DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

TRABAJO DE TITULACIÓN PREVIO LA OBTENCIÓN DEL TÍTULO DE INGENIERÍA EN SISTEMAS E INFORMÁTICA

TEMA: ANÁLISIS TAXONÓMICO PREDICTIVO UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS MEDIANTE LA METODOLOGÍA CRISP-DM PARA PREDECIR POSIBLES CASOS DE DESERCIÓN DE ALUMNOS EN LA UNIVERSIDAD DE LAS FUERZAS ARMADAS - ESPE MATRIZ

AUTORES: AVALOS SERRANO, KATHERINE IRINA
PAGUAY FLORES, SANDRA ELIZABETH

DIRECTOR: ING. ALMACHE, MARIO

SANGOLQUÍ, 2017

CERTIFICADO

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
CARRERA DE INGENIRÍA EN SISTEMAS E INFORMÁTICA

CERTIFICACIÓN

Certifico que el trabajo de titulación "ANÁLISIS TAXONÓMICO PREDICTIVO UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS MEDIANTE LA METODOLOGÍA CRISP-DM PARA PREDECIR POSIBLES CASOS DE DESERCIÓN DE ALUMNOS EN LA UNIVERSIDAD DE LAS FUERZAS ARMADAS - ESPE MATRIZ" realizado por las señoritas AVALOS SERRANO KATHERINE IRINA y PAGUAY FLORES SANDRA ELIZABETH, ha sido revisada en su totalidad y analizado por el software anti-plagio, el mismo que cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de la Fuerzas Armadas ESPE, por lo tanto me permito acreditarlo y autorizar a las señoritas AVALOS SERRANO KATHERINE IRINA y PAGUAY FLORES SANDRA ELIZABETH para que lo sustenten públicamente.

Sangolquí, 27 de julio de 2017

Atentamente,

Ing. Mario Almache Cueva
Director de Proyecto

AUTORÍA DE RESPONSABILIDAD

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
CARRERA DE INGENIRÍA EN SISTEMAS E INFORMÁTICA

AUTORÍA DE RESPONSABILIDAD

Nosotras, AVALOS SERRANO KATHERINE IRINA, con cédula de identidad N° 1719238121, y PAGUAY FLORES SANDRA ELIZABETH, con cédula de identidad N° 1718325523, declaramos que este trabajo de titulación "ANÁLISIS TAXONÓMICO PREDICTIVO UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS MEDIANTE LA METODOLOGÍA CRISP-DM PARA PREDECIR POSIBLES CASOS DE DESERCIÓN DE ALUMNOS EN LA UNIVERSIDAD DE LAS FUERZAS ARMADAS - ESPE MATRIZ" ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citadas bibliografías.

Consecuentemente declaramos que este trabajo es de nuestra autoría, en virtud de ello nos declaramos responsables del contenido, veracidad y alcance de la investigación mencionada.

Sangolquí, 27 de julio de 2017

AVALOS SERRANO KATHERINE IRINA
1719238121

PAGUAY FLORES SANDRA ELIZABETH
1718325523

AUTORIZACIÓN

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
CARRERA DE INGENIRÍA EN SISTEMAS E INFORMÁTICA

AUTORIZACIÓN

Nosotras, AVALOS SERRANO KATHERINE IRINA y PAGUAY FLORES SANDRA ELIZABETH, autorizamos a la Universidad de la Fuerzas Armadas ESPE publicar en la Biblioteca Virtual de la institución el presente trabajo de titulación "ANÁLISIS TAXONÓMICO PREDICTIVO UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS MEDIANTE LA METODOLOGÍA CRISP-DM PARA PREDECIR POSIBLES CASOS DE DESERCIÓN DE ALUMNOS EN LA UNIVERSIDAD DE LAS FUERZAS ARMADAS - ESPE MATRIZ" cuyo contenido, ideas y criterios son de nuestra autoría y responsabilidad.

Sangolquí, 27 de julio de 2017

AVALOS SERRANO KATHERINE IRINA
1719238121

PAGUAY FLORES SANDRA ELIZABETH
1718325523

DEDICATORIA

A Dios, por habernos dado la vida y ser una fuente de fe y fortaleza para no desfallecer durante toda nuestra vida universitaria, guiándonos para superar y afrontar todas las adversidades que se nos ha presentado en el camino y por regalarnos unas maravillosas familias.

A nuestras familias, que con amor han forjado nuestro carácter con valores y principios, siempre brindándonos su comprensión, consejos, ejemplo y apoyo incondicional motivándonos a alcanzar nuestras metas y a ser perseverantes en la vida, y también por ayudarnos con todos los recursos necesarios para culminar con éxito esta fase de nuestras vidas.

A nuestros amigos con los que hemos compartido grandiosos momentos que de una u otra manera forman parte importante en nuestras vidas y han contribuido para alcanzar este objetivo.

AGRADECIMIENTOS

Nuestros más sinceros agradecimientos al Ingeniero Mario Almache Cueva que en calidad de Director del Proyecto con su conocimiento y experiencia ha sabido orientarnos clara y acertadamente en el desarrollo de este trabajo gracias a su participación activa y disponibilidad de tiempo. A Verito Avalos experta en DataMining que ha sido parte fundamental del proyecto que gracias a su guía, consejos y recomendaciones se pudo culminar con éxito.

A nuestras padres y hermanos por darnos todo su amor durante esta etapa y enseñarnos que el camino es duro, pero con trabajo y sacrificio todo lo que nos propongamos lo podremos alcanzar. A nuestros sobrinos que con su ternura y sonrisa han aliviado e iluminado nuestras vidas. A Marco y Diego por su paciencia y apoyo incondicional durante todos estos años convirtiéndose en un pilar esencial en nuestro día a día.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El índice de deserción estudiantil en el Ecuador es un problema que afecta a la sociedad, a la economía y a la Universidad por lo tanto este proyecto se enfoca en el desarrollo de un modelo que permita predecir la posibilidad que tiene un alumno de reprobar una materia o finalmente desertar, de tal forma que se pueda crear estrategias y tomar decisiones que lo solventen utilizando la minería de datos que es un conjunto de técnicas que permiten descubrir patrones de comportamiento a partir de un gran conjunto de datos. Se recuperó la información personal y académica de los alumnos de la Universidad de las Fuerzas Armadas – ESPE Matriz durante los últimos cinco años. Para el desarrollo del proceso de minería de datos se ha utilizado la metodología CRISP-DM que está constituida por seis etapas: Comprensión del Negocio, Comprensión de los Datos, Preparación de los Datos, Modelado, Evaluación y Despliegue. En las dos primeras etapas se utilizó las herramientas Pentaho Data Integration para la creación del Data Warehouse y R Studio para la exploración estadística de los datos, en la tercera etapa se utilizó Open Refine y Pentaho Data Integration y finalmente en las dos últimas etapas la herramienta WEKA que permite la extracción del conocimiento utilizando distintas técnicas de minería de datos.

Palabras clave:

- **DESERCIÓN UNIVERSITARIA**
- **MINERÍA DE DATOS.**
- **CRISP-DM**
- **WEKA.**
- **PENTAHO DATA INTEGRATION.**
- **R STUDIO.**

ABSTRACT

The dropout rate in Ecuador is a problem that affects society, the economy and the university. Therefore, this project focuses on the development of a model that allows predicting the possibility of a student failing a subject. Finally Deserting, in such a way that you can create strategies and make decisions that solvent using data mining is a set of techniques that allow you to detect patterns of behavior from a large dataset. The personal and academic information of the students of the Universidad de las Fuerzas Armadas - ESPE Matriz during the last five years was recovered. For the development of the data mining process, the CRISP-DM methodology has been used, which consists of six stages: Business Understanding, Data Comprehension, Data Preparation, Modeling, Evaluation and Deployment. In the first two stages we used the Pentaho data integration tools for the creation of the data warehouse and R Study for the statistical exploration of the data, in the third stage we used Open data integration of Pentaho and Pentaho in the Two last stages of the WEKA tool that allow the extraction of knowledge using various techniques of data mining.

Keywords:

- **UNIVERSITY DROPOUTS**
- **DATA MINING.**
- **CRISP-DM**
- **WEKA.**
- **PENTAHO DATA INTEGRATION.**
- **R STUDIO.**

CAPÍTULO I
INTRODUCCIÓN

1.1 ANTECEDENTES

La deserción universitaria se define como el proceso de abandono voluntario o forzoso de la carrera en que se matricula un alumno, por la influencia positiva o negativa de circunstancias internas o externas a él (Gonzáles, 2005), según lo define un estudio sobre Repitencia y deserción en la educación superior de Guatemala, realizado por el Instituto Internacional para la Educación Superior en América Latina y el Caribe - IESALC.

La Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura indica que el abandono universitario llega a 40% (UNESCO, 2015). En el Ecuador al 2014, se inscribieron alrededor de 400.000 alumnos en universidades públicas y cofinanciadas, de los cuales "el 26% abandonó su carrera en los primeros semestres según la Secretaría Nacional de Educación Superior" (EL COMERCIO, 2016). En ambos tipos de universidades se prevé que continuará aumentando en los años siguientes, pero a menores tasas.

En septiembre de 2015, el Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la Educación Superior (CEAACES) inició un proceso de Evaluación acreditación y re-categorización institucional, donde se definió una serie de indicadores cuyo objetivo es: "...determinar que su desempeño cumple con las características y estándares de calidad de las instituciones de educación superior, verificando que sus actividades se realicen en concordancia con la misión, visión, propósitos y objetivos institucionales o de carrera..." (Asamblea Nacional, 2010).

En el apartado 1.6.Criterio: Alumnos del documento denominado Adaptación del Modelo de Evaluación Institucional de Universidades y Escuelas Politécnicas 2013 al Proceso de Evaluación, Acreditación y Categorización de Universidades y Escuelas Politécnicas 2015 se evalúa la Eficiencia Académica a través de la Tasa de Retención Grado, es decir, el nivel de permanencia e indirectamente el nivel de deserción (Tasa

de Deserción) de los alumnos de la institución al inicio de su carrera y la Tasa de Titulación grado, que considera el nivel de permanencia de los alumnos hasta el final de su carrera. (CEAACES, 2015).

1.2 PROBLEMÁTICA

Hoy en día la deserción estudiantil en los programas de pregrado es un problema de magnitud global, que tiene un impacto multidimensional en el desarrollo social y económico de un país (Timarán, 2010), debido a que no hay profesionales bien capacitados para servir al mismo, incrementando los niveles de trabajo informal o no especializado, recursos económicos irrecuperables que invierte el Estado en las universidades y que no es retribuido por los alumnos que desertan, siendo muy necesarios en otras áreas como salud, trabajo y defensa. Además, influye con fuerza en varios estamentos, que se detallan a continuación:

- La universidad que, dentro del proceso de evaluación para la acreditación establecido por el CEAACES, debe llegar a una tasa de alumnos egresados por carrera y a nivel de institución, y al no cumplirse las carreras pueden ser cerradas o a su vez la institución, lo que daría como consecuencia desprestigio y desconfianza de la misma, pérdida del presupuesto y recursos en la formación de capital humano que no se incorpora al mercado laboral.
- Los alumnos que, al tener una interrupción de su desarrollo personal y profesional, deriva en la disminución de oportunidades laborales, al desenvolverse en un mercado laboral competitivo, la pérdida económica y de tiempo en estudios infructuosos y el sentimiento de fracaso y de frustración personal (González y Uribe, 2002).
- La sociedad donde se aumenta los niveles de violencia, pues las frustraciones que sienten los jóvenes se trasladan con ellos a otros sectores de la comunidad como la familia, el trabajo y las amistades (González y Uribe, 2002).

Es importante que las universidades implementen una adecuada orientación y

estrategias para disminuir la tasa de deserción y que los alumnos puedan culminar la carrera con éxito, entendiéndose como tasa de deserción, el total de alumnos que estando en condiciones de cursar un determinado nivel en el sistema académico, no lo cursan, en relación a la matrícula del siguiente periodo. Se contabiliza como deserción a la que ocurre durante el periodo escolar como también la que se produce al pasar de un periodo a otro (Román, 2013). En base a las cifras presentadas por el Instituto Nacional de Estadística y Censos (INEC) hubo un incremento de 5,9% de esta cifra del año 2010 al año 2014 siendo 19.5% y 25.4% respectivamente. (El Comercio, 2016). Por esta razón si no se toman medidas a esta problemática provocará un incremento de las cifras de la población que deserta de la universidad y no consigue oportunidades de trabajo convirtiéndose en un problema social.

Actualmente la Universidad de las Fuerzas Armadas "ESPE" no cuenta con un sistema formal que le permita identificar los casos de deserción de los alumnos. Al no existir un diagnóstico oportuno, conlleva a que los alumnos no sean guiados por un proceso de tutorías previo a presentarse el caso de deserción.

1.3 JUSTIFICACIÓN

Los países de América Latina enfrentan desafíos similares en la educación superior, los cuales constituyen el contexto de la deserción estudiantil como son: financiamiento, incremento de la cobertura, aseguramiento de la calidad, mejoramiento de la equidad en el acceso y permanencia, diversificación de la oferta, intereses y necesidades (ciencia, tecnología, sector productivo, investigación, artes y formación integral) y mayor vinculación con el sector laboral y productivo (Pereira, Romero y Toledo, 2013).

Es importante que se realice una investigación profunda, aplicando técnicas de minería de datos, con el objetivo de encontrar relaciones existentes entre atributos, para poder predecir la probabilidad de deserción de un alumno, por ejemplo disminuir la tasa de deserción registrada en el Departamento de Ciencias de la Computación e incrementar el número de graduados con el fin de que la universidad pueda impulsar

nuevas estrategias de retención de alumnos y que siga manteniendo su categoría “A” dentro del proceso de evaluación para la acreditación y disminuya la pérdida del presupuesto invertido en la educación de alumnos desertores. La Universidad como ente indispensable en el desarrollo profesional de los alumnos debe generar un desarrollo personal y profesional de calidad, permitiéndoles el acceso a un mejor nivel de vida y a una vinculación con la sociedad.

“El aprendizaje puede ser definido como cualquier proceso a través del cual un sistema mejora su eficiencia. La habilidad de aprender es considerada como una característica central de los sistemas inteligentes” (García-Martínez & Borrajo, 2000), por lo tanto, la utilización de las Tecnologías de la Información y la Comunicación (TIC) son indispensables para que la información almacenada sea transformada en conocimiento, brindando soluciones eficientes y sustentadas en la realidad y en base a ello se pueda tomar medidas necesarias y oportunas para resolver el problema.

Si se implementa un sistema que permita gestionar de forma adecuada los casos de deserción, los docentes podrán implementar medidas para garantizar y promover condiciones adecuadas que permitan a los alumnos alcanzar resultados exitosos en su carrera académica.

Es importante que antes de la selección de la técnica de minería de datos el analista realice un arduo trabajo en el pre procesamiento de los datos, realizando un análisis exploratorio y gráfico de las variables implicadas y tomando decisiones con criterio de los datos ausentes o atípicos con el objetivo de desplegar datos de calidad para generar modelos, patrones o reglas de mayor certidumbre. Según Zhang & Yang la preparación de los datos es un proceso indispensable porque:

- Los datos obtenidos del mundo real puede ser incompletos, inconsistentes, o presentar algún tipo de ruido.
- La preparación genera grupos de datos que son menores que el grupo original, mejorando la eficiencia del algoritmo.
- La preparación genera datos de calidad, al obtener instancias incompletas, corregir errores o resolver conflictos.

1.4 OBJETIVOS

1.4.1 Objetivo General

Realizar un análisis taxonómico predictivo utilizando técnicas de minería de datos mediante la metodología CRISP-DM para predecir los casos de deserción de alumnos en la Universidad de las Fuerzas Armadas - ESPE matriz.

1.4.2 Objetivos Específicos

- Identificar las causas que afectan e influyen en la deserción universitaria.
- Analizar las distintas técnicas de minería de datos para elegir la adecuada de acuerdo al problema planteado.
- Analizar los datos generados por la minería de datos y proponer un modelo predictivo.
- Desarrollar e implementar el modelo predictivo mediante una interfaz web de consulta.

1.5 ALCANCE

El presente proyecto se enfoca en aplicar cada una de las seis etapas de la metodología CRISP-DM que se encuentran claramente delimitadas, sobre los datos académicos almacenados por la universidad en su Data Warehouse de alumnos de la ESPE - Matriz de todas las carreras de pregrado de modalidad presencial en el intervalo del año 2010 - 2016, y al mismo tiempo aplicar y comparar las diferentes técnicas de minería de datos que se adapten a los objetivos del proyecto. Este trabajo tiene el siguiente alcance:

1. Analizar la información académica obtenida del periodo Septiembre 2010 a Octubre 2016, usando la herramienta libre WEKA.

2. Analizar desde la dimensión de alumno: número de materias por periodo, número de créditos, número de materias aprobadas y reprobadas a lo largo de su periodo estudiantil, tipo de matrícula (primera, segunda o tercera matrícula) del alumno en la materia y otras variables que durante el desarrollo del proyecto de investigación resulten del modelo predictivo.
3. No se van a implementar todas las técnicas de minería de datos existentes, se escogerán tres de las cuales se realizará una comparación de los modelos generados por cada una de estas técnicas y se seleccionará la que proponga una predicción con mayor certidumbre.
4. Desarrollar e implementar una interfaz web de consulta, que permita visualizar los alumnos que tienen probabilidad de deserción por materia, tomando como caso de estudio los alumnos del último período Marzo 2017 – Agosto 2017.

CAPÍTULO II
MARCO TEÓRICO

2.1. DESERCIÓN UNIVERSITARIA

La deserción se define como el abandono prematuro de un programa de estudios antes de alcanzar el título o grado, y considera un tiempo suficientemente largo como para descartar la posibilidad de que el alumno se reincorpore (Himmel, 2002). En Colombia, el Ministerio de Educación considera que un alumno ha desertado cuando este no logra cumplir las aspiraciones de su proyecto educativo, y presenta inactividad académica por un año o más. Es decir, una vez que transcurra dos semestres seguidos sin registrar materias se considera un alumno desertor.

En Ecuador, es una realidad que afecta de forma directa a todas las Instituciones de Educación Superior (IES) de todos los alumnos que ingresan a la universidad y se matriculan para diversas carreras, gran cantidad decide no continuar con los estudios escogidos, según Allauca, J. (2012) son varias las causas que provocan la deserción académica, pero con un solo resultado, alumnos fracasados y expectativas de logros bajos en su desarrollo profesional, personal y económico.

Según René Ramírez titular de la Secretaría Nacional de Educación Superior (Senescyt) ocho de cada diez alumnos, que ingresaron a una universidad o a una escuela politécnica pública, continuaron sus estudios en primer año disminuyendo así el porcentaje de deserción que era del 52% que ahora llega al 20% después de la implementación del ENES. (EL COMERCIO, 2016)

Dentro del documento de Adaptación del Modelo de Evaluación Institucional de Universidades y Escuelas Politécnicas 2013 al Proceso de Evaluación, Acreditación y Re-categorización de Universidades y Escuelas Politécnicas 2015 hace referencia a los índices con los cuales se califica a una institución el grado de deserción las cuales se describen a continuación.

2.1.1. TASA DE RETENCIÓN DE GRADO

Tiene como "propósito medir la capacidad de retención y la eficiencia interna de un sistema educativo. Este indicador proporciona información sobre la retención de alumnos de un grado a otro y, a la inversa, la magnitud del abandono escolar por grado" (UNESCO, 2009). El CEAACES lo define como la tasa de retención que evalúa el nivel de permanencia e indirectamente el nivel de deserción de los alumnos de la institución al inicio de su carrera (CEAACES, 2015). Para calcular la tasa de Retención de Grado según el CEAACES se debe aplicar la siguiente fórmula:

- Tipo de indicador: Cuantitativo
- Período de evaluación: El período se determina por la definición de las cohortes.

Ecuación 1. Tasa de Retención Grado.

Fuente: (CEAACES, 2015)

$$\mathrm{TR} = 100 * \frac{\mathrm{NEMA}}{\mathrm{NTEA}}$$

Donde:

- TR: Tasa de retención.
- NEMA: Número de alumnos matriculados durante el período académico ordinario en el que se efectúa la evaluación de la institución, que fueron admitidos dos años antes.
- NTEA: Número total de alumnos que fueron admitidos en la carrera dos años antes del período de evaluación.

Se aplicó la Ecuación 1 sobre la Tasa Retención Grado en la Universidad de las Fuerzas Armadas – ESPE con la información de los últimos 5 años para lo cual se obtuvo que la universidad tiene 54.95% de retención grado, por lo tanto se puede asumir que el resto, es decir, aproximadamente un 50% deserta.

$$TR = 100 * \frac{3163}{5751} = 54.947\%$$

2.1.2. TASA DE TITULACIÓN GRADO

El indicador complementa la evaluación de la eficiencia académica considerando el nivel de permanencia de los alumnos hasta el final de su carrera (CEAACES, 2015). Para calcular la tasa de Retención de Grado según el CEAACES se debe aplicar la siguiente fórmula:

- Tipo de indicador: Cuantitativo.
- Período de evaluación: El período se determina por la definición de las cohortes.

Ecuación 2. Tasa de Titulación Grado.

Fuente: (CEAACES, 2015)

$$TT = 100 * \frac{N\ EG_p}{N\ EC_P}$$

Donde:

- TT: Tasa de titulación.
- NEGp: Número de alumnos de grado que ingresaron en la (s) cohorte(s) definidas y se graduaron hasta el final del último período académico regular concluido antes del inicio del proceso de evaluación.
- NECp: Número de alumnos de grado que ingresaron en la(s) cohorte(s) definidas.

Se aplicó la Ecuación 2 sobre la Tasa Titulación en la Universidad de las Fuerzas Armadas – ESPE con la información de los últimos 5 años para lo cual se obtuvo que la universidad tiene 20.731% de tasa de titulación que es un porcentaje relativamente bajo de graduados.

$$\mathrm{TT} = 100 * \frac{590}{2846} = 20.731\%$$

2.2. MINERÍA DE DATOS

La minería de datos surge de dos ramas, "la inteligencia artificial y la estadística, dichas técnicas son plasmadas en algoritmos, que después se aplican sobre un conjunto de datos para obtener resultados" (García et al, 2001). Estos resultados permiten detectar fácilmente patrones en los datos en grandes cantidades con el objetivo de encontrar información predecible y dar respuesta a preguntas que tradicionalmente requerían un intenso análisis manual (Vallejos, 2006).

2.2.1. TÉCNICAS DE MINERÍA DE DATOS

Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados-predictivos y no supervisados o de descubrimiento del conocimiento (Weiss y Indurkhya, 2010).

- Algoritmos supervisados: Predicen el valor de un atributo de un conjunto de datos a partir de datos cuyo resultado ya se conoce, por lo que busca una relación entre dichos atributos.
- Algoritmos no supervisados: Son algoritmos que no utilizan datos históricos y descubren patrones en los datos actuales que sirve para tomar decisiones y acciones sobre el conocimiento, obteniendo un beneficio (científico o de negocio) de ellas.

Tabla 1.
Clasificación de las técnicas de minería de datos.

SUPERVISADOS	NO SUPERVISADOS
Árboles de decisión	Segmentación
Inducción neuronal	Agrupamiento ("clustering")
Regresión	Reglas de asociación
Series temporales	Patrones secuenciales

a) Árboles de Decisión

Son herramientas analíticas muy potentes que generan representaciones gráficas en forma de estructura de árbol, para el descubrimiento de reglas, relaciones y patrones mediante una base de datos establecida, generalmente utilizan una técnica de aprendizaje supervisado. En la minería de datos es empleada para dar solución a problemas de predicción, clasificación y segmentación por medio de una serie de condiciones sucesivas (reglas).

Según BERMEO, M (2013) los árboles de decisión están conformados por:

- Nodo de decisión: Representado por un cuadrado, es el punto donde se toma una decisión dentro del proceso.
- Nodo de probabilidad: Representado por un círculo es el punto del proceso donde ocurre un evento aleatorio.
- Rama: Son los caminos que puede tomar un proceso cuando se toma una decisión.

Todos los caminos del árbol tanto las ramas como los nodos establecen una regla de clasificación. Un claro ejemplo de árbol de decisión para saber si se le otorga a una persona un crédito como se muestra en la Figura 1.

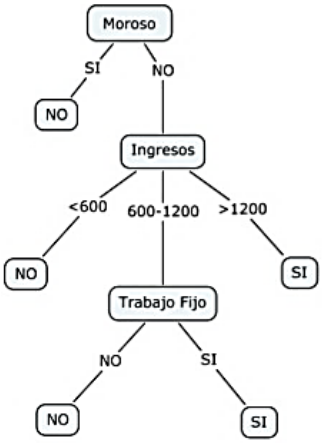

Figura 1. Árbol de decisión para préstamo de crédito.
Fuente: BERMEO, M (2013)

b) Inducción Neuronal

Es un conjunto de elementos que procesan la información de forma automática comportándose de forma parecida a nuestro cerebro, con su método de aprender de la experiencia y el pasado (Molina, J. & García, J. 2006) de este modo aplica tal conocimiento a la resolución de problemas nuevos. Creando un sistema altamente interconectado de neuronas en una red son capaces de detectar y aprender de patrones y sus características de los datos analizados, además, una vez entrenada una red neuronal, ésta puede hacer clasificaciones y segmentación (Aguilar, J. & Estrada, C. 2012). La red neuronal está constituida por tres capas, entrada oculta y salida como se muestra en la Figura 2. Se puede aplicar las redes neuronales en varios ámbitos como son: Reconocimiento de imágenes, Reconocimiento de voz, Análisis y filtrado de señales, Análisis financiero, Predicción dinámica, etc. Algunos ejemplos de red neuronal son:

- El Perceptrón multicapa.
- Los Mapas Auto-organizados, (redes de Kohonen).
- El Perceptrón.
- Máquinas de Boltzmann

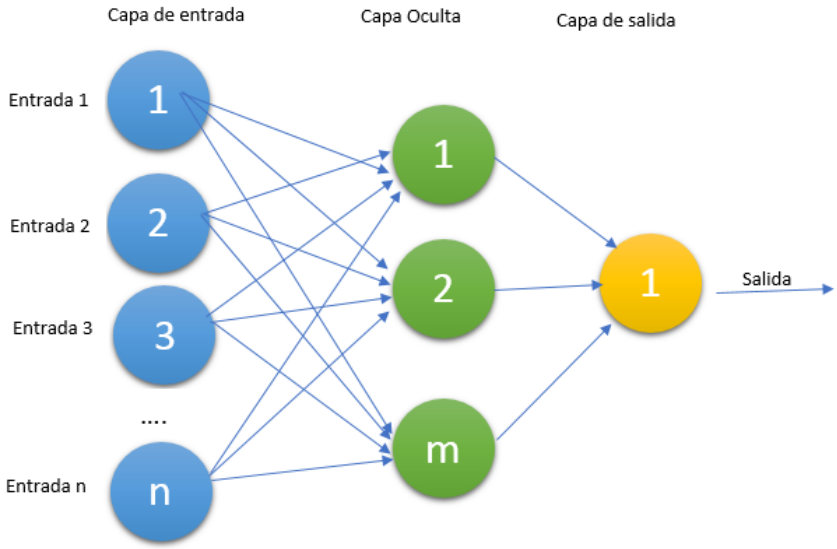

Figura 2 Ejemplo de Red Neuronal

c) Regresión Lineal

La regresión lineal es un modelo simple, que puede predecir una o más variables continuas y ausentes basándose en su relación con otras dentro de una tabla de datos. El objetivo del modelo de minería de datos es representar estos datos como nodo único, definiendo una fórmula de regresión a la que mejor se ajusten, es decir se "aproxima la relación de dependencia entre una variable dependiente Y, las variables independientes Xi y un término aleatorio ε" (Walpole et al, 1999). Este modelo puede ser expresado como se muestra en la Ecuación 3.

Ecuación 3. Modelo de regresión lineal.

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \varepsilon$$

Se representa la regresión lineal a partir de una recta que corresponde a la estimación obtenida a partir de 20 pares de observaciones donde x representa la temperatura fijada en un recinto cerrado e Y el ritmo cardíaco de un vertebrado (ver Figura 3).

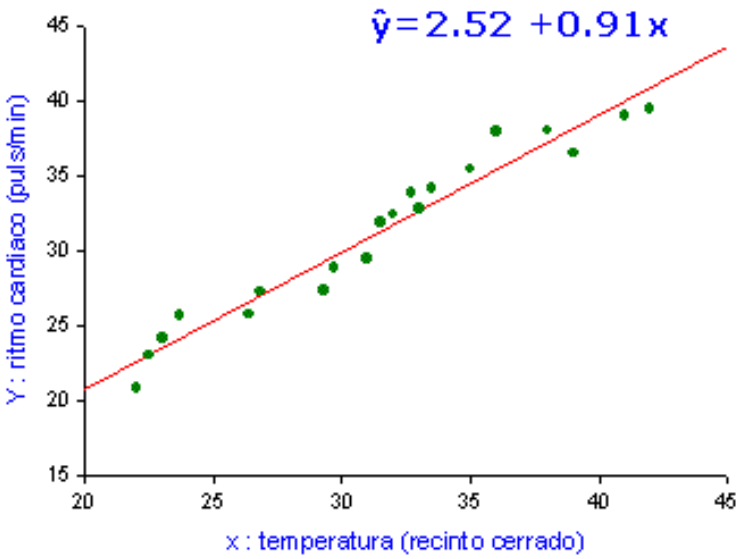

Figura 3. Ejemplo de Regresión Lineal
Fuente: Dpto. de Matemática Aplicada (Biomatemática)

d) Series Temporales

Según Chris Chatfield (2003), una serie temporal consiste en una colección de observaciones realizadas de manera secuencial en el tiempo. Existen dos tipos de series temporales continuas y discretas, produciéndose la primera cuando las observaciones son hechas de forma continua en el tiempo y se consideran discretas si las observaciones tienen lugar solo en momentos específicos. En la Figura 4 se grafica la cantidad de pasajeros que realizaron vuelos internacionales en una aerolínea durante el período de tiempo especificado siendo una serie de tiempo discreta. El principal objetivo de una serie de tiempo es interpretar eventos ocurridos ya que está constituida por observaciones históricas de uno o varias variables conteniendo información valiosa para su dominio de procedencia.

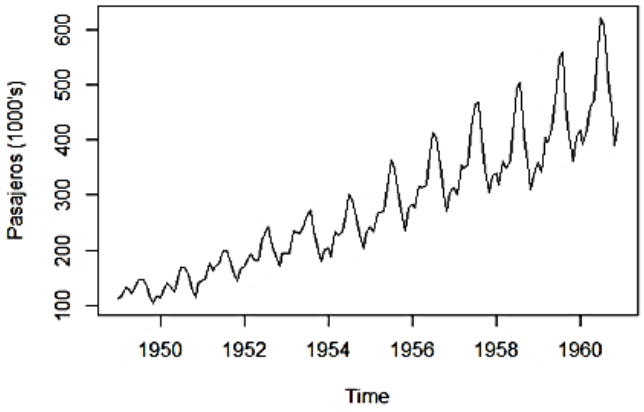

Figura 4. Ejemplo de Series Temporales.
Fuente: Valero C. S.

e) Segmentación

Según Marques (2013) la segmentación se define como el proceso de dividir un universo en grupos uniformes más pequeños que tengan características comunes denominados segmentos. La segmentación se debe realizar según los valores de determinadas variables que son los que determinan sus características.

La clasificación de técnicas de segmentación se divide en técnicas predictivas, en las que las variables que intervienen en el proceso pueden clasificarse inicialmente en dependientes e independientes y técnicas descriptivas, en las que todas las variables tienen inicialmente el mismo estatus. Las técnicas de segmentación también se conocen como técnicas de clasificación porque que permiten extraer perfiles de comportamiento o clases. Se usa para identificar grupos que tienen características comunes como se muestra en la Figura 5. Las técnicas de segmentación permiten identificar claramente el comportamiento de un grupo de casos que difiere de otros grupos o conjuntos.

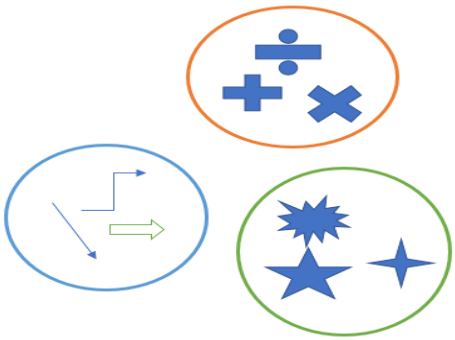

Figura 5 Ejemplo de Segmentación

f) Agrupamiento (Clustering)

Esta técnica agrupa datos en un número de clases que se pueden establecer o que previamente estaban establecidas, utilizando criterios de distancia o similitud como se muestra en la Figura 6, de tal manera que permita que las clases sean similares entre sí y distintas con las otras. Además el Clustering divide la información en grupos que a diferencia de la segmentación, no se conoce donde se tendrá un cluster o de que datos se crearán los clusters su objetivo es determinar grupos y su pertenencia. Este a su vez tiene diversas técnicas como, por ejemplo:

- Algoritmo K-means.
- Algoritmo K-medoids.
- Reglas de asociación: Busca un patrón que se descubre sobre la base de una relación entre los elementos en la misma transacción, como, por ejemplo: “Cuando una mujer compra un vestido necesariamente va a comprar unos zapatos nuevos”.

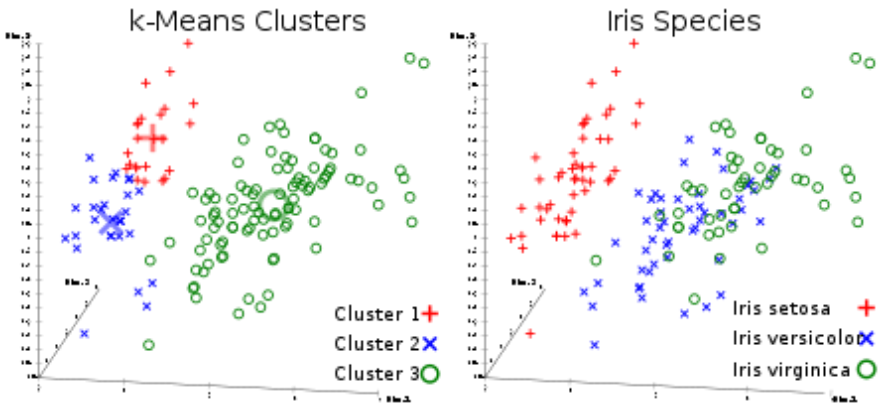

Figura 6 Ejemplos de Clusters
Fuente: Amorim, R. C.

g) Patrones Secuenciales

La minería de secuencias es un caso particular de la minería de datos estructurados. Consiste en encontrar patrones estadísticamente relevantes en colecciones de datos que están representados de forma secuencial. (MABROUKEH, 2010). En este caso la técnica de patrones secuenciales busca patrones en forma de un condicional que frecuentemente están relacionados con el tiempo u otras consecuencias, por ejemplo: "Si una persona compra la película Rápidos y Furiosos 1, después de un tiempo comprará Rápidos y Furiosos 2" (ver Figura 7), por ello en este tipo de técnica es indispensable el orden, ya que dependiendo del primer suceso ocurrirá el siguiente como consecuencia del anterior. Este tipo de técnica resulta ventajosa en el mercadeo, predicción y otros.

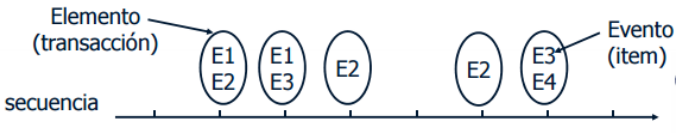

Figura 7 Ejemplo de Secuencia
Fuente: Berzal, F.

2.2.2. HERRAMIENTAS DE MINERÍA DE DATOS

El siguiente apartado contiene una breve descripción sobre algunas de las herramientas empleadas en la minería de datos, que facilitan la extracción de modelos, patrones y tendencias para la generación de conocimiento a través de datos y para predecir futuros comportamientos. Algunas de las herramientas de minería de datos se mencionan a continuación:

a) SPSS Clementine

SPSS Clementine es una potente herramienta para análisis estadísticos y gestión de información que cuenta con un entorno gráfico que solicita información al usuario para realizar el trabajo fuerte (ver Figura 8), además permite analizar archivos con grandes cantidades de datos sin utilizar grandes cantidades de espacio de almacenamiento temporal en disco. También, permite la utilización de un servidor que aumenta la velocidad de acceso y procesamiento de la información compleja, debido que utiliza un análisis en modo distribuido de la información. Entre las características que ofrece se tienen las siguientes:

- Distintas herramientas de minería de datos: correlación, reglas de asociación (GRI, a priori), patrones secuenciales (regresión), segmentación (Kohonen, Two-step y k-means), clasificación (redes neuronales, reglas y árboles de decisión).
- Manipulación de datos (pick & mix, muestreo, combinación y separación). · Combinación de modelos.
- Visualización anterior (datos).
- Exportación de modelos a distintos lenguajes (C, SPSS, SAS).
- Exportación de datos integrada a otros programas (XLS). · Generación de informes.

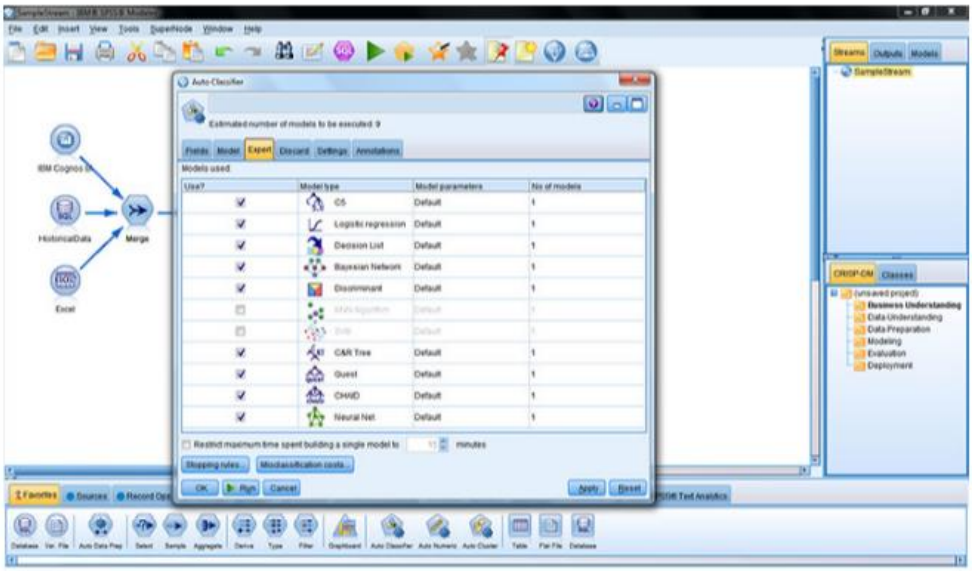

Figura 8 Interfaz de SPSS Clementine
Fuente: IBM, (2017)

b) WEKA

WEKA (Entorno Waikato para el Análisis del Conocimiento) es una herramienta de libre distribución escrita en lenguaje Java con acceso a bases de datos SQL procesando los resultados como una consulta de base de datos. Es decir, ofrece un entorno de experimentación de análisis de datos en función del pre procesamiento, agrupamiento, clasificación, regresión, visualización y características de selección (ver Figura 9), a partir de un análisis y evaluación de las técnicas más relevantes de análisis de datos alineados con el aprendizaje automático.

Sus técnicas se basan en la hipótesis de que los datos están disponibles en un único archivo plano en formato ARFF (archivo plano organizado en filas y columnas), además cuenta con un preprocesador de datos donde se selecciona y transforma los atributos y aplica técnicas de aprendizajes tanto supervisados como no supervisados. Weka contiene herramientas para diferentes tareas básicas como:

- Preprocess: Multitud de herramientas para el preprocesamiento de los datos (como por ejemplo discretización de variables).
- Classify: Algoritmos de clasificación, distribuidos por paquetes, como por

ejemplo ID3 o C4.5

- Cluster: Diferentes algoritmos de segmentación como el simple k-means.
- Associate: Algoritmos para encontrar relaciones de asociación entre variables (Apriori entre otros).
- Select atributtes: Aquí, una vez cargados los datos, Weka es capaz de buscar por nosotros las mejores variables del modelo.
- Visualize: Herramienta de visualización de datos en los ejes cartesianos, con muchas posibilidades.

Figura 9 Interfaz de WEKA
Fuente: De la Calle, J.

c) KEPLER

Es una herramienta comercial distribuida por Dialogis que tiene múltiples modelos de análisis con un pre procesado de datos que integra varias herramientas de aprendizaje como se observa en la Figura 10, para seleccionar un modelo o manipular la representación gráfica de los modelos obtenidos como:

- Árboles de decisión
- Redes neuronales
- Regresión no lineal
- Aplicaciones estadísticas

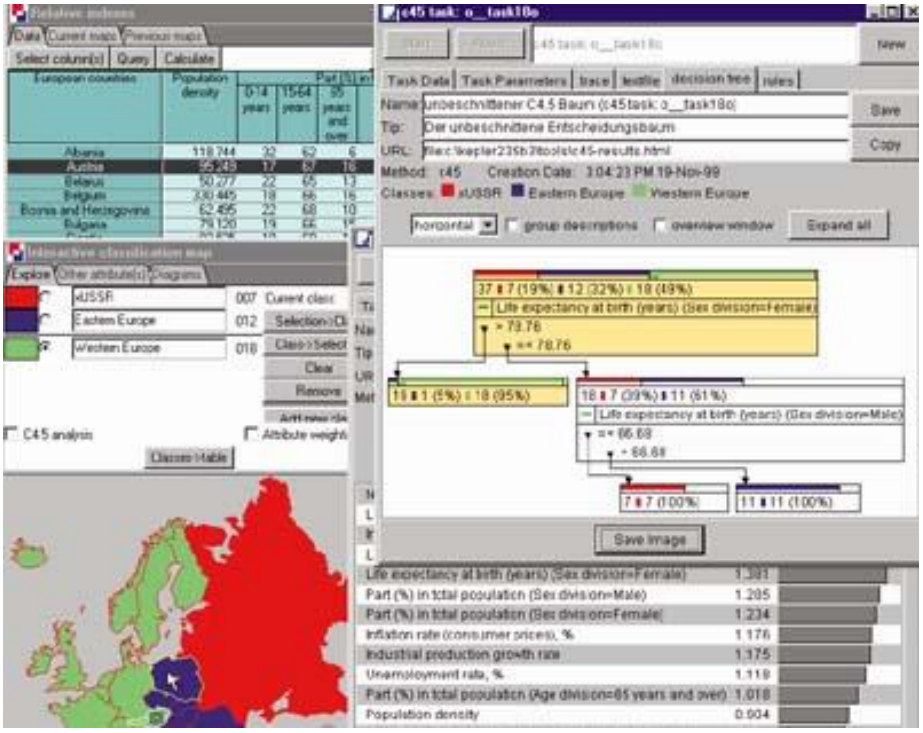

Figura 10 Interfaz de Kepler
Fuente: Andrienko et al

d) ODMS

ODMS (Oracle Data Mining Suite) es una herramienta comercial diseñada en base a una arquitectura cliente servidor que brinda una gran capacidad de adaptarse con rapidez al acceso a grandes volúmenes de información (ver Figura 11). Sus principales características son:

- Acceso a datos en diversos formatos: almacenes de datos, bases de datos relacionales, archivos planos etc.
- Preprocesado de datos: muestreo de datos, patrones de datos.
- Modelos de aprendizaje: redes neuronales, regresión lineal.
- Herramientas de visualización.
- Importación de datos a varias plataformas.

