FEBRERO 2016
VOL. 3, NO. 1

revie

Revista de Investigación y Evaluación Educativa

ISSN 2409-1553

ideice
Instituto Dominicano de Evaluación e Investigación de la Calidad Educativa

revie
Revista de Investigación y Evaluación Educativa

Revista Digital de suscripción gratuita del Instituto Dominicano de Evaluación e Investigación de la Calidad Educativa (IDEICE)

Periodicidad Semestral

Edición
Febrero 2016, Vol.3, No. 1

Dirección Ejecutiva
Julio Leonardo Valeirón Ureña

Consejo Editorial
Dinorah de Lima Jiménez
Julián Álvarez Acosta
Luis Camilo Matos De León

Correción de estilos
Ramón Fari Rosario

Coordinación General
Dilcia Armesto Núñez

Diseño y Diagramación
Natasha Mercedes Arias

ISSN: 2409-1553

IDEICE
Ave. César Nicolás Penson No. 30, Gazcue
Santo Domingo, D.N.
Teléfono: +1 (809) 732-7152
www.ideice.gob.do

Santo Domingo, Rep. Dom.

Esta obra está bajo una licencia de Licencia Creative CommonsAtribución-No-Comercial-SinDerivar 4.0 Internacional.

EDITORIAL

La presente edición de **REVIE** contiene cuatro investigaciones, las tres primeras auspiciada por el IDEICE y la cuarta corresponde a un artículo de un reconocido investigador internacional. Con este nuevo número reafirmamos el firme compromiso con la investigación de calidad como soporte científico, aportando evidencias y conocimientos pertinentes en la toma de decisiones en el ámbito de la educación.

Domínguez Ruiz y colaboradoras estudian la tasa de retorno de la educación en la población dominicana entre 18 y 65 años que reciben ingresos por remuneraciones laborales; ofrecen informaciones relevantes para la elaboración de políticas públicas relacionadas a la oferta laboral, el nivel de salarios y la equidad en la distribución del ingreso.

Oscar Amargós en el artículo *Evaluación de resultados e impacto de la política de educación secundaria en República Dominicana*, realiza una comparación entre los egresados de la modalidad general y los titulados de la modalidad técnico profesional, con el propósito de aportar evidencias objetivas para sustentar las decisiones de las autoridades educativas nacionales evaluando las variables que permiten determinar los efectos e imparto de las políticas de educación secundaria en el desarrollo económico y social del país.

El estudio de González y su colaborador sobre *Un modelo predictivo de deserción escolar para la República Dominicana* nos ofrece una importante herramienta para la predicción del riesgo de deserción en los estudiantes de nivel básico y medio del sistema educativo nacional.

Este número concluye con la entrega del investigador Díaz Esteve, profesor de la Universidad de Valencia, del artículo sobre la *Importancia de utilizar la teoría de la respuesta al Item (TRI) en la construcción de pruebas de aptitud y conocimiento* donde nos presenta los fundamentos teóricos y principios básicos sobre los que se ha construido esta teoría y elementos, así como también utiliza los datos obtenidos en la aplicación de una prueba de aptitud donde se pueden visualizar los valores paramétricos de los ítems y su interpretación.

Finalmente, el IDEICE reafirma su vocación investigativa para con ello no solo conocer la realidad educativa, sino propiciar su transformación, estando seguros de hacer presente nuestra total convicción de que las investigaciones y reflexiones presentadas, serán un aporte que nutrirá el conocimiento acerca de la educación y sus procesos.

Julio Leonardo Valeirón Ureña
Director Ejecutivo

CONTENIDO

RENATO R. GONZÁLEZ

r.gonzalez@codetel.net.do

Maestría en Ciencias Económicas de la Pontificia Universidad Católica Madre y Maestra (PUCMM).

FELIPE ANT. LLAUGEL

fllaugel@hotmail.com

Maestría en Economía, Decano de Ingeniería de la Universidad Dominicana O&M.

UN MODELO PREDICTIVO DE DESERCIÓN ESCOLAR PARA LA REPÚBLICA DOMINICANA

RESUMEN

La motivación de la investigación, cuyos resultados se recogen en este artículo, es la de dotar de una herramienta que permita predecir el riesgo de deserción de los estudiantes del nivel básico y medio en el sistema educativo dominicano. Mediante modelos predictivos de minería de datos es posible determinar patrones de comportamiento del alumno(a), analizando la historia académica del estudiante junto a los factores socio económicos y ambientales, que determinan su condición de potencial desertor, asociándole un índice de deserción como probabilidad de abandono del sistema educativo.

Se han elegido 72 centros educativos del distrito escolar de Los Alcarrizos, así como una cohorte conformada por cinco periodos académicos comprendidos entre el 2009 y el 2014, como plan de prueba piloto de esta tecnología predictiva.

A partir de este pronóstico, y al ser extendido en el futuro el estudio a nivel nacional, las autoridades gestoras de cada centro educativo y del Ministerio de Educación podrían elaborar políticas de intervención efectivas y puntuales encaminadas a la retención de los alumnos y alumnas en el sistema educativo nacional (educación básica y media) y al mejoramiento de los procesos de los centros educativos en pro de la disminución del índice de deserción escolar.

PALABRAS CLAVE

Deserción escolar; Retención escolar; Modelo predictivo; Data mining; Modelo de datos; Algoritmos; Árbol de decisión; Cohorte; Condición académica; Precisión y exactitud del modelo.

ABSTRACT

The motivation of the research, whose results are reported in this article, is to provide a tool to predict the risk of students dropping out from the basic and medium level education in the Dominican education system. Using predictive data mining models can determine patterns of student behavior, analyzing student academic history with the socio-economic and environmental factors that determine their potential defector status, associating a dropout rate as likely school leaving.

We have chosen 72 schools in the school district of Los Alcarrizos and a cohort made up of five academic periods between 2009 and 2014 as a pilot scheme of this technology predictive test.

From this forecast, and to be extended in the future nationwide study, the managing authorities of each school and the Ministry of Education should make policies effective and timely intervention aimed at the retention of students in the national education system (primary and secondary education) and the improvement of the processes of schools towards decreasing the dropout rate.

KEYWORDS

Dropout; Retention; Predictive modeling; Data mining; Data model; Algorithms; Attrition risk; Cohort; Academic status, Precision and accuracy of the model.

INTRODUCCIÓN

El abandono y deserción escolar es un indicador que busca medir el fenómeno provocado por los(as) alumnos(as) que dejan sus estudios antes de concluirlos. A pesar de los avances alcanzados en cuanto al acceso a la educación primaria y los esfuerzos realizados para retener a los niños, niñas y adolescentes para que culminen los estudios tanto del nivel básico como del nivel medio del sistema educativo pre-universitario, el país presenta elevados porcentajes de abandono intra-anual y de deserción antes de la conclusión del ciclo educativo.

La deserción es un fenómeno que impacta el ciclo completo de los 12 años de estudios preuniversitarios, y se da antes de concluir tanto el nivel básico como el nivel medio. De acuerdo a datos del último Censo Nacional de Población y Vivienda 2010, el 45.9% de los desertores corresponden al nivel básico. En el nivel medio el porcentaje de alumnos que desertan antes de concluirlo es de alrededor de un 16.0% (ONE, 2014).

En el país, al igual que en muchos países de la región, la pobreza y la inequidad de género son motivos importantes que conducen a la deserción. Los bajos ingresos de las familias, así como la desigualdad de género se conjugan en factores sociales que empujan a niños y niñas a abandonar el sistema educativo para insertarse de forma prematura al sistema laboral y/o a realizar actividades, tradicionalmente asociadas a estereotipos de género (vulnerabilidad del hogar).

El 64% de los niños que desertaron lo hizo por razones económicas; de igual manera el 18% de las niñas. Los demás factores influyentes en el fenómeno están asociados a factores demográficos, condiciones ambientales, condiciones del sistema educativo como tal y de los centros educativos en particular (ONE, 2014).

La dedicación del 4% del PIB para la educación, que se traduce en planes para la inversión en mejoras de los planteles escolares, condiciones curriculares nuevas, capacitación de maestros, desayuno escolar, tanda extendida, debe contribuir al mejoramiento de las condiciones generales de calidad del sistema educativo nacional. Pero a nivel micro se requieren de políticas y planes de acción específicos que vayan en la dirección de revertir los niveles de deserción escolar y contribuir a reforzar en forma directa las condiciones para el desarrollo intelectivo, psicológico, material y familiar de los niños y niñas como factores humanos relevantes en la problemática para producir el efecto de retención en el sistema educativo nacional.

La motivación de esta investigación es la de dotar de una herramienta que permita predecir el riesgo de deserción de los estudiantes antes de que ocurra el evento. Mediante modelos predictivos de minería de datos es posible determinar patrones de comportamiento del alumno(a), analizando la historia académica del estudiante junto a los factores socio económicos y ambientales, que determinan su condición de potencial desertor, asociándole un índice de deserción como probabilidad de abandono del sistema educativo.

A partir de este pronóstico, las autoridades gestoras del centro educativo y del sistema nacional podrían elaborar políticas de intervención efectivas y puntuales encaminadas a la retención de los alumnos y alumnas en el sistema educativo nacional (educación básica y media) y al mejoramiento de los procesos de los centros educativos en pro de la disminución del índice de deserción escolar.

ASPECTOS CONCEPTUALES DEL MODELO PREDICTIVO DE DESERCIÓN ESCOLAR

La minería de datos es entendida como el proceso de descubrir conocimientos ocultos, como patrones, asociaciones, cambios, anomalías y estructuras significativas a partir de grandes cantidades de datos almacenadas en bases de datos, datawarehouses, o cualquier otro medio de almacenamiento de información (Witten & Frank, 2000, p. 7). La aplicación de algoritmos de minería de datos requiere de actividades previas destinadas a preparar los datos de manera homogénea. La minería de datos podría identificar varios grupos, que luego pueden ser utilizados para obtener resultados más precisos de predicción por un sistema de soporte de decisiones. La recolección, preparación de datos, y la interpretación de los resultados son la etapa preliminar de la minería de datos, y pertenecen a todo el proceso de descubrimiento de patrones como pasos importantes.

El modelo de datos de entrada elaborado para producir el modelo de minería de datos, se entiende como el entrenamiento del algoritmo predictivo seleccionado mediante técnica supervisada, y representa un componente fundamental para que el algoritmo pueda aprender correctamente y generar resultados con cierta precisión estadística (Duda, Hart & Stock, 2001, p.14).

Algunas de las aplicaciones de la minería se centran en detectar patrones de comportamiento de individuos o entidades de datos, como son los agrupamientos de registros de datos por ciertas similitudes de sus atributos (análisis de clúster), registros poco usuales (la detección de anomalías), dependencias de las diferentes instancias (reglas de asociación), deserción de individuos de una actividad recurrente (educación, consumo y uso de servicios, etc.). Estos patrones pueden entonces ser vistos como una especie de resumen de los datos de entrada, y pueden ser utilizados en un análisis adicional en la máquina de aprendizaje y predictiva.

