Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico

A. G. Hernández, R. A. Meléndez, L.A. Morales, A. García, J. L. Tecpanecatl and I. Algredo

***Abstract*— In this paper, a comparative study of predicting student dropout risks at the ITSM-Mexico (Instituto Tecnologico Superior de Misantla-Mexico) using SQL server is presented. This system uses the personal and academic information from the students at the ITSM. The comparative study is using four algorithms: logistic regression, clustering, decision trees and neuronal network, these consider the information in the database of the control school system of the institute. The results show that the logistic regression algorithm has a good agreement with experimental results.**

***Keywords*— logistic regression, clustering, decision trees, neuronal network, student dropout risks.**

I. INTRODUCCIÓN

EN México y en otros países, el índice de deserción en el sector educativo es un fenómeno que afecta a la población estudiantil en general. Los efectos que tienen estos fenómenos traen consigo una disminución en la eficiencia terminal y por lo tanto aumenta el rezago educativo nacional, que finalmente se puede convertir en un problema social y económico. El Instituto Nacional para la Evaluación de la Educación (INEE) en México advirtió que la deserción escolar aumenta el desempleo. La Organización para la Cooperación y el Desarrollo Económico (OCDE) dejó en claro que México ocupó el primer lugar en el número de desertores escolares de 15 a 18 años. La Secretaria de Educación Pública (SEP) señaló recientemente, que la deserción escolar en México provoca pérdidas de más de 34 millones de pesos al año, por el más de un millón de estudiantes que abandonaron sus estudios en los diferentes niveles de educación [24,29]. El desarrollo de un sistema automatizado para detectar a los alumnos con un alto porcentaje de abandonar sus estudios a nivel superior, no existe como tal [1-5]. Sin embargo, este es un parámetro importante para cualquier programa educativo a nivel superior que desee una certificación. Por tal razón, autoridades de las instituciones de educación superior han implementado estrategias que tienen como objetivo disminuir el índice de deserción en los alumnos de todos los niveles y ciclos escolares. Estas estrategias se realizan con base a la información que tiene el departamento de tutorías, ya que este último es el encargado de reducir los indicies de deserción de cualquier institución de educación superior, además de poder guiar al alumno y de esta forma superar cualquiera dificultad y así poder cumplir con su objetivo de concluir sus estudios. Sin embargo, el departamento de tutorías, con base a las herramientas que utiliza, obtiene la información del alumno de manera tardía, es decir que el departamento se entera de la deserción una vez que el alumno ya decidió abandonar sus estudios, por alguna razón que a veces se desconoce y ya solo se tiene la oportunidad de aplicar una estrategia correctiva, como ofrecer una beca alimenticia, de transporte, etc. Por lo que, con esto método es imposible aplicar alguna estrategia preventiva. De ahí la necesidad de un sistema automatizado o semi-automatizado que nos proporcione la probabilidad de que un alumno tenga un alto porcentaje de deserción de manera temprana, y así poder aplicar una estrategia preventiva para evitar su deserción.

Algunos esfuerzos, en la realización de sistemas automatizados, han sido llevados a cabo para detectar de manera temprana las altas probabilidades que tiene un alumno de desertar de sus estudios, tal como se muestra en [6-7]. En la referencia [6] se muestran los estudios más recientes sobre este tema, sin embargo lo único que desea en este artículo es demostrar que puede obtener un sistema para los alumnos que tienden a desertar. En la referencia [7], se lleva a cabo un estudio analítico, sin embargo es para una población muy pequeña. Ambas investigaciones utilizan técnicas de minería de datos para encontrar algunas predicciones, éstas se basan principalmente en variables, que son a simple vista las que más influyen en el desempeño de los alumnos. Aun más, en la referencia [9] se ha llevado a cabo un análisis completo, del 2002-2014, del uso de minería de datos para la predicción en los estudiantes. Así como también, en la referencia [10] se ha realizado una revisión de los algoritmos de agrupamiento de datos aplicados en minería de datos educativa. La mayoría de los resultados de estas investigaciones han sido publicados en la conferencia de IEEE-Frontiers in Education Conference, por ejemplo un análisis para estudiantes de ingeniería se muestran en las referencias [5-8]. Todos estos esfuerzos han sido exitosos, más aun en el área de matemáticas, sin embargo un sistema que tome como base la información de los alumnos que está concentrada en los departamentos de información de la institución y que por sí misma nos informe, qué estudiantes tienen altas probabilidades de desertar de sus estadios, no se ha realizado hasta ahora. En la Tabla I se mencionan los trabajos ya reportados, sobre este tema. Aunque, los 3 primeros se basan principalmente en la experiencia que ya se tiene de casos anteriores, mientras los 3 últimos utilizan un modelo predictivo, tanto para evaluar el desempeño de los alumnos como su posible deserción. Sin embargo, ninguno de ellos muestra un perfil de deserción y más aún, este perfil nos puede aportar información relevante de un grupo de estudiantes y así poder utilizarlo para las siguientes generaciones de los planteles educativos.

