UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INFORMÁTICA

TESIS

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

PRESENTADA POR:

LUIS ALBERTO HOLGADO APAZA

PARA OPTAR EL GRADO ACADÉMICO DE:

MAGISTER SCIENTIAE EN INFORMÁTICA

MENCIÓN EN GERENCIA DE TECNOLOGÍA DE INFORMACIÓN Y COMUNICACIONES

PUNO, PERÚ

2018

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INFORMÁTICA

TESIS

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

PRESENTADA POR:

LUIS ALBERTO HOLGADO APAZA

PARA OPTAR EL GRADO ACADÉMICO DE:

MAGISTER SCIENTIAE EN INFORMATICA

MENCIÓN EN GERENCIA DE TECNOLOGÍA DE INFORMACIÓN Y *COMUNICACIONES*

APROBADA POR EL SIGUIENTE JURADO:

PRESIDENTE

M. Sc. ERNERTO NAYER TUMI FIGUEROA

PRIMER MIEMBRO

M. Sc. FREDY HERIC VILLASANTE SARAVIA

SEGUNDO MIEMBRO

M. Sc. RAMIRO PEDRO LAURA MURILLO

ASESOR DE TESIS

Dr. EDGAR ELOY CARPIO VARGAS

Puno 06 de febrero de 2019

ÁREA: Escalabilidad y rendimiento en análisis inteligente de datos y distribuidos
TEMA: Técnica de minería de datos
LINEA: Detección de patrones mediante minería de datos

DEDICATORIA

A Dios, por darme la vida y sabiduría.

A la memoria de mi querido padre Ricardo Holgado Rojas, quien me inculcó el valor del trabajo y la disciplina.

A mi madre Dolores Apaza Rocca, por haber hecho de mi un hombre de bien.

A mi familia: mi hija Luciana, mi esposa Solinka por motivarme en todo momento.

A mi hermano Alfredo y hermanas: Doris, Georgina, Lisbeth y Nélida.

AGRADECIMIENTOS

- A los docentes y personal administrativo de la escuela de Posgrado-Maestría en informática de la Universidad Nacional del Altiplano por la labor sacrificada que cumplen en beneficio de los distintos profesionales.

- A los miembros del jurado M. Sc. Ernesto Nayer Tumi Figueroa, M. Sc. Fredy Heric Villasante Saravia, M. Sc. Ramiro Pedro Laura Murillo y al Dr. Edgar Eloy Carpio Vargas; por todos sus alcances y observaciones en el presente trabajo.

- Al M. Sc. Remo Choquejahua Acero por la guía y alances en el presente trabajo.

- Al personal directivo de la Dirección Universitaria de Asuntos Académicos de la Universidad Nacional Amazónica de Madre de Dios por haber dado las facilidades para realizar esta investigación, de manera muy especial al M. Sc. Elias Gutierrez Paredes.

ÍNDICE GENERAL

Pág.

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

CAPÍTULO III

MATERIALES Y MÉTODOS

CAPÍTULO IV

RESULTADOS Y DISCUSIONES

ÍNDICE DE TABLAS

Pág.

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE ANEXOS

Pág.

RESUMEN

El presente estudio se llevó a cabo en la Universidad Nacional Amazónica de Madre de Dios (UNAMAD), ubicado en el departamento de Madre de Dios, durante el año 2018, tuvo como objetivo general detectar los patrones de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios, para su desarrollo se empleó la metodología de minería de datos denominado: CRISP-DM; el algoritmo Random Forest permitió identificar que las variables: cantidad de asignaturas cursadas, el servicio de comedor universitario, la carrera profesional, deuda con la universidad, son las variables que más influyen en la prediccion del rendimiento académico, en relación a los tres algoritmos empleados: Random Forest, C5.0 y CART, el algoritmo que obtuvo mejor desempeño para el modelo 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, por otra parte 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: “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”

Palabras clave: C5.0, clasificación, CRISP-DM, data mining, data science, KDD, RandomForest, RPART.

ABSTRACT

The present study was carried out in the Universidad Nacional Amazónica de Madre de Dios (UNAMAD), located in the department of Madre de Dios, during the year 2018. Its general objective was to detect the patterns of low academic performance of the students of the Universidad Nacional Amazónica de Madre de Dios, for its development the data mining methodology called: CRISP-DM; The Random Forest algorithm made it possible to identify which variables: number of subjects studied, university canteen service, professional career, debt with the university, are the variables that most influence the prediction of academic performance, in relation to the three algorithms used : Random Forest, C5.0 and CART, the algorithm that obtained the best performance for the classification model for the low academic performance of the students of the Universidad Nacional Amazónica de Madre de Dios, was C5.0, with an accuracy of 77.8% and with a kappa coefficient of 0.56, on the other hand the application of the algorithms CART and C5.0 allowed to identify the profile that the students with low academic performance have in the Universidad Nacional Amazónica de Madre de Dios is the following: "students who passed more than 6 courses, but with less than 62 courses in total, also they do not have a cafeteria service access and have some debts with the university "

Keywords: C5.0, classification, CRISP-DM, data mining, data science, KDD, RandomForest, RPART.

INTRODUCCIÓN

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, estos datos permitieron descubrir conocimiento oculto mediante la aplicación de técnicas predictivas de minería de datos, específicamente los árboles de clasificación Random Forest, C5.0 y CART.

El objetivo del presente estudio estuvo orientado a detectar los patrones de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios.

Este trabajo se desarrolló en cuatro capítulos que detallamos a continuación:

Capítulo I, en este capítulo se desarrolla el marco teórico, abarcando distintas técnicas descriptivas y predictivas de minería de datos, metodologías de minería de datos más usadas en los últimos tiempos, también se aborda los antecedentes de estudio.

Capítulo II, se ha considerado la problemática de investigación, permitiéndonos verificar los problemas que atraviesa la universidad en relación a la información que almacena y el desaprovechamiento para extraer conocimiento oculto en ello, así también en este capítulo se consideró los objetivos del presente estudio.

Capítulo III, se detalla los materiales y métodos utilizados, el lugar de estudio, la población y el método de investigación.

Capítulo IV, se presenta los resultados obtenidos luego de aplicar los algoritmos Random Forest, C5.0 y CART, estos resultados se presentan en forma de gráficos de barra, matriz de confusión, de árboles teóricos y árboles de modo gráfico.

Finalmente se presentan las conclusiones a las que se llegaron, recomendaciones, bibliografía y anexos.

CAPÍTULO I

REVISIÓN DE LITERATURA

1.1. Marco Teórico

1.1.1. El proceso de descubrimiento de conocimiento en bases de datos (KDD)

Mondragon (2007) citando a Fayyad *et al*. (1996), define a KDD como un proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles, y entendibles en los datos. En este contexto, los *datos* se refieren a un conjunto de hechos (ejemplos en una base de datos), y los *patrones* son expresiones en algún lenguaje que describen de manera compacta los datos. El término *proceso* implica que KDD comprende muchos pasos, entre los que se encuentran: la preparación de datos, búsqueda de patrones, evaluación del conocimiento, y 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 confiablemente el comportamiento futuro de alguna entidad. También se desea que los patrones sean *novedosos* (al menos para el sistema y preferentemente para el usuario) y *potencialmente 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 (p. 6).

El proceso de KDD es interactivo e iterativo, involucrando numerosos pasos con muchas decisiones tomadas por el usuario. Brachman y Anand (1996) ofrecen una visión práctica del proceso KDD enfatizando la naturaleza interactiva del proceso como se observa en la Figura 1.

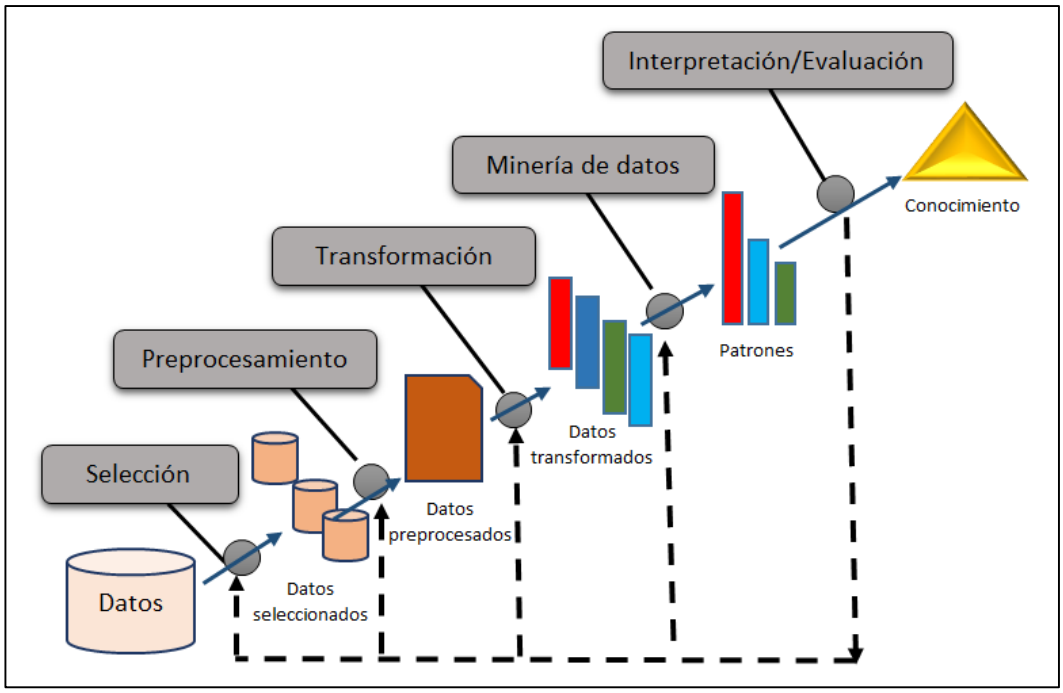


Figura 1. Esquema del proceso KDD

Fuente: Adaptado de (Fayyad *et al.*, 1996)

Como se observa en la Figura 1, el proceso de KDD consta de 5 etapas, Webmining Consultores (2011) describe cada etapa de la siguiente manera:

1. **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.

2. **Preprocesamiento.** 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.

3. **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, consolidando los datos de una forma necesaria para la fase siguiente.

4. **Data Mining.** Es la fase de modelamiento propiamente tal, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u "ocultos" en los datos.
5. **Interpretación y evaluación.** Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

Para Perez & Santin (2007) el proceso de extracción de extracción del conocimiento KDD consta de las siguientes fases:

Figura 2. Esquema del proceso de extracción del conocimiento KDD
Fuente: Adaptado de (Perez & Santin, 2007)

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, buscando 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. Es 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, generalmente 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. Finalmente, en la ase difusión se hace uso del nuevo conocimiento y se hace partícipe de él a todos los posibles usuarios (p. 5).

En este sentido podemos clasificar las fases de proceso de extracción del conocimiento en el siguiente esquema:

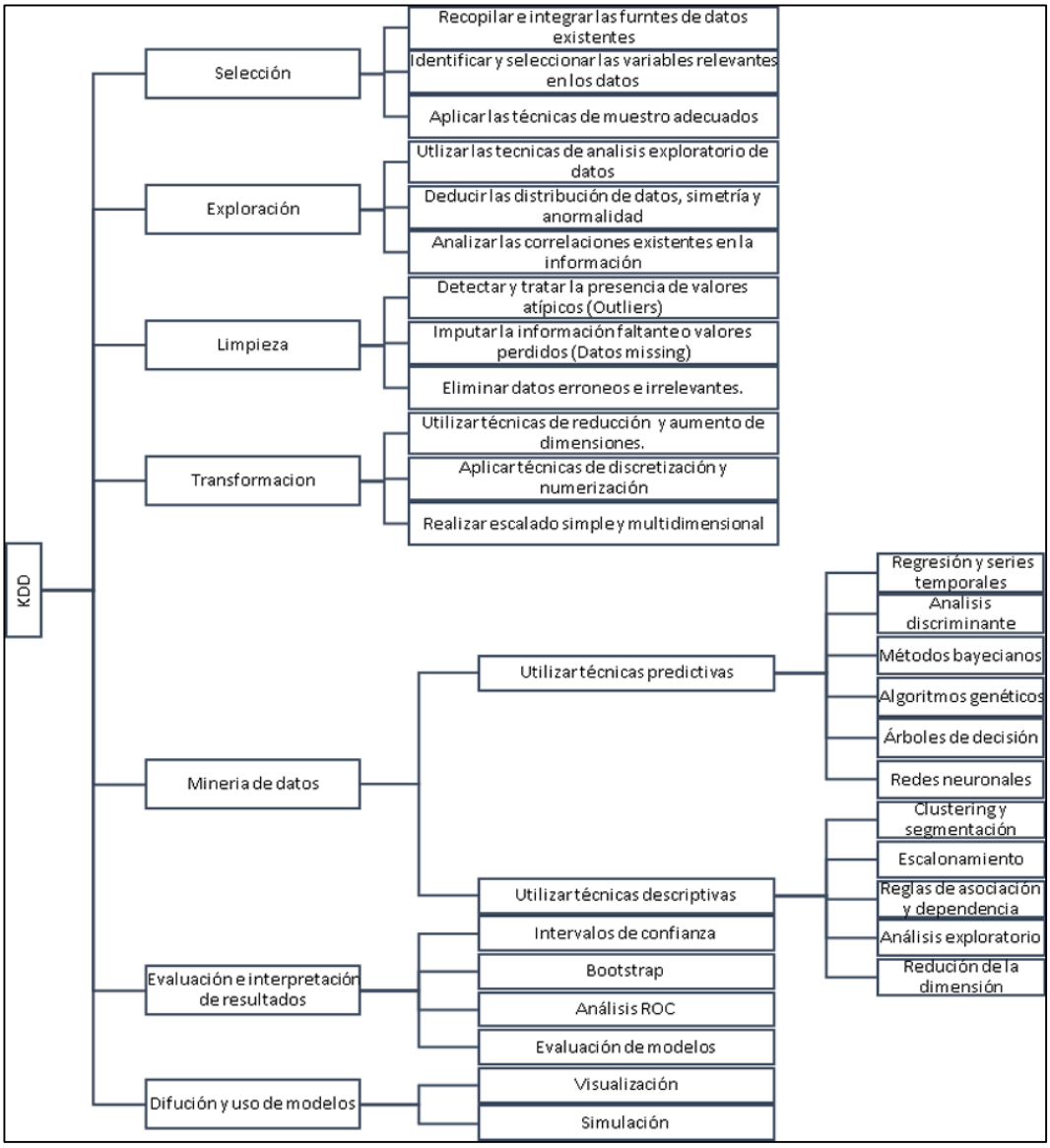


Figura 3. Esquema de clasificación del proceso de extracción del conocimiento

Fuente: Adaptado de (Perez & Santin, 2007)

1.1.2. Minería de datos

Gil (2009) 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.

Básicamente, 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 (p. 87).

Por otro lado, Microsoft (2018) 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.

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.

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 Mondragon (2007) 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 automáticamente 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 (p.10).

1.1.3. Técnicas de minería de datos

Las técnicas de minería de minería de datos de acuerdo con Perez & Santin (2007) podemos clasificarlo como: técnicas predictivas, técnicas descriptivas y técnicas auxiliares. A continuación, se detallaremos estas técnicas:

1.1.4. Clasificación de las técnicas de Data Mining

A continuación, se muestra una *clasificación de las técnicas de Data Mining.*

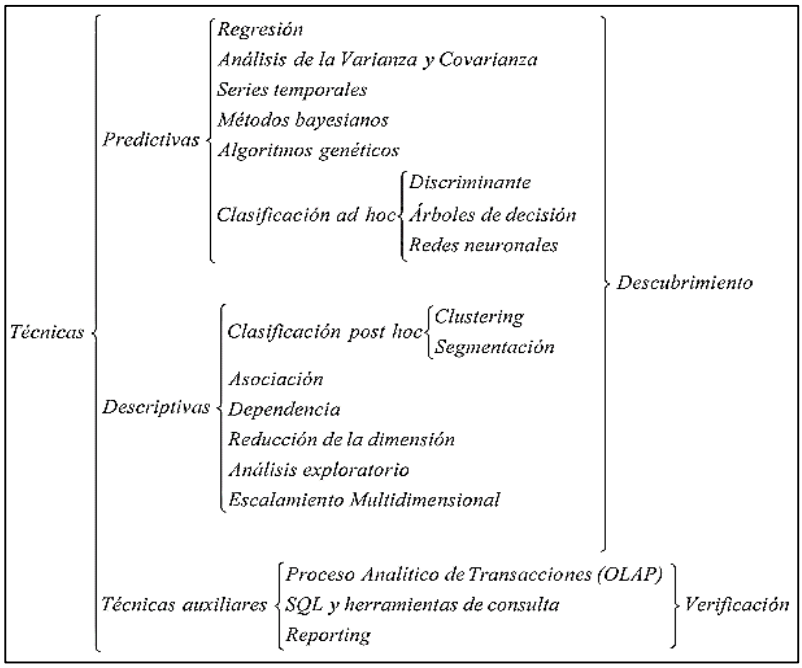


Figura 4. Esquema de clasificación de las técnicas de Data Mining

Fuente: Perez & Santin (2007)

En la figura 4 podemos observar que 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 previamente

definidos. Las técnicas descriptivas se denominan *técnicas de clasificación post hoc* porque realizan clasificación sin especificación previa de los grupos.

1.1.4.1. Técnicas predictivas o supervisadas

De acuerdo con Moreno *et al.* (2001) estos algoritmos predicen el valor de un atributo (*etiqueta*) de un conjunto de datos, conocidos otros atributos (*atributos descriptivos*). A partir de 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 usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos) (p. 3).

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 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 (Perez & Santin, 2007, pág. 8)

De acuerdo con Rosado & Verjel (2014) las técnicas predictivas tienen las tareas de clasificación y regresión, por otra parte (Valcárcel, 2004) 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 (Aluja, 2001) 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.

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 puede ser dicotómica o politómica y una o más variables independientes, llamémoslas X, que pueden ser de cualquier naturaleza, cualitativas o cuantitativas. Si la variable Y es dicotómica, podrá tomar el valor "0" si el hecho no ocurre y "1" si el hecho ocurre. Este proceso es denominado binomial ya que solo sólo tiene dos posibles resultados, siendo la probabilidad de cada uno de ellos constante en una serie de repeticiones. (Alderete, 2006, pág. 54)

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 esquemáticamente 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 (implementación de hardware de RNAs). (Flóres & Fernández, 2008, pág. 11)

Villada *et al.* (2014) menciona 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 delante. La figura 5 muestra una red de propagación hacia delante con dos capas ocultas. El número de entradas es directamente 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 unidireccionalmente 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.

