REINGTEC

Revista de Ingeniería y Tecnologías para el Desarrollo Sustentable 11 (2023) 51 -54

ISSN: 2448 - 7198

MODELO DE CLUSTERIZACIÓN APLICADO AL DEPARTAMENTO DE PSICOLOGÍA DEL ITSOEH

CLUSTERIZATION MODEL APPLIED TO THE PSYCHOLOGY DEPARTMENT OF THE ITSOEH

Oropeza, José Martin[a], Sánchez, Aurelia[b], y Neri, Giovani Humberto[c]

[a, b, c] Tecnológico Nacional de México/ ITS del Occidente del Estado de Hidalgo, División de Ingeniería en Tecnologías de la Información y Comunicaciones. Mixquiahuala de J. Hidalgo, México. *jmoropeza@itsoeh.edu.mx .

RESUMEN. *El presente trabajo se enfoca en la extracción de conocimiento mediante clusterización aplicada al departamento de psicología del Instituto Tecnológico Superior del Occidente del Estado de Hidalgo (ITSOEH). Haciendo uso de gran cantidad de datos recopilados por medio de una encuesta realizada a los estudiantes de los 8 programas educativos para descubrir nuevas relaciones entre instancias, visualizarlos de forma clara y relevante que permita tomar decisiones informadas y diseñar intervenciones más efectivas.*

El objetivo de esta investigación es presentar los hallazgos que se obtuvieron al utilizar datos de estudiantes de manera anónima, para identificar los motivos más recurrentes de la canalización de estudiantes así como el programa educativo al que pertenecen y su relación con otras instancias como la edad y el género, de tal manera que se obtenga una caracterización precisa de la situación de los estudiantes que solicitan (son canalizados) al departamento de psicología para conocer a los estudiantes más allá de la información que se obtiene con los datos a simple vista.

Palabras clave: *minería de datos, k-means, estudiantes*

ABSTRACT. *This work focuses on the extraction of knowledge through clustering applied to the psychology department of the Higher Technological Institute of the West of the State of Hidalgo (ITSOEH). Using a large amount of data collected through a survey of students from the 8 educational programs to discover new relationships between instances, visualize them in a clear and relevant way that allows informed decisions to be made and more effective interventions to be designed.*

The objective of this research is to present the findings that were obtained by using anonymous student data, to identify the most recurrent reasons for the channeling of students as well as the educational program to which they belong and its relationship with other instances such as age and gender, in such a way that a precise characterization of the situation of the students who request (are channeled) to the psychology department is obtained to know the students beyond the information obtained with the data at a glance.

Key words*: data mining, k-means, students.*

INTRODUCCIÓN

El concepto de analítica descriptiva se utiliza para extraer información de datos históricos que dé respuesta a preguntas como ¿Qué sucedió? O ¿qué está sucediendo? en una organización, de tal manera que a partir de este análisis se logren entender las razones detrás del éxito o fracaso de ciertas iniciativas llevadas a cabo. En el ITSOEH, el departamento de psicología recaba, almacena y utiliza gran cantidad de datos provenientes de los estudiantes que atiende, algunos de estos datos son de conocimiento público como la matrícula o el programa educativo, pero algunos otros son de tipo confidencial o sensible como el motivo de canalización o el diagnóstico.

El modelo utilizado fue la clusterización o clustering (segmentación) el cual busca patrones de datos apoyado en algoritmos y fórmulas matemáticas, en particular se utilizó k-means. Este algoritmo agrupa las instancias a partir de la distancia entre ellas y un punto central o principal el cual representa el promedio means de cada grupo. A partir de este proceso se obtuvieron datos de interés en el ámbito de la psicología donde se identificaron tendencias y conexiones significativas para respaldar la toma de decisiones informadas en el Departamento de Psicología del ITSOEH.

METODOLOGÍA

La metodología utilizada sigue las etapas descritas por Tomar D. y Agarwal S.[1] la cual inicia con los datos

almacenados, la selección y preprocesamiento, la obtención de patrones (en este caso agrupación) y finalmente visualización e interpretación de los patrones identificados. Figura 1.