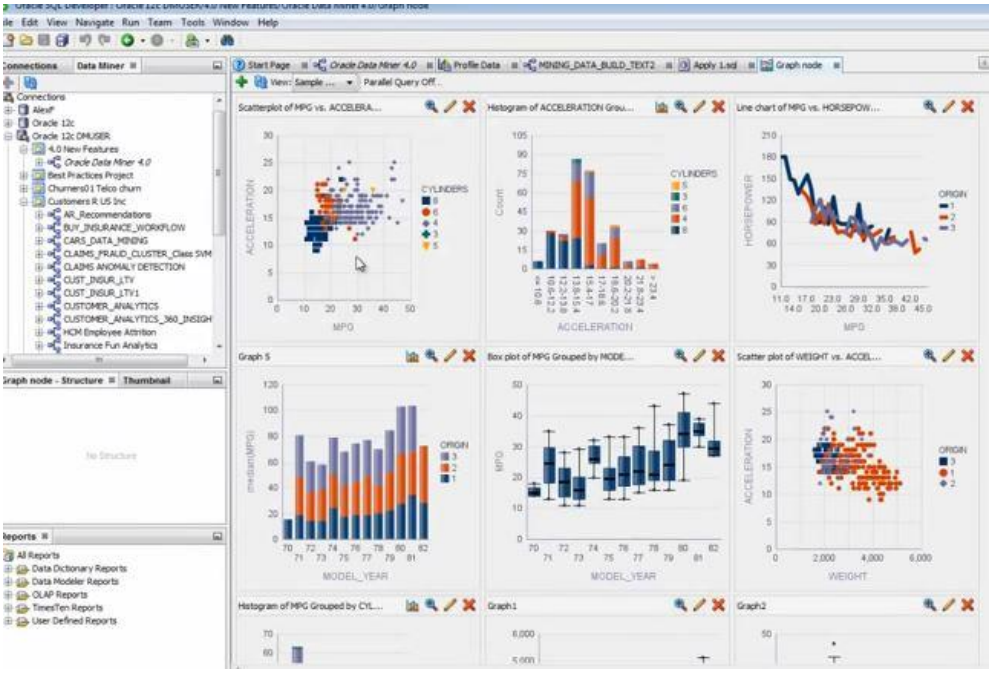

Figura 11 Interfaz de ODMS
Fuente: Oracle

e) DBMINER

DBMINER es un sistema interactivo de minería de datos desarrollado por la Universidad de Simon Fraser (ver Figura 12). Que tiene dos tipos de licencias pública y comercial para el uso empresarial. Fue creado para la obtención de conocimiento de bases de datos relacionales, almacenes de datos y Web. Su arquitectura está basada en los siguientes tipos de análisis: OLAP (Procesamiento Analítico en línea) y OLAM (Minería Analítica en Línea) y sus módulos de trabajo pueden a través de una interfaz gráfica o vía interfaz de script. Entre las principales características se encuentran:

- Un conjunto de algoritmos bastante amplio y basado en consultas definidas en base a jerarquías conceptuales las cuales describen determinados conceptos a diferentes niveles de generalidad (desde valores muy específicos hasta conceptos generales). Por medio de las jerarquías conceptuales es posible extraer conocimiento a diferentes niveles de granularidad.
- Integración de los algoritmos con el SGBD orientado hacia la obtención de una

mayor eficiencia.

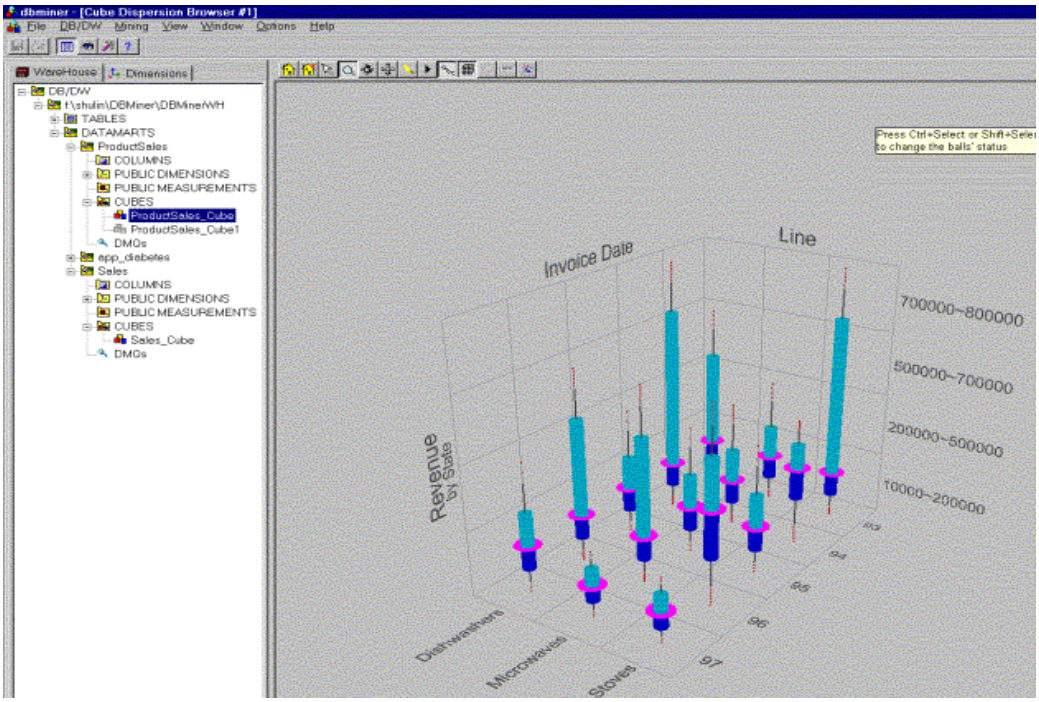

Figura 12 Interfaz de DBMINER
Fuente: LIN, S. y ANTON, C

f) RAPID MINER

Es una herramienta de aprendizaje desarrollada en lenguaje Java por el departamento de inteligencia artificial de la Universidad de Dormund y se mantiene una versión open-source (ver Figura 13). El sistema cuenta con operaciones para la importación y pre-procesamiento de datos, aprendizaje automático, validación de modelos y aplica técnicas de minería de datos supervisadas y no supervisadas. RapidMiner ofrece minería de datos y procesos de aprendizaje automático que incluye:

- Carga y transformación de datos (extracción, transformación, carga ETL)
- Pre-procesamiento de datos y visualización
- Análisis predictivo y modelos estadísticos
- Evaluación y despliegue

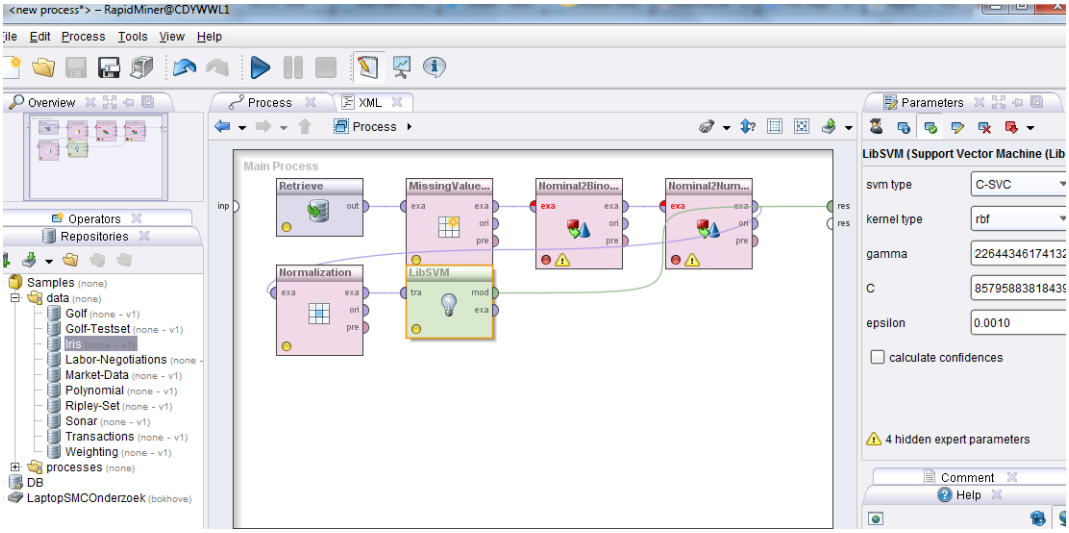

Figura 13 Interfaz de RapidMiner
Fuente: Whatasoftware

g) R Studio

R en sus inicios fue desarrollado en el año 1993 en Nueva Zelanda por el Departamento de Estadística de la Universidad de Auckland. A pesar de entregar un entorno estadístico potente y completo, su desventaja radica en la dificultad de manejo, porque no cuentan con una interfaz amigable como se muestra en la Figura 14. "Las llamadas a R se realizan en línea de comando y sus paquetes, por desgracia, no siempre se utilizan de la misma forma (al provenir de desarrolladores diferentes, lo que dificulta la realización de muchas tareas)" (Cubero y Berzal, 2014). Las principales características que brinda R son:

- Orientado al proceso y análisis de datos (con data.frames).
- Gráficos potentes
- El más utilizado en análisis de datos
- Visualización del proceso de exploración.

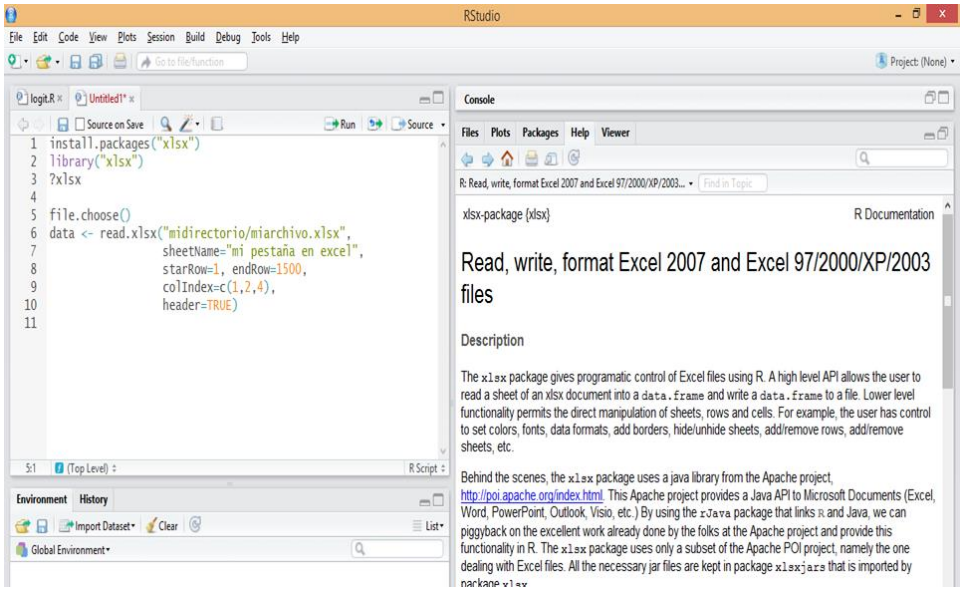

Figura 14 Interfaz de R
Fuente: Temabetametrica

2.2.3. ÁREAS DE APLICACIÓN DE MINERÍA DE DATOS

Con el tiempo la minería de datos ha ido extendiendo sus áreas de aplicación debido a la rapidez del avance tecnológico, además la minería de datos actualmente es aplicada en distintos sectores, ya que les permite beneficiarse de la obtención de datos fiables. Entre algunas áreas que abarca la minería de datos son:

- Financiera/Marketing: Es utilizada para encontrar datos que permitan tener un nivel de confianza que aseguren que es posible practicar análisis sistemáticos en condiciones avanzadas y con garantías de fiabilidad. Algunos ejemplos son:
    - Diseño y construcción de almacenes de datos para el análisis multidimensional de datos y minería de datos.
- Económica:
    - Préstamo de pago, predicción y análisis de políticas de crédito del cliente.

  - La clasificación y agrupación de los clientes para la comercialización de destino.
  - La detección de lavado de dinero y otros delitos financieros.

- Telecomunicaciones: Debido al rápido desarrollo de la tecnología las telecomunicaciones han crecido a grandes pasos llegando a ofrecer diversos servicios como: redes sociales, teléfonos inteligentes, mensajería, correo electrónico, transmisión de datos web, etc. Por este motivo la minería de datos se vuelve muy importante para ayudar y entender el negocio permitiendo:
  - Identificar los patrones de telecomunicaciones.
  - Detectar fraudes.
  - Mejorar el uso de recursos
  - Mejorar la calidad del servicio
  - Análisis de servicios móviles, etc.
- Salud: La minería de datos permite mejorar los sistemas de salud, debido al análisis que realiza para identificar y mejorar procesos tales como:
  - Diagnóstico y tratamiento: ayudando a los médicos a identificar los tratamientos más eficaces y definir mejores prácticas exportables.
  - Predecir el volumen de pacientes en cada categoría.
  - Monitorear procesos que aseguren de que los pacientes reciban la atención adecuada en el lugar correcto y en el momento adecuado.
  - La minería de datos también puede ayudar a las aseguradoras de salud para detectar el fraude y el abuso.
- Educación: “La minería de datos en la educación es una disciplina emergente que desarrolla métodos para explorar los datos que vienen de entornos educativos, y los usa para entender mejor a los alumnos y los entornos en los que aprenden.” (CHO et al, 2004). La minería de datos en esta área se denomina MDE. La MDE ayuda a:
  - Promover la creación de nuevos modelos y métodos educativos.
  - Análisis de los alumnos y su comportamiento.
  - Análisis del aprendizaje asistido por computador, etc.

La aplicación de la minería de datos en el presente proyecto se centra en la educación, en esta área "La MDE tiene como objetivo obtener una mejor comprensión del proceso de aprendizaje de los alumnos y de su participación global en el proceso, orientado a la mejora de la calidad y rentabilidad del sistema educativo" (Winters, T, 2006), como punto específico de la investigación es la aplicación de la Minería de datos para analizar la deserción estudiantil.

2.2.4. APLICACIÓN DE LA MINERÍA DE DATOS PARA ANALIZAR LA DESERCIÓN UNIVERSITARIA

La aplicación de la minería de datos en la índole educativa ya lleva varios años desarrollándose por lo que ha sido relevante su aplicación permitiéndole predecir los fenómenos que se presenten. "De esta forma, utilizando las técnicas que nos ofrece la minería de datos, se puede predecir, con un porcentaje muy alto de confiabilidad la probabilidad de desertar de cualquier alumno" (Orea et al, 2005). En el entorno internacional se han desarrollado algunos proyectos de investigación aplicando la minería de datos al descubrimiento de patrones de deserción estudiantil, a continuación, se mencionan algunos:

- Heredia, Amaya y Barrietos han desarrollado un modelo predictivo, que permite determinar la probabilidad que un alumno abandone la universidad, teniendo en cuenta las reglas de conducta y el entorno del alumno. (Heredia et al, 2015).
- El estudio titulado, "Predicción del Factor de Abandono de Alumnos usando Técnicas EDM" analiza los factores que contribuyen el rendimiento académico y la predicción de su fracaso y abandono utilizando dos técnicas de minería de datos: clasificación (inducción) y árboles de decisión a través de la herramienta WEKA. (Pradeep et al, 2015).
- Otro artículo propuesto por Shaleena y Shaiju sobre "Data Mining Techniques for Predicting Student Performance", propone un método en el que se utilizan clasificadores de árboles de decisión para clasificar a los alumnos de acuerdo a sus datos académicos, personales y familiares. (Shaleena y Shaiju, 2015).

La aplicación de estos modelos ha sido minoritaria en el contexto de la Universidad de la Fuerzas Armadas “ESPE”, debido a que se tiene problemas con la información ya que ésta, por ejemplo se encuentra con formatos incorrectos, existen registros incompletos de alumnos, además éstas se han desarrollado en un contexto y entorno diferente, porque durante los últimos cinco años han existido varios cambios políticos y sociales a nivel de educación, por lo tanto, es importante que se realice un estudio en el contexto del Ecuador.

2.3. METODOLOGÍAS DE MINERÍA DE DATOS

Las metodologías permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial, además ayudan a las organizaciones a entender el proceso de descubrimiento de conocimiento y proveen una guía para la planificación y ejecución de proyectos, por ello en los últimos años se ha enfatizado en la creación de metodologías para Minería de Datos. Con el tiempo son diversas las metodologías propuestas para el desarrollo de proyectos de Minería de Datos como lo son: SEMMA (Sample, Explore, Modify, Model, Assess), KDD Process o CRISP-DM Cross Industry Standard Process for Data Mining), siendo la metodología CRISP-DM la más utilizada en ambientes académicos e industriales como se muestra en la Figura 15.

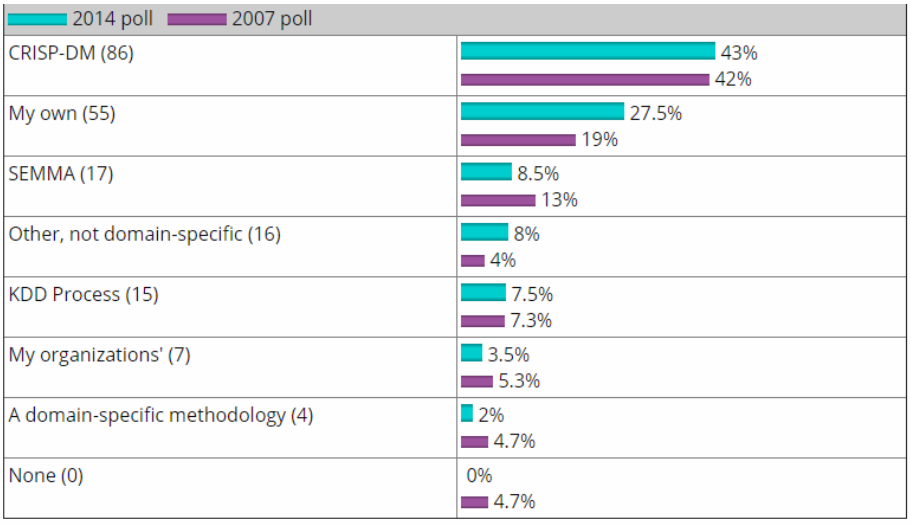

Figura 15. Comparación de las encuestas realizadas por la KDnuggets en el año 2007 y 2014.
Fuente: (KDnuggets, 2014)

Para el presente proyecto se optó por utilizar CRISP-DM por su flexibilidad y porque ofrece etapas cada una con tareas claras y específicas que pueden interactuar con gran fluidez entre sí, permitiendo tener una guía apropiada del proceso de minería de datos. Además, tiene la ventaja de que no solo ha sido construida de manera teórica sino que se basa en experiencias reales aplicándose en varios proyectos de minería de datos.

2.3.1. METODOLOGÍA SEMMA

SEMMA (Sample, Explore, Modify, Model, Assess) es una metodología propuesta por SAS que a comparación con la metodología CRISP-DM es más corta porque se centra en el desarrollo del proceso de Minería de datos y no toma en cuenta los objetivos del negocio al que se esté aplicando la minería de datos, ni el despliegue o la explotación de modelos resultantes. Está compuesta por varias fases (ver Figura 16) las cuales no están descrita de forma rígida, lo que quiere decir que no es necesario terminar una fase antes de comenzar por otra fase.

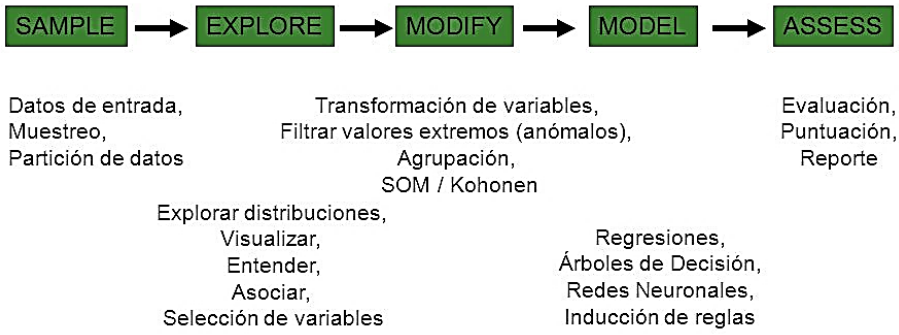

Figura 16. Fases de la Metodología SEMMA.
Fuente: (Oquendo,2016)

A continuación, se presenta una breve descripción de cada una de las fases de la metodología SEMMA que lo realiza Ian et al en su libro Practical Machine Learning Tools and Techniques.

- Muestreo: En esta fase se selecciona una muestra aleatoria y representativa del problema de estudio que tenga un nivel confianza.
- Exploración: Su objetivo es simplificar y optimizar la eficiencia del modelo a través de herramientas de visualización y técnicas estadísticas para seleccionar las variables explicativas o entradas del modelo.
- Modifica: Se encarga de formatear los datos, filtra los valores anómalos, los agrupa y transforma las variables para ser utilizado por el modelo.
- Modela: Establece una relación entre las variables explicativas y variables objetivo a través de la implementación de técnicas de minería de datos supervisadas o no supervisadas.
- Evalúa: Entrega un reporte con los resultados obtenidos del análisis de bondad del modelo, contrasta con otros métodos estadísticos o con nuevas muestras.

2.3.2. PROCESO DE EXTRACCIÓN DEL CONOCIMIENTO (KDD)

"El proceso no trivial de identificar patrones válidos, nuevos, potencialmente útiles y en última instancia comprensible en los datos" (Usama Fayyad, 1996). En la Figura 17 se ilustra las fases del proceso KDD, el mismo que se constituye en 7 fases del proceso KDD:

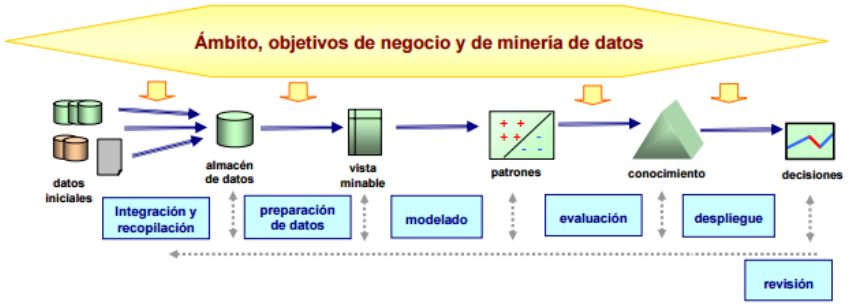

Figura 17. Fases del KDD.
Fuente: (Hernández Universidad Politécnica de Valencia)

1. Identificar las fuentes de información que pueden ser útiles y dónde conseguirlas.
2. Diseñar el esquema de un almacén de datos (Data Warehouse) con el objetivo de integrar toda la información.
3. Establecer un almacén de datos que permita la "navegación" y visualización previa de sus datos, para filtrar los aspectos que resulten interesantes ser estudiados.
4. Selección, limpieza y transformación de los datos analizados.
5. Seleccionar y aplicar el método de minería de datos apropiado para el entorno de estudio.
6. Evaluación, interpretación, transformación y representación de los patrones extraídos.
7. Difusión y uso del nuevo conocimiento.

2.3.3. METODOLOGÍA CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) es la metodología más utilizada en el desarrollo de proyectos de minería de datos en los ambientes académico e industrial (Gallardo, 2010).

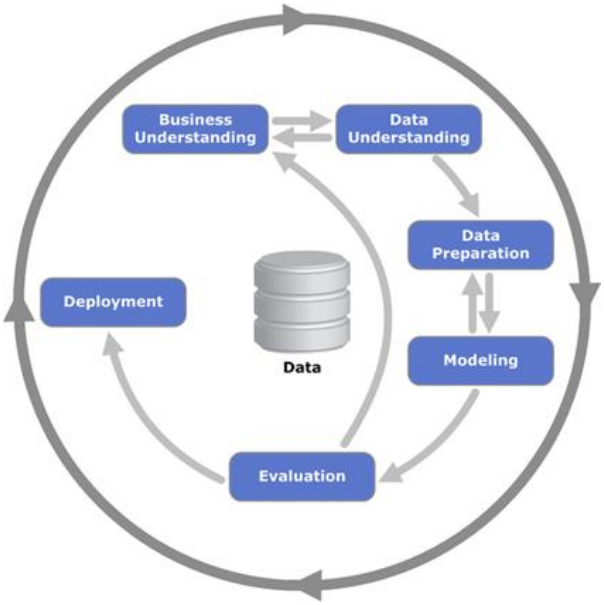

Figura 18. Metodología CRISP-DM
Fuente: (Rodríguez, 2010)

En la Figura 18 se ilustra el ciclo de vida de un proyecto de minería de datos el cual consta de seis fases que no son rígidas, permitiéndoles un movimiento bidireccional hacia adelante y hacia atrás entre las fases. Las flechas muestran las interacciones más importantes y frecuentes entre fases. (DATAPRIX, 2015).

a) Comprensión del Negocio

Según (Rodríguez, 2010), en esta primera y más importante etapa de la metodología, principalmente se enfoca en comprender claramente el problema a resolver, los requisitos, identificando los objetivos desde la perspectiva del cliente o

interesado, para lograr convertirlos en objetivos técnicos los mismos que serán los objetivos de la minería de datos, ya que solo con una comprensión clara y correcta del negocio se podrá recolectar datos en información correcta para el proyecto que se pretenda desarrollar. Esta esta etapa cuenta con cuatro tareas fundamentales que se muestran en la Figura 19.

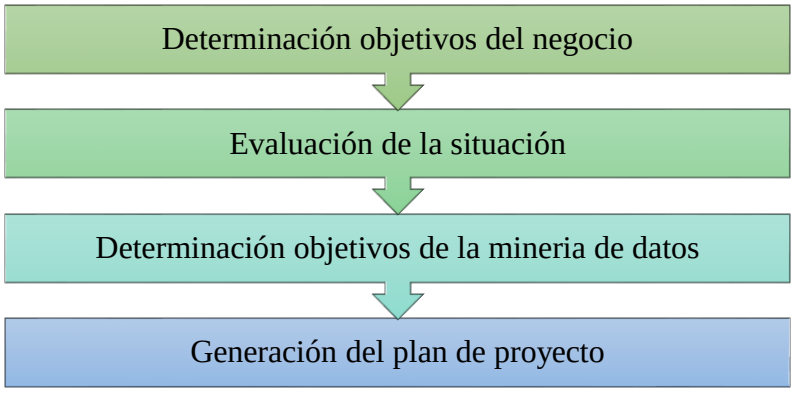

Figura 19. Tareas de la Etapa: Comprensión del Negocio

1. **Determinación objetivos del negocio.**

   Se entiende el problema a resolver, es decir, lo que el cliente quiere lograr en sus términos, además se establece que objetivos tienen mayor prioridad unos sobre otros y finalmente determinar porqué se utilizará la minería de datos en la resolución del problema, una vez hecho esto se procede a registrarlos y detallarlos a cada y de estos y a su vez a establecer criterios de éxito que deben ser específicos y capaces de ser medidos.

2. **Evaluación de la situación.**

   Aquí se debe calificar los antecedentes, el estado de la situación antes del proceso de minería de datos y los requisitos del problema en términos del negocio como en términos de la minería de datos. Esta tarea es similar a la anterior con la diferencia que esta se amplía más sobre los detalles, entre algunos aspectos que deben ser considerados son:

- Recursos disponibles para el proyecto: personal, datos, recursos hardware y software.
- Requerimientos del proyecto
- Presunciones
- Restricciones
- Riesgos

Finalmente se realiza un análisis de costo-beneficio para el proyecto.

3. **Determinación objetivos de la minería de datos.**

   En esta tarea se representan los objetivos del negocio en términos de los objetivos de la minería de datos y finalmente se establecen los criterios de éxito de los objetivos propuestos.

4. **Generación del plan de proyecto.**

   En esta tarea se establece un plan para alcanzar los objetivos de la minería de datos, dicho plan debe especificar:

   - Los pasos a ser realizados durante el proyecto,
   - Selección de herramientas,
   - Selección de técnicas,
   - Selección de métodos.

b) Comprensión de los datos

En esta etapa inicia el contacto con los datos, donde se identificará la calidad, así como las relaciones evidentes, con los cuales se podrán establecer las primeras hipótesis. Es importante crear una nueva base de datos para el desarrollo del proyecto de minería de datos debido a que durante el desarrollo es posible que existan múltiples accesos a la base de datos con el fin de realizar consultas y probablemente se produzcan modificaciones. Esta esta etapa cuenta con cuatro tareas fundamentales que se muestran en la Figura 20.

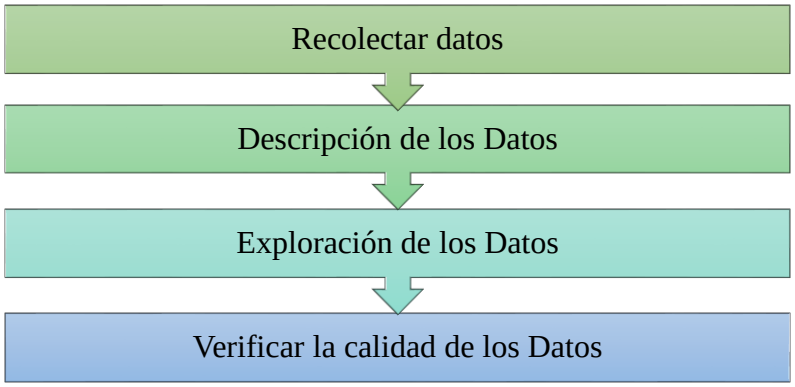

Figura 20. Tareas de la Etapa: Comprensión de los Datos

1. **Recolectar datos.**

   Se recolectan los datos iniciales del proyecto, así como también se realiza la adecuación para el procesamiento de los mismos. Se elaborará una lista con los datos obtenidos con la siguiente información:

   - Localización,
   - Técnicas utilizadas en la recolección,
   - Problemas encontrados y soluciones alcanzadas (evitar problemas similares a futuro).

2. **Descripción de los datos.**

   Una vez obtenidos los datos se procede al proceso de descripción, en el cual se incluyen datos como:

   - Formato de los datos,
   - Volúmenes de datos,
   - Significado e identificadores de campos,
   - En otros datos que sean considerados relevantes para el experto.

3. **Exploración de los datos.**

   En la exploración se los datos se encontrará la estructura de los datos, para esta tarea se hace uso de pruebas básicas estadísticas para revelar las propiedades de los datos obtenidos en las tareas anteriores. Esta tarea tiene como entregable un informe de exploración de los datos.

4. **Verificar la calidad de los datos.**

   Esta tarea determinará la consistencia, es decir, si los datos que se obtuvieron son completos de los datos, la cantidad y distribución de valores nulos, valores fuera de rango que pueden desviar o nublar el proceso. Si los datos tienen problemas de calidad se deben listar las soluciones de dichos problemas.

c) Preparación de los Datos

La preparación que se realiza en esta etapa guarda relación con la técnica de minería de datos que se vaya a utilizar, a su vez esta etapa está relacionada con la etapa de modelado, porque en función del modelado los datos son procesados de distintas maneras, por esta razón las etapas de preparación y de modelado interactúan de forma permanente. Las cinco tareas que contiene esta se muestran en la Figura 21.

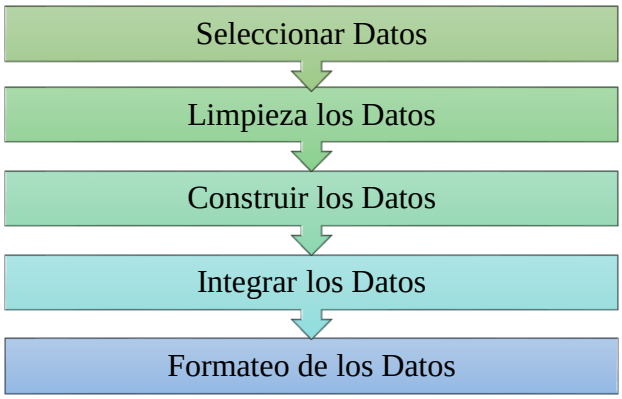

Figura 21. Tareas de la Etapa: Preparación de los datos.

1. **Seleccionar datos.**

   Se inicia con la selección de un subconjunto de datos que cumplan con los criterios de calidad de los adquiridos en las etapas anteriores, es importante tomar en cuenta que esta etapa cubre la selección de columnas o atributos y la selección de registros o filas, así como las restricciones en cuanto a cantidad

y/o volumen dependiendo de la técnica de minería de datos seleccionada

2. **Limpieza de datos.**

   Se mejora la calidad de los datos de tal forma que cumplan con el requisito establecido por la técnica de minería de datos seleccionada, esta tarea es la que más tiempo requiere, debido a que puede implicar la selección de subconjuntos de los subconjuntos previamente seleccionados o la inserción de nuevo datos. Algunas de las técnicas que se utilizan son:

   - La normalización de los datos.
   - Discretización de campos.
   - Tratamiento de valores faltantes.
   - Reducción del volumen de datos.

3. **Construir los datos.**

   En esta tarea se pueden generar derivados que son la fusión o el resultado de combinar dos atributos, unión de nuevos registros o transformación de atributos existentes.

4. **Integrar los datos.**

   Su objetivo consiste en la creación de nuevas estructuras, a partir de los datos seleccionados como:

   - Generación de nuevos campos a partir de otros existentes.
   - Creación de nuevos registros.
   - Fusión de tablas campo.
   - Nuevas tablas, etc.

5. **Formateo de los datos.**

   Esta tarea transforma sintácticamente los datos sin modificar su significado con el fin de facilitar el empleo de la técnica de minería de datos seleccionada, solamente si así lo requiere. Entre una de las transformaciones que pueden sufrir se tiene el reordenamiento de los registros.

d) Modelado

Se elige la técnica de modelado de minería de datos más apropiada para el proyecto la cual debe estar basada en los siguientes criterios:

- Ser apropiada al problema.
- Disponer de los datos adecuados.
- Cumplir requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

En muchas ocasiones la técnica se selecciona en base a la herramienta que se haya seleccionado, además junto con la selección de la técnica de modelado se debe seleccionar una técnica de evaluación para el modelo generado. Las cuatro tareas con las que cuenta se muestran en la Figura 22.

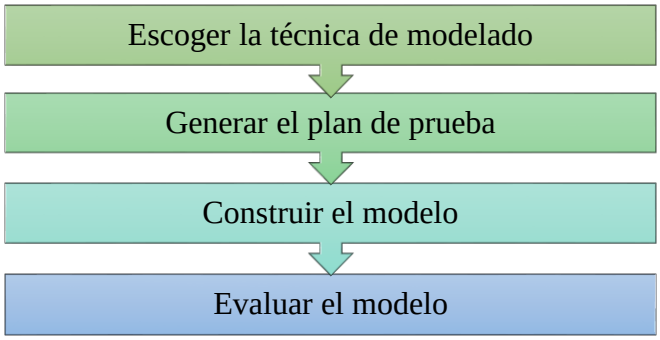

Figura 22. Tareas de la Etapa: Modelado.

1. **Escoger la técnica de modelado.**

   Esta esta tarea se selecciona la técnica de modelado más apropiada basándose en el objetivo del proyecto, las herramientas con las que se cuenta y de acuerdo a los criterios que establece esta etapa. Algunas de las técnicas se detallan en el punto 3.1. del presente documento.

2. **Generar el plan de prueba.**

   El plan generado en esta tarea se encargará específicamente de comprobar la calidad y validez del modelo una vez que este esté construido.

3. **Construir el modelo.**

Utiliza los datos previamente preparados a los cuales se les aplica técnica de minería de datos seleccionada para lograr la generación de un modelo el cual requiere distintos parámetros dependiendo la de técnica de minería de datos seleccionada. Esta tarea tiene como salida un informe de la interpretación y el rendimiento del modelo generado.

4. **Evaluar el modelo.**

En esta tarea los expertos en minería de datos son los encargados de la evaluación en base a todo lo antes establecido como son los objetivos del proceso de minería de datos y criterios de éxito, además los expertos podrán evaluar en base a criterios que ellos consideren que son relevantes para dicha evaluación.

h) Evaluación

La evaluación se hace verificando el cumplimiento de los criterios preestablecidos, si el modelo es generado resulta válido se procede a la explotación del mismo y a la vez a interpretarlo, para lo cual se sugiere emplear herramientas para la interpretación de resultados, en el caso que el modelo no sea válido se debe revisar el proceso para identificar el paso en dónde se cometió el error, además se debe tomar en cuenta que la fiabilidad calculada se aplica únicamente para los datos con los cuales se haya trabajado. Las cuatro tareas con las que cuenta se muestran en la Figura 23.

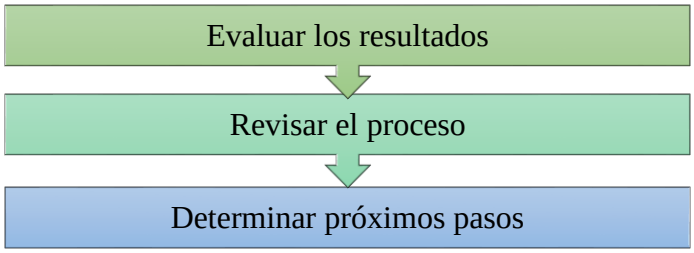

Figura 23. Tareas de la Etapa: Evaluación

1. **Evaluar los resultados.**

   Esta evaluación se realiza basándose en los objetivos del negocio y busca determinar si en algún factor el modelo generado no cumple con la expectativa propuesta. Si se cuenta con las condiciones y necesarias y el tiempo suficiente probar el modelo en tiempo real.

2. **Revisar el proceso.**

   Esta tarea se ejecuta una vez que el modelo se aceptado como válido, se revisa todo el proceso de la minería de datos para verificar si hay factores que no se hayan tomado en cuenta pero que sin embargo son relevantes en el proceso o identificar si hay procesos que se pudieren mejorar para obtener resultados más fiables.

3. **Determinar próximos pasos.**

   Como se mencionó en la descripción de esta metodología existen etapas que son bidireccionales, es decir, que si el proceso no ha sido satisfactorio se puede elegir a que etapa regresar para la verificación de errores, caso contrario se podrá continuar con la siguiente etapa.

i) Despliegue

Con un modelo construido y validado, el experto debe transformar el modelo en acciones o decisiones dentro de la organización, como por ejemplo puede aplicar el modelo a distintos conjuntos de datos. Esta etapa igual que las anteriores debe documentarse y entregar resultados de forma clara al usuario, con el fin de incrementar el conocimiento. Las cuatro tareas con las que cuenta se muestran en la Figura 24.

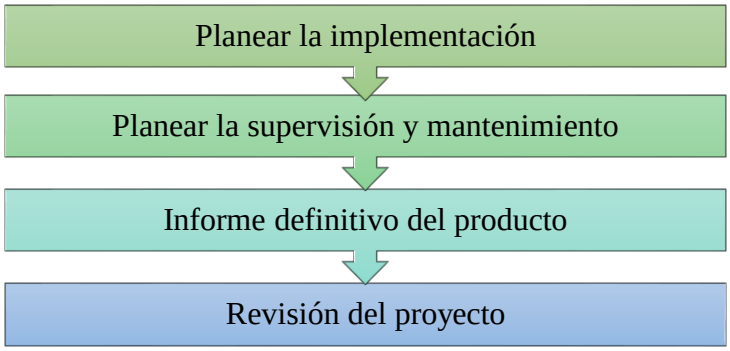

Figura 24. Tareas de la Etapa: Despliegue

1. **Planear la implementación:** Esta tarea se basa en la etapa de la evaluación concluyendo con las estrategias de la implementación.
2. **Planear la supervisión y mantenimiento:** Se deben establecer las estrategias para supervisar y mantener al modelo generado con el fin de evitar lapsos de tiempo de uso erróneo del modelo generado como producto de la minería de datos.
3. **Informe definitivo del producto:** Contiene las conclusiones del proceso de minería de datos junto con los resultados alcanzados, también deben incluirse los aciertos y desaciertos y finalmente cosas que se pueden mejorar.
4. **Revisión del proyecto:** Se realiza una revisión total de proyecto con el fin de detectar posibles errores o mejorar.

CAPÍTULO III
DESARROLLO DEL PROYECTO

4.1. COMPRENSIÓN DEL NEGOCIO

En esta etapa se cumplirá con cada una de las cuatro tareas de las que consta esta primera fase, cuyo objetivo es comprender claramente el problema para determinar los objetivos del proyecto desde el punto de vista del negocio, para en el desarrollo del proyecto poder convertirlos en objetivos de la minería de datos traducido en un plan de proyecto.

4.1.1. DETERMINACIÓN OBJETIVOS DEL NEGOCIO

El objetivo del negocio para la Universidad es disminuir la problemática y justificación descrita en los puntos 1.2 y 1.3 respectivamente para mejorar el servicio de educación entregado a los alumnos, para lo cual se han planteado los siguientes objetivos:

- Identificar posibles causas que influyen en el rendimiento académico.
- Mejorar el rendimiento académico.
- Mejorar el servicio de educación pública entregada a los alumnos.

Criterios de éxito del negocio

Desde enfoque de minería de datos se tiene como criterio de éxito la probabilidad de predecir la deserción estudiantil con el menor grado de error, para nuevos alumnos, de tal forma que se puedan dar y crear nuevas medidas de ayuda, como segundo criterio de éxito del negocio se tiene la posibilidad de incrementar el porcentaje de alumnos aprobados en las materias con mayor índice de deserción.

a) Evaluación de la Situación

Los recursos disponibles con los que se cuenta para el desarrollo del proyecto se sub divide en varias líneas las cuales se detallan a continuación:

- Personal: El proyecto es viable operativamente debido a que el personal que va a desarrollar la investigación va estar conformado por dos alumnos, las señoritas Katherine Irina Avalos Serrano y Sandra Elizabeth Paguay Flores. Además, se cuenta con varios interesados dentro de la Universidad de las Fuerzas Armadas "ESPE" como son las autoridades, las facultades, los docentes y los alumnos. Finalmente se cuenta con la tutoría del Ingeniero Mario Giovanny Almache Cueva, director de tesis del presente proyecto y docente de la Facultad de Ingeniería en Sistemas e Informática.

- Datos: La Unidad de Tecnologías de la Información UTIC's de la Universidad de las Fuerzas Armadas "ESPE", dispone del sistema de gestión de alumno y docentes BANNER desde el año 2010 hasta la actualidad, por lo que a priori se puede afirmar que se dispone de una cantidad de datos necesarios para poder cumplir con los objetivos planteados. Para el desarrollo del proyecto de investigación se utilizará los datos obtenidos de la base de datos del periodo Marzo 2011- Marzo 2016 de los alumnos de pre-grado presencial. La información con la que se cuenta en dicha base de datos de manera general es:

  - Información personal del alumno. (edad, sexo, residencia, colegio, discapacidad, etc.).
  - Registro de notas finales por materia de cada alumno durante cada período.
  - Registro de horarios por materia durante cada período.
  - Registro de docentes que impartieron la materia durante cada período.

- Recursos Hardware y Software: El proyecto es viable técnicamente debido a que los recursos de TI necesarios para el desarrollo y la implementación

están a disposición del grupo de proyecto debido a que tienen licencia GPL como se evidencia en la Tabla 2:

Tabla 2.
Factibilidad Técnica

	Herramienta	Licencia
SOFTWARE	Weka v3.8	GPL
	Oracle 11G R2	
	SqlDeveloper v4.1.5	
	Netbeans v7.4	GPL
	Glassfish v4.1	GPL
	Pentaho Data Integration v 7.0	GPL
	Open Refine v2.0	GPL
	R Studio v3.2.2	GPL
HARDWARE	Toshiba CORE i7 SATELLITE S855-S5381	NA
	Laptop Hp Pavilion dv6 Core i7	NA
	Impresora Epson	NA

- **Requisitos, supuestos y restricciones**

  Requisitos

  - Disponer de la autorización emitida por el Director de las UTIC's, el Ingeniero Rommel Asitimbay, para obtener la información requerida.
  - Disponer con una amplia gama de información necesaria para realizar minería de datos.
  - Contar con las tutorías de expertos en el área de minería de datos con la finalidad de generar un modelo de calidad que cumpla con los objetivos planteados.

Restricciones

- Los datos para el desarrollo del proyecto se limitan a la información registrada dentro de las bases de datos de las UTIC's.

- **Riesgos y contingencias**
  - Si el proyecto dura más de lo programado es necesario que se actualice el cronograma inicial para poder optimizar el tiempo y cumplir con las tareas propuestas.
  - Si los datos obtenidos son de escasa calidad y no cubren con las expectativas para obtener modelos de predicción precisos.

- **Análisis de costes/beneficios**

El coste del proyecto oscila en un monto de $4692.00 los cuales se desglosa a continuación en la Tabla 3.