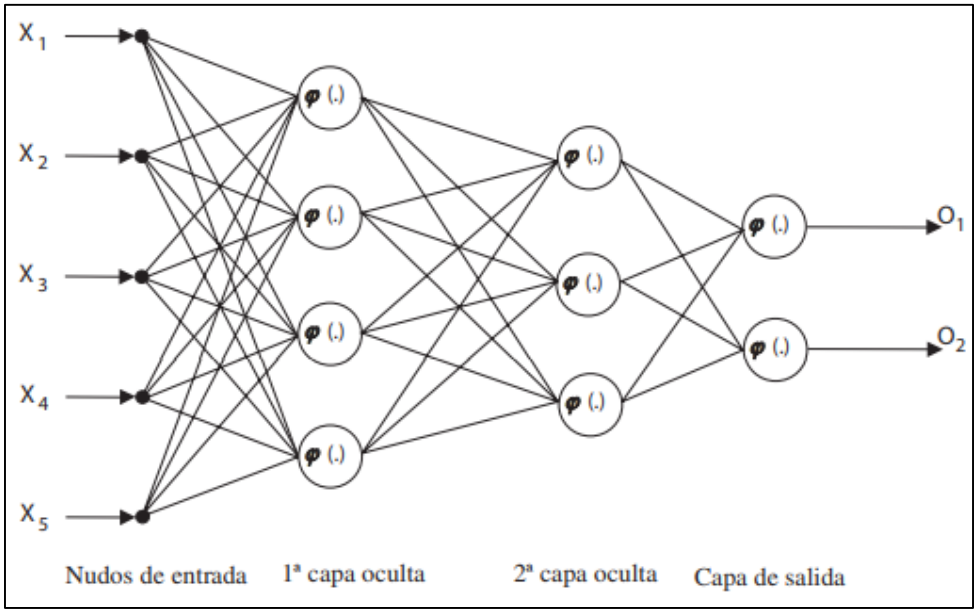


Figura 5. Red neuronal de propagación hacía adelante
Fuente: Villada *et al.* (2014)

Arboles de decisión

De acuerdo con Barrientos *et al.* (2009). 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 gráficamente 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 hoja corresponden a una decisión, la cual coincide con una de las variables clase del problema a resolver. (p. 20)

De acuerdo con Diaz (2007). Los arboles 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 frecuentemente "arboles 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 exactamente dos hijos. Estos son los más usados, debido a su simplicidad, aunque tampoco son infrecuentes los árboles que exhiben nodos con más de dos hijos. (p. 15)

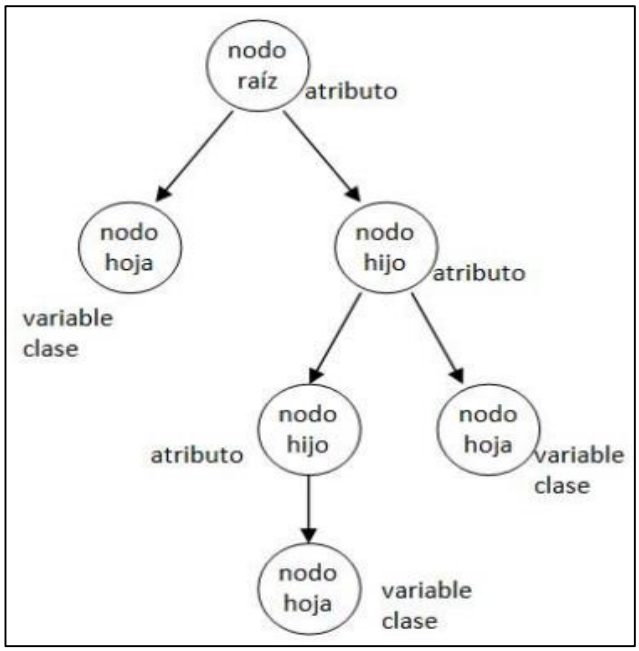


Figura 6. Estructura de un árbol de decisión

Fuente: Barrientos *et al.* **(**2009**)**

Boostrap

Esta técnica se enmarca entre los procedimientos de remuestreo, consistentes en generar un elevado número de muestras como base para estudiar el comportamiento de determinados estadísticos. 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 a modo general:

-A partir de una muestra original $\{X_1, X_2, X_3, X_4, \dots, X_n\}$, se extraen una nueva muestra $\{X_1^*, X_2^*, X_3^*, X_4^*, \dots, X_n^*\}$, por medio de muestreo con reposición. Es decir, tras la estación de un primer elemento, éste 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 infinitamente grande construido a partir de la información que provee la muestra.

-Para la muestra obtenida se calcula el valor de un determinado estadístico $\hat{?}$ que se utiliza como estimador del parámetro poblacional ?, en cuyo estudio estamos interesados.

-Repetimos los dos pasos anteriores, hasta obtener un elevado número de estimaciones $\hat{?}^*$. En este punto el recurso a herramientas informáticas que desarrollen las tareas de selección de muestras y determinación de las estimaciones resultará ineludible.

-Se construye una distribución empírica del estadístico $\hat{?}$, que representa una buena aproximación a la verdadera distribución de la probabilidad para ese estadístico. 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 formulas analíticas para determinar los correspondientes parámetros de distribución (Gil J. , 2005).

Bagging

De acuerdo a lo afirmado por Medina & Ñique (2017), "una forma natural de reducir la varianza, y por consiguiente aumentar la precisión de la predicción de un método de aprendizaje estadístico, es seleccionar una gran cantidad de conjuntos de entrenamiento de la población y construir un modelo de predicción independiente utilizando cada conjunto de entrenamiento. En otras palabras, se puede calcular, $\widehat{h^1}(x)$, $\widehat{h^2}(x), \dots,$ $\widehat{h^B}(x)$ utilizando *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 $\widehat{h^{*b}}(x)$, este promedio se obtiene:

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

Algoritmo CART

Según Kovalevski & Maca (2012), este algoritmo funciona de la siguiente manera: "Dado un conjunto de datos D = (X, Y), donde Y es la variable a explicar y X = (X1,. . ., Xp) es un vector de p variables 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 aplicar en muchos contextos distintos. En el caso en que la variable dependiente Y sea cualitativa, se dice que CART es un árbol de clasificación, y el 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".

Algoritmo Random Forest

Según afirman Villa *et al.* (2017), "esta técnica se basa en la construcción de árboles de predicción mediante el empleo de Bootstrap y Bagging, lo que garantiza la estabilidad del proceso. Cada árbol es construido usando muestras bootstrap con reposición a fin de corregir el error de predicción que se genera a consecuencia de la selección específica 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 que aproximadamente un tercio de la muestra original queda excluida de cada muestra generada por bootstrap. 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 previamente establecido y se restringe la selección de la variable de división a dicho conjunto. De esta forma se incluye una mayor variabilidad 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 de bootstrap 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 *et al.* (2017) 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 aleatoriamente 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 $\{T_b\}_1^B$ 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.

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 divisiones en más de dos subgrupos (Vallejo & Tenalanda, 2012)

Por otra parte, Joaquín (2017) menciona que: "C5.0 es el algoritmo sucesor de C4.5, ambos publicados por Quinlan, con el objetivo de crear 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".

Máquinas de soporte vectorial

De acuerdo con Montt *et al.* (2011). Las Máquinas de Soporte Vectorial (SVM) son una técnica de reconocimientos de patrones basada en la metodología de aprendizaje, generando resultados robustos y satisfactorios. Fueron desarrolladas como una herramienta robusta y sólida para regresión

y clasificación en dominios complejos, por Vladimir Vapnik y su equipo en los laboratorios AT&T. 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, Krikorian *et al.* (2011). Afirman que debido a que los problemas tomados de la realidad son difíciles 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 éste si es perfectamente separable, o separable bajo una cota de error aceptable.

Entre los kernels más utilizados por las SVM tenemos: Función de base radial o gaussiana, Lineal, polinómica, sigmoid.

1.1.4.2. 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 automáticamente partiendo 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, etc.) y de escalonamiento multidimensional.

Tanto las técnicas predictivas como descriptivas están enfocadas al descubrimiento del conocimiento embebido en los datos (Perez & Santin, 2007, pág. 9).

De acuerdo con (Moreno, Miguel, García, & Polo, 2001) 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.

Clustering

De acuerdo con García & Gómez (2012), es el 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" (p. 7).

Por otro lado, Garre *et al.* (2007) afirman que el proceso de clustering consiste 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, etc. (…). 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 (p. 9).

De acuerdo con Hair *et al.* (1999) los algoritmos para la obtención de conglomerados más utilizados pueden clasificarse en dos categorías generales: (1) jerárquicos y (2) no jerárquicos. A continuación, detallaremos estas dos técnicas (p. 510).

Clusters jerárquico: Dendograma

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

Hair *et al.* (1999) explican que, en los *métodos de aglomeración*, cada objeto u observación empieza dentro de su propio conglomerado. En 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. Eventualmente, 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 (p. 510).

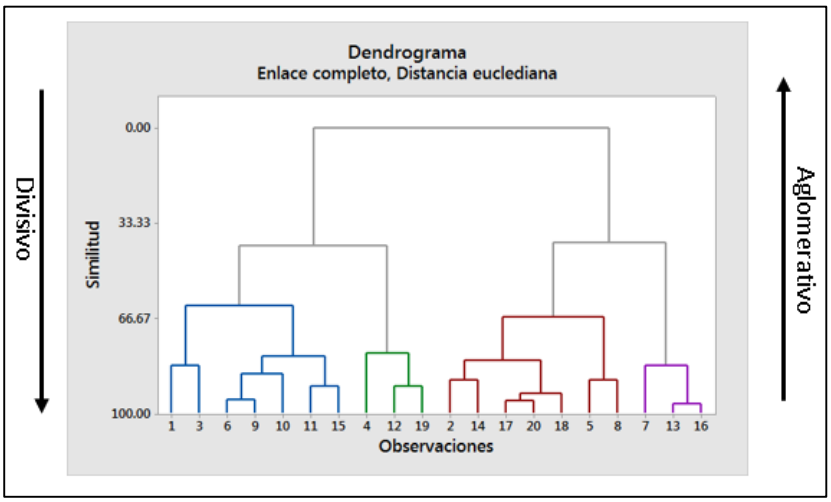


Figura 7. Dendograma ilustrativo de la obtención de conglomerados jerárquicos.

Fuente: Adaptado de (Hair *et al.*, 1999)

Entre los métodos aglomerativos tenemos los siguientes:

Método de las distancias mínimas o Enlace simple (single linkage)

Hair *et al.* (1999) afirman que este método se basa en la distancia mínima. Encuentra los dos objetos 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 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 (p. 511), 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*.

Método de las distancias máximas o Enlace completo (complete linkage)

Pérez (2004) 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" (p. 429).

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

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

Método del promedio entre grupos o Enlace promedio (average linkage)

Ato *et al.* (1990) mencionan que este algoritmo define la distancia como la media aritmética de todas las posibles entre dos puntos de dos conglomerados (p. 191).

Por otro parte Marín (2014) 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:

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

Método del centroide o Enlace centroide (centrod method)

Ato *et al.* (1990) afirma 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 (p. 191).

(Marín, 2014, pág. 489) menciona que, la distancia entre el conglomerado *AB* y el conglomerado C se calcula como:

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

Método de la mediana (median method).

Marín (2014) afirma que, en el método de agrupación de medianas, los dos conglomerados (o elementos) que se combinan reciben idéntica ponderación en el cálculo del nuevo centroide combinado, independientemente 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 (p. 490):

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

Método de Ward o Enlace por mínima varianza (momento central de orden dos o pérdida de inercia mínima).

Pérez (2004) afirma que, en el método de 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 aquéllos que resultan con el menor incremento en la suma total de las distancias al cuadrado intracluster. La métrica normalmente 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 (p. 430).

(Álvarez R. , 1995) menciona 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 la diagonal principal de la matriz de varianzas y covarianzas) sea mínima" (p. 208).

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.

Pascual (2010) menciona que los algoritmos de partición tratan de descubrir clusters reubicando iterativamente puntos entre subconjuntos. Por ejemplo, los métodos k-Medias y el de k-Medoides (PAM, CLARA, CLARANS), también pueden tener un enfoque probabilístico (EM, autoClass, MClust) (p. 49).

K Means

Es el método más utilizado más utilizado actualmente para realizar clustering.

Según Camana (2012) es un algoritmo de clasificación no supervisado, inventado por J. 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 euclideas entre los patrones de entrada y el centro más cercano (p. 70).

Flores (2014) 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 dentro del cluster no cambien, o equivalentemente, hasta que los centroides dejen de cambiar (p. 21).

De acuerdo con Peña (2012) este algoritmo funciona de la siguiente marera:

Supongamos una muestra de *n* elementos con *p* variables. El objetivo es dividir esta muestra en un número de grupos prefijado, G. El algoritmo de k-medias (que con nuestra notación debería ser de G−medias) requiere las cuatro etapas siguientes:

1) Seleccionar G puntos como centros de los grupos iniciales. Esto puede hacerse:
   a) Asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos así formados;
   b) Tomando como centros los G puntos más alejados entre sí;
   c) Construyendo los grupos con información a priori, o bien seleccionando los centros a priori.
2) Calcular las distancias euclídeas de cada elemento al centro de los G grupos, y asignar cada elemento al grupo más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.
3) Definir un criterio de optimalidad y comprobar si reasignando uno a uno cada elemento de un grupo a otro mejora el criterio.
4) Si no es posible mejorar el criterio de optimalidad, terminar el proceso (p. 228).

PAM (Partitioning Around Medoids)

Benítez (2005) 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 necesariamente, 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(p. 18).

EM (Expectation-Maximization)

De acuerdo con Benítez (2005) 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 (p. 22).

1.1.5. Metodologías para la minería de datos

Moine *et al.* (2011) 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 (p. 2).

De acuerdo con revisión de la literatura científica tres son los modelos de proceso de minería de datos actualmente utilizados: el modelo KDD, CRISP-DM (*Cross Industry Standard Process for Data Mining*) y SEMMA (Sample, Explore, Modify, Model, Assess). A continuación, detallamos los modelos de proceso de minería de datos CRISP-DM y SEMMA, dado que el modelo KDD lo detallamos al inicio de este capítulo.

1.1.5.1. CRISP-DM (Cross Industry Standard Process for Data Mining)

Moine *et al.* (2012) afirman que este modelo fue creado por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en 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 necesariamente 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.

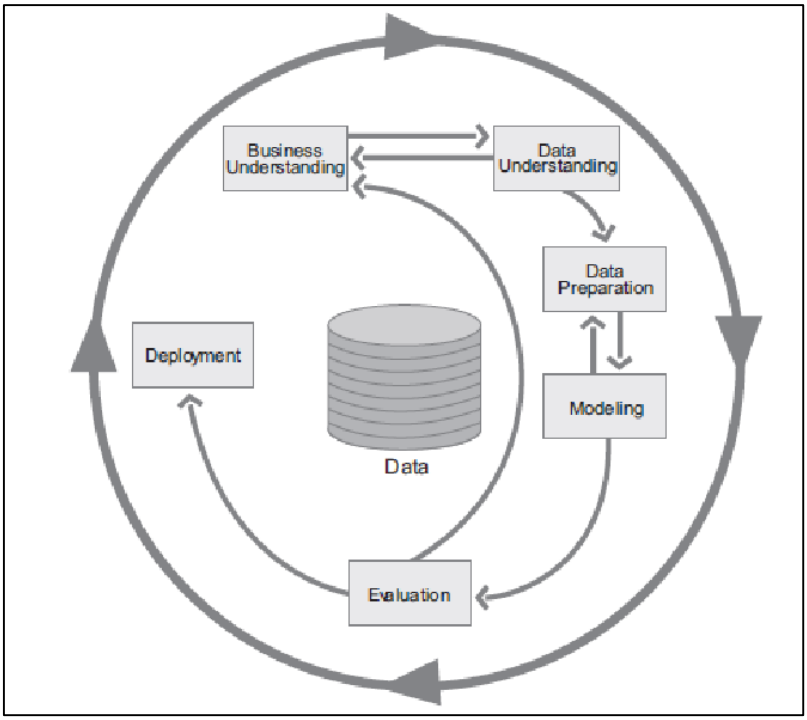


Figura 8. Fases del modelo CRISP-DM

Fuente: Tomado de (Chapman, 2000)

A continuación, describimos cada una de las seis fases, las mismas que tienen un conjunto de tareas detalladas en cuatro niveles de abstracción que van de lo más general a lo más específico: fase, tarea genérica, tarea especializada, e instancia de procesos.

1. Fase de compresión del problema o negocio

Como afirma Gallardo (2009) 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" (p. 17).

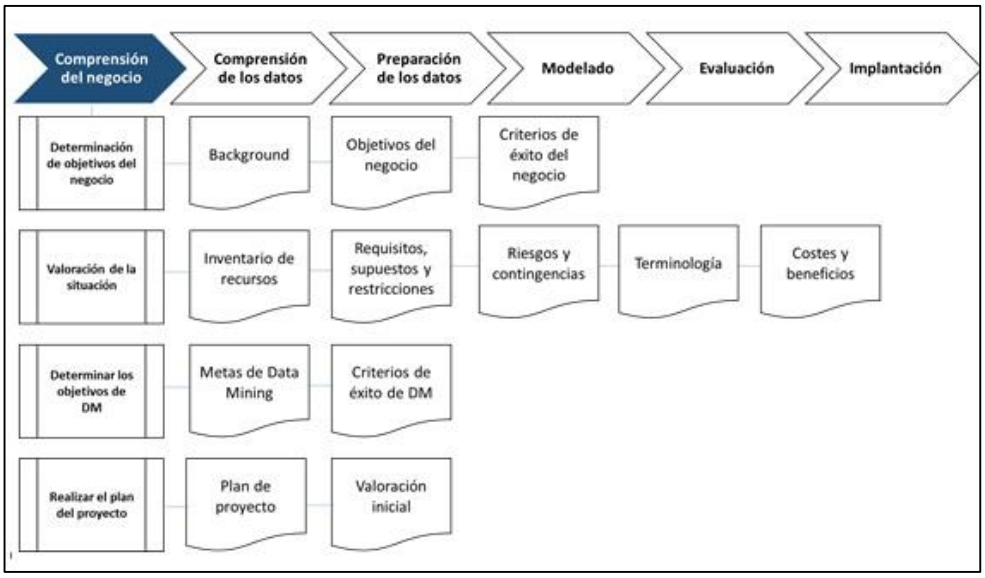


Figura 9. Comprensión del negocio
Fuente: Adaptado de (Chapman, 2000)

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 determinamos 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 debemos 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 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 definimos los requisitos tanto en términos de negocio como en términos de minería de datos.

Determinación de los objetivos de minería de datos

Gallardo (2009): afirma textualmente: "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" (p. 18).

Producción del plan de proyecto

Siendo esta la última tarea de la primera fase tiene como fin del desarrollo de un plan para el proyecto, que detalle los pasos a seguir y las técnicas a emplear en cada uno de los pasos.

2. Fase de compresión de los datos

Esta segunda fase (Figura 10) está relacionada con recolección inicial de los datos, podemos decir que la idea principal aquí es familiarizarse con los datos e información que se va a manejar, evaluar su calidad e identificar relaciones.

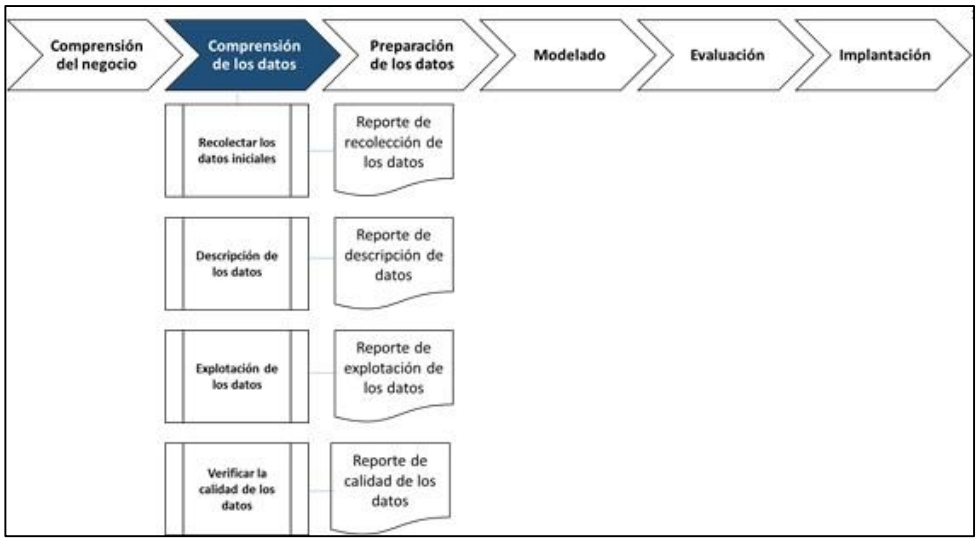


Figura 10. Comprensión de los datos

Fuente: Adaptado de (Chapman, 2000)

Comprende las siguientes tareas:

Recolección inicial de datos

Constituye la primera tarea en esta segunda fase de la metodología CRISP-DM, como menciona Rodríguez (s. f.) "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." (p. 6).