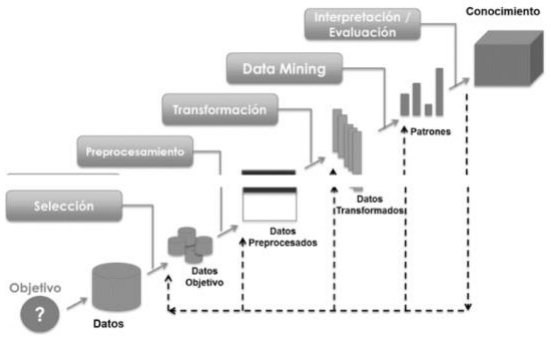

Figura 1. Etapas de minería de datos: pre-procesamiento y post-procesamiento de datos. (D. Tomar, 2013)

Los datos se obtuvieron mediante encuestas aplicadas vía formulario de internet a un total de 1273 estudiantes. Importando esta información a hojas de cálculo.

En la etapa de preprocesamiento se llevó a cabo un riguroso proceso de exploración y preparación de datos, sustituyendo la información sensible por claves que permitan el procesamiento matemático, pero manteniendo el anonimato. Figura 2.

	A	B	C	D
1	Tipo_ID	Fue canalizado a algu	Situación Académica	Alumno_ID
2	t1	Indefinido	Regular	a1
3	t2	Indefinido	Regular	a2
4	t3	taller de hábitos de es	Regular	a3
5	t4	Becas	Regular	a4
6	t5	Psicología, Taller de a	Regular	a5
7	t6	Psicología	Regular	a6
8	t7	Asesorías Academica	Regular	a7
9	t8	Indefinido	Regular	a8
10	t9	Indefinido	Irregular	a9
11	t10	Becas	Irregular	a10
12	t11	Tutoría individual	Regular	a11
13	t12	Psicología, Servicio m	Irregular	a12

Figura 2. Vista parcial de data set de alumnos-canalización-situación académica. (fuente propia)

Para la etapa de data mining y búsqueda de patrones se utilizó K-means el cual es un algoritmo de clasificación no supervisado (clusterización) que agrupa objetos de acuerdo con sus características, para lo cual se utiliza la distancia cuadrática. Fórmula 1.

$$E(\mu i) = \sum_{i=1}^{k} \sum_{xj \epsilon Si} |xj - \mu i|^2$$

Fórmula 1.- Distancia cuadrática del algoritmo k-means

Donde el conjunto de datos S lo integran objetos representados por *x*j, *k* es el número de grupos (clústeres) y μi es el centroide correspondiente de cada grupo. De tal manera que el algoritmo se repite hasta que los centroides no se modifiquen (se minimice la distancia) con lo cual se considera que el algoritmo está optimizado y entonces se tienen los grupos organizados por características similares.

Finalmente, en la etapa de interpretación se utilizó la versión de prueba de un software de aplicación que facilita la visualización de resultados [2]. El software ya tiene implementado como parte de sus librerías el algoritmo k-means por lo tanto, al alimentar el conjunto de datos se obtienen los primeros resultados. Figura 3.

Figura 3. Vista general del grupo de datos. (fuente propia).

Como puede observarse, el software representa cada instancia en forma de barras, en este primer resultado se pueden visualizar los datos de acuerdo con su valor estadístico simple, por ejemplo, la cantidad de canalizaciones al departamento de psicología por tipo o por programa educativo.

Para obtener nuevo conocimiento es necesario aplicar técnicas de minería de datos, en esta etapa se crearon 6 clústeres. Los cuales se agruparon de la siguiente manera. Figura 4.

Figura 4. Vista general de la creación de clusters. (fuente propia)

Sin embargo, el uso de clústeres proporciona mayor información que describe y caracteriza a las diferentes instancias. A continuación, se presenta el análisis de resultados con el algoritmo k-means.

RESULTADOS Y DISCUSIÓN

En primer lugar, se analizaron los histogramas de frecuencias de interés general por ejemplo ¿de cuál semestre se canalizan más estudiantes? Se obtuvo que al analizar el clúster por semestre en todos los PE. Resultó que los semestres en los cuales más canalizaciones se realizan son de 2do. a 4to con un 62.07 % y van disminuyendo gradualmente hasta llegar a un porcentaje de 10.3 % en los últimos 2 semestres. Figura 5.

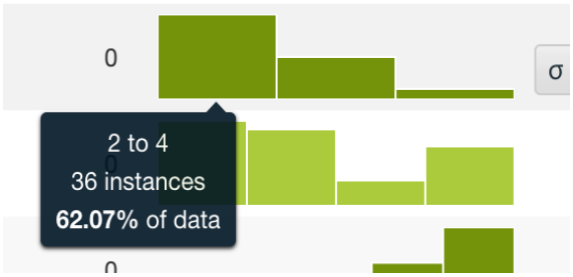

Figura 5. Número de canalizaciones por semestre (fuente propia)

El resultado más significativo se obtuvo al analizar el clúster por tipo de canalización. Partiendo de la idea que las principales causas de consultas psicológicas en el nivel superior en México se refieren a depresión, seguido de trastornos alimenticios y las relaciones de pareja[3].

