DESARROLLO DE UN MODELO PREDICTIVO DE DESERCIÓN ESCOLAR

Santo Domingo, República Dominicana.
2018

Desarrollo de un Modelo Predictivo de Deserción escolar

Santo Domingo,
República Dominicana
2018

DESARROLLO DE UN MODELO PREDICTIVO DE DESERCIÓN ESCOLAR

Consultor
Renato R. González Disla

Coordinadora de proyecto IDEICE-PNUD
Rita Licelot Cruz

Corrección de estilo
Ramón Fari Rosario

Diseño y Diagramación
Yeimy Rosa Olivier Salcedo
Natasha Mercedes Arias

Centro de Gestión de la Información y Documentación
Dilcia Armesto Núñez

ISBN
978-9945-499-36-0 (impreso)
978-9945-499-37-7 (digital)

Santo Domingo, D.N.
República Dominicana

Danilo Medina Sánchez
Presidente de la República

Margarita Cedeño de Fernández
Vicepresidenta de la República

Andrés Navarro García
Ministro de Educación

Denia Burgos
Viceministra de Educación, Encargada de Servicios Técnicos y Pedagógicos

Freddy Radhamés Rodríguez
Viceministro de Educación, Encargado de Asuntos Administrativos y Financieros

Manuel Ramón Valerio Cruz
Viceministro de Educación, Encargado de Certificación Docente

Víctor Ricardo Sánchez
Viceministro de Educación, Encargado de Planificación y Desarrollo Educativo

Adarberto Martínez
Viceministro de Educación, Encargado de Supervisión y Evaluación de la Calidad Educativa

Luis de León
Viceministro de Educación, Encargado de Descentralización

Julio Leonardo Valeirón
Director Ejecutivo del Instituto Dominicano de Evaluación e Investigación de la Calidad Educativa

Contenido

1. Introducción

1.1 Antecedentes del proyecto

El propósito del Proyecto de "Desarrollo de un Modelo Predictivo de Deserción Escolar mediante Técnicas de Minería de Datos" es elaborar e implementar un modelo y sistema predictivo de deserción escolar para los ciclos Básico y Medio del Sistema Educativo nacional público, con el propósito de contribuir a la elaboración de políticas y planes de acción encaminados 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.

El proyecto tiene como objetivos específicos:

1. Desarrollar un sistema para predecir qué tipo de estudiantes tienen riesgo de deserción escolar antes de concluir los ciclos de educación Básica y Media, determinando las variables explicativas de mayor incidencia y una función de riesgo de fracaso escolar.
2. Generar una base de datos para poder seleccionar grupos de alumnos y alumnas que tengan carencias educativas y condiciones socio-económicas, demográficas y de género específicas (vulnerabilidad ambiental y a nivel de hogar) que representen un vector de riesgo significativo en el fracaso escolar.
3. Determinar qué tipo de escuelas o centros educativos, poseen el mayor índice de riesgo de deserción, tomando en cuenta su ubicación geográfica, condiciones cualitativas docentes, condiciones físicas del plantel, etc.

Estos objetivos se enmarcan dentro de la línea de investigación del Programa de Procesos y Logros de Aprendizajes y Factores asociados a los logros de aprendizaje del Proyecto 00077918: "Fortalecimiento institucional y operativo del IDEICE con miras a contribuir a la mejora de la evaluación e investigación de la calidad educativa dominicana".

Para llevar a cabo estos objetivos se han contratado los servicios de consultoría de quien suscribe, de parte de IDEICE-PNUD, mediante el contrato No. 010-2015.

Este tercer informe se corresponde con el **Producto 3-Final**, establecido por el contrato en su Artículo 5 y descrito en el documento de "Convocatoria a presentación de Propuestas de Investigación de IDEICE", que en el capítulo **XII. IDENTIFICACION DE LOS PRODUCTOS ENTREGABLES DE LA INVESTIGACION** define:

PRODUCTO NO. 3: Presentar el informe en su versión "final" del trabajo realizado, el cual debe incluir un esquema que abarque y dé respuesta a los objetivos de la investigación, con una descripción sobre la metodología aplicada y la descripción del trabajo realizado. Asimismo el informe ha de contener una presentación de los resultados y un análisis detallado de los mismos con gráficos y tablas que faciliten su comprensión e interpretación.

En este informe, de acuerdo al cronograma del proyecto (anexo I), se expone la fase de modelación, prueba e implementación del modelo de deserción escolar. En el capítulo 2 el informe contiene una sección que explica el contexto del problema, los objetivos y alcances del mismo así como la conceptualización de solución usando tecnología de data *mining.* En el capítulo 3 se explica la metodología a ser usada en el proyecto y el ciclo de vida del modelo. En el capítulo 4 se explica el proceso de extracción y preparación de datos que fue profundizado en el

Informe 2 de esta consultoría. El capítulo 5 contiene la explicación del proceso de modelación y desarrollo del modelo predictivo usando las herramientas de SPSS Modeler y RapidMiner. El capítulo 6 contiene el análisis de los resultados de aplicar el modelo predictivo con la data de la cohorte 2009-2014, usada para entrenamiento, prueba y verificaciones del modelo. El capítulo 7 contiene un resumen de los resultados y procesos ejecutados del modelo, con las recomendaciones de los próximos pasos a realizar en el sistema educativo nacional. Incluimos la bibliografía usada en el proyecto y los anexos con los archivos concernientes a resultados.

- En este proyecto hemos contado con la colaboración estrecha del equipo técnico del Viceministerio de Planificación Educativa del MINERD, en especial del equipo de la Dirección de Información, quienes nos proveyeron de los datos e informaciones del historial académico de los estudiantes y centros del Distrito escolar de Los Alcarrizos, dichos datos se usaron como modelo piloto para la cohorte seleccionada. El Departamento de Estadística que nos facilitó su tiempo e informaciones para la evaluación de las tasas de deserción escolar calculadas. La consultora Damaris Lara fue quien nos facilitó la base de datos de índice de pobreza de centros, resultante del Estudio sobre uso del tiempo en los Centros Educativos Dominicanos (EDUCA-UE, 2014).
- En el proceso de desarrollo del modelo predictivo de deserción escolar he contado con la importante colaboración del consultor asociado Ing. Felipe Llaugel.

1.2 Términos y Definiciones Relevantes

- **Condición académica del alumno(a):** Representa el estado del estudiante al finalizar el año escolar y posterior a las evaluaciones. Estas son:
  - Abandono (intra anual): alumno que se retira de la escuela y que no finaliza el año escolar, pero puede o no inscribirse el año siguiente.
  - Reprobado: Alumno que permanece en el mismo grado por más de un año.
  - Promovido: alumno que completa exitosamente un grado y prosigue al finalizar el año escolar.
  - Otras: otras condiciones académicas que posteriormente se traducen a una de las anteriores después de corregida la situación que la determina.
- **Deserción escolar:** Alumno que en cualquiera de los estados de condición académica anterior que interrumpe un año escolar y luego no continúa al año siguiente (queda fuera del sistema escolar). Para este estudio se considera deserción el alumno que habiendo cursado un año escolar no continúa en ninguno de los años siguientes dentro del rango de estudio 2009-2015.
- **La tasa de deserción escolar** en el nivel Básico y Media estima cuántos alumnos y alumnas de cada cien inscritos en un año escolar y que no concluyeron el nivel no se inscribieron en el siguiente año para continuar su educación. El indicador es una medida aproximada de la deserción escolar entre dos años escolares consecutivos (definida más adelante).
- **Riesgo de deserción escolar** es la proporción o porcentaje que se obtiene al dividir la deserción de alumnos de un grado específico en un año escolar entre el total de matriculados en ese año. El riesgo acumulado a un grado específico a un nivel (Básico o Media) es la frecuencia acumulada de deserciones hasta ese grado o nivel dividido entre el total de matriculados en ese año escolar.
- **Reincorporación escolar:** todo alumno que habiendo desertado se reintegra al sistema en uno o más años posteriores.
- **Retención escolar:** es la capacidad que tiene el sistema para retener los alumnos en las aulas bajo cualquiera de las condiciones académicas que posea al finalizar el año escolar.

2. Contexto del Problema, Objetivos y Alcances

2.1 Contexto del problema

El abandono y la deserción escolar es un indicador que busca medir el fenómeno provocado por los alumnos y alumnas que dejan sus estudios antes de concluirlos. A pesar de los avances alcanzados con relación 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 pre-universitarios, 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 y alumnas que desertan antes de concluirlo es alrededor de un 16.0%.

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 con 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 con factores demográficos, condiciones ambientales, condiciones del sistema educativo como tal y de los centros educativos en particular.

Los principales desafíos que enfrenta la República Dominicana para avanzar hacia el logro de los Objetivos de Desarrollo del Milenio y para el cumplimiento de las metas en el año 2015, son evitar que los niños y niñas abandonen la escuela antes de terminar el ciclo Básico y disminuir lo más posible la deserción en el ciclo Medio. (Tomado del *Boletín* de la ONE, P*anorama Estadístico*, enero del 2014).

La asignación del 4% del PIB para la educación, que se traduce en planes para la inversión en mejoras de los planteles escolares, de condiciones curriculares nuevas, de capacitación de maestros, desayuno escolar, tanda extendida, debe contribuir al mejoramiento de las condiciones generales relativas a la 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 de contribuir a reforzar, de forma directa, las condiciones para el desarrollo cognoscitivo, psicológico, material y familiar de los niños y niñas ya que son factores humanos relevantes para frenar la problemática y producir el efecto de retención en el sistema educativo nacional.

Para estos propósitos es importante tener 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 alumnado analizando la historia académica del estudiante junto a los factores socio económicos y ambientales, que también 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 intervenciones 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.

2.2 Objetivos e hipótesis

GENERAL(ES)	ESPECÍFICO(S)
Desarrollar un modelo y sistema predictivo de deserción escolar para los ciclos Básico y Medio con el propósito de contribuir a la elaboración de políticas y planes de acción encaminados a la retención de los alumnos y alumnas en el sistema educativo nacional y al mejoramiento de los procesos de los centros educativos en pro de la disminución del índice de deserción escolar.	Desarrollar un sistema para predecir qué tipo de estudiantes tienen riesgo de deserción escolar antes de concluir los ciclos de educación Básica y Media, determinando las variables explicativas de mayor incidencia y una función de riesgo de fracaso escolar.
	Generar una base de datos para poder seleccionar grupos de alumnos y alumnas que tengan carencias educativas y condiciones socio-económicas, demográficas y de género específicas (vulnerabilidad ambiental y a nivel de hogar) que representen un vector de riesgo significativo en el fracaso escolar.
	Determinar qué tipo de escuelas o centros educativos, poseen el mayor índice de riesgo de deserción, tomando en cuenta sus factores ambientales y niveles de pobreza.

Las hipótesis fundamentales del estudio son:

Además de los factores demográficos **(edad, genero, etc.) del estudiante**

1. Los niveles de vulnerabilidad socio-económicos del hogar son factores determinantes en el abandono escolar de los alumnos y alumnas del ciclo Básico y Medio en las escuelas públicas.
2. La desigualdad de género es un factor social que empuja 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.
3. Las condiciones del centro educativo (físicas, ambientales, recursos, gestión de centros, etc.) y los procesos de enseñanza (programas, capacitación de maestros y maestras, materiales educativos, laboratorios, etc.) son factores que determinan el rendimiento escolar, los resultados de los alumnos y alumnas, la esperanza de vida escolar y la deserción escolar.

2.3 Alcance del proyecto

Se ha determinado una población escolar distrital para el estudio piloto, conformada por escuelas públicas de nivel Básico y Medio, consideradas de importancia para los propósitos del Ministerio de Educación y para el IDEICE de acuerdo con los planes estratégicos elaborados y la disponibilidad de información existente.

Este modelo podrá ser replicado gradualmente en las restantes provincias del territorio nacional en posteriores proyectos. Esta gradualidad permite 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 piloto y como efecto de demostración, acompañado de acciones específicas sobre los centros del Distrito escolar elegido.

Además cada Distrito escolar lo constituye una población estudiantil y un conjunto de centros con sus propias características diferenciables de los demás distritos escolares, por las condiciones geográficas, socio-económicas y ambientales del sector, lo que podríamos definir como un conglomerado propenso a una modelación particular.

A partir de la discusión de los objetivos del proyecto fue seleccionado el Distrito escolar de Los Alcarrizos en el cual se incluyeron 72 centros educativos públicos pertenecientes al nivel Básico y Medio del país. Se observó la información de los centros educativos y escuelas públicas y se han incluido inicialmente todos los alumnos y alumnas de estos centros, con el objeto de integrar al modelo los datos de factores demográficos y educativos registrados en el Sistema de Gestión de Centros (SIGERD), que contiene la base de datos del Registro Estudiantil del MINERD y la base de datos de centros educativos.

Estos datos serán complementados mediante el uso de la base de datos demográfica y los 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 la 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 tomó en consideración el recién creado, aún en fase experimental, del Índice de Vulnerabilidad a nivel de hogar, auspiciado por PNUD-SIUBEN, que fue obtenido a partir del estudio sobre uso del tiempo en los Centros Educativos Dominicanos (EDUCA-UE 2014).

En este estudio de EDUCA, para lograr sus objetivos, se estableció un índice de vulnerabilidad de centros educativos a nivel nacional, que fue usado para clasificar los 72 centros de Los Alcarrizos bajo estudio y como atributo de vulnerabilidad del estudiante como variable explicativa en el modelo predictivo.

Adicionalmente, la elección de esta población del sistema educativo del Distrito de los Alcarrizos se correspondió con los siguientes criterios:

1. Este Distrito cuenta con centros educativos públicos cuya incidencia de educandos provienen de un sector con alta vulnerabilidad socioeconómica. El 80% de los centros pertenecen a la zona urbana-marginal.
2. Los centros educativos están provistos de equipamiento informático y de capacitación del personal docente y administrativo en el uso de dichos sistemas.
3. Existe un alto grado de completitud de la información académica de los educandos y la identificación de los estudiantes y centros de según su nivel de vulnerabilidad y el impacto de las condiciones de pobreza.

Los años escolares para la cohorte de estudio son 2009-2010, como año inicial, hasta el 2013-2014, este último como marco de matriculación actual y los 4 restantes como referencia de historial académico de los estudiantes (con condición académica). Se tiene como referencia de población activa a todos los estudiantes que en algún momento de estos años escolares hayan estado matriculados en algún centro de los 72 centros seleccionados del Distrito de Los Alcarrizos, Santo Domingo Oeste.

El punto de referencia inicial, o año base 2009-2010, se usó por entender, según recomendaciones técnicas del MINERD, que en ese año se inició un registro riguroso con un sistema de información automatizado, lo cual garantiza un buen margen de calidad de los datos, factor vital para este proyecto.

2.4 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. La aplicación de algoritmos de minería de datos requiere de actividades previas destinadas a preparar los datos de manera homogénea. Esta primera etapa es también conocida como ETL (*Extract, Transform and Load*). Un proceso completo de aplicación de técnicas de minería, mejor conocido como proceso de *descubrimiento del conocimiento (KDD)* en bases de datos lo que establece la minería de datos como una etapa del mismo.

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 minería de datos, y pertenecen a todo el proceso KDD como pasos importantes. El modelo de datos de entrada elaborado para producir el modelo de minería de datos, se entiende como el entrenamiento al 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.

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):

- 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).
- 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 de forma conjunta (modelos multivariados) de los elementos bajo estudio.

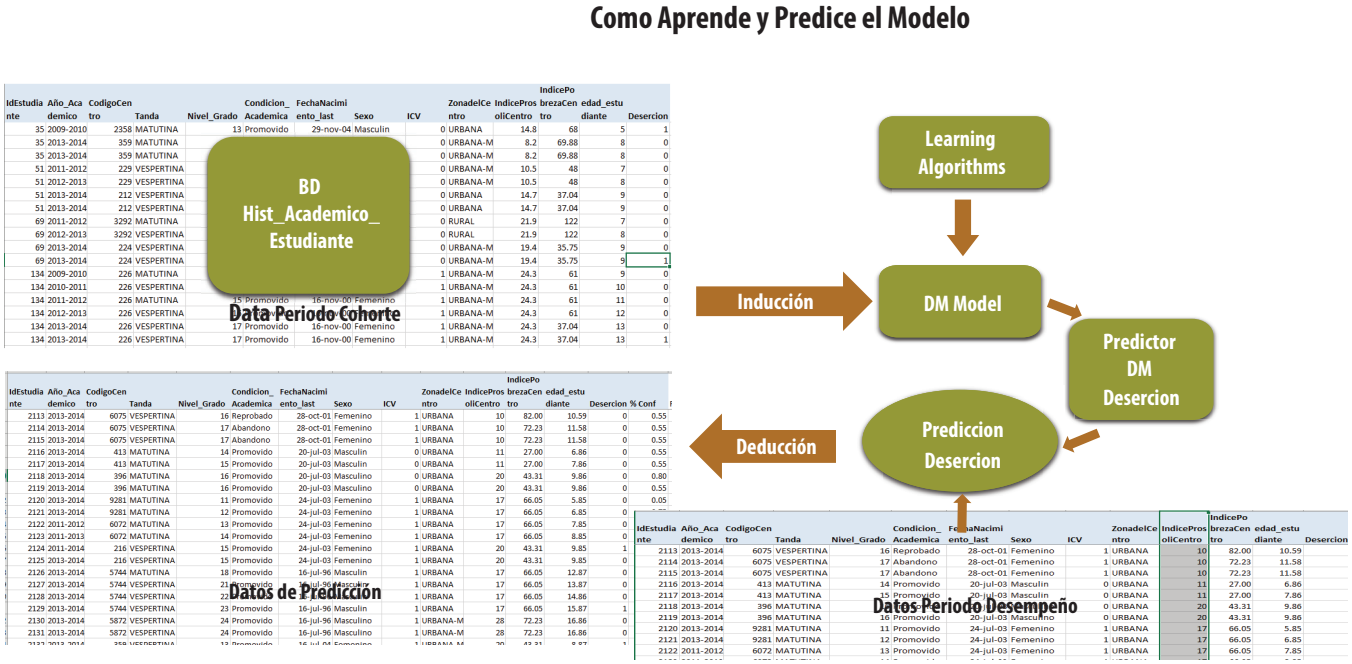
Las técnicas más representativas son:

- **Redes neuronales:** Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.
- **Regresión lineal y logística:** Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.
- **Reglas de asociación:** Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.
- **Árboles de decisión:** Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial y el análisis predictivo, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.
- **Modelos estadísticos:** Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.
- **Agrupamiento o Clustering:** Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes.

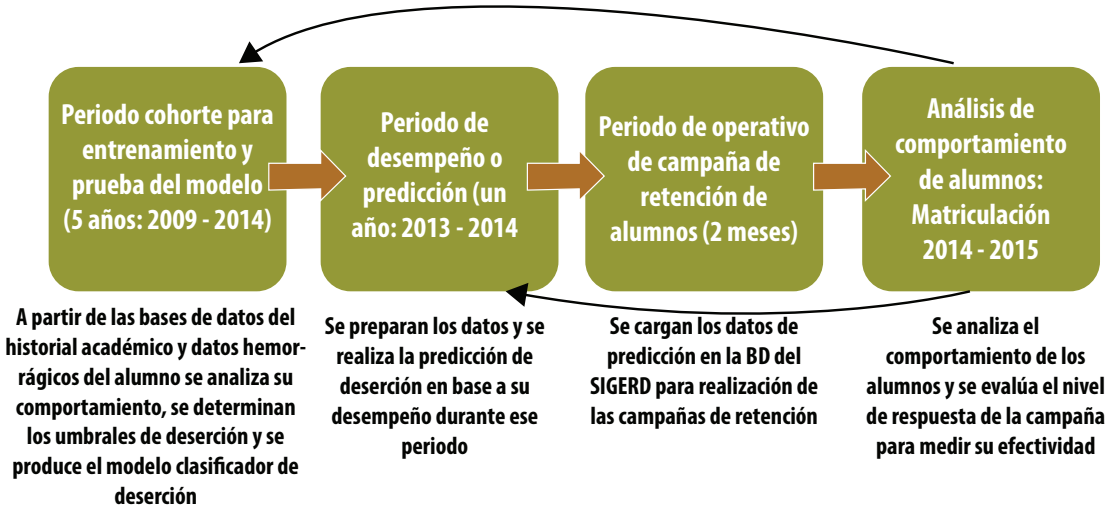
Los modelos predictivos de deserción se inscriben como técnicas de minería de datos supervisadas no-paramétricas, donde la variable respuesta u objetivo es binaria (deserción o no deserción) y cuyo valor está dado en función de un conjunto de variables observables, denominadas explicativas o "predictoras". La serie histórica de estos registros determinan 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 período siguiente al período 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 período determinado futuro.

La aplicación de los algoritmos predictivos de minería de datos a la deserción escolar es una actividad analítica de aplicación reciente. En el sistema educativo nacional se ve la necesidad de contar con mecanismos que ayuden a determinar la deserción de los estudiantes en los centros educativos, antes de que suceda el evento. Es por esto que este estudio se basa en el análisis de técnicas que utilizan los diferentes modelos predictivos con la finalidad de buscar la que mejor se comporte para predecir la probabilidad de que un estudiante deserte, comprobándose con el menor margen de error.

El diagrama siguiente explica este proceso:



Las variables explicativas usadas son de tipo académica de rendimiento escolar del grado o nivel cursante del alumno(a) y su condición académica final de cada período (promoción, reprobación, abandono, etc.), sus factores demográficos (edad, sexo, etc.), sus factores socioeconómicos (índice de pobreza del hogar, vulnerabilidad del sector geográfico, etc.). La variable respuesta, objetivo o dependiente es la condición de deserción (1) o no deserción (0), observada durante el período de cohorte. Como se nota en el diagrama anterior se selecciona un período de colección de datos o cohorte (historial de varios años académicos) que sirven de entrenamiento y prueba del modelo. Estamos usando 2009-2014 como período de cohorte, 70% *dataset* de entrenamiento y 30% *data set* para prueba del modelo (véase diagrama siguiente).



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 el de prueba, en una tabla de contingencia con pruebas estadísticas de significación (usaremos la exactitud y *F1 Score* como medidas de precisión del modelo). Luego se procede a probar el modelo simulando un período académico próximo, donde aplicaremos dichas prueba de igual manera.

Se usa el último año académico 2013-2014 como período 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 comprobada 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.

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

a) **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 período de tiempo específico para predecir su efecto o no en el paciente. También el enfoque de regresión logística binomial usado como clasificador binario.

b) **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.

c) **El enfoque de reglas asociativas predictivas** basado en árboles de decisión y en reglas de dependencia que dependen del contenido de información del modelo de datos de entrada y categorizan una serie de condiciones que suceden de forma sucesiva. En esta técnica se identifican los árboles de decisión binarios, árboles CHAID, Árboles C5.0, redes neuronales, reglas bayesianas, y otros.

Como veremos, 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*, descritas en la sección 5.4.

En Amaya-Barrientos-Heredia (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 algorítmico adaptado a la realidad de información de cada país o región. De aquí se desprende que el estudio de los 72 centros de Los Alcarrizos es considerado un conglomerado poblacional particular y no pretende realizarse 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 modo, 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.