A. G. Hernandez, Instituto Tecnológico Superior de Misantla, México, gamaliel_hg@hotmail.com

R. A. Melendez, Instituto Tecnológico Superior de Misantla, México, ramelendeza@itsm.edu.mx

L.A. Morales, Universidad Michoacana de San Nicolás de Hidalgo, México, lamorales@conacyt.mx

A. Garcia, Universidad Autónoma de San Luís Potosí México, abel.garcia@uaslp.mx

J. L. Tecpanecatl, Universidad Autónoma de San Luís Potosí México, jluis@ieee.org.

I. Algredo, Universidad Politécnica de Tlaxcala, México, ignacio.algredo@uptlax.edu.mx

En el presente trabajo de investigación se utiliza la información personal y académica de los alumnos, la cual existente en la base de datos del sistema de información escolar (SIE) del mismo instituto. El sistema se desarrolló con la herramienta SQL Server Data Tools de Microsoft Visual Studio 2012. Además, este mismo sistema utiliza la minería de datos y realiza un estudio comparativo utilizando cuatro algoritmos para realizar la predicción: el algoritmo de regresión logística, de clústeres de Microsoft, de árboles de decisión, y por último el algoritmo basado en redes neuronales. Los resultados indican que el algoritmo de regresión logística es el más adecuado para realizar la predicción y más aún para definir un perfil en los alumnos con alto índice de deserción.

TABLA I
TRABAJOS REPORTADOS SOBRE MODELOS PARA EVALUAR EL DESEMPEÑO DEL ESTUDIANTE CON LA POSIBILIDAD DE DESERCIÓN

Algoritmo	Aplicado	Referencia
Acciones	Deserción y Evaluación	[2]
Modelo Cuantitativo	Deserción y Evaluación,	[3]
Difuso	Deserción	[4]
Difuso	Evaluación	[5]
Modelo Predictivo	Deserción y Evaluación	[6]
Modelo Predictivo	Deserción	[7]
Modelo Predictivo	Evaluación	[8]

II. DESARROLLO DEL SISTEMA

Para llevar a cabo el desarrollo de un sistema predictivo que permita detectar al alumno con alta probabilidad de deserción y más aún proporcionar el perfil de los alumnos desertores, utilizando información personal y académica de alumnos almacenada en el SIE del ITSM. La solución propuesta consiste en cinco fases: la primera fase consiste en migrar los datos del sistema SIE a un formato basado en Microsoft Excel, también se transforman los datos y se migran al servidor SQL Server 2012. En la segunda fase se seleccionan las variables de entrada al sistema predictivo para preparar un origen de datos con un formato adecuado para el proceso de minería de datos. En la tercera fase, se crea la estructura del sistema de minería de datos y se agregan los sistemas de minería a utilizar, los cuales son validados mediante la matriz de clasificación, el gráfico de elevación y la validación cruzada de cada sistema. En la cuarta fase, se aplica el sistema de minería de datos basado en el algoritmo de regresión logística de Microsoft para predecir la deserción, y para obtener el perfil de deserción escolar. Por último, se adicionó una quinta fase que aunque ésta se encuentra fuera del área de investigación del presente trabajo se menciona ya que con la información proporcionada por el sistema la institución podrá emplear acciones preventivas o correctivas para evitar la deserción, es decir el sistema puede tomar una decisión y sugerir que acción debe ser implementada, tal como se muestra en la Fig. 1. Para llevar a cabo el desarrollo del sistema se utilizó Microsoft SQL Server Analysis Services, las cuales proporciona las siguientes herramientas que se pueden utilizar para crear soluciones de minería de datos: 1) El asistente para minería de datos de SQL Server Data Tools (SSDT) facilita la creación de estructuras y de sistemas de minería de datos, usando orígenes de datos relacionales o datos multidimensionales en cubos. En el asistente, se eligen los datos que desee utilizar y, a continuación se aplican técnicas de minería de datos específicas, como agrupación en clústeres, redes neurales o modelado de series temporales. 2) SQL Server Management Studio y SQL Server Data Tools disponen de visores de sistemas para explorar los sistemas de minería de datos una vez creados. Es posible examinar los sistemas mediante visores adaptados a cada algoritmo o analizar con mayor profundidad utilizando el visor de contenido del sistema. 3) El generador de consultas de predicción, se proporciona en SQL Server Management Studio y SQL Server Data Tools para ayudarle a crear consultas de predicción. También se puede probar la exactitud de los sistemas respecto a un conjunto de datos de exclusión o datos externos, o utilizar validación cruzada para evaluar la calidad del conjunto de datos. 4) SQL Server Management Studio es la interfaz en la que administra las soluciones de minería de datos implementadas en una instancia de Analysis Services. Además, es posible procesar nuevamente las estructuras y el sistema para actualizar los datos que contienen. 5) SQL Server Integration Services contiene herramientas para limpiar datos, automatizar tareas como la creación de predicciones y actualización de sistemas y para crear soluciones de minería de datos.

