33

Fecha de presentación: octubre, 2018
Fecha de aceptación: diciembre, 2018
Fecha de publicación: febrero, 2019

DESCUBRIMIENTO DE REGLAS

DE CLASIFICACIÓN PARA ESTUDIANTES QUE SE INSCRIBEN DEL BACHILLERATO A CARRERAS UNIVERSITARIAS

DISCOVERY OF RULES OF CLASSIFICATION FOR STUDENTS WHO REGISTER FROM BACHELOSHIP TO UNIVERSITY CAREERS

Jorge Guanín[1]
E-mail: jorgeguanin@uteq.edu.ec
ORCID: http://orcid.org/0000-0001-9150-4009
Raúl Díaz[1]
E-mail: rdiaz@uteq.edu.ec
Byron Oviedo[1]
E-mail: boviedo@uteq.edu.ec
[1] Universidad Técnica Estatal de Quevedo. Ecuador.

Cita sugerida (APA, sexta edición)

Guanín, J., Díaz, R., & Oviedo, B. (2019). Descubrimiento de Reglas de clasificación para estudiantes que se inscriben del bachillerato a carreras universitarias. ***Universidad y Sociedad, 11(2)***, 220-226. Recuperado de http://rus.ucf.edu.cu/index.php/rus

RESUMEN

El uso de datos que extraen técnicas que usan supervisado proporciona el conocimiento para arreglar las opciones que ellos hacen sus directores (gerentes) académicos. Este papel (periódico) habla de los resultados de los algoritmos de software de QUILLA, los algoritmos son usados 14 extraído incluyendo las reglas que predicen a aspirantes de instituto en el éxito o el fracaso del curso de nivelación.

Palabras clave:

Minería de datos, técnicas supervisadas, algoritmos, conocimiento, técnicas de muestreo, reglas.

ABSTRACT

The application of data mining techniques using supervised techniques provide knowledge to arrange the choices that make the academic managers. This paper discusses the results of the KEEL software algorithms, that are used 14 extracted, including rules that predict high school applicants in the success or failure of the levelling course.

Keywords:

Data mining, supervised techniques, algorithms, knowledge, sampling techniques, rules.

INTRODUCCION

La minería de datos educativos EDM (siglas en inglés) (Holte, 1993) 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.

En Ecuador actualmente el ingreso a las universidades se lo está efectuando a través de un examen de selectividad que lo establece la Secretaria Nacional de Ciencia y Tecnología (SENESCYT) y es coordinado por el Sistema Nacional de Nivelación y Admisión (SNNA), entes que programan calendarios académicos, centros de recepción de exámenes, etc. Para tal efecto los exámenes son elaborados por la empresa Educational Testing Service (Quinlan, 2006) entidad de reconocimiento mundial por más de 60 años en la elaboración de todo tipo de evaluaciones.

La preparación académica de los estudiantes del bachillerato para optar por un cupo a la enseñanza de tercer nivel hasta ahora son exigentes debido a las disposiciones del Gobierno de turno cuyo objetivo es el garantizar la igualdad de oportunidades, la meritocracia, transparencia y acceso a la Educación Superior (Vilardi, 2011) los datos que se toman para esta investigación provienen de los resultados de los exámenes aplicados a los aspirantes de las carreras que oferta la universidad.

La naturaleza de los datos se fundamenta en el entorno de estudio que ha mantenido el bachiller desde la secundaria hasta optar por una carrera universitaria, en este sentido las cualidades académicas y demás información es limitada.

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. Llama la atención que teniendo información de los estudiantes no se han realizado estudios para el apoyo de la Gestión Académica a efectos de brindar la ayuda necesaria y oportuna a los bachilleres aspirantes para que tengan mayor superación en sus estudios con el fin de acceder a las carreras universitarias.

MATERIALES Y MÉTODOS

El Sistema Administrador de Base de Datos (DBMS-siglas en inglés) que dispone la Universidad donde se alojan los datos es la plataforma Microsoft SQL Server 2008, en consecuencia se los extrae a través de un procedimiento almacenado con las características de datos descritos en la sección de anexos A1 se detallan los atributos.

Proceso de descubrimiento del conocimiento

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining) estructura el ciclo de vida de un proyecto en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo de la investigación (Chellatamilan, Ravichandran, Suresh & Kulanthaivel, 2011).