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. Es una aplicación de software orientada a la solución de problemas mediante técnicas predictivas y de minería de datos. Proporciona un rango de algoritmos y técnicas avanzadas, incluidos el análisis de texto, el análisis de entidad, la gestión y optimización de decisiones, para ayudarle a seleccionar las acciones que dan como resultado un mejor resultado. Disponible en varias ediciones, incluida una versión basada en la nube, SPSS Modeler puede escalar desde despliegues de escritorio a la integración dentro de los sistemas de redes.

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. En una encuesta realizada por *KDnuggets*, un periódico de minería de datos, RapidMiner ocupó el segundo lugar en herramientas de analítica y de minería de datos utilizadas para proyectos reales en 20091 y fue el primero en 2010. La versión inicial fue desarrollada por el departamento de inteligencia artificial de la Universidad de Dortmund en 2001. Se distribuye bajo licencia AGPL y está hospedado en SourceForge desde el 2004. RapidMiner proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, pre procesamiento de datos y visualización. También permite utilizar los algoritmos incluidos en *Weka*.

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. Mientras que la aplicación del modelo usando el algoritmo seleccionado (árbol de decisión) se ha realizado usando la herramienta RapidMiner. 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, como sería la aplicación del modelo a nivel regional o nacional del sistema educativo nacional, puede pensarse en el uso de SPSS Modeler como herramienta de mayor potencia de manejo de datos.

3. Metodología del Proyecto y Ciclo de Vida del Modelo

3.1 Metodología de gestión del proyecto

La minería de datos sigue una metodología de ciclo de vida de proyectos con características particulares según el tipo de herramienta y objetivos de proyectos definidos en el alcance del proyecto. Se ha establecido un ciclo de proyecto basado en la metodología **CRISP-DM**.

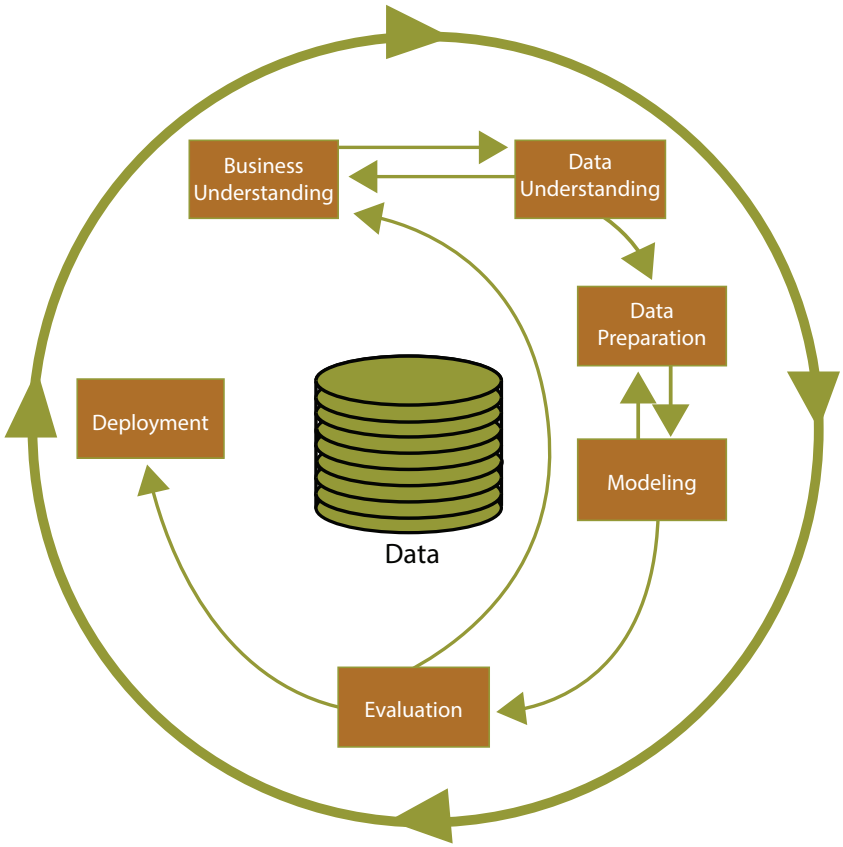
CRISP-DM, de Cross Industry Standard Process for Data Mining, trata de un modelo de proceso de minería de datos que describe los enfoques comunes que utilizan los expertos en minería de datos. Encuestas realizadas en 2002, 2004 y 2007 muestran que es la principal metodología utilizada para esta tarea. 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 factor para el desarrollo de la minería de datos y los proyectos de descubrimiento de conocimiento".

CRISP-DM fue concebido en 1996. En 1997 se puso en marcha como un proyecto de la Unión Europea bajo la iniciativa de financiación ESPRIT. El proyecto fue dirigido por cinco empresas: SPSS, Teradata, Daimler AG, NCR Corporation y Ohra, una compañía de seguros.

Este consorcio original trajo diferentes experiencias al proyecto: ISL, más tarde adquirió y se fusionó con SPSS, Inc. El gigante de la informática NCR Corporation produjo el data *warehouse de Teradata* y su propio software de minería de datos. Daimler-Benz tenía un equipo de minería de datos relevante. OHRA estaba empezando a explorar el uso potencial de la minería de datos. La primera versión de la metodología se presentó en el cuarto CRISP-DM SIG taller en Bruselas en marzo de 1999, y fue publicada, más tarde ese año, como una guía paso a paso de minería de datos.

Entre 2006 y 2008 se formó un 2.0 SIG CRISP-DM y hubo discusiones acerca de la actualización del modelo de proceso CRISP-DM. Aunque muchos de los profesionales de minería de datos que utilizan CRISP-DM no son colaboradores de IBM, sin embargo esta empresa lleva la voz cantante actualmente ya que promueve el modelo de proceso CRISP-DM; hace disponibles algunos de los viejos documentos CRISP-DM para su descarga y ha incorporado el modelo de proceso a su producto SPSS Modeler (Tomado de Wikipedia).

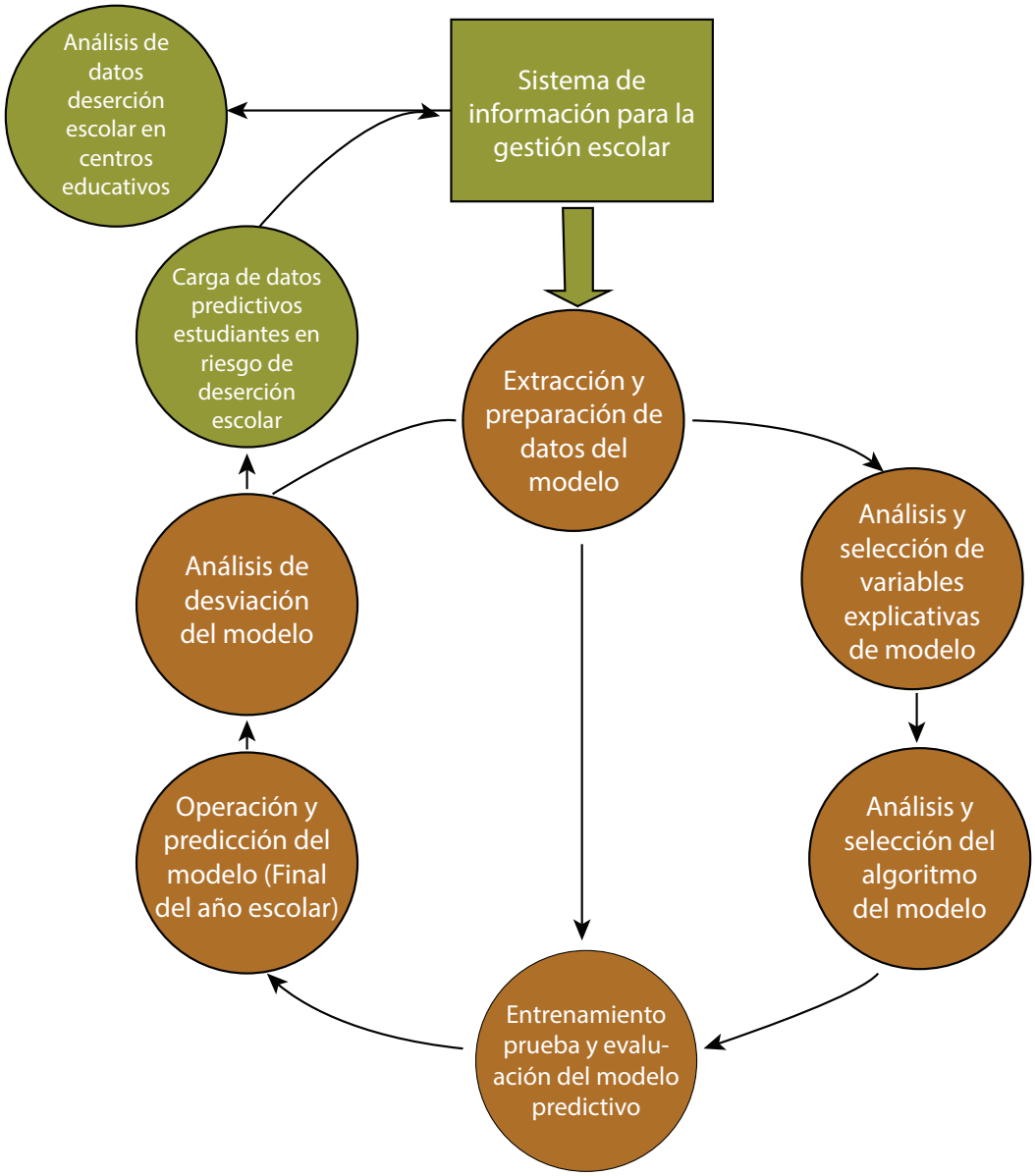
El ciclo de proyecto CRISP-DM se muestra en el siguiente gráfico (tomado de IBM SPSS Modeler Reference Manual):



El estándar incluye un modelo y una guía, estructurados en siete fases, algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores (ver cronograma de Proyecto anexo I y próxima sección 3.2):

3.2 Ciclo de Vida del Modelo

El Modelo tiene un ciclo de vida de seis etapas y dos actividades de servicio de información que se cumplen al final de cada año escolar. En el gráfico siguiente se muestran dichas etapas, que serán explicadas en las secciones siguientes.



I. Extracción y Preparación de Datos del Modelo

a) **Comprensión de los datos**

Consiste en familiarizarse con los datos teniendo presente los objetivos del proyecto. Los datos de proyecto se almacenan en diversas bases de datos por los sistemas institucionales, tanto los datos básicos demográficos y personales de las entidades envueltas como la información transaccional. Consiste en extraer los datos relevantes para crear un repositorio usado posteriormente por el analizador de datos de SPSS Modeler denominado *datamart*.

b) **Preparación de los datos**

El objetivo es obtener la vista o *data set* de la minería de datos para el desarrollo de los modelos. La limpieza y transformación de datos es el más intensivo de los recursos del proceso de proyecto de minería de datos. El propósito de la limpieza de datos es eliminar el ruido e información irrelevante del conjunto de datos y modificar la fuente de datos en diferentes formatos en función de los tipos de datos y valores.

En el informe anterior se hizo un **análisis de datos descriptivo** aplicando los conceptos de indicadores de deserción y retención escolar (promedios y tasas), fundamentados a partir de los resultados de la preparación de datos y que sirven de guía para establecer cuál es el estado de situación de los estudiantes y centros del Distrito de Los Alcarrizos, tomado como población de estudio. Se han incluido tablas analíticas y gráficas de barra de las series de los cinco años escolares y de los 12 grados para la matriculación, la condición académica y la deserción escolar (ver anexo 8.8).

II. Análisis y Selección de Variables Explicativas

La precisión del algoritmo depende de la naturaleza de los datos como del número de los estados del atributo de predicción o factor objetivo (en este caso deserción o no del estudiante), la distribución del valor de cada atributo, las relaciones entre los atributos, la correlación y auto correlación. Para la determinación de la variable o atributos significativos para el modelo se usan los test de significación de Pearson provistos por las herramientas de modelación y análisis estadístico (SPSS Y Rapid Miner).

En esta etapa se aplican las técnicas de minería de datos al *dataset* Historial_Academico-acumulado_2009_2014 preparado, que consiste en la cohorte de cinco años académicos.

Aplicaremos los recursos algorítmicos provistos por las herramientas como RapidMiner y SPSS Modeler.

III. Análisis y Selección del Algoritmo del Modelo

Se han de seleccionar los algoritmos 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.

IV. Modelación: Entrenamiento, Prueba y Predicción del Modelo

En esta etapa se aplican las técnicas de minería de datos al modelo de datos o *dataset*. Una vez que los datos se limpian y las variables son transformadas, podemos empezar a construir modelos basados en los recursos algorítmicos provistos por IBM SPSS Modeler y RapidMiner. Antes de construir cualquier modelo, tenemos que entender el objetivo del proyecto de minería de datos y el tipo de la tarea de minería de datos, como se ha definido en la sección anterior.

Una vez que entienda el tipo de tarea de minería de datos, se seleccionan los algoritmos de análisis de datos correctos. 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 que es el mejor algoritmo de ajuste para los datos antes del entrenamiento del modelo.

La precisión del algoritmo depende de la naturaleza de los datos como el número de los estados del atributo de predicción, la distribución del valor de cada atributo, las relaciones entre los atributos y así sucesivamente.

Comprende:

- Selección de la técnica de modelado
- Diseño de la evaluación
- Construcción del modelo
- Evaluación del modelo

Entregable: Modelo de DM) y documento descriptivo del modelo y sus componentes: Rutas, algoritmo empleado y los parámetros de precisión del mismo y las variables explicativas y objetivas usadas y evaluadas.

V. Implementación del Modelo

El próximo paso es el de **Despliegue e implementación**. Consiste en explotar la utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización. La implementación de un sistema de minería de datos cubre los siguientes componentes:

- Planificación de despliegue
- Planificación de la monitorización y del mantenimiento
- Capacitación de las áreas de usuarios
- Generación del informe final
- Revisión del proyecto

VI. Gestión del Modelo y Análisis de Desviación

Es difícil mantener el estado de los modelos de minería de datos. Cada modelo de minería tiene un ciclo de vida. En algunas instituciones, los patrones de datos son relativamente estables y los modelos no requieren de reciclaje con frecuencia. Sin embargo, en los patrones de muchas instituciones varían con frecuencia, esto significa que las nuevas reglas de asociación aparecen cada día. En el caso de la deserción escolar los cambios pueden ser muy dinámicos debido a que el MINERD implementa políticas de inversión en el sistema de educación y de intervención en los centros educativos y en los estudiantes que revertirían las tendencias de la deserción escolar. En un proceso dinámico como este el modelo de predicción de la deserción escolar debe ser evaluado anualmente al cierre del año escolar. En última instancia, determinar la exactitud del modelo y la creación de nuevas versiones de este debe llevarse a cabo mediante el uso de procesos automatizados, Rapid Miner e IBM SPSS Modeler ofrece una herramienta versátil de gestión de contenidos y versiones.

Entregables:

- Modelo de DM en RapidMiner y documento descriptivo del modelo y sus componentes: Rutas, algoritmo empleado y los parámetros de precisión del mismo y las variables explicativas y objetivas usadas y evaluadas.
- Documento evaluativo del resultado del modelo por parte del equipo de proyecto
- Documento de evaluación de modelo
- Entregable: Documento de cierre del proyecto.

4. Extracción y Preparación de Datos del Modelo

4.1 Comprensión de los datos

Consiste en familiarizarse con los datos teniendo presente los objetivos del proyecto. Los datos de proyecto se almacenan en diversas bases de datos por los sistemas institucionales, tanto los datos básicos demográficos y personales de las entidades envueltas como la información transaccional. Consiste en extraer los datos relevantes para crear un repositorio que, posteriormente, usará el analizador de datos de SPSS Modeler denominado datamart.

Para este proyecto, el MINERD debe disponer de los datos demográficos y del registro escolar del alumnado, así como de los datos de centros educativos e información de docentes.

Comprende:

- Recopilación inicial de datos
- Descripción de los datos
- Exploración de los datos
- Verificación de calidad de datos

4.2 Preparación de los datos

El objetivo es obtener el modelo de datos de entrada o data set de la minería de datos para el desarrollo de los modelos predictivos. La limpieza y transformación de datos es el más intensivo de los recursos del proceso de proyecto de minería de datos. El propósito de la limpieza de datos es eliminar el ruido e información irrelevante del conjunto de datos, y modificar la fuente de datos en diferentes formatos en función de los tipos de datos y valores.

La primera consistió en seleccionar el marco poblacional del estudio, familiarizarse con los datos, y validación de calidad teniendo presente los objetivos del proyecto. Se realizó una validación y aseguramiento de la calidad de los atributos de datos (corrección de valores perdidos, recodificación, etc.) por cada archivo o entidad de datos. Los datos del proyecto se almacenan en diversas bases de datos por los sistemas institucionales del MINERD, tanto los datos básicos demográficos, socioeconómicos y personales de los estudiantes y centros educativos, como la información transaccional del historial académico de cada estudiante.

La segunda fase: su objetivo fue obtener la vista o data set de la minería de datos para el desarrollo del modelo de deserción escolar. La limpieza y transformación de estos datos es el más intensivo de los recursos del proceso de proyecto de minería de datos. El propósito de la limpieza de datos es eliminar el ruido e información irrelevante del conjunto de datos y modificar la fuente de datos en diferentes formatos en función de los tipos de datos y valores, que sirvan para elaborar un modelo efectivo de deserción escolar. Para tales fines se elaboró un diagrama de estado de la deserción escolar que sirve como guía de preparación de datos y de análisis de información.

Comprende las siguientes actividades:

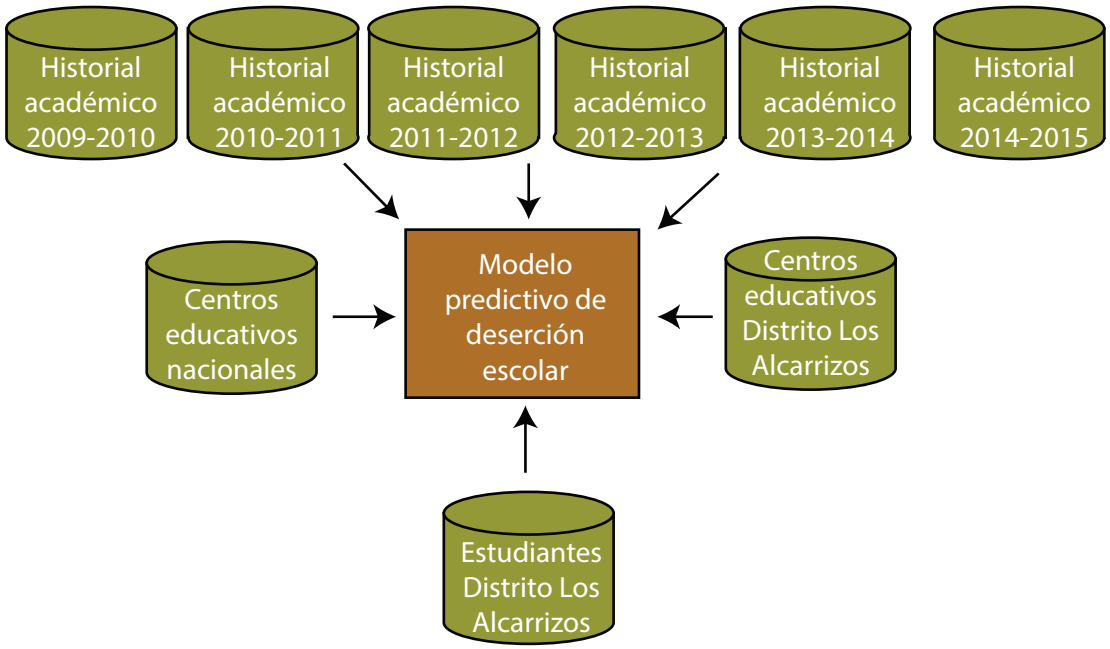
- Modelo físico de los datos
- Descripción de los datos
- Validación y control de calidad
- Diagrama de Estado de los datos

- Selección de los datos
- Limpieza de datos
- Construcción de datos
- Integración de datos
- Formateo de datos

4.3 Modelo Físico de Datos Provisto

El diagrama anexo representa el conjunto de archivos físicos que contienen las informaciones para el modelo de datos anterior y que sirven de insumos para la construcción del modelo predictivo de deserción escolar.

Estos archivos fueron provistos por el MINERD (Dirección de Información del Viceministerio de Planificación) a partir del requerimiento de datos y tuvieron como fuente las bases de datos del Sistema de Gestión de Centros.



4.4 Descripción, Validación y Control de calidad de datos

En el Informe del Producto 2 de esta consultoría se procedió a desarrollar los resultados de estas etapas que consistió en:

a) La descripción de datos contiene una explicación sobre cada variable o atributo del archivo componente del modelo físico de datos. A continuación aparece un gráfico con la estructura de los archivos del modelo físico, los registros del archivo y la descripción de sus atributos y sus tipos de datos, tal como se despliega en SPSS *Statistics*.

   i. Historial académico del estudiante
   ii. Estudiantes Distrito de Los Alcarrizos
   iii. Centros Educativos Distrito de Los Alcarrizos
   iv. Centros Educativos Nacionales

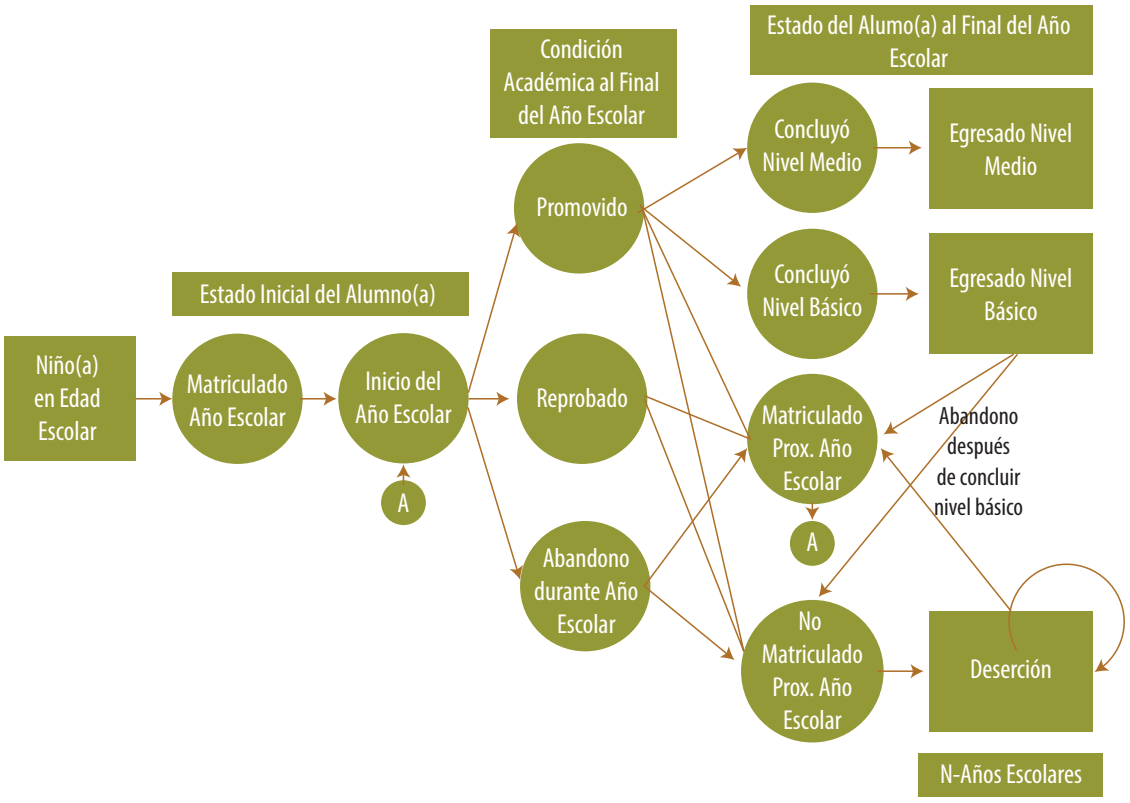
a) Para cada archivo y Mediante SPSS Statistics y sus funciones de auditar, se ha realizado una auditoría y verificación de los datos de los archivos de cada modelo.

   i. Verificar rangos de valores de variables medibles y valores de variables nominales.
   ii. Validar atributos numéricos y rangos de campos nominales y de fechas, valores perdidos, mal codificados, en blanco, o *null* de sistema
   iii. Determinar frecuencias de ocurrencias de valores
   iv. Hacer gráficos histogramas de frecuencia y otros.
   v. Asegurar la calidad y completitud de los datos.