Las técnicas de la minería de datos provienen de la inteligencia computacional (inteligencia artificial, machine learning, clasificación de patrones, etc.) y de la estadística (análisis multivariado, inferencia, teoría estadística del aprendizaje, regresión, etc.). Dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados.

Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en supervisados y no supervisados (Theodoris & Koutroumbas, 2006, p. 6): (1) Algoritmos supervisados (o predictivos): predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos y de la historia de una entidad registrada en sus atributos explicativos y de una variable dependiente u objetivo, también denominada variable respuesta. La variable respuesta es usada como un clasificador de las entidades de datos bajo estudio (región de deserción o región de no deserción del individuo). (2) Algoritmos no supervisados (o del descubrimiento del conocimiento): se descubren patrones y tendencias en los datos por lo que son usados como técnicas de agrupamiento y clasificación mediante patrones similares de sus factores o atributos en forma conjunta (modelos multivariados) de los elementos bajo estudio.

Existen tres enfoques diferentes para abordar los modelos predictivos de deserción:

I. El enfoque basado en modelos de regresión: En este enfoque se inscriben el modelo de sobrevivencia de entidades de Cox, que ha sido exitosamente usado en estudios demográficos y de grupos de individuos sometidos a tratamientos médicos durante un periodo de tiempo específico para predecir su efecto o no en el paciente. También el enfoque de regresión logística binomial usado como clasificador binario.

II. El enfoque de clasificadores con discriminantes lineales o no lineales: logran la clasificación del objeto tomando una decisión de clasificación basada en el valor de una combinación lineal o función no lineal de las características explicativas de la variable de respuesta como superficie de separación de las clases. En esta técnica se inscriben Support Vector Machine (SVM), Linear Discriminant Analysis, etc.

III. El enfoque de reglas asociativas predictivas basado en arboles de decisión y en reglas de dependencia que se definen a partir del contenido de información del modelo de datos de entrada y categorizan una serie de condiciones que se presentan de forma sucesiva. En esta técnica se identifican los arboles de decisión binarios, Arboles CHAID, Arboles C5.0, redes neuronales, reglas bayesianas, y otros.

Como veremos más adelante, hemos adoptado este último enfoque por considerarlo el más adecuado a la situación del modelo piloto de los 72 centros educativos del distrito escolar de Los Alcarrizos. Esta decisión está basada en las pruebas de ajuste de los algoritmos predictivos realizada con el auxilio de las herramientas de software de modelación y data mining.

En Amaya et al. (2012) se referencia la ejecución de 19 proyectos de predicción de deserción escolar en diferentes países usando técnicas de minería de datos diversas. Es importante destacar que cada situación de estudio responde a un modelo específico de algorítmico adaptado a la realidad de información de cada país o región.

De aquí que el estudio de los 72 centros de los Alcarrizos es considerado un conglomerado poblacional

particular y no se pretende realizar ninguna inferencia, generalización o expansión de este modelo a otros distritos escolares del sistema educativo nacional. Es decir, cada región escolar, debe ser considerada una población particular propensa de generar un modelo particular predictivo. De igual manera, al adoptar estas técnicas de aprendizaje automático no paramétrico, no se presume ningún tipo de comportamiento de las variables envueltas ni de su distribución de probabilidad.

METODOLOGÍA DEL PROYECTO Y HERRAMIENTAS EMPLEADAS

La minería de datos sigue una metodología de ciclo de vida de proyectos con características particulares por el tipo de herramienta y objetivos definidos en el alcance del proyecto. Se ha establecido un ciclo de proyecto basado en la metodología CRISP-DM (SPSS, 2008, p. 3).

CRISP-DM, de Cross Industry Standard Process for Data Mining, trata de un modelo de proceso de minería de datos que describe los enfoques comunes que utilizan los expertos en minería de datos. Encuestas realizadas en 2002, 2004 y 2007 muestran que es la principal metodología utilizada para esta tarea. El único otro estándar de data mining nombrado en estas encuestas era el SEMMA. No obstante, 3-4 veces más personas reportaron optar por CRISP-DM. Una revisión y crítica de los modelos de minería de datos en 2009 llamó a CRISP-DM el "estándar de facto para el desarrollo de la minería de datos y los proyectos de descubrimiento de conocimiento".

El estándar incluye un modelo y una guía, estructurados en seis fases, algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores. Estas son:

1. Entendimiento del problema
2. Comprensión de los datos
3. Preparación de los datos
4. Modelación
5. Evaluación del modelo
6. Despliegue o distribución del modelo.

Las herramientas de software de minería de datos que hemos usado en este estudio son: a) IBM SPSS Modeler que es una plataforma de análisis predictivo diseñada para aportar inteligencia predictiva a decisiones llevadas a cabo por personas, grupos, sistemas y la empresa; b) RapidMiner (anteriormente, YALE, Yet Another Learning Environment) es un programa informático para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación de educación, capacitación, creación rápida de prototipos y en aplicaciones empresariales.

El análisis de variables usado para determinar el nivel de influencia de los factores explicativos en el modelo, así como el análisis de algoritmo predictivo usado para determinar cuál algoritmo de minería de datos mejor se ajusta a los datos de entrada de la deserción escolar se ha realizado con SPSS Modeler (SPSS, 2003). Mientras que la aplicación del modelo usando el algoritmo seleccionado (árbol de decisión) se ha realizado usando la herramienta RapidMiner (Hofmann & Klinkenberg, 2013).

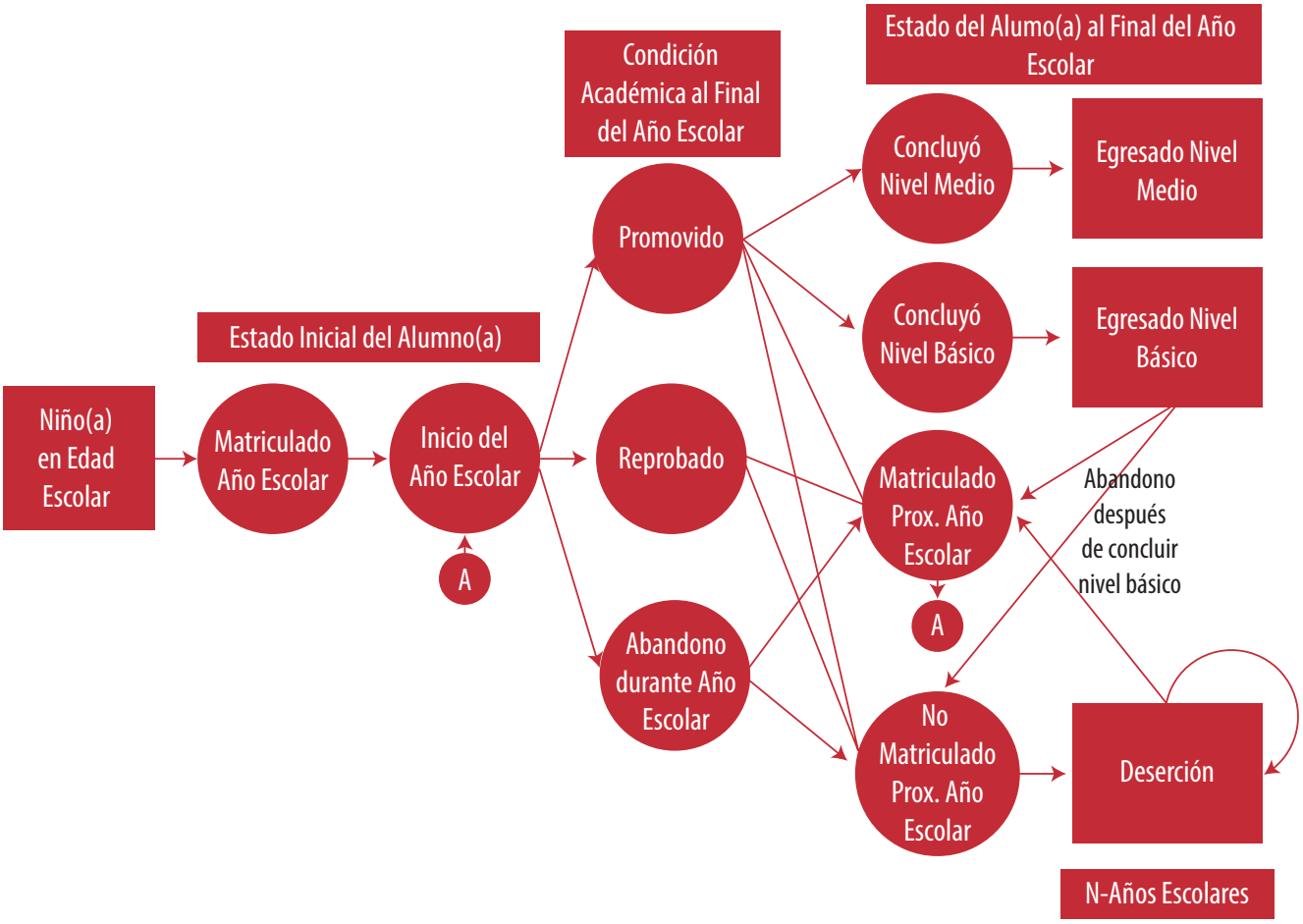
Esto así por considerar en esta etapa piloto de baja escala de los datos, el uso de una herramienta económica y fácil de usar que es RapidMiner. Para un proyecto de mayor escala y complejidad, como sería la aplicación del modelo a nivel regional o nacional del sistema educativo, se puede pensar en el uso de SPSS Modeler como herramienta de mayor potencia.

PREPARACIÓN DEL MODELO DE DATOS DE DESERCIÓN ESCOLAR

Para poder realizar un proceso de preparación de datos efectivo y de calidad partimos de un modelo de Deserción Escolar el cual lo representamos en el diagrama un estado (ver gráfico 1), cuyo objeto es establecer todos los posibles estados de un alumno(a) durante su tiempo de vida en el sistema educativo. Este modelo de estado nos provee los componentes fundamentales para poder entender las condiciones iniciales, intermedias y finales del alumno inscrito en un centro educativo específico dentro del sistema

educativo nacional. Nos va a servir para dos propósitos fundamentales: a) definir factores de medición o variables explicativas de la deserción escolar; b) establecer los procedimientos de preparación de los datos que van a servir de entrada al modelo predictivo de deserción escolar.

DIAGRAMA DE ESTADO DESERCIÓN ESCOLAR



En este modelo los estados se representan en círculos y rectángulos. Los rectángulos representan los estados inicial y finales y los círculos los estados intermedios. El primer componente importante es el niño o niña en edad escolar, que se matricula en un centro educativo, previo al inicio de un año escolar, y que lo convierte en un alumno del centro y del sistema educativo nacional. Esta acción representa su estado inicial en el diagrama.

Al producirse el proceso de evaluación del alumno(a) al final del año escolar este obtiene una condición académica final que puede tener tres estados posibles: Promovido, Reprobado o de Abandono durante el año escolar.