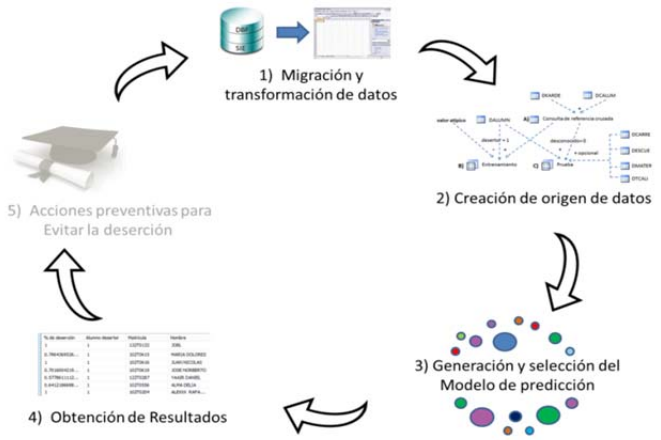

Figura 1. Diagrama esquemático para el desarrollo del sistema propuesto.

Una vez obtenida la metodología para llevar a cabo el sistema, se llevó a cabo una comparación de los algoritmos más usados para estudiar la deserción de los estudiantes, los cuales se describen a continuación:

a) Regresión Logística

La regresión logística [9-11] es una técnica estadística ya muy bien conocida que se usa para modelar los resultados de encuestas. Existen varias implementaciones de regresión logística en la investigación estadística, que utilizan diferentes técnicas de aprendizaje. El algoritmo de regresión logística de Microsoft es una variación del algoritmo de red neuronal de Microsoft. Este algoritmo comparte muchas de las cualidades de las redes neurales, el cual tienen una característica importante, la cual es fácil su entrenamiento. Una de las

ventajas de la regresión logística es que el algoritmo es muy flexible, pues puede tomar cualquier tipo de entrada y admite varias tareas analíticas diferentes. La regresión logística analiza datos distribuidos binomialmente de la forma:

$$Y_i \sim B(p_{i,} n_i),\ para\ i = 1, \dots, m, \quad (1)$$

donde los números de ensayos Bernoulli n_i son conocidos y las probabilidades de éxito p_i son desconocidas. El sistema es entonces obtenido con base al resultado de cada ensayo (valor de *i*) y el conjunto de variables explicativas/independientes puedan informar acerca de la probabilidad final. De esta manera se plantea el algoritmo de regresión logística, el cual tiene infinidad de aplicaciones en tecnologías modernas, ya que nos ayudan a identificar y predecir el resultado de una variable categórica.

b) Algoritmo de Clústeres

El algoritmo de clústeres de Microsoft [15, 18, 25-27] es un algoritmo de segmentación suministrado por Analysis Services. El algoritmo utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares [25]. El algoritmo de clústeres se diferencia de otros algoritmos de minería de datos, como el algoritmo de árboles de decisión de Microsoft, en no tener que designar una columna de predicción para generar un sistema de agrupación en clústeres. El algoritmo de clústeres entrena al sistema de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo. Una característica principal de este algoritmo es que necesita una columna de predicción opcional; el algoritmo no necesita una columna de predicción para generar el sistema, pero puede agregar una columna de predicción de casi cualquier tipo de datos. Los valores de la columna de predicción se pueden tratar como entradas del sistema de agrupación en clústeres, o se puede especificar que solo se utilicen para las predicciones.