4.5 Diagrama de Estado de Deserción Escolar y Factores Derivados

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 de estado siguiente, 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.



En este modelo los estados se representan en círculos y rectángulos. Los rectángulos representan los estados inicial y final 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 o alumna 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. El otro estado posible es el de concluir la educación Básica si aprueba sus exámenes estando en el 8vo. grado. Su estado sería 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 sería **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. grado) **o de deserción escolar de nivel Medio** si está cursando entre el primero y el cuarto del nivel Medio.

El **estatus de deserción escolar** considera un período 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 pueda matricularse).

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 descritos en la sección XXX. Estos son:

a) 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.
b) 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 períodos o años escolares ha durado en el sistema.
c) Último grado alcanzado antes de pasar a su condición de deserción.
d) 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.
e) Tiempo de deserción transcurrido, es decir, la cantidad de años escolares sin retornar al sistema.

f) Cantidad de reprobaciones tenidas antes de pasar a su condición de deserción o egresado.

g) Cantidad de promociones tenidas antes de pasar a la condición de deserción o egresado.

h) Si se ha transferido de centro educativo durante su estadía en el sistema antes de la condición de deserción o egresado (movilidad).

4.6 Modelo de Datos para el Entrenamiento, Prueba y Predicción

a) **Descripción del Modelo**

Los procedimientos de preparación de datos consisten en un conjunto de pasos para lograr el esquema de datos final que se ha de usar para la derivación del modelo predictivo de deserción escolar y el modelo analítico de las condiciones del centro educativo. Estos procedimientos consisten en:

i) Unificación de los historiales académicos

ii) Archivo de estado últimos registros historial académico (acumulados)

iii) Preparación del archivo de Centros Educativos del Distrito Los Alcarrizos

Se ha procedido a unir los cinco archivos de historiales académicos entre el 2009 y el 2014. El resultante es un archivo con el mismo formato de datos de los anteriores.

Los atributos que componen tanto los archivos de los cinco años escolares entre 2009 y 2014 como el resultante del proceso de unión son los siguientes, como se describió en la sección de Comprensión, recolección y calidad de datos.

El cuadro siguiente contiene las cantidades de registros por archivo y el total.

CANTIDAD DE REGISTROS POR AÑO ACADÉMICO		FREQUENCY	PERCENT	VALID PERCENT	CUMULATIVE PERCENT
Valid	2009-2010	48,710	16.70%	16.70%	16.70%
	2010-2011	48,774	16.73%	16.73%	33.43%
	2011-2012	48,794	16.73%	16.73%	50.16%
	2012-2013	49,258	16.89%	16.89%	67.05%
	2013-2014	49,767	17.07%	17.07%	84.12%
	2014-2015	46,309	15.88%	15.88%	100.00%
	Total	291,612	100.00%	100.00%	

- Luego se agregan los datos demográficos de los estudiantes incluyendo si pertenece al programa PROSOLI.
- En el archivo de Estudiantes Distrito Los Alcarrizos, se sabe la condición, o sea, si el estudiante está en el plan de Subsidios de Solidaridad ILAE (nivel Básico) o BEEP para nivel Medio.
- Estos subsidios son otorgados a los estudiantes con nivel de pobreza 1 y 2 de acuerdo a la clasificación solicitada al SIUBEN.
- Esta información ha de servir como variable explicativa del nivel de pobreza del estudiante, como factor socioeconómico potencial que influye en el índice de deserción. En el atributo ICV1 se le coloca la condición 1 si está en uno de los programas.
- El archivo de Centros Educativos nacionales contiene los atributos con los datos de vulnerabilidad del centro educativo: Indicador de pobreza e indicador de vulnerabilidad.
- Estos indicadores se obtuvieron a partir de la clasificación solicitada al SIUBEN sobre el entorno geográfico y barrial del centro educativo, determinando la proporción de hogares pobres (véase estudio de EDUCA-UE)
- Esta información debe servir como variable explicativa del nivel de vulnerabilidad del entorno del centro como factor socioeconómico potencial que influye en el índice de deserción.

Para poder establecer el estado del estudiante en el sistema educativo para los centros seleccionados del Distrito de Los Alcarrizos se procedió a seleccionar el último registro del historial académico y acumular los factores explicativos relativos a la condición académica y al grado escolar del estudiante para los cinco años escolares comprendidos entre en 2009 y 2014.

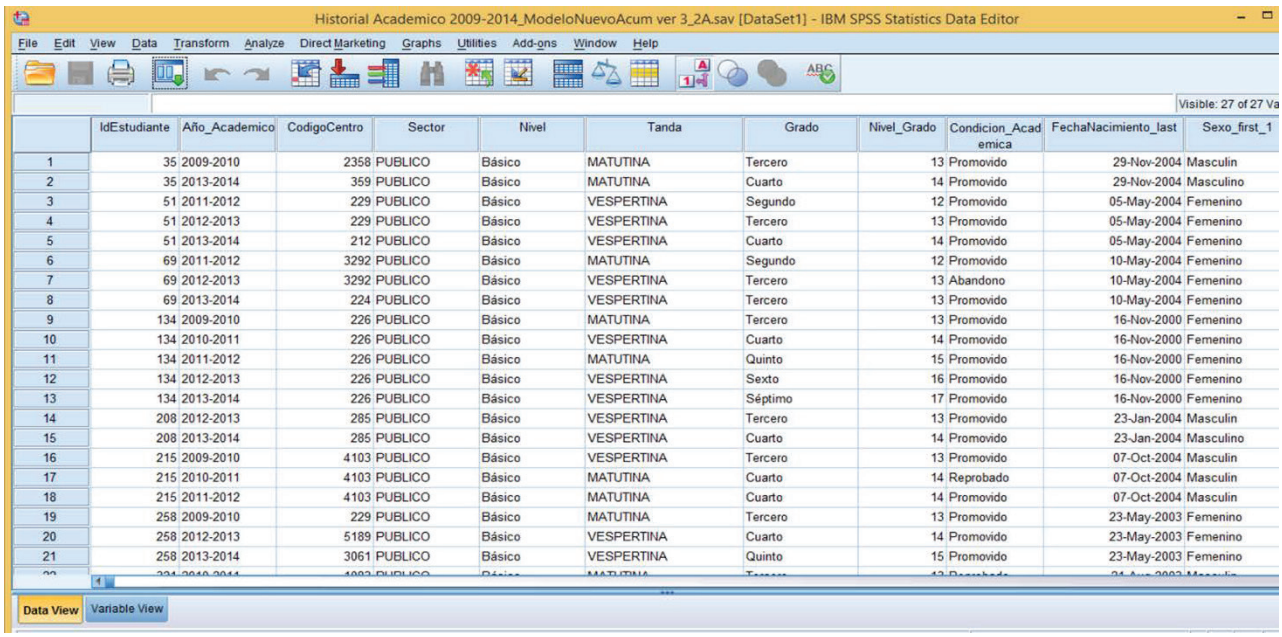
En ese sentido, se determinan los siguientes estatus de estudiantes:

- Estudiantes que se inscribieron para el próximo año escolar, es decir, para el 2014-2015. Estos son los considerados estudiantes activos.
- Estudiantes que salieron del sistema por haber completado el ciclo preuniversitario (nivel Básico y nivel Medio), es decir, habiendo aprobado el 4to. grado del nivel Medio. Para los estudiantes que concluyeron el ciclo Básico (aprobación del 8vo. grado) se le pone un estatus de conclusión del nivel Básico. Si no se inscriben para el siguiente año escolar se considera deserción del nivel Medio.
- Estudiantes que no se inscribieron o se reinscribieron para el próximo año escolar teniendo como tope el 2014-2015, estos son los considerados desertores del sistema. Se identifican dos tipos: los que retornan antes del período tope y los que no retornaron.

El siguiente diagrama muestra el proceso arriba descrito:

En la primera etapa se procede a producir un archivo seleccionando el último registro de cada estudiante en el historial 2009-2014 en los cinco períodos o años escolares bajo estudio. En la segunda etapa se procede a registrar los acumulados de factores de medición a ser usados en el modelo de predicción de deserción escolar. En la tercera etapa se realiza una comparación (maching record) de este archivo de historial académico acumulado y el archivo de historial 20014-2015 con el objeto de determinar cuáles estudiantes se inscribieron para ese último período, luego se procede a identificar los que no lo hicieron como deserciones. Además los estudiantes egresados (condición académica de aprobación del 4to. del bachillerato) son indicados como egresados del nivel Medio y los de 8vo. grado egresado del ciclo Básico.

La descripción de registros y los atributos del archivo resultante es como sigue:


Historial Academico 2009-2014_ModeloNuevoAcum ver 3_2A.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Visible: 27 of 27 Va

	IdEstudiante	Año_Academico	CodigoCentro	Sector	Nivel	Tanda	Grado	Nivel_Grado	Condicion_Academica	FechaNacimiento_last	Sexo_first_1
1	35	2009-2010	2358	PUBLICO	Básico	MATUTINA	Tercero	13	Promovido	29-Nov-2004	Masculin
2	35	2013-2014	359	PUBLICO	Básico	MATUTINA	Cuarto	14	Promovido	29-Nov-2004	Masculino
3	51	2011-2012	229	PUBLICO	Básico	VESPERTINA	Segundo	12	Promovido	05-May-2004	Femenino
4	51	2012-2013	229	PUBLICO	Básico	VESPERTINA	Tercero	13	Promovido	05-May-2004	Femenino
5	51	2013-2014	212	PUBLICO	Básico	VESPERTINA	Cuarto	14	Promovido	05-May-2004	Femenino
6	69	2011-2012	3292	PUBLICO	Básico	MATUTINA	Segundo	12	Promovido	10-May-2004	Femenino
7	69	2012-2013	3292	PUBLICO	Básico	VESPERTINA	Tercero	13	Abandono	10-May-2004	Femenino
8	69	2013-2014	224	PUBLICO	Básico	VESPERTINA	Tercero	13	Promovido	10-May-2004	Femenino
9	134	2009-2010	226	PUBLICO	Básico	MATUTINA	Tercero	13	Promovido	16-Nov-2000	Femenino
10	134	2010-2011	226	PUBLICO	Básico	VESPERTINA	Cuarto	14	Promovido	16-Nov-2000	Femenino
11	134	2011-2012	226	PUBLICO	Básico	MATUTINA	Quinto	15	Promovido	16-Nov-2000	Femenino
12	134	2012-2013	226	PUBLICO	Básico	VESPERTINA	Sexto	16	Promovido	16-Nov-2000	Femenino
13	134	2013-2014	226	PUBLICO	Básico	VESPERTINA	Séptimo	17	Promovido	16-Nov-2000	Femenino
14	208	2012-2013	285	PUBLICO	Básico	VESPERTINA	Tercero	13	Promovido	23-Jan-2004	Masculin
15	208	2013-2014	285	PUBLICO	Básico	VESPERTINA	Cuarto	14	Promovido	23-Jan-2004	Masculino
16	215	2009-2010	4103	PUBLICO	Básico	VESPERTINA	Tercero	13	Promovido	07-Oct-2004	Masculin
17	215	2010-2011	4103	PUBLICO	Básico	MATUTINA	Cuarto	14	Reprobado	07-Oct-2004	Masculin
18	215	2011-2012	4103	PUBLICO	Básico	MATUTINA	Cuarto	14	Promovido	07-Oct-2004	Masculin
19	258	2009-2010	229	PUBLICO	Básico	MATUTINA	Tercero	13	Promovido	23-May-2003	Femenino
20	258	2012-2013	5189	PUBLICO	Básico	VESPERTINA	Cuarto	14	Promovido	23-May-2003	Femenino
21	258	2013-2014	3061	PUBLICO	Básico	VESPERTINA	Quinto	15	Promovido	23-May-2003	Femenino

Data View Variable View

	ICV1	ZonadelCentro	IndiceProsoliCentro	IndicePobrezaCentro	edad_estudiante	matriculado	Desercion	Promovido_sum	Reprobado_sum	Abandono_sum	Condicion_otra_s...
1	0	URBANA-M	8.20	69.88	8.50	1	0	2	0	0	0
2	0	URBANA	14.80	68.00	4.50	0	1	1	0	0	0
3	0	URBANA	14.70	37.04	9.07	1	0	3	0	0	0
4	0	URBANA-M	10.50	48.00	7.07	1	0	1	0	0	0
5	0	URBANA-M	10.50	48.00	8.07	1	0	2	0	0	0
6	0	URBANA-M	19.40	35.75	9.05	1	0	2	0	1	0
7	0	RURAL	21.90	122.00	7.05	1	0	1	0	0	0
8	0	RURAL	21.90	122.00	8.05	1	0	1	0	1	0
9	1	URBANA-M	24.30	61.00	8.53	1	0	1	0	0	0
10	1	URBANA-M	24.30	61.00	9.53	1	0	2	0	0	0
11	1	URBANA-M	24.30	61.00	10.53	1	0	3	0	0	0
12	1	URBANA-M	24.30	61.00	11.53	1	0	4	0	0	0
13	1	URBANA-M	24.30	37.04	12.53	1	0	5	0	0	0
14	1	URBANA-M	15.70	88.00	8.35	1	0	1	0	0	0
15	1	URBANA-M	15.70	72.23	9.35	1	0	2	0	0	0
16	0	RURAL	11.10	71.00	4.65	1	0	1	0	0	0
17	0	RURAL	11.10	71.00	5.64	1	0	1	1	0	0
18	0	RURAL	11.10	71.00	6.64	0	1	2	1	0	0
19	1	URBANA-M	9.60	26.00	6.02	0	1	1	0	0	0
20	1	URBANA	12.50	36.32	10.02	1	0	3	0	0	0
21	1	URBANA-M	9.60	26.00	9.02	1	0	2	0	0	0

*Historial Academico 2009-2014_ModeloNuevoAcum ver 3_2.sav [DataSet3] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	IdEstudiante	Numeric	12	0		None	None	9	Right	Scale	Input
2	Año_Academico	String	9	0		None	None	11	Left	Nominal	Input
3	CodigoCentro	Numeric	12	0		None	None	10	Right	Scale	Input
4	Sector	String	11	0		None	None	11	Left	Nominal	Input
5	Nivel	String	7	0		None	None	11	Left	Nominal	Input
6	Tanda	String	17	0		None	None	17	Left	Nominal	Input
7	Grado	String	12	0		None	None	12	Left	Nominal	Input
8	Nivel_Grado	Numeric	8	0		None	None	9	Right	Scale	Input
9	Condicion_Academica	String	12	0		None	None	10	Left	Nominal	Input
10	FechaNacimiento_last	Date	11	0		None	None	15	Right	Scale	Input
11	Sexo_first_1	String	12	0		None	None	11	Left	Nominal	Input
12	ICV1	Numeric	1	0		None	None	9	Right	Nominal	Input
13	ZonadelCentro	String	8	0		None	None	12	Left	Nominal	Input
14	IndiceProsoliCentro	Numeric	7	2	SMEAN(Indice...	None	None	14	Right	Scale	Input
15	IndicePobrezaCentro	Numeric	7	2	SMEAN(Indice...	None	None	15	Right	Scale	Input
16	edad_estudiante	Numeric	8	2		None	None	11	Right	Scale	Input
17	matriculado	Numeric	1	0	Case source is ...	None	None	8	Right	Nominal	Input
18	Desercion	Numeric	8	0		None	None	11	Right	Nominal	Input
19	Promovido_sum	Numeric	8	0		None	None	11	Right	Nominal	Input
20	Reprobado_sum	Numeric	8	0		None	None	12	Right	Nominal	Input
21	Abandono_sum	Numeric	8	0		None	None	11	Right	Nominal	Input
22	Condicion_otra_sum	Numeric	8	0		None	None	12	Right	Nominal	Input
23	Cambio_Centro	Numeric	8	0		None	None	11	Right	Nominal	Input
24	resumen	Numeric	1	0	Case source is	None	None	9	Right	Nominal	Input

Data View | Variable View

Los cuadros siguientes muestran la tasa de deserción para los diferentes niveles de Básico y del nivel Medio y para los años de la cohorte seleccionada (2009-2014), así como 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 cinco períodos 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.

	12 GRADOS				TASA DE EGRESADOS	
TASAS 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 Geom			**9.00%**	**8.84%**	**9.09%**	**10.87%**
	Retorno Total	**1834**	**7.43%**			

	NIVEL BASICO					NIVEL MEDIO				
TASAS DE DESERCIÓN ESCOLAR	MATRÍCULA NIVEL BÁSICO	EGRESADOS NIVEL BÁSICA	DESERCIO-NES NIVEL BÁSICA	INGRESA-DOS 1ER GRADO	TASA DE DESERCION NIVEL BASICO	MATRICULA NIVEL MEDIO	EGRE-SADOS MEDIA	DESER-CIONES MEDIA	INGRESA-DOS 1ERO DE MEDIA	TASA DESERCION 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		17861	25655		66108		6521	20441	
Promedio	**38445**		**3572**	**4421**	**8.79%**	**10616**		**1304**	**3242**	**9.97%**
Media Geom					**8.49%**					**9.79%**

En el informe 2 se hizo un **análisis de datos descriptivo** aplicando los conceptos de indicadores de deserción y retención escolar (promedios y tasas), fundamentados a partir de los resultados de la preparación de los datos y del modelo de datos resultantes y que sirven de guía para establecer cuál es el estado de situación de los estudiantes y centros del Distrito de los Alcarrizos, tomado como población bajo estudio. Se han incluido tablas analíticas y gráficas de barra de las series de los cinco años escolares de la cohorte 2009-2014 y de los 12 grados para la matriculación, la condición académica y la deserción escolar (ver anexo 8.8 del Informe II).

5. Desarrollo del Modelo de Deserción Escolar

5.1 Objetivos

En esta etapa se aplican las técnicas de minería de datos al modelo de datos o dataset. Una vez que los datos se limpian, las variables son transformadas y se crea un modelo de datos con cierta consistencia lo que permite empezar a construir modelos basados en los recursos algorítmicos provistos por RapidMiner e IBM SPSS Modeler. Antes de construir cualquier modelo, tenemos que entender el objetivo del proyecto de minería de datos y el tipo de la tarea de minería de datos, como se ha definido en la sección anterior.

Una vez que se entienda el tipo de tarea de minería de datos, se seleccionan los algoritmos de análisis de datos correctos. 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.

La precisión del algoritmo depende de la naturaleza de los datos como el número de los estados del atributo de predicción, la distribución del valor de cada atributo, las relaciones entre los atributos y así sucesivamente.

Comprende:

- Análisis y selección de variables
- Análisis y selección del algoritmo de modelado
- Construcción, entrenamiento y prueba del modelo
- Evaluación del modelo

Entregable: Modelo de DM en SPSS Modeler (rutas) y documento descriptivo del modelo y sus componentes: Rutas, algoritmo empleado y los parámetros de precisión del mismo y las variables explicativas y de respuesta usadas y evaluadas.

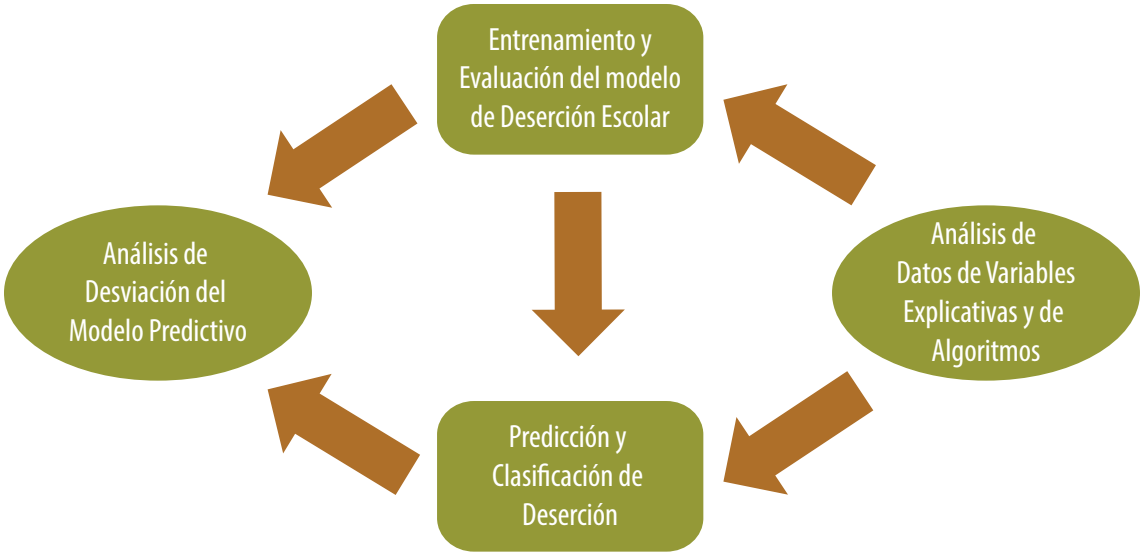
5.2 Componentes del Modelo Predictivo

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 período siguiente al período 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, pero también para determinar el riesgo de que un cliente deje de usar el servicio dentro de un período determinado en el futuro.

En la siguiente gráfica pueden verse los cuatro componentes funcionales del modelo de deserción escolar implementado mediante las herramientas RapidMiner y SPSS Modeler, y que son explicados en las secciones siguientes.

En la siguiente gráfica pueden verse los componentes funcionales del modelo que son explicados en las secciones siguientes.



