Diseño de un Modelo predictivo aplicando Minería de Datos para identificar causas de Deserción Estudiantil Universitaria

Predictive Model Design applying Data Mining to identify causes of Dropout in University Students

Maya Pérez, Petra Norma[*]
Aguilar C, Jorge R.[†]
Zamora R., Rosa A.[‡]
Barron A., J. Miguel[§]

Recibido el 13 de Octubre de 2018. Aceptado el 31 de Diciembre de 2018.

Resumen

En este artículo se presenta el diseño de un Modelo de predicción desarrollado con técnicas y métodos de Minería de Datos (DM) basadas en reglas de clasificación y selección de atributos. El objetivo es identificar patrones relacionados con los aspectos de mayor influencia en la deserción estudiantil de una Institución de Educación Superior (IES) del estado de México. La investigación inicia con un análisis exploratorio, correlacional y explicativo, guiado por la estrategia de cebolla y el ciclo de vida de un proyecto de DM. Posteriormente, se realiza el entrenamiento del modelo con una muestra de 170 estudiantes, en el que se aplican diferentes algoritmos de clasificación (JRIP, OneR, ZeroR, J48, REPTree) y selección (CfsSubsetEval y BestFirst). Los mejores resultados se obtienen en la identificación de las causas que impactan en la deserción y reprobación escolar en un 66%, con respecto a las causas que reporta la IES y un margen de error del 47% del algoritmo J48 con factor de confianza del 0.60, 0.75 y 1.0. Con el método implementado se logra una aproximación satisfactoria para abordar el fenómeno de la deserción o la reprobación en las Universidades.

[*] Profesora Investigadora Universidad Tecnológica del Valle de Toluca. Doctorado en Universidad Popular Autónoma de Puebla. petranorma.maya@upaep.edu.mx, mpn10_utvt@yahoo.com.mx
[†] Profesor Investigador Universidad Popular Autónoma de Puebla. jorge.aguilar@upaep.mx
[‡] Asesora Externa Universidad Popular Autónoma de Puebla. rosaangelica.zamora@upaep.mx
[§] Profesor Investigador de Universidad Tecnológica del Valle de Santiago, Guanajuato. mbarrona@utsoe.edu.mx

Palabras Clave: Minería de Datos, Deserción Escolar, Modelo de predicción, error de clasificación.

Abstract

This paper presents a prediction model developed with techniques and methods Data Mining (DM) with classification rules and selection of attributes. The objective is to identify patterns related to the aspects of greater influence in the for school dropout in Higher Education Institutions (IES) in the Mexico State. The research begins with an exploratory, correlational and explanatory analysis, guided by the strategy and the life cycle of a DM project. Subsequently, the training of the model is carried out with a sample of 170 students, in which different classification algorithms are applied (JRIP, OneR, ZeroR, J48, REPTree) and selection (CfsSubsetEval and BestFirst). The best results are obtained in the identification of the causes that impact school dropout and failure by 66%, with respect to the causes reported by the IES and error margin of 47% in J48 algorithm with a confidence factor of 0.60, 0.75 and 1.0. The implemented method obtains, a satisfactory approach is achieved to address the phenomenon of dropout or failure in the Universities.

Keywords: Data Mining, School Dropout, Predictive Model, Classification Error.

1 Introducción

En los últimos diez años, en el campo de investigación en minería de datos por sus siglas en inglés Data Mining (DM), se ha incrementado el número de aplicaciones en diferentes disciplinas y áreas del conocimiento, encontrándose entre las más citadas: la predicción de ventas de una organización, la predicción de pagos completos de créditos financieros por parte de un cliente, los diagnósticos sobre medicina, robótica, mecatrónica, sistemas de toma de decisiones y predicción de deserción escolar o rendimiento académico, entre otros. Calders y Pechenizkiy (2011), citado en Maya, Cisneros, Zamora y Barron (2016) mencionan que han sido pocos los estudios en entornos educativos aplicados en fenómenos de deserción o abandono escolar, rendimiento académico, estadísticos, análisis de textos académicos, entre otros. Después de la revisión literaria en la temática de DM en el sector educativo, surge la idea de diseñar un modelo predictivo denominado "PredATIS" en su primera versión, en la que se aplican técnicas

de clasificación y selección de DM, el método científico deductivo y la estrategia de cebolla; con el propósito de analizar e identificar los factores o aspectos de mayor influencia, en los estudiantes que no logran culminar su carrera profesional en una Institución de Educación Superior (IES) del Estado de México y que representa un fenómeno de interés que se debe abordar.

De esta manera, el trabajo presenta una revisión del marco teórico, la metodología y el desarrollo del modelo, así como los resultados obtenidos de una primera aproximación con datos reales a través de pruebas y entrenamientos realizados en el software (SW) de WEKA, siendo una ventaja competitiva en análisis predictivos, con respecto a sistemas tradicionales que resultan tardíos; por lo que esta investigación presenta un modelo desarrollado con técnicas y algoritmos de DM como una herramienta tecnológica que aporte o genere información útil y oportuna a la IES seleccionada, que coadyuvará para aplicar estrategias de retención escolar.

2 Marco Teórico

2.1 Estadísticos Educativos

La Secretaria de Educación Pública (SEP) y el Sistema de Información y Gestión Educativa (SIGED) define que un indicador educativo es un "instrumento que nos permite medir y conocer la tendencia o desviación de las acciones educativas, que determinan el éxito o grado de avance de otras metas del sistema educativo y registrando mediciones sistemáticas en cada inicio y fin de cursos" (SEP-SIGED, 2014, Sección de Estadística - Indicadores, párr. 1). El reporte de deserción escolar en México por AtlantiaSearch (2014) menciona que los tres indicadores más representativos para evaluar la eficiencia del sistema educativo son la deserción, la reprobación y la eficiencia terminal, que se encuentran relacionados al momento de analizar las posibles causas

y consecuencias del desempeño de los estudiantes y el rol del propio sistema educativo. AtlantiaSearch (2014) describe que la deserción es el número o porcentaje de estudiantes que abandonan las actividades escolares antes de terminar algún grado o nivel educativo. En tanto que la reprobación es la proporción de estudiantes que finalizaron el ciclo escolar pero que no cumplieron con los requisitos para ser promovidos del grado o nivel educativo que finaliza y la eficiencia terminal es la relación porcentual que resulta de dividir el número de egresados de un nivel educativo determinado, entre el número de estudiantes de nuevo ingreso que entraron al primer grado de ese nivel educativo años antes.

Actualmente el fenómeno de deserción y reprobación escolar en los diferentes niveles educativos básico, medio y superior, es un tema que se está abordando en muchos países con el objeto de determinar los múltiples factores que influyen en él (Álvarez, 2009), además del rendimiento académico de los estudiantes (Araque, Roldán y Salguero, 2009). Valero, Vargas y García (2010) así como las estadísticas reportadas por el SIGED, el Sistema Nacional de Información Estadística Educativa SNIEE y la Organización para la Cooperación y el Desarrollo Económico (OCDE) (SEP-SIGED, 2014 y SEP-SNIEE, 2014) sobre altos indicadores de deserción y reprobación, bajos índices de eficiencia terminal en todos los niveles educativos, mencionan que tales indicadores se encuentran entre los problemas más complejos y frecuentes en las IES de México. En este sentido, las IES se han preocupado por disminuir estos índices a través de programas de tutorías, asesorías especializadas, talleres, congresos, entre otros, con la intención de involucrar directamente y aumentar el compromiso del estudiante; sin embargo esto, no es suficiente y se repite en cada ciclo escolar como lo mencionan, principalmente en los primeros semestres de estudios universitarios, señalando que este elevado fracaso se ve impactado en la

mayoría de los casos por aspectos sociales, económicos y humanos que la sociedad no debe ignorar (Más-Estellés, Alcover, Dapena, Valderruten, Satorre, Llopis y otros (2009).

