regresión Logística, Análisis Discriminante, Máquina de Soporte Vectorial.

ABSTRACT

To access to higher education is a concern for students who are finish high school. The educational process is conditioned by a great number of factors, the most preponderant is the social, economic, and family environment in which the student develops. This research study uses Big Data techniques that allow, to forecast the access to higher education of high school graduates throughout the country, with information based on: Ser Bachiller 2019 exams results, and the survey of Associated Factors 2019 INEVAL. With the data provided, single matrix was constructed. The matrix was depurated through the elimination of similar variables: those who does not have information in at least 50 % of the observations, observations without partial evaluations results were eliminated.

The matrix experienced a new depuration when we match used in similar studies carried out by countries in the region such as Colombia, Argentina, and Costa Rica. In addition, the variables used by Pisa and Terce are reviewed so that the most recurrent indicators are taken as the significant variables for this study, resulting in a matrix of 265 915 observations for 48 variables; of these 48 variables, 45 correspond to the Associated Factors. Applying Principal Component Analysis, the 45 Associated Factors are represented in a condensed set of smaller dimensions. This is the base on which the predictive model that determined the entrance or not of a certain individual to higher education is worked, using the techniques of Logistic Regression, Discriminant Analysis and Support Vector Machine, establishing that Logistic Regression produce a forecast with 70 % efficiency, after we chose this model to work, the study also determined which associated factors are the most influential in a positive or negative way for the University access of the student.

Key words: Big Data, Principal Component Analysis, Logistic Regression, Discriminant Analysis, Support Vector Machine.