Descubrimiento de reglas de clasificación para estudiantes que se inscriben del bachillerato a carreras universitarias.

Jorge H. Guanín Fajardo
Universidad Técnica Estatal de Quevedo, Quevedo, Quevedo, Los Ríos, Ecuador, jorgeguanin@uteq.edu.ec

Iván Ponce Vélez
Universidad Técnica Estatal de Quevedo, Quevedo, Quevedo, Los Ríos, Ecuador, ivanpv1403@gmail.com

ABSTRACT

The application of data mining techniques using supervised provides information for decision making from senior academic. This article discusses the results of software algorithms Keel, 14 algorithms are used to extract rules the success or failure of the leveling course..

Keywords: Data Mining, supervised techniques, algorithms, knowledge, sampling techniques.

RESUMEN

La aplicación de la minería de datos utilizando técnicas supervisadas proporciona información para la toma de decisiones de los responsables académicos. Este artículo discute los resultados de los algoritmos de software Keel, se utilizan 14 algoritmos extrayendo reglas con el éxito o el fracaso del curso de nivelación.

Palabras claves: Minería de datos, técnicas supervisadas, algoritmos, conocimiento, técnicas de muestreo.

1. INTRODUCCIÓN

La minería de datos educativos EDM[1] (siglas en inglés) se concentra en métodos computacionales para el uso de los datos con el fin de abordar importantes cuestiones educativas, uno de los objetivos de la EDM consiste en la mejora de los sistemas de educación personalizada. La minería de datos educativos puede mejorar la eficacia, la personalización y/o la adaptabilidad de estos entornos de aprendizaje. A su vez, los datos de alumnos procedentes de sistemas personalizados son semánticamente más relevantes, que los datos de la web tradicional basada en sistemas educativos, lo que puede llevarnos a un análisis más profundo. En la actualidad se pueden destacar varios trabajos realizados en el ámbito de la minería de datos educativos. Sin embargo, nos enfocaremos a los más recientes de esta nueva área utilizando técnicas de aprendizaje automático supervisadas.

1.1 PROBLEMA

La carencia de conocimiento relevante para encaminar las buenas prácticas académicas en los estudiantes del bachillerato dificulta el acceso a las carreras que oferta la Universidad. Uno de los problemas que afrontan los bachilleres es la indecisión respecto a las carreras que debe seguir en su vida universitaria, sin embargo, la mayoría optan por carreras que son guiadas por grupos de estudios que forman ellos mismos en su etapa colegial, ofertas de amigos, marketing de las carreras promocionando una buena profesión, etc.

2. PROPÓSITO DEL TRABAJO

Descubrir el conocimiento oculto en los datos almacenados en el sistema de informático relacionado con los bachilleres que optan por las carreras que oferta la Universidad Técnica Estatal de Quevedo, a través de las técnicas de clasificación supervisadas disponible en la minería de datos.

3. DESCUBRIMIENTO DEL CONOCIMIENTO (KDD)

La aplicación de las técnicas de minería de datos en la investigación está relacionada al aprendizaje supervisado utilizando para esto 14 algoritmos de clasificación de tres grupos diferentes "Crisp Rule Learning", "Decision Trees" y "Evolutionary crisp

[1] Educational Data Mining

rule learning". El conjunto de datos que se adquiere para la ejecución de los algoritmos cuenta con 1126 instancias y 15 atributos entre numéricos y categóricos incluida la clase, dada la realidad de los datos con los que se trabajará la clase cuenta ejemplos distintos, es decir tenemos un problema con ejemplos no balanceados (Imbalanced) Un método popular para resolver el problema del conjunto de datos no balanceados es volver a muestrear el conjunto de entrenamiento. Sin embargo, pocos estudios en el pasado han considerado el re-muestrear (resampling) algoritmos en los conjuntos de datos con alta dimensión (Witten et al, 2000).

4. ANÁLISIS DE LOS RESULTADOS

Al resultado de los algoritmos propuestos en este trabajo se les aplicó la prueba de Friedman[2] para determinar si existe solapamiento entre ellos además de la prueba de Tstudent[3] para que se determine estadísticamente cuál de las técnicas de muestreo utilizada en el conjunto de datos tiene mejores resultados. El entrenamiento de los algoritmos de clasificación elegidos en este trabajo demuestra tener mejores incidencias aplicando la técnica de muestreo "OverSampling". Por otra parte es importante destacar que la generación de reglas de clasificación o modelo, está relacionado con el tamaño del conjunto de datos utilizado.

5. CONCLUSIONES

Precisemos antes que nada, que los resultados obtenidos en el entrenamiento de los algoritmos pueden tener un mejor ajuste y alcanzar resultados más apropiados utilizando atributos con mayor relevancia, debido a que la Universidad no cuenta con información suficiente de los bachilleres postulantes en cuanto a su desarrollo académico, entorno social, situación económica y otra información que contribuyan para obtener un mejor modelo. Sin embargo, uno de los componentes más importantes es la extracción del conocimiento que servirá para la gestión universitaria para concertar sus decisiones. En síntesis, las técnicas aplicadas y algoritmos que se han entrenado para este caso en particular del que se obtiene conocimiento (reglas) con el propósito de ayudar a la Universidad en la aplicación de estrategias focalizadas a grupos de intereses.

6. AUTORIZACIÓN Y RENUNCIA

Los Autores autorizan a LACCEI para publicar el documento en las actas del congreso. LACCEI, los editores no son responsables ni por el contenido ni por las implicaciones de lo que se expresa en el documento.

REFERENCIAS

Amo, J. (2004). Reducción de Datos basada en Selección Evolutiva de Instancias para Minería de Datos.

Au et al. (2003). In Anovel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation.*, 532-545.

Breiman et al. (1984). Classification and Regresion Trees. . *Chapma and Hall.*

Carvalho et al. (2004). A hibryd decision tree/genetic algorithm method for data mining. . *In Information Sciences*, 13-35.

Chellatamilan et al. (2011). Effect of Mining educational data to improve Adaptatino of learning in e-learning. *Second International Conference on Sustentable Energy and Intelligent System*, 922-927.

Clark et al. (1989). The CN2 induction algoritthm. *Machine Learning Journa*, 261-283.

Granada. (2008). KEEL:a software tool to asses evolutionary algorithms for data mining problems. *University of Granada*.

Gray et al. (2008). Classification tree analysis using TARGET. *In Computational Stadistic $ Data Analysis*, 1362-1372.

Greene et al. (1993). In Competition-based induction of decision models from examples. *MAchine Learning*, 229-257.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. . *Machine Learning Journal 11*, 63-91.

KOTSIANTIS. (2007). Supervised Machine Learning: A review of classification Techniques. . *In Emerging Artificial Intelligence Applications in computer Engineering*, 249–268.

[2]En estadística la prueba de Friedman es una prueba no paramétrica desarrollado por el economista Milton Friedman.

[3] En estadística, una prueba t de Student o Test-T es cualquier prueba en la que el estadístico utilizado tiene una distribución t de Student.