c) Árboles de Decisión

El algoritmo de árboles de decisión [12-17,19-24] de Microsoft es un algoritmo de clasificación y regresión proporcionado por Microsoft SQL Server Analysis Services para el modelado de predicción de atributos discretos y continuos. Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos, utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. El árbol de decisión realiza predicciones basándose en la tendencia hacia un resultado concreto. Los requisitos para un sistema de árboles de decisión son los siguientes: 1) Una columna key; cada sistema debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas, 2) Una columna de predicción; se requiere al menos una columna de predicción. Puede incluir varios atributos de predicción en un sistema y pueden ser de tipos diferentes, numérico o discreto. Sin embargo, el incremento del número de atributos de predicción puede aumentar el tiempo de procesamiento. 3) Columnas de entrada; se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entrada afecta el tiempo de procesamiento, y esto se notó en las pruebas o ejecuciones que se realizaron para encontrar los estudiantes con alto porcentaje de deserción.

d) Red Neuronal de Microsoft

En SQL Server Analysis Services, el algoritmo de red neuronal [28-32] de Microsoft combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción, y usa los datos de entrenamiento para calcular las probabilidades. Posteriormente, puede usar estas probabilidades para la clasificación o la regresión, así como para predecir un resultado del atributo de predicción basándose en los atributos de entrada. Los modelos de minería de datos construidos con el algoritmo de red neuronal de Microsoft pueden contener varias redes, en función del número de columnas que se utilizan para la entrada y la predicción, o solo para la predicción. Para nuestro caso, los parámetros definieron la forma como se muestran los datos, la manera como se distribuyen o como se espera que estén distribuidos en cada columna, y cuando se invoca la selección de características para limitar los valores usados en el modelo final.

III. RESULTADOS

En esta sección se muestran los resultados después de aplicar los algoritmos descritos en la sección 2. Los resultados muestran que sí es posible predecir a los alumnos que tienen altas posibilidades de desertar de sus estudios a nivel superior. En nuestro caso, fue un análisis sobre los alumnos del programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones del Instituto Tecnológico Superior de Misantla, en el estado de Veracruz, donde después de un análisis de 134 alumnos, aproximadamente, y utilizando la información de la base de datos del mismo instituto y aplicando los algoritmos descritos anteriormente, se logró determinar los posibles alumnos desertores, y más aún obtener un perfil de deserción de los mismos estudiantes desertores. La tarea de identificar alumnos desertores durante el proceso de su formación académica no es tan fácil de realizarlo manualmente, esto es debido a que el proceso involucra una cantidad importante de información y por tal razón es necesario un sistema automatizado, el cual ha sido llevado a cabo en este trabajo de investigación. Los algoritmos empleados deben de tomar decisiones, para esto se consideraron las siguientes métricas de clasificación. Una matriz de clasificación (Microsoft, Matriz de clasificación (Analysis Services - Minería de datos)) ordena todos los casos del sistema en categorías, determinando si el valor de predicción coincide con el valor real, tal como se muestra en la Tabla II, donde TP es verdadero positivo, FP falso positivo, TN verdadero negativo y FN falso negativo. La matriz de clasificación es una herramienta estándar de evaluación de sistemas estadísticos a la que se denomina matriz de confusión.

Además, seleccionar el sistema de minería de datos a utilizar no es una tarea trivial, pues implica un análisis exhaustivo con la finalidad de encontrar el sistema que mejor clasifica, con la más alta probabilidad de predicción, con menores errores y que se acerque lo más posible al sistema ideal. Así que, el sistema se debe de determinar considerando los resultados arrojados por la matriz de confusión, el gráfico de elevación y la validación cruzada, la precisión y la sensibilidad.

TABLA II
ESQUEMA DE MATRIZ DE CLASIFICACIÓN

		Clase Actual	
Categorías		0	1
Clase Hipotética	0	TN	FN
	1	FP	TP
Columnas Totales		N=FP+TN	P=TP+FN

La matriz de clasificación nos sirve para validar los algoritmos, ya que estos entregan información importante sobre los resultados que arrogan a la aplicación de los algoritmos con la base de datos del sistema de información, en este caso del programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones del ITSM. La Tabla III muestra la matriz de clasificación para cada algoritmo aplicado, la de clúster EM (Esperanza-Maximización) escalable, la de árbol de decisión, regresión logística y red neuronal, con los valores reales y sus porcentajes correspondientes.