5.3 Análisis y Selección de Variables Explicativas

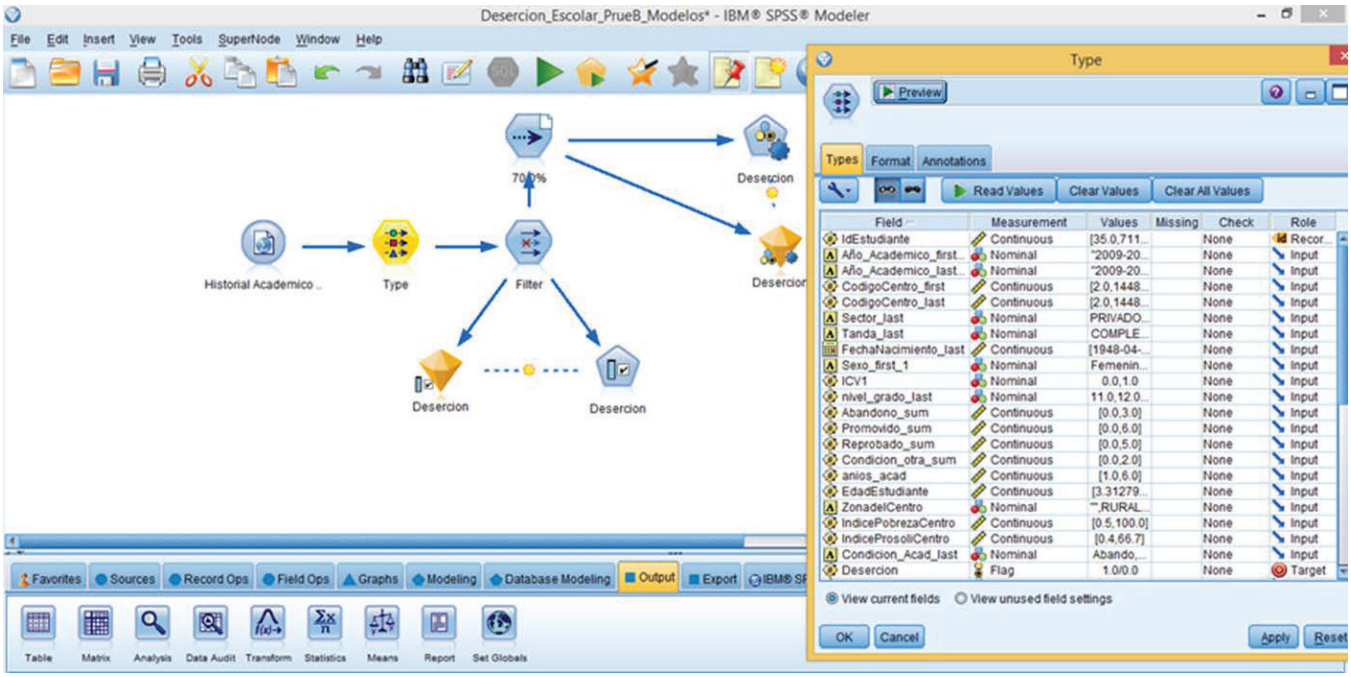
La precisión del algoritmo depende de la naturaleza de los datos como del número de los estados del atributo de predicción o factor objetivo (en este caso de la deserción o no del estudiante), la distribución del valor de cada atributo, las relaciones entre los atributos, la correlación y auto-correlación. Para la determinación de la variable o atributos significativos para el modelo se usan los test de significación de Pearson provistos por las herramientas de modelación y análisis estadístico (SPSS Y Rapid Miner).

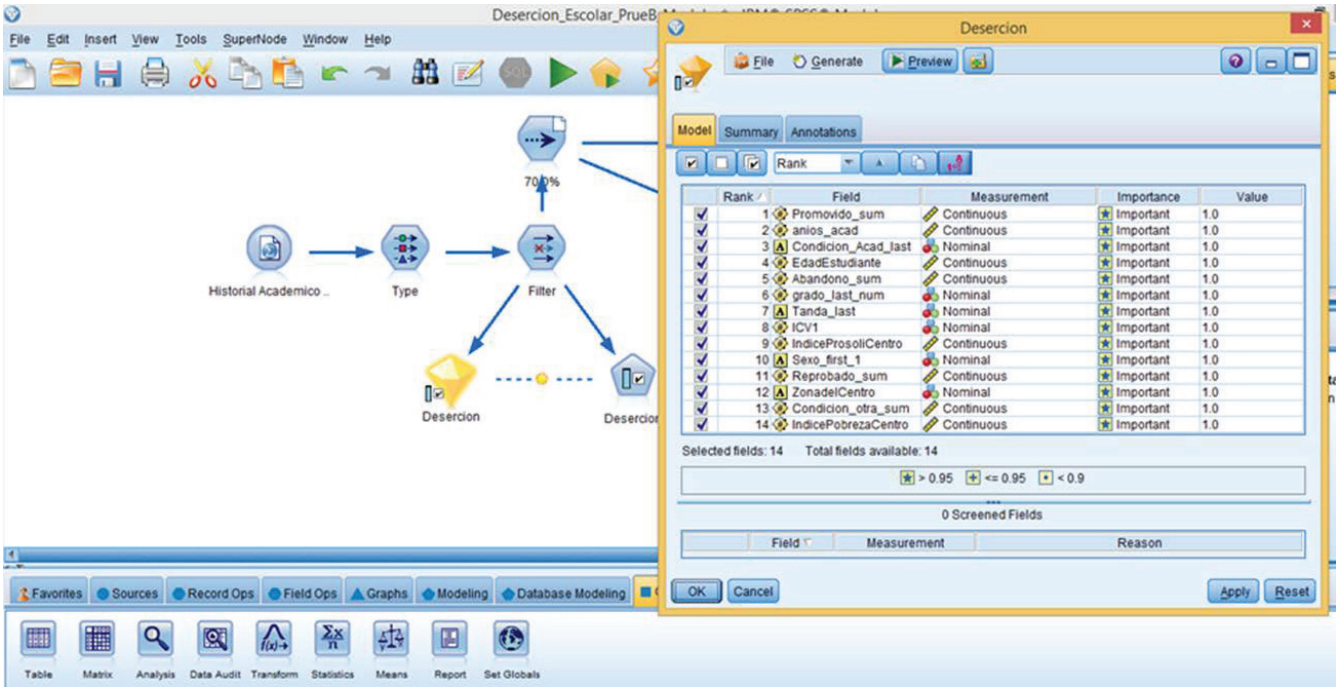
En esta etapa se aplican las técnicas de minería de datos al *dataset* Historial_Academico-acumulado_2009_2014 preparado, que describen la cohorte de cinco años académicos. Aplicaremos los recursos algorítmicos provistos por la herramienta de SPSS Modeler.

El proceso de aprendizaje y de predicción está basado en los factores demográficos del estudiante, en su condición de vulnerabilidad social (medido por el índice de calidad de vida ICV del SIUBEN), el nivel de vulnerabilidad del centro educativo medido por el índice de pobreza del sector geográfico, el historial académico del estudiante (cohorte de 5 años académicos entre 2009 y 2014).

Estos factores se usan como variables explicativas para la estimación de la deserción escolar (cero (0) no deserción y 1 deserción y su probabilidad de ocurrencia). **[Deserción, Riesgo] = F(factor1, factor2, …,factorn).**

Los factores explicativos se seleccionan de acuerdo con su nivel de significación, eliminación de factores auto correlacionados, basados en índices de correlación y pruebas de hipótesis estadísticas realizadas mediante el nodo de Selección de Características de SPSS Modeler y SPSS Statistics (Features Selecction Node). El nodo selección de características filtra los campos de entrada para su eliminación en función de un conjunto de criterios (como el porcentaje de valores perdidos); a continuación, clasifica el grado de importancia del resto de entradas de acuerdo con un objetivo específico.


Desercion_Escolar_PrueB_Modelos* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Window Help
Historial Academico ..
Type
Filter
Desercion
Desercion
Desercion
Type
Preview
Types Format Annotations
Read Values Clear Values Clear All Values
Field Measurement Values Missing Check Role
IdEstudiante Continuous [35.0,711.. None Recor...
Año_Academico_first... Nominal "2009-20... None Input
Año_Academico_last... Nominal "2009-20... None Input
CodigoCentro_first Continuous [2.0,1448... None Input
CodigoCentro_last Continuous [2.0,1448... None Input
Sector_last Nominal PRIVADO... None Input
Tanda_last Nominal COMPLE... None Input
FechaNacimiento_last Continuous [1948-04-... None Input
Sexo_first_1 Nominal Femenin... None Input
ICV1 Nominal 0.0,1.0 None Input
nivel_grado_last Nominal 11.0,12.0... None Input
Abandono_sum Continuous [0.0,3.0] None Input
Promovido_sum Continuous [0.0,6.0] None Input
Reprobado_sum Continuous [0.0,5.0] None Input
Condicion_otra_sum Continuous [0.0,2.0] None Input
anios_acad Continuous [1.0,6.0] None Input
EdadEstudiante Continuous [3.31279... None Input
ZonadelCentro Nominal "","RURAL... None Input
IndicePobrezaCentro Continuous [0.5,100.0] None Input
IndiceProsoliCentro Continuous [0.4,66.7] None Input
Condicion_Acad_last Nominal Abando,... None Input
Desercion Flag 1.0/0.0 None Target
View current fields View unused field settings
OK Cancel Apply Reset
Favorites Sources Record Ops Field Ops Graphs Modeling Database Modeling Output Export
Table Matrix Analysis Data Audit Transform Statistics Means Report Set Globals


File Edit Insert View Tools SuperNode Window Help
Historial Academico ..
Type
Filter
Desercion
Desercion
Desercion
File Generate Preview
Model Summary Annotations
Rank
Rank Field Measurement Importance Value
1 Promovido_sum Continuous Important 1.0
2 anios_acad Continuous Important 1.0
3 Condicion_Acad_last Nominal Important 1.0
4 EdadEstudiante Continuous Important 1.0
5 Abandono_sum Continuous Important 1.0
6 grado_last_num Nominal Important 1.0
7 Tanda_last Nominal Important 1.0
8 ICV1 Nominal Important 1.0
9 IndiceProsoliCentro Continuous Important 1.0
10 Sexo_first_1 Nominal Important 1.0
11 Reprobado_sum Continuous Important 1.0
12 ZonadelCentro Nominal Important 1.0
13 Condicion_otra_sum Continuous Important 1.0
14 IndicePobrezaCentro Continuous Important 1.0
Selected fields: 14 Total fields available: 14
> 0.95 <= 0.95 < 0.9
0 Screened Fields
Field Measurement Reason
OK Cancel Apply Reset
Favorites Sources Record Ops Field Ops Graphs Modeling Database Modeling
Table Matrix Analysis Data Audit Transform Statistics Means Report Set Globals

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 siguiente los atributos que tienen menos peso son: grado, zona del centro y año académico first. Se excluyó año académico first del modelo.

Poder de clasificación de los atributos

ATRIBUTO	PESO
Matriculado	0.7009
N_años_acad	0.5446
Fecha Nacimiento_last	0.1781
edad_estudiante	0.1781
Año_Académico_last	0.163
Cambio_Centro	0.0824
Abandono_sum	0.0764
Nivel_Grado	0.0689
Condición_Académica	0.0632
Condición_otra_sum	0.0385
Índice Pobreza Centro	0.0295
Índice Prosoli Centro	0.0105
ICV1	0.0096
Resumen	0.0077
Reprobado_sum	0.0077
Código Centro	0.007
Tanda	0.0055
Sector	0.0055
Promovido_sum	0.0038
Nivel	0.0025
Año_Académico	0.0013
Sexo_first_1	0.0013
Grado	0.0006
Zona del Centro	0.0003
Año_Académico_first_1	0

El siguiente resultado digno de analizar es la tabla de correlaciones para ver qué atributos están correlacionados y eliminar aquellos que tengan menos peso en el poder de predicción. La tabla siguiente resume los resultados y como puede verse, 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 (0.880), el modelo presentó mejores resultados incluyendo ambos atributos.

Matriz de correlación de los atributos

MATRIZ DE CORRELACION	AÑO_ ACADE-MICO	TANDA	NIVEL_ GRADO	CONDI-CION_ACA-DEMICA	SEXO_ FIRST_1	ICV1	ZONADEL-CENTRO	INDICEPRO-SOLICENTRO	INDICEPOBRE-ZACENTRO	EDAD_ ESTU-DIANTE	PROMO-VIDO_ SUM	REPRO-BADO_ SUM	ABANDO-NO_SUM	CONDICION_ OTRA_SUM	CAMBIO_ CENTRO
Año_Academico	1.000000	0.019631	0.149690	0.007791	0.005006	0.039411	0.000769	0.000487	-0.006932	0.132143	0.321565	0.090108	0.036301	-0.016320	0.094970
Tanda	0.019631	1.000000	0.303318	0.055530	0.005789	-0.038405	0.021679	-0.007523	0.026340	0.292320	0.042650	0.037107	0.035339	-0.005920	0.178359
Nivel_Grado	0.149690	0.303318	1.000000	0.093752	0.072778	-0.071969	0.014478	0.309595	0.190549	0.878957	0.203950	-0.005703	0.013495	-0.011709	0.060934
Condicion_Academica	0.007791	0.055530	0.093752	1.000000	-0.055557	-0.015377	0.002778	0.021800	0.004129	0.112070	0.205972	0.584046	0.188221	0.091082	0.051099
Sexo_first_1	0.005006	0.005789	0.072778	-0.055557	1.000000	0.012386	0.005150	0.047965	0.031506	0.017157	0.009409	-0.087977	-0.026808	-0.003981	-0.023556
ICV1	0.039411	-0.038405	-0.071969	-0.015377	0.012386	1.000000	0.022925	0.105249	0.048442	-0.071925	0.035546	0.022877	-0.026314	-0.012216	-0.020337
ZonadelCentro	0.000769	0.021679	0.014478	0.002778	0.005150	0.022925	1.000000	0.075602	0.193030	0.019411	0.004387	-0.012868	0.005852	0.010581	0.061287
IndiceProsoliCentro	0.000487	-0.007523	0.309595	0.021800	0.047965	0.105249	0.075602	1.000000	0.520881	0.241426	0.003178	-0.006199	-0.045855	-0.001814	-0.043691
IndicePobrezaCentro	-0.006932	0.026340	0.190549	0.004129	0.031506	0.048442	0.193030	0.520881	1.000000	0.147777	-0.013001	-0.016707	-0.046692	-0.001332	0.014398
edad_estudiante	0.132143	0.292320	0.878957	0.112070	0.017157	-0.071925	0.019411	0.241426	0.147777	1.000000	0.167018	0.068464	0.057620	-0.006595	0.073723
Promovido_sum	0.321565	0.042650	0.203950	0.205972	0.009409	0.035546	0.004387	0.003178	-0.013001	0.167018	1.000000	0.117744	0.105918	0.050631	0.157791
Reprobado_sum	0.090108	0.037107	-0.005703	0.584046	-0.087977	0.022877	-0.012868	-0.006199	-0.016707	0.068464	0.117744	1.000000	0.000482	-0.007040	0.077983
Abandono_sum	0.036301	0.035339	0.013495	0.188221	-0.026808	-0.026314	0.005852	-0.045855	-0.046692	0.057620	0.105918	0.000482	1.000000	-0.004576	0.040644
Condicion_otra_sum	-0.016320	-0.005920	-0.011709	0.091082	-0.003981	-0.012216	0.010581	-0.001814	-0.001332	-0.006595	0.050631	-0.007040	-0.004576	1.000000	0.010110
Cambio_Centro	0.094970	0.178359	0.060934	0.051099	-0.023556	-0.020337	0.061287	-0.043691	0.014398	0.073723	0.157791	0.077983	0.040644	0.010110	1.000000

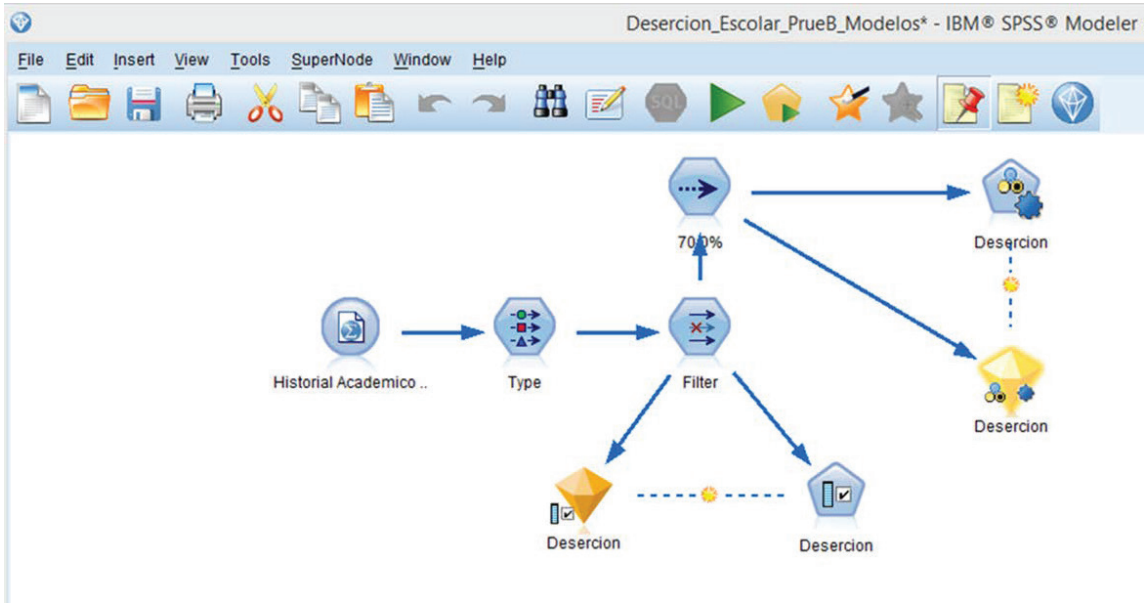
5.4 Análisis y Selección del Algoritmo del Modelo

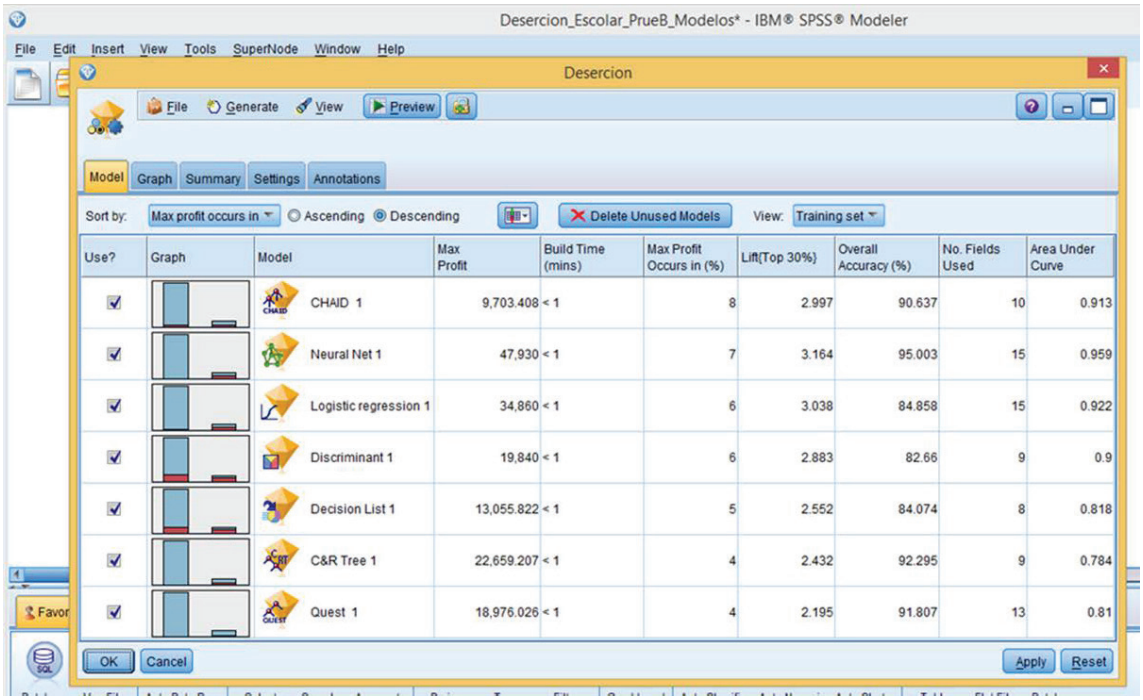
En la etapa de modelación se construye un conjunto de modelos que utilizan diferentes algoritmos y ajustes de parámetros. Entonces, ¿cuál es el mejor modelo en términos de precisión? ¿Cómo se evalúan estos modelos? IBM SPSS Modeler provee herramientas para evaluar la calidad de un modelo. Se utiliza un modelo de formación para predecir los valores del conjunto de datos de prueba, sobre la base del valor de predicción y la probabilidad.

Se han de seleccionar los algoritmos 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 crea y compara varios modelos diferentes para obtener resultados binarios (sí o no, abandono o no de clientes, 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.

- El algoritmo predictivo se seleccionó a partir de las pruebas de precisión, usando la variable objetivo o de respuesta Deserción y las variables explicativas descritas anteriormente y seleccionadas como significativas en la prueba de selección de variables del punto 4.3 anterior.





- Se hicieron pruebas comparativas del clasificador automático para 7 modelos algorítmicos con una muestra del 70% de los datos.

En la opción de clasificar modelos se especifican los criterios utilizados para comparar y clasificar los modelos. Las opciones incluyen la precisión global, área debajo de la curva ROC, beneficio, elevación y número de campos. Al calcular beneficios, elevación y estadísticos relacionados, se supone que el valor True definido para el campo objetivo representa un acierto. En el caso nuestro un 1 para el campo objetivo Deserción.

- **Precisión global:** Porcentaje de registros predichos correctamente por el modelo respecto al número total de registros.
- **Área debajo de la curva ROC:** La curva ROC proporciona el índice de rendimiento de un modelo. Cuanto más se encuentre la curva sobre la línea de referencia, más exacta será la prueba.
- **Beneficio (acumulado):** Suma de los beneficios de los percentiles acumulados (clasificados en términos de confianza para la predicción), calculados en base a los costes, ingresos y criterios de ponderación especificados. Normalmente, el beneficio comienza cerca del cero (0), aumenta rápidamente y a continuación, desciende. Para obtener un modelo válido, los beneficios deben mostrar un pico bien definido junto con el percentil donde aparece. En el caso de un modelo que no proporciona ninguna información, la curva de beneficio será bastante recta y puede ascender, descender o permanecer en el mismo nivel en función de la estructura de costes/ingresos que se aplique.
- **Elevación (acumulado):** Tasa de aciertos en cantidades acumuladas con respecto a la muestra global (donde los cuantiles se clasifican en función de la confianza para la predicción). Por ejemplo, un valor de elevación de 3 para el cuartil superior indica una tasa de aciertos tres veces más alta que la de la muestra global. Para obtener un modelo válido, la elevación debe comenzar muy por encima de 1,0 para los cuantiles superiores y, a continuación, descender rápidamente hasta 1,0 para los cuantiles inferiores. En el caso de un modelo que no proporciona ninguna información, la elevación se mantendrá alrededor de 1,0.

- **Número de campos:** Ordena los modelos en función de los campos de entrada utilizados.

