Rev. Soc. cient. Parag. 2019;24(2):293-304

ARTÍCULO ORIGINAL – ORIGINAL ARTICLE

Análisis de la Educación Inicial en Paraguay a través de las Técnicas de Aprendizaje Automático

Analysis of Pre-school Education from Paraguay through Machine Learning Techniques

Viviana Elizabeth Jiménez Chaves[1], Miguel García Torres[2]
[1] Universidad Americana, Centro de Investigación. Asunción, Paraguay.
[2] Universidad Pablo de Olavide, Division of Computer Science. Sevilla, España.
Autor de correspondencia: viviana.jimenez@americana.edu.py
DOI: https://doi.org/10.32480/rscp.2019-24-2.293-304
Recibido: 23/09/2019. Aceptado: 1/12/2019.

Resumen: La educación inicial es fundamental para sentar las bases de conocimiento y habilidades que serán de vital importancia en el futuro. En Paraguay, el departamento de Estadística del Ministerio de Educación y Ciencias ha recolectado una base de datos con información sobre la escolarización de los niños. En este trabajo se hace un análisis descriptivo de dichos datos mediante una aproximación de minería de datos educativa (EDM del inglés) haciendo uso de técnicas de aprendizaje no supervisado. Para ello se consideraron las siguientes variables: Departamento, año, sexo, zona geográfica (rural o urbana), sector (oficial, subvencionada o privada) y etapa (primera o segunda). Los resultados obtenidos son prometedores y ayudan a entender la situación actual del país en relación con la Educación Inicial. Fruto del análisis se concluye, por una parte, que EDM es una aproximación que proporciona potentes técnicas de análisis en el ámbito educativo y, por otra, al contextualizar los resultados con las necesidades y realidades socioeconómicas del país, que la matricula en el nivel inicial tiene como principal factor de deserción el económico.

Palabras clave: Educación inicial, Paraguay, minería de datos educativa.

Abstract: Pre-school education is the key to establish the knowledge base and skills that will be of vital importance in the future. In Paraguay, the Department od Statististics at the Ministry of Education and Science ha collected data about the school enrolment of children throughout the contry. In this work a descriptive analysis is carried out using an unsupervised

learning approach. For such purpose, the following vairables were selected: Department, year, sex, geographical area (rural or urban), sector (official, charter school or private) y level (first or second). Results achieved are promising and help to understand the current status of Paraguay regarding Pre-school education. The general conclusions of this works are, on one side, that EDM is an apporach that provides pwerfoul tools to analyse data in education. On the other hand, when including the context of needs and socio-economic reality of the country, the major cause of school dropout is the economic factor.

Keywords: Pre-school education, Paraguay, educational data mining.

1. INTRODUCCIÓN

La República del Paraguay es un estado multicultural y plurilingüe, con cerca de 7 millones de habitantes (1). Es un país joven, en donde cerca del 70 por ciento de la población es menor de 30 años (dos millones de habitantes tiene menos de 14 años). Con un crecimiento anual de la población del 1,30%, mantiene una elevada tasa de fertilidad de 2.5 hijos por mujer (2).

En el Paraguay, al igual que en los demás países de la región, la educación inicial es de primer nivel del Sistema Educativo Nacional. En las dos últimas décadas el sistema educativo paraguayo ha planteado objetivos y estrategias tendientes a garantizar el acceso a la educación, como la universalización del preescolar.

El artículo N° 2 de la Ley N° 4088, sancionada en 2010, establece la obligatoriedad y gratuidad de la Educación Inicial en las escuelas públicas de gestión oficial. La Educación Inicial se compone por tres ciclos (maternal, inicial y preescolar) que concluyen cuando el escolar tiene cinco años y comienza la etapa de Educación Escolar Básica.

En Paraguay existe un gran número de escuelas pequeñas en zonas rurales, lo que suponen el 67% de las instituciones educativas que dan cobertura a la demanda educativa en la etapa obligatoria, pero cubren tan solo el 35% del total de la oferta educativa nacional (3).