Tabla 3.
Factibilidad Real

HARDWARE - NO APLICA	
Software	
Weka v3.8	Open Source
Oracle 11G R2	$0 (Licencia Estudiantil).
Netbeans v7.4	Open Source
Glassfish v4.1	Open Source
Pentaho Data Integration v 7.0	Open Source
Open Refine v2.0	Open Source
R Studio v3.2.2	Open Source
Operativo	
2 Ingenieros en Sistemas	$366 * 6 meses = $2196 * 2 Personas= $4392
Artículos de Oficina	$300
Total	$4692

Debido a que el proyecto de investigación se lo realiza con un propósito estudiantil, se cuenta a disposición con el personal para realizarlo, por lo que el coste

del proyecto oscila en un monto de $300.00 los cuales se desglosa a continuación en la Tabla 4:

Tabla 4.
Factibilidad Proyecto

Software	
Weka v3.8	Open Source
Oracle 11G R2	$0 (Licencia Estudiantil).
Netbeans v7.4	Open Source
Glassfish v4.1	Open Source
Pentaho Data Integration v 7.0	Open Source
Open Refine v2.0	Open Source
R Studio v3.2.2	Open Source
Operativo	
Artículos de Oficina	$300
Total	$300

b) Determinación objetivos de la minería de datos

Los objetivos en términos de minería de datos son:

- Identificar las variables que influyen con mayor frecuencia en la deserción de alumnos.
- Identificar las asignaturas en las que los alumnos desertan con mayor frecuencia.
- Refinar, limpiar los datos recopilados y prepararlos para el modelado.
- Seleccionar y comparar los resultados de las técnicas de minería de datos que genere mejores resultados para la predicción de deserción estudiantil.
- Encontrar el comportamiento y patrones de los estudiantes desertores.

Criterios de éxito de minería de datos

Desde enfoque de minería de datos se tiene como criterio de éxito la predicción de la deserción estudiantil con el menor grado de error, para nuevos alumnos. El porcentaje de error en las predicciones se calculará a través del registro histórico (casos de prueba).

c) Generación del plan de proyecto

Ver Anexo 1: Cronograma del proyecto

4.1.2. COMPRENSIÓN DE LOS DATOS

Esta segunda etapa comprende la recolección inicial de los datos con el objetivo de tener el primer contacto con los mismos y con los que se trabajará, además implica estudiar más de cerca los datos disponibles para la minería accediendo a los datos y explorarlos con la ayuda de tablas y gráficos. También se identificará la calidad de los mismos, con los cuales se podrán establecer las primeras hipótesis. La fase de comprensión de datos es importante para evitar posibles problemas en la fase de preparación de datos.

a) Recolectar Datos

Los datos han sido recolectados completamente del Sistema BANNER con la ayuda de la Ingeniera Anita Torres que forma parte de la Unidad de Tecnología de la Información, dichos datos se encuentran registrados en varias tablas de una misma base de datos.

1. La recolección de datos se realizó a través de la herramienta SQL Manager para Oracle 2007 para acceder a la base de datos Oracle.
2. La información que se obtuvo con los distintos querys se exportó a un formato xlsx.
3. Los datos se exportaron en dos distintos archivos .xlsx, el primero que contiene una tabla con la información personal de todos los alumnos durante los períodos solicitados, el segundo contiene las notas del alumno por materia.
4. Los datos obtenidos son únicamente de los períodos de Marzo 2010 – Octubre 2016.

b) Descripción de los datos

En las Tablas 5 y 6 que se presenta a continuación se detallan los atributos para las tablas Alumnos y Notas.

Tabla 5.
Variables de la Tabla Alumnos

TABLA	DESCRIPCIÓN	N ° REGISTROS
Alumnos	Esta tabla contiene la información personal de los alumnos.	16220
CAMPO	**DESCRIPCIÓN**	
CEDULA	Cédula del alumno.	
NOMBRES	Nombres y Apellidos del alumno.	
ETNIA	Conjunto de comunidad lingüística, cultural o raza al que pertenece el alumno.	
EDAD	Edad de vida del alumno.	
ESTADO_CIVIL	Condición del alumno según el registro civil en función de si tiene o no pareja y su situación legal	
GÉNERO	Género del alumno.	
DIRECCIÓN	Calles de residencia del alumno.	
DISCAPACIDAD	Discapacidad del alumno, variable con valor SI o NO.	
TIPO_DISCAPACIDAD	Tipo de Discapacidad del alumno será un valor vacío en caso de tener una discapacidad.	
CARRERA	Carrera que sigue el alumno.	
PERIODO_COHORTE	Período en que el alumno ingresa a primer nivel de la carrera.	
PERIODO_INGRESO	Período en que el alumno ingresa a Prepolitécnico.	
NACIONALIDAD	País de nacimiento del alumno.	
PROVINCIA_NAC	Provincia de nacimiento del alumno.	
CANTON_NAC	Cantón de nacimiento del alumno.	
PROVINCIA_RES	Provincia de residencia del alumno.	
CANTON_RES	Cantón de residencia del alumno.	
PARROQUIA_RES	Parroquia de residencia del alumno.	
MILITAR	Variable con valor SI o NO en caso de que el alumno pertenezca a las Fuerzas Armadas.	
COLEGIO	Colegio en el que se graduó el alumno.	
PROMEDIO_COLEGIO	Promedio con el que se graduó el alumno.	
INGRESOS_FAMILIARES	Ingresos familiares del alumno en dólares.	

Tabla 6.
Variables de la Tabla Nota

TABLA	**DESCRIPCIÓN**	**NÚMERO DE REGISTROS**
Nota	Información de las notas de los alumnos por cada materia	604893
CAMPO	**DESCRIPCIÓN**	
SMRPRLE_PROGRAM_DESC	Nombre de la carrera a la que pertenece.	
CEDULA	Cédula del alumno	
NOMBRES	Nombres y apellidos de los alumnos	
PERIODO	Período en el que el alumno tomó la materia	
CODIGO_MATERIA	Código de la materia	
MATERIA	Nombre de la materia	
NOTA	Nota con la cual el alumno terminó la materia	
COMENTARIO	Indica si el alumno aprobó o reprobó la materia	
HORA_INICIO	Hora de inicio de la clase de la materia	
HORA_FIN	Hora de fin de la clase de la materia	
LUNES	Selecciona si el horario de la materia fue impartido en este día	
MARTES	Selecciona si el horario de la materia fue impartido en este día	
MIERCOLES	Selecciona si el horario de la materia fue impartido en este día	
JUEVES	Selecciona si el horario de la materia fue impartido en este día	
VIERNES	Selecciona si el horario de la materia fue impartido en este día	
SABADO	Selecciona si el horario de la materia fue impartido en este día	
DOMINGO	Selecciona si el horario de la materia fue impartida en este día	
DOCENTE	Docente que dictó dicha materia al alumno	

c) Exploración de los datos

A continuación, se realizó el análisis estadístico de los datos descritos en el punto anterior para revelar las características de los datos demográficos y académicos y obtener gráficos de la distribución de los datos, dónde el tamaño de la muestra fue de 8264 alumnos de la Universidad de las Fuerzas Armadas – ESPE durante los últimos 5 años que equivale desde el período Septiembre 2010-Enero 2011 al período Octubre 2016 – Febrero 2017 de los alumnos cuyo período de cohorte fue desde el 2011 al 2016. En la Figura 25 se observa que del total de alumnos el 54.28% corresponden a alumnos de género masculino y el 45.72% a género femenino existiendo una cifra relativamente equitativa entre hombre y mujeres.

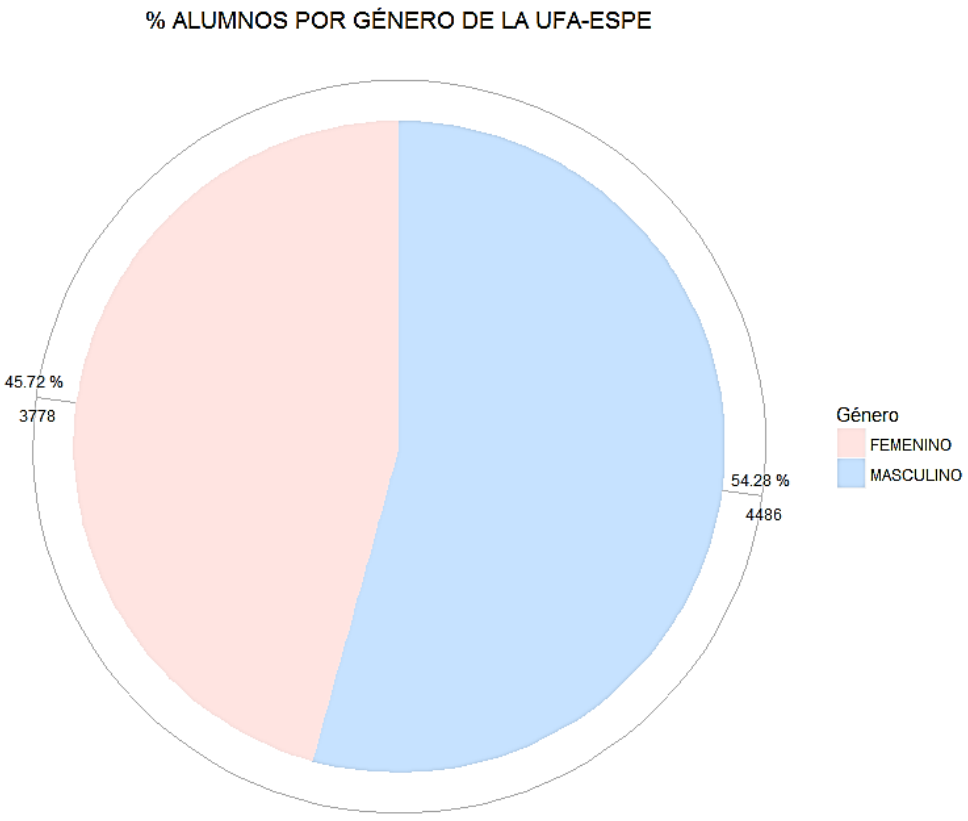

Figura 25. Porcentaje de Alumnos por género de la Universidad de las Fuerzas Armadas – ESPE

El rango de edad donde existe mayor población de alumnos con el 71.09% es entre 20 y 25 años seguido de los alumnos entre 17 y 20 años como se puede observar en la Figura 26, por lo general las personas de mayor edad son militares o alumnos que retomaron sus estudios después de haberse retirado o a su vez alumnos que solo egresaron que por lo general están obligados a matricularse en el Plan de Actualización de Conocimientos.

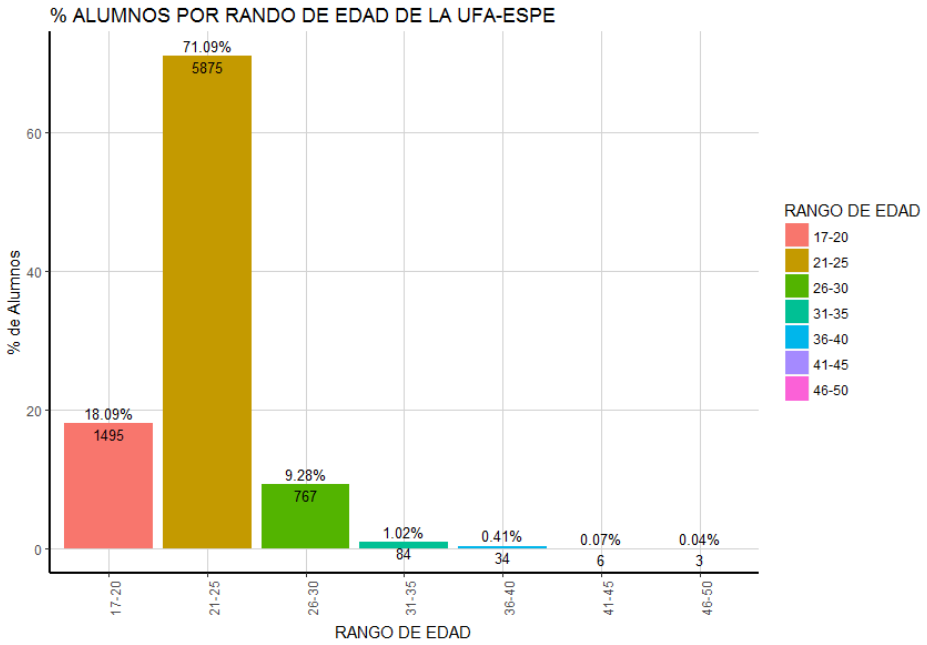

Figura 26. Cantidad de alumnos distribuidos por rango de edad de la Universidad de las Fuerzas Armadas – ESPE

En la Figura 27 se observa que existe una cantidad mínima de alumnos extranjeros, tan solo 25 alumnos del total de 8264, es decir el 0.30% de la población.

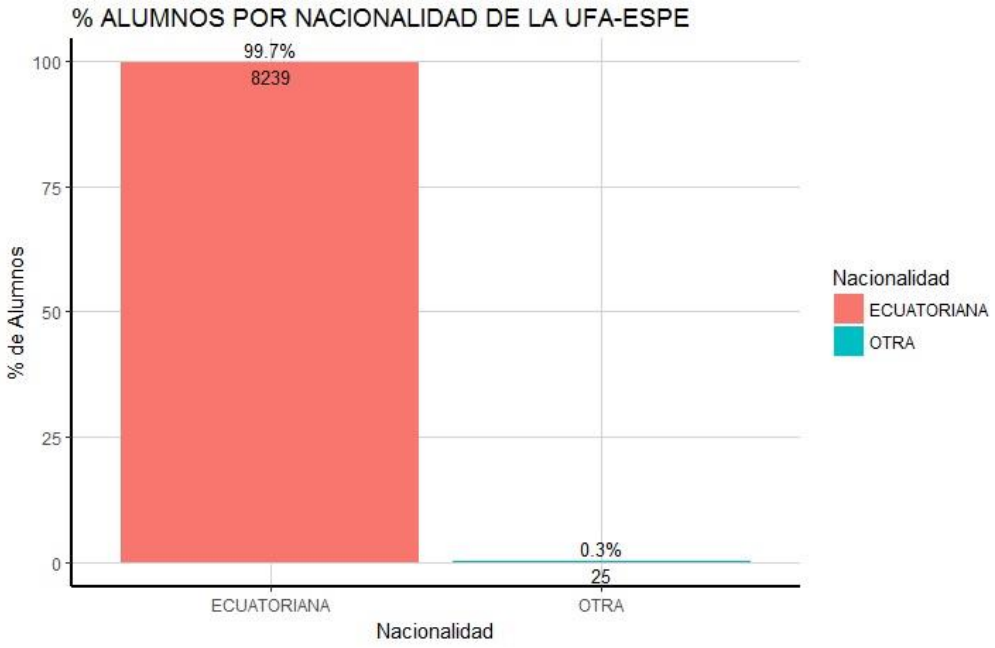

Figura 27. Porcentaje de alumnos por nacionalidad de la Universidad de las Fuerzas Armadas – ESPE

De los 25 alumnos extranjeros como se observa en la Figura 27, la mayor parte (16 alumnos) no se conoce su procedencia, sin embargo, se conoce que el 16% (4 alumnos) son colombianos mientras que el resto (5 alumnos) se divide equitativamente entre las demás nacionalidades como se observa en la Figura 28.

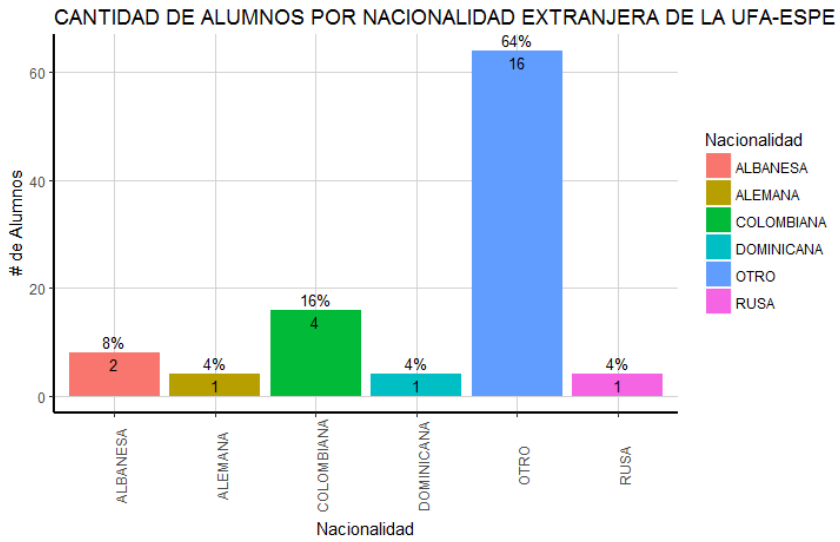

Figura 28. Cantidad de alumnos por nacionalidad extranjera de la Universidad de las Fuerzas Armadas - ESPE

Para analizar la provincia de procedencia se ha tomado la provincia del colegio de donde proviene el alumno, debido a que la provincia de nacimiento no es una variable real que indique de donde provienen los alumnos que ingresan a la Universidad de las Fuerzas Armadas - ESPE. Como se observa en la Figura 29, del total de 8264 alumnos se ha reducido la muestra a 8261 datos debido a que el resto no registraban el colegio de procedencia, de los cuales en Pichincha han ingresado 6900 alumnos, siendo la provincia con mayor número de alumno, seguido por Imbabura y Tungurahua con 253, y 179 alumnos respectivamente. Mientras que las provincias con menor incidencia son Zamora Chinchipe, Cañar, Galápagos y Santa Elena con 2, 5, 8 y 8 alumnos respectivamente.

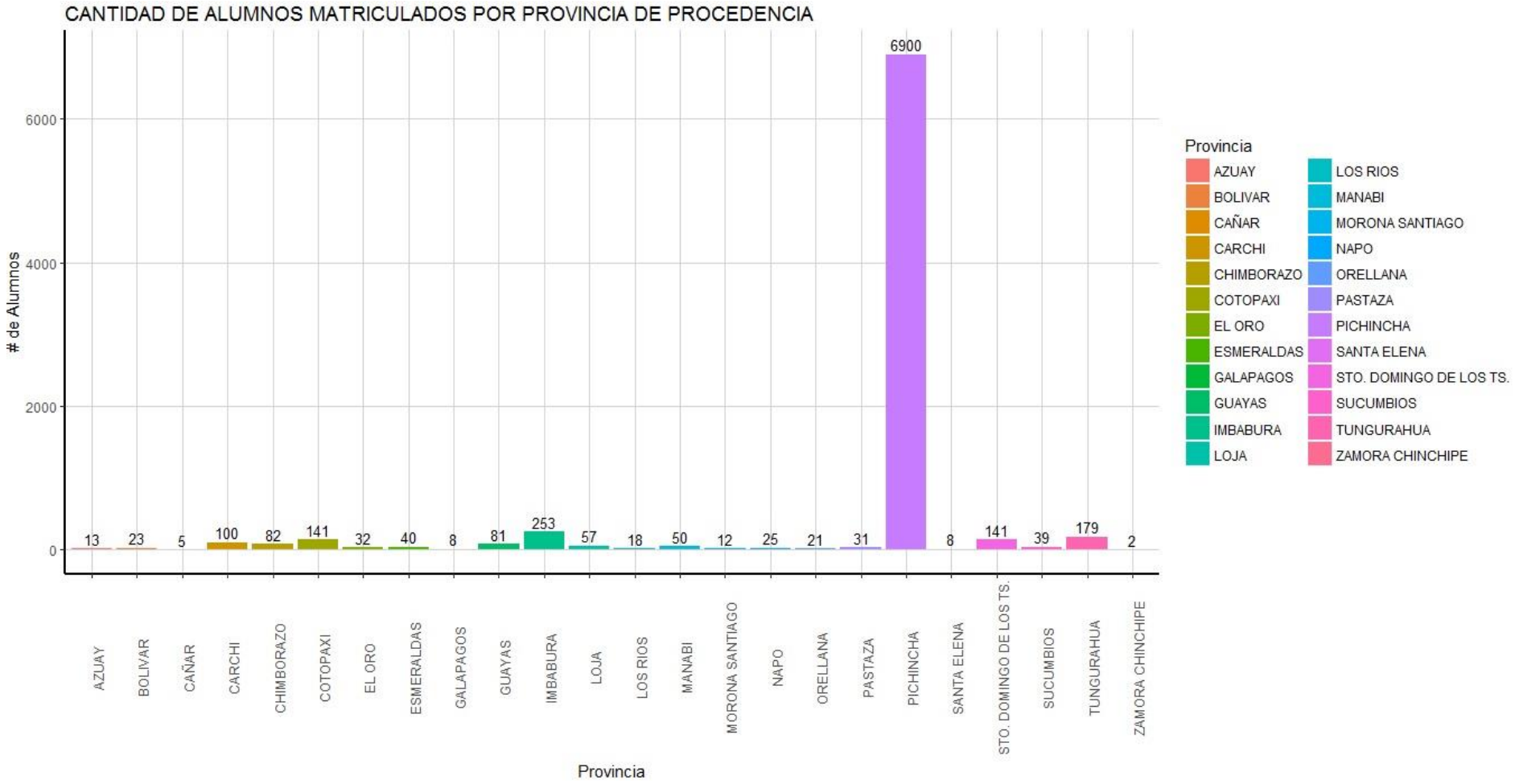

Figura 29. Cantidad de alumnos por provincia de procedencia.

Existe una gran diferencia entre alumnos mestizos en comparación con otras etnias como se observa en la Figura 30 del total de 8264 alumnos, el 96,9% de la población (8026 alumnos) son mestizos, mientras que el resto (238 alumnos) de se divide en Afro-Ecuatorianos, Blancos, Indígenas, Montubios, Mulatos y Otras etnias.

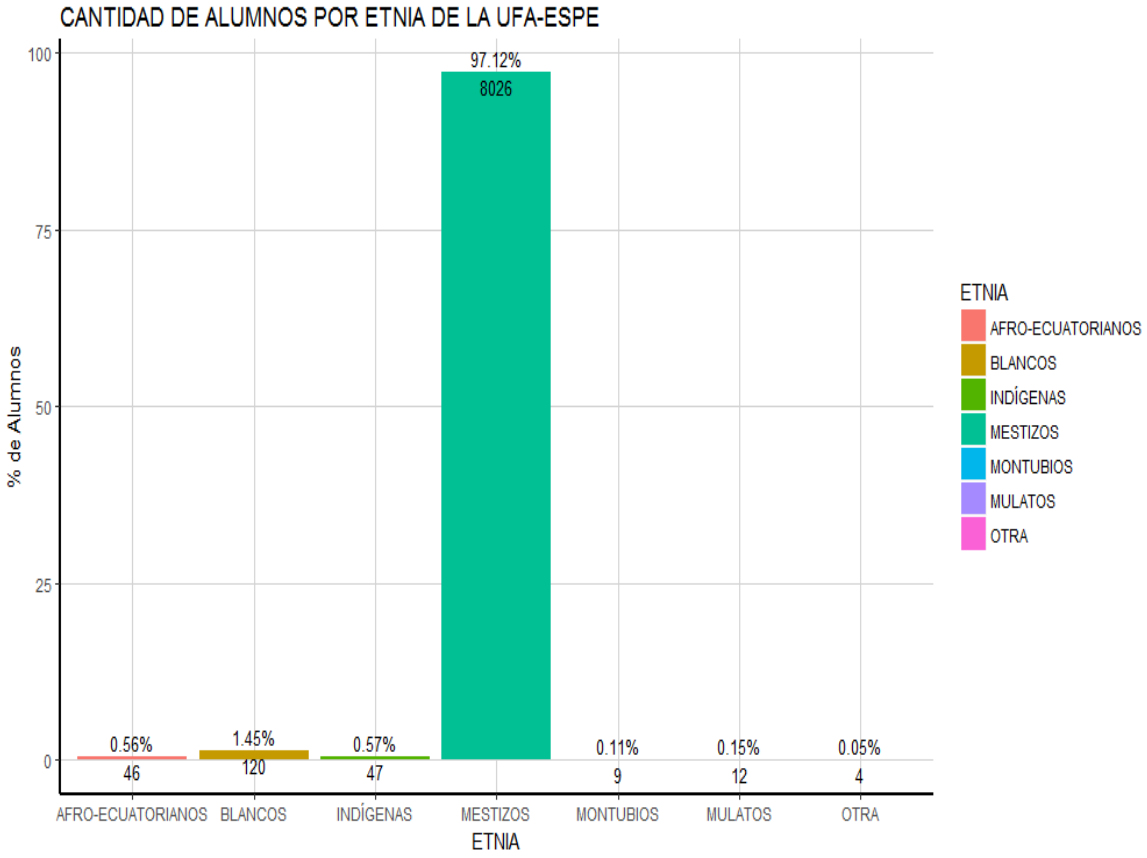

Figura 30. Cantidad de alumnos por etnia de la Universidad de las Fuerzas Armadas-ESPE

En la Figura 31. se observa que la mayor parte de alumnos de la Universidad de las Fuerzas Armadas – ESPE son solteros, el 96.8% de la población seguido por un mínimo de alumnos casados con 2.64%, existiendo una diferencia sustancial de 7773 alumnos entre solteros y casados, mientras que una mínima cantidad de alumnos se divide en los otros tipos de estado civil.

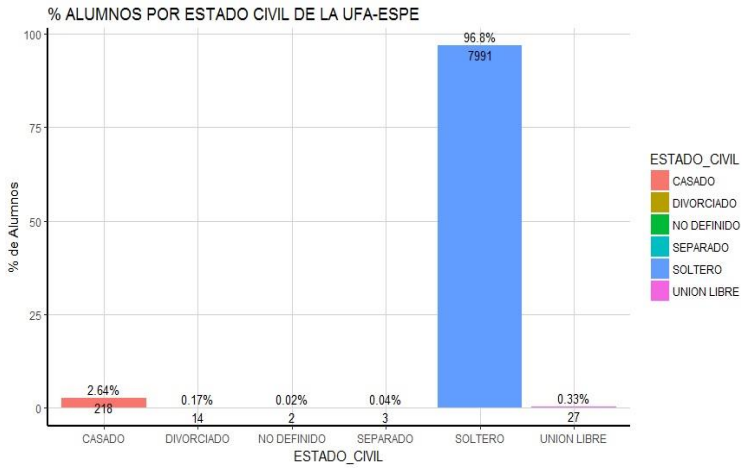

Figura 31. Porcentaje de alumnos por estado civil de la Universidad de las Fuerzas Armadas – ESPE

En la Figura 32 se observa que la inclusión de personas con discapacidad dentro de la Universidad de las Fuerzas Armadas – ESPE es mínima, tan solo el 0.25% de la población (21 alumnos) durante los últimos 5 años.

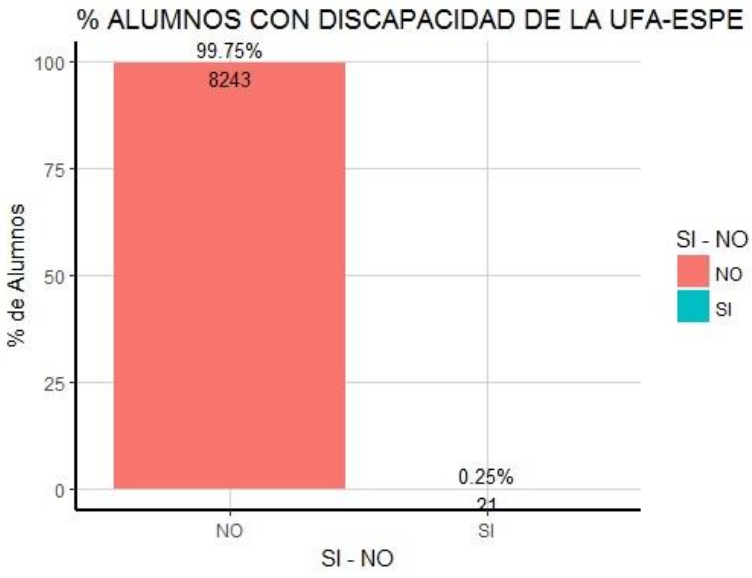

Figura 32. Porcentaje de alumnos con Discapacidad de la Universidad de las Fuerzas Armadas – ESPE

De los 21 alumnos con discapacidad existen 5 tipos de dificultades: auditivas, físicas, motoras, psicológicas y visuales como se muestra en la Figura 33. donde la mayoría, el 52,38% (11 alumnos) de los 21 alumnos, tienen dificultades físicas.

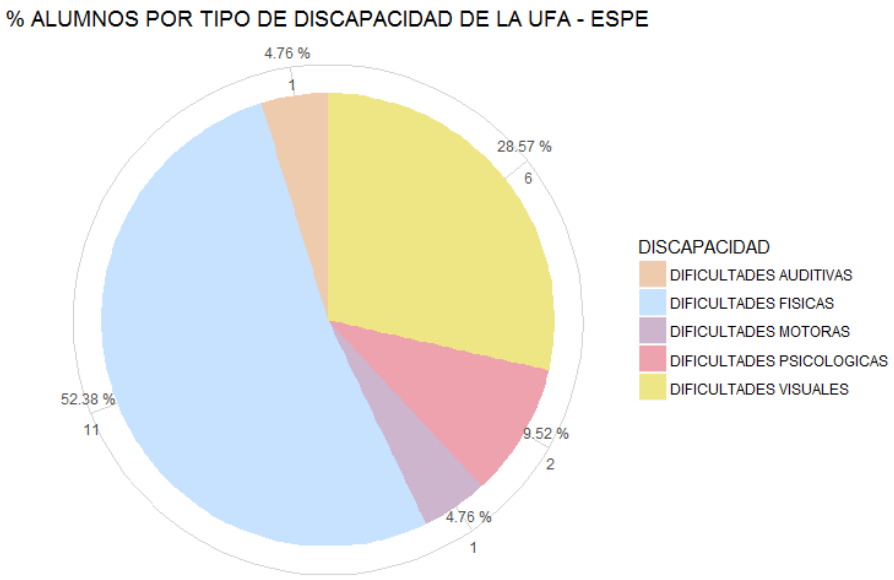

Figura 33. Porcentaje de alumnos por tipo de discapacidad de la Universidad de las Fuerzas Armadas – ESPE

La Universidad de las Fuerzas Armadas – ESPE está conformada en su mayoría por alumnos civiles 98,29 % (14983 alumnos) de 15244 alumnos, mientras que un mínimo porcentaje 1.71% (261 alumnos) son alumnos militares como se indica en la Figura 34.

En la Universidad de las Fuerzas Armadas – ESPE han ingresado alumnos de 807 colegios diferentes, que según su régimen de estudios la mayoría provienen del régimen Sierra, es decir 7755 alumnos, seguido por 368 alumnos del régimen Costa, 130 alumnos del régimen Oriente y con 8 alumnos del régimen Insular como se encuentra en la Figura 35.

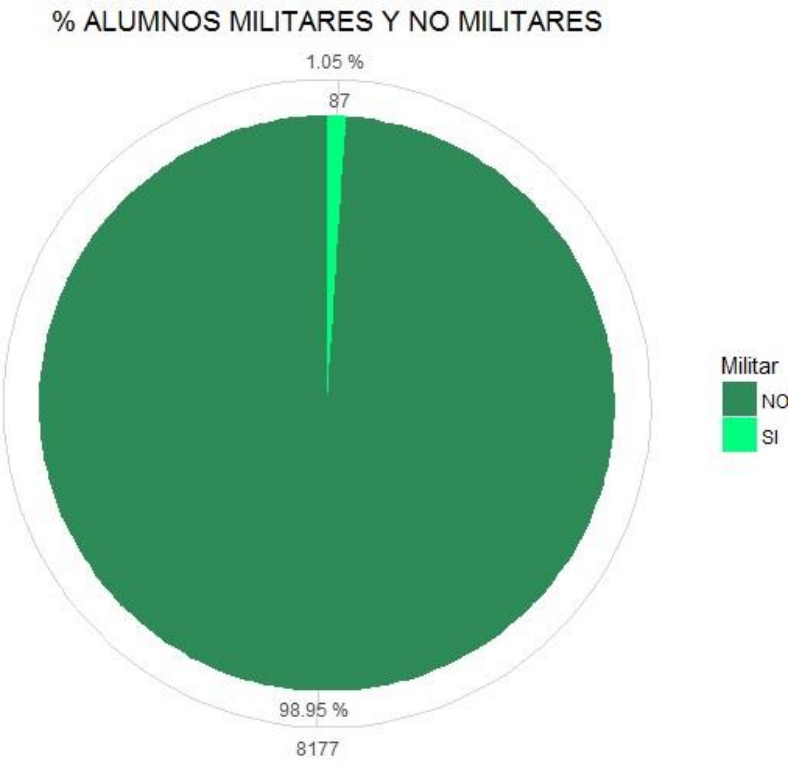

Figura 34. Porcentaje de alumnos civiles y militares de la Universidad de las Fuerzas Armadas – ESPE

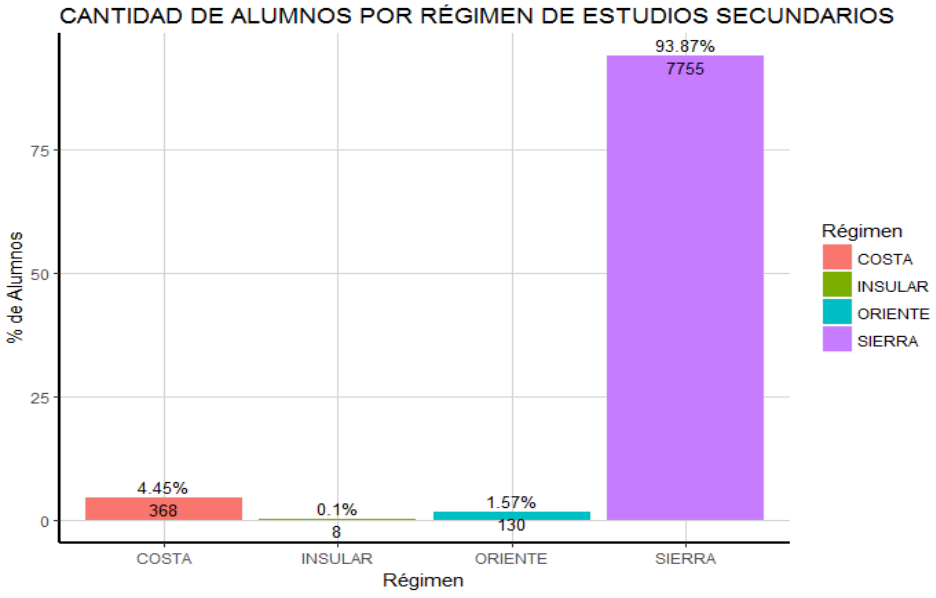

Figura 35. Cantidad de Alumnos por Régimen de Estudios Secundarios de la Universidad de las Fuerzas Armadas - ESPE

De los 973 colegios de donde provienen los alumnos, la mayoría son de colegios donde se entrega cierta cantidad de dinero por la educación, como son colegios: Particulares, Municipales y Fiscomisionales con un 52,8% de los alumnos (4362 alumnos), mientras que 3899 alumnos provienen de colegios fiscales, siendo el 47.2 % del resto de la población como se puede apreciar en la Figura 36.

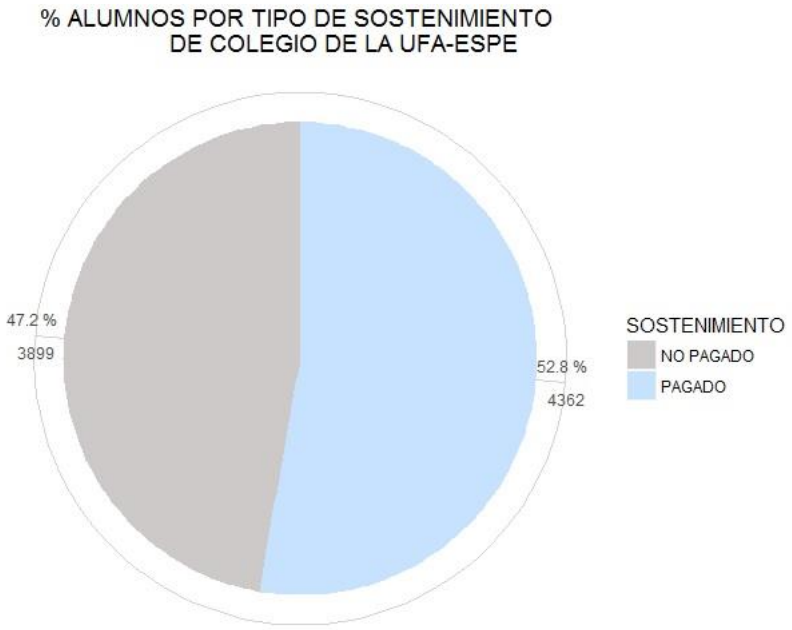

Figura 36. Porcentaje de Alumnos por tipo de sostenimiento del colegio de procedencia

De todos los colegios de los cuales provienen los alumnos se puede apreciar en la Figura 37, los que encabezan la lista son el colegio “Juan Pio Montufar” y “24 de Mayo” dentro de los colegios con sostenimiento no pagado y los colegios “Fernandez Madrid” y “Sebastián de Benalcázar” dentro de los colegios con sostenimiento pagado, esto debido a que como se muestra en la Figura 29 la mayor cantidad de alumnos provienen de la provincia de Pichincha.

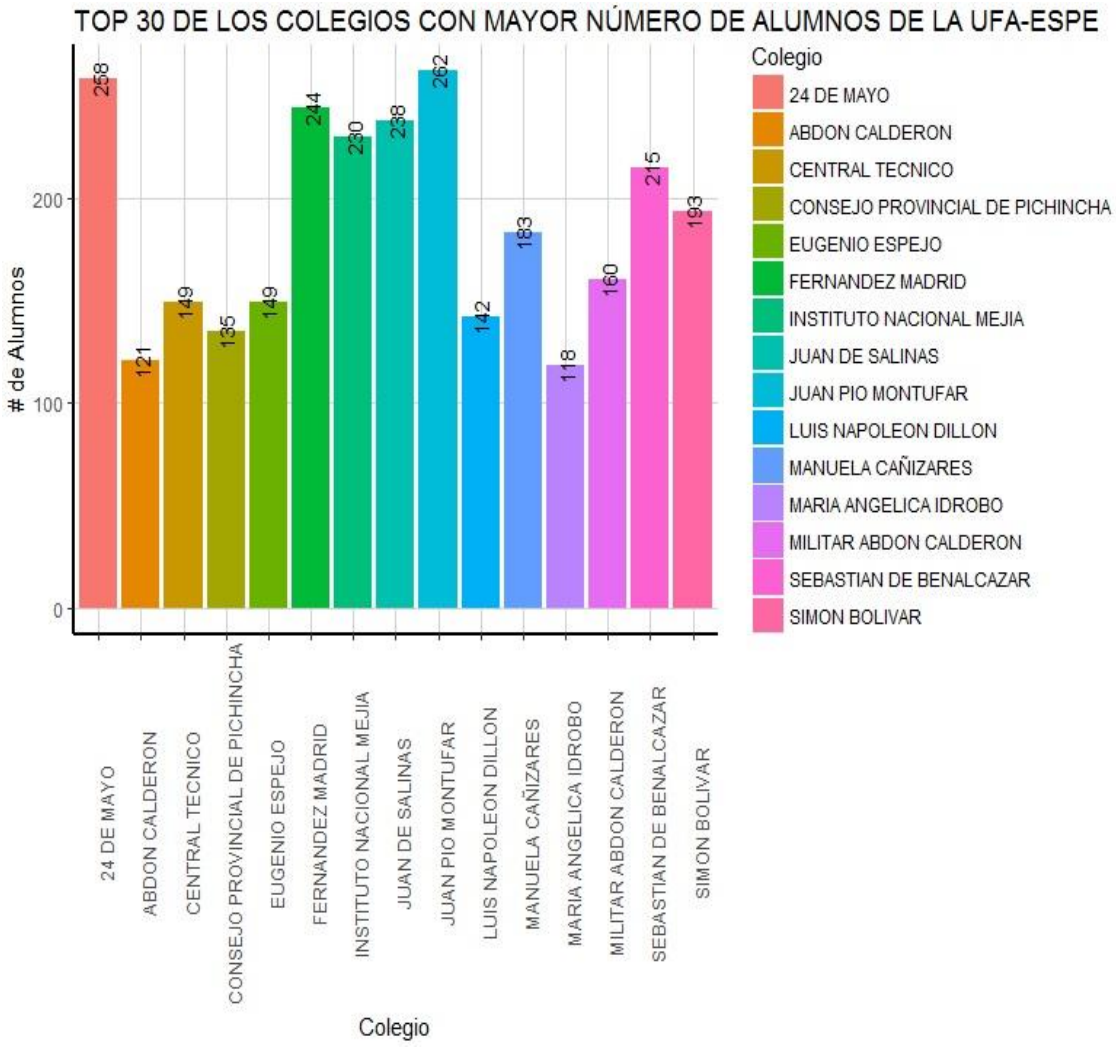

Figura 37. Cantidad de Alumnos por colegio que ingresaron a la Universidad de las Fuerzas Armadas - ESPE

De los 8264 alumnos tomados como muestra, en la Figura 38 se aprecia la cantidad de alumnos que ingresó a la Universidad de las Fuerzas Armadas - ESPE dividido en períodos académicos, donde el período 201310 (Septiembre 2012-Enero 2013) fue la primera promoción que ingresó con el examen del SENECYT , la cantidad de alumnos que ingresó disminuyó notablemente en relación a los períodos anteriores, sin embargo en los períodos posteriores la cantidad de alumnos aumentó de forma considerable, además se puede observar que en el año 2016 la cantidad de alumnos que ingresan a la universidad tiende a crecer superando a todos los períodos anteriores.

Figura 38 . Cantidad de alumnos matriculados en los últimos 5 años por período académico

En la Figura 39 se muestra el porcentaje de personas que ingresaron con el examen de la Universidad de las Fuerzas Armadas-ESPE y el examen del SENECYT, como se muestra el porcentaje que ingresó con el examen del SENECYT es mayor debido a que de los 5 años de estudio, los 3 últimos son períodos en donde los alumnos ingresaron con la prueba del SENECYT.

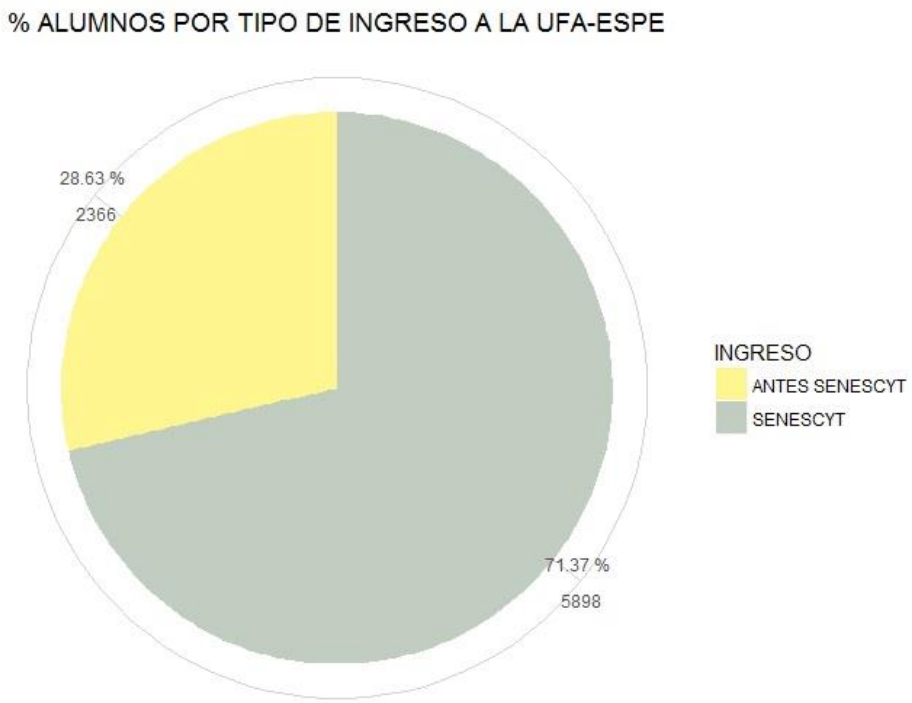

Figura 39. Porcentaje de alumnos por tipo de ingreso a la Universidad de las Fuerzas Armadas – ESPE

Del total de los alumnos tomados como muestra se dividió a las carreras en tres tipos: administrativas, humanísticas y técnicas dónde se pudo conocer que el tipo de carrera que más demanda tiene son las técnicas, alcanzando un 52.01%, así como lo muestra la Figura 40.

Las carreras que mayor demanda tienen son Ingeniería en Finanzas y Auditoría e Ingeniería Comercial que pertenecen al grupo de carreras administrativas, ocupando

el primero y segundo lugar respectivamente en la lista, mientras que en las carreras técnicas las que lideran son Ingeniería en Biotecnología y Electrónica Automatización y Control ocupando el tercero y cuarto lugar, como se puede apreciar en la Figura 41.