En el ITSOEH se encontró que la situación académica fue la principal causa para solicitar atención psicológica. Figura 6a.

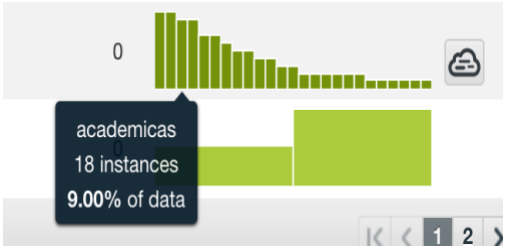

Figura 6a. Número de canalizaciones por tipo. (fuente propia)

En segundo lugar, se encontró el aspecto económico, el cual incluye a estudiantes que trabajan lo que significa reducción de tiempo para realizar actividades académicas. Figuras 6b.

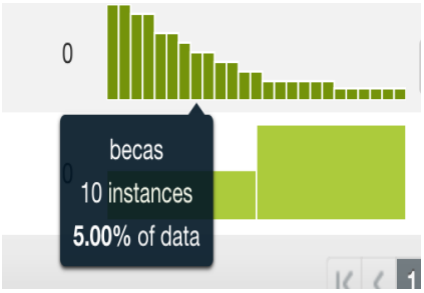

Figura 6b. Número de canalizaciones por tipo. (fuente propia)

Y en tercer lugar el aspecto psicológico como la depresión o déficit de atención. Figura 6c.

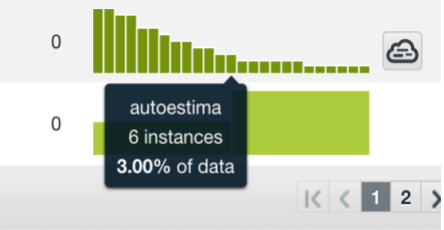

Figura 6c. Número de canalizaciones por tipo. (fuente propia)

Finalmente se analizó la correlación entre dos variables de interés, el sexo y el motivo de canalización, los resultados indican que es mayor el número de mujeres que asisten al departamento de psicología por motivos económicos y que de ellas en su mayoría son estudiantes regulares. Figura 7.

Figura 7.- Correlación entre sexo y motivo de canalización (fuente propia).

El conjunto de resultados puede brindar una idea más clara y precisa de los aspectos que pueden tomarse en cuenta para orientar los esfuerzos del departamento de psicología[4]. ¿Qué grupo de estudiantes requiere mayor atención sobre aspectos económicos? La clusterización indica que el deberían ser estudiantes femeninos de los semestres 2 a 4to.

CONCLUSIONES

El análisis de datos mediante la clusterización arrojó nuevos hallazgos sobre todo en relación al tipo de canalización más recurrente, que resultó ser la situación académica, lo cual incluye hábitos de estudio, perfiles de egreso del nivel medio superior y déficit de atención. Ante estos resultados es claro que la canalización al departamento de psicología en ocasiones se trata más de casos de asesoría académica y de cuestiones económicas que de aspectos psicológicos.

AGRADECIMIENTOS Y/O RECONOCIMIENTOS

Los autores agradecen el apoyo proporcionado por personal del departamento de psicología del Instituto Tecnológico Superior del Occidente del Estado de Hidalgo

REFERENCIAS

1.- D. Tomar, S Agarwal(2013) *A survey on Data Mining approaches for Healthcare*. International Journal of Bio-Scince and Bio-Technology vol. 5. (5), 241-266. DOI://dx.doi.org/10.14257/ijbsbt.2013.5.5.25

2. Ferrari, A., & Russo, M. (2016). Introducing Microsoft Power BI. Microsoft Press. https:// download .microsoft.com

3. Riveros Rosas, Angelica (2018) *Los estudiantes universitarios: vulnerabilidad, atención e intervención en su desarrollo.* Revista digital Universitaria. (RDU)UNAM. Vol19,(1) DOI:/doi.org/10.22201/codeic.16076079e.2018.v19n1.a6.

4. Izar, J. M., Ynzunza, C. V., López , H. A (2011). Factores que afectan el desempeño de los estudiantes de nivel superior en Rioverde SLP. *Revista investigación Educativa*, *17*(1), 32-41. Instituto de Investigación en Educación. Universidad Veracruzana