Márquez, Romero y Ventura (2013) resumen que los factores que influyen en el abandono o fracaso escolar en diferentes niveles educativos por el estudiante son: los aspectos personales, académicos, físicos(enfermedades, discapacidad, problemas visuales o auditivos), económicos, familiares, sociales, institucionales, pedagógicas, laborales, adicciones (exceso de tiempo dedicado a la televisión, videojuegos o juegos en computadora, acceso a redes sociales, mensajería instantánea y el consumo de alcohol y/o drogas), interrupción de estudios, entre otros. Siendo éstos los que se analizan en la investigación y se describen en la tabla 1.

Treviño, Ibarra, Castán, Laria, y Guzmán (2013) mencionan que estos factores y variables tienen una gran influencia en el rendimiento académico del estudiante de las IES, que requiere de un análisis multifactorial y se apoya en áreas de descubrimiento del conocimiento como DM e Inteligencia Artificial (IA), entre otras.

2.2 El papel de las tecnologías de la información en la inteligencia organizacional

La inteligencia organizacional, se relaciona con la capacidad de reunir, analizar y diseminar información interna y externa a las IES, que no sería posible sin el uso de las Tecnologías de la Información y Comunicación (TIC), ya que se requiere de las bases de datos, técnicas y herramientas para el manejo y el análisis del gran volumen de datos disponibles. “De igual forma contar con acceso libre a internet, que permita el acceso a esa enorme fuente de información que posibilite la realización de búsquedas eficaces y establecer una comunicación interpersonal y grupal” dentro de las IES (Cendejas, 2014:47). En la última década, ha existido grandes avances de investigación centrada en el entorno educativo donde se aplican cuatro áreas de las TIC,

consideradas como la base para la explotación de la información y pieza fundamental en la inteligencia organizacional, entre las que se agrupan: Minería de datos (Data Mining - DM), Minería Web (Web Mining- WM), Minería textual (Text Mining - TM) y Minería de datos Educacional (Educational Data Mining EDM).

2.3 Minería de Datos (DM)

KDD por sus siglas en inglés (Knowledge Discovery from Databases) es un proceso no trivial para identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos. La diferencia es que busca el descubrimiento del conocimiento sin una hipótesis predefinida o preconcebida con respecto a sistemas tradicionales de explotación de datos basados en la existencia de una hipótesis o modelos previos. El proceso KDD define una metodología que provee una representación completa del ciclo de vida de un proyecto DM (Formia, 2013). El proceso de KDD se divide en cinco fases principales como lo describen Hernández, Ramírez y Ferri (2014), citados en Galán (2015).

- Integración y recopilación de datos: aquí se decide de dónde se van a obtener los datos que se utilizarán más adelante, es decir qué fuentes de información resultan ser útiles. Después se transforman todos los datos a un formato en común, debido a que pueden provenir de fuentes heterogéneas, generalmente esto se consigue usando un almacén de datos.
- Selección, limpieza y transformación: los datos recopilados en el almacén pueden contener errores en sus valores, o incluso puede que a algunos de estos les falte algún valor, que estos sean erróneos. En esta fase se trata de corregir o incluso eliminar estos datos y se decide qué hacer con aquellos datos que estén incompletos. También se realiza una selección de aquellos datos que son relevantes para el proceso de extracción de conocimiento que se desea realizar.

- Minería de datos: es la fase principal que se trata en esta investigación, en la que, se debe decidir cuál es la tarea (agrupar, clasificar, etc.) que se va a realizar, se elige el método y algoritmo por aplicar.
- Evaluación e interpretación: en la fase de minería de datos debe dar unos resultados, por ejemplo unos patrones observados en los datos. Aquí se evalúa e interpreta estos patrones, con el fin de poder entender el resultado obtenido. (Galán, 2015:10)
- Difusión y uso: es la última fase del proceso de KDD tiene como objetivo utilizar el nuevo conocimiento adquirido y hacer que dicho conocimiento sea empleado por todos los usuarios posibles. (Galán, 2015:11)

La DM es la etapa de descubrimiento en el proceso de KDD: siendo un paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos pre procesados; aunque se suelen usar indistintamente los términos KDD y Minería de Datos (Hernández et al., 2014).

Por su parte Galán (2015) menciona que la DM es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo, se usan diferentes aplicaciones software en cada etapa, que pueden ser estadísticas, de visualización de datos o IA principalmente. En la actualidad existen aplicaciones o herramientas comerciales de DM muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto, sin embargo, casi siempre se complementan con otras herramientas de SW. Además el autor, señala que una vez que se han recopilado los datos necesarios y están bien organizados y limpios, se continúa con el proceso de DM, cuyo objetivo es descubrir patrones qué deben ser válidos, novedosos y por supuesto, comprensibles. Para ello existen diversas tareas y métodos o técnicas que permiten resolver.

- Las tareas de DM se definen como un tipo de problema a ser resuelto por algoritmos de DM; implicando en cada tarea sus propios requisitos e información que se obtiene empleando una tarea en concreto que puede ser muy distinta a la obtenida, si se emplea otra tarea diferente. Se identifican dos tipos de tareas de DM, predictivas y descriptivas (Galán, 2015):
    a) Predictivas: el objetivo es estimar valores futuros o desconocidos de algunas variables de interés a partir de otras variables independientes (variables predictivas), que incluyen la clasificación o discriminación (en estadística), la clasificación suave, la estimación de probabilidad de clasificación, la categorización y la regresión.
    b) Descriptivas: identifican patrones en los datos que los explican o resumen. En este tipo de tareas se incluyen: agrupamiento (clustering), correlaciones y factorizaciones y reglas de asociación.
- Los métodos permiten resolver cualquiera de las tareas anteriores, donde requieren aplicar diferentes técnicas, algoritmos y no sólo uno sino la combinación de varios de ellos; puntualizando que un método puede servir para resolver más de una tarea. Galán (2015:19) cita algunas Técnicas Bayesianas que incluyen el clasificador bayesiano naive, entre otros, basadas en árboles de decisión y sistemas de aprendizaje de reglas (ID3/C4.5 o el CART, M5P, etc.) y redes neuronales artificiales (RNA) que utilizan los algoritmos más comunes: retropropagación y redes neuronales de función de base radial (FBR), entre otras.

2.4 Técnicas y Método de Minería de Datos

Rodríguez y Díaz (2009: 77); Girones, Casas, Minguillón y Cauhuelas (2017) referencian que las técnicas y métodos de DM, aplican algoritmos de aprendizaje supervisados. Los No Supervisados no toman como base una variable dependiente, endógena o variable a predecir, en

la que integra el Clustering o agrupación (EM, Simple k-Means, Cobweb, etc.) y Reglas de asociación (A priori, FilteredAssociator y FPGrowth). En tanto que los Supervisados tienen una variable dependiente, endógena o variable a predecir, en las que se agrupan:

a) Técnicas de Clasificación: árboles de decisión, tabla de decisión, inducción de reglas, bayesianas, redes neuronales, lógica difusa, técnicas genéticas. En los árboles de decisión la variable a predecir o dependiente es categórica y por los resultados obtenidos pueden representar una regresión logística (Witten, Frank, y Hall, 2011, citado en Herrero y Molina, 2012).

b) Técnicas de Predicción: árbol de regresión (predicción), regresión, estimador de núcleos. Un árbol de regresión es análogo a una regresión lineal donde la variable a predecir o dependiente es numérica (discreta o continua) (Witten, et al., 2011, citado en Herrero y Molina, 2012).

> Un Árbol de decisión, representa el conjunto de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos, son fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos y los valores faltantes. Su principal ventaja es la facilidad de interpretación. (Rodríguez y Díaz, 2009: 77 y Girones et al., 2017).

Actualmente existe una gran variedad de algoritmos de árboles de Decisión (Tree) como: Decisión Stump (árbol de un solo nivel), ID3 y C4.5 (J48), LMT, M5 (M5P), REPTree, etc., así como de reglas de clasificación entre las que se mencionan Tablas de Decisión, JRip, M5Rules, OneR, PART, ZeroR, señalando que estos nombres pueden diferir dependiendo del Software que se utilice para el procesamiento de DM, tal es el caso de WEKA (Waikato University, s.f.), Rapid Miner, SPSS, R, Matlab, Oracle Data Mining, entre otros (Rodríguez y Díaz, 2009 y Machine Learning, s.f.).

3 Metodología

La investigación inicia utilizando el método científico deductivo, tomando como base la estrategia "cebolla" propuesta por Saunders, Lewis y Thornhill, (2009) y un análisis de tipo exploratorio, dado que ha sido poco abordada la aplicación de modelos de DM en fenómenos del ámbito educativo, en relación al desarrollo aplicado en áreas de medicina, negocios, mecatrónica, sistemas de decisiones entre otras (Calders y Pechenizkiy, 2011); además se incluye un estudio correlacional y explicativo en las etapas subsecuentes, como lo describen Hernández, Fernández y Baptista (2010), que identifique los factores y variables que inciden con mayor frecuencia en los indicadores de desempeño (deserción, reprobación y eficiencia terminal). Derivado del análisis exploratorio, correlacional y explicativo, se desarrolla un modelo DM para identificar patrones relacionados con los aspectos de mayor influencia en la deserción estudiantil universitaria, con base en las fases del ciclo de vida de un proyecto DM presentado en la figura 1, que permite identificar los espacios que no han sido cubiertos por los modelos desarrollados en esta temática de DM.

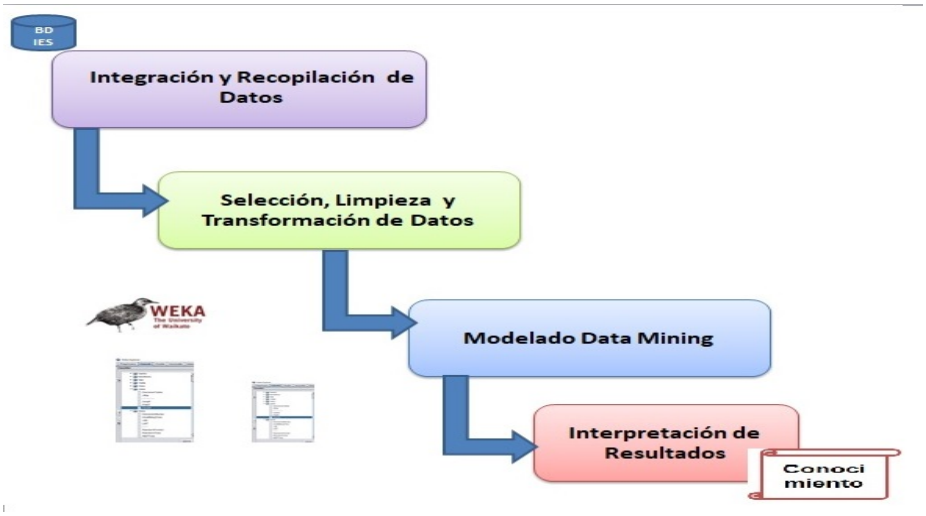

Fig. 1 Metodología de desarrollo del Modelo PredATISv1

Fuente: Elaboración Propia.

La investigación se inicia con la recopilación de datos de fuentes existentes de una Universidad del estado de México, seleccionando una muestra representativa de un programa educativo del tercer período escolar del año 2017, donde se realiza la selección, limpieza y trasformación de datos que son pre procesados antes de ejecutar el entrenamiento con métodos, técnicas y algoritmos de DM en el SW de WEKA (Waikato University, s.f.), para su validación e interpretación de resultados obtenidos del modelo propuesto.

4 Modelo DM de PredATISv1

El diseño del modelo propuesto se guía en el ciclo de vida de un proyecto de DM descrito en la figura 1 y el desarrollo de PredATIS en su primera versión, aplica algoritmos de aprendizaje generados con la selección de mejores atributos, reglas de clasificación y árboles de decisión, utilizando el software WEKA (Waikato University, s.f.). Finalmente se verifica y valida con respecto a las estadísticas reportadas por la IES en estudio.

4.1 Integración y Recopilación de Datos

En esta fase se realizan primero las actividades previas al proceso de DM que se describen a continuación:

a) Recopilación de Datos. Los datos son importados a través de archivos de tipo CSV, TXT y SQL que se generan y exportan del Sistema de Automatización de Información Integral de la IES sobre la muestra seleccionada y son pre procesados en primera instancia, revisando los datos requeridos para análisis de variables o factores que inciden en la deserción estudiantil universitaria.

b) Determinación de variables. De los datos recopilados, se examinaron 106 variables, seleccionando 39 variables representativas de los estudiantes que incluyen los siguientes aspectos, donde se asigna un número de ítem para ser relacionadas con un nombre de variable que se describe en la tabla 1:

- Personales: Edad (1), Género (2), Estado civil (3), Número de Hijos (4), Planes de matrimonio a corto plazo o en su estancia en la IES (5), Relación familiar (6), Planes de su futuro profesional (7), Ocupación de su tiempo libre (8).
- Vocacionales: Titulo Técnico o Bachillerato General (9), La IES es su primera opción (10), El PE es su primera opción (11), Tiene planeado presentar examen en otra IES (12), Ha interrumpido estudios (13), Ha cursado otras carreras en otras IES diferentes al PE que estudia (14).
- Académicos: Promedio de Nivel Medio Superior (15), Nuevo Ingreso al período escolar /Reingreso (16), Materias con dificultad en nivel medio superior (17), Número de asignaturas extraordinarias en períodos escolares anteriores (18), Asignaturas extraordinarias en períodos escolares anteriores(19), Técnicas de Estudio(20), Tiene libros de su perfil de carrera como apoyo para su estudio en casa (21), Tiene computadora en casa como apoyo a sus estudios (22), Tiene acceso a internet en su domicilio (23).
- Socioeconómicos: Ingreso Mensual Familiar (24), Estudiante Trabaja Si/NO - número de horas semanales (25), ¿De quién depende económicamente el estudiante? (26), ¿Quién depende económicamente de usted? (27), Gastos de transportación semanales a la IES (28), Tiempo de transportación (hrs) Casa – IES (29), Tiene Beca (Si/No) (30), Tipo de Beca (31).

- De salud: Estado de salud (32), Problemas de salud (obesidad, delgadez extrema, manchas en la piel, falta de energía, dentadura, visual, auditivo, discapacidad, Ninguna) (33), Padece alguna enfermedad, tratamiento o alergia (Si/No, especifica) (34), Frecuencia con que presenta enfermedades menores como gripe, infecciones estomacales, dolores de cabeza. (especifica enfermedad y frecuencia) (35) y si ha recibido atención psicológica (36).
- Otros: Adicciones de fumar (37), de ingerir bebidas alcohólicas (38) y adicción de drogas (39).

4.2 Selección, Limpieza y Transformación de Datos

En esta fase se preparan los datos del archivo de valores separados por comas (.CSV) importado del sistema integral de datos de la IES en estudio, con el objeto de generar la vista minable y patrones de entrada del entrenamiento para el diseño del Modelo PredATISv1 donde se realizan las siguientes actividades.

- Selección y Limpieza: se revisa el archivo de datos importado CSV, para que no existan datos nulos o erróneos, donde se recopilan, modifican y eliminan los datos innecesarios, que no participarán en el entrenamiento del modelado de PredATISv1.
- Transformación a variables, se preparan los datos con las 39 variables seleccionadas descritas en la primera columna e identificadas con un nombre significativo, en la segunda columna de la tabla 1, que representan las Variables de Entrada, siendo la base del entrenamiento para el modelo, donde se aplican algoritmos de árboles de decisión, reglas de clasificación y selección de atributos en el modelado desarrollado en el SW WEKA (Waikato University, s.f.).

- Pre procesamiento de variables, consiste en categorizar las variables identificadas con valores alfanuméricos descritos en la tercera columna de la tabla 1, que son clasificadas categóricamente y que a su vez representan los patrones de entrenamiento del aprendizaje supervisado en el diseño del modelo propuesto, a través de técnicas y algoritmos de árboles de decisión, reglas de clasificación y selección de atributos.

Tabla 1 Variables de entrada y categorizadas para el Modelado

Descripción de la Variable	Item - Identificación de Variable de Entrada Modelado- Clasificación DM	Valores extraídos y categorizados
Edad	1.Edad	Menor de 18 años, De 18 a 21, De 22 a 25 años, >25 años, No Respondió Pregunta (NRP) y No realizó encuesta (NRE)
Género	2.Género	Femenino, Masculino, Otro, NRP y NRE
Estado civil	3.EstadoCivil	Soltero, Casado, Divorciado, Viudo, Madre/Padre Soltero(a), Unión libre, NRP y NRE
Número de Hijos	4.NumHijos	Ninguno o 0, De 1 a 2, De 3 a 4, más de 4, NRP y NRE
Planes de matrimonio a corto plazo o en su estancia en la IES	5.PlanesMatrimCtoPlzo	No, Si, NRP y NRE
Relación familiar	6.RelacionFamiliar	Buena, Regular, Mala, Con Problemas Padres, Prefiero no responder y NRE
Planes de su futuro profesional	7.PlanesFuturoProf	Trabajar en perfil Empresa o Negocio propio, Graduarse TSU e Ingeniería, Especialidades y/o certificaciones del perfil, seguir estudios posgrado, Otros, Ninguno, NRP y NRE
Ocupación de su tiempo libre	8.Ocupacióntiempolibre	Estudiar o visitar bibliotecas, actividad cultural o Deportiva, Varios entretenimiento (visitar plazas comerciales, cines, parques, Juegos, música, videojuegos, TV, teatro, leer revistas de entretenimiento, etc), Chat o Redes Sociales/Navegar Internet, Platicar con amigos o Hablar por teléfono, otras actividades (domésticas /familiares/ religiosas/trabajo), NRP y NRE
Título Técnico o Bachillerato General	9.TituloTécn/BachGral	Técnico a fin al PE, Técnico no afín a la Carrera, Bachillerato, NRP y NRE
La IES es su primera opción	10.La IES1a.Opc	Si, No aprobé en otra IES, Indeciso, Obligado por padres/Cercanía, NRP y NRE

Descripción de la Variable	Item - Identificación de Variable de Entrada Modelado- Clasificación DM	Valores extraídos y categorizados
El PE es su primera opción	11.El PEPrimOpc	Si, No aprobé en otras IES o no me inscribí al Perfil deseado, Indeciso, Obligado por padres/Cercanía, mi última opción, No sin especificar motivo, NRP y NRE
Tiene planeado presentar examen en otra IES	12.TienePlaneadoExamOtIES	No, Indeciso, Si, NRP y NRE
Ha interrumpido estudios	13.InterrupcionEstudios	No, De 1 a 2 años por motivos personales, De 1 a 2 años por motivos económicos, De 1 a 2 años por motivos académicos o vocacionales, De 1 a 2 años por otros motivos, >2 años por motivos personales/Económicos, >2 años por motivos académicos o vocacionales, >2 años por otros motivos, NRP y NRE
Ha cursado otras carreras en otras IES diferentes al PE que estudia	14.CursadoOtrasIES	No, Si la del perfil, Si perfil diferente, NRP y NRE
Promedio de Nivel Medio Superior	15.PromedioNivelMedioSup	De 9.6-10, De 9-9.5, De 8-8.9, De 7-7.9, De 6-6.9, < 5.9, NRP y NRE.
Nuevo Ingreso al período escolar /Reingreso	16.NuevoIngr/Reingreso	Nuevo Ingreso al ciclo escolar, Reingreso 1 vez, Reingreso 2 veces, Reingreso más de 2 veces, NRP y NRE
Materias con dificultad en nivel medio superior	17.MateriasDificultadMSsup /IES	Ninguna, Matemáticas y Lógica, Áreas Informáticas, Inglés, Matemáticas e Informática (BD/Programación/Redes/SO), Inglés y Asignaturas de Informática (Programación/Redes/SO/BD), otras no afines al perfil, NRP y NRE
Número de Asignaturas extraordinarias en períodos escolares anteriores	18.NAsigExtraPdosAnt	Ninguna, una asignatura, dos asignaturas, >= 3 asignaturas, NRP y NRE.
Asignaturas extraordinarias en períodos escolares anteriores	19.AsignExtraordinariasAnt	Ninguna, Asignaturas no técnicas del perfil (formación sociocultural, expresión oral y escrita. etc.), Desarrollo de Habilidades lógico-matemáticas-Física, Inglés, Asignaturas Técnicas del perfil (Programación, BD, Redes, Soporte Técnico, SO, etc.), Inglés-Técnicas Perfil, Ingles-Matemáticas, NRP y NRE

Descripción de la Variable	Item - Identificación de Variable de Entrada Modelado- Clasificación DM	Valores extraídos y categorizados
Técnicas de estudio	20.TecnicasEstudio	Asesorías especiales, video tutoriales, páginas web/Internet, manuales/Libros/revistas académicas, Asesoría/Video Tutoriales/Web, Varias técnicas de estudio posterior, Repaso Notas/Prácticas Clases, Ninguna solo las clases, NRP y NRE
Tiene libros de su perfil de carrera como apoyo para su estudio en casa	21.TieneLibrosPerfilCasa	Si, Ninguno/No, NRP y NRE
Tiene computadora en casa como apoyo a sus estudios	22.TieneCcomputCasa	Si, No, NRP y NRE
Tiene acceso a internet en su domicilio	23.TieneInternetCasa	Si, No, NRP y NRE
Ingreso mensual familiar	24.IngreMensualFamiliar	> $6000, De 4001 a 6000, De 3001 a 4000, De $2000 a 3000, < $2000, NRP y NRE
Alumno trabaja Si/NO, Número de horas semanales	25.AlTtrabajaSi/NO-HrSem	No trabaja, Si < 20 horas (hrs), Si de 20 a 30 hrs, Si de 30 a 40 hrs, Si >40 hrs, NRP y NRE
¿De quién depende económicamente el estudiante?	26.DeQuienDepEcoEst	Padres, Esposa (o), Familiar, Hijos, Otro, Nadie, él mismo solventa gastos, NRP y NRE
¿Quién depende económicamente de usted?	27.DependEconomicos	Nadie, Esposa (o), Hijos, Padres, Familia, Otros, NRP y NRE
Gastos de transportación semanales a la IES	28.GastosTransport-IES	<= $100, $101 a $300, $301 a $500, 501 a 800, > $800, NRP y NRE
Tiempo de transportación (hrs) Casa – IES	29.TiempoTrapCasaIES	<1 hr, 1 a 2 hr, >2 y <3 hrs, >3 hrs, NRP y NRE
Tiene Beca (Si/No)	30.TieneBeca	Si, No, Prefiero no contestar, NRP y NRE
Tipo de Beca	31.TipoBeca	Manutención, Programas Sociales(Prospera, otra), Excelencia, Laptop, otras, No Tiene, NRP y NRE
Estado de salud	32.EdoSalud	Buena, regular, mala, pésima, NRP y NRE

Descripción de la Variable	Item - Identificación de Variable de Entrada Modelado- Clasificación DM	Valores extraídos y categorizados
Problemas de salud (obesidad, delgadez extrema, manchas en la piel, falta de energía, dentadura, visual, auditivo, discapacidad o ninguna	33.ProblemasSalud	Ninguna, dentadura, Obesidad, Delgadez extrema/falta de energía, visual, auditivo, discapacidad, Varios problemas, NRP y NRE
Padece alguna enfermedad, tratamiento o alergia (Si/No, especifica)	34.PadeceEnfermedad	No, Si enfermedad crónica, Si enfermedad con tratamiento, Si otro tipo de enfermedad, Si es Alérgico a algún medicamento, NRP y NRE
Frecuencia con que presenta enfermedades menores como gripe, infecciones estomacales, dolores de cabeza. (especifica enfermedad y frecuencia)	35.FrecPresEnfMenores	Rara vez, Alguna vez al mes, alguna vez a la semana, muy frecuentemente, Otras Enfermedades pocas veces, NRP y NRE
Ha recibido atención psicológica(Si, No, cuanto tiempo)	36.AtencPsicologica	No, Si de 1 mes, Si de 2 meses a menor de 1 año, Si de 1 a 2 años, Si >2 años, Prefiero no contestar y NRE
Adicciones de fumar	37.AdiccFumar	No Fuma, rara vez, Si una vez al mes, Si Algunas veces al mes, Si varias veces a la semana, Si una vez la semana, Si una vez al día, Si varias veces al día, NRP y NRE
Adicciones de ingerir bebidas alcohólicas	38.AdiccBebAlcoh	No, rara vez ingiere, Si una vez al mes, Si Algunas veces al mes, Si varias veces a la semana, Si una vez la semana, Si una vez al día, Si varias veces al día, NRP y NRE
Adicciones a drogas	39.AdiccDrog	No, rara vez, Si una vez al mes, Si Algunas veces al mes, Si varias veces a la semana, Si una vez la semana, Si una vez al día, Si varias veces al día, NRP y NRE

Fuente: Elaboración Propia.

4.3 Desarrollo del Modelo

El modelo PredATISv1 se desarrolló con técnicas de clasificación de árbol de decisión, que permiten investigar las correlaciones existentes entre las variables de mayor impacto en la deserción de los estudiantes; así como de reglas de clasificación y selección de mejores atributos para identificar las variables predictivas del análisis del fenómeno en estudio. El entrenamiento se realizó con el SW open source de WEKA (Waikato University, s.f.), basado en algoritmos de árboles de decisión J48, con variable dependiente categórica y REPTree que acepta datos de entrada y variable dependiente numérica y categórica que es la que se reporta en este trabajo. Adicionalmente se utilizan reglas de clasificación basadas en JRIP y OneR con datos categóricos.

5 Resultados del Modelo

Los resultados obtenidos del diseño del modelo PredATISv1 en la muestra seleccionada de 170 estudiantes del tercer período escolar mayo-agosto'2017 de un PE, en la que se validan los algoritmos aplicados en relación a causas de deserción y reprobación reportadas por la IES.

a) Deserción: nunca se inscribió al cuatrimestre inmediato, nunca se presentó pero si se inscribió, sólo se presentó los primeros 15 días del inicio de cuatrimestre, por cambio a otra universidad, por cambio de carrera en la misma IES, por cambio de domicilio, por incumplimiento de expectativas del modelo educativo, por problemas de vocación del estudiante, por la distancia de la universidad, por problemas de salud, por problemas económicos, por problemas familiares, por embarazo, por problemas laborales, matrimonio, motivos personales, faltas al reglamento (más del 20% de inasistencias), falta de certificado de bachillerato, sin causas conocidas, cambio de especialidad o área de la misma carrera.

b) Reprobación: faltas al reglamento (reprobó más de 2 asignaturas), reprobación en el extraordinario y reprobación en prácticas profesionales.

Después de generar la Vista Minable se realiza el entrenamiento del modelo en el SW WEKA, iniciando con algoritmos de selección de atributos aplicando el método de búsqueda (BestFirst) y atributo evaluador (CfsSubsetEval), en el que se identifican 15 atributos más representativos del total de 39 variables de estudio como se visualiza en la figura 2, donde se representa el nivel de porcentaje calculado para cada variable entre el 10 y el 100% de participación como variable de impacto en la deserción estudiantil.

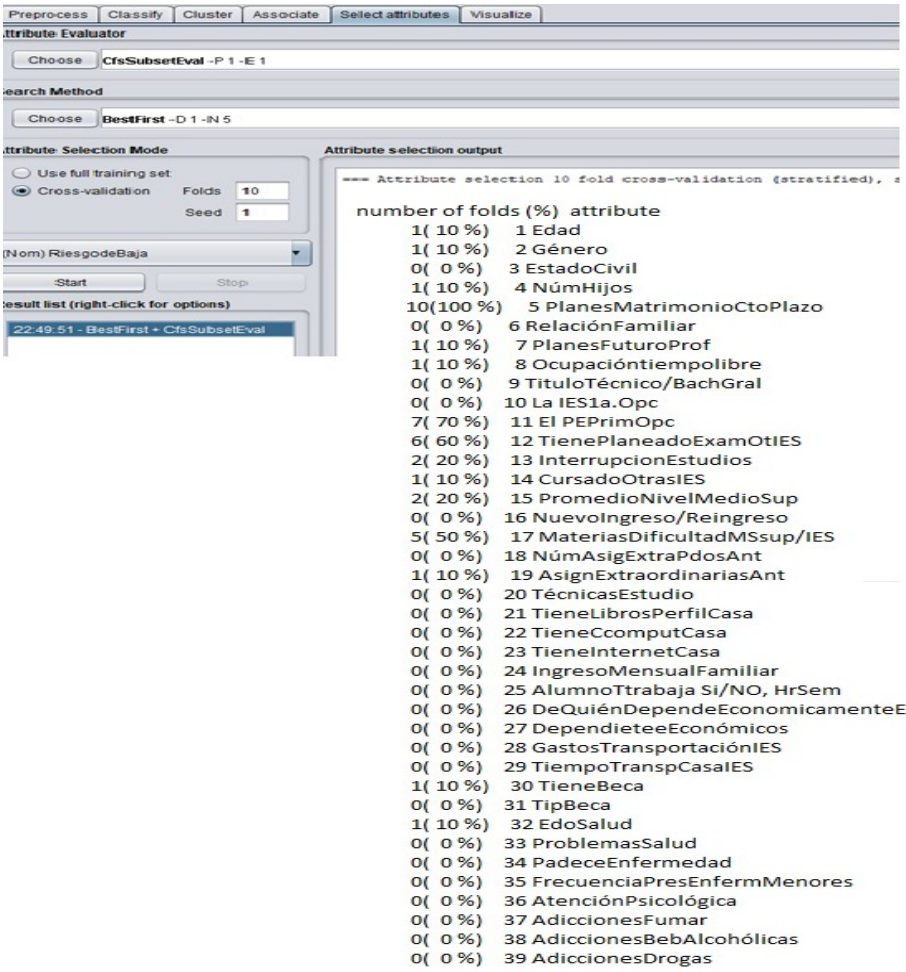

Figura 2 Aplicación de algoritmo de Selección de Atributos.

Fuente: Elaboración Propia.

En la figura 2, se observan 15 variables seleccionadas en el modelo, verificando que 10 de ellas se relacionan con las causas de deserción y reprobación de estudiantes, reportadas por la IES en estudio, representando con ello el 66% de validación de los algoritmos aplicados como se observa en la tabla 2.

Tabla 2 Factores identificados por ModeloPredATISv1

Variable de Entrada Modelado-Clasificación DM	**Causa que se reporta en la IES**
10(100 %) 5 PlanesMatrimonioCtoPlazo	Matrimonio Por embarazo
7(70 %) 11 El PEPrimOpc	Por cambio de carrera en la misma IES Cambio de especialidad de la misma carrera
6(60 %) 12 TienePlaneadoExamOtIES	Por cambio a otra Universidad
5(50 %) 17 MateriasDificultadMSsup/IES	Por problemas de vocación del estudiante
2(20 %) 15 PromedioNivelMedioSup	
2(20 %) 13 InterrupcionEstudios	
1(10 %) 32 EdoSalud	Por problemas de salud
1(10 %) 30 TieneBeca	Por problemas económicos
1(10 %) 19 AsignExtraordinariasAnt	Faltas al reglamento (Reprobó más de 2 materias) Reprobación en extraordinario
1(10 %) 14 CursadoOtrasIES	• Nunca se presentó pero si se inscribió • Sólo se presentó los primeros 15 días del inicio de cuatrimestre •
1(10 %) 8 Ocupacióntiempolibre	
1(10 %) 7 PlanesFuturoProf	Por incumplimiento de expectativas del modelo educativo
1(10 %) 4 NumHijos	Matrimonio
1(10 %) 2 Género	
1(10 %) 1 Edad	

Fuente: Elaboración Propia.

Posteriormente en el desarrollo del modelo de PredATISv1, se aplican los siguientes algoritmos de clasificación.

a) Reglas de clasificación JRIP, One R, ZeroR: el entrenamiento se realizó en sus tres tipos de validación: Porcentaje split 66, Cross-Validation Test y Training set, cuyos resultados se muestran en la tabla 3; donde se observa que la precisión del nivel de error absoluto menor es de 56.6650% con la validación del conjunto total de datos (Training set) y clasificador OneR – B.6, reflejando los resultados lógicos de la diagonal de matriz de confusión con 4 y 140 de clasificación correctamente positiva y negativa respectivamente.

Tabla 3 Resultados obtenidos con Reglas de Clasificación

Tipo Training/Test	Clasificación correcta		Clasificación Incorrecta		Número Total de Instancias	Error Absoluto Relativo
	Instancias	%	Instancias	%		
Aplicando el clasificador JRIP (classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1)						
Percentage split 66	44	75.8621%	14	24.1379%	58	107.3899%
Cross-Validation Test	139	81.7647%	31	18.2353%	170	100.1004%
Training set	147	86.4706%	29	13.5294%	170	86.3629%
Aplicando el clasificador OneR (classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1)						
Percentage split 66	44	75.8621%	14	24.1379%	58	87.5000%
Cross-Validation Test	137	80.5882%	33	19.4118%	170	71.9996%
Training set	144	84.7059%	26	15.2941%	170	56.6650%
Aplicando el clasificador ZeroR (classifiers.rules.ZeroR)						
Percentage split 66	48	82.7586%	10	17.2414%	58	100.0000%
Cross-Validation Test	143	84.1176%	33	15.8824%	170	100.0000%
Training set	143	84.1176%	33	15.8824%	170	100.0000%

Fuente: Elaboración Propia.

b) Árbol de decisión J48 y REPtree. Son técnicas de clasificación que más se han utilizado en estudios de predicción; por ello, se aplican los algoritmos J48 y REPTree y las tres formas de validación; Percentage split 66, Cross-Validation Test y Training set; mostrando en la tabla 4 los resultados obtenidos, en la que se pretende dar un refinamiento al análisis para el modelo predictivo, cuya variable a predecir o endógena es Riesgo de Baja categorizada en: Si o No es probable la deserción del estudiante en una segunda etapa de esta investigación.