Si la condición académica al final del año escolar del alumno(a) es de Promovido, el mismo pudo haber concluido sus estudios de nivel medio si está cursando el 4to grado de media. En este caso su estado final es de egresado del nivel medio y por tanto de la educación preuniversitaria. En este caso el alumno ha alcanzado el estado ideal, objetivo de nuestro sistema educativo al cumplir el ciclo completo de 12 años exitosamente. El otro estado posible es el de concluir la educación básica si aprueba sus exámenes estando en el 8vo grado. Su estado seria de egresado del nivel básico. Pero en este punto tiene dos alternativas: (1) se matricula para el próximo año escolar en un centro educativo y prosigue así sus estudios secundarios

o de nivel medio y su estatus seria matriculado para el próximo año escolar; (2) no se matricula para el próximo año escolar, por lo que obtiene un estatus de deserción del nivel medio.

Un alumno en estatus de promovido para cualquier otro grado escolar o en estatus de reprobado o abandono durante el año escolar puede pasar a uno de los estatus de matriculado o de no-matriculado para el próximo año escolar. En el primer caso este estudiante pasa a continuar sus estudios en el próximo año. En el segundo caso de no matriculado para el próximo año pasaría al estatus final de deserción escolar del nivel básico si está cursando un grado de este nivel (entre 1ro y 8vo) o de deserción escolar de nivel medio si eta cursando entre el 1ero y 4to grado del nivel medio.

El estatus de deserción escolar considera un periodo de uno o más años fuera del sistema educativo nacional al momento del corte del estudio (año académico 2013-2014) sin que el estudiante se haya reintegrado al sistema educativo nacional (independientemente del centro en que se pueda matricular).

A partir de este modelo de estado del estudiante en el sistema educativo nacional podemos derivar un conjunto de mediciones que han de servir como parámetros o factores explicativos de deserción escolar, en adición a los demás factores demográficos y socio económicos. Estos son:

- Condición académica del alumno(a) al final del año escolar cuando pasa a condición de deserción, es decir, cuando no se matricula para el próximo año académico.
- Tiempo de permanencia del alumno en el sistema educativo al momento del corte del estudio y antes de pasar a condición de deserción o de egresado, es decir, cuantos periodos o años escolares ha durado en el sistema.
- Ultimo grado alcanzado antes de pasar a su condición de deserción.
- Cantidad de abandonos tenidos antes de pasar a su condición de deserción o de egresado. Entendemos por abandono el retiro voluntario o no de un estudiante durante el año escolar, denominado también abandono intra anual. El alumno puede retornar al sistema el siguiente año escolar.
- Tiempo de deserción transcurrido, es decir, la cantidad de años escolares sin retornar al sistema
- Cantidad de reprobaciones tenidas antes de pasar a su condición de deserción o egresado.
- Cantidad de promociones tenidas antes de pasar a la condición de deserción o egresado.
- Si se ha transferido de centro educativo durante su estadía en el sistema antes de la condición de deserción o egresado (movilidad).

En la siguiente tabla se muestran el orden, nombre y descripción de los atributos de datos del archivo de entrada al modelo.

LISTA DE ATRIBUTOS DE ENTRADA

NO.	NOMBRE ATRIBUTO	DESCRIPCIÓN	EJEMPLO
1	IdEstudiante		35
2	Año_Academico_first_1		2009-2010
3	Año_Academico_last_1		2013-2014
4	CodigoCentro_first		2358
5	CodigoCentro_last		359
6	Sector_last		PUBLICO
7	Tanda_last		MATUTINA
8	FechaNacimiento_last		29-Nov-04
9	Sexo_first_1		Masculin
10	ICV1	Pertenece o no al programa Prosoli	0
11	nivel_grado_last		14
12	Abandono_sum		0
13	Promovido_sum		2
14	Reprobado_sum		0
15	Condicion_otra_sum		0
16	anios_acad		2
17	EdadEstudiante		10.502396
18	ZonadelCentro		URBANA-M
19	IndicePobrezaCentro	vulnerabilidad del sector geográfico	69.88
20	IndiceProsoliCentro	Indice de cantidad de estudiantes en el programa Prosoli	8.2
21	Condicion_Acad_last		Promov
22	Desercion		0 o 1

Las tablas 2 y 3 siguientes muestran la tasa de deserción (UNESCO, 2009) para los diferentes niveles de básico y medio y para los años de la cohorte seleccionada (2009-2014) para el total de estudiantes que han cursado al menos un año académico en uno de los 72 centros de los Alcarrizos. El promedio es de 9.09% para los 5 periodos escolares bajo estudio y los estudiantes que registramos desde el 2009 usados para entrenamiento del modelo. Para nivel básico la tasa es de 8.79% y para nivel medio de 9.97%; las cuales compiten con las tasas de egresados de 9.15% y 10.94% respectivamente. Es muy significativo notar que del total de deserciones entre el 2009 y el 2014 solo retornaron al sistema educativo el 7.43% de los estudiantes (1,834 alumnos).

El total de estudiantes es de 74,291 que representa la cantidad de estudiantes que han cursado al menos un grado en los 72 centros del Distrito de Los Alcarrizos tomado como piloto de estudio.

TABLA 2

	12 Grados				Tasa de Egresados	
TASA DE DESERCIÓN ESCOLAR	**TOTAL MATRICULA GRADOS 12**	**TOTAL DESERCIÓN GRADOS 12**	**TASA DESERCIÓN GRADOS 12**	**TASA PROM PONDERADA DESERCIÓN GRADOS 12**	**TASA DE EGRESADOS BÁSICA**	**TASA DE EGRESADOS MEDIA**
2009-2010	48710	5915	10.82%	10.84%	8.18%	10.11%
2010-2011	48774	4669	8.73%	8.73%	8.09%	10.59%
2011-2012	48794	3869	7.28%	7.23%	8.99%	9.88%
2012-2013	49258	3771	7.04%	6.48%	9.34%	10.74%
2013-2014	49767	6446	12.22%	12.17%	11.12%	13.36%
2014-2015	46309					
Total	291612	24670				
Promedio	49061	4934	9.22%	9.09%	9.15%	10.94%
Media Geométrica			9.00%	8.84%	9.09%	10.87%
	Retorno Total	1834	7.43%			

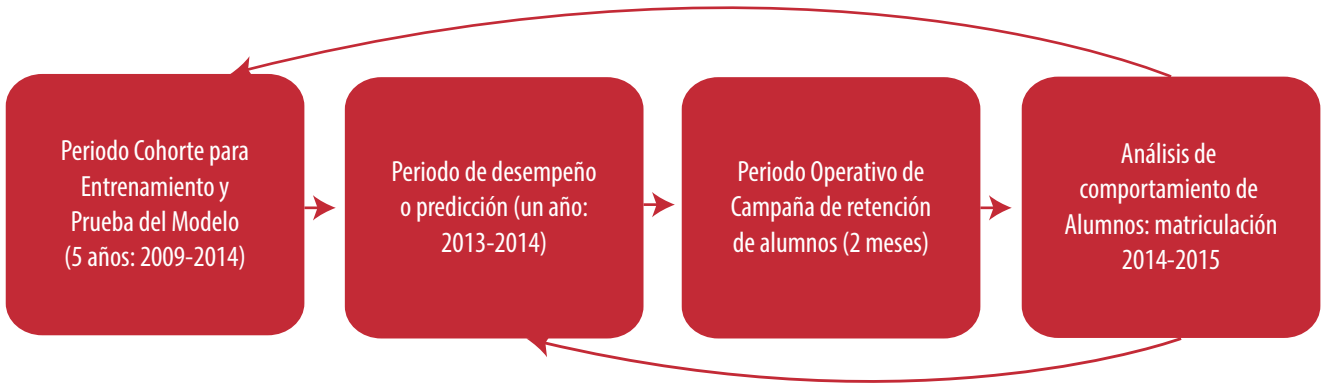
TABLA 3

	Nivel Básico					Nivel Medio				
TASAS DE DESERCIÓN ESCOLAR	**MATRICULA NIVEL BÁSICO**	**EGRESADOS NIVEL BÁSICA**	**DESERCIONES NIVEL BÁSICA**	**INGRESADOS 1ER GRADO**	**TASA DE DESERCIÓN NIVEL BÁSICO**	**MATRICULA NIVEL MEDIO**	**EGRESADOS MEDIA**	**DESERCIONES MEDIA**	**INGRESADOS 1ERO DE MEDIA**	**TASA DESERCIÓN NIVEL MEDIO**
2009-2010	40860	3523	4862	5480	10.95%	7850	1039	1053	2988	10.23%
2010-2011	39531	3353	3638	4498	8.72%	9243	1270	1031	3635	8.78%
2011-2012	38076	3661	2738	4166	6.88%	10718	1339	1131	3767	8.48%
2012-2013	37046	3784	2259	4074	5.80%	12212	1571	1224	1610	8.57%
2013-2014	36711	4164	4364	3885	11.59%	13056	2009	2082	4211	13.78%
2014-2015	33280			3552		13029			4230	
Total	225504	18485	17861	25655		66108	7228	6521	20441	
Promedio	38445	3697	3572	4421	8.79%	10616	1446	1304	3242	9.97%
Media Geométrica					8.49%					9.79%

CÓMO FUNCIONA UN MODELO PREDICTIVO DE DESERCIÓN ESCOLAR

Todo modelo predictivo de data mining cuenta de cuatro periodos en su ciclo de vida, como se muestra en el gráfico 2. Se selecciona un periodo de colección de datos o cohorte (historial de varios años académicos) que sirven de entrenamiento y prueba del modelo. Estamos usando los 5 años 2009-2014 como periodo de cohorte, 70% dataset de entrenamiento y 30% data set para prueba del modelo.

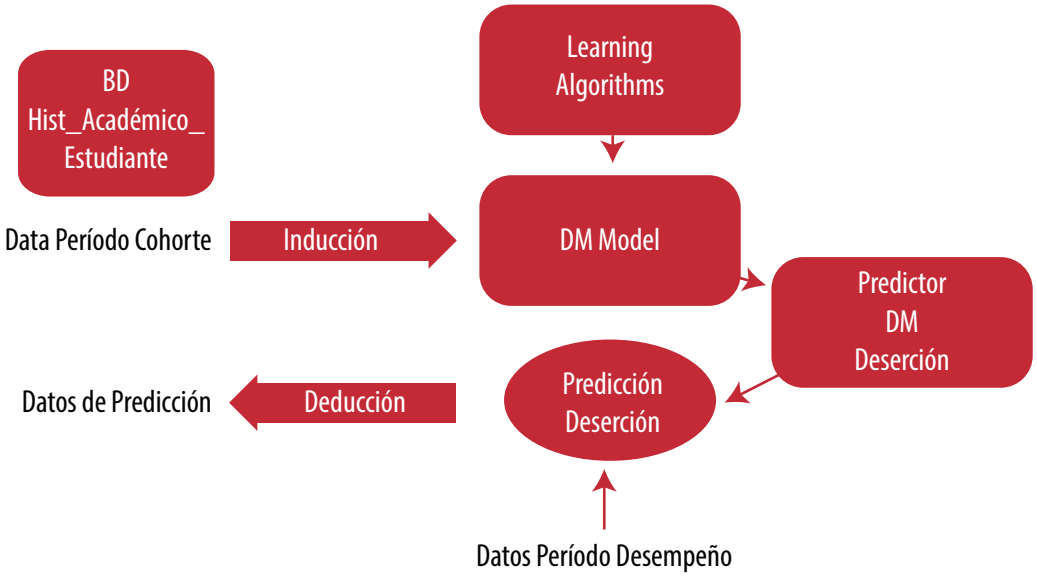
GRAFICO 2



Luego de la generación o entrenamiento del modelo basado en el algoritmo seleccionado se procede a observar su precisión y exactitud, tanto del set de entrenamiento como del de prueba, en una tabla de contingencia con pruebas estadísticas de significación (usaremos F1 Score como medidas de precisión del modelo). Luego se procede a verificar el modelo simulando un periodo académico próximo, donde aplicaremos dichas prueba de igual manera. El próximo año académico se registra la matriculación del alumno y se contrasta con la predicción para ver el nivel de exactitud y precisión predictiva del modelo y como prueba de ajuste o desviación del modelo.