TABLA III
MATRIZ DE CLASIFICACIÓN PARA CADA ALGORITMO

Clúster EM escalable				
	0 (inscrito Real)	1 (desertor Real)	0 (inscrito Real)	1 (desertor Real)
0 (inscrito)	32	2	100.00%	25.00%
1 (desertor)	0	6	0.00%	75.00%
correctas	32	6	100%	75%
incorrectas	0	2	0%	25%
Árbol de Decisión				
	0 (inscrito Real)	1 (desertor Real)	0 (inscrito Real)	1 (desertor Real)
0 (inscrito)	32	8	100.00%	100.00%
1 (desertor)	0	0	0.00%	0.00%
correctas	32	0	100%	0%
incorrectas	0	8	0%	100%
Regresión Logística				
	0 (inscrito Real)	1 (desertor Real)	0 (inscrito Real)	1 (desertor Real)
0 (inscrito)	32	0	100.00%	0.00%
1 (desertor)	0	8	0.00%	100.00%
correctas	32	8	100%	100%
incorrectas	0	0	0%	0%
Red Neuronal				
	0 (inscrito Real)	1 (desertor Real)	0 (inscrito Real)	1 (desertor Real)
0 (inscrito)	32	1	100.00%	12.50%
1 (desertor)	0	7	0.00%	87.50%
correctas	32	7	100%	88%
incorrectas	0	1	0%	13%

La Tabla IV presenta las medidas de precisión y sensibilidad para los 4 algoritmos utilizados en el sistema desarrollado, estas medidas están calculadas a partir de las matrices de clasificación, mostradas en la Tabla II. La precisión se refiere a la fracción de ejemplares que se han clasificado como de la clase correspondiente y que, en realidad, son de esa clase. Por otro lado, la sensibilidad se refiere a la fracción de los ejemplos de la clase de todo el conjunto que se clasifican correctamente, es decir, mide la probabilidad de que si un alumno pertenece a una categoría, el sistema lo asigne a esa categoría. Sin embargo existe otra métrica muy común para comparar sistemas medida-F que es una combinación de ambas, y se refiere a la media armónica de precisión y sensibilidad.

TABLA IV
MEDIDAS DE PRECISIÓN Y SENSIBILIDAD PARA CADA ALGORITMO

	Clúster	Árbol de Decisión	Regresión Logística	Red Neuronal
Precisión	100%	0%	100%	100%
Sensibilidad	75%	0%	100%	88%
Exactitud	95%	80%	100%	98%
Medida-F	86%	0%	100%	93%
Elevación	5	0	5	5

En la Tabla IV, claramente se puede observar que cuando el sistema utiliza el algoritmo de regresión logística, tiene el 100% en precisión, sensibilidad, exactitud y la medida-F. Por tal razón, se puede decir que el sistema da mejores resultados utilizando el algoritmo de regresión logística. Así mismo, se puede decir que el sistema desarrollado da buenos resultados si se utiliza el algoritmo de red neuronal ya que tienen los siguientes porcentajes en precisión 100 %, sensibilidad 88%, exactitud 98% y la medida-F 93%.

Además de esto, se realizó la validación cruzada para corroborar que los resultados del sistema no se encuentran sesgados y de esta manera tener una métrica confiable y poder seleccionar el algoritmo a utilizar para este trabajo de investigación, tal como se muestra en la Tabla V. Donde se comparan los 4 algoritmos utilizados en el sistema, donde TP es verdadero positivo, FP falso positivo, TN verdadero negativo, FN falso negativo, LS es la puntuación larga, Lift es la elevación y RMSE es el error cuadrático medio.

TABLA V
RESULTADOS DE LA MATRIZ DE CONFUSIÓN (STANDARD DEVIATION)

	TP	FP	TN	FN	LS	Lift	RMSE
Clúster	0.499	0.294	0.441	0.308	0.098	0.110	0.058
Árbol de Decisión	0.488	0.939	0.937	0.739	0.158	0.114	0.0597
Regresión logística	0.499	0.308	0.447	0.308	0.069	0.076	0.0601
Red Neuronal	0.480	0.466	0.499	0	0.652	0.634	0.0281

En esta investigación lo que nos interesa es que el sistema desarrollado clasifique correctamente a los alumnos desertores, así que el algoritmo de regresión logística arroja los mejores resultados, según la Tabla V. También, utilizando los resultados de la matriz de clasificación se puede obtener la gráfica para medir la precisión del modelo de minería de datos, ver Fig. 2. Ésta compara los valores reales con los

valores de predicción para cada algoritmo de predicción especificado. Este gráfico de elevación se obtuvo de la pestaña gráfico de elevación, la cual compara todos los modelos de minería seleccionados en la estructura de minería de datos seleccionada.

Para los resultados experimentales, se utilizaron la base de datos de los alumnos del ITSM de las generaciones 2010 hasta la 2013. En las primeras tres generaciones, el sistema acertó entre un 95 y 100 % a los alumnos desertores. Para la generación 2013, el sistema aserto el 100%, indicando que dos alumnos serían los desertores, tal como se muestra en la Tabla VI.

Figura 2. Comparación de los valores reales (línea azul) con los valores de predicción para cada algoritmo de predicción especificado, regresión logística (rojo) y la de red neuronal (verde).