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.

3. 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 técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para la exploración de datos. Tal como se observa en la (figura 11) comprende las siguientes tareas:

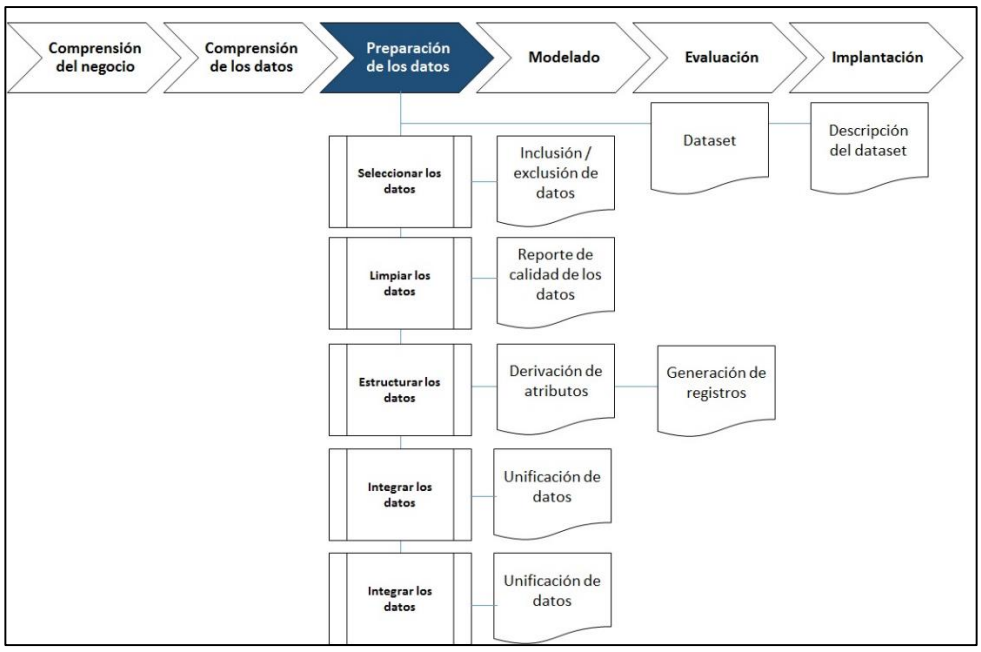


Figura 11. Preparación de los datos

Fuente: Adaptado de (Chapman, 2000)

Selección de los datos

Esta tarea consiste en seleccionar un subconjunto de los datos que se obtuvieron en la etapa anterior, apoyándose en criterios previamente establecidos en las fases anteriores (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

Con esta tarea consiste en la transformación de datos para realzar un análisis correcto de estos. Un ejemplo de esto podría ser, en el caso que deseamos predecir un valor numérico los datos necesariamente deben estar en dicho formato.

4. Fase de modelado

En esta fase, seleccionamos 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
- Conocimiento de la técnica

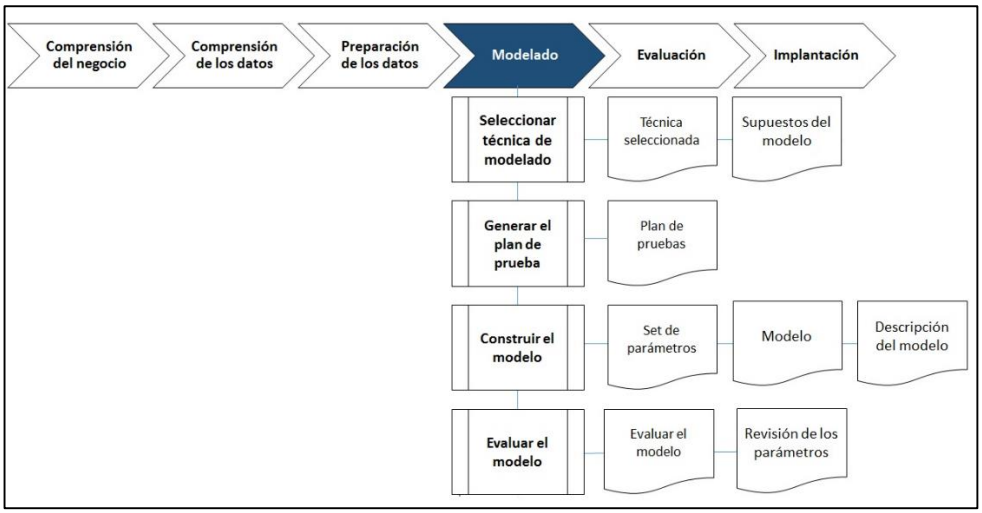


Figura 12. Modelado

Fuente: Adaptado de (Chapman, 2000)

Como se observa en la figura 12 esta fase tiene 4 tareas generales que a continuación detallamos:

Selección de la técnica de modelado

Esta tarea se refiere específicamente 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 de y su relación con las herramientas de minería de datos.

Generación del plan de prueba

Consiste en generar un procedimiento para probar la calidad y la eficiencia del modelo construido. Habitualmente se divide los datos en dos conjuntos uno de entrenamiento y otro de prueba, para posteriormente 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 anteriormente 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 de contexto del dominio.

5. 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.

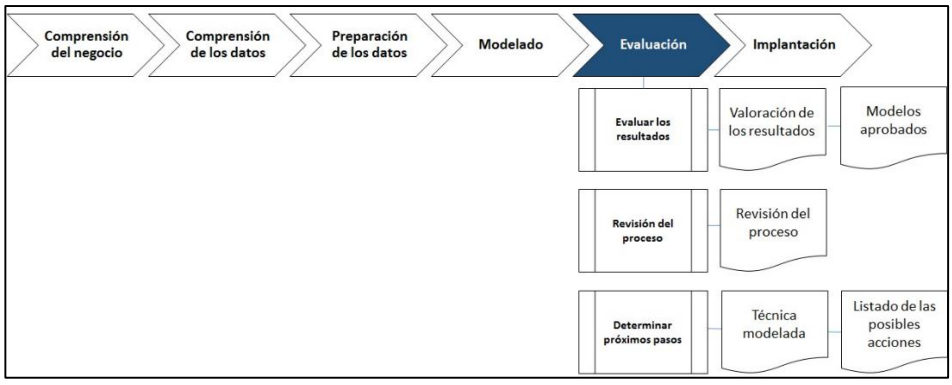


Figura 13. Evaluación

Fuente: Adaptado de (Chapman, 2000)

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.

Proceso de revisión

Consiste en la revisión de todo el proceso de minería de datos, con el objetivo de identificar elementos que pudieran ser mejorados.

Determinación de futuras fases

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 decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. En ocasiones es posible que en esta fase se decida partir desde cero con un nuevo proyecto de minería de datos.

6. 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, aplicando el modelo a diferentes conjuntos de datos o como parte del proceso.

Los Proyecto 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.

Como se puede apreciar en la figura 14, 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.

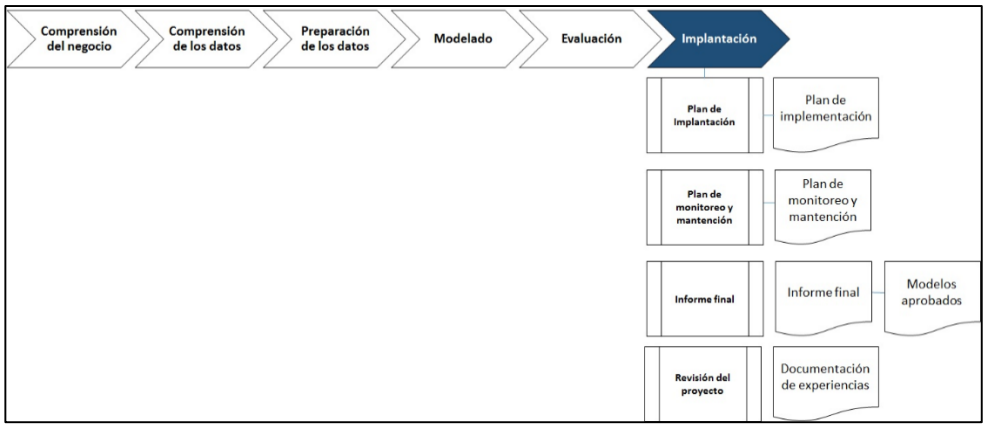


Figura 14. Despliegue

Fuente: Adaptado de (Chapman, 2000)

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 fase (Gallardo, 2009) 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" (p. 24).

Informe final

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

Revisión del proyecto

Esta tarea involucre la evaluación de lo que fue correcto y que fue incorrecto, es decir que es lo que se hizo bien y que es lo que se necesita mejorar.

1.1.5.2. SEMMA (Sample, Explore, Modify, Model, Assess)

Britos (2008) define a esta metodología como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos. El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso (Figura 15).

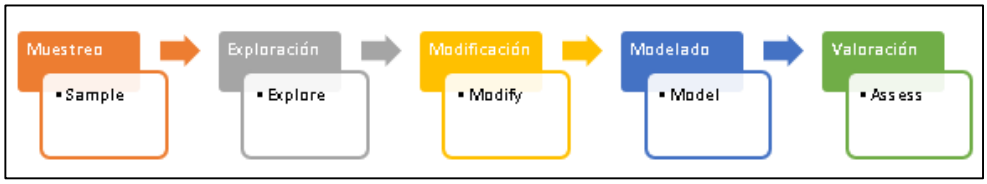


Figura 15. Fases de la metodología SEMMA
Fuente: Adaptado de (Britos, 2008)

De acuerdo con Hernández & Dueñas (2009) detallamos cada uno de las fases:

Muestreo

En esta etapa se realiza la extracción de una muestra de los datos que permita representar características comunes de la población para posteriormente comenzar el análisis de los mismos. Con esta fase 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 logrando detectar, identificar y posteriormente 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 de acuerdo con éstas 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 asociaciones o combinaciones entre los datos, logrando 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.

1.1.6. 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, la accuracy, Classification-error 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

Donde VP son los verdaderos positivos y VN verdaderos negativos que vienen a ser la cantidad de observaciones que el clasificador predijo correctamente como la clase positiva y negativa. El FP son los falsos positivos y FN falsos negativos que vienen a ser la cantidad de observaciones que el clasificador predice incorrectamente como clase positiva siendo la clase negativa y como negativa siendo la clase positiva respectivamente. A partir de esta tabla se puede calcular la exactitud, el error de clasificación.

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

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

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

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

Donde:

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

Coeficiente de Kappa (k)

De acuerdo con Cerda & Villarroel (2008): “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 and Koch, 1977)

Tabla 2

Valoración del coeficiente de kappa (Landis y Koch, 1977)

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: (Cerda & Villarroel, 2008)

1.1.7. Rendimiento académico

De acuerdo con Reyes (2003), 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, etc.

Para el presente estudio se consideró como bajo rendimiento académico cuando un estudiante obtiene el promedio ponderado menor que el mínimo exigido por la Universidad Nacional Amazónica de Madre de Dios que es 11, en una escala de 0 a 20.

1.2. Antecedentes

Los estudios previos que guardan relación con este trabajo de investigación son los siguientes:

Bacallao *et al.* (2004). 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, exclusivamente, 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.

Dapozo *et al.* (2006), mencionan que la Minería de Datos abarca una variedad de métodos 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). 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.

Pereira (2009), 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 aplicando 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 minería 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 su utilizó el algoritmo C4.5 y para las reglas de Asociación, el algoritmo EquipAsso.

Spositto *et al.* (2010), presentan los resultados de la evaluación del rendimiento académico y de la deserción estudiantil de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas (DIIT) de la Universidad Nacional de La Matanza (UNLAM). La investigación se realizó aplicando el proceso de descubrimiento de conocimiento sobre los datos de alumnos del período 2003-2008. La implementación de este proceso se realizó con el software MS SQL Server para la generación de un almacén de datos, el software SPSS para realizar un preprocesamiento de los datos y el software Weka (Waikato Environment for Knowledge Analysis) para encontrar un clasificador del rendimiento académico y para detectar los patrones determinantes de la deserción estudiantil. En la fase de minería de datos se aplicó la tarea de clasificación de la minería de datos, la técnica elegida fue árboles de decisión utilizando para ello el algoritmo C4.5 cuya implementación en la herramienta Weka es conocido como J48.

La Red Martinez *et al*. (2010), mencionan que aplicaron la técnicas de almacenes de datos (DataWarehouses: DW) y de minería de datos (Data Mining: DM) basadas en clustering, entre otras, para la búsqueda de perfiles de los alumnos de la asignatura de sistemas operativos, según su rendimiento académico, situación demográfica y socio económica, con el propósito de determinar a priori situaciones potenciales de éxito o de fracaso académico, lo cual permitiría encarar las medidas tendientes a minimizar los fracasos.

Gatica *et al.* (2010), mencionan que los estudiantes ingresan a la universidad con diversas habilidades y características personales, familiares y académicas que influyen en su desempeño escolar, y se enfrentan a mayores compromisos que los del bachillerato pues las metas académicas son más exigentes. Los 2 primeros años de la licenciatura son decisivos para el alumno, ya que en este periodo define la continuación o abandono de sus estudios. Este estudio tuvo como objetivo: Identificar las variables de factores académicos, personales y socioeconómicos asociadas al éxito académico durante los 2 primeros años de la carrera, en los estudiantes de la Facultad de Medicina de la UNAM. Material y métodos: Estudio observacional retrospectivo, donde se utilizaron las evaluaciones diagnósticas de primer ingreso en conocimientos generales, español e inglés, la encuesta socioeconómica de ingreso a la UNAM, las bases de datos

de Servicios Escolares de la Facultad y los promedios porcentuales de todos los exámenes departamentales de primero y segundo año. Se trabajaron variables agrupadas en factores académicos, socioeconómicos y personales. Análisis estadístico: ANOVA de un factor, t de student para muestras independientes, chi cuadrada, regresión lineal simple y árboles de clasificación jerárquica. Resultados: Se estudiaron 945 estudiantes (626 mujeres y 319 hombres), con edad promedio de 18.4 años. El alumno académicamente exitoso es quien cursa en su primera ocasión sus asignaturas y acredita con una puntuación superior a la media más una desviación estándar en los exámenes departamentales de primer y segundo año de la carrera. Se analizaron 3 grupos de variables. Variables académicas: rendimiento en el examen diagnóstico de español ≥ 75, bachillerato de procedencia, rendimiento académico en la evaluación de inglés (≥ 51.79), de conocimientos generales (≥ 61) y un promedio de egreso del bachillerato ≥ 9. Variable personal: sexo femenino. Variable socioeconómica: escolaridad del padre (licenciatura o posgrado).

Berlanga *et al*. (2013), afirman que un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento. Nos ayudan a tomar la decisión más "acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones. Estos árboles permiten examinar los resultados y determinar visualmente cómo fluye el modelo. Los resultados visuales ayudan a buscar subgrupos específicos y relaciones que tal vez no encontraríamos con estadísticos más tradicionales. Los árboles de decisión son una técnica estadística para la segmentación, la estratificación, la predicción, la reducción de datos y el filtrado de variables, la identificación de interacciones, la fusión de categorías y la discretización de variables continuas. La función árboles de decisión (Tree) en SPSS crea árboles de clasificación y de decisión para identificar grupos, descubrir las relaciones entre grupos y predecir eventos futuros.

Formia *et al*. (2013), presentan un estudio realizado en la Universidad Nacional de Río Negro (UNRN), y en particular en la Sede Atlántica desde la Licenciatura en Sistemas, donde abordaron el fenómeno de deserción y desgranamiento que se han podido apreciar en los cuatro primeros años de vida de la Institución. En su trabajo describen el proceso de identificación de las características más relevantes del problema a través de las cuales, utilizando técnicas de Minería de Datos (DM), puede obtenerse un modelo de la deserción universitaria en la unidad académica mencionada. Para identificar las

características más relevantes se propone analizar, luego del preprocesamiento de los datos, las proyecciones de los atributos en las clases o respuestas esperadas. Su aplicación a los datos de los alumnos de las carreras de grado de la UNRN ha ofrecido resultados satisfactorios permitiendo efectuar recomendaciones tendientes a reducir el porcentaje de alumnos que abandona la carrera. En este estudio utilizaron el algoritmo predictivo de árboles de decisión denominado C4.5.

Heredia *et al*. (2015). Manifiestan que la minería de datos permite descubrir información oculta en grandes cantidades de datos, lo cual es muy difícil de visualizar con 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, caracterizando 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.

La Red Martínez *et al*. (2015), manifiestan que el rendimiento académico es un factor crítico teniendo en cuenta que, frecuentemente, el bajo rendimiento académico está asociado a una alta tasa de deserción. Esto se ha observado en asignaturas del primer nivel de la carrera de Ingeniería en Sistemas de Información (ISI) de la Universidad Tecnológica Nacional Facultad Regional Resistencia (UTN-FRRe), situada en la ciudad de Resistencia, provincia del Chaco, Argentina, entre ellas Algoritmos y Estructura de Datos, donde el bajo rendimiento académico se observa en proporciones muy altas (entre el 60% y el 80% aproximadamente en los últimos años). En este trabajo se propone la utilización de técnicas de minería de datos sobre información del desempeño de los alumnos de la asignatura mencionada con el propósito de caracterizar los perfiles de alumnos exitosos (buen rendimiento académico) y de aquellos que no lo son (bajo rendimiento académico). La determinación de estos perfiles permitiría a futuro definir acciones específicas tendientes a revertir el bajo rendimiento académico, una vez detectadas las variables asociadas al mismo. En este artículo se describen los modelos de datos y de minería de datos utilizados y se comentan los principales resultados obtenidos.

Salinas (2016), manifiesta que en los últimos semestres el número de estudiantes desaprobados en el curso Estadística General ha correspondido a un 41%. Por ello, en el presente estudio 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 de dichos alumnos, y que tal dependencia puede expresarse a través de un modelo estadístico. Usando 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 con un aproximado de 1500 alumnos, y se encontraron las principales variables sociodemográficas y académicas que determinan la situación del rendimiento académico (aprobado y desaprobado). Usando esta información se puede predecir la situación final del alumno (aprobado o desaprobado) apenas el alumno se matricule en el curso sin haber rendido ningún tipo de evaluación.

Álvarez & Cuji (2016), presentan un estudio sobre deserción estudiantil que tuvo como objetivo primario detectar el porcentaje de abandono escolar que presenta la carrera de Docencia en Informática, a partir del año 2006 hasta el año 2015, con base en ésta información se aplicó el algoritmo de Arboles de decisión para diseñar un prototipo de modelo predictivo de Deserción Estudiantil, la metodología usada se basa en el método KDD(Knowledge Discovery in Database), detallado en cinco etapas , selección, procesamiento, transformación, minería de datos, e interpretación de la información. Posterior a la aplicación del algoritmo se obtuvo un árbol de decisión de cuatro niveles de profundidad, evidenciando que las variables nivel y notas tienen mayor influencia en la Deserción Estudiantil dentro de la Carrera. Finalmente se obtuvieron cuatro reglas que fueron programadas y visualizadas en una interfaz web, que evalúa a los nuevos posibles desertores de la Carrera de Docencia en Informática.

(Yamao, 2018), presenta un estudio donde se realiza la predicción del rendimiento académico de los alumnos que ingresaron a la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martin de Porres en el primer ciclo utilizando minería de datos. Se extrajeron datos de 1304 ingresantes que fueron clasificados en tres factores: sociales, económicos y académicos. Se realizaron predicciones a través de tres técnicas: regresión lineal, árbol de decisiones y support vector machines, y el mejor resultado de 82.87% se obtuvo utilizando árbol de decisiones. De los diferentes factores, los que más influyeron en el rendimiento

académico fueron los siguientes: nota de examen de admisión, género, edad, modalidad de ingreso y distancia desde su casa hasta el centro de estudios. Utilizando minería de datos fue posible realizar predicciones del rendimiento académico de los ingresantes. Esto permitió la detección de ingresantes que podrían enfrentarse a problemas en sus estudios.