El diagrama siguiente explica este proceso:

GRAFICO 3
COMO APRENDE Y PREDICE EL MODELO



Se usa el último año académico 2013-2014 como periodo de verificación del modelo simulando la predicción de este año basado en el entrenamiento de la cohorte seleccionada. Se proveen los datos de las variables explicativas con la condición de deserción en blanco, que es la variable a ser predicha.

El sistema proveerá un valor de cero o uno, según sea no o si la deserción esperada en la matriculación del siguiente año escolar 2014-2015. El valor del % de confiabilidad será usado como valor en riesgo de la respuesta. Esta información deberá ser cargada en el sistema de Gestión de Centros Educativos del MINERD con el objeto de servir a los planes de retención de alumnos con el mayor riesgo de deserción.

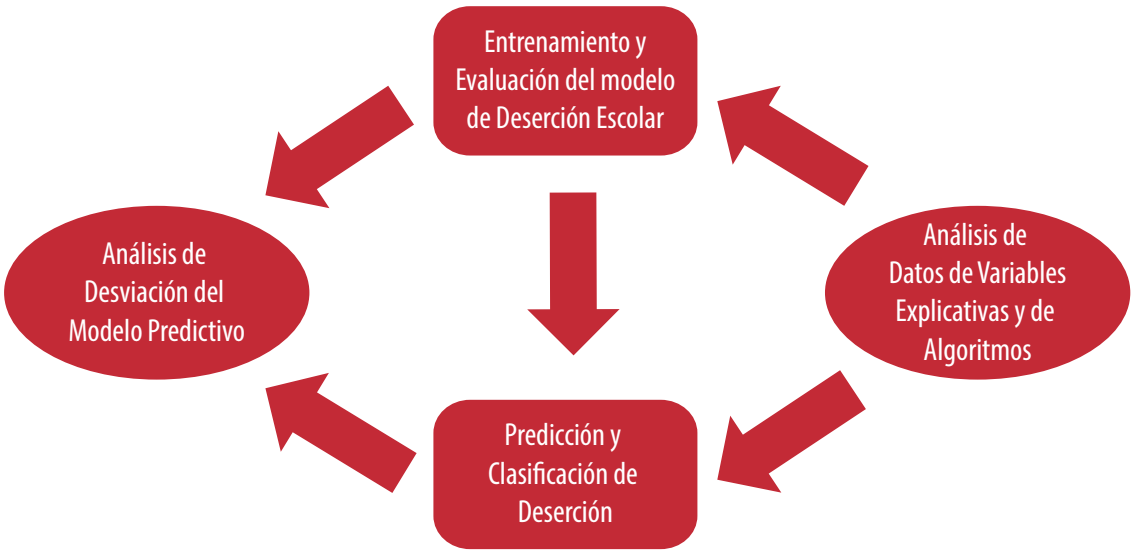
COMPONENTES DEL MODELO PREDICTIVO DE DESERCIÓN ESCOLAR

Como hemos explicado, los modelos predictivos de deserción se inscriben como técnicas de minería de datos supervisadas no-paramétricas. La serie histórica de registros o cohorte, denominada data set de entrenamiento y prueba, determina el patrón de comportamiento analizado por el algoritmo predictivo, que crea las reglas de inferencia (denominada también generalización) para nuevos individuos que no pertenecen al conjunto original de entrenamiento del modelo y para un periodo siguiente al periodo usado como entrenamiento.

Esta técnica ha sido empleada con mucho éxito en el ámbito de las aplicaciones bancarias, de tarjetas de crédito y de telecomunicaciones, para determinar el riesgo de que un cliente deje de usar el servicio dentro de un periodo determinado futuro.

En siguiente gráfico 4 se pueden ver los cuatro componentes funcionales del modelo de deserción escolar implementado mediante las herramientas Rapid-Miner y SPSS Modeler, y que son explicados en las secciones siguientes.

GRAFICO 4



ANÁLISIS DE DATOS DE VARIABLES EXPLICATIVAS Y DE ALGORITMOS

Los factores se usan como variables explicativas para la estimación de la deserción escolar o variable objetivo (0 no deserción y 1 deserción y su probabilidad de ocurrencia) [Deserción, Riesgo] = F(factor1, factor2, …,factorn). Estos se seleccionan de acuerdo a su nivel de significación, eliminación de factores auto correlacionados, basados en índices de correlación y pruebas de hipótesis.

El primer reporte que se analiza es la tabla de clasificación de atributos según la capacidad de predicción que tiene cada uno de ellos. Según los resultados de la tabla 4 siguiente los atributos que no se repiten por estudiantes son los que tienen más peso, por eso solo se incluyen los que están debajo del atributo FechaNacimiento_Last, incluyéndolo también.

Poder de clasificación de los atributos:

TABLA 4

ATRIBUTO	PESO
matriculado	0.7009
N_anios_acad	0.5446
FechaNacimiento_last	0.1781
edad_estudiante	0.1781
Año_Academico_last	0.1630
Cambio_Centro	0.0824
Abandono_sum	0.0764
Nivel_Grado	0.0689
Condicion_Academica	0.0632
Condicion_otra_sum	0.0385
IndicePobrezaCentro	0.0295
IndiceProsoliCentro	0.0105
ICV1	0.0096
resumen	0.0077
Reprobado_sum	0.0077
CodigoCentro	0.0070
Tanda	0.0055
Sector	0.0055
Promovido_sum	0.0038
Nivel	0.0025
Año_Academico	0.0013
Sexo_first_1	0.0013
Grado	0.0006
ZonadelCentro	0.0003
Año_Academico_first_1	0.0000

El siguiente resultado es la tabla de correlaciones para ver que atributos están correlacionados y eliminar aquellos que tenga menos peso en el poder de predicción (ver tabla 5). Como se puede ver, los atributos incluidos en el modelo, después de pasarles el operador que remueve los atributos inútiles, tienen muy poca correlación. El único caso a resaltar fue el de edad del estudiante con el nivel académico alcanzado, que aunque tiene una alta correlación (0.872) el modelo presento mejores resultados incluyendo ambos atributos.

TABLA 5.
MATRIZ DE CORRELACIÓN DE LOS ATRIBUTOS

ATTRIBUTES	TANDA	NIVEL_ GRADO	SEXO_ FIRST_1	ICV1	ZONADELCEN...	INDICEPROSO...	INDICEPOBRE...	EDAD_ESTUD...	PROMOVIDO_...	CAMBIO_C
Tanda	1	0.255	0.010	-0.021	0.012	0.005	-0.025	0.247	0.072	0.150
Nivel_Grado	0.255	1	0.099	-0.041	0.004	0.297	0.063	0.872	0.275	0.066
Sexo_first_1	0.010	0.099	1	0.020	0.002	0.065	-0.056	0.040	0.115	0.033
ICV1	-0.021	-0.041	0.020	1	0.026	0.106	0.033	-0.043	0.046	-0.000
ZonadelCen	0.012	0.004	0.002	0.026	1	0.067	0.168	0.012	-0.002	0.049
IndiceProsol	0.005	0.297	0.065	0.106	0.067	1	0.408	0.238	0.066	-0.016
IndicePobreza	-0.025	0.063	-0.056	0.033	0.168	0.408	1	0.038	-0.242	-0.134
edad_estudi	0.247	0.872	0.040	-0.043	0.012	0.238	0.038	1	0.226	0.082
Promovido_	0.072	0.275	0.115	0.046	-0.002	0.066	-0.242	0.226	1	0.295
Cambio_Ce	0.150	0.066	0.033	-0.000	0.049	-0.016	-0.134	0.082	0.295	1

Luego de analizar los atributos se selecciona el algoritmo de análisis predictivo supervisado de acuerdo a un objetivo de precisión con el set de entrenamiento y de prueba. Para cada tarea de minería, hay algunos algoritmos adecuados como lo hemos definido en la sección anterior. Se seleccionan un conjunto de datos preliminares de prueba. En muchos casos, no sabemos cuál es el mejor algoritmo de ajuste para los datos antes del entrenamiento del modelo. En SPSS Modeler usamos el Clasificador Automático.

El nodo Clasificador automático compara varios modelos diferentes para obtener resultados binarios (sí o no, abandono o no de estudiante, etc.), lo que le permite seleccionar el mejor enfoque para un análisis determinado. Son compatibles varios algoritmos de modelado, por lo que es posible seleccionar los métodos que desee utilizar, las opciones específicas para cada uno y los criterios para comparar los resultados. El nodo genera un conjunto de modelos basado en las opciones especificadas y clasifica los mejores candidatos en función de los criterios que especifique en nuestro caso de precisión, es decir cuántos desertores fueron predichos versus la deserción observada y cuantos no desertores por igual.

El árbol de decisión es el mejor algoritmo de predicción estimado partir de la medida de precisión arrojada para este conjunto de datos. Esto coincide con estudios anteriores de países latinoamericanos con realidades similares a las nuestras (Valero Orea, Salvador Vargas & García Alonso. 2014).