Figura 40. Porcentaje de alumnos por tipo de carrera de la Universidad de las Fuerzas Armadas - ESPE

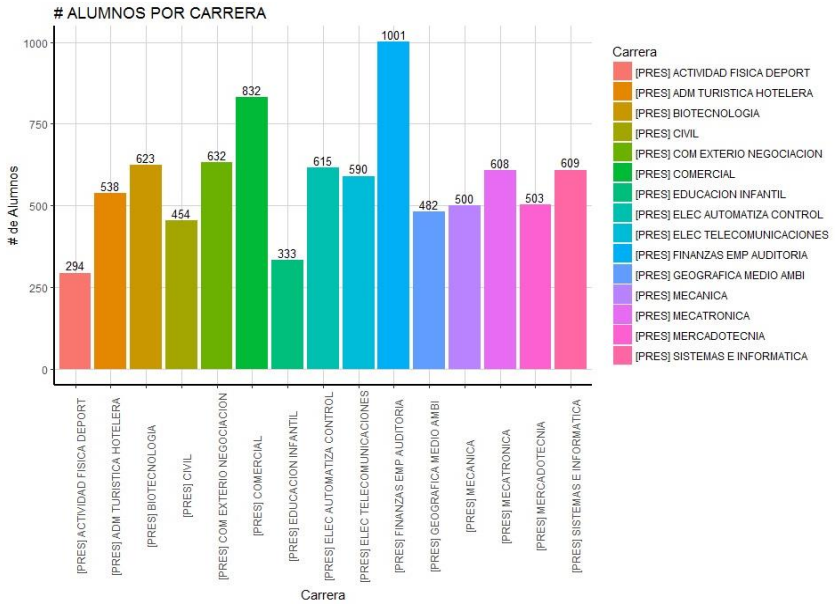

Figura 41. Cantidad de alumnos por carrera de la Universidad de las Fuerzas Armadas - ESPE

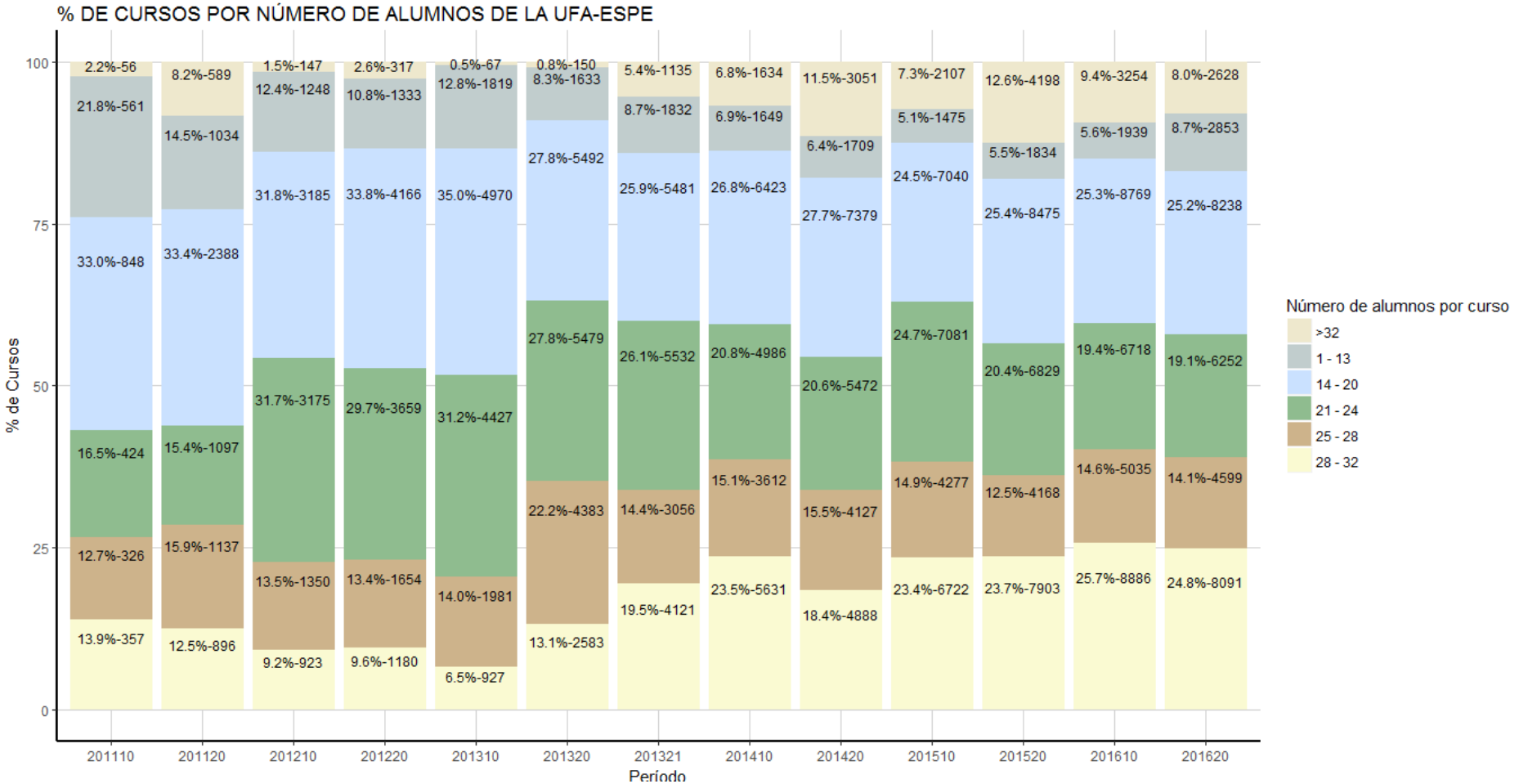

Figura 42. Porcentaje de cursos por rango de cantidad de alumnos

La Figura 42 muestra que en los años 2011 y 2012 la Universidad de las Fuerzas Armadas - ESPE creaba cursos en su mayoría con una cantidad entre 14 y 20 alumnos, pero a medida que han pasado los años y la Universidad se volvió pública, ha visto la necesidad de crear cursos con mayor número de alumnos que van de 21 a 28 alumnos por curso hasta que en los últimos dos años los cursos cuentan con más de 28 alumnos por curso, esto es debido al crecimiento de alumno que ingresan a la universidad que se presenta en la Figura 38

En la Figura 43 se puede observar el porcentaje de alumnos que han aprobado y reprobado de materias por género dónde se afirma que los alumnos de género masculino tienden a reprobar más, porque del 100% de hombres el 14.8% reprueba, mientras que del 100% de mujeres solo reprueba el 8.50%.

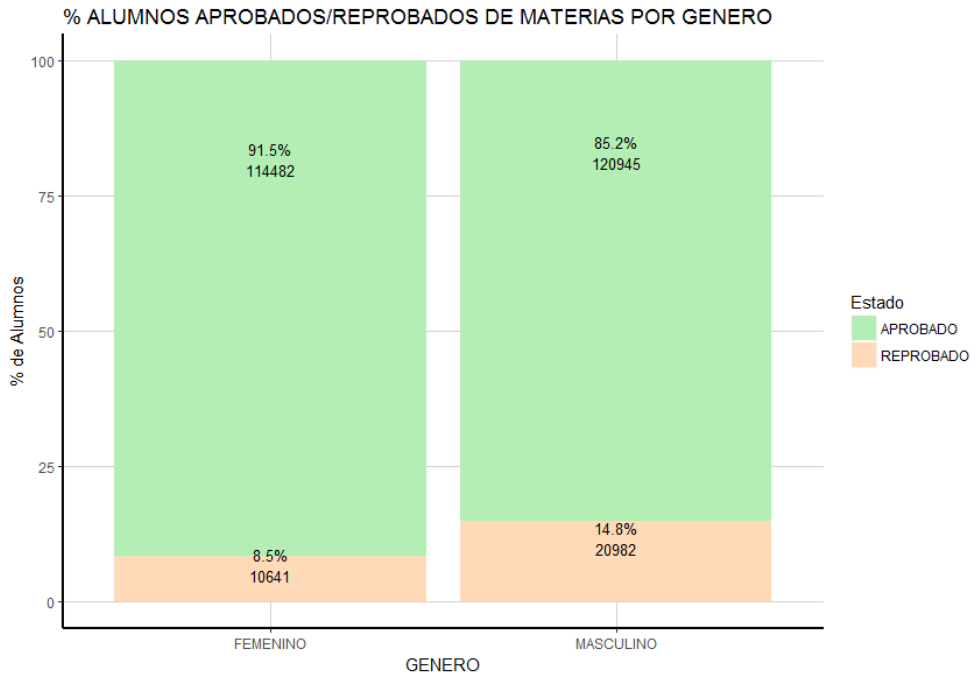

Figura 43. Porcentaje de alumnos que han aprobado y reprobado de materias por género.

Los rangos de edad en donde existe una cantidad mayor de aprobación de materias son los de 31 a 45 años siendo este el rango de edad del cual existe menor cantidad de alumnos (114 alumnos), mientras que los de 17 a 20 años tienden a tener un 85.1% que es un porcentaje menor de aprobación como se puede observar en la Figura 44.

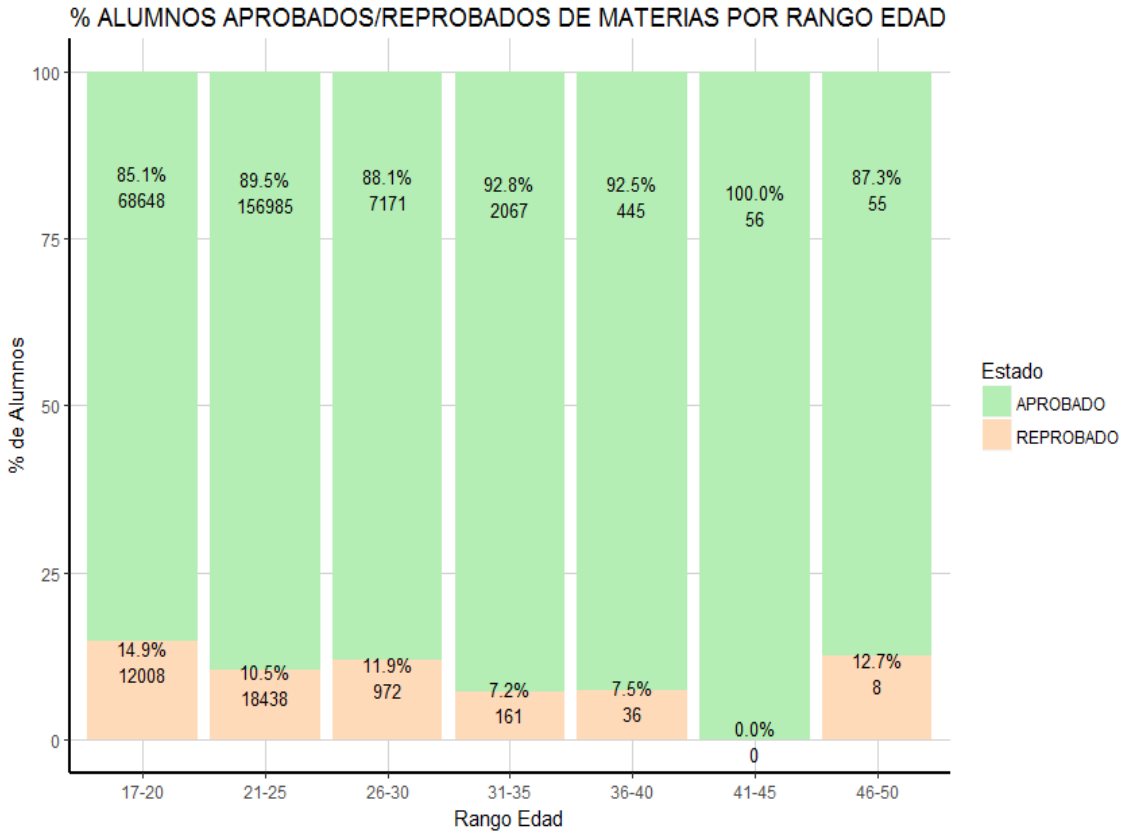

Figura 44. Porcentaje de alumnos que han aprobado y reprobado de materias por rango de edad.

Respecto a la nacionalidad del total de alumnos recogidos como muestra se separó a los de las distintas nacionalidades excluyendo a los de nacionalidad ecuatoriana como se muestra en la Figura 45, que del total de alumnos de nacionalidad colombiana reprueban el 21.7% siendo esta la nacionalidad con mayor cantidad de

alumnos como se muestra en la Figura 28, a esto le siguen las de una nacionalidad albanesa con el 66.7%.

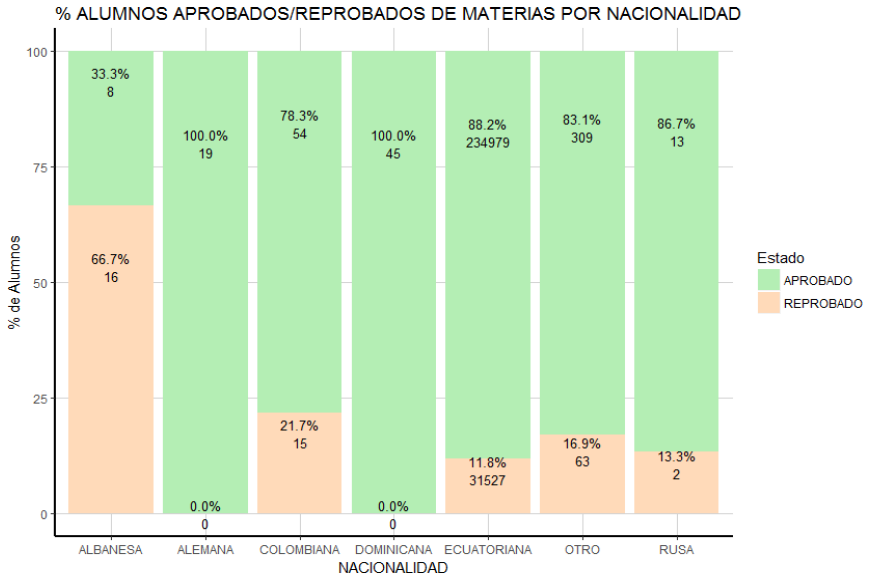

Figura 45. Porcentaje de alumnos que han aprobado y reprobado de materias de nacionalidad extranjera.

Del total de alumnos seleccionados como muestra, 6900 son de Pichincha y 1361 son se otra provincia, a su vez en la Figura 46 se observa que el porcentaje de aprobación de alumnos de otras provincias es 86.6% siendo este menor a los de la provincia de Pichincha que tienen el 88.5% teniendo una diferencia de alrededor del 2% en porcentaje de aprobación, a pesar de ello los alumnos de otras provincias tienden a reprobar con mayor frecuencia.

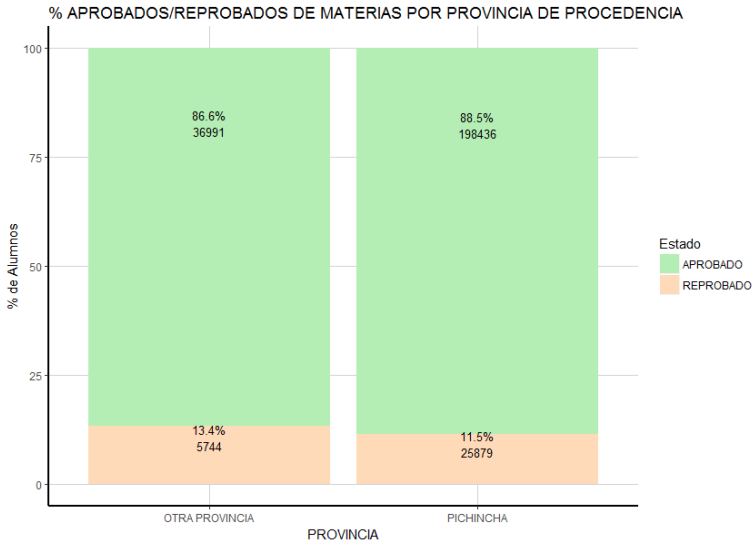

Figura 46. Porcentaje de alumnos que han aprobado y reprobado de materias por provincia de procedencia.

Las provincias Imbabura, Tungurahua, Sto. Domingo de los Tsáchilas y Cotopaxi que son las provincias distintas Pichincha que tienen mayor cantidad de alumnos se observa en la Figura 47 que su porcentaje de reprobación es del 14.2%, 14.3% , 12.9% y 11.6% respectivamente siendo valores relativamente similares entre ellos y en general para las demás provincias con excepción de las provincias de Galápagos y Zamora Chinchipe que registran un incremento de materias reprobadas.

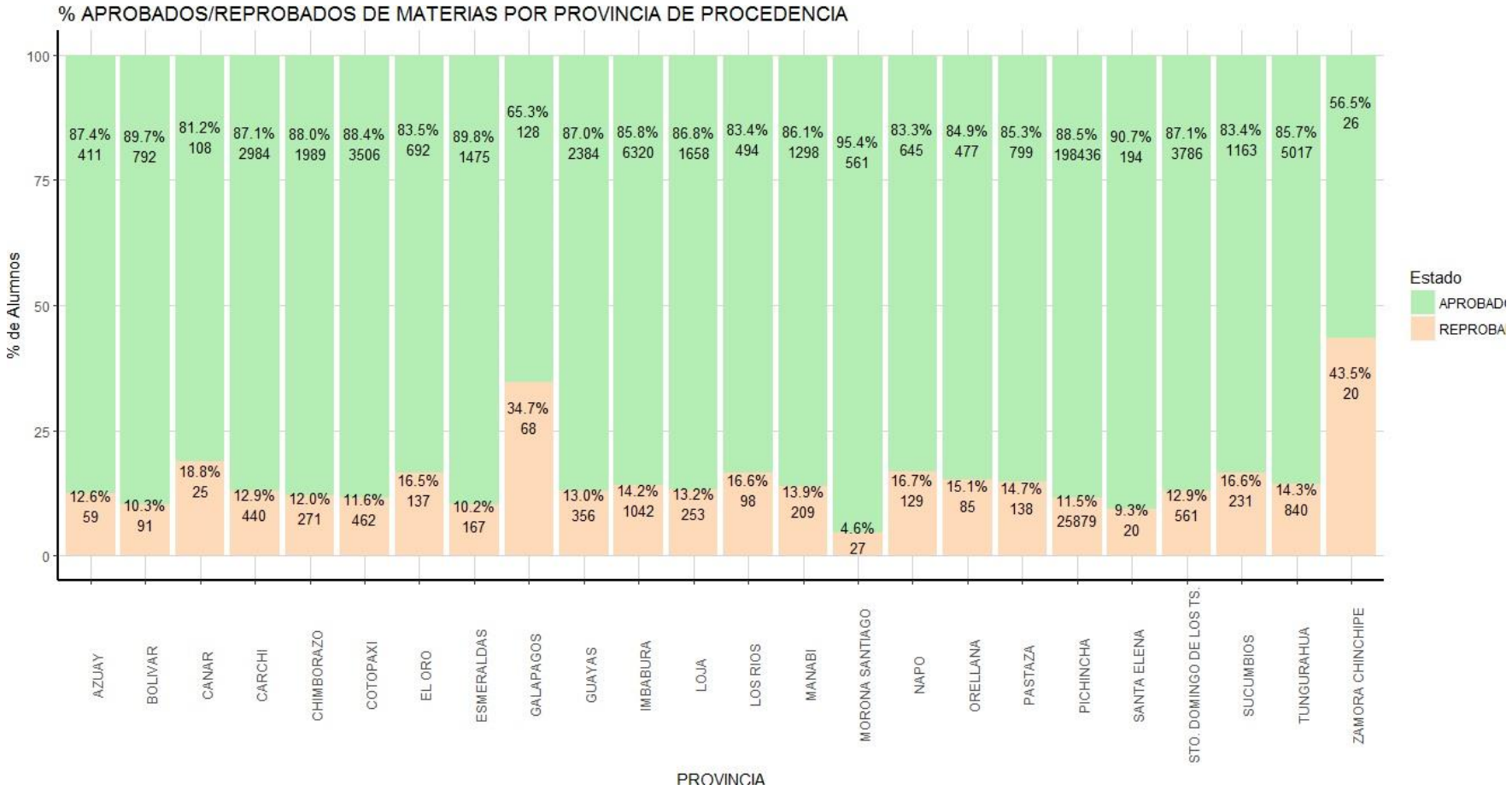

Figura 47. Porcentaje de alumnos que han aprobado y reprobado de materias por provincia de procedencia excluido Pichincha

De los alumnos de distintas etnias el porcentaje de reprobación es alrededor del 17%, en cuanto al porcentaje de reprobación los de etnia indígena tienen el 19.3% siendo el mayor sobre todas las etnias como se muestra en la Figura 48.

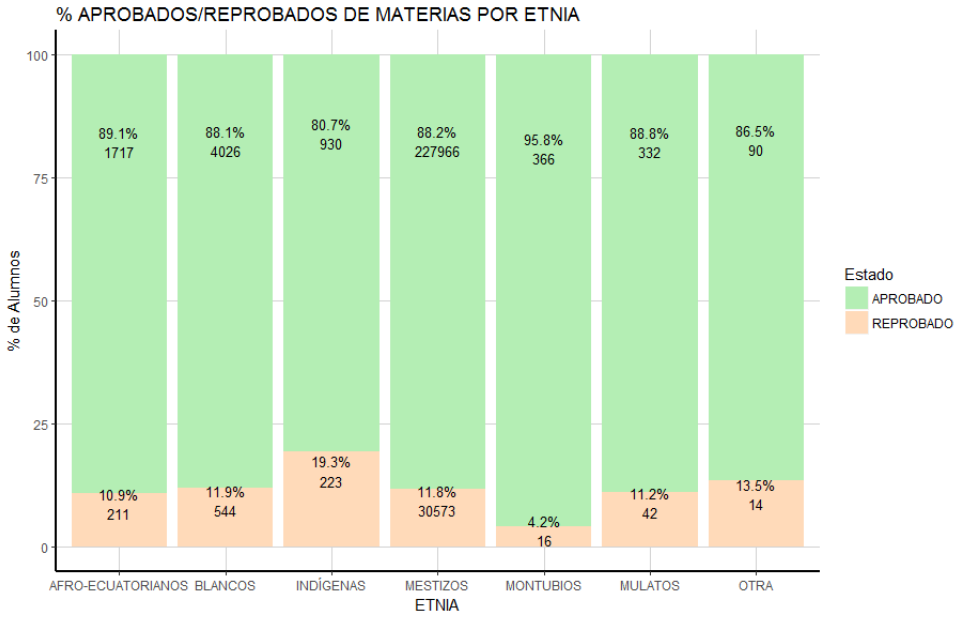

Figura 48. Porcentaje de alumnos que han aprobado y reprobado de materias por etnia.

La mayor cantidad de alumnos son solteros los cuales están seguidos por los casados abarcando el 96.8% y 2.64% respetivamente, dónde los casados son los que tienden a reprobar menos que los solteros como se muestra en la Figura 45. Al hablar de los demás estados civiles de los cuales existe una cantidad de alumnos relativamente pequeña el porcentaje de reprobación está entre el 15%.

Como muestra la Figura 50 de los alumnos sin discapacidad el porcentaje de reprobación ha variado en los últimos 5 años entre 8% y 14%, pero al mismo tiempo el porcentaje ha ido incrementado desde que la primera promoción que rindió el examen SENECYT ingresó a la Universidad de las Fuerzas Armadas – ESPE lo cual ocurrió en el período 201310. En cuanto al porcentaje de aprobación en el año 2012 se registró el menor porcentaje de reprobados.

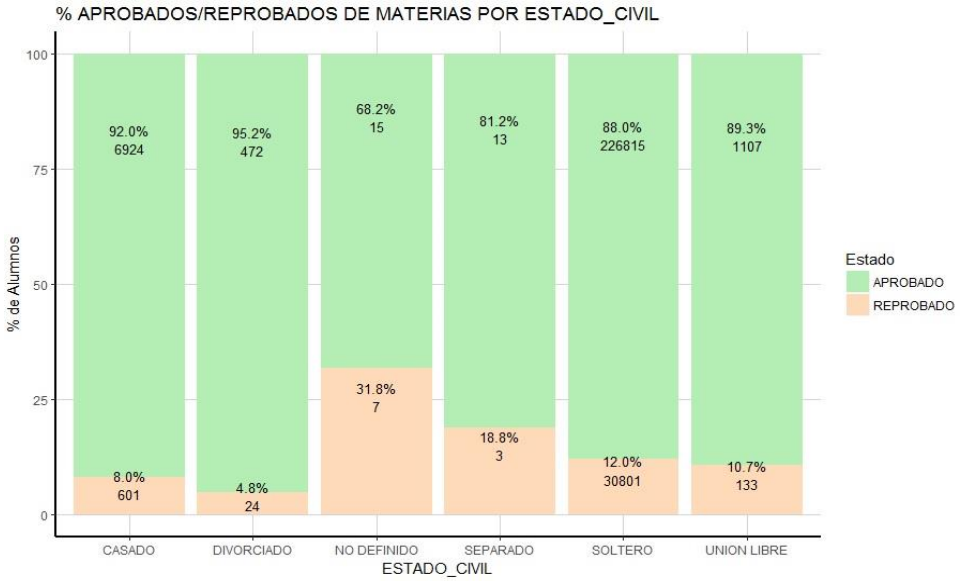

Figura 49. Porcentaje de alumnos que han aprobado y reprobado de materias por estado civil.

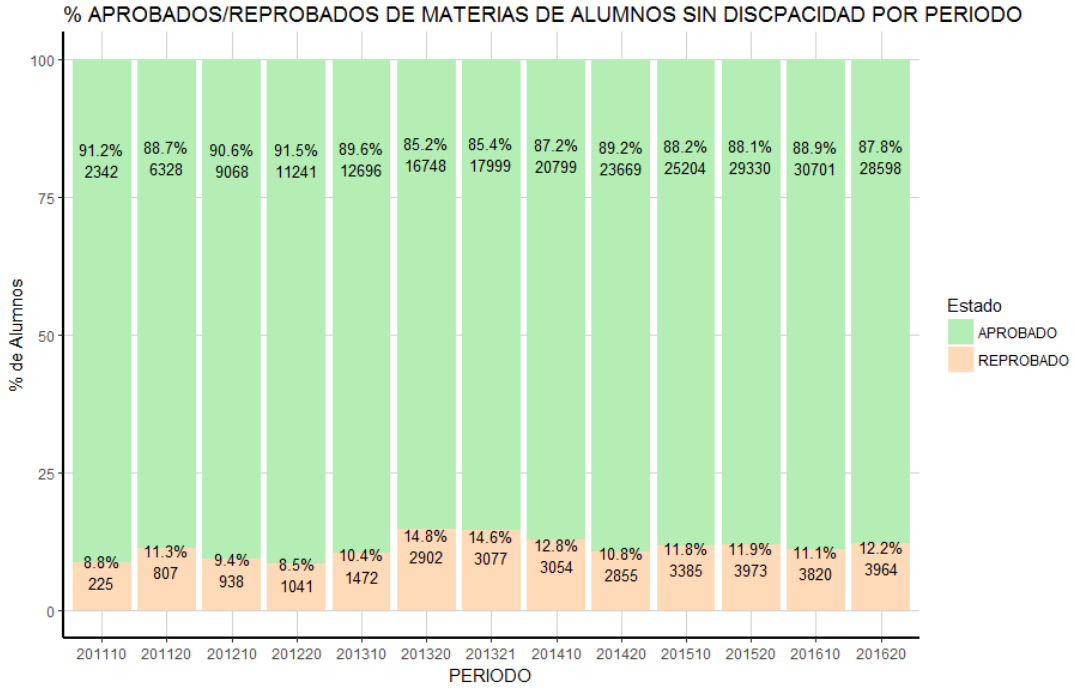

Figura 50. Porcentaje de alumnos que han aprobado y reprobado de materias de alumnos sin discapacidad.

Al hablar de alumnos con discapacidad los porcentajes de aprobación y reprobación varía mucho en los últimos cinco años como se observa en la Figura 51, cabe mencionar que el período 201120 no registra ningún reprobado.

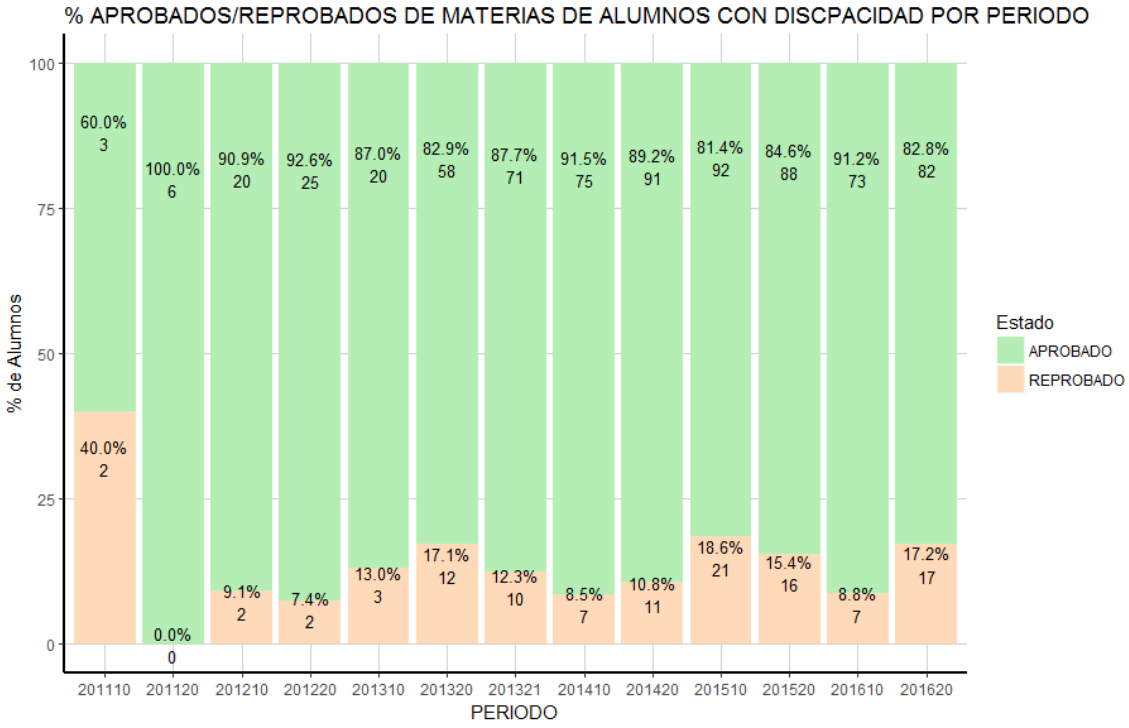

Figura 51. Porcentaje de alumnos que han aprobado y reprobado de materias de alumnos con discapacidad.

La Figura 52 muestra el Porcentaje de alumnos que han aprobado y reprobado de materias de alumnos por tipo de capacidades especiales el mayor porcentaje de reprobación son de los alumnos con dificultades físicas con 19.7%, seguidos por los de discapacidad visual con 10.7% materias reprobadas.

A continuación, en la Figura 53 se observa que el porcentaje de reprobación de materias supera considerablemente a los alumnos militares respecto de los civiles. Los alumnos militares en aprobación tienen 96.9% y los civiles 88.1% lo cual implica más o menos 10% de diferencia entre ambas mostrando como resultado que los alumnos militares tienen más probabilidades de aprobar sobre los alumnos civiles.

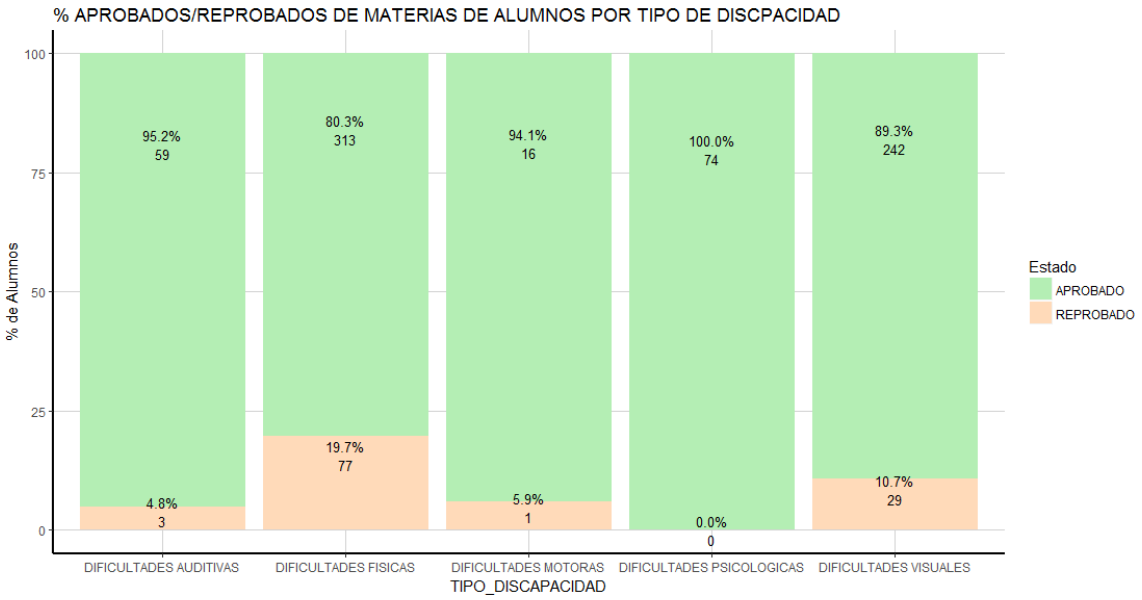

Figura 52. Porcentaje de alumnos que han aprobado y reprobado de materias de personas por tipo de capacidades especiales.

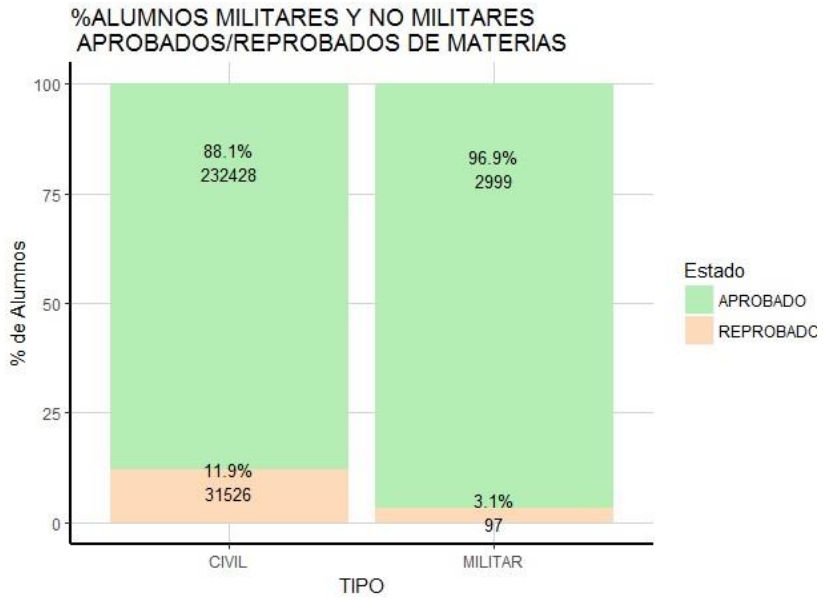

Figura 53. Porcentaje de alumnos que han aprobado y reprobado de materias de alumnos militares vs civiles.

El régimen del cual provienen la mayor cantidad de alumnos es sierra, seguido por el régimen costa, en la Figura 59 se observa que los de régimen Insular tienden a

reprobar con mayor frecuencia en comparación con los otros. Y los que reprueban menos son los que vienen del régimen escolar de la Sierra.

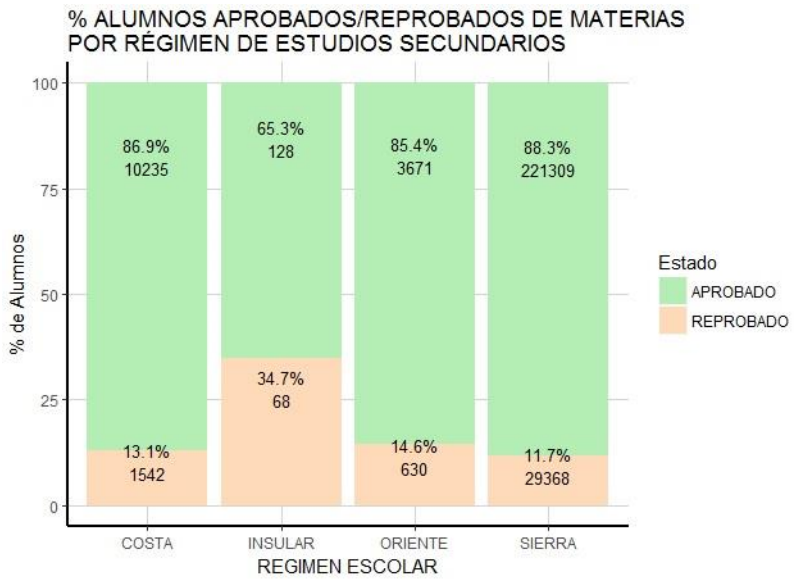

Figura 54. Porcentaje de alumnos que han aprobado y reprobado de materias por régimen del colegio.

La Figura 55 muestra el porcentaje de alumnos aprobados por sostenimiento del colegio donde se observa que a medida que pasan los años el porcentaje de materias aprobadas entre los dos se ha ido igualando.

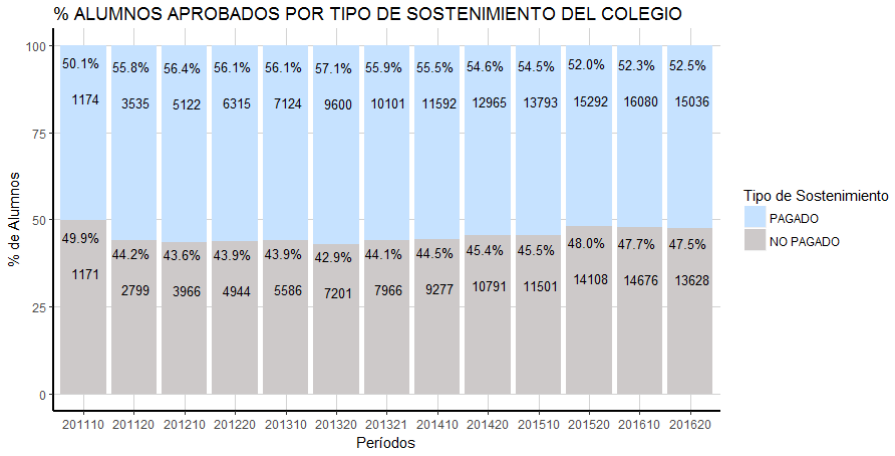

Figura 55. Porcentaje de alumnos aprobados por sostenimiento del colegio.

La Figura 56 muestra el porcentaje de alumnos reprobados por sostenimiento del colegio donde se observa que a medida que pasan los años el porcentaje entre los dos se ha ido igualando llegando a un 50% de cada uno.

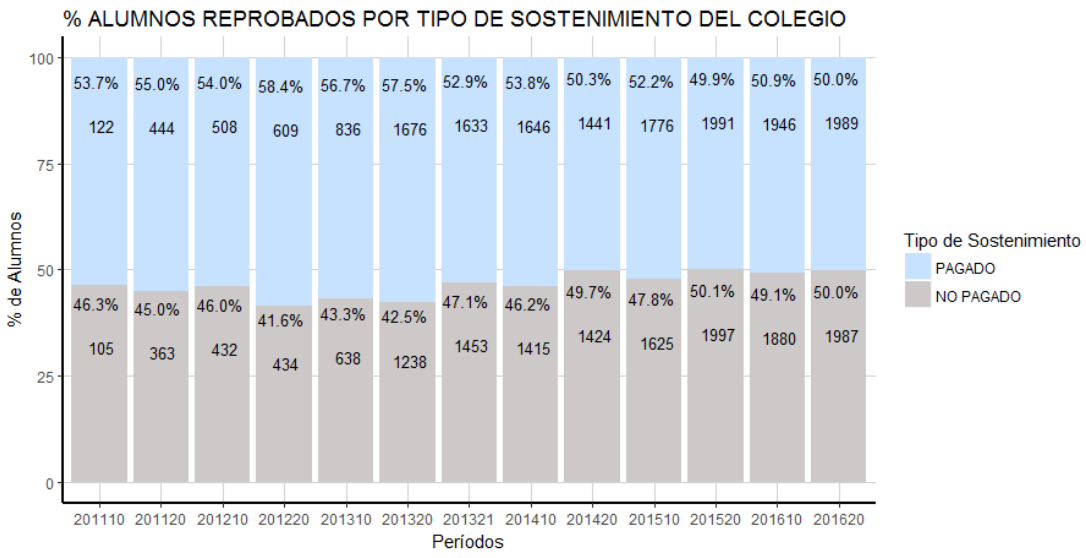

Figura 56. Porcentaje de alumnos reprobados en materias por tipo de sostenimiento del colegio.

Cómo se observa en la Figura 57 los colegios con mayor cantidad de aprobados en la Universidad de las Fuerzas Armadas - ESPE Luis Napolleon Dillon, Fernandez Madrid. De manera general los porcentajes de aprobados, reprobados y retirados varían mucho entre 6% y 8% entre los quince colegios con mayor demanda.

En la Figura 58 se puede observar de manera general que en los tres últimos años dónde se ingresó con examen SENECYT, es decir, en los últimos nueve períodos académicos el porcentaje de materias aprobadas ha ido disminuyendo mientras que el de materias reprobadas ha ido incrementando paulatinamente.

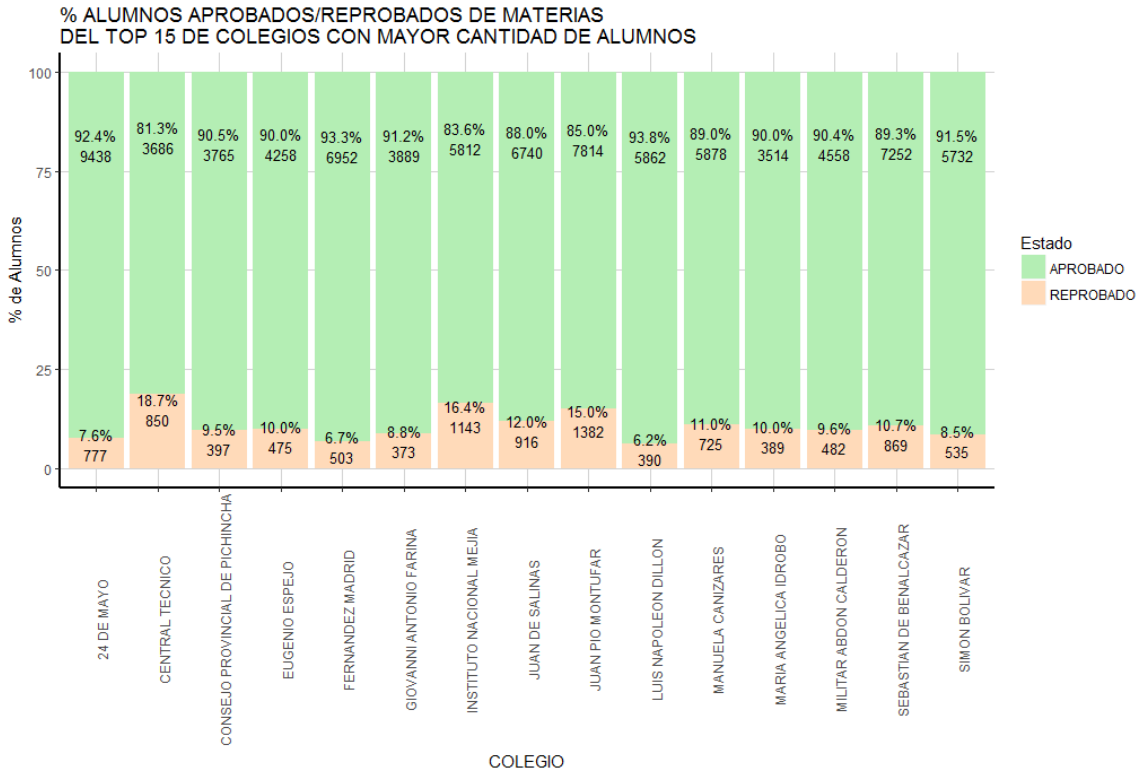

Figura 57. Porcentaje de alumnos que han aprobado y reprobado de materias del top 15 de colegios con mayor número de alumnos.