Jimenez (2017), en su tesis doctoral desarrollado en la ciudad de Puno, entre los años 2016-2017, cuyo objetivo principal es 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 CRISP-DM (Cross Industry Standard Process for Data Mining), 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 Ing. 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.

Coyla (2017), presentó los resultados de la exploración de los datos de los postulantes e ingresantes a la Universidad Nacional del Altiplano de Puno, que tuvo como objetivo general, Identificar características y patrones de comportamiento con el desempeño académico, utilizando Bigdata, manifestando que para ello se ha empleado el diseño cuasi experimental, tomando un grupo experimental de 18 ingresantes a la E. P. de Ingeniería de Sistemas – proceso de admisión CEPREUNA enero marzo 2015. Para la implementación del bigdata se utilizó la metodología SEMMA. El paquete Rattle diseñado en el Lenguaje R son útiles para explorar, analizar y manipular base de datos gigantes, analizando los datos de los ingresantes a la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional del Altiplano del proceso de admisión CEPREUNA Enero Marzo del año 2015, se llegó a los siguientes resultados, el 51 % de ingresantes conocen problemas de matemáticos I. Los ingresantes razonan y demuestran

proposiciones matemáticas, representan, analizan e interpretan datos matemáticos contextualizados y resuelven problemas matemáticos contextualizados. El 62% de ingresantes no resuelven problemas de matemática II. El 68 % de ingresantes marcaron erradamente las alternativas de las preguntas referidos a Física, el 67 % de ingresantes marcaron erradamente las alternativas de las preguntas referidos a Química, el 53 % de ingresantes conocen problemas de razonamiento matemático y el 64 % de ingresantes conocen problemas de razonamiento verbal.

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1. Identificación del problema

Más del 90% de la información de todo el mundo para el 2007 ya estaba en formato digital, en este mismo año, 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 (Hilbert & López, 2011), mucha de esta información es generado producto de las operaciones que a diario realizamos 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 capturado por dispositivos móviles, computadores personales para luego ser almacenados en grandes bases de datos.

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

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. Considerando 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”

(UNAMAD, 2018), uno de los aspectos importantes a ser analizados y evaluados para cumplir con esta misión es el rendimiento académico de los estudiantes, como lo afirma (Garbanzo, 2007) textualmente: “el rendimiento académico constituye un indicador importante a la hora de valorar la calidad educativa en la educación superior”.

Estos datos almacenados en los repositorios de la DUAA, representan un desafío a tener que enfrentarnos para analizar y descubrir nuevos conocimientos ocultos que nos permitan explicar de mejor manera el rendimiento académico. En este contexto nos formulamos la siguiente interrogante general:

- ¿Cuáles son los patrones de bajo rendimiento académico de los estudiantes de la universidad nacional amazónica de madre de dios 2018?

Y las siguientes interrogantes específicas:

- ¿Cuáles son 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?
- ¿Cuáles el modelo de clasificación que permite predecir las condiciones que cumplen los estudiantes con bajo rendimiento académico de la universidad nacional amazónica de madre de dios 2018?
- ¿Cuáles son los perfiles de bajo rendimiento académico de los estudiantes de la universidad nacional amazónica de madre de dios 2018?

2.2. Justificación

La minería de datos en sector educativo o minería de datos educativa es uno temas emergentes debido a la gran cantidad de datos que se generan a diario en las instituciones de educación básica y educación superior públicas o privadas de nuestro país, la Universidad Nacional Amazónica de Madre de Dios no es ajena a ello, en la dirección universitaria de asuntos académicos de esta casa superior de estudios continuamente se generan y se almacenan datos de los estudiantes matriculados, retirados, aprobados y desaprobados, sin embargo, estos datos no están siendo aprovechados para la extracción del conocimiento oculto en ellos, la importancia del presente estudio radica en que mediante ella podremos analizar el rendimiento

académico y determinar los perfiles de estudiantes con bajo rendimiento académico mediante técnicas de clasificación, específicamente mediante arboles de clasificación.

2.3. Objetivos

2.3.1. 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.

2.3.2. 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.

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1. 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.

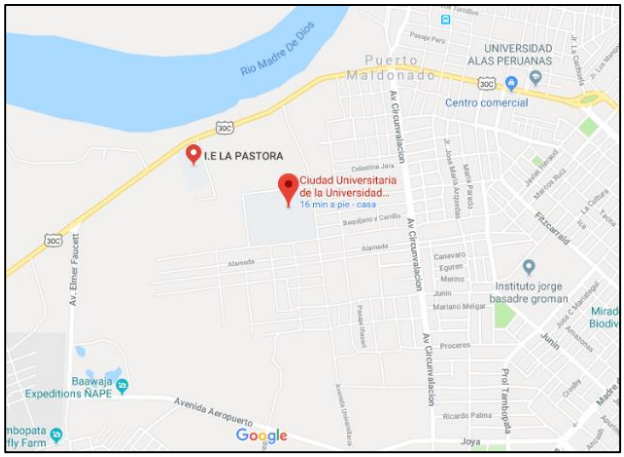


Figura 16. Ubicación geográfica-UNAMAD

Fuente: https://www.google.com/maps/@-12.5896145,-69.212753,16z

3.2. Población y tamaño de muestra

3.2.1. Población

La población para el 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 9545 registros.

Tabla 3

Registros de proceso de matrícula UNAMAD del 2001 al 2018

Id	Semestres	Total de registros
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	1005
17	2017-I, 2017-II	1060
18	2018-I	528
	Total	9545

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

3.2.2. Muestra

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

3.3. Método de investigación

Dado que la naturaleza del presente estudio implica 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 estudio es: no experimental-documental (Arias, 2006), la metodología optada para el logro de los objetivos es CRISP-DM.

3.4. Descripción detallada de métodos por objetivos específicos

Para el logro del objetivo específico Nro. 1: "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", se empleó 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.

Par cumplir con el objetivo específico Nro. 2: "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", se empleó 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 paquete C50 y rpart al IDE RStudio respectivamente.

La Identificación 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.

Se debe aclarar que una descripción más detallada se encuentra en el siguiente capítulo, donde se observará la aplicación de la metodología CRISP-DM.

CAPÍTULO IV

RESULTADOS Y DISCUSIONES

4.1. Aplicación de la metodología CRISP-DM

En el presente estudio, se utilizó el modelo de procesos Cross Industry Standard Process for Data Mining (CRISP-DM) ampliamente abordado en el marco teórico.

4.1.1. Fase 1: Comprensión del negocio

(UNAMAD, 2016) 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 veterinario y zootecnista, en las diversas especialidades de acuerdo con las demandas esenciales de la región" (p. 17)

1) Determinar los objetivos del negocio

a. 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.

b. Objetivos del negocio

(UNAMAD, 2016) 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" (p. 18).

En la figura 17 se presenta los ejes estratégicos institucionales:

Figura 17. Ejes estratégicos institucionales

Fuente: Tomado del plan estratégico institucional (UNAMAD, 2016)

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" (UNAMAD, 2016).

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.

2) Evaluación de la situación

a. Inventario de recursos

Los recursos con los que se cuenta para el desarrollo del proyecto son: 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 cuenta es el histórico de datos de los procesos de matrícula y evaluación de la oficina de la Dirección Universitaria de Asuntos Académicos de la UNAMAD.

b. Requisitos, supuestos y restricciones

El personal del proyecto cuenta con los conocimientos necesarios de minería de datos para lograr los objetivos planteados, no se presentan restricciones, por cuanto no se expondrá información personal de los estudiantes.

c. Costos y beneficios

Costo de hardware para el proyecto de minería de datos:

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**

Costo de software para el proyecto de minería de datos:

Tabla 5

Costo de software

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

Recursos humanos para el proyecto de minería de datos:

Tabla 6

Recursos humanos

Recurso humano	Mes 1	Mes 2	Mes 3	Total
Analista de datos	S/. 4200,00	S/. 4200,00	S/. 4200,00	S/. 12600,00
		Total		**S/. 12600,00**

Fuente: http://unete.sunat.gob.pe/images/2018/CAS/095/095.pdf recuperado de CONVOCATORIA CAS N° 095 - 2018 "ANALISTA DE DATOS JUNIOR"-SUNAT

Total, de inversión

Tabla 7

Cuadro resumen de costos

Detalle	Costos
Costo de hardware para el proyecto de minería de datos.	S/. 2950,00
Costo de software para el proyecto de minería de datos.	S/. -00,00
Recursos humanos para el proyecto de minería de datos.	S/. 12600,00
Total	**S/. 15550,00**

El proyecto asciende a 15500 soles, los cuales serán asumidos por el tesista.

El presente proyecto no genera beneficios económicos, el beneficio que presenta es descubrir los perfiles de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios, esto permitirá de 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 son 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 de 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 realiza la solicitud de la base de datos histórica de los procesos académicos a la Oficina del Vicerrectorado académico, dado que la DUAA depende jerárquicamente de esta (Ver Anexo 01).

Segunda etapa: Análisis los datos que emite 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 son:

Tabla 8

Herramientas para la minería de datos empleadas

Herramienta	Descripción
	Entorno y lenguaje de programación
 **RStudio**	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: Tomado de (RStudio, 2018)

Las técnicas de minería de datos empleadas son:

Tabla 9

Técnicas de minería de datos empleadas

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

4.1.2. Fase 2: Compresión de los datos

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

a. 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 9923 instancias) en formato xlsx, que se detalla a continuación:

Monto_matricu	universi	ditos_m	editos_a	ditos_de	mestre_i	Edad_ingre	Modalidad_ingreso(CEPRE/Primera opcion/Ord	Satisfacion_carrera(Primera/segunda o	Nota_ingre	Tipo_preparacion(CEPRE/Part	Prodio_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

Figura 18. Reporte de datos acumulado

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

b. 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 opcion/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.

c. 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:

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

```
1  library("readxl")
2  setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
3  datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
```

Untitled1* | datos

Filter

	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

Figura 19. Datos cargados en RStudio

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:

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
tablaDep<-as.data.frame(table(Departamento=datos$departamento))
tablaDep
transform(tablaDep,
          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	AMAZONAS	17	17	0.002	0.002	0.2	0.171
2	ANCASH	8	25	0.001	0.003	0.1	0.252
3	APURIMAC	176	201	0.018	0.020	1.8	2.026
4	AREQUIPA	230	431	0.023	0.043	2.3	4.343
5	AYACUCHO	60	491	0.006	0.049	0.6	4.948
6	CAJAMARCA	18	509	0.002	0.051	0.2	5.129
7	CALLAO	23	532	0.002	0.054	0.2	5.361
8	CUSCO	2075	2607	0.209	0.263	20.9	26.272
9	HUANCAVELICA	13	2620	0.001	0.264	0.1	26.403
10	HUANUCO	45	2665	0.005	0.269	0.5	26.857
11	ICA	45	2710	0.005	0.273	0.5	27.310
12	JUNIN	60	2770	0.006	0.279	0.6	27.915
13	LA LIBERTAD	15	2785	0.002	0.281	0.2	28.066
14	LAMBAYEQUE	21	2806	0.002	0.283	0.2	28.278
15	LIMA	322	3128	0.032	0.315	3.2	31.523
16	LORETO	42	3170	0.004	0.319	0.4	31.946
17	MADRE DE DIOS	5742	8912	0.579	0.898	57.9	89.812
18	MOQUEGUA	21	8933	0.002	0.900	0.2	90.023
19	PASCO	13	8946	0.001	0.902	0.1	90.154
20	PIURA	13	8959	0.001	0.903	0.1	90.285
21	PUNO	456	9415	0.046	0.949	4.6	94.881
22	SAN MARTIN	104	9519	0.010	0.959	1.0	95.929
23	SIN DATOS	271	9790	0.027	0.987	2.7	98.660
24	TACNA	44	9834	0.004	0.991	0.4	99.103
25	UCAYALI	89	9923	0.009	1.000	0.9	100.000

Figura 20. Distribución de frecuencias de la variable departamento

Diagrama de barras para la variable departamento:

Script en el lenguaje R:

```
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"))
```

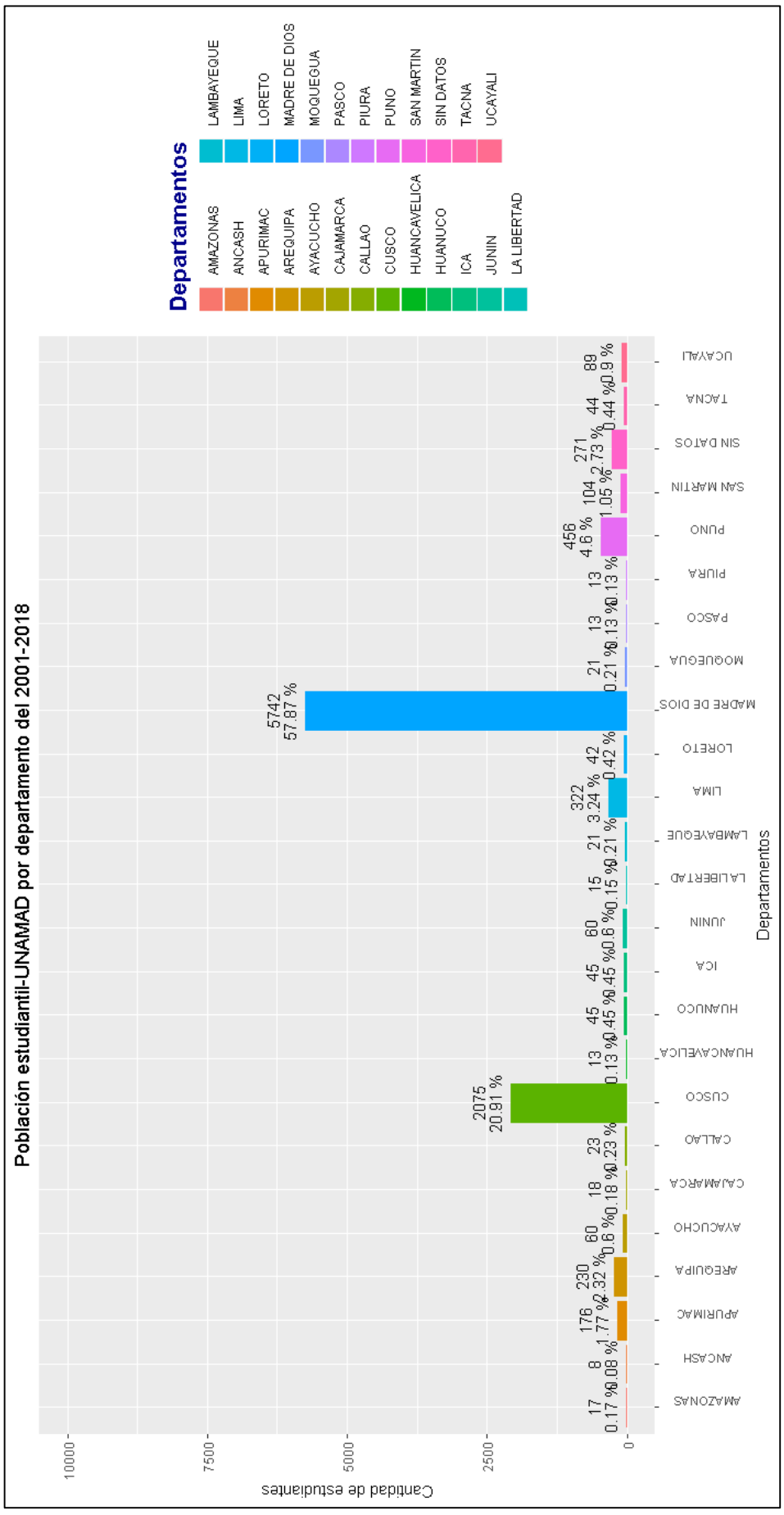


Figura 21. Población estudiantil-UNAMAD por departamentos del 2001-2018

En la figura 21 se puede observar que la población estudiantil de la Universidad Nacional Amazónica de Madre de Dios, está formado principalmente por un 57.87% de estudiantes procedentes del departamento de Madre de Dios, un 20.91% procedentes de la región del Cusco, otro 4.6% proceden del departamento de Puno.

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:

```
 1  library("readxl")
 2  setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
 3  datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
 4  #cargamos dplyr para filtrar
 5  library(dplyr)
 6  provMDD<-filter(datos,datos$departamento=="MADRE DE DIOS")
 7  provMDD
 8  tablaProv<-as.data.frame(table(Provincia=provMDD$provincia))
 9  transform(tablaProv,
10            FreqAc=cumsum(Freq),
11            Rel=round(prop.table(Freq),3),
12            RelAc=round(cumsum(prop.table(Freq)),3),
13            Porcentaje=round(prop.table(Freq),3)*100,
14            PorcentajeAc=round(cumsum(prop.table(Freq)*100),3)
15  )
  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
```

Figura 22. Tabla de frecuencias: estudiantes por provincias-Madre de Dios

Diagrama de barras departamento de Madre de Dios:

Script en el lenguaje R:

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
library(dplyr)
provMDD<-filter(datos,datos$departamento=="MADRE DE DIOS")
provMDD
library (ggplot2)
qplot(data=provMDD, factor(provMDD$provincia), geom="bar",
  ylab="Cantidad de estudiantes",xlab="Provincias",
  fill=factor(provMDD$provincia))+
  theme(axis.text.x=element_text(size=8, angle=0))+
  scale_fill_discrete(name = "Provincias")+
  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, 7000))+
  ggtitle("Población estudiantil-UNAMAD 2001-2018\n Distribución por provincias del
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4", size=16, face="bold"))
```

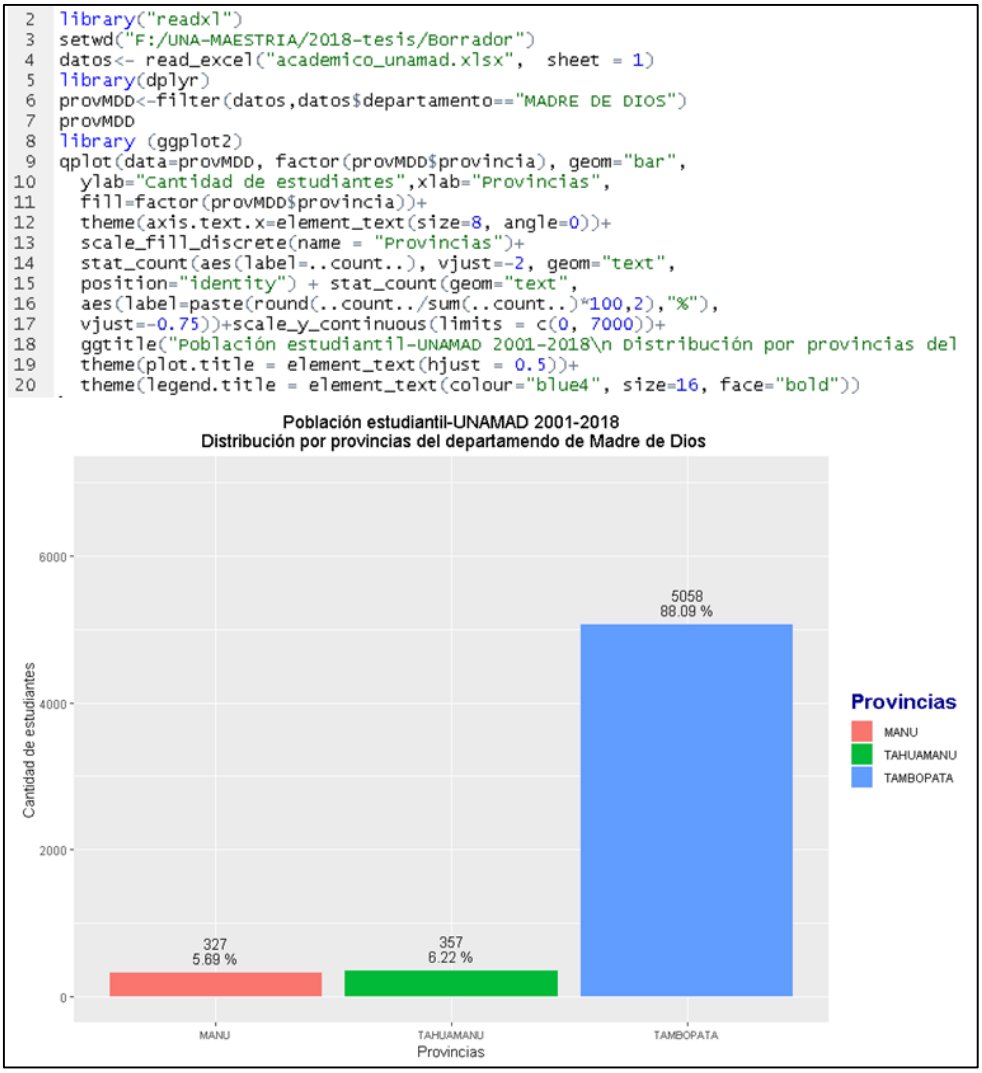


Figura 23. Distribución de estudiantes por provincias-Madre de Dios (UNAMAD 2001-2018)

La figura 23, representa 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:

```
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
```

Figura 24. Tabla de frecuencias: estudiantes por provincias-Cusco (UNAMAD 2001-2018)

Diagrama de barras departamento de Cusco:

Script en el lenguaje R:

```
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,datos$departamento=="CUSCO")
provC
library (ggplot2)
qplot(data=provC, factor(provC$provincia), geom="bar",
  ylab="Cantidad de estudiantes",xlab="Provincias",
  fill=factor(provC$provincia))+
  theme(axis.text.x=element_text(size=8, angle=90))+
  scale_fill_discrete(name = "Provincias")+
  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, 1000))+
  ggtitle("Población estudiantil-UNAMAD 2001-2018\n Distribución por provincias
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4", size=16, face="bold"))
```

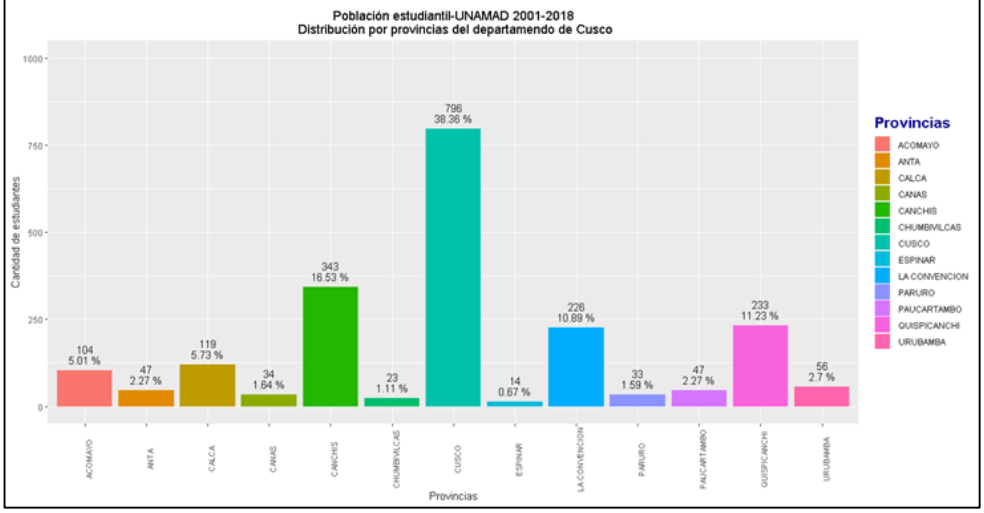


Figura 25. Distribución de estudiantes por provincias-Cusco (UNAMAD 2001-2018)

La figura 25, representa 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 frecuencias departamento de Puno

Script en el lenguaje R:

```
#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

Figura 26. Tabla de frecuencias: estudiantes por provincias-Puno (UNAMAD 2001-2018)

Diagrama de barras departamento de Puno:

Script en el lenguaje R:

```
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,datos$departamento=="PUNO")
provP
library (ggplot2)
qplot(data=provP, factor(provP$provincia), geom="bar",
  ylab="Cantidad de estudiantes",xlab="Provincias",
  fill=factor(provP$provincia))+
  theme(axis.text.x=element_text(size=8, angle=90))+
  scale_fill_discrete(name = "Provincias")+
  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, 300))+
  ggtitle("Población estudiantil-UNAMAD 2001-2018\n
  Distribución por provincias del departamendo de Puno") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4",
  size=16, face="bold"))
```

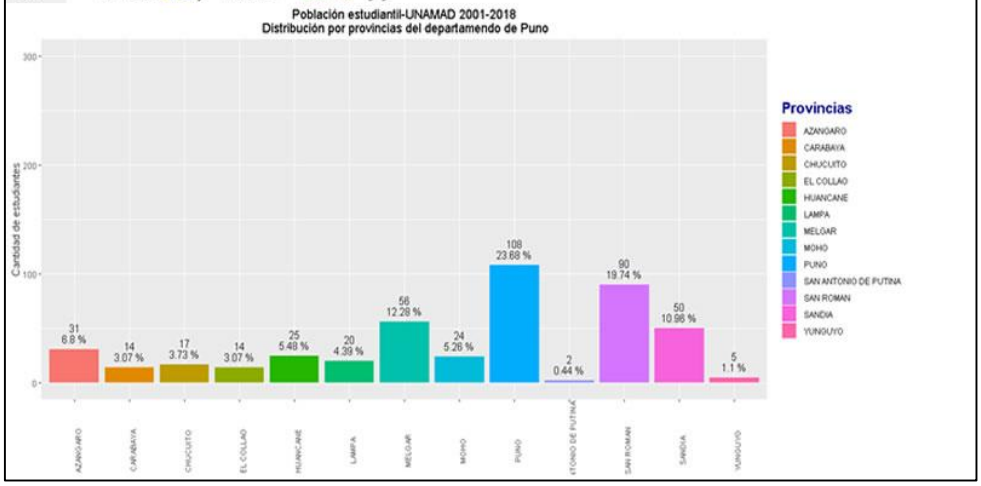


Figura 27. Distribución de estudiantes por provincias-Puno (UNAMAD 2001-2018)

La figura 27, representa la totalidad de estudiantes pertenecientes al departamento de Puno, de estos un 23.68% son originario de la provincia de Puno, seguido de un 19.74% que proceden de la provincia de San Román, otro 12.28% procedentes de la provincia de Melgar y un 10.96% de la provincia de Sandia.

Distribución de frecuencias de la variable sexo

Script en el lenguaje R:

```
#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
```

Figura 28. Tabla de frecuencias: estudiantes por género (UNAMAD 2001-2018)

Diagrama de barras para la variable sexo:

Script en el lenguaje R:

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
library (ggplot2)
qplot(data=datos, factor(datos$sexo), geom="bar",
  ylab="Cantidad de estudiantes",xlab="Género",
  fill=factor(datos$sexo))+
  theme(axis.text.x=element_text(size=12, angle=0))+
  scale_fill_discrete(name = "Género")+
  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, 7000))+
  ggtitle("Población estudiantil-UNAMAD 2001-2018\n Distribución por Género") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4", size=16, face="bold"))
```

Figura 29. Distribución de estudiantes por Género (UNAMAD 2001-2018)

En la figura 29, se observa 52.43% de la población estudiantil son de género masculino y un 47.55 % de sexo femenino.

Distribución de frecuencias de la variable carrera profesional

Script en el lenguaje R:

```
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 INFORMÀTICA	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

Figura 30. Tabla de frecuencias: estudiantes por carrera profesional (UNAMAD 2001-2018)

Diagrama de barras para la variable carrera profesional:

Script en el lenguaje R:

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
library (ggplot2)
qplot(data=datos, factor(datos$carrera), geom="bar",
  ylab="Cantidad de estudiantes",
  xlab="Carrera Profesional",fill=factor(datos$carrera))+
  theme(axis.text.x=element_text(size=8, angle=45))+
  scale_fill_discrete(name = "Carreras")+
  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, 2000))+
  ggtitle("Población estudiantil-UNAMAD por Carrera del 2001-2018") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4",
  size=16, face="bold"))
```

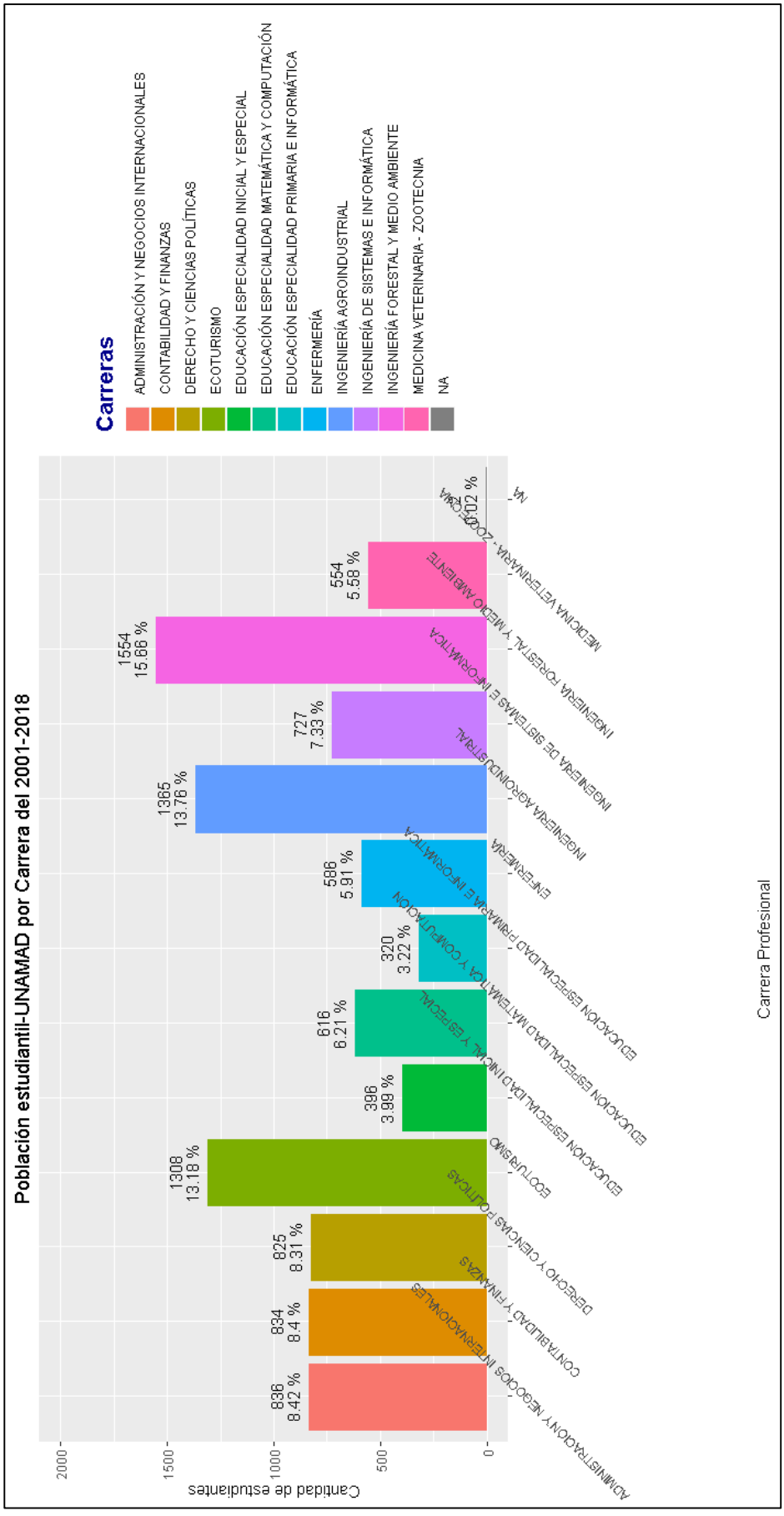


Figura 31. Población estudiantil-UNAMAD por carrera profesional del 2001-2018

En la figura 31, presenta la cantidad de matriculado desde año 2001 al 2018, donde se aprecia que el 15.66% de estudiantes matriculados pertenecen a la carrera de profesional de ingeniería forestal y medio ambientes, un 13.76% a la carrera profesional de ingeniería agroindustrial, un 13.18% a la carrera profesional de ecoturismo que vienen a ser las carreras con mayor antigüedad de creación en esta casa superior de estudios.

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

Script en el lenguaje R:

```
#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

Figura 32. Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito de Tambopata-Madre de Dios

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

Script en el lenguaje R:

```
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$escuela_ubigeo_distrito==1)
library (ggplot2)
qplot(data=dataDisMDD, factor(dataDisMDD$escuela), geom="bar",
  ylab="Cantidad de estudiantes",xlab="Institución educativa",
  fill=factor(dataDisMDD$escuela))+
  theme(axis.text.x=element_text(size=7, angle=90))+
  scale_fill_discrete(name = "Institución educativa")+
  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, 1000))+
  ggtitle("Procedencia de estudiantes UNAMAD 2001-2018\n por
  institución educativa del distrito de Tambopata-Madre de Dios") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4", size=12, face="bold"))+
  theme(legend.text = element_text(colour="black", size=8))
```

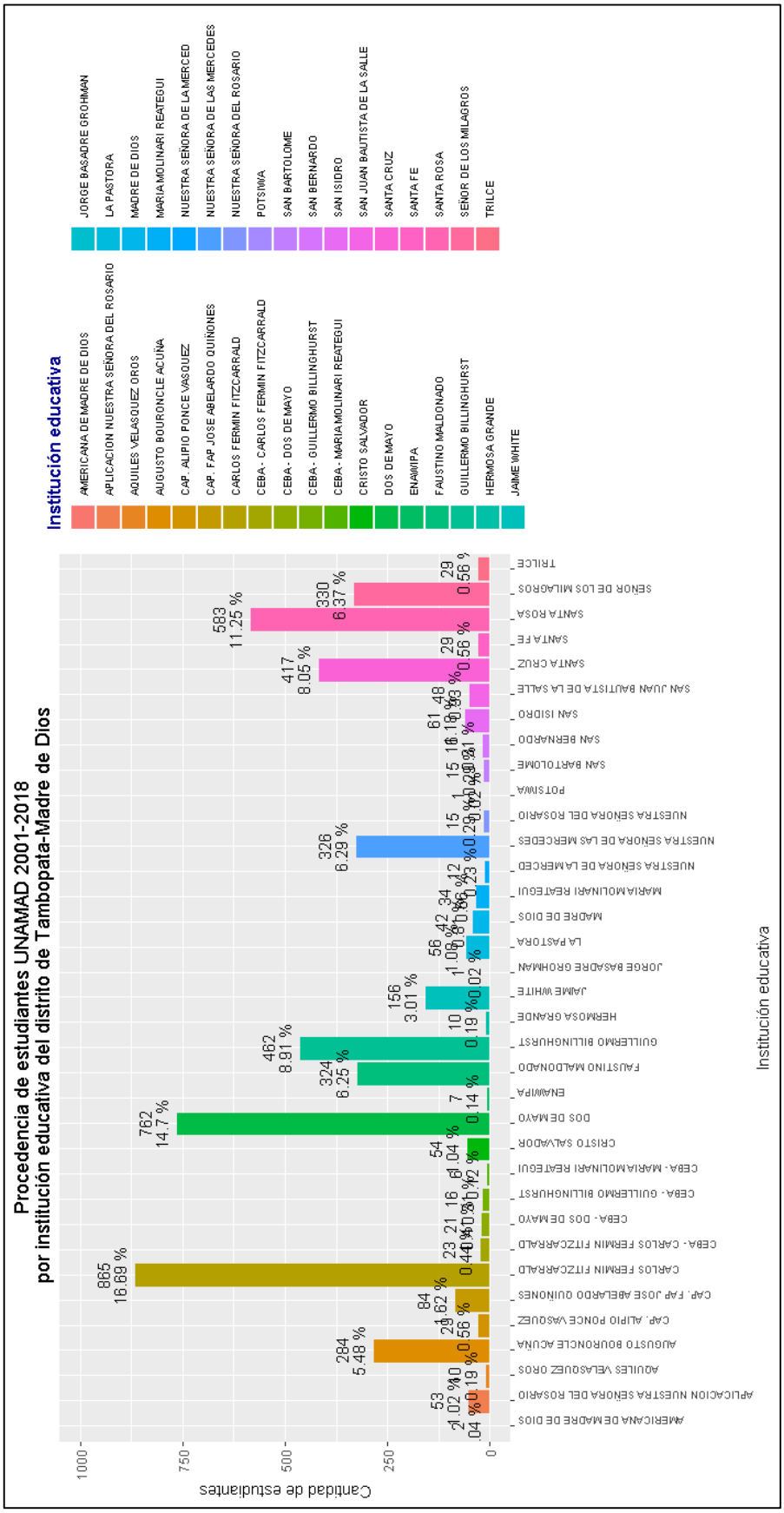


Figura 33. Distribución de estudiantes UNAMAD 2001-2018 por institución educativa del distrito de Tambopata-Madre de Dios

La figura 33, presenta la procedencia de los estudiantes por instituciones educativas del distrito de Tambopata, provincia de Tambopata, del departamento de Madre de Dios, en la UNAMAD desde el año 2001 al 2018, donde se aprecia que el 16.69% de estos proceden de la institución educativa Carlos Fermín Fitzcarrald, un 14.7% proceden de la institución educativa Dos de Mayo, un 11.25% proceden de la institución educativa Santa Rosa, otro 8.5% de la institución educativa Santa Cruz, el 6.29% son procedentes de la institución educativa Nuestra Señora de las Mercedes, el 6.25% proceden de la institución educativa Faustino Maldonado y un 5.48% procedentes de la institución educativa Faustino Maldonado, el 30% restante proceden de otras instituciones de la provincia Tambopata.