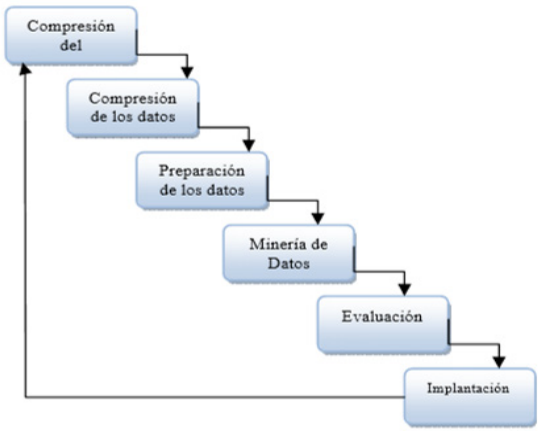

Figura 1. Fases del KDD según el modelo iterativo CRISP-DM.

Descripción de los datos

La información que servirá para el objeto de estudio de este trabajo corresponde a los bachilleres que se inscribieron en el periodo 2011.

En la actualidad existen diferentes programas que se utilizan para la minería de datos WEKA es un programa con una colección de algoritmos de aprendizaje automático para la extracción de conocimiento, es un software de código abierto publicado bajo la licencia GNU (General Public License)

También se puede citar a otro sistema que hace tareas similares a WEKA pero que es mucho más potente y dispone de varios algoritmos que el programa anteriormente mencionado no posee.

KEEL es una herramienta de software para evaluar los algoritmos evolutivos para problemas de minería de datos incluidos de regresión, la clasificación de la agrupación, la explotación sistemática y así sucesivamente. Contiene una gran colección de algoritmos de extracción clásica del conocimiento, las técnicas de pre-procesamiento, Inteligencia Computacional algoritmos de aprendizaje, incluyendo la evolución, incluyendo algoritmos de regla de aprendizaje basadas en enfoques diferentes (Pittsburgh, Michigan e IRL), e híbridos como los modelos de sistemas difusos genéticos, la evolución de redes neuronales, etc. Nos permite realizar un análisis completo de cualquier modelo de aprendizaje en comparación con los existentes, incluyendo un módulo de prueba estadística para la comparación (Kotsiantis, 2007). Sin embargo, otros programas tratan la extracción de conocimientos con la aplicación de algoritmos similares a los anteriormente descritos pero estos no son distribuciones libres, se cita a los otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. ***Conference'10***, Month 1–2, 2010, City, State, Country. Copyright 2010 ACM 1-58113-000-0/00/0010 15.00.

http://www.cs.waikato.ac.nz/ml/weka/

http://www.keel.es/siguientes: SPSS, Clementine, Oracle Data Mining, KnowledgeSTUDIO, RapidMiner.

Técnicas clasificación supervisadas

Tabla 1. Algoritmos para el experimento.

Ítem	Algoritmo	Enfoque	Referencia
1	OneR	Crisp Rule Learning.	[8]
2	C45Rules	Crisp Rule Learning.	[9]
3	PART	Crisp Rule Learning.	[10]
4	Ripper	Crisp Rule Learning.	[11]
5	Slipper	Crisp Rule Learning.	[12]
6	CN2	Crisp Rule Learning.	[13]
7	CART	Decision Trees	[14]
8	ID3	Decision Trees	[15]
9	DT_GA	Decision Trees	[16]
10	TARGET	Decision Trees	[17]
11	REPSO	Evolutionary crisp rule learning	[18]
12	SIA	Evolutionary crisp rule learning	[19]
13	DMEL	Evolutionary crisp rule learning	[20]
14	COGIN	Evolutionary crisp rule learning	[21]

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 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), para recopilar resultados lo suficientemente confiables nos valemos

de la técnica de sobre-muestreo (oversampling) y sub-muestreo (undersampling).

Según Yu-Chung, el problema de la clasificación cuando una clase tiene una probabilidad mucho más baja en el conjunto de entrenamiento se llama el problema conjunto de datos desequilibrado. Un método popular para resolver el problema del conjunto de datos desequilibrado es volver a muestrear el conjunto de entrenamiento. Sin embargo, pocos estudios en el pasado han considerado el remuestreo (resampling) de algoritmos en los conjuntos de datos con alta dimensionalidad (Witten & Eibe, 2000).