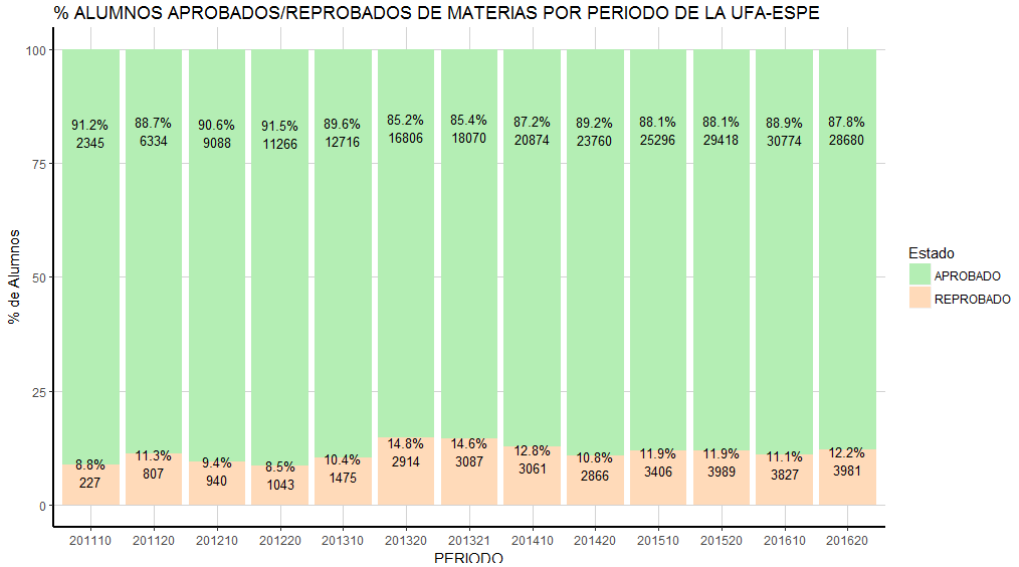

Figura 58. Porcentaje de alumnos que han aprobado y reprobado de materias por período.

La Figura 59 muestra que en el primer período académico de alumnos que ingresaron con la prueba del SENESCYT la deserción tuvo porcentajes bastante altos, pero a medida que avanzaron los períodos está ha ido disminuyendo hasta llegar a un porcentaje menor que el del inicio.

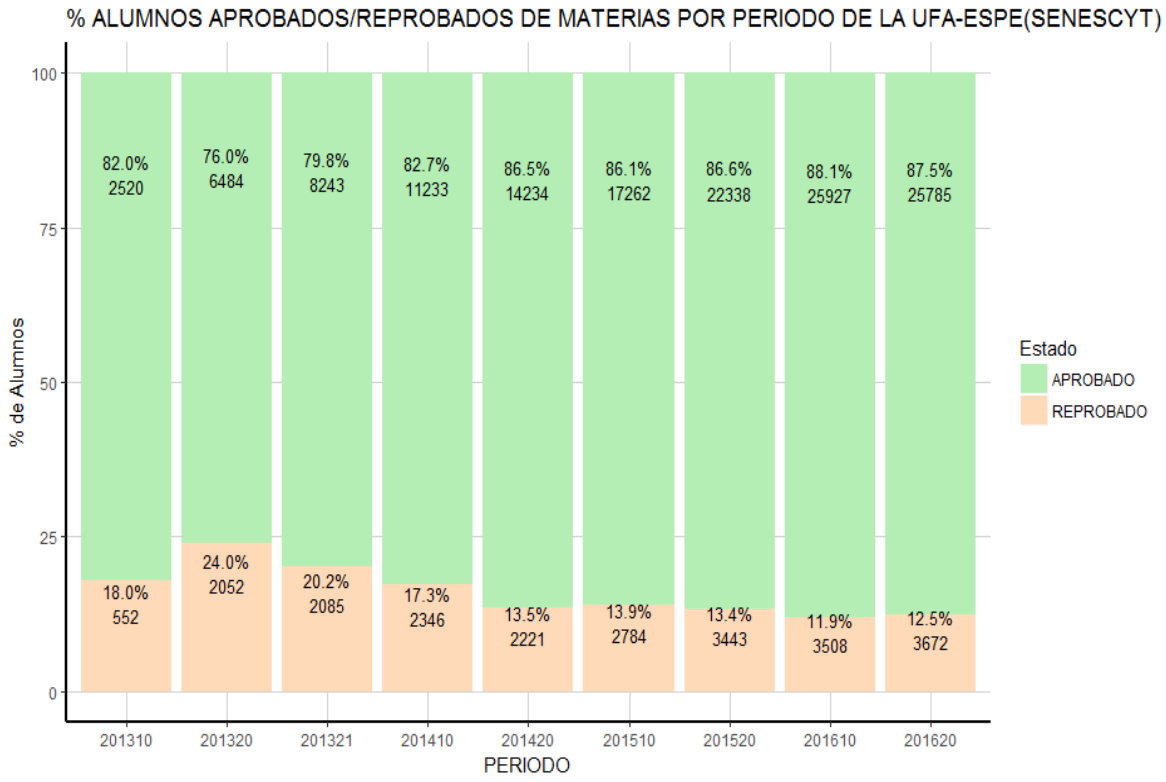

Figura 59. Porcentaje de alumnos que han aprobado y reprobado de materias de matriculados con la prueba del SENESCYT.

En la Figura 60 se observa que los porcentajes de aprobados y reprobados de materias que no ingresaron con el examen SENECYT tenían como porcentaje de reprobados un valor similar al de los alumnos de examen SENECYT que obtuvieron a finales del 2016 y que la medida que pasaron los períodos estos iban disminuyendo con pequeños porcentajes por lo cual su cambio se ha mantenido casi uniforme.

La Figura 61 muestra el top 10 de las materias con mayor número de alumnos reprobados de carreras técnicas, donde casi la totalidad de las materias mostradas pertenecen al departamento de Ciencias Exactas al ser carreras técnicas tienen mayor cantidad de materias de este departamento.

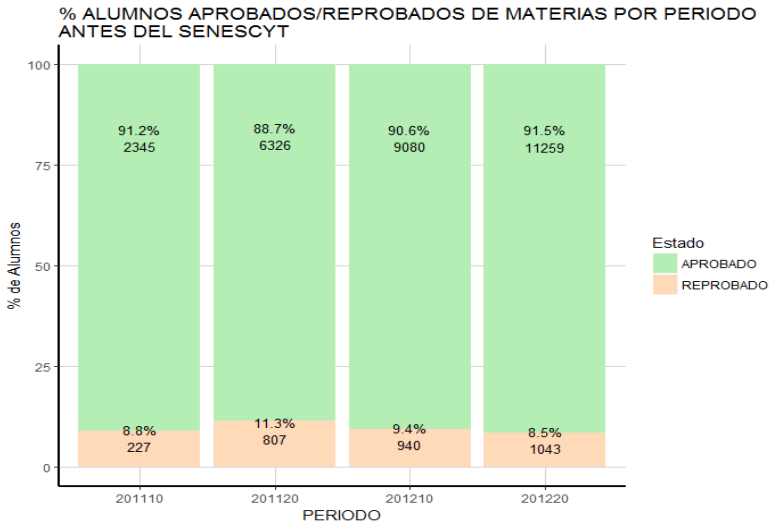

Figura 60. Porcentaje de alumnos que han aprobado y reprobado de materias de matriculados con la prueba antes del SENESCYT.

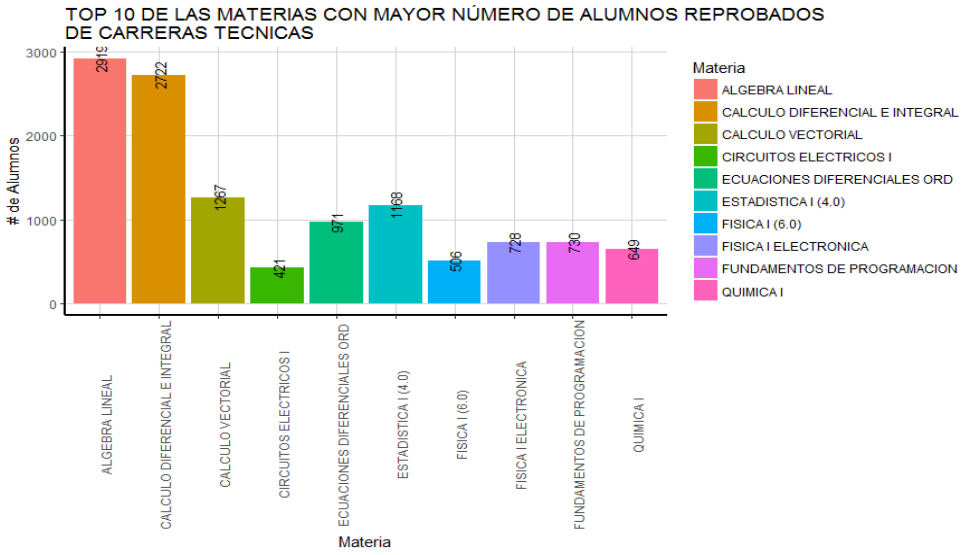

Figura 61. Top 10 de las materias con mayor número de alumnos reprobados de carreras técnicas

En las carreras administrativas las materias con mayor cantidad de reprobados pertenecen en un casi el 50% al departamento de ciencias exactas de igual forma que en la gráfica anterior (ver Figura 62).

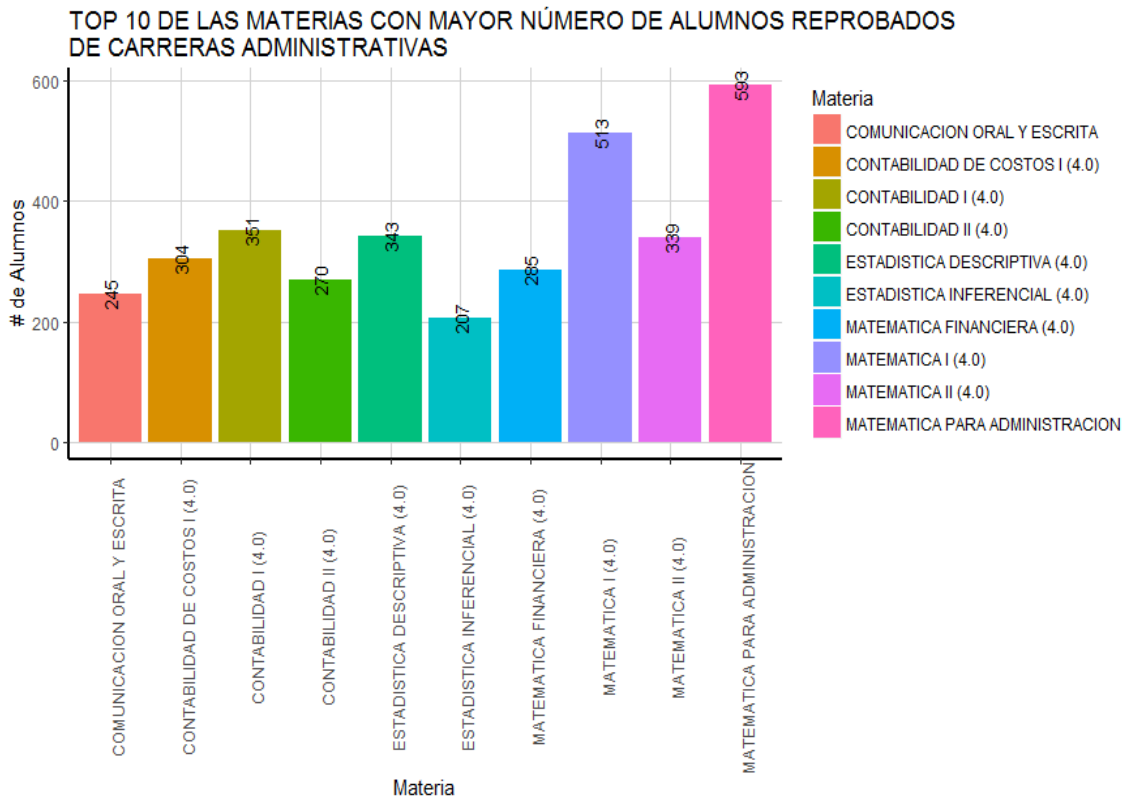

Figura 62. Top 10 de las materias con mayor número de alumnos reprobados de carreras administrativas

De las carreras Humanísticas donde se encuentran las carreras de Actividad Física y del Deporte y Educación infantil la que mayor demanda tiene es la primera mencionada, por lo que las materias con mayor cantidad de reprobados son materias deportivas como se observa en la Figura 63, además se puede observar que de las pocas materias del departamento de Ciencias Exactas que toman dichas carreras estás se encuentran incluidas entre las que más reprueban.

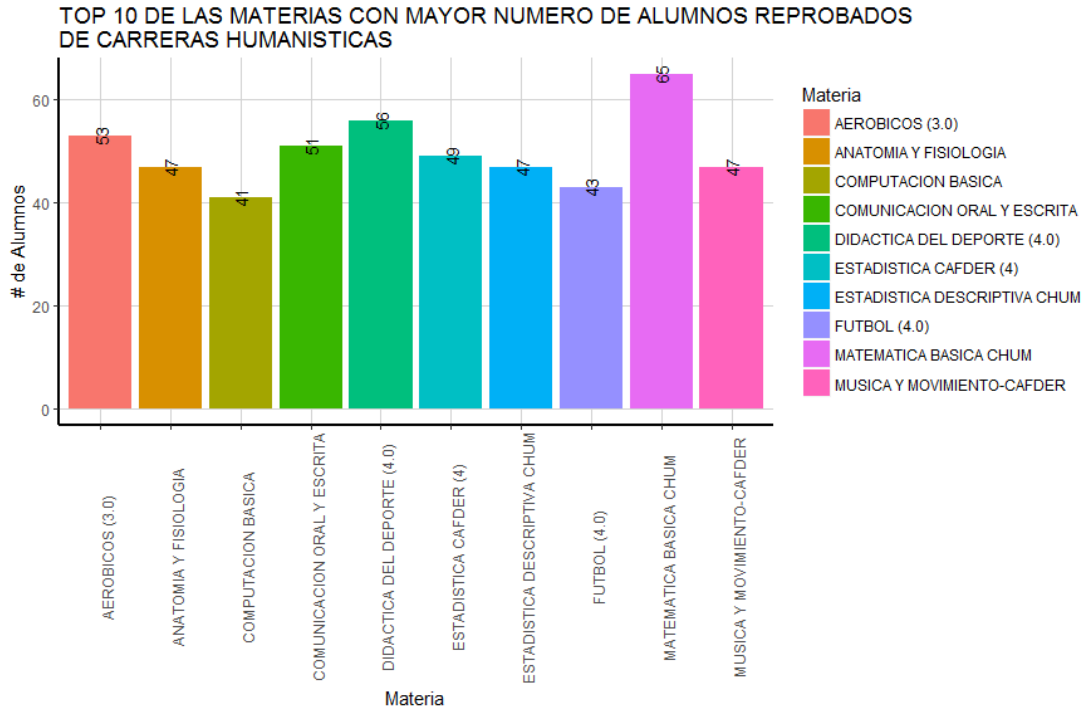

Figura 63. Top 10 de las materias con mayor número de alumnos reprobados de carreras humanísticas.

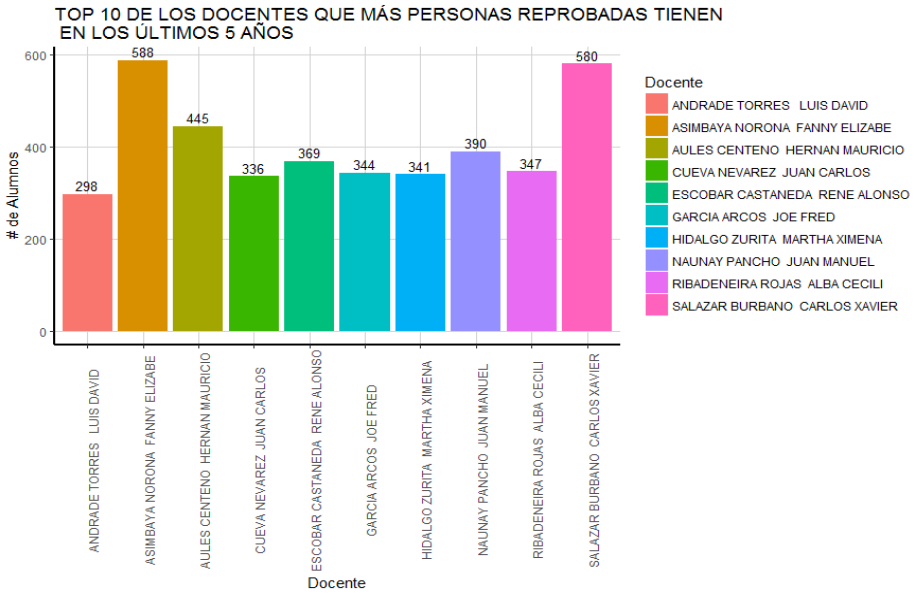

Figura 64. Top 10 de los docentes con mayor número de alumnos reprobados.

La Figura 64 muestra los docentes que más cantidad de alumnos reprobados tienen en los últimos cinco años los cuales casi su totalidad pertenecen al departamento de Ciencias Exactas exceptuando a uno que pertenece al Ciencias Humanas y Sociales.

Las materias de las distintas carreras se encuentran asociadas a un departamento, por ello se ha obtenido el porcentaje de alumnos aprobados y reprobados de cada uno de estos, dónde la Figura 65 muestra que las materias del departamento de Ciencias Exactas con el mayor porcentaje de reprobación con 28.9%, seguido por los departamentos de Eléctrica y Electrónica, Energía de ciencias Mecánicas y Ciencias de la Computación con 12.6%, 16.6%, 11.4% y 11.4% respectivamente.

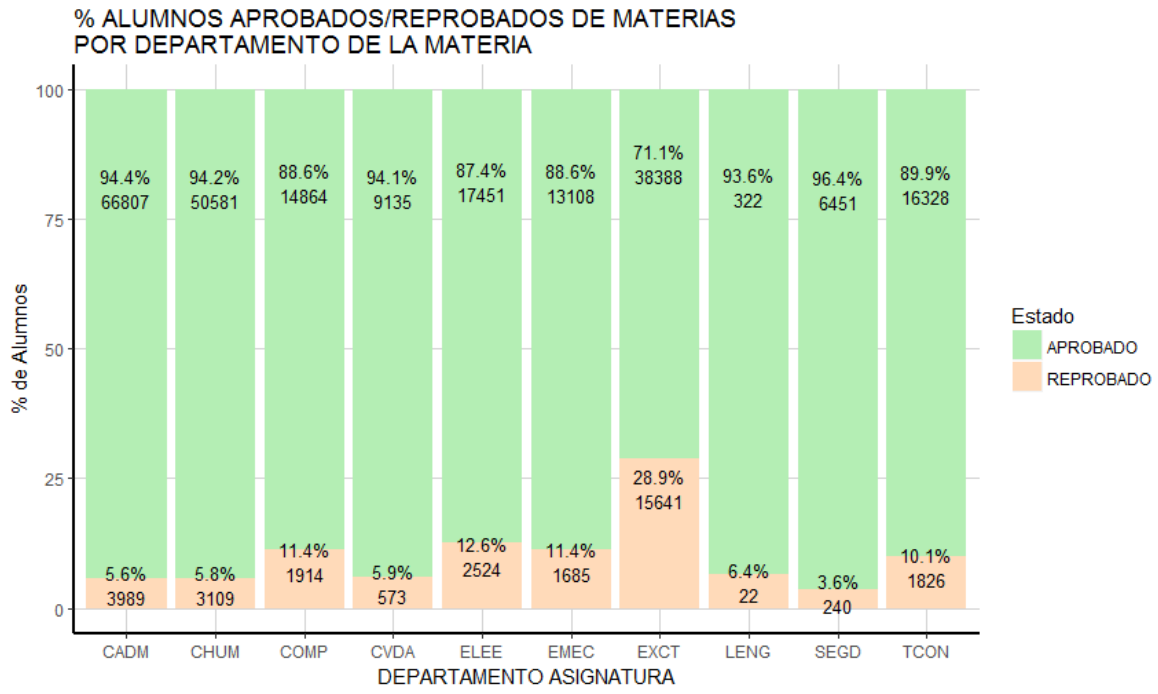

Figura 65. Porcentaje alumnos con materias aprobadas y reprobadas por departamento

El promedio general de los estudiantes por carrera varía entre 14.27 y 16.68 sobre 20. Las carreras con mayor puntaje son Educación Infantil y Hotelera y

Comercio Exterior y Negociación con (16,68 y 16,59) y las que menor puntaje tienen son las carreras Electrónica y Telecomunicaciones, Electrónica Automatización y Control y Mecánica con (14.27, 14.48 y 14.48) como se observa en la Figura 66.

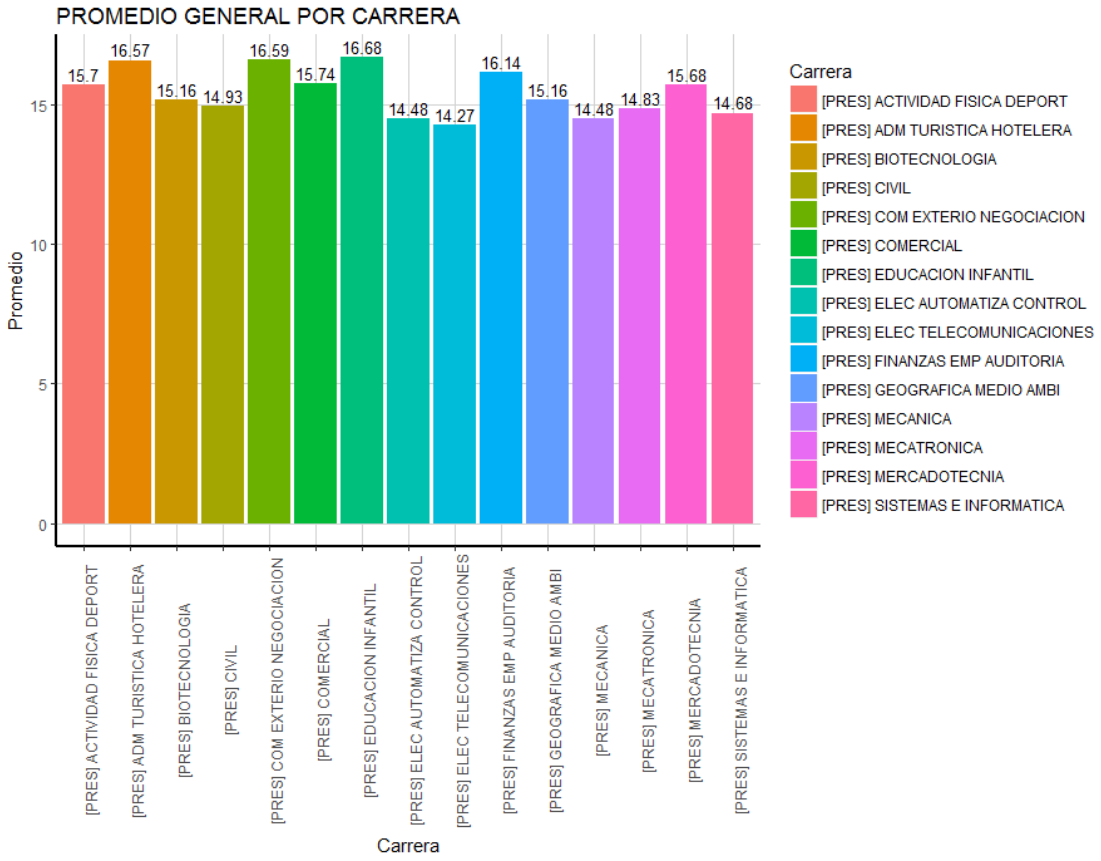

Figura 66. Promedio General por carrera de la Universidad de las Fuerzas Armadas – ESPE

Los períodos posteriores al 201321 (Agosto 2013 – Diciembre 2013) de los alumnos que pertenecen al departamento de Ciencias Económicas, Administrativas y de Comercio, el porcentaje de reprobados ha incrementado considerablemente en comparación con el de períodos anteriores. A pesar de esto, el promedio que tienen de aprobar una materia es mayor al 90% y la cantidad de matrículas reprobadas no superan las 1042 por período como se muestra en la Figura 67.

De los alumnos que pertenecen al departamento de Ciencias Humanas y

Sociales, a partir del período 201310 ha incrementado considerablemente el número de reprobados de 37 a 184. En general el porcentaje para aprobar es superior a 89.8%. y la cantidad de matrículas reprobadas no superan 184 por período como se observa en la Figura 68.

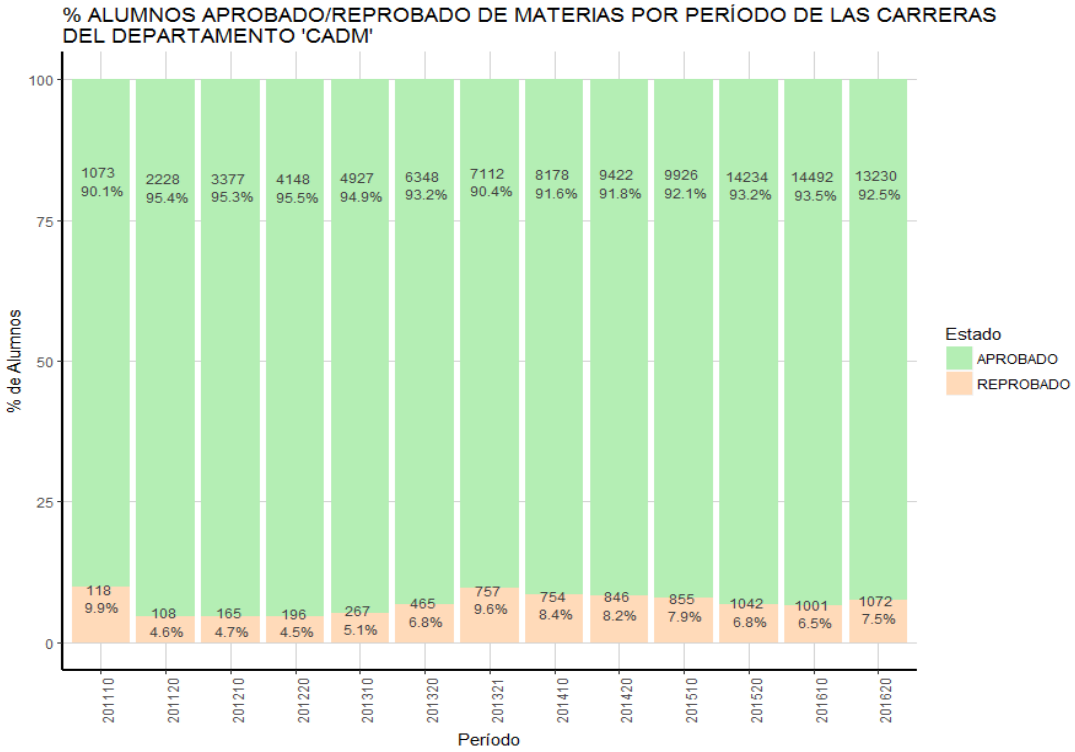

Figura 67. Porcentaje de alumnos del departamento de CADM con materias aprobadas y reprobadas por período.

De los alumnos que pertenecen al departamento de Ciencias de la Computación, el porcentaje de materias reprobadas no supera el 22.3%, la cantidad de reprobados durante estos 5 años ha seguido con un patrón en el que incrementa luego disminuye a continuación aumenta y finalmente vuelve a disminuir como se muestra en la Figura 69.

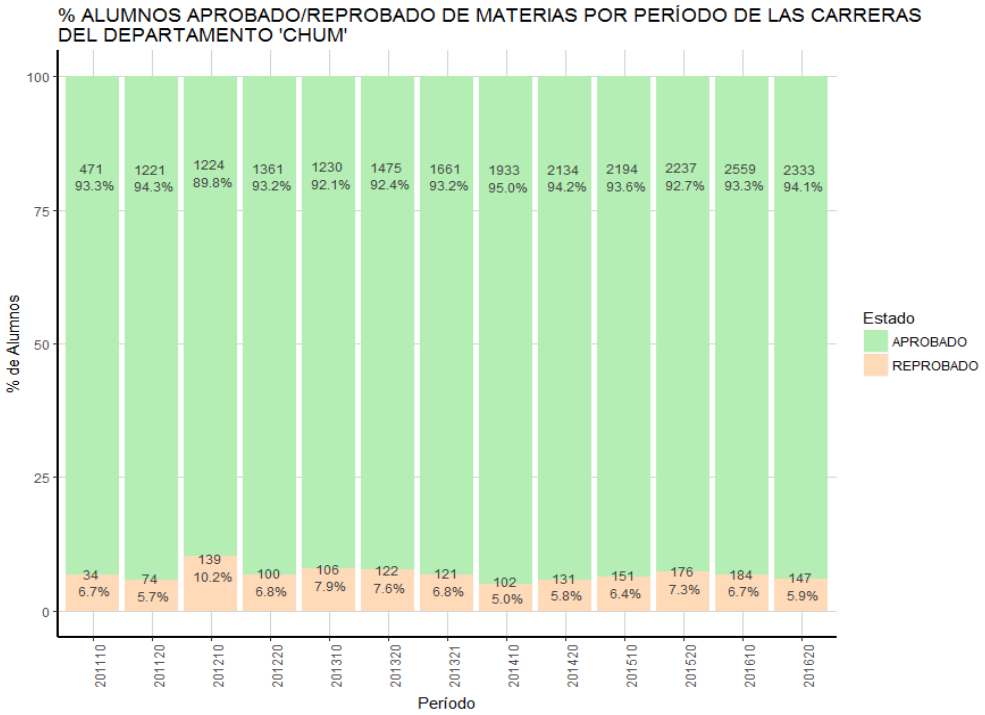

Figura 68. Porcentaje de alumnos del departamento de Ciencias Humanas y Sociales con materias aprobadas y reprobadas por período

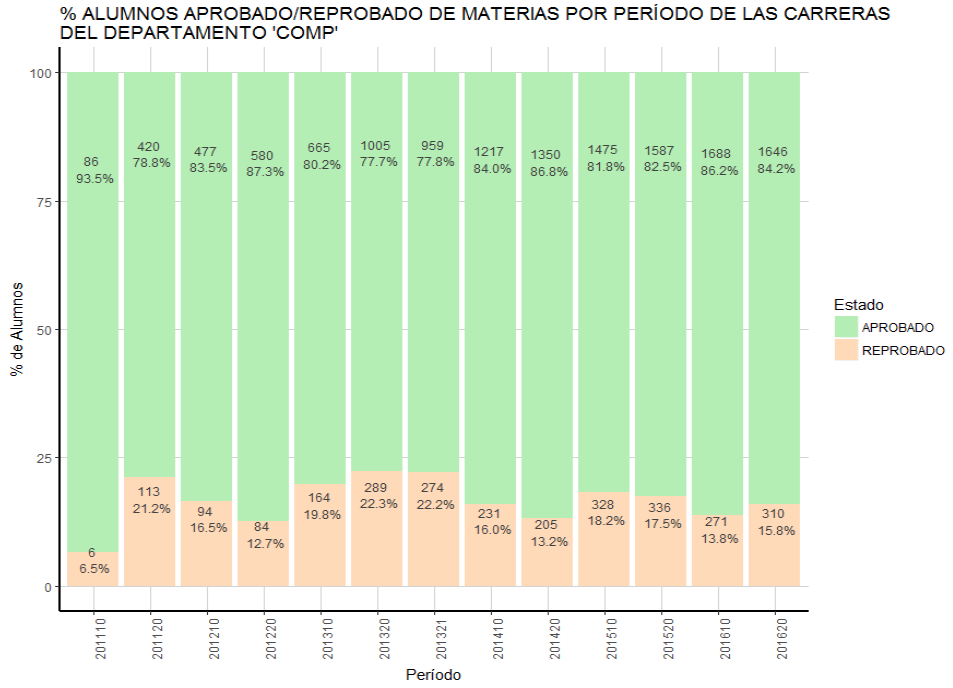

Figura 69. Porcentaje de alumnos del departamento de Ciencias de la Computación con materias aprobadas y reprobadas por período

Respecto a los alumnos que pertenecen al departamento de Ciencias de la Vida, el porcentaje de materias reprobadas ha ido incrementando de forma radical con el paso de los períodos de 11 a 398 como se muestra en la Figura 70 a continuación.

De los alumnos que pertenecen al departamento de Eléctrica y Electrónica, el porcentaje de materias reprobadas ha ido incrementando con el paso de los períodos de 15 a 787 como se muestra en la Figura 71. El mayor porcentaje de materias retiradas fue en el período 201320 con el 24.4% del total de registros en el mismo.

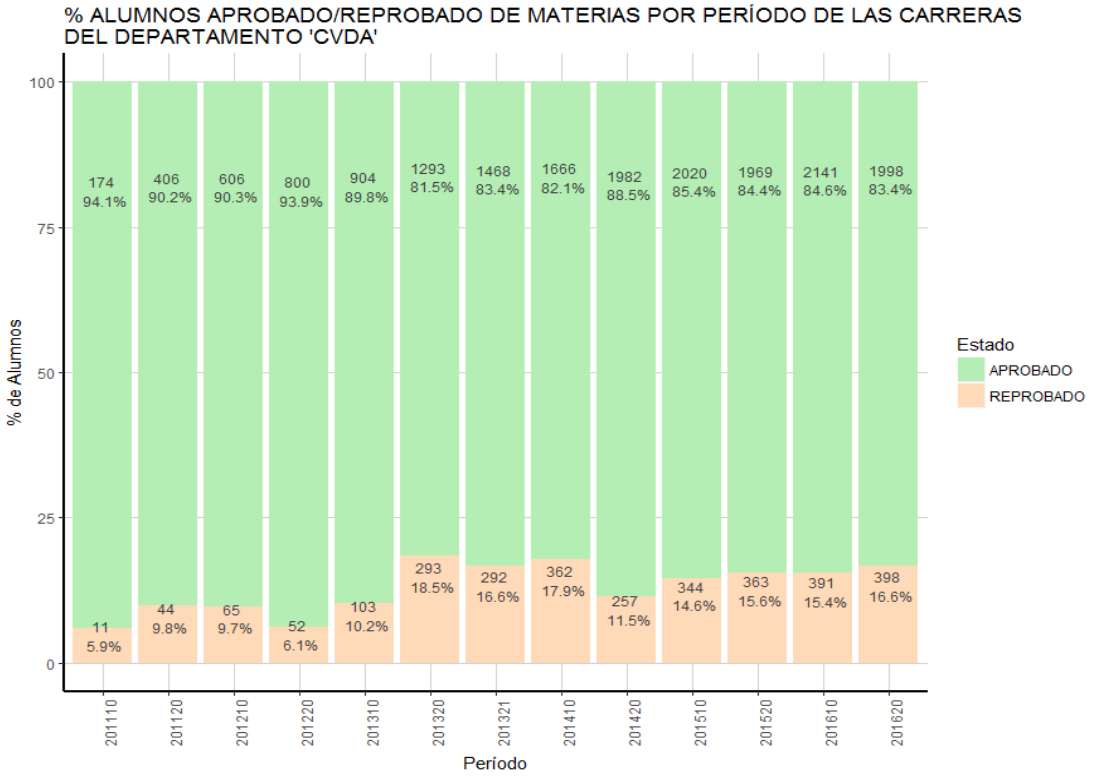

Figura 70. Porcentaje de alumnos del departamento de Ciencias de la Vida con materias aprobadas y reprobadas por período

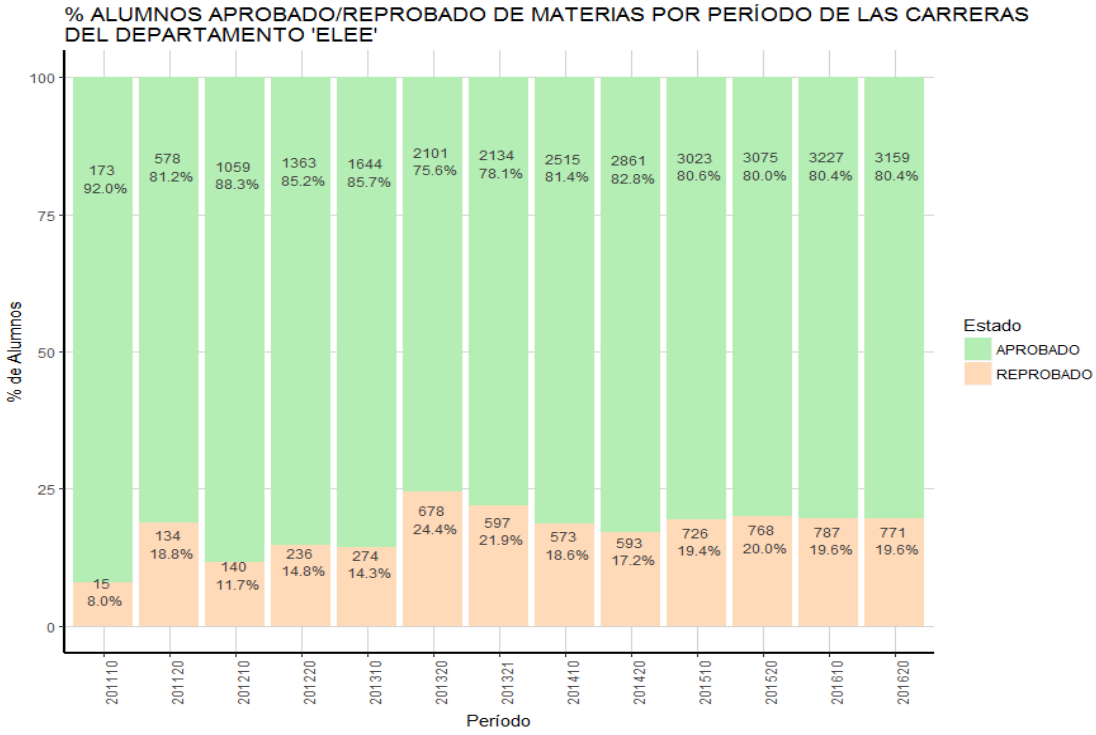

Figura 71. Porcentaje de alumnos del departamento de Eléctrica y Electrónica con materias aprobadas y reprobadas por período

De los alumnos que pertenecen al departamento de Energía y Mecánica, el mayor porcentaje de materias reprobadas fue en el 201321 con el 19.9%, sin embargo, los tres períodos consecutivos este porcentaje disminuyo y volvió a incrementar en del 201520 al 201620 como se observa en la Figura 72. Además, muestra que durante los últimos 5 años ha seguido con un patrón similar al de los alumnos del Departamento de Ciencias de la Computación donde incrementa luego disminuye a continuación aumenta y finalmente vuelve a disminuir.

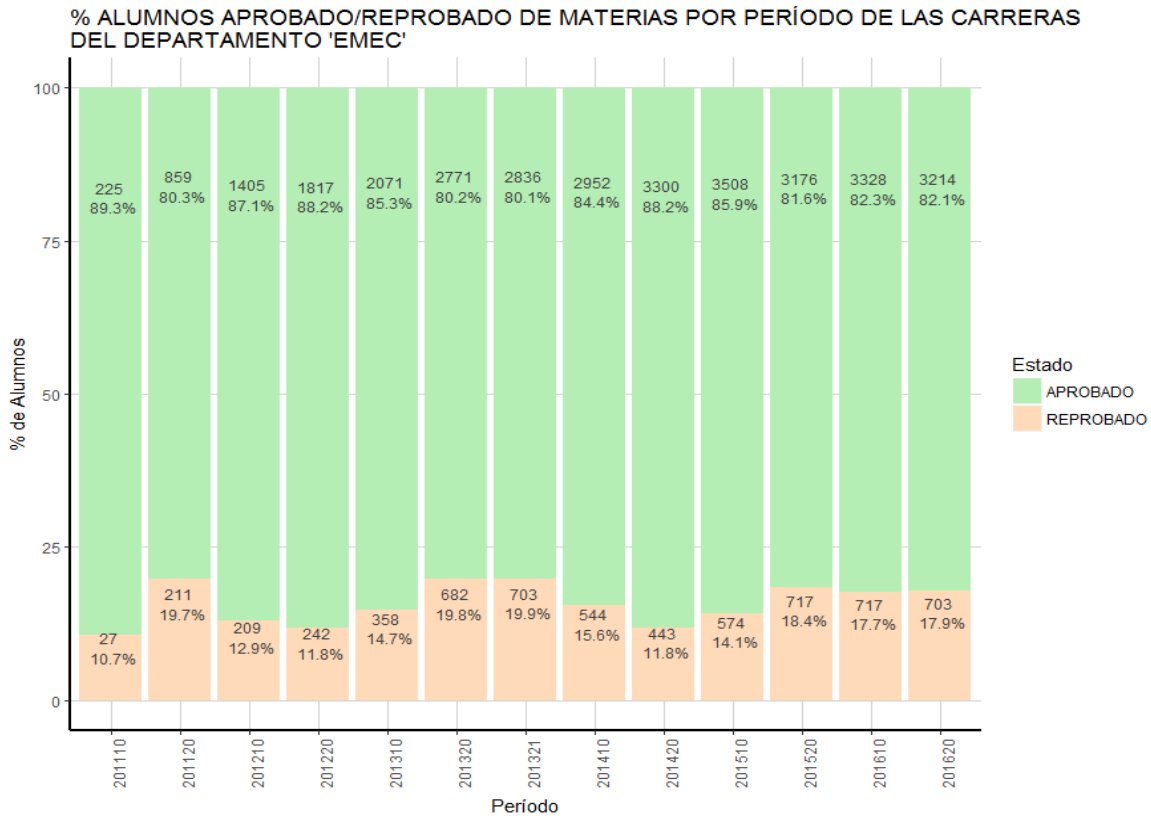

Figura 72. Porcentaje de alumnos del departamento de Energía y Mecánica con materias aprobadas y reprobadas por período

De los alumnos que pertenecen al departamento de Ciencias de la Tierra y de la Construcción, el porcentaje de materias reprobadas varía de 16 a 587 que con el paso de los años ha ido incrementando, mientras que el porcentaje de materias aprobadas es superior al 83% como se observa en la Figura 73.

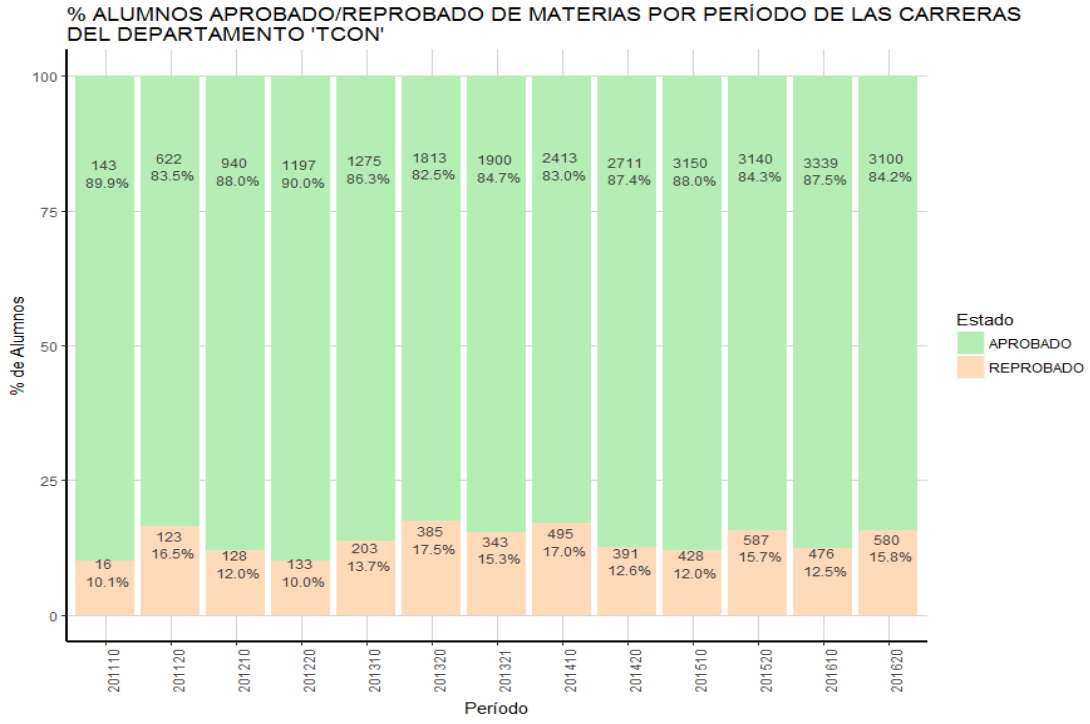

Figura 73. Porcentaje de alumnos del departamento de Ciencias de la Tierra y de la Construcción con materias aprobadas y reprobadas por período

En la Figura 74 se puede observar que, del total de 8264 alumnos, tan solo 374, es decir el 4.53%, se han cambiado de carrera mientras que el resto 7890 alumnos nunca se han cambiado de carrera dentro de la Universidad de las Fuerzas Armadas – ESPE.