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

Script en el lenguaje R:

```
#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

Figura 34. Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Las Piedras-Madre de Dios

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

Script en el lenguaje R:

```
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$escuela_ubigeo_distrito==1)
library (ggplot2)
qplot(data=dataDisMDD, factor(dataDisMDD$escuela), geom="bar",
  ylab="Cantidad de estudiantes",xlab="Institución educativa",
  fill=factor(dataDisMDD$escuela))+
  theme(axis.text.x=element_text(size=7,angle=90,hjust=1))+
  scale_fill_discrete(name = "Institución educativa")+
  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, 100))+
  ggtitle("Procedencia de estudiantes UNAMAD 2001-2018\n por
  institución educativa del distrito Las Piedras-Madre de Dios") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4", size=12, face="bold"))+
  theme(legend.text = element_text(colour="black", size=7))
```

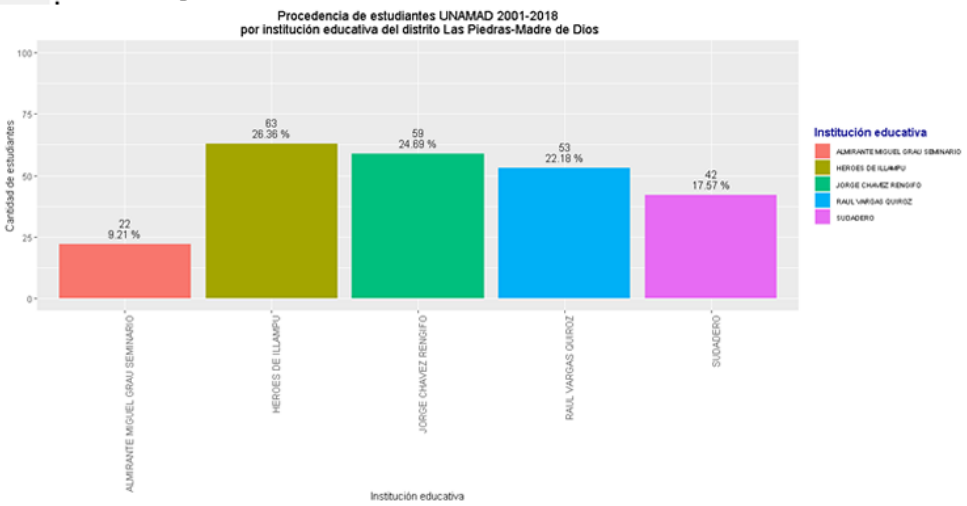


Figura 35. Distribución de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Las Piedras-Madre de Dios

La figura 35, presenta la procedencia de los estudiantes por instituciones educativas del distrito Las Piedras, provincia Tambopata, del departamento de Madre de Dios, en la UNAMAD desde el año 2001 al 2018, donde se aprecia que el 26.36% de estos proceden de la institución educativa Héroes de Illampu, un 24.69% proceden de la institución educativa Jorge Chávez Rengifo, un 22.18% proceden de la institución educativa Raúl Vargas Quiroz, otro 17.57% de la institución educativa Sudadero, y un 9.21% proceden 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:

```
#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
```

Figura 36. Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Laberinto-Madre de Dios

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

Script en el lenguaje R:

```
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$escuela_ubigeo_distrito==1)
library (ggplot2)
qplot(data=dataDisMDD, factor(dataDisMDD$escuela), geom="bar",
  ylab="Cantidad de estudiantes",xlab="Institución educativa",
  fill=factor(dataDisMDD$escuela))+
  theme(axis.text.x=element_text(size=10, angle=90,hjust=1))+
  scale_fill_discrete(name = "Institución educativa")+
  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, 200))+
  ggtitle("Procedencia de estudiantes UNAMAD 2001-2018\n
  por institución educativa del distrito Laberinto-Madre de Dios") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4", size=12, face="bold"))+
  theme(legend.text = element_text(colour="black", size=7))
```

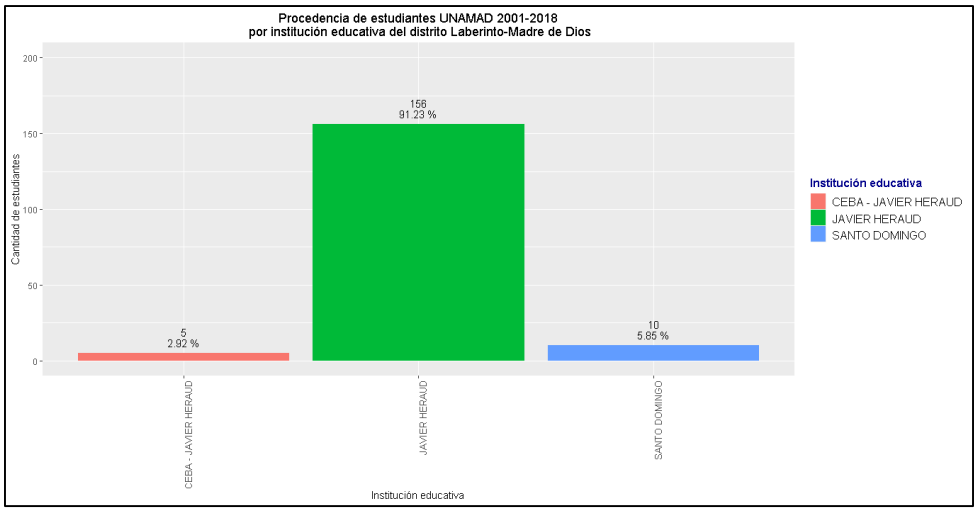


Figura 37. Distribución de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Laberinto-Madre de Dios**.**

La figura 37, presenta la procedencia de los estudiantes por instituciones 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% de estos proceden 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:

```
#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

Figura 38. Procedencia de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Inambari-Madre de Dios

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

Script en el lenguaje R:

```
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$escuela_ubigeo_distrito==1)
library (ggplot2)
qplot(data=dataDisMDD, factor(dataDisMDD$escuela), geom="bar",
  ylab="Cantidad de estudiantes",xlab="Institución educativa",
  fill=factor(dataDisMDD$escuela))+
  theme(axis.text.x=element_text(size=10, angle=90,hjust=1))+
  scale_fill_discrete(name = "Institución educativa")+
  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, 200))+
  ggtitle("Procedencia de estudiantes UNAMAD 2001-2018\n
  por institución educativa del distrito Inambari-Madre de Dios") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4", size=12, face="bold"))+
  theme(legend.text = element_text(colour="black", size=12))
```

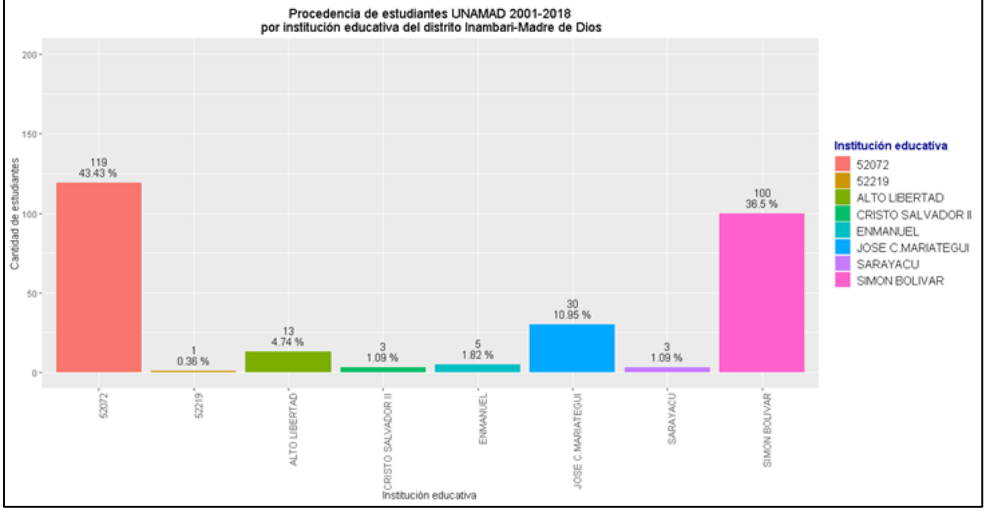


Figura 39. Distribución de estudiantes UNAMAD 2001-2018 por institución educativa del distrito Inambari-Madre de Dios

La figura 39, presenta la procedencia de los estudiantes por instituciones educativas del distrito Inambari, provincia Tambopata, del departamento de Madre de Dios, en la UNAMAD desde el año 2001 al 2018, donde se aprecia que el 43.43% de estos proceden de la institución educativa 52072, un 36.5% proceden de la institución educativa Simón Bolívar, un 10.95% proceden de la institución educativa José Carlos Mariátegui.

Ingresantes UNAMAD por semestre del 2001-2018

Script en el lenguaje R:

```
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_ingreso))
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
37	2051-1	1	9921	0.000	1.000	0.0	100.000

Figura 40. Ingresantes UNAMAD por semestre del 2001-2018

Diagrama de barras de ingresantes UNAMAD por semestre del 2001-2018

Script en R:

```
library("readxl")
setwd("F:/UNA-MAESTRIA/2018-tesis/Borrador")
datos<- read_excel("academico_unamad.xlsx",  sheet = 1)
library (ggplot2)
qplot(data=datos, factor(datos$semestre_ingreso), geom="bar",
  ylab="Cantidad de estudiantes",
  xlab="Semestre",fill=factor(datos$semestre_ingreso))+
  theme(axis.text.x=element_text(size=8, angle=90))+
  scale_fill_discrete(name = "Semestre")+
  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, 1000))+
  ggtitle("Población estudiantil-UNAMAD por semestre del 2001-2018") +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(legend.title = element_text(colour="blue4", size=16, face="bold"))
```

F

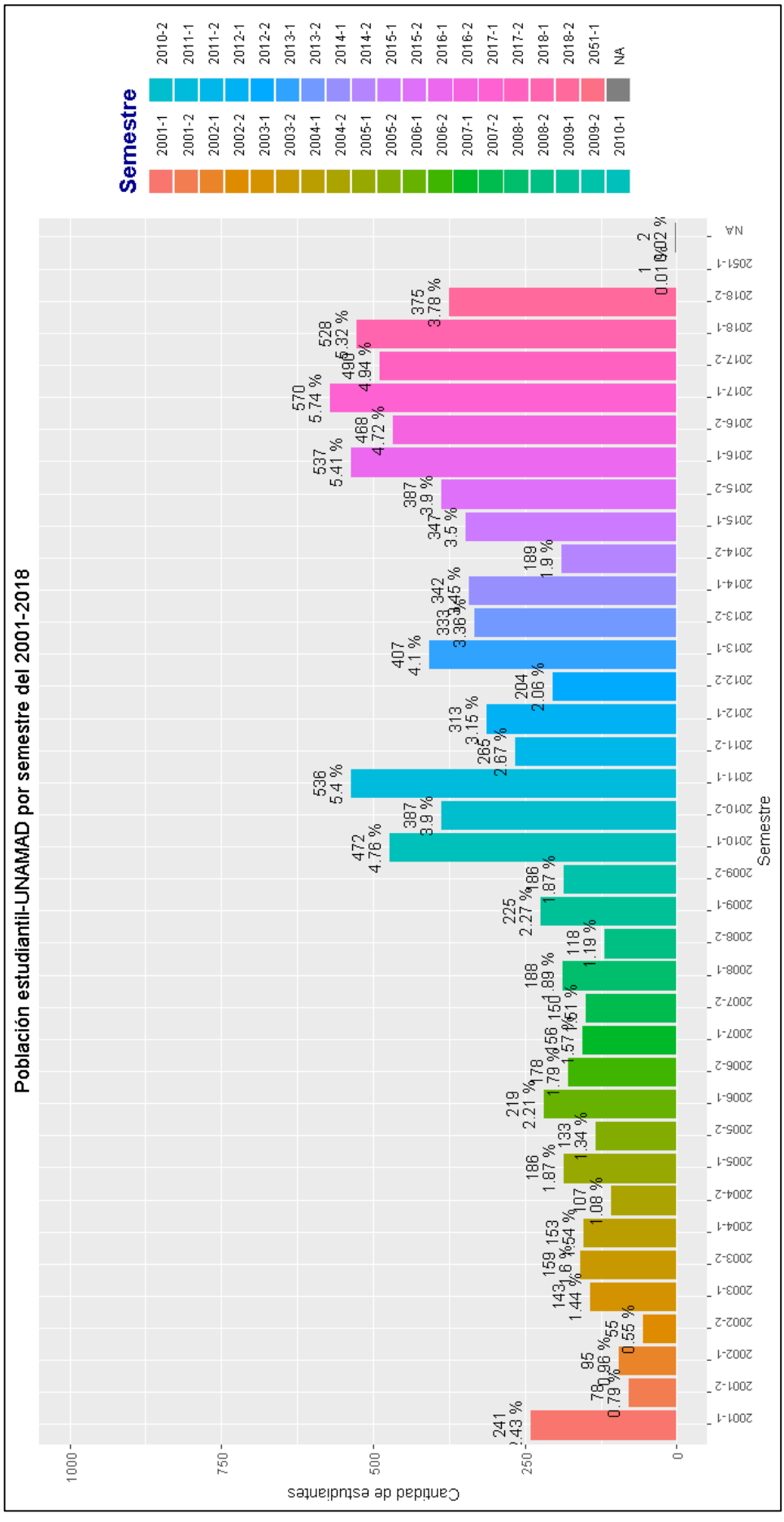


Figura 41. Ingresantes-UNAMAD por semestre del 2001-2018

La figura 41, presenta la cantidad de estudiantes ingresantes desde el semestre 2001-I al 2018-II, donde se aprecia que la mayor cantidad de estudiantes ingresantes se ubican en el semestre 2017-I con un total de 570 estudiantes, el semestre 2016-I con 537 estudiantes, el 2011-I con 536 estudiantes y el 2018-I con un total de 528 estudiantes.

d. Verificación de la calidad de los datos

De acuerdo con el análisis exploratorio de los datos, se observa que respecto a la variable departamento (figura 21), existe 271 registros con el valor SIN DATOS, respecto a la variable sexo (figura 29) existen 2 registros con el valor SIN DATOS, respecto a la variable carrera (figura 31) existen dos registros con el valor NA, respecto a la variable semestre_ingreso (Figura 41) existen 2 registros con el valor NA y un registro con el valor 2051-I. estas incoherencias serán corregidas en la siguiente fase.

4.1.3. 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 nuestra variable objetivo debe ser categórica, mientras que nuestras variables predictoras pueden ser continuas o categóricas, se utilizó las funciones filter(), select() y mutate() del paquete dplyr de R.

a) Selección de los 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

b) 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, a continuación, se detalla el script en el lenguaje R:

```
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",
                            TRUE ~ "others")
)
```

```r
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=="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$clase)
data_unamad1$t_escuela<-factor(data_unamad1$t_escuela)
data_unamad1$m_ingreso<-factor(data_unamad1$m_ingreso)
```

c) Estructuración de los datos

Con los resultados de la tarea anterior (limpieza de datos), con 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: Ingeniería Agroindustrial 10: Ingeniería Forestal y Medio Ambiente 11: Ingeniería 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.	SI NO
d_universidad	Indica si el estudiante adeuda a la universidad.	0: NO 1: SI
m_ingreso	Indica la modalidad de ingreso del estudiante.	1: Centro preuniversitario 2: CEPRE ordinario 3: Deportistas calificados 4: Examen especial para secundaria 5: Examen ordinario 6: Federación agraria departamental de madre de dios 7: Federación nativa de rio 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: victimas del terrorismo o plan de
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

La variable clase se generó de la transformación del atributo numérico promedio_ponderado_acumulado, tomando como referencia (MINEDU, 2009) donde define la escala de evaluación como se detalla en la siguiente tabla:

Tabla 13

Escala de evaluación de los aprendizajes

Variable	Clase	Valores	Descripción
promedio_ponderado_acumulado	C	0-10	Cuando el estudiante evidencia el logro de los aprendizajes previstos, demostrando 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 está empezando a desarrollar los aprendizajes previstos o evidencia dificultades para el desarrollo de éstos y necesita mayor tiempo de acompañamiento e intervención del docente de acuerdo con su ritmo y estilo de aprendizaje.

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

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

```
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")))
)
```

d) 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.

e) Formateo de los datos

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

```
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)
```

La vista minable quedó de la siguiente manera:

	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

Figura 42. Vista minable

Esta tabla quedó con 7309 instancias y 9 columnas.

4.1.4. Fase 4: Modelamiento

a) Selección de la técnica del 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 CART (Classification And Regression Trees) implementado en el paquete rpart,

C5.0 que es una extensión del algoritmo de árboles de decisión C4.5, implementado en el paquete C50 y finalmente Random Forest implementado en el paquete randomForest de RStudio.

b) Generación del plan de pruebas

Para esta tarea se realiza 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% respectivamente.

Script en el lenguaje R:

```
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
```

Figura 43. Resumen del conjunto de datos de entrenamiento

```
t_sexo   escuela_ubigeo_departamento   t_carrera    cant_cursos_cursados Servicio_comedor d_universidad
0:1063   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
```

Figura 44. Resumen del conjunto de datos de prueba

En la figura 43 y 44 se observan que los conjuntos de datos se encuentran no balanceados, dado que existen clases con número de instancias mayores que otras. De acuerdo (Espinar, 2018) “existen dos métodos de remuestreo: Downsampling y Upsampling. Upsampling es una técnica que simula o

atribuye datos adicionales para mejorar el equilibrio de las clases, mientras Downsampling es una técnica que reduce el tamaño de la muestra para mejorar el equilibrio de dichas clases, también puede darse la hibridación de ambas.”.

c) 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 utilizando el algoritmo Random Forest: *Experimento 1*

Identificación del número óptimo de predictores por cada partición

Script en el lenguaje R:

```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 del out-of-bag-error vs mtry",
  x = "nº predictores empleados", y="out-of-bag clasification error") +
  theme_bw()
```

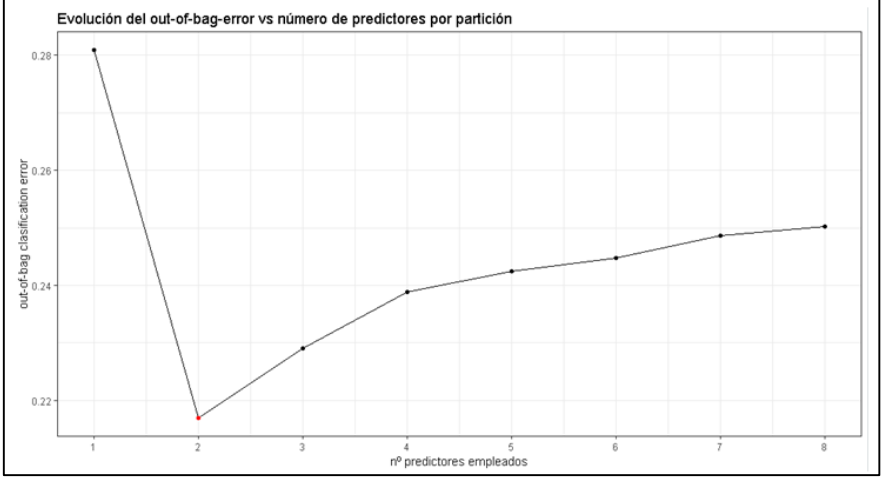


Figura 45. Evolución del out-of-bag-error versus número de predictores por partición.

En la figura 45 se observa la evolución del out-bag-error en función del número de predictores, se observa que el valor de este error es mínimo para 2 predictores por partición.

Identificación del tamaño óptimo de los nodos finales.

Script en el lenguaje 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()
```

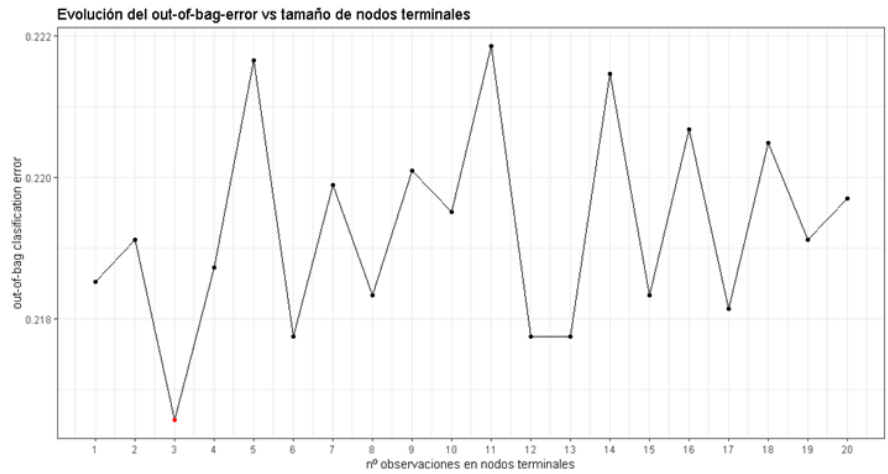


Figura 46. Evolución del out-of-bag-error versus tamaño de nodos

En la figura 46 se observa la evolución del out-bag-error en función del número de observaciones en los nodos terminales, se observa que el error se minimiza para 3 observaciones en los nodos terminales.

Identificación del número óptimo de árboles

Script en el lenguaje R:

```
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()
```

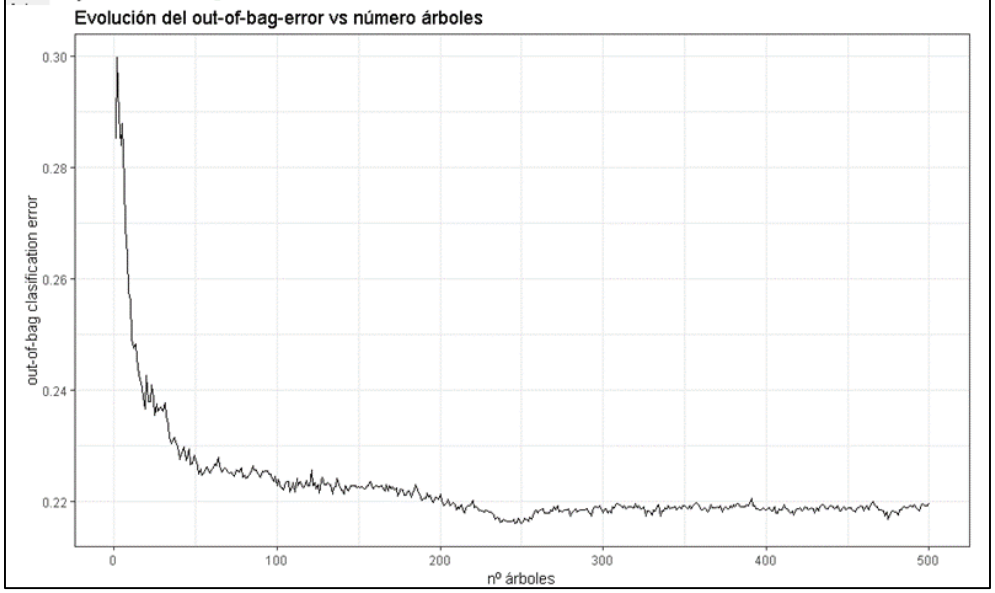


Figura 47. Evolución del out-of-bag-error versus número de arboles

En la figura 47 se observa que el out-of-bag-error del modelo se logra estabilizar, para 250 árboles, lo cual indica que el modelo óptimo de para nuestro 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:

```r
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 índice de Gini") +
  geom_col() +
  coord_flip() +
  theme_bw() +
  theme(legend.position = "bottom")
  ggarrange(figura_rf1, figura_rf2)
```

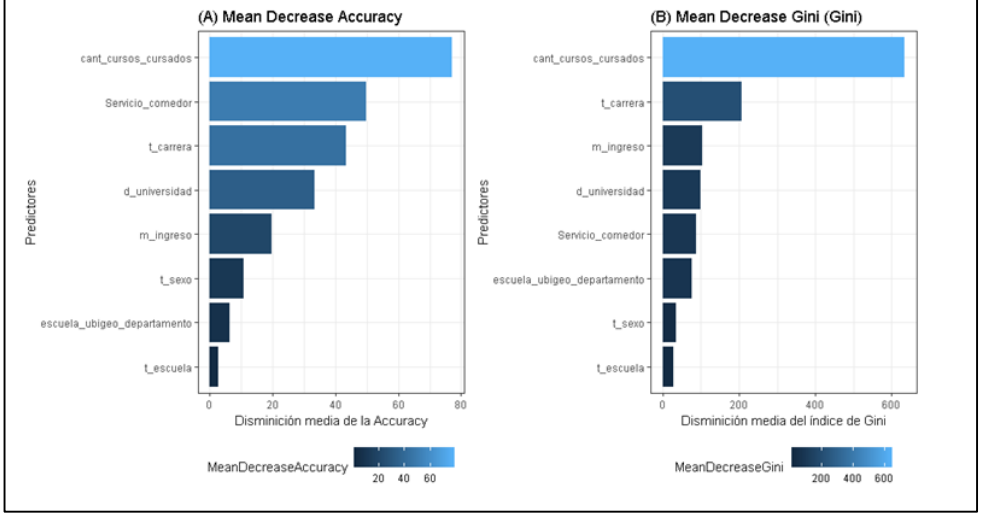


Figura 48. Influencia de las variables en el modelo de clasificación-Random Forest.

En la figura 48 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 (Servicio_comedor), la carrera profesional (t_carrera), deuda con la universidad (d_universidad) son las variables que más influyen en la prediccion del rendimiento académico, la gráfica de barras muestra cuánto disminuye la precisión del modelo si dejamos de lado esas 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 que reducen más el índice de impureza de Gini. (Sarría, 2016) menciona que: 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.

Discusión

Estos resultados guardan relación con lo que sostienen La Red Martinez *et al*. (2015) en su estudio denominado: Perfiles de rendimiento académico: Un modelo basado en minería de datos, quienes señalan que el tipo de escuela que cursó el alumno no está relacionado con el rendimiento académico logrado por el mismo. Ello es acorde a lo que en este estudio se halla.

Matriz de confusión y estadísticas

Script en el lenguaje R:

```
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
```

Figura 49. Matriz de confusión del modelo construido con el algoritmo Random Forest

En la figura 49 se observa que la exactitud (Accuracy) del modelo de clasificación es 77.5%, de donde se desprende que la tasa de error de clasificación el del 22.5%, por otro parte el coeficiente de kappa es 0.55, lo que indica de acuerdo a la tabla de valoración del coeficiente de kappa propuesta por (Landis & Koch, 1997) que la clasificación observada concuerda moderadamente con la clasificación predecida por el clasificador.

Árbol de clasificación utilizando 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:

```
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)
```

Figura 50. Árbol de clasificación para el rendimiento académico - C5.0

En la figura 50 se observa el árbol de clasificación generado por el algoritmo C5.0, en la hoja 3 de arriba hacia abajo, se aprecia que los estudiantes clasificados en la categoría C ascienden a 1275, estos tienen el

siguiente perfil: *estudiantes que no poseen servicio de comedor universitario, que cursaron más de 6 cursos, pero menos de 62, que poseen deuda con la universidad*, en la hoja 4 se aprecia que fueron clasificados 901 estudiantes en la categoría C, estos estudiantes además de tener el mismo perfil anterior *pertenecen a las carreras de Ingeniería Forestal y Medio Ambiente, Ingeniería de Sistemas e Informática, Medicina Veterinaria y Zootecnia, Ecoturismo y también a la carrera de Ingeniería Agroindustrial.*

Matriz de confusión y estadísticas

```
1 prediccion_1 <- predict(modelo, newdata = datos.test, type = "class")
2 summary(prediccion_1)
3 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
```

Figura 51. Matriz de confusión del modelo construido con el algoritmo C5.0

En la figura 51 se observa que la exactitud (Accuracy) del modelo de clasificación es 77.8%, siendo la tasa de error de clasificación el del 22.2%, por otro parte el coeficiente de kappa es 0.56, lo que indica de acuerdo a la tabla de valoración del coeficiente de kappa propuesta por (Landis & Koch,

1997) que la clasificación observada concuerda moderadamente con la clasificación predicha por el clasificador.

Identificación de las variables más influyentes en el modelo predictivo utilizando el algoritmo C5.0: *Experimento 3*

Script en el lenguaje R:

```
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)
```

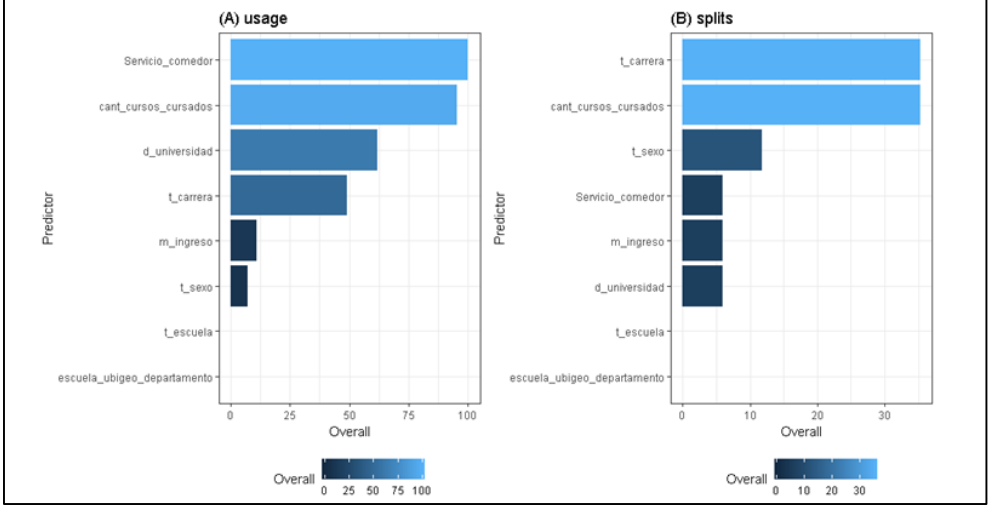


Figura 52. Influencia de las variables en el modelo predictivo de clasificación-C5.0

En la figura 52 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 (cant_cursos_cursados) y género (t_sexo) son las variables que más participan en las divisiones del árbol de clasificación.

Árbol de clasificación utilizando el algoritmo CART: *Experimento 4*

Script en R para entrenar el modelo:

```
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 (0.005668491 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 1206 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) *
```

Figura 53. Reglas obtenidas por el algoritmo CART

La figura 53 muestra el esquema del árbol de clasificación. Cada inciso nos indica un nodo y la regla de clasificación que le corresponde. Siguiendo estos nodos, podemos llegar a las hojas del árbol, que corresponde a la clasificación de nuestros datos.

A continuación, presentamos el árbol de clasificación de manera gráfica:

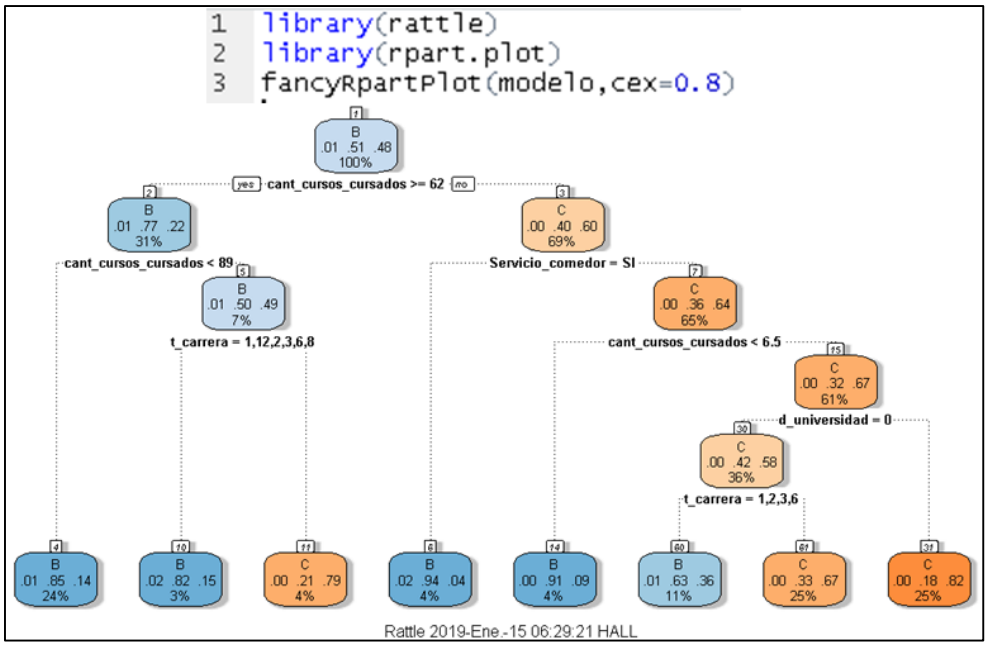


Figura 54. Árbol de clasificación para el rendimiento académico - CART

La figura 54 representa gráficamente el modelo de árbol de clasificación, en ello 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 representan 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 podemos 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% respectivamente, estos estudiantes tienen una calificación de 0 a 10, resumiendo la hoja 8 podemos afirmar 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, además de la hoja 7 podemos afirmar que este grupo de estudiantes no poseen deuda con la universidad y no pertenecen a las carreras de: Administración y 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 las Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres, Lima-Perú, quien señala que se realizaron predicciones para el rendimiento académico y se obtuvieron resultados de 82.87% utilizando árbol de decisiones. En el presente estudio 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: nota de examen de admisión, género, edad, modalidad de ingreso y distancia desde su casa hasta el centro de estudios". En este estudio los factores género, modalidad de ingreso no figuran el árbol de clasificación, perdiendo así influencia significativa en el modelo predictivo de clasificación.

Matriz de confusión y estadísticas

Script en el lenguaje R:

```
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
```

Figura 55. Matriz de confusión del modelo construido con el algoritmo CART

En la figura 55 se observa que la exactitud (Accuracy) del modelo de clasificación es 76.9%, siendo la tasa de error de clasificación el del 21.1%, por otro parte el coeficiente de kappa es 0.54, lo que indica de acuerdo a la tabla de valoración del coeficiente de kappa propuesta por (Landis & Koch, 1997) que la clasificación observada concuerda moderadamente con la clasificación predicha por el clasificador.

4.1.5. Fase 5: Evaluación

Evaluación de los resultados

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	

4.1.6. 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.

CONCLUSIONES

- Tras la aplicación de Minería de datos mediante la metodología CRISP-DM, con el uso del 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 48), seguidamente 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 asimismo 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 54): “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.

RECOMENDACIONES

- Se recomienda considerar más variables predictoras para determinar su grado de influencia en los modelos de clasificación para el rendimiento académico tomando 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, aplicando otras técnicas predictivas de minería de datos como la regresión logística binaria y máquinas de soporte vectorial, utilizando 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.

BIBLIOGRAFÍA

Alderete, A. M. (2006). Fundamentos del Análisis de Regresión Logística en la Investigación Psicológica . *Revista evaluar, 6*(1), 52-67. Recuperado el 2018 de Octubre de 2018, de https://revistas.unc.edu.ar/index.php/revaluar/article/view/534/474

Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial. *Qüestiió: quaderns d'estadística i investigació operativa, 25*(3), 479-498. Recuperado el 30 de Agosto de 2018, de https://www.raco.cat/index.php/Questiio/article/download/27009/26843

Álvarez, E., & Cuji, B. (2016). Las Técnicas de Predicción y su incidencia en la detección de patrones de Deserción Estudiantil en la Carrera de Docencia en Informática de la Facultad de Ciencias Humanas y de la Educación de la Universidad Técnica de Ambato. *(Tesis de maestría).* Universidad Técnica de Ambato, Ambato, Ecuador. Recuperado el 05 de Octubre de 2018, de http://repositorio.uta.edu.ec/bitstream/123456789/23839/1/Tesis_t1164mbd.pdf

Álvarez, R. (1995). *Estadística multivariante y no paramétrica con SPSS.* Madrid: Ediciones Díaz de Santos.

Amat, J. (Febrero de 2017). *https://rpubs.com/*. Recuperado el 16 de Enero de 2019, de https://rstudio-pubs-static.s3.amazonaws.com/255596_79f26327969342c3b4a15f38b5b9f8f3.html#ejemplo_clasificaci%C3%B3n34

Arias, F. (2006). *El proyecto de investigación.* Caracas: EPISTEME.

Ato, M., Lopez, J. A., Velandrino, A., & Sanchez, J. (1990). *Estadística Avanzada con el paquete Systat.* Murcia: Universidad de Murcia.

Bacallao, J., Parapar, J. M., Roque, M., & Bacallao, J. (2004). Árboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico. *Educación Médica Superior, 18*(3), 1. Recuperado el 30 de Setiembre de 2018, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21412004000300002

Barrientos, R., Cruz, N., Acosta, H., Rabatte, I., & Gogeascoechea, M. d. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista Médica de la Universidad Veracruzana, 9*(2), 19-24. Recuperado el 07 de Octubre de

2018, de http://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=27872

Basogain, X. (s.f.). *Redes neuronales artificiales y sus aplicaciones.* Recuperado el 06 de Octubre de 2018, de https://ocw.ehu.eus/file.php/102/redes_neuro/contenidos/pdf/libro-del-curso.pdf

Benítez, I. J. (Setiembre de 2005). *Researchgate.* Recuperado el 31 de Agosto de 2018, de 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

Berlanga, V., Rubio Hurtado, M. J., & Vilà Baños, R. (2013). Cómo aplicar árboles de decisión en SPSS. *Revista d'Innovació i Recerca en Educació, 6*(1), 65-79. Recuperado el 30 de Setiembre de 2018, de http://diposit.ub.edu/dspace/bitstream/2445/43762/1/618361.pdf

Britos, P. (2008). Procesos de explotación de información basados en sistemas inteligentes. *(Tesis para obtar el grado de doctor en ciencias informáticas).* UNIVERSIDAD NACIONAL DE LA PLATA FACULTAD DE INFORMÁTICA, Buenos Aires. Recuperado el 02 de Setiembre de 2018, de http://sedici.unlp.edu.ar/bitstream/handle/10915/4142/Documento_completo.pdf?sequence=1&isAllowed=y

Camana, R. (2012). Aplicación de Técnicas de Minería de Datos para la Indagación y Estudio de Resultados Electorales. *CienciAmérica: Revista de divulgación científica de la Universidad Tecnológica Indoamérica, 1*(1), 85-94. Recuperado el 26 de Agosto de 2018, de https://dialnet.unirioja.es/servlet/articulo?codigo=6163757

Cerda, J., & Villarroel, L. (2008). Evaluación de concordancia inter-observador en investigación pediátrica: coeficiente de Kappa. *Revista chilena de pediatría, 79*(1), 54-58. Recuperado el 17 de Enero de 2019, de https://scielo.conicyt.cl/pdf/rcp/v79n1/art08.pdf

Chapman, P. C. (2000). CRISP-DM 1.0 Step-by-step data mining guide. Recuperado el 02 de Setiembre de 2018, de http://www.crisp-dm.org/CRISPWP-0800.pdf

Coyla, E. (2017). Análisis de datos con Bigdata en procesos de admisión de la Universidad Nacional del Altiplano de Puno, 2016. *(Tesis de doctorado).* Universidad Nacional del Altiplano de Puno, Puno, Perú. Recuperado el 05 de Octubre de 2018, de http://repositorio.unap.edu.pe/handle/UNAP/6211

Dapozo, G. N., Porcel, E., López, M. V., Bogado, V. S., & Bargiela, R. (Junio de 2006). 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*. Recuperado el 30 de Setiembre de 2018, de http://sedici.unlp.edu.ar/handle/10915/20797

Diaz, Z. (2007). *Predicción de crisis empresariales en seguros no vida, mediante árboles de decisión y reglas de clasificación.* Madrid, España: Complutense. Recuperado el 07 de Octubre de 2018

Digital Guide. (30 de Enero de 2018). *IONOS*. Recuperado el 02 de Enero de 2019, de https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas/

Espinar, R. (2018). Modelos de clasificación con datos no balanceados. *Trabajo de fin de grado.* Universidad de Sevilla, Sevilla. Recuperado el 14 de Enero de 2019, de https://idus.us.es/xmlui/bitstream/handle/11441/77518/Espinar%20Lara%20Roc%C3%ADo%20TFG.pdf?sequence=1

Fayyad, U., Piatetsky-Shapiro, G., & Padhraic, S. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *ASOCIACIÓN PARA EL AVANCE DE LA INTELIGENCIA ARTIFICIAL*, 83-84.

