Calidad Universitaria mediante Técnicas del Data Mining

Myriam Herrera[1], María Inés Lund[2], Susana Beatriz Ruiz[1], María Gema Romagnano[2]

[1]Departamento de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan
[2]Instituto de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan

mherrera, mlund, mromagnano@{iinfo.unsj.edu.ar}, sbruizr@yahoo.com.ar

RESUMEN

En la actualidad, la mayoría de las instituciones, empresas u organizaciones miden la calidad de sus productos y/o servicios. De igual forma las instituciones educativas se ven obligadas a medir la calidad educativa. Para ello se necesita conocer los factores que influyen en la calidad de la institución, entre ellos, los relacionados al rendimiento académico de sus alumnos y al grado de satisfacción de sus egresados. Para lograr esto se utilizarán valiosas técnicas estadísticas que permitirán clasificar sujetos u objetos a partir de características similares. Estas técnicas se pueden diferenciar por la manera de extraer conocimiento útil escondido en los datos. Por un lado, el Análisis Discriminante, también referido como reconocimiento de patrones supervisado o asistido o aprendizaje con guía. Por otro lado, el Análisis de Conglomerados, referido como reconocimiento de patrón no supervisado o conocimiento sin guía. Como es común recopilar grandes conjuntos de datos, de distinta naturaleza, en voluminosas bases de datos, es que se utilizarán los análisis de Datos Simbólicos empleando la Lógica Difusa, que son también herramientas para Data Mining. En este proyecto se aplicarán las técnicas mencionadas para analizar los factores influyentes en la calidad universitaria como así también se detectarán tipologías básicas de grupos, obtenidos de los alumnos universitarios y egresados de la Facultad de Ciencias Exactas, Físicas y Naturales de la UNSJ.

Palabras clave: Calidad Universitaria, Clasificación, Data Mining.

CONTEXTO

Este proyecto ha sido presentado en la Convocatoria 2017 de la Universidad Nacional de San Juan, para el período comprendido entre 01/01/2018 al 31/12/2019 y se encuentra en proceso de evaluación externa.

Se desarrollará en el ámbito de la Facultad de Ciencias Exactas, Físicas y Naturales de la Universidad Nacional de San Juan, con el apoyo de la Secretaría de Asuntos Estudiantiles, el Instituto de Informática y los Departamentos de Biología, de Geofísica y Astronomía, de Informática y de Geología, que abarcan las distintas carreras que se dictan en esta Facultad.

Es un proyecto que continúa en línea con las investigaciones desarrolladas en proyectos anteriores, entre los que se mencionan:

- "Técnicas de Clasificación aplicadas al rendimiento académico". Acreditado por el CICITCA. Vigencia: 01/01/2016 – 31/12/2017. Código: 21 E / 1011.
- "Algoritmos de Clasificación de Procesos Multivariados utilizando Medidas de Asociación Espacial". Acreditado por el CICITCA. Vigencia: 01/01/2014 – 31/12/2015. Código: 21 E / 948.
- "Determinación y Comparación de Perfiles Sociales y Culturales de Estudiantes Universitarios a través de Técnicas Estadísticas Multivariadas". Acreditado por el CICITCA. Vigencia:

01/01/2014 – 31/12/2015. Código 21 F/ 982.

- "Clasificación Espacial Multivariada". Acreditado por el CICITCA. Vigencia: 01/01/2011 – 31/12/2013. Código: 21 E/ 878
- "Reducción y Selección de Variables en la Clasificación Digital". Acreditado por el CICITCA. Vigencia: 1/01/2008 - 31/12/2010. Código: 21 E/ 820.
- "Aplicación de una metodología en la medición de la calidad del proceso enseñanza aprendizaje en la universidad". Acreditado por el CICITCA. Vigencia: 01/05/2003 al 31/12/2005.

1. OBJETIVOS

Objetivo General

Determinar factores influyentes en los alumnos y egresados universitarios que caractericen la Calidad Universitaria.