El desafío actual más importante de la educación paraguaya es mejorar la calidad y la equidad de la educación. Los dos factores que marcan las desigualdades educativas más significativas son la ubicación geográfica y el nivel económico. Cerrar las brechas en el acceso a la Educación Inicial incluyendo el preescolar debería ser una prioridad para poder atacar el

problema del abandono escolar. El acceso los programas de atención integral a la primera infancia es aún bajo. Según datos del MEC, la tasa bruta de cobertura en el prejardín es de 8%, del jardín es de 38% y del preescolar es del 73% (4).

La apuesta a la primera infancia para edades de 3 y 4 años se refleja en el incremento de la matrícula del Pre escolar y Jardín, especialmente en los últimos años de la serie analizada. En Preescolar, en contrapartida, decreció en un 15 %, entre 2006 y 2016.En general, teniendo en cuenta todos los ciclos y niveles, desde 2006, se produce un decrecimiento sostenido de la matrícula total (-9%). Esto es principalmente provocado por la caída de la matriculación en Preescolar y la Educación Escolar Básica (1º, 2º y 3º ciclo), cuando el acceso gratuito a estos ciclos de estudios debe estar garantizado por el Estado (5).

Este estudio interdisciplinario, que aplica técnicas de minería de datos para los datos educativos, se conoce como Minería de Datos Educativos (EDM) (6). En EDM se han aplicado técnicas de agrupamiento como el K-means o agrupamiento Jerárquico para obtener perfiles. En en el cual el agrupamiento es aplicado para agrupar estudiantes que tuvieran estilos de aprendizaje similares (7). Otro trabajo interesante es (8), en el cual el análisis descriptivo es usado para encontrar perfiles de estudiantes que pudieran hacer equipos colaborativos de mayor éxito. Los datos de la herramienta Moodle también han sido considerados para obtener perfiles de estudiantes (9). También cabe destacar que en (10) se hace una revisión de las distintas técnicas de agrupamiento que han sido aplicadas al ámbito educativo.

En este trabajo se analizan diversas variables asociadas a los datos de matriculación del nivel inicial de los años comprendidos entre 2013 y 20116 de niños de todo el País. Además, se estudian los distintos perfiles de estudiante en base a las variables seleccionadas aplicando técnicas de agrupamiento. Esto nos permitirá entender mejor el estado actual del nivel inicial en Paraguay.

El resto del artículo se organiza del siguiente modo. La Sección 2 describen los datos y se introducen las técnicas de minería de datos. Los experimentos son descritos y discutidos en la Sección 3. Finalmente, las conclusiones son expuestas en la Sección 4.

2. METODOLOGÍA

2.1. Recolección de los datos

En este contexto se realizó la investigación basada en los datos del sistema de estadísticas del Ministerio de Educación y Ciencias el cual realizo el relevamiento de la situación del nivel inicial a través de los datos obtenidos por el departamento de estadísticas del Ministerio de Educación y Ciencias el cual realizo el relevamiento de los años 2013,2014,2015 y 2016 sumando en total 146.244 alumnos del nivel inicial en los 17 departamentos de todo el Paraguay.

Los datos fueron recogidos a través de cuestionarios relacionados a la matriculación en cada institución educativa que contaba con el Nivel Inicial en total 5840 instituciones tanto públicas, privadas y subvencionadas en todo el país.

Las variables que fueron seleccionadas son las siguientes: Departamento, año, sexo, zona geográfica (rural y urbana) Sector (oficial, subvencionada, privada) etapa (primera y segunda).

2.2. Técnicas de minería de datos

En esta sección se introducen las distintas técnicas empleadas para desarrollar un análisis en torno al ámbito de la minería de datos.

Agrupamiento

La tarea de agrupamiento se define formalmente del siguiente modo. Sea A un conjunto de n observaciones u objetos $a_i, i = 1, \ldots, n$, descrito mediante d atributos $X_j, j = 1, \ldots, d$. Para cada objeto a_i el valor del atributo X_j viene representado por a_{ij}. El objetivo del agrupamiento es obtener una partición de los objetos en p grupos $C_{1,\ldots}, C_p$ de modo que los objetos pertenecientes a un mismo grupo son más similares entre sí que aquellos perteneciente a grupos distintos.