CONCEPTOS TEÓRICOS SOBRE EL ALGORITMO DE ÁRBOL DE DECISIÓN

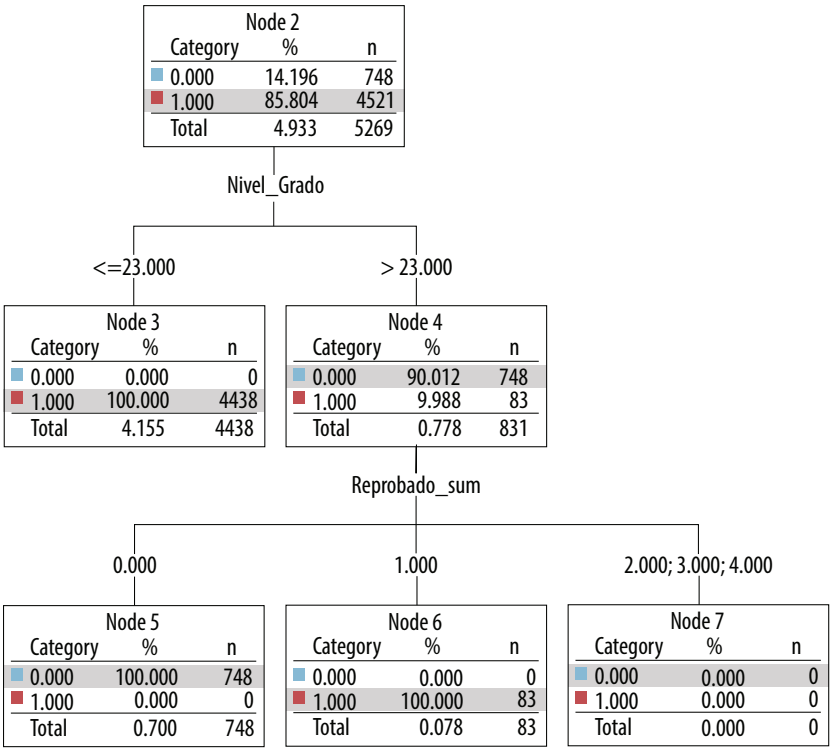
A los arboles de decisión también se les conoce como arboles de clasificación y se emplean en muchas áreas del saber tales como medicina, finanzas e ingeniería. Su uso es muy popular debido a su simplicidad y transparencia. Son auto-explicativos y no se necesita un experto para entender la estructura de un árbol de decisión. Cuando las ramificaciones del árbol son muchas, se dificulta su interpretación, y en esos casos se puede recurrir a otras técnicas de representación, como la regla de decisión.

El árbol de decisión es un clasificador recursivo de la información suministrada, de manera tal que a través de las ramificaciones del árbol se logre asignar una clase a cada una de las instancias (ejemplos), en este caso, estudiante. El árbol tiene nodos y ramas. Cada nodo tiene una rama de entrada y dos o más de salida. La cantidad de ramas que salen de un nodo dependerán de la evaluación de un atributo en ese nodo (Han, Kamber & Pei. 2012).

Hay un nodo inicial (root) del cual parten todas las demás ramificaciones del árbol. Hay nodos internos que a su vez se siguen ramificando, y nodos externos, llamados hojas, (nodos terminales o nodos de decisión) que es donde se hace la clasificación. Cada hoja asigna una clase y una probabilidad de pertenencia a esa clase, denominada "confidence".

En gráfico 5 siguiente se muestra un ejemplo de una sección del árbol de decisión predictivo del modelo resultante de la predicción escolar. El primer nodo es el inicial donde se puede ver que su ocurrencia ha dependido de la probabilidad de eventos cero (no deserción) en un 15% y de eventos uno (deserción) 85%, a partir de aquí se aplica la condición de si es un alumno del 4to grado de media o no lo es. Si no lo es se clasifica como una deserción, si es de 4to de media se verifica su acumulado histórico de condición académica reprobado. Si no ha reprobado se clasifica como cero (no deserción) de lo contrario se clasifica como uno (deserción).

GRAFICO 5



El pre procesamiento de los datos resulta sencillo cuando se usan arboles de decisión, ya que no está influenciado por las medidas usadas en los atributos, si hay una gran diferencia en los valores de los mismos. Otros algoritmos de clasificación requieren usar normalización para evitar la falta de convergencia.

El tamaño del árbol es crucial en su interpretación. La complejidad del árbol mejora la precisión de la clasificación. Hay varias métricas para medir la complejidad del árbol: a) el número total de nodos; b) número total de hojas; y c) Profundidad o número de atributos usados.

El funcionamiento del árbol de decisión se basa en dividir los datos (estudiantes) en función de la homogeneidad de los mismos. Se define una métrica de impureza que cumpla con cierto criterio, basada en calcular la proporción de los estudiantes que pertenece a una clase. El criterio es el siguiente:

La métrica de impureza es máxima cuando todas las clases (desertores y no desertores) están igualmente representadas. La métrica de impureza es cero cuando solo una clase está representada.

Entre las métricas empleadas están la Entropía, el Índice de Gini, y el Information Gain.

Entropía: C. Shannon el creador de la teoría de información, define la entropía como log(1/P) o –log(P) donde P es la probabilidad de que ocurra un evento. Si la probabilidad de todos los eventos posibles no es la misma, se necesita un factor ponderador y entonces la entropía será:

$$H = -\sum_{k=1}^{m} P_k log_2(P_k)$$

Donde m representa las diferentes clases a ser usadas, en nuestro caso dos, desertores y no desertores. Mientras más alta la entropía, mayor el contenido de información. La entropía es 0 (mínima impureza) cuando todos los elementos de la data son de la misma clase, y 1 (máxima impureza) cuando todas las clases tienen la misma proporción.

Índice de Gini: Este índice, que no es el mismo que se usa en economía para medir la concentración, varía entre 0 y 0.5 y se calcula con la siguiente formula:

$$G = \sum_{k=1}^{m} (1 - P_k)^2$$

Information Gain: Es el cambio en la entropía. Este criterio calcula la entropía de todos los atributos al momento de hacer la selección del atributo para dividir el árbol y aquel que presente la mayor Information Gain es seleccionado. La fórmula es la siguiente:

$$IG(y|x) = H(y) - H(y|x)$$

En resumen el algoritmo de árbol de decisión funciona así:

- Usando el criterio de entropía, ordenar los datos en homogéneos y no homogéneos por atributo. Atributos homogéneos tienen baja entropía y atributos no homogéneos tienen gran entropía.
- Asignar una ponderación a cada atributo en función de la entropía.
- Computar la Information gain para cada atributo.
- El atributo con la mayor Ingormation Gain será el root o punto de partida del árbol.
- Repetir los pasos anteriores hasta para cada atributo con entropía diferente de cero, si la entropía es cero, entonces esa rama se convierte en terminal.

En la vida real es difícil hacer que los nodos terminales del árbol o las hojas sean 100% homogéneos, es decir que solo haya una clase. De ocurrir eso, estaríamos produciendo el fenómeno de Overfitting, por lo tanto hay que evitar la ramificación excesiva del árbol cuando se cumplan los siguientes criterios:

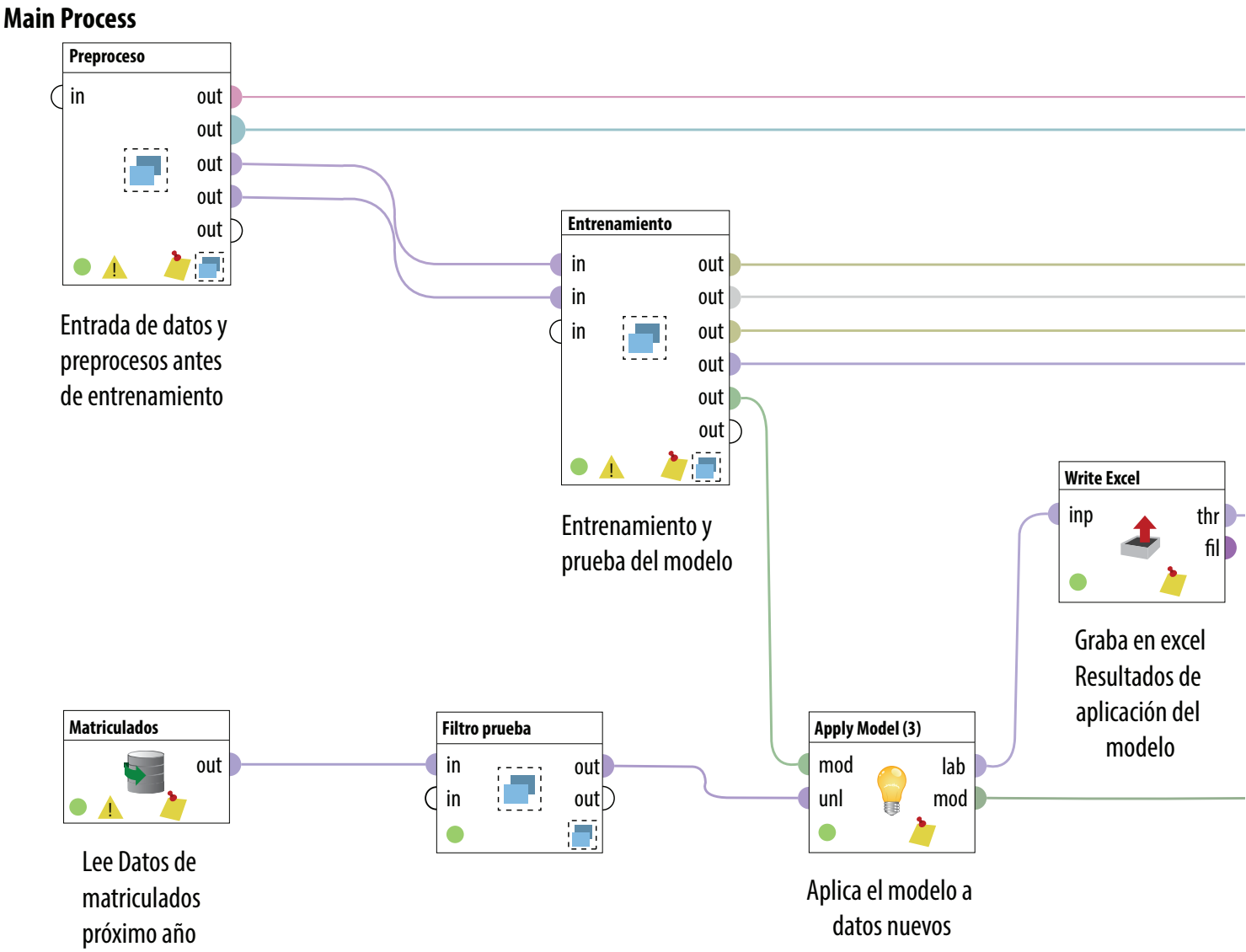
- Ningún atributo satisface el umbral de la mínima ganancia de información (Information Gain).
- Se ha alcanzado la máxima profundidad del árbol. Mientras más grande se hace el árbol, más difícil es su interpretación.
- Hay al menos un cierto número de casos en una parte del árbol.

El fenómeno denominado "overfitting", consiste en que el modelo aprende solo los ejemplos del training set, pero no tiene la capacidad de clasificar correctamente ejemplos nuevos, es decir, pierde capacidad de generalización predictiva, demostrado en la exactitud y precisión del test set.

ESTRUCTURA DEL MODELO EN RAPIDMINER

Para hacer más fácil la documentación y el mantenimiento del modelo de predicción de deserción escolar, el mismo fue construido en una estructura modular jerárquica, donde los componentes principales son los de: a) lectura y pre-procesamiento de datos, b) Entrenamiento y evaluación del modelo, y c) Aplicación del modelo para predecir la probabilidad de deserción de los estudiantes. En el siguiente gráfico 6 se puede ver el modelo a nivel superior.