En este último, se comparan los resultados de nuestros ensayos y descubrir que, mientras que la mejor técnica a utilizar es a menudo dependiente conjunto de datos, no todas tienden a desempeñarse consistentemente cuando se combina con ciertos algoritmos.

Conjunto de datos para el entrenamiento de algoritmos

Los datos que son extraídos a través del script cuentan con 15 atributos y 1126 instancias, la clase posee 1053 elementos que son seleccionados como APROBADO y 73 que REPRUEBAN (Suspenden o pierden el curso). Previo al entrenamiento y testeo de los algoritmos se realizó una selección de atributos para optar por los más relevantes con el fin de obtener un mejor conjunto de reglas que muestren nitidez y sean legibles, para tal efecto, la ejecución de los algoritmos con lo anteriormente descrito logrará que tengan un mejor desempeño al entrenar y probar el modelo de reglas adquirido y al mismo tiempo que logrará una mayor precisión de sus resultados.

En la Figura 2 se muestra el conjunto de datos en su forma original (Desbalanceado) por lo que presentado este escenario se aplican las técnicas de balanceo "*OverSamplig*" y "*UnderSampling";* la aplicación de esta técnica de balanceo es para evitar que los algoritmos generen las reglas de forma sesgada.

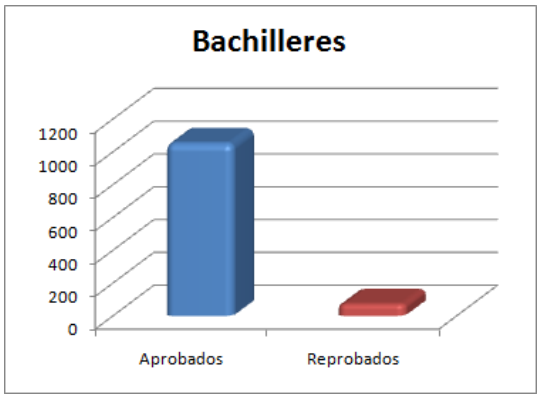

Figura 2. Total de bachilleres que aprueban y reprueban.

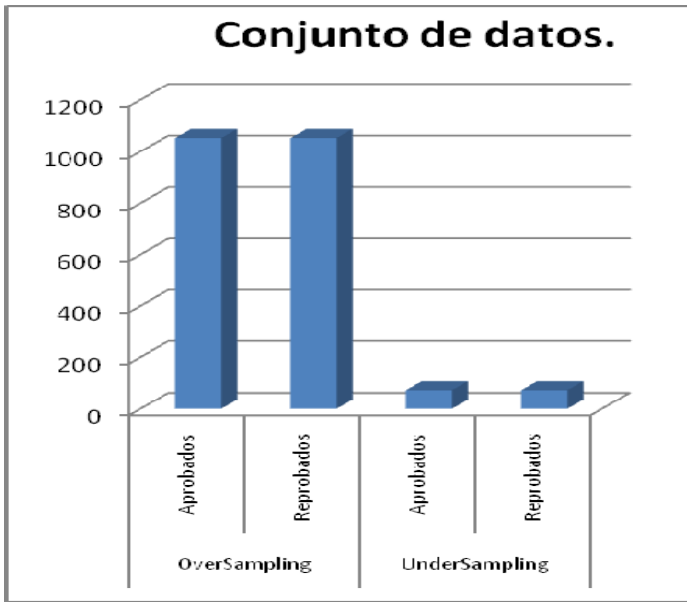

Figura 3. Conjunto de datos balanceados.

Tabla 2. Resultados del Sub-muestreo (Undersampling).