En este trabajo el conjunto de datos tiene variables categóricas y numéricas. Por tanto, se usa una extensión del algoritmo K-medias (7) propuesto por Huang (8) adaptado para tratar datos mixtos. Dicho algoritmo difiere, básicamente, en la medida de similitud.

El algoritmo de agrupamiento está basado en particiones e identifica, con cada grupo, un centroide, el cual es un objeto real o hipotético que se calcula

como el centro de masa del propio grupo. Por tanto, podemos interpretar que el objetivo es, en este caso, determinar los centroides $\{c_1,\dots,c_p\}$ de los$p$grupos. Una vez tengamos estos centroides, se puede asignar cada uno de los objetos $a_i$al grupo $C_{k(i)}$ cuyo centroide sea más próximo. Por tanto, se busca la partición que minimice la distancia de cada objeto con su centroide.

$$min \sum_{i=1}^{n} D\left(a_i, C_{k(i)}\right)$$

con $k(i) = 1,\dots,n$. En el caso de datos mixtos, supongamos que tenemos t variables numéricas y $p - t$ variables categóricas. Entonces la medida de distancia$D$viene definida del siguiente modo

$$D\left(a_i, C_{k(i)}\right) = \sum_{j=1}^{p}\left(a_{ij} - c_{kj}\right)^2 + \gamma \sum_{j=p+1}^{d} \delta\left(a_{ij}, c_{kj}\right)$$

donde el primer término no es más que el cuadrado de la distancia Euclídea mientras que el segundo término es una medida de solapamiento sopesado por el término γpara evitar un sesgo en la medida por el tipo de datos. Para saber más sobre este término puede consultarse el trabajo (9). La medida de solapamiento viene definida de modo que vale 0 si el valor del atributo categórico$X_j$del objeto$i$no coincide con el del centroide$k(i)$y l en caso de que coincidan.

Para los experimentos se usa el paquete clustMixType (10) de R (11).

Índice de Silhouette

El valor de Silhouette es una medida de cohesión entre un objeto y el grupo al que pertenece comparado con su separación con otros grupos. Dicha medida puede usarse como una medida de separación entre grupos al extender su cálculo a todos los objetos. Este cálculo, denominado índice de Silhouette, permite estimar el número óptimo de grupos. Para ello, considerando que los$n$objetos han sido agrupados en $p$grupos, el índice de Silhouette viene definido por la expresión

$$Silhouette = \frac{1}{n}\frac{\sum_{i=1}^{n} b_i - q_i}{max(q_i, b_i)}$$

con$q_i = \frac{1}{p-1}\sum_{j\in C_{k(i)}, i\neq j} D\left(a_i, a_j\right)$la medida de disimilitud promedio del objeto i-ésimo con el resto de objetos pertenecientes al mismo grupo, $b_i = \min_{z\neq i} \frac{1}{|C_z|}\sum_{j\in C_z} D\left(a_i, a_z\right)$, donde $D(a_i, a_z)$ es la disimilitud del objeto a_i con el resto de objetos a_z que no pertenezcan al grupo $C_{k(i)}$. El máximo valor del índice indica el número óptimo de grupos. Para calcularlo se usó el paquete clustMixType.

Análisis de Componentes Principales (PCA) de datos mixtos

PCA hace una proyección lineal de los datos a unas nuevas coordenadas, denominadas Componentes Principales (PC), de modo que apuntan a las direcciones de máxima dispersión de los datos. En este sentido, una mayor dispersión se considera que tenemos una mayor información de la información contenida en los datos. Dichos PC están ordenados de modo que el primero es el que apunta a la dirección de máxima dispersión de los datos. La segunda es la siguiente dirección, ortogonal con la primera, con máxima dispersión y así sucesivamente.