De los 525 alumnos que se han cambiado de carrera 521 lo han realizado dos veces, de los cuales 210 alumnos se han cambiado de una carrera técnica a otra técnica, seguidos por los alumnos que se cambian entre dos carreras administrativas y 69 alumnos que se han cambiado de carreras técnicas a administrativas como se observa en la Figura 75.

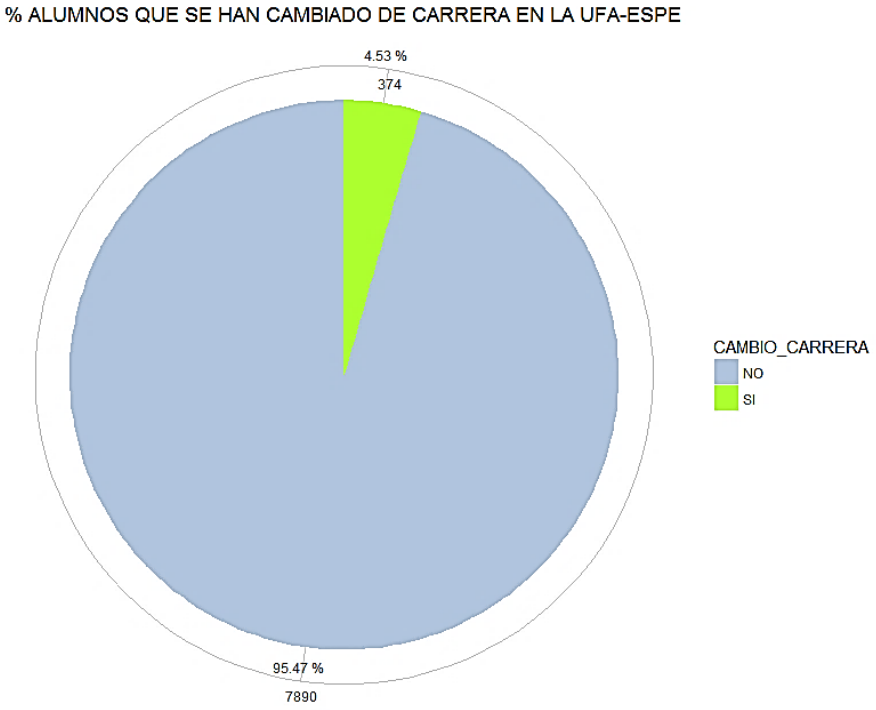

Figura 74. Porcentaje de alumnos que se han cambiado de carrera.

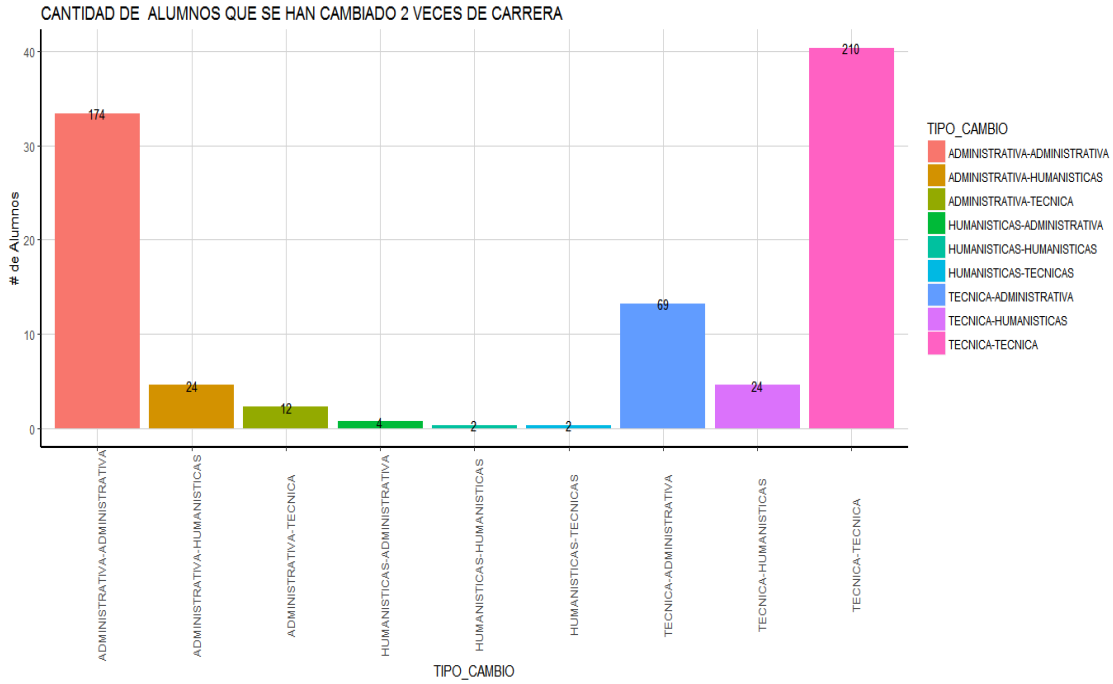

Figura 75. Cantidad de alumnos que se han cambiado dos veces de carrera de la Universidad de las Fuerzas Armadas – ESPE

De los 525 alumnos que se han cambiado de carrera 4 lo han realizado tres veces, como se observa en la Figura 76.

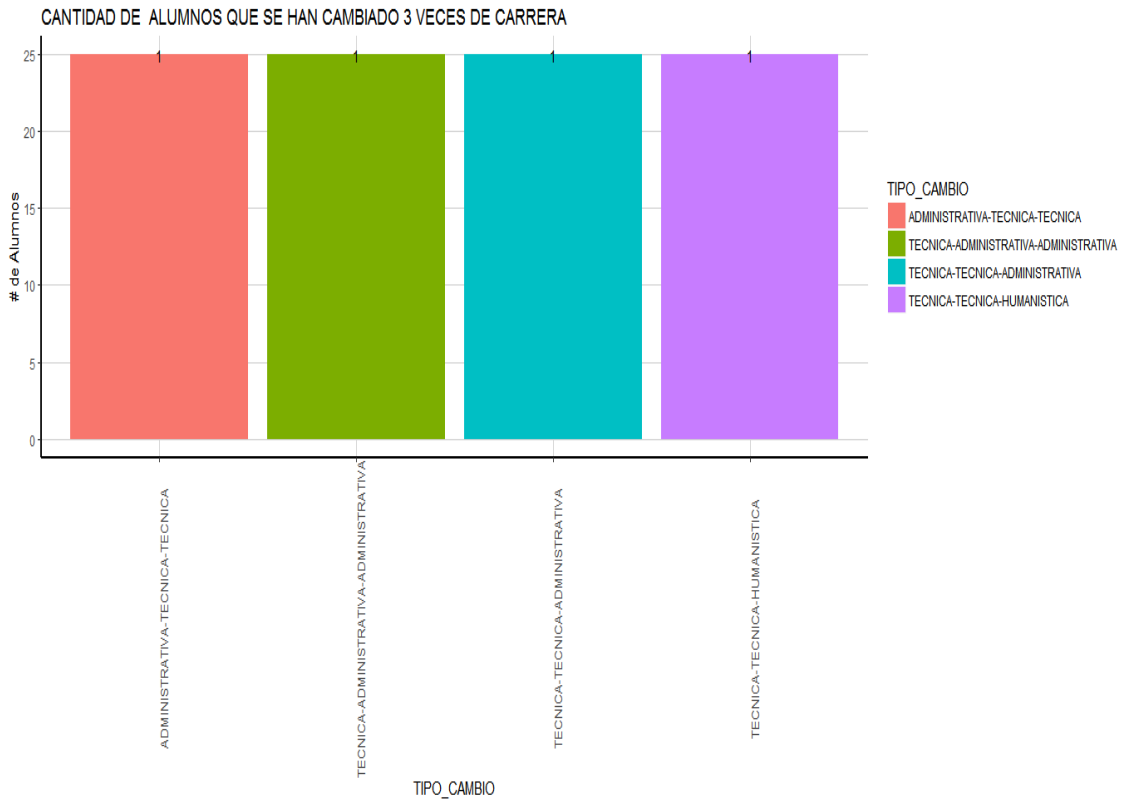

Figura 76. Cantidad de alumnos que se han cambiado tres veces de carrera de la Universidad de las Fuerzas Armadas - ESPE

Los cursos que tienen mayor cantidad de alumnos tienen mayor probabilidad de reprobar como se observa en la Figura 77. Los alumnos que toman en un período de 13 a 21 créditos son los que tienen mayor porcentaje de reprobar como se muestra en la Figura 78 sin embargo, este porcentaje disminuye en 8.8% a 9.8% para los alumnos que toman de 0 a 12 créditos y más de 31 créditos, el mismo porcentaje que para los alumnos que toman de 5 a 8.

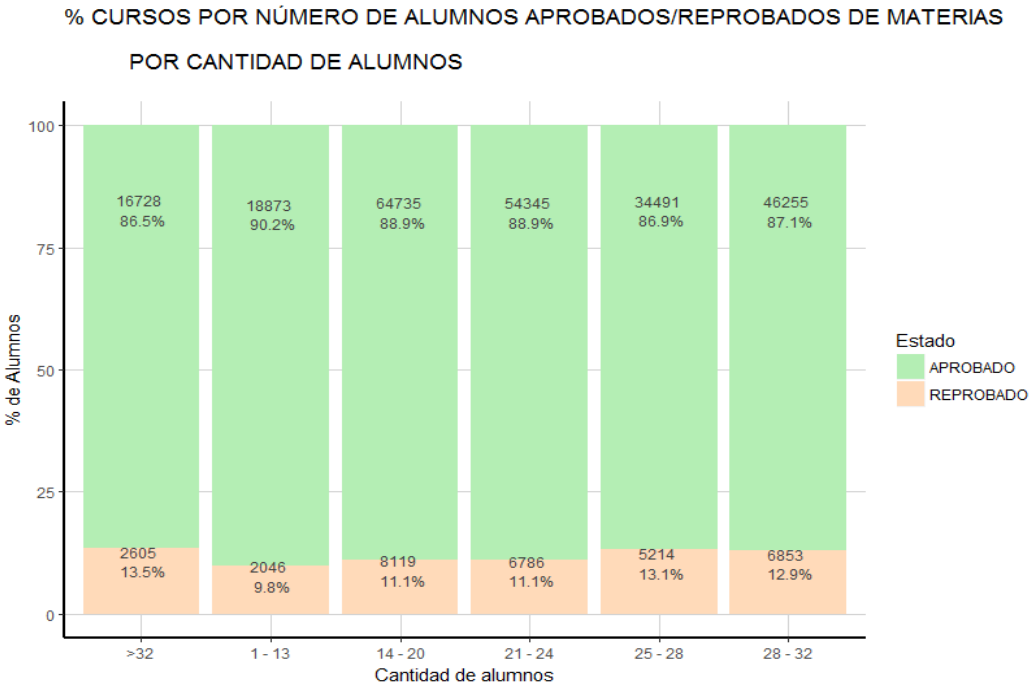

Figura 77. Porcentaje de cursos por número de alumnos con materias aprobadas y reprobadas.

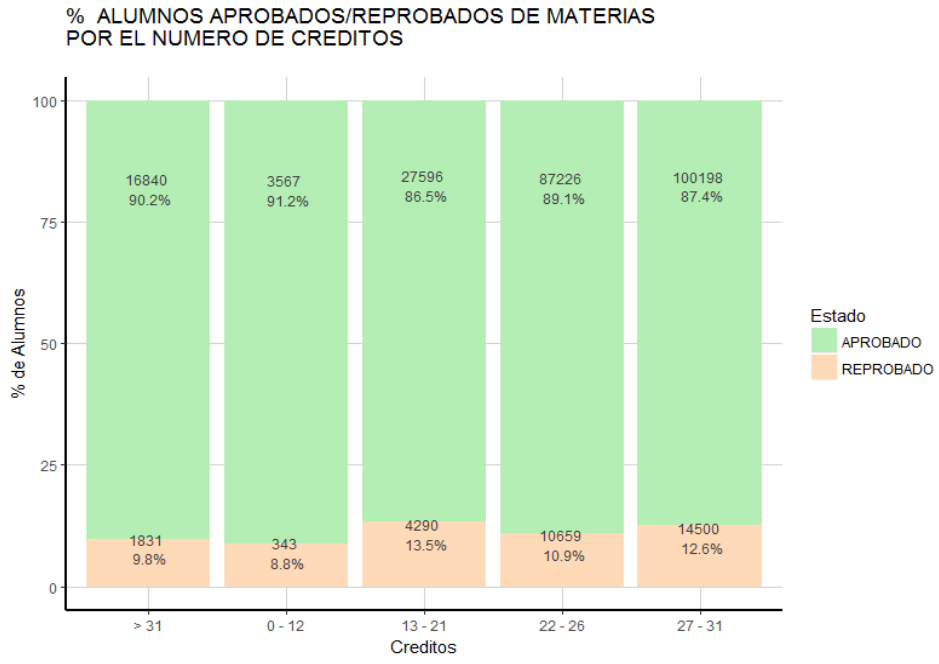

Figura 78. Porcentaje de alumnos con materias aprobadas y reprobadas por el número de créditos

Los alumnos que toman materias en la mañana tienen aproximadamente 8% más probabilidad de reprobar que los alumnos que estudian en la noche. Este fenómeno se presenta debido a que en la tarde y noche se encuentra los horarios pertenecientes a carreras administrativas donde se observó que el porcentaje de reprobados era mínimo como se indica en la Figura 79.

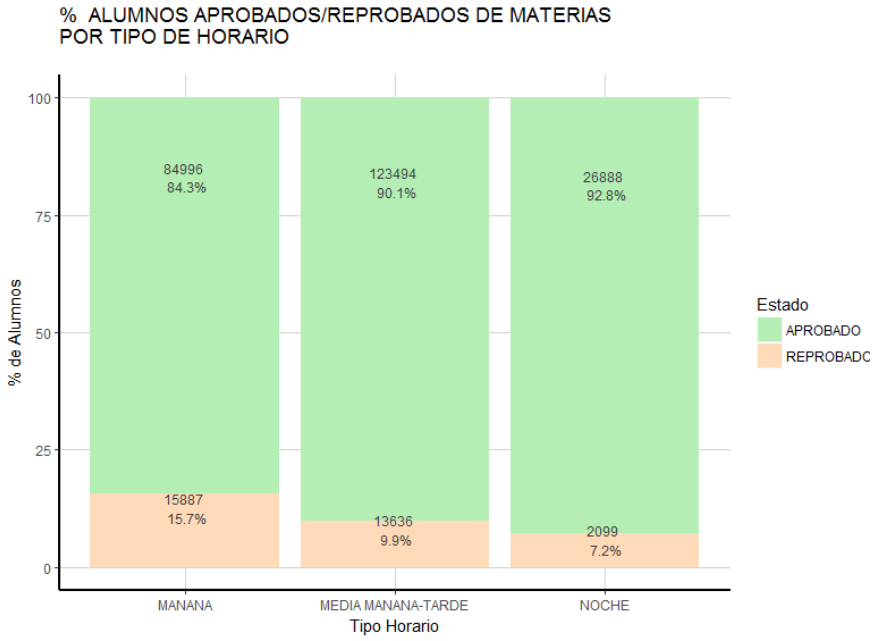

Figura 79. Porcentaje de alumnos aprobadas y reprobadas de materias tomadas en horario de la mañana o de la tarde

Las materias que son impartidas los días viernes se detallan a continuación en la Figura 80, donde se observa que el período 201320 y 201321 es donde ocurre los más altos porcentajes de reprobados, mientras que los períodos anteriores al 201320 son los que tienen menores reprobados.

En estos últimos cinco años han existido 18439 registros de segunda matrícula y 852 de tercera matrícula. La cantidad de alumnos que han realizado segunda matrícula ha ido incrementando exponencialmente de 45 a 2625 entre los períodos 201110 y 201620 (años 2011 y 2016). Respecto a los alumnos con tercera matrícula

se muestra un gran incremento en los dos últimos períodos de 4 a 278 matrículas, debido a que durante estos dos últimos períodos la Universidad permitió realizar tercera matrícula a todos los alumnos que lo necesitaran ver Figura 81.

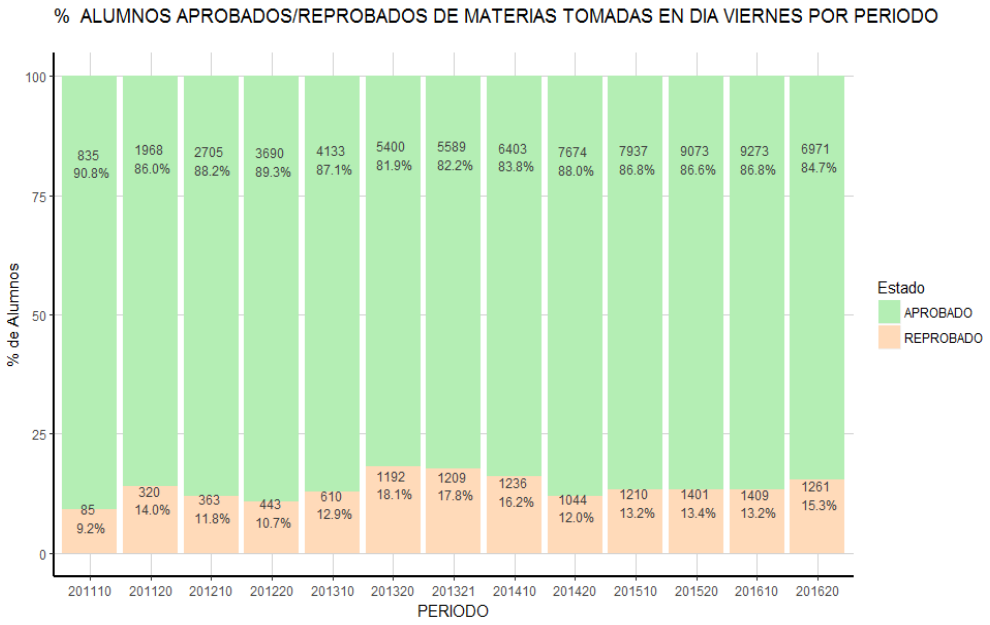

Figura 80. Porcentaje de alumnos aprobadas y reprobadas de materias tomadas el día viernes

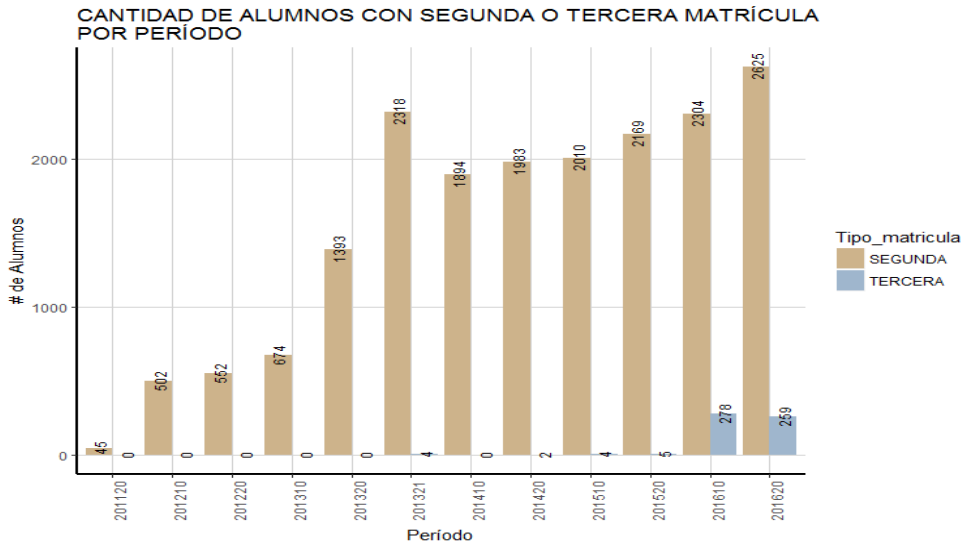

Figura 81. Cantidad de alumnos con segunda o tercera matrícula por período.

Del total de 19291 registros entre segunda y tercera matrícula, 1067 son alumnos desertores de la Universidad debido a que reprobaron segunda o tercera matrícula. Y es en el período 201321 donde presenta un alto número de alumnos desertores como se observa en la Figura 82.

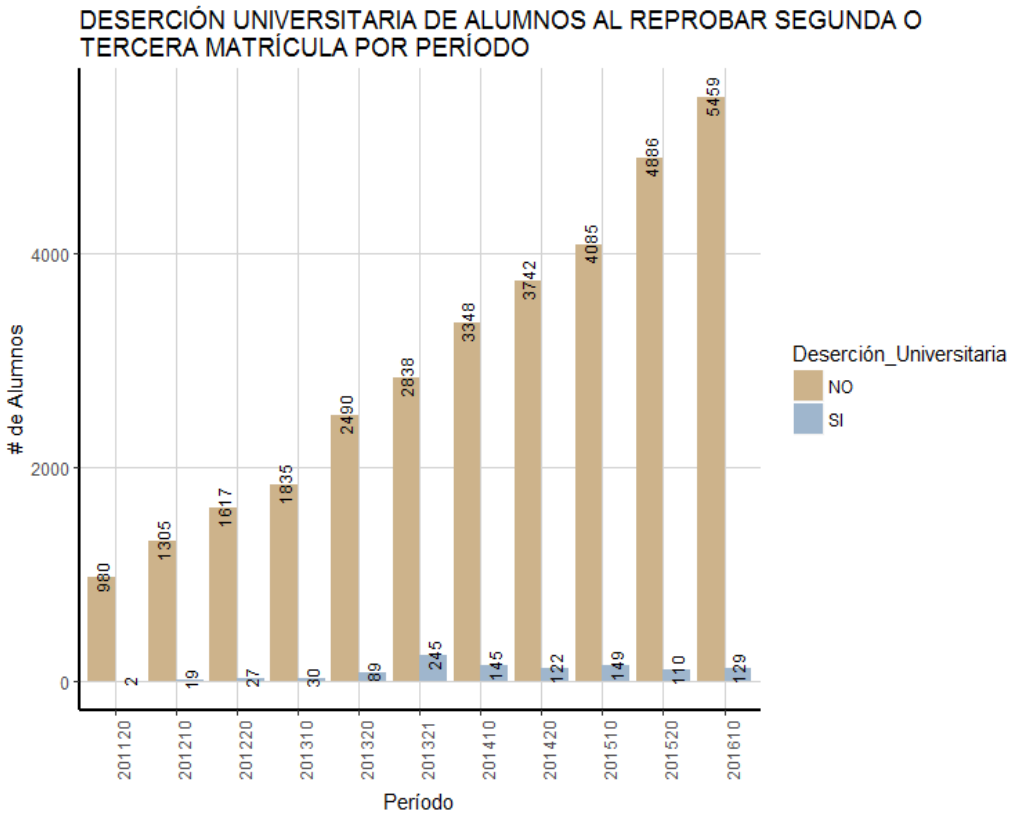

Figura 82. Deserción universitaria de alumnos al reprobar segunda o tercera matrícula por período.

En su mayoría del total de 19291 de segunda y tercera matrícula un gran porcentaje se matricula en el período consecutivo del que reprueba este porcentaje varía entre 363 y 1854 y que con el paso de los períodos ha ido incrementando este valor como se observa en la Figura 83.

Normalmente los alumnos toman en el período consecutivo la misma materia que reprobaron, con excepción del período 201110 esto debido a que en los últimos

períodos se cambiaron las restricciones para matricularse en materia debido a que cada una tiene seguimiento como se muestra en la Figura 84.

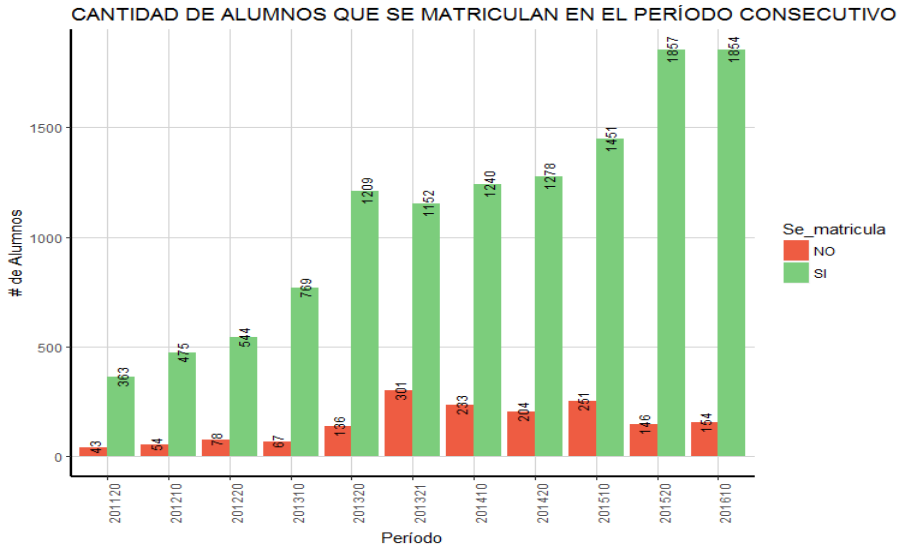

Figura 83. Cantidad de alumnos que matriculan en el período consecutivo del que reprueban una materia.

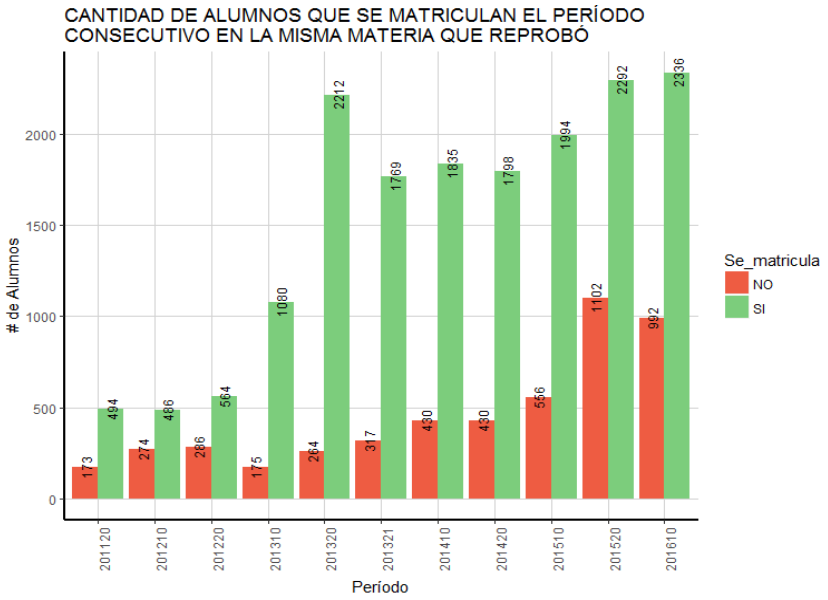

Figura 84. Cantidad de alumnos que se matriculan en el período consecutivo en la misma materia que reprobó

Si un alumno ha reprobado al menos una materia tiene mayor probabilidad de continuar reprobando como se muestra en la Figura 85, el 82.89% ha continuado reprobando mientras que tan solo 17.11% no volvió a reprobar nunca más.

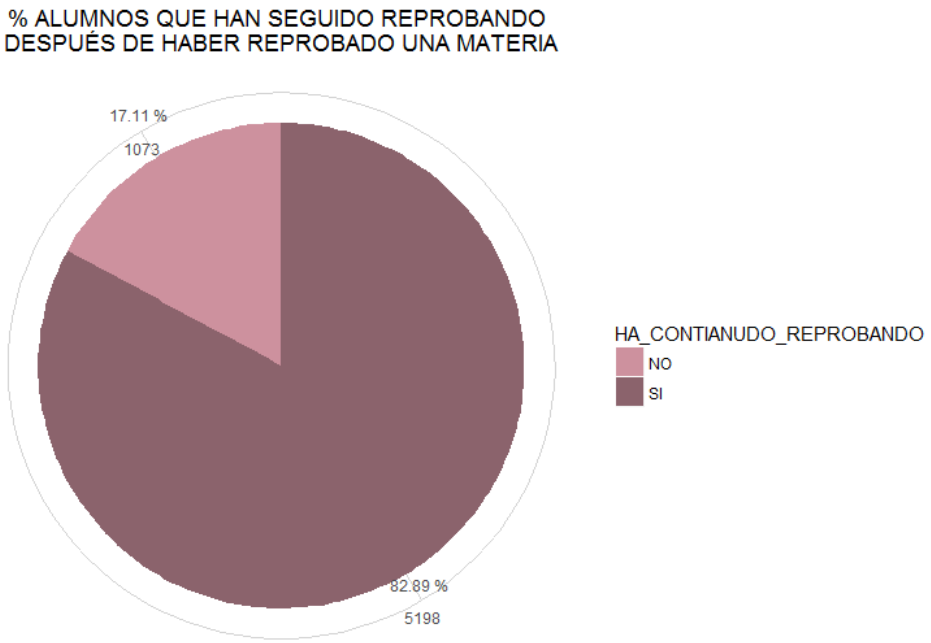

Figura 85. Porcentaje de alumnos que han seguido reprobando después de haber reprobado una materia.

e) Verificar la calidad de los datos

Para verificar la calidad de los datos se ha utilizado las herramientas Open Refine que permite agrupar valores con representaciones alternativas de lo mismo, haciendo que los datos tengan características manejables para su análisis y kettle4-profiling-datacleaner que es una herramienta que facilita la limpieza de los datos permitiendo detectar valores nulos y duplicados utilizando lógica difusa y los pesos de umbrales.

Para detectar los valores nulos por cada atributo, se debe exportar el archivo

(ORIGINAL TODOS ALUMNOS.xlsx) en la herramienta DataCleaner como se muestra en la Figura 86 y se selección la opción analizar. A continuación, se selecciona el archivo importado, se da click derecho y seleccionar la opción Quick Análisis como se observa en la Figura 100, después de unos segundos se desplegará varia información de los atributos cargados las cuales se detallan en las Tablas 7 y 8.

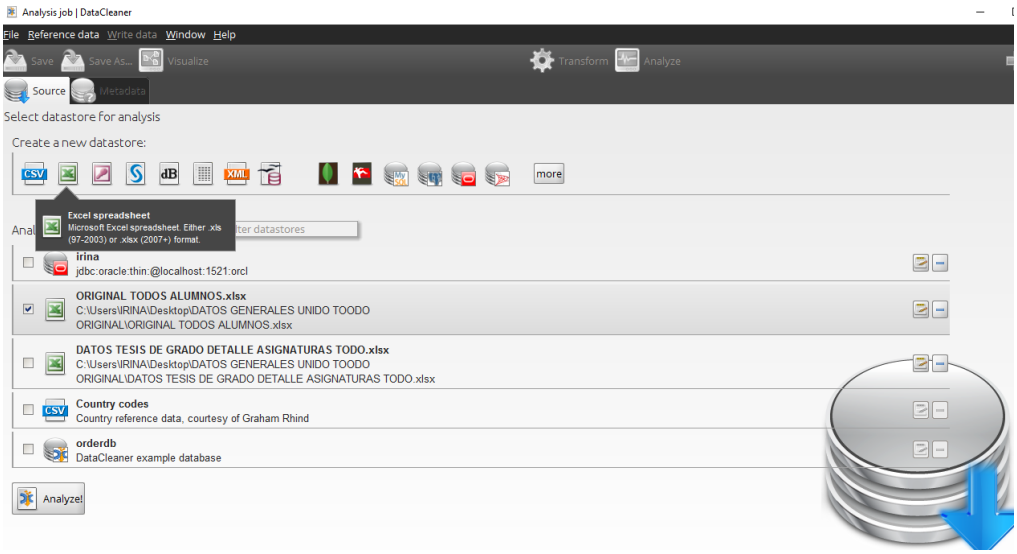

Figura 86. Exportación de archivos en la herramienta DataCleaner.

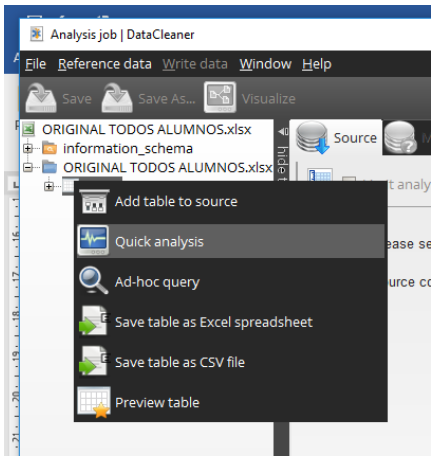

Figura 87. Análisis del archivo Excel.

Datos Generales del Alumno

El archivo original que tiene los datos generales del alumno tiene algunos atributos nulos como se detalla en la tabla 7, a pesar de ellos esta es una cantidad mínima en comparación con el total de 16220 alumnos.

Tabla 7.
Cantidad de atributos nulos de la tabla Alumnos

Cantidad de Filas = 16220			
ATRIBUTO	**NULOS**	**ATRIBUTO**	**NULOS**
Cédula	1	Dirección	1372
Nombres	0	Discapacidad	0
Edad	3	Tipo de Discapacidad	16182
Provincia de Residencia	1351	Carrera	0
Cantón de Residencia	1355	Período de Cohorte	1186
Parroquia de Residencia	1349	Período de Ingreso	0
Militar	0	Nacionalidad	15
Colegio	2352	Provincia de Nacimiento	2257
Etnia	22	Cantón de Nacimiento	2259
Estado Civil	177	Promedio del Colegio	4107
Género	0	Ingresos Familiares	16217

A continuación, se utiliza la herramienta Open Refine para verificar la similitud entre valores de un mismo atributo, para ello se exporta el archivo (ORIGINAL TODOS ALUMNOS.xlsx) y se crea el proyecto como se muestra en la Figura 88.

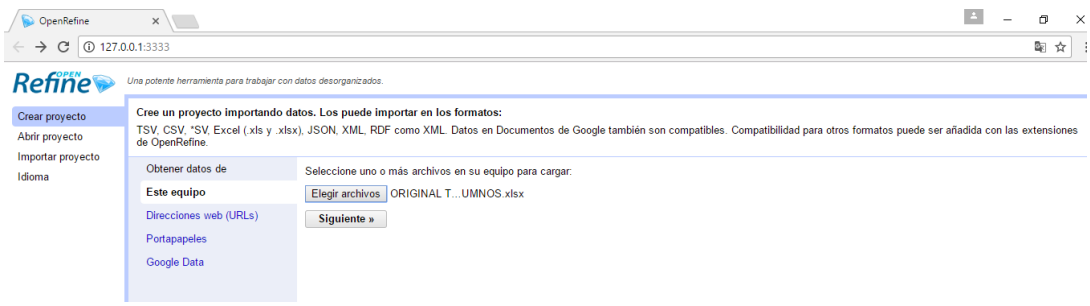

Figura 88. Crear un nuevo proyecto en OpenRefine

Una vez creado el proyecto se realiza una verificación de facetas de texto para cada atributo como se muestra en la Figura 89 y se selecciona la opción cluster o agrupación, la que genera un archivo con la agrupación de datos, se puede seleccionar entre 4 tipos de agrupación en este estudio se ha seleccionado metaphone3 que es un algoritmo que permite indexar palabras similares.

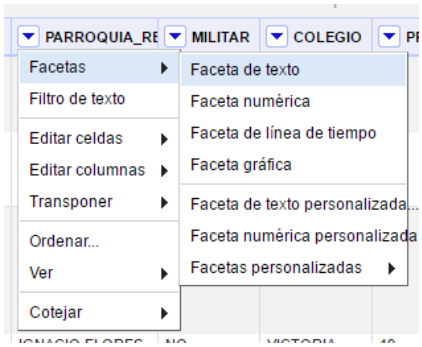

Figura 89. Facetas de texto para el atributo Parroquia en OpenRefine

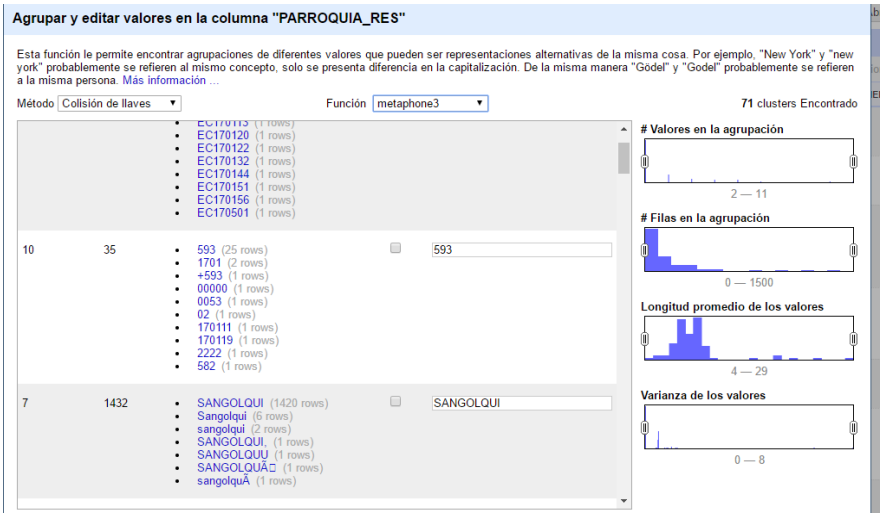

Figura 90. Cluster de parroquia de residencia del alumno en la herramienta OpenRefine

Como se observa en la Figura 90 la herramienta despliega 71 clusters., algunos de ellos

se detallan en la tabla 8, donde se ha encontrado que ciertos datos han sido ingresado al como códigos, datos basura o error en la escritura de datos.

Tabla 8.
Clusters encontrados de parroquia de resiendencia del alumno con la herramienta OpenRefine

PARROQUIA	
EC170105, EC071150, EC170103, EC170113, EC170120, EC170122, EC170132, EC170144, EC170151, EC170156, EC170501, 170111, 170119.	Códigos de parroquias
593, 1701, +593, 00000, 0053, 02, 2222, 582, NC, NO SE, nose, CHA, SP, CP	Datos basura
SANGOLQUI, Sangolqui, sangolqui, SANGILQUI, SANGOLQUI, SANGOLQUU, SANGOLQUÃ, sangolquÃ-	Error en la escritura de Datos
AMAGUA?A, AMAGUAÃ‘A, AMAGUAÑA, AMAGUAÃ?A, AMAGUA, ANAGUAÑA, AmaguaÃ±a	Error en la escritura de Datos
CHILLOGALLO, CHILLOGALL, Chillogall, Chillogallo, CHOLLOGALL, chillogall, CHIÃ‘Ã‘OGALL	Error en la escritura de Datos

Este proceso de agrupación se realizó para cada uno de los atributos del archivo de datos generales del alumno, sin embargo, tan solo se encontró agrupaciones para Parroquia de Residencia y Colegio, algunas de estas agrupaciones se detallan en la Tabla 11.

Tabla 9.
Clusters encontrados de colegio de residencia del alumno con la herramienta OpenRefine

COLEGIO	
ABDON CALDERON MUÃ‘OZ [JIPIJA], ABDON CALDERON MUÑOZ [JIPIJA]	Error en escritura de Datos
ABDON CALDERON MUÃ‘OZ LOJA, ABDON CALDERON MUÃ‘OZ [LOJA], ABDON CALDERON MUÑOZ [LOJA]	Error en escritura de Datos
ABDON CALDERON MUÃ‘OZ [EL ORO], ABDON CALDERON MUÑOZ [EL ORO]	Error en escritura de Datos
NUESTRA SEÃ‘ORA DEL ROSARIO, NUESTRA SEÑORA DEL ROSARIO	Error en escritura de Datos

Datos Académicos del Alumno

De la exploración realizada en DataCleaner para estos datos se encontró que los períodos 201120, 201210, 201321, 201410, 201510, 201620 tienen 52025, 51495, 43455, 41957, 44247, 43045 filas respectivamente y no poseen ningún atributo nulo, para los demás períodos que si tienen nulos se detalla a continuación en la Tabla 10.

Tabla 10.
Atributos nulos por período académico

Período 201110 *Cantidad de Filas = 51083*			
ATRIBUTO	**NULOS**	**ATRIBUTO**	**NULOS**
Docente	0	Período	0
Días	178	Campus	0
Código de Materias	0	Carrera	0
Materia	0	Hora de Fin	0
Nota	0	Cédula	0
Comentario	2	Nombres	0
Hora Inicio	0		
Período 201220 *Cantidad de Filas = 47616*			
ATRIBUTO	**NULOS**	**ATRIBUTO**	**NULOS**
Docente	0	Período	0
Días	1	Campus	0
Código de Materias	0	Carrera	0
Materia	0	Hora de Fin	1
Nota	0	Cédula	0
Comentario	0	Nombres	0
Hora Inicio	1		
Período 201310 *Cantidad de Filas = 46501*			
ATRIBUTO	**NULOS**	**ATRIBUTO**	**NULOS**
Docente	0	Período	0
Días	10	Campus	0
Código de Materias	0	Carrera	0
Materia	0	Hora de Fin	0
Nota	0	Cédula	0
Comentario	0	Nombres	0
Hora Inicio	0		
Período 201320 *Cantidad de Filas = 48315*			
ATRIBUTO	**NULOS**	**ATRIBUTO**	**NULOS**
Docente	0	Período	0
Días	2	Campus	0
Código de Materias	0	Carrera	0
Materia	0	Hora de Fin	0
Nota	0	Cédula	0

Comentario	0	Nombres	0
Hora Inicio	0		

Período 201420
Cantidad de Filas = 43045

ATRIBUTO	**NULOS**	**ATRIBUTO**	**NULOS**
Docente	0	Período	0
Días	1	Campus	0
Código de Materias	0	Carrera	0
Materia	0	Hora de Fin	0
Nota		Cédula	0
Comentario	18	Nombres	0
Hora de Inicio	0		

Período 201520
Cantidad de Filas = 49064

ATRIBUTO	**NULOS**	**ATRIBUTO**	**NULOS**
Docente	0	Período	0
Días	50	Campus	0
Código de Materias	0	Carrera	0
Materia	0	Hora de Fin	37
Nota	0	Cédula	0
Comentario	0	Nombres	0
Hora de Inicio	37		

Período 201610
Cantidad de Filas = 43045

ATRIBUTO	**NULOS**	**ATRIBUTO**	**NULOS**
Docente	0	Período	0
Días	30	Campus	0
Código de Materias	0	Carrera	0
Materia	0	Hora de Fin	30
Nota	0	Cédula	0
Comentario	0	Nombres	0
Hora de Inicio	30		0

Las horas de inicio y fin de las asignaturas se encuentran mal registradas, porque se toma como diferentes las horas que cambian de 1 a 5 minutos, por ejemplo: 0700 = 0701 = 0706 o 0715 = 0716 sin embargo, su hora de inicio sería 0700 y 0715 respectivamente. Otro problema que se encontró dentro de los datos académicos es la existencia de valores duplicados para las materias en las que su hora de inicio y fin cambia entre días, dependiendo de la cantidad de créditos puede existir entre 2 a 4 duplicados de una misma materia, esto también ocurre en las materias que tienen el mismo día, pero por los recesos entre horas se dividen en varios registros. Por ejemplo, Calculo Vectorial empieza el día martes a la 9h45 am y el día miércoles inicia a las

09h31 am por lo que existen dos registros diferentes para el mismo alumno como se muestra en la Tabla 11.