ítem	Algoritmo	#Reg	% Ac. Trn	% Ac. Tst
1	OneR	9	65.24 ± 0.0082	60.28 ± 0.0851
2	C45Rules	9	60.88 ± 0.0341	59.71 ± 0.0830
3	PART	4	58.90 ± 0.0371	53.47 ± 0.0525
4	Ripper	24	71.16 ± 0.0405	56.09 ± 0.1414
5	Slipper	37	76.94 ± 0.0216	52.14 ± 0.0871
6	CN2	25	79.21 ± 0.0431	52.61 ± 0.0928
7	CART	90	72.44 ± 0.0242	61.66 ± 0.0939
8	ID3	50	100.00 ± 0.00	73.90 ± 0.1160
9	DT_GA	18	66.74 ± 0.0242	63.85 ± 0.1061
10	TARGET	2	61.94 ± 0.0079	60.33 ± 0.0689
11	REPSO	3	56.61 ± 0.0257	48.71 ± 0.0617
12	SIA	60	77.55 ± 0.0229	50.09 ± 0.0972
13	DMEL	27	71.45 ± 0.0382	62.95 ± 0.1392
14	COGIN	48	76.70 ± 0.0339	52.09 ± 0.1594

Tabla 3. Resultados del Sobre-muestreo (Oversampling).

ítem	Algoritmo	#Reg	% Ac. Trn	% Ac. Tst
1	OneR	5	58.34 ± 0.0036	57.12 ± 0.0253
2	C45Rules	115	62.60 ± 0.0113	62.06 ± 0.0373
3	PART	4	56.50 ± 0.027	55.93 ± 0.0226
4	Ripper	80	75.77 ± 0.0614	72.64 ± 0.0781
5	Slipper	75	84.18 ± 0.0094	81.72 ± 0.0313
6	CN2	33	65.28 ± 0.0097	63.86 ± 0.0299
7	CART	90	58.31 ± 0.0000	58.76 ± 0.0000
8	ID3	165	100.00 ± 0.00	98.66 ± 0.0096
9	DT_GA	375	85.88 ± 0.0052	83.19 ± 0.0195
10	TARGET	8	61.94 ± 0.0079	60.33 ± 0.0689
11	REPSO	3	50.78 ± 0.0092	49.90 ± 0.0088
12	SIA	149	53.77 ± 0.0120	52.04 ± 0.0117
13	DMEL	46	56.13 ± 0.0070	55.65 ± 0.0175
14	COGIN	3	50.78 ± 0.0092	49.90 ± 0.0088

Aplicó la prueba de Friedman11 para determinar si existe solapamiento entre ellos.

Tabla 4. Estadísticos de contraste.

Estadísticos	valor
N	28
Chi-cuadrado	70.500
Gl	3
Sig. asintót.	0.000
a. Prueba de Friedman	

Habiéndose determinado en el método estadístico antes mencionado que no existían diferencias significativas entre los algoritmos que se propusieron ya que su valor es 0.000. A continuación, se utiliza la prueba de Tstudent (Cohen, 1999) para que determine estadísticamente cuál de las técnicas de muestreo utilizada en el conjunto de datos tiene mejores resultados. En la tabla que a continuación se muestra se detalla el informe del paquete estadístico.

Tabla 5. Estadísticos de grupos entre las técnicas de muestreo.

	Técnica_ Muestreo	N	Medi a	Desv. Std.	Error típ. de la media
reglas	UnderSam-pling	14	29.00	25.637	6,852
	OverSumpling	14	82.21	100.805	26.941
Desv. Std	UnderSam-pling	14	0.098	0.031	0.008
	OverSumpling	14	0.026	0.022	0.006
precisió n UnderSampling		14	57.70	6.842	1.828
	OverSumpling	14	64.41	14.473	3.868

Al hacer énfasis a la ***Tabla 5*** como resultado de la prueba Tstudent notamos que entre las dos técnicas de muestro y los índices que tomados en consideración para el análisis nos revela que la técnica de "***Undersampling***" tiene resultados inferiores frente a los obtenidos por la otra técnica, para tal efecto 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 del número de reglas de clasificación o modelo, está relacionado con el tamaño del conjunto de datos utilizado.

Para el análisis de resultados se cuenta con la ayuda del software SPSS.10 La información de las dos tablas anteriores ***Tabla 2 y Tabla 3***, muestran la precisión y desviación estándar tanto del

grupo de entrenamiento (% Ac.Trn) y las de prueba (% Ac. Tst). al resultado de los algoritmos propuestos en este trabajo.

CONCLUSIONES

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. A futuro y con el objeto de mejorar los resultados se recolectará información concreta para usar técnicas de análisis inteligente personalizadas.