Para poder aplicar PCA a datos con mezcla de variables categóricas y numéricas se usa PCAMIX (12), una extensión del método estándar PCA que introduce pesos en las observaciones y en los atributos. El algoritmo consta de tres etapas:

- Preprocesamiento de los datos. Se construye la matriz de datos con las variables categóricas y numéricas, así como las matrices de pesos de las observaciones y los atributos.
- Factorización de las coordenadas. En esta etapa se descompone la matriz de datos, usando las matrices de pesos como métricas, mediante una generalización del método de descomposición en valores singulares (SVD).
- Procesamiento de las cargas cuadráticas. Dichas cargas se definen como la contribución de cada variable a la varianza de los componentes principales.

Se consideró la librería PCAmixdata (13) para llevar a cabo el PCA.

3. RESULTADOS Y DISCUSIÓN

A lo largo de esta sección se describirán los distintos experimentos llevados a cabo. El objetivo que se persigue es demostrar que existen distintos perfiles de alumnos matriculados. Para ello analizaremos los datos. En primer lugar, se realizará un análisis preliminar de los datos para entender mejor cómo se distribuyen las matrículas en base al año y al sexo. Posteriormente se hará un análisis descriptivo haciendo uso de técnicas de minería de datos. para identificar los distintos perfiles de estudiantes.

3.1. Análisis exploratorio de los datos

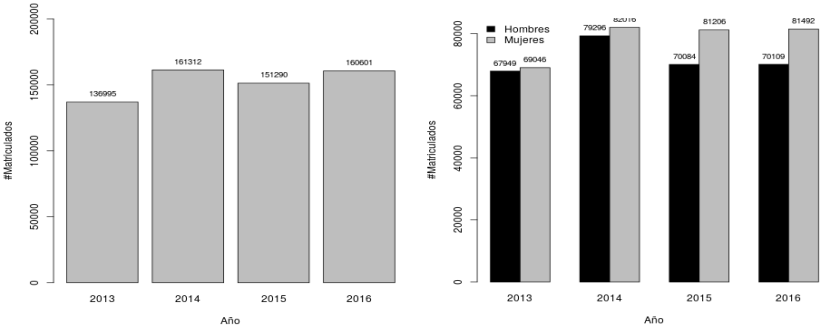

Figura 1: Número total de matriculados por año y número total de matriculados por sexo y año.

El número de matriculados puede verse en la Figura 1. La gráfica de la izquierda muestra el total de matriculados por año mientras que la de la derecha el número total por sexo y por año. Como puede apreciarse, el total de matriculados crece de 2013 a 2014 y se mantiene estable en los siguientes tres años del estudio. Sin embargo, al separar por sexo se aprecia que mientras el número mujeres matriculadas supera ligeramente al de hombres en 2013, dicha diferencia empieza a acuciarse a partir de 2014 debido al descenso del número de hombres matriculados mientras se mantiene estable el de mujeres. Entre los factores que se cita en el Informe Nacional Paraguay Educación para Todos 2000-2015 se menciona que el principal factor de disminución de la matrícula es el económico, la familia al no tener recursos

económicos opta por retirar del sistema escolar a sus hijos en especial los niños con el motivo de ayudar a llegar a un ingreso económico. Separando los matriculados por regiones, tal y como puede verse en la Figura 2, se aprecia que la región occidental sigue un patrón distinto a la oriental. El número de matrículas asociado a los hombres tiene un comportamiento similar en ambas regiones mientras que el asociado a las mujeres difiere. Mientras que en el oriental el número aumenta en 2014 y se mantiene constante, en la región occidental disminuye. Los factores asociados a la deserción de los matriculados son el económico, la lucha por la supervivencia diaria obliga a muchos padres e hijos a darle prioridad a la búsqueda de medios materiales para seguir subsistiendo. Con esta dolorosa realidad, la educación se vuelve prescindible.