Siendo el árbol de decisión CHAID el mejor algoritmo de predicción estimado a partir de la medida de precisión arrojada para este conjunto de datos. Como se puede notar el criterio de orden de importancia es el que corresponde al Beneficio Acumulado o "Max Profit Occurs (%)", que para CHAID es de 8.

5.5 Entrenamiento, Prueba y Predicción del Modelo

a) **Conceptos Teóricos sobre el Algoritmo de Árbol de Decisión**

El problema de la deserción escolar se enfocó como un problema de clasificación, donde se necesita un modelo que sea capaz de clasificar a los estudiantes en dos clases: desertores y no desertores, en función de las estadísticas escolares y otras informaciones socio-económicas suministradas. Las razones para la deserción escolar son variadas y los detalles de las mismas deben ser estudiados para su evaluación. Sin embargo, con la información disponible pueden buscarse patrones que permitan identificar cuáles son los estudiantes con mayor propensión a dejar la escuela con el objeto de elaborar políticas e intervenciones preventivas de acuerdo a cada caso particular.

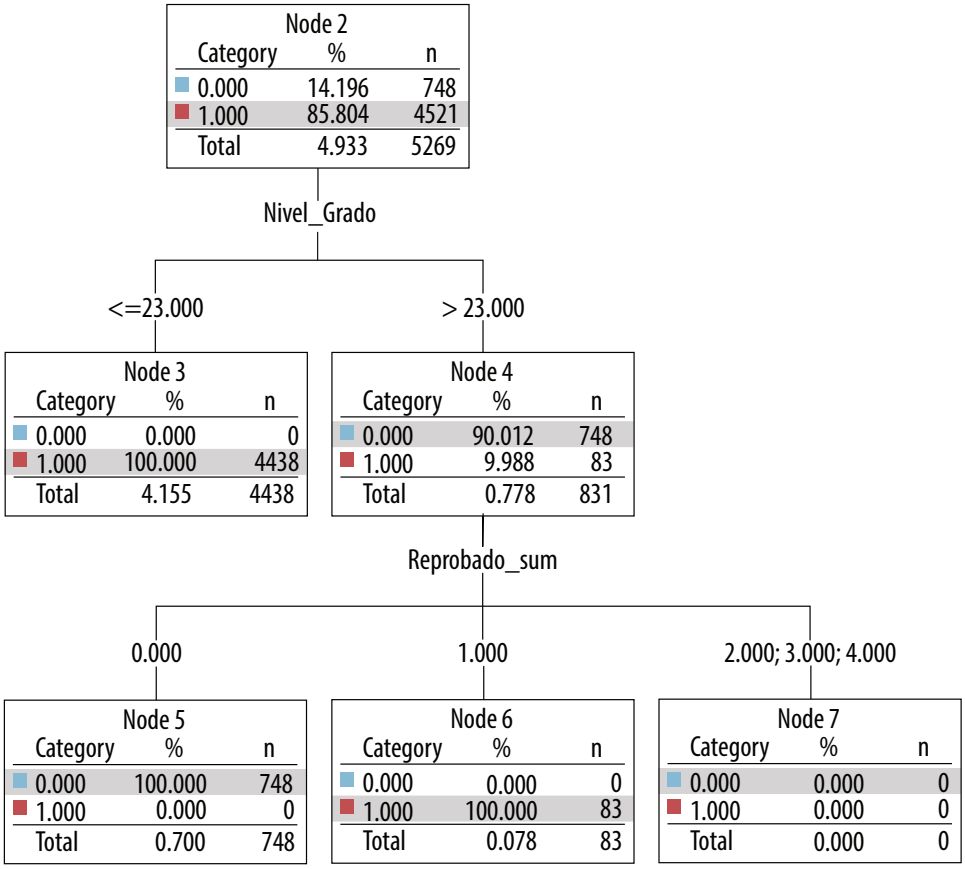
Debido a las características de las variables disponibles y al conjunto de datos de la cohorte 2009-2014 del historial académico del estudiante perteneciente a uno de los 72 centros del Distrito de los Alcarrizos, el algoritmo más adecuado para hacer la clasificación de los estudiantes fue el de árbol de Decisión CHAID, seleccionado en el punto 4.4.

A los árboles de decisión también se les conoce como árboles 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 tal como la que se muestra en el anexo 1.

El árbol de decisión es un clasificador expresado como 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) o 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.

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.

En la gráfica 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 puede verse que su ocurrencia ha dependido de la probabilidad de ocurrencia 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).



El pre-procesamiento de los datos resulta sencillo cuando se usan árboles 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: Claude 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 es las diferentes clases a ser clasificadas, en nuestro caso dos: desertores y no desertores. Mientras más alta es 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 fórmula:

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

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, lo que significa 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, hacer 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.

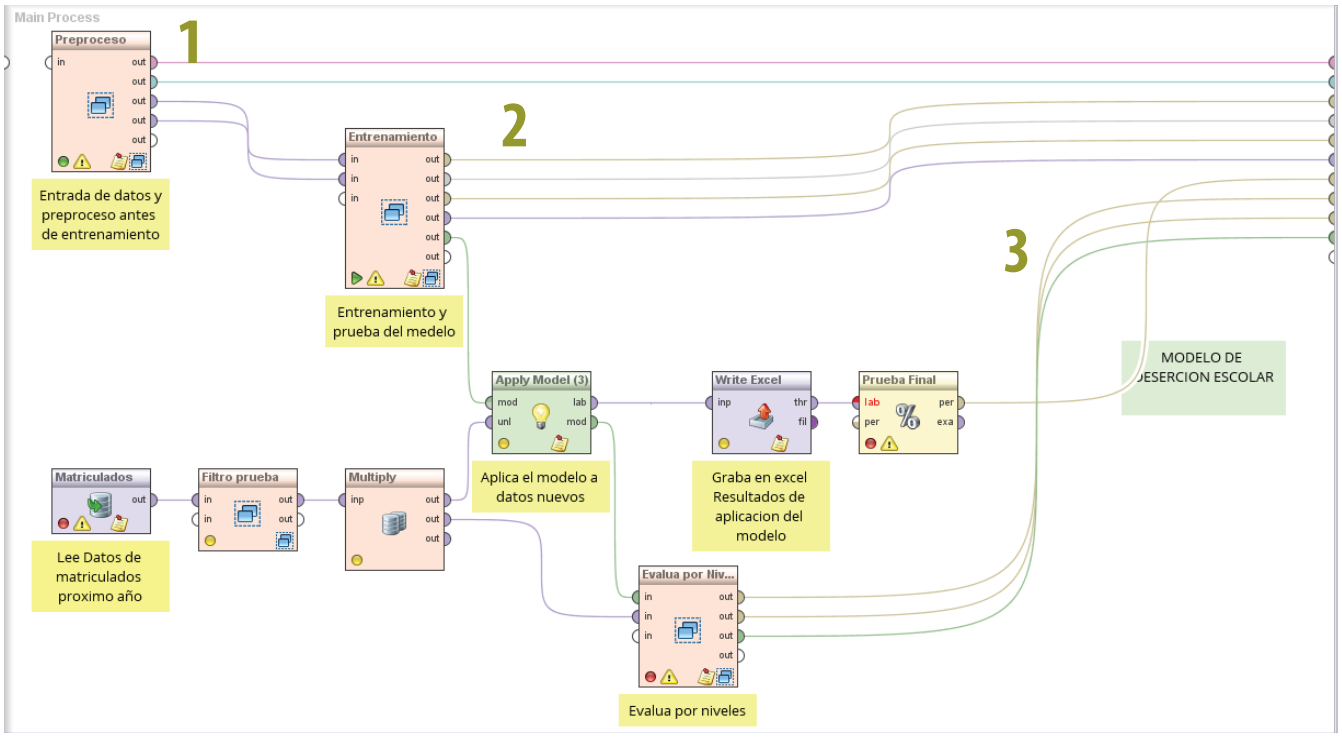
El fenómeno denominado *overfitting* consiste en que el modelo aprende solo los ejemplos del *training set,* pero no tiene la capacidad de clasificar correctamente ejemplos nuevos, es decir, pierde capacidad de generalización predictiva.

b) **Estructura del Modelo Desarrollado con RapidMiner**

Después de realizar la selección de las técnicas y modelos procederemos a la Evaluación del Modelo, para esto se usa la herramienta RapidMiner Studio, versión 6.4.000, la cual es un software especializado en el desarrollo de modelos de minería de datos y machine learning, muy flexible, expandible y de bajo costo.

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 la siguiente gráfica puede verse el modelo a nivel superior.

Gráfica 5.1 Nivel superior de modelo de deserción escolar



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

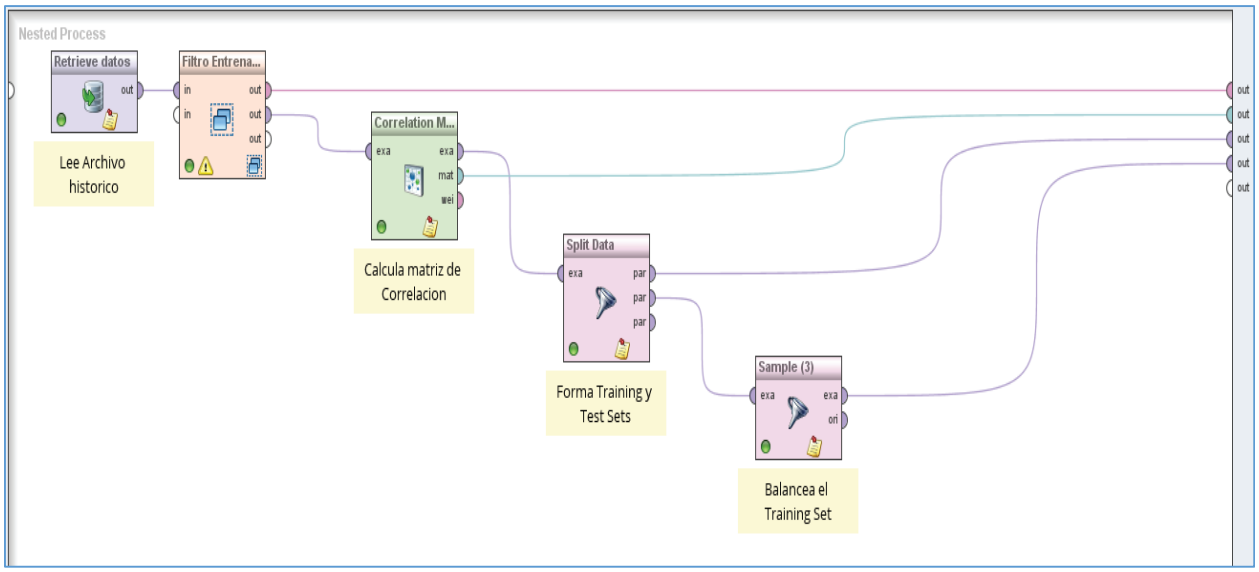
El módulo 2: Es donde residen las operaciones más importantes del modelo, ya que es ahí donde se prueban los algoritmos de clasificación y predicción. En este módulo también se hace la optimización de los algoritmos de clasificación.

El módulo 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, lo que significa que el módulo 3 puede usarse 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.

c) Pre-Procesamiento

El pre-procesamiento es el primer módulo del modelo de predicción de deserción escolar. El objetivo de este módulo es leer y preparar la data para la posterior etapa de entrenamiento y prueba del modelo. En la siguiente figura pueden apreciarse sus componentes.

Gráfica 5.2 Módulo de pre-procesamiento



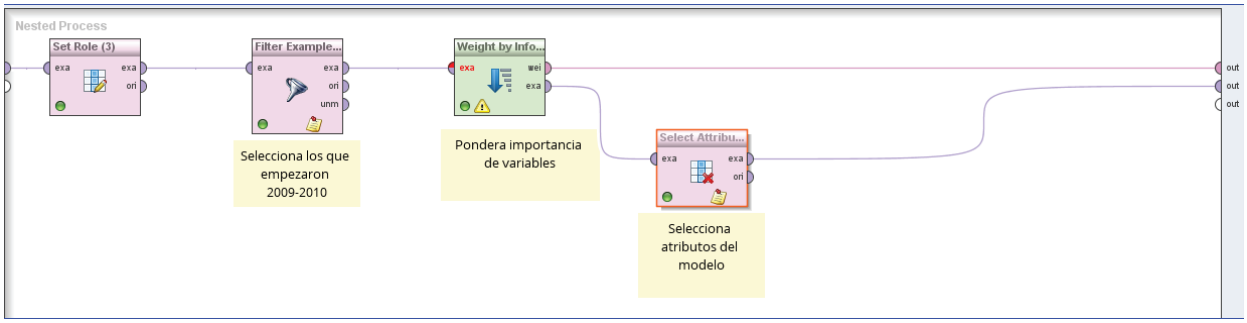
El módulo de pre-procesamiento tiene 4 salidas como puede apreciarse en la parte derecha de la gráfica 5.2, y ninguna entrada. Esto es así porque el módulo lee los datos del archivo SPSS previamente preparado y produce los resultados que se usarán en el siguiente módulo para el entrenamiento del modelo. El módulo de pre-procesamiento lo componen los siguientes operadores:

Retrieve datos: Este operador tiene un solo parámetro, que es el nombre de la base de datos donde residen las informaciones importadas desde SPSS, en este caso, esa base de datos se llama "datos". En la siguiente gráfica se muestra el parámetro del operador.

Gráfica 5.3 Parámetros del operador Retrieve datos

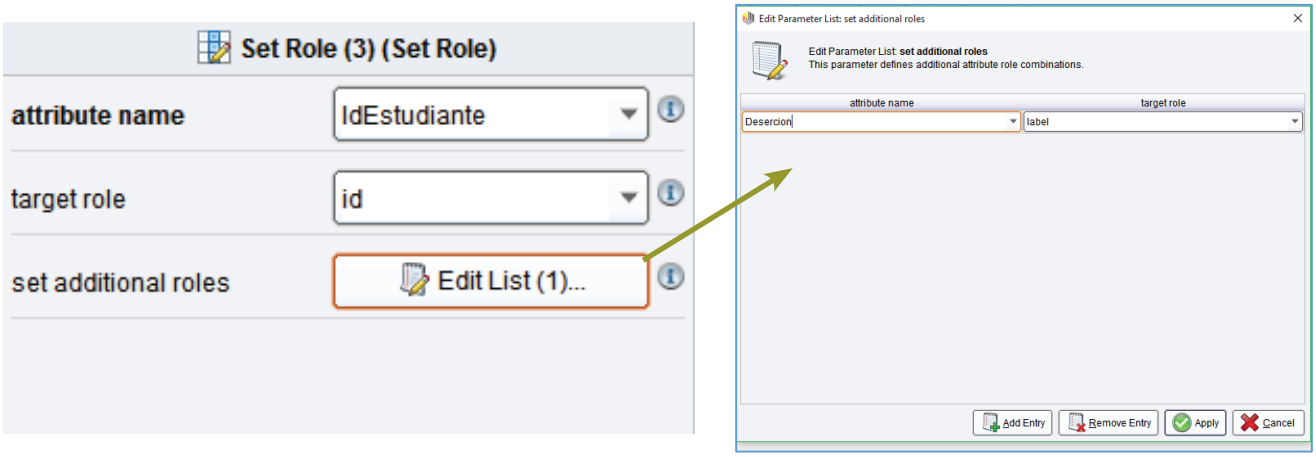
Filtro entrenamiento: Este es un operador subproceso el cual agrupa otros operadores mostrados a continuación.

Gráfica 5.4 Operador Filtro Entrenamiento



Set Role: Este operador asigna el rol de id a la identificación del estudiante y label al campo deserción. Estos son atributos especiales para saber que el campo deserción es el que el algoritmo tratará de predecir. A continuación la parametrización de este operador.

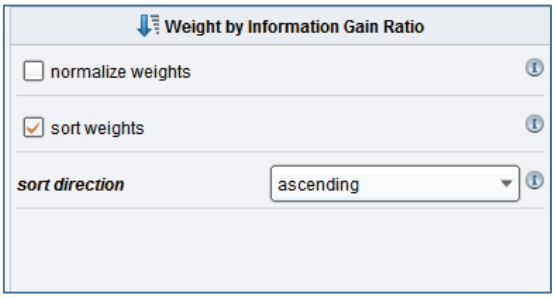
Gráfica 5.5 Parametrización Operador Set Role



Filter Examples: A continuación tres operadores Filter Example, cuyo objetivo es depurar los ejemplares que se usarán en el entrenamiento y la prueba del modelo.

El segundo incluye en la muestra solo a los estudiantes que estaban en la muestra en el primer año académico del estudio, 2009-2010. Se configuran de manera similar al primer filtro.

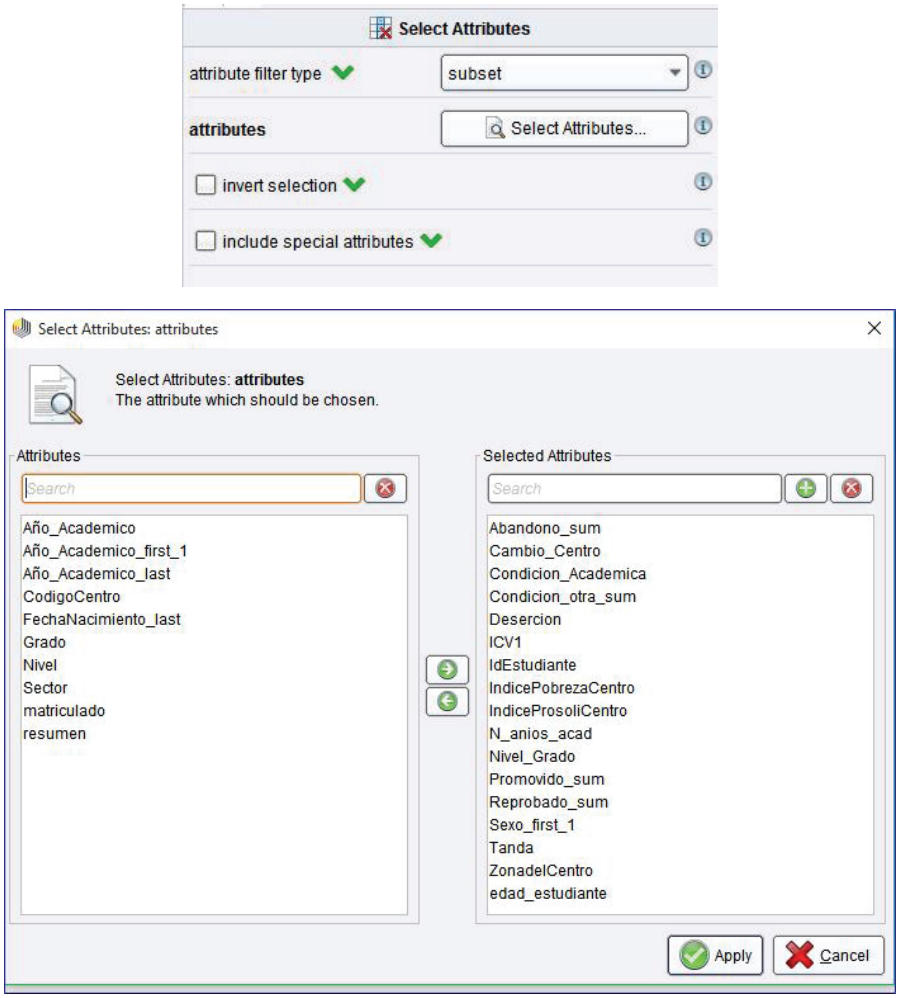
Gráfica 5.6 Parámetros del operador Filter Examples



Select Attributes: En este operador se seleccionan los atributos que serán usados en el modelamiento y se descartan los demás. La información provista por el operador Wright by Information Gain Ratio es útil para hacer esta discriminación.

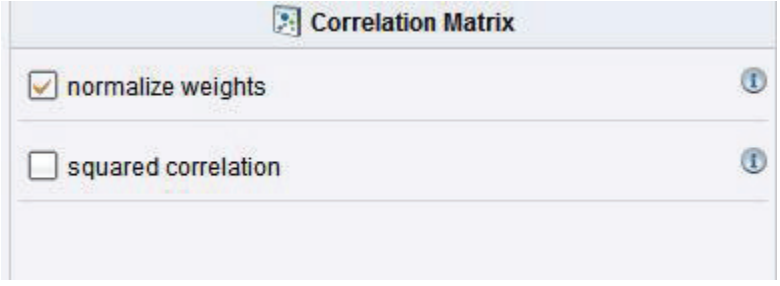
La modalidad escogida fue la de Subset en el parámetro "attribute filter type", por lo tanto se asigna Select Attribute al parámetro Attributes. Luego se deben asignar los atributos de lugar. En la gráfica 5.8 se muestran estas asignaciones y como puede apreciarse se excluyeron 11 atributos de los 27 que, originalmente, contiene la base de datos. Los atributos excluidos son los que se muestran en la parte izquierda de la parte inferior de la gráfica 5.8.

Gráfica 5.8 Parámetros del operador Select Attributes



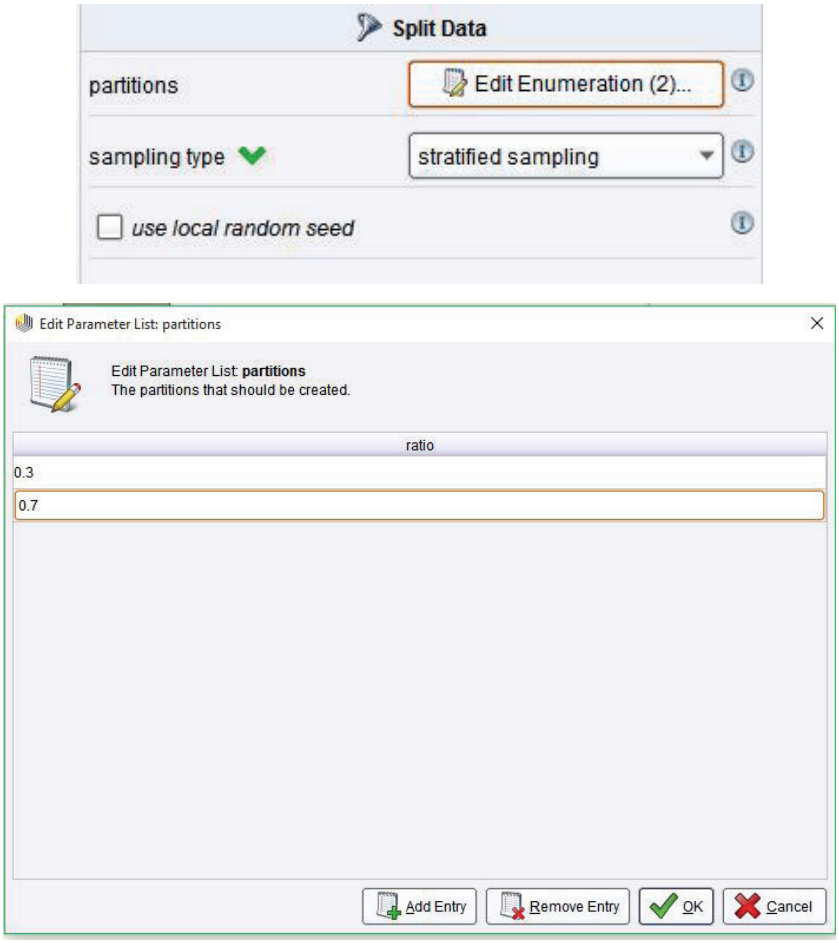
Correlation Matrix: Este operador construye una matriz de correlación de los atributos seleccionados en el operador anterior, con la finalidad de verificar si existe alta correlación entre esos atributos, y si es el caso proceder a descartar alguno de ellos. En este operador simplemente se normalizan las ponderaciones.

Gráfica 5.9 Parámetros del operador Correaltion Matrix



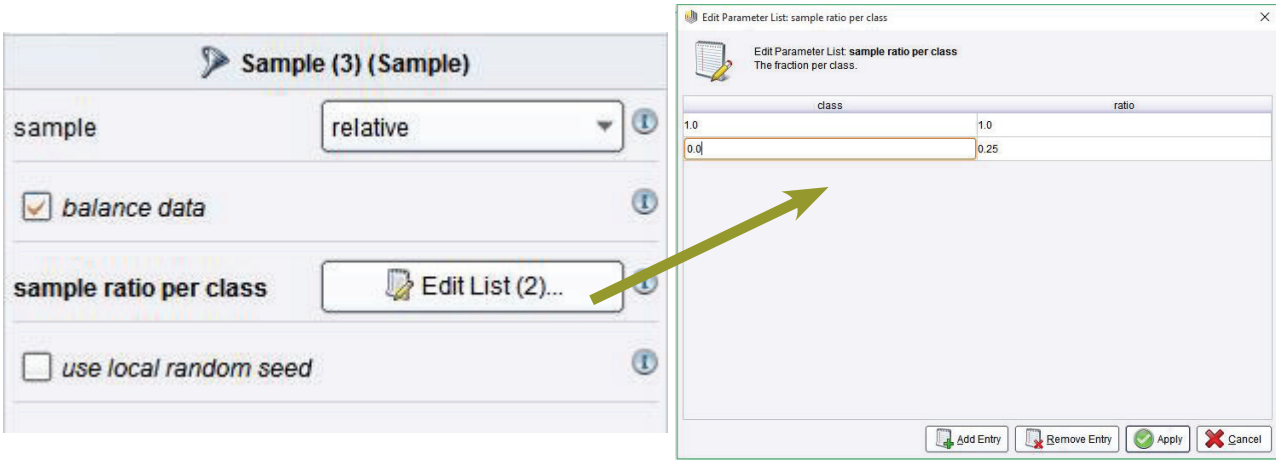
Split Data: Este operador divide la base de datos en dos partes: una para el entrenamiento del modelo (70% de los datos) y otra para la prueba del modelo (30% de los datos). En la gráfica 5.10 se muestran los parámetros asignados.

Gráfica 5.10 Parámetros del operador Split data



Sample (Balancea el Training Set): Este operador asegura una buena representación en la muestra del training set de los estudiantes que desertaron, ya que son minoría, esto tiende a que el modelo aprenda a identificar los no desertores con más exactitud que los desertores. En el training set se incluye el 100% de los ejemplos que son desertores y solo el 25% de los no desertores. Esto resuelve el problema del desbalance entre ambas clases de registros.

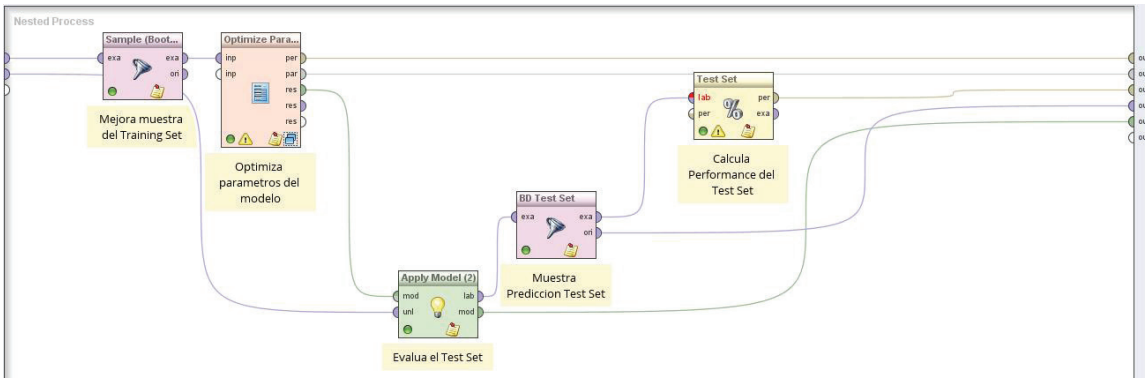
Gráfica 5.11Parámetros del Sample operador para Balancear Training Set



d) **Componentes del Entrenamiento y Prueba del Modelo**

En este módulo se usan los insumos del módulo de pre-procesamiento y se construye el modelo que se usará para en la predicción de la deserción escolar de los estudiantes. A continuación se describen los operadores del módulo.

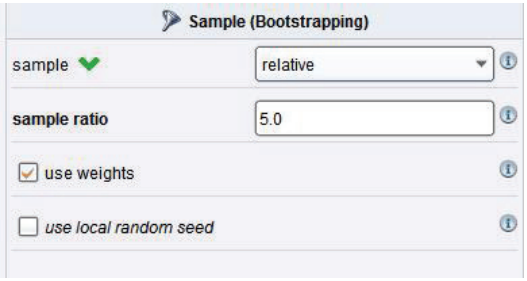
Gráfica 5.12 Módulo de entrenamiento y prueba del modelo



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.

Sample (Bootstrapping): Debido a que el porcentaje de estudiantes que desertan es relativamente pequeño con respecto al de que no lo hacen, para mejorar el poder de predicción del modelo, se hace un muestreo con reposición para así aumentar la cantidad de casos de deserción usados en el entrenamiento del modelo.

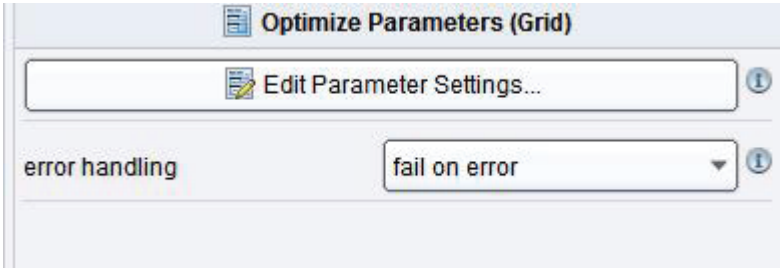
Gráfica 5.13 Parámetros del operador Sample Bootstrapping



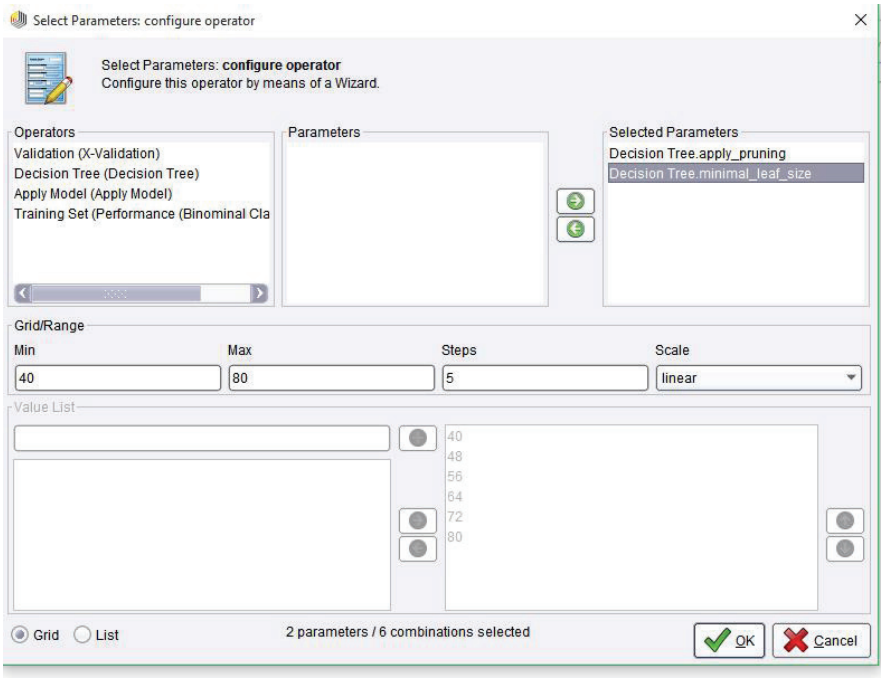
Se usó la opción *"relative"* en el parámetro simple, indicando que la muestra sería un porcentaje de los datos, y en este caso, un 5% tal como indica el *sample ratio.*

Optimize Parameters (Grid): La función de este operador es buscar entre una serie de opciones, la configuración más apropiada para el algoritmo de clasificación (árbol de decisión). En este caso se hace click en *"Edit Parameters Settings"* en la gráfica 5.14 y tenemos la gráfica 5.15.

Gráfica 5.14 Parámetros del operador Optimize Parameters (Grid)



Gráfica 5.15 Configuración de parámetros del operador Optimize Parameters (Grid)



Los parámetros del algoritmo árbol de decisión a configurar y el rango de variación fueron los siguientes:

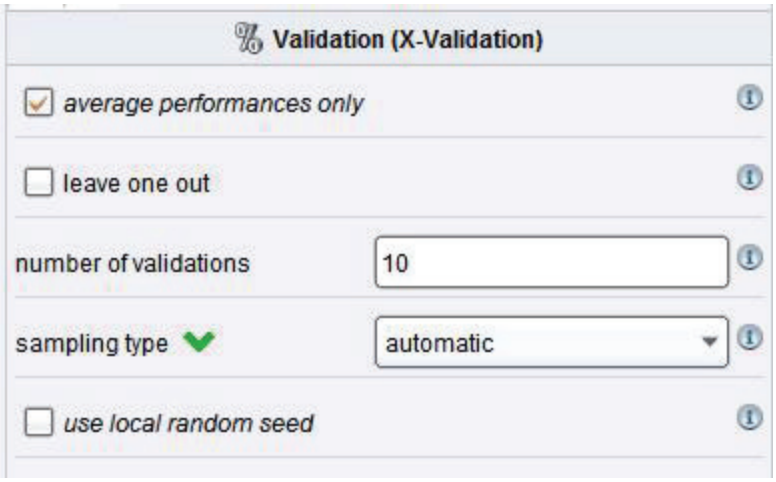
a) Decision Tree.apply_prunning: false.

b) Decision Tree.minimal_leaf_size: Mínimo 40, máximo 80, numero de pasos 5, escala lineal.

Haciendo click en el operador Optimize Parameters (Grid), el cual es un subproceso, se encuentra otro operador, que a su vez es también otro subproceso.

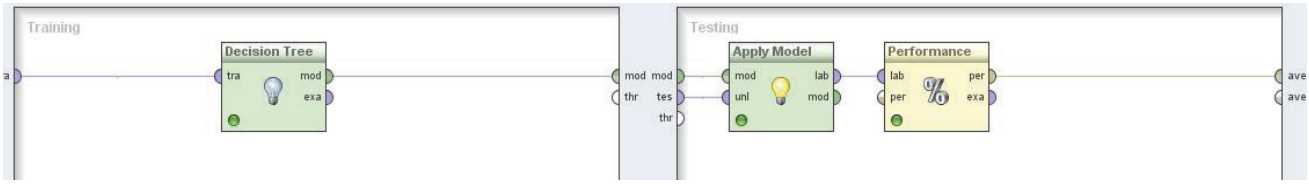
% Validation (X-validation): Este operador tiene como entrada el *training set,* y como salida: a) el modelo ajustado; b) el *training set*; y c) La tabla de contingencia con los resultados de la prueba del modelo. Este operador hace una validación cruzada, es decir, con los datos del *training set*, separa aleatoriamente una parte para entrenamiento y otra validación (diferente del *test set*). En la gráfica 5.16 se aprecia la configuración de este operador.

Gráfica 5.16 Configuración del operador % Validation (X-calidation)



Se hicieron 10 validaciones con muestreo automático para la validación, lo cual implica muestreo estratificado. Haciendo click en este operador, podemos ver su contenido mostrado en la gráfica 5.17.

Gráfica 5.17 Contenido del operador % Validation (X-validation)



Este operador a su vez, tiene dos partes. En la de la izquierda está el operador del algoritmo que se desea entrenar (árbol de decisión). En la derecha se colocan dos operadores, uno para aplicar el modelo y otro para evaluar el desempeño en la clasificación. A estos tres operadores se les asignó la configuración de parámetros que traen por *default*.

Continuando con los demás operadores del módulo de entrenamiento y prueba del modelo tenemos:

Apply Model: Este operador tiene como entrada el test set y el modelo entrenado y como salida: a) el test set con las predicciones de deserción para cada estudiante; y b) el modelo ajustado para que se muestre en la perspectiva de resultados.

BD test Set: Este operador se usa como comodín para producir dos salidas: una para la base de datos del test set con la predicción de deserción para los estudiantes y otra para usarla para la evaluación del desempeño del modelo con los datos del test set.

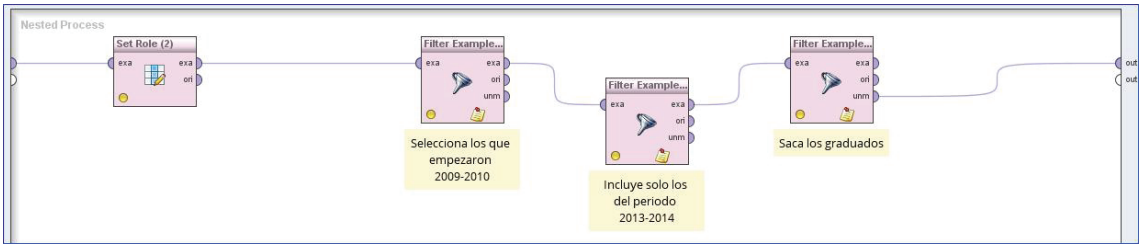
Performance (Binomial Classification): Este operador hace la evaluación del modelo con el test set y tiene como salida la tabla de contingencia con los resultados.

En este módulo se alimenta el modelo ajustado en los módulos anteriores con los datos del último año académico para hacer la predicción de los estudiantes que son propensos a desertar. Está compuesto por tres operadores:

Retrieve datos (Matriculados): Este operador lee la base de datos del último año académico en formato similar al usado en el training set.

Filtro Prueba: Este es un operador del tipo subproceso el cual contiene los operadores mostrados en la siguiente figura.

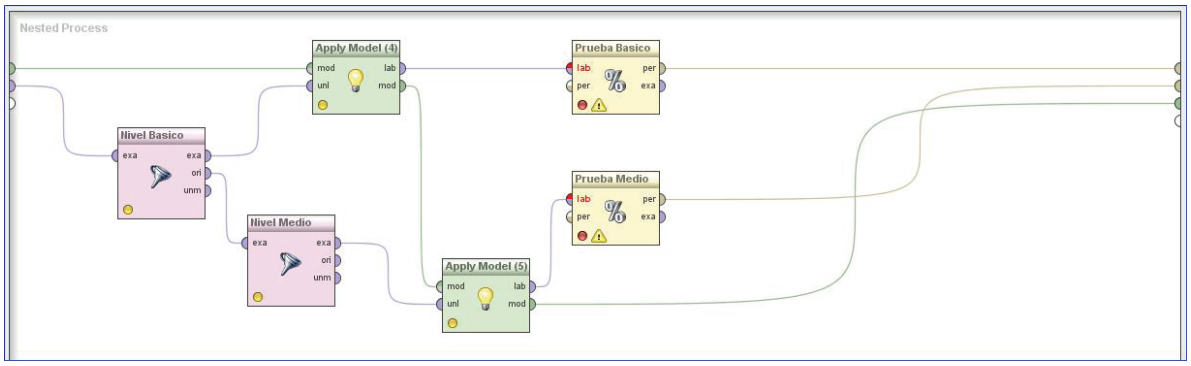
Gráfica 5.18 Operador Filtro Prueba



Como puede observarse, usa los mismos filtros que se usaron en la preparación de la base de datos de entrenamiento y prueba, con la diferencia de que ahora se incluyen solamente los estudiantes del año académico 2013-2014, para probar la exactitud de predicción del modelo y también se excluyen los graduandos de la base de datos de predicción.

Para probar la bondad del modelo se somete la base de datos de del período 2013-2014 al modelo de predicción completa por un lado, y en el subproceso se *"evalúa por niveles"* y se hace la misma predicción, pero separada para los estudiantes del nivel Básico y el nivel Medio.

Gráfica 5.19 Subproceso Evalúa por Niveles



Como puede apreciarse en la gráfica, se filtra la base de datos en los niveles Básico y el Medio y se aplica y evalúa el modelo por separado. De este subproceso salen las evaluaciones de los dos niveles y el modelo para ser mostrado en la perspectiva de resultados.

Apply Model: Usando el modelo ajustado con los datos nuevos, hace la predicción de deserción.

Write Excel: Este operador escribe a un archivo Excel en el folder especificado, las predicciones de deserción para los estudiantes de la base de datos del último año académico.

Como puede apreciarse en la tabla 5.1, se agregaron tres atributos: a) *Confidence* (0) que es la probabilidad de que la deserción sea cero (0) (no deserción); b) *Confidence* (1) que es la probabilidad de que la deserción sea 1 (si deserción); y c) predicción (Desercion) que será 1 o cero (0).

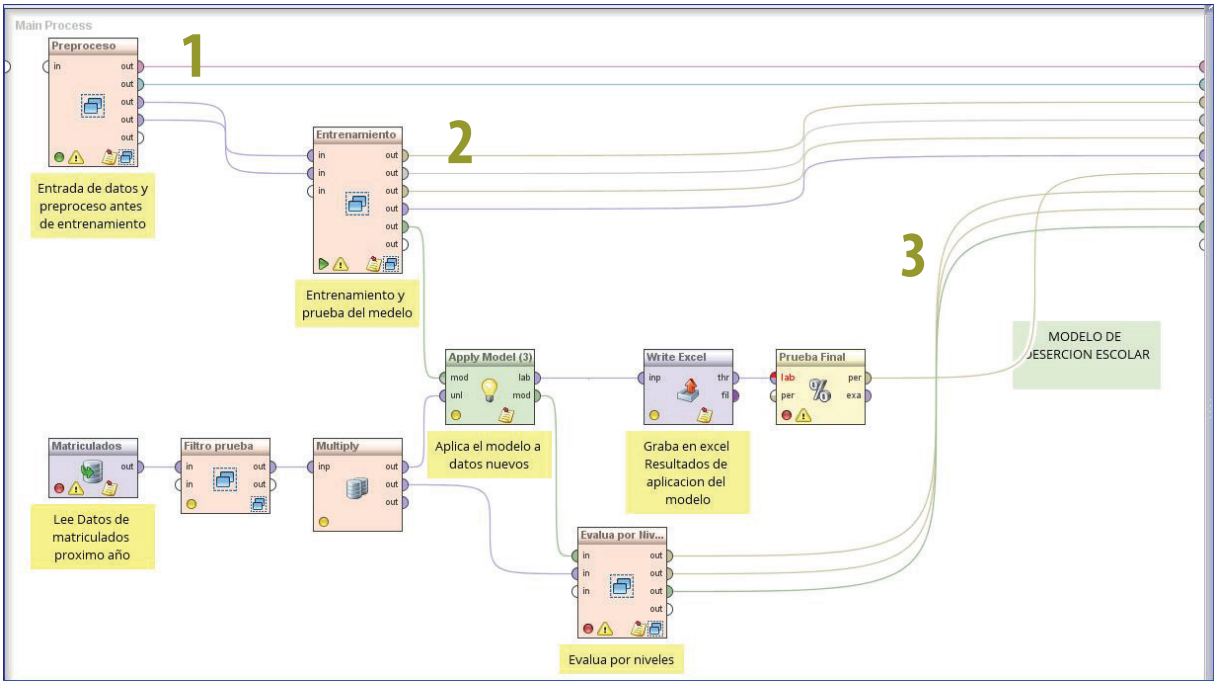
Tabla 5.1 Lista de atributos de salida

NO.	COLUMNA	ATRIBUTO	EJEMPLO
1	A	Año_Academico_first_1	2009-2010
2	B	Año_Academico_last_1	2013-2014
3	C	CodigoCentro_first	2358.0
4	D	CodigoCentro_last	359
5	E	Sector_last	PÚBLICO
6	F	Tanda_last	MATUTINA
7	G	FechaNacimiento_last	2004-11-29 00:00:00
8	H	Sexo_first_1	Masculin
9	I	ICV1	0
10	J	nivel_grado_last	14
11	K	Abandono_sum	.0
12	L	Promovido_sum	2.0
13	M	Reprobado_sum	0
14	N	Condicion_otra_sum	0
15	O	anios_acad	2.0
16	P	EdadEstudiante	10.5
17	Q	ZonadelCentro	URBANA-M
18	R	IndicePobrezaCentro	69.9
19	S	IndiceProsoliCentro	8.2
20	T	Condicion_Acad_last	Promov
21	U	IdEstudiante	35.0
22	**V**	**confidence(0)**	**.9**
23	**W**	**confidence(1)**	**.1**
24	**X**	**prediction(Desercion)**	**0**

e) **Componente de Predicción del Modelo**

El módulo 3 muestra el componente del modelo que se usará cada año para hacer la predicción de los estudiantes que desertarán y luego se probará, cuando se conozcan los resultados de las inscripciones, con el objeto de producir un despliegue de la información de deserción por centro educativo, con el propósito de que la dirección del centro tome medidas de intervención con los estudiantes de alto riesgo de deserción. Además de que se podrá medir qué tan preciso fue el modelo en su predicción.

Gráfica 7.1 Nivel superior de modelo de deserción escolar



5.6 Resultados del Entrenamiento, Prueba y Predicción del Modelo

En esta sección de explicarán los resultados del proceso de entrenamiento y prueba del modelo de predicción de deserción escolar que arroja los resultados evaluativos en función de los datos de entrenamiento y prueba.

El entrenamiento del modelo se realizó con la cohorte de años escolares 2009-2014 y con una muestra del 70% de los estudiantes que iniciaron en el año escolar 2009-2010, esto así por entender que el mejor modelo de información es el de los estudiantes que están comprendidos dentro del período de cohorte desde el inicio, y que muestran el evento de salida del sistema educativo nacional por deserción o conclusión de niveles (Básico o Medio). La prueba del modelo se realizó con el 30% de los estudiantes restantes de esa cohorte, como se muestra en el cuadro siguiente. Ver los cuadros con las cifras de estudiantes que muestran el entrenamiento y la prueba en la sección 4.6.

Además procedemos a introducir los datos de los estudiantes matriculados en el año escolar 20013-2014 y que también iniciaron el ciclo en el 2009-2010 con su condición académica final con el objeto de predecir su estado de desertor o no en el próximo año académico 2014-2015. En esta predicción se excluyó a los estudiantes graduados del ciclo de los 12 años de escolaridad.

a) **Prueba de Bondad del Modelo *(Trainning)***

Para evaluar la bondad del modelo y para que prediga la probabilidad de que un estudiante deje la escuela, se entrena el modelo con los datos del *training* set (período escolar 2009-2014 de la cohorte), cuidando que no se produzca el fenómeno denominado *"overfitting"*, que consiste en que el modelo aprende solo los ejemplos del *training set*, pero no tiene la capacidad de clasificar correctamente ejemplos nuevos. En la tabla 6.3 se muestra 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 6.3 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.

b) **Prueba de la Capacidad Predictiva del Modelo (Test)**

Se hizo el mismo análisis con la base de datos del test set. En la tabla 6.4 se resumen los resultados. El test set contiene 54,359 casos, es decir, el 30% de la base de datos completa.

Tabla 6.4 Evaluación del modelo con datos del Test Set

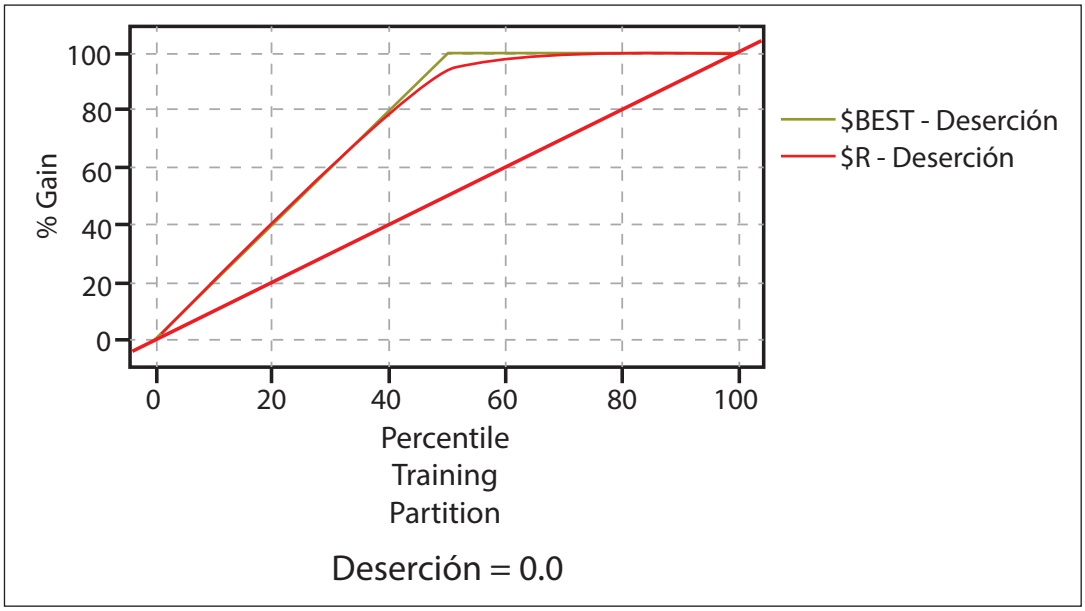
ACCURACY: 94.56%			
	TRUE 1.0	TRUE 0.0	CLASS PRECISION
pred 1.0	5409	2182	71.26%
pred 0.0	775	45993	98.34%
class recall	87.47%	95.47%	

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 clasificada, el 95.47% (*Specificity*) de los que no desertaron que se clasificó correctamente. Estos son resultados muy satisfactorios considerando que el test set contiene datos de estudiantes que el modelo no usó durante el entrenamiento, por lo que puede deducirse 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 sí lo harían), 775 (12.53% de los que si desertaron) fueron falsos negativos (estudiantes que sí desertaron y no se pronosticaron como tal).

La gráfica 6.1 muestra la curva ROC (Receiver Operating Characteristic) para los datos del test set. Esta curva es un indicador de qué 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 puede observarse es muy cercana a la curva roja mostrando una gran bondad en el ajuste.

Gráfica 6.1 Curva ROC para datos del Test Set



c. Predicción del Modelo (Verificación)

En la tabla 7.1 se muestran los resultados de usar el modelo para predecir los estudiantes que desertarían en el período 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 módulo se usará cuando se tengan los resultados del período 2014-2015 y la matriculación del 2015-2016.

Tabla 7.1 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 de 3,939 falsos positivos.

Los resultados de la predicción por niveles académicos (Básico y Medio) se muestran en las siguientes tablas.

Tabla 7.2 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%	83.63%	

Tabla 7.3 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 nos da una muestra de la información de predicción de cada estudiante del set de 26,465, 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) o uno (1).

IDESTUDIANTE	AÑO_ACADEMICO	CODIGOCENTRO	SECTOR	TANDA	NIVEL	NIVEL_GRADO	CONDICION_ ACADEMICA	SEXO_FIRST_1	ICV1	ZONADELCENTRO	AÑO_ ACADEMICO_LAST	N_ANIOS_ACAD	$C-DESERCION (PREDICTION)
35	2013-2014	359.0	PUBLICO	MATUTINA	Básico	14.0	Promovido	Masculino	0.0	URBANA-M	2013-2014	2	1
134	2013-2014	226.0	PUBLICO	VESPERTINA	Básico	17.0	Promovido	Femenino	1.0	URBANA-M	2013-2014	5	0
258	2013-2014	3061.0	PUBLICO	VESPERTINA	Básico	15.0	Promovido	Femenino	1.0	URBANA	2013-2014	3	0
356	2013-2014	230.0	PUBLICO	MATUTINA	Básico	16.0	Promovido	Femenino	0.0	URBANA-M	2013-2014	5	1
429	2013-2014	224.0	PUBLICO	VESPERTINA	Básico	12.0	Promovido	Masculino	0.0	URBANA-M	2013-2014	4	0
1605	2013-2014	5710.0	PUBLICO	MATUTINA	Medio	21.0	Promovido	Femenino	1.0	URBANA-M	2013-2014	5	1
1637	2013-2014	229.0	PUBLICO	MATUTINA	Básico	16.0	Promovido	Femenino	0.0	URBANA-M	2013-2014	2	0
1777	2013-2014	13402.0	PUBLICO	VESPERTINA	Medio	21.0	Promovido	Femenino	1.0	RURAL	2013-2014	4	0
1800	2013-2014	4852.0	PUBLICO	MATUTINA	Básico	17.0	Promovido	Femenino	0.0	URBANA-M	2013-2014	3	1
1959	2013-2014	528.0	PUBLICO	VESPERTINA	Básico	16.0	Promovido	Masculino	0.0	RURAL	2013-2014	4	0
1976	2013-2014	2251.0	PUBLICO	MATUTINA	Básico	17.0	Promovido	Masculino	0.0	URBANA	2013-2014	4	0
2005	2013-2014	216.0	PUBLICO	VESPERTINA	Básico	16.0	Promovido	Masculino	0.0	URBANA	2013-2014	4	0
2062	2013-2014	12.0	PUBLICO	MATUTINA	Básico	15.0	Promovido	Masculino	1.0	URBANA-M	2013-2014	4	0
2113	2013-2014	6075.0	PUBLICO	VESPERTINA	Básico	17.0	Abandono	Femenino	1.0	URBANA	2013-2014	5	0
2231	2013-2014	216.0	PUBLICO	VESPERTINA	Básico	15.0	Promovido	Femenino	1.0	URBANA	2013-2014	5	0
2985	2013-2014	5735.0	PUBLICO	NOCTURNA	Medio	23.0	Promovido	Femenino	0.0	URBANA-M	2013-2014	5	1
3017	2013-2014	5738.0	SEMIOFICIAL	MATUTINA	Medio	23.0	Promovido	Femenino	0.0	URBANA	2013-2014	5	0
3169	2013-2014	4286.0	PUBLICO	VESPERTINA	Básico	17.0	Promovido	Masculino	0.0		2013-2014	3	0
3348	2013-2014	228.0	PUBLICO	MATUTINA	Básico	17.0	Promovido	Femenino	1.0	URBANA-M	2013-2014	5	1
3378	2013-2014	5726.0	PUBLICO	MATUTINA	Básico	18.0	Promovido	Femenino	0.0	URBANA-M	2013-2014	5	1
3718	2013-2014	5710.0	PUBLICO	MATUTINA	Medio	24.0	Reprobado	Femenino	0.0	URBANA-M	2013-2014	5	0

d. 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 (unos o verdaderos positivos) es muy inferior a la de no deserciones (ceros 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 con relación al total observado. Para lograrlo el modelo ha tenido que equilibrar, mediante la aplicación de técnicas de *boosting* y balanceo de parámetros, 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.

Este fenómeno contribuye a sesgar los resultados de la totalidad de unos (1) predichos (true positive) con 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 con 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 puede interpretarse 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:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{recall}}{\text{Precision} + \text{recall}}$$

Para el caso de la predicción del año 2013-2014 (ver el cuadro siguiente) 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. 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 trabajo 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*), 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 ceros 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 (ver los cuadros siguientes).

CONTIGENCY TABLE: PREDICCION 2013-2014						
Decision Tree for Schollar Churn			**Predicted**			
			Desercion			**Percentage Correct**
			0	1	Totals	
Observed	Desercion	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%**

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 = True Pos/#Observed Pos**
01	643	False Negative	**82.69%**	**Accuracy = % Overall Predicted**
00	18,827	True Negative	**43.69%**	**Precision = True Pos/#Predicted Pos**
			0.5715	**F1 Score**

Cuadro 14

CONTIGENCY TABLE: PREDICCION 2013-2014 ESTUDIANTES NIVEL BASICO						
Decision Tree for Schollar Churn			**Predicted**			
			Desercion			**Percentage Correct**
			0	1	Totals	
Observed	Desercion	0	12,513	2,451	14,964	83.62%
		1	396	1,638	2,034	**80.53%**
	Totals		12,909	4,089	16,998	82.08%
		Percentage Correct	96.93%	**40.06%**	68.50%	**83.25%**

LEYENDA			INDICADORES DE RESULTADOS DE PRUEBA	
10	2,451	False Positive	**83.62%**	**Specificity = True Neg/#Observed Neg**
11	1,638	True Positive	**80.53%**	**Sensitivity = True Pos/#Observed Pos**
01	396	False Negative	**83.25%**	**Accuracy = % Overall Predicted**
00	12,513	True Negative	**40.06%**	**Precision = True Pos/#Predicted Pos**
			0.5350	**F1 Score**

CONTIGENCY TABLE: PREDICCION 2013-2014 ESTUDIANTES NIVEL BASICO						
Decision Tree for Schollar Churn			**Predicted**			
			Desercion			**Percentage Correct**
			0	1	Totals	
Observed	Desercion	0	6,314	1,488	7,802	80.93%
		1	247	1,418	1,665	**85.17%**
	Totals		6,561	2,906	9,467	83.05%
		Percentage Correct	96.24%	**48.80%**	72.52%	**81.67%**

LEYENDA			INDICADORES DE RESULTADOS DE PRUEBA	
10	1,488	False Positive	**80.93%**	**Specificity = True Neg/#Observed Neg**
11	1,418	True Positive	**85.17%**	**Sensitivity = True Pos/#Observed Pos**
01	247	False Negative	**81.67%**	**Accuracy = % Overall Predicted**
00	6,314	True Negative	**48.80%**	**Precision = True Pos/#Predicted Pos**
			0.6204	**F1 Score**

NOTAS:
F1 Score >= 0.80 muy buena o excelente precisión (verde)
< 0.80 pero >= 0.70 buena precisión (Amarillo subiendo)
< 0.70 pero >= 0.50 regular precisión (Amarillo bajando)
< 0.50 mala precisión (rojo)
Accuracy = % Overall Predicted - Proporción de predicciones acertadas True Negative y True Possitives
>= 0.80 verde muy buena (verde)
>= 0.60 y <0.80 buena (Amarillo)
<0.60 mala (rojo)

5.7 Gestión del Modelo y Análisis de Desviación

Es difícil mantener el estado de los modelos de minería de datos. Cada modelo de minería tiene un ciclo de vida. En algunas instituciones, los patrones de datos son relativamente estables y los modelos no requieren de reciclaje con frecuencia. Sin embargo, en los patrones de muchas instituciones varían con frecuencia, esto significa que las nuevas reglas de asociación aparecen cada día. En el caso de la deserción escolar los cambios pueden ser muy dinámicos debido a que el MINERD implementa políticas de inversión en el sistema de educación y de intervención en los centros educativos y los estudiantes, que revertirían las tendencias a la deserción escolar.

En un proceso dinámico como este, el modelo de predicción de la deserción escolar debe ser evaluado anualmente al cierre del año escolar. En última instancia, determinar la exactitud del modelo y la creación de nuevas versiones del mismo conlleva el uso de procesos automatizados: RapidMiner y SPSS Modeler ofrecen una herramienta versátil de gestión de contenidos y versiones.

Este proyecto presenta una gran ventaja desde el punto de vista del monitoreo, y es que cada año cuando se hace 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.

Eso por un lado, por otro, puede medirse 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.

En el ciclo de vida del modelo se observa que una vez realizado el test de desviación del modelo se debe exportar el archivo de salida del modelo hacia la base de datos del Sistema de Gestión de Centros Educativos, con el objeto de que sirva a los directivos y al personal docente de cada centro para el seguimiento y la aplicación de políticas de retención de los alumnos y alumnas.

6. Análisis de Datos de los Resultados del Modelo

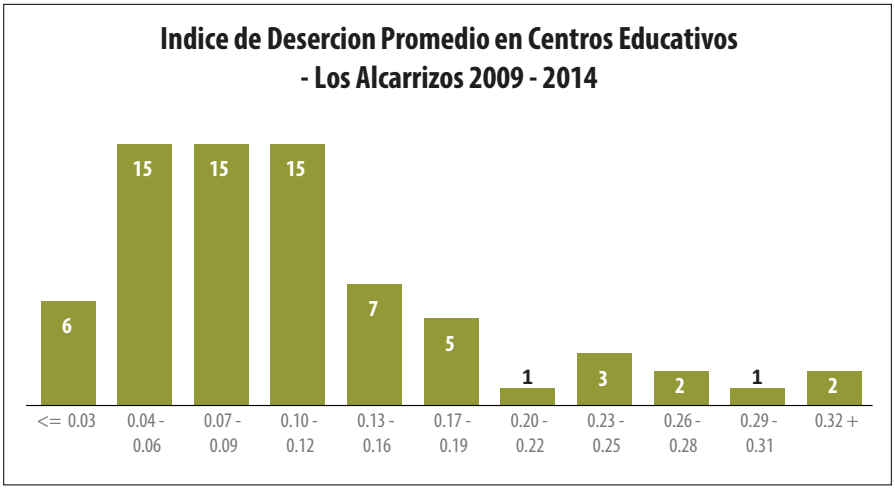
6.1 Objetivo

El objetivo es presentar un conjunto de reflexiones basadas en análisis de datos de las variables de deserción y otros factores en los 72 centros de Los Alcarrizos para el período 2009-2014.

6.2 Análisis de riesgo de deserción por centros educativos

En el cuadro y gráfico siguientes se muestran la distribución del índice de deserción de estudiantes en los cinco períodos de la cohorte (2009-2014). El índice de deserción es un promedio de la proporción de deserción de cada centro entre el total de estudiantes matriculados en los períodos mencionados. Los intervalos de clase seleccionados facilitan la representación de este índice y la frecuencia de centros que la poseen. Como puede notarse la mayor frecuencia de deserción de los centros se produce entre 0.03 y 0.12 acumulando un 70.8% (51 centros del total de 72). El 29.2% (21 centros) restante está entre 0.12 y 0.32.

INDICE DE DESERCION PROMEDIO EN CENTROS EDUCATIVOS - LOS ALCARRIZOS 2009 - 2014			
Interval	**Frequency**	**Percent**	**Cumulative Percent**
<= 0.03	6	8.33%	8.33%
0.04 - 0.06	15	20.83%	29.17%
0.07 - 0.09	15	20.83%	50.00%
0.10 - 0.12	15	20.83%	70.83%
0.13 - 0.16	7	9.72%	80.56%
0.17 - 0.19	5	6.94%	87.50%
0.20 - 0.22	1	1.39%	88.89%
0.23 - 0.25	3	4.17%	93.06%
0.26 - 0.28	2	2.78%	95.83%
0.29 - 0.31	1	1.39%	97.22%
0.32 +	2	2.78%	100.00%
Total	**72**	**100.00%**	



La cantidad de estudiantes inscritos en los 72 centros durante el año escolar 2013-2014 fue de 62,004, como se muestra en el cuadro siguiente.

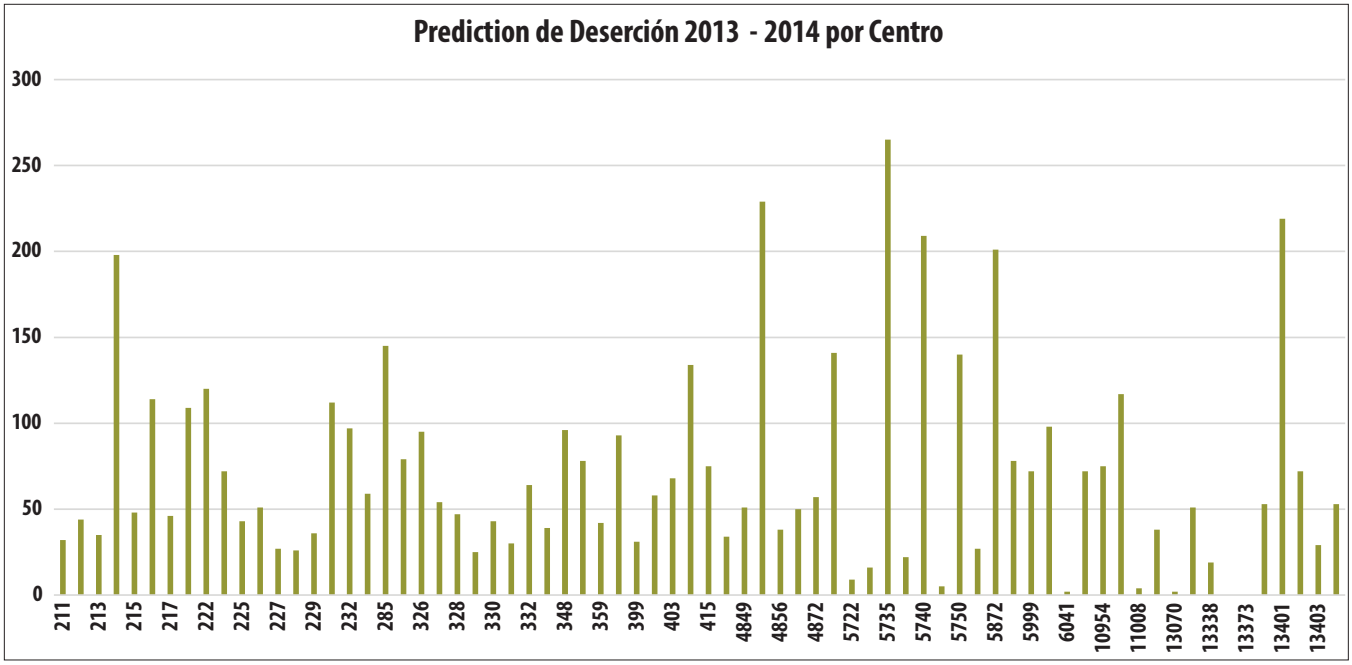
CENTROS LOS ALACRRIZOS		
TOTAL_ESTUDIANTES 2013-2014		
N	Valid	72
	Missing	0
Mean		861.17
Median		785.00
Mode		7[a]
Std. Deviation		575.572
Minimum		3
Maximum		2348
Sum		62004

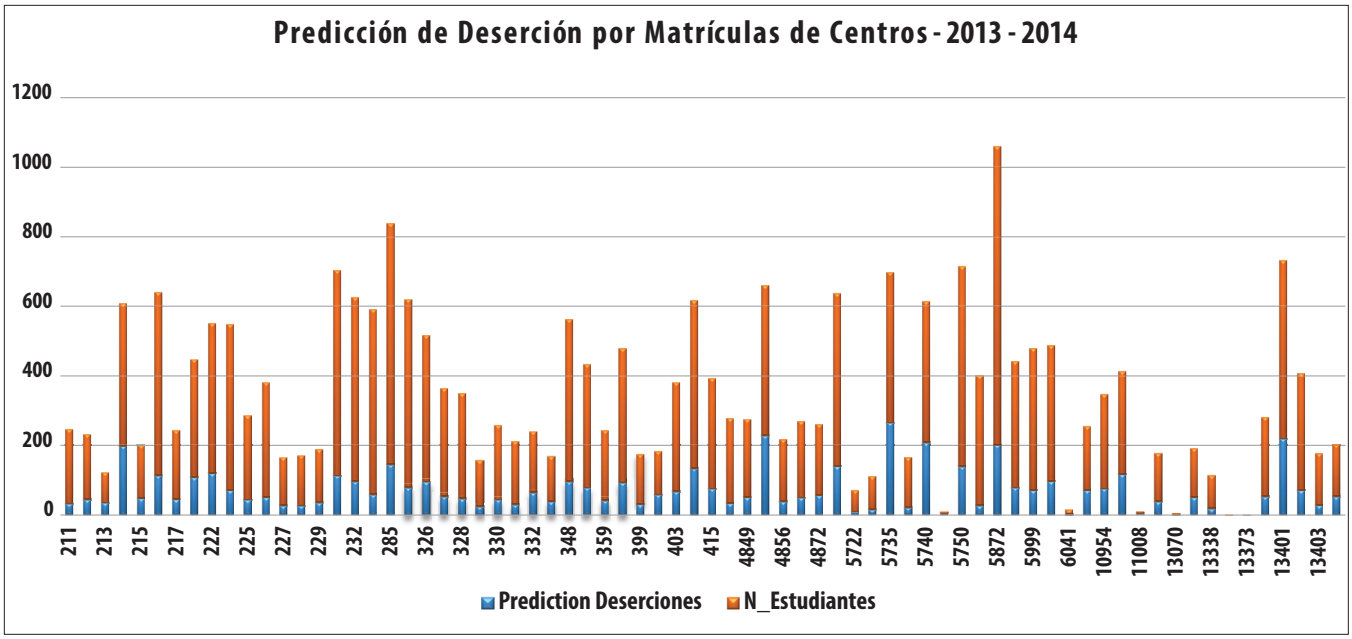
Para el análisis predictivo del modelo se seleccionaron solo los estudiantes registrados en el inicio de la cohorte, es decir, el año 2009-2010. Esa cantidad asciende a 20,202 como se muestra en el siguiente cuadro. Sobre la base de estos estudiantes se realizó la predicción de los estudiantes que desertarían en el año 2013-2014, según lo descrito en la sección 5.6, los que ascienden a 5,113 para un porcentaje de 25.31%. Esta cifra es el doble de lo esperado debido a que el proceso predictivo incluyo los falsos negativos (no desertores clasificados como desertores) según lo que se explicó en la sección anterior.

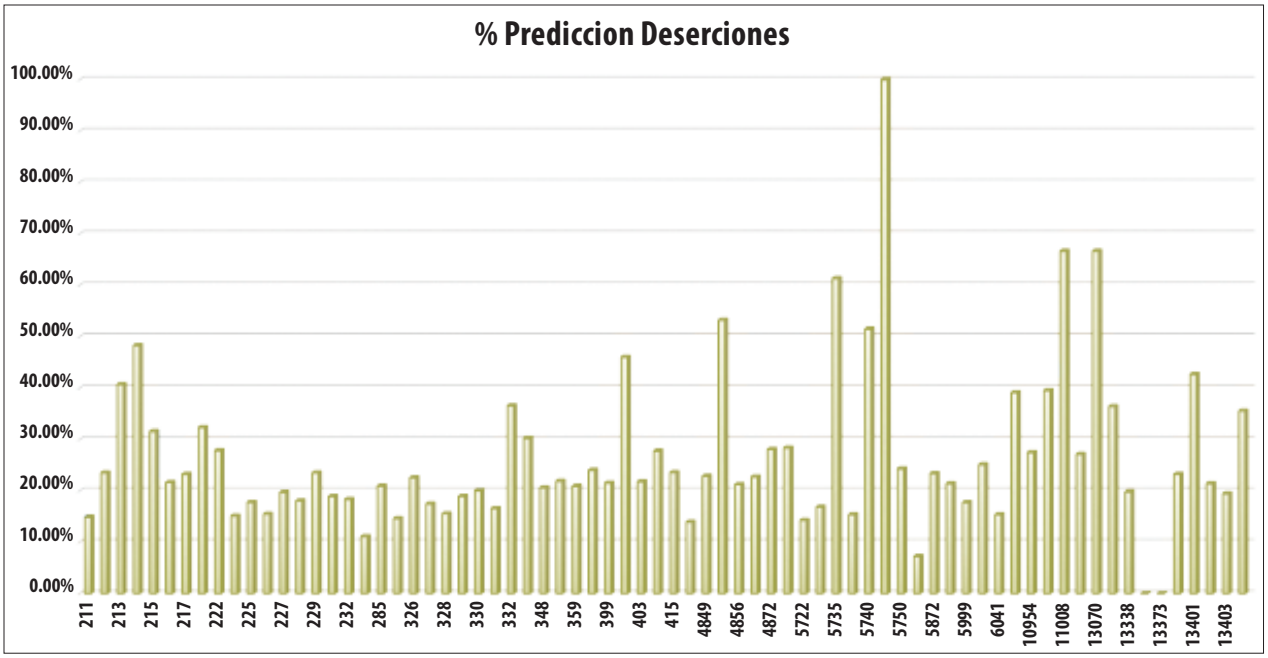
En los cuadros siguientes puede observarse la distribución de la predicción de deserción por centro educativo, su porcentaje con relación a la matrícula 2013-2014 (ingresados registrados 2009-2010).

PREDICCION DE DESERCIONES CENTROS LOS ALCARRIZO - 2013-2014			
CodigoCentro	**Prediction Deserciones**	**N_Estudiantes**	**% Prediccion Deserciones**
211	32	215	14.88%
212	44	187	23.53%
213	35	86	40.70%
214	198	410	48.29%
215	48	152	31.58%
216	114	527	21.63%
217	46	198	23.23%
218	109	337	32.34%
222	120	431	27.84%
224	72	476	15.13%
225	43	242	17.77%
226	51	329	15.50%
227	27	137	19.71%
228	26	144	18.06%
229	36	153	23.53%
230	112	592	18.92%
232	97	529	18.34%
259	59	531	11.11%
285	145	693	20.92%
319	79	540	14.63%
326	95	421	22.57%
327	54	309	17.48%
328	47	302	15.56%
329	25	132	18.94%
330	43	215	20.00%
331	30	181	16.57%
332	64	175	36.57%
333	39	129	30.23%
348	96	466	20.60%
349	78	356	21.91%
359	42	201	20.90%
396	93	387	24.03%
399	31	144	21.53%
400	58	126	46.03%
403	68	312	21.79%
404	134	483	27.74%
415	75	318	23.58%
4847	34	244	13.93%
4849	51	223	22.87%
4854	229	430	53.26%
4856	38	179	21.23%
4862	50	220	22.73%
4872	57	203	28.08%

PREDICCION DE DESERCIONES CENTROS LOS ALCARRIZO - 2013-2014			
CodigoCentro	**Prediction Deserciones**	**N_Estudiantes**	**% Prediccion Deserciones**
5710	141	497	28.37%
5722	9	63	14.29%
5726	16	95	16.84%
5735	265	432	61.34%
5738	22	143	15.38%
5740	209	406	51.48%
5741	5	5	100.00%
5750	140	576	24.31%
5799	27	373	7.24%
5872	201	859	23.40%
5873	78	364	21.43%
5999	72	406	17.73%
6000	98	390	25.13%
6041	2	13	15.38%
6075	72	184	39.13%
10954	75	273	27.47%
10955	117	296	39.53%
11008	4	6	66.67%
11171	38	140	27.14%
13070	2	3	66.67%
13158	51	140	36.43%
13338	19	96	19.79%
13349	0	1	0.00%
13373	0	1	0.00%
13400	53	228	23.25%
13401	219	513	42.69%
13402	72	336	21.43%
13403	29	149	19.46%
14125	53	149	35.57%
Totales	**5113**	**20202**	**25.31%**


Prediction de Deserción 2013 - 2014 por Centro
300
250
200
150
100
50
0
211 213 215 217 222 225 227 229 232 285 326 328 330 332 348 359 399 403 415 4849 4856 4872 5722 5735 5740 5750 5872 5999 6041 10954 11008 13070 13338 13373 13401 13403


Predicción de Deserción por Matrículas de Centros - 2013 - 2014
1200
1000
800
600
400
200
0
211 213 215 217 222 225 227 229 232 285 326 328 330 332 348 359 399 403 415 4849 4856 4872 5722 5735 5740 5750 5872 5999 6041 10954 11008 13070 13338 13373 13401 13403
Prediction Deserciones
N_Estudiantes



6.3 Factores de vulnerabilidad y riesgo de deserción

Para medir los factores de vulnerabilidad se tomó como referencia la deserción registrada en el año escolar 2013-2014. Como podremos observar en los cuadros siguientes existe una correlación positiva de 0.68 del índice de estudiantes con tarjeta Prosoli y el nivel de deserción dentro de los 72 centros educativos, medido por la correlación de Pearson. El indicador de índice de estudiantes con tarjeta Prosoli representa la proporción de estudiantes en condición de pobreza que inciden en el centro educativo.

La correlación del índice de pobreza del centro educativo también es positiva con relación a la deserción del centro educativo, 0.23.

CORRELATIONS			
		DESERCION_SUM	INDICE ESTUDIANTES CON TARJETA PROSOLI - CV1_SUM
Desercion_sum	Pearson Correlation	1	.677**
	Sig. (2-tailed)		.000
	N	72	72
Indice Estudiantes Con Tarjeta Prosoli - CV1_sum	Pearson Correlation	.677**	1
	Sig. (2-tailed)	.000	
	N	72	72
**. Correlation is significant at the 0.01 level (2-tailed).			

CORRELATIONS			
		DESERCION_SUM	INDICE POBREZA DEL CENTRO
Desercion_sum	Pearson Correlation	1	.225
	Sig. (2-tailed)		.057
	N	72	72
Indice Pobreza del Centro	Pearson Correlation	.225	1
	Sig. (2-tailed)	.057	
	N	72	72

6.4 Factores género y riesgo de deserción

La proporción de deserción de los estudiantes masculinos en el período 2013-2014 superó a la de estudiantes del género femenino (14% masculino y 12% femenino).

CUADRO 20-DESERCION POR GENERO 2013-2014				
Genero	**Desercion**	**Estudiantes**	**Totales**	**%Desercion**
Femenino	2940	24936	27876	11.79%
Masculino	3506	24831	28337	14.12%
Totales	**6446**	**49767**	**56213**	**12.95%**

7. Conclusión

7.1 Resumen de las etapas realizadas

En este proyecto hemos seguido la metodología CRISP_DM. La validación y preparación de datos es la etapa inicial y más intensiva en trabajo manual y en la estrategia de producción de modelos de datos porque de ella depende la bondad de los resultados de los modelos predictivos. La deserción escolar la hemos determinado a partir de la cohorte del historial académico de 5 años escolares (del 2009 al 2014). Se verificó el último registro de cada estudiante en la base de datos del sistema de gestión de centros. Si aparece para el año académico 2013-2014 posee una condición de no desertor en la cohorte, de lo contrario al no aparecer más registros en su historial académico ha desertado del sistema educativo nacional.

En ese sentido, además de las variables demográficas y socioeconómicas del estudiante, se derivaron un conjunto de factores determinantes explicativos de la deserción escolar, a partir del historial de datos del estudiante y de su último registro. El diagrama de estado realizado previo al proceso de preparación de datos nos arrojó ese conjunto de factores, que consisten en:

- 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 períodos o años escolares ha durado en el sistema.
- Último 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).

La determinación de los factores de vulnerabilidad del estudiante tiene dos fuentes principales:

1. El registro en la base de datos del sistema de gestión de centros de la condición de pobreza del estudiante, provisto por el SIUBEN y usado por el MINERD y PROSOLI para la emisión de la tarjeta Solidaridad para en subsidio escolar (ILAE y BEEP). Estos estudiantes constituyen el 18% de la población de los 72 centros de Los Alcarrizos, tomados como referencia de estudio.
2. La base de datos de centros nacionales provista por el estudio EDUCA_UE que nos da un índice de vulnerabilidad del centro a partir del índice de pobreza de los hogares de su entorno geográfico. Esta información fue proporcionada a dicho estudio por el SIUBEN lo que permitió determinar el índice de vulnerabilidad del centro. Esta información ha de servir como variable explicativa de deserción escolar para el modelo predictivo del estudiante como para el análisis de datos de los niveles de prospección y riesgo de deserción de los centros educativos.

Todas estas informaciones nos han llevado a una resultante que es el archivo de **Historial_ Académico-Acumulado_2009_2014**, que contiene los datos de los 82,000 estudiantes con sus variables explicativas, como preparación de datos de entrada a la elaboración del modelo predictivo de deserción escolar.

En adición a esta preparación del archivo fundamental, hemos elaborado un análisis descriptivo de las informaciones para determinar los promedios y tasas de deserción y retención escolar para los períodos de estudio. Estos cuadros y gráficos nos proveen de informaciones claves que servirán de guía para los resultados predictivos de deserción de estudiantes y de riesgos de centros.

La determinación de una tasa promedio de deserción de 9.09% para los períodos del 2009-2014 en el Distrito educativo de Los Alcarrizos nos llevan a la idea de que el modelo predictivo debe servir como herramienta de medición del riesgo de deserción escolar para tomar medidas preventivas durante el año escolar con intervenciones del MINERD que tiendan a evitar la deserción del estudiante el siguiente año escolar y así reducir la tasa promedio establecida. De igual manera conociendo los riesgos de deserción por centro educativo pueden elaborarse estrategias de centro para disminuir la deserción y aumentar la tasa de retención escolar.

Luego de preparar el modelo de datos de entrada, hemos procedido a desarrollar el modelo predictivo con las herramientas de SPSS Modeler y RapidMiner. Los resultados de este proceso son:

- **Los modelos predictivos de deserción** se inscriben como técnicas de minería de datos supervisadas no-paramétricas, donde la variable respuesta u objetivo es binaria (deserción o no deserción) y cuyo valor está dado en función de un conjunto de variables observables, denominadas explicativas o "predictoras".
- La variable respuesta, objetivo o dependiente es la condición de deserción (1) o no deserción cero (0), observadas durante el período de cohorte. Como se nota en el diagrama anterior se selecciona un período de colección de datos o cohorte (historial de varios años académicos) que sirven de entrenamiento y prueba del modelo (estamos usando 2009-2014 como período de cohorte, 70% *dataset* de entrenamiento y 30% *data set* para prueba del modelo).

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 el de prueba, en una tabla de contingencia con pruebas estadísticas de significación (F1 y Chi Cuadrado). Luego se procede a probar el modelo simulando un período académico próximo.

- Se usa el último año académico 2013-2014 como período de verificación del modelo simulando la predicción de este año basado en el entrenamiento de la cohorte seleccionada.
- 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 de cuatro componentes básicos:

    - Análisis de datos variables explicativas y de algoritmo
    - Entrenamiento y evaluación del modelo
    - Aplicación del modelo para predecir la probabilidad de deserción de los estudiantes.
    - Análisis de desviación del modelo

- Los factores se usan como variables explicativas para la estimación de la deserción escolar (0 no deserción y 1 deserción y su probabilidad de ocurrencia). Es decir, **[Deserción, Riesgo] = F(factor1, factor2, …,factorn)**. Se determinan cuáles factores explicativos entran en el modelo de acuerdo con su nivel de significación, eliminación de factores auto-correlacionados, basados en índices de correlación y pruebas de hipótesis estadísticas realizadas mediante el nodo de Selección de Características de SPSS Modeler (Features Selecction Node). El nodo Selección de características filtra los campos de entrada para su eliminación en función de un conjunto de criterios (como el porcentaje de valores perdidos). A continuación, se clasifica el grado de importancia del resto de entradas de acuerdo con un objetivo específico.
- Se han seleccionado los algoritmos de análisis predictivo supervisado de acuerdo a un objetivo de precisión con el set de entrenamiento y de prueba, denominado árbol de decisión. El nodo Clasificador automático crea y compara varios modelos diferentes para obtener resultados binarios (sí o no, abandono o no de estudiantes, etc.), lo que le permite seleccionar el mejor enfoque para un análisis determinado, de acuerdo al criterio de exactitud *(accuracy)*, que indica cuántos *unos* fueron aceptados como uno y cuántos ceros fueron aceptados como cero del total de registros del modelo de entrada en el proceso de aprendizaje.
- El problema de la deserción escolar se enfocó como un problema de clasificación, donde se necesita un modelo que sea capaz de clasificar a los estudiantes en dos clases: desertores y no desertores, en función de las estadísticas escolares y otras informaciones socio-económicas suministradas. Debido a las características de las variables disponibles y al conjunto de datos de la cohorte 2009-2014 del historial académico del estudiante perteneciente a uno de los 72 centros del Distrito educativo de Los Alcarrizos, el algoritmo más adecuado para hacer la clasificación de los estudiantes fue el de Árbol de Decisión CHAID, seleccionado en el punto 4.4.
- El módulo de pre-procesamiento es el primer módulo del modelo de predicción de deserción escolar. El objetivo de este módulo es leer y preparar la data para la posterior etapa de entrenamiento y prueba del modelo.
- El módulo de entrenamiento se usan los insumos del módulo de pre-procesamiento y se construye el modelo que se usará para la predicción de la deserción escolar de los estudiantes.
- 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.

- 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* (período escolar 2009-2014 de la cohorte), cuidando que no se produzca el fenómeno denominado "overfitting", que consiste en que el modelo aprende solo los ejemplos del *training set*, pero no tiene la capacidad de clasificar correctamente ejemplos nuevos. En la tabla 6.3 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.
- Se hizo el mismo análisis con la base de datos del *test set*. En la tabla 6.4 se resumen los resultados. El *test set* contiene 54,359 casos, es decir el 30% de la base de datos completa. 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 clasificó correctamente. Estos son resultados muy satisfactorios considerando que el *test set* contiene datos de estudiantes que el modelo no usó durante el entrenamiento, por lo que puede deducirse que el modelo no sufre de *Overfitting*.
- Debido al fenómeno que contribuye a sesgar los resultados de la totalidad de 1s predichos (*true positive*) con relación a los falsos negativos (ceros predichos como unos), lo que no puede ser detectado mediante la medida de *Accuracy*, se ha requerido 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.
- Para los 72 centros educativos de Los Alcarrizos y basado en las predicciones del modelo, se muestran la distribución del índice de deserción de estudiantes en los 5 períodos de la cohorte (2009-2014). El índice de deserción es un promedio de la proporción de deserción de cada centro entre el total de estudiantes matriculados en los períodos mencionados. Los intervalos de clase seleccionados facilitan la representación de este índice y la frecuencia de centros que la poseen. Como se puede notar, la mayor frecuencia de deserción de los centros se produce entre 0.03 y 0.12 acumulando un 70.8% (51 centros del total de 72). El 29.2% (21 centros) restante está entre 0.12 y 0.32.
- Para el análisis predictivo del modelo se seleccionaron solo los estudiantes registrados en el inicio de la cohorte, es decir, el año 2009-2010. Sobre la base de estos estudiantes se realizó la predicción de los estudiantes que desertarían en el año 2013-2014, según lo descrito en la sección 5.6, los que ascienden a 5,113 para un porcentaje de 25.31%. Esta cifra es el doble de lo esperado debido a que el proceso predictivo incluyó los falsos negativos (no desertores clasificados como desertores) según se ha explicado en este estudio.
- Para medir los factores de vulnerabilidad se tomó como referencia la deserción registrada en el año escolar 2013-2014. Existe una correlación positiva de 0.68 del índice de estudiantes con tarjeta Prosoli y el nivel de deserción dentro de los 72 centros

educativos, medido por la correlación de Pearson. El índice de estudiantes con tarjeta Prosoli representa la proporción de estudiantes en condición de pobreza que inciden en el centro educativo.

- La correlación del índice de pobreza del centro educativo también es positiva con relación a la deserción del centro educativo, 0.23. La proporción de deserción de los estudiantes masculinos en el período 2013-2014 superó a la de estudiantes del género femenino (14% masculino y 12% femenino).
- El modelo de deserción producido en RapidMiner ha sido entrenado con la cohorte 2013-2015, lo que indica que está preparado para realizar la predicción de deserción del año escolar 2015-2016. Esta información aún no está disponible debido al proceso de implementación del sistema SIGERD del MINERD. Cuando se tenga el registro de la condición académica 2014-2015 podrá realizarse la predicción de deserción. Cuando se registre la matriculación para el 2015-20 de los estudiantes se podrá realizar la prueba de desviación del modelo.

7.2 Próximos pasos

Este proyecto se ha realizado en torno a una población escolar distrital como estudio piloto, dicha población se conforma de 72 escuelas públicas del nivel Básico y Medio de Los Alcarrizos, y es considerada de importancia para los propósitos del Ministerio de Educación e IDEICE según 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 piloto y de efecto demostración, acompañado de acciones específicas sobre los centros del 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 considera el recién creado, aún en fase experimental, del Índice de Vulnerabilidad a nivel de hogar, auspiciado por PNUD-SIUBEN, sobre todo el 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 sea considerado un conglomerado poblacional particular, por lo que no debe realizarse alguna 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 al adoptarse de estas técnicas de aprendizaje automático no paramétrico, no se presume ningún tipo de comportamiento de las variables envueltas ni de su distribución de probabilidad.

Para llevar a cabo esta escala de proyecto se ha de requerir un plan amplio de proyecto que incluiría la dotación de recursos de servidores, bases de datos y herramientas de software para la gestión de recursos de información y modelación (SQL SerVer, SPSS Modeler, SPSS Statistics,

RapidMiner, etc.). El equipo de proyecto requerido debe ser conformado por personal de las áreas de sistemas de información, estadísticas y las dependencias funcionales y docentes del MINERD (Educación Básica, Educación Media, calidad educativa y otras.

La creación de un mecanismo de transferencia de datos hacia el sistema de gestión de centros SIGERD es recomendable para que la predicción de deserción sea integrada al tablero de control de cada centro, para que sirva eficientemente a los planes de intervención para la prevención de deserción dentro de los planes de cada centro educativo del sistema de República Dominicana.

8. Bibliografía Referenciada

- Barrientos-Heredia, A. (2012). Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos. Barranquilla: Universidad Simón Bolívar.
- Centro de Estudios del Ministerio de Educación de Chile. (2013). Medición de la deserción escolar en Chile (Serie Evidencias). Santiago de Chile: el autor.
- Han-Kamber-Pei; Data Mining – Concepts and Techniques (3th Ediction); Elseiver Inc. 2012.
- Rockach L. & Maimon, O. Data Mining with Decision Trees, Theory and Applications. World Scientific Publishing Co. 2008.
- Hofmann, M. & Klinkenberg, R. (2013). RapidMiner, Data Mining Use cases and Business Analytics Applications. New York: CRC Press.
- International Busibess Machines. (2003). Modelación Avanzada con IBM SPSS Modeler. New York: IBM.
- International Busibess Machines. (2008). Introducción al SPSS Modeler y Data Mining. New York: el IBM.
- Ministerio de Educación. (2014). Boletín Estadístico 2013-2014. Santo Domingo: Ministerio de Educación.
- Iniciativa Dominicana por una Educación de Calidad. (2014). Informe anual de seguimiento y monitoreo 2014. Santo Domingo: IDEC.
- Ministerio de Educación de República Dominicana. (2008). Plan Decenal de Educación 2008-2018. Santo Domingo: el autor.
- Oficina Nacional de Estadísticas. (2014). Boletín de la ONE: Panorama Estadístico. Santo Domingo: ONE.
- Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2009). Education Indicators. Technical Guidelines. París: UNESCO Institute of Statistics.
- República Dominicana. (1997). Ley General de Educación No. 66-97. Santo Domingo: Congreso Nacional.
- Secretaría de Estado de Educación. (2006). Modelo de gestión de la calidad para los centros educativos. Santo Domingo: SEE.
- Theodoris, S. & Koutroumbas, K. (2006). Pattern Recognition (3th Ed.). London: Academic Press.
- University of Minnesota. (2010). Essential Tools-Increasing Rates of School Completion: Moving From Policy and Research to Practice-A Manual for Policymakers, Administrators, and Educators. Minesota, USA: College of Education and Human Development.

- Valero, 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. Ciudad de Puebla, México: Universidad Tecnológica de Izúcar de Matamoros.
- Yegny Amaya, Edwin Barrientos, Diana Heredia Vizcaíno. (2014). Modelo Predictivo de Deserción Escolar. Bogota, Colombia: Universidad Simón Bolívar y Universidad Francisco de Paula Santander.
- Ministerio de Educación de Costa Rica. (Octubre, 2004). CUARTO INFORME INDICADORES DE EQUIDAD DE GÉNERO: PROMOCIÓN, REPITENCIA, DESERCIÓN, ALFABETIZACIÓN Y COBERTURA DEL SISTEMAEDUCATIVO COSTARRICENSE. San Jose, Costa Rica: Ministerio de Educación.
- Iniciativa Dominicana por una Educación de Calidad (IDEC). (2013). Documento final. Resultados de las mesas y marco de acción y monitoreo. Santo Domingo, República Dominicana: IDEC.
- Poder Ejecutivo República Dominicana. (2014). Decreto No. 228-13 Pacto Educativo. Santo Domingo: Presidencia de la Republica Dominicana.
- Sergio Martinic. (2015). Uso del tiempo en centros educativos de jornada extendida y media jornada en República Dominicana-Informe final. Santo Domingo: PAPSE II.
- Instituto Nacional de Evaluación de la Educación INEE. (2015). Tasa de deserción total (2008/2009)- Panorama Educativo de México. Ciudad de México: INEE.

ideice

Instituto Dominicano de Evaluación e Investigación de la Calidad Educativa

www.ideice.gob.do

ISBN 978-9945-499-37-7