Objetivos Específicos

- Formar recursos humanos, al nivel de grado y posgrado, en la temática que involucra la Gestión de Calidad en Educación y específicamente en la metodología elaborada.
- Analizar los datos otorgados por los sistemas SIU Kolla (Egresados) y SIU Guaraní (Alumnos).
- Determinar las variables influyentes que caractericen a alumnos y egresados universitarios.
- Identificar qué variables tienen mayor poder de discriminación y de predicción en la clasificación de sujetos.
- Determinar los datos simbólicos que determinarán reglas lógicas y taxonomías de las variables influyentes en alumnos y egresados.
- Transferir la herramienta metodológica obtenida a instituciones educativas del medio.

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

Los datos se han convertido en un recurso crítico en muchas organizaciones e instituciones con diversos objetivos y por lo tanto, el acceso eficiente a estos, el compartirlos, extraer información de los mismos y hacer uso de la información se transforma en una urgente necesidad. Existen varios enfoques de investigación que han aportado en ésta temática [1], [2], [3], [4]. El objetivo de analizar y comprender grandes y complejos conjuntos de datos, que posteriormente conducen a valiosa información, es común a todos los campos de los negocio, ciencia, ingeniería, entre otros [5]. La habilidad para extraer conocimiento útil escondido en esos datos y actuar sobre ése conocimiento está transformándose en algo cada vez más importante en el mundo competitivo actual. Como resultado hay muchos esfuerzos, no sólo para integrar varias fuentes de datos dispersos a través de sitios diferentes, sino también extraer información de esas bases de datos en la forma de patrones y tendencias. Dentro de la Inteligencia Artificial, la Minería de Datos, comúnmente conocida como Data Mining, analiza conjuntos de datos para encontrar relaciones y resúmenes de datos útiles para el propietario de los datos. Estas relaciones y resúmenes derivados a través del ejercicio del Data Mining se refieren a modelos y patrones.

La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos. Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento [6], [7].

Los modelos supervisados o predictivos requieren de un conjunto de pruebas y de interacciones de entrenamiento. Las técnicas usadas son la clasificación (Análisis Discriminante) y la predicción de valores. Los modelos no supervisados o descriptivos descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). Las técnicas usadas son: Asociación, Segmentación o 'Clustering' (Análisis de Conglomerados) [8], [9].

El Reconocimiento de Patrones tiene como objetivo la **clasificación** de objetos dentro de un número de categorías o clases. Dependiendo de la aplicación estos objetos pueden ser imágenes, señales o cualquier tipo

de medidas que necesitan ser clasificadas. Esas medidas se llaman patrones.
Las medidas usadas para la clasificación de objetos o patrones son conocidas como características. El conjunto de todas las características forman el vector que identifica únicamente a un patrón (objeto).
Las cuestiones que preocupan en el diseño de un sistema de clasificación que ejecuta la tarea de un reconocimiento de patrones dados son: (a) la generación de características para lo cual es importante la elección del mejor número de características; (b) el diseño del clasificador y finalmente, cuando el clasificador está diseñado, (c) la evaluación del rendimiento del clasificador diseñado mediante el error de clasificación [10], [11], [12].
Análisis de Datos Simbólicos es también una herramienta para **Data Mining** que generaliza los métodos clásicos exploratorios e informáticos. En muchas actividades humanas es común recopilar considerables conjuntos de datos en grandes bases de datos, por lo cual es importante resumir estos datos en términos de sus conceptos con el sentido de extraer nuevos conocimientos. Estos conceptos se pueden describir por tipos de datos más complejos, llamados Datos Simbólicos, que contienen variación interna y son estructurados. Es en este contexto que surge la necesidad de extender los métodos de análisis de datos estándar (exploratorio, representaciones gráficas, clustering, análisis factorial, discriminación, etc.) a estos datos simbólicos. Los datos simbólicos implican tablas de datos más complejas llamadas tablas de datos simbólicos. Una celda de tales tablas no necesariamente contiene valores categóricos o cuantitativos simples, sino muchos valores, que pueden tener pesos o estar unidos por reglas lógicas y taxonomías. Por ejemplo, una celda puede contener un intervalo o una distribución. Este tipo de Análisis será generado para analizar y tomar decisiones sobre grandes bases de datos, especialmente para datos de encuestas. Si bien resumen, en gran proporción, las bases, éstas preservan lo esencial o la información de interés. Además, permiten visualizar, comparar y clasificar objetos [10], [13], [14].
También el uso de la Lógica Difusa puede ser de vital importancia en cualquier proceso de Minería de Datos ya que es habitual que el conjunto de datos a analizar se haya obtenido con un propósito distinto al de la extracción de conocimiento. Es común la presencia de información numérica junto con información textual, con ambigüedades por el uso de diferentes símbolos con igual significado, redundancia, términos perdidos, imprecisos o erróneos, etc.[15]. La inclusión de la Lógica Difusa dentro del Soft Computing, constituye una herramienta de representación del conocimiento que permite modelar incertidumbre e imprecisión de una forma sencilla y directamente interpretable por el usuario [16]. En este proyecto se aplicarán las técnicas mencionadas, o una combinación de ellas.