Flores, C. (2014). Exigencias de calidad de suministro en base a densidad de consumo mediante técnicas de minería de datos. *(Memoria para optar al título de ingeniero civil electricista).* Universidad de Chile, Santiago de Chile, Chile. Recuperado el 27 de Agosto de 2018, de http://repositorio.uchile.cl/bitstream/handle/2250/115571/cf-flores_cc.pdf?sequence=1

Flóres, R., & Fernández, J. (2008). *Las redes neuronales artificiales.* La Coruña: NETBIBLO. Recuperado el 07 de Octubre de 2018

Formia, S., Lanzarini, L. C., & Hasperué, W. (2013). Caracterización de la deserción universitaria en la UNRN utilizando Minería de Datos. *TE & ET*(11), 92-98. Recuperado el 30 de Setiembre de 2018, de http://sedici.unlp.edu.ar/handle/10915/32401

Gallardo, J. A. (2009). Metodología para la definición de requisitos en proyectos de data mining. *Tesis(Doctoral).* Universidad Politécnica de Madrid, Madrid. Recuperado el 01 de Enero de 2019, de http://oa.upm.es/1946/1/JOSE_ALBERTO_GALLARDO_ARANCIBIA.pdf

Garbanzo, G. M. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación, 31*(1), 43-63. Recuperado el 2018 de Setiembre de 2018, de http://www.redalyc.org/pdf/440/44031103.pdf

García, C., & Gómez, I. (2012). Algoritmos de Aprendizaje: KNN & KMEANS. *Inteligencia en Redes de Telecomuncicación*, 6-7. Recuperado el 26 de Agosto de 2018, de http://blogs.ujaen.es/barranco/wp-content/uploads/2012/02/Algoritmos-de-aprendizaje-knn-y-kmeans.pdf

Garre, M., Cuadrado, J. J., Sicilia, M., Rodríguez, D., & Rejas, R. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. *Revista Española de Innovación, Calidad e Ingeniería del Software,*

3(1), 6-22. Recuperado el 28 de Agosto de 2018, de http://www.redalyc.org/html/922/92230103/

Gatica, F., Méndez, I., Sánchez, M., & Martínez, A. (2010). Variables asociadas al éxito académico en estudiantes de la Licenciatura en Medicina de la UNAM. *Revista de la Facultad de Medicina de la UNAM, 53*(5), 9-18. Recuperado el 30 de Setiembre de 2018, de http://www.pve.unam.mx/informacion/medicina/REVFACMEDSEPTIEMBRE.pdf#page=9

Gil, G. (2009). *DATA MINING.* Lima: Megabyte.

Gil, J. (2005). Aplicación del métodos de bootstrap al contraste de hipotesis en la nvestigación educativa. *Revista de Educación*, 251-265. Recuperado el 19 de Enero de 2019, de https://idus.us.es/xmlui/handle/11441/77873

Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). *Analisis Multivariante.* Madrid: Prentice Hall.

Heredia, D., Amaya, Y., & & Barrientos, E. (2015). Student Dropout Predictive Model Using Data Mining Techniques. *IEEE Latin America Transactions, 13*(9), 3127-3134. Recuperado el 30 de Setiembre de 2018, de http://www.revistaieeela.pea.usp.br/issues/vol13issue9Sept.2015/13TLA9_45Heredia.pdf

Hernández, C. L., & Dueñas, M. X. (2009). Hacia una metodología de gestión del conocimiento basada en minería de datos. *COMTEL*, 80-96. Recuperado el 02 de Setiembre de 2018, de http://repositorio.uigv.edu.pe/bitstream/handle/20.500.11818/982/COMTEL-2009-80-96.pdf?sequence=1

Hilbert, M., & López, P. (01 de Abril de 2011). La capacidad tecnológica mundial para almacenar, comunicar y calcular información. *Science, 332*(6025), 60-65. doi:10.1126 / science.1200970

Jimenez, A. C. (2017). Análisis predictivo para los procesos de admisión de la Universidad Nacional del Altiplano - Puno. *(Tesis de doctorado).* Universidad Nacional del Altiplano de Puno, Puno, Perú. Recuperado el 05 de Octubre de 2018, de http://repositorio.unap.edu.pe/handle/UNAP/6212

Joaquín, R. (Febrero de 2017). *rpubs*. Recuperado el 20 de Enero de 2019, de https://rpubs.com/Joaquin_AR/255596

Kovalevski, L., & Macat, P. (2012). Alternativas no paramétricas de clasisicación multivariada. *Decimoséptimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística*. Recuperado el 20 de Enero de 2019, de http://biblioteca.puntoedu.edu.ar/bitstream/handle/2133/7467/Kovalevski_Macat_alternativas%20no%20parametricas.pdf?sequence=3&isAllowed=y

Krikorian, M., Ruedin, A., & Seijas, L. (2011). Reconocimiento de patrones utilizando transformadas wavelet sin submuestreo y máquinas de soporte vectorial. *XIV Reunión de Trabajo Procesamiento de la Información y Control*, 839-844.

La Red Martínez, D. L., Karanik, M., Giovannini, M., & Pinto, N. (2015). Perfiles de Rendimiento Académico: Un Modelo Basado en Minería de Datos. *Campus Virtuales, 4*(1), 12-30. Recuperado el 30 de Setiembre de 2018, de http://www.uajournals.com/ojs/index.php/campusvirtuales/article/view/66

La Red Martinez, D., Acosta, J. C., Cutro, L., Uribe, V., & Rambo, A. (2010). Data warehouse y data mining aplicados al estudio del rendimiento académico y de perfiles de alumnos. *In XII Workshop de Investigadores en Ciencias de la Computación*. Recuperado el 30 de Setiembre de Setiembre, de http://sedici.unlp.edu.ar/handle/10915/19461

Marín, J. M. (08 de Abril de 2014). *Universidad Carlos III de Madrid.* Recuperado el 31 de Agosto de 2018, de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/GuiaSPSS/22conglj.pdf

Medina, R., & Ñique, C. (2017). Bosques aleatorios como extensión de arboles de clasificación con los programas R y Python. *Portal de reivistas Ulima*, 165-189. Recuperado el 19 de Enero de 2019, de http://revistas.ulima.edu.pe/index.php/Interfases/article/view/1775/1828

Microsoft. (30 de Abril de 2018). *Microsoft*. Obtenido de https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017

MINEDU. (2009). *Diseño Curricular Nacional de Educación Básica Regular*. LIMA, Perú: Santillana. Recuperado el 12 de Enero de 2019

Moine, J. M., Gordillo, S., & Haedo, A. S. (18 de Julio de 2012). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. *SEDICI*. Recuperado el 02 de Setiembre de 2018, de http://hdl.handle.net/10915/18749

Moine, J. M., Haedo, A. S., & Gordillo, S. (2011). Estudio comparativo de metodologías para minería de datos. *SEDICI*. Recuperado el 02 de Setiembre de 2018, de http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1&isAllowed=y

Mondragon, R. (2007). EXPLORACIONES SOBRE EL SOPORTE MULTI-AGENTE BDI EN EL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS. *(Tesis para Optar el Grado de Magister).* UNIVERSIDAD VERACRUZANA, Mexico.

Montt, C., Castro, F., & Rodríguez, N. (2011). Análisis de Accidentes de Tránsito con Máquinas de Soporte Vectorial LS-SVM. *Revista de ingenieria de transporte, 15*(2), 7-14. Recuperado el 07 de Octubre de 2018, de http://ingenieriadetransporte.org/ojs/index.php/sochitran/article/view/119/22

Moreno, M., Miguel, L., García, F., & Polo, M. J. (2001). 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.

Researchgate. Recuperado el 29 de Agosto de 2018, de 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, D. (2010). Algoritmos de Agrupamiento basados en densidad y Validación de clusters. *(Tesis Doctoral).* Universitat Jaume I, Castellón de la Plana, España. Recuperado el 31 de Agosto de 2018, de http://www.cerpamid.co.cu/sitio/files/DamarisTesis.pdf

Peña, D. (2012). *Analisis de Datos Multivariantes.* Madrid: University Carlos III de Madrid. Recuperado el 27 de Agosto de 2018, de https://www.researchgate.net/profile/Daniel_Pena4/publication/40944325_Analisis_de_Datos_Multivariantes/links/549154880cf214269f27ffae/Analisis-de-Datos-Multivariantes.pdf

Pereira, R. T. (2009). Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos. *En Memorias de la 8ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI*. Recuperado el 30 de Setiembre de 2018, de http://www.iiis.org/cds2008/cd2009cSc/CISCI2009/PapersPdf/C692YV.pdf

Pérez, C. (2004). *Técnicas de Análisis Multivariante de Datos.* Madrid: Pearson Prentice Hall.

Perez, C., & Santin, D. (2007). *Minería de datos. Técnicas y herramientas.* Madrid: Thomson.

Reyes, Y. (2003). Relación entre el rendimiento académico, la ansiedad ante los exámenes, los rasgos de personalidad, el autoconcepto y la asertividad en estudiantes del primer año de psicología de la UNMSM. Recuperado el 08 de Octubre de 2018, de http://sisbib.unmsm.edu.pe/bibvirtual/tesis/salud/reyes_t_y/cap2.htm

Rodriguez, O. (s.f.). *oldemarrodriguez.com.* Recuperado el 31 de Diciembre de 2018, de http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037

Rosado, A. A., & Verjel, A. (2014). Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander. *Revista Tecnura, 19*(45), 101-113. doi:http://dx.doi.org/10.14483/udistrital.jour.tecnura.2015.3.a08

RStudio. (2018). *rstudio*. Recuperado el 18 de Enero de 2019, de https://www.rstudio.com/resources/webinars/whats-new-with-readxl/

Salinas, J. W. (2016). 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*, 115-122. Recuperado el 05 de Octubre de 2018, de http://funes.uniandes.edu.co/9282/1/Salinas2016Deteccion.pdf

Sarría, F. G. (2016). Modelos predictivos para el estudio del abandono agrícola. *researchgate*, 161-180. Recuperado el 16 de Enero de 2019, de https://www.researchgate.net/publication/311589338_Modelos_predictivos_para_el_estudio_del_abandono_agricola

Spositto, O., Etcheverry, M., Ryckeboer, H., & Bossero, J. (2010). Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. *Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática, CISC. I International Institute of Informatics and Systemics, Orlando (Florida, EE. UU.)*. Recuperado el 30 de Setiembre de 2018, de http://www.iiis.org/cds2010/cd2010csc/cisci_2010/paperspdf/ca156fk.pdf

UNAMAD. (Abril de 2016). *unamad.edu.pe.* Recuperado el 01 de Enero de 2019, de http://www.unamad.edu.pe/index.php/descargas/send/24-institucionales/5468-pei-unamad-2017-2019

UNAMAD. (29 de Setiembre de 2018). *UNAMAD.* Obtenido de http://www.unamad.edu.pe/index.php/universidad/mision-y-vision

Valcárcel, V. (2004). Datamining y eldescubrimiento del conocimiento. *Revista de la Facultad de Ingeniería Industrial, 7*(2), 8386. Recuperado el 30 de Agosto de 2018, de http://www.redalyc.org/html/816/81670213/

Vallejo, D., & Tenalanda, G. (2012). Minería de datos aplicada en la detección de intrusos. *Ingenierías USBMed, 3*(1). Recuperado el 20 de Enero de 2019, de https://revistas.usb.edu.co/index.php/IngUSBmed/article/view/264/178

Villa, A., Carrión, A., & Sozzi, A. (2017). Optimización del diseño de parámetros:Método Forest-Genetic univariante. *Publicaciones en ciencias y tecnologías, 10*(1), 12-24. Recuperado el 19 de Enero de 2019, de https://riunet.upv.es/bitstream/handle/10251/103728/2016_PCYT%202016-1-Art2-Forest%20Genetic.pdf?sequence=1&isAllowed=y

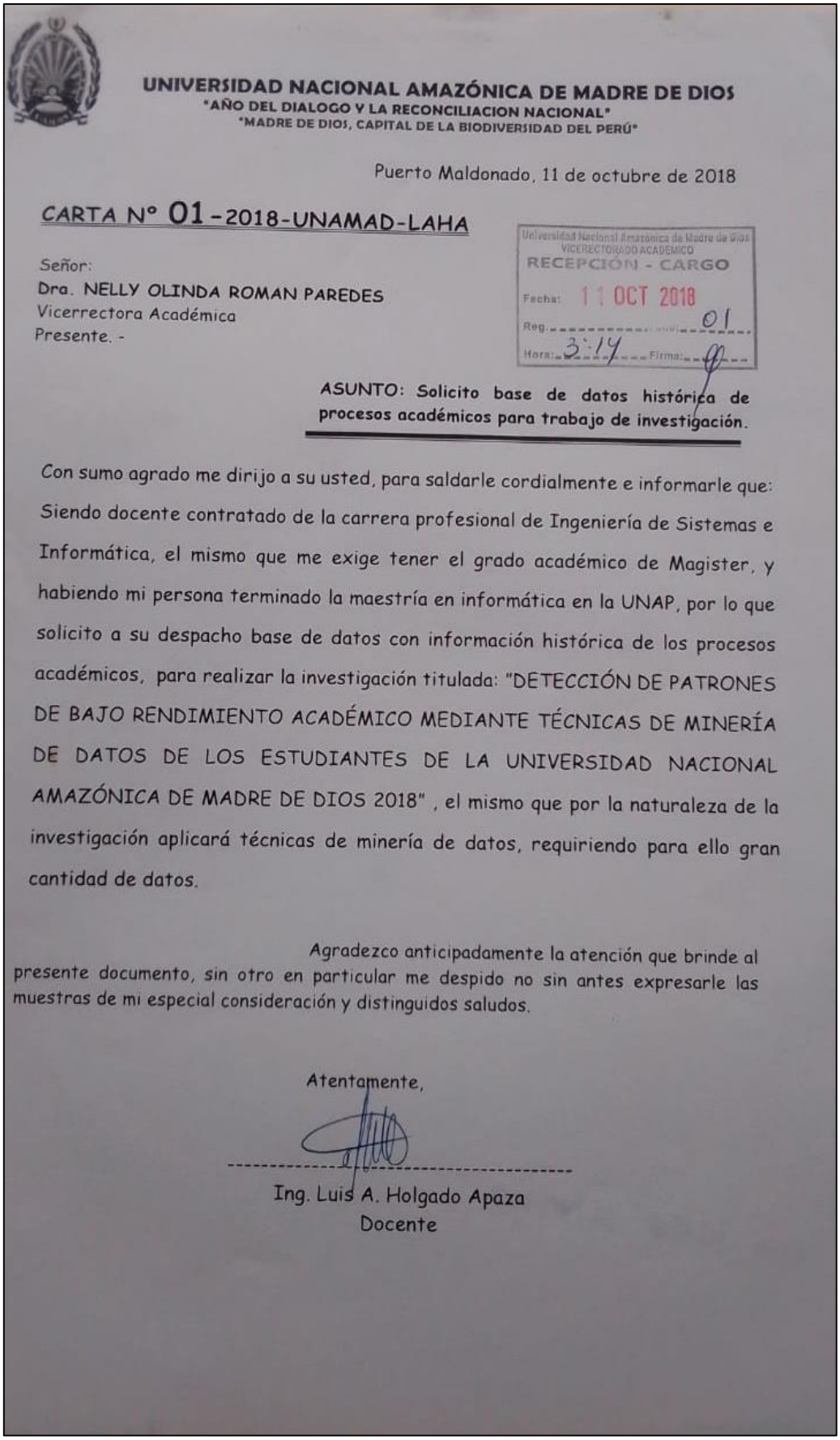
Villada, F., Cadavid, D. R., & Molina, J. D. (2014). Pronóstico del precio de la energía eléctrica usando redes neuronales artificiales. *Revista facultad de ingeniería*(44), 111-118. Recuperado el 07 de Octubre de 2018, de http://aprendeenlinea.udea.edu.co/revistas/index.php/ingenieria/article/view/18508/15905

Webmining Consultores. (10 de Enero de 2011). *WEBMINING*. Obtenido de http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/

Yamao, E. (2018). Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de las Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres, Lima-Perú. *(Tesis para optar el grado de maestro).* Universidad de San Martin de Porres, Lima, Perú. Recuperado el 17 de Enero de 2019, de http://www.repositorioacademico.usmp.edu.pe/handle/usmp/3555

ANEXOS

Anexo 1. Carta de solicitud de base de datos histórica de procesos académicos



UNIVERSIDAD NACIONAL AMAZÓNICA DE MADRE DE DIOS
"AÑO DEL DIALOGO Y LA RECONCILIACION NACIONAL"
"MADRE DE DIOS, CAPITAL DE LA BIODIVERSIDAD DEL PERÚ"

Puerto Maldonado, 11 de octubre de 2018

CARTA N° 01 -2018-UNAMAD-LAHA

Universidad Nacional Amazónica de Madre de Dios
VICERECTORADO ACADEMICO
RECEPCIÓN - CARGO
Fecha: 11 OCT 2018
Reg. 01
Hora: 3:14 Firma:

Señor:
Dra. NELLY OLINDA ROMAN PAREDES
Vicerrectora Académica
Presente. -

ASUNTO: Solicito base de datos histórica de procesos académicos para trabajo de investigación.

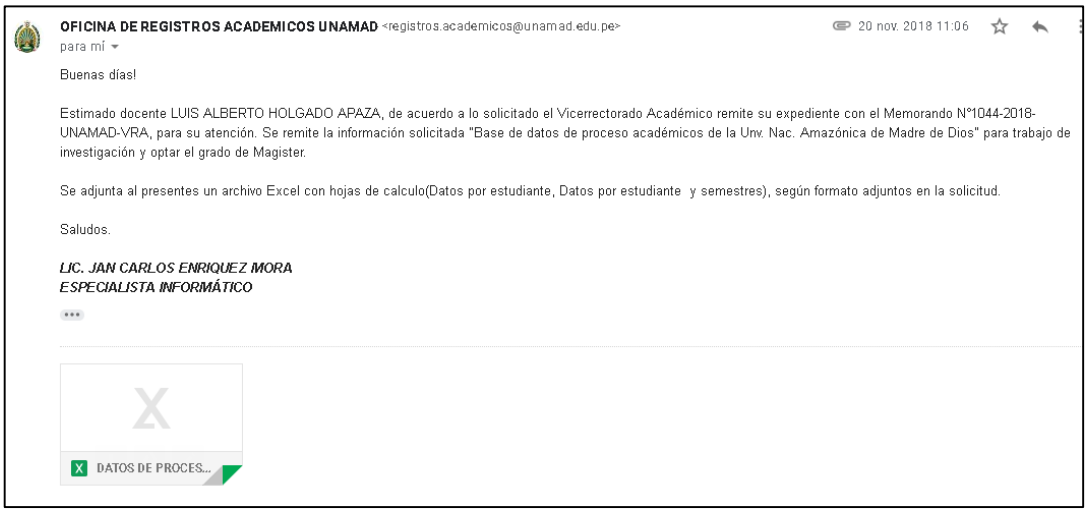
Con sumo agrado me dirijo a su usted, para saldarle cordialmente e informarle que: Siendo docente contratado de la carrera profesional de Ingeniería de Sistemas e Informática, el mismo que me exige tener el grado académico de Magister, y habiendo mi persona terminado la maestría en informática en la UNAP, por lo que solicito a su despacho base de datos con información histórica de los procesos académicos, para realizar la investigación titulada: "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" , el mismo que por la naturaleza de la investigación aplicará técnicas de minería de datos, requiriendo para ello gran cantidad de datos.

Agradezco anticipadamente la atención que brinde al presente documento, sin otro en particular me despido no sin antes expresarle las muestras de mi especial consideración y distinguidos saludos.

Atentamente,

Ing. Luis A. Holgado Apaza
Docente

Anexo 2. Respuesta de carta de solicitud de base de datos histórica de procesos académicos



OFICINA DE REGISTROS ACADEMICOS UNAMAD <registros.academicos@unamad.edu.pe> 20 nov. 2018 11:06

para mí

Buenas días!

Estimado docente LUIS ALBERTO HOLGADO APAZA, de acuerdo a lo solicitado el Vicerrectorado Académico remite su expediente con el Memorando N°1044-2018-UNAMAD-VRA, para su atención. Se remite la información solicitada "Base de datos de proceso académicos de la Unv. Nac. Amazónica de Madre de Dios" para trabajo de investigación y optar el grado de Magister.

Se adjunta al presentes un archivo Excel con hojas de calculo(Datos por estudiante, Datos por estudiante y semestres), según formato adjuntos en la solicitud.

Saludos.

LIC. JAN CARLOS ENRIQUEZ MORA
ESPECIALISTA INFORMÁTICO

DATOS DE PROCES...