UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA TECNOLÓGICA

CENTRO DE POSGRADO

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TEMA:

"ANÁLISIS DE PATRONES DE DESERCIÓN ESTUDIANTIL DE LA UNIDAD EDUCATIVA LENIN SCHOOL APLICANDO MINERÍA DE DATOS".

AUTOR: ING. GALLARDO CORRALES DIEGO EDUARDO

DIRECTOR: ING. MOLINA BUSTAMANTE MARCO E. PhD.

SANGOLQUI

2017

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**ANÁLISIS DE PATRONES DE DESERCIÓN ESTUDIANTIL DE LA UNIDAD EDUCATIVA LENIN SCHOOL APLICANDO MINERÍA DE DATOS**" realizado por el señor Ing. **DIEGO EDUARDO GALLARDO CORRALES**, ha sido revisado en su totalidad y analizado por el software anti-plagio, el mismo cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, por lo tanto me permito acreditarlo y autorizar al señor Ing. **DIEGO EDUARDO GALLARDO CORRALES** para que lo sustente públicamente.

Sangolquí, 24 de julio del 2017

Ing. MARCO EDUARDO MOLINA BUSTAMANTE PhD.

DIRECTOR

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORÍA DE RESPONSABILIDAD

Yo, **DIEGO EDUARDO GALLARDO CORRALES**, con cédula de identidad N° 0503201717, declaro que este trabajo de titulación "**ANÁLISIS DE PATRONES DE DESERCIÓN ESTUDIANTIL DE LA UNIDAD EDUCATIVA LENIN SCHOOL APLICANDO MINERÍA DE DATOS**" ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citas bibliográficas.

Consecuentemente declaro que este trabajo es de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance de la investigación mencionada.

Sangolquí, 24 de julio del 2017

Ing. Diego Eduardo Gallardo Corrales

C.C 0503201717

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORIZACIÓN

Yo, **DIEGO EDUARDO GALLARDO CORRALES**, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar en la biblioteca Virtual de la institución el presente trabajo de titulación "**ANÁLISIS DE PATRONES DE DESERCIÓN ESTUDIANTIL DE LA UNIDAD EDUCATIVA LENIN SCHOOL APLICANDO MINERÍA DE DATOS**" cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Sangolquí, 24 de julio del 2017

Ing. Diego Eduardo Gallardo Corrales

C.C: 0503201717

DEDICATORIA

Este trabajo está dedicado a mis padres Eduardo y Fanny quienes dejaron a un lado sus metas para ver cumplidas las mías; y, a mi compañera de vida Lady, por tus motivaciones y ayuda para culminar este escrito.

Diego

AGRADECIMIENTO

A Dios por sus bendiciones mostradas todos los días de mi vida.

A mi amada Lady, por tu apoyo y empuje mostrado durante esta etapa.

Al Ing. Marco Molina PhD por su valioso tiempo en la revisión de este trabajo y por su orientación brindada con todo su profesionalismo y conocimiento.

A mis familiares, quienes fueron, son y siempre serán fuente de inspiración y respaldo moral en mis actividades, sepan que este trabajo también es su logro.

Diego

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

El abandono o deserción escolar es una problemática actual que afecta directamente al sistema educativo y a los involucrados en la educación de un individuo como la familia, la institución o el país, por tanto, la detección temprana del abandono escolar es una estrategia clave para las instituciones académicas a todo nivel. Como medida actual para esta detección se encuentra el uso de técnicas de Minería de Datos Educacional (EDM) que permite descubrir patrones en los datos demográficos y académicos de los alumnos; siendo uno de los casos el uso de modelos de clasificación para predecir los alumnos que abandonen la institución; en este estudio se utilizó este marco conceptual con un conjunto de datos de tres años lectivos de la Unidad Educativa Lenin School ubicada en Latacunga-Cotopaxi para evidenciar la efectividad de los modelos de clasificación de EDM. Por medio del software estadístico R se ejecutaron 8 modelos de clasificación sobre los datos de los alumnos y se realizó la evaluación de estos por medio de la matriz de confusión y el área bajo la curva ROC. Como resultado se obtuvo que el modelo NNET tiene una exactitud del 98% sobre los datos recogidos del alumnado para la predicción de los posibles abandonos en un año lectivo. Adicionalmente también se comprobó que el mismo conjunto de datos puede servir para explicar la retención de alumnos de la institución, aunque en un menor porcentaje de exactitud.

Palabras Clave:

ABANDONO ESCOLAR

MINERÍA DE DATOS EDUCACIONAL

CLASIFICACIÓN

NNET

ROC

ABSTRACT

Early Leaving or drop-out school is a current problem that directly affects the education system and people who are involved in the education process such as family, institution or country, so early detection of school drop-out is a key strategy for academic institutions at all levels. A current solve for this detection is the use of Educational Data Mining (EDM) techniques that allows to discover patterns in demographic and academic data of the students; being one of the cases the use of classification models to predict the students that drop-out the institution. In this study, this conceptual framework was used with a data set of three academic years of the "Lenin School" Educational Unit located in Latacunga-Cotopaxi to demonstrate the effectiveness of the EDM classification models. Using the statistical software R, 8 models of classification were executed on the data of the students and the evaluation of these was made by the confusion matrix and the area under the ROC curve. As a result of the study was obtained the NNET model which has an accuracy of 98% on the data collected from the students for the prediction of possible drop-outs in a school year. In addition, it was also verified that the same data set can be used to explain the retention of students in the institution, even though in a lower percentage of accuracy.

Keywords:

DROPOUT SCHOOL

EDUCATIONAL DATA MINING

CLASSIFICATION

NNET

ROC

CAPÍTULO 1.

INTRODUCCIÓN

Este trabajo de investigación está estructurado de cinco capítulos que se desarrollan de la siguiente manera:

Capítulo 1: Este capítulo contiene el planteamiento del problema en donde se detalla la situación actual, la descripción del problema, la justificación e importancia, los objetivos general y específicos a cubrirse en esta investigación, el alcance y la hipótesis a verificarse.

Capítulo 2: Trata sobre el estado del arte referente a la deserción escolar a nivel internacional y nacional, conceptos de inteligencia de negocios y minería de datos y cómo estas metodologías son utilizadas en el ámbito educativo.

Capítulo 3: Describe la propuesta de un modelo de prevención de abandono escolar en base a los datos de la unidad educativa "Lenin School", también se describe cual fue el pre-procesamiento de datos realizado, la selección de técnicas de modelados, así como la generación y evaluación de los modelos.

Capítulo 4: Expone los resultados de predicción de los modelos, y la interpretación que se le puede dar a estos resultados.

Capítulo 5: Finaliza el escrito con las conclusiones obtenidas de la investigación junto con los trabajos futuros que pueden desarrollarse como continuación de esta tesis.

1.1 Planteamiento del problema

1.1.1 Situación actual del problema

Empíricamente se delimita a la deserción/abandono escolar como el abandono del centro escolar y de los estudios por parte del alumno/a debido a diversos factores; por lo tanto, el abandono escolar como un fenómeno de la educación secundaria tiene que ver con el hecho de que un número de jóvenes se retiran de las aulas o del sistema educativo de una institución.

Un estudio de abandono escolar en el Ecuador de Sánchez en 2013 detalla lo siguiente:

> "Del análisis acerca del abandono escolar por tipo de sostenimiento, se evidencia que las menores tasas de abandono escolar se encuentran en la educación particular, con un 2,0% para el año lectivo 2013-2014, mientras que las mayores tasas se encuentran en la educación fiscal, con un 3,8% para el mismo año lectivo (…)" (pág. 53).

En el mismo estudio; Sánchez en 2013, describe varias estadísticas de las cuales entre las más relevantes declara:

> "Se desprende de los datos que la mayor tasa de abandono se encuentra en Bachillerato, con un 6,1%, frente al 2,8% de la Educación General Básica. De manera adicional, los hombres presentan una mayor tasa de abandono, con un 3,7%, respecto al 3,1% de las mujeres (…)" (pág. 64).

Otro análisis realizado por Antamba-Chacua del Ministerio de Educación del Ecuador en 2015 indica que:

> "El abandono al 1ro de bachillerato subió en un punto porcentual en los dos últimos años, llegando al 9.5% de estudiantes que abandonaron el sistema educativo en el 2013-2014, lo que significa que cerca de 27 mil estudiantes no continuaron con sus estudios, el 57% corresponde a hombres y el 43% a mujeres" (pág. 16).

Esta realidad educativa claramente perjudica el cumplimiento del objetivo 4.3 planteado en el Plan Nacional del Buen Vivir 2013-2017 (PNBV) que busca: "Reducir el abandono escolar en 8° de educación básica general y 1° de bachillerato al 3,0%" (Secretaría Nacional de Planificación y Desarrollo, 2013, pág. 176).

En este contexto de ideas, la detección temprana de estudiantes que abandonen sus estudios durante el año escolar se convierte en una de las herramientas necesarias de las instituciones educativas para prevenir la deserción escolar. Si bien la deserción escolar es un problema social, este se agrava al hablar en términos de oportunidades educativas ya que el estudiante que deserta en sus estudios deja un puesto que pudo haber sido ocupado por otro estudiante que con éxito hubiese concluido la carrera estudiantil. También la deserción tiene sus efectos económicos negativos porque el estado, la institución educativa y la familia del estudiante habrá usado parte de sus recursos en el estudiante para educarlo sin resultados finales adecuados. Con estos antecedentes y desde el entorno social existen varios estudios referentes a las diferentes problemáticas que originan el abandono escolar, así como también estudios sobre la detección temprana de posibles estudiantes desertores. Sin embargo, el enfoque de uso

de técnicas innovadoras de análisis de datos para la predicción de deserción estudiantil es un tema en auge en la comunidad científica.

1.1.2 Descripción del problema

En la Unidad Educativa "Lenin School" de la ciudad de Latacunga, existe preocupación respecto a la deserción de alumnos ya que empíricamente detectan que el abandono de estudios en el último año escolar fue mayor que los últimos 3 años. Más aún carecen de herramientas que ayuden en la detección temprana de alumnos con tendencias al abandono escolar. Con el objeto de que factores como bajo rendimiento en los estudios o problemas de conducta no sean los causantes de deserción, la institución trabaja con psicólogos y docentes propios que brindan ayuda a estos estudiantes detectados por los maestros; pero para la institución no es posible detectar todos los casos, a la vez que quienes reciben la ayuda de los profesionales terminan desertando debido a que no fueron detectados de una manera temprana u oportuna. Para agravar esta situación las autoridades aseveran que un factor para este fenómeno fue la reactivación del volcán Cotopaxi en agosto de 2015.

En ese aspecto la pregunta a ser resuelta en este proyecto es: ¿Aplicando técnicas de minería de datos en la información histórica de la Unidad Educativa Lenin School, será posible dar a conocer los patrones de deserción estudiantil en la misma?

1.2 Justificación e Importancia

Uno de los objetivos fundamentales descritos ampliamente en el PNBV es el fortalecimiento de las capacidades y potencialidades de la ciudadanía; y, para lograrlo uno de sus lineamientos impulsa el promover la culminación de los estudios en todos los niveles educativos; para ello indica el mismo PNBV que se debe "Investigar, prevenir y combatir los elementos que causan la expulsión involuntaria y el abandono escolar en los segmentos sociales de atención prioritaria, con acciones focalizadas e intersectoriales y con pertinencia cultural y territorial" (Plan Nacional para el Buen Vivir 2013-2017, 2013, pág. 168).

En este contexto es de importancia para las instituciones identificar de manera temprana aquellos estudiantes que posiblemente deserten durante el año lectivo; en este aspecto el uso de técnicas de minería de datos como son el descubrimiento de

patrones y la predicción de registros o la clasificación de los mismos permitirán encontrar información contenida en los datos demográficos y educativos de los estudiantes; a esta aplicación de técnicas sobre los datos educativos se conoce como Minería de Datos Educacional (EDM por sus siglas en inglés).

Romero & Ventura indican que el EDM es: "El desarrollo, investigación y aplicación de métodos computarizados para detectar patrones en grandes colecciones de datos educativos que de otro modo serían difíciles o imposibles de analizar debido a la enorme cantidad de datos dentro de la cual existen" (citado en Papamitsiou & Economides, 2014). Varios estudios e investigaciones como los de (Yukselturk, Ozekes, & Türel, 2014), (Pal, 2012), (Abu-Oda & El-Halees, 2015), (Sivakumar, Venkataraman, & Selvaraj, 2016), (Márquez-Vera, Romero M., & Ventura S., 2013), y muchos más en compilaciones como (Papamitsiou & Economides, 2014) proporcionan casos reales en los que estas técnicas de minería de datos permiten obtener dichas predicciones por medio de datos demográficos de los estudiantes, inclusive en (da Cunha, Moura, & Cesar, 2016, pág. 189) se propone el uso de estas técnicas con el fin de que los educadores puedan diagnosticar a tiempo las causas de la deserción escolar y la desaprobación, permitiendo tomar acciones pedagógicas oportunas.

Debido a esto, el trabajo a realizarse se desarrollará en conjunto con la Unidad Educativa Lenin School, que se encuentra ubicada en la ciudad de Latacunga provincia de Cotopaxi, esta Institución cuenta con los tres años básicos medios (octavos, novenos y décimos años de educación básica) y los tres años de bachillerato.

Además, el uso de técnicas de minería de datos para la educación en el país es un tema innovador a la vez que goza de alto interés en la comunidad científica internacional debido a que aborda temas de actualidad; así como también los hallazgos derivados de esta investigación proporcionarán información para la toma de decisiones de la institución que pueden trascender a la comunidad educativa en Cotopaxi y al país.

1.3 Objetivos

1.3.1 Objetivo General

Descubrir y analizar los patrones de deserción estudiantil de los alumnos en la Unidad Educativa Lenin School, utilizando técnicas de minería de datos educacionales.

1.3.2 Objetivos Específicos

1) Elaborar un modelo para descifrar el problema de deserción estudiantil en la Unidad Educativa Lenin School, en el que se incluyan tanto aspectos de Business Intelligence como técnicas de minería de datos.
2) Utilizar los datos aportados por la Unidad Educativa Lenin School para aplicar las técnicas planteadas en un modelo de clasificación para descubrir los patrones de deserción.
3) Proponer una alternativa de solución al problema de deserción estudiantil para la Unidad Educativa Lenin School y plantear algunos trabajos futuros.

1.4 Alcance

El presente proyecto culmina con la presentación de los resultados obtenidos de la aplicación de técnicas de minería de datos educacionales respecto a la predicción de alumnos que podrían abandonar sus estudios durante el año escolar. Así como también se presentarán detalles técnicos sobre la metodología utilizada en el proceso de minería de datos.

1.5 Hipótesis

H0: Aplicando técnicas de minería de datos en la información histórica de la Unidad Educativa “Lenin School”, es posible predecir los patrones de deserción estudiantil en la misma.

CAPÍTULO 2.

ESTUDIO DEL ESTADO DEL ARTE

En este capítulo se recopila el estado del arte con respecto a la deserción escolar, se presentan varios enfoques y la fórmula de cálculo que se utiliza en el Ecuador. Además, se recopila todo el marco teórico actual referente a inteligencia de negocios y minería de datos, recolectando para ambos conceptos tales como *Self Services Business Intelligence*, Minería de Datos Educacional, *Data Warehouse*, Técnicas de aprendizaje supervisado y no supervisado entre otros; y, metodologías actuales utilizadas como Kimball y CRISP-DM. Se finaliza con un análisis de los principales estudios de investigación relacionados de estas tecnologías en su uso a nivel educativo principalmente en la detección de patrones de deserción escolar.

2.1 Deserción escolar

2.1.1 Conceptualización

En cuanto a la definición de "deserción escolar" el profesor Stephen Lamb et al. (2016) propuso una premisa a considerar en este estudio:

> "Los mismos términos[1] presentan problemas importantes cuando se comparan los sistemas nacionales. Por ejemplo, el término "deserción" (*dropout*) se usa principalmente en los Estados Unidos y Canadá para referirse a los jóvenes que abandonan la escuela sin obtener un diploma de escuela secundaria. Es un término utilizado raramente por los organismos de estadística, las autoridades educativas y los centros de investigación de otros países. Otras naciones tienen conceptos similares, como el "abandono escolar temprano" (*Early school leaving*) y "no en la educación, el empleo o la formación" (*Not in education, employment or training - NEET*) (pág. 4).

Además, el profesor Stephem Lamb et al. (2016) también asevera que las definiciones al igual que los términos usados varían entre las naciones; entonces las investigaciones sobre deserción escolar deben estar contextualizadas y muy bien marcadas por las convenciones establecidas por cada país alrededor de este tema.

Para el Ecuador, en 2016 la Comisión Especial de Estadísticas de Educación define el término a utilizarse como "Abandono Escolar" y establece el cálculo de la "Tasa de

[1] Abandono, deserción. (Esta nota al pie le corresponde al investigador del proyecto)

Abandono Escolar" como el "Número de estudiantes contabilizados al final de un período escolar que abandonan un determinado grado o curso de estudios, expresado como porcentaje del total de estudiantes matriculados al final del mismo grado o curso de estudios y periodo escolar" (Comisón Especial de Estadísticas de Eduación, 2016).

FÓRMULA DE CÁLCULO:

$$TAB = \frac{Est\ Ab_g^t}{Est\ M_g^t} \times 100 \qquad \textbf{(1)}$$

Donde:

$\boldsymbol{TAB}$ = Tasa de abandono escolar

$\boldsymbol{Est\ Ab_g^t}$ = Número de estudiantes que abandonan el grado o curso **g** en el período escolar **t**.

$\boldsymbol{Est\ M_g^t}$ = Total de estudiantes matriculados en el grado o curso **g** en el periodo escolar **t**.

$\boldsymbol{g}$ = Grado o curso de estudio. (este indicador es aplicable para todos los grados de educación general básica y bachillerato)

$\boldsymbol{t}$ = Período escolar.

2.1.2 Razones de deserción en el Ecuador

Espínola citado en (Cortez & Pérez, 2016) establece que el éxito o el fracaso escolar dependen de un cúmulo de factores individuales, familiares, sociales y/o culturales que se afectan simultáneamente. En (Lamb, Markussen, Teese, Sandberg, & Polese, 2011) se describe que estos factores pueden aparecer temprano en la vida escolar como en etapas posteriores pero que pueden llegar a ser evidentes.

Son muchas las causas de la deserción, pero para el caso presente se pueden resaltar las siguientes (Yépez, 2013): la situación geográfica, la crisis económica y algunos factores sociales especialmente los relacionados con la desintegración familiar, la migración de los niños del campo a las ciudades en busca de mejores oportunidades.

Las **causas geográficas**: Los cambios domiciliarios de la familia obligan a continuar el período escolar en otros establecimientos, sin desertar necesariamente del sistema

educativo. En cambio, la emigración familiar puede aparecer como deserción de los estudios en el país.

Las **causas económicas**: una de las causas del abandono escolar es la difícil situación económica en la que viven muchos niños, que les lleva a priorizar la búsqueda de trabajo para sobrevivir antes que ir a la escuela. El Coordinador Nacional del Movimiento Contrato Social por la Educación Milton Luna citado en (Parra, 2010) manifiesta que:

> "En el Ecuador, uno de cada tres niños no llega a completar los seis años de educación primaria, uno de cada cinco niños abandona la escuela en quinto de básica y tres de cada diez niños de séptimo de básica desertan de la escuela."

Las **causas sociales**: En (Espíndola & León, 2002) explican que:

> "La deserción escolar genera elevados costos sociales y privados. La baja productividad del trabajo, y su efecto en el crecimiento de las economías, se considera también como un costo social del bajo nivel educacional que produce el abandono de la escuela durante los primeros años del ciclo escolar".

Las escuelas cumplen una función muy importante en la prevención del abandono escolar, siempre que se la entienda como un protector de riesgo para los estudiantes, como una comunidad de compañerismo y compromiso.

Estas causas o factores de abandono son mencionadas en varios estudios y reportes a nivel nacional como los motivos que tiene el estudiante para abandonar los estudios. Cada investigación realizada tiende a determinar varias causas; sin embargo, dependiendo de la naturaleza y resultados de la investigación se tiende a encontrar diferentes predominancias de cada una, entre ellos por ejemplo (GUAMÁN, 2016), (Antamba-Chacua, 2015), (Pérez, 2014), demuestran que el factor socio-económico es uno de los determinantes para la culminación exitosa del curso, debido a que este influye en el nivel de asistencia del alumno (Antamba-Chacua & Quituisaca-Samaniego, 2015), (Atamba Chacua, 2015), los estudios concuerdan además que este no es el único factor. En otras investigaciones se señala al embarazo como el causante del abandono (Valdivieso, 2013), (De la Vega, 2014), la nutrición y hábitos de estudio (Hidalgo, 2012), el contexto familiar (Cortez & Pérez, 2016), (Sánchez, 2015), (Pacho & Chiqui, 2011), también son analizados los niveles de escolaridad, tales como asistencia y desempeño escolar (GUAMÁN, 2016), (Pacho & Chiqui, 2011),

(Hinojosa & Zambrano, 2012), infraestructura y cuerpo docente (Rodríguez J. , 2016). Existen estudios (Sánchez, 2015), (Bravo & Jacqueline, 2011), (Montenegro L. & Taco C., 2012), (Yépez, 2013) que ubican a las razones personales, familiares y pedagógicas como las principales causas de abandono.

Las causas revisadas en los estudios locales son coincidentes con varias investigaciones internacionales (Instituto de Estadística de la UNESCO, 2012), (Valmyr B., Lydersen, & Kvernmo, 2016), (Mendoza & Zúñiga, 2017), (De Witte, Cabus, Thyssen, Groot, & Maassen van den Brink, 2013), (European Commission/EACEA/Eurydice/Cedefop, 2014), (Tanja & Van der Velden, 2008). En el libro (Lamb, Markussen, Teese, Sandberg, & Polese, 2011) se encuentran publicados más de quince artículos científicos referentes al abandono escolar, los estudios son compilaciones de investigadores internacionales y cada uno hace referencia a las distintas realidades y enfoques que cada país aporta con respecto a esta problemática; a pesar de ser de contextos diferentes al Ecuador, los factores que rondan alrededor del abandono son similares en los estudios realizados.

En (Bruce & Bridgeland, 2011) se indica que el abandono escolar es un proceso acumulativo; en su publicación los investigadores demuestran en sus hallazgos que existen tres indicadores que pueden predecir una posible deserción. Estos indicadores son conocidas como A, B, C (*attendance, behavior, course performance*) por sus siglas en inglés: Ausentismo, Conductas desviadas y Rendimiento.

Por lo tanto, crear mecanismos para observar estos tres indicadores a tiempo en un estudiante permite una intervención oportuna.

2.2 Inteligencia de Negocios

2.2.1 Conceptualización y definiciones técnicas

La información estratégica es esencial para cualquier industria (George, Vijayakumar, & Santhosh Kumar, 2015). Según la misma publicación se requiere una derivación de la combinación de datos históricos y actuales. Por ello en (Grossmann & Rinderle-Ma, 2015) se identifica a la Inteligencia de Negocios como una tecnología necesaria en las empresas para realizar reconocimiento, análisis, modelamiento, estructuración y optimización de procesos de negocios. Según (Gallardo D. , 2012)

estos modelos utilizan conceptos de bases de datos estadísticos que siguen siendo utilizadas en componentes de inteligencia de negocios como OLAP y Data Warehouse.

En el marco de la inteligencia de negocios, BI, *Business Intelligence* (Gallardo D. , 2012) menciona que:

> "Es necesario soportar el procesamiento para transformar datos en información "accionable". Es decir, aprovechar todos los datos que se tienen almacenados y que son producto de la labor diaria, para convertirlos en información como base de las decisiones y acciones a ejecutar." (pág. 1).

2.2.1.1 Data Warehouse

Un *data warehouse* ofrece datos coherentes de alta calidad y que soporten diversas actividades de BI; en particular reportes estándar (Grossmann & Rinderle-Ma, 2015). Esto se logra debido a que un *data warehouse* es una base de datos que soporta procesamiento analítico y asistencia en el proceso de toma de decisiones (Martinho¸Bruno & Yasmina S., 2016).

Se considera al *data warehouse* como el corazón de un sistema de inteligencia de negocios (Dedić & Stanier, 2016), por tanto las consideraciones de calidad de datos, promedio de respuestas de consultas, actualización de datos, y los acuerdos de nivel de servicio son factores clave para determinar la satisfacción del usuario.

Kimball citado en (Taniar, 2009) indica que el *data warehouse* puede ser organizado como bases de datos multidimensionales denominados *data marts*. Según (Grossmann & Rinderle-Ma, 2015) las estructuras multidimensionales pueden ser implementadas en tres diferentes formas: La forma OLAP multidimensional (MOLAP), la forma OLAP relacional (ROLAP), y la forma *Hybrid* OLAP (HOLAP). MOLAP aplica estructuras nativas multidimensionales. ROLAP mapea estructuras multidimensionales en tablas relacionales. HOLAP se refiere a la aplicación combinada de estructuras relacionales y multidimensionales (pág. 101).

En ROLAP existen dos tipos de estructuras prominentes representativas que son el modelo en copo de nieve y el modelo en estrella (Grossmann & Rinderle-Ma, 2015); en ambos modelos existen dos tipos de tablas, las tablas de hechos (*fact tables*) y de dimensiones (*dimension tables*). La diferencia entre ambos modelos radica en la forma en que se relacionan las tablas; en un modelo en estrella las tablas de dimensiones se

relacionan únicamente con las tablas de hechos mientras que en un modelo en copo de nieve las dimensiones extienden sus relaciones a otras dimensiones; en ambos modelos no existen relaciones entre tablas de hechos (Gallardo D. , 2012).

2.2.1.2 Self-Service BI

(Gartner, 2016) define este paradigma como usuarios finales diseñando e implementando sus propios informes y análisis dentro de una cartera de herramientas y arquitectura aprobada y soportada.

En (Imhoff & White, 2011), (Alpar & Schulz, 2016), (Pour, 2014) se determina además que el *Self-Service BI* (SSBI) es una tendencia actual en la industria; donde las áreas de negocio buscan descentralizar del área de IT la generación de reportes avanzados. Y como menciona (Jacobsson, Ransnäs, & Runnström, 2015) esto hace que el análisis sea más accesible a un grupo más amplio de usuarios; y, también ha dado lugar a una mayor necesidad de funcionalidad de autoservicio.

El SSBI cumple con cuatro objetivos (Imhoff & White, 2011):

1. Fácil acceso a las fuentes de datos.
2. Facilidad de uso de herramientas BI.
3. Fácil consumo e interpretación de resultados.
4. Rápido manejo y fácil administración de la solución *data warehouse*.

Actualmente destacan tres fabricantes de software líderes en el área de BI cuyas herramientas cumplen los objetivos del SSBI. Ellos son: Power BI de Microsoft, Tableu de Tableu Inc. Y Qlik de QlikTech (Gartner, 2017).

La consultora Forrester también menciona en su reporte (Forrester, 2015) a los líderes en Agil BI Platforms a Micrsoft, Qlik y Tableu.

2.2.2 Metodologías

En desarrollo de software existen varias metodologías de desarrollo y cada una puede ser aplicada según la conveniencia del proyecto; así mismo, para una solución de inteligencia de negocios existen diferentes metodologías y ciclos de vida a ser utilizados; sin embargo, en (Jukic, 2006) se determinan dos metodologías contemporáneas más usadas al momento de diseñar un *data warehouse*, estas son las

metodologías de Inmon y Kimball. Cada metodología tiene diferentes enfoques, la Tabla 1 resumen las principales diferencias.

Tabla 1

Diferencia de enfoques Inmon y Kimball

EL ENFOQUE INMON	EL ENFOQUE KIMBALL
Implementación de arriba hacia abajo	Implementación de abajo hacia arriba
El *data warehouse* se desarrolla sobre la base del modelo de datos de toda la empresa	Comienza con un *data mart* (ej: ventas), luego se adicionan más *data mart* (ej: marketing, cartera, etc.)
El *data warehouse* con un único repositorio alimenta datos a los data marts	Los datos fluyen desde la fuente hacia los *data marts*; y, luego al *data warehouse*
De implementación larga	Implementación rápida. La implementación se hace por etapas
Puede fallar debido a la falta de paciencia y compromiso	Necesidad de asegurar la consistencia de la metadata

Fuente: Traducido de (Gabriel, 2017)

Ambos enfoques son utilizados dependiendo del entorno del proyecto. Por la naturaleza de esta investigación y en base a (Revelo C., Hinojosa, & Duque, 2013) nos centraremos en el enfoque Kimball. Como se aprecia en (Grossmann & Rinderle-Ma, 2015), (Kimball & Ross, 2008) el enfoque Kimball, también conocido como metodología Kimball o "*Kimball Lifecycle*" por el *Kimball Group*, ha sido utilizado en una gran variedad de proyectos de inteligencia de negocio; inclusive a nivel regional esta metodología es utilizada como referente para proyectos de BI.

La clave de la metodología Kimball se basa en el entendimiento y enfoque del negocio; como se menciona en (Kimball & Ross, 2008) no es suficiente tener el modelo de datos perfecto o la mejor tecnología. Se necesita coordinar las múltiples facetas de un proyecto BI.

La Figura 1 muestra el flujo general de implementación de un *data warehouse*. El "*Kimball Lifecycle*" identifica la secuencia de tareas a realizarse en un alto nivel; el

detalle de las tareas dependerá del entorno del proyecto, así como el hecho de si todas las etapas serán utilizadas ya que cualquier tarea puede obviarse dependiendo de las necesidades del negocio.

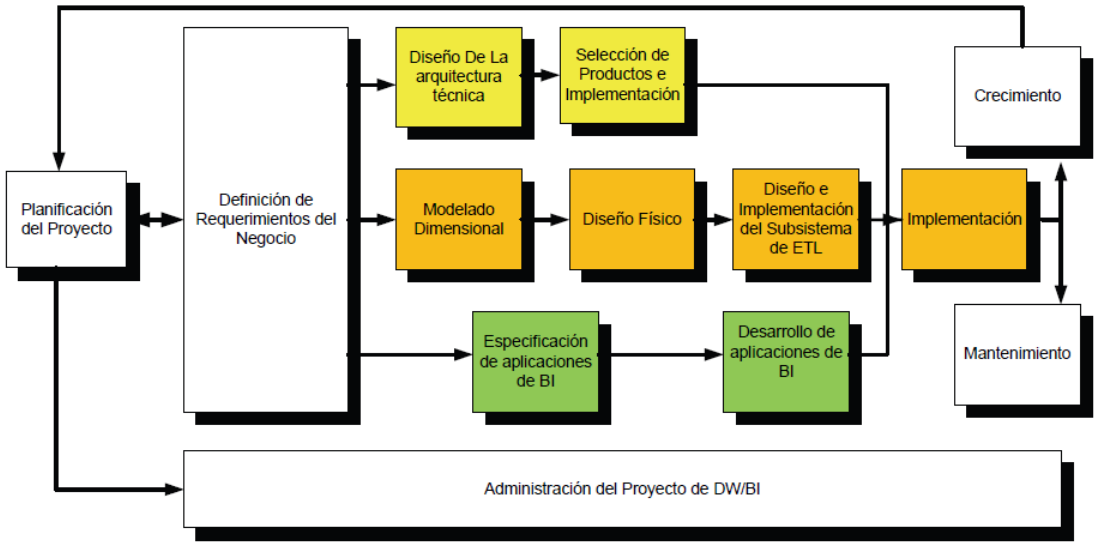


Figura 1. Diagrama del *Kimball Lifecycle*

Fuente: *(Kimball & Ross, 2008)*

Se debe resaltar el rol central de la tarea de definición de requerimientos. Los requerimientos del negocio son el soporte inicial de las tareas subsiguientes. También tiene influencia en el plan de proyecto (de ahí la doble fecha entre la caja de definición de requerimientos y la de planificación).

En el Diagrama del *Kimball Lifecycle*, se puede evidenciar tres rutas cuyos enfoques están enmarcados en tres diferentes áreas que se combinan para construir la solución:

a. Tecnología (Parte superior).
b. Datos (Parte media).
c. Aplicaciones de Inteligencia de Negocios (Parte inferior).

2.2.3 Principales estudios relacionados

Las publicaciones referentes a los modelos de inteligencia de negocios aplicados en la industria son sumamente extensas; como se puede apreciar en estos artículos

científicos del *International Journal of Data Warehousing and Mining* (IJDWM): (Campbell, Mao, Pei, & Al-Barakati, 2017), (Xu, 2016), (Ravat, Song, & Teste, 2016), (Bimonte, Saulot, Journaux, & Faivre, 2017), o en el *Business Intelligence Journal*: (Watson, 2017), (Cerqueira & Brandão, 2017), (Michel, 2017), (Armstrong, Lewandowski, & Loftis, 2017).

En relación a la temática del uso de los métodos de inteligencia de negocios en el ámbito educativo podemos iniciar citando a (Campbell & Oblinger, 2007) donde se demuestra el uso del *data warehouse* como repositorio de datos de diversas fuentes que permite la ejecución de consultas complejas y análisis sin interrumpir o ralentizar los sistemas de producción.

Otro caso se describe en (Bernhardt, 2005) donde se evidencia las bondades del uso de herramientas de datos para las escuelas, estas herramientas como los *data warehouses* educacionales permiten realizar seguimientos a los datos históricos educativos de los estudiantes; inclusive a través de diversos lugares geográficos realizando cruces de información entre diferentes bases de datos.

Un caso de mayor extensión es la creación de la plataforma "*Wisconsin Information System for Education Data Dashboard*" o WISEdash (Wisconsin Department of Public Instruction, 2016). WISEdash es un portal que utiliza *dashboards*, o colecciones visuales de gráficas y tablas, que proveen múltiples años de datos educacionales de las escuelas de Wisconsin. Cada año las escuelas distritales de Wisconsin recogen información acerca de los alumnos, docentes y cursos para luego almacenarlos en un *data warehouse* que sirve como fuente de reportes y acceso a datos (Wisconsin Department of Public Instruction, 2016).

Sin embargo, el uso de un *data warehouse* educativo es como lo menciona la literatura antes citada, un repositorio de datos para realizar el análisis de los datos; en varias publicaciones citadas se puede apreciar que la etapa final de las investigaciones consiste en la ejecución de técnicas y métodos de minería de datos con el fin de encontrar información oculta en los datos.

2.3 Minería de Datos

2.3.1 Conceptualización y definiciones técnicas

El campo de inteligencia de negocios aborda soluciones informáticas que generen valor agregado a la organización basados en datos o más específicamente en hechos. Desde Kaplan y Norton como precursores del Cuadro de Mando Integral (BSC por sus siglas en inglés) las técnicas y modelos de inteligencia de negocios han tomado varios caminos por medio de los que se puede realizar análisis de datos con el fin de exponer a nivel gerencial y de forma automática los resultados reales de la empresa por medio de indicadores al igual que permite realizar un profundo análisis de la variabilidad de los datos a través del tiempo (Kaplan, 2009). Sin embargo, dado los últimos avances en tecnología y la importancia del conocimiento en la industria por la era de la información en la que nos encontramos; nuevas técnicas y modelos emergen en la comunidad científica; una de estas es la denominada minería de datos (Ale M. J., Data Mining - Revision, 2015). A continuación, se presentan tres conceptos de minería de datos:

a. La minería de datos en ciencias de la computación, es el proceso de descubrir patrones usables e interesantes en grandes volúmenes de datos relacionados. (Clifton, 2015)
b. Según (Gartner Inc., 2015), la minería de datos es el proceso de descubrir correlaciones significativas, patrones y tendencias buscando a través de grandes cantidades de datos almacenados en repositorios.
c. (Ale M. J., Data Mining - Revision, 2015) Define la minería de datos como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil, desde grandes volúmenes de datos.

Son cuantiosas las definiciones encontradas y varían en sus detalles con cada autor, sin embargo, el punto central de estas es el descubrimiento de información importante, útil y sobre todo descubierta en grandes volúmenes de datos; es decir el término "Minería de Datos" aparece en las técnicas usadas para indagar en grandes volúmenes de datos (Jaramillo B. , 2015).

La minería de datos es realmente sólo el siguiente paso en el proceso de análisis de datos en el mundo de la inteligencia de negocios (Ver Figura 2). En lugar de obtener

consultas sobre las relaciones especificadas o usuarios estándar, la minería de datos va un paso más allá mediante la búsqueda de relaciones significativas en los datos. Las relaciones que se pensaba que no han existido, o las que le dan una visión más detallada de los datos (Hema & Malik, 2010).

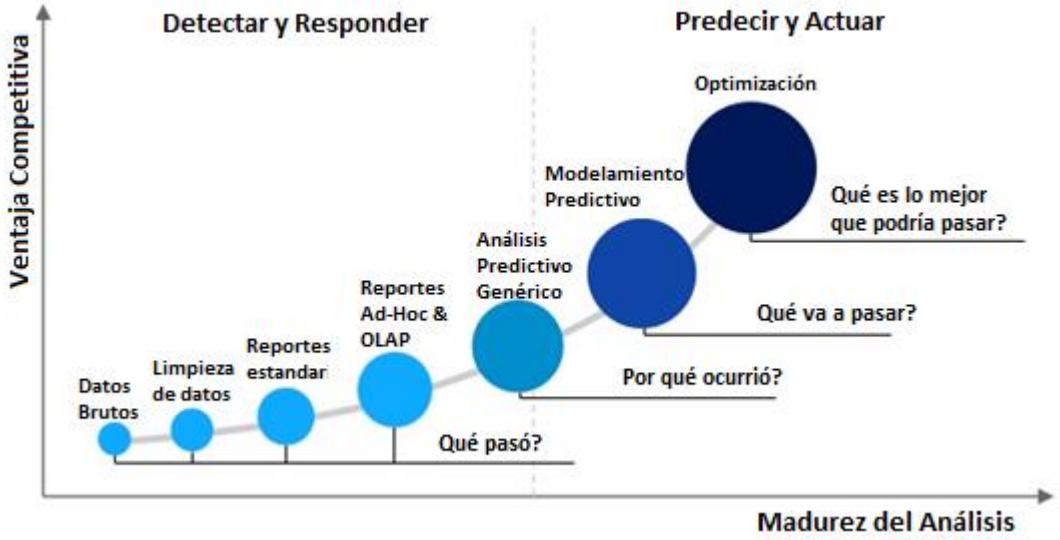


Figura 2. Evolución del análisis predictivo

Fuente: *(Vinueza, 2015)*

2.3.1.1 Fundamentos

Minería de datos también conocido como "*Knowledge Discovery in Databases*" (KDD) tiene tres raíces o fundamentos genéricos (Gorunescu, 2011) (Ver Figura 3); la asociación de estos fundamentos y tras un largo proceso de investigación y desarrollo de productos dieron como resultado múltiples técnicas utilizadas en la manipulación masiva de datos.

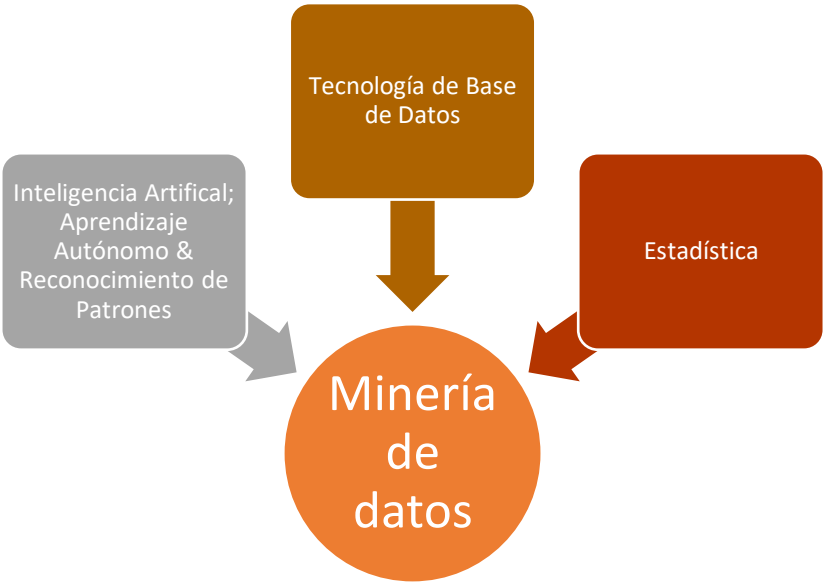


Figura 3. Fundamentos de minería de datos

Fuente: *(Gorunescu, 2011)*

2.3.1.2 Modelamiento Estadístico

Las estadísticas clásicas usan técnicas que podemos resumir en lo que se conoce como Análisis Exploratorio de Datos (EDA por sus siglas en inglés) (Gorunescu, 2011), que se utiliza para identificar las relaciones sistemáticas entre diferentes variables de información acerca de su naturaleza.

Utilizando técnicas estadísticas se obtiene información útil de los datos brutos; basados en teorías probabilísticas, el análisis de datos estadístico provee hechos históricos desde datos empíricos; varias técnicas en modelos estadísticos han sido desarrolladas en base a esta premisa; estos modelos permiten a los expertos descubrir resultados históricos; pero también existen trabajos realizados en predicción usando técnicas analíticas (Holmes & Jain, 2012).

2.3.1.2.1 Análisis Predictivo

Algunos algoritmos permiten la predicción de tendencias futuras (Holmes & Jain, 2012); uno de ellos por ejemplo es el uso de árboles de decisión, los cuales se enfocan en observaciones de un conjunto de variables que son estadísticamente analizadas para generar una base de predicción sobre un atributo cualitativo predefinido por valores

previos; a este proceso se lo conoce como aprendizaje dirigido, y puede utilizar varias técnicas de inteligencia artificial y aprendizaje autónomo.

2.3.1.3 Proceso de Minería de datos

El proceso de minería de datos es secuencial e iterativo; el camino a seguir en un proyecto de este tipo contempla típicamente los siguientes pasos (Aggarwal, 2015):

- **Recolección de datos:** Puede requerir de software o hardware especializado dependiendo del problema; o únicamente una conexión estable a una base de datos; en muchos casos este es el típico ejemplo; sin embargo, puede requerirse la ayuda de sensores o incluso de trabajo manual para realizar la recolección de datos; parte de este proceso también contempla el entendimiento del negocio; ya que en esta etapa es necesario recolectar los datos requeridos para la resolución del problema.
- **Pre-procesamiento de datos (Limpieza):** Una vez recolectados los datos; estos deben pasar por un proceso de limpieza y transformación; debido a que si se utiliza no solo una sino varias fuentes los datos pueden requerir incluso hasta un proceso de federación de datos. La transformación de datos es obligatoria ya que de esta depende colocar la información en un formato legible para los diferentes algoritmos; por ejemplo, el algoritmo entenderá datos categóricos en la edad como “De 18 a 25 años” que los datos numéricos como tales. Es posible que el algoritmo se comporte de una u otra manera dependiendo de datos nulos o vacíos, por lo que es importante controlar que los datos sean lo más completos posibles.
- **Procesamiento analítico (Modelamiento):** En esta etapa el analista diseña un modelo de análisis con las diferentes técnicas estadísticas, las cuales les permitirán responder al problema planteado o en su defecto encontrar las tendencias y patrones contemplados como parte del problema. Es necesario una evaluación del modelo donde se mida la eficiencia del mismo; si el modelo no cumple con el umbral esperado o si los resultados obtenidos del proceso no son consistentes es necesario realizar una iteración a pasos anteriores para ajustar el modelo (Olson & Delen, Advanced Data Mining Techniques, 2008).

Al finalizar el proceso de minería de datos, como entregable se despliegan los resultados encontrados en torno al modelamiento. Esto se lo realiza por medio de un informe escrito que contempla el análisis e interpretación de los resultados.

2.3.1.4 Modelos

Los modelos generados para el análisis de datos pueden ser de dos tipos:

- **Modelos de Análisis Descriptivos:** El principal objetivo de este tipo de modelamiento es hallar patrones interpretables que describan los datos. Estos pueden ser reglas asociativas, agrupaciones de datos, similitudes entre los datos; se puede encontrar además conjuntos de segmentos.
- **Modelos de Análisis Predictivos:** Este modelamiento utiliza algoritmos estadísticos y de análisis para predecir valores desconocidos o futuros de unas variables en base a datos existentes de otras variables.

2.3.1.5 Técnicas

En minería de datos, las técnicas pueden corresponder a dos tipos; de aprendizaje supervisado y aprendizaje no supervisado; las técnicas de aprendizaje supervisado utilizan datos previos o de entrenamiento con el fin de autoajustar sus funciones hasta obtener un resultado equivalente o aceptable a la realidad (Rivero Pérez, 2014); por otro lado las técnicas de aprendizaje no supervisado no utilizan estos datos de entrenamiento, por lo que tratan a los datos ingresados como un conjunto de variables aleatorias sobre los cuales puede encontrar patrones (Ghahramani, 2004).

A continuación, se listan técnicas comúnmente utilizadas según (Bramer, 2013):

2.3.1.5.1 Aprendizaje Supervisado: Clasificación

Dentro del aprendizaje supervisado se encuentra la categoría “clasificación”, que es una de las aplicaciones más comunes para la minería de datos. Corresponde a una tarea que se repite con frecuencia en la vida cotidiana. Por ejemplo: un hospital puede querer clasificar a los pacientes que se encuentran en las siguientes condiciones de riesgo de adquirir una determinada enfermedad: alto, medio o bajo. Una compañía de encuestas de opinión podría clasificar a las personas entrevistadas en: los que son propensos a votar por un candidato de determinado partido político o están indecisos; o, podemos desear clasificar un proyecto estudiantil en las siguientes calidades: distinción, mérito,

pasar o fallar. En todos los casos mencionados la clasificación se utiliza con el fin de predecir una etiqueta.

Existen muchas formas de realizar esto incluyendo:

- *Algoritmo del vecino más cercano:* Es un método de clasificación supervisada que sirve para estimar la función de densidad $F(x/Cj)$ de las predictoras x por cada clase C_j (Fix & Hodges, 1989). Se utilizan para el reconocimiento de patrones; un ejemplo de este algoritmo es el: k-vecinos más Cercanos (k-nn) (Altman, 1992).
- *Reglas de clasificación:* Halla un modelo para el atributo clase como una función de los valores de otros atributos en un conjunto de datos. Un ejemplo de este algoritmo es el *Quote Rule* (Ellis, Michaely, & O'Hara, 2000)
- *Árboles de decisión:* Técnica que permite analizar decisiones secuenciales basada en el uso de resultados y probabilidades asociadas. Un ejemplo de este algoritmo es el C 4.5 (Quinlan, 2014).

2.3.1.5.2 Aprendizaje Supervisado: Predicción numérica o Regresión

La regresión es utilizada con la finalidad de realizar una predicción numérica; tal como la predicción de ventas de una compañía o el comportamiento de las acciones de una empresa.

Algunas maneras de hacerlo son:

- *Regresión lineal:* Predicen una o más variables continuas, como las pérdidas o los beneficios, basándose en otros atributos del conjunto de datos. Un ejemplo de algoritmos de regresión son las series temporales (Hamilton, 1994).
- *Series temporales:* Una serie temporal se define como una colección de observaciones de una variable recogidas secuencialmente en el tiempo. Estas observaciones se suelen recoger en intervalos iguales de tiempo. Si los datos se recogen en instantes temporales de forma continua, se debe: bien digitalizar la serie, es decir, recoger sólo los valores de intervalos iguales de tiempo; o, bien acumular los valores sobre intervalos de tiempo (Marin, 2015).

- *Redes neuronales artificiales:* Las redes neuronales artificiales se han hecho muy populares debido a la facilidad en su uso e implementación y su habilidad para aproximar cualquier función matemática. Las Redes de Neuronas Artificiales, son eficientes obteniendo resultados de conjuntos de datos complicados e imprecisos, por lo que pueden utilizarse para extraer patrones y detectar tramas que son muy difíciles de apreciar por humanos u otras técnicas computacionales (Gestal P., 2015).

2.3.1.5.3 Aprendizaje No Supervisado: Reglas de asociación

Dado un conjunto de registros, cada uno de los cuales contiene un cierto número de ítems, tomados de una cierta colección se genera reglas de dependencia que predecirán la ocurrencia de un ítem basado en las ocurrencias de otros ítems (Ale M. J., Data Mining - Revision, 2015).

Una forma de aplicación de esta técnica es el "Análisis de canastas de mercado", donde se busca obtener los productos relacionados que los consumidores compran en un negocio a través del tiempo.

El nombre del algoritmo más utilizado para realizar esta aplicación es: "*A priori"* (Agrawal & Srikant, 1994).

2.3.1.5.4 Aprendizaje No Supervisado: Clustering

Los algoritmos de clúster examinan los datos en búsqueda de grupos de ítems que son similares; estas similitudes son medidas por medio de distancias euclidianas si los atributos son continuos, pero pueden utilizar otros tipos de medidas específicas para problemas específicos.

Como características cabe señalar que los objetos dentro de un clúster son similares entre sí; y los objetos en clústeres separados son menos similares entre sí.

Un ejemplo de este algoritmo es el *K-Medias* (Hartigan & Wong, 1979).

2.3.1.6 Minería de datos educacional

La minería de datos educacional ha emergido como un área de investigación independiente en años recientes, culminando en 2008 con el establecimiento de la

Conferencia Internacional anual sobre Minería de Datos Educacional y el "*Journal of Educational Data Mining*" (Baker R. , in press).

La minería de datos educacional (EDM) como menciona Romero y Ventura se refiere a "desarrollo, investigación, y aplicación de métodos computarizados para detectar patrones en grandes colecciones de datos educacionales que de otro modo sería muy difícil o imposible analizar debido al enorme volumen de datos existentes." (citado en Papamitsiou & Economides, 2014).

Baker et. al. también mencionan que: "EDM desarrolla y adapta métodos estadísticos, *machine-learning* y de minería de datos para estudiar los datos educacionales generados básicamente por estudiantes e instructores. Sus aplicaciones pueden ayudar a analizar el proceso de aprendizaje del estudiante considerando sus interacciones con el ambiente" (citado en Calvet L. & Juan P., 2015). Por esta razón es común encontrar modelos extraídos de la literatura psicométrica en publicaciones de EDM (Baker & Yacef, 2009).

(Nithya, Umamaheswari, & Umadevi, 2016) resume 4 objetivos para el EDM del estudio (Baker & Yacef, 2009):

- **Predecir el futuro comportamiento del estudiante**: Este objetivo se puede lograr mediante la creación de modelos de estudiantes que incorporen las características del alumno, incluyendo información detallada como su conocimiento, comportamientos y motivación para aprender.
- **Descubrir o mejorar los modelos de dominio**: A través de varios métodos y aplicaciones de EDM, es posible descubrir nuevos modelos y mejoras a los existentes.
- **Estudiar los efectos del apoyo educacional**: Se puede lograr a través de sistemas de aprendizaje.
- **Conocimiento científico avanzado sobre el aprendiz y el aprendizaje:** Mediante la construcción e incorporación de modelos de estudiantes, el campo de la investigación EDM, la tecnología y el software utilizado.

Los métodos de trabajo propuestos en (Romero & Ventura, 2013) son abstraídos en (Calvet L. & Juan P., 2015) y presentados en la Tabla 2:

Tabla 2

Métodos comunes en EDM

Método	Meta/Descripción	Aplicaciones clave	Ejemplo
Predicción	Para inferir una variable objetivo de alguna combinación de otras variables. Clasificación, regresión y estimación de la densidad son tipos de métodos de predicción.	Predecir el rendimiento de los estudiantes y detectar los comportamientos de los estudiantes	(Yadav & Pal, 2012)
Clustering	Identificar grupos de observaciones similares.	Agrupar materiales o estudiantes similares basados en sus patrones de aprendizaje e interacción.	(Antonenko, Toy, & Niederhauser, 2012)
Minería de relaciones	Estudiar las relaciones entre variables y codificar reglas. La minería de reglas de asociación, la minería secuencial de patrones, la minería de correlación y la minería de datos causales son los principales tipos.	Identificar las relaciones en los patrones de comportamiento de los estudiantes y diagnosticar las dificultades de los estudiantes.	(Kinnebrew & & Biswas, 2012)
Destilación de datos para el juicio humano	Representar los datos de manera inteligible mediante el resumen, la visualización y las interfaces interactivas.	Ayudar a los instructores a visualizar y analizar las actividades en curso de los	(Baker, Corbett, & Wagner, 2006)

		estudiantes y el uso de la información.	
Descubrimiento con modelos	Utilizar un modelo previamente validado de un fenómeno como componente en otro análisis.	Identificación de las relaciones entre comportamientos y características de los estudiantes o variables contextuales. Integración de los marcos de modelización psicométrica en modelos de aprendizaje automático.	(Jeong & Biswas, 2008)
Detección de valores atípicos	Para señalar a individuos significativamente diferentes.	Detección de alumnos con dificultades o procesos de aprendizaje irregulares.	(Ueno, 2004)
Análisis de redes sociales	Analizar las relaciones sociales entre entidades en una red de información	Interpretación de la estructura y las relaciones en actividades de colaboración e interacciones con herramientas de comunicación.	(Palazuelos, García-Saiz, & Zorrilla, 2013)
Minería de procesos	Para obtener conocimiento de los	Reflejar el comportamiento del estudiante en	(Trćka, Pechenizkiy,

	procesos de los *event logs*	términos de examinación de sus trazas, que consiste en una secuencia de curso, grado y marca de tiempo.	& Aalst, 2011)
Minería de texto	Extraer información de alta calidad del texto.	Analizar los contenidos de foros, chats, páginas web y documentos.	(Tane, Schmitz, & Stumme, 2004)
Rastreo de conocimiento	Para estimar el dominio de habilidades del estudiante, empleando tanto un modelo cognitivo que asigna un tema de resolución de problemas a las habilidades requeridas; y, registros de las respuestas correctas e incorrectas de los estudiantes como evidencia de su conocimiento en una habilidad particular.	Control del conocimiento de los estudiantes a lo largo del tiempo.	(Lee & Brunskill, 2012)
Matriz de factorización no negativa	Definir una matriz M de números positivos con los resultados de las pruebas de los estudiantes que se pueden descomponer en dos matrices: Q, que representa una matriz	*Assessment* de habilidades de estudiantes	(Desmarais, 2011)

	de ítems, y S, que representa el dominio de habilidades del estudiante.

Fuente: (Calvet L. & Juan P., 2015)

2.3.2 Metodologías

Existen varias metodologías utilizadas en proyectos de minería de datos; sin embargo, una encuesta realizada por KDnuggets en 2014 y 2007 indica que la más ampliamente usada es CRISP-DM (*CRoss Industry Standard Process for Data Mining*), como se muestra en la Figura 4. Una revisión y crítica de los modelos de minería de datos en 2009 llamó a CRISP-DM el "estándar de facto para el desarrollo de la minería de datos y los proyectos de descubrimiento de conocimiento" (Marbán, Mariscal, & Segovia, 2009).

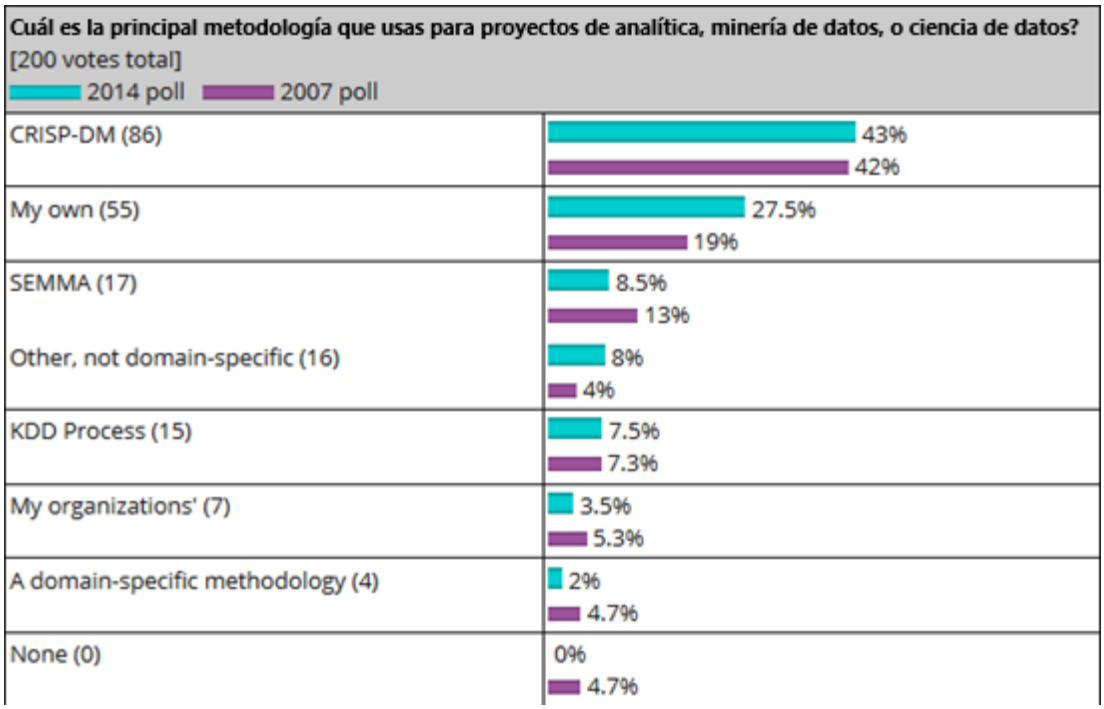


Figura 4. Encuesta de uso de metodologías en proyectos de minería de datos

Fuente: *(KDnuggets, 2016)*

Según las opiniones recopiladas en la encuesta de KDnuggets, CRISP-DM es la más ampliamente recomendada debido a que inicia con el reconocimiento y entendimiento

del negocio y de sus datos; y termina con la construcción de modelos que describen o perciben el comportamiento de las variables del negocio entorno al problema.

Interpretando a (Shearer, 2000) CRISP-DM es una metodología iterativa; debido a que antes de finalizar el proceso se debe realizar una evaluación al modelo; es decir se mide la calidad del mismo y dependiendo de esta medición se procede al despliegue de resultados o se comienza nuevamente el proceso de análisis. Como se observa en la Figura 5; la metodología recoge los pasos típicos mencionados en el punto 2.3.1.3 Proceso de Minería de Datos.

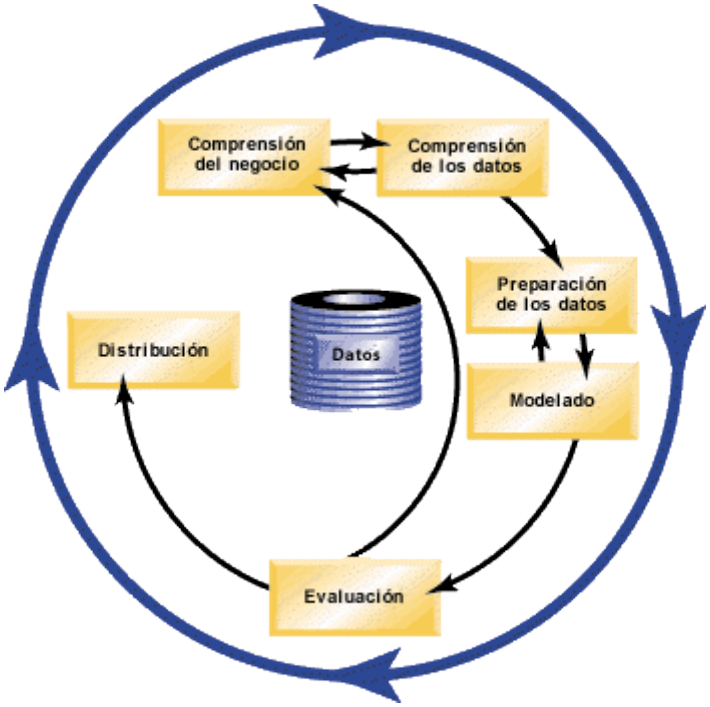


Figura 5. Modelo de proceso CRISP-DM

Fuente: *(IBM, 2012)*

CRISP-DM contempla 6 etapas en su proceso (Olson, 2007):

1. *La comprensión de negocios:* La comprensión de negocios incluye determinar los objetivos del negocio, la evaluación de la situación actual, el establecimiento de minería de datos, metas y desarrollar un plan de proyecto.
2. *La comprensión de datos:* Una vez que los objetivos de negocio y el plan del proyecto son establecidos, la comprensión de datos considera las necesidades de datos. Este paso puede incluir la recolección inicial de datos, descripción de los datos, la exploración de datos y la verificación de la calidad de los datos.

Exploración de datos, como resumen de visión estadística (que incluye la presentación visual de las variables categóricas). Modelos como el análisis de conglomerados también pueden ser aplicados durante esta fase, con la intención de identificación de patrones en los datos.

3. *Preparación de datos:* Una vez que los recursos de datos disponibles se identifican, estos necesitan ser seleccionados, limpiados, construidos en la forma deseada, y con formato. Esta fase contempla la limpieza de datos y la transformación de datos. También puede aplicarse una exploración de datos a mayor profundidad; y, los modelos adicionales utilizados, proporcionando de nuevo la oportunidad de ver los patrones basados en la comprensión del negocio.
4. *Modelado:* En esta etapa se utiliza herramientas de software de minería de datos como la visualización y análisis de conglomerados (identificar qué variables tienen mayor sentido juntas) son útiles para el análisis inicial. Herramientas como la inducción de reglas generalizadas puede desarrollar reglas iniciales de asociación. Una vez que se obtiene una mayor comprensión de los datos (a menudo a través de reconocimiento de patrones), los modelos más detallados y apropiados al tipo de datos pueden ser aplicados. La división de datos en aprendizaje y de prueba también es necesaria para el modelado.
5. *Evaluación:* Los resultados del modelo se deben evaluar en el contexto de los objetivos del negocio establecidos en la primera fase (comprensión del negocio). Esto dará lugar a la identificación de otras necesidades, y con frecuencia volviendo a fases anteriores de CRISP-DM. Ganar comprensión del negocio es un procedimiento iterativo en la minería de datos, donde los resultados de diversas visualizaciones, estadística, y las herramientas de inteligencia artificial muestran al usuario nuevas relaciones que proporcionan una comprensión más profunda de operaciones de la organización.
6. *Despliegue:* La minería de datos se puede utilizar tanto para verificar hipótesis preexistentes, o para el descubrimiento de conocimiento (identificación de inesperado y relaciones útiles). A través del conocimiento descubierto en las fases anteriores del proceso CRISP-DM, algunos modelos de predicción pueden ser obtenidos y podrían ser aplicados a las operaciones de negocio para muchos propósitos, incluyendo predicción o la identificación de situaciones

clave. Estos modelos tienen que ser monitorizados debidos a los cambios en las condiciones de operación, debido a lo que podría ser verdad hoy puede no ser cierto dentro de un año. Si se producen cambios significativos, el modelo debe ser hecho de nuevo.

2.3.3 Principales estudios relacionados

Como se aprecia en la sección de minería de datos existen muchos trabajos de investigación realizados en diversos campos de la industria; también como se menciona en la Tabla 2 existen muchos trabajos relacionados al EDM.

Para el caso concreto de abandono estudiantil podemos citar varios trabajos realizados; entre ellos destacan: (Sivakumar, Venkataraman, & Selvaraj, 2016) , donde se utiliza técnicas de clasificación probando que el uso de árboles de decisión genera mejores resultados sobre otros algoritmos; el algoritmo permitió identificar los atributos relevantes de datos socio-demográficos, académicos e institucionales de estudiantes sin graduarse de una universidad en India. Otro ejemplo se muestra en esta publicación (Yukselturk, Ozekes, & Türel, 2014) donde se realiza un análisis de deserción con alumnos que se inscribieron en una plataforma de estudios online; se recolectó información de los *logs* de la plataforma y se complementó con varias encuestas *online* a los estudiantes; se aplicaron 4 clasificadores *k-Nearest Neighbour* (k-NN), *Decision Tree* (DT), *Neural Network* (NN) y *Naive Bayes* (NB) de los cuales se obtuvo una detección de sensibilidad del 87%, 79.7%, 76.8% y 73.9 respectivamente.

Otro caso es (Abu-Oda & El-Halees, 2015) quienes aplicaron los clasificadores *Decision Tree* (DT), *Naive Bayes* (NB) y donde la exactitud encontrada fue de 98.14% y 96.86% respectivamente para datos de estudiantes de la universidad de ALAQSA; por su parte (Pal, 2012) prueba varios algoritmos de tipo árbol de decisión siendo ID3 el que mejor exactitud logra obtener con 85.7% utilizando datos académicos y demográficos de los estudiantes de ingeniería del Instituto de Ingeniería y Tecnología de la Universidad de Purvanchal. Por otro lado (da Cunha, Moura, & Cesar, 2016) utiliza algoritmos de *clustering* para generar grupos de atributos que permitan identificar el estado de desertores en las bases de datos del IFRN (*Federal Institute of Rio Grande do Norte*) de Brasil. (Dekker, Pechenizkiy, & Vleeshouwers, 2009) utiliza

de igual forma árboles de decisión con exactitud entre el 75 y 80% sobre datos de estudiantes de primer semestre de Ingenieria Eléctrica en la Universidad Técnica de Eindhoven.

(Luan, 2002) propone por otro lado la predicción de la persistencia del alumno en la institución académica; siendo este un factor un significado de mejores programas académicos y de mejores ingresos; de igual manera utiliza los algoritmos *Neural Networks* (NN), C5.0 y C&RT; siendo el mejor puntuado el NN con 85.1% para los No Persistentes y 27.5% para los Persistentes.

A nivel nacional existe la publicación (García-Tinizaray, Ordoñez-Briceño, & Torres-Diaz, 2014) donde se realiza una regresión logística sobre variables extraídas de una plataforma web de estudios a distancia (LMS – *Learning Management System*); se tomaron en cuenta 6 cursos de un semestre de clases (Oct 2012 – Feb 2013). Por otro lado (CÓRDOVA G., 2014) aplicó los algoritmos *K-Means* y J48 sobre los datos de una muestra de estudiantes del año lectivo 2013-2014 teniendo un mejor resultado el algoritmo *K-Means*. (Ordoñez B., 2013) también utilizó *K-Means* y J48 sobre la muestra de estudiantes utilizando las plataformas de educación a distancia de la universidad como son: Entorno Virtual de Aprendizaje (EVA), Sistema Académico (Syllabus); en este caso igualmente tuvo mejor resultado el algoritmo *K-Means*.

CAPÍTULO 3.

PROPUESTA DEL MODELO

El capítulo inicia con una descripción de la situación de abandono escolar en la unidad educativa “Lenin School” y una descripción detallada de la propuesta del modelo de análisis; se realiza una revisión de las fuentes de información, definición de variables descriptivas y variables objetivo con un análisis exploratorio de estas. Se genera los modelos de clasificación y se efectúan dos pruebas de rendimiento sobre estos; por último, se muestran los resultados de clasificación obtenidos en cada modelo y los resultados de las pruebas.

3.1 Modelo de prevención de abandono escolar

3.1.1 El abandono escolar en la Unidad Educativa “Lenin School”

Las autoridades de la Unidad Educativa “Lenin School” requieren ayuda en el proceso de prevención de abandono escolar en sus estudiantes; las autoridades aseveran que el abandono escolar en la institución es un problema que requiere asistencia previa; para ello la institución cuenta con el apoyo de un psicólogo en el departamento de consejería estudiantil quien mantiene entrevistas con todos los estudiantes a fin de detectar problemas familiares o personales que desemboquen en deserción; pero conforme avanza el período académico se hace poco evidente detectar a los posibles estudiantes que abandonarán el colegio.

Adicionalmente un problema que aparece desde el año 2013 es la baja tasa de retención de los estudiantes; al empezar un nuevo año lectivo las autoridades evidenciaron un declive en el número de alumnos matriculados con respecto al año anterior; este problema se agudizó aún más con la activación del volcán Cotopaxi en agosto de 2015.

Es importante la prevención temprana del abandono escolar porque por medio de ello es posible que la institución pueda generar estrategias para contrarrestar este fenómeno. El Capítulo 2 de este trabajo de investigación aclara que una de las herramientas actuales para la detección temprana de abandono escolar es el uso de

EDM. Siendo las técnicas de clasificación las más aceptadas por la comunidad científica para la predicción del abandono escolar.

3.1.2 Descripción de la propuesta

En base a la necesidad de la unidad educativa de realizar una indagación objetiva en los datos de sus alumnos; se propone utilizar la metodología CRISP-DM con técnicas de EDM de tipo clasificación y de aprendizaje supervisado sobre los datos demográficos y académicos de los alumnos matriculados en los años lectivos 2013-2014, 2014-2015 y 2015-2016 con el fin de evidenciar cuales son las variables significativas y las posibles asociaciones entre las mismas que puedan predecir el abandono estudiantil.

Para los datos de entrenamiento se utilizará la información de los alumnos de los años 2013-2014 y 2014-2015; y los datos de prueba serán del año 2015-2016. La selección del modelo se realizará una vez se evalúen las matrices de confusión generadas de 10 algoritmos de clasificación; el modelo ganador será el que mayor exactitud de aciertos tenga en su matriz de confusión con los datos de prueba.

Una vez realizada la selección del modelo se recogerán las variables significativas que la predicción haya arrojado; y, se elaborará un informe ejecutivo que exponga los hallazgos encontrados; este informe será entregado a la unidad educativa como constancia de fin del proyecto y el mismo servirá para que las autoridades puedan confeccionar medidas de prevención de abandono escolar en el nuevo año lectivo.

Para la exitosa ejecución del proyecto, la unidad educativa se compromete en brindar todos los recursos que satisfagan la recolección de datos; tales como la ubicación de ex alumnos, docentes y estudiantes de apoyo para la recolección de datos, uso de salones para entrevistas, aula de informática con acceso a internet, proyector, etc. Una vez recopilados los datos se utilizará el computador del investigador para las tareas posteriores y el software especializado será de licenciamiento gratuito. Se utilizará la herramienta de SSBI de Microsoft: “Power BI” para la generación de tableros analíticos y el análisis exploratorio de datos; posteriormente, se utilizará el software estadístico R para las tareas de análisis y predicción.

Igualmente, la unidad educativa cuenta con información demográfica y académica de sus alumnos; sin embargo, existe cierta información que deberá ser recolectada por medio de entrevistas y encuestas. Para ello, se realizará una entrevista previa con secretaría, inspectoría y el departamento de orientación vocacional con el fin de conocer a fondo cual es la información que la unidad educativa posee de sus alumnos y que variables son las faltantes.

Como parte de la gestión del proyecto se presenta además la siguiente Tabla 3 con los principales riesgos y contingencias que podrían presentarse en esta investigación:

Tabla 3

Tabla de principales riesgos del proyecto

Ord	Descripción	Probabilidad	Impacto	Evitación	Mitigación	Plan de contingencia
1	No encontrar a todos los ex alumnos de la unidad educativa durante los años 2013-2014, 2014-2015 y 2015-2016 para completar las variables faltantes	80%	1. Catastrófico	Utilizar únicamente las variables extraídas de los datos que el psicólogo haya levantado en sus entrevistas con los alumnos de la institución únicamente.	Cotejar datos de vivienda y teléfonos entre secretaría, el inspector general y el psicólogo de la unidad educativa	De no ser posible ubicar a todos los alumnos se deberá trabajar con una muestra significativa
2	No encontrar un modelo de predicción con un índice de exactitud sobre el 80%	80%	2. Crítico		Estimar una predicción con varios algoritmos de clasificación y seleccionar el que tenga mejor exactitud de predicción	

Fuente: Elaboración propia.

3.2 Pre-procesamiento de datos

3.2.1 Fuentes de información

El siguiente listado enumera las fuentes de información utilizadas en el presente proyecto:

1. Base de datos escolástica: Es la fuente primaria de información; de aquí derivan los datos de matrícula de los tres años de análisis; incluyendo datos escolares del alumno, como promedio general y disciplina; estos datos fueron contrastados con información pública del Archivo Maestro de Instituciones Educativas (AMEI).
2. Fichas del registro acumulativo general del Departamento de Consejería Estudiantil: Son fichas impresas llenadas en las sesiones de trabajo del psicólogo con los estudiantes, en estas fichas se detalla la información demográfica de los estudiantes incluyendo situación de vivienda, datos familiares, y socio-económica.
3. Encuestas a alumnos del colegio: Con motivo de completar la información faltante de las variables escogidas para la elaboración del modelo de predicción se elaboraron encuestas a alumnos y exalumnos del colegio; la ejecución de la encuesta se la realizó en conjunto con el Inspector General del colegio previo una sesión de trabajo explicativo del objetivo de esta y por menores de las preguntas; este mismo conocimiento fue reproducido hacia los encuestados. Las encuestas fueron aplicadas a alumnos del periodo 2015-2016 y vía email (Previo llamada telefónica) a los ex-alumnos que el colegio logró ubicar en primera instancia.
4. Entrevistas con ex alumnos: Las entrevistas consistieron en completar de forma presencial la encuesta provista en el punto 3 a los estudiantes que no pudieron ser ubicados directamente por el colegio; entre ellos constaron especialmente los alumnos que abandonaron el colegio durante los años de estudio. Se elaboró una lista de 67 estudiantes que no constaban en la encuesta; en este punto se determinó que el Riesgo 1 de la Tabla 3 no fue crítico ya que de los 67 estudiantes ninguno cambió de domicilio por lo que

en los fines de semana entre los meses de noviembre 2016 a febrero 2017 se cubrió toda la información requerida.

3.2.2 Definición de variables

Las siguientes variables fueron determinadas en base al análisis bibliográfico de los siguientes trabajos relacionados (Bacher, Tamesberger, Leitgöb, & Lankmayer, 2014) (Valmyr B., Lydersen, & Kvernmo, 2016), (Lamb, Markussen, Teese, Sandberg, & Polese, 2011), (Narváez & Barragán, 2015), (Sánchez, 2015), también fueron contrastadas e inferidas del "*Educational Longitudinal Study (2002) Ranked Reasons for Dropout in 2006 by Student Dropouts*" (National Dropout Prevention Center/Network AT CLEMSON UNIVERSITY, 2016) junto con algunas recomendaciones dadas por el Psicólogo de la Unidad Educativa (Cepeda & Gallardo, 2016). Todas las variables tienen relación directa con el alumno:

1. Variables Demográficas

1.1. Cambio_Domicilio

Descripción: Determina si el alumno tuvo un cambio de domicilio en alguno de los años lectivos

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Si
- No

1.2. Edad

Descripción: Se determina en base a la fecha de nacimiento del alumno

Origen: Base de datos

Tipo de dato: Int

1.3. Enfermedades_Cronicas

Descripción: Determina si el alumno tiene enfermedades crónicas en alguno de los años lectivos

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Si
- No

1.4. Estado_Civil

Descripción: Determina el estado civil del alumno durante un período lectivo

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Soltero
- Casado
- Viudo
- Divorciado
- Unión Libre

1.5. Estructura_Familiar:

Descripción: Determina la estructura familiar de la que proviene el alumno

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango: Las siguientes clasificaciones de estructuras familiares son compendios obtenidos de diferentes fuentes: (Rodríguez & Sánchez, 2006), (Vargas M., 2014)**,** (Reyna, Salcido, & Arredondo, 2013)**,** (Waite, 2006)**,** incluido la entrevista realizada al psicólogo de la institución quien confirmó y amplió esta clasificación (Cepeda & Gallardo, 2016)

- *Familia Nuclear:* formada por la madre, el padre y los hijos, es la típica familia clásica.
- *Familia Extendida:* formada por parientes cuyas relaciones no son únicamente entre padres e hijos. Una familia extendida puede incluir abuelos, tíos, primos y otros consanguíneos o afines.
- *Familia Monoparental:* formada por uno solo de los padres (la mayoría de las veces la madre) y sus hijos. Puede tener diversos orígenes: padres separados o divorciados donde los hijos quedan viviendo con uno de los padres, por un embarazo precoz donde se constituye la familia de madre soltera y por último el fallecimiento de uno de los cónyuges.

- *Familia Adoptiva:* hace referencia a los padres que adoptan a un niño.
- *Familia Ensamblada:* está formada por agregados de dos o más familias (ejemplo: madre sola con hijos se junta con padre viudo con hijos)
- *Familia de Hecho:* este tipo de familia tiene lugar cuando la pareja convive sin ningún enlace legal.
- *Familia Homoparental:* formada por una pareja homosexual (hombres o mujeres) y sus hijos biológicos o adoptados.
- *Familia de Padres Separados:* los progenitores se han separado tras una crisis en su relación. A pesar de que se nieguen a vivir juntos siguen cumpliendo con sus deberes como padres.
- *Familia de Soporte:* Se delega autoridad a los hijos más grandes, para que cuiden a los hermanos más pequeños.
- *Familia de Migrantes:* las cabezas del hogar (madre y/o padre) migraron por diferentes razones y los hijos quedaron en manos de familiares directos (abuelos, tíos, etc.).

1.6. Grupo Étnico

Descripción: Auto denominación del grupo étnico del alumno

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Indígena
- Mestizo
- Afrodescendeinte
- Blanco
- Otro

1.7. Genero

Descripción: Género del alumno

Origen: Base de datos

Tipo de dato: String

Rango:

- M: Masculino

- F: Femenino

1.8. Ingresos_Familiares

Descripción: Determina el nivel de ingresos familiares.

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- 0 - 700
- 701-1400
- 1401-2100
- 2101-2800
- 2801-3500
- 3501-5000
- 5001-10000
- Más de 10000

1.9. Lengua_Materna

Descripción: Especifica la lengua materna del estudiante

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Español
- Quechua
- Otro

1.10. Madre_Condicion_Actividad

Descripción: Determina la condición económica de la madre del alumno. Está basado en la variable Madre_Ocupación.

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango: Los rangos están basados en (INEC, 2014)

- PEA - Población económicamente activa
- PEI - Población económicamente inactiva
- No Contesta (En caso de no conocer la ocupación del progenitor)

1.11. Madre_Educacion

Descripción: Determina el nivel de educación de la madre del alumno

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Ninguno
- Primaria
- Secundaria
- Técnico Superior
- Tecnología
- Superior (Universitario)
- Posgrado
- Doctorado

1.12. Madre_Ocupacion

Descripción: Determina el grupo ocupacional en el que se encuentra la Madre.

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango: Los rangos para el PEA están basados en la Clasificación Internacional Uniforme de Ocupaciones (CIUO-08) mencionados en (INEC, 2014), mientras que los rangos del PEI son fuente directa de (INEC, 2014). Todos los subniveles correspondientes a cada grupo del PEA se encuentran especificados en el ANEXO I estos subniveles fueron compartidos a los alumnos encuestados y están basados en (Organización Internacional del Trabajo, 2008).

- PEA - Directores y gerentes
- PEA - Profesionales científicos e intelectuales
- PEA - Técnicos y profesionales de nivel medio
- PEA - Personal de apoyo administrativo
- PEA - Trabajadores de los servicios y vendedores de comercios y mercados
- PEA - Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros
- PEA - Oficiales, operarios y artesanos de artes mecánicas y de otros oficios

- PEA - Operadores de instalaciones y máquinas y ensambladores
- PEA - Ocupaciones elementales
- PEA - Ocupaciones militares
- PEI - Rentista
- PEI - Jubilado o pensionado
- PEI - Estudiante
- PEI - Ama de casa
- PEI - Incapacitado
- PEI - Otro
- No Contesta (En caso de no conocer la ocupación del progenitor)

1.13. Movilizacion_Colegio

Descripción: Determina los medios de transporte utilizados para movilizarse al colegio por el estudiante

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Vehículo familiar
- Transporte público
- Recorrido
- Bicicleta
- Caminata
- Otro

1.14. Num_Hermanos

Descripción: Especifica el número de hermanos del estudiante

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- 0
- 1
- 2
- 3
- más de 3

1.15. Padre_Condicion_Actividad

Descripción: Determina la población económica a la que pertenece el padre del alumno. Está basado en la variable Padre_Ocupación.

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango: Los rangos están basados en (INEC, 2014)

- PEA - Población económicamente activa
- PEI - Población económicamente inactiva
- No Contesta (En caso de no conocer la ocupación del progenitor)

1.16. Padre_Educacion

Descripción: Determina el nivel de educación del padre del alumno

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Ninguno
- Primaria
- Secundaria
- Técnico Superior
- Tecnología
- Superior (Universitario)
- Posgrado
- Doctorado

1.17. Padre_Ocupacion

Descripción: Determina el grupo ocupacional en el que se encuentra el padre.

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango: Los rangos para el PEA están basados en la Clasificación Internacional Uniforme de Ocupaciones (CIUO-08) mencionados en (INEC, 2014), mientras que los rangos del PEI son fuente directa de (INEC, 2014). Todos los subniveles correspondientes a cada grupo del PEA se encuentran especificados en el ANEXO I estos subniveles fueron compartidos a los alumnos encuestados y están basados en (Organización Internacional del Trabajo, 2008).

- PEA - Directores y gerentes

- PEA - Profesionales científicos e intelectuales
- PEA - Técnicos y profesionales de nivel medio
- PEA - Personal de apoyo administrativo
- PEA - Trabajadores de los servicios y vendedores de comercios y mercados
- PEA - Agricultores y trabajadores calificados agropecuarios, forestales y pesqueros
- PEA - Oficiales, operarios y artesanos de artes mecánicas y de otros oficios
- PEA - Operadores de instalaciones y máquinas y ensambladores
- PEA - Ocupaciones elementales
- PEA - Ocupaciones militares
- PEI - Rentista
- PEI - Jubilado o pensionado
- PEI - Estudiante
- PEI - Ama de casa
- PEI - Incapacitado
- PEI - Otro
- No Contesta (En caso de no conocer la ocupación del progenitor)

1.18. Posicion_Hermanos

Descripción: Determinar la posición que ocupa el estudiante entre sus hermanos

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Mayor
- Intermedio
- Menor

1.19. Residencia_Canton

Descripción: Nombre del cantón de residencia del estudiante durante el año en curso. Está basado en la variable Residencia_Sector.

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Latacunga
- Pujilí
- Salcedo
- Saquisilí
- Otro

1.20. Residencia_Sector

Descripción: Nombre de la parroquia de residencia del estudiante durante el año en curso.

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Latacunga - La Matriz
- Latacunga - Eloy Alfaro (San Felipe)
- Latacunga - Ignacio Flores (La Laguna)
- Latacunga - Juan Montalvo (San Sebastián)
- Latacunga - San Buenaventura
- Latacunga - Toacaso
- Latacunga - Pastocalle
- Latacunga - Mulaló
- Latacunga - Tanicuchí
- Latacunga - Guaytacama
- Latacunga - Aláquez
- Latacunga - Poaló
- Latacunga - Once de Noviembre
- Latacunga - Belisario Quevedo
- Latacunga - Joseguango Bajo
- Pujilí - Pujilí
- Pujilí - Angamarca
- Pujilí - Guangaje

- Pujilí - La Victoria
- Pujilí - Pilaló
- Pujilí - El Tingo La Esperanza
- Pujilí - Zumbahua
- Salcedo - San Miguel de Salcedo
- Salcedo - Antonio José Holguín (Sta. Lucía)
- Salcedo - Cusubamba
- Salcedo - Mulalillo
- Salcedo - Mulliquindil (Santa Ana)
- Salcedo - Panzaleo
- Saquisilí - Saquisilí
- Saquisilí - Cochapamba
- Saquisilí - Canchagua
- Saquisilí - Chantilín
- Otro

1.21. Tiempo_Movilizacion_Colegio

Descripción: Especifica el tiempo de movilización que le toma al estudiante trasladarse de su vivienda hacia el colegio

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- 1-15 min
- 16-30 min
- 31-45 min
- 45-60 min
- Más de 60 min

1.22. Tiene_Hijos

Descripción: Determina si el estudiante tiene hijos durante el año escolar

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Si
- No

1.23. Tiene_trabajo

Descripción: Determina si el estudiante tenía trabajo durante el año escolar

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Si
- No

1.24. Tipo_Vivienda

Descripción: Determina la condición de propiedad de la vivienda en la que vive la familia del alumno

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Propia
- Arrendada
- Donada
- Prestada por parientes o amigos
- Otro

1.25. Total_Amigos

Descripción: Número de amigos que el estudiante considera tener durante los años lectivos

Origen: Encuesta & Fichas de registro

Tipo de dato: int

2. Variables Escolares

2.1. Actual_Dias_Asistidos

Descripción: Total de días asistidos en un Quimestre del curso actual, el quimestre utilizado depende de la fecha de ingreso; es decir, se toma en consideración el quimestre al que se ingresa

Origen: Base de datos

Tipo de dato: Int

2.2. Actual_Faltas_Injustificadas

Descripción: Total de faltas injustificadas en un Quimestre del curso actual, el quimestre utilizado depende de la fecha de ingreso; es decir, se toma en consideración el quimestre al que se ingresa

Origen: Base de datos

Tipo de dato: Int

2.3. Actual_Faltas_Justificadas

Descripción: Total de faltas justificadas en un Quimestre del curso actual, el quimestre utilizado depende de la fecha de ingreso

Origen: Base de datos

Tipo de dato: Int

2.4. Actual_Promedio_Disciplina

Descripción: Promedio de disciplina de un Quimestre del curso actual, el quimestre utilizado depende de la fecha de ingreso

Origen: Base de datos

Tipo de dato: String

2.5. Actual_Promedio_General

Descripción: Promedio general de un Quimestre del curso actual, el quimestre utilizado depende de la fecha de ingreso

Origen: Base de datos

Tipo de dato: Decimal

2.6. Anio_Lectivo

Descripción: Año lectivo de matrícula. Esta variable es utilizada para el análisis exploratorio de datos; y, no se la utiliza en el proceso de predicción

Origen: Base de datos

Tipo de dato: String

Rango

- 2013-2014
- 2014-2015
- 2015-2016

2.7. Anterior_Colegio

Descripción: Identifica el tipo de colegio del que proviene el alumno.

Origen: Base de datos

Tipo de dato: String

Rango

- Academia
- Centro Educativo
- Colegio
- Escuela
- Instituto Tecnológico
- Unidad Educativa
- Unidad Educativa "Lenin School"

2.8. Apreciacion_Colegio

Descripción: Determina la apreciación general que tuvo el alumno del colegio en el año lectivo en curso

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Excelente
- Bueno
- Regular

2.9. Apreciacion_Docentes

Descripción: Determina la apreciación general que tuvo el alumno de los docentes en el año lectivo en curso

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango

- Excelente
- Bueno
- Regular

2.10. Apreciacion_Infraestructura

Descripción: Determina la apreciación general que tuvo el alumno de la infraestructura del colegio en el año lectivo en curso

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Excelente

- Bueno
- Regular

2.11. Clases_Particulares

Descripción: Determina si el alumno asistió a clases particulares en alguno de los años lectivos

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- **Si**
- **No**

2.12. Codigo_Matricula

Descripción: Código único de matrícula del estudiante

Origen: Base de datos

Tipo de dato: Int

2.13. Curso

Descripción: Determina el nivel educativo Octavo (1), Noveno (2), Décimo (3), Primero de Bachillerato (4), Segundo de Bachillerato (5) y Tercero de Bachillerato (6) con la jornada de estudios (Diurno, Nocturno)

Origen: Base de datos

Tipo de dato: String

2.14. Desayuno

Descripción: Determina si el alumno ingiere regularmente alimentos antes de asistir al colegio

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango

- **Si**
- **No**

2.15. Especialidad

Descripción: Especialidad a seguir del alumno

Origen: Base de datos

Tipo de dato: String

Rango:

- Administración de sistemas
- Básico común
- Ciencias
- Ciencias 2
- Contabilidad y administración

2.16. Estudiante

Descripción: Código único del estudiante. Esta variable se utiliza para general medidas calculadas o en funciones para determinar otras variables como la variable Retorno_Al_Colegio; no se utiliza directamente en el análisis exploratorio o en el proceso predictivo.

Origen: Base de datos

Tipo de dato: Int

2.17. Fecha de Ingreso

Descripción: Fecha en la que ingreso al colegio el alumno. De aquí se desglosa únicamente el Mes de Ingreso en el proceso de predicción, no utilizamos el año.

Origen: Base de datos

Tipo de dato: Date

2.18. Ha_Abandonado_Causa

Descripción: Determina la principal causa de abandono del estudiante, depende de la variable Ha_Abandonado.

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango: Las siguientes opciones fueron obtenidas de los análisis de (Cepeda & Gallardo, 2016), (Sánchez, 2015), (National Dropout Prevention Center/Network AT CLEMSON UNIVERSITY, 2016):

- Ningún motivo, no he abandonado un curso
- Problemas familiares

- Enfermedad en el hogar
- Mal entorno del colegio
- Problemas de socialización con compañeros
- Falta de recursos económicos
- Cambio de colegio
- Problemas académicos
- Problemas de disciplina
- Otro

2.19. Ha_Abandonado

Descripción: Identifica si el estudiante ha abandonado un curso alguna vez, indistintamente si el abandono se dio en los años de análisis.

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango

- Si
- No

2.20. Ha_Reprobado

Descripción: Identifica si el estudiante ha reprobado un curso alguna vez, indistintamente si la reprobación se dio en los años de análisis.

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Si
- No

2.21. Horas_Estudio

Descripción: Determina las horas de estudio dedicadas por el estudiante fuera del colegio

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- 0

- 1-2
- 3-4
- Más de 4

2.22. Lugar_Estudio

Descripción: Determina el lugar de estudio del estudiante

Origen: Encuesta & Fichas de registro

Tipo de dato: String

Rango:

- Cuarto de estudio
- Comedor
- Sala
- Cocina
- Dormitorio
- Patio
- Biblioteca
- Otro

2.23. Tipo_Matricula

Descripción: Especifica el tipo de matrícula con la que el estudiante ingresa al colegio

Origen: Base de datos

Tipo de dato: String

Rango

- Ordinaria
- Extraordinaria
- Excepción

2.24. Total_Estudiantes_Grupo

Descripción: Número total del grupo de estudiantes del curso del alumno

Origen: Base de datos

Tipo de dato: int

3. Medidas Calculadas

3.1. # Estudiantes

Descripción: Recuento de estudiantes

Fórmula: COUNT(Estudiante)

Tipo de dato: Int

3.2. % Abandono Curso Actual

Descripción: Determina el porcentaje de estudiantes que abandonaron los estudios del colegio; puede utilizarse para segregar el análisis a nivel de año lectivo, curso, etc. Abstraído de (Comisón Especial de Estadísticas de Eduación, 2016).

Fórmula: Número de estudiantes abandonados / total de estudiantes en un período dado

Tipo de dato: Decimal

3.3. % Regreso colegio

Descripción: Determina el porcentaje de estudiantes que del año en análisis regresaron el siguiente año a la unidad educativa; puede utilizarse para segregar el análisis a nivel de año lectivo, curso, etc. Abstraído de (Comisón Especial de Estadísticas de Eduación, 2016).

Fórmula: Número de estudiantes que regresaron el año siguiente / total de estudiantes en un período dado

Tipo de dato: Decimal

4. Variable objetivo de predicción

4.1. Abandono_Curso_Actual

Descripción: Identifica si el alumno abandonó los estudios en el año lectivo de análisis; esta variable es la que se desea predecir.

Origen: Base de datos

Tipo de dato: Bit

Rango

- 1: Si
- 0: No

4.2. Regreso_Al_Colegio

Descripción: Determina si el estudiante del año lectivo de análisis regresó a la institución el año siguiente de estudios (Por ejemplo, si el estudiante 123 en el

año 2013-2014 regresó a la institución en el año lectivo 2014-2015 entonces la variable tendrá el valor SI en el período 2013-2014 para ese estudiante).

En conversación con las autoridades del plantel, se considera oportuno aprovechar los detalles del modelo y variables construidas para cubrir una necesidad adicional de la unidad educativa y es el perfilamiento de estudiantes que retornarán al colegio el año siguiente de estudios excluyendo a los alumnos de Tercero de Bachillerato.

Origen: Base de datos

Tipo de dato: Bit

Rango

- 1: Si
- 0: No

3.2.3 Exploración de datos

La Unidad Educativa Lenin School contó con 677 alumnos en los años lectivos 2013-2014, 2014-2015 y 2015-2016; de los cuales el 42,54% de alumnos se encuentra en el año 2013-2014, los siguientes años lectivos descienden al 33,83% y al 23,63% en los años 2014-2015 y 2015-2016 respectivamente. Estos datos confirman la preocupación de las autoridades con respecto a la disminución de estudiantes; con respecto a la tasa de abandono en la unidad educativa va del 9,72%, 8,30% y 5% en los años 2013-2014, 2014-2015 y 2015-2016 respectivamente; si bien la tasa de abandono escolar es decreciente; la misma, sigue sobre los niveles estipulados en el PNBV. Aún más preocupante es analizar la tasa de retención y se observa un índice del 47,57%, 45,85% y 51,88% para los años 2013-2014, 2014-2015 y 2015-2016 respectivamente; es decir, del total de alumnos en un año lectivo, al menos la mitad no continuará en la institución. Estos datos se presentan en la Figura 6.

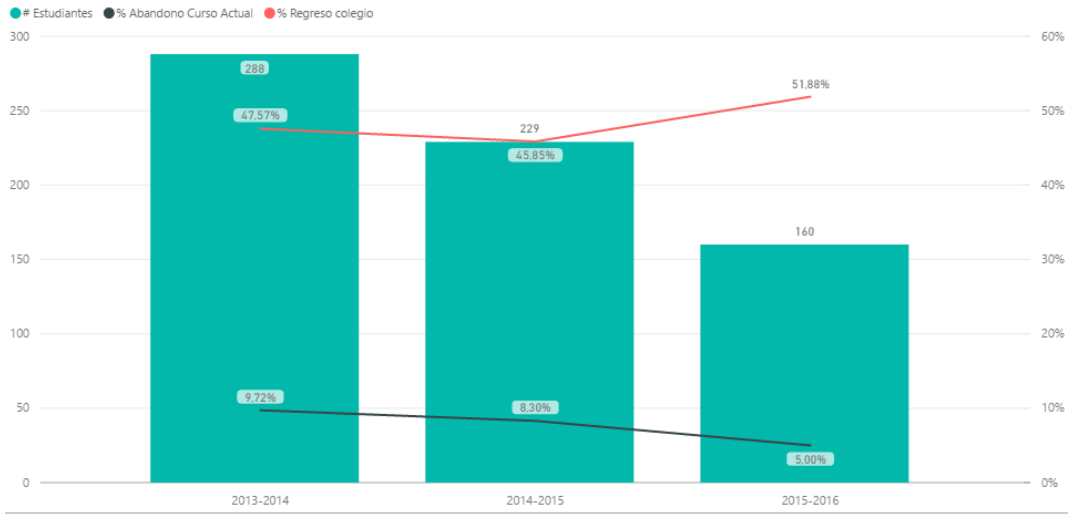


Figura 6. Total de Estudiantes, Tasa de Abandono y Tasa de retención por año lectivo

Como se observa en la Figura 7, la tasa de abandono a nivel de curso por todos los años lectivos de estudio tiene los índices más altos en los cursos de Primero de Bachillerato (4), Segundo de Bachillerato (5) y Décimo de Educación Básica (3) con el 13,57%, 12,96% y el 7,84% respectivamente.

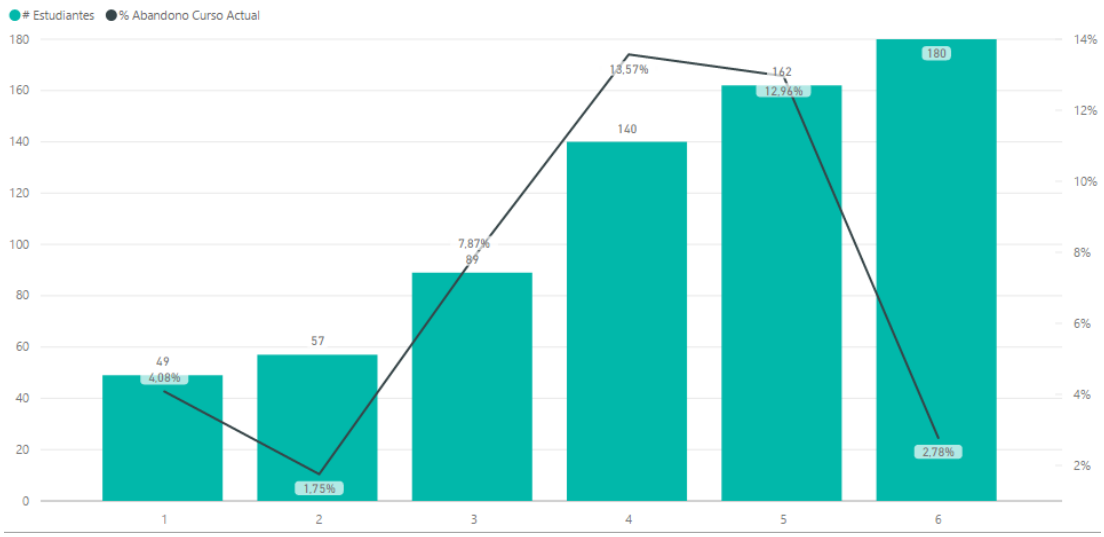


Figura 7. Total de estudiantes y tasa de abandono por curso

Igualmente, analizada por la variable curso, se aprecia en la Figura 8 que los niveles con menor tasa de retención son los cursos de Octavo (1), Décimo (3) y Primero de Bachillerato (4) con el 55,1%, 57,3% y 61,43%. Como se esperaba el nivel con menor tasa de retención corresponde al curso de Tercero de Bachillerato (6) con el 1,11%.

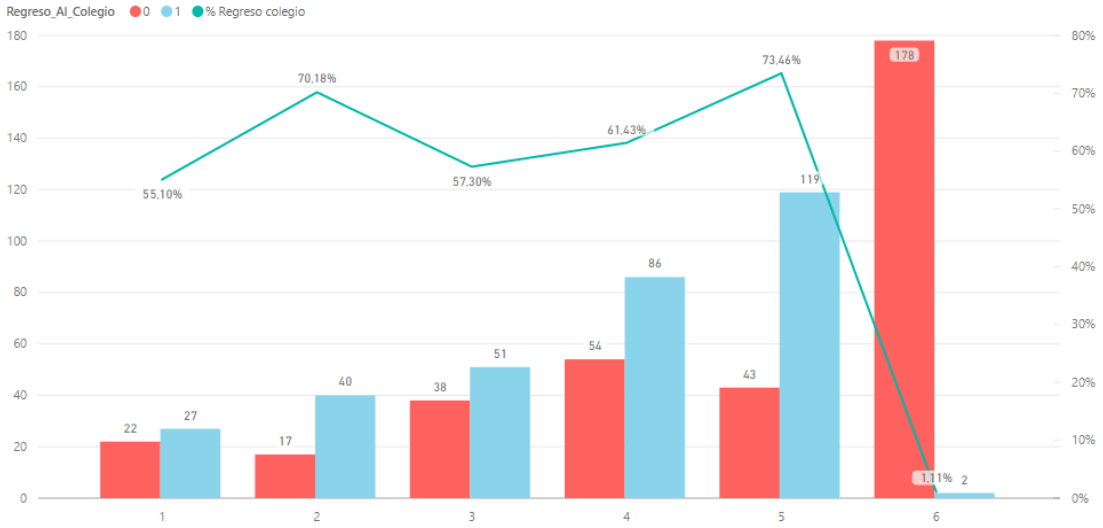


Figura 8. Tasa de retención por curso

Entre los datos demográficos tenemos que del total de estudiantes analizados en los años de estudio el 61,89% son de género masculino; el 95,27% se auto consideran mestizos; el 3,25% de alumnos tiene al quechua como lengua materna; solo el 1,92% tuvo un trabajo remunerado; el 0,59% tiene hijos; el 4,43% tiene enfermedades crónicas; y, el 3,69% de estudiantes se cambió de domicilio. El estado civil más frecuente es el de Soltero con el 98.52%; sin embargo, también están presentes estudiantes casados y en unión libre en un 0,74% equitativamente en ambos casos. Las edades más comunes del colegio son 16, 17 y 18 años con el 15,51%, 14,18% y 13,73% respectivamente; sin embargo, se aceptaron alumnos de edades mayor a los 20 años, la edad mínima es de 10 años y la máxima es de 42 años.

Con respecto a la familia de los estudiantes, el 70,9% proviene de una familia nuclear y el 83,90% de familias cuenta con vivienda propia. De los padres de los estudiantes el 20,68% corresponden al grupo de ocupación: “Trabajadores de los servicios y vendedores de comercios y mercados”, el 16,84% al grupo “Oficiales, operarios y artesanos de artes mecánicas y de otros oficios”; y el 15,81% al de “Ocupaciones elementales”; también, el 41,36% terminó la primaria y el 38,11% la secundaria. Mientras que las madres el 31,02% corresponden al grupo de ocupación: “Ocupaciones elementales”, el 16,84% al grupo “Trabajadores de los servicios y vendedores de comercios y mercados” y el 14,92% son “Amas de casa”; además, el 42,10% terminó la primaria y el 39,59% la secundaria; de los ingresos familiares el 55,98% tienen máximo 700 dólares de ingresos mensuales; y, el 85,08% de alumnos

tienen más de 1 hermano en casa; si cotejamos estos datos con los estudios realizados en (INEC, 2011) podríamos conjeturar que al menos el 50% de las familias de los alumnos analizados se encuentra en una estratificación socioeconómica media.

Con respecto a la distribución geográfica de los alumnos el 92,61% se encuentran en el cantón de Latacunga; y de este conjunto, el 53,59% se encuentra en la parroquia de Eloy Alfaro la cual se ubica muy cercana a la ubicación del colegio en el período de análisis. Para la movilización a la unidad educativa el 70,01% de estudiantes utiliza el transporte público y el 19,50% prefiere caminar. Al 62,63% de alumnos le toma hasta 15 minutos movilizarse al colegio, mientras que al 3.69% le toma de 30 a 60 minutos; con respecto a la alimentación el 66,62% si ingiere alimentos antes de ir a clases.

De los hábitos de estudio tenemos que el 62,19% de los estudiantes tienen como lugar de estudio sus dormitorios, y el 72,37% dedica entre 1 a 3 horas de estudio fuera del horario de clases. Las notas quimestrales más altas se encuentran en los cursos de Noveno, Tercero de bachillerato y Décimo con valores de 7,71; 7,64 y 7,41 respectivamente; además, el año lectivo con mejor puntaje es el 2015-2016 con 7,70; en todos los años lectivos la nota máxima excede el 9.8 sin llegar al 10 y existen alumnos con notas asentadas en cero; se indagó a secretaria los casos con notas menores de cinco; en total son 78 casos de los cuales 46 son abandonos, 25 corresponden a cambio de matrícula es decir se cambiaron de sección de diurno a nocturno o cambio de especialidad con nueva matrícula, en todos estos casos no hubo cambio de colegio por lo que no se considera abandono; y, los 8 restantes son casos de chicos problema según el inspector los cuales pasaron por supletorio al fin del año lectivo. Sobre la disciplina se tiene que el 54,80% tiene B y un 29,39% se calificó en D; solo el 12.85% de estudiantes tiene A en conducta y el restante se calificó como C. De los colegios de procedencia se tiene que el 53,62% provienen de la misma unidad educativa; este dato difiere del 48,01% calculado para el nivel de retención; debido a que para el análisis de la tasa de retención se contabilizaron aquellos estudiantes que inmediatamente el año siguiente continuaron sus estudios en la unidad educativa; mientras que el campo de colegio anterior indaga la procedencia del alumno, sin importar que este haya tenido un período sin estudios, es decir salió un año de la unidad

educativa, dedicó otro año a otras actividades sin haberse inscrito en otra institución educativa y luego regreso a Lenin School y continuó sus estudios.

Con respecto al análisis de abandono de los estudiantes se detecta que un 6,35% de estudiantes ha abandonado anteriormente los estudios; de ese grupo se aprecia que en los tres años de evaluación la mayor causa de abandono corresponde a un mal entorno del colegio; en el último año de estudio se evidencia un incremento considerable del factor "Problemas familiares" como se puede apreciar en la Figura 9.

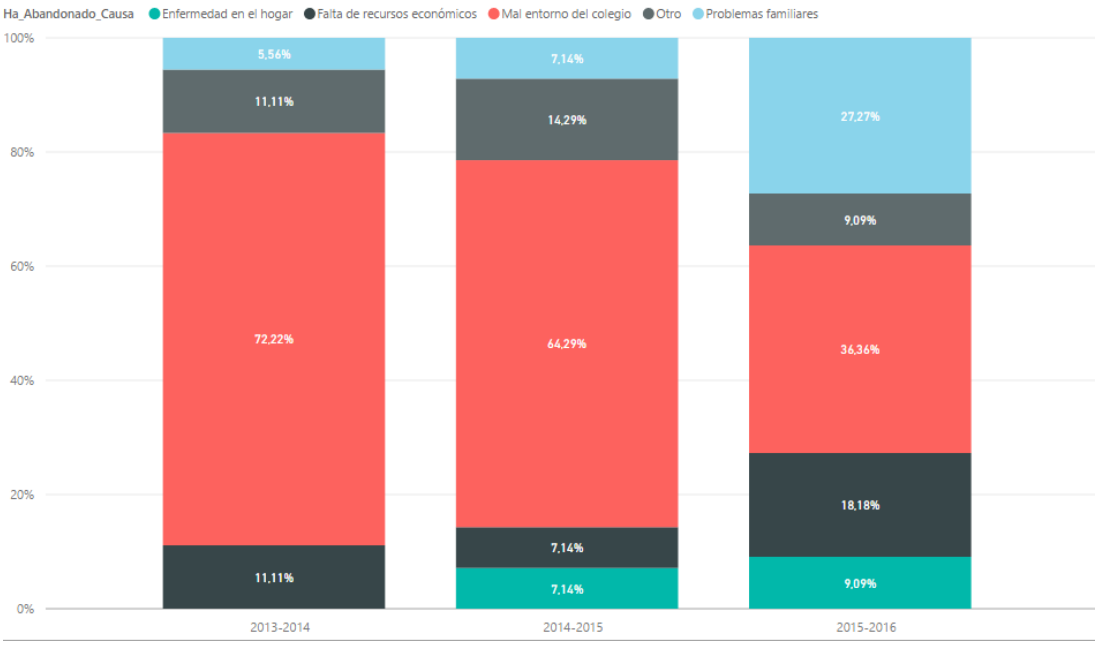


Figura 9. Causas de abandono de estudiantes que indicaron haber abandonado los estudios alguna vez

Referente a la evaluación del alumno al colegio, se observa en la Figura 10 que la apreciación general es buena con una tasa promedio del 59% y en promedio el 32% lo considera excelente en los tres años de estudio; mientras que en la Figura 11 la apreciación de los docentes en los tres años lectivos tiende a ser excelente por más del 58% de estudiantes y el 35% lo considera bueno; no obstante, como se verifica en la Figura 12 cerca del 70% considera que la infraestructura del colegio es buena y alrededor del 15% la considera regular.

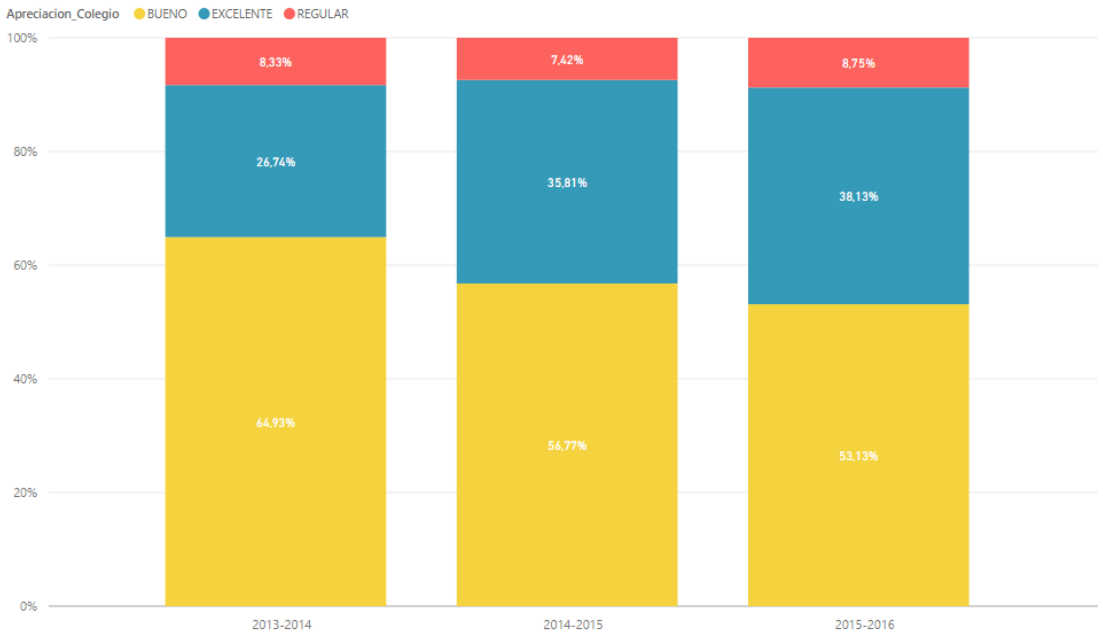


Figura 10. Apreciación general del colegio por años lectivos

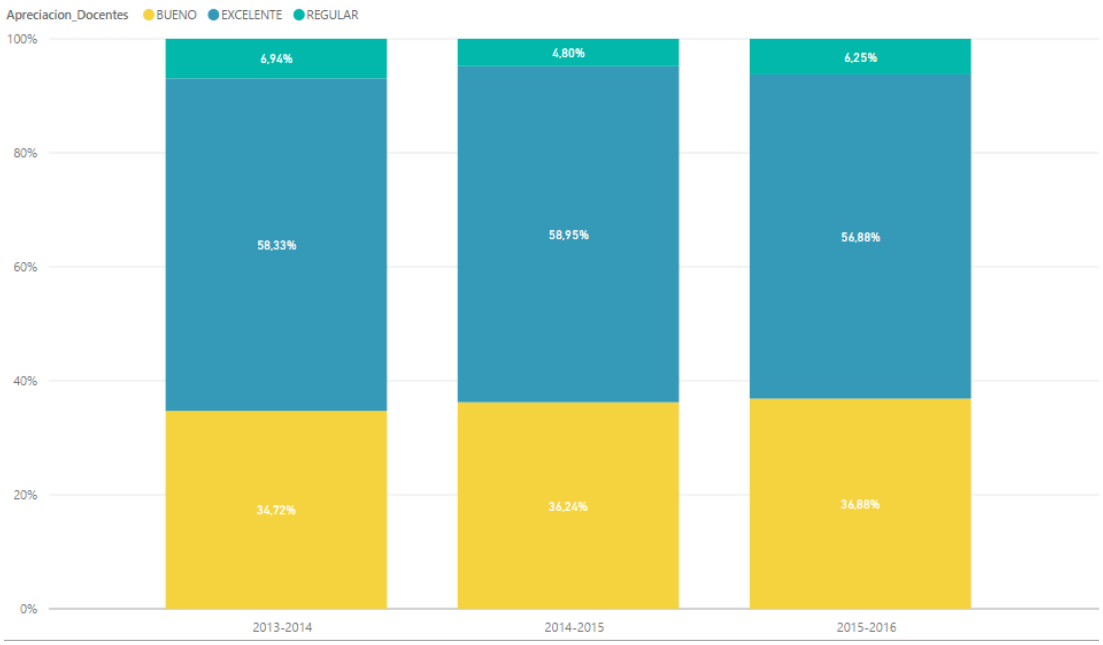


Figura 11. Apreciación de los docentes por años lectivos

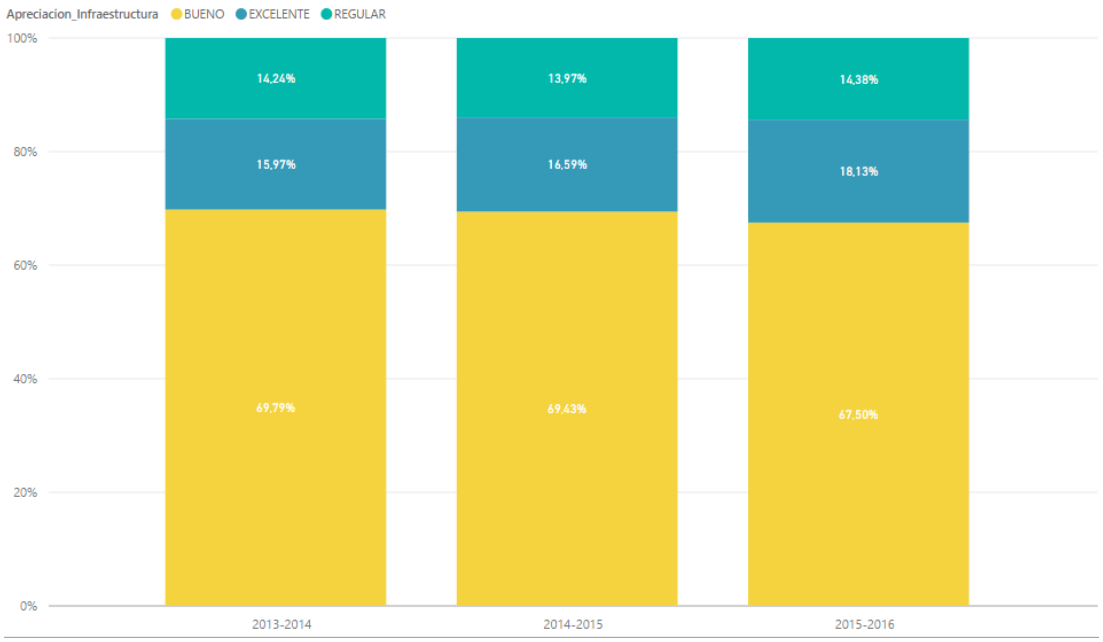


Figura 12. Apreciación de la infraestructura del colegio por años lectivos

3.2.4 Limpieza y construcción de datos

Los primeros pasos de limpieza de datos se dieron en la confirmación o corrección de información obtenida de las encuestas de los alumnos; principalmente en los campos de ocupación del padre y del total de amigos del alumno. En ambos casos fue necesario ubicar al estudiante y corroborar o corregir los datos proporcionados, de todas las encuestas contabilizamos 27 de este tipo; 14 fueron receptadas por email y las restantes fueron encuestas escritas de los alumnos que lamentablemente tenían tachones en las hojas.

Con respecto a la base de datos se detectó anomalías en los datos de asistencias; debido a que se encontraron registros con asistencias mayores a 200 días que es la duración total del año lectivo en la unidad educativa según confirmó el Inspector General. Dado que el número de casos fueron únicamente dos entonces se procedió a corregir la información cotejada con la bitácora del Inspector General. Se pensó en utilizar el total de asistencias en los tres años; sin embargo, fue recomendación del Inspector General realizar únicamente con la asistencia de un Quimestre¸ debido a que existen alumnos que llegan a la institución para el segundo quimestre del año lectivo, esto por traspaso de colegio; debido a esto la obtención de estos datos fue realizada únicamente con la información del primer quimestre para aquellos con fecha de

inscripción hasta diciembre; y con fecha de inscripción a partir de enero se consideró únicamente las asistencias del segundo quimestre, la misma regla se aplicó para las variables del promedio general y promedio de disciplina.

Para la obtención de los estudiantes que abandonaron el colegio en los años lectivos de estudio, la secretaría del colegio proporcionó vía Excel el listado de estos alumnos; ya que mantiene los registros y actas de los mismos en archivo; este Excel se subió como tabla a un respaldo de la base de datos escolástica proporcionada por IT para ser cotejada.

Para obtener la variable de retorno al colegio fue necesario cotejar por código de persona (código único del estudiante en cualquier año de estudio) que estudiante se matriculó el año posterior al de análisis; aquellos que si lo habían hecho se marcaron como "1". Esta variable pudo haberse obtenido de la variable Colegio_Anterior; pero, no fue posible debido a que secretaría notificó que existen casos en los que se marca el colegio anterior como la unidad educativa cuando el alumno finalizó un año en el colegio, luego el alumno no siguió con sus estudios en ninguna institución educativa y volvió en uno o más años nuevamente a seguir con sus estudios.

Por último, de las variables obtenidas por medio de las encuestas; una vez tabuladas en Excel se pasó esta información a una tabla en la base de datos del colegio; aquí básicamente se realizó una depuración de mala digitalización; ya que se contó con el apoyo de un grupo de 3 personas quienes digitalizaron las entrevistas y las enviadas por email.

Del total de columnas se construyen dos variables adicionales; estas son:

1. Mes_Ingreso:

   Descripción: Se deriva de la variable Fecha_Ingreso; obtiene el mes de la fecha de ingreso del alumno.

   Variable Origen: Fecha_Ingreso

   Tipo de dato: String
2. Cat_Edad:

   Descripción: Se categoriza la variable edad.

   Variable Origen: Edad

   Tipo de dato: String

Rango:

- < 14 años
- 14-17 años
- 18-21 años
- 22-25 años
- 26-29 años
- > 29 años

El conjunto de datos de análisis se construye por medio de un script sql que se puede observar en el ANEXO III; el script genera una tabla de salida que será almacenada en Excel para su consumo en R. La Figura 13 representa el conjunto de tablas utilizadas en el script sql, entre ellas se incluye además la tabla que almacena la información recopilada de las encuestas:

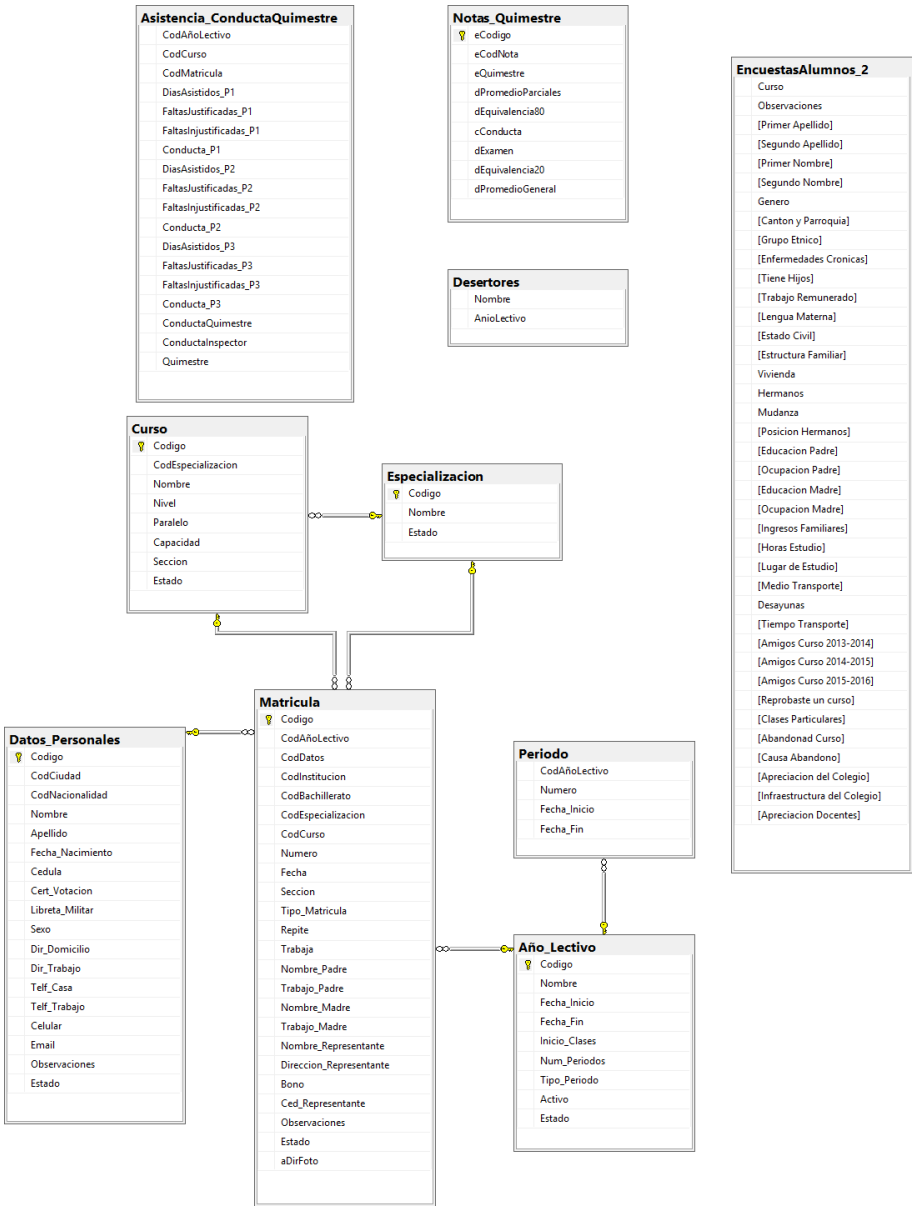


Figura 13. Diagrama de principales tablas de la base de datos escolástica

3.3 Esquematización del modelo

3.3.1 Selección de técnicas de modelado

Del total de variables a utilizarse para el modelamiento se tienen 41 variables categóricas; y, 9 variables numéricas, entre las variables categóricas se encuentra justamente la variable de predicción: Abandono_Curso_Actual.

Dado el objetivo de minería de datos de la unidad educativa descrito en el apartado 1.3 Objetivos, se sigue la línea menciona en la propuesta que consiste en utilizar algoritmos de clasificación con aprendizaje supervisado para el modelo de predicción;

esto “debido a que los casos del conjunto de entrenamiento aparecen etiquetados con la clase a la que corresponden” (Berzal, 2016).

Los modelos seleccionados cumplen con las características de los datos; estos modelos son:

1. Redes neuronales (NNET)

(Aggarwal, 2015) indica que las redes neuronales son una técnica de modelado para la aproximación de funciones teniendo su origen en el perceptrón. De las diversas topologías de redes neuronales se tiene como principal a la red perceptrón multicapa cuyo modelo contiene una capa oculta y es definido en tres capas como se muestra en la Figura 14.

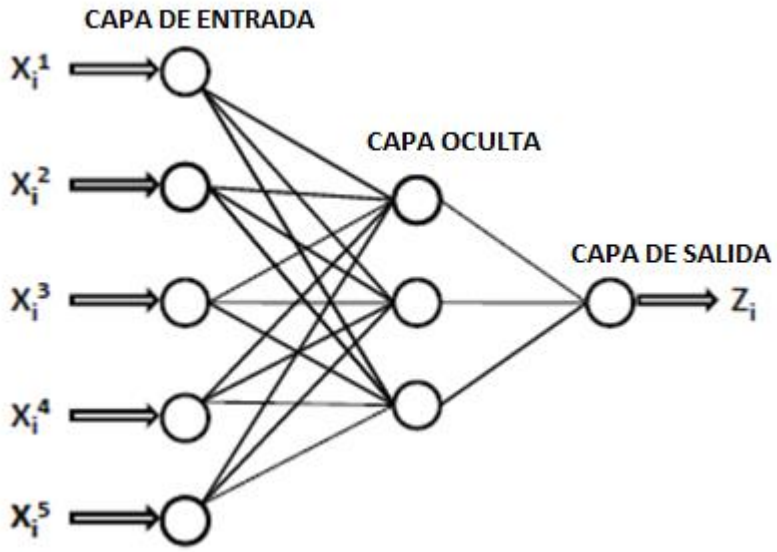


Figura 14. Red neuronal perceptrón multicapas

Fuente: *(Aggarwal, 2015)*

2. Quinlan's C5.0 (C5.0)

Según (Kuhn, Weston, Coulter, & Quinlan, 2015) este modelo extiende los algoritmos de clasificación C4.5 descritos en la publicación de Quinlan de 1992. El modelo puede tomar la forma de un árbol de decisión completo o una colección de reglas.

3. Support Vector Machine (SVM)

El SVM es una máquina de aprendizaje para problemas de clasificación de dos grupos. Según (Cortes & Vapnik, 1995) esta máquina implementa conceptualmente la siguiente idea: “Los vectores de entrada se asignan no linealmente a un espacio de características de muy alta dimensión. En este espacio de características se construye una superficie de decisión lineal. Las propiedades especiales de la superficie de decisión garantizan una alta capacidad de generalización de la máquina de aprendizaje”. (Platoš, 2016) explica que en el SVM las características categóricas pueden ser binarizadas para su uso; y la clase etiqueta se asumen pueden ser del conjunto {-1;1}. En (Gavoyiannis, Vogiatzis, Georgiadis, & Hatziargyriou, 2001) se aclara que la idea principal del SVM es “construir un hiperplano como superficie de decisión de tal manera que se maximice el margen de separación entre ejemplos positivos y negativos. La máquina adquiere esta propiedad deseable siguiendo un enfoque basado en principios enraizado en la teoría del aprendizaje estadístico”. Un ejemplo de la función hiperplano del SVM se presenta en la Figura 15.

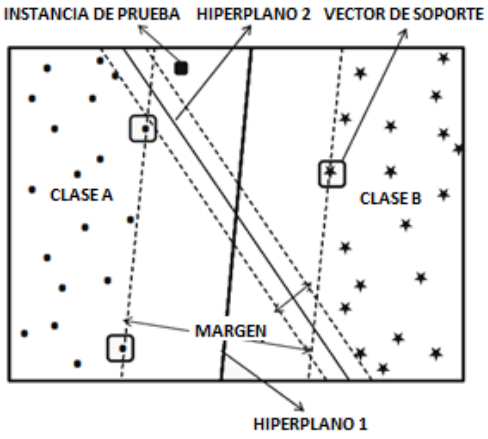


Figura 15. SVM. Caso de separación lineal

Fuente: *(Platoš, 2016)*

4. Bootstrapped Aggregation (Bagging)

Bootstrapped Aggregation o *Bagging* es un método para generar múltiples versiones de un predictor y usarlas para obtener un predictor agregado. En (Breiman L. , Bagging Predictors, 1996) se explica que la agregación promedia las versiones al predecir un resultado numérico (regresión) y hace una pluralidad de votos al predecir una clase (clasificación) como se aprecia en la Figura 16. Las versiones se forman haciendo

réplicas bootstrap (muestreo con reemplazo) del conjunto de aprendizaje y usando estos como nuevos conjuntos de aprendizaje.

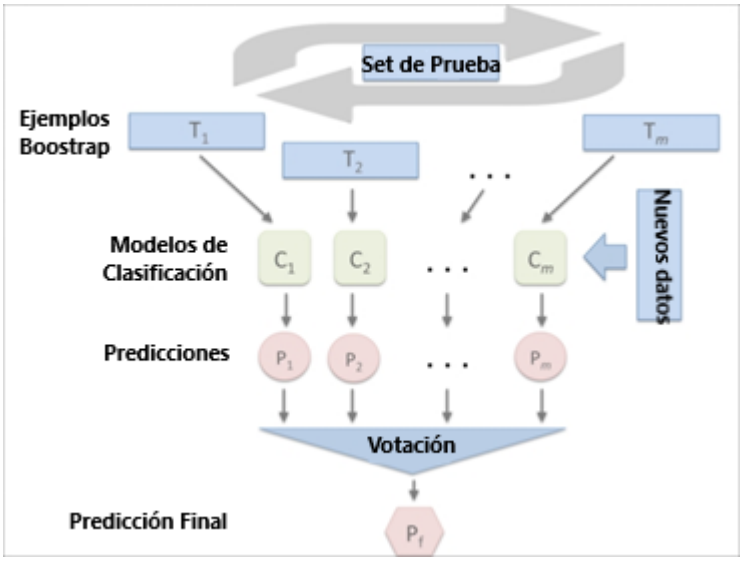


Figura 16. Metodología bagging para un clasificador

Fuente: *(Raschka, 2015)*

5. Random Forest

Random Forest o *Random Forests* como se especifica en (Breiman L. , 2001) son una combinación de predictores de árboles de tal manera que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para todos los árboles en el bosque. Este algoritmo es una mejora sustancial de bagging ya que el error de generalización de un Random Forest depende de la fuerza de los árboles individuales en el bosque y la correlación entre ellos.

6. Adaptative Boosting (AdaBoost)

Según (Zhi-Hua, 2012) "el termino *boosting* hace referencia a una familia de algoritmos que son capaces de convertir aprendices débiles en fuertes". (Aggarwal, 2015) explica que en *boosting* se asocia un peso a cada instancia de entrenamiento; y, los clasificadores son entrenados con el uso de sus pesos; luego cada peso es modificado iterativamente basado en el rendimiento del clasificador. Sus autores en (Freund & Schapire, 1996) indican que AdaBoost: "puede ser utilizado para reducir significativamente el error de cualquier algoritmo de aprendizaje que genera

consistentemente clasificadores cuyo rendimiento es un poco mejor que la suposición aleatoria".

7. Árboles de clasificación: rpart

RPART o *Recursive Partitioning and Regression Trees* es una implementación de la mayoría de funcionalidades publicadas en (Breiman, Friedman, Stone, & Olshen, 1984). En ejemplo simple se muestra en la Figura 17. (Grossmann & Rinderle-Ma, 2015) explica que la definición de este modelo para datos de formación realistas requiere la definición formal de dos estrategias:

1) **Reglas de división:** Una estrategia para cultivar el árbol definiendo en cada nodo que variable debe usarse para dividir junto al umbral para la división.
2) **Reglas de poda:** Una estrategia para podar el árbol con el fin de evitar el *overfitting* de los datos de entrenamiento.

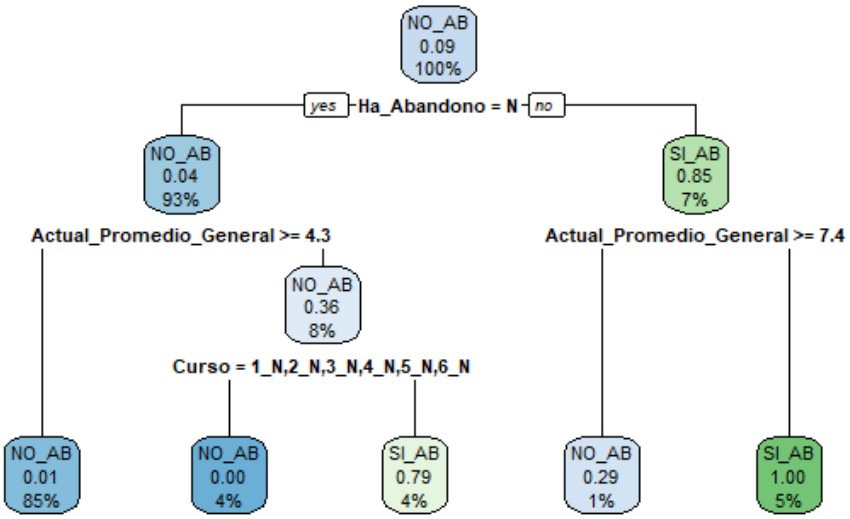


Figura 17. Ejemplo simple de un árbol de decisión

Fuente: (R packages rpart; rpart.plot)

8. Clasificador bayesiano ingenuo (Naive Bayes classifier): naiveBayes

Un clasificador bayesiano ingenuo pertenece a la familia de clasificadores probabilísticos fundamentados en el Teorema de Bayes (Russell & Norvig, 1995). Como se describe en (Lemaire, Christophe, & Bondu, 2015) el algoritmo supone que las variables explicativas son independientes condicionalmente a la variable objetivo.

El desempeño del clasificador bayesiano ingenuo depende de la calidad de la estimación de las distribuciones condicionales univariadas y de una selección eficiente de las variables explicativas informativas. En (Langley, Iba, & Thompson, 1992) se describe que el método almacena un resumen probabilístico para cada clase; este resumen contiene la probabilidad adicional de cada valor de atributo dada la clase, así como la probabilidad (o tasa de base) de la clase. Cuando se le da una instancia de prueba, el clasificador utiliza una función de evaluación para clasificar las clases alternativas en base a sus resúmenes probabilísticos y asigna la instancia a la clase de puntuación más alta.

3.3.2 Generación del modelo

Toda la fase de generación, pruebas y validación de los modelos se lo realizó en el software estadístico R; como se aprecia en la Figura 18 es la herramienta más utilizada por la comunidad científica a la fecha de desarrollo del proyecto de investigación.

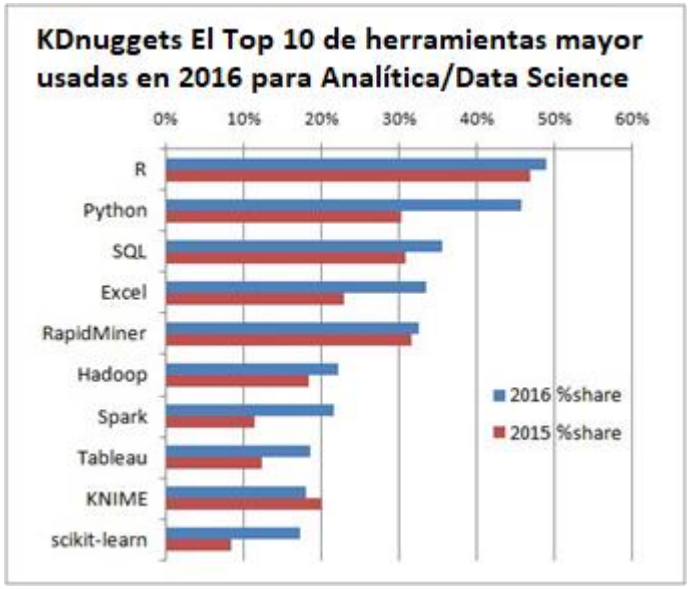


Figura 18. El Top 10 de herramientas mayor usadas en 2016 para Analítica/Data Science

Fuente: *(KDnuggets, 2016)*

Con el fin de corroborar la validación de cada modelo se utilizó una función de R modificada de (TES Global Limited, 2017); esta función ajusta los 8 modelos propuestos en la sección 3.3.1 Selección de técnicas de modelado, el script se encuentra en el ANEXO II.

La función en un primer paso genera un vector con los nombres de las columnas que son redundantes para la predicción, estas columnas se retirarán antes de ejecutar los modelos; estas son:

- Columnas de código: El conjunto de datos contiene además columnas de código y descripción de algunas variables; dado que la columna a utilizarse en la predicción es la que guarda la descripción se retiró aquellas columnas con los códigos. Entre ellas tenemos Codigo_Anterior_Colegio, Cod_Anio_Lectivo, Codigo_Especialidad, Codigo_Curso.
- Anio_Lectivo: Fue una variable principal en la exploración de datos, pero como se puede observar en (Suganya & Narayani, 161-166) esta variable se puede tomar como una variable temporal no se recomienda utilizarla en la predicción actual ya que en cuyo caso los modelos corresponderían de mejor forma a series temporales.
- Estudiante: Variables de referencia estadística;
- Codigo_Matricula: Variables de referencia estadística;
- Ciudadania_Ecuatoriana: Todas la población analizada es ecuatoriana;
- Fecha_Ingreso: Tiene cierta similitud con la variable de año lectivo; pero lo interesante con esta variable es el desglose del mes de ingreso; ya que sin contar con el año de ingreso puede convertirse en una variable categórica para el análisis.
- Edad: Fue transformada en variable categórica.
- Variables objetivo: Dependiendo de la variable que se desee predecir se excluye una variable objetivo u otra (Si es Abandono_Curso_Actual, se excluye la variable Retorno_Al_Colegio; y viceversa).

Seguidamente, se realiza una división del conjunto de datos: entre un 75% para entrenamiento y un 25% para pruebas; para ello utilizamos los datos de los alumnos de los años lectivos 2013-2014 y 2014-2015 como los datos de entrenamiento y los alumnos del año lectivo 2015-2016 como datos de pruebas. Con esto se busca enfatizar una linealidad en los patrones de abandono de forma anual, dado que de ser efectivo alguno de los modelos se comprobaría la existencia de patrones recurrentes entre los años de estudio. Para la predicción de la variable Regreso_Al_Colegio se excluye a los alumnos de Tercero de Bachillerato.

Por último, se entrenan los modelos descritos en la sección 3.3.1 Selección de técnicas de modelado y se realiza la predicción de prueba con los datos del año 2015-2016, una vez realizada la predicción de todos los modelos, se almacenan los resultados, los modelos y los valores de evaluación en una lista que será devuelta al script principal.

El script principal por otro lado, consta de la asignación de una semilla, la definición de los paths de almacenamiento/recuperación de los modelos y de lectura de la fuente de datos (Excel); la preparación de columnas especificado en la sección 3.2.4 Limpieza y construcción de datos; y, por último la ejecución de la función modificada de comparación de modelos. Este script se resume en el siguiente algoritmo, el script en R se encuentra en el ANEXO IV:

1. Generación de semilla para reproducir resultados
2. Establecer la dirección física de almacenamiento y lectura de modelos
3. Establecer la dirección física del archivo consolidado de información
4. Leer la fuente de datos
5. Preparar columnas existentes
6. Crear columnas para análisis
7. Probar modelos para abandono escolar
8. Probar modelos para regreso al colegio
9. Salvar modelos de abandono
10. Salvar modelos de regreso al colegio
11. Recuperación de modelos

3.3.3 Evaluación del modelo

Para las evaluaciones de los modelos utilizaremos la evaluación de matriz de confusión la cual según (Grossmann & Rinderle-Ma, 2015) “Suma los conteos correctos e incorrectos para un cierto conjunto de datos”. Las filas de la matriz corresponden al valor de predicción asignado por el modelo, y las columnas al valor real del conjunto de datos, este método de evaluación como se menciona en (Grossmann & Rinderle-Ma, 2015), es especialmente útil en matrices de dos variables o también llamado problema de clasificación binaria; la Tabla 4 muestra las estadísticas que pueden ser obtenidas de la matriz de confusión:

Tabla 4

Terminología y conexión de matriz de confusión a un problema de clasificación binaria

<table>
<tr><td colspan="2" rowspan="2">Matriz de Confusión</td><td colspan="2">Objetivo / Real</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Positivo</td><td>Negativo</td></tr>
<tr><td rowspan="2">Modelo/ Asignado</td><td>Positivo</td><td>a</td><td>b</td><td>Valor Predictivo Positivo</td><td>a / (a + b)</td></tr>
<tr><td>Negativo</td><td>c</td><td>d</td><td>Valor Predictivo Negativo</td><td>d / (c + d)</td></tr>
<tr><td colspan="2" rowspan="2"></td><td>Sensibilidad</td><td>Especificidad</td><td colspan="2" rowspan="2">Exactitud = (a + d) / (a + b + c + d)</td></tr>
<tr><td>a / (a + c)</td><td>d / (b + d)</td></tr>
</table>

Fuente: (Grossmann & Rinderle-Ma, 2015)

Según el Dr. Saed Sayad cada término en la matriz indica (Sayad, 2017):

- **Precisión:** La proporción del número total de predicciones que eran correctas.
- **Valor predictivo positivo (TPR):** la proporción de casos positivos que se identificaron correctamente.
- **Valor Predictivo Negativo (FPR):** la proporción de casos negativos que fueron correctamente identificados.
- **Sensibilidad o Recuperación:** la proporción de casos positivos reales que están correctamente identificados.
- **Especificidad:** la proporción de casos negativos reales que se identifican correctamente.

Además, se realizará una última evaluación de los modelos por medio de la curva ROC (*Receiver Operating Characteristics*) y del área bajo la curva ROC denominada por sus siglas en inglés AUC (*Area Under the ROC Curve*) (Fawcett, 2006). Este modelo de evaluación utiliza el TPR y el FPR de la matriz de confusión para generar una curva, como se presenta en la Figura 19 los FPR se colocan en el eje X y los TPR

en el eje Y; y, el gráfico resultante muestra la curva ROC correspondiente con cada punto etiquetado por el umbral que lo produce.

El AUC es una porción del área del cuadrado unitario, por ello su valor oscila entre 0 y 1; además, se considera que un buen clasificador tiene un valor AUC mayor a 0,5. Con esta medida se considera que un clasificador tiene mejor rendimiento que otro al obtener un mejor valor AUC.

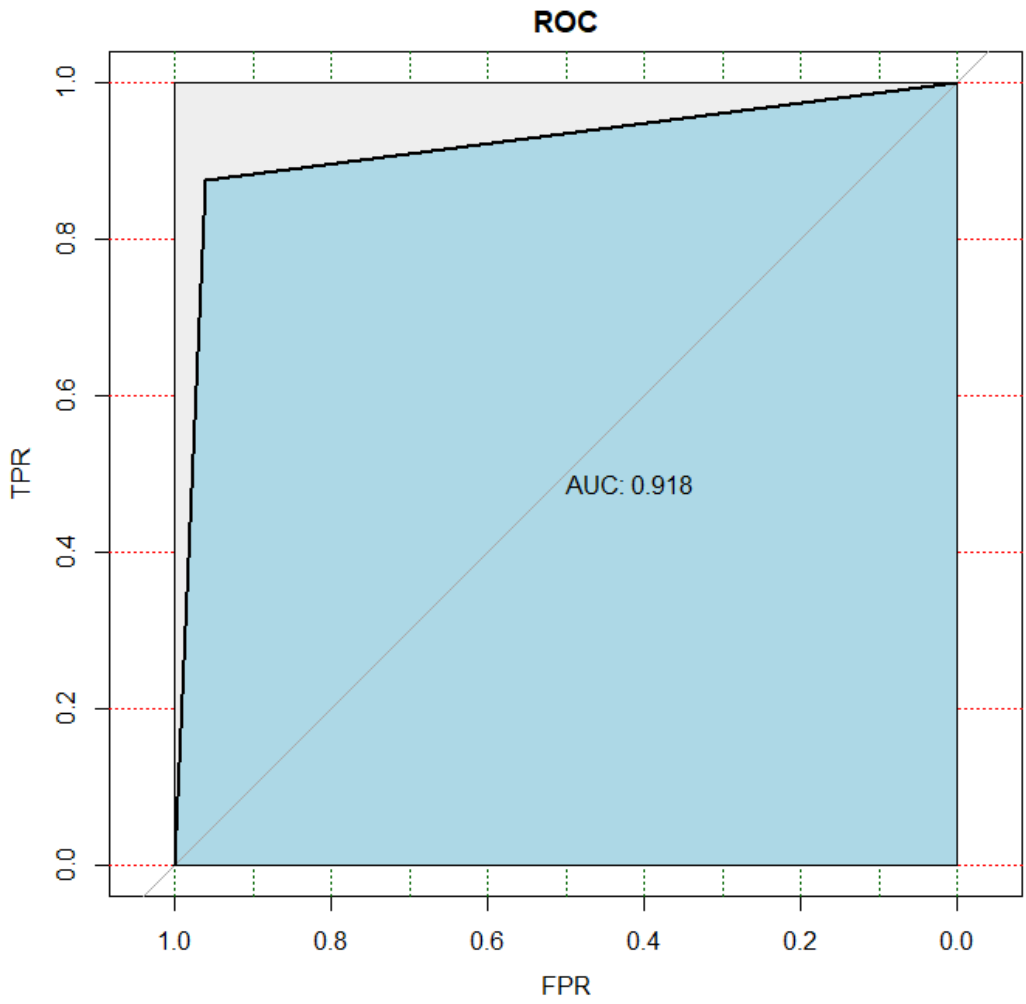


Figura 19. Curva ROC y cálculo AUC

A continuación, desde la Figura 20 hasta la Figura 35, se presenta las estadísticas y matrices de confusión obtenidas desde R para cada modelo evaluado por cada variable objetivo (Abandono_Curso_Actual, Retorno_Al_Colegio):

1. Redes neuronales:

```
> caret::confusionMatrix(Abandono.TestModels$mc$nnet)
Confusion Matrix and Statistics

          Real
Asignado NO_AB SI_AB
   NO_AB   151     1
   SI_AB     1     7

               Accuracy : 0.9875
                 95% CI : (0.9556, 0.9985)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.01218

                  Kappa : 0.8684
 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9934
            Specificity : 0.8750
         Pos Pred Value : 0.9934
         Neg Pred Value : 0.8750
             Prevalence : 0.9500
         Detection Rate : 0.9437
   Detection Prevalence : 0.9500
      Balanced Accuracy : 0.9342

       'Positive' Class : NO_AB
```

Figura 20. Matriz de confusión. Modelo: NNET. Objetivo: Abandono_Curso_Actual

```
> caret::confusionMatrix(Regreso.TestModels$mc$nnet)
Confusion Matrix and Statistics

          Real
Asignado  0  1
       0  9  2
       1 19 81

               Accuracy : 0.8108
                 95% CI : (0.7255, 0.8789)
    No Information Rate : 0.7477
    P-Value [Acc > NIR] : 0.0746182

                  Kappa : 0.3722
 Mcnemar's Test P-Value : 0.0004803

            Sensitivity : 0.32143
            Specificity : 0.97590
         Pos Pred Value : 0.81818
         Neg Pred Value : 0.81000
             Prevalence : 0.25225
         Detection Rate : 0.08108
   Detection Prevalence : 0.09910
      Balanced Accuracy : 0.64867

       'Positive' Class : 0
```

Figura 21. Matriz de confusión. Modelo: NNET. Objetivo: Regreso_Al_Colegio

2. Quinlan's C5.0

```
> caret::confusionMatrix(Abandono.TestModels$mc$C5.0)
Confusion Matrix and Statistics

          Real
Asignado NO_AB SI_AB
   NO_AB   146     1
   SI_AB     6     7

               Accuracy : 0.9562
                 95% CI : (0.9119, 0.9822)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.4494

                  Kappa : 0.6447
 Mcnemar's Test P-Value : 0.1306

            Sensitivity : 0.9605
            Specificity : 0.8750
         Pos Pred Value : 0.9932
         Neg Pred Value : 0.5385
             Prevalence : 0.9500
         Detection Rate : 0.9125
   Detection Prevalence : 0.9187
      Balanced Accuracy : 0.9178

       'Positive' Class : NO_AB
```

Figura 22. Matriz de confusión. Modelo: C5.0. Objetivo: Abandono_Curso_Actual

```
> caret::confusionMatrix(Regreso.TestModels$mc$C5.0)
Confusion Matrix and Statistics

          Real
Asignado  0  1
       0 14 17
       1 14 66

               Accuracy : 0.7207
                 95% CI : (0.6276, 0.8017)
    No Information Rate : 0.7477
    P-Value [Acc > NIR] : 0.7800

                  Kappa : 0.2851
 Mcnemar's Test P-Value : 0.7194

            Sensitivity : 0.5000
            Specificity : 0.7952
         Pos Pred Value : 0.4516
         Neg Pred Value : 0.8250
             Prevalence : 0.2523
         Detection Rate : 0.1261
   Detection Prevalence : 0.2793
      Balanced Accuracy : 0.6476

       'Positive' Class : 0
```

Figura 23. Matriz de confusión. Modelo: C5.0. Objetivo: Regreso_Al_Colegio

3. Support Vector Machine

```
> caret::confusionMatrix(Abandono.TestModels$mc$svm)
Confusion Matrix and Statistics

          Real
Asignado NO_AB SI_AB
   NO_AB   152     4
   SI_AB     0     4

               Accuracy : 0.975
                 95% CI : (0.9372, 0.9931)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.09385

                  Kappa : 0.6552
 Mcnemar's Test P-Value : 0.13361

            Sensitivity : 1.0000
            Specificity : 0.5000
         Pos Pred Value : 0.9744
         Neg Pred Value : 1.0000
             Prevalence : 0.9500
         Detection Rate : 0.9500
   Detection Prevalence : 0.9750
      Balanced Accuracy : 0.7500

       'Positive' Class : NO_AB
```

Figura 24. Matriz de confusión. Modelo: SVM. Objetivo: Abandono_Curso_Actual

```
> caret::confusionMatrix(Regreso.TestModels$mc$svm)
Confusion Matrix and Statistics

          Real
Asignado  0  1
       0  7  1
       1 21 82

               Accuracy : 0.8018
                 95% CI : (0.7154, 0.8714)
    No Information Rate : 0.7477
    P-Value [Acc > NIR] : 0.1128

                  Kappa : 0.3117
 Mcnemar's Test P-Value : 5.104e-05

            Sensitivity : 0.25000
            Specificity : 0.98795
         Pos Pred Value : 0.87500
         Neg Pred Value : 0.79612
             Prevalence : 0.25225
         Detection Rate : 0.06306
   Detection Prevalence : 0.07207
      Balanced Accuracy : 0.61898

       'Positive' Class : 0
```

Figura 25. Matriz de confusión. Modelo: SVM. Objetivo: Regreso_Al_Colegio

4. Bootstrapped Aggregation: Bagging

```
> caret::confusionMatrix(Abandono.TestModels$mc$bagging)
Confusion Matrix and Statistics

          Real
Asignado NO_AB SI_AB
   NO_AB   147     1
   SI_AB     5     7

               Accuracy : 0.9625
                 95% CI : (0.9202, 0.9861)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.3071

                  Kappa : 0.6809
 Mcnemar's Test P-Value : 0.2207

            Sensitivity : 0.9671
            Specificity : 0.8750
         Pos Pred Value : 0.9932
         Neg Pred Value : 0.5833
             Prevalence : 0.9500
         Detection Rate : 0.9187
   Detection Prevalence : 0.9250
      Balanced Accuracy : 0.9211

       'Positive' Class : NO_AB
```

Figura 26. Matriz de confusión. Modelo: Bagging. Objetivo: Abandono_Curso_Actual

```
> caret::confusionMatrix(Regreso.TestModels$mc$bagging)
Confusion Matrix and Statistics

          Real
Asignado  0  1
       0  8 10
       1 20 73

               Accuracy : 0.7297
                 95% CI : (0.6372, 0.8096)
    No Information Rate : 0.7477
    P-Value [Acc > NIR] : 0.7118

                  Kappa : 0.1874
 Mcnemar's Test P-Value : 0.1003

            Sensitivity : 0.28571
            Specificity : 0.87952
         Pos Pred Value : 0.44444
         Neg Pred Value : 0.78495
             Prevalence : 0.25225
         Detection Rate : 0.07207
   Detection Prevalence : 0.16216
      Balanced Accuracy : 0.58262

       'Positive' Class : 0
```

Figura 27. Matriz de confusión. Modelo: Bagging. Objetivo: Regreso_Al_Colegio

5. Random Forest

```
> caret::confusionMatrix(Abandono.TestModels$mc$randomForest)
Confusion Matrix and Statistics

          Real
Asignado NO_AB SI_AB
   NO_AB   150     1
   SI_AB     2     7

               Accuracy : 0.9812
                 95% CI : (0.9462, 0.9961)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.03882

                  Kappa : 0.8137
 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9868
            Specificity : 0.8750
         Pos Pred Value : 0.9934
         Neg Pred Value : 0.7778
             Prevalence : 0.9500
         Detection Rate : 0.9375
   Detection Prevalence : 0.9437
      Balanced Accuracy : 0.9309

       'Positive' Class : NO_AB
```

Figura 28. Matriz de confusión. Modelo: Random Forest. Objetivo: Abandono_Curso_Actual

```
> caret::confusionMatrix(Regreso.TestModels$mc$randomForest)
Confusion Matrix and Statistics

          Real
Asignado  0  1
       0  8  3
       1 20 80

               Accuracy : 0.7928
                 95% CI : (0.7055, 0.8639)
    No Information Rate : 0.7477
    P-Value [Acc > NIR] : 0.1627439

                  Kappa : 0.3124
 Mcnemar's Test P-Value : 0.0008492

            Sensitivity : 0.28571
            Specificity : 0.96386
         Pos Pred Value : 0.72727
         Neg Pred Value : 0.80000
             Prevalence : 0.25225
         Detection Rate : 0.07207
   Detection Prevalence : 0.09910
      Balanced Accuracy : 0.62478

       'Positive' Class : 0
```

Figura 29. Matriz de confusión. Modelo: Random Forest. Objetivo: Regreso_Al_Colegio

6. Adaptative Boosting

```
> caret::confusionMatrix(Abandono.TestModels$mc$boosting)
Confusion Matrix and Statistics

          Real
Asignado NO_AB SI_AB
   NO_AB   151     1
   SI_AB     1     7

               Accuracy : 0.9875
                 95% CI : (0.9556, 0.9985)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.01218

                  Kappa : 0.8684
 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9934
            Specificity : 0.8750
         Pos Pred Value : 0.9934
         Neg Pred Value : 0.8750
             Prevalence : 0.9500
         Detection Rate : 0.9437
   Detection Prevalence : 0.9500
      Balanced Accuracy : 0.9342

       'Positive' Class : NO_AB
```

Figura 30. Matriz de confusión. Modelo: Boosting. Objetivo: Abandono_Curso_Actual

```
> caret::confusionMatrix(Regreso.TestModels$mc$boosting)
Confusion Matrix and Statistics

          Real
Asignado  0  1
       0 10 12
       1 18 71

               Accuracy : 0.7297
                 95% CI : (0.6372, 0.8096)
    No Information Rate : 0.7477
    P-Value [Acc > NIR] : 0.7118

                  Kappa : 0.2288
 Mcnemar's Test P-Value : 0.3613

            Sensitivity : 0.35714
            Specificity : 0.85542
         Pos Pred Value : 0.45455
         Neg Pred Value : 0.79775
             Prevalence : 0.25225
         Detection Rate : 0.09009
   Detection Prevalence : 0.19820
      Balanced Accuracy : 0.60628

       'Positive' Class : 0
```

Figura 31. Matriz de confusión. Modelo: Boosting. Objetivo: Regreso_Al_Colegio

7. Árboles de clasificación: rpart

```
> caret::confusionMatrix(Abandono.TestModels$mc$rpart)
Confusion Matrix and Statistics

          Real
Asignado NO_AB SI_AB
   NO_AB   150     1
   SI_AB     2     7

               Accuracy : 0.9812
                 95% CI : (0.9462, 0.9961)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.03882

                  Kappa : 0.8137
 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9868
            Specificity : 0.8750
         Pos Pred Value : 0.9934
         Neg Pred Value : 0.7778
             Prevalence : 0.9500
         Detection Rate : 0.9375
   Detection Prevalence : 0.9437
      Balanced Accuracy : 0.9309

       'Positive' Class : NO_AB
```

Figura 32. Matriz de confusión. Modelo: rpart. Objetivo: Abandono_Curso_Actual

```
> caret::confusionMatrix(Regreso.TestModels$mc$rpart)
Confusion Matrix and Statistics

          Real
Asignado  0  1
       0 15 10
       1 13 73

               Accuracy : 0.7928
                 95% CI : (0.7055, 0.8639)
    No Information Rate : 0.7477
    P-Value [Acc > NIR] : 0.1627

                  Kappa : 0.4305
 Mcnemar's Test P-Value : 0.6767

            Sensitivity : 0.5357
            Specificity : 0.8795
         Pos Pred Value : 0.6000
         Neg Pred Value : 0.8488
             Prevalence : 0.2523
         Detection Rate : 0.1351
   Detection Prevalence : 0.2252
      Balanced Accuracy : 0.7076

       'Positive' Class : 0
```

Figura 33. Matriz de confusión. Modelo: rpart. Objetivo: Regreso_Al_Colegio

8. Clasificador bayesiano ingenuo (Naive Bayes classifier): naiveBayes

```
> caret::confusionMatrix(Abandono.TestModels$mc$naiveBayes)
Confusion Matrix and Statistics

          Real
Asignado NO_AB SI_AB
   NO_AB   149     1
   SI_AB     3     7

               Accuracy : 0.975
                 95% CI : (0.9372, 0.9931)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.09385

                  Kappa : 0.7647
 Mcnemar's Test P-Value : 0.61708

            Sensitivity : 0.9803
            Specificity : 0.8750
         Pos Pred Value : 0.9933
         Neg Pred Value : 0.7000
             Prevalence : 0.9500
         Detection Rate : 0.9313
   Detection Prevalence : 0.9375
      Balanced Accuracy : 0.9276

       'Positive' Class : NO_AB
```

Figura 34. Matriz de confusión. Modelo: Naive Bayes. Objetivo: Abandono_Curso_Actual

```
> caret::confusionMatrix(Regreso.TestModels$mc$naiveBayes)
Confusion Matrix and Statistics

          Real
Asignado  0  1
       0 12 12
       1 16 71

               Accuracy : 0.7477
                 95% CI : (0.6565, 0.8254)
    No Information Rate : 0.7477
    P-Value [Acc > NIR] : 0.5506

                  Kappa : 0.2981
 Mcnemar's Test P-Value : 0.5708

            Sensitivity : 0.4286
            Specificity : 0.8554
         Pos Pred Value : 0.5000
         Neg Pred Value : 0.8161
             Prevalence : 0.2523
         Detection Rate : 0.1081
   Detection Prevalence : 0.2162
      Balanced Accuracy : 0.6420

       'Positive' Class : 0
```

Figura 35. Matriz de confusión. Modelo: Naive Bayes. Objetivo: Regreso_Al_Colegio

CAPÍTULO 4.

EXPERIMENTACIÓN Y RESULTADOS

Aquí se presentan los resultados de evaluación y clasificación de cada modelo para las dos variables objetivo enfatizando los modelos con mejor rendimiento en base a las pruebas de evaluación. Para finalizar, se presenta la interpretación de los resultados con las recomendaciones del investigador para prevenir el abandono de los estudiantes.

4.1 Experimentación con datos de la unidad educativa Lenin School

4.1.1 Variable Objetivo: Abandono_Curso_Actual

Se analiza primero los resultados con respecto a la variable Abandono_Curso_Actual. Conforme los resultados de evaluación para cada modelo expuestas en la sección 3.3.3 Evaluación del modelo se presenta la Tabla 5 que ordena los modelos en orden descendente por el porcentaje de exactitud de sus respectivas matrices de confusión:

Tabla 5

Medidas de exactitud de los modelos para variable Abandono_Curso_Actual

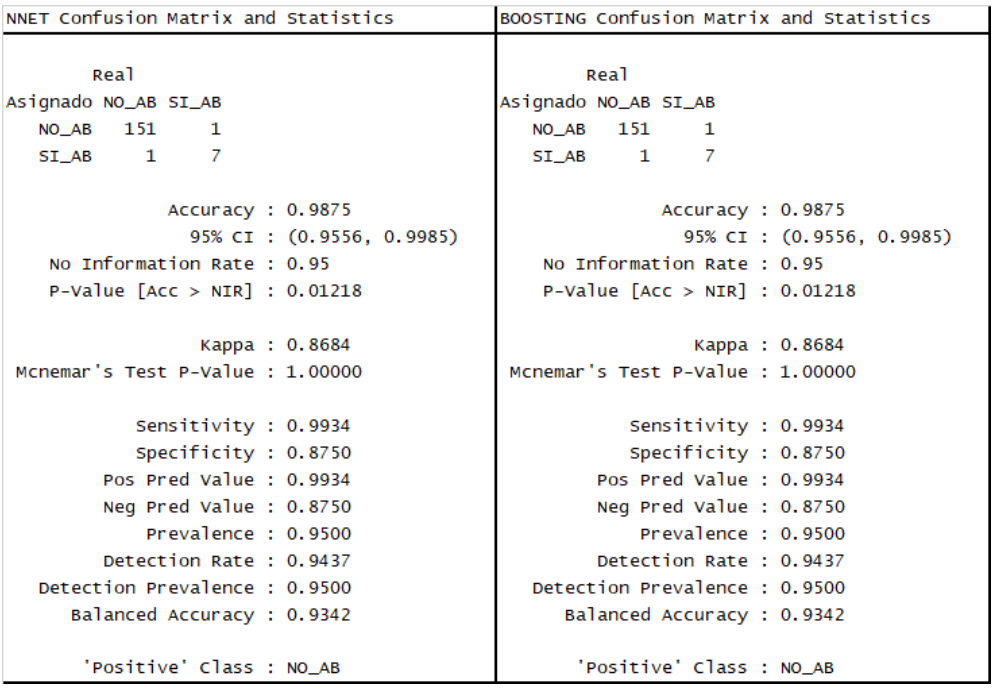
Modelos	Negativos *	Positivos **	Falsos Negativos	Falsos Positivos	% Exactitud	Sensitividad	Especificidad
NNET	7	151	1	1	98.750	0.9934	0.8750
Boosting	7	151	1	1	98.750	0.9934	0.8750
Random Forest	7	150	1	2	98.125	0.9868	0.8750
rpart	7	150	1	2	98.125	0.9868	0.8750
SVM	4	152	4	0	97.500	1.0000	0.5000
Naive Bayes	7	149	1	3	97.500	0.9671	0.8750
Bagging	7	147	1	5	96.250	0.9803	0.8750
C5.0	7	146	1	6	95.625	0.9605	0.8750

* La clase predictiva es SI_AB
** La clase predictiva es NO_AB
Fuente: Elaboración propia. Resumen de matrices de confusión de modelos predictivos

La Tabla 5 evidencia que los modelos con mejor exactitud de clasificación para la variable Abandono_Curso_Actual son: NNET y Boosting. El modelo con mejor

sensitividad es SVM, pero también es el modelo con la especificidad más baja; mientras que los modelos NNET y Boosting siguen manteniendo mucha fortaleza en ambas medidas.

En la Figura 36 se presenta una comparativa entre ambos modelos ganadores; dado que la matriz de confusión es similar entre ambos modelos; las estadísticas obtenidas son similares para ambos casos.



NNET Confusion Matrix and Statistics

```
          Real
Asignado NO_AB SI_AB
   NO_AB   151     1
   SI_AB     1     7

               Accuracy : 0.9875
                 95% CI : (0.9556, 0.9985)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.01218

                  Kappa : 0.8684
 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9934
            Specificity : 0.8750
         Pos Pred Value : 0.9934
         Neg Pred Value : 0.8750
             Prevalence : 0.9500
         Detection Rate : 0.9437
   Detection Prevalence : 0.9500
      Balanced Accuracy : 0.9342

       'Positive' Class : NO_AB
```

BOOSTING Confusion Matrix and Statistics

```
          Real
Asignado NO_AB SI_AB
   NO_AB   151     1
   SI_AB     1     7

               Accuracy : 0.9875
                 95% CI : (0.9556, 0.9985)
    No Information Rate : 0.95
    P-Value [Acc > NIR] : 0.01218

                  Kappa : 0.8684
 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9934
            Specificity : 0.8750
         Pos Pred Value : 0.9934
         Neg Pred Value : 0.8750
             Prevalence : 0.9500
         Detection Rate : 0.9437
   Detection Prevalence : 0.9500
      Balanced Accuracy : 0.9342

       'Positive' Class : NO_AB
```

Figura 36. Comparación de estadísticas de matriz de confusión entre modelos NNET y BOOSTING

En la Tabla 6 se presentan los valores de AUC para cada uno de los modelos; nuevamente para los modelos NNET y Boosting sus valores AUC coinciden en 0.9342 siendo los mejores puntuados nuevamente.

Tabla 6

Valor AUC de los modelos para la variable Abandono_Curso_Actual

Modelos	AUC
NNET	0.9342
Boosting	0.9342
Random Forest	0.9309
rpart	0.9309
Naive Bayes	0.9276

Bagging	0.9211
C5.0	0.9178
SVM	0.75

Fuente: Elaboración propia. Resumen ROC de modelos predictivos

Como evidencian las Tablas 5 y 6 los 8 modelos propuestos son efectivos en la predicción de la variable Abandono_Curso_Actual.

Al analizar el modelo NNET, se detecta que el algoritmo binarizó las variables de tipo factor como se muestra en la Figura 37 para la variable curso; cada una de las variables binarizadas generaron más columnas al conjunto de datos según los niveles o rangos de cada variable.

Codigo_Matricula	Estudiante	Curso
1004	167	1_D
944	227	1_D
1081	623	2_D
1015	526	2_D
1017	585	3_D
1012	427	3_D
980	458	1_D

Codigo_Matricula	Estudiante	Curso1_D	Curso2_D	Curso3_D
1004	167	1	0	0
944	227	1	0	0
1081	623	0	1	0
1015	526	0	1	0
1017	585	0	0	1
1012	427	0	0	1
980	458	1	0	0

Figura 37. Ejemplo de binarización de la variable Curso

El modelo NNET es un perceptrón multicapa similar al de la Figura 14; tiene 154 entradas que son resultado de las variables categóricas binarizadas y las variables numéricas; se configuró por defecto 5 nodos en la capa oculta y 1 salida como se muestra en la Figura 38.

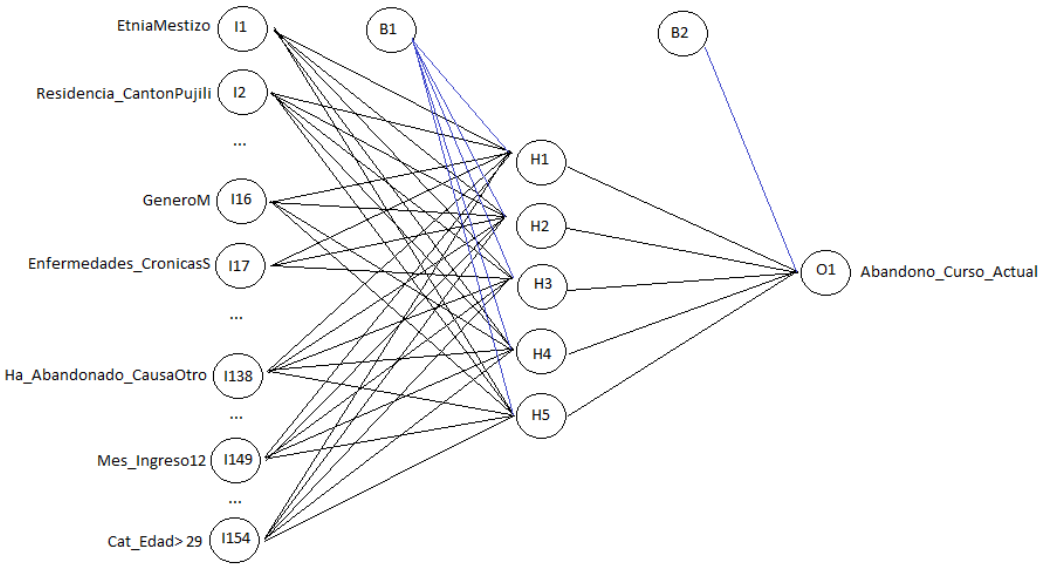


Figura 38. Gráfica resumida de NNET para Abandono_Curso_Actual

Al buscar las variables importantes para la predicción, se localiza que estas pueden ser halladas por medio de los pesos que el algoritmo NNET aplica sobre las diferentes capas. El paquete NeuralNetworkTools expone dos funciones que permiten obtener la importancia relativa de las variables explicativas para las variables de respuesta, estas funciones son Olden y Garson y se basan en (Olden, Joy, & Death, 2004) y (Garson, 1991) respectivamente. Según la bibliografía utilizaremos la función Olden al tener mejores ventajas sobre Garson como se menciona en (Olden, Joy, & Death, 2004).

La Figura 39 se forma filtrando la importancia de las variables explicativas mayores al valor 3.5 y menores a -3.5; teniendo en cuenta la dirección de los pesos (positivo, negativo) las variables presentadas se consideran las más relevantes en la predicción de la variable Abandono_Curso_Actual.

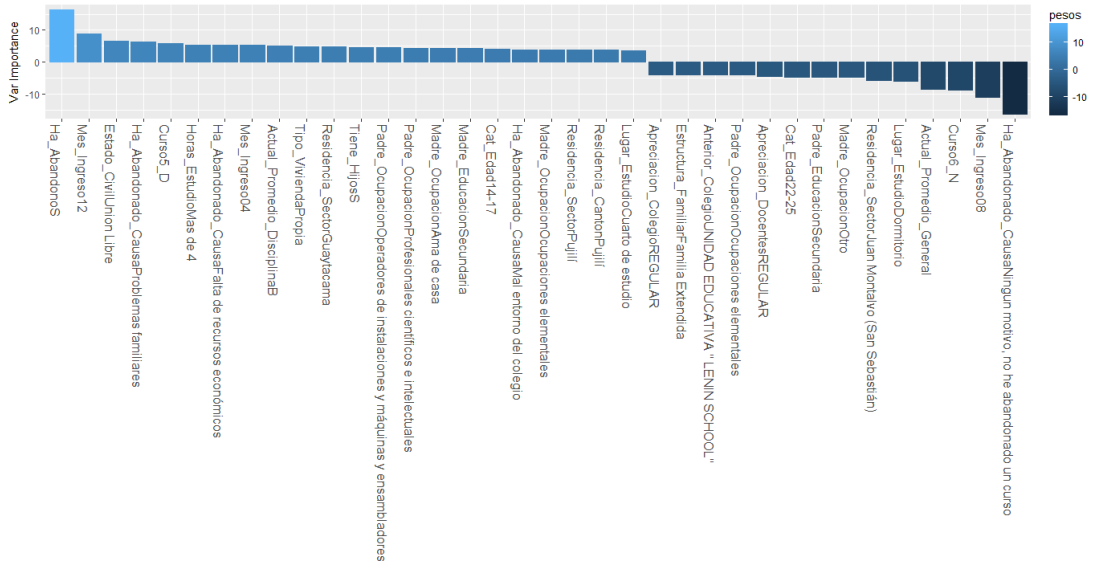


Figura 39. Resumen de Importancia de las variables usadas en el modelo NNET para Abandono_Curso_Actual

Al analizar el modelo Boosting, por otro lado, tenemos que este generó 100 árboles de decisión; el peso máximo para las instancias es de 2.1876 el cual pertenece al árbol 70 que se lo puede observar en la Figura 40; y, el peso mínimo es de 1.0735 y corresponde al árbol 76. A diferencia del modelo NNET no se binarizaron las variables; por lo que el conjunto de datos sigue siendo de 46 variables.

Al mismo tiempo, el modelo también genera automáticamente un vector de los valores de importancia relativa de cada variable en la tarea de clasificación; que como

se menciona en (Alfaro, Gámez, & García, 2013) esta medida tiene en cuenta la ganancia del índice de Gini cada variable en un árbol y el peso de ese árbol; es así como en la Figura 41 se muestra estos valores en orden descendente.

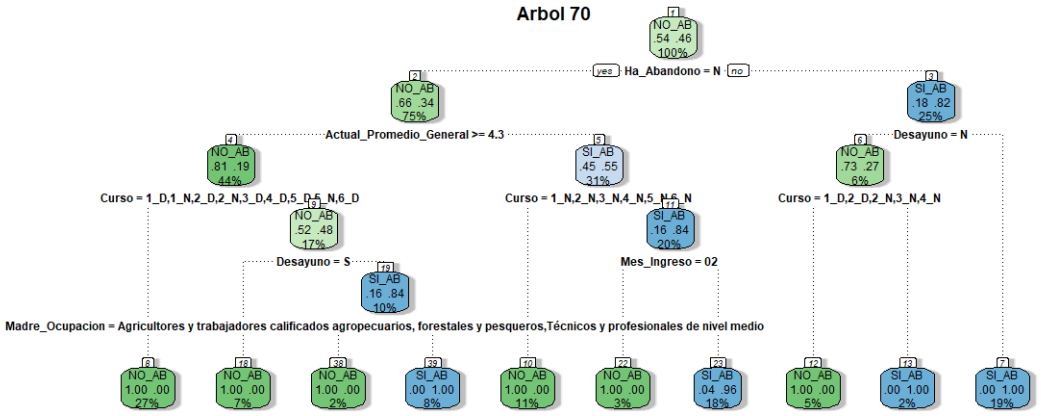


Figura 40. Boosting Instancia de Predicción 70 para Abandono_Curso_Actual

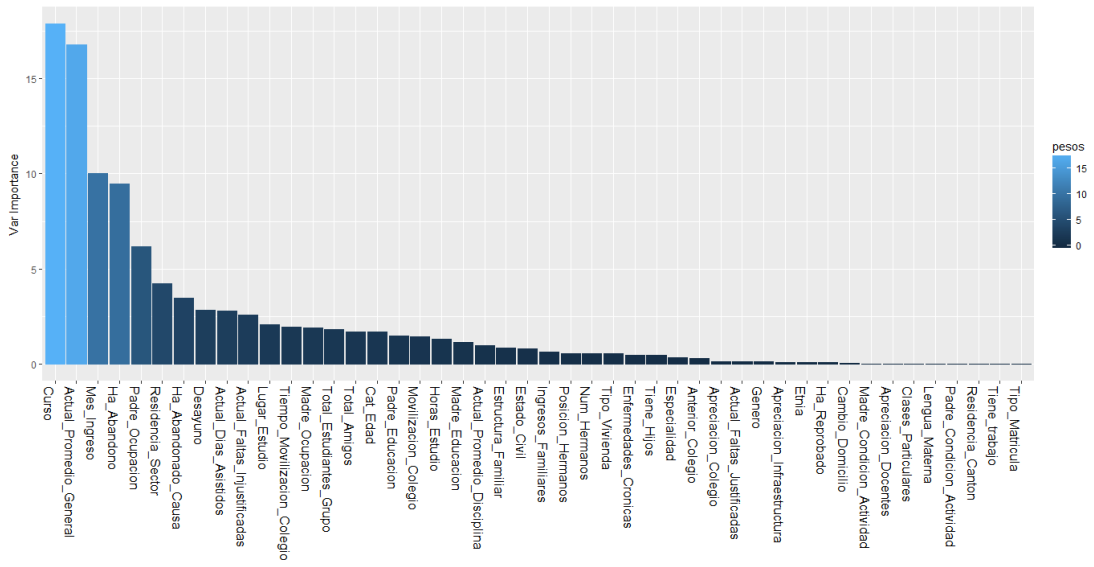


Figura 41. Resumen de Importancia de las variables usadas en el modelo Boosting para Abandono_Curso_Actual

4.1.2 Variable Objetivo: Regreso_Al_Colegio

Los resultados con respecto a la variable Regreso_Al_Colegio conforme los resultados de evaluación para cada modelo descrito en la sección 3.3.3 Evaluación del

modelo se resumen en la Tabla 7 que ordena los modelos en orden descendente por el porcentaje de exactitud de sus respectivas matrices de confusión:

Tabla 7

Medidas de exactitud de los modelos para la variable Regreso_Al_Colegio

Modelos	Negativos *	Positivos **	Falsos Negativos	Falsos Positivos	% Exactitud	Sensitividad	Especificidad
NNET	81	9	2	19	81.010	0.3214	0.9759
SVM	82	7	1	21	80.180	0.2500	0.9880
rpart	73	15	10	13	79.279	0.5357	0.8795
Random Forest	78	8	5	20	79.279	0.2857	0.9636
Naive Bayes	71	12	12	16	74.775	0.4286	0.8554
Bagging	75	10	8	18	72.973	0.2857	0.8795
Boosting	72	9	11	19	72.973	0.3571	0.8554
C5.0	66	14	17	14	72.072	0.5000	0.7952

*La clase predictiva es 1
** La clase predictiva es 0
Fuente: Elaboración propia. Resumen de matrices de confusión de modelos predictivos

La Tabla 7 evidencia que los modelos con mejor exactitud de clasificación para la variable Regreso_Al_Colegio son: NNET y SVM. Sin embargo, al analizar la medida de sensitividad observamos que el modelo con mejor puntaje es rpart; mientras que el modelo con mejor especificidad sigue siendo SVM. Asimismo, se obtuvo la medida AUC la cual se resume en la Tabla 8; los mejores modelos rpart y C5.0 sus valores AUC son en esta prueba los mejores puntuados con 0.7076 y 0.6476 respectivamente.

Tabla 8

Valor AUC de los modelos para la variable Regreso_Al_Colegio

Modelos	AUC
rpart	0.7076
NNET	0.6487
C5.0	0.6476
Naive Bayes	0.6420

Random Forest	0.6248
SVM	0.6190
Boosting	0.6063
Bagging	0.5826

Fuente: Elaboración propia. Resumen ROC de modelos predictivos

Dado los valores AUC en la Tabla 8 se evidencia que el modelo con mejor rendimiento de clasificación es rpart y NNET.

Como evidencian las Tablas 7 y 8 los 8 modelos propuestos son efectivos en la predicción de la variable Regreso_Al_Colegio.

Para el análisis se opta por los modelos NNET y rpart dado que ambos son los que poseen mejores valores de evaluación en las pruebas de matriz de confusión y de AUC respectivamente.

Al analizar el modelo NNET, se detecta que el algoritmo binarizó las variables de tipo factor como se muestra en la Figura 37 para la variable curso; cada una de las variables binarizadas generaron más columnas al conjunto de datos según los niveles o rangos de cada variable.

El modelo NNET es un perceptrón multicapa similar al de la Figura 14; tiene 154 entradas que son resultado de las variables categóricas binarizadas y las variables numéricas; se configuró por defecto 5 nodos en la capa oculta y 1 salida como se muestra en la Figura 42.

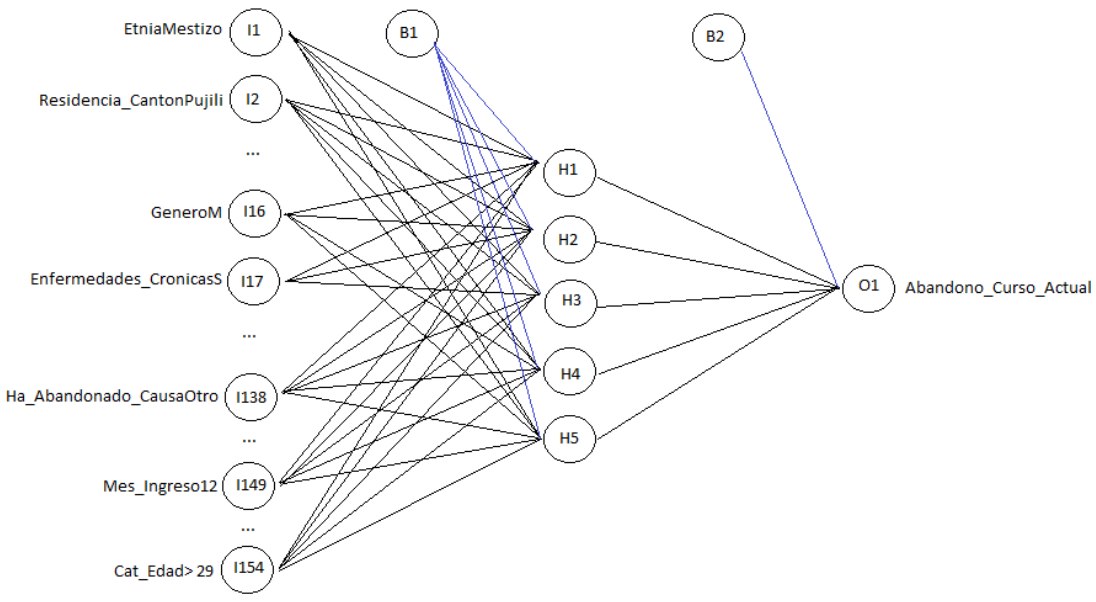


Figura 42. Gráfica resumida de NNET para Regreso_Al_Colegio

En la Figura 43 se presentan las variables explicativas utilizadas por el modelo NNET ordenadas por su peso de importancia, esta gráfica se forma filtrando la importancia de las variables explicativas mayores al valor 3.5 y menores a -3.5; teniendo en cuenta la dirección de los pesos (positivo, negativo) las variables presentadas se consideran las más relevantes en la predicción de la variable Regreso_Al_Colegio.

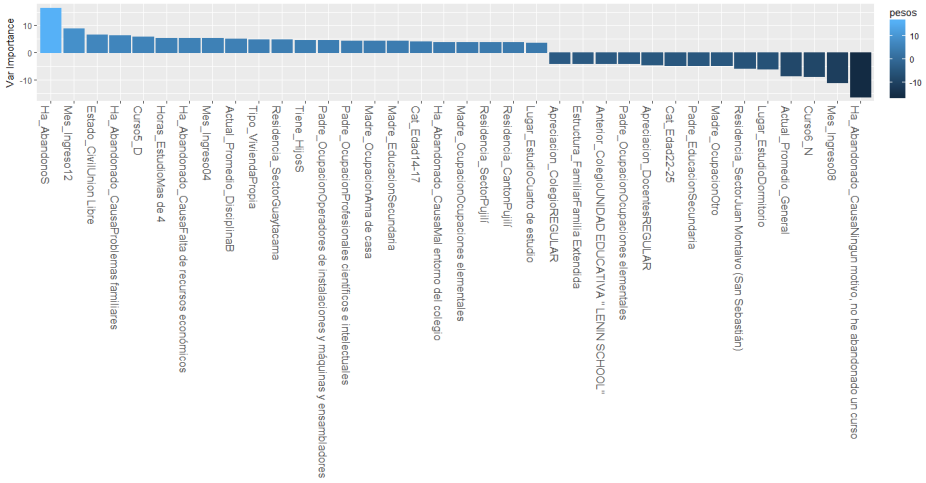


Figura 43. Resumen de Importancia de las variables usadas en el modelo NNET para Regreso_Al_Colegio

Al analizar el modelo rpart a diferencia del modelo NNET no se binarizaron las variables; por lo que el conjunto de datos sigue siendo de 46 variables. Además, el modelo también genera automáticamente un vector de los valores de importancia relativa de cada variable en la tarea de clasificación; es así como en la Figura 44 se muestra estos valores en orden descendente. Por otro lado, tenemos que este generó un árbol de decisión de 10 niveles; y, se lo puede observar en la Figura 45.

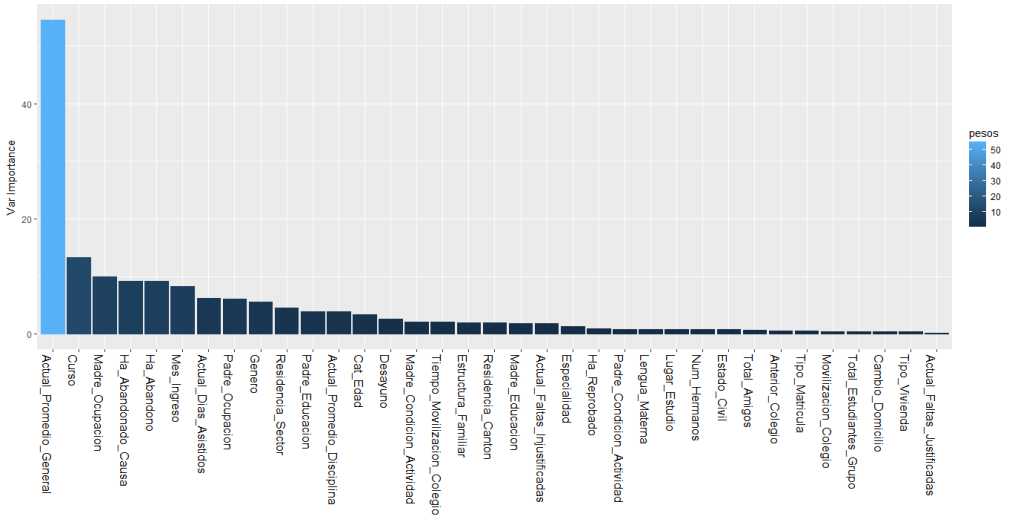


Figura 44. Resumen de Importancia de las variables usadas en el modelo RPART para Regreso_Al_Colegio

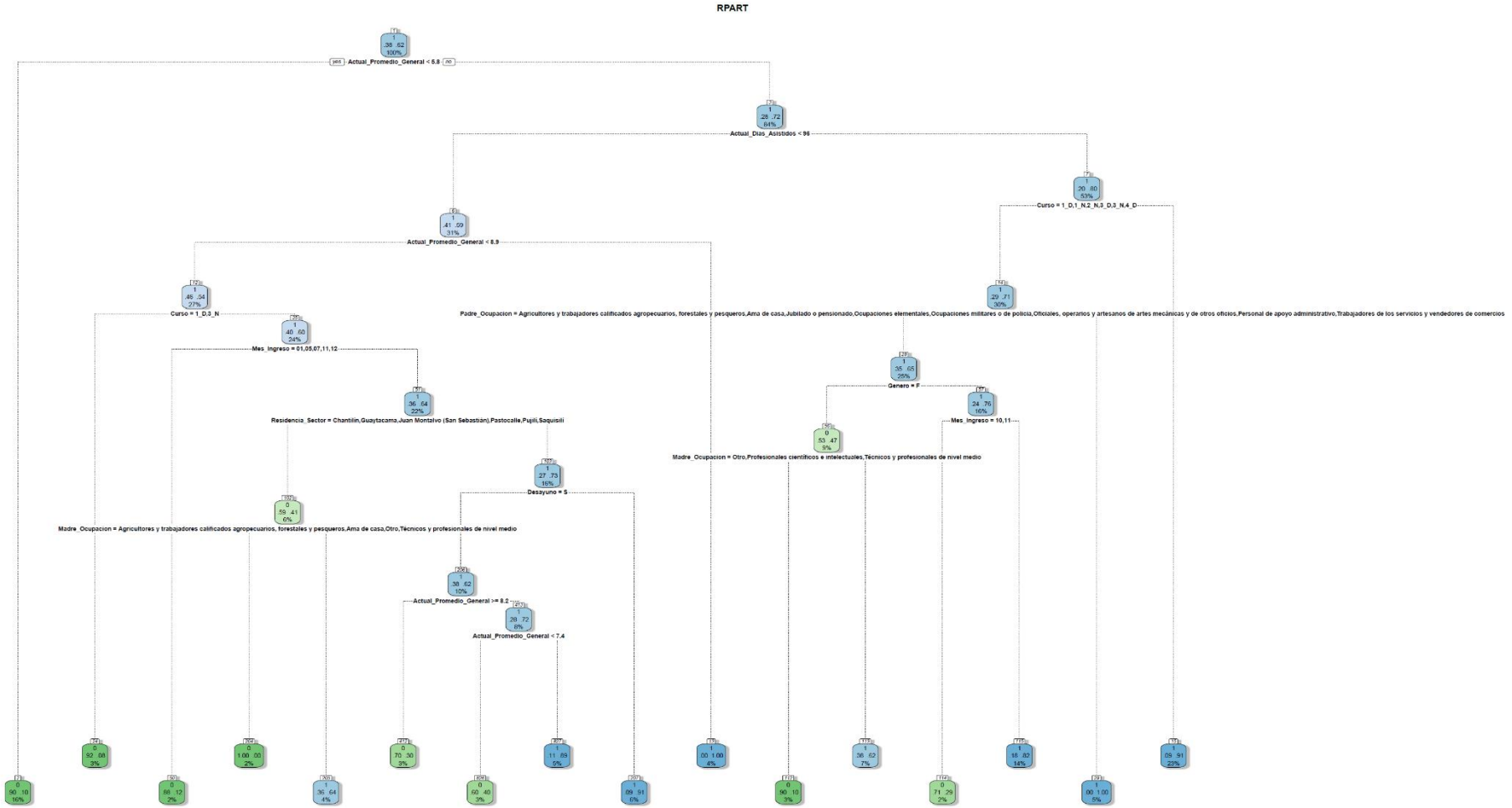


Figura 45. Árbol de decisión del modelo RPART para Regreso_Al_Colegio

4.2 Interpretación y exposición de resultados

A continuación, y según la metodología CRISP-DM, se describen los hallazgos resultantes del proceso de predicción. Los resultados forman parte del informe entregado a los directivos de la unidad educativa.

4.2.1 Variable Objetivo Abandono_Curso_Actual

Con respecto al indicador del abandono escolar en la unidad educativa es evidente un descenso a nivel global en los tres años de análisis. Sin embargo, para el año 2015-2016 este indicador se estableció en el 5% a nivel colegio y el valor máximo es de 13.04% para el curso de Primero de Bachillerato sección Diurno en este año lectivo; sin tener en cuenta la sección (Diurna, Nocturna) se ubica un 9.30% para los cursos de Segundos de Bachillerato; dando como resultado que el indicador aún se encuentra sobre la base establecida en el objetivo 4.3 planteado en el Plan Nacional del Buen Vivir 2013-2017 el cual busca reducir este índice al 3%.

Como parte del análisis exploratorio de los datos de los estudiantes que abandonaron el colegio en los tres años de análisis se tiene que el 74.55% de alumnos están en edades entre 15-18 años; el 72.73% de ellos pertenecen a familias nucleares, el 92.73% residen en del cantón Latacunga. El 49.09% provienen de otros colegios y el 36.36% provienen de la misma unidad educativa. Los estudiantes tienen un promedio de asistencia de 54 días en un quimestre; con respecto al promedio académico se evidencia notas bajas, pero en su mayoría los alumnos se retiran sin notas asentadas. Todos los estudiantes tienen matricula ordinaria, el principal lugar de estudio es el dormitorio y el 24% de estudiantes no desayuna antes de ir al colegio. Sobre la apreciación del colegio más de la mitad considera que el colegio es bueno, con profesores excelentes y con buena infraestructura.

De los modelos utilizados para realizar la predicción se describen dos con el mejor rendimiento los cuales son NNET y Boosting con un 98.75% de exactitud para ambos modelos; de las pruebas de evaluación encontramos que ambos modelos son capaces de predecir la variable de abandono en un 87% de exactitud para los casos de SI_AB y un 99% de exactitud para los casos de NO_AB. De los algoritmos se desprende las variables importantes al momento de evaluar el abandono escolar en la unidad educativa (ver Tabla 9):

Tabla 9

Top 10 de variables explicativas para el abandono escolar

Variables	Pesos
Curso	17,871
Actual_Promedio_General	16,788
Mes_Ingreso	10,011
Ha_Abandono	9,469
Padre_Ocupacion	6,155
Residencia_Sector	4,230
Ha_Abandonado_Causa	3,468
Desayuno	2,822
Actual_Dias_Asistidos	2,803
Actual_Faltas_Injustificadas	2,567

Fuente: Elaboración propia. Resultados del modelo Boosting para la variable Abandono_Curso_Actual

Dado que el modelo Boosting genera árboles de decisión se presenta además el árbol con mejor peso en la predicción en la Figura 46.

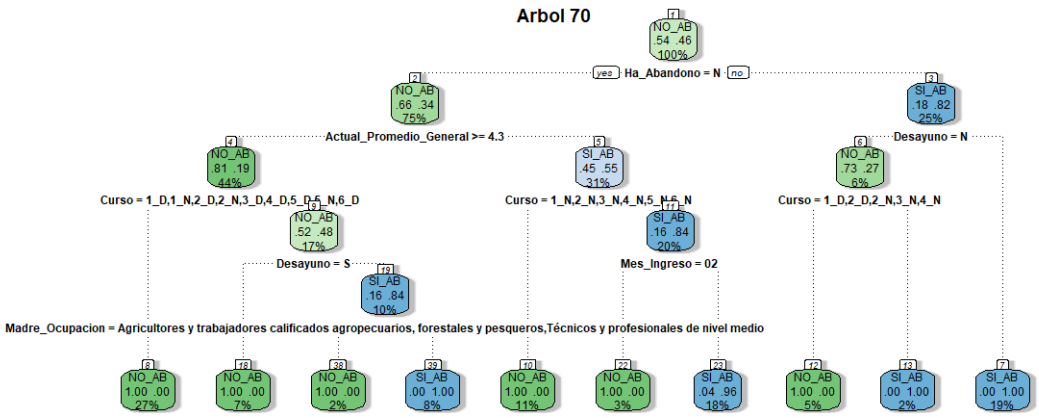


Figura 46. Árbol de decisión con mayor peso para la predicción de abandono escolar

Con los datos analizados y los modelos generados se pone a consideración de las autoridades de la unidad educativa las siguientes estrategias para mitigar el abandono escolar en la institución sin discriminación de alumnos:

1) Generar una base de datos demográfica de los alumnos basado en las variables descritas en este estudio; las variables pueden ser obtenidas en la matrícula del alumno y enviadas a una base de datos separada de la base escolástica si así se lo desea.
2) Generar un proceso automático que realice la predicción de los alumnos inscritos una vez por quimestre.
3) Sin perder el seguimiento que el Departamento de Orientación Vocacional realiza sobre todos los alumnos se recomienda dar seguimiento prioritario a los alumnos que el clasificador marque como posibles detectores.
4) Realizar una vez al año el ajusto del modelo predictivo con datos de al menos tres años lectivos anteriores al nuevo periodo.

Las siguientes estrategias son tomadas del análisis comparativo entre los porcentajes dados en el análisis exploratorio, el árbol de decisión en la Figura 46 y la importancia de las variables explicativas descritas en la Tabla 9. Con el criterio conjunto del Inspector General y al psicólogo se considera que una acción directa sobre las variables explicativas puede inducir en un cambio de decisión del abandono escolar en la unidad educativa. De forma inmediata se recomienda tener a consideración las siguientes estrategias:

5) Realizar juntas de medio quimestre con los docentes para detectar bajo rendimiento académico y desmotivación del alumno; en especial con los cursos de Primero y Segundo de bachillerato dar un seguimiento mensual.
6) Crear un plan de ambientación a los alumnos con traspaso de colegio a mediados de año que incluya una intervención previa con el psicólogo y los padres de familia.
7) Informar al psicólogo de la unidad educativa sobre los alumnos con antecedentes de abandono para realizar el seguimiento debido según la causa de abandono.
8) Impartir charlas de concientización a los padres de familia sobre “La importancia de un desayuno nutritivo en los alumnos”.

4.2.2 Variable Objetivo Regreso_Al_Colegio

Con respecto al indicador de retención de alumnos para la unidad educativa y sin considerar a los alumnos de Tercer año de Bachillerato se destaca que el índice de

alumnos que no regresan a la unidad educativa supera fácilmente el 30% del alumnado. El período más bajo de retención es el de 2014-2015 con el 60% de retención; esto puede deberse al fenómeno natural de reactivación del volcán Cotopaxi presentado en agosto de 2015; sin embargo, dado que para los años lectivos 2013-2014 y 2015-2016 tienen un 63% y 75% de retención respectivamente se considera importante analizar las variables que puedan predecir la anomalía de este indicador.

Como parte del análisis exploratorio de los datos de los estudiantes que no volvieron al colegio en los tres años de análisis se especifica que los cursos con mayor porcentaje son los cursos de Octavo y Décimo año de educación básica con un 44.90% y 42.70% respectivamente. El 58.03% de alumnos están en edades entre 14-17 años, el 55.17% son de género Masculino; el 75.86% de ellos pertenecen a familias nucleares y el 54.02% tiene un promedio de ingresos menor a 700 dólares. El 91.38% residen en el cantón Latacunga y el 84.48% tiene vivienda propia. El 43.10% del alumnado provienen de la misma unidad educativa y el 33.91% provienen de otros colegios; los estudiantes tienen una media de asistencia de 58 días en un quimestre; con respecto a las notas se tiene una media de 5.7 y únicamente el 7.47% tiene un promedio de disciplina de A en el quimestre. Sobre la apreciación del colegio más de la mitad considera que el colegio es bueno, con profesores excelentes y con buena infraestructura.

De los modelos utilizados para realizar la predicción se describen dos ganadores los cuales son NNET con 81.08% de exactitud en la matriz de confusión y rpart con 0.7076 en la medida AUC; el algoritmo con mejor predicción para el caso de no regreso (Cero) es rpart con un 53% de exactitud frente a un 32% del modelo nnet; mientras que para el caso de si regreso (Uno) se denota el 97% de exactitud del modelo nnet en contraste con el 87% del modelo rpart. De los algoritmos se desprende las variables importantes (Ver Tabla 10) al momento de evaluar el abandono escolar en la unidad educativa:

Tabla 10

Top 10 de variables explicativas para la tasa de retención escolar

Variables	Pesos
Actual_Promedio_General	54,507
Curso	13,299
Madre_Ocupacion	9,933
Ha_Abandonado_Causa	9,194
Ha_Abandono	9,194
Mes_Ingreso	8,276
Actual_Dias_Asistidos	6,177
Padre_Ocupacion	6,062
Genero	5,567
Residencia_Sector	4,574

Fuente: Elaboración propia. Resultados del modelo rpart para la variable Regreso_Al_Colegio

Además, se presenta el árbol de decisión generado por el algoritmo rpart en la Figura 47.

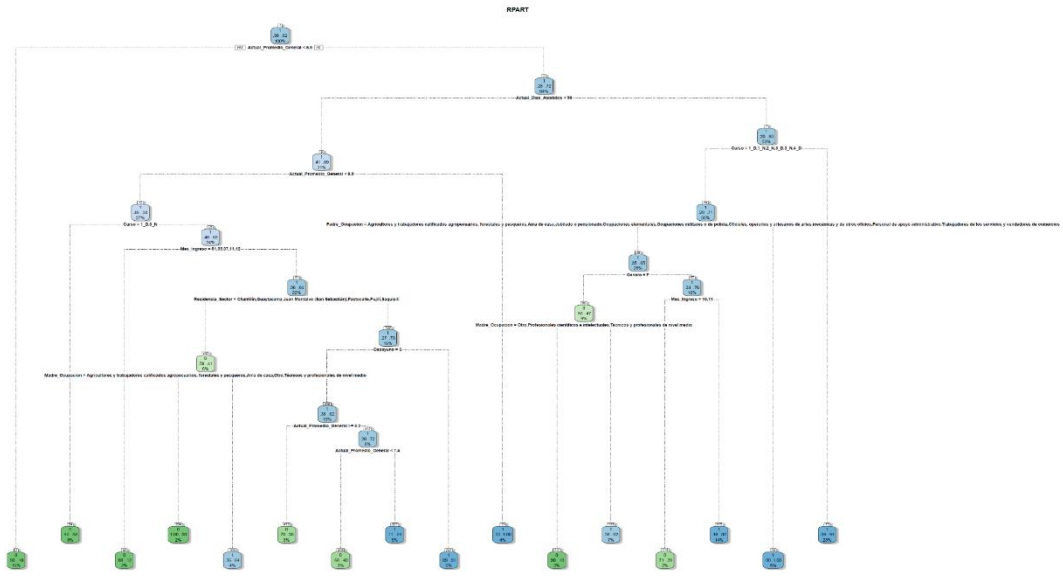


Figura 47. Árbol de decisión del algoritmo rpart para la predicción de la retención de alumnos

Con los datos analizados y los modelos generados se pone a consideración de las autoridades de la unidad educativa las siguientes estrategias para vigorizar la retención escolar en la institución; algunos de estas estrategias forman parte de las recomendaciones realizadas para la atenuación del abandono escolar:

1) Generar una base de datos demográfica de los alumnos basado en las variables descritas en este estudio; las variables pueden ser obtenidas en la matrícula del alumno y enviadas a una base de datos separada de la base escolástica.
2) Generar un proceso automático que realice la predicción de los alumnos inscritos una vez por quimestre.
3) Sin perder el seguimiento que el Departamento de Orientación Vocacional realiza sobre todos los alumnos se recomienda generar un plan de acción sobre los alumnos que el modelo prediga que no regresarán el año siguiente.
4) Realizar una vez al año el ajusto del modelo predictivo con datos de al menos tres años lectivos anteriores al nuevo periodo.

Las siguientes estrategias son tomadas del análisis comparativo entre los porcentajes dados en el análisis exploratorio, el árbol de decisión en la Figura 47 y la importancia de las variables explicativas descritas en la Tabla 10. Con el criterio conjunto del Inspector General y al psicólogo se considera que una acción directa sobre ciertas variables explicativas puede inducir en un cambio de decisión en el retorno del alumno a la unidad educativa. De forma inmediata se recomienda tener a consideración las siguientes estrategias:

1) Elaborar y ejecutar un plan de mejora académica para aquellos alumnos con notas inferiores a 5.8 en el primer quimestre; detectando las materias con menor puntaje y realizando un seguimiento adecuado para cada alumno.
2) Elaborar un sistema de registro de asistencia que emita alertas de alumnos con asistencias menores a 5 días después de los 90 días escolares del quimestre. Una vez detectados el grupo de alumno y luego de una sesión previa de trabajo con el inspector general y el psicólogo, el departamento de Orientación Vocacional deberá elaborar y ejecutar un plan de trabajo con estos alumnos.

a. De ser posible permitir que el sistema de alertas envíe notificaciones vía SMS a los padres de familia o representantes del alumno al detectar una falta en el día de clases.

3) Informar al psicólogo de la unidad educativa sobre los alumnos con antecedentes de abandono para realizar el seguimiento debido según la causa de abandono.

CAPÍTULO 5.

CONCLUSIONES Y TRABAJOS FUTUROS

Se exponen las principales conclusiones del trabajo de investigación junto con los trabajos futuros que pueden desarrollarse como continuación de esta tesis.

5.1 Conclusiones

1. Se acepta la Hipótesis **H0** planteada.
2. Se verifica que el uso de técnicas de clasificación de EDM son un método eficaz para la detección temprana de abandono escolar en la unidad educativa Lenin School en base a los porcentajes de exactitud de predicción de los modelos utilizados.
3. El uso de la metodología CRISP-DM permitió orientar la investigación hacia las necesidades del negocio y evidenciar los mejores modelos para cada necesidad; es por ello que se escoge para cada variable de predicción dos modelos con mejores puntuaciones en las evaluaciones.
4. De los modelos entrenados y probados para la detección temprana de abandono escolar se comprueba que los mejores clasificadores corresponden a los árboles de decisión y a las redes neuronales según los rendimientos de cada uno, además y en contraste con estudios internacionales mencionados en este trabajo también estos modelos son especificados como los mejores.
5. La exhaustiva investigación sobre los modelos de datos utilizados en estudios previos permitió generar un conjunto de datos robusto para la detección temprana de estudiantes que abandonen la unidad educativa “Lenin School”. Sin embargo, es necesario que la unidad educativa genere una base de datos depurada y actualizada que permita mantener una trazabilidad adecuada para la implementación de los modelos.
6. El uso y contraste de las técnicas de validación: matriz de confusión y la curva ROC permitieron evaluar a los modelos desde dos puntos de vista basados en la exactitud general de predicción y en los verdaderos positivos y negativos generados por los modelos.
7. Al usar el mismo conjunto de datos para la variable objetivo de la tasa de retención estudiantil se comprueba que los modelos tienen un buen

rendimiento pero que podría mejorar con otro conjunto de datos más acorde a la explicación de esta variable.

8. Para la predicción del abandono escolar y de la tasa de retención estudiantil se evidencia que el modelo con mejor exactitud corresponde a las redes neuronales.
9. Se podría prevenir eficazmente tanto el abandono escolar como la tasa de retención al evidenciar problemas académicos de forma temprana, en cada caso el enfoque debe orientarse a los cursos con mayor índice de abandono o menor tasa de retención.
10. El uso de herramientas líderes en el mercado permitió un fácil desarrollo de este proyecto; el lenguaje R permitió generar de manera rápida los algoritmos de predicción a su vez que la herramienta Power BI proporcionó un reporte (Anexo V) que por su dinamismo brindó valor agregado a la unidad educativa.
11. No se evidencia una acción directa de la reactivación del volcán Cotopaxi sobre la tasa de abandono escolar en la unidad educativa; sin embargo, para la variable de la tasa de retención existe una certeza empírica en la institución de que este indicador bajó debido a este fenómeno, según la estadística el año lectivo 2014-2015 fue el de menor valor de los tres años analizados, pero se requiere mayor indagación para determinar al fenómeno natural como la causa primaria.

5.2 Trabajos Futuros

Como continuación de este trabajo de tesis existen varias líneas de investigación que quedan abiertas y en las que es posible seguir trabajando, también se pueden ejecutar algunos desarrollos específicos para mejorar los modelos propuestos:

1. Se evidenció que los clasificadores tuvieron un buen rendimiento sobre la variable de retención de alumnos; sin embargo, se considera que la ampliación de esta temática a partir de esta investigación puede mejorar el rendimiento de los clasificadores respecto a la tasa de retención.
2. Replicar este estudio sobre el abandono escolar en una nueva institución educativa; con el fin de evidenciar si los patrones encontrados en este estudio puedan ser generalizados a nivel zonal o incluso nacional.

3. Generar perfiles psicológicos de los estudiantes que abandonaron los estudios junto con las razones de abandono para generar un modelo de clasificación de razones de abandono con técnicas de EDM.
4. Se propone generar una base de datos histórica de los alumnos de una institución educativa para desarrollar un sistema de detección automática de abandono estudiantil que permita la detección en tiempo real de los alumnos que se matriculen en un nuevo año lectivo.
5. Diseñar, desarrollar e implementar un modelo de *forecast* de matriculación de estudiantes para las instituciones educativas basado en los datos históricos y dividido por curso, sección y especialización.
6. Realizar un análisis sobre la afectación en la tasa de retención que se generó debido a la reactivación del volcán Cotopaxi en las instituciones educativas que se encuentran en zona de riesgo. Esto debido a que durante las entrevistas no se evidenció la influencia directa de este fenómeno natural sobre la tasa de retención.
7. El profesor Stephen Lamb citado en este trabajo indica que en otros países se considera en abandono escolar a los jóvenes que superen la mayoría de edad sin haber obtenido un título de escuela secundaria, con esta premisa se propone realizar un estudio de investigación sobre el rumbo que toman los alumnos que abandonan la secundaria en un año lectivo predeterminado, a fin de conocer si estos alumnos finalizan el colegio en algún momento antes de la mayoría de edad. Un convenio con alguna institución educativa o relacionada, permitiría el éxito de la realización de este trabajo propuesto.
8. Diseñar, desarrollar e implementar tableros de control educacionales que permitan visualizar indicadores de rendimiento basados en el PNBV y que permita visualizar la evolución histórica de los mismos.
9. Utilizando los datos de notas parciales por materia e históricos de los alumnos de una institución educativa generar un modelo de detección de alumnos con dificultades en el proceso de aprendizaje.

CAPÍTULO 6.

BIBLIOGRAFÍA

Abu-Oda, G. S., & El-Halees, A. M. (2015). Data Mining In Higher Education : University Student Dropout Case Study. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5*, 15-27.

Aggarwal, C. C. (2015). *Data Mining: The Textbook.* Nueva York: Springer International Publishing Switzerland .

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215*, 487-499.

Ale, M. J. (2015). Data Mining - Revision [Grabado por D. Gallardo]. La Plata, La Plata, Argentina.

Alfaro, E., Gámez, M., & García, N. (2013). adabag: An R Package for Classification with Boosting and Bagging. *Journal of Statistical Software, 54(2)*, 1-35.

Alpar, P., & Schulz, M. (2016). Self-Service Business Intelligence. *Bus Inf Syst Eng 58(2)*, 151-155.

Altman, S. N. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician. 46 (3)*, 175-185.

Amat Rodrigo, J. (Septiembre de 2016). *Análisis discriminante lineal (LDA) y Análisis discriminante cuadrático (QDA).* Obtenido de JoaquinAmatRodrigo/Estadistica-con-R: https://github.com/JoaquinAmatRodrigo/Estadistica-con-R/blob/master/28_Linear_Discriminant_Analysis_LDA_y_Quadratic_Discriminant_Analysis_QDA.pdf

Antamba-Chacua, L. (2015). Desagregación Socio-Demográfica De La Tasa De Abandono A Primer Año Del Bachillerato. *Contexto, Análisis de indicadores educativos Vol. 1*, 15-29.

Antamba-Chacua, L., & Quituisaca-Samaniego, L. (2015). *Caracterización de la tasa neta de asistencia a bachillerato.* Quito: Ministerio de Educación del Ecuador.

Antonenko, P. D., Toy, S., & Niederhauser, D. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development, 60(3),* 383-398.

Armstrong, R., Lewandowski, G., & Loftis, L. (2017). BI Expert's Perspective: Setting Up to Sell Your Analytics. *Business Intelligence Journal, 22(2)*.

Atamba Chacua, L. (2015). *Estadística Educativa, Reporte de Indicadores.* Quito: Ministerio de Educación del Ecuador.

Bacher, J., Tamesberger, D., Leitgöb, H., & Lankmayer, T. (2014). *Not in Education, Employment or Training: Causes, Characteristics of NEET-affected Youth and Exit Strategies in Austria.* Linz, Austria: Institut für Sozial- und Wirtschaftswissenschaften.

Baker, R. (in press). *Data Mining for Education.* Oxford, UK: To appear in McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition).

Baker, R. S., Corbett, A. T., & Wagner, A. (2006). Human classification of low-fidelity replays of student actions. *M. Ikeda, K. Ashlay, & T. Chan (Eds.), Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 29-35.

Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *JEDM - Journal of Educational Data Mining, 1(1),* 5-18.

Bernhardt, V. L. (2005). Data Tools for School Improvement. *Educational Leadership. 62(5),* 66-69.

Berzal, F. (16 de 12 de 2016). *Departamento de Ciencias de la Computación e I.A.* Obtenido de Universidad de Granada Web Site: http://elvex.ugr.es/decsai/intelligent/slides/dm/D4%20Classification.pdf

Bimonte, S., Saulot, L., Journaux, L., & Faivre, B. (2017). Multidimensional Model Design using Data Mining: A Rapid Prototyping Methodology. *International Journal of Data Warehousing and Mining (IJDWM) 13(1),* 1-35.

Bramer, M. (2013). *Principles of Data Mining, Second Edition.* London: Springer-Verlag.

Bravo, S., & Jacqueline, A. (2011). *Causas de la deserción escolar de los estudiantes de octavo año de básica del colegio fiscal Palestina.* Guayaquil: Universidad de Guayaquil.

Breiman, L. (1996). Bagging Predictors. *Machine learning, 24(2)*, 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning 45 (1)*, 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* CRC press.

Bruce, M., & Bridgeland, J. (2011). On track for Success: the use of early warning indicator and intervention systems of build a Grand Nation.

Calvet L., L., & Juan P., Á. A. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. *Learning Analytics: Intelligent Decision Support Systems for Learning Environments*, 98-112.

Campbell, A., Mao, X., Pei, J., & Al-Barakati, A. (2017). Multidimensional Business Benchmarking Analysis on Data Warehouses. *International Journal of Data Warehousing and Mining (IJDWM) 13(1)*, 51-75.

Campbell, J., & Oblinger, D. (2007). Academic Analytics. *EDUCASE*.

Cepeda, P., & Gallardo, W. E. (15 de 10 de 2016). Situación demográfica de los estudiantes en los años 2013 a 2016 en la Unidad Educativa Lenin School. (D. Gallardo, Entrevistador)

Cerqueira, P., & Brandão, J. (2017). Adapting Design Thinking to Agile Scrum DW/BI Development. *Business Intelligence Journal, 22(2)*.

Clifton, C. (15 de 12 de 2015). *Data Mining*. Obtenido de Enciclopedia Britanica: http://www.britannica.com/technology/data-mining

Comisón Especial de Estadísticas de Eduación. (2016). *Resolución de la Comisón Especial de Estadísticas de Eduación - CEEE 015-2016.* Quito: Ecuador en Cifras.

CÓRDOVA G., J. C. (2014). *Aplicación De Técnicas De Minería De Datos Para Predecir La Deserción De Los Estudiantes Que Pertenecen Al Colegio Fiscomisional "San Francisco" De La Ciudad De Ibarra (Tesis de pregrado).* Ambato: Universidad Regional Autónoma de Los Andes.

Cortes, C., & Vapnik, V. (1995). Support-vector Networks. *Machine Learning. 20 (3)*, 273-297.

Cortez, V. E., & Pérez, J. I. (2016). *Abandono Escolar De Los Adolescentes De Educación General Básica Del Sector Rural De Gualaceo*. Cuenca: Universidad de Cuenca.

da Cunha, J. A., Moura, E., & Cesar, A. (2016). Data Mining in Academic Databases to Detect Behaviors of Students Related to School Dropout and Disapproval . *New Advances in Information Systems and Technologies*, 189-198.

De la Vega, M. (2014). *El embarazo en adolescentes y su relación con el rendimiento académico de las estudiantes del nivel básico del Colegio "Provincia de Cotopaxi" Ciudad de Pujilí, Cantón Pujilí, Parroquia La Matriz año lectivo 2013-2014.* Latacunga: Universidad Técnica Particular de Loja.

De Witte, K., Cabus, S., Thyssen, G., Groot, W., & Maassen van den Brink, H. (2013). A Critical Review of the Literature on School Dropout. *TIER*.

Dedić, N., & Stanier, C. (2016). Measuring the Success of Changes to Existing Business Intelligence Solutions to Improve Business Intelligence Reporting. *Research and Practical Issues of Enterprise Information Systems*, 225-236.

Dekker, G., Pechenizkiy, M., & Vleeshouwers, J. (2009). Predicting students drop out : a case study. *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009*, 41-50.

Desmarais, M. C. (2011). Mapping question items to skills with non-negative matrix factorization. *SIGKDD Exploration Newsletter, 13(2)*, 30-36.

Ellis, K., Michaely, R., & O'Hara, M. (2000). The accuracy of trade classification rules: Evidence from Nasdaq. *Journal of Financial and Quantitative Analysis, 35(04)*, 529-551.

Espíndola, E., & León, A. (2002). La deserción escolar en América Latina: un tema prioritario para la agenda regional. *Revista Iberoamericana de Educación, No. 30*.

European Commission/EACEA/Eurydice/Cedefop. (2014). *Tackling Early Leaving from Education and Training in Europe: Strategies, Policies and Measures. Eurydice and Cedefop Report.* Luxembourg: Publications Office of the European Union.

Fawcett, T. (2006). An introduction to ROC Analysis. *Patter Recognition Letters 27*, 861-874.

Fix, E., & Hodges, J. (1989). An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges. *International Statistical Review*, 233-238.

Forrester. (2015). *The Forrester Wave™: Agile Business Intelligence Platforms, Q3 2015.* Forrester Inc.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *icml Vol. 96*, 148-156.

Gabriel, I. P. (21 de 02 de 2017). *Lecture Notes & Examples.* Obtenido de Isaac J. Gabriel, Ph.D. : http://isaac.doctor-gabriel.com/MSIS620_Touro/IntroToDW_DWDevelopmentLifecycle_Kimball.ppt

Gallardo, D. (2012). *Diseño e implementación de un sistema de administración de tiempos en proyectos de desarrollo de software y control de desempeño mediante cubos de información para toma de decisiones gerenciales. Caso práctico: cubos de información para el control del des*. Latacunga: Escuela Politécnica del Ejercito.

García-Tinizaray, D., Ordoñez-Briceño, K., & Torres-Diaz, J. C. (2014). Learning analytics para predecir la deserción de estudiantes a distancia. *Campus virtuales, 3(1)*, 120-126.

Garson, G. (1991). Interpreting neural network connection weights. *Artificial Intelligence Expert. 6(4)*, 46–51.

Gartner. (8 de 12 de 2016). *IT Glossary*. Obtenido de Gartner: http://www.gartner.com/

Gartner. (2017). *Magic Quadrant for Business Intelligence and Analytics Platforms.* Gartner Inc.

Gartner Inc. (15 de 12 de 2015). *Gartner*. Obtenido de It Glossary: http://www.gartner.com/it-glossary/data-mining

Gavoyiannis, A. E., Vogiatzis, D. G., Georgiadis, D. P., & Hatziargyriou, N. D. (2001). Combined Support Vector Classifiers using Fuzzy Clustering for Dynamic Security Assessment. *In Power Engineering Society Summer Meeting, (Vol. 2)*, 1281-1286.

George, J., Vijayakumar, B., & Santhosh Kumar, V. (2015). Data Warehouse Design Considerations for a Healthcare Business Intelligence System. *Proceedings of the World Congress on Engineering 2015 Vol I*.

Gestal P., M. (17 de December de 2015). *Sistemas Adaptativos y Bioinspirados en Inteligencia Artificial.* Obtenido de Universidade da Coruña: http://sabia.tic.udc.es/mgestal/cv/RNATutorial/TutorialRNA.pdf

Ghahramani, Z. (2004). Unsupervised Learning. En O. Bousquet, G. Raetsch, & U. Von Luxburg, *Advanced Lectures on Machine Learning.* Springer-Verlag.

Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques.* Berlin: Springer-Verlag Berlin Heidelberg.

Grossmann, W., & Rinderle-Ma, S. (2015). *Fundamentals of Business Intelligence.* New York: Springer Heidelberg.

GUALACATA, J. (2015). *Desarrollo De Una Solución De Business Intelligence Orientado Al Análisis Del Rendimiento Académico Y La Planificación De Los Cursos De La Carrera De Ingeniería Informática En La Facultad De Ingeniería, Ciencias Físicas Y Matemática (Tesis de pregrado).* Quito: Universidad Central Del Ecuador.

GUAMÁN, J. A. (2016). *La Inversión En Educación Y Su Incidencia En El Crecimiento Económico De Ecuador, Periodo: 2000-2014.* Riobamba: Universidad Nacional de Chimborazo.

Hamilton, J. D. (1994). *Time series analysis (Vol. 2).* Princeton: Princeton university press.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28(1)*, 100-108.

Hema, R., & Malik, N. (2010). Data Mining and Business Intelligence. *INDIACom-2010.*

Hidalgo, O. (2012). *Factores escolares y extra-escolares que inciden en el fracaso escolar de los/as estudiantes de la "Unidad Educativa Experimental Manuela Cañizares" de la ciudad de quito. Año lectivo 2010-2011.* Quito: Universidad Central del Ecuador.

Hinojosa, S. E., & Zambrano, H. R. (2012). *Repitencia y deserción de los estudiantes de pregrado de la Facultad de Ciencias Psicológicas y la Facultad de Ingeniería, Escuela Ciencias, Carrera Ingeniería Informática de la Universidad Central del Ecuador, durante el período 2003-2009.* Quito: Universidad Central del Ecuador.

Holmes, D. E., & Jain, L. C. (2012). *Data Mining: Foundations and Intelligent Paradigms.* Berlin: Springer-Verlag.

IBM. (2012). *Manual CRISP-DM de IBM SPSS Modeler.* IBM Corporation 1994.

Imhoff, C., & White, C. (2011). *Self-Service Business Intelligence. Empowering Users to Generate Insights.* Renton: TDWI Best Practices Report.

INEC. (2011). *Encuesta de Estratificación del Nivel Socioeconómico NSE 2011.* Quito: INEC.

INEC. (2014). *Metología para la medición del empleo en Ecuador.* Quito: INEC.

Instituto de Estadística de la UNESCO. (2012). *Compendio mundial de la educación 2012. Oportunidades perdidas: El impacto de la repetición y de la salida prematura de la escuela.* Québec: Instituto de Estadística de la UNESCO.

Jacobsson, M., Ransnäs, A., & Runnström, C. (2015). *Self-Service Business Intelligence Konsekvenserna av flexibilitet.* Lund, Suecia: Lund University. Institutionen för informatik.

Jaramillo, B. (2015). Clase Magistral Data Mining [Grabado por D. Gallardo]. Sangolqui, Pichincha, Ecuador.

Jeong, H., & Biswas, G. (2008). Mining student behavior models in Learning-by-teaching environments. *R. S. J. D. Baker, T. Barnes, & J. Beck (Eds.), Proceedings of the 1st International Conference on Educational Data Mining*, 127-136.

Jukic, N. (2006). Modeling strategies and alternatives for data warehousing projects. *Communications of the ACM, 49(4)*, 83-88.

Kaplan, R. (2009). Entrevista a Robert Kaplan, creador del BSC. (E. HSM, Entrevistador)

KDnuggets. (14 de Diciembre de 2016). *Polls*. Obtenido de KDnuggets: http://www.kdnuggets.com/polls/

Kimball, R., & Ross, M. (2008). *The Data Warehouse Lifecycle toolkit, 2nd Edition.* Wiley.

Kinnebrew, J., & & Biswas, G. (2012). Identifying learning behaviours by contextualizing differential sequence mining with action features and performance evolution. *K. Yacef, O. Zaïane, H. Hershkovitz, M. Yudelson, & J. Stamper (Eds.), Proceedings of the 5th International Conference on Educational Data Mining*, 57-64.

Kuhn, M., Weston, S., Coulter, N., & Quinlan, R. (2015). *C5.0 Decision Trees and Rule-Based Models.* Obtenido de R package version 0.1. 0-24: https://cran.r-project.org/web/packages/C50/C50.pdf

Lamb, S., Markussen, E., Teese, R., Sandberg, N., & Polese, J. (2011). *School Dropout and Completion International Comparative Studies in Theory and Policy.* Netherlands: Springer Netherlands.

Langley, P., Iba, W., & Thompson, K. (1992). An Analysis of Bayesian Classifiers. *Proceedings of the Tenth National Conference on Artificial Intelligence*, 223-228.

Lee, J. I., & Brunskill, E. (2012). The impact on individualizing student models on necessary practice opportunities. *K. Yacef, O. Zaïane, H. Hershkovitz, M. Yudelson, & J. Stamper (Eds.), Proceedings of the 5th International Conference on Educational Data Mining*, 118-125.

Lemaire, V., Christophe, S., & Bondu, A. (2015). A Survey on Supervised Classification on Data Streams. *eBISS 2014, LNBIP 205*, 88-125.

Luan, J. (2002). Data Mining and Knowledge Management in Higher Education Potential Applications. *Annual Forum for the Association for Institutional Research (42nd, Toronto, Ontario, Canada,June 2-5, 2002).*

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A Data Mining & Knowledge Dicovery Process Model. En J. Ponce, & A. Karahoca, *Data Mining and Knowledge Discovery in Real Life Applications* (pág. 438). Vienna: I-Tech Education and Publishing.

Marin, J. M. (17 de December de 2015). *Departamento de Estadística.* Obtenido de Universidad Carlos III de Madrid: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/EDescrip/tema7.pdf

Márquez-Vera, C., Romero M., C., & Ventura S., S. (2013). Predicting School Failure and Dropout by Using Data Mining Techniques. *IEEE Journal Of Latin-American Learning Technologies, VOL. 8*, 7-14.

Martinho¸Bruno, & Yasmina S., M. (2016). An Architecture for Data Warehousing in Big Data Environments. *Research and Practical Issues of Enterprise Information Systems*, 241-250.

Mendoza, E., & Zúñiga, M. (2017). Intra and extra school factors associated with educational backwardness in vulnerable communities. *Alteridad. Revista de Educación. Vol. 12*, 79-91.

Michel, M. (2017). Managing Technical Debt in the Data Warehouse. *Business Intelligence Journal, 22(2)*.

Montenegro L., C. M., & Taco C., M. B. (2012). *Repitencia y deserción de los estudiantes de pregrado de las Facultades de Arquitectura y Administración: carrera de Administración Pública presencial y semi-presencial de la Universidad Central del Ecuador, causas, consecuencias y costos económico finan.* Quito: Universidad Central del Ecuador.

Narváez, M., & Barragán, G. (2015). *Análisis sobre la deserción estudiantil en la Universidad Politécnica Salesiana, sede Guayaquil: Caso de las carreras de Administración de Empresas y Contabilidad y Auditoría. Período de aplicación 2007 - 2012.* Guayaquil: Universidad Politécnica Salesiana.

National Dropout Prevention Center/Network At Clemson University. (12 de 2016). *Why Students Drop Out*. Obtenido de Dropout Prevention: http://dropoutprevention.org/resources/statistics/quick-facts/why-students-drop-out/

Nithya, P., Umamaheswari, B., & Umadevi, A. (2016). A Survey on Educational Data Mining in Field of Education. *International Journal of Advanced Research in Computer Engineering & Technology, 5(1)*, 69-78.

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling 178*, 389–397.

Olson, D. L. (2007). Data mining in business services. *Service Business* , 181-193.

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques.* Berlín: Springer-Verlag.

Ordoñez B., K. (2013). *Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL.* Loja: Universidad Técnica Particular de Loja.

Organización Internacional del Trabajo. (2008). *Resolución sobre la actualización de la Clasificación Internacional Uniforme de Ocupaciones.* Organización Internacional del Trabajo.

Pacho, F., & Chiqui, D. (2011). *Estudio de las causas de la deserción escolar.* Cuenca: Universidad de Cuenca.

Pal, S. (2012). Mining Educational Data to Reduce Dropout Rates of Engineering Students. *I.J. Information Engineering and Electronic Business*, 1-7.

Palazuelos, C., García-Saiz, D., & Zorrilla, M. (2013). Social Network Analysis and Data Mining: An Application to the E-learning Context. *C. Badica, N. T. Nguyen, M. Brezovan (Eds.), Proceedings of the 5th International Conference on Computational Collective Intelligence*, 651-660.

Papamitsiou, Z., & Economides, A. A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society, 17(4)*, 49-64.

Parra, J. (2010). *Estrategias para el mejoramiento de la calidad de la educación básica: el caso de las escuelas y colegios del ciclo costa, de los cantones San Miguel de los Bancos, Pedro Vicente Maldonado y Puerto Quito.* Quito: Instituto de Altos Estudios Nacionales.

Pérez, E. (2014). *Determinantes socioeconómicos del avance y deserción del nivel primario al secundario en el sistema educativo del Ecuador : ENEMDU 2012.* Quito: Pontificia Universidad Católica del Ecuador.

Platoš, J. (23 de Octubre de 2016). *Department of Computer Science | Data Analysis 3. Support Vector Machines.* Obtenido de VŠB - Technical University of Ostrava: http://homel.vsb.cz/~pla06/files/mad3/mad3_04.pdf

Pour, J. (2014). Self-service business intelligence. *Systémová Integrace 1-2*, 135-146.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning.* Elseiver.

Raschka, S. (2015). *Python Machine Learning.* Packt Publishing Limited.

Ravat, F., Song, F., & Teste, O. (2016). OLAP Analysis Operators for Multi-State Data Warehouses. *International Journal of Data Warehousing and Mining (IJDWM) 12(4)*, 20-53.

Revelo C., D., Hinojosa, C., & Duque, L. (2013). Desarrollo De Cubos De Información Y Una Aplicación Web Orientados A La Toma De Decisiones Con El Uso De La Plataforma De Bussiness Inteligence De Microsoft Para Generar Un Ranking De Instituciones Financieras A Nivel Nacional. *Universidad de las Fuerzas Armadas ESPE. Carrera de Ingeniería en Sistemas e Informática*.

Reyna, J., Salcido, M., & Arredondo, A. (2013). Análisis del ciclo vital de la estructura familiar y sus principales problemas en algunas familias mexicanas. *Altern. psicol. [online], 17(28)*, 73-91.

Rivero Pérez, J. L. (2014). Técnicas de aprendizaje automático para la detección de intrusos. *Revista Cubana de Ciencias Informáticas*, 52-73.

Rodríguez, J. (2016). *Determinantes de la deserción escolar en la educación secundaria asociados a las características de las instituciones educativas en el Ecuador.* Quito: Pontificia Universidad Católica del Ecuador.

Rodríguez, T. L., & Sánchez, J. M. (2006). Estructura familiar y satisfacción parental: propuestas para la intervención. *Acciones e investigaciones sociales, (1)*, 455-490.

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery, 3*, 12-27.

Russell, S., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach (2nd ed.).* Prentice Hall.

Sánchez, D. (2015). La tendencia del abandono escolar en Ecuador: período 1994-2014. *VALOR AGREGADO* , 37-66.

Sayad, S. (7 de 3 de 2017). *Model Evaluation - Classification*. Obtenido de An Introduction to Data Mining: http://www.saedsayad.com/model_evaluation_c.htm

Secretaría Nacional de Planificación y Desarrollo. (2013). *Plan Nacional para el Buen Vivir 2013-2017.* Quito: Secretaría Nacional de Planificación y Desarrollo – Senplades. Obtenido de http://documentos.senplades.gob.ec/Plan%20Nacional%20Buen%20Vivir%202013-2017.pdf

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing, 5(4)*, 13-22.

Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree. *Indian Journal of Science and Technology, Vol 9(4)*, 1-5.

Statnikov, A., Hardin, D., & Aliferis, C. (2006). Using SVM Weight-Based Methods to Identify Causally Relevant and Non-Causally Relevant Variables. *sign, 1(4)*.

Suganya, S., & Narayani, V. (161-166). Analysis of students dropout forecasting using data mining. *3rd Internaational Conference on Lastest Trends in Engineering, Science, Humanities and Management*, 2017.

Tane, J., Schmitz, C., & Stumme, G. (2004). Semantic resource management for the web: an e-learning application. *Proceedings of the 13th International Conference of the WWW*, 1-10.

Taniar, D. (2009). *Progressive Methods in Data Warehousing and Business Intelligence: Concepts and Competitive Analytics.* Australia: IGI Global.

Tanja, T., & Van der Velden, R. K. (2008). Early school-leaving in the Netherlands. The role of student-, family- and school factors for early school-leaving in lower secondary education. *ROA-RM-2008/3*.

TES Global Limited. (15 de 3 de 2017). *Wikispaces*. Obtenido de Rpro: https://rpro.wikispaces.com/Comparativa+de+modelos+de+clasificaci%C3%B3n

Trćka, N., Pechenizkiy, M., & Aalst, W. V. (2011). Process mining from educational data. *C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. D. Baker (Eds.), Handbook of Educational Data Mining*, 123-142.

Ueno, M. (2004). Online Outlier Detection System for Learning Time Data in E-Learning and Its Evaluation. *Proceedings of the International Conference on Computers and Advanced Technology in Education*, 248-253.

Valdivieso, C. (2013). *El embarazo adolescente: un problema de salud pública.* Quito: Pontificia Universidad Católica del Ecuador.

Valmyr B., E., Lydersen, S., & Kvernmo, S. (2016). Non-completion of upper secondary school among female and male young adults in an Arctic sociocultural context; the NAAHS study. *BMC Public Health*.

Vargas M., H. (2014). Tipo de familia y ansiedad y depresión. *Rev Med Hered [online], 25(2)*, 57-59.

Vinueza, J. (2015). Clase magistral Data Mining [PowerPoint slides]. Sangolqui, Pichincha, Ecuador.

Waite, L. (2006). Marriage and Family. *Poston DL, Micklin M, editors Handbook of Population. New York: Springer*, 87–108.

Watson, H. (2017). The Cognitive Decision-Support Generation. *Business Intelligence Journal, 22(2)*.

Wisconsin Department of Public Instruction. (27 de 11 de 2016). *Welcome to the Data Warehouse & Decision Support Team page!* Obtenido de Wisconsin Department of Public Instruction: https://dpi.wi.gov/dwds

Wisconsin Department of Public Instruction. (29 de 11 de 2016). *Welcome to WISEdash — where you can compare and explore statistics about Wisconsin public schools*. Obtenido de WISEdash: http://wisedash.dpi.wi.gov/Dashboard/portalHome.jsp

Xu, X. (2016). Data Approximation for Time Series Data in Wireless Sensor Networks. *International Journal of Data Warehousing and Mining (IJDWM) 12(3)*, 1-13.

Yadav, S., & Pal, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. *World of Computer Science and Information Technology Journal, 2(2)*, 51-56.

Yépez, V. (2013). *Deserción Escolar De Los Estudiantes De Educación Básica De La Unidad Educativa "Sagrado Corazón De Jesús" De La Ciudad De Latacunga Durante Los Últimos Tres Años Lectivos.* Quito: Universidad Andina Simón Bolivar.

Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting Dropout Student: An Application Of Data Mining Methods In An Online Education Program. *European Journal of Open, Distance and e-Learning Vol. 17*, 118-133.

Zhi-Hua, Z. (2012). *Ensemble Methods: Foundations and Algorithms.* Ney York: Chapman and Hall/CRC.