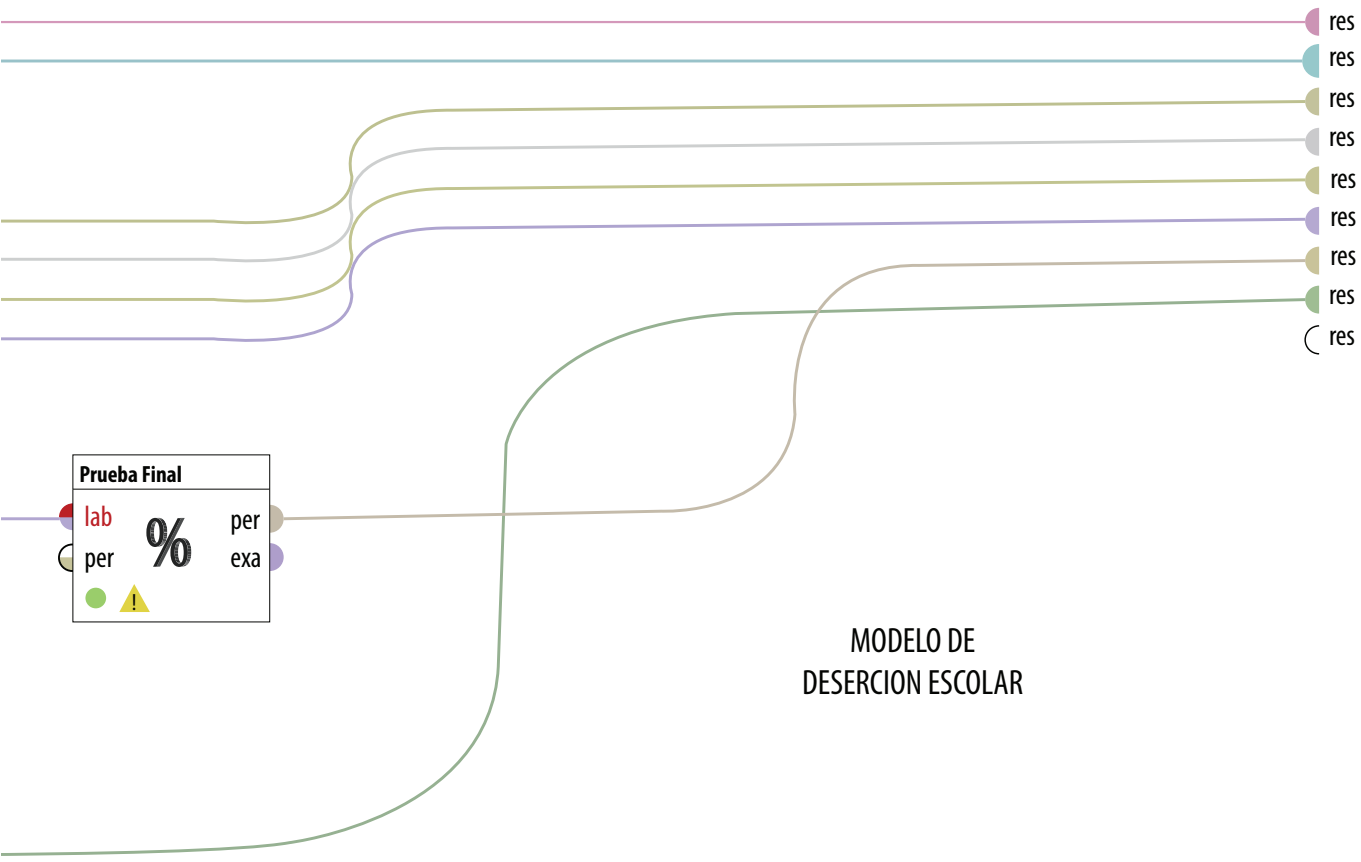
GRAFICO 6



El módulo 1: Contiene los operadores donde se leen los datos de los estudiantes correspondientes a los últimos periodos académicos. Además se hacen algunas operaciones para evaluar la utilidad de los atributos y preparar los datos para el entrenamiento y prueba del modelo.

El módulo 2: Es donde residen las operaciones más importantes del modelo, ya que es ahí donde se prueban los algoritmos de clasificación y predicción. En este módulo también se hace la optimización del algoritmo de clasificación.

El Modulo 3: Lo componen los operadores para leer los datos del último año académico, el de filtro de los registros, otro para alimentar estos datos al modelo optimizado de predicción de deserción escolar, uno para copiar a un archivo en formato Excel los resultados y uno final para evaluar el desempeño del modelo con esos datos nuevos. Es decir, el módulo 3 se puede usar tanto para predecir la deserción del año siguiente, como para evaluar la bondad de la predicción cuando se tengan los resultados de las inscripciones de los estudiantes.

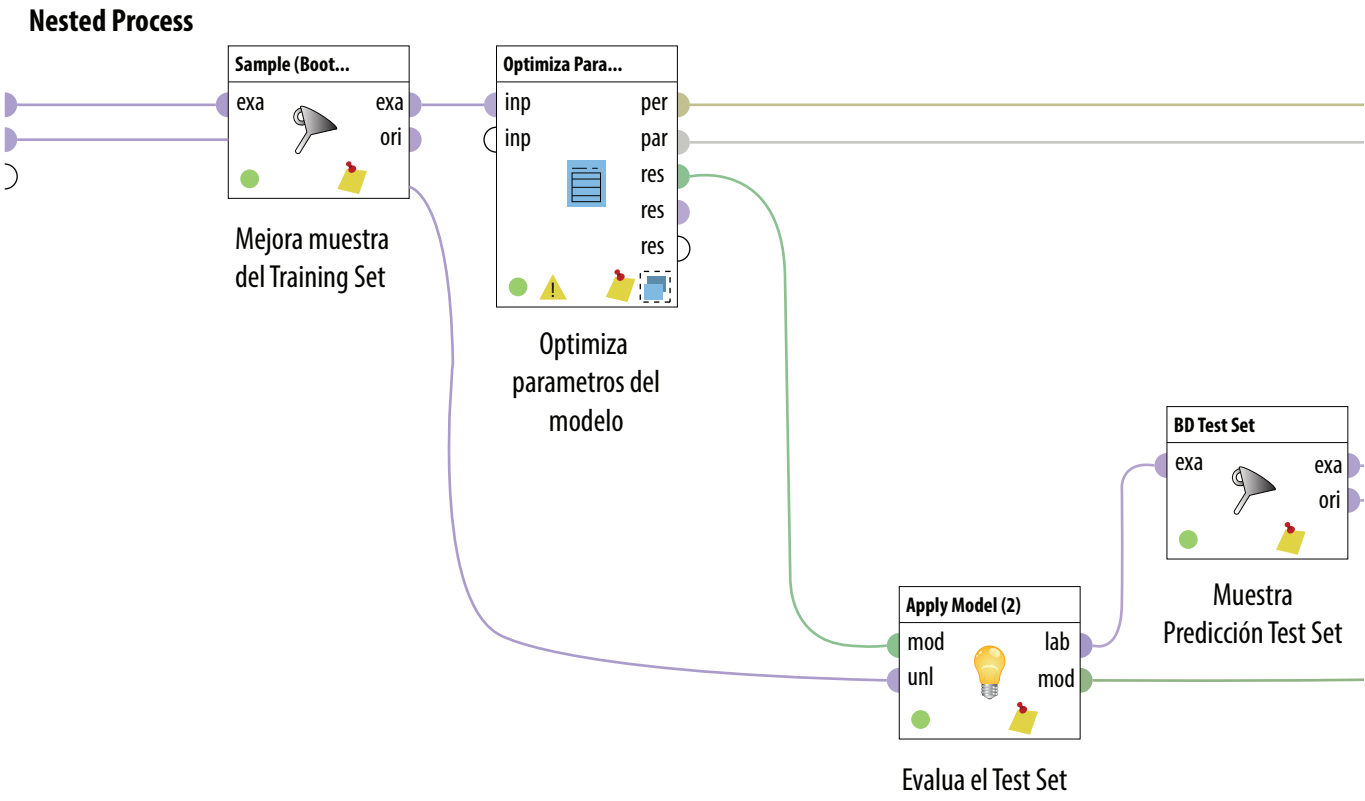


MODELO DE
DESERCION ESCOLAR

A. MÓDULOS DE ENTRENAMIENTO, PRUEBA Y PREDICCIÓN DEL MODELO

En este módulo se usan los insumos del módulo de pre-procesamiento y se construye el modelo que se usara para la predicción de la deserción escolar de los estudiantes (ver gráfico 7). A continuación se describen los operadores del módulo.

GRAFICO 7



Nótese que el módulo de entrenamiento tiene 2 entradas (el training set y el test set) y 5 salidas: a) La tabla de contingencia con los resultados de la evaluación del training set; b) El reporte de los parámetros óptimos del algoritmo de clasificación (árbol de decisión); c) La tabla de contingencia con los resultados del test set; d) La tabla con la base de datos original y la predicción de deserción para cada estudiante; e) El modelo de clasificación entrenado.

A continuación la descripción de los operadores del módulo de entrenamiento y prueba del modelo:

B. ENTRENAMIENTO DEL MODELO

Para evaluar la bondad del modelo para predecir la probabilidad de que un estudiante deje la escuela, se entrena el modelo con los datos del training set (periodo escolar 2009-2014 de la cohorte), cuidando que no se produzca el fenómeno denominado "overfitting". En la siguiente Tabla 6 se ve el resultado, y esta indica que el modelo pudo predecir el 99.55% de los casos del training set. Esto quiere decir que la predicción del modelo coincide con lo que tiene el atributo "deserción" en el training set en ese mismo porcentaje.