REFERENCIAS BIBLIOGRAFICAS

Chellatamilan, T., Ravichandran, M., Suresh, R. M., & Kulanthaivel, G. (2011). Effect of Mining educational data to improve Adaptatino of learning in e-learning. Second International Conference on Sustentainable Energy and Intelligent System, India.

Cohen, W. W. (1999). A Simple, Fast, and Effective Rule Learner. Sixteenth National Conference on Artificial Intelligence. Orlando.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. Machine Learning Journal, 11, 63-91. Recuperado de https://www.mlpack.org/papers/ds.pdf

Kotsiantis, S. B. (2007). Supervised Machine Learning: A review of classification Techniques. Emerging Artificial Intelligence Applications in computer Engineering, 31(3), 249–268. Ecuperado de https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

Quinlan, J. R. (2006). MDL and categorical Theories (Continued). Recuperado de https://pdfs.semanticscholar.org/cb94/e3d981a5e1901793c6bfedd93ce9cc07885d.pdf

Vilardi, C., et al. (2011). A data mining approach to guide students through the enrollment process based on academic performance. Berlin: Springer.

Witten, I. H., & Eibe, F. (2000). Data Mining: Practical Machine Learning tolls and techniques with java implementations. San Francisco: Morgan Kaufmann.

Zang, H., Jiang, L., & Su, J. (2005). Haidden naive Bayes. American Association for artificial Inteligent - AAAI, 919-924. Recuperado de https://www.aaai.org/Papers/AAAI/2005/AAAI05-145.pdf

ANEXOS

Anexo 1. Tipos de Variable.

Variable	Tipo	Descripción	Valores
Sexo	categórica	Define el sexo del estudiante.	F=Femenino, M=Masculino
Sostenimiento	Categórica	Sostenimiento económico y administrativo del centro educativo.	1=Fiscal 2=Particular 3=Otro
A_graduacion	Numérica.	Número de años que tiene el estudiante desde que se gradúa hasta que se inscribe como aspirante.	0…25
Nota_graduacion	Numérica.	Puntaje final obtenido en el bachillerato.	12…20
Localización_colegio	Categórica.	Describe la zona donde se encuentra situado el colegio de procedencia del estudiante.	LOCAL FUERA_QUEVEDO OTRO
Edad.	Numérica.	Edad con la que cuenta el estudiante al momento de inscribirse como aspirante.	15...99
trabaja	Categórica	Establece la dependencia laboral del estudiante.	N=No S=Si
Pregunta 17	Categórica	¿Sin considerar las horas que asistió al Colegio, Cuántas horas de estudio emplea al día?	0= No tengo tiempo 1=1 2=2 3=3 4=4 5=5 6=6 7=7 8=8 9=9
Pregunta 18	Categórica	¿Por lo general usted estudia en?	1=En dormitorio 2=Cuarto de estudio 3=Sala de mi casa 4=En cualquier Lugar

Pregunta 19	Categórica	¿Con quién estudia?	1=Grupo de compañeros 2=Con profesor =Particular 3=Solo con apuntes de libros 4=Cursos previo al ingreso 5=No me he preparado
Pregunta 20	Categórica	¿Considera usted que dispone de tiempo suficiente para estudiar?	1=Si 2=No 3=A veces
Pregunta 21	Categórica	¿Las personas con quién vive se interesan que usted alcance un buen desempeño en la Universidad?	1=Si 2=No 3=A veces
Pregunta 22	Categórica	¿Durante el bachillerato, como era su ritmo de estudio?	1=Estudiaba de forma continua 2=Estudiaba ocasionalmente 3=Estudiaba solo para las evaluaciones 4=No estudiaba

Estilo	Categórica	Estilo de aprendizaje del estudiante	Auditivo Lectura Visual Quinésico
observación	Categórica		Aprobado Reprobado

Anexo 2. Discretización de valores de los atributos año graduación, nota graduación y edad.

corte	Año graduación
0	1.90
1	3.80
2	5.69
3	7.60
4	9.50
5	11.40
6	13.30
7	15.20
8	17.10

corte	nota graduación
0	12.80
1	13.60
2	14.40
3	15.20
4	16.00
5	16.8
6	17.60
7	18.40
8	19.20

corte	Edad ingreso
0	18.20
1	20.40
2	22.59
3	24.79
4	26.99
5	29.19
6	31.39
7	33.59
8	35.80