Para la generación 2013, se contó con un grupo de 26 estudiantes del programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones del ITSM. Los resultados muestran que dos alumnos tienen el 100% de probabilidad de desertar, los otros 24 están dentro del mismo porcentaje de deserción. En la Tabla VI, se muestra los resultados de los primeros 8 estudiantes analizados.

TABLA VI
RESULTADOS DE LA PROBABILIDAD DE DESERCIÓN

Matricula del Alumno	Edad	% de Deserción
132TO126	22	100%
132TO132	21	100%
112TO360	25	98.55
112TO365	22	98.54
112TO367	22	98.54
102TO181	25	98.54
102TO190	23	98.54
102TO204	23	98.53

Obtener el perfil de deserción de los estudiantes desertores no es una tarea fácil, ya que el fenómeno del abandono escolar es multifactorial, los estudios indican como posibles causas a las condiciones económicas, salud, orientación escolar, lugar de procedencia, asignaturas específicas, entre otros factores. Sin embargo, en el sistema desarrollado aquí, se logró obtener los factores que más influyen en la deserción escolar, que son principalmente el lugar de procedencia y asignaturas en específico, lo cual concuerda con algunos otras investigaciones [7-9].

IV. CONCLUSIONES

El sistema desarrollado en este trabajo de investigación, pretende detectar de forma temprana y oportuna a los alumnos con altos índices de deserción. Para esto, se ha llevado a cabo un análisis y comparación de 4 algoritmos: el de árbol de decisión, regresión logística, el de clústers, y el red neuronal, en conjunto con técnicas de minería de datos y con el único objetivo, el de hallar a los alumnos con un mayor porcentaje de desertar de sus estudios. Sin embargo, los resultados muestran que el algoritmo con mejores resultados son el de regresión logística, con este algoritmo se pueden encontrar los posibles alumnos desertores; también se puede hallar un perfil de los estudiantes con las mismas probabilidades de deserción. Todos estos resultados fueron comparados con resultados reales que tiene el departamento de sistemas de información del Instituto Tecnológico Superior de Misantla. En este caso, solo fue para el programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones del mismo instituto, sin embargo sin ningún problema puede ser implementado para cualquier otro programa del Instituto o cualquier plantel de educación superior de México.

AGRADECIMIENTOS

Este trabajo ha sido realizado gracias al apoyo de PROMEP-México, mediante el apoyo a UASLP-PTC-553 y CONACYT-México, mediante los proyectos con número de financiamiento 169062 y 204419.

REFERENCIAS

[1] Horacio Kuna, Ramón García Martínez And Francisco R. Villatoro, Pattern Discovery In University Students Desertion Based On Data Mining, Advances and Applications in Statistical Sciences, *Proceedings of The IV Meeting on Dynamics of Social and Economic Systems*,Vol. 2, Issue 2, Pages 275-285, 2010.

[2] Guillermo López, María Posada, Claudia Cardozo, Diego José Cuartas, Specific actions for desertion reduction, competence identification and guidance for new students of an engineering program, a case study, *International Congress on Engineering Education* (ICEED), 2010.

[3] Jesus Alfonso Perez Gama, Martha Isabel Rozo Arteaga; Roger Smith Londono Buritica ; Alejandro Marulanda Quinche, Quantitative models and software architecture, facing student Desertion and Permanence, *IEEE International Conference on Teaching, Assessment and Learning for Engineering* (TALE), 2013.

[4] Alfonso Pérez-Gama, Guillermo Hoyos, Leyini Parra-Espitia, Miguel Ortegón, Luis Giovanny Rozo-Pardo, Byron Perez-Gutierrez, Education software architecture: Facing student desertion in Colombia higher education with an intelligent knowledge based coaching system, *IEEE ANDESCON*, 2010

[5] Luis Felipe Zapata Rivera; Jorge Luis Restrepo Ochoa; Jaime L. Barbosa Perez, Improving student results in a statics course using a computer-based training and assessment system, *IEEE Frontiers in Education Conference*, 2013

[6] Sandra Milena Merchan Rubiano; Jorge Alberto Duarte Garcia, Formulation of a predictive model for academic performance based on students academic and demographic data, *IEEE Frontiers in Education Conference (FIE)*, 2015.