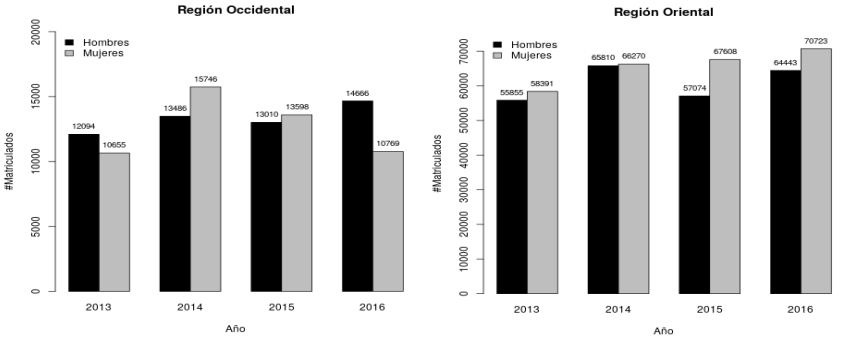

Figura 2: Número de matriculados por sexo y año para las regiones occidental y oriental de Paraguay.

3.2. Aplicación de las técnicas de minería de datos

A lo largo de esta sección estudiaremos el perfil de los distintos tipos de estudiantes matriculados. Sin embargo, *a priori* no hay ninguna información que nos indique cuántos perfiles hay; de modo que haremos uso del Índice de Silhouette para estimar el número adecuado de perfiles. Como puede verse en la Figura 2, el índice indica que el número óptimo es 2; de modo que haremos el análisis de conglomerado considerando 2 grupos.

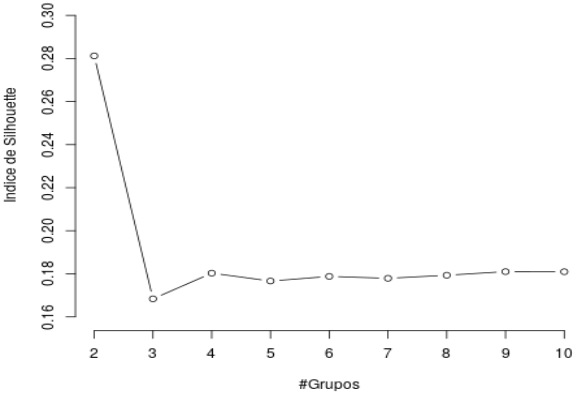

Figura 3: Número óptimo de grupos en base al índice de Silhouette.

Aplicando la técnica de agrupamiento encontramos los perfiles de cada uno de los grupos. Dichos perfiles se detallan en la Tabla 1. Cabe destacar que en ambos perfiles la etapa de estudio más representativa es la segunda infiriendo en el resto de las variables. Por ejemplo, el estudiante tipo del primer grupo sería de zona rural que cursa estudios oficiales y es hombre y proviene del departamento Central mientras que el del segundo curso proviene de una zona urbana que cursa estudio privados subvencionados y es preferentemente mujer. El número de matriculados suele ser mayor en el perfil correspondiente al primer grupo que al del segundo. En este contexto se puede concluir que la matricula suele ser mayor por tratarse de una institución del sector oficial donde la misma es de carácter gratuito, mientras que la institución privada subvencionada tiene un costo para el ingreso a la misma lo cual hace que la familia invierta un monto en la educación de sus hijos en este caso las hijas.

Tabla 1: Perfiles

Departamento	Zona	Sector	Etapa	Sexo	#Matriculados
Central	Rural	Oficial	Segunda	Hombre	1028
Caaguazú	Urbana	Privado-Subv.	Segunda	Mujer	272

La representación de los perfiles proyectada aplicando PCA puede verse en la Figura 3. En dicha proyección puede apreciarse que ambos perfiles tienen valores diferenciados en los valores de las variables estudiadas y podría separarse linealmente a pesar de que haya ciertos perfiles que solapen. Cuando hablamos de factores que pueden estar asociados al comportamiento de las variables estudiadas podemos citar como una de las principales el económico que influye tanto en los sectores, zonas y departamentos. En este caso Caaguazú es uno de los departamentos más pobres según el último Censo Nacional lo cual nos hace suponer que al tratarse de una institución privada subvencionada la familia está en una situación económica estable para enviar a sus hijos teniendo en cuenta el costo que implica la educación de estos.