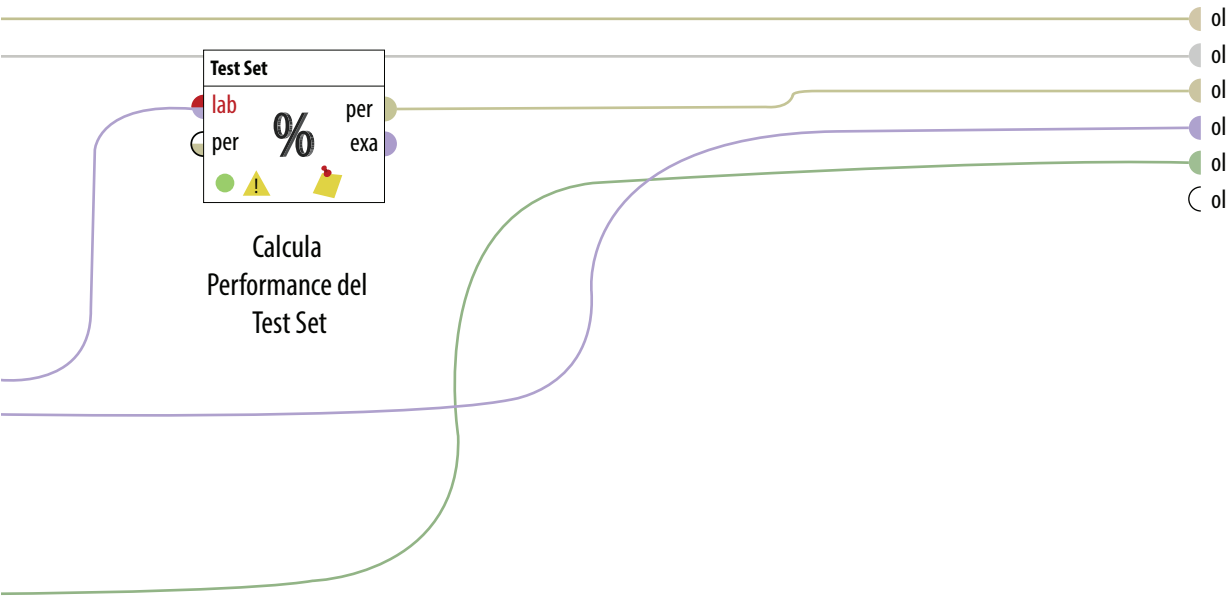


TABLA 6.
EVALUACIÓN DEL MODELO CON DATOS DEL TRAINING SET

ACCURACY:99.55% +/- 0.05 (MIKRO:99.55%)			
	true 1.0	true 0.0	class precision
pred. 1.0	42854	432	99.00%
pred. 0.0	372	134515	99.72%
class recall	99.14%	99.68%	

Es importante resaltar que el modelo pudo predecir el 99.14% de los estudiantes que desertaron en la muestra y el 99.68% de los que no desertaron incluidos en la muestra del training set. La cantidad de estudiantes contenidos en el training set es más del 70% de la muestra, debido al efecto que produce la aplicación del operador para balancear la muestra y el operador de Bootstrapping.

Otro aspecto importante a destacar es que solo el 0.86% (1 - 99.14%) de los casos del training set que desertaron, fueron mal clasificados por el modelo. Esto era de esperarse, ya que el modelo no puede clasificar correctamente el 100% de los casos, ya que esto podría quitarle poder de generalización, es decir, la capacidad para poder clasificar correctamente casos nuevos.

C. PRUEBA DEL MODELO

Se hizo el mismo análisis con la base de datos del test set. En la tabla 7 siguiente se resumen los resultados. El test set contiene 54,359 casos, es decir el 30% de la base de datos completa.

TABLA 7.
PRUEBA DEL MODELO

ACCURACY:94.56%			
	true 1.0	true 0.0	class precision
pred. 1.0	5409	2182	71.26%
pred. 0.0	775	45993	98.34%
class recall	87.47%	95.47%	

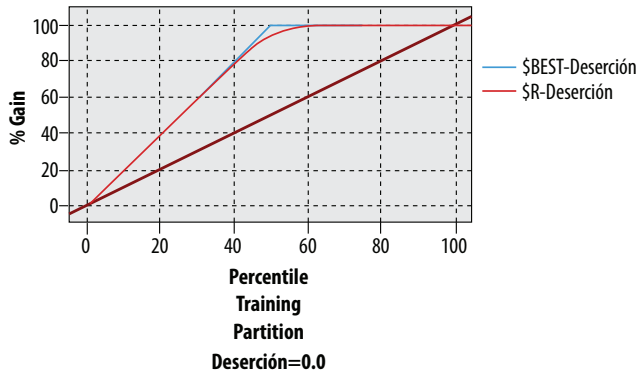
La exactitud del modelo en el test set fue de 94.56%, es decir, un 5.44% de error de clasificación, siendo 87.47% (Sensitivity) la cantidad de estudiantes que desertaron que fue bien clasificado, y el 95.47% (Specificity) de los que no desertaron que se clasifico correctamente. Estos son resultados muy satisfactorios considerando que el test set contiene datos de estudiantes que el modelo no uso durante el entrenamiento, por lo que se puede deducir que el modelo no sufre de Overfitting.

De los resultados de la prueba también se deduce que 2,182 estudiantes fueron falsos positivos (estudiantes que no desertaron y se pronosticó que si lo harían), 775 (12.53% de los que si desertaron) fueron falsos negativos (estudiantes que si desertaron y no se pronosticaron como tal).

El gráfico 8 siguiente muestra la curva ROC (Receiver Operating Characteristic) para los datos del test set. Esta curva es un indicador de que tan bueno es el modelo en su capacidad de predicción. El área bajo la curva ROC (la curva roja) es de 0.911, mientras más cerca de 1.0 mejor. Esta curva se construye con la tasa de positivos verdaderos contra la tasa de falsos positivos para varios puntos de corte. Esta es una de las principales herramientas para medir la exactitud de clasificadores binarios como en este caso (1=desertores, 0= no desertores).

La curva azul representa el comportamiento teórico del ajuste del modelo, que como se puede observar es muy cercana a la curva roja mostrando un gran bondad en el ajuste.

GRAFICO 8



D. PREDICCIÓN DEL MODELO (VERIFICACIÓN)

En la tabla 8 siguiente se muestran los resultados de usar el modelo para predecir los estudiantes que desertarían en el periodo 2013-2014, y que iniciaron el ciclo en 2009-2010, excluyendo del análisis los que se graduaron (aprobaron el 4 grado de nivel medio). El mismo modulo se usara cuando se tengan los resultados del periodo 2014-2015 y la matriculación del 2015-2016.

TABLA 8. PRUEBA DEL MODELO CON LOS ESTUDIANTES DEL AÑO ACADÉMICO 2013–2014

ACCURACY:82.69%			
	true 1.0	true 0.0	class precision
pred. 1.0	3056	3939	43.69%
pred. 0.0	643	18827	96.70%
class recall	82.62%	82.70%	

La precisión del modelo fue del 82.69%, con una precisión del 82.62% en la predicción de los estudiantes que desertaron, y 3,939 falsos positivos. Los resultados de la predicción por niveles académicos básicos y medio se muestran en las siguientes tablas 9 y 10.

TABLA 9. PRUEBA DEL MODELO NIVEL BÁSICO DEL AÑO ACADÉMICO 2013–2014

ACCURACY:83.25%			
	true 1.0	true 0.0	class precision
pred. 1.0	1638	2451	40.06%
pred. 0.0	396	12513	96.93%
class recall	80.53%	83.62%	

TABLA 10. PRUEBA DEL MODELO NIVEL MEDIO DEL AÑO ACADÉMICO 2013–2014

ACCURACY:81.67%			
	true 1.0	true 0.0	class precision
pred. 1.0	1418	1488	48.80%
pred. 0.0	247	6314	96.24%
class recall	85.17%	80.93%	

La matriz siguiente (tabla 11) nos da una muestra de la información de predicción de cada estudiante del set de 20,000, que representan el conjunto de estudiantes de los 72 centros de los Alcarrizos que tienen registros desde el 2009, como predicción 2013-2014. En la columna $C-Deserción aparece el valor cero o uno indicando no deserción o deserción, al lado aparece la columna $CC-Desercion o "confidence" que indica el grado de riesgo de deserción y de retención dependiendo de si es cero o uno.

TABLA 11. MATRIZ DE RESULTADOS PREDICTIVOS 2013–2014

ID ESTUDIANTE	AÑO ACADEMICO	CODIGO CENTRO	TANDA	NIVEL_GRADO	CONDICIO ACADEMICA	SEXO	ICV1	INDICE PROSOLI CENTRO	INDICE POBREZA CENTRO	EDAD ESTUDIANTE	PROMOVIDO_SUM	REPROBADO_SUM	ABANDONO_SUM	CONDICION_OTRA	CAMBIO_CENTRO	N_ANIOS_ACAD	$R-DESERCION	$RC-DESERCION
35	2013-2014	359	MATUTINA	14	Promovido	Masculino	0	8.2	69.88	8.5	2	0	0	0	0	2	1	79.50%
134	2013-2014	226	VESPERTINA	17	Promovido	Femenino	1	24.3	37.04	12.53	5	0	0	0	0	5	0	59.50%
258	2013-2014	3061	VESPERTINA	15	Promovido	Femenino	1	12.5	36.32	10.02	3	0	0	0	0	3	1	75.80%
356	2013-2014	230	MATUTINA	16	Promovido	Femenino	0	13.5	37.04	11.02	5	0	0	0	0	5	0	71.00%
429	2013-2014	224	VESPERTINA	12	Promovido	Masculino	0	20.01	43.31	9.57	2	1	1	0	0	4	0	82.70%
1605	2013-2014	5710	MATUTINA	21	Promovido	Femenino	1	29.6	35.75	15.67	5	0	0	0	0	5	0	75.50%
1637	2013-2014	229	MATUTINA	16	Promovido	Femenino	0	10.5	37.04	11.79	2	0	0	0	0	2	1	72.90%
1777	2013-2014	13402	VESPERTINA	21	Promovido	Femenino	1	20.01	43.31	14.3	4	0	0	0	0	4	0	75.10%
1800	2013-2014	4852	MATUTINA	17	Promovido	Femenino	0	7.7	22.24	12.32	3	0	0	0	0	3	1	75.90%
1959	2013-2014	528	VESPERTINA	16	Promovido	Masculino	0	18	59.59	10.62	3	1	0	0	0	4	0	59.50%
1976	2013-2014	2251	MATUTINA	17	Promovido	Masculino	0	25.7	60.42	12.63	4	0	0	0	0	4	1	72.80%
2005	2013-2014	216	VESPERTINA	16	Promovido	Masculino	0	20.01	43.31	10	4	0	0	0	0	4	0	73.80%
2062	2013-2014	12	MATUTINA	15	Promovido	Masculino	1	10.6	35.74	12.27	4	0	0	0	0	4	0	68.90%
2113	2013-2014	6075	VESPERTINA	17	Abandono	Femenino	1	10.1	72.23	11.58	3	1	1	0	0	5	1	80.60%
2231	2013-2014	216	VESPERTINA	15	Promovido	Femenino	1	20.01	43.31	9.85	5	0	0	0	0	5	0	78.00%
2937	2013-2014	5872	VESPERTINA	24	Promovido	Masculino	1	27.6	72.23	16.86	5	0	0	0	0	5	0	79.80%
2985	2013-2014	5735	NOCTURNA	23	Promovido	Femenino	0	20.01	43.31	16.28	4	1	0	0	0	5	0	63.20%
3017	2013-2014	5738	MATUTINA	23	Promovido	Femenino	0	11.1	37.04	15.91	5	0	0	0	0	5	0	69.40%

ID ESTUDIANTE	AÑO ACADEMICO	CODIGO CENTRO	TANDA	NIVEL_ GRADO	CONDICIO ACADEMICA	SEXO	ICV1	INDICE PROSOLI CENTRO	INDICE POBREZA CENTRO	EDAD ESTUDIANTE	PROMOVIDO_SUM	REPROBADO_ SUM	ABANDONO_SUM	CONDICION_ OTRA	CAMBIO_ CENTRO	N_ ANIOS_ ACAD	$R-DESERCION	$RC-DESERCION
3169	2013-2014	4286	VESPERTINA	17	Promovido	Masculino	0	36.6	28.22	11.95	3	0	0	0	0	3	1	81.40%
3348	2013-2014	228	MATUTINA	17	Promovido	Femenino	1	14.6	37.04	11.27	5	0	0	0	0	5	0	67.70%
3378	2013-2014	5726	MATUTINA	18	Promovido	Femenino	0	21.4	35.75	11.41	5	0	0	0	0	5	1	70.60%

E. PRUEBAS DE PRECISIÓN PREDICTIVA DEL MODELO

En los modelos de deserción escolar se presenta el fenómeno de que la proporción de la especifity (los verdaderos negativos) como la proporción de sensitivity o recall (verdaderos positivos) juega un rol especial, en el sentido de que la cantidad de deserciones (1s o verdaderos positivos) es muy inferior a la de no deserciones (0s o verdaderos negativos). En el caso bajo estudio es de un promedio del 10% del total de observaciones. Esto induce el riesgo de que el modelo no pueda identificar correctamente los desertores convirtiéndolos en falsos positivos en una proporción muy alta y por otro lado no identificar una proporción muy alta de no desertores (verdaderos negativos).