Tabla 4 Resultados obtenidos con árbol de decisión

Tipo Training/Test	Clasificación correcta		Clasificación Incorrecta		Número Total de Instancias	Error Absoluto Relativo	Matriz de Confusión
	Instancias	%	Instancias	%			
Árbol decisión J48 con factor 0.25							
Percentage split 66	48	82.7586%	10	17.2414%	58	98.5491%	a b classified as 0 10 | a = Si 0 48 | b = No
Cross-Validation Test	143	84.1176%	27	15.8824%	170	98.8897%	a b classified as 0 27 | a = Si 0 143 | b = No
Training set	143	84.1176%	27	15.8824%	170	98.9970%	a b classified as 0 27 | a = Si 0 143 | b = No
Árbol decisión J48 con factor 0.50							
Percentage split 66	48	82.7586%	10	17.2414%	58	98.5491%	a b classified as 0 10 | a = Si 0 48 | b = No
Cross-Validation Test	141	82.9412%	29	17.0588%	170	86.5505%	a b classified as 2 25 | a = Si 4 139 | b = No
Training set	147	86.4706%	23	13.5294%	170	79.6133%	a b classified as 4 23 | a = Si 0 143 | b = No

Tipo Training/Test	Clasificación correcta		Clasificación Incorrecta		Número Total de	Error Absoluto	Matriz de Confusión
Árbol decisión J48 con factor 0.60, .75 y 1.0							
Percentage split 66	40	68.9655%	18	31.0345%	58	106.4340%	a b classified as 1 9 | a = Si 9 39 | b = No
Cross-Validation Test	127	74.7059%	43	25.2941%	170	91.4633%	a b classified as 5 22 | a = Si 21 122 | b = No
Training set	155	91.1765%	15	8.8235%	170	47.5863%	a b classified as 13 14 | a = Si 1 142 | b = No
Árbol decision REPTree con trees. REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0							
Percentage split 66	48	82.7586%	10	17.2414%	58	98.5991%	a b classified as 0 10 | a = Si 0 48 | b = No
Cross-Validation Test	140	82.3529%	30	17.6471%	170	99.9465%	a b classified as 0 27 | a = Si 3 140 | b = No
Training set	143	84.1176%	27	15.8824%	170	98.9970%	a b classified as 0 27 | a = Si 0 143 | b = No

Fuente: Elaboración Propia.

Derivado de los resultados de la tabla 3 y 4, se observa que la mejor clasificación es, en el algoritmo J48 con un factor de confianza de 0.60, 0.75 y 1.0 del árbol de decisión aplicado al conjunto total de datos de entrenamiento, que se muestra en la tabla 4 y el diagrama en la tabla 5, comparado con REPTree, JRIP, OneR y ZeroR como reglas de clasificación.

Tabla 5 J48 pruned tree

```
PlanesMatrimonioCtoPlazo = Si
|  MateriasDificultadMSsup/IES = Áreas Informáticas: Si (4.0)
|  MateriasDificultadMSsup/IES = Matemáticas y lógica: Si (0.0)
|  MateriasDificultadMSsup/IES = otras no afines al perfil: Si (0.0)
|  MateriasDificultadMSsup/IES = Ninguna: No (3.0)
|  MateriasDificultadMSsup/IES = Inglés y Asignaturas de Informática: Si (0.0)
|  MateriasDificultadMSsup/IES = NRP: Si (0.0)
|  MateriasDificultadMSsup/IES = Matemáticas e Informática: Si (0.0)
|  MateriasDificultadMSsup/IES =  Inglés: Si (0.0)
|  MateriasDificultadMSsup/IES = NRE: Si (0.0)
PlanesMatrimonioCtoPlazo = No
|  El PEPrimOpc = Si: No (30.0)
|  El PEPrimOpc = No sin especificar motivo
|  |  MateriasDificultadMSsup/IES = Áreas Informáticas
|  |  |  Género = Masculino: No (4.0)
|  |  |  Género = Femenino
|  |  |  |  AsignExtraordinariasAnt = Asignaturas Técnicas del perfil: No (3.0/1.0)
|  |  |  |  AsignExtraordinariasAnt = Ninguna: Si (1.0)
|  |  |  |  AsignExtraordinariasAnt = NRP: Si (0.0)
|  |  |  |  AsignExtraordinariasAnt = Inglés: Si (0.0)
|  |  |  |  AsignExtraordinariasAnt = Asignaturas no técnicas del perfil (formación sociocultural,
expresión oral y escrita. etc.): Si (0.0)
|  |  |  |  AsignExtraordinariasAnt = Inglés - Técnicas Perfil: Si (2.0)
|  |  |  |  AsignExtraordinariasAnt = Ingles/Matemáticas: Si (0.0)
|  |  |  |  AsignExtraordinariasAnt = Desarrollo de Habilidades lógico-matemáticas-Física: Si (0.0)
|  |  |  |  AsignExtraordinariasAnt = NRE: Si (0.0)
|  |  |  Género = NRE: No (0.0)
|  |  MateriasDificultadMSsup/IES = Matemáticas y lógica: No (8.0/1.0)
|  |  MateriasDificultadMSsup/IES = otras no afines al perfil: No (0.0)
|  |  MateriasDificultadMSsup/IES = Ninguna
|  |  |  Género = Masculino
|  |  |  |  Ocupacióntiempolibre = Estudiar o visitar bibliotecas: No (6.0/1.0)
|  |  |  |  Ocupacióntiempolibre = actividad cultural o Deportiva: Si (4.0/1.0)
|  |  |  |  Ocupacióntiempolibre = Varios entretenimiento (visitar plazas comerciales, cines, parques,
Juegos, música, videojuegos, TV, teatro, leer revistas de entretenimiento, etc): Si (2.0)
|  |  |  |  Ocupacióntiempolibre = otras actividades(domésticas/familiares/religiosas/trabajo): No (3.0)
|  |  |  |  Ocupacióntiempolibre = NRP: No (0.0)
|  |  |  |  Ocupacióntiempolibre = Platicar con amigos o Hablar por teléfono: No (0.0)
|  |  |  |  Ocupacióntiempolibre = NRE: No (0.0)
|  |  |  |  Ocupacióntiempolibre = Chat o Redes Sociales/Navegar Internet: No (0.0)
|  |  |  Género = Femenino: No (5.0)
|  |  |  Género = NRE: No (0.0)
|  |  MateriasDificultadMSsup/IES = Inglés y Asignaturas de Informática: Si (1.0)
|  |  MateriasDificultadMSsup/IES = NRP: No (0.0)
|  |  MateriasDificultadMSsup/IES = Matemáticas e Informática: No (4.0)
|  |  MateriasDificultadMSsup/IES =  Inglés: No (5.0)
|  |  MateriasDificultadMSsup/IES = NRE: No (0.0)
|  El PEPrimOpc = No aprobé en otras IES o no me inscribe al Perfil deseado: No (57.0/5.0)
|  El PEPrimOpc = Obligado por padres/Cercanía: No (2.0)
|  El PEPrimOpc = Prefiero no responder: No (10.0)
|  El PEPrimOpc = NRE: No (0.0)
PlanesMatrimonioCtoPlazo = Prefiero no contestar: No (3.0)
PlanesMatrimonioCtoPlazo = NRE: No (13.0/6.0)
Number of Leaves  : 44
Size of the tree : 52
```

Fuente: elaboración propia.

6 Discusión y Conclusiones

Los resultados del entrenamiento aplicado en la muestra seleccionada, representan una primera aproximación del modelo predictivo en la versión inicial PredATISv1, desarrollado con técnicas supervisadas, basadas en el algoritmo CfsSubsetEval y BestFirst para la selección de atributos, donde se identificaron 15 variables de entrada en el modelo descritas en tabla 1 y que 10 de ellas, se relacionan con las causas reales que se reportan en la IES descritas en la tabla 2 en la que representa el 66.6% de aproximación identificada y el 33.4%, de no identificada, en este sentido se evidencia que el algoritmo J48 con factor de confianza del 0.60, .75 y 1.0 aplicado, permitió identificar las causas que más influyen en la deserción y reprobación escolar con un margen del 47.59% de error absoluto relativo, el 91.18% de clasificación correcta y 8.83% de incorrecto en el conjunto total (Training set). Así también se observa en la tabla 5, la estructura del árbol de decisión bajo el esquema *SI – entonces* sucede evento. Un estudiante está en riesgo de baja SI tiene planes de matrimonio, las asignaturas de Informática son las que más se le dificultan, ha presentado exámenes extraordinarios de las Asignaturas Técnicas del perfil como Programación, Base de Datos, Redes de cómputo, Soporte Técnico, Sistemas Operativos e Inglés, su mayor tiempo libre lo ocupa en realizar actividad cultural o deportiva o en Varios entretenimiento (visitar plazas comerciales, cines, parques, Juegos, música, videojuegos, TV, teatro, leer revistas de entretenimiento, etc. y en otras actividades (domésticas/familiares/religiosas/trabajo, etc.) y poco tiempo en Estudiar o visitar bibliotecas, Adicional a que el PE No es su primera opción, debido a que No aprobó en otras IES, NO se inscribe al Perfil deseado y es Obligado por padres o por Cercanía a la Universidad, en tanto que el Género del estudiante no es un factor representativo en la deserción o reprobación.

Finalmente, se concluye que los resultados obtenidos en esta investigación validan la precisión de los algoritmos DM aplicados en la muestra de la IES seleccionada, evidenciando que se cumple con el objetivo de este trabajo y se logra una aproximación satisfactoria del modelo PredATISv1 para abordar este fenómeno con respecto al reporte real emitido por la IES, donde se obtuvo una clasificación correcta de más del 90%, que identifica las causas o aspectos que más influyen en la deserción o reprobación estudiantil, como se describe en la tabla 2.

Referencias

Álvarez, L. (2009). Comportamiento de la Deserción y Reprobación en el Colegio de Bachilleres del Estado de Baja California: Caso Plantel Ensenada. *X Congreso Nacional de Investigación Educativa.* Veracruz, Veracruz, México.

Araque, F., Roldán, C., y Salguero, A. (2009). Factors Influencing University Drop Out Rates. *Computers y Education, 53*(3), 563 - 574.

Atlantia, S. (2014). *Reporte Cualitativo. Investigación sobre las causas de la deserción escolar en México*. Recuperado de http://atlantiasearch.com/wp-content/uploads/2014/12/CO-RS-2013-04-Reporte-Deserci%C3%B3n-escolar-en-M%C3%A9xico-1.pdf

Calders, T., y Pechenizkiy, M. (2011). Introduction to the special section on educational data mining. *ACM SIGKDD Explorations Newsletter, 13*(2), 3 - 6. doi:10.1145/2207243.2207245.

Cendejas, J. (2014). Implementación del modelo integral colaborativo (MDSIC) como fuente de innovación para el desarrollo ágil de software en las empresas de la zona centro - occidente en México. *(Tesis Doctoral. Centro Interdisciplinario de Posgrados, Investigación y Consultoría UPAEP).* Puebla, México.

Formia, S. (2013). La deserción en cursos universitarios. Construcción de modelos sobre datos de la UNRN usando técnicas de Extracción de Conocimiento. *(Tesis Magsiter, Facultad de Informática, UNLP)*. Argentina.

Galán, V. (2015). Aplicación de la Metodología CRISP-DM a un proyecto de Minería de Datos en el entorno Universitario. *(Tesis de grado.Escuela Politecnica Superior, Ingeniería en Informática, Universidad Carlos III de Madrid)*. Madrid: Recuperado de http://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf?sequence=1.

Girones J., Casas J., Minguillón J. y Cauhuelas R. (2017). Minería de datos. Modelos y Algoritmos. Recuperado de https://edoc.site/mineria-de-datos-modelos-y-algoritmospdf-pdf-free.html

Hernández, J., Ramírez, M.J., y Ferri, C. (2014). *Introducción a la Minería de Datos. España:* Pearson.

Hernández, R., Fernández, C., y Baptista, M. (2010). *Metodología de la investigación.* México: Mc Graw-Hill.

Herrero, J., y Molina, J. (2012). Técnicas de análisis de datos. *Aplicaciones Prácticas utilizando Microsoft Excel y WEKA*. Madrid, España: Universidad Carlos III.

Machine Learning at Waikato University. (s.f.). Isla Norte, New Zealand. Recuperado de https://www.cs.waikato.ac.nz/ml/weka/book.html

Márquez, C., Romero, C., y Ventura, S. (2013). Predicting School Failure and Dropout by Using Data Mining Techniques. IEEE Journal of Latin-American Learning Technologies, 8(1), 7-14.

Más-Estellés, J., Alcover, R., Dapena, A., Valderruten, A., Satorre, F., Llopis, F., y otros. (2009). Rendimiento Académico de los Estudios de Informática en Algunos Centros Españoles. *XV Jornadas de Enseñanza Universitaria de la Informática*, 5 - 12 Recuperado de: http://upcommons.upc.edu/bitstream/handle/2099/7904/p156.pdf.

Maya, P., Aguilar , C., Zamora , R., y Barron, A. (2016). Propuesta del Diseño de un Modelo Predictivo de alerta temprana en indicadores educativos de Nivel Superior aplicando Minería de Datos. *Revista de Aplicación Científica y Técnica 2016, 2*(4), 29-40.

Rodríguez, Y., y Díaz, A. (2009). Herramientas de Minería de Datos. *Revista Cubana de Ciencias Informáticas, 3*(3 - 4), 73 - 80.

Saunders, Lewis, P., y Thornhill, A. (2009). *Research methods for business students (Fifth ed.).*England: Pearson Education.

SEP-SIGED. (2014). Sistema de Información y Gestión Educativa. Recuperado de: http://www.siged.sep.gob.mx/.

SEP-SNIEE. (2014). Sistema Nacional de Información Estadística Educativa. *Estadísticas Educativas*. Recuperado de http://www.snie.sep.gob.mx/estadisticas_educativas.html.

Treviño, M., Ibarra, S., Castán, J., Laria, J., y Guzmán, J. (2013). A Framework to avoid Scholar Desertion using Artificial Intelligence. Proceedings of the World Congress on Engineering, 3, 1493-1497. Recuperado de http://www.iaeng.org/publication/WCE2013/WCE2013_pp1493-1497.pdf

Valero, S., Vargas, A., y García , M. (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. En M. Prieto, Tecnologías del Aprendizaje . Merida: Kaambal.

Waikato University. (s.f.). Weka 3: Data Mining Software in Java. Isla Norte, New Zealand. Recuperado de https://www.cs.waikato.ac.nz/ml/weka/index.html.

Witten, I., Frank, E., y Hall, M. (2011). Data mining, Practical machine learning tools and techniques. USA: Morgan Kaufmann.