3. FORMACION DE RECURSOS HUMANOS

En el grupo de trabajo se encuentran dos doctorandos. Se prevé la incorporación al presente proyecto de profesionales que actualmente están cursando posgrados en Computación, y también no docentes y alumnos ayudantes.

4. REFERENCIAS

[1] M. de M. Diaz, P. A. Urquijo, J. M. Arias Blanco, T. Escudero Escorza, S. Rodriguez Espinar, and J. Vidal García, “Evaluación del rendimiento en la enseñanza superior. Comparación de resultados entre alumnos procedentes de la LOGSE y del COU,” in *Revista de Investigación Educativa*, vol. 20, no. 2, M. de Miguel, P. A. Urquijo, J. M. A. Blanco, T. E. Escorza, S. R. Espinar, and J. V. García, Eds. 2002, pp. 357–383.

[2] T. Escudero Escorza, “La evaluación y mejora de la enseñanza en la Universidad: otra perspectiva,” in *Revista de Investigación Educativa*, vol. 18, no. 2, 2000, pp. 405–416.

[3] G. M. Garbanzo Vargas, “Factores asociados al rendimiento académico en

estudiantes universitarios, una reflexión desde la calidad de la educación superior pública," *Rev. Educ.*, vol. 31, no. 1, pp. 43–63, 2007.

[4] M. H. Efrón and A. Pérez Lindo, Eds., *Aportes al debate sobre la gestión universitaria I*. Buenos Aires, Argentina: Editorial De los cuatro vientos, 2005.

[5] N. Moscoloni, "Las Nubes de Datos. Métodos para analizar la complejidad," 2005.

[6] A. . Anaya and J. G. Boticario, "A data mining approach to reveal representative collaboration indicators in open collaboration frameworks," 2009.

[7] M. de L. Gurmendi, "De la información al conocimiento: factores que ayudan a un mejor uso de la tecnología en la gestión - Nülan," in *Aportes al debate sobre la gestión universitaria II*, R. I. In Efrón, Marcelo Héctor y Vega, Ed. Buenos Aires: De los cuatro vientos, 2005, pp. 66–75.

[8] I. González López, "Realización de un Análisis discriminante explicativo del rendimiento académico en la Universidad," in *Revista de Investigación Educativa*, vol. 22, no. 1, 2004, pp. 43–59.

[9] M. M. Torrado-Fonseca and V. Berlanga-Silvente, "Revista d'innovació i recerca en educació.," *Rev. d'Innovació i Recer. en Educ.*, vol. 6, no. 2, pp. 150–166, 2013.

[10] H.-H. Bock and E. Diday, *Analysis of Symbolic Data : Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Berlin Heidelberg, 2000.

[11] D. Peña, *Análisis de datos multivariantes*. McGraw-Hill/Interamericana, 2002.

[12] "Spad - Software Informer. SPAD,a data analytics software, uses company data to anticipate risks." [Online]. Available: http://spad.software.informer.com/. [Accessed: 13-Mar-2017].

[13] E. Diday, "Análisis de Datos Simbólicos," *Rev. IRICE*, vol. 11, 1997.

[14] R. A. (Richard A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*. Pearson Prentice Hall, 2007.

[15] J. Hernández Orallo, M. J. Ramírez Quintana, and C. Ferri Ramírez, *Introducción a la Minería de Datos*. Editorial Pearson, 2004.

[16] L. A. Zadeh, "Fuzzy Sets," *Inf. Control*, vol. 8, pp. 338–353, 1965.