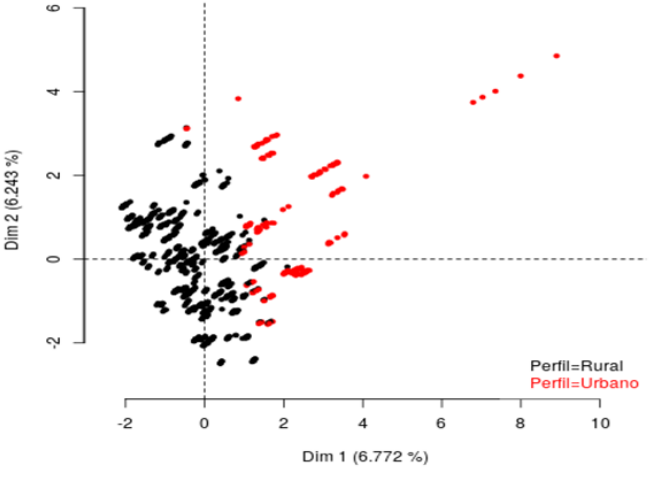

Figura 4: Representación de los perfiles usando Análisis de Componentes Principales.

4. CONCLUSIONES

En este trabajo se analizaron datos de educación inicial de Paraguay recolectados por el departamento de Estadística del Ministerio de Educación y Ciencia. Con el objetivo de entender la realidad el país en ese ámbito se hizo un análisis descriptivo de los datos aplicando técnicas de agrupamiento (aprendizaje no supervisado). El análisis se centró en los datos de

matriculación de los años comprendidos entre 2013 y 2016 de niños de todo el país. Como resultado del análisis se identificaron dos perfiles bien diferenciados. Uno de ellos correspondiente a un tipo de alumno más urbano frente a otro rural. El de perfil urbano se caracteriza por cursar estudios subvencionados frente al rural que elige el oficial. Otro resultado obtenido es la variación de los niveles de matriculación. Una posible causa de dicha variación en el nivel inicial en Paraguay es el factor económico ya que las familias deciden retirar a sus niños de este nivel por considerarlo un costo elevado e innecesario al no ser obligatorio la escolarización de los niños en ese nivel.

REFERENCIAS BIBLIOGRÁFICAS

1. DGEEC. Proyección de la Población Nacional, Áreas Urbana y Rural por sexo y edad. Fernando de la Mora: DGEEC, 2017.
2. UNESCO. Institute of Statistics. Paraguay country statistical data, 2018.
3. Banco Mundial. Informe del Banco Mundial. Washington DC: Banco Mundial, 2018. Disponible en: www.worldbank.org
4. Ministerio de Educación y Ciencias. Datos abiertos matriculaciones en Educación Inicial, 2016. Disponible en: https://datos.mec.gov.py/data/matriculaciones_inicial
5. Juntos por la Educación. Financiamiento público de la educación en el Paraguay. Notas para el debate y construcción de políticas públicas. Asunción, Paraguay, 2019.
6. Hegazi MO, Abugroon MA. The state of the art on educational data mining in higher education. International. Journal of Computer Trends and Technology. 2016;31(1):46-56.
7. MacQueen JB. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967:281–297.
8. Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Variables. Data Mining and Knowledge Discovery. 1998;2:283-304.
9. Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. Proceedings of the SIGMOD. Workshop on Research Issues on Data Mining and Knowledge Discovery. 1997:1–8.
10. Szepannek G. ClustMixType: User-Friendly Clustering of Mixed-Type Data in R. The R Journal. 2018;10(2):200-208.
11. Core Team R. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2019.

12. Kiers HAL. Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables. Psychometrika. 1991;56:197-212.
13. Chavent M, Kuentz-Simont V, Labenne A, Saracco J. Multivariate Analysis of Mixed Data: The PCAMix data R package, aXiv. 2017; 1411(4911).