La idea es minimizar tanto los falsos positivos (Error tipo I) como los falsos negativos (Error Tipo II), relacionándolos entre sí mediante el accurarcy, que es la proporción de true positive más true negativo en relación al total observado. Para lograrlo el modelo ha tenido que equilibrar, mediante la aplicación de técnicas de boostrapping y balanceo de la muestra, con el objeto de hacer un buen reconocimiento o aprendizaje de las características de las deserciones para poder aumentar la exactitud (accuracy) de ambos grupos.

Por otro lado, este fenómeno contribuye a sesgar los resultados de la totalidad de 1s predichos (true positive) en relación a los falsos negativos, lo que no puede ser detectado mediante la medida de Accuracy, requiriéndose por ende una medida de precisión relativa. Para detectar este aspecto las medidas tradicionales para el ajuste en tablas de contingencias de clasificación binaria, para pruebas de contraste, como lo es la Chi cuadrado, no suelen ser muy efectivas. Por tal razón se ha creado una puntuación o scoring de medición de la prueba denominado F1 Score.

En el análisis estadístico de clasificación binaria, la puntuación F1 (también F-Score o el F-medida) es una medida de la precisión de una prueba. Se considera tanto la precisión como la sensitivity (recall) de la prueba para calcular la puntuación. Precisión es el número de resultados positivos correctos (true-positive) dividido por el número de todos los resultados positivos predichos o estimados, y mide la proporción de positivos correctos en relación a los negativos falsos predichos. Por otro lado, el sensitivity o recall es el número de resultados positivos correctos (true-positive) dividido por el número de positivos observados. La puntuación de F1 se puede interpretar como un promedio ponderado de la precisión y el sensitivity o recall, donde una puntuación de F1 alcanza su mejor valor en 1 y el peor a 0. La fórmula es como sigue:

$$F_1 = 2. \frac{precision \,.\, recall}{precision + recall}$$

Para el caso de la predicción del año 2013-2014 (ver tablas 12a y 12b siguientes) de la deserción de estudiantes, las medidas de Specificty, Sencitivity y Accuracy son altas (alrededor de un 83%) lo que nos provee un alerta verde, indicando una muy buena exactitud o Accuracy, debido a la gran potencia predictiva del modelo mostrado en la sección anterior.

Sin embargo si observamos la medida de precisión notamos que es de un 44% en rojo (ver cuadro siguiente). Esto significa que la proporción de falsos negativos es superior a la de verdaderos positivos en la predicción. La medida F1 es de 0.57 ligeramente por encima del 50%, dándole una precisión de regular a la predicción. Un esfuerzo adicional en la modelación podría bajar los estudiantes falsos negativos (3,939) y así aumentar la precisión.

Desde el punto de vista de las políticas de retención educativas se requiere de un mayor esfuerzo de tra-

bajo con los 6,995 estudiantes en condición de riesgo de deserción predicho. No obstante, la respuesta a esta actividad será positiva puesto que en ellos están contenidos casi la totalidad de los 1s (true-positive desertores verdaderos), es decir 3,056 de los 3,699 observados, debido a la exactitud del modelo y los 3,939 alumnos en falso negativo tienen bajo riesgo de deserción por ser 0s en la realidad.

Igual análisis se hace para la predicción de la deserción en el nivel básico y medio, vistos separadamente, donde se hereda lógicamente este patrón.

TABLA 12A.
CONTIGENCY TABLE: PREDICCIÓN 2013–2014

DECISION TREE FOR SCHOLLAR CHURN			PREDICTED			
			Deserción			Percentage Correct
			0	1	Totals	
Observed	Deserción	0	18,827	3,939	22,766	82.70%
		1	643	3,056	3,699	82.62%
	Totals		19,470	6,995	26,465	82.66%
		Percentage Correct	96.70%	43.69%	70.19%	82.69%

TABLA 12B

LEYENDA		INDICADORES DE RESULTADOS DE PRUEBA		
10	3,939	False Positive	82.70%	Specificity=True Neg/#Observed Neg
11	3,056	True Positive	82.62%	Sensitivity(Recall)=True Pos/#Observed Pos
01	643	False Negative	82.69%	Accuracy=%Overall Predicted (True Positive+ True negative)
00	18,827	True Negative	43.69%	Precision=True Pos/#Predicted Pos
			0.5715	F1 Score

GESTIÓN DEL MODELO Y ANÁLISIS DE DESVIACIÓN

Este proyecto presenta una gran ventaja desde el punto de vista del monitoreo, y es que cada año cuando se va a hacer la predicción de la deserción por estudiante, se toman los datos más recientes para recalibrar el modelo.

El modelo tiene incorporado un proceso de optimización que ayuda a la mejor determinación de los parámetros del algoritmo de clasificación-predicción (árbol de decisión), y entre los reportes que genera el mismo, están las tablas de contingencia tanto para el training set como para el test set.

Cuando los resultados de esas dos tablas no sean satisfactorios, entonces es momento de revisar el modelo, y en este caso, pudiera limitarse a cambiar el algoritmo de predicción, lo cual requerirá de conocimientos más especializados para poder ajustar el mejor algoritmo.

Por otro lado, se puede medir con los resultados del año escolar la precisión en la predicción del modelo usado para predecir la deserción, comparando para cada estudiante de la base de datos, la predicción con el resultado real. Esta sería la precisión de la predicción con los resultados reales. Si el patrón de comportamiento de los estudiantes no ha variado, no debe haber mucha diferencia entre ese resultado y lo que se obtuvo en el test set de los datos usados en el desarrollo del modelo. Se debe especificar un valor para la precisión (ejemplo 80%), tal que, si el resultado de la prueba es menor que ese número, se proceda a reajustar el modelo.

CONCLUSIÓN

Este proyecto se ha realizado sobre una población escolar distrital cómo estudio piloto, conformada por 72 escuelas públicas de nivel básico y medio de Los Alcarrizos, consideradas de importancia para los propósitos del Ministerio de Educación e IDEICE y de acuerdo a los planes estratégicos elaborados y la disponibilidad de información existente.

La recomendación es que este modelo sea replicado gradualmente en las restantes provincias y distritos escolares del territorio nacional en posteriores proyectos. Esta gradualidad permitiría tener mayor efectividad y control en el alcance de los objetivos (a diferencia de hacerlo a escala nacional de una vez) y actúa como modelo de efecto demostración, acompañado de acciones específicas sobre los centros del región o distrito escolar elegido.

Estos datos deberán ser complementados mediante el uso de la base de datos demográfica y de resultados escolares del sistema de Gestión de Centros Educativos del MINERD. La clasificación por nivel de carencia de los hogares de los alumnos y alumnas, según el Índice de Condiciones de Vida (ICV) está registrado en esta base de datos usada para la emisión de la Tarjeta Solidaridad del PROSOLI, para los estudiantes con un nivel de pobreza alto. Por otro lado, se ha de considerar el recién creado, aún en fase experimental, Índice de Vulnerabilidad a nivel de hogar, auspiciado por PNUD-SIUBEN, aplicado al mapa educativo nacional de las 10 regiones y sus distritos escolares.

Es importante destacar que cada situación de estudio responde a un modelo específico algorítmico adaptado a la realidad de información de cada región geográfica por sus características socioeconómicas y ambientales. De aquí que el estudio de los 72 centros de los Alcarrizos es considerado un conglomerado poblacional particular y no se debe realizar ninguna inferencia, generalización o expansión de este modelo a otros distritos escolares del sistema educativo nacional. Es decir, cada región escolar, debe ser considerada una población particular propensa de generar un modelo particular predictivo. De igual manera que por adoptar estas técnicas de aprendizaje automático no paramétrico, no se presume ningún tipo de comportamiento de las variables envueltas ni de su distribución de probabilidad, hecho que se recoge en el modelo de data mining exploratorio construido mediante el set de entrenamiento de datos.

REFERENCIAS

Amaya, Y., Barrientos, E. & Heredia, D. (2014). *Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos. Colombia*: Universidad Francisco de Paula Santander, Universidad Simón Bolívar.

Duda, R.O., Hart, P.E. & Stock, D.G. (2001).P*attern Classification (2nd. Ed.)*. New York, NY: John Wiley & Sons. Inc.

Han J., Kamber M. & Pei J. (2012). *Data Mining – Concepts and Techniques (3th Edition)*. Waltham, MA, USA: Elseiver Inc.

Hofmann M. & Klinkenberg R. (2013). *RapidMiner, Data Mining Use cases and Business Analytics Applications*. Boca Raton, FL, USA: Chapman and Hal/CRC Press.

ONE. (2014). Boletín Panorama Estadístico, Santo Domingo, Rep. Dom. : ONE.

Rockach L. & Maimon O. (2008). *Data Mining with Decision Trees, Theory and Applications.* Singapore: World Scientific Publishing Co.

SPSS Inc. (2008). *Introducción al SPSS Modeler y Data Mining*. Chicago, IL, USA: SPSS.

SPSS Inc. (2003). *Modelación Avanzada con IBM SPSS Modeler*. Chicago, IL, USA: SPSS.

Theodoris, S. & Koutroumbas, K. (2006). Pattern Recognition (3th Edition). San Diego, CA, USA: Academic Press.

UNESCO. (2009). *Education Indicators. Technical Guidelines*. UNESCO Institute of Statistics.

University of Minnesota. (2013). *Essential Tools-Increasing Rates of School Completion: Moving From Policy and Research to Practice-A Manual for Policymakers, Administrators, and Educators*. USA: College of Education and Human Development.

Valero Orea, S., Salvador Vargas, A. & García Alonso, M. (2014). *Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos.*

Witten, I.H. & Frank, E. (2000). *Data Mining.* San Diego, CA, USA: Morgan Kaufmann Publishers.