[7] John M. Mativo; Shaobo Huang, Prediction of students academic performance: Adapt a methodology of predictive modeling for a small sample size, *IEEE Frontiers in Education Conference (FIE)*, 2014

[8] Patrick D. Schalk; David P. Wick; Peter R. Turner; Michael W. Ramsdell, Predictive assessment of student performance for early

strategic guidance, *IEEE Frontiers in Education Conference (FIE)*, 2011.
[9] Poojar Thakar, Anil Mehta, and Manisha, Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue, *International Journal of Computer Applications*, Vol. 110 – No. 15, 2015.
[10] Ashish Dutt, Saeed Aghabozrgi,Maizatul Akmal Binti Ismal, and Hamidreza Mahroeian, Clustering Algorithms Applied in Educational Data Mining, *International Journal of Information and Electronics Engineering*, Vol. 5, No. 2, March 2015.
[11] Setiono, R. ; Sch. of Comput., Nat. Univ. of Singapore, Singapore ; Azcarraga, A., An effective method for generating multiple linear regression rules from artificial neural networks, *IEEE 13th International Conference on Tools with Artificial Intelligence*, 2001
[12] D. D. Finlay, C. D. Nugent, P. J. McCullagh, N. D. Black, J. A. Lopez, Evaluation of a Statistical Prediction Model Used in the Design of Neural Network Based ECG Classifiers: A Multiple Linear Regreession Approach, *Procc. of the 4th Annual IEEE Conf. on Information Technology Applications in Biomedicine*, UK, 2003
[13] Zhi-cheng Zheng , Xin LIU, Analysis of regional logistics demand prediction based on support vector non-linear multiple regression, *International Conference on Management and Service Science*, 2009.
[14] Amany Abdelhalim, Issa Traore, A New Method for Learning Decision Trees from Rules, *International Conference on Machine Learning and Applications*, 2009
[15] R. S. Michalski and I. F. Imam, "Learning problem-oriented decision structures from decision rules: the AQDT-2 system", In Proceedings of 8th International Symposium Methodologies for Intelligent Systems. *Lecture Notes in Artificial Intelligence*, 869, Springer Verlag, Heidelberg, 1994, pp. 416-426.
[16] Y. Akiba, S. Kaneda, and H. Almuallim, "Turning majority voting classifiers into a single decision tree", *In Proceedings of the 10th IEEE International Conference on Tools with Artificial Intelligence*, 1998, pp. 224-230.
[17] Masaki Kurematsu and Hamido Fujita, A Framework for Integrating a Decision Tree Learning Algorithm and Cluster Analysis, *12th IEEE International Conference on Intelligent Software Methodologies, Tools and Techniques*, Budapest, Hungary, September 22-24, 2013
[18] Saori Amanuma, Masaki Kurematsu and Hamido Fujita, "An Idea of Improvement Decision Tree Learning Using Cluster Analysis ", *The 11th International Conference on Software Methodologies, Tools and Techniques*, pp.351-360, 2012
[19] Quinlan, J. R. "Induction of Decision Trees", *Machine Learning*, Vol.1, No.1,pp.81-106(1986).
[20] G. Bortolan, C. Brohet, S. Fusaro, "Possibilities of using neural networks for ECG classification," *Journal of Electrocardiology*, vol. 162, pp. 10-16, 1996
[21] Sam Chao, Fai Wong, An incremental decision tree learning methodology regarding attributes in medical data mining, *Procc. International Conference on Machine Learning and Cybernetic*s, 2009
[22] P.E. Utgoff, N.C. Berkman, and J.A. Clouse, "Decision Tree Induction Based on Efficient Tree Restructuring", *Machine Learning, Kluwer Academic Publishers*, Vol. 29, pp. 5-44, 1997.
[23] Y.L. Chen, C.L. Hsu, and S.C. Chou, "Constructing a Multi-Valued and Multi-Labeled Decision Tree", *Expert Systems with Applications*, Vol. 25, pp. 199-209, 2003.
[24] I. Kononenko, "Inductive and Bayesian Learning in Medical Diagnosis", Applied Artificial Intelligence, Vol. 7, pp. 317-337, 1993
[25] Shen Bin,Liu Yuan, Wang Xiaoyi, *Research on Data Mining Models for the Internet of Things, International Conference on Image Analysis and Signal Processing* (IASP), 2010
[26] http://www.sems.gob.mx/work/models/sems/Resource/11390/1/images/000_INTRODUCCION_Movimiento_contra_Abandono.pdf
[27] [25] H. Haggag, M. Hossny, S. Haggag, S. Nahavandi, D. Creighton, Efficacy comparison of clustering systems for limb detection*, 9th International Conference on System of Systems Engineering*, 2014
[28] Lei Qiu; Yongqing Zheng; Yuliang Shi; Chengliang Sang, CET: Clustering Extension Table research in multi-tenant database for SaaS applications, *International Conference on Information Science and Technology (ICIST*), 2013
[29] D. Fetterly; M. Manasse; M. Najork, On the evolution of clusters of near-duplicate Web pages, Proceedings. *First Latin American Web Congress*, 2003.
[30] Callejas Ivan; Pineros Juan; Rocha Juan; Hernandez Ferney; Delgado Fabio, Implementación de una red neuronal artificial tipo SOM en una FPGA para la resolución de trayectorias tipo laberinto, *II International Congress of Engineering Mechatronics and Automation (CIIMA),* 2013
[31] J. A. Blakeley, C. Cunningham, N. Ellis, Balaji Rathakrishnan, M. -C. Wu Distributed/heterogeneous query processing in Microsoft SQL, *Proceedings 21st International Conference on Data Engineering (ICDE)* 2005.
[32] Osamu Araki; Kazuyuki Aihara, Dual Information Representation with Stable Firing Rates and Chaotic Spatiotemporal Spike Patterns in a Neural Network Model, *Neural Computation*, Volume: 13, Issue: 12, 2001.

MSc. Arnulfo Gamaliel Hernández González was born in Misantla, Veracruz, México, in 1976. He received the Licenciatura in Informatics and Master in Computer Science degree from the Instituto Tecnológico Superior de Misantla, México, en el 2000, and 2016, respectively. Since 2006, . Since 2006, MSc. Hernandez Gonzalez is an assistant professor at the Information Technologies and Communications Engendering at the Instituto Tecnológico Superior de Misantla. His research efforts are in the mobile computing, internet of things and embebided sytems.

MIA. Roberto Ángel Meléndez Armenta received the B. Eng in Computer Systems from Benemérita Universidad Autónoma de Puebla in 2007 and Master's degree in Artificial Intelligence from Universidad Veracruzana in 2011. Since 2012, he has been full time professor Instituto Tecnológico Superior de Misantla. His scientific interests are focused on security, intelligent computing, mobile computing, parallel and distributed systems for specific applications, and artificial vision.

Dr. Luis Alberto Morales Rosales received the B.Eng in Computer Systems from Technologic Institute of Colima (ITC) in 2002 and the M.Sc and Ph.D degrees in Computer Science from National Institute for Astrophysics, Optics and Electronics (INAOE) in 2005 and 2009, respectively. Since 2010, he is the Head of the Master Program in Computer Systems at the Superior Technological Institute of Misantla. His research efforts focus on security, intelligent computing, mobile computing, parallel and distributed systems for specific applications.

Dr. Abel Garcia Barrientos was born in Tenancingo, Tlaxcala, Mexico, in 1979. He received the Licenciatura degree in Electronics from the Autonomous University of Puebla, Mexico, in 2000, and the M.Sc. and Ph.D. degree in Electronics from the National Institute for Astrophysics, optics, and Electronics (INAOE), Tonantzintla, Puebla, in 2003 and 2006, respectively. Since 2007 he joined as a researcher at the Mechatronics Department at the Polytechnic University of Pachuca, Mexico. In 2009 he was a Post-Doctoral Fellow at the Micro- and Nano-Systems Laboratory at the McMaster University, Ontario, Canada and in 2010 Dr. Garcia-Barrientos was a Post- Doctoral Fellow at the Advanced Materials and Device Analysis group of Institute for Microelectronics, Technische Universitat Wien, too. Since January 2016, Dr. Garcia Barrientos is a full time professor at the Faculty of Sciences at the Universidad Autonoma de San Luis Potosi. He has been member of SNI since 2008, level 1.

Dr. José Luis Tecpanecatl Xihuitl was born in Puebla, Mexico. He received the Licenciatura degree in Electronics from the Autonomous University of Puebla, Mexico, in 1999, the M.Sc. degree in Electronics from the National Institute for Astrophysics, Optics, and Electronics (INAOE), Tonantzintla, Puebla, in 2001 and Ph.D. degree in Computational Engineering from the University of Louisiana at Lafayette, in 2008. Since 2009, Dr. Tecpanecatl Xihuitl is a full time professor at the faculty of science at the Universidad Autonoma de San Luis Potosi.

Dr. Ignacio Algredo-Badillo received the B.Eng in Electronic Engineering from Technologic Institute of Puebla (ITP) in 2002 and the M.Sc and Ph.D degrees in Computer Science from National Institute for Astrophysics, Optics and Electronics (INAOE) in 2004 and 2008, respectively. In 2009, he was professor of Computer Engineering at University of Istmo, and since 2014, he has been professor of Technology Information Engineering at Polytechnic University of Tlaxcala. He has led several projects and he is member of the SNI level 1.