Tabla 11.
Datos duplicados en los archivos académicos del alumno por la diferencia de horario entre días

CEDULA	NOMBRES	PERÍODO	MATERIA	NOTA	COMENTARIO	HORA INICIO	HORA FIN	DOCENTE	DIAS
1725299703	ACOSTA CALERO, ALEJANDRO RAFAE	PREGRADO S-II OCT16-FEB17	CALCULO VECTORIAL	17.7	APROBADO/ EVALUACION	0945	1145	ALVEAR VARGAS, MARCO ANTONIO	M
1725299703	ACOSTA CALERO, ALEJANDRO RAFAE	PREGRADO S-II OCT16-FEB17	CALCULO VECTORIAL	17.7	APROBADO/ EVALUACION	0931	1129	ALVEAR VARGAS, MARCO ANTONIO	W

4.1.3. PREPARACIÓN DE LOS DATOS

Esta permite preparar los datos para llevar a cabo el proceso de minería de datos, es decir, seleccionar, mejorar la calidad, fusionar y eliminar registros y/o atributos para obtener el conjunto de datos que será tomado para aplicar la técnica de minería de datos, de tal manera que se pueda obtener mayor precisión en la generación del modelo.

a) <u>Seleccionar Datos</u>

La selección de datos se realizó en base a campos y registros de las distintas tablas debido a que algunos de estos no son necesarios para los objetivos de minería de datos del presente proyecto, por lo que no serán tomados en cuenta; éstos se detallan a continuación:

En la tabla ALUMNO, los campos: Promedio_colegio, Canton_nac, Canton_res, Dirección e Ingresos_familiares son variables que no influyen en la deserción de los alumnos, como se observa en el punto anterior no están completos al 100% generando así demasiado ruido en la construcción del modelo. En cuanto a los registros de alumnos se eliminaron a todos los que poseían carreras distintas

a las ofertadas en ESPE-MATRIZ y a distancia, se eliminó además los colegios nulos y alumnos de los cuales no se registraban notas. Además se consideró únicamente a los alumnos cuyo período de cohorte fue mayor o igual al 201110 con el objetivo de tener los registros históricos completos de los estudiantes desde su ingreso a primer nivel, después de haber realizado está selección se disminuyó el tamaño de muestra a 8264 estudiantes.

En la tabla NOTA se suprimieron los campos: días_semana los cuales fueron sustituidos por el campo días y la variable Campus porque todos los alumnos pertenecen a ESPE-MATRIZ. En cuanto a registros se eliminaron los que no registraban información personal en la tabla alumno, reduciéndose la muestra a 267050 registros de notas.

b) Limpieza de Datos

A continuación, se establece una solución a los registros que poseen valores erróneos, inconsistentes o faltantes encontrados en la calidad de los datos y que tienen las distintas tablas y prepararlos para el modelado.

Primero se procedió a convertir en mayúsculas todos los datos para mantener un mismo formato y se realizaron cambios de las palabras al no haber sido exportados en el formato UTF-8 las "Ñ" y tildes se encontraron distorsionadas en las siguientes formas (Ã‘),(Ã±),(?), (Ã), (Ã“), (Ã?), (Ã- í) por lo cual se corrigieron estos errores en todos los registros. Adicionalmente se cambió todas las tildes por su letra normal debido a que existe problemas con las Ñ y tildes al importar en formato .arf para ser utilizado por WEKA.

Para los datos nulos de la tabla ALUMNO del campo edad se procedió a obtenerlo en base a su cedula de identidad al igual que los de edades incoherentee. Para los alumnos que no tienen discapacidad y en Tipo_discapacidad son nulos se los llenó con el valor "ninguno". Se estandarizó las nacionalidades debido a que algunas se encontraban en femenino y otras en masculino como por ejemplo "Colombiana" y

"Colombiano". Tomando en cuenta la clusterización realizada en la verificación de la calidad de los datos, para el campo Colegios se modificó a los que tenían error en su escritura como se puede ver en la Tabla 8.

En la tabla NOTA se modificó las horas de inicio y fin de las asignaturas como en el caso de las 0716 por 0715 o 0731 por 0730 etc. Además, se fusionó los registros duplicados de los alumnos dónde una materia del mismo período académico podía tener desde 2 hasta 4 registros como se explica en la verificación de la calidad de los datos, a los cuales se les aplicó el formato como se explica en las Tablas 12 y 13 para tener un solo registro de alumno por materia en un período académico. En el campo materia se cambió el nombre de las materias que estaban escritas de diferente manera por ejemplo "EDU. FÃ• SICA-BÃ• SQUET" y "EDU. FÃ• SICA-BÃ• SKET" se le asignó EDU. FISICA BASQUET, el mismo cambio se realizó para el campo comentario y docente.

Tabla 12.
Antes de la unión de los datos duplicados en la tabla Nota

CÉDULA	PERÍODO	MATERIA	NOTA	H-I	H-F	D	DOCENTE
0202181806	PREGRADO S-II OCT16-FEB17	INST. INDUSTRIAL MECANICA	14.19	1200	1259	F	ECHEVERRIA YANEZ, LUIS MANUEL
0202181806	PREGRADO S-II OCT16-FEB17	INST. INDUSTRIAL MECANICA	14.19	0930	1130	F	ECHEVERRIA YANEZ, LUIS MANUEL

Tabla 13.
Después de la unión de los datos duplicados en la tabla Nota

CÉDULA	PERÍODO	MATERIA	NOTA	H-I	H-F	D	DOCENTE
0202181806	PREGRADO S-II OCT16-FEB17	INST. INDUSTRIAL MECANICA	14.19	**0930**/1200	**1130**/1300	F/**F**	ECHEVERRIA YANEZ, LUIS MANUEL

c) Construir E Integrar Los Datos

Una vez finalizada la fase de limpieza de datos, se ha realizado la agrupación y normalización de los datos. Para lo cual se empezó creando un modelo relacional de base de datos donde se crearon 18 nuevas tablas (Asignatura, Colegio, Carrera, Comentario, Departamento, Docente, Estado Civil, Etnia, Género, Cantón, Horario, Nacionalidad, Parroquia, Período, Provincia, Régimen escolar, Sostenimiento, Tipo de Discapacidad) que se originan y relacionan con las tablas Alumno y Nota las cuales se describen a continuación en las tablas 14 – 33. En esta fase de uso la herramienta Pentaho Data Integration para crear el Data Warehouse empleando Kettle mediante archivos ETL (Extract – Transform – Load).

En primer lugar, se debe crear una transformación donde se irá asignando los pasos o Steps necesarios construir nuestra base de datos, los steps son la unidad mínima de trabajo de una Transformación encargados de ejecutar tareas específicas como leer ficheros o tablas, validaciones, transformar datos, insertar datos en la base, etc.

A continuación, se explica el proceso que se realiza para crear la tabla Nacionalidad. La transformación inicia con una entrada, que en este caso sería la lectura desde un archivo Excel, después se agrupa por una o más variables específicas para esta tabla sería “nacionalidad”, se filtran los registros donde la nacionalidad no sea nula, luego se ordena los datos alfabéticamente y se agrega una secuencia ascendente para cada valor llamada ID_NACIONALIDAD. Finalmente, la salida se conecta con la base de datos crea la tabla NACIONALIDAD e inserta los datos recuperados y transformados como se observa en la Figura 91. Este proceso se realizó para crear cada una de las nuevas tablas antes mencionadas.

Figura 91. Entrada Proceso y Salida de una transformación en Kettle

Finalmente para relacionar las nuevas tablas creadas con la tabla Alumno y Notas se tiene dos entradas, la primera es la nueva tabla Nacionalidad y la segunda es la tabla Nota se ordena cada una en función de la variable que se va a relacionar (Nacionalidad) y finalmente se realiza un merge join que realiza operaciones lógicas de combinación interna (inner join) o externa izquierda (left outer join), se realiza el mismo proceso para cada una de las variables que se quiera relacionar con otra tablas.

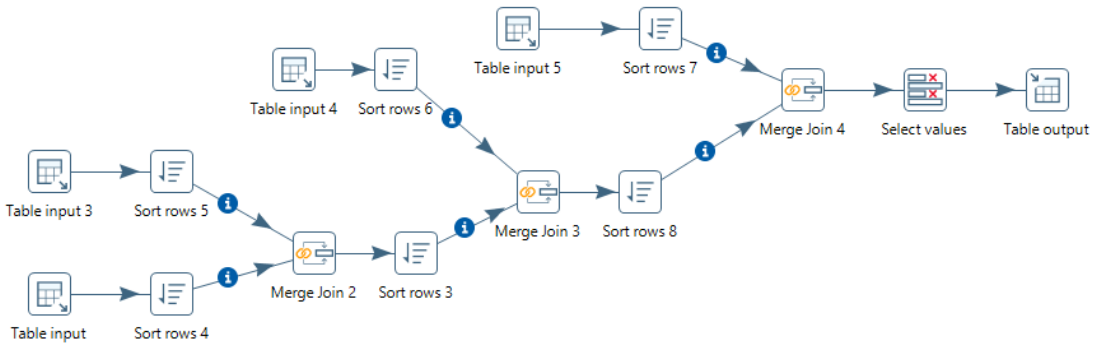

Figura 92. Relación entre las tablas creadas y la tabla nota

Tabla 14.
Tabla Alumno

TABLA	DESCRIPCIÓN	Nº REGISTROS
ALUMNO	Esta tabla contiene la información personal de los alumnos.	8264
CAMPO	**DESCRIPCIÓN**	
NOMBRES	Nombres y Apellidos del alumno.	
CEDULA	Cédula del alumno.	
ID_COLEGIO	Código del colegio en el que se graduó el alumno.	
DISCAPACIDAD	Discapacidad del alumno, variable con valor SI o NO.	
EDAD	Edad de vida del alumno.	
MILITAR	Variable con valor SI o NO en caso de que el alumno pertenezca a las Fuerzas Armadas.	
PERIODO_COHORTE	Período en que el alumno ingresa a primer nivel de la carrera.	
PERIODO_INGRESO	Período en que el alumno ingresa a Prepolitécnico.	
PROMEDIO_COLEGIO	Promedio con el que se graduó el alumno.	
ID_ETNIA	Código del conjunto de comunidad lingüística, cultural o raza al que pertenece el alumno.	
ID_GÉNERO	Código del género del alumno.	
ID_TIPO_DISCAPACIDAD	Código del tipo de Discapacidad del alumno será un valor vacío en caso de tener una discapacidad.	
ID_ESTADO_CIVIL	Código de la condición del alumno según el registro civil en función de si tiene o no pareja y su situación legal	
ID_NACIONALIDAD	Código del País de nacimiento del alumno.	
ID_CARRERA	Código de la Carrera que sigue el alumno.	

Tabla 15.
Variables de la tabla Nota

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
NOTA	Información de las notas de los alumnos por cada materia	267050
CAMPO	**DESCRIPCIÓN**	
ID_NOTA	Código del registro de la nota	
CEDULA	Cédula del alumno	

ID_CARRERA	Código de la carrera del alumno en la que tomó la materia.
ID_PERIODO	Código del período en el que el alumno tomó la materia
ID_ASIGNATURA	Código de la materia
ID_HORARIO	Código del horario de la materia
ID_DOCENTE	Código del docente que impartió dicha materia al alumno
NOTA	Nota con la cual el alumno terminó la materia
ID_COMENTARIO	Código del comentario que indica si el alumno aprobó, reprobó o se retiró de la materia
ID_DEPARTAMENTO	Código del departamento al que pertenece la materia

La tabla Colegio tiene todos las Instituciones del Ecuador al 2016, tanto jardines, escuelas y colegios, que han sido obtenidas de AMIE (2016) y cuyas características se encuentran registradas en la tabla Colegio.

Tabla 16.
Tabla Colegio

TABLA	**DESCRIPCIÓN**	**NÚMERO DE REGISTROS**
COLEGIO	Información de los todos los colegios del Ecuador	29065
CAMPO	**DESCRIPCIÓN**	
NOMBRE	Nombre del colegio	
ID_PROVINCIA	Código de la provincia de localización.	
ID_CANTON	Código del cantón de localización	
ID_PARROQUIA	Código de la parroquia de localización	
NUMDOCENTES	Cantidad de docentes con las que cuenta	
ID_JORNADA	Código de la jornada en las que trabaja matutina o vespertina	
ID_REGIMENESCOLAR	Código del régimen escolar con el cual trabaja	
ID_SOSTENIMIENTO	Código del sostenimiento privada, fiscal,	

	ficomisional
ID_COLEGIO	Código del colegio
NUMALUMNOS	Cantidad de alumnos para los cuales el colegio tiene capacidad.

Tabla 17.
Tabla Carrera

TABLA	**DESCRIPCIÓN**	**NÚMERO DE REGISTROS**
CARRERA	Información de las carreras que oferta la UFA-ESPE	50
CAMPO	**DESCRIPCIÓN**	
ID_CARRERA	Código de la carrera	
CARRERA	Nombre de la carrera	
ID_DEPARTAMENTO	Código del departamento al que pertenece la carrera	
TIPO_CARRERA	Tipo de la carrera(Técnica/Administrativa/Humanística)	

Tabla 18.
Tabla Comentario

TABLA	**DESCRIPCIÓN**	**NÚMERO DE REGISTROS**
COMENTARIO	Información de los distintos comentarios con los cuales el alumno aprueba, reprueba o se retira de una materia	19
CAMPO	**DESCRIPCIÓN**	
ID_COMENTARIO	Código del comentario	
COMENTARIO	Descripción del comentario	
STATUS	Tipo del comentario	

Tabla 19.
Tabla Departamento

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
DEPARTAMENTO	Información de los departamentos por los cuales está dividida la UFA-ESPE matriz.	10
CAMPO	**DESCRIPCIÓN**	
ID_DEPARTAMENTO	Código del departamento	
DESCRIPCIÓN	Nombre del departamento	

Tabla 20.
Tabla Docente

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
DOCENTE	Información de los docentes que impartes las distintas asignaturas.	1227
CAMPO	**DESCRIPCIÓN**	
ID_DOCENTE	Código del docente	
DOCENTE	Nombre del docente	

Tabla 21.
Tabla de Estado Civil

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
ESTADO CIVIL	Información de estados civiles que tiene una persona.	8
CAMPO	**DESCRIPCIÓN**	

ID_ ESTADOCIVIL	Código del estado civil
ESTADO_CIVIL	Descripción del estado.

Tabla 22.
Tabla Etnia

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
ETNIA	Información de las etnias del Ecuador	8
CAMPO	**DESCRIPCIÓN**	
ID_ ETNIA	Código de la etnia.	
ETNIA	Descripción de la etnia.	

Tabla 23.
Tabla Género

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
GÉNERO	Información de los géneros de los alumnos	2
CAMPO	**DESCRIPCIÓN**	
ID_ GENERO	Código del género	
GENERO	Nombre del género	

Tabla 24.
Tabla Asignatura

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
ASIGNATURA	Información de las materias que toman los alumnos.	1064
CAMPO	**DESCRIPCIÓN**	
ID_ASIGNATURA	Código de la materia	
MATERIA	Nombre de la materia	

ID_DEPARTAMENTO	Código del departamento al que pertenece la materia

Tabla 25.
Tabla Cantón

TABLA	**DESCRIPCIÓN**	**NÚMERO DE REGISTROS**
CANTON	Información de los cantones del Ecuador.	224
CAMPO	**DESCRIPCIÓN**	
ID_CANTON	Código del cantón	
ID_PROVINCIA	Código de la provincia a la que pertenece el cantón	
CANTON	Nombre del cantón	

El campo días de la nueva tabla Horario se generó después de la unión de los campos Lunes, Martes, Miércoles, Jueves, Viernes, Sábado y Domingo descritos en la Tabla 7.

Tabla 26.
Tabla Horario

TABLA	**DESCRIPCIÓN**	**NÚMERO DE REGISTROS**
HORARIO	Información de los horarios de las distintas materias.	7016
CAMPO	**DESCRIPCIÓN**	
ID_HORARIO	Código del horario	
ID_ASIGNATURA	Código de la materia	
MATERIA	Nombre de la materia	
HORA_INICIO	Hora en la que comienza la clase	
HORA_FIN	Hora en la que termina la clase	

DIAS	Días en los que se imparte

Tabla 27.
Tabla Nacionalidad

TABLA	**DESCRIPCIÓN**	**NÚMERO DE REGISTROS**
NACIONALIDAD	Información de nacionalidades de los alumnos de la UFA-ESPE de los últimos cinco años	16
CAMPO	**DESCRIPCIÓN**	
ID_ NACIONALIDAD	Código de la nacionalidad	
NACIONALIDAD	Nombre de la nacionalidad	

Tabla 28.
Tabla Parroquia

TABLA	**DESCRIPCIÓN**	**NÚMERO DE REGISTROS**
PARROQUIA	Información de las parroquias del Ecuador	1399
CAMPO	**DESCRIPCIÓN**	
ID_PARROQUIA	Código de la parroquia	
ID_PROVINCIA	Código de la provincia a la que pertenece la parroquia	
ID_CANTON	Código del cantón a la que pertenece la parroquia	
PARROQUIA	Nombre de la parroquia	

Tabla 29.
Tabla Período

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
PERÍODO	Información de todos los períodos que ha tenido la UFA-ESPE	452
CAMPO	**DESCRIPCIÓN**	
INICIO	Fecha de inicio del período	
FIN	Fecha de finalización del período	
ANIO	Año del período	
TIPO_PERIODO	Especifica si el período fue semestral o anual	
ID_PERIODO	Código del período	
DESCRIPCION	Nombre del período	

Tabla 30.
Tabla Provincia

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
PROVINCIA	Información de las provincias del Ecuador	25
CAMPO	**DESCRIPCIÓN**	
ID_ PROVINCIA	Código de la provincia	
PROVINCIA	Nombre de la provincia	

Tabla 31.
Tabla Régimen Escolar

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
REGIMEN_ESCOLAR	Información del régimen escolar de los colegios del Ecuador	5
CAMPO	**DESCRIPCIÓN**	
ID_ REGIMENESCOLAR	Código del régimen escolar	
REGIMENESCOLAR	Descripción del régimen escolar	

Tabla 32.
Tabla Sostenimiento

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
SOSTENIMIENTO	Información del sostenimiento de los colegios del Ecuador	4
CAMPO	**DESCRIPCIÓN**	
ID_ SOSTENIMIENTO	Código del sostenimiento	
SOSTENIMIENTO	Descripción del sostenimiento del colegio	
TIPO_ SOSTENIMIENTO	Tipo del sostenimiento del colegio	

Tabla 33.
Tabla Tipo de Discapacidad

TABLA	DESCRIPCIÓN	NÚMERO DE REGISTROS
Tipo de Discapacidad	Información de los distintos tipos de discapacidad que puede tener un alumno	5
CAMPO	**DESCRIPCIÓN**	
ID_TIPODISCAPACIDAD	Código del tipo de discapacidad	
TIPO_DISCAPACIDAD	Descripción del tipo de discapacidad	

A continuación, se procedió a integrar en una sola tabla todas las variables para que puedan ser utilizadas en el modelado a la cual llamaremos MACRO_ENTRENAMIENTO, también se creó una tabla llamada MACRO_PRUEBA la misma que contiene los datos de los alumnos actuales con el último período 201620 y que servirán para la implementación del proyecto. Se calcularon algunas variables a partir de la información recopilada las cuales son:

- Cantidad Cambio: Son el número de veces que el alumno se ha cambiado de carrera dentro de la misma universidad.
- Créditos: Es el número de créditos actuales que toma el alumno en el semestre en el que transcurre.
- Créditos Materia: Es la cantidad de créditos que tiene la materia en la que se matriculó el alumno.
- Curso Alumno: Es el total de alumnos por curso de la materia matriculada.
- Deserción: Es una variable de si o no que se calculó en base a los alumnos que reprueban segunda o tercera matrícula y que no se han matriculado de nuevo en la universidad.
- Día Viernes: Es una variable de si o no que generaliza si la materia se imparte el día viernes o no.
- Km: En base a la parroquia de residencia se calculó los kilómetros aproximados que existe entre la universidad y la parroquia.
- Materias Cursadas: Es le número de materias que han sido aprobadas por el alumno antes del período actual.
- Materias Repetidas: Es le número de materias que han sido reprobadas por el alumno antes del período actual.
- Matricula Siguiente: Es una variable cuyos valores son si - no que se calculó de los alumnos que toman al semestre consecutivo una materia que la habían reprobado.
- Nivel: Es el total de crédito de materias aprobadas que ha toma el alumno durante toda su permanencia en la universidad.

- Período Tipo: Puede ser primer período o segundo período del año.
- Promedio Alumno: Es el promedio del semestre anterior del alumno.
- Tiempo: En base a la parroquia de residencia se asignó un tiempo aproximado que se tardaría el alumno en llegar a la universidad.
- Tipo Alumno: Puede ser Senescyt en el caso de que su período de cohorte sea superior o igual al 201310 o No Senescyt para los que su período de cohorte es menor o igual 201220 y alumnos militares.
- Tipo Carrera: Dependiendo de la carrera pueden ser Administrativas, Humanísticas o Técnicas.
- Tipo Horario: Considerando las horas de inicio de la materia la materia puede tener un horario en la mañana (07h00 – 12h00), media mañana-tarde (12h01 - 1700) o noche (17h01-21h30).
- Tipo matricula: Puede ser primera, segunda o tercera dependiendo de la materia que tome.

Adicionalmente se discretizaron algunas variables numéricas ya que éstas al ser procesadas en la herramienta WEKA en los árboles de decisión van a generar únicamente dos ramas, sin embargo, si se discretiza, es decir si se crean rangos o intervalos se va a generar un camino de decisión por cada opción. Dichas variables se describen a continuación:

Tabla 34.
Discretización de la variable créditos

CLUSTER	Rango 1	Rango 2	Rango 3	Rango 4	Rango 5
CRÉDITOS	0 – 12	13 – 21	22 - 26	27 – 31	>31

Nota: Hace referencia a la cantidad de créditos que el alumno toma en el semestre

Tabla 35.
Discretización de la variable Créditos Materia

CLUSTER	Rango 1	Rango 2
CRED_MATERIA	1 - 4	>4

Nota: Hace referencia a la cantidad de créditos de una materia

Tabla 36.
Discretización de la variable curso

CLUSTER	Rango 1	Rango 2	Rango 3	Rango 4	Rango 5	Rango 6
CURSO	1 - 13	14 – 20	21 - 24	25– 28	28 - 32	>32

Nota: Hace referencia a la cantidad de alumnos que existe en un curso

Tabla 37.
Discretización de la variable edad

CLUSTER	Rango 1	Rango 2	Rango 3	Rango 4
EDAD	< 23	23 - 25	26 - 30	>30

Nota: Las edad está dada en años.

Tabla 38.
Discretización de la variable materias cursadas

CLUSTER	Rango 1	Rango 2	Rango 3	Rango 4	Rango 5	Rango 6
MATERIAS CURSADAS	0 – 4	5 - 18	19 - 32	33 - 45	46 - 58	>58

Nota: No se consideran las materias aprobadas en prepolitécnico.

Tabla 39.
Discretización de la variable materias repetidas

CLUSTER	Rango 1	Rango 2	Rango 3	Rango 4	Rango 5
MATERIAS REPETIDAS	0	1 - 2	3 - 6	7 - 11	>11

Nota: No se consideran las materias reprobadas en prepolitecnico.

Tabla 40.
Discretización de la variable nivel

CLUSTER	Rango 1	Rango 2	Rango 3	Rango 4	Rango 5	Rango 6
MATERIAS CURSADAS	0 – 14	15 - 67	68 - 117	118 - 165	166 - 215	>215

Nota: Se considera la suma de créditos aprobados del alumno.

Tabla 41.
Discretización de la variable promedio alumno

CLUSTER PROMEDIO ALUMNO	Rango 1	Rango 2	Rango 3	Rango 4	Rango 5	Rango 6
	EXCELENTE >17.44	MUY BUENO 16.12 - 17.44	BUENO 14.99 - 16.11	MALO 13.56 - 14.98	MUY MALO 10.60 - 13.55	REGULAR 0 -10.59

Tabla 42.
Discretización de la variable provincia colegio y provincia nacimiento

CLUSTER PROV_COLEGIO Y PROV_NAC	Rango 1	Rango 2
	PICHINCHA	OTRA

4.1.4. MODELADO

a) Escoger la técnica de modelado

Se utilizará la herramienta WEKA para generar el modelo de minería de datos, para lo cual se seleccionarán las técnicas de clasificación que nos ofrece esta herramienta, las cuales van acorde a los objetivos de la minería de datos. Los modelos que mejor se adaptan son los supervisados, porque la predicción se va a realizar a partir de un conjunto de datos cuyo resultado ya se conoce, sin embargo, se realizarán pruebas con las dos técnicas (supervisadas y no supervisadas) con el fin de constatar que los atributos seleccionados tienen un margen aceptable de precisión.

De las técnicas supervisas se utilizarán Árboles de Decisión y Regresión Lineal, porque son técnicas de clasificación que generan resultados que permiten realizar pruebas de caja blanca entregando un modelo comprensible para el usuario y de las técnicas no supervisadas se utilizará Clasificadores Lazy (Ktar).

b) Generar el plan de prueba

Para evaluar el modelo, se seleccionará Percentage Split que permite dividir previamente en dos partes (entrenamiento y prueba) al conjunto de datos ingresado, para este proyecto el 80% servirá para el entrenamiento de los modelos y el 20% restante será para el testeo del modelo.

WEKA provee algunas métricas para medir la calidad del modelo, sin embargo, las que se tomarán en cuenta son: las instancias correctamente clasificadas, error cuadrático medio, error absoluto medio y la matriz de confusión. A continuación, se describe cada uno de ellos para su mejor comprensión.

- Instancias correctamente clasificadas: Expresa en porcentaje el total de aciertos del modelo, es decir es la suma de la diagonal que entrega la matriz de confusión en porcentaje.
- El error cuadrático medio (RMSE) calcula la diferencia entre valores predichos por el modelo y los valores reales de donde se partió para la creación del modelo y los valores reales a partir de los cuales se ha creado el modelo y se calcula de la siguiente manera:

Ecuación 4. Error cuadrático medio

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Donde:

- y_j es un vector de valores calculados.
- $\hat{y}_j$ es un vector de valores reales
- n es el número de observaciones

- El error absoluto medio (MAE) al igual que la fórmula anterior sirve para

calcular la diferencia entre las predicciones realizadas por un estimador y los valores reales. La diferencia entre las dos aparece del problema que tiene el error cuadrático medio, que cuando de al cuadrado la diferencia se da más peso a los errores más extremos, afectando al resultado final, con la utilización del error absoluto medio se elimina dicho problema." (GALÁN, 2016) y se lo calcula de la siguiente manera:

Ecuación 5. Error absoluto medio

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

Donde:

- y_j es un vector de valores calculados.
- $\hat{y}_j$ es un vector de valores reales
- n es el número de observaciones

- Matriz de Confusión: las columnas representan el número de atributos, en donde los valores de la diagonal corresponden a los clasificados correctamente mientras que los demás reflejan los errores producidos en la clasificación.

c) Construir el modelo y evaluar el modelo

En este apartado se construirán dos tipos de modelos, el primero podrá predecir si un alumno es propenso a desertar (reprobar) o no una materia y el segundo podrá predecir si un alumno desertará de la universidad.

Para predecir si un alumno es propenso a desertar (reprobar) o no una materia se creará un modelo por cada departamento del estudiante el mismo que se dividirá en dos subconjunto de datos que los llamaremos Etapas como se observa en las Tablas 43 - 50 con el objetivo de disminuir el conjunto de datos (atributos irrelevantes o redundantes) de tal forma que no afecte significativamente al modelo y se pueda reducir el tiempo de procesamiento de

entrenamiento y coste computacional incrementando la precisión de dicho modelo, basándose en la selección de atributos que ofrece la herramienta WEKA como se observa en la Figura 93.

Como atributo evaluador se seleccionó CorrelationAttributeEval como se muestra en la Figura 93, que evalúa la correlación que tienen cada uno de los atributos con la clase seleccionada, para este modelo será la relación que tienen los atributos con la clase STATUS (Reprobado o Aprobado). Y como método de búsqueda se seleccionó Ranker que se encarga ordenar los atributos y luego construye un subconjunto de manera jerárquica del más al menos influyente como se observa en las Figuras 94 y 95.

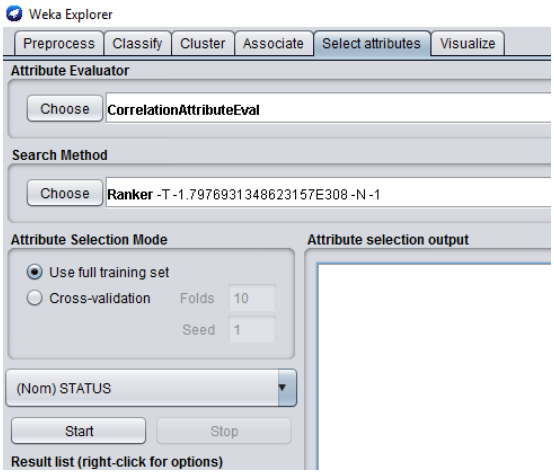

Figura 93. Selección de Atributos Evaluador y Búsqueda en WEKA

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 44 STATUS):
        Information Gain Ranking Filter

Ranked attributes:
 0.5704229244560655    39 PROMEDIO_ALUMNO
 0.5239778700610621    31 MATERIAS_REP_ANT
 0.5147779035212585    12 CLUSTER_PROMEDIO
 0.514549797935328      9 CLUSTER_MATERIAS_REP_ANT
 0.3467429595001299    13 COLEGIO
 0.31120477286140324   22 DOCENTE
 0.2578565339409725    29 MATERIA
 0.09614816210196031   17 DEP_ASIGNATURA
 0.08456308619275055    3 CARRERA
 0.07909109692709426   37 PERIODO_INGRESO
 0.07149002660174897   35 PARROQUIA
 0.06498851635664538    1 CANT_COLEGIO
 0.05878641289678965   18 DEPCARRERA

   0.05614483177373508   28 KM
   0.05459668966123066   27 HORA_INICIO
   0.05391197661250002   20 DIAS
   0.05168882042629275   36 PERIODO_COHORTE
   0.05043847965657811   47 TIPO_CARRERA
   0.04984349841801983   34 NIVEL
   0.04928215496203103   30 MATERIAS_CURSADAS
   0.04615871978406672   50 TIPO_MATRICULA
   0.04465347643199657   14 CRED_MATERIA
   0.04186918819357166   41 PROVINCIA_NAC
   0.04179183623214633    4 CLUSTER_CRED_MATERIA
   0.03579326533961869   40 PROV_COLEGIO
   0.03358722146703019   15 CREDITOS
   0.03335240363092229   49 TIPO_HORARIO
   0.03266745482334377   46 TIPO_ALUMNO
   0.02430630196598293    5 CLUSTER_CREDITOS
   0.02393499820205369   23 EDAD
   0.02166537512642719   32 MILITAR
   0.02005372322448185    7 CLUSTER_EDAD
   0.01683921656155662   26 HORA_FIN
   0.01669426982271693    8 CLUSTER_MATERIAS_CURSADAS
    0.02005372322448185    7 CLUSTER_EDAD
    0.01683921656155662   26 HORA_FIN
    0.01669426982271693    8 CLUSTER_MATERIAS_CURSADAS
    0.01456689205318518   11 CLUSTER_NIVEL
    0.00636534053782922    2 CANTIDAD_CAMBIO
    0.00564437736267998   16 CURSO_ALUMNO
    0.00533551473585914   25 ETNIA
    0.00475450330848959   45 TIEMPO
    0.00410488578135892   43 SOSTENIMIENTO
    0.00409059581303894   19 DIA_VIERNES
    0.00375734771694425   51 TIPO_SOSTENIMIENTO
    0.00281140692154569   24 ESTADO_CIVIL
    0.00228992838989783   42 REGIMENESCOLAR
    0.00153684971241996   38 PERIODO_TIPO
    0.0013487782567333    10 CLUSTER_NACIONALIDAD
    0.0013487782567333    33 NACIONALIDAD
    0.00116941381659885    6 CLUSTER_CURSO
    0.00001181188960109   48 TIPO_DISCAPACIDAD
   -0.00000000000000255   21 DISCAPACIDAD

  Selected attributes: 39,31,12,9,13,22,29,17,3,37,35,1,18,28,
    ,27,20,36,47,34,30,50,14,41,4,40,15,49,46,5,23,32,7,26,8,11,2,16,25,45,43,19,51,
                         24,42,38,10,33,6,48,21 : 50
```

Figura 94. Selección de Atributos mediante la herramienta WEKA para el Modelo1

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 18 DESERCION):
        Correlation Ranking Filter
Ranked attributes:
 0.53745   37 PROMEDIO_ALUMNO
 0.43782   31 MATERIAS_REP_ANT
 0.24455   38 PROV_COLEGIO
 0.2341    12 CLUSTER_PROMEDIO
 0.23224   39 PROVINCIA_NAC
 0.22294   47 TIPO_MATRICULA
 0.22139    9 CLUSTER_MATERIAS_REP_ANT
 0.20421    2 CANTIDAD_CAMBIO
 0.18765    1 CANT_COLEGIO
 0.175     30 MATERIAS_CURSADAS
 0.16613   34 NIVEL
 0.16144   48 TIPO_SOSTENIMIENTO
 0.12681   32 MILITAR
 0.11746   14 CREDITOS
 0.11437    4 CLUSTER_CRED_MATERIA
 0.10996   41 SOSTENIMIENTO
 0.10571   43 TIPO_ALUMNO
 0.09214    8 CLUSTER_MATERIAS_CURSADAS
 0.09197   11 CLUSTER_NIVEL
 0.08774   40 REGIMENESCOLAR
 0.08498   16 DEP_ASIGNATURA
 0.07591   35 PERIODO_COHORTE
 0.06948   42 TIEMPO
 0.06616   28 KM
 0.06465   44 TIPO_CARRERA
 0.06175    5 CLUSTER_CREDITOS
 0.06103   23 ESTADO_CIVIL
 0.06044   25 GENERO
 0.05924   13 CRED_MATERIA
 0.05556   17 DEPCARRERA
 0.05272   22 EDAD
 0.04462   33 NACIONALIDAD
 0.04462   10 CLUSTER_NACIONALIDAD
 0.04413   46 TIPO_HORARIO
 0.04343    3 CARRERA
 0.0377    45 TIPO_DISCAPACIDAD
0.03696    36 PERIODO_TIPO
 0.03245   24 ETNIA
 0.02456   26 HORA_FIN
 0.02377   27 HORA_INICIO
 0.02342   20 DIAS
 0.02027   19 DIA_VIERNES
 0.0192    21 DISCAPACIDAD
 0.01849   29 MATERIA
 0.018      7 CLUSTER_EDAD
 0.01711    6 CLUSTER_CURSO
 0.00589   15 CURSO_ALUMNO

Selected attributes: 37,31,38,12,39,47,9,2,1,30,34,48,32,14,4,41,4

        40,16,35,42,28,44,5,23,25,13,17,22,33,10,46,3,45,

             36,24,26,27,20,19,21,29,7,6,15 : 47
```

Figura 95. Selección de Atributos mediante la herramienta WEKA para el Modelo2

El proceso para generar las distintas Etapas se muestra en la Figura 96, donde primero se cuenta con el conjunto original, es decir, con todos los atributos, luego se realiza la selección de atributos eliminando los menos influyentes según el resultado selección de atributos de las Figuras 94 y 95 para los modelos 1 y 2 respectivamente, llegando así a tener un nuevo subconjunto y posteriormente evaluándolo de acuerdo al porcentaje de instancias clasificadas correctamente, en el caso de que este valor sea admisible es decir, mayor al 80% y que no afecte significativamente el porcentaje de clasificación anterior (Etapa Anterior) se crea una etapa, caso contrario se vuelva a generar un nuevo subconjunto de atributos.

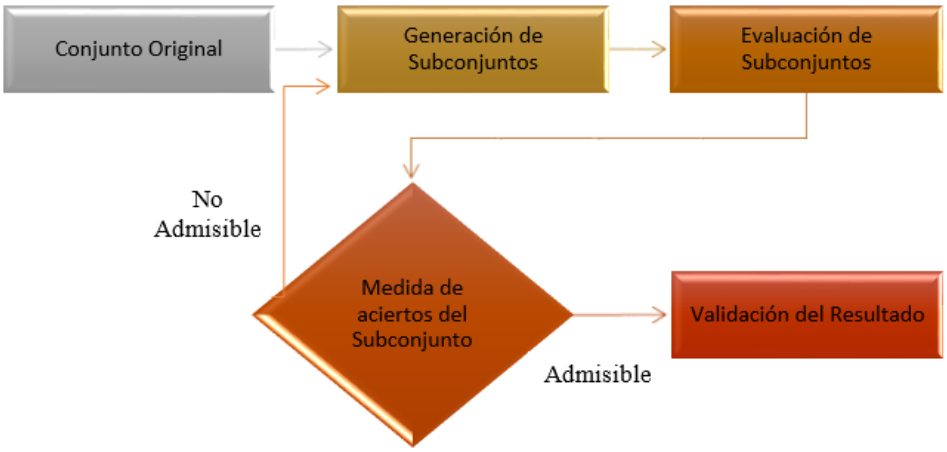

Figura 96. Proceso de Selección de Subconjuntos de atributos.

MODELO 1: Predice si un alumno aprobará o reprobará una materia.

Tabla 43.
Atributos de las etapas para el modelo 1 del Departamento de Ciencias de la Computación

.

ETAPAS	ATRIBUTOS		N°
ETAPA I	CLUSTER_CRED_MATERIA CLUSTER_PROV_COLEGIO CLUSTER_PROV_NAC COLEGIO CREDITOS CURSO_ALUMNO DEP_ASIGNATURA EDAD ESTADO_CIVIL ETNIA GENERO	MATERIAS_CURSADAS MATERIAS_REP_ANT MILITAR NIVEL PERIODO_TIPO PROMEDIO_ALUMNO SOSTENIMIENTO STATUS TIPO_ALUMNO TIPO_HORARIO TIPO_MATRICULA	22
ETAPA II	STATUS CLUSTER_CRED_MATERIA CLUSTER_PROV_NAC COLEGIO CREDITOS CURSO_ALUMNO DEP_ASIGNATURA EDAD	ESTADO_CIVIL GENERO MATERIAS_CURSADAS MATERIAS_REP_ANT NIVEL PROMEDIO_ALUMNO TIPO_ALUMNO TIPO_HORARIO TIPO_MATRICULA	17

Observación: Las variables que no se toman en cuenta en la segunda etapa son: cluster_prov_colegio, etnia, militar, periodo_tipo, sostenimiento.

ETAPA I

ÁRBOLES DE DECISIÓN

En el caso de los árboles de decisión utilizando el algoritmo J48 se obtuvo que, de 1030 instancias de prueba 840 las clasificó correctamente, es decir, el 81.5534% La matriz de confusión muestra que el algoritmo confundió 105 reprobados como aprobados y 85 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.384 y el error absoluto medio sea de 0.2629 como se detalla en la Figura 97.

```
=== Summary ===

Correctly Classified Instances         840               81.5534 %
Incorrectly Classified Instances       190               18.4466 %
Kappa statistic                          0.6312
Mean absolute error                      0.2629
Root mean squared error                  0.384
Relative absolute error                 52.5793 %
Root relative squared error             76.8    %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,832    0,201    0,801      0,832    0,816      0,632    0,857     0,827     APROBADO
                 0,799    0,168    0,831      0,799    0,815      0,632    0,857     0,828     REPROBADO
Weighted Avg.    0,816    0,184    0,816      0,816    0,816      0,632    0,857     0,827

=== Confusion Matrix ===

   a   b   <-- classified as
 422  85 |   a = APROBADO
 105 418 |   b = REPROBADO
```

Figura 97. Evaluación del Modelo 1 para el Departamento de Ciencias de la Computación con árboles de decisión J48 en la ETAPA I

REGRESIÓN LINEAL

El algoritmo Regresión Lineal clasificó de 1030 instancias de prueba 832 las clasificó correctamente, es decir, el 80.7767%. La matriz de confusión muestra que el algoritmo confundió 99 reprobados como aprobados y 99 aprobados como reprobados. El error cuadrático medio es de 0.359 y el error absoluto medio sea de 0.257 como se detalla en la Figura 98.

```
=== Summary ===

Correctly Classified Instances         832               80.7767 %
Incorrectly Classified Instances       198               19.2233 %
Kappa statistic                          0.6154
Mean absolute error                      0.257
Root mean squared error                  0.359
Relative absolute error                 51.387  %
Root relative squared error             71.7965 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,805    0,189    0,805      0,805    0,805      0,615    0,899     0,902     APROBADO
                 0,811    0,195    0,811      0,811    0,811      0,615    0,899     0,897     REPROBADO
Weighted Avg.    0,808    0,192    0,808      0,808    0,808      0,615    0,899     0,900

=== Confusion Matrix ===

   a   b   <-- classified as
 408  99 |   a = APROBADO
  99 424 |   b = REPROBADO
```

Figura 98. Evaluación del Modelo 1 para el Departamento de Ciencias de la Computación con regresión lineal utilizando el algoritmo Regresión Lineal en la ETAPA I

CLASIFICACIÓN LAZY

En este caso se utilizó el algoritmo KStar donde, de 1030 instancias de prueba 838 las clasificó correctamente, es decir, el 81.3592%. La matriz de confusión muestra que el algoritmo confundió 103 reprobados como aprobados y 89 aprobados como reprobados. El error cuadrático medio es de 0.3847 y el error absoluto medio sea de 0.2029 como se detalla en la Figura 99.

```
=== Summary ===

Correctly Classified Instances         838               81.3592 %
Incorrectly Classified Instances       192               18.6408 %
Kappa statistic                          0.6273
Mean absolute error                      0.2029
Root mean squared error                  0.3847
Relative absolute error                 40.5681 %
Root relative squared error             76.9439 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,824    0,197    0,802      0,824    0,813      0,627    0,889     0,895     APROBADO
               0,803    0,176    0,825      0,803    0,814      0,627    0,889     0,858     REPROBADO
Weighted Avg.  0,814    0,186    0,814      0,814    0,814      0,627    0,889     0,876

=== Confusion Matrix ===

   a   b   <-- classified as
 418  89 |   a = APROBADO
 103 420 |   b = REPROBADO
```

Figura 99. Evaluación del Modelo 1 con clasificación lazy para el Departamento de Ciencias de la Computación utilizando el algoritmo KStar en la ETAPA I

ETAPA II

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 17 atributos obtuvo que de 1030 instancias de prueba 837 las clasificó correctamente, es decir, el 81.2621%. La matriz de confusión muestra que el algoritmo confundió 113 reprobados como aprobados y 80 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.3842 y el error absoluto medio sea de 0.2663 como se detalla en la Figura 100.

```
=== Summary ===

Correctly Classified Instances         837               81.2621 %
Incorrectly Classified Instances       193               18.7379 %
Kappa statistic                          0.6255
Mean absolute error                      0.2663
Root mean squared error                  0.3842
Relative absolute error                 53.2622 %
Root relative squared error             76.8277 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,842    0,216    0,791      0,842    0,816      0,627    0,859     0,826     APROBADO
               0,784    0,158    0,837      0,784    0,809      0,627    0,859     0,835     REPROBADO
Weighted Avg.  0,813    0,186    0,814      0,813    0,813      0,627    0,859     0,830

=== Confusion Matrix ===

   a   b   <-- classified as
 427  80 |   a = APROBADO
 113 410 |   b = REPROBADO
```

Figura 100. Evaluación del Modelo 1 para el Departamento de Ciencias de la Computación con árboles de decisión J48 en la ETAPA II

REGRESIÓN LINEAL

El algoritmo Regresión Lineal esta vez si generó un modelo donde, de 1030 instancias de prueba 835 las clasificó correctamente, es decir, el 81.068%. La matriz de confusión muestra que el algoritmo confundió 102 reprobados como aprobados y 93 aprobados como reprobados. El error cuadrático medio es de 0.3596 y el error absoluto medio sea de 0.2613 como se detalla en la Figura 101.

```
=== Summary ===

Correctly Classified Instances         835               81.068  %
Incorrectly Classified Instances       195               18.932  %
Kappa statistic                          0.6214
Mean absolute error                      0.2613
Root mean squared error                  0.3596
Relative absolute error                 52.254  %
Root relative squared error             71.9148 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,817    0,195    0,802      0,817    0,809      0,621    0,900     0,901     APROBADO
                 0,805    0,183    0,819      0,805    0,812      0,621    0,900     0,901     REPROBADO
Weighted Avg.    0,811    0,189    0,811      0,811    0,811      0,621    0,900     0,901

=== Confusion Matrix ===

   a   b   <-- classified as
 414  93 |   a = APROBADO
 102 421 |   b = REPROBADO
```

Figura 101. Evaluación del Modelo 1 del Departamento de Ciencias de Computación con regresión lineal utilizando el algoritmo Regresión Lineal en la ETAPA II

CLASIFICACIÓN LAZY

El algoritmo KStar tiene que, de 1030 instancias de prueba 831 las clasificó correctamente, es decir, el 80.6796%. La matriz de confusión muestra que el algoritmo confundió 99 reprobados como aprobados y 100 aprobados como reprobados. El error cuadrático medio es de 0.3827 y el error absoluto medio sea de 0.2187 como se detalla en la Figura 102.

```
=== Summary ===

Correctly Classified Instances         831               80.6796 %
Incorrectly Classified Instances       199               19.3204 %
Kappa statistic                          0.6135
Mean absolute error                      0.2187
Root mean squared error                  0.3827
Relative absolute error                 43.7383 %
Root relative squared error             76.5276 %
Total Number of Instances             1030

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,803    0,189    0,804      0,803    0,804      0,613    0,887     0,893     APROBADO
                 0,811    0,197    0,809      0,811    0,810      0,613    0,887     0,856     REPROBADO
Weighted Avg.    0,807    0,193    0,807      0,807    0,807      0,613    0,887     0,874

=== Confusion Matrix ===

   a   b   <-- classified as
 407 100 |   a = APROBADO
  99 424 |   b = REPROBADO
```

Figura 102. Evaluación del Modelo 1 para el Departamento de Ciencias de la Computación con clasificación lazy utilizando el algoritmo KStar en la ETAPA II

CIENCIAS DE LA TIERRA Y DE LA CONSTRUCCION

Tabla 44.
Atributos de las etapas para el modelo 1 del Departamento de Ciencias de la Tierra y de la Construcción

ETAPAS	ATRIBUTOS		N°
ETAPA I	CLUSTER_CRED_MATERIA CLUSTER_NIVEL CLUSTER_PROV_NAC COLEGIO CREDITOS CURSO_ALUMNO DEP_ASIGNATURA DOCENTE EDAD ESTADO_CIVIL ETNIA GENERO	MATERIAS_CURSADAS MATERIAS_REP_ANT MILITAR NIVEL PERIODO_TIPO PROMEDIO_ALUMNO SOSTENIMIENTO STATUS TIPO_ALUMNO TIPO_HORARIO TIPO_MATRICULA TIPO_SOSTENIMIENTO	24
ETAPA II	STATUS CLUSTER_CRED_MATERIA CLUSTER_NIVEL CLUSTER_PROV_NAC COLEGIO CREDITOS CURSO_ALUMNO DEP_ASIGNATURA	GENERO MATERIAS_CURSADAS MATERIAS_REP_ANT MILITAR PERIODO_TIPO PROMEDIO_ALUMNO SOSTENIMIENTO TIPO_ALUMNO	23

	DOCENTE EDAD ESTADO_CIVIL ETNIA	TIPO_HORARIO TIPO_MATRICULA TIPO_SOSTENIMIENTO	

Observación: La variable que no se tomó en cuenta en la segunda etapa es nivel

ETAPA I

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 24 atributos obtuvo que de 1544 instancias de prueba 1227 las clasificó correctamente, es decir, el 79.4689%. La matriz de confusión muestra que el algoritmo confundió 163 reprobados como aprobados y 154 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.4013y el error absoluto medio sea de 0.2773 como se detalla en la Figura 103.

```
=== Summary ===

Correctly Classified Instances        1227               79.4689 %
Incorrectly Classified Instances       317               20.5311 %
Kappa statistic                          0.5894
Mean absolute error                      0.2773
Root mean squared error                  0.4013
Relative absolute error                 55.4509 %
Root relative squared error             80.2556 %
Total Number of Instances             1544

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,801    0,212    0,792      0,801    0,797      0,589    0,837     0,799     APROBADO
                 0,788    0,199    0,797      0,788    0,793      0,589    0,837     0,805     REPROBADO
Weighted Avg.    0,795    0,205    0,795      0,795    0,795      0,589    0,837     0,802

=== Confusion Matrix ===

   a   b   <-- classified as
 621 154 |   a = APROBADO
 163 606 |   b = REPROBADO
```

Figura 103. Evaluación del Modelo 1 del Departamento de la Tierra y de la Construcción con árboles de decisión J48 en la ETAPA I

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 1544 instancias de prueba 1318 las clasificó correctamente, es decir, el 85.3627%. La matriz de confusión muestra que el algoritmo confundió 121 reprobados como aprobados y 105 aprobados como reprobados. El error cuadrático medio es de 0.3221 y el error absoluto medio sea de 0.1959 como se detalla en la Figura 104.

```
=== Summary ===

Correctly Classified Instances        1318               85.3627 %
Incorrectly Classified Instances       226               14.6373 %
Kappa statistic                          0.7072
Mean absolute error                      0.1959
Root mean squared error                  0.3221
Relative absolute error                 39.1878 %
Root relative squared error             64.4114 %
Total Number of Instances             1544

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,865    0,157    0,847      0,865    0,856      0,707    0,931     0,917     APROBADO
               0,843    0,135    0,861      0,843    0,852      0,707    0,931     0,936     REPROBADO
Weighted Avg.  0,854    0,146    0,854      0,854    0,854      0,707    0,931     0,927

=== Confusion Matrix ===

   a   b   <-- classified as
 670 105 |   a = APROBADO
 121 648 |   b = REPROBADO
```

Figura 104. Evaluación del Modelo 1 para el Departamento de la Tierra y de la Construcción con regresión lineal utilizando el algoritmo Regresión Lineal en la ETAPA I

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 1544 instancias de prueba 1270 las clasificó correctamente, es decir, el 82.2539%. La matriz de confusión muestra que el algoritmo solo confundió 154 reprobados como aprobados y 120 aprobados como reprobados. El error cuadrático medio es de 0.382 y el error absoluto medio sea de 0.1822 como se detalla en la Figura 105.

```
=== Summary ===

Correctly Classified Instances        1270               82.2539 %
Incorrectly Classified Instances       274               17.7461 %
Kappa statistic                          0.645
Mean absolute error                      0.1822
Root mean squared error                  0.382
Relative absolute error                 36.4348 %
Root relative squared error             76.399  %
Total Number of Instances             1544

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,845    0,200    0,810      0,845    0,827      0,646    0,908     0,903     APROBADO
                 0,800    0,155    0,837      0,800    0,818      0,646    0,908     0,912     REPROBADO
Weighted Avg.    0,823    0,178    0,823      0,823    0,822      0,646    0,908     0,907

=== Confusion Matrix ===

   a   b   <-- classified as
 655 120 |   a = APROBADO
 154 615 |   b = REPROBADO
```

Figura 105. Evaluación del Modelo 1 para el Departamento de la Tierra y de la Construcción con clasificación lazy utilizando el algoritmo KStar en la ETAPA I

ETAPA II

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 23 atributos obtuvo que de 1544 instancias de prueba 1236 las clasificó correctamente, es decir, el 80.0518%. La matriz de confusión muestra que el algoritmo confundió 152 reprobados como aprobados y 156 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.3929 y el error absoluto medio sea de 0.2827 como se detalla en la Figura 106.

```
=== Summary ===

Correctly Classified Instances        1236               80.0518 %
Incorrectly Classified Instances       308               19.9482 %
Kappa statistic                          0.601
Mean absolute error                      0.2827
Root mean squared error                  0.3929
Relative absolute error                 56.5367 %
Root relative squared error             78.5769 %
Total Number of Instances             1544

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,799    0,198    0,803      0,799    0,801      0,601    0,848     0,830     APROBADO
               0,802    0,201    0,798      0,802    0,800      0,601    0,848     0,822     REPROBADO
Weighted Avg.  0,801    0,199    0,801      0,801    0,801      0,601    0,848     0,826

=== Confusion Matrix ===

   a   b   <-- classified as
 619 156 |   a = APROBADO
 152 617 |   b = REPROBADO
```

Figura 106. Evaluación del Modelo 1 para el Departamento de la Tierra y de la Construcción con árboles de decisión J48 en la ETAPA II

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 1544 instancias de prueba 1329 las clasificó correctamente, es decir, el 86.0751%. La matriz de confusión muestra que el algoritmo confundió 117 reprobados como aprobados y 98 aprobados como reprobados. El error cuadrático medio es de 0.318 y el error absoluto medio sea de 0.1947 como se detalla en la Figura 107.

```
=== Summary ===

Correctly Classified Instances        1329               86.0751 %
Incorrectly Classified Instances       215               13.9249 %
Kappa statistic                          0.7215
Mean absolute error                      0.1947
Root mean squared error                  0.318
Relative absolute error                 38.9385 %
Root relative squared error             63.6095 %
Total Number of Instances             1544

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,874    0,152    0,853      0,874    0,863      0,722    0,934     0,922     APROBADO
               0,848    0,126    0,869      0,848    0,858      0,722    0,934     0,940     REPROBADO
Weighted Avg.  0,861    0,139    0,861      0,861    0,861      0,722    0,934     0,931

=== Confusion Matrix ===

   a   b   <-- classified as
 677  98 |   a = APROBADO
 117 652 |   b = REPROBADO
```

Figura 107. Evaluación del Modelo 1 para el Departamento de la Tierra y de la Construcción con regresión lineal utilizando el algoritmo Regresión Lineal en la ETAPA II

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 1544 instancias de prueba 1278 las clasificó correctamente, es decir, el 82.772%. La matriz de confusión muestra que el algoritmo solo confundió 150 reprobados como aprobados y 116 aprobados como reprobados. El error cuadrático medio es de 0.3807 y el error absoluto medio sea de 0.1838 como se detalla en la Figura 108.

```
=== Summary ===

Correctly Classified Instances        1278               82.772  %
Incorrectly Classified Instances       266               17.228  %
Kappa statistic                          0.6554
Mean absolute error                      0.1838
Root mean squared error                  0.3807
Relative absolute error                 36.7519 %
Root relative squared error             76.1373 %
Total Number of Instances             1544

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,850    0,195    0,815      0,850    0,832      0,656    0,907     0,899     APROBADO
                 0,805    0,150    0,842      0,805    0,823      0,656    0,907     0,911     REPROBADO
Weighted Avg.    0,828    0,172    0,828      0,828    0,828      0,656    0,907     0,905

=== Confusion Matrix ===

   a   b   <-- classified as
 659 116 |   a = APROBADO
 150 619 |   b = REPROBADO
```

Figura 108. Evaluación del Modelo 1 para el Departamento de la Tierra y de la Construcción con clasificación lazy utilizando el algoritmo KStar en la ETAPA II

CIENCIAS ADMINISTRATIVAS ECONÓMICAS Y DE COMERCIO

Tabla 45.
Atributos de las etapas para el modelo 1 del Departamento de Ciencias Administrativas Económicas y de Comercio

ETAPAS	ATRIBUTOS		N°
ETAPA I	CLUSTER_CRED_MATERIA CLUSTER_CREDITOS CLUSTER_CURSO CLUSTER_MATERIAS_CURSADAS CLUSTER_NIVEL CLUSTER_PROV_COLEGIO CLUSTER_PROV_NAC CREDITOS CURSO_ALUMNO DEP_ASIGNATURA DOCENTE EDAD ESTADO_CIVIL ETNIA GENERO MATERIAS_CURSADAS MATERIAS_REP_ANT MILITAR	NIVEL PARROQUIA PERIODO_TIPO PROMEDIO_ALUMNO PROV_COLEGIO PROVINCIA_NAC REGIMENESCOLAR SOSTENIMIENTO STATUS TIEMPO TIPO_ALUMNO TIPO_DISCAPACIDAD TIPO_HORARIO TIPO_MATRICULA TIPO_SOSTENIMIENTO	33
ETAPA II	STATUS CLUSTER_CRED_MATERIA CREDITOS CURSO_ALUMNO DEP_ASIGNATURA DOCENTE EDAD ESTADO_CIVIL ETNIA GENERO	MATERIAS_CURSADAS MATERIAS_REP_ANT NIVEL PERIODO_TIPO PROMEDIO_ALUMNO SOSTENIMIENTO TIPO_ALUMNO TIPO_HORARIO TIPO_MATRICULA	19

Observación: Las variables que no fueron tomadas en cuenta en la segunda etapa son: cluster_creditos, cluster_curso, cluster_materias_cursadas, cluster_nivel, cluster_prov_nac, cluster_prov_colegio, militar, parroquia, prov_colegio, provincia_nac, tiempo, tipo_discapacidad, tipo_sostenimiento.

ETAPA I

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 33 atributos obtuvo que de 2920 instancias de prueba 2655 las clasificó correctamente, es decir, el 90.9247%. La matriz de confusión muestra que el algoritmo confundió 188 reprobados como aprobados y 77 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.2796 y el error absoluto medio sea de 0.1359 como se detalla en la Figura 109.

```
=== Summary ===

Correctly Classified Instances        2655               90.9247 %
Incorrectly Classified Instances       265                9.0753 %
Kappa statistic                          0.8186
Mean absolute error                      0.1359
Root mean squared error                  0.2796
Relative absolute error                 27.1769 %
Root relative squared error             55.9216 %
Total Number of Instances             2920

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,947    0,127    0,879      0,947    0,912      0,821    0,934     0,898     APROBADO
                 0,873    0,053    0,944      0,873    0,907      0,821    0,934     0,932     REPROBADO
Weighted Avg.    0,909    0,090    0,912      0,909    0,909      0,821    0,934     0,915

=== Confusion Matrix ===

    a    b   <-- classified as
 1368   77 |    a = APROBADO
  188 1287 |    b = REPROBADO
```

Figura 109. Evaluación del Modelo 1 para el Ciencias Administrativas Económicas y de Comercio con árboles de decisión J48 en la ETAPA I

REGRESIÓN LINEAL

Se utilizó el algoritmo Regresión Lineal que ofrece WEKA, sin embargo, no se puedo obtener el resultado de este, porque tiene una gran cantidad de atributos que impide generar un modelo.

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 2920 instancias de prueba 2682 las clasificó correctamente, es decir, el 91.8493%. La matriz de confusión

muestra que el algoritmo solo confundió 159 reprobados como aprobados y 79 aprobados como reprobados. El error cuadrático medio es de 0.2643 y el error absoluto medio sea de 0.0872 como se detalla en la Figura 110.

```
=== Summary ===

Correctly Classified Instances        2682               91.8493 %
Incorrectly Classified Instances       238                8.1507 %
Kappa statistic                          0.8371
Mean absolute error                      0.0872
Root mean squared error                  0.2643
Relative absolute error                 17.4315 %
Root relative squared error             52.8649 %
Total Number of Instances             2920

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,945    0,108    0,896      0,945    0,920      0,838    0,965     0,951     APROBADO
                 0,892    0,055    0,943      0,892    0,917      0,838    0,965     0,964     REPROBADO
Weighted Avg.    0,918    0,081    0,920      0,918    0,918      0,838    0,965     0,957

=== Confusion Matrix ===

    a    b   <-- classified as
 1366   79 |    a = APROBADO
  159 1316 |    b = REPROBADO
```

Figura 110. Evaluación del Modelo 1 para en Departamento de Ciencias Administrativas Económicas y de Comercio con clasificación lazy utilizando el algoritmo KStar en la ETAPA I

ETAPA II

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 19 atributos obtuvo que de 2920 instancias de prueba 2658 las clasificó correctamente, es decir, el 91.0274%. La matriz de confusión muestra que el algoritmo confundió 183 reprobados como aprobados y 79 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.279 y el error absoluto medio sea de 0.1352 como se detalla en la Figura 111.

```
=== Summary ===

Correctly Classified Instances        2658               91.0274 %
Incorrectly Classified Instances       262                8.9726 %
Kappa statistic                          0.8207
Mean absolute error                      0.1352
Root mean squared error                  0.279
Relative absolute error                 27.0479 %
Root relative squared error             55.7941 %
Total Number of Instances             2920

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,945    0,124    0,882      0,945    0,912      0,823    0,934     0,897     APROBADO
                 0,876    0,055    0,942      0,876    0,908      0,823    0,934     0,934     REPROBADO
Weighted Avg.    0,910    0,089    0,912      0,910    0,910      0,823    0,934     0,915

=== Confusion Matrix ===

    a    b   <-- classified as
 1366   79 |    a = APROBADO
  183 1292 |    b = REPROBADO
```

Figura 111. Evaluación del Modelo 1 para el Ciencias Administrativas Económicas y de Comercio con árboles de decisión J48 en la ETAPA II

REGRESIÓN LINEAL

Se utilizó el algoritmo Regresión Lineal que ofrece WEKA, sin embargo, no se puedo obtener el resultado de este, porque tiene una gran cantidad de atributos que impide generar un modelo.

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 2920 instancias de prueba 2717 las clasificó correctamente, es decir, el 93.0479%. La matriz de confusión muestra que el algoritmo solo confundió 141 reprobados como aprobados y 62 aprobados como reprobados. El error cuadrático medio es de 0.2408 y el error absoluto medio sea de 0.0899 como se detalla en la Figura 112.

```
=== Summary ===

Correctly Classified Instances        2717               93.0479 %
Incorrectly Classified Instances       203                6.9521 %
Kappa statistic                          0.861
Mean absolute error                      0.0899
Root mean squared error                  0.2408
Relative absolute error                 17.9788 %
Root relative squared error             48.1547 %
Total Number of Instances             2920

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,957    0,096    0,907      0,957    0,932      0,862    0,975     0,970     APROBADO
                 0,904    0,043    0,956      0,904    0,929      0,862    0,975     0,977     REPROBADO
Weighted Avg.    0,930    0,069    0,932      0,930    0,930      0,862    0,975     0,973

=== Confusion Matrix ===

    a    b   <-- classified as
 1383   62 |    a = APROBADO
  141 1334 |    b = REPROBADO
```

Figura 112. Evaluación del Modelo 1 para en Departamento de Ciencias Administrativas Económicas y de Comercio con clasificación lazy utilizando el algoritmo KStar en la ETAPA II

CIENCIAS DE LA VIDA

Tabla 46.
Atributos de las etapas para el modelo 1 del Departamento de Ciencias de la Vida

ETAPAS	ATRIBUTOS		N°
ETAPA I	STATUS CLUSTER_CURSO COLEGIO CRED_MATERIA CREDITOS CURSO_ALUMNO DEP_ASIGNATURA EDAD ESTADO_CIVIL GENERO	MATERIAS_CURSADAS MATERIAS_REP_ANT PARROQUIA PERIODO_TIPO PROMEDIO_ALUMNO SOSTENIMIENTO TIPO_ALUMNO TIPO_HORARIO TIPO_MATRICULA ETNIA	20
ETAPA II	STATUS COLEGIO CRED_MATERIA CREDITOS DEP_ASIGNATURA EDAD GENERO MATERIAS_CURSADAS	MATERIAS_REP_ANT PARROQUIA PERIODO_TIPO PROMEDIO_ALUMNO TIPO_ALUMNO TIPO_MATRICULA	14

Observación: Las variables que no fueron tomadas en cuenta en la segunda etapa son: cluster_curso, curso_alumno, estado_civil,etnia, sostenimiento y tipo_horario.

ETAPA I

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 20 atributos obtuvo que de 1092 instancias de prueba 961 las clasificó correctamente, es decir, el 88.0037%. La matriz de confusión muestra que el algoritmo confundió 74 reprobados como aprobados y 57 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.1492 y el error absoluto medio sea de 0.3087 como se detalla en la Figura 113.

```
=== Summary ===

Correctly Classified Instances         961               88.0037 %
Incorrectly Classified Instances       131               11.9963 %
Kappa statistic                          0.76
Mean absolute error                      0.1492
Root mean squared error                  0.3087
Relative absolute error                 29.837  %
Root relative squared error             61.745  %
Total Number of Instances             1092

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,897    0,137    0,870      0,897    0,883      0,760    0,935     0,911     APROBADO
                 0,863    0,103    0,891      0,863    0,877      0,760    0,935     0,934     REPROBADO
Weighted Avg.    0,880    0,120    0,880      0,880    0,880      0,760    0,935     0,922

=== Confusion Matrix ===

   a   b   <-- classified as
 495  57 |   a = APROBADO
  74 466 |   b = REPROBADO
```

Figura 113. Evaluación del Modelo 1 para el Ciencias de la Vida con árboles de decisión J48 en la ETAPA I

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 1092 instancias de prueba 947 las clasificó correctamente, es decir, el 86.7216%. La matriz de confusión muestra que el algoritmo confundió 83 reprobados como aprobados y 62 aprobados como reprobados. El error cuadrático medio es de 0.3056 y el error absoluto medio sea de 0.1729como se detalla en la Figura 114.

```
=== Summary ===

Correctly Classified Instances         947               86.7216 %
Incorrectly Classified Instances       145               13.2784 %
Kappa statistic                          0.7343
Mean absolute error                      0.1729
Root mean squared error                  0.3056
Relative absolute error                 34.5767 %
Root relative squared error             61.1144 %
Total Number of Instances             1092

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,888    0,154    0,855      0,888    0,871      0,735    0,941     0,928     APROBADO
                 0,846    0,112    0,881      0,846    0,863      0,735    0,941     0,948     REPROBADO
Weighted Avg.    0,867    0,133    0,868      0,867    0,867      0,735    0,941     0,938

=== Confusion Matrix ===

   a   b   <-- classified as
 490  62 |   a = APROBADO
  83 457 |   b = REPROBADO
```

Figura 114. Evaluación del Modelo 1 para el Ciencias de la Vida con Regresión Lineal en la ETAPA I

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 1092 instancias de prueba 923 las clasificó correctamente, es decir, el 84.5238%. La matriz de confusión muestra que el algoritmo solo confundió 94 reprobados como aprobados y 75 aprobados como reprobados. El error cuadrático medio es de 0.3559 y el error absoluto medio sea de 0.1648 como se detalla en la Figura 115.

```
=== Summary ===

Correctly Classified Instances         923               84.5238 %
Incorrectly Classified Instances       169               15.4762 %
Kappa statistic                          0.6903
Mean absolute error                      0.1648
Root mean squared error                  0.3559
Relative absolute error                 32.9657 %
Root relative squared error             71.1815 %
Total Number of Instances             1092

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,864    0,174    0,835      0,864    0,850      0,691    0,925     0,930     APROBADO
                 0,826    0,136    0,856      0,826    0,841      0,691    0,925     0,920     REPROBADO
Weighted Avg.    0,845    0,155    0,846      0,845    0,845      0,691    0,925     0,925

=== Confusion Matrix ===

   a   b   <-- classified as
 477  75 |   a = APROBADO
  94 446 |   b = REPROBADO
```

Figura 115. Evaluación del Modelo 1 para el Departamento de Ciencias de la Vida con clasificación lazy utilizando el algoritmo KStar en la ETAPA I

ETAPA II

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 14 atributos obtuvo que de 1092 instancias de prueba 972 las clasificó correctamente, es decir, el 89.011%. La matriz de confusión muestra que el algoritmo confundió 65 reprobados como aprobados y 55 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.3001 y el error absoluto medio sea de 0.1409 como se detalla en la Figura 116.

```
=== Summary ===

Correctly Classified Instances         972               89.011  %
Incorrectly Classified Instances       120               10.989  %
Kappa statistic                          0.7801
Mean absolute error                      0.1409
Root mean squared error                  0.3001
Relative absolute error                 28.1878 %
Root relative squared error             60.0168 %
Total Number of Instances             1092

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,900    0,120    0,884      0,900    0,892      0,780    0,940     0,917     APROBADO
               0,880    0,100    0,896      0,880    0,888      0,780    0,940     0,940     REPROBADO
Weighted Avg.  0,890    0,110    0,890      0,890    0,890      0,780    0,940     0,928

=== Confusion Matrix ===

   a   b   <-- classified as
 497  55 |   a = APROBADO
  65 475 |   b = REPROBADO
```

Figura 116. Evaluación del Modelo 1 para el Departamento de Ciencias de la Vida con árboles de decisión J48 en la ETAPA II

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 1092 instancias de prueba 966 las clasificó correctamente, es decir, el 88.4615 %. La matriz de confusión muestra que el algoritmo confundió 72 reprobados como aprobados y 54 aprobados como reprobados. El error cuadrático medio es de 0.2996 y el error absoluto medio sea de 0.164 como se detalla en la Figura 117.

```
=== Summary ===

Correctly Classified Instances         966               88.4615 %
Incorrectly Classified Instances       126               11.5385 %
Kappa statistic                          0.7691
Mean absolute error                      0.164
Root mean squared error                  0.2996
Relative absolute error                 32.804  %
Root relative squared error             59.9112 %
Total Number of Instances             1092

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,902    0,133    0,874      0,902    0,888      0,770    0,944     0,935     APROBADO
               0,867    0,098    0,897      0,867    0,881      0,770    0,944     0,949     REPROBADO
Weighted Avg.  0,885    0,116    0,885      0,885    0,885      0,770    0,944     0,942

=== Confusion Matrix ===

   a   b   <-- classified as
 498  54 |   a = APROBADO
  72 468 |   b = REPROBADO
```

Figura 117. Evaluación del Modelo 1 para el Departamento de Ciencias de la Vida con Regresión Lineal en la ETAPA II

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 1092 instancias de prueba 956 las clasificó correctamente, es decir, el 87.5488%. La matriz de confusión muestra que el algoritmo solo confundió 83 reprobados como aprobados y 53 aprobados como reprobados. El error cuadrático medio es de 0.3248 y el error absoluto medio sea de 0.1496 como se detalla en la Figura 118.

```
=== Summary ===

Correctly Classified Instances         956               87.5458 %
Incorrectly Classified Instances       136               12.4542 %
Kappa statistic                          0.7507
Mean absolute error                      0.1496
Root mean squared error                  0.3248
Relative absolute error                 29.9182 %
Root relative squared error             64.9622 %
Total Number of Instances             1092

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,904    0,154    0,857      0,904    0,880      0,752    0,937     0,937     APROBADO
               0,846    0,096    0,896      0,846    0,870      0,752    0,937     0,939     REPROBADO
Weighted Avg.  0,875    0,125    0,877      0,875    0,875      0,752    0,937     0,938

=== Confusion Matrix ===

   a   b   <-- classified as
 499  53 |   a = APROBADO
  83 457 |   b = REPROBADO
```

Figura 118. Evaluación del Modelo 1 para en Departamento de Ciencias de la Vida con clasificación lazy utilizando el algoritmo KStar en la ETAPA II

CIENCIAS HUMANAS Y SOCIALES

Tabla 47.
Atributos de las etapas para el modelo 1 del Departamento de Ciencias Humanas y Sociales

ETAPAS	ATRIBUTOS		N°
ETAPA I	CLUSTER_CRED_MATERIA CLUSTER_PROV_NAC CRED_MATERIA CREDITOS CURSO_ALUMNO DEP_ASIGNATURA EDAD ESTADO_CIVIL ETNIA GENERO	MATERIAS_CURSADAS MATERIAS_REP_ANT MILITAR NIVEL PERIODO_TIPO PROMEDIO_ALUMNO SOSTENIMIENTO STATUS TIPO_ALUMNO TIPO_HORARIO	20
ETAPA II	STATUS CLUSTER_CRED_MATERIA CLUSTER_PROV_NAC CREDITOS CURSO_ALUMNO DEP_ASIGNATURA EDAD GENERO MATERIAS_CURSADAS	MATERIAS_REP_ANT MILITAR NIVEL PERIODO_TIPO PROMEDIO_ALUMNO SOSTENIMIENTO TIPO_ALUMNO TIPO_HORARIO	17

Observación: Las variables que no fueron tomadas en cuenta en la segunda etapa son: cred_materia, estado_civil y etnia.

ETAPA I

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 20 atributos obtuvo que de 560 instancias de prueba 519 las clasificó correctamente, es decir, el 92.6786%. La matriz de confusión muestra que el algoritmo confundió 18 reprobados como aprobados y 23 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.2465 y el error absoluto medio sea de 0.1026 como se detalla en la Figura 119.

```
=== Summary ===

Correctly Classified Instances         519               92.6786 %
Incorrectly Classified Instances        41                7.3214 %
Kappa statistic                          0.8536
Mean absolute error                      0.1026
Root mean squared error                  0.2465
Relative absolute error                 20.5166 %
Root relative squared error             49.3072 %
Total Number of Instances              560

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,919    0,065    0,935      0,919    0,927      0,854    0,957     0,944     APROBADO
               0,935    0,081    0,918      0,935    0,927      0,854    0,957     0,946     REPROBADO
Weighted Avg.  0,927    0,073    0,927      0,927    0,927      0,854    0,957     0,945

=== Confusion Matrix ===

   a   b   <-- classified as
 260  23 |   a = APROBADO
  18 259 |   b = REPROBADO
```

Figura 119. Evaluación del Modelo 1 para el Departamento Ciencias Humanas y Sociales con árboles de decisión J48 en la ETAPA I

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 560 instancias de prueba 507 las clasificó correctamente, es decir, el 90.5357%. La matriz de confusión muestra que el algoritmo confundió 20 reprobados como aprobados y 33 aprobados como reprobados. El error cuadrático medio es de 0.2707 y el error absoluto medio sea de 0.1569 como se detalla en la Figura 120.

```
=== Summary ===

Correctly Classified Instances         507               90.5357 %
Incorrectly Classified Instances        53                9.4643 %
Kappa statistic                          0.8108
Mean absolute error                      0.1569
Root mean squared error                  0.2707
Relative absolute error                 31.3861 %
Root relative squared error             54.1434 %
Total Number of Instances              560

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,883    0,072    0,926      0,883    0,904      0,812    0,966     0,969     APROBADO
               0,928    0,117    0,886      0,928    0,907      0,812    0,966     0,959     REPROBADO
Weighted Avg.  0,905    0,094    0,906      0,905    0,905      0,812    0,966     0,964

=== Confusion Matrix ===

   a   b   <-- classified as
 250  33 |   a = APROBADO
  20 257 |   b = REPROBADO
```

Figura 120. Evaluación del Modelo 1 para el Departamento Ciencias Humanas y Sociales con Regresión Lineal en la ETAPA I

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 560 instancias de prueba 523 las clasificó correctamente, es decir, el 93.3929%. La matriz de confusión muestra que el algoritmo solo confundió 22 reprobados como aprobados y 15 aprobados como reprobados. El error cuadrático medio es de 0.2311 y el error absoluto medio sea de 0.0824 como se detalla en la Figura 121.

```
=== Summary ===

Correctly Classified Instances         523               93.3929 %
Incorrectly Classified Instances        37                6.6071 %
Kappa statistic                          0.8678
Mean absolute error                      0.0824
Root mean squared error                  0.2311
Relative absolute error                 16.4863 %
Root relative squared error             46.2118 %
Total Number of Instances              560

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,947    0,079    0,924      0,947    0,935      0,868    0,981     0,981     APROBADO
                 0,921    0,053    0,944      0,921    0,932      0,868    0,981     0,981     REPROBADO
Weighted Avg.    0,934    0,066    0,934      0,934    0,934      0,868    0,981     0,981

=== Confusion Matrix ===

   a   b   <-- classified as
 268  15 |   a = APROBADO
  22 255 |   b = REPROBADO
```

Figura 121. Evaluación del Modelo 1 para el Departamento Ciencias Humanas y Sociales con clasificación lazy utilizando el algoritmo KStar en la ETAPA I

ETAPA II

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 17 atributos obtuvo que de 560 instancias de prueba 523 las clasificó correctamente, es decir, el 93.3929%. La matriz de confusión muestra que el algoritmo confundió 18 reprobados como aprobados y 19 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.243 y el error absoluto medio sea de 0.0963 como se detalla en la Figura 122.

```
=== Summary ===

Correctly Classified Instances         523               93.3929 %
Incorrectly Classified Instances        37                6.6071 %
Kappa statistic                          0.8678
Mean absolute error                      0.0963
Root mean squared error                  0.243
Relative absolute error                 19.2584 %
Root relative squared error             48.606  %
Total Number of Instances              560

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,933    0,065    0,936      0,933    0,935      0,868    0,959     0,947     APROBADO
                 0,935    0,067    0,932      0,935    0,933      0,868    0,959     0,947     REPROBADO
Weighted Avg.    0,934    0,066    0,934      0,934    0,934      0,868    0,959     0,947

=== Confusion Matrix ===

   a   b   <-- classified as
 264  19 |   a = APROBADO
  18 259 |   b = REPROBADO
```

Figura 122. Evaluación del Modelo 1 para el Departamento Ciencias Humanas y Sociales con árboles de decisión J48 en la ETAPA II

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 560 instancias de prueba 505 las clasificó correctamente, es decir, el 90.1786 %. La matriz de confusión muestra que el algoritmo confundió 21 reprobados como aprobados y 34 aprobados como reprobados. El error cuadrático medio es de 0.2736y el error absoluto medio sea de 0.1596 como se detalla en la Figura 123.

```
=== Summary ===

Correctly Classified Instances         505               90.1786 %
Incorrectly Classified Instances        55                9.8214 %
Kappa statistic                          0.8036
Mean absolute error                      0.1596
Root mean squared error                  0.2736
Relative absolute error                 31.919  %
Root relative squared error             54.7095 %
Total Number of Instances              560

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,880    0,076    0,922      0,880    0,901      0,805    0,964     0,967     APROBADO
                 0,924    0,120    0,883      0,924    0,903      0,805    0,964     0,956     REPROBADO
Weighted Avg.    0,902    0,098    0,903      0,902    0,902      0,805    0,964     0,962

=== Confusion Matrix ===

   a   b   <-- classified as
 249  34 |   a = APROBADO
  21 256 |   b = REPROBADO
```

Figura 123. Evaluación del Modelo 1 para el Departamento Ciencias Humanas y Sociales con Regresión Lineal en la ETAPA II

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 560 instancias de prueba 527 las clasificó correctamente, es decir, el 94.1071%. La matriz de confusión muestra que el algoritmo solo confundió 21 reprobados como aprobados y 12 aprobados como reprobados. El error cuadrático medio es de 0.2288 y el error absoluto medio sea de 0.0819 como se detalla en la Figura 124.

```
=== Summary ===

Correctly Classified Instances         527               94.1071 %
Incorrectly Classified Instances        33                5.8929 %
Kappa statistic                          0.8821
Mean absolute error                      0.0819
Root mean squared error                  0.2288
Relative absolute error                 16.3763 %
Root relative squared error             45.7594 %
Total Number of Instances              560

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,958    0,076    0,928      0,958    0,943      0,883    0,982     0,982     APROBADO
                 0,924    0,042    0,955      0,924    0,939      0,883    0,982     0,982     REPROBADO
Weighted Avg.    0,941    0,059    0,942      0,941    0,941      0,883    0,982     0,982

=== Confusion Matrix ===

   a   b   <-- classified as
 271  12 |   a = APROBADO
  21 256 |   b = REPROBADO
```

Figura 124. Evaluación del Modelo 1 para el Departamento Ciencias Humanas y Sociales con el algoritmo KStar en la ETAPA II

DEPARTAMENTO DE ELÉCTRICA Y ELECTRÓNICA

Tabla 48.
Atributos de las etapas para el modelo 1 del Departamento de Eléctrica y Electrónica

ETAPAS	ATRIBUTOS		N°
ETAPA I	STATUS CLUSTER_CURSO CLUSTER_NIVEL CLUSTER_PROV_COLEGIO CLUSTER_PROV_NAC COLEGIO CRED_MATERIA CREDITOS CURSO_ALUMNO DEP_ASIGNATURA EDAD ESTADO_CIVIL	ETNIA GENERO MATERIAS_CURSADAS MATERIAS_REP_ANT MILITAR NIVEL PARROQUIA PROMEDIO_ALUMNO PROV_COLEGIO PROVINCIA_NAC SOSTENIMIENTO TIPO_ALUMNO TIPO_HORARIO TIPO_MATRICULA	26
ETAPA II	STATUS CLUSTER_PROV_NAC COLEGIO CRED_MATERIA CREDITOS DEP_ASIGNATURA EDAD ETNIA GENERO	MATERIAS_CURSADAS MATERIAS_REP_ANT MILITAR NIVEL PARROQUIA PROMEDIO_ALUMNO TIPO_ALUMNO TIPO_HORARIO TIPO_MATRICULA	18

Observación: Las variables que no fueron tomadas en cuenta en la segunda etapa son: cluster_curso, cluster_nivel, cluster_prov_colegio, curso_alumno, estado_civil, prov._colegio, provincia_nac y sotenimiento.

ETAPA I

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 26 atributos obtuvo que de 2260 instancias de prueba 1756 las clasificó correctamente, es decir, el 77.6991%. La matriz de confusión muestra que el algoritmo confundió 229 reprobados como aprobados y 275 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.4141 y el error absoluto medio sea de 0.3026 como se detalla en la Figura 125.

```
=== Summary ===

Correctly Classified Instances        1756               77.6991 %
Incorrectly Classified Instances       504               22.3009 %
Kappa statistic                          0.5539
Mean absolute error                      0.3036
Root mean squared error                  0.4141
Relative absolute error                 60.7111 %
Root relative squared error             82.8273 %
Total Number of Instances             2260

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,756    0,202    0,788      0,756    0,772      0,554    0,813     0,768     APROBADO
                 0,798    0,244    0,767      0,798    0,782      0,554    0,813     0,785     REPROBADO
Weighted Avg.    0,777    0,223    0,777      0,777    0,777      0,554    0,813     0,776

=== Confusion Matrix ===

   a   b   <-- classified as
 852 275 |   a = APROBADO
 229 904 |   b = REPROBADO
```

Figura 125. Evaluación del Modelo 1 para el Departamento de Eléctrica y Electrónica con árboles de decisión J48 en la ETAPA I

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 2260 instancias de prueba 1798 las clasificó correctamente, es decir, el 79.5575%. La matriz de confusión muestra que el algoritmo confundió 255 reprobados como aprobados y 207 aprobados como reprobados. El error cuadrático medio es de 0.3805 y el error absoluto medio sea de 0.279 como se detalla en la Figura 126.

```
=== Summary ===

Correctly Classified Instances        1798               79.5575 %
Incorrectly Classified Instances       462               20.4425 %
Kappa statistic                          0.5912
Mean absolute error                      0.279
Root mean squared error                  0.3805
Relative absolute error                 55.7908 %
Root relative squared error             76.0937 %
Total Number of Instances             2260

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,816    0,225    0,783      0,816    0,799      0,592    0,871     0,842     APROBADO
                 0,775    0,184    0,809      0,775    0,792      0,592    0,871     0,880     REPROBADO
Weighted Avg.    0,796    0,204    0,796      0,796    0,795      0,592    0,871     0,861

=== Confusion Matrix ===

   a   b   <-- classified as
 920 207 |   a = APROBADO
 255 878 |   b = REPROBADO
```

Figura 126. Evaluación del Modelo 1 para el Departamento de Eléctrica y Electrónica con Regresión Lineal en la ETAPA I

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 2260 instancias de prueba 1794 las clasificó correctamente, es decir, el 79.3805%. La matriz de confusión muestra que el algoritmo solo confundió 254 reprobados como aprobados y 212 aprobados como reprobados. El error cuadrático medio es de 0.4224 y el error absoluto medio sea de 0.2185 como se detalla en la Figura 127.

```
=== Summary ===

Correctly Classified Instances        1794               79.3805 %
Incorrectly Classified Instances       466               20.6195 %
Kappa statistic                          0.5876
Mean absolute error                      0.2185
Root mean squared error                  0.4224
Relative absolute error                 43.6938 %
Root relative squared error             84.4886 %
Total Number of Instances             2260

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,812    0,224    0,783      0,812    0,797      0,588    0,856     0,835     APROBADO
                 0,776    0,188    0,806      0,776    0,790      0,588    0,856     0,846     REPROBADO
Weighted Avg.    0,794    0,206    0,794      0,794    0,794      0,588    0,856     0,841

=== Confusion Matrix ===

   a   b   <-- classified as
 915 212 |   a = APROBADO
 254 879 |   b = REPROBADO
```

Figura 127. Evaluación del Modelo 1 para el Departamento de Eléctrica y Electrónica con clasificación lazy utilizando el algoritmo KStar en la ETAPA I

ETAPA II

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 18 atributos obtuvo que de 2260 instancias de prueba 1758 las clasificó correctamente, es decir, el 77.7876%. La matriz de confusión muestra que el algoritmo confundió 229 reprobados como aprobados y 273 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.416 y el error absoluto medio sea de 0.3055 como se detalla en la Figura 128.

```
=== Summary ===

Correctly Classified Instances        1758               77.7876 %
Incorrectly Classified Instances       502               22.2124 %
Kappa statistic                          0.5557
Mean absolute error                      0.3055
Root mean squared error                  0.416
Relative absolute error                 61.0949 %
Root relative squared error             83.1933 %
Total Number of Instances             2260

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,758    0,202    0,789      0,758    0,773      0,556    0,810     0,762     APROBADO
                 0,798    0,242    0,768      0,798    0,783      0,556    0,810     0,780     REPROBADO
Weighted Avg.    0,778    0,222    0,778      0,778    0,778      0,556    0,810     0,771

=== Confusion Matrix ===

   a   b   <-- classified as
 854 273 |   a = APROBADO
 229 904 |   b = REPROBADO
```

Figura 128. Evaluación del Modelo 1 para el Departamento de Eléctrica y Electrónica con árboles de decisión J48 en la ETAPA II

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 2260 instancias de prueba 1802 las clasificó correctamente, es decir, el 79.7345%. La matriz de confusión muestra que el algoritmo confundió 241 reprobados como aprobados y 271 aprobados como reprobados. El error cuadrático medio es de 0.3835 y el error absoluto medio sea de 0.2818 como se detalla en la Figura 129.

```
=== Summary ===

Correctly Classified Instances        1802               79.7345 %
Incorrectly Classified Instances       458               20.2655 %
Kappa statistic                          0.5947
Mean absolute error                      0.2818
Root mean squared error                  0.3835
Relative absolute error                 56.3537 %
Root relative squared error             76.6901 %
Total Number of Instances             2260

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,807    0,213    0,791      0,807    0,799      0,595    0,868     0,842     APROBADO
                 0,787    0,193    0,804      0,787    0,796      0,595    0,868     0,874     REPROBADO
Weighted Avg.    0,797    0,203    0,797      0,797    0,797      0,595    0,868     0,858

=== Confusion Matrix ===

   a   b   <-- classified as
 910 217 |   a = APROBADO
 241 892 |   b = REPROBADO
```

Figura 129. Evaluación del Modelo 1 para el Departamento de Eléctrica y Electrónica con Regreasión Lineal utilizando el algoritmo KStar en la ETAPA II

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 2260 instancias de prueba 1812 las clasificó correctamente, es decir, el 80.177%. La matriz de confusión muestra que el algoritmo solo confundió 248 reprobados como aprobados y 200 aprobados como reprobados. El error cuadrático medio es de 0.4016 y el error absoluto medio sea de 0.2218 como se detalla en la Figura 130.

```
=== Summary ===

Correctly Classified Instances        1812               80.177  %
Incorrectly Classified Instances       448               19.823  %
Kappa statistic                          0.6036
Mean absolute error                      0.2218
Root mean squared error                  0.4016
Relative absolute error                 44.3543 %
Root relative squared error             80.3243 %
Total Number of Instances             2260

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,823    0,219    0,789      0,823    0,805      0,604    0,869     0,863     APROBADO
                 0,781    0,177    0,816      0,781    0,798      0,604    0,869     0,857     REPROBADO
Weighted Avg.    0,802    0,198    0,802      0,802    0,802      0,604    0,869     0,860

=== Confusion Matrix ===

   a   b   <-- classified as
 927 200 |   a = APROBADO
 248 885 |   b = REPROBADO
```

Figura 130. Evaluación del Modelo 1 para el Departamento de Eléctrica y Electrónica con clasificación lazy utilizando el algoritmo KStar en la ETAPA II

DEPARTAMENTO DE ENERGÍA Y MECÁNICA

Tabla 49.
Atributos de las etapas para el modelo 1 del Departamento de Energía y Mecánica

ETAPAS	ATRIBUTOS		N°
ETAPA I	CLUSTER_CRED_MATERIA CLUSTER_CREDITOS CLUSTER_CURSO CLUSTER_EDAD CLUSTER_NIVEL CLUSTER_PROV_COLEGIO COLEGIO CREDITOS CURSO_ALUMNO DEP_ASIGNATURA DOCENTE EDAD ESTADO_CIVIL ