TÉCNICAS DE MINERÍA DE DATOS PARA DETECTAR EL BAJO RENDIMIENTO ACADÉMICO

Nelly Jacqueline Ulloa Gallardo
Ralph Miranda Castillo
Luis Alberto Holgado Apaza

ilae
Instituto Latinoamericano de Altos Estudios

Técnicas de minería de datos para detectar patrones de bajo rendimiento académico

Técnicas de minería de datos para detectar patrones de bajo rendimiento académico

Nelly Jacqueline Ulloa Gallardo
Ralph Miranda Castillo
Luis Alberto Holgado Apaza

Instituto Latinoamericano de Altos Estudios –ILAE–

Publicación sometida a evaluación de pares académicos (*Peer Review Double Blinded*).

ISBN 978-958-5535-46-6

Cra. 18 # 39A-46, Teusquillo, Bogotá, Colombia
pbx: (571) 232-3705, fax (571) 323 2181
www.ilae.edu.co

Diseño de carátula y composición: Jesús Alberto Chaparro Tibaduiza
Edición electrónica: Editorial Milla Ltda. (571) 702 1144
editorialmilla@telmex.net.co

Editado en Colombia
Published in Colombia

Contenido

Índice de tablas

ÍNDICE DE FIGURAS

Introducción

En una ciudad inteligente, los individuos necesitan de una educación inclusiva en relación a la variedad cultural de sus ciudadanos. Las instituciones educativas deben apuntar a lograr una formación integral de la población al tomar en cuenta la heterogeneidad del proceso de aprendizaje. No obstante, lo anterior resulta impedido de realizarse debido a diferentes situaciones que acarrean los estudiantes como los problemas individuales, la falta de recursos económicos, el bajo desempeño, entre otros[1].

En relación a lo anterior, el abandono estudiantil cobra especial interés debido a su relación con el desempeño intelectual, con las etapas de selección y con el rendimiento académico en el estudiante. En ese sentido, el grado de deserción de los estudios es el resultado de la mezcla y efecto de diversas variables[2].

Las instituciones de nivel superior por lo general contratan docentes que cuentan con un alto grado de preparación, pero es constante hallar que el desarrollo de los contenidos en los salones no se realice en consideración de los diferentes estilos de aprendizaje que requieren los estudiantes, lo que de manera probable derive a incrementar el porcentaje de fracaso estudiantil por tanta deserción.

---

1 Cruz Vergara, Ana Oviedo, Claudia Carmona, Gloria Vélez e Iván Amón. "Estilos de aprendizaje y minería de datos: un estudio preliminar en el contexto universitario", en *Ingeniería e Innovación*, vol. 6, n.º 1, junio de 2018, pp. 13 a 18, disponible en [https://revistas.unicordoba.edu.co/index.php/rii/article/view/1534/1803].

2 Karina B. Eckert y Roberto Suénaga. "Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos", en *Formación Universitaria*, vol. 8, n.º 5, 2015, disponible en [https://www.researchgate.net/publication/281671104_Analisis_de_Desercion-Permanencia_de_Estudiantes_Universitarios_Utilizando_Tecnica_de_Clasificacion_en_Mineria_de_Datos/fulltext/5681256408ae051f9aec2b62/Analisis-de-Desercion-Permanencia-de-Estudiantes-U].

En Mercedes Segarra Ciprés, Marta Estrada Guillén y Diego Monferrer Tirado[3] se dio cuenta de los estilos de aprendizaje de universitarios pertenecientes a programas de diferentes áreas de conocimiento como las ciencias sociales, humanas, experimentales y económicas. En el estudio se profundizó en la comprensión del estilo de aprendizaje abordado por los universitarios, así como en el manejo de fórmulas para su mejora y la mejora de sus rendimientos. Una vez tenida las puntuaciones para cada cuadrante, se señaló que los alumnos con mejor desempeño son aquellos que cuentan con un perfil de dominancia alterna, o sea, que en su pensamiento interconectan los dos hemisferios cerebrales.

Más del 90% de la información de todo el mundo para el 2007 ya estaba en formato digital. La humanidad pudo almacenar 2.9×10^{20} bytes comprimidos de manera óptima, comunicar casi 2×10^{21} bytes, y llevar a cabo 6.4×10^{18} instrucciones por segundo en computadoras de uso general[4]. Mucha de esta información es producto de las operaciones que a diario se realizan como búsquedas en internet, compras de artículos, noticias que gustamos leer, mensajes que enviamos mediante las redes sociales, correos electrónicos, entre otros, los mismos que son para luego ser almacenados en grandes bases de datos.

Las universidades no son ajenas al fenómeno referido, dado que como organizaciones están formadas por distintas dependencias que generan gran cantidad de datos –propios de las operaciones transaccionales que realizan a diario–, los más resaltantes que se gestan, en su mayoría, son de carácter administrativo y académico, fruto de los procesos de admisión, matrículas, enseñanza-aprendizaje (aulas virtuales), evaluaciones, entre otros.

---

3 Mercedes Segarra Ciprés, Marta Estrada Guillén y Diego Monferrer Tirado. "Estilos de aprendizaje en estudiantes universitarios: lateralización vs. Interconexión de los hemisferios cerebrales", en *Revista española de pedagogía*, n.º 262, septiembre-diciembre de 2015, pp. 583 a 600, disponible en [https://revistadepedagogia.org/wp-content/uploads/2015/11/Estilos-de-aprendizaje-en-estudiantes-universitarios-lateralizaci%C3%B3n-vs.-interconexi%C3%B3n-de-los-hemisferios-cerebrales.pdf].

4 Martin Hilbert y Priscila López. "The World's Technological Capacity to Store, Communicate, and Compute Information", en *Science*, vol. 332, n,º 6025, abril de 2011, pp. 60 a 66.

Por ejemplo, Jesús Walter Salinas[5] manifiesta que en los últimos semestres el número de estudiantes desaprobados en el curso Estadística General ha correspondido a un 41%. Por ello, se planteó la hipótesis de que existe dependencia entre el rendimiento académico (aprobado y desaprobado) de los alumnos con las variables socio-demográficas y académicas, tal dependencia puede expresarse a través de un modelo estadístico. Al usar las técnicas estadísticas de minería de datos se estudiaron a los alumnos de pre-grado de la Universidad Nacional Agraria La Molina que hayan llevado el curso durante tres semestres académicos en un número aproximado de 1.500 alumnos. De lo anterior, se encontraron las principales variables sociodemográficas y académicas que determinan la situación del rendimiento académico. Al usar esta información, se puede predecir la situación final del alumno apenas se matricule en el curso sin haber rendido ningún tipo de evaluación.

5 Jesús Walter Salinas. "Detección de patrones de los alumnos de pregrado desaprobados en el curso de estadística general de la Universidad Nacional Agraria La Molina usando técnicas de minería de datos", *Memorias del II Encuentro Colombiano de Educación Estocástica*, 2016, pp. 115 a 122, disponible en [http://funes.uniandes.edu.co/9282/1/Salinas2016Deteccion.pdf].

Capítulo primero
El proceso de descubrimiento de conocimiento en base de datos –KDD–

El proceso KDD en la educación es un término referido desde hacía varios años atrás y su investigación y atención ha sido trascendente en la actualidad, el uso de este proceso autoriza revisar cuantiosos volúmenes de datos al encontrar vínculos y patrones no triviales.

Diferentes universidades han puesto a prueba proyectos de investigación en referencia al abandono estudiantil según la aplicación de KDD. En Colombia, por ejemplo, se emplearon las técnicas de minería de datos para evaluar los factores que determinaron grandes índices de deserción universitaria en la Universidad de Nariño y la Institución Universitaria CESMAG[6].

En referencia a ello, con más antelación, Ricardo Timarán Pereira[7] presenta los resultados de la investigación realizada en la Universidad de Nariño (Colombia) cuyo objetivo fue determinar en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil al aplicar técnicas de descubrimiento de conocimiento, a partir de los datos almacenados en las bases de datos durante los últimos 15 años. Este proceso se apoyó con TariyKDD, una herramienta de mine-

6 Leidy Carolina Calvache Fernández, Valentina Álvarez Vallejo y Jorge Iván Triviño Arbeláez. "Proceso KDD como apoyo a las estrategias del proyecto SARA (Sistema de Acompañamiento para el Rendimiento Académico)", *Revista Educación en Ingeniería*, vol. 13, n.º 26, julio de 2018, pp. 82 a 89, disponible en [https://educacioneningenieria.org/index.php/edi/article/view/916/365].

7 Ricardo Timarán Pereira. "Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos", en *Memorias de la 8.ª* Conferencia Iberoamericana en Sistemas, Cibernética e Informática *CISCI*, 2009, disponible en [http://www.iiis.org/cds2008/cd2009cSc/CISCI2009/PapersPdf/C692YV.pdf].

ría de datos de distribución libre, desarrollada en los laboratorios de DCBD del Departamento de Ingeniería. Las técnicas de minería de datos utilizadas para el descubrimiento de patrones de deserción estudiantil y bajo rendimiento académico las de clasificación y asociación. Para generar las reglas de clasificación se utilizó el algoritmo C4.5 y para las reglas de Asociación, el algoritmo EquipAsso.

Por otro lado, Jorge Bacallao Gallestey *et al.*[8] realizaron un estudio para detectar estudiantes con alto riesgo de fracaso académico e identificar los mejores predictores del rendimiento. Se caracterizaron los estudiantes que ingresaron en el primer año en el ICBP "Victoria de Girón" durante el curso 2001-2002 de acuerdo con su índice académico del preuniversitario, índice escalafonario, exámenes de ingreso, prueba de inteligencia y un indicador de su motivación profesional. Se emplearon árboles de clasificación para identificar los predictores relevantes y sus puntos de corte óptimos. Se utilizó un modelo de regresión ordinal para evaluar la importancia relativa de los predictores y proponer el algoritmo de predicción. A partir del índice escalafonario, de forma exclusiva, se obtuvo un procedimiento de clasificación, que permitió identificar a los estudiantes de mayor riesgo de fracaso académico. Los puntos de corte fueron 87 y 91 puntos, que definen una tricotomía para el pronóstico del rendimiento.

En Argentina, en la Universidad de Misiones una tesis centró su atención en el empleo de técnicas de minería de datos para determinar y englobar a los estudiantes según sus cualidades académicas, temas sociales y demográficos, con la finalidad de disminuir el porcentaje de abandono en los modelos académicos de esta universidad. En la tesis, además, se sostiene que los mejores resultados de acuerdo con el estudio de la deserción se lograron mediante la técnica de árboles de decisión y han evaluado la posibilidad de apreciar más variables socioeconómicas[9].

---

8 Jorge Bacallao Gallestey, José Mario Parapar de la Riestra, Mercedes Roque Gil y Jorge Bacallao Guerra. "Árboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico", en *Educación Médica Superior*, vol. 18, n.º 3, julio-septiembre de 2004, disponible en [http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21412004000300002].

9 Jesús Germán Andrés Pautsh. "Minería de datos aplicada al análisis de la deserción en la carrera de analista en sistemas de computación", tesis de pregrado, Posadas, Univer-

La incorporación del proceso KDD cimentado en minería de datos hace posible hallar conocimiento útil y original que no es apreciable a simple vista. Al mostrar patrones, un proceso de minería de datos confiere el control adecuado de la información y soporta la toma de decisiones en el seno de una organización o empresa, de manera que otorga la probabilidad de mejorar detalles de su contexto. Al hallar información trascendente y no trivial, un proceso de minería consentiría otorgar un valor diferenciador para la organización.

En lo que respecta a la aplicación de las técnicas de minería de datos en terreno educativo, varios de los patrones abordados a lo largo de esta investigación se pueden transformar en un punto de inicio para promover estrategias tempranas de retención estudiantil, que tienen que ver con la disponibilidad horaria, asesorías académicas, apoyos psicológicos y colaboración académica que conceda a los estudiantes a proseguir con su ciclo educativo.

I. El proceso de KDD en retrospectiva

Rosibelda Mondragon Becerra[10] citando a Usama Fayyad, Gregory Piatetsky Shapiro y Smyth Phadhraic[11], define KDD como un proceso no trivial de identificación de patrones válidos, novedosos, en potencia útiles, y entendibles en los datos. En este contexto, los datos se refieren a un conjunto de hechos (ejemplos en una base de datos); los patrones, por otro lado, son expresiones en algún lenguaje que describen de manera compacta los datos. El término proceso implica que KDD

---

sidad Nacional de Misiones, 2009, disponible en [https://www.lawebdelprogramador.com/pdf/6566-Mineria-de-Datos-aplicada-al-analisis-de-la-desercion-en-la-Carrera-de-Analista-en-Sistemas-de-Computacion.html].

10 Rosibelda Mondragón Becerra. "Exploraciones sobre el soporte multi-agente BDI en el proceso de descubrimiento de conocimiento en bases de datos", tesis de maestría, Veracruz, Universidad Veracruzana, 2007, disponible en [https://www.uv.mx/personal/aguerra/files/2013/06/2007-mondragon-becerra.pdf].

11 Usama Fayyad, Gregory Piatetsky Shapiro y Smyth Phadhraic. "Knowledge Discovery and Data Mining: Towards a Unifying Framework", en *KDD*, n.º 96, 1996, pp. 82 a 88, disponible en [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwj5saPW26DrAhUpU98KHTBHCV8QFjACegQIAxAB&url=https%3A%2F%2Fwww.aaai.org%2FPapers%2FKDD%2F1996%2FKDD96-014.pdf&usg=AOvVaw0RmxCEkOXM9afVEgY_m5Y2].

comprende muchos pasos, entre los que se encuentran: la preparación de datos, la búsqueda de patrones, la evaluación del conocimiento, y el refinamiento; los cuales pueden ser repetidos en múltiples iteraciones. Por no trivial, debe entenderse que alguna búsqueda o inferencia es llevada a cabo; es decir, involucra la búsqueda de estructuras, modelos, patrones o parámetros. Los patrones descubiertos deben ser válidos sobre nuevos datos con algún grado de certeza, para que puedan describir y/o predecir de manera confiable el comportamiento futuro de alguna entidad. También se desea que los patrones sean novedosos (al menos para el sistema y de forma preferente para el usuario) y en potencia útiles, es decir, que proporcionen algún beneficio al usuario o a la tarea. Por último, los patrones deben ser entendibles, en otro caso, será necesario algún post procesamiento.

El proceso de KDD es interactivo e iterativo que involucra numerosos pasos con muchas decisiones tomadas por el usuario. Ronald Jay Brachman y Tej Anand[12] ofrecen una visión práctica del proceso KDD al enfatizar la naturaleza interactiva del proceso.

II. El proceso de la KDD

El proceso de KDD consta de cinco etapas. Webmining Consultores[13] describe cada etapa de la siguiente manera.

A. Selección de datos

En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos.

---

12 Ronald Jay Brachman y Tej Anand. "The process of knowledge Discovery in databases", en Brachman (ed.), *Workshop on knowledge discovery in databases*, 1994, pp. 37 a 53, disponible en [https://pdfs.semanticscholar.org/2db5/ec88e07974242eb8f8de867275bec8f29e3a.pdf].

13 Webmining Consultores. *Webmining*, 10 de enero de 2011, disponible en [http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/].

B. Procesamiento

Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

C. Transformación

Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, al consolidar los datos de una forma necesaria para la fase siguiente.

D. Data mining

Es la fase de modelamiento propiamente tal, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones antes desconocidos, válidos, nuevos, de manera potencial útiles y comprensibles y que están contenidos u "ocultos" en los datos.

E. Interpretación y evaluación

Se identifican los patrones obtenidos y que son en realidad interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

Para César Pérez López y Daniel Santín[14] el proceso de extracción del conocimiento KDD consta de las siguientes fases:

- Selección
- Exploración

14 César Pérez López y Daniel Santín. *Minería de datos. Técnicas y herramientas*, Madrid, Thomson, 2007.

- Limpieza
- Transformación
- Minería de datos
- Evaluación
- Difusión

En la fase de selección se integran y recopilan los datos, se determinan las fuentes de información que pueden ser útiles y donde conseguirlas, se identifican y seleccionan las variables relevantes en los datos y se aplican las técnicas de muestreo adecuadas. Todo ello se facilita disponiendo de un almacén de datos con la información en formato común y sin inconsistencias. Dado que lo datos provienen de diferentes fuentes, es necesario su exploración mediante técnicas de análisis exploratorio de datos, al buscar entre otras cosas la distribución de los datos, su simetría y normalidad y las correlaciones existentes en la información. A continuación, es necesaria la limpieza de los datos, ya que pueden contener valores atípicos, valores faltantes y valores erróneos. En esta fase se analizan la influencia de los datos atípicos, se imputan los valores faltantes y se eliminan o corrigen los datos incorrectos. A continuación, si es necesario, se lleva a cabo la transformación de los datos, por lo general mediante técnicas de reducción o aumento de la dimensión y escalado simple y multidimensional, en otras palabras. Las cuatro primeras fases se suelen englobar bajo el nombre de preparación de datos. En la fase de minería de datos, se decide cuál es la tarea a realizar (clasificar, agrupar, etc.) y se elige la técnica descriptiva o predictiva que se va a utilizar. En la fase de evaluación e interpretación se evalúan los patrones y se analizan por los expertos, y si es necesario se vuelve a las fases anteriores para la nueva iteración. Por último, en la fase de difusión se hace uso del nuevo conocimiento y se hace partícipe de él a todos los posibles usuarios.

III. La extracción de conocimientos en bases de datos: las redes sociales

La extracción de conocimientos en bases de datos es un desarrollo dificultoso que tiene por meta otorgarles relevancia a los datos. La minería de datos solo se constituye como un paso de este proceso

cuya finalidad radica en la obtención de patrones y modelos por medio de la aplicación de métodos estadísticos y técnicas de aprendizaje automático.

En los últimos años, la difusión de la KDD se utiliza mediante el concepto de Big Data, aunque de manera equivocada. En 2001, LANEY señaló los tres grandes retos que implicaba el estudio de grandes datos y definió "las tres V" que diferenciaron a Big Data: volumen, variedad y velocidad. En primera instancia, volumen está referido a la magnitud de los datos, ello implica por ejemplo a millones de publicaciones en Facebook o al estudio de billones de críticas de *movies*. Por otro lado, variedad se refiere a la diversidad de los datos a indagar: textos, imágenes, audios, videos, entre otros. Por último, velocidad quiere decir que enormes cantidades de flujos de información son estudiados en tiempo real, por ejemplo, los datos que proceden de los *smartphones*. De forma novedosa, algunos autores han señalado otras particularidades de Big Data: veracidad, variabilidad (o complejidad) y valor[15].

De acuerdo con ello, H. ANDREW SCHWARTZ y LYLE H. UNGAR[16] proponen que con normalidad la psicología ha estudiado los pensamientos, los sentimientos y los rasgos de personalidad por medio de formularios dados a cantidades ínfimas de pacientes. Al contrario, destacan las nuevas opciones en favor de la valoración psicológica que otorgan el estudio de contenido dado por los daots (*data-driven content analyses*) o la perspectiva del vocabulario abierto (*open vocabulary approach*) de emplearse enormes cantidades de volúmenes de información disponibles en las redes sociales.

En esa dirección, otras investigaciones han llegado a la conclusión de que el estudio del lenguaje subyacente de las redes sociales es muy beneficioso para realizar informes epidemiológicos a gran escala o con el objeto de señalar las particularidades psicológicas que

---

15 AMIR GANDOMI y MURTAZA HAIDER. "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management*, vol. 35, n.º 2, 2015, pp. 137 a 144, disponible en [https://www.sciencedirect.com/science/article/pii/S0268401214001066/pdfft?md5=83a9e41c2aa8141394ce1a998ed61553&pid=1-s2.0-S0268401214001066-main.pdf].

16 H. ANDREW SCHWARTZ y LYLE H. UNGAR. "Data-Driven Content Analysis of Social Media: A Systematic Overview of Atomated Methods", en *The ANNALS of the American Academy of Political and Social Science*, vol. 659, n.º 1, abril de 2015, pp. 78 a 94.

se ponen de relieve en distintas zonas geográficas, aquellas relacionadas con el bienestar, por poner un caso. A este respecto, el método de investigación mencionado es más veloz y con un valor monetario inferior que las convencionales encuestas llevadas a cabo por las agencias de gobierno[17].

En ese sentido Luciana Mariñelarena Dondena, Marcelo Luis Errecalde y Alejandro Castro Solano señalan que:

> Las investigaciones interdisciplinarias de las ciencias de la computación, sociales y de la salud constituyen sin lugar a dudas un campo promisorio. En nuestra disciplina en particular, estos enfoques abren la puerta para la medición o evaluación de los constructos psicológicos mediante la aplicación de técnicas de *minería de textos*. Esta perspectiva podría convertirse a futuro en un nuevo método de evaluación psicológica a nivel individual y para la realización de estudios epidemiológicos a gran escala[18].

Al respecto, se ha verificado que la investigación de las peculiaridades del lenguaje empleado en las redes sociales permite conocer las señales de personalidad, el género y el sexo de los participantes[19] como así también anticipar el rango de bienestar de la población que habitan en diferentes lugares de Estados Unidos. Además, se han generado investigaciones epidemiológicas a gran escala que hallan en el lenguaje de las redes sociales aquellas particularidades psicológicas dadas en la comunidad vinculadas con el riesgo de mortalidad por arteriosclerosis[20].

---

17 Hansen Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michael Kosinski, David Stillwell, Martín E. P. Seligman y Lyle H. Ungar. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulay Approach", en *plos one*, vol. 8, n.º 9, e73791, septiembre de 2013, disponible en [https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791].

18 Luciana Mariñelarena Dondena, Marcelo Luis Errecalde y Alejandro Castro Solano. "Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología", *Revista Argentina de Ciencias del Comportamiento*, vol. 9, n.º 2, enero-diciembre de 2017, disponible en [https://revistas.unc.edu.ar/index.php/racc/article/view/12701/Mari%C3%B1elarena-Dondena], p. 74.

19 Schwartz, Eichstaedt, Kern, y otros. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulay Approach", cit.

20 Johannes C. Eichstaedt, Hansen Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, Megha Agrawal, Lukasz A. Dziurzynski, Maarten Sap, Christopher Weeg, Emily E. Larson, Lyle H. Ungar y Martin E. P. Seligman. Psychological Languaje on Twitter Predicts County-Level Heart Disease

Por otro lado, por medio de la información proporcionada por *Twitter* se han seguido los casos de *bullying*. Los *posts* hicieron posible dar con quienes estuvieron envueltos en ellos, cuál fue el tipo de daño y quiénes denuncian estos hechos. A la misma vez, es factible registrar el lugar de donde provienen los mensajes y el día y la hora en que fueron realizados[21].

Mortality, en *Psychol Sci.,* vol. 26, n.º 2, 2015, pp. 159 a 169, disponible en [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4433545/#].

21 Amy Bellmore, Angela J. Calvin, Jun-Ming Xu y Xiaojin Zhu. "The five W's of "bullying" on Twitter: Who, What, Why, Where, and When", en *Computers in Human Behavior*, vol. 44, marzo de 2015, pp. 305 a 314.

Capítulo segundo
Minería de datos

Guillermo Gil Albarrán[22] define a la minería de datos como el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos de un determinado contexto.

De modo esencial, el *datamining* surge para intentar ayudar a comprender el contenido de un repositorio de datos o almacén de datos (*Data Warehouse*). Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmo de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales.

Por otro lado, Microsoft menciona que:

> La minería de datos es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiados datos[23].

Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos. Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

Pronóstico**:** cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.

---

22 Guillermo Gil Albarrán. *Data Mining*, Lima, Megabyte, 2009.

23 Microsoft. "Conceptos de Minería de datos", 2020, disponible en [https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions&viewFallbackFrom=sql-server-ver15].

Riesgo y probabilidad**:** elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.

Recomendaciones**:** determinación de los productos que se pueden vender juntos y generación de recomendaciones.

Búsqueda de secuencias**:** análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.

Agrupación**:** distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

De acuerdo con Mondragón Becerra[24] se pueden identificar dos objetivos del proceso de KDD: (1) *verificación* y (2) *descubrimiento.* En el primer caso, el sistema se limita a verificar la hipótesis del usuario, mientras que, con el descubrimiento, el sistema de forma automática encuentra patrones nuevos, siempre que sea posible. El objetivo de descubrimiento se puede subdividir, en *descripción* y *predicción.* Con la descripción, el sistema obtiene patrones que presenta al usuario de forma entendible, y con la predicción, el sistema encuentra patrones para predecir el comportamiento futuro de alguna entidad.

I. Clasificación de las técnicas de *Data Mining*

Las técnicas de minería de minería de datos de acuerdo con Pérez López y Santín[25] podemos clasificarlo como: técnicas predictivas, técnicas descriptivas y técnicas auxiliares.

A continuación, se muestra una clasificación de las técnicas de *Data Mining*. En primer lugar, se tiene a las técnicas predictivas que comprenden procesos como la regresión, las redes neuronales artificiales, *bosstrap*, entre otros; por otro lado, las técnicas descriptivas que cuenta con el *clustering*, enlace promedio, enlace simple, entre otros.

---

24 Mondragón Becerra. "Exploraciones sobre el soporte multi-agente BDI en el proceso de descubrimiento de conocimiento en bases de datos", cit.

25 Pérez López y Santín. *Minería de datos. Técnicas y herramientas*, cit.

Las técnicas de clasificación pueden pertenecer tanto al grupo de técnicas predictivas como a las descriptivas. Las técnicas de clasificación predictivas suelen denominarse técnicas de clasificación *ad hoc* ya que clasifican individuos u observaciones dentro de grupos con antelación definidos. Las técnicas descriptivas se denominan técnicas de clasificación *post hoc* porque realizan clasificación sin especificación previa de los grupos.

II. Técnicas predictivas

De acuerdo con María N. Moreno García, Luis Antonio Miguel Quintales, Francisco José García Peñalvo y María José Polo Martín[26], los algoritmos de las técnicas predictivas anticipan el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de los datos, cuya etiqueta se conoce, se induce a una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: Entrenamiento (construcción de un modelo al usar un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).

Por ejemplo, las redes neuronales permiten descubrir modelos más complejos y afinarlos a medida que progresa la exploración de los datos. Gracias a su capacidad de aprendizaje, permiten descubrir relaciones complejas entre variables sin ninguna intervención externa. Podemos incluir en estas técnicas todos los tipos de regresión, series temporales, análisis de la varianza, y la covarianza, análisis discriminante, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas. Tanto los arboles de decisión como las redes neuronales y el análisis discriminante son a su vez técnicas de clasificación que

---

26 María N. Moreno García, Luis Antonio Miguel Quintales, Francisco José García Peñalvo y María José Polo Martín. "Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software", en *Researchgate*, 2001, disponible en [https://www.researchgate.net/publication/220958273_Aplicacion_de_Tecnicas_ de_Mineria_de_Datos_en_la_Construccion_y_Validacion_de_Modelos_Predicti vos_y_Asociativos_a_Partir_de_Especificaciones_de_Requisitos_De_Software].

pueden extraer perfiles de comportamiento o clases. Los árboles de decisión permiten clasificar los datos en grupos basados en los valores de las variables. El mecanismo de base consiste en elegir un atributo como raíz y desarrollar el árbol según las variables más significativas[27].

De acuerdo con Alveiro Alonso Rosado Gómez y Alejandra Verjel Ibáñez[28] las técnicas predictivas tienen las tareas de clasificación y regresión, por otra parte, Violeta Valcárcel Asencios[29] afirma que las tareas de regresión persiguen la obtención de un modelo que permita predecir el valor numérico de alguna variable (modelos de regresión logística). En relación con la clasificación, Tomas Aluja Banet[30] menciona que si la respuesta es categórica (p. e. la compra o no de un producto) diremos que se trata de un problema de clasificación.

A continuación, detallamos algunas de las técnicas supervisadas de minería de datos.

A. Análisis de regresión logística

La regresión logística, al igual que otras técnicas estadísticas multivariadas, da la posibilidad de evaluar la influencia de cada una de las variables independientes sobre la variable dependiente o de respuesta y controlar el efecto del resto. Tendremos, por tanto, una variable dependiente, llamémosla Y, que ser dicotómica o politónica y una o más variables independientes, llamémoslas X, que pueden ser de cualquier

---

27 Pérez López y Santín. *Minería de datos. Técnicas y herramientas*, cit.

28 Alveiro Alonso Rosado Gómez y Alejandra Verjel Ibáñez. "Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander", *Revista Tecnura*, vol. 79, n.º 45, 2014, pp. 101 a 113, disponible en [http://www.scielo.org.co/pdf/tecn/v19n45/v19n45a08.pdf].

29 Violeta Valcárcel Asencios. "Datamining y el descubrimiento del conocimiento", *Revista de la Facultad de Ingeniería Industrial*, vol. 7, n.º 2, 2004, pp. 83 a 86, disponible en [https://www.researchgate.net/publication/307181857_DATA_MINING_Y_EL_DESCUBRIMIENTO_DEL_CONOCIMIENTO/fulltext/57c432b908aee5141be5bc8f/DATA-MINING-Y-EL-DESCUBRIMIENTO-DEL-CONOCIMIENTO.pdf].

30 Tomas Aluja Banet. "La minería de datos, entre la estadística y la inteligencia artificial", en *Qüesiió: quaderns d'estadística i investigació operativa*, vol. 25, n.º 3, 2001, pp. 479 a 498, disponible en [https://www.researchgate.net/profile/Tomas_Aluja-Banet/publication/28177489_La_mineria_de_datos_entre_la_estadistica_y_la_inteligencia_artificial/links/00b7d53b3b091899b7000000/La-mineria-de-datos-entre-la-estadistica-y-la-inteligencia-artificial.pdf].

naturaleza, cualitativas o cuantitativas. Si la variable Y es dicotómica, podrá tomar el valor "0" su el hecho no ocurre y "1" su el hecho ocurre. Este proceso es denominado binomial ya que solo tiene dos posibles resultados, siendo la probabilidad de cada uno de ellos constante en una serie de repeticiones[31].

B. Redes neuronales artificiales

Las RNAs tratan de emular el comportamiento del cerebro humano, caracterizado por el aprendizaje a través de la experiencia y la extracción de conocimiento genérico a partir de un conjunto de datos. Estos sistemas imitan en resumen la estructura neuronal del cerebro, bien mediante un programa de ordenador (simulación), bien mediante su modelado a través de estructuras de procesamiento con cierta capacidad de cálculo paralelo (emulación), o bien mediante la construcción física de sistemas cuya arquitectura se aproxima a la estructura de la red neuronal biológica o implementación de *hardware* de RNAs[32].

Fernando Villada, Diego Raúl Cadavid y Juan David Molina[33] mencionan que las redes neuronales artificiales son muy efectivas para resolver problemas complicados de clasificación y reconocimiento de patrones. La más utilizada es la llamada de propagación hacia adelante. La figura 3 muestra una red de propagación hacia adelante con dos capas ocultas. El número de entrada es de primera mano dependiente de la información disponible para clasificar mientras que el número de neuronas de salida es igual al número de clases a separar. Las unidades de una capa se conectan de forma unidireccional con las de la siguiente, en general todas con todas, sometiendo sus salidas a la multiplicación por un peso que es diferente para cada una de las conexiones.

---

31 Ana María Alderete. "Fundamentos del análisis de regresión logística en la investigación psicológica", *Revista evaluar*, vol. 6, n.º 1, 2006, pp. 52 a 67, disponible en [https://revistas.unc.edu.ar/index.php/revaluar/article/view/534/474].

32 Raquel Flóres López y José Miguel Fernández. *Las redes neuronales artificiales*, La Coruña, netbiblo, 2008.

33 Fernando Villada, Diego Raúl Cadavid y Juan David Molina. "Pronóstico del precio de la energía eléctrica usando redes neuronales artificiales", *Revista Facultad de Ingeniería Universidad de Antioquia*, n.º 44, junio de 2014, pp. 111 a 118, disponible en [http://www.scielo.org.co/pdf/rfiua/n44/n44a11.pdf].

C. Árboles de decisión

De acuerdo con Rocío Erandi Barrientos Martínez, Nicandro Cruz Ramírez, Héctor Gabriel Acosta Mesa *et al*.[34] un árbol de decisión es un modelo de predicción cuyo objetivo principal es el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Son muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva para la solución de un problema. El conocimiento obtenido durante el proceso de aprendizaje inductivo se representa mediante un árbol. Un árbol en detalle se representa por un conjunto de nodos, hojas y ramas. El nodo principal o raíz es el atributo a partir del cual se inicia el proceso de clasificación; los nodos internos corresponden a cada una de las preguntas acerca del atributo en particular del problema. Cada posible respuesta a los cuestionamientos se representa mediante un nodo hijo. Las ramas que salen de cada uno de estos nodos se encuentran etiquetadas con los posibles valores del atributo. Los nodos finales o nodos de hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver.

De acuerdo con Zuleyka Díaz Martínez[35], los árboles de decisión que se usan para problemas de clasificación son denominados a menudo "árboles de clasificación", y cada nodo terminal contiene una etiqueta que indica la clase predicha de un vector de características dado. Los árboles de decisión utilizados para problemas de regresión se denominan con frecuencia "árboles de regresión", y las etiquetas de los nodos terminales deben ser constantes o ecuaciones que especifican el valor output predicho de un vector *input* dado.

Un árbol de decisión se denomina "binario" cuando cada nodo interno tiene de manera exacta dos hijos. Estos son los más usados, debido

---

34 Rocío Erandi Barrientos Martínez, Nicandro Cruz Ramírez, Héctor Gabriel Acosta Mesa, Ivonne Rabatte Suárez, María del Carmen Gogeascoechea Trejo, Patricia Pavón León y Sobeida L. Blázquez Morales. "Árboles de decisión como herramiento en el disgnóstico médico", *Revista Médica de la Universidad Veracruzana*, vol. 9, n.º 2, 2009, pp. 19 a 24, disponible en [https://www.medigraphic.com/pdfs/veracruzana/muv-2009/muv092c.pdf].

35 Zuleyka Díaz Martínez. *Predicción de crisis empresariales en seguros no vida, mediante árboles de decisión y reglas de clasificación*, Madrid, Complutense, 2007, disponible en [https://eprints.ucm.es/48680/1/9788474918823.pdf].

a su simplicidad, aunque tampoco son infrecuentes los árboles que exhiben nodos con más de dos hijos.

D. Bosstrap

Esta técnica se enmarca entre los procedimientos de remuestreo, consistentes en generar un elevado número de muestras como base para estudiar el comportamiento de determinadas estadísticas. A nivel práctico, la actual facilidad para realizar procedimientos iterativos de manera informatizada elimina los posibles obstáculos que la aplicación de este tipo de métodos pudiera presentar. Esto implica desarrollar los siguientes pasos de modo general:

- A partir de una muestra original $\{X_1, X_2, X_3, X_4, ..., X_n\}$ se extrae una nueva muestra $\{X^*_1, X^*_2, X^*_3, X^*_4,..., X^*_n\}$ por medio de muestreo con reposición. Es decir, tras la estación de un primer elemento, este se repone en la muestra original de tal forma que podría ser elegido de nuevo como segundo elemento de la muestra extraída. De este modo, cada observación individual tiene una probabilidad 1/n de ser elegida cada vez, como si el muestreo se realizara sin reposición en un universo construido a partir de la información que provee la muestra.

– Para la muestra obtenida se calcula el valor de un determinado estadístico que se utiliza como estimador del parámetro poblacional, en cuyo estudio se muestra interés por parte de esta investigación.

– Repetir los dos pasos anteriores hasta obtener un elevado número de estimaciones. En este punto, el recurso a herramientas informáticas y determinación de las estimaciones resultará ineludible.

– Se construye una distribución empírica del estadístico, que representa una buena aproximación a la verdadera distribución de la probabilidad para ella. Es decir, se determina de este modo la distribución muestral de un estadístico sin haber hecho suposiciones sobre la distribución teórica a que esta se ajusta y sin manejar fórmulas analíticas para determinar los correspondientes parámetros de distribución[36].

---

36 Javier Gil Flores. "Aplicación del método de Bootstrap al contraste de hipótesis en la investigación educativa", *Revista de Educación*, n.º 336, 2005, pp. 251 a 261.

E. Bagging

De acuerdo a lo afirmado por Rosa Fátima Medina Merino y Carmen Ismelda Ñique Chacón[37] una manera de disminuir la varianza, y en consecuencia aumentar la certeza de la predicción de un método de aprendizaje estadístico, es seleccionar un elevado número de conjuntos de entrenamiento de la población y edificar un modelo de predicción independiente y emplear cada conjunto de entrenamiento. En otras palabras, se puede calcular, $h^1(x)$, $h^2(x)$..., $hB(x)$ al utilizar B conjuntos de entrenamiento por separado, y un promedio de ellas con el fin de obtener un único modelo de aprendizaje estadístico de varianza pequeña. Esto es:

$$\widehat{h_{avg}(x)} = \frac{1}{B}\sum_{b=1}^{B} \widehat{h^b}\,(x)$$

Bajo este enfoque se forman B distintos conjuntos de datos de entrenamiento siguiendo la técnica de *bootstrap*, para luego entrenar el modelo con el b^{avo} conjunto *bootstrap* de entrenamiento con el objetivo de conseguir $h^{*b}(x)$, de este promedio se obtiene:

$$\widehat{h_{bag}(x)} = \frac{1}{B}\sum_{b=1}^{B} \widehat{h^{+b}}\,(x)$$

37 Rosa Fátima Medina Merino y Carmen Ismelda Ñique Chacón. "Bosques aleatorios como extensión de árboles de clasificación con los programas R y Python", en *Portal de revistas Ulima*, diciembre de 2017, pp. 165 a 189, disponible en [file:///Users/optimus-prime/Downloads/Bosques_aleatorios_como_extension_de_los_arboles_d.pdf].

F. Algoritmo Cart

Según Leandro Kovalevski y Paula Macat[38], este algoritmo funciona tal y como se explica a continuación. Dado un conjunto de datos D = (X, Y), donde Y es la variable a explicar y X = (X1, …, Xp) es un vector de p cariables que describe a los individuos. El objetivo de CART es predecir los valores de Y a partir de los valores observados de las variables Xi, i = 1, …, p. Tanto la variable dependiente Y, como cada una de las variables explicativas Xi puede ser cuantitativa o cualitativa, esto dota a CART de una gran flexibilidad, pues se puede explicar en diferentes contextos. En el caso en que la variable dependiente Y sea cualitativa, se dice que CART es un árbol de clasificación cuyo objetivo es predecir la clasificación que le correspondería a un individuo con cierto perfil de valores en las variables explicativas. Por otra parte, si Y es cuantitativa, CART es llamado árbol de regresión y el objetivo es idéntico al de un modelo lineal, obtener una estimación del valor de Y asociado a cada nicho o perfil de predictores.

G. Algoritmo Random Forest

Según afirman Adriana Villa Murillo, Andrés Carrión García y Antonio Sozzi Rodríguez[39] esta técnica se sustenta en la elaboración de árboles de predicción mediante el empleo de *Bootstrap* y *Bagging*, lo que garantiza la estabilidad del proceso. Cada árbol es edificado por medio de muestras *bootstrap* con reposición a fin de corregir el error de predicción que se genera a raíz de la selección característica de una muestra y para disponer, por cada árbol, de una muestra independiente un *out-of-bag* para la estimación del error de clasificación, puesto

---

38 Leandro Kovalevski y Paula Macat. "Alternativas no paramétricas de clasificación multivariada", *Décimoséptimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística*, noviembre de 2012, Rosario, disponible en [https://www.fcecon.unr.edu.ar/web-nueva/sites/default/files/u16/Decimocuartas/kovalevski_macat_alternativas_no_parametricas.pdf].

39 Adriana Villa Murillo, Andrés Carrión García y Antonio Sozzi Rodríguez. "Optimización del diseño de parámetros: método Forest-Genetic univariante", en *Publicaciones en ciencias y tecnologías*, vol. 10, n.º 1, 2017, pp. 12 a 24, disponible en [https://dialnet.unirioja.es/descarga/articulo/6501229.pdf].

que alrededor de un tercio de la muestra original queda excluida de cada muestra generada por *boostrap*. Para cada división de un nodo, no se selecciona la mejor variable de entre todas como en CART, sino que se selecciona al azar un conjunto de variables de un tamaño prestablecido y se limita la selección de la variable de división a dicho conjunto. De esta forma se introduce una mayor diversidad de árboles y se reduce la dependencia del resultado con las divisiones precedentes.

El algoritmo usa el conjunto de datos de entrenamiento T, para luego crear k muestras mediante la técnica T_k, con estas muestras se construyen los árboles $h(x, T_k)$ y el promedio de ellos será el predictor bagget en el caso de regresión y el más botado para el caso de clasificación. En adelante, para cada (y, x) de T se construyen los árboles en cada T_k que no contienen a (y, x), esto son las muestras que quedaron fuera de las muestras *bootstrap*.

VILLA MURILLO, CARRIÓN GARCÍA y SOZZI RODRÍGUEZ[40] afirma que este algoritmo consta de los siguientes pasos:

– Se toman B muestras *bootstrap* de tamaño N del conjunto de entrenamiento.

– Se crean T_b (b = 1, ..., B) árboles con las muestras hasta que se obtiene el tamaño mínimo en el nodo terminal. Esto se logra de forma recursiva mediante los siguientes pasos:

1. Seleccionar de forma aleatoria m_{try} variables del conjunto total de P variables.
2. Seleccionar la óptima variable de división entre las p variables.
3. Dividir el nodo en dos nodos hijos.

– El conjunto de salida es el ensamble (promedio) de los árboles, es decir:

$$\hat{f}_{RF}^{B} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

– La estimación de la tasa de error o error de clasificación se obtiene mediante el conjunto OOB.

---

40 Ídem.

H. Algoritmo C5.0

Este algoritmo genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico. Se permiten varias decisiones en más de dos subgrupos[41].

Por otra parte, el C5.0 es el algoritmo sucesor de C4.5 que tiene como meta, la creación de árboles de clasificación. Entre sus características, destacan la capacidad para generar árboles de predicción simples, modelos basados en reglas, ensembles basados en *boosting* y asignación de distintos pesos a los errores. Este algoritmo ha resultado de una gran utilidad a la hora de crear modelos de clasificación y todas sus capacidades en r mediante el paquete C50.

I. Máquinas de soporte vectorial

De acuerdo con Cecilia Montt, Felix Castro y Nibaldo Rodríguez[42], las Máquinas de Soporte Vectorial –svm– constituyen una técnica de reconocimiento de patrones basada en la metodología de aprendizaje, al generar resultados robustos y satisfactorios. Fueron desarrolladas como una herramienta robusta y sólida para regresión y clasificación en dominios complejos. Su proceso de aprendizaje es supervisado, es decir, del ámbito predictivo. En clasificación supervisada, los casos pertenecientes al conjunto de datos tienen asignada una clase o etiqueta a priori, siendo el objetivo encontrar patrones o tendencias de los casos pertenecientes a una misma clase.

Por otro lado, Mauro Krikorian, Ana Ruedin y Leticia Seijas[43] afirman que debido a que los problemas tomados de la realidad son difíci-

---

41 Diego Vallejo P. y Germán Tenalanda V. "Minería de datos aplicada en la detección de intrusos", en *Ingenieras usbMed*, vol. 3, n.º 1, 2012, disponible en [https://dialnet.unirioja.es/descarga/articulo/4694116.pdf].

42 Cecilia Montt, Felix Castro y Nibaldo Rodríguez. "Análisis de accidentes de tránsito con máquinas de soporte vectorial LS-svm", *Revista de ingeniería de transporte*, vol. 15, n.º 2, 2011, pp. 7 a 14, disponible en [https://pdfs.semanticscholar.org/8416/eefa57a0d491f57bb9c7686cc10cc383949f.pdf].

43 Mauro Krikorian, Ana Ruedin y Leticia Seijas. "Reconocimiento de patrones utilizando transformadas wavelets sin submuestreo y máquinas de soporte vectorial", *xiv Reunión*

les de resolver con un clasificador lineal, el modelo es extendido para utilizar superficies de decisión no lineal. Se introduce la utilización de "kernels" con la idea de transformar el conjunto de datos a un espacio de dimensión superior donde este si es en lo perfecto separable, o separable bajo una cota de error aceptable.

Entre los kernels más utilizados por las SVM tenemos: función de base radial o gasussiana, lineal, polinómica, sigmoid.

III. Técnicas descriptivas o no supervisadas

En las técnicas descriptivas no se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables dependientes ni independientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean de forma automática a partir del reconocimiento de patrones. En este grupo se incluyen las técnicas de *clustering* y segmentación (que también son técnicas de clasificación en cierto modo), las técnicas de asociación y reducción de la dimensión (factorial, componentes principales, correspondencias, entre otros) y de escalonamiento multidimensional.

Tanto las técnicas predictivas como descriptivas están enfocadas al descubrimiento del conocimiento embebido en los datos[44]. En ese sentido, de acuerdo con Moreno García, Miguel Quintales, García Peñalvo y Polo Martín[45] estas técnicas descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas.

---

de Trabajo Procesamiento de la Información y Control, 2011, pp. 839 a 844, disponible en [http://dc.sigedep.exactas.uba.ar/media/academic/grade/thesis/krikorian.pdf].

44 Pérez López y Santín. *Minería de datos. Técnicas y herramientas*, cit,

45 Moreno García, Miguel Quintales, García Peñalvo y Polo Martín. "Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software", cit.

A. Clustering

De acuerdo con Cristina García Cambronero y Irene Gómez Moreno[46] se refiere al proceso de agrupar datos en clases o clusters de tal forma que los objetos de un *cluster* tengan una similaridad alta entre ellos y baja (sean muy diferentes) con objetos de otros clusters. Los mismos autores definen al cluster o grupo como un conjunto de objetos que son "similares" entre ellos y "diferentes" de los objetos que pertenecen a los otros grupos.

La palabra "cluster" viene del inglés y significa agrupación. Desde un punto de vista general, el *cluster* puede considerarse como la búsqueda automática de una estructura o de una clasificación en una colección de datos no etiquetados. Por otro lado, Miguel Garre, Juán José Cuadrado, Miguel Ángel Sicilia, Daniel Rodríguez y Ricardo Rejas[47] afirman que el proceso de *clustering* consisten en la división de los datos en grupos de objetos similares. Para medir la similitud entre objetos se suelen utilizar diferentes formas de distancia: distancia euclídea, de Manhatan, de Mahalanobis, entre otros. *Clustering* es una técnica más de aprendizaje automático, en la que el aprendizaje realizado es no supervisado. Desde un punto de vista práctico, el *clustering* juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales (tales como GIS o datos procedentes de astronomía), aplicaciones web, marketing, diagnóstico médico, análisis de ADN en biología computacional, y muchas otras.

De acuerdo con Joseph Hair, Rolph E. Anderson, Ronald L. Tatham y William C. Black[48], los algoritmos para la obtención de conglome-

---

46 Cristina García Cambronero y Irene Gómez Moreno. "Algoritmos de aprendizaje: KNN y KMEANS", en *Inteligencia en Redes de Telecomunicación*, 2012, pp. 6 y 7, disponible en [http://www.it.uc3m.es/jvillena/irc/practicas/08-09/06.pdf].

47 Miguel Garre, Juán José Cuadrado, Miguel Ángel Sicilia, Daniel Rodríguez y Ricardo Rejas. "Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software", *Revista Española de Innovación, Calidad e Ingeniería de Software*, vol. 3, n.º 1, 2007, pp. 6 a 22, disponible en [https://www.redalyc.org/pdf/922/92230103.pdf].

48 Joseph Hair, Rolph E. Anderson (comp.), Ronald L. Tatham (trad.) y William C. Black (trad.). *Analisis Multivariante*, Madrid, Prentince Hall, 1999.

rados más utilizados pueden clasificarse en dos categorías generales: (1) jerárquicos y (2) no jerárquicos. A continuación, se detallarán estas dos técnicas.

B. Clusters jerárquico: Dendograma

Estas técnicas consisten en la construcción de una estructura en forma de árbol. Existen en principio dos tipos de procedimientos de obtención de conglomerados jerárquicos: de aglomeración y divisivos.

Hair, Anderson, Tatham y Black[49] explican que, en los métodos de aglomeración, cada objeto u observación empieza dentro de su propio conglomerado. En estas etapas ulteriores, los dos conglomerados más cercanos (o individuos) se combinan en un nuevo conglomerado agregado, reduciendo así el número de conglomerados paso a paso. En algunos casos, un tercer individuo se une a los dos primeros en un conglomerado. En otros, dos grupos de individuos formados en un paso anterior pueden unirse en un nuevo conglomerado. En ocaciones, todos los individuos se agrupan en un único conglomerado; por esta razón, los procedimientos de aglomeración son denominados a veces como métodos de construcción.

En los métodos divisivos, empezamos con un gran conglomerado que contiene todas las observaciones (objetos). En los pasos sucesivos, las observaciones que son más diferentes se dividen y se construyen conglomerados más pequeños. Este proceso continúa hasta que cada observación es un conglomerado en sí mismo. Entre los métodos aglomerativos se encuentran los siguientes:

1. Enlace simple (*single linkage*)

En Hair, Anderson, Tatham y Black[50] se afirma que el Enlace simple se basa en la distancia mínima. Encuentra los dos objetivos separados por la distancia más corta y las coloca en el primer conglomerado. A continuación, se encuentra la distancia más corta, y o bien un tercer objeto

49 Ídem.
50 Ídem.

se une a los dos primeros para formar un conglomerado o se forma un nuevo conglomerado de dos miembros. El proceso continúa hasta que todos los objetos se encuentran en un conglomerado. Este procedimiento también se ha denominado como el enfoque del vecino más cercano, este algoritmo es conocido también como *nearest neighbor*.

Así, la distancia d_{AB} entre los conglomerados A y B se calculan mediante:

$$d_{AB} = min\ (d_{ij})$$

Donde (d_{ij}) es la distancia entre los elementos i y j, el primero pertenece al conglomerado A y el segundo al conglomerado B.

2. Enlace completo (*complete linkage*)

César Pérez López[51] afirma que este método considera como distancia entre dos grupos la existente entre vecinos más lejanos (*furthest neighbor*), es decir, entre los individuos más separados de ambos grupos (máxima distancia que es posible encontrar entre un caso de un cluster y un caso de otro). Presenta una excesiva tendencia a producir grupos de igual diámetro, y se ve muy distorsionado ante valores atípicos moderados.

La distancia entre dos conglomerados A y B se calcula como:

$$d_{AB} = max\ (d_{ij})$$

3. Enlace promedio (*avarege linkage*)

Manuel Ato García, José Antonio López Pina, Antonio Pablo Velandrino Nicolás y Julio Sánchez Meca[52] mencionan que este al-

---

51 César Pérez López. *Técnicas de análisis multivariante de datos*, Madrid, Pearson Prentice Hall, 2004, disponible en [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwj8xvP6r6HrAhVykuAKHcYKBgsQFjABegQIAhAB&url=https%3A%2F%2Fwww.academia.edu%2F39613182%2FT%25C3%25A9cnicas_de_an%25C3%25A1lisis_multivariante_de_datos_Aplicaciones_con_].

52 Manuel Ato García, José Antonio López Pina, Antonio Pablo Velandrino Nicolás y Julio Sánchez Meca. *Estadística avanzada con el paquete systat*, Murcia, Universidad de Murcia, 1990.

goritmo define la distancia como la media aritmética de todas las posibles entre dos puntos de dos conglomerados. Por otro lado, Juan José Marín Hernández[53] afirma que en este método la distancia entre dos conglomerados se calcula como la distancia promedio existente entre todos los pares de elementos de ambos conglomerados:

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij}$$

4. Enlace centroide (*centrod method*)

Ato García, López Pina, Velandrino Nicolás y Sánchez Meca[54] sostienen que este método define la distancia entre conglomerados como la distancia entre centroides de los dos conglomerados, siendo el centroide el promedio de todos los valores dentro del conglomerado. Tanto este método como el método del enlace simple tienen tendencia a generar conglomerados esféricos. Se dice que son métodos que imponen una estructura más que buscarla. J. Marín[55] señala que la distancia entre el conglomerado AB y el conglomerado C se calcula como:

$$d_{(AB)C} = \frac{n_A}{n_A + n_B} d_{AC} + \frac{n_B}{n_A + n_B} d_{BC} - \frac{n_A n_B}{(n_A + n_B)^2} d_{AB}$$

5. La mediana (*median method*)

J. Marín[56] afirma que el método de agrupación de medianas, los dos conglomerados (elementos) que se combinan reciben idéntica ponderación en el cálculo del nuevo centroide combinado, independiente

---

53 Juan José Marín Hernández. "Análisis de conglomerados (II): El procedimiento Conglomerados jerárquicos", Universidad Carlos III de Madrid, 2014, disponible en [http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/GuiaSPSS/22conglj.pdf].

54 Ídem.

55 Ídem.

56 Ídem.

del tamaño cada uno de los conglomerados (o elementos). La matriz de distancias utilizada en cada etapa para los cálculos es la matriz del paso previo. Dado un conglomerado AB y un elemento C, la nueva distancia del conglomerado al elemento se calcula como:

$$d_{(AB)C} = \frac{d_{AC} + d_{BC}}{2} - \frac{d_{AB}}{4}$$

6. Enlace por mínima varianza o de Ward

César Pérez López[57] llega a la determinación de que en el método Ward se calcula la media de todas las variables de cada cluster, luego se calcula la distancia euclídea al cuadrado entre cada individuo y la media de su grupo y después se suman las distancias de todos los casos. En cada paso, los clusters que se forman son aquellos que resultan con el menor incremento en la suma total de las distancias al cuadrado intracluster. La métrica por lo general considerada en los métodos hasta aquí descritos es la euclídea o la euclídea al cuadrado. Esta última se suele usar por omisión en programas estadísticos.

Rafael Álvarez Cáceres[58] menciona que con respecto al método Ward al unir dos grupos, la varianza aumenta. El método de Ward calcula cuál sería la varianza de dos grupos, en caso de unirlos, uniendo en el paso siguiente aquellos grupos cuya varianza sea mínima. En caso de tener en cuenta más de una variable en lugar de la varianza, se unen los grupos cuya inercia (suma de diagonal principal de la matriz de varianzas y covarianzas) sea mínima.

---

57 César Pérez López. *Técnicas de análisis multivariante de datos*, Madrid, Pearson Prentice Hall, 2004, disponible en [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwj8xvP6r6HrAhVykuAKHcYKBgsQFjABegQIAhAB&url=https%3A%2F%2Fwww.academia.edu%2F39613182%2FT%25C3%25A9cnicas_de_an%25C3%25A1lisis_multivariante_de_datos_Aplicaciones_con_].

58 Rafael Álvarez Cáceres. *Estadística multivariante y no paramétrica con SPSS*, Madrid, Ediciones Díaz de Santos, 1995.

C. Clusters no jerárquicos

Estos algoritmos permiten clasificar individuos (no son válidos para variables) en una clasificación de K *clusters*, donde K se especifica a priori. Damaris Pascual González[59] menciona que los algoritmos de partición tratan de descubrir *clusters* al reubicar en forma iterativa puntos entre subconjuntos. Por ejemplo, los métodos K-Medias y el K-Medoides (pam, clara, clarans), también pueden tener un enfoque probabilístico (EM, autoClass, MClust).

1. K Means

Es el método más utilizado en la actualidad para realizar *clustering*. Según Roberto Camana Fiallos[60] es un algoritmo de clasificación no supervisado, inventado por James B. MacQueen en 1967, mediante el cual el espacio de patrones de entrada se divide en K clases o regiones, cada una representada por un punto llamado centroide. Dichos centros se determinan con el objetivo de minimizar las distancias euclídeas entre los patrones de entrada y el centro más cercano.

Claudio Nicolás Flores Cartes[61] explica que en este algoritmo primero se eligen K centroides iniciales, donde K es un parámetro especificado por el usuario y corresponde al número de *clusters* deseados. Cada punto es asignado a su centroide más cercano y cada colección de puntos asignado a un centroide representa un cluster. El centroide de cada *cluster* se actualiza basado en la asignación de puntos al *cluster*. Se repiten los pasos de asignación y actualización hasta que los puntos

---

59 Damaris Pascual González. "Algoritmos de agrupamiento basados en densidad y validación de clusters", tesis doctoral, Castellón de Plana, Universidad Jaume I, marzo de 2010, disponible en [http://www.cerpamid.co.cu/sitio/files/DamarisTesis.pdf].

60 Roberto Camana Fiallos. "Aplicación de técnicas de minería de datos para la indagación y estudio de resultados electorales", en *CienciAmérica: revista de divulgación científica de la Universidad Tecnológica Indoamérica*, vol. 7, n.º 1, enero de 2012, pp. 85 a 94, disponible en [http://cienciamerica.uti.edu.ec/openjournal/index.php/uti/article/view/10/8].

61 Claudio Nicolás Flores Cartes. "Exigencias de calidad de suministro en base a densidad de consumo mediante técnicas de minería de datos", tesis de pregrado, Santiago de Chile, Universidad de Chile, 2014, disponible en [http://repositorio.uchile.cl/bitstream/handle/2250/115571/cf-flores_cc.pdf?sequence=1&isAllowed=y].

dentro del *cluster* no cambien, o en forma equivalente, hasta que los centroides dejen de cambiar.

2. Partiotining Around Medoids –PAM–

Ignacio Benítez y José Luis Diez[62] afirma que PAM es una extensión del algoritmo *K-means*, en donde cada grupo o cluster está representado por un medoide en vez de un centroide. El medoide es el elemento más céntrico posible del cluster al que pertenece; similar al centroide, pero no por necesidad, ya que el centroide representa el valor patrón o medio del conjunto, que no siempre coincide con el más céntrico. El procedimiento para el agrupamiento es similar al del *K-means*.

3. Expectation-Maximization –EM–

De acuerdo con Benítez[63] este algoritmo asigna cada objeto a un *cluster* predefinido, según la probabilidad de pertenencia del objeto a ese grupo concreto. Como modelo se usa una función de distribución gaussiana, siendo el objetivo el ajuste de sus parámetros, según cómo los distintos objetos del conjunto se ajustan a la distribución en cada *cluster*.

---

62 Ignacio Benítez y José Luis Diez. "Técnicas de agrupamiento para el análisis de datos cuantitativos y cualitativos", en *Reaserchgate*, Valencia, septiembre de 2005, disponible en [https://www.researchgate.net/profile/Ignacio_Benitez/publication/239526131_Tecnicas_de_Agrupamiento_para_el_Analisis_de_Datos_Cuantitativos_y_Cualitativos/links/00b7d51c15cca2cb1f000000/Tecnicas-de-Agrupamiento-para-el-Analisis-de-Datos-Cuantitativos-y-Cu].

63 Ídem.

Capítulo tercero
Metodologías para la minería de datos

Juan Miguel Moine, Silvia Ethel Gordillo y Ana Silvia Haedo[64] afirman que las metodologías permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Ayudan a las organizaciones a entender el proceso de descubrimiento de conocimiento y proveen una guía para la planificación y ejecución de los proyectos

Algunos modelos conocidos como metodologías son en realidad un modelo de proceso: un conjunto de actividades y tareas organizadas para llevar a cabo un trabajo. La diferencia fundamental entre metodología y modelo de proceso radica en que el modelo de proceso establece qué hacer, y la metodología especifica cómo hacerlo. Una metodología no solo define las fases de un proceso sino también las tareas que deberían realizarse y cómo llevar a cabo las mismas.

De acuerdo con revisión de la literatura científica tres son los modelos de proceso de minería de datos en el presente utilizados: el modelo *Cross Industry Standard Process for Data Mining* –KDD, CRISP-DM– y *Sample, Explore, Modify, Model, Assess* –SEMMA–. A continuación, se detallarán los modelos de proceso de minería de datos CRISP-DM y SEMMA.

I. *Cross Industry Standard Process for Data Mining* –CRISP-DM–

Moine, Gordillo y Haedo[65] afirman que este modelo fue creado por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es la

64 Juan Miguel Moine, Silvia Ethel Gordillo y Ana Silvia Haedo. "Análisis comparativo de metodologías para la gestión de proyectos de minería de datos", en *SEDICI*, 2012, disponible en [http://hdl.handle.net/10915/18749].

65 Ídem.

guía de referencia más utilizada para el desarrollo de proyectos de minería de datos. Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación. La sucesión de fases, no es por necesidad rígida. Cada fase se descompone en varias tareas generales de segundo nivel. CRISP-DM establece un conjunto de tareas y actividades para cada fase del proyecto, pero no especifica cómo llevarlas a cabo.

A. Fase de comprensión del problema o negocio

Como afirma José Alberto Gallardo Arancibia esta fase es:

> probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables[66].

Las principales tareas de esta fase son:

– Determinar los objetivos del negocio

Involucra determinar cuál es el alcance del proyecto, para ello nos formulamos las siguientes interrogantes: ¿cuál es el problema que queremos resolver?, ¿qué es lo que queremos lograr?, ¿qué beneficio ofreceremos a al cliente?, ¿por qué es necesario aplicar la minería de datos? Así también se determinan los criterios de éxito del objetivo del negocio. Estos criterios pueden ser de carácter cuantitativo o cualitativo.

– Evaluación de la situación actual

En el desarrollo de esta tarea se debe considerar el estado de la situación antes de iniciar el proyecto de minería de datos. En ese sentido, las siguientes interrogantes serán de ayuda: ¿cuáles son los diversos recursos

---

66 José Alberto Gallardo Arancibia. "Metodología para la definición de requisitos en proyectos de data mining", tesis doctoral, Madrid, Universidad Politécnica de Madrid, 2009, p. 17.

o requerimientos (*software*, *hardware* o recursos humanos) que se van a necesitar o con los que vamos a trabajar? ¿cuál es el conocimiento previo sobre el problema? ¿cuáles son los supuestos y limitaciones? ¿cuál es la relación beneficio-costo del proyecto de minería de datos?

En esta parte se definen los requisitos tanto en términos de negocio como en términos de minería de datos.

– Determinación de los objetivos de minería de datos

Gallardo Arancibia, afirma de manera textual:

> esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de DM, como, por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de DM será, por ejemplo, determinar el perfil de los clientes respecto de su capacidad de endeudamiento[67].

– Producción del plan de proyecto

Siendo esta la última tarea de la primera fase, su finalidad es implementar el desarrollo de un plan para el proyecto, que tome en cuenta los pasos a seguir y las técnicas a emplear en cada uno de los mencionados.

B. Fase de comprensión de los datos

Esta segunda fase está relacionada con la recolección inicial de los datos. Se puede decir que la idea principal aquí es la familiarización con los datos e información que se va a manejar, así como evaluar su calidad e identificar relaciones.

Comprende las siguientes tareas:

– Recolección inicial de datos

Constituye la primera tarea en esta segunda fase de la metodología crisp-dm, como menciona Oldemar Rodríguez:

---

67 Ibíd., p. 18.

> comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de DM. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos ad-hoc al proyecto de DM, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y abundantes accesos a la base de datos a objeto de realizar consultas y probablemente modificaciones, lo cual podría generar muchos problemas[68].

– Descripción de los datos

Esta tarea consiste en detallar los datos iniciales, su verificación, el significado de cada campo y la descripción de su formato inicial.

– Exploración de los datos

Esta tarea consiste en la exploración cuyo fin es encontrar una estructura general para los datos. Esto implica la aplicación de pruebas estadísticas como tablas de distribución de frecuencias, gráficos de distribución que revelen propiedades de los datos recién adquiridos.

– Verificación la calidad de los datos

Consiste en la verificación de los datos con la finalidad de determinar la consistencia de los valores de cada uno de los campos, la cantidad, distribución de los valores nulos y valores atípicos, los mismos que pueden ocasionar ruido en el proceso. La finalidad de esta tarea es garantizar la completitud y corrección de los datos.

C. Fase preparación de los datos

Esta fase tiene como finalidad la preparación de los datos en función a las técnicas de minería de datos que se aplicarán, tales como las de vi-

---

68 Oldemar Rodríguez (s. f.). *Metodología para el desarrollo de proyectos en Minería de datos CRISP-DM,* disponible en [http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037], p. 6.

sualización de datos, de búsqueda de relaciones entre variables u otras medidas para la exploración de datos.

Comprende las siguientes tareas:

– Preparación de los datos

Esta tarea consiste en seleccionar un subconjunto de los datos que se obtuvieron en la etapa anterior, apoyándose en criterios antes establecidos en las fases pasadas (calidad de los datos, corrección de los datos, limitaciones en el volumen y que los tipos de datos estén acorde a las técnicas de minería de datos a utilizar).

– Limpieza de los datos

Esta tarea comprende la aplicación de técnicas orientadas a la optimización de la calidad de los datos con el objetivo de prepararlos para la fase siguiente (modelado). Algunas de las técnicas utilizadas en esta tarea son: normalización de los datos, discretización de los campos numéricos, tratamiento de valores faltantes, reducción del volumen de datos, entre otros.

– Estructuración de los datos

Esta tarea consiste en la generación de nuevos atributos, a partir de los atributos ya existentes, un ejemplo de esto es: si poseemos una tabla histórica mensual de ventas, podemos crear nuevos atributos en ella como el promedio.

– Integración de los datos

Tarea que consiste en combinar información de diferentes tablas o registros con el fin de generar nuevos campos o registros.

– Formateo de los datos

Esta tarea se basa en la transformación de datos para realzar un análisis correcto de estos. Un ejemplo de esto podría ser en el caso en que

se desee predecir un valor numérico. Para ello, los datos deben presentarse en dicho formato.

D. Fase de modelado

En esta fase se seleccionan las técnicas de modelado que más se ajusten al proyecto de minería de datos. Estas técnicas se eligen teniendo en cuenta los siguientes criterios: ser apropiada al problema, disponer de datos adecuados, cumplir con los requisitos del problema, tiempo adecuado para obtener el modelo y conocimiento de la técnica.

Esta fase tiene cuatro tareas generales que se detallan a continuación:

– Selección de la técnica de modelado

Esta tarea se refiere de manera específica a la selección de la técnica de minería de dato más apropiada al problema a resolver. Para ello se debe considerar el objetivo principal del proyecto y su relación con las herramientas de minería de datos.

– Generalización del plan de prueba

Consiste en generar un procedimiento para probar la calidad y la eficiencia del modelo construido. De forma habitual se divide los datos en dos conjuntos uno de entrenamiento y otro de prueba, para después construir un modelo basado en el conjunto de entrenamiento y medir la calidad de este en el conjunto de prueba.

– Construcción del modelo

Consiste en ejecutar la herramienta de modelado sobre los datos preparados antes con el fin de crear uno o más modelos. Todas las técnicas de modelado poseen un conjunto de parámetros que determinan las características del modelo a generar.

– Evaluación del modelo

Esta tarea consiste en la evaluación y revisión de parámetros del modelo. En esta tarea participan los ingenieros de minería de datos y los

expertos en el dominio del problema los cuales juzgan los modelos dentro del contexto del dominio.

E. Fase de evaluación

En esta fase se verifica el cumplimiento de los criterios de éxito preestablecidos, ya sean del negocio como los de minería de datos. Esto es, evaluar todos los resultados u observaciones del proceso de modelamiento.

Comprende las siguientes etapas:

– Evaluación de los resultados

Esta tarea involucra la evaluación del modelo en términos de los objetivos del negocio y pretende determinar si existe alguna razón del negocio para la cual en modelo aun es deficiente. Es recomendable probar el modelo en situaciones reales siempre en cuando el tiempo y las restricciones lo permitan.

– Determinación de futuras clases

Esta tarea consiste verificar los resultados generados hasta el momento. De ser así, es posible pasar a la fase siguiente; de lo contrario, se podría decidir por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. En ocasiones es probable que se decida otra vez con un nuevo proyecto de minería de datos.

F. Fase de implementación

En esta fase se transforma el conocimiento obtenido en acciones dentro del proceso de negocio. Esto pude darse como recomendación del analista al aplicar el modelo a diferentes conjuntos de datos o como parte del proceso.

Los proyectos de minería de datos no concluyen con la implantación del modelo, por lo tanto, se deben documentar y presentar los resultados de manera entendible para el usuario.

Las tareas que se llevan a cabo en esta fase son: plan de implementación, monitoreo y mantenimiento, informe final y por último la revisión del proyecto.

Comprende las siguientes etapas:

– Plan de implementación

Esta tarea consiste en elaborar una estrategia de implementación con base en los resultados obtenidos en la fase evaluación.

– Monitorización y mantenimiento

Respecto a esta etapa, Gallardo Arancibia afirma:

> si los modelos resultantes del proceso de Data Mining son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente[69].

– Informe final

Esta tarea consiste en preparar un informe que contenga un resumen de los hitos del proyecto, así como los resultados obtenidos del mismo.

– Revisión del proyecto

Esta labor involucra a la evaluación de lo que fue correcto e incorrecto, es decir, que es lo que se está bien y que es lo que se necesita mejorar.

II. *Sample, Explore, Modify, Model, Access* –semma–

De acuerdo con Paola Verónica Britos[70], esta metodología es el proceso de selección, exploración y modelado de grandes cantidades de

69 Ibíd., p. 24.

70 Paola Verónica Britos. "Procesos de explotación de información basados en sistemas inteligentes", tesis doctoral, Buenos Aires, Universidad Nacional de la Plata, agosto de 2008, disponible en [http://sedici.unlp.edu.ar/bitstream/handle/10915/4142/Documento_completo.pdf?sequence=1&isAllowed=y].

datos para descubrir patrones de negocio desconocidos. El nombre de esta terminología corresponde con el acrónimo –en inglés–correspondiente a las cinco fases básicas del proceso.

De acuerdo con Claudia L. Hernández G. y María Ximena Dueñas R.[71] detallamos cada uno de las fases:

– Muestreo

En esta fase se realiza la extracción de una muestra de los datos que permita representar características comunes de la población para más adelante comenzar el análisis de los mismos. Se logra facilitar los procesos de minado sobre los datos, reduciendo costes y tiempo para la organización.

– Exploración

La exploración de datos a través de técnicas estadísticas permite realizar un seguimiento a los mismos al lograr detectar, identificar y, más adelante, eliminar datos que representen anomalías o deficiencias en las fases siguientes hacia el descubrimiento de información.

– Modificación

En esta fase se realiza una selección y transformación de los datos de acuerdo con las variables seleccionadas para el proceso de minado, la cual permitirá adaptar el enfoque de selección y diseño del modelo.

– Modelado

En este punto de la metodología se hace uso de herramientas de *software* que permitan la utilización de técnicas y métodos propios de la minería de datos, las cuales tiendan hacia el descubrimiento de asocia-

71 Claudia L. Hernández G. y María Ximena Dueñas R. *Hacia una metodología de gestión del conocimiento basada en minería de datos*, comtel, 2009, pp. 79 a 96, disponible en [http://repositorio.uigv.edu.pe/bitstream/handle/20.500.11818/982/COMTEL-2009-80-96.pdf?sequence=1&isAllowed=y].

ciones o combinaciones entre los datos, al lograr así la predicción de resultados con un alto nivel de confianza. Entre las técnicas más utilizadas para el modelado de datos, se encuentran: métodos estadísticos, de agrupamiento, redes neuronales, árboles de decisión, lógica difusa, reglas de asociación, entre otros.

– Evaluación

Uno de los pasos principales dentro de una metodología es la valoración de la solución. A partir del modelo obtenido en la fase anterior se realiza una evaluación de resultados para verificar el éxito del proyecto. Una buena práctica para comprobar la validez del modelo es seleccionar otra muestra de datos y aplicarlo para verificación de resultados, si este resulta optimo se procede con el proceso de producción, en caso contrario se desarrollará otro modelo.

III. Técnicas para evaluar clasificadores

Existen medidas de desempeño para clasificadores binarios y multiclase como la *Accuracy* (Exactitud), *Classification-error* (Error de clasificación), *kappa coefficient* (Coeficiente de kappa), estas métricas permiten comparar entre varias técnicas de clasificación y seleccionar la que tenga mayor precisión. En este estudio se propone la evaluación mediante la Matriz de confusión y *kappa coefficient*.

– Matriz de confusión

Es una tabla de doble entrada donde se muestra la clasificación observada (real) y la clasificación predicha (mediante el clasificador propuesto) para las distintas clases de la variable objetivo. En la tabla 1 se observa una matriz de confusión para dos clases:

Tabla 1
Matriz de confusión

Clasificación observada	Clasificación predicha		Total, observado
	Positiva (clase 0)	Negativa (clase 1)	
Positiva (clase 0)	VP	FN	VP + FN
Negativa (clase 1)	FP	VN	FP + VN
Total, predicho	VP + FP	FN + VN	N

Fuente: elaboración propia.

Donde VP son los verdaderos positivos y VN verdaderos negativos que vienen a ser la cantidad de observaciones que el clasificador predijo de forma correcta como la clase positiva y negativa. Los FP son los falsos positivos y los FN, falsos negativos los cuales componen la cantidad de observaciones que el clasificador predice de manera incorrecta como clase positiva siendo la clase negativa y como negativa siendo la clase positiva en lo respectivo. A partir de esta tabla se puede calcular la exactitud, el error de clasificación.

$$\text{Exactitud} = \frac{VP + VN}{N}$$

Este valor mide la proporción de las observaciones que fueron clasificados de forma correcta por el modelo predictivo de clasificación.

$$\text{Tasa de error} = \frac{FP + FN}{N}$$

Este valor mide la proporción de las observaciones que fueron clasificados de manera incorrecta por el modelo predictivo de clasificación.

Donde:

$$N = VP + VN + FP + FN$$

– Coeficiente de *Kappa* (k)

De acuerdo con Jaime Cerda Lorca y Luis Villarroel[72], se razona que el coeficiente kappa refleja la concordancia inter-observador y puede ser calculado en tablas de cualquier dimensión, siempre y cuando se contrasten dos observadores (para la evaluación de concordancia de tres o más observadores se utiliza el coeficiente *kappa* de Fleiss, cuya explicación excede el propósito del presente artículo). El coeficiente kappa puede tomar valores entre - 1 y + 1. Mientras más cercano a + 1, mayor es el grado de concordancia inter-observador, por el contrario, mientras más cercano a - 1, mayor es el grado de discordancia inter-observador.

En la Tabla 2, se presenta la valoración del valor de k propuesta por Landis y Koch en 1977.

Tabla 2
Valoración del coeficiente de kappa

Coeficiente de kappa	**Fuerza de concordancia**
0,00	Pobre (*Poor*)
0,01-0,20	Leve (*Slight*)
0,21-0,40	Aceptable (*Fair*)
0,41-0,60	Moderada (*Moderate*)
0,61-0,80	Considerable (*Subtantial*)
0,81-1,00	Casi perfecta (*Almost perfect*)

Fuente: Landis y Koch (como se cita en Cerda Lorca y Villarroel, 2008).

72 Jaime Cerda Lorca y Luis Villarroel. "Evaluación de concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa", *Revista chilena de pediatría*, vol. 79, n.º 1, 2008, pp. 54 a 58.

IV. Rendimiento académico

De acuerdo con Yesica Noelia Reyes Tejada[73], el rendimiento académico es un indicador del nivel de aprendizaje alcanzado por el alumno, por ello, el sistema educativo brinda tanta importancia a dicho indicador. En tal sentido, el rendimiento académico se convierte en una tabla imaginaria de medida para el aprendizaje logrado en el aula, que constituye el objetivo central de la educación. Sin embargo, en el rendimiento académico, intervienen muchas otras variables externas al sujeto, como la calidad del maestro, el ambiente de clase, la familia, el programa educativo, etc., y variables psicológicas o internas, como la actitud hacia la asignatura, la inteligencia, la personalidad, el autoconcepto del alumno, la motivación, entre otros factores.

73 Yesica Noelia Reyes Tejada. Relación entre el rendimiento académico, la ansiedad ante los exámenes, los rasgos de personalidad, el autoconcepto y la asertividad en estudiantes de primer año de psicología de la UNMSM, Lima, SISBIB, 2003, disponible en [http://sisbib.unmsm.edu.pe/bibvirtual/tesis/salud/reyes_t_y/cap2.htm].

Capítulo cuarto
Detección de patrones de bajo rendimiento académico mediante técnicas de minería de datos de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios 2018

Diana Heredia, Yegny Amaya y Edwin Barrientos[74] manifiestan que la minería de datos permite descubrir información oculta en grandes cantidades de datos, lo cual es muy difícil de visualizar con la aplicación de un proceso tradicional. Este tema de la informática permite la manipulación y clasificación de grandes cantidades de datos. Por ejemplo, se ha demostrado que el árbol de decisión C4.5 e ID3 es eficiente para casos de predicción específicos. Este artículo muestra la construcción de un modelo predictivo de deserción estudiantil, que caracteriza a los estudiantes de la Universidad Simón Bolívar para predecir la probabilidad de que un estudiante abandone su programa académico, mediante dos técnicas de minería de datos y comparación de resultados. Para crear el modelo se utilizó el software WEKA que contiene múltiples herramientas eficientes para el procesamiento de datos.

En referencia a lo anterior, Gladys N. Dapozo, Eduardo Porcel, María Victoria López, Verónica S. Bogado, y Roberto Bargiela[75] mencionan que la minería de datos abarca una variedad de métodos

74 Diana Heredia, Yegny Amaya y Edwin Barrientos. "Student Dropout Predictive Model Using Data Mining Techniques", en *ieee Latin America Transactions*, vol. 13, n.º 9, 2015, pp. 3127 a 31234.

75 Gladys N. Dapozo, Eduardo Porcel, María Victoria López, Verónica S. Bogado, y Roberto Bargiela. "Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de Sistemas de la facena-unne", *sedici*, Buenos Aires, junio de 2016, disponible en [http://sedici.unlp.edu.ar/handle/10915/20797].

estadísticos y computacionales para investigar la existencia de relaciones y patrones de comportamiento en almacenamientos electrónicos de datos. En este trabajo se presenta un estudio a través de técnicas de minería de datos que permiten determinar, a través de un clasificador, el rendimiento académico de los alumnos ingresantes de la carrera de Licenciatura en Sistemas de Información de la Facultad de Ciencias Exactas de la Universidad Nacional del Nordeste –FACENA-UNNE–, Corrientes, Argentina. Se llevó a cabo un estudio comparativo de diferentes algoritmos clasificadores disponibles en el *software* Weka, de libre distribución, y se seleccionó el que ofrecía mejores resultados.

Desde un punto de vista regional, Adolfo Carlos Jiménez Chura[76], en su tesis doctoral desarrollado en la ciudad de Puno, departamento de Puno, en Perú, entre los años 2016-2017, tuvo como objetivo principal predecir la tendencia de postulantes e ingresantes a las Escuelas Profesionales y su formación en las escuelas de educación secundaria, públicas y privadas, y en función a estos resultados establecer políticas adecuadas en las Escuelas Profesionales de la Universidad Nacional del Altiplano y las escuelas de educación secundaria de la región de Puno. Se usó como referencia la metodología *Cross Industry Standard Process for Data Mining* –CRISP-DM–, para los modelos de predicción se usó el *software* R y paquetes adicionales como *RMySQL*, *dplyr*, *ggplot*, *polynom*, entre otros. Estos permitieron procesar, analizar, graficar e interpretar la información acerca de los postulantes e ingresantes en los procesos de admisión general y cepreuna. El resultado obtenido, con los modelos lineales y polinómicos permitieron predecir y confirmar el nivel de crecimiento de las escuelas profesionales como ingeniería civil, ciencias contables, etc. La escuela de educación secundaria de la Gran Unidad Escolar San Carlos cuenta con una mayor cantidad de postulantes, sin embargo, la mayor cantidad de ingresantes es de la escuela Santa Rosa, lo que indica que sus estudiantes poseen una mejor formación.

76 Adolfo Carlos Jiménez Chura. "Análisis predictivo para los procesos de admisión de la Universidad Nacional del Altiplano - Puno", tesis de doctorado, Puno, Universidad Nacional del Altiplano, 2017, disponible en [https://1library.co/document/q2nod6jq-analisis-predictivo-procesos-admision-universidad-nacional-altiplano-puno.html?tab=pdf].

La Dirección Universitaria de Asuntos Académicos –DUAA– de la Universidad Nacional Amazónica de Madre de Dios, en Perú, en la actualidad cuenta con un sistema de información denominado "OPULUS-Sistema Académico" para la gestión de los procesos como: matricula, carga académica, evaluación docente, gestión de notas y rendimiento académico, estos datos permitieron descubrir conocimiento oculto mediante la aplicación de técnicas predictivas de minería de datos, en específico los árboles de clasificación *Random Forest*, C5.0 y CART.

La Dirección Universitaria de Asuntos Académicos –DUAA– de la Universidad Nacional Amazónica de Madre de Dios en la actualidad cuenta con un sistema de información denominado OPULUS-Sistema Académico para la gestión de los procesos como: matricula, carga académica, evaluación docente, gestión de notas y rendimiento académico. Este sistema viene almacenando los datos desde el año 2015. Al considerar que la misión de esta universidad es la de formar profesionales con orientación humanística, científica y tecnológica en el estudiante, contribuyendo al desarrollo sostenible de la biodiversidad con identidad cultural y responsabilidad social[77]. En ese sentido, uno de los aspectos más importantes a ser analizados y evaluados para cumplir con esta misión es el rendimiento académico de los estudiantes, debido a que se constituye un indicador importante a la hora de valorar la calidad educativa en la educación superior[78].

I. Método de investigación

Dado que la naturaleza del presente estudio implicó la aplicación del conocimiento en la solución de problemas prácticos, se ubica dentro de la investigación aplicada, debido a que no habrá manipulación de variables dependientes y, dado que el origen de los datos proviene de base de datos electrónicas, el diseño de investigación del presente es-

77 UNAMAD. *Plan estratégico institucional UNAMAD 2017-2019*, Puerto Maldonado, abril de 2016, disponible en [http://www.unamad.edu.pe/index.php/descargas/send/24-institucionales/5468-pei-unamad-2017-2019].

78 Guiselle María Garbanzo Vargas. "Factores asociados al rendimiento académico en estudiantes universitarios. Una reflexión desde la calidad de la educación superior pública", *Revista Educación*, vol. 37, n.º 1, 2007, pp. 43 a 63, disponible en [https://www.redalyc.org/pdf/440/44031103.pdf].

tudio fue: no experimental-documental[79], la metodología optada para el logro de los objetivos fue CRISP-DM.

II. Métodos por objetivos

Para el logro del objetivo específico n.º 1: Se identificaron las variables más influyentes en el modelo de clasificación para el bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios, se emplearon las técnicas de clasificación de minería de datos, para tal efecto se utilizó el algoritmo *Random Forest*, las herramientas utilizadas fueron el lenguaje de programación R y el entorno de desarrollo integrado RStudio, para el uso del algoritmo *Random Forest* se tuvo que agregar el paquete *randomForest* al IDE RStudio.

Para cumplir con el objetivo específico n.º 2: Se estableció el modelo de clasificación que permitió predecir las condiciones que cumplieron los estudiantes con bajo rendimiento académico de la Universidad Nacional Amazónica de Madre de Dios. Se emplearon las técnicas de clasificación de minería de datos, para tal efecto se utilizaron los algoritmos *Random Forest*, C5.0 y CART las herramientas utilizadas fueron el lenguaje de programación R y el entorno de desarrollo integrado *RStudio*, para el uso del algoritmo C5.0 y CART se tuvo que agregar los paquetes C50 y rpart al IDE *RStudio*, a respecto.

La identificación de los perfiles de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios fue posible mediante las técnicas de clasificación de minería de datos, se empleó el árbol de clasificación CART implementado en el paquete *rpart* de *RStudio*, para la representación gráfica del árbol se utilizó la librería rpart.plot del mismo.

---

79 Fidias G. Arias. *El proyecto de investigación. Introducción a la metodología científica*, Caracas, Episteme, 2006, disponible en [https://ebevidencia.com/wp-content/uploads/2014/12/EL-PROYECTO-DE-INVESTIGACI%C3%93N-6ta-Ed.-FIDIAS-G.-ARIAS.pdf].

III. Lugar de estudio

El presente estudio se realizó en la Universidad Nacional Amazónica de Madre de Dios, ubicado en la ciudad de Puerto Maldonado, departamento de Madre de Dios, provincia Tambopata y distrito Tambopata.

IV. Población

La población del presente estudio estuvo constituida por las instancias de los estudiantes matriculados desde el año 2001 hasta el semestre 2018-I, de la Universidad Nacional Amazónica de Madre de Dios, que ascienden a 9.545 registros.

Tabla 3
Registros de proceso de matrícula UNAMAD del 2001 al 2018

01	2001-I, 2001-ii	319
02	2002-I, 2002- ii	150
03	2003-I, 2003- ii	302
04	2004-I, 2004- ii	260
05	2005-I, 2005- ii	319
06	2006-I, 2006- ii	397
07	2007-I, 2007- ii	306
08	2008-I, 2008- ii	306
09	2009-I, 2009- ii	411
10	2010-I, 2010- ii	859
11	2011-I, 2011-ii	801
12	2012-I, 2012- ii	517
13	2013-I, 2013- ii	740
14	2014-I, 2014- ii	531
15	2015-I, 2015- ii	734
16	2016-I, 2016- ii	1.005
17	2017-I, 2017- ii	1.060
18	2018-i	528
	Total	9.545

Fuente: Dirección Universitaria de Asuntos Académicos –DUAA– de la Universidad Nacional Amazónica de Madre de Dios.

V. Muestra

Dado que el presente estudio utilizó técnicas de minería de datos para descubrir patrones en grandes volúmenes de datos, se optó por trabajar con toda la población.

VI. Objetivo general

Detectar los patrones de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios, mediante el uso de minería de datos.

VII. Objetivos específicos

- Identificar las variables más influyentes en el modelo de clasificación para el bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios.

- Establecer el modelo de clasificación que permita predecir las condiciones que cumplen los estudiantes con bajo rendimiento académico de la Universidad Nacional Amazónica de Madre de Dios.

- Identificar los perfiles de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios.

VIII. Análisis de resultados

En el presente estudio, se utilizó el modelo de procesos *Cross Industry Standard Process for Data Mining* –CRISP-DM– de manera amplia abordado en el marco teórico.

A. Fase 1: Comprensión del negocio

La UNAMAD en su plan estratégico institucional 2017-2019 menciona: la Universidad Nacional Amazónica de Madre de Dios –UNAMAD– es

una comunidad socioeducativa nacional, científica y democrática, integrada por docentes, estudiantes, egresados, autoridades universitarias y personal administrativo.

La UNAMAD se dedica al estudio, la investigación y la enseñanza; la transmisión, difusión y reproducción del conocimiento y la cultura considerando su proyección y extensión social; así como a la producción de bienes o servicios para servir al desarrollo del país, para la formación de profesionales de calidad.

La UNAMAD forma ingenieros, abogados, licenciados (enfermería, administración y negocios Internacionales, ecoturismo, educación matemáticas y computación), médicos veterinarios y zootecnistas, en las diversas especialidades de acuerdo con las demandas esenciales de la región[80].

1. Determinar los objetivos del negocio

– Contexto

El presente estudio se realiza en la oficina de la dirección universitaria de asuntos académicos de la Universidad Nacional Amazónica de Madre de Dios, en este contexto la información es de carácter académico, resultado de los procesos de matrícula de los estudiantes de pregrado.

– Objetivos del negocio

UNAMAD[81] en su plan estratégico institucional 2017-2019, tiene como misión:

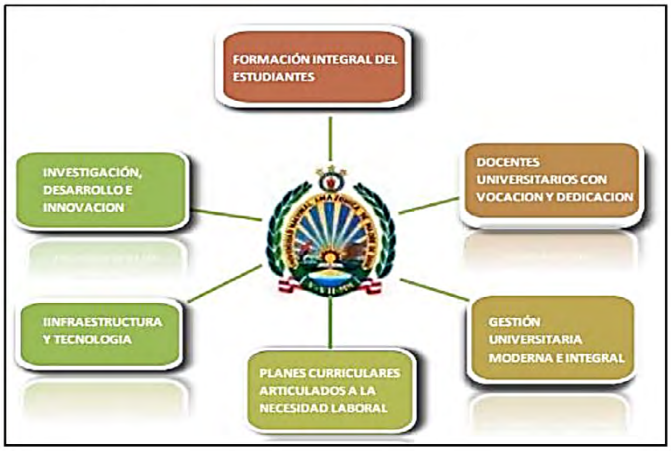
> formar profesionales de alta calidad de manera integral, humanista, científica y tecnológica a estudiantes, capaces de contribuir al desarrollo sostenible con responsabilidad social que valoren su biodiversidad y afirmen su identidad cultural.

En la figura 1 se presenta los ejes estratégicos institucionales:

---

80 UNAMAD. *Plan estratégico institucional UNAMAD 2017-2019*, cit.
81 Ibíd., p. 18.

Figura 1
Ejes estratégicos institucionales



Fuente: [http://www.unamad.edu.pe/index.php/descargas/send/24-institucionales/5468-pei-unamad-2017-2019].

El eje estratégico 1: "Formación integral del estudiante", está fundamentado en asegurar la educación de calidad y avanzar en la internacionalización de la universidad, han sido preocupaciones centrales de la gestión institucional durante los últimos años. El servicio educativo universitario garantiza en sus estudiantes el desarrollo de competencias para el ejercicio profesional, producción científica y un sentido de identidad comprometido con el desarrollo del país[82].

Este eje estratégico 1, establece el objetivo estratégico 1: Desarrollar competencias de los estudiantes para su ejercicio profesional.

En este escenario el objetivo de este proyecto de minería de datos es detectar los patrones de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios.

82 UNAMAD. *Plan estratégico institucional UNAMAD 2017-2019*, cit.

2. Evaluación de la situación

– Inventario de los recursos

Los recursos con los que se contó para el desarrollo del proyecto fueron: lenguaje de programación R, IDE RStudio. Estas herramientas incorporan algoritmos de minería de datos y aprendizaje automático.

La fuente de datos con la que se contó fue el historial de datos de los procesos de matrícula y evaluación de la oficina de la Dirección Universitaria de Asuntos Académicos de la UNAMAD.

– Requisitos, supuestos y restricciones

El personal del proyecto tuvo los conocimientos necesarios de minería de datos para lograr los objetivos planteados, no se presentaron restricciones, por cuanto no se expuso información personal de los estudiantes.

– Costos y beneficios

Tabla 4
Costo de *hardware*

Equipos	Costo unitario	Costo total
01 computara portátil	S/. 2500,00	S/. 2500,00
01 impresora	S/. 450,00	S/. 450,00
Total		S/. 2950,00

Fuente: elaboración propia.

Tabla 5
Costo de *software*

Software	**Costos**
Lenguaje de programación R	S/. 0.00
ide RStudio	S/. 0.00
Total	**S/. 00,00**

Fuente: elaboración propia.

Tabla 6
Recursos humanos

Recurso humano	**Mes 1**	**Mes 2**	**Mes 3**	**Total**
Analista de datos	S/. 4.200,00	S/. 4.200,00	S/. 4.200,00	S/. 12.600,00
	Total			S/. 12.600,00

Fuente: [http://unete.sunat.gob.pe/images/2018/CAS/095/095.pdf].

Tabla 7
Total de inversión

Detalle	Costos
Costo de hardware para el proyecto de minería de datos. Costo de software para el proyecto de minería de datos. Recursos humanos para el proyecto de minería de datos.	S/. 2950,00 S/. -00,00 S/. 12600,00
Total	S/. 15.550,00

Fuente: elaboración propia

El proyecto ascendió a 15.500 soles, los cuales fueron asumidos por el tesista.

El presente proyecto no generó beneficios económicos, el beneficio que presentó fue descubrir los perfiles de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios, esto permitió tomar acciones correctivas por parte de los directivos de la DUAA para la mejora del rendimiento académico.

3. Determinar los objetivos de la minería de datos

Los objetivos del proyecto de minería de datos fueron los siguientes:

Identificar las variables más influyentes en el modelo de clasificación para el bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios.

Establecer el modelo de clasificación que permita predecir las condiciones que cumplen los estudiantes con bajo rendimiento académico de la Universidad Nacional Amazónica de Madre de Dios.

Identificar los perfiles de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios.

4. Producción del plan del proyecto

A continuación, se detalla las etapas del proyecto con el fin de una mejor organización y cumplimiento los objetivos del proyecto.

Primera etapa: Se realizó la solicitud de la base de datos histórica de los procesos académicos a la Oficina del Vicerrectorado académico, dado que la DUAA depende en forma jerárquica de esta.

Segunda etapa: Análisis los datos que emitió la oficina de la DUAA.

Tercera etapa: Preparado de los datos: limpieza y transformación de los datos emitidos por la oficina de la DUAA.

Cuarta etapa: Elección de la técnica de minería de datos que se ajuste al problema que se quiere resolver.

Quinta etapa: Evaluación de los modelos obtenidos al aplicar las distintas técnicas de minería de datos.

Sexta etapa: Generación de informes alineados a los objetivos del negocio y criterios de éxito planteados.

Sexta etapa: Presentación de resultados finales a los directivos de la DUAA.

5. Evaluación inicial de herramientas y técnicas

Las herramientas de minería de datos empleadas en el presente proyecto fueron:

Tabla 8
Herramientas para la minería de datos empleadas

Herramienta	Descripción
	Entorno y lenguaje de programación
	Entorno de desarrollo integrado para el lenguaje de programación R
	Librería para realizar gráficos estadísticos
	Librería para para el proceso de limpieza de datos
	Librería para lectura de archivos en formato xlsx

Fuente: [https://www.rstudio.com/resources/webinars/whats-new-with-readxl/].

Las técnicas de minería de datos empleadas fueron:

Tabla 9
Técnicas de minería de datos empleadas

Técnicas	**Algoritmo**	**Paquete en R**
Predictivas de clasificación	*Random Forest* C5.0 cart	randomForest C50 rpart

Fuente: elaboración propia.

B. Fase 2: comprensión de los datos

A continuación, se detalla las tareas realizadas en esta fase del proyecto.

1. Recolección inicial de datos

Esta tarea se realizó con ayuda del personal autorizado para el acceso a los datos de la DUAA a solicitud del tesista, siendo resultado de ello un reporte de acumulado (con 9.923 instancias) en formato xlsx, que se detalla a continuación:

Figura 2
Reporte de datos acumulado

Monto_matric	universi	idtos_m	idtos_a	tos_de	mestre_i	Edad_ingre	Modalidad_ingreso(CEPRE/Primera opcion/Otro	Satisfacion_carrera(Primera/segunda o	Nota_ingr	Tipo_preparacion(CEPRE/Par	Priodo_ponderado_semest
SIN DATOS	NO	22	22	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	16.77
SIN DATOS	NO	23	23	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	15.43
SIN DATOS	NO	23	23	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	16.43
SIN DATOS	NO	23	23	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	14.22
SIN DATOS	NO	24	24	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	17.46
SIN DATOS	NO	24	24	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	17.25
SIN DATOS	NO	24	24	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	16.63
SIN DATOS	NO	24	24	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	16
SIN DATOS	NO	21	21	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	17
SIN DATOS	NO	23	23	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	17.13
SIN DATOS	SI	22	22	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	13.09
SIN DATOS	SI	23	0	23	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	0
SIN DATOS	SI	16	9	7	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	10
SIN DATOS	SI	17	9	8	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	9.53
SIN DATOS	SI	7	7	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	12
SIN DATOS	SI	20	13	7	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	8.95
SIN DATOS	SI	16	6	10	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	7.38
SIN DATOS	SI	4	0	4	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	5
SIN DATOS	SI	17	3	14	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	2.35
SIN DATOS	NO	8	8	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	13.88
SIN DATOS	NO	19	19	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	15.79
SIN DATOS	NO	22	22	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	14.59
SIN DATOS	NO	22	22	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	14.32
SIN DATOS	NO	24	24	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	15.29
SIN DATOS	NO	24	24	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	15.63
SIN DATOS	NO	24	24	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	14.25
SIN DATOS	NO	24	24	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	15
SIN DATOS	NO	21	21	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	15.57
SIN DATOS	NO	23	23	0	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	16.22
SIN DATOS	SI	22	8	14	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	4.86
SIN DATOS	SI	15	0	15	2010-1	SIN DATOS	EXAMEN ORDINARIO	SIN DATOS	0	SIN DATOS	1.2

Fuente: Dirección Universitaria de Asuntos Académicos –DUAA– de la Universidad Nacional Amazónica de Madre de Dios.

2. Descripción de los datos

A continuación, describimos cada uno de los campos de la tabla, así como el formato inicial:

Tabla 10
Descripción de campos de la tabla de datos

Campo	Formato inicial	Descripción
id Alumno	Numérico	Número correlativo
codigo alumno	Cadena de caracteres	Código de estudiantes
id_departamento	Numérico	Código de departamento
Departamento	Cadena de caracteres	Nombre del departamento
id_provincia	Numérico	Código de provincia
Provincia	Cadena de caracteres	Nombre de provincia
id_distrito	Numérico	Código de distrito
distrito	Cadena de caracteres	Nombre de distrito
fecha_nacimiento	Tipo fecha	Fecha de nacimiento
sexo	Booleano	Sexo de estudiante
id_carrera	Numérico	Código de carrera
carrera	Cadena de caracteres	Carrera
cant_cursos_cursados	Numérico	Número de asignaturas cursadas
cant_cursos_aprobados	Numérico	Número de asignaturas aprobadas
cant_cursos_desaprobados	Numérico	Número de asignaturas desaprobadas
id_escuela	Numérico	Código de escuela
escuela	Cadena de caracteres	Nombre de escuela de procedencia
tipo_escuela(publico/privado)	Numérico	Tipo de escuela
escuela_ubigeo_departamento	Numérico	Código de ubicación geográfica
escuela ubigeo_provincia	Numérico	Código de ubicación geográfica
escuela_ubigeo_distrito	Numérico	Código de ubicación geográfica
Servicio_comedor(Si/No)	Booleano	Servicio de comedor universitario
deuda_universidad(Si/No)	Booleano	Adeudo con la universidad
nro_creditos_matriculados	Numérico	Número de créditos matriculados
nro_creditos_aprobados	Numérico	Número de créditos aprobados
nro_creditos_desaprobados	Numérico	Número de créditos desaprobados
semestre_ingreso	Cadena de caracteres	Semestre de ingreso
modalidad_ingreso(cepre/ Primera opción/Ordinario)	Cadena de caracteres	Modalidad de ingreso
promedio_ponderado_ acumulado	Numérico	Promedio semestral

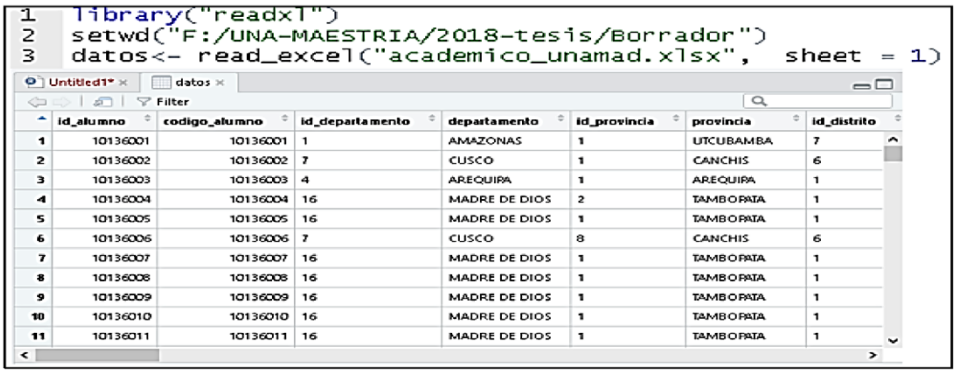
Fuente: Dirección Universitaria de Asuntos Académicos –DUAA– de la Universidad Nacional Amazónica de Madre de Dios.

3. Exploración de los datos

Durante esta tarea se procedió a realizar la lectura del archivo llamado academico_unamad.xlsx donde se encuentra toda la información de los procesos académicos, a continuación, se presenta las líneas de comando utilizado en el lenguaje r para realizar las primeras exploraciones:

Establecemos conexión con el dataset, que se encuentra en formato xlsx, para leer este formato cargamos el paquete readxl:

Figura 3
Datos cargados en *RStudio*



	id_alumno	codigo_alumno	id_departamento	departamento	id_provincia	provincia	id_distrito
1	10136001	10136001	1	AMAZONAS	1	UTCUBAMBA	7
2	10136002	10136002	7	CUSCO	1	CANCHIS	6
3	10136003	10136003	4	AREQUIPA	1	AREQUIPA	1
4	10136004	10136004	16	MADRE DE DIOS	2	TAMBOPATA	1
5	10136005	10136005	16	MADRE DE DIOS	1	TAMBOPATA	1
6	10136006	10136006	7	CUSCO	8	CANCHIS	6
7	10136007	10136007	16	MADRE DE DIOS	1	TAMBOPATA	1
8	10136008	10136008	16	MADRE DE DIOS	1	TAMBOPATA	1
9	10136009	10136009	16	MADRE DE DIOS	1	TAMBOPATA	1
10	10136010	10136010	16	MADRE DE DIOS	1	TAMBOPATA	1
11	10136011	10136011	16	MADRE DE DIOS	1	TAMBOPATA	1

Fuente: elaboración propia.

A continuación, se procede a realizar el análisis descriptivo para las diferentes variables de la base de datos.

– Distribución de frecuencias de la variable departamento

Script en el lenguaje R

Figura 4
Población estudiantil por departamento

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
library (ggplot2)
qplot(data=datos, factor(datos$departamento), geom="bar",
  ylab="Cantidad de estudiantes",xlab="Departamentos",
  fill=factor(datos$departamento))+
  theme(axis.text.x=element_text(size=8, angle=90))+
  scale_fill_discrete(name = "Departamentos")+
  stat_count(aes(label=..count..), vjust=-2, geom="text", position="identity") +
  stat_count(geom="text", aes(label=paste(round(..count../sum(..count..)*100,2),"%"),
  vjust=-0.75))+  scale_y_continuous(limits = c(0, 10000))+
  ggtitle("Población estudiantil-UNAMAD por departamento del 2001-2018") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4", size=16, face="bold"))
```

Fuente: elaboración propia.

– Distribución de frecuencias de la variable provincia

A continuación, se realiza la distribución de frecuencias y gráfico de barras, para las provincias de los departamentos que más estudiantes tienen en esta casa superior de estudios.

– Distribución de frecuencias departamento de Madre de Dios

Script en el lenguaje R:

Figura 5
Tabla de frecuencias: estudiantes por provincias de Madre de Dios

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
#cargamos dplyr para filtrar
library(dplyr)
provMDD<-filter(datos,datos$departamento=="MADRE DE DIOS")
provMDD
tablaProv<-as.data.frame(table(Provincia=provMDD$provincia))
transform(tablaProv,
          FreqAc=cumsum(Freq),
          Rel=round(prop.table(Freq),3),
          RelAc=round(cumsum(prop.table(Freq)),3),
          Porcentaje=round(prop.table(Freq),3)*100,
          PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
)
```

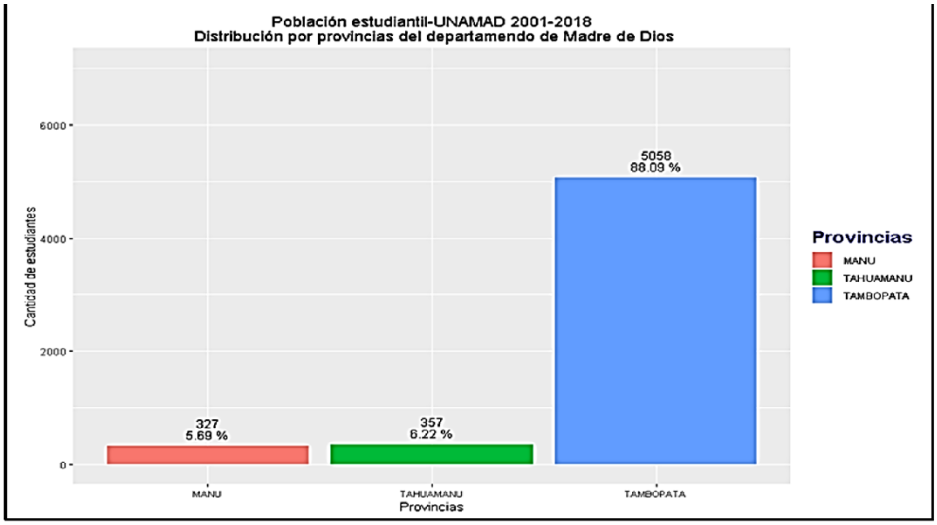
	Provincia	Freq	FreqAc	Rel	RelAc	Porcentaje	PorcentajeAc
1	MANU	327	327	0.057	0.057	5.7	5.695
2	TAHUAMANU	357	684	0.062	0.119	6.2	11.912
3	TAMBOPATA	5058	5742	0.881	1.000	88.1	100.000

Fuente: elaboración propia.

– Diagrama de barras departamento de Madre de Dios

Script en el lenguaje R:

Figura 6
Distribución de estudiantes por provincias de Madre de Dios



Fuente: UNAMAD 2001-2018.

La figura 6 representó la totalidad de estudiantes pertenecientes al departamento de Madre de Dios, de estos un 88% son originario de la provincia de Tambopata, seguido de un 6.22% que proceden de la provincia de Tahuamanu y otro 5.69% de la provincia del Manu.

– Distribución de frecuencias departamento de Cusco

Script en el lenguaje R:

Figura 7
Tabla de frecuencias: estudiantes por provincias de Cusco

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
#cargamos dplyr para filtrar
library(dplyr)
provC<-filter(datos,departamento=="CUSCO")
provC
#Tabla de distribución de frecuencias
tablaProv<-as.data.frame(table(Provincia=provC$provincia))
transform(tablaProv,
          FreqAc=cumsum(Freq),
          Rel=round(prop.table(Freq),3),
          RelAc=round(cumsum(prop.table(Freq)),3),
          Porcentaje=round(prop.table(Freq),3)*100,
          PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
)
```

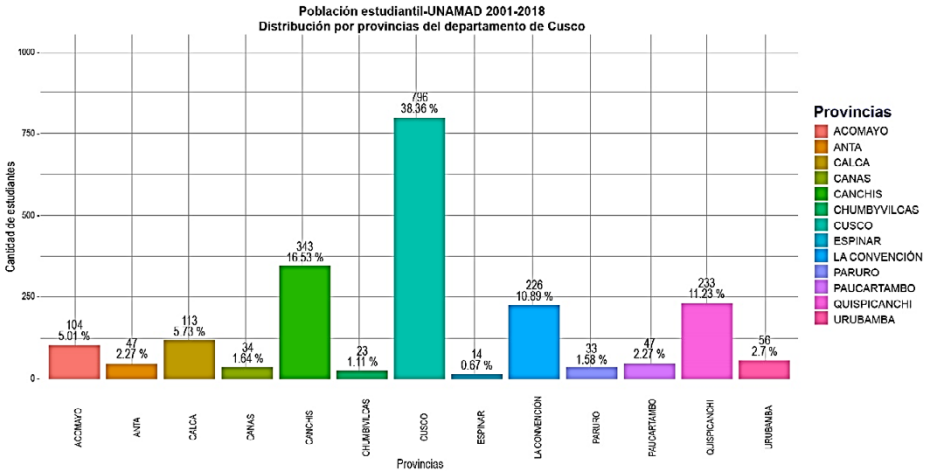
	Provincia	Freq	FreqAc	Rel	RelAc	Porcentaje	PorcentajeAc
1	ACOMAYO	104	104	0.050	0.050	5.0	5.012
2	ANTA	47	151	0.023	0.073	2.3	7.277
3	CALCA	119	270	0.057	0.130	5.7	13.012
4	CANAS	34	304	0.016	0.147	1.6	14.651
5	CANCHIS	343	647	0.165	0.312	16.5	31.181
6	CHUMBIVILCAS	23	670	0.011	0.323	1.1	32.289
7	CUSCO	796	1466	0.384	0.707	38.4	70.651
8	ESPINAR	14	1480	0.007	0.713	0.7	71.325
9	LA CONVENCION	226	1706	0.109	0.822	10.9	82.217
10	PARURO	33	1739	0.016	0.838	1.6	83.807
11	PAUCARTAMBO	47	1786	0.023	0.861	2.3	86.072
12	QUISPICANCHI	233	2019	0.112	0.973	11.2	97.301
13	URUBAMBA	56	2075	0.027	1.000	2.7	100.000

Fuente: unamad 2001-2018.

– Diagrama de barras departamento de Cusco

Script en lenguaje R:

Figura 8
Distribución de estudiantes por provincias de Cusco



Fuente: UNAMAD 2001-2018.

La figura 8 representó la totalidad de estudiantes pertenecientes al departamento de Cusco, de estos un 38.36% son originario de la provincia de Cusco, seguido de un 16.54% que proceden de la provincia de Canchis, otro 11.23% procedentes de la provincia de Quispicanchi y un 10.89% de la provincia La Convención.

– Distribución de frecuencia departamento de Puno

Script en el lenguaje R:

Figura 9
Tabla de frecuencias: estudiantes por provincias de Puno

```
#cargamos readxl para leer archivos de excel
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
#cargamos dplyr para filtrar
library(dplyr)
provP<-filter(datos,departamento=="PUNO")
provP
#Tabla de distribución de frecuencias
tablaProv<-as.data.frame(table(Provincia=provP$provincia))
transform(tablaProv,
          FreqAC=cumsum(Freq),
          Rel=round(prop.table(Freq),3),
          RelAc=round(cumsum(prop.table(Freq)),3),
          Porcentaje=round(prop.table(Freq),3)*100,
          PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
)
```

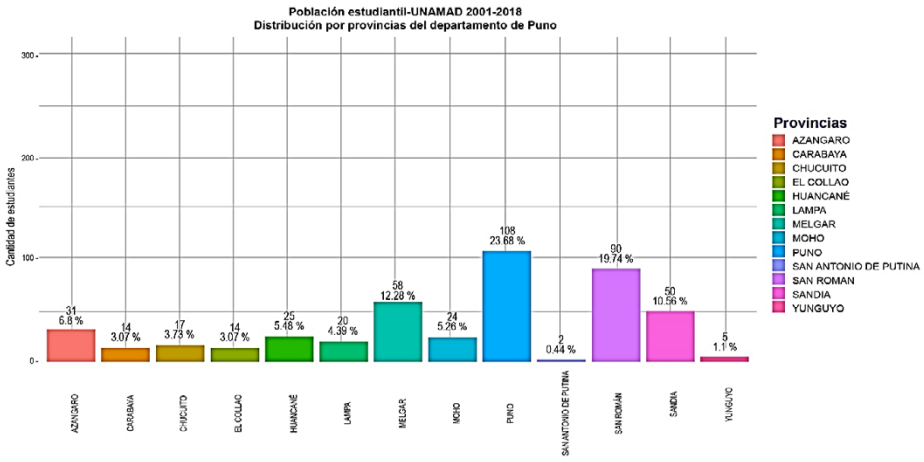
	Provincia	Freq	FreqAc	Rel	RelAc	Porcentaje	PorcentajeAc
1	AZANGARO	31	31	0.068	0.068	6.8	6.798
2	CARABAYA	14	45	0.031	0.099	3.1	9.868
3	CHUCUITO	17	62	0.037	0.136	3.7	13.596
4	EL COLLAO	14	76	0.031	0.167	3.1	16.667
5	HUANCANE	25	101	0.055	0.221	5.5	22.149
6	LAMPA	20	121	0.044	0.265	4.4	26.535
7	MELGAR	56	177	0.123	0.388	12.3	38.816
8	MOHO	24	201	0.053	0.441	5.3	44.079
9	PUNO	108	309	0.237	0.678	23.7	67.763
10	SAN ANTONIO DE PUTINA	2	311	0.004	0.682	0.4	68.202
11	SAN ROMAN	90	401	0.197	0.879	19.7	87.939
12	SANDIA	50	451	0.110	0.989	11.0	98.904
13	YUNGUYO	5	456	0.011	1.000	1.1	100.000

Fuente: unamad 2001-2018.

– Diagramas de barras departamento de Puno

Script en el lenguaje R:

Figura 10
Distribución de estudiantes por provincias Puno



Fuente: unamad 2001-2018.

La figura 10 representó la totalidad de estudiantes pertenecientes al departamento de Puno. De estos, un 23.68% son originarios de la provincia de Puno, seguido de un 19.74% que proviene de la provincia de San Román; otro 12.28%, de la provincia de Melgar y, por último, un 10.96%, de la provincia de Sandia.

– Distribución de frecuencias de la variable sexo

Script en el lenguaje R:

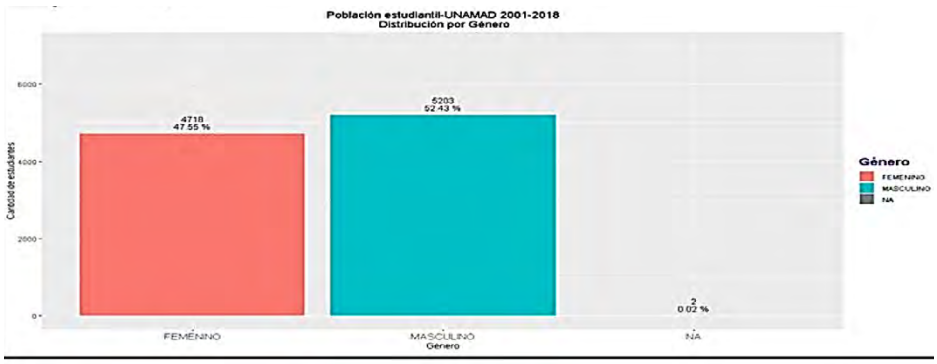
Figura 11
Tabla de frecuencias: estudiantes por género

```
#cargamos readxl para leer archivos de excel
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
#Tabla de distribución de frecuencias
tablaS<-as.data.frame(table(Sexo=datos$sexo))
transform(tablaS,
          FreqAC=cumsum(Freq),
          Rel=round(prop.table(Freq),3),
          RelAc=round(cumsum(prop.table(Freq)),3),
          Porcentaje=round(prop.table(Freq),3)*100,
          PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
)
       Sexo Freq FreqAC   Rel RelAc Porcentaje PorcentajeAc
1  FEMENINO 4718    4718 0.476 0.476       47.6       47.556
2 MASCULINO 5203    9921 0.524 1.000       52.4      100.000
```

Fuente: unamad 2001-2018.

– Diagrama de barras para la variable sexo:

Figura 12
Distribución de estudiantes por género



Fuente: unamad 2001-2018.

En la figura 12 se observa que el 52.43% de la población estudiantil pertenece al género masculino y; un 47.55%, al sexo femenino.
– Distribución de frecuencias de la variable carrera profesional

Figura 13
Tabla de frecuencias: estudiantes por carrera profesional

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
tablaCar<-as.data.frame(table(Departamento=datos$carrera))
tablaCar
transform(tablaCar,
          FreqAc=cumsum(Freq),
          Rel=round(prop.table(Freq),3),
          RelAc=round(cumsum(prop.table(Freq)),3),
          Porcentaje=round(prop.table(Freq),3)*100,
          PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
)
```

	Departamento	Freq	FreqAc	Rel	RelAc	Porcentaje	PorcentajeAc
1	ADMINISTRACIÓN Y NEGOCIOS INTERNACIONALES	836	836	0.084	0.084	8.4	8.427
2	CONTABILIDAD Y FINANZAS	834	1670	0.084	0.168	8.4	16.833
3	DERECHO Y CIENCIAS POLÍTICAS	825	2495	0.083	0.251	8.3	25.149
4	ECOTURISMO	1308	3803	0.132	0.383	13.2	38.333
5	EDUCACIÓN ESPECIALIDAD INICIAL Y ESPECIAL	396	4199	0.040	0.423	4.0	42.324
6	EDUCACIÓN ESPECIALIDAD MATEMÀTICA Y COMPUTACIÓN	616	4815	0.062	0.485	6.2	48.533
7	EDUCACIÓN ESPECIALIDAD PRIMARIA E INFORMÀTICA	320	5135	0.032	0.518	3.2	51.759
8	ENFERMERÍA	586	5721	0.059	0.577	5.9	57.666
9	INGENIERÍA AGROINDUSTRIAL	1365	7086	0.138	0.714	13.8	71.424
10	INGENIERÍA DE SISTEMAS E INFORMÁTICA	727	7813	0.073	0.788	7.3	78.752
11	INGENIERÍA FORESTAL Y MEDIO AMBIENTE	1554	9367	0.157	0.944	15.7	94.416
12	MEDICINA VETERINARIA - ZOOTECNIA	554	9921	0.056	1.000	5.6	100.000

Fuente: UNAMAD 2001-2018.

– Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito de Tambopata-Madre de Dios

Script en el lenguaje R:

Figura 14

Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Tambopata-Madre de Dios

```
#cargamos readxl para leer archivos de excel
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
#cargamos dplyr para filtrar
library(dplyr)
dataDep<-filter(datos,datos$escuela_ubigeo_departamento==16)
dataProvMDD<-filter(dataDep,dataDep$escuela_ubigeo_provincia==1)
dataDisMDD<-filter(dataProvMDD,dataProvMDD$escuela_ubigeo_distrito==1)
#Tabla de distribución de frecuencias
tablaProv<-as.data.frame(table(Escuela=dataDisMDD$escuela))
transform(tablaProv,
          FreqAc=cumsum(Freq),
          Rel=round(prop.table(Freq),3),
          RelAc=round(cumsum(prop.table(Freq)),3),
          Porcentaje=round(prop.table(Freq),3)*100,
          PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
)
```

	Escuela	Freq	FreqAc	Rel	RelAc	Porcentaje	PorcentajeAc
1	AMERICANA DE MADRE DE DIOS	2	2	0.000	0.000	0.0	0.039
2	APLICACION NUESTRA SEÑORA DEL ROSARIO	53	55	0.010	0.011	1.0	1.061
3	AQUILES VELASQUEZ OROS	10	65	0.002	0.013	0.2	1.254
4	AUGUSTO BOURONCLE ACUÑA	284	349	0.055	0.067	5.5	6.734
5	CAP. ALIPIO PONCE VASQUEZ	29	378	0.006	0.073	0.6	7.293
6	CAP. FAP JOSE ABELARDO QUIÑONES	84	462	0.016	0.089	1.6	8.914
7	CARLOS FERMIN FITZCARRALD	865	1327	0.167	0.256	16.7	25.603
8	CEBA - CARLOS FERMIN FITZCARRALD	23	1350	0.004	0.260	0.4	26.047
9	CEBA - DOS DE MAYO	21	1371	0.004	0.265	0.4	26.452
10	CEBA - GUILLERMO BILLINGHURST	16	1387	0.003	0.268	0.3	26.761
11	CEBA - MARIA MOLINARI REATEGUI	6	1393	0.001	0.269	0.1	26.876
12	CRISTO SALVADOR	54	1447	0.010	0.279	1.0	27.918
13	DOS DE MAYO	762	2209	0.147	0.426	14.7	42.620
14	ENAWIPA	7	2216	0.001	0.428	0.1	42.755
15	FAUSTINO MALDONADO	324	2540	0.063	0.490	6.3	49.006
16	GUILLERMO BILLINGHURST	462	3002	0.089	0.579	8.9	57.920
17	HERMOSA GRANDE	10	3012	0.002	0.581	0.2	58.113
18	JAIME WHITE	156	3168	0.030	0.611	3.0	61.123
19	JORGE BASADRE GROHMAN	1	3169	0.000	0.611	0.0	61.142
20	LA PASTORA	56	3225	0.011	0.622	1.1	62.223
21	MADRE DE DIOS	42	3267	0.008	0.630	0.8	63.033
22	MARIA MOLINARI REATEGUI	34	3301	0.007	0.637	0.7	63.689
23	NUESTRA SEÑORA DE LA MERCED	12	3313	0.002	0.639	0.2	63.921
24	NUESTRA SEÑORA DE LAS MERCEDES	326	3639	0.063	0.702	6.3	70.210
25	NUESTRA SEÑORA DEL ROSARIO	15	3654	0.003	0.705	0.3	70.500
26	POTSIWA	1	3655	0.000	0.705	0.0	70.519
27	SAN BARTOLOME	15	3670	0.003	0.708	0.3	70.808
28	SAN BERNARDO	16	3686	0.003	0.711	0.3	71.117
29	SAN ISIDRO	61	3747	0.012	0.723	1.2	72.294
30	SAN JUAN BAUTISTA DE LA SALLE	48	3795	0.009	0.732	0.9	73.220
31	SANTA CRUZ	417	4212	0.080	0.813	8.0	81.266
32	SANTA FE	29	4241	0.006	0.818	0.6	81.825
33	SANTA ROSA	583	4824	0.112	0.931	11.2	93.074
34	SEÑOR DE LOS MILAGROS	330	5154	0.064	0.994	6.4	99.440
35	TRILCE	29	5183	0.006	1.000	0.6	100.000

Fuente: unamad 2001-2018.

– Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Las Piedras-Tambopata-Madre de Dios

Script en el lenguaje R:

Figura 15
Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Las Piedras-Madre de Dios

```
#cargamos readxl para leer archivos de excel
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
#cargamos dplyr para filtrar
library(dplyr)
dataDep<-filter(datos,datos$escuela_ubigeo_departamento==16)
dataProvMDD<-filter(dataDep,dataDep$escuela_ubigeo_provincia==3)
dataDisMDD<-filter(dataProvMDD,dataProvMDD$escuela_ubigeo_distrito==1)
#Tabla de distribución de frecuencias
tablaProv<-as.data.frame(table(Escuela=dataDisMDD$escuela))
transform(tablaProv,
          FreqAc=cumsum(Freq),
          Rel=round(prop.table(Freq),3),
          RelAc=round(cumsum(prop.table(Freq)),3),
          Porcentaje=round(prop.table(Freq),3)*100,
          PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
)
```

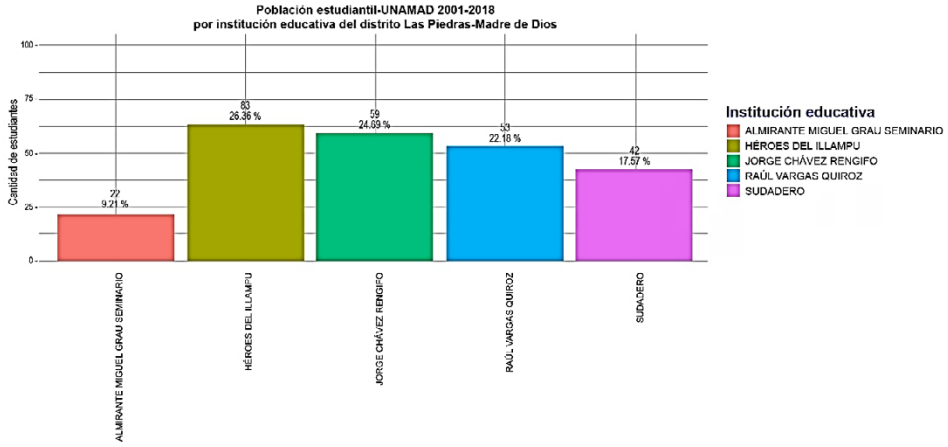
	Escuela	Freq	FreqAc	Rel	RelAc	Porcentaje	PorcentajeAc
1	ALMIRANTE MIGUEL GRAU SEMINARIO	22	22	0.092	0.092	9.2	9.205
2	HEROES DE ILLAMPU	63	85	0.264	0.356	26.4	35.565
3	JORGE CHAVEZ RENGIFO	59	144	0.247	0.603	24.7	60.251
4	RAUL VARGAS QUIROZ	53	197	0.222	0.824	22.2	82.427
5	SUDADERO	42	239	0.176	1.000	17.6	100.000

Fuente: unamad 2001-2018.

– Diagrama de barras estudiantes UNAMAD 2001-2018 por institución educativa del distrito Las Piedras-Tambopata-Madre de Dios

Script en el lenguaje R:

Figura 16
Distribución de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Las Piedras-Madre de Dios



Fuente: unamad 2001-2018La figura 16 presentó la procedencia de los estudiantes por instituciones educativas del distrito Las Piedras, provincia Tambopata del departamento Madre de Dios, en la unamad desde el año 2001 al 2018, donde se aprecia que el 26.36% surge de la institución educativa Héroes de Illampu, un 24.69% procede de la institución educativa Jorge Chávez Rengifo, un 22.18% proviene de la institución educativa Raúl Vargas Quiroz, otro 17.57% de la institución educativa Sudadero y, por último, un 9.21% de la institución educativa Almirante Miguel Grau Seminario.

– Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Laberinto-Tambopata-Madre de Dios

Script en el lenguaje R:

Figura 17
Procedencia de estudiantes UNAMAD 2001-2018
por institución educativa del distrito Laberinto-Madre de Dios

```
#cargamos readxl para leer archivos de excel
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
#cargamos dplyr para filtrar
library(dplyr)
dataDep<-filter(datos,datos$escuela_ubigeo_departamento==16)
dataProvMDD<-filter(dataDep,dataDep$escuela_ubigeo_provincia==4)
dataDisMDD<-filter(dataProvMDD,dataProvMDD$escuela_ubigeo_distrito==1)
#Tabla de distribución de frecuencias
tablaProv<-as.data.frame(table(Escuela=dataDisMDD$escuela))
transform(tablaProv,
            FreqAc=cumsum(Freq),
            Rel=round(prop.table(Freq),3),
            RelAc=round(cumsum(prop.table(Freq)),3),
            Porcentaje=round(prop.table(Freq),3)*100,
            PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
)
```

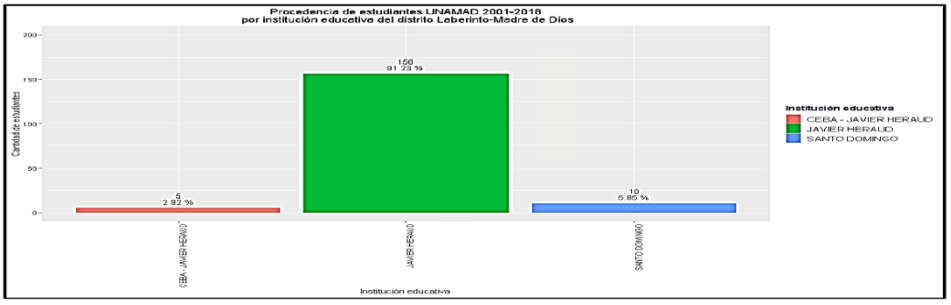
	Escuela	Freq	FreqAc	Rel	RelAc	Porcentaje	PorcentajeAc
1	CEBA - JAVIER HERAUD	5	5	0.029	0.029	2.9	2.924
2	JAVIER HERAUD	156	161	0.912	0.942	91.2	94.152
3	SANTO DOMINGO	10	171	0.058	1.000	5.8	100.000

Fuente: unamad 2001-2018.

– Diagrama de barras estudiantes UNAMAD 2001-2018 por institución educativa del distrito Laberinto-Tambopata-Madre de Dios

Script en el lenguaje R:

Figura 18
Distribución de estudiantes UNAMAD 2001-2018
por institución educativa del distrito Laberinto-Madre de Dios



Fuente: unamad 2001-2018.

La figura 18 presentó la procedencia de los estudiantes por institu-

ciones educativas del distrito de Laberinto, provincia Tambopata, del departamento de Madre de Dios, en la UNAMAD desde el año 2001 al 2018, donde se aprecia que el 91.23% procede de la institución educativa Javier Heraud.

Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Inambari-Tambopata-Madre de Dios

Script en el lenguaje R:

Figura 19

Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Inambari-Madre de Dios

```
#cargamos readxl para leer archivos de excel
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
#cargamos dplyr para filtrar
library(dplyr)
dataDep<-filter(datos,datos$escuela_ubigeo_departamento==16)
dataProvMDD<-filter(dataDep,dataDep$escuela_ubigeo_provincia==2)
dataDisMDD<-filter(dataProvMDD,dataProvMDD$escuela_ubigeo_distrito==1)
#Tabla de distribución de frecuencias
tablaProv<-as.data.frame(table(Escuela=dataDisMDD$escuela))
transform(tablaProv,
            FreqAc=cumsum(Freq),
            Rel=round(prop.table(Freq),3),
            RelAc=round(cumsum(prop.table(Freq)),3),
            Porcentaje=round(prop.table(Freq),3)*100,
            PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
)
```

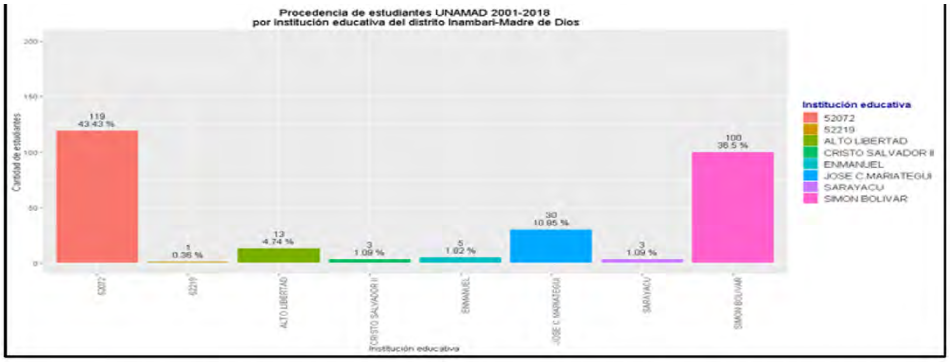
	Escuela	Freq	FreqAc	Rel	RelAc	Porcentaje	PorcentajeAc
1	52072	119	119	0.434	0.434	43.4	43.431
2	52219	1	120	0.004	0.438	0.4	43.796
3	ALTO LIBERTAD	13	133	0.047	0.485	4.7	48.540
4	CRISTO SALVADOR II	3	136	0.011	0.496	1.1	49.635
5	ENMANUEL	5	141	0.018	0.515	1.8	51.460
6	JOSE C.MARIATEGUI	30	171	0.109	0.624	10.9	62.409
7	SARAYACU	3	174	0.011	0.635	1.1	63.504
8	SIMON BOLIVAR	100	274	0.365	1.000	36.5	100.000

Fuente: unamad 2001-2018.

– Diagrama de barras estudiantes UNAMAD 2001-2018 por institución educativa del distrito Laberinto-Tambopata-Madre de Dios

Script en el lenguaje R:

Figura 20
Distribución de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Inambari-Madre de Dios



Fuente: unamad 2001-2018.

La figura 20 presentó la procedencia de los estudiantes por instituciones educativas del distrito Inambari, provincia de Tambopata, del departamento de Madre de Dios, en la UNAMAD desde el año 2001 al 2018, donde se aprecia que el 43.43% procede de la institución educativa 52072; mientras que un 36.5%, de la institución educativa Simón Bolívar y; por último, un 10.95%, de la institución educativa José Carlos Mariátegui.

– Ingresantes UNAMAD por semestre del 2001-2018

Script en el lenguaje R:

Figura 21
Ingresantes UNAMAD por semestre del 2001-2018

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
tablaSemestre<-as.data.frame(table(Semestre=datos$semestre_ingre
tablaSemestre
transform(tablaSemestre,
          FreqAc=cumsum(Freq),
          Rel=round(prop.table(Freq),3),
          RelAc=round(cumsum(prop.table(Freq)),3),
          Porcentaje=round(prop.table(Freq),3)*100,
          PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
)
```

	Semestre	Freq	FreqAc	Rel	RelAc	Porcentaje	PorcentajeAc
1	2001-1	241	241	0.024	0.024	2.4	2.429
2	2001-2	78	319	0.008	0.032	0.8	3.215
3	2002-1	95	414	0.010	0.042	1.0	4.173
4	2002-2	55	469	0.006	0.047	0.6	4.727
5	2003-1	143	612	0.014	0.062	1.4	6.169
6	2003-2	159	771	0.016	0.078	1.6	7.771
7	2004-1	153	924	0.015	0.093	1.5	9.314
8	2004-2	107	1031	0.011	0.104	1.1	10.392
9	2005-1	186	1217	0.019	0.123	1.9	12.267
10	2005-2	133	1350	0.013	0.136	1.3	13.607
11	2006-1	219	1569	0.022	0.158	2.2	15.815
12	2006-2	178	1747	0.018	0.176	1.8	17.609
13	2007-1	156	1903	0.016	0.192	1.6	19.182
14	2007-2	150	2053	0.015	0.207	1.5	20.693
15	2008-1	188	2241	0.019	0.226	1.9	22.588
16	2008-2	118	2359	0.012	0.238	1.2	23.778
17	2009-1	225	2584	0.023	0.260	2.3	26.046
18	2009-2	186	2770	0.019	0.279	1.9	27.921
19	2010-1	472	3242	0.048	0.327	4.8	32.678
20	2010-2	387	3629	0.039	0.366	3.9	36.579
21	2011-1	536	4165	0.054	0.420	5.4	41.982
22	2011-2	265	4430	0.027	0.447	2.7	44.653
23	2012-1	313	4743	0.032	0.478	3.2	47.808
24	2012-2	204	4947	0.021	0.499	2.1	49.864
25	2013-1	407	5354	0.041	0.540	4.1	53.966
26	2013-2	333	5687	0.034	0.573	3.4	57.323
27	2014-1	342	6029	0.034	0.608	3.4	60.770
28	2014-2	189	6218	0.019	0.627	1.9	62.675
29	2015-1	347	6565	0.035	0.662	3.5	66.173
30	2015-2	387	6952	0.039	0.701	3.9	70.074
31	2016-1	537	7489	0.054	0.755	5.4	75.486
32	2016-2	468	7957	0.047	0.802	4.7	80.204
33	2017-1	570	8527	0.057	0.859	5.7	85.949
34	2017-2	490	9017	0.049	0.909	4.9	90.888
35	2018-1	528	9545	0.053	0.962	5.3	96.210
36	2018-2	375	9920	0.038	1.000	3.8	99.990

Fuente: unamad 2001-2018.

4. Verificación de la calidad de los datos

De acuerdo con el análisis exploratorio de los datos, se observó que respecto a la variable departamento, existieron 271 registros con el valor SIN DATOS. Respecto a la variable sexo (figura 12), hubo dos re-

gistros con el valor SIN DATOS. En tanto la variable carrera, se dieron dos registros con el valor NA. Por último, la variable semestre_ingreso presentó dos registros con el valor NA y un registro con el valor 2051-I. estas incoherencias se corrigieron en la segunda fase.

C. Fase 3: preparación de los datos

Durante esta tarea se preparó las variables de acuerdo al algoritmo de árboles de clasificación conocido como CART: Classification and Regression Trees. Para esta técnica, la variable objetivo tuvo que ser categórica; mientras que las variables predictoras, continuas o categóricas. Se emplearon las funciones filter (), select () y mutate () del paquete dplyr de R.

1. Selección de datos

Los atributos seleccionados para este algoritmo fueron:}

Tabla 11
Atributos seleccionados para el modelo

Atributo	Tipo	Descripción
edad_actual	Cuantitativa- discreta	Edad del estudiante
Sexo	Cualitativa- dicotómica	Género del estudiante
escuela_ubigeo_provincia	Cualitativa-politómica	Número de abigeo provincial de la institución educativa de origen.
id_carrera	Cualitativa-politómica	Código de carrera profesional
cant_cursos_cursados	Cuantitativa-discreta	Cantidad de asignaturas cursadas
deuda_universidad	Cualitativa-dicotómica	Especifica deuda con universidad (si/no)
modalidad_ingreso	Cualitativa-politómica	Establece la modalidad de ingreso a la universidad
tipo_escuela	Cualitativa-politómica	Tipo de institución educativa de origen
promedio_ponderado_ semestral	Cuantitativa-continua	Promedio ponderado semestral del estudiante

Fuente: elaboración propia.

95

2. Limpieza de los datos

Esta tarea se realizó con la ayuda de las funciones *filter*, *select* y mutate del paquete *dplry* de R. para el tratamiento de valores faltantes, la discretización de variables numéricas se detalla en el *script* en el lenguaje R como se nota continuación:

Figura 22
Discretización de variables numéricas

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
library(dplyr)
datos<-filter(datos,datos$Prodio_ponderado_acumulado!="SIN DATOS")
datos<-filter(datos,datos$Prodio_ponderado_acumulado!="EN PROCESO")
datos<-filter(datos,datos$Prodio_ponderado_acumulado!="")
datos1<-filter(
  datos,as.integer(as.character(datos$Prodio_ponderado_acumulado))>=0 &
    as.integer(as.character(datos$Prodio_ponderado_acumulado))<=20
)
datos1<-datos %>% mutate(
  t_escuela = case_when(is.na(datos$tipo_escuela) ~ "missing",
                        tipo_escuela=="Pública - Sector Educación" ~ "1",
                        tipo_escuela=="Privada - Particular" ~ "2",
                        tipo_escuela=="Pública - En convenio" ~ "3",
                        tipo_escuela=="Privada - Parroquial" ~ "4",
                        tipo_escuela=="Privada - Instituciones Benéficas" ~ "5",
                        tipo_escuela=="Pública - Municipalidad" ~ "6",
                        TRUE ~ "others")
)
datos1<-datos1 %>% mutate(
  d_universidad = case_when(is.na(datos1$deuda_universidad) ~ "missing",
                            deuda_universidad=="SI" ~ "1",
                            deuda_universidad=="NO" ~ "0",
```

```
datos1<-datos1 %>% mutate(
  t_sexo = case_when(is.na(datos1$sexo) ~ "missing",
                     sexo=="FEMENINO" ~ "0",
                     sexo=="MASCULINO" ~ "1",
                     TRUE ~ "others")
)
datos1<-datos1 %>% mutate(
  t_carrera = case_when(is.na(datos1$id_carrera) ~ "missing",
                        id_carrera=="AN" ~ "1",
                        id_carrera=="CF" ~ "2",
                        id_carrera=="DC" ~ "3",
                        id_carrera=="EC" ~ "4",
                        id_carrera=="ED" ~ "5",
                        id_carrera=="EI" ~ "6",
                        id_carrera=="EN" ~ "7",
                        id_carrera=="EP" ~ "8",
                        id_carrera=="IA" ~ "9",
                        id_carrera=="IF" ~ "10",
                        id_carrera=="IS" ~ "11",
                        id_carrera=="MV" ~ "12",
                        TRUE ~ "others")
)
datos1<-datos1 %>% mutate(
  m_ingreso= case_when(is.na(datos1$modalidad_ingreso) ~ "missing",
                       modalidad_ingreso=="CENTRO PREUNIVERSITARIO" ~ "1",
                       modalidad_ingreso=="CEPRE ORDINARIO" ~ "2",
                       modalidad_ingreso=="DEPORTISTAS CALIFICADOS" ~ "3",
                       modalidad_ingreso=="EXAMEN ESPECIAL PARA SECUNDARIA" ~ "4",
                       modalidad_ingreso=="EXAMEN ORDINARIO" ~ "5",
                       modalidad_ingreso=="FEDERACION AGRARIA DEPARTAMENTAL DE MADRE DE DIOS" ~ "6",
                       modalidad_ingreso=="FEDERACION NATIVA DE RIO MADRE DE DIOS Y AFLUENTES" ~ "7",
                       modalidad_ingreso=="PERSONAS CON DISCAPACIDAD" ~ "8",
                       modalidad_ingreso=="PRIMEROS PUESTOS" ~ "9",
                       modalidad_ingreso=="PROGRAMA NACIONAL DE BECAS" ~ "10",
                       modalidad_ingreso=="RESOLUCION DE CONSEJO UNIVERSITARIO" ~ "11",
                       modalidad_ingreso=="SIN DATOS" ~ "12",
```

```
                    modalidad_ingreso=="RESOLUCION DE CONSEJO UNIVERSITARIO" ~ "11",
                    modalidad_ingreso=="SIN DATOS" ~ "12",
                    modalidad_ingreso=="TITULADOS Y/O GRADUADOS" ~ "13",
                    modalidad_ingreso=="TRASLADO INTERNO DIFERENTE FACULTAD" ~ "14",
                    modalidad_ingreso=="TRASLADO INTERNO MISMA FACULTAD" ~ "15",
                    modalidad_ingreso=="TRASLADOS EXTERNOS NACIONAL" ~ "16",
                    modalidad_ingreso=="VICTIMAS DEL TERRORISMO O PLAN DE" ~ "17",
                    TRUE ~ "others")
)
datos1<-datos1 %>% mutate(
  clase=ifelse(round(as.integer(as.character(Prodio_ponderado_acumulado))) %in% 0:10, "C",
        ifelse(round(as.integer(as.character(Prodio_ponderado_acumulado))) %in% 11:15, "B",
        ifelse(round(as.integer(as.character(Prodio_ponderado_acumulado))) %in% 14:17, "A","AD")))
)
data_unamad<-select(
  datos1,-id_alumno,fecha_nacimiento,-codigo_alumno,-departamento,
  -distrito,-provincia,-id_distrito,-carrera,-id_escuela,-escuela,
  -tipo_escuela,-escuela_ubigeo_distrito,-semestre_ingreso,
  -nro_creditos_desaprobados,-nro_creditos_aprobados
)
data_unamad1<-filter(
  data_unamad,!is.na(id_departamento),!is.na(id_provincia),
  !is.na(modalidad_ingreso),!is.na(sexo),
  !is.na(id_carrera),!is.na(escuela_ubigeo_departamento),
  !is.na(escuela_ubigeo_provincia),
  !is.na(Deuda_universidad),!is.na(t_escuela)
)
data_unamad1<-filter(data_unamad1,escuela_ubigeo_departamento!="SIN DATOS")
data_unamad1<-filter(data_unamad1,data_unamad1$t_escuela!="others")
data_unamad1<-filter(data_unamad1,data_unamad1$m_ingreso!="SIN DATOS")
data_unamad1<-filter(data_unamad1,data_unamad1$m_ingreso!="others")
data_unamad1<-filter(
  data_unamad1,data_unamad1$nro_creditos_matriculados>=0
)
data_unamad1$clase<-factor(data_unamad1$clase)
```

Fuente: elaboración propia.

3. Estructuración de los datos

Con los resultados de la tarea interior (limpieza de datos), los valores filtrados, reemplazados y eliminados, se presenta la estructura del dataset definitivo:

Tabla 12
Estructura del *dataset*

Atributo	**Descripción**	**Valores**
t_sexo	Género del estudiante	0: Femenino 1: Masculino
escuela_ubigeo_departamento	Código Ubigeo de la escuela de origen del estudiante	Valores numéricos

t_carrera	Carrera profesional del estudiante	1: Administración y negocios internacionales 2: Contabilidad y Finanzas 3: Derecho y ciencias políticas 4: Ecoturismo 5: Educación 6: Educación Inicial y Primaria 7: Enfermería 8: Educación Primaria e Informática 9: Ingienería Agroindustrial 10: Ingienería Forestal y Medio Ambiente 11: Ingienería de Sistemas e Informática 12: Medicina Veterinaria
cant_cursos_cursados	Cantidad de asignaturas cursadas por el estudiante	Valores numéricos
servicio_comedor	Indica si el estudiante cuenta con servicio de comedor universitario	Sí No
d_universidad	Indica si el estudiante adeuda a la universidad	No Sí 1: Centro preuniversitario 2: CEPRE ordinario 3: Deportistas calificados 4: Examen especial para secundaria 5: Examen ordinario 6: Federación agraria 7: Federación nativa de río Madre de Dios y afluentes 8: Personas con discapacidad 9: Primeros puestos 10: Programa Nacional de Becas 11: Resolución de consejo universitario 12: Titulados y/o graduados 13: Traslado interno diferente facultad 14: Traslado interno misma facultad 15: Traslados externos nacional 16: Víctimas del terrorismo
t_escuela	Indica el tipo de escuela de procedencia del estudiante	1: Pública – Sector Educación 2: Privada – Particular 3: Pública – En convenio 4: Privada – Parroquial 5: Privada – Instituciones benéficas 6: Pública - Municipalidad
clase	Variable a predecir	C: 0 – 10 B: 11 – 13 A: 14 – 17 AD: 18 – 20

Fuente: elaboración propia.

La variable clase se generó de la transformación del atributo numérico promedio_ponderado_acumulado, al tomar como referencia a MINEDU[83] donde se define la escala de evaluación como se detalla en la siguiente tabla.

Tabla 13
Escala de evaluación de aprendizajes

Variable	Clase	Valores	Descripción
promedio_ponderado_acumulado	C	0 – 10	Cuando el estudiante evidencia el logro de los aprendizajes previstos, al demostrar incluso un manejo solvente y muy satisfactorio en todas las tareas propuestas
	B	11 – 13	Cuando el estudiante evidencia el logro de los aprendizajes previstos en el tiempo programado
	A	14 – 17	Cuando el estudiante está en camino de lograr los aprendizajes previstos, para lo cual requiere acompañamiento durante un tiempo razonable para lograrlo
	AD	18 – 20	Cuando el estudiante empieza a desarrollar los aprendizajes previstos o evidencia dificultades para el desarrollo de estos y necesita mayor tiempo de acompañamiento e intervención del docente de acuerdo con su ritmo y estilo de aprendizaje

Fuente: Adaptado de MINEDU (2009, p. 53).

El script utilizado en el lenguaje R, para esta tarea fue:

83 MINEDU. *Diseño Curricular Nacional de Educación Básica Regular*, Lima, Santillana, 2009, disponible en [https://www.yumpu.com/es/document/read/4471316/diseno-curricular-nacional-santillana].

Figura 23
Script en lenguaje R para la escala de evaluación de aprendizajes

```
datos1<-datos1 %>% mutate(
  clase=ifelse(round(as.integer(
        as.character(Prodio_ponderado_acumulado))) %in% 0:10, "C",
        ifelse(round(as.integer(
        as.character(Prodio_ponderado_acumulado))) %in% 11:15,"B",
        ifelse(round(as.integer(
        as.character(Prodio_ponderado_acumulado))) %in% 14:17, "A","AD")))
)
```

Fuente: elaboración propia.

4. Integración de los datos

Dado que la fuente de datos para el presente estudio fue resultado de una consulta a la base de datos de procesos de matrícula de oficina de la DUAA en formato *Excel*, no se tuvo la necesidad de fusionar múltiples tablas.

5. Formateo de los datos

En esta tarea se realizó la selección de variables: t_sexo, escuela_ubigeo_departamento, t_carrera, cant_cursos_cursados, servicio_comedor, d_universidad, m_ingreso, t_escuela, y clase de tabla data_unamad1, preparados durante la tarea anterior, se consideró a la variable cat_cursos_cursados como variable numérica, las demás variables incluyendo la variable objetivo (clase) se consideraron como factor. A continuación, se detalla el Script en el lenguaje R:

Figura 24
Script en lenguaje R: formateo de datos

```
data_unamad2<-select(
  data_unamad1, t_sexo,
  escuela_ubigeo_departamento,
  t_carrera, cant_cursos_cursados,
  Servicio_comedor,
  d_universidad,
  m_ingreso,
  t_escuela,
  clase
)
data_unamad2$t_sexo<-factor(data_unamad2$t_sexo)
data_unamad2$escuela_ubigeo_departamento<-factor(data_unamad2$escuela_ubigeo_departamento)
data_unamad2$t_carrera<-factor(data_unamad2$t_carrera)
data_unamad2$cant_cursos_cursados<-as.integer(as.character(data_unamad2$cant_cursos_cursados))
data_unamad2$Servicio_comedor<-factor(data_unamad2$Servicio_comedor)
data_unamad2$d_universidad<-factor(data_unamad2$d_universidad)
data_unamad2$m_ingreso<-factor(data_unamad2$m_ingreso)
data_unamad2$t_escuela<-factor(data_unamad2$t_escuela)
data_unamad2$clase<-factor(data_unamad2$clase)
```

Fuente: elaboración propia.

La vista minable quedó de la siguiente manera:

Figura 25
Vista minable

	t_sexo	escuela_ubigeo_departamento	t_carrera	cant_cursos_cursados	Servicio_comedor	d_universidad	m_ingreso	t_escuela	clase
1	0	16	1	80	NO	0	5	1	A
2	0	7	1	49	NO	1	5	2	C
3	0	4	1	80	NO	0	5	3	B
4	0	16	1	22	NO	1	5	1	C
5	1	16	1	55	NO	0	5	1	B
6	1	7	1	86	NO	0	5	2	B
7	1	16	1	87	NO	0	5	1	B
8	1	16	1	81	NO	0	5	1	B
9	0	16	1	91	NO	1	5	1	C
10	1	16	1	13	NO	1	5	1	C
11	1	14	1	87	NO	0	5	1	B
12	0	16	1	107	NO	1	5	1	C
13	1	16	1	16	NO	1	5	1	C
14	1	21	1	86	NO	0	5	1	B

Showing 1 to 14 of 7,309 entries

Fuente: elaboración propia.

Esta tabla quedó con 7309 instancias y 9 columnas.

D. Fase 4: Modelamiento

1. Selección de la técnica de modelado

De acuerdo a los objetivos del presente estudio, las técnicas de modelado que mejor se ajustan para el logro de estos son: el algoritmo *Classification and Regression Trees* –CART– implementado en paquete rpart C5.0 que es una extensión del algoritmo de árboles de decisión C4.5, implementado en el paquete C50 y por último Random Forest puesto en el paquete *Random Forest* de *RStudio*.

2. Generación del plan de pruebas

Para esta tarea se realizó la separación de los datos en un conjunto de entrenamiento y otro de prueba, el primero para el proceso de entrenamiento del modelo y el segundo para probar el modelo entrenado en una proporción de 70% y 30% a proporción.

Script en el lenguaje R:

Figura 26
Resumen del conjunto de datos de entrenamiento

```
tamano.total <- nrow(data_unamad2)
tamano.entreno <- round(tamano.total*0.7)
datos.indices <- sample(1:tamano.total , size=tamano.entreno)
datos.entreno <- data_unamad2[datos.indices,]
str(datos.entreno$clase)
datos.test <- data_unamad2[-datos.indices,]
summary(datos.entreno)
```

```
t_sexo   escuela_ubigeo_departamento   t_carrera     cant_cursos_cursados servicio_comedor d_universidad
0:2392   16      :4048                10      : 837  Min.   :  1.00       NO:4877          0:3485
1:2724   7       : 530                4       : 760  1st Qu.: 12.00       SI: 239          1:1631
         20      : 122                9       : 734  Median : 26.00
         14      : 111                1       : 444  Mean   : 39.09
         4       :  89                2       : 410  3rd Qu.: 70.00
         3       :  51                3       : 405  Max.   :122.00
         (other): 165                 (other):1526
  m_ingreso    t_escuela clase
5       :3001   1:4582    A:  29
1       :1427   2: 477    B:2609
4       : 221   3:  38    C:2478
9       : 128   4:  16
6       : 105   5:   2
7       :  64   6:   1
(other): 170
```

Fuente: elaboración propia.

Figura 27
Resumen del conjunto de datos de prueba

```
t_sexo   escuela_ubigeo_departamento   t_carrera     cant_cursos_cursados servicio_comedor d_universidad
0:1068   16      :1748                10      :350   Min.   :  1.00       NO:2078          0:1501
1:1130   7       : 223                4       :306   1st Qu.: 12.00       SI: 115          1: 692
         14      :  59                9       :303   Median : 25.00
         4       :  45                1       :188   Mean   : 38.26
         20      :  41                2       :178   3rd Qu.: 69.00
         3       :  18                3       :175   Max.   :128.00
         (other):  59                 (other):693
  m_ingreso    t_escuela clase
5       :1267   1:1940    A:  16
1       : 621   2: 228    B:1154
4       : 115   3:  14    C:1023
9       :  48   4:   9
6       :  44   5:   1
3       :  29   6:   1
(other):  69
```

Fuente: elaboración propia.

En las figuras 26 y 27 se observan que los conjuntos de datos se encuentran no balanceados, dado que existen clases con número de instancias mayores que otras. De acuerdo con Rocío Espinar Lara:

> existen dos métodos de remuestreo: Downsampling y Upsampling. La última es

una técnica que simula o atribuye datos adicionales para mejorar el equilibrio de las clases, mientras que la primera es una técnica que redice el tamaño de la muestra para mejorar el equilibrio de dichas clases, tambipen puede darse la hibridación de ambas[84].

3. Construcción del modelo

A continuación, se presenta el procedimiento para la construcción del modelo de clasificación:

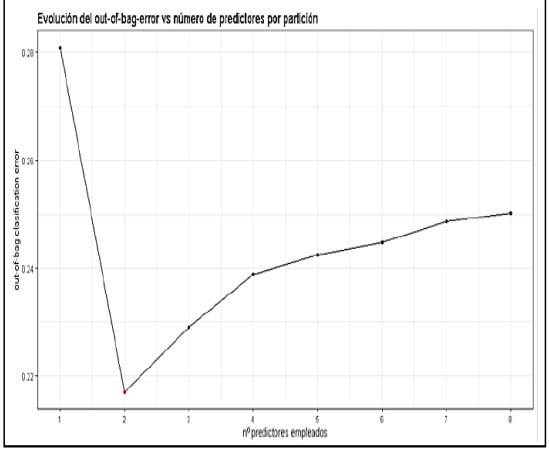
– Identificación de las variables más influyentes en el modelo predictivo al utilizar el algoritmo *Random Forest*: experimento 1
– Identificación del número óptimo de predictores por cada partición

Script en el lenguaje R:

84 Rocío Espinar Lara. "Modelos de clasificación con datos no balanceados", tesis de pregrado, Sevilla, Universidad de Sevilla, junio de 2018, disponible en [https://idus.us.es/bitstream/handle/11441/77518/Espinar%20Lara%20Roc%c3%ado%20TFG.pdf?sequence=1&isAllowed=y].

Figura 28
Evolución del out-of-bag-error versus número de predictores por partición

```r
library(randomForest)
#Función que de devuelve el número de predictores con la tasa de clasificación
get_mean_prediction_error <- function(p_data, p_y, p_ntree){
  require(dplyr)
  t_predictores <- ncol(p_data) - 1
  n_predictores <- rep(NA, t_predictores)
  tasa_err_oob  <- rep(NA, t_predictores)
  for (i in 1:t_predictores) {
    set.seed(123)
    f <- formula(paste(p_y,"~ ."))
    modelo_rf <- randomForest(formula = f, data = p_data, mtry = i, ntree=p_ntree)
    n_predictores[i] <- i
    tasa_err_oob[i] <- tail(modelo_rf$err.rate[, 1], n = 1)
  }
  result <- data_frame(n_predictores, tasa_err_oob)
  return(result)
}
#Obtenemos el número de predictores por partición
var_mtry <-  get_mean_prediction_error(datos.entreno,"clase",250)
var_mtry %>% arrange(tasa_err_oob )
#Graficamos la tasa de error de clasifición vr número de predictores por partición
ggplot(data = var_mtry, aes(x = n_predictores, y = tasa_err_oob)) +
  scale_x_continuous(breaks = var_mtry$n_predictores) +
  geom_line() +geom_point() +
  geom_point(data = var_mtry %>% arrange(tasa_err_oob) %>% head(1),
  color = "red") +labs(title = "Evolución dal out-of-bag-error vs mtry",
  x = "nº predictores empleados", y="out-of-bag clasification error") +
  theme_bw()
```



Fuente: elaboración propia.

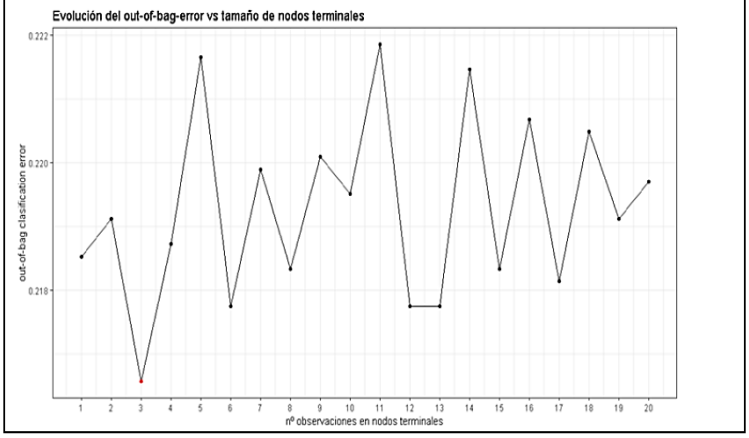
En la figura 28 se observa la evolución del *out-bag-error* en función del número de predictores. Además, se nota que el valor de este error es mínimo para dos predictores por partición.

– Identificación del tamaño óptimo de los nodos finales

Script en el lenguaje R:

Figura 29
Evolución del ot-of-bag-error versus tamaño de nodos

```r
get_nodesize <- function(p_data, p_y, p_size = NULL, p_ntree){
  require(dplyr)
  if (is.null(p_size)){
    p_size <- seq(from = 1, to = nrow(p_data), by = 5)
  }
  tasa_err_oob <- rep(NA, length(p_size))
  for (i in seq_along(p_size)) {
    set.seed(321)
    f <- formula(paste(p_y,"~ ."))
    modelo_rf <- randomForest(formula = f, data = p_data, mtry = 2, ntree = p_ntree,nodesize = i)
    tasa_err_oob[i] <- tail(modelo_rf$err.rate[, 1], n = 1)
  }
  results <- data_frame(p_size, tasa_err_oob)
  return(results)
}
var_nodesize<-get_nodesize(datos.entreno,"clase",p_size = c(1:20),250)
var_nodesize %>% arrange(tasa_err_oob)
ggplot(data = var_nodesize, aes(x = p_size, y = tasa_err_oob)) +
  scale_x_continuous(breaks = var_nodesize$p_size) +
  geom_line() +
  geom_point() +
  geom_point(data = var_nodesize %>% arrange(tasa_err_oob) %>% head(1),color = "red") +
  labs(title = "Evolución del out-of-bag-error vs tamaño de nodos terminales",
  x = "nº observaciones en nodos terminales", y="out-of-bag clasification error") +
  theme_bw()
```



Fuente: elaboración propia.

En la figura 29 se observa la evolución del *out-of-bag-error* en función del número de observaciones en los nodos terminales. Además, se nota que el error se minimiza para 3 observaciones en los nodos terminales.

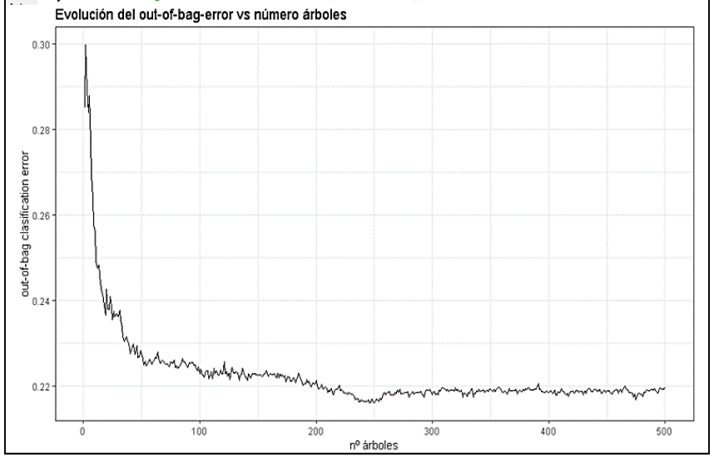
– Identificación del número óptimo de árboles

Script en el lenguaje R:

105

Figura 30
Evolución del *out-of-bag-error* versus número de árboles

```
library(randomForest)
get_nro_arboles <- function(p_data){
  modelo_rf <- randomForest(clase ~ ., data = p_data, mtry = 2, ntree = 500,
                            importance = TRUE, nodesize = 3)
  tasa_err_oob <- data.frame(arboles = seq_along(modelo_rf$err.rate[, 1]),
                             tasa_err_oob = modelo_rf$err.rate[, 1])
  return(tasa_err_oob)
}
var_nro_arboles<-get_nro_arboles(datos.entreno)
var_nro_arboles %>% arrange(tasa_err_oob)
ggplot(data = var_nro_arboles, aes(x = arboles, y = tasa_err_oob ))+geom_line()+
  labs(title = "Evolución del out-of-bag-error vs número árboles",x = "nº árboles",
  y="out-of-bag clasification error")+theme_bw()
```



Fuente: elaboración propia.

En la figura 30 se observa que el out-of-bag-error del modelo se logra estabilizar para 250 árboles, lo cual indica que el modelo óptimo en favor del set de datos se puede dar a partir de este valor.

– Modelo de clasificación final con los valores obtenidos

Script en el lenguaje R:

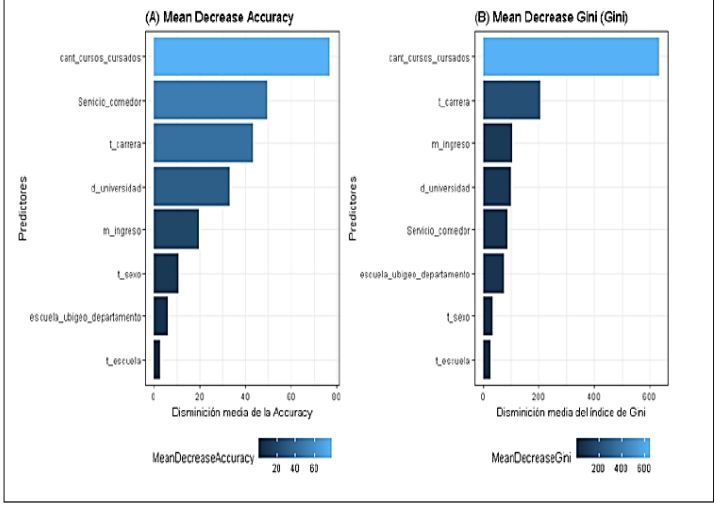
```
modelo ← randomForest (as.factor(clase) ~ . , data = datos.entreno, mtry = 2, ntree = 250,
importance = TRUE, nodesize = 3,
norm.votes = TRUE )
```

– Variables más influyentes en el modelo de clasificación

Script en el lenguaje R:

Figura 31
Influencia de las variables en el modelo de clasificación *Random Forest*

```
library(tidyverse)
library(ggpubr)
library(tibble)
var_importancia <- as.data.frame(importance(modelo, scale = TRUE))
var_importancia <- rownames_to_column(var_importancia, var = "variable")
figura_rf1 <- ggplot(data=var_importancia,
                     aes(x=reorder(variable, MeanDecreaseAccuracy),
                         y = MeanDecreaseAccuracy,fill = MeanDecreaseAccuracy)) +
labs(x = "Predictores", title = "Mean Decrease Accuracy",
     y="Disminición media de la Accuracy") +
  geom_col() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "bottom")
figura_rf2 <- ggplot(data = var_importancia,
              aes(x = reorder(variable, MeanDecreaseGini),
                  y = MeanDecreaseGini,fill = MeanDecreaseGini)) +
labs(x = "Predictores", title = "Mean Decrease Gini (Gini)",
     y="Disminición media del indice de Gini") +
  geom_col() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "bottom")
  ggarrange(figura_rf1, figura_rf2)
```



Fuente: elaboración propia.

En la figura 31 se observa: (A) De entre todos los predictores utilizados en el modelo de clasificación, la cantidad de asignaturas cursadas (cant_cursos_cursados), el servicio de comedor universitario (servi-

cio_comedor), la carrera profesional (t_carrera), deuda con la universidad (d_universidad) son las variables que más influyen en la predicción del rendimiento académico. La gráfica de barras muestra cuánto disminuye la precisión del modelo si dejamos de lado estas variables. (B) las variables cantidad de asignaturas cursadas (cant_cursos_cursados), carrera profesional (t_carrera), modalidad de ingreso (m_ingreso), deuda con la universidad (d_universidad) son las variables que más reducen el índice de impureza de Gini. La importancia de los predictores se evalúa teniendo en cuenta el número de veces que han sido utilizados por los diversos árboles y su capacidad para reducir el índice de Gini[85].

– Discusión

Estos resultados guardan relación con lo que sostienen David Luis La Red Martínez, Marcelo Karanik, Mirtha Giovannini y Noelia Pinto[86] en su estudio *Perfiles de rendimiento académico: un modelo basado en minería de datos*. Los autores señalan que el tipo de escuela que cursó el alumno no está relacionado con el rendimiento académico logrado por el mismo. Ello concuerda con lo hallado en este trabajo de investigación.

– Matriz de confusión y estadísticas

Script en el lenguaje R:

85 Francisco Alonso Sarría y Fulgencio Cánovas García. "Modelos predictivos para el estudio del abandono agrícola", en *Researchgate*, junio de 2016, pp. 161 a 180, disponible en [https://www.researchgate.net/publication/311589338_Modelos_predictivos_para_el_estudi o_del_abandono_agricola].

86 David Luis La Red Martínez, Marcelo Karanik, Mirtha Giovannini y Noelia Pinto. "Perfiles de Rendimiento Académico: un modelo basado en minería de datos", en *Campus Virtuales*, vol. 4, n.º 1, 2015, pp. 12 a 30, disponible en [http://uajournals.com/ojs/index.php/campusvirtuales/article/view/66].

Figura 32
Matriz de confusión del modelo construido con el algoritmo *Random Forest*

```
1  prediccion<- predict(modelo, newdata = datos.test, type = "class")
2  confusionMatrix(prediccion, datos.test[["clase"]])

                 Reference
Prediction   A   B   C
         A   0   0   0
         B  11 857 191
         C   1 290 843

Overall Statistics

               Accuracy : 0.7752
                 95% CI : (0.7571, 0.7925)
    No Information Rate : 0.523
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5536
 Mcnemar's Test P-Value : 4.36e-07

Statistics by Class:

                     Class: A Class: B Class: C
Sensitivity          0.000000   0.7472   0.8153
Specificity          1.000000   0.8069   0.7489
Pos Pred Value            NaN   0.8093   0.7434
Neg Pred Value       0.994528   0.7443   0.8196
Prevalence           0.005472   0.5230   0.4715
Detection Rate       0.000000   0.3908   0.3844
Detection Prevalence 0.000000   0.4829   0.5171
Balanced Accuracy    0.500000   0.7770   0.7821
```

Fuente: elaboración propia.

En la figura 32 se observa que la exactitud (*Accuracy*) del modelo de clasificación es 77.5%. De dicho porcentaje se desprende que la tasa de error de clasificación es de 22.5%. Por otra parte, el coeficiente de kappa es 0.55, de acuerdo con la tabla de su valoración. En ese sentido, la clasificación observada concuerda en forma moderada con la clasificación precedida por el clasificador.

– Árbol de clasificación al utilizar el algoritmo C5.0: experimento 2

Script en R para entrenar el modelo:

```
1  library (C50)

2  modelo ← C5.0 (clase~., data = datos.entreno)
```

Resultado obtenido:

Figura 33
Árbol de clasificación para el rendimiento académico – C5.0.

```
Call:
C5.0.formula(formula = clase ~ ., data = datos.entreno)


C5.0 [Release 2.07 GPL Edition]        Tue Jan 15 10:34:59 2019
-------------------------------

Class specified by attribute 'outcome'

Read 5116 cases (9 attributes) from undefined.data

Decision tree:

Servicio_comedor = SI: B (239/20)
Servicio_comedor = NO:
:...cant_cursos_cursados <= 62:
    :...cant_cursos_cursados <= 6: B (223/23)
    :   cant_cursos_cursados > 6:
    :   :...d_universidad = 1: C (1275/244)
    :       d_universidad = 0:
    :       :...t_carrera in {10,11,12,4,9}:
    :           :...m_ingreso in {1,12,13,14,15,16,2,3,5,6,7,8}: C (901/252)
    :           :   m_ingreso in {10,11,4,9}: B (88/37)
    :           t_carrera in {1,2,3,5,6,7,8}:
    :           :...m_ingreso in {12,6,7}: C (33/8)
    :               m_ingreso in {10,11,13,14,15,16,2,4,8,9}: B (84/20)
    :               m_ingreso = 3:
    :               :...t_carrera in {1,5,6,7,8}: C (5)
    :               :   t_carrera in {2,3}: B (3)
    :               m_ingreso = 5:
    :               :...cant_cursos_cursados <= 13: C (133/47)
    :               :   cant_cursos_cursados > 13: B (327/142)
    :               m_ingreso = 1:
    :               :...cant_cursos_cursados > 21: B (134/27)
    :                   cant_cursos_cursados <= 21:
    :                   :...escuela_ubigeo_departamento in {1,10,11,12,13,15,
    :                       :                              17,18,19,2,20,21,
    :                       :                              22,23,24,25,3,4,5,
    :                       :                              6,7,
    :                       :                              9}: B (17/6)
    :                       escuela_ubigeo_departamento = 8: C (1)
    :                       escuela_ubigeo_departamento = 14:
    :                       :...t_carrera = 1: C (2)
    :                       :   t_carrera in {2,3,5,6,7,8}: B (3)
    :                       escuela_ubigeo_departamento = 16:
    :                       :...t_carrera in {1,3,6}: B (39/12)
    :                           t_carrera in {2,7,8}: C (40/16)
    :                           t_carrera = 5:
    :                           :...cant_cursos_cursados <= 14: C (7/1)
    :                               cant_cursos_cursados > 14: B (7)
    cant_cursos_cursados > 62:
    :...cant_cursos_cursados <= 85: B (1106/155)
        cant_cursos_cursados > 85:
        :...t_carrera in {1,12,2,3,6,7,8}: B (203/24)
            t_carrera in {10,11,4,5,9}:
            :...cant_cursos_cursados > 95:
                :...t_carrera in {10,4,9}: C (77)
                :   t_carrera in {11,5}:
                :   :...cant_cursos_cursados <= 99: B (5/1)
                :       cant_cursos_cursados > 99: C (18/1)
                cant_cursos_cursados <= 95:
                :...t_carrera = 9: C (41/5)
```

Fuente: elaboración propia.

En la figura 33 se observa el árbol de clasificación generado por el algoritmo C5.0. En la hoja 3 de arriba hacia abajo se nota que los estudiantes clasificados en la categoría C ascienden a 1.275. Estos tienen el siguiente perfil: estudiantes que no poseen servicio de comedor universitario, que cursaron más de seis cursos, pero menos de 62, que poseen deuda con la universidad. En la hoja 5 se aprecia que fueron clasificados 901 estudiantes en la categoría C. Estos estudiantes, además de contar con el perfil anterior, pertenecen a las carreras de Ingeniería Forestal y Medio Ambiente, Ingeniería de Sistemas e Informática, Medicina Veterinaria y Zootecnia, Ecoturismo e Ingeniería Industrial.

– Matriz de confusión y estadísticas

Figura 34
Matriz de confusión del modelo construido con el algoritmo C5.0.

```
prediccion_1 <- predict(modelo, newdata = datos.test, type = "class")
summary(prediccion_1)
confusionMatrix(prediccion_1, datos.test[["clase"]])
```

```
Confusion Matrix and Statistics

          Reference
Prediction   A   B   C
         A   0   0   0
         B  15 875 225
         C   1 244 833

Overall Statistics

               Accuracy : 0.7788
                 95% CI : (0.7609, 0.7961)
    No Information Rate : 0.5103
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5607
 Mcnemar's Test P-Value : 0.0007881

Statistics by Class:

                     Class: A Class: B Class: C
Sensitivity          0.000000   0.7819   0.7873
Specificity          1.000000   0.7765   0.7841
Pos Pred Value            NaN   0.7848   0.7727
Neg Pred Value       0.992704   0.7737   0.7982
Prevalence           0.007296   0.5103   0.4824
Detection Rate       0.000000   0.3990   0.3798
Detection Prevalence 0.000000   0.5084   0.4916
Balanced Accuracy    0.500000   0.7792   0.7857
```

Fuente: elaboración propia.

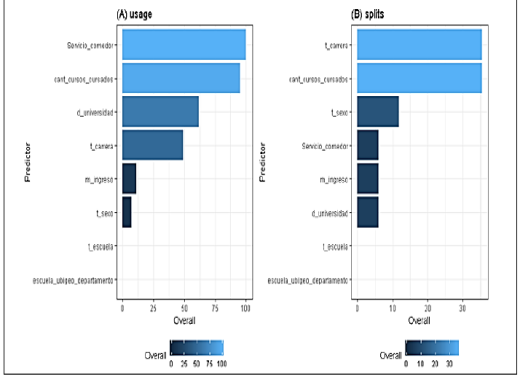
En la figura 34 se observa que la exactitud (*Accuracy*) del modelo de clasificación es 77.8% siendo la tasa de error de clasificación de 22.2%. Por otra parte, el coeficiente de *kappa* es 0.56. La clasificación observada concuerda con la clasificación predicha por el clasificador.

– Identificación de las variables más influyentes en el modelo predictivo al utilizar el algoritmo C5.0: experimento 3

Script en el lenguaje R:

Figura 35
Influencia de las variables en el modelo predictivo de clasificación-C5.0

```
library(C50)
modelo <- C5.0(clase ~ .,data = datos.entreno)
summary(modelo)
library(tibble)
library(dplyr)
importancia_usage <- C5imp(modelo, metric = "usage")
importancia_usage <- importancia_usage %>% rownames_to_column(var = "Predictor")
importancia_usage
importancia_splits <- C5imp(modelo, metric = "splits")
importancia_splits <- importancia_splits %>% rownames_to_column(var = "Predictor")
importancia_splits
library(ggplot2)
figura1 <- ggplot(data = importancia_usage, aes(x = reorder(Predictor, Overall),
                                                y = Overall, fill = Overall)) +
  labs(x = "Predictor", title = "(A) usage") +
  geom_col() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "bottom")
figura2 <- ggplot(data = importancia_splits, aes(x = reorder(Predictor, Overall),
                                                 y = Overall, fill = Overall)) +
  labs(x = "Predictor", title = "(B) splits") +
  geom_col() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "bottom")
ggarrange( figura1,figura2)
```



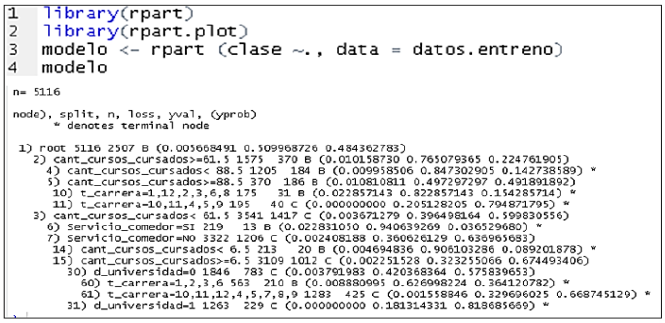
Fuente: elaboración propia.

En la figura 35 se observa: (A) de entre todos los predictores utilizados en el modelo de clasificación servicio de comedor universitario (servicio_comedor), cantidad de asignaturas cursadas (cant_cursos_cursados), deuda con la universidad (d_universidad), carrera profesional a la que pertenece (t_carrera) son las variables que más influyen. (B) las variables carrera profesional a la que pertenece (t_carrera), cantidad de asignaturas cursadas (cat_cursos_cursados) y género (t_sexo) son las variables que más participan en las divisiones del árbol de clasificación.

– Árbol de clasificación al utilizar el algoritmo CART: experimento 4

Script en R para entrenar el modelo:

Figura 36
Reglas obtenidas por el algoritmo CART



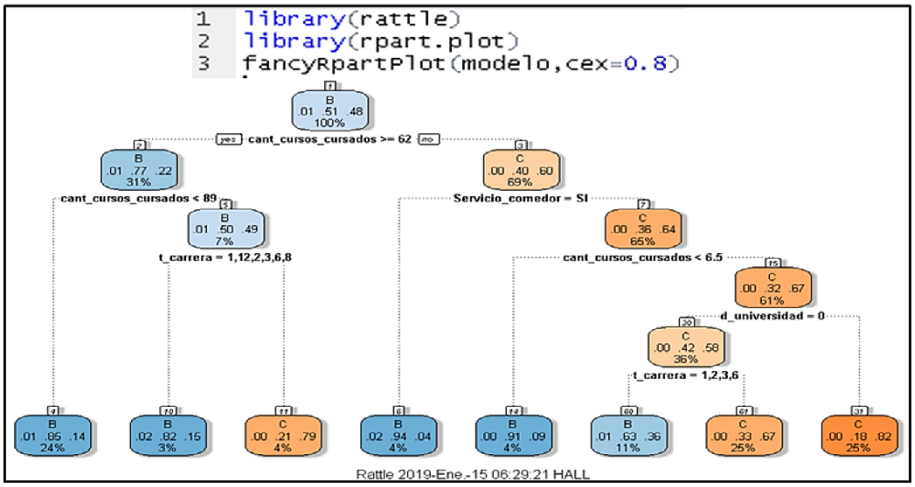
```
1  library(rpart)
2  library(rpart.plot)
3  modelo <- rpart (clase ~., data = datos.entreno)
4  modelo

n= 5116

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 5116 2507 B (0.003668491 0.509968726 0.484362783)
   2) cant_cursos_cursados>=61.5 1575  370 B (0.010158730 0.765079365 0.224761905)
     4) cant_cursos_cursados< 88.5 1205  184 B (0.009958506 0.847302905 0.142738589) *
     5) cant_cursos_cursados>=88.5 370  186 B (0.010810811 0.497297297 0.491891892)
      10) t_carrera=1,12,2,3,6,8 175   31 B (0.022857143 0.822857143 0.154285714) *
      11) t_carrera=10,11,4,5,9 195   40 C (0.000000000 0.205128205 0.794871795) *
   3) cant_cursos_cursados< 61.5 3541 1417 C (0.003671279 0.396498164 0.599830556)
     6) servicio_comedor=SI 219   13 B (0.022831050 0.940639269 0.036529680) *
     7) servicio_comedor=NO 3322 1209 C (0.002408188 0.360626129 0.636965683)
      14) cant_cursos_cursados< 6.5 213   20 B (0.004694836 0.906103286 0.089201878) *
      15) cant_cursos_cursados>=6.5 3109 1012 C (0.002251528 0.323255066 0.674493406)
        30) d_universidad=0 1846  783 C (0.003791983 0.420368364 0.575839653)
          60) t_carrera=1,2,3,6 563  210 B (0.008880995 0.626998224 0.364120782) *
          61) t_carrera=10,11,12,4,5,7,8,9 1283  425 C (0.001558846 0.329696025 0.668745129) *
        31) d_universidad=1 1263  229 C (0.000000000 0.181314331 0.818685669) *
```

Fuente: elaboración propia.

La figura 36 muestra el esquema del árbol de clasificación. Cada inciso indica un nodo y la regla de clasificación que le corresponde. Siguiendo estos nodos, se puede llegar a las hojas del árbol, que corresponde con la clasificación de los datos manejados en esta investigación.

A continuación, se presenta el árbol de clasificación de manera gráfica:

Figura 37
Árbol de clasificación para el rendimiento académico –CART–



Fuente: elaboración propia

La figura 37 representa en detalle el modelo de árbol de clasificación, en ella se observa la hoja 7 de izquierda a derecha, que el 33% de estudiantes fueron clasificados en la categoría B y un 67% en la categoría C, que representa el 25% del total de los datos. En la hoja 8 se observa que el 18% de estudiantes fueron clasificados en la categoría B, mientras que un 82% en la categoría C y estos representan otro 25% del total de los datos.

De este árbol de clasificación se puede afirmar que el 50% del total de estudiantes se encuentran en las hojas 7 y 8 donde predominan estudiantes de la categoría C en proporción de 67% y 82% en forma respectiva. Estos estudiantes tienen la clasificación de 0 a 10. Resumiendo, la hoja 8 se llega a que el perfil que poseen es el siguiente: estudiantes que aprobaron más de 6 cursos, pero menos de 62 cursos, que no poseen servicio de comedor universitario y que poseen alguna deuda con la universidad. En tanto la hoja 7, este grupo de estudiantes no poseen deuda con la universidad y no pertenecen a las carreras de: Administración, Negocios Internacionales, Contabilidad y Finanzas, Derecho y Ciencias Políticas, Educación Inicial y Especial.

– Discusión

Estos resultados guardan relación con lo que sostiene Yamao (2018) en su estudio denominado Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres, Lima-Perú. En dicha investigación se señala que se realizaron predicciones para el rendimiento académico y se obtuvieron resultados de 82.87% al utilizar árbol de decisiones. En el presente libro se halla un 77.8% de exactitud con el algoritmo C5.0.

Pero en lo que no concuerda este estudio con la referida autora es que ella menciona que, de los factores, los que más influyeron en el rendimiento académico fueron los siguientes: notas de examen de admisión, género, edad, modalidad de ingreso y distancia desde su casa hasta el centro de estudios. En este estudio los factores de género, modalidad de ingreso no figuran en el árbol de clasificación, perdiendo así la influencia significativa en el modelo predictivo de clasificación.

– Matriz de confusión y estadísticas

Script en el lenguaje R:

Figura 38
Matriz de confusión del modelo construido con el algoritmo –CART–

```
1  prediccion<- predict(modelo, newdata = datos.test, type = "class", lebel=0.95)
2  library(caret)
3  confusionMatrix(prediccion, datos.test[["clase"]])

Confusion Matrix and Statistics

          Reference
Prediction   A   B   C
         A   0   0   0
         B  10 847 183
         C   6 307 840

Overall Statistics

               Accuracy : 0.7693
                 95% CI : (0.7511, 0.7868)
    No Information Rate : 0.5262
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5433
 Mcnemar's Test P-Value : 2.886e-10

Statistics by Class:

                     Class: A Class: B Class: C
Sensitivity          0.000000   0.7340   0.8211
Specificity          1.000000   0.8142   0.7325
Pos Pred Value            NaN   0.8144   0.7285
Neg Pred Value       0.992704   0.7337   0.8240
Prevalence           0.007296   0.5262   0.4665
Detection Rate       0.000000   0.3862   0.3830
Detection Prevalence 0.000000   0.4742   0.5258
Balanced Accuracy    0.500000   0.7741   0.7768
```

Fuente: elaboración propia.

En la figura 38 se observa que la exactitud (*Accuracy*) del modelo de clasificación es 76.9%, siendo la tasa de error de clasificación de 21.1.%. Por otra parte, el coeficiente de *kappa* es 0.54.

– Fase 5: evaluación

A continuación, se procede a verificar el cumplimiento de los objetivos del proyecto de minería de datos:

Tabla 14
Objetivos del Proyecto de minería de datos

Objetivos	**Si**	**No**
Identificar las variables más influyentes en el modelo de clasificación para el bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios.	X	
Establecer el modelo de clasificación que permita predecir las condiciones que cumplen los estudiantes con bajo rendimiento académico de la Universidad Nacional Amazónica de Madre de Dios.	X	
Identificar los perfiles de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios.	X	

Fuente: elaboración propia.

– Fase 6: implantación

En esta fase de la metodología CRISP-DM se hará la entrega de los resultados obtenidos a las autoridades universitarias para que tomen acciones en mejora del rendimiento académico de los estudiantes.

IX. Conclusiones

– Tras la aplicación de Minería de datos mediante la metodología CRISP-DM, el algoritmo *Random Forest* permitió identificar como variables más influyentes en el modelo de clasificación para el bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios: primero se puede considerar a la *cantidad de asignaturas cursadas* como uno de las variables que más influyen en el bajo rendimiento académico (figura 31), de seguido por la variable *servicio de comedor universitario,* esta nos indica que si el estudiante cuenta con servicio de comedor universitario influye en el rendimiento académico así mismo se puede considerar la *carrera profesional,* como una variable influyente de donde se deduce que la elección acertada de la carrera profesional también influye en el rendimiento académico.

– En relación a los tres algoritmos empleados: *Random Forest, C5.0* y *CART,* el algoritmo que obtuvo mejor desempeño para el modelo predictivo de clasificación para el bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios fue C5.0, con una medida de exactitud de clasificación (*Accuracy*) del 77.8% y el coeficiente de *kappa* del 0.56, pero el que más explica y se acerca a la realidad es *Random Forest* cabe mencionar que la diferencia es insignificante frente al modelo C5.0.

– La aplicación de los algoritmos *CART* y *C5.0* permitió identificar que el perfil que poseen los estudiantes con de bajo rendimiento académico en la Universidad Nacional Amazónica de Madre de Dios es el siguiente (figura 33): "estudiantes que aprobaron más de 6 cursos, pero menos de 62 cursos, que no poseen servicio de comedor universitario y que poseen alguna deuda con la universidad".

– Al culminar el presente estudio se logró obtener un patrón general de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios, determinado por las variables: cantidad de asignaturas cursadas, el servicio de comedor universitario, la deuda que posee el estudiante con la universidad y la carrera profesional al que pertenece.

X. Recomendaciones

– Se recomienda considerar más variables predictoras para determinar su grado de influencia en los modelos de clasificación para el rendimiento académico al tomar como métricas la exactitud (*Accuracy*) y la reducción del índice impureza Gini.

– Se recomienda como trabajos futuros continuar con el estudio del éxito o fracaso del rendimiento académico de los estudiantes de la universidad Nacional Amazónica de Madre de Dios, al aplicar otras técnicas predictivas de minería de datos como la regresión logística binaria y máquinas de soporte vectorial, al utilizar el lenguaje de programación R.

– Se recomienda a la Universidad Nacional Amazónica de Madre de Dios implementar acciones para la mejora del rendimiento académico poniendo especial énfasis en estudiantes que pasaron los cursos del primer semestre de las carreras de Ecoturismo, Educación: Matemática y

Computación, Enfermería, Educación Primaria e Informática, Ingeniería Agroindustrial, Ingeniería Forestal y Medio Ambiente, Ingeniería de Sistemas e informática y Medicina Veterinaria.

Se recomienda a los directivos de la Universidad Nacional Amazónica de Madre de Dios, implementar mecanismos de control de calidad de los datos en los sistemas de información en la oficina de la DUAA.

Capítulo quinto
Sobre la conveniencia de la aplicación de la *data mining* en casos de bajo rendimiento académico

La vertiginosa rapidez de la evolución tecnológica en la sociedad de la información trae a colación una serie de preguntas de nivel técnico, económico, sociológico, cultural y político. Una de las más destacadas es si los sistemas educativos estarán en la posibilidad de brindar la cantidad y calidad de profesionales con el fin de corresponder con los requerimientos de personal calificado de esta sociedad del conocimiento.

En ese sentido, en el contexto educativo universitario se pone a colación el incesante desafío de conservar a diario la calidad académica hasta el punto de mejorarla. En referencia a lo anterior, de manera ininterrumpida se verifican contenidos, fórmulas y sistemas de enseñanza en mor de asegurar convenientes estándares de calidad que den como consecuencia la instrucción de profesionales de alta calidad necesarios para la sociedad.

Es evidente que la productividad académica se constituye como un elemento crítico al tomar en consideración además que el bajo rendimiento académico se liga a un alto índice de deserción. Esto es ni más ni menos lo que con frecuencia se ha notado en estudiantes de la Universidad Nacional Amazónica de Madre de Dios en 2018.

En especial, el rendimiento académico se explica como la productividad del sujeto, características y la aprehensión regular de los propósitos asignados. El rendimiento académico puede llegar a ser objeto de debate debido a que se dan muchas variables que influyen en el desempeño del alumno y delimitan de forma exacta que la productividad estudiantil no es una labor trivial.

El rendimiento académico es afectado por una serie factores heterogéneos tanto internos como externos que limitan la dedicación del

estudiante. En razón de esto, tomar en cuenta el desempeño del estudiantado en general de igual forma no arroja información que pueda ser empleada para hallar problemas cognitivos, de captación, de diferenciamiento, entre otros. Una posibilidad acertada, entonces, es procurar delimitar si se dan excepcionalidades comunes a colectivos de estudiantes. De esta forma, la elección de perfiles se torna en una habilidad de carácter relevante en el momento de tomar acciones correspondientes a la performance de aquellos.

Determinar perfiles resulta una labor divulgada en varias áreas y es equivalente al proceso de clasificación de patrones. En nuestros días se dan varias formas para establecer y seleccionar modelos que se empleen en la Inteligencia Artificial y del Aprendizaje de máquinas. Estos algoritmos retornan datos de importancia en el instante de la toma de decisiones. La información se ordena en considerables almacenes de datos (*Data Warehouse*) que es revisada por algoritmos que ejecutan minería de datos después de una limpieza previa.

Por otro lado, se ha llegado a la determinación de que el factor más vinculado con la calidad educativa es el mismo alumno, dado por medio del nivel socioeconómico del hogar de origen y se ha notado que el rendimiento del estudiante es mayor para las mujeres, para los estudiantes de más corta edad y para aquellos que provienen de núcleos familiares más educados, teniendo gran relevancia el vínculo entre horas ocupadas y performance académica.

De manera notable, tomar en cuenta métodos matemáticos acertados es un provecho en el momento de la evaluación del desempeño, el tema es emplearlos de la forma correcta. La multiplicidad de estudios acerca del rendimiento académico señala que no hay una manera única para evaluarlo. La organización adecuada de los datos, agregada a un modelo adecuado de manejo de los mismos brinda un panorama evidente de los problemas en el desempeño de los estudiantes. De tal forma, hay métodos propios de la Inteligencia Artificial, tales como la minería de datos (*Data Mining*) empleada para el hallazgo de conocimiento escondido en grandes volúmenes de datos que determinan patrones en forma acertada.

La *Data Mining* es la etapa de la detección de saber en bases de datos. Ello se refiere a la aplicación objetiva de algoritmos concretos que

ocasionan una enumeración de patrones con respecto a los datos en estado de procesamiento.

Al emplear la minería de datos se pueden llegar a producir grandes volúmenes de información cuyo análisis requerirá de una gran cantidad de tiempo debido a la influencia del rendimiento académico por el medio socio-económico y cultural del estudiante además de temas actitudinales del mencionado en tanto el estudio y el empleo de TIC.

El éxito y el fracaso académico se relaciona con el tipo de institución de educación secundaria donde el estudiante llevó a cabo sus estudios, el empeño del sujeto en cuestión medida en tiempo de estudio, la relevancia brindada al estudio en comparación a la diversión y al trabajo, el grado de instrucción de los padres y madres y la consideración que de las TIC poseen los estudiantes.

En el presente libro de investigación se ha abordado un modelo eficiente para la elaboración de perfiles de alumnos en consideración del rendimiento económico al emplear almacenes de datos y técnicas de minería de datos. Estas reflexiones harán posible tomar medidas en favor de la reducción del fracaso académico, al actuar de manera anticipada en compañía de los estudiantes cuyo perfil de señales de fracaso académico.

En consecuencia, la toma en cuenta de los métodos de minería de datos empleados para la reducción del fracaso académico de estudiantes universitarios es pertinente para la elaboración de perfiles y se manifiesta como una herramienta adecuada de alta utilidad en favor de la gestión académica. En ese sentido, la presente investigación puede ser implementada por diferentes instituciones.

Bibliografía

Alderete, Ana María. "Fundamentos del análisis de regresión logística en la investigación psicológica", *Revista evaluar*, vol. 6, n.º 1, 2006, pp. 52 a 67, disponible en [https://revistas.unc.edu.ar/index.php/revaluar/article/view/534/474].

Alonso Sarría, Francisco y Fulgencio Cánovas García. "Modelos predictivos para el estudio del abandono agrícola", en *Researchgate*, junio de 2016, pp. 161 a 180, disponible en [https://www.researchgate.net/publication/311589338_Modelos_predictivos_para _el_estudi o_del_abandono_agricola].

Aluja Banet, Tomas. "La minería de datos, entre la estadística y la inteligencia artificial", en *Qüesiió: quaderns d'estadística i investigació operativa*, vol. 25, n.º 3, 2001, pp. 479 a 498, disponible en [https://www.researchgate.net/profile/Tomas_Aluja-Banet/publication/28177489_La_mineria_de_datos_entre_la_estadistica_y_la_inteligencia_artificial/links/00b7d53b3b091899b7000000/La-mineria-de-datos-entre-la-estadistica-y-la-inteligencia-artificial.pdf].

Álvarez Cáceres, Rafael. *Estadística multivariante y no paramétrica con spss*, Madrid, Ediciones Díaz de Santos, 1995.

Arias, Fidias G. *El proyecto de investigación. Introducción a la metodología científica*, Caracas, Episteme, 2006, disponible en [https://ebevidencia.com/wp-content/uploads/2014/12/EL-PROYECTO-DE-INVESTIGACI%C3%93N-6ta-Ed.-FIDIAS-G.-ARIAS.pdf].

Ato García, Manuel; José Antonio López Pina, Antonio Pablo Velandrino Nicolás y Julio Sánchez Meca. *Estadística avanzada con el paquete systat*, Murcia, Universidad de Murcia, 1990.

Bacallao Gallestey, Jorge; José Mario Parapar de la Riestra, Mercedes Roque Gil y Jorge Bacallao Guerra. "Árboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico", en *Educación Médica*

Superior, vol. 18, n.º 3, julio-septiembre de 2004, disponible en [http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21412004000300002].

Barrientos Martínez, Rocío Erandi; Nicandro Cruz Ramírez, Héctor Gabriel Acosta Mesa, Ivonne Rabatte Suárez, María del Carmen Gogeascoechea Trejo, Patricia Pavón León, Sobeida L. Blázquez Morales. "Árboles de decisión como herramiento en el disgnóstico médico", *Revista Médica de la Universidad Veracruzana*, vol. 9, n.º 2, 2009, pp. 19 a 24, disponible en [https://www.medigraphic.com/pdfs/veracruzana/muv-2009/muv092c.pdf].

Bellmore, Amy; Angela J. Calvin, Jun-Ming Xu y Xiaojin Zhu. "The five W's of "bullying" on Twitter: Who, What, Why, Where, and When", en *Computers in Human Behavior*, vol. 44, marzo de 2015, pp. 305 a 314.

Benítez, Ignacio y José Luis Diez. "Técnicas de agrupamiento para el análisis de datos cuantitativos y cualitativos", en *Reaserchgate*, Valencia, septiembre de 2005, disponible en [https://www.researchgate.net/profile/Ignacio_Benitez/publication/239526131_Tecnicas_de_Agrupamiento_para_el_Analisis_de_Datos_Cuantitativos_y_Cualitativos/links/00b7d51c15cca2cb1f000000/Tecnicas-de-Agrupamiento-para-el-Analisis-de-Datos-Cuantitativos-y-Cu].

Brachman, Ronald Jay y Tej Anand. "The process of knowledge Discovery in databases", en Brachman (ed.), *Workshop on knowledge discovery in databases*, 1994, pp. 37 a 53, disponible en [https://pdfs.semanticscholar.org/2db5/ec88e07974242eb8f8de867275bec8f29e3a.pdf].

Britos, Paola Verónica. "Procesos de explotación de información basados en sistemas inteligentes", tesis doctoral, Buenos Aires, Universidad Nacional de la Plata, agosto de 2008, disponible en [http://sedici.unlp.edu.ar/bitstream/handle/10915/4142/Documento_completo.pdf?sequence=1&isAllowed=y].

Calvache Fernández, Leidy Carolina; Valentina Álvarez Vallejo y Jorge Iván Triviño Arbeláez. "Proceso kdd como apoyo a las estrategias del proyecto sara (Sistema de Acompañamiento para el Rendimiento Académico)", *Revista Educación en Ingeniería*, vol. 13, n.º 26, julio de 2018, pp. 82 a 89, disponible en [https://educacioneningenieria.org/index.php/edi/article/view/916/365].

Camana Fiallos, Roberto. "Aplicación de técnicas de minería de datos para la indagación y estudio de resultados electorales", en *CienciAmérica: revista de divulgación científica de la Universidad Tecnológica Indoamérica*, vol. 7, n.º 1, enero de 2012, pp. 85 a 94, disponible en [http://cienciamerica.uti.edu.ec/openjournal/index.php/uti/article/view/10/8].

Cerda Lorca, Jaime y Luis Villarroel. "Evaluación de concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa", *Revista chilena de pediatría*, vol. 79, n.º 1, 2008, pp. 54 a 58.

Dapozo, Gladys N.; Eduardo Porcel, María Victoria López, Verónica S. Bogado, y Roberto Bargiela. "Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de Sistemas de la facena-unne", Buenos Aires, sedici, junio de 2016, disponible en [http://sedici.unlp.edu.ar/handle/10915/20797].

Díaz Martínez, Zuleyka. *Predicción de crisis empresariales en seguros no vida, mediante árboles de decisión y reglas de clasificación*, Madrid, Complutense, 2007, disponible en [https://eprints.ucm.es/48680/1/9788474918823.pdf].

Eckert, Karina B. y Roberto Suénaga. "Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos", en *Formación Universitaria*, vol. 8, n.º 5, 2015, disponible en [https://www.researchgate.net/publication/281671104_Analisis_de_Desercion-Permanencia_de_Estudiantes_Universitarios_Utilizando_Tecnica_de_Clasificacion_en_Mineria_de_Datos/fulltext/5681256408ae051f9aec2b62/Analisis-de-Desercion-Permanencia-de-Estudiantes-U].

Eichstaedt, Johannes C.; Hansen Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, Megha Agrawal, Lukasz A. Dziurzynski, Maarten Sap, Christopher Weeg, Emily E. Larson, Lyle H. Ungar y Martin E. P. Seligman. Psychological Languaje on Twitter Predicts County-Level Heart Disease Mortality, en *Psychol Sci.,* vol. 26, n.º 2, 2015, pp. 159 a 169, disponible en [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4433545/#].

Espinar Lara, Rocío. "Modelos de clasificación con datos no balanceados", tesis de pregrado, Sevilla, Universidad de Sevilla, junio de 2018, disponible en [https://idus.us.es/bitstream/handle/11441/77518/Espinar%20Lara%20Roc%c3%ado%20TFG.pdf?sequence=1&isAllowed=y].

Fayyad, Usama; Gregory Piatetsky Shapiro y Smyth Phadhraic. "Knowledge Discovery and Data Mining: Towards a Unifying Framework", en *kdd*, n.º 96, 1996, pp. 82 a 88, disponible en [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwj5saPW26DrAhUpU98KHTBHCV8QFjACegQIAxAB&url=https%3A%2F%2Fwww.aaai.org%2FPapers%2FKDD%2F1996%2FKDD96-014.pdf&usg=AOvVaw0RmxCEkOXM9afVEgY_m5Y2].

Flores Cartes, Claudio Nicolás. "Exigencias de calidad de suministro en base a densidad de consumo mediante técnicas de minería de datos", tesis de pregrado, Santiago de Chile, Universidad de Chile, 2014, disponible en [http://repositorio.uchile.cl/bitstream/handle/2250/115571/cf-flores_cc.pdf?sequence=1&isAllowed=y].

Flóres López, Raquel y José Miguel Fernández. *Las redes neuronales artificiales*, La Coruña, netbiblo, 2008.

Gallardo Arancibia, José Alberto. "Metodología para la definición de requisitos en proyectos de data mining", tesis doctoral, Madrid, Universidad Politécnica de Madrid, 2009.

Gandomi, Amir y Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management*, vol. 35, n.º 2, 2015, pp. 137 a 144, disponible en [https://www.sciencedirect.com/science/article/pii/S0268401214001066/pdfft?md5=83a9e41c2aa8141394ce1a998ed61553&pid=1-s2.0-S0268401214001066-main.pdf].

Garbanzo Vargas, Guiselle María. "Factores asociados al rendimiento académico en estudiantes universitarios. Una reflexión desde la calidad de la educación superior pública", *Revista Educación*, vol. 37, n.º 1, 2007, pp. 43 a 63, disponible en [https://www.redalyc.org/pdf/440/44031103.pdf].

García Cambronero, Cristina y Irene Gómez Moreno. "Algoritmos de aprendizaje: knn y kmeans", en *Inteligencia en Redes de Telecomunicación*, 2012, pp. 6 y 7, disponible en [http://www.it.uc3m.es/jvillena/irc/practicas/08-09/06.pdf].

Garre, Miguel; Juán José Cuadrado, Miguel Ángel Sicilia, Daniel Rodríguez y Ricardo Rejas. "Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software", *Revista Española de Innovación, Calidad e Ingeniería de Software*, vol. 3, n.º 1, 2007, pp. 6 a 22, disponible en [https://www.redalyc.org/pdf/922/92230103.pdf].

Gil Albarrán, Guillermo. *Data Mining*, Lima, Megabyte, 2009.

Gil Flores, Javier. "Aplicación del método de Bootstrap al contraste de hipótesis en la investigación educativa", *Revista de Educación*, n.º 336, 2005, pp. 251 a 261.

Hair, Joseph; Rolph E. Anderson (comp.), Ronald L. Tatham (trad.) y William C. Black (trad.). *Analisis Multivariante*, Madrid, Prentince Hall, 1999.

Heredia, Diana; Yegny Amaya y Edwin Barrientos. "Student Dropout Predictive Model Using Data Mining Techniques", en *ieee Latin America Transactions*, vol. 13, n.º 9, 2015, pp. 3127 a 31234.

Hernández G., Claudia L. y María Ximena Dueñas R. *Hacia una metodología de gestión del conocimiento basada en minería de datos*, comtel, 2009, pp. 79 a 96, disponible en [http://repositorio.uigv.edu.pe/bitstream/handle/20.500.11818/982/COMTEL-2009-80-96.pdf?sequence=1&isAllowed=y].

Hilbert, Martin y Priscila López. "The World's Technological Capacity to Store, Communicate, and Compute Information", en *Science*, vol. 332, n,º 6025, abril de 2011, pp. 60 a 66.

Jiménez Chura, Adolfo Carlos. "Análisis predictivo para los procesos de admisión de la Universidad Nacional del Altiplano - Puno", tesis de doctorado, Puno, Universidad Nacional del Altiplano, 2017, disponible en [https://1library.co/document/q2nod6jq-analisis-predictivo-procesos-admision-universidad-nacional-altiplano-puno.html?tab=pdf].

Johannes C. Eichstaedt; Hansen Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, Megha Agrawal, Lukasz A. Dziurzynski, Maarten Sap, Christopher Weeg, Emily E. Larson, Lyle H. Ungar, Martin E. P. Seligman. "Psychological Language on Twitter Predicts County-Level Heart Disease Mortality", en *Psychological Science*, vol. 26, n.º 2, pp. 159 a 169, disponible en [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4433545/].

Kovalevski, Leandro y Paula Macat. "Alternativas no paramétricas de clasificación multivariada", *Décimoséptimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística*, noviembre de 2012, Rosario, disponible en [https://www.fcecon.unr.edu.ar/web-nueva/sites/default/files/u16/Decimocuartas/kovalevski_macat_alternativas_no_parametricas.pdf].

Krikorian, Mauro; Ana Ruedin y Leticia Seijas. "Reconocimiento de patrones utilizando transformadas wavelets sin submuestreo y máquinas de soporte vectorial", *xiv Reunión de Trabajo Procesamiento de la Información y Control*, 2011, pp. 839 a 844, disponible en [http://dc.sigedep.exactas.uba.ar/media/academic/grade/thesis/krikorian.pdf].

Laney, Doug. "3-D data management: Controlling data volume, velocity and variety", en *Application Delivery Strategies*, meta Group Inc., febrero de 2001, disponible en [https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf].

La Red Martínez, David Luis; Marcelo Karanik, Mirtha Giovannini y Noelia Pinto. "Perfiles de Rendimiento Académico: un modelo basado en minería de datos", en *Campus Virtuales*, vol. 4, n.º 1, 2015, pp. 12 a 30, disponible en [http://uajournals.com/ojs/index.php/campusvirtuales/article/view/66].

Marín Hernández, Juan José. "Análisis de conglomerados (ii): El procedimiento Conglomerados jerárquicos", Universidad Carlos iii de Madrid, 2014, disponible en [http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/GuiaSPSS/22conglj.pdf].

Mariñelarena Dondena, Luciana; Marcelo Luis Errecalde y Alejandro Castro Solano. "Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología", *Revista Argentina de Ciencias del Comportamiento*, vol. 9, n.º 2, enero-diciembre de 2017, pp. 65 a 76, disponible en [https://revistas.unc.edu.ar/index.php/racc/article/view/12701/Mari%C3%B1elarena-Dondena].

Medina Merino, Rosa Fátima y Carmen Ismelda Ñique Chacón. "Bosques aleatorios como extensión de árboles de clasificación con los programas R y Python", en *Portal de revistas Ulima*, diciembre de 2017, pp. 165 a 189, disponible en [file:///Users/optimusprime/Downloads/Bosques_aleatorios_como_extension_de_los_arboles_d.pdf].

Microsoft. "Conceptos de Minería de datos", 2020, disponible en [https://docs.microsoft.com/es- es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server- 2017].

minedu. *Diseño Curricular Nacional de Educación Básica Regular*, Lima, Santillana, 2009, disponible en [https://www.yumpu.com/es/document/read/4471316/diseno-curricular-nacional-santillana].

Moine, Juan Miguel; Silvia Ethel Gordillo y Ana Silvia Haedo. "Análisis comparativo de metodologías para la gestión de proyectos de minería de datos", en *sedici*, 2012, disponible en [http://hdl.handle.net/10915/18749].

Moine, Juan Miguel; Silvia Ethel Gordillo y Ana Silvia Haedo. "Estudio comparativo de metodologías para minería de datos", en *sedici*, 2011, disponible en [http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1&isAllowed=y].

Mondragón Becerra, Rosibelda. "Exploraciones sobre el soporte multi-agente bdi en el proceso de descubrimiento de conocimiento en bases de datos", tesis de maestría, Veracruz, Universidad Veracruzana, 2007, disponible en [https://www.uv.mx/personal/aguerra/files/2013/06/2007-mondragon-becerra.pdf].

Montt, Cecilia; Felix Castro y Nibaldo Rodríguez. "Análisis de accidentes de tránsito con máquinas de soporte vectorial LS-svm", *Revista de ingeniería de transporte*, vol. 15, n.º 2, 2011, pp. 7 a 14, disponible en [https://pdfs.semanticscholar.org/8416/eefa57a0d491f57bb9c7686cc10cc383949f.pdf].

Moreno García, María N.; Luis Antonio Miguel Quintales, Francisco José García Peñalvo y María José Polo Martín. "Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software", en *Researchgate*, 2001, disponible en [https://www.researchgate.net/publication/220958273_Aplicacion_de_Tecnicas_ de_Mineria_de_Datos_en_la_Construccion_y_Validacion_de_Modelos_Predictivos_y_Asociativos_a_Partir_de_Especificaciones_de_Requisitos_De_Software].

Pascual González, Damaris. "Algoritmos de agrupamiento basados en densidad y validación de clusters", tesis doctoral, Castellón de Plana, Universidad Jaume i, marzo de 2010, disponible en [http://www.cerpamid.co.cu/sitio/files/Damaris-Tesis.pdf].

Pautsh, Jesús Germán Andrés. "Minería de datos aplicada al análisis de la deserción en la carrera de analista en sistemas de computación", tesis de pregrado, Posadas, Universidad Nacional de Misiones, 2009, disponible en [https://www.lawebdelprogramador.com/pdf/6566-Mineria-de-Datos-aplicada-al-analisis-de-la-desercion-en-la-Carrera-de-Analista-en-Sistemas-de-Computacion.html].

Ricardo Timarán Pereira. "Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos", en *Memorias de la 8.ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática cisci*, 2009, disponible en [http://www.iiis.org/cds2008/cd2009cSc/CISCI2009/PapersPdf/C692YV.pdf].

Pérez López, César. *Técnicas de análisis multivariante de datos*, Madrid, Pearson Prentice Hall, 2004, disponible en [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwj8xvP6r6HrAhVykuAKHcYKBgsQFjABegQIAhAB&url=https%3A%2F%2Fwww.academia.edu%2F39613182%2FT%25C3%25A9cnicas_de_an%25C3%25A1lisis_multivariante_de_datos_Aplicaciones_con_].

Pérez López, César y Daniel Santín. *Minería de datos. Técnicas y herramientas*, Madrid, Thomson, 2007.

Reyes Tejada, Yesica Noelia. *Relación entre el rendimiento académico, la ansiedad ante los exámenes, los rasgos de personalidad, el autoconcepto y la asertividad en estudiantes de primer año de psicología de la unmsm*, Lima, sisbib, 2003, dispo-

nible en [http://sisbib.unmsm.edu.pe/bibvirtual/tesis/salud/reyes_t_y/cap2.htm].

Rodríguez, Oldemar. (s. f.). *Metodología para el desarrollo de proyectos en Minería de datos CRISP-DM,* disponible en [http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037].

Rosado Gómez, Alveiro Alonso y Alejandra Verjel Ibáñez. "Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander", *Revista Tecnura*, vol. 79, n.º 45, 2014, pp. 101 a 113, disponible en [http://www.scielo.org.co/pdf/tecn/v19n45/v19n45a08.pdf].

Salinas, Jesús Walter. "Detección de patrones de los alumnos de pregrado desaprobados en el curso de estadística general de la Universidad Nacional Agraria La Molina usando técnicas de minería de datos", *Memorias del II Encuentro Colombiano de Educación Estocástica*, 2016, pp. 115 a 122, disponible en [http://funes.uniandes.edu.co/9282/1/Salinas2016Deteccion.pdf].

Schwartz, Hansen Andrew; Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michael Kosinski, David Stillwell, Martín E. P. Seligman y Lyle H. Ungar. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulay Approach", en *PLOS ONE*, vol. 8, n.º 9, e73791, septiembre de 2013, disponible en [https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791].

Schwartz, H. Andrew y Lyle H. Ungar. "Data-Driven Content Analysis of Social Media: A Systematic Overview of Atomated Methods", en *The ANNALS of the American Academy of Political and Social Science*, vol. 659, n.º 1, abril de 2015, pp. 78 a 94.

Segarra Ciprés, Mercedes; Marta Estrada Guillén y Diego Monferrer Tirado. "Estilos de aprendizaje en estudiantes universitarios: lateralización vs. Interconexión de los hemisferios cerebrales", en *Revista española de pedagogía*, n.º 262, septiembre-diciembre de 2015, pp. 583 a 600, disponible en [https://revistadepedagogia.org/wp-content/uploads/2015/11/Estilos-de-aprendizaje-en-estudiantes-universitarios-lateralizaci%C3%B3n-vs.-interconexi%C3%B3n-de-los-hemisferios-cerebrales.pdf].

unamad. *Plan estratégico institucional UNAMAD 2017-2019*, Puerto Maldonado, abril de 2016, disponible en [http://www.unamad.edu.pe/index.php/descargas/send/24-institucionales/5468-pei-unamad-2017-2019].

Valcárcel Asencios, Violeta. "Datamining y el descubrimiento del conocimiento", *Revista de la Facultad de Ingeniería Industrial*, vol. 7, n.º 2, 2004, pp. 83 a 86,

disponible en [https://www.researchgate.net/publication/307181857_DATA_MINING_Y_EL_DESCUBRIMIENTO_DEL_CONOCIMIENTO/fulltext/57c432b908aee5141be5bc8f/DATA-MINING-Y-EL-DESCUBRIMIENTO-DEL-CONOCIMIENTO.pdf].

Vallejo P., Diego y Germán Tenalanda V. "Minería de datos aplicada en la detección de intrusos", en *Ingenieras usbMed*, vol. 3, n.º 1, 2012, disponible en [https://dialnet.unirioja.es/descarga/articulo/4694116.pdf].

Vergara Cruz; Ana Oviedo, Claudia Carmona, Gloria Vélez e Iván Amón. "Estilos de aprendizaje y minería de datos: un estudio preliminar en el contexto universitario", en *Ingeniería e Innovación*, vol. 6, n.º 1, junio de 2018, pp. 13 a 18, disponible en [https://revistas.unicordoba.edu.co/index.php/rii/article/view/1534/1803].

Villa Murillo, Adriana; Andrés Carrión García y Antonio Sozzi Rodríguez. "Optimización del diseño de parámetros: método Forest-Genetic univariante", en *Publicaciones en ciencias y tecnologías*, vol. 10, n.º 1, 2017, pp. 12 a 24, disponible en [https://dialnet.unirioja.es/descarga/articulo/6501229.pdf].

Villada, Fernando, Diego Raúl Cadavid y Juan David Molina. "Pronóstico del precio de la energía eléctrica usando redes neuronales artificiales", *Revista Facultad de Ingeniería Universidad de Antioquia*, n.º 44, junio de 2014, pp. 111 a 118, disponible en [http://www.scielo.org.co/pdf/rfiua/n44/n44a11.pdf].

Webmining Consultores. *Webmining*, 10 de enero de 2011, disponible en [http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/].

Yamao, Eiriku. "Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la Escuela Académico Profesional de Ingienería de Computación y Sistemas, Universidad de San Martín de Porres, Lima-Perú", tesis de posgrado, Lima, Universidad de San Martín de Porres, 2018, disponible en [http://repositorio.usmp.edu.pe/bitstream/handle/usmp/3555/yamao_e.pdf?sequence=3&isAllowed=y].

Los autores

Nelly Jacqueline Ulloa Gallardo
jaulga44@gmail.com

Docente asociada de la Facultad de Ingeniería. Nelly Ulloa Gallardo estudió Ingeniería de Sistemas E Informática en la Universidad Privada Antenor Orrego, Perú. Obtuvo la licenciatura en la misma universidad. Obtuvo la maestría en Ciencias de la Educación por la Universidad Nacional de Educación Enrique Guzmán y Valle. Recibió el doctorado en Ciencias de la Educación por la misma Universidad. Actualmente se desempeña como docente de la Universidad Nacional Amazónica de Madre de Dios.

Perfil académico:
[http://directorio.concytec.gob.pe/appDirectorioCTI/VerDatosInvestigador.do;jsessionid=eff173d9453755c2638cb90ed19a?id_investigador=47823].

Ralph Miranda Castillo
ralphi2010@hotmail.com

Ingeniero Electrónico, de la especialidad de Telecomunicaciones, con Maestría en Ciencias de la Ingeniería Electrónica mención en Automatización e instrumentación; Doctor en ciencias de la Ingeniería Mecatrónica, cuenta con conocimientos en robótica con microcontroladores, hardware libre, redes de datos y Telemática, sistemas de video vigilancia, telecomunicaciones, electrónica de potencia, sistemas SCADA, lenguajes de programa-

ción, aplicaciones de inteligencia artificial, con experiencia en elaboración, ejecución, gestión, evaluación y supervisión de proyectos tecnológicos de inversión público y privado.

Perfil académico:
[http://directorio.concytec.gob.pe/appDirectorioCTI/VerDatosInvestigador.do;jsessionid=6575044c616fa15a3c8e4108e2e0?id_investigador=47353].

Luis Alberto Holgado Apaza
luisholgadoapaza@gmail.com

Luis Alberto Holgado Apaza es Ingeniero en Sistemas, egresado de las Universidad Andina de Cusco, con conocimientos de lenguajes de programación como Java, PHP y R. Actualmente desarrolla proyectos de minería de datos mediante la modelo de procesos CRISP-DM y lenguaje de programación R.

Ingeniero en Sistemas y Mg. Scientiae en Informática con Mención en Gerencia de Tecnologías de Información y Comunicaciones.

Cargo: Docente Contratado

Líneas principales de Investigación: Ciencias de la Computación

Editado por el Instituto Latinoamericano de Altos Estudios –ILAE–,
en septiembre de 2020

Se compuso en caracteres Cambria de 12 y 9 ptos.

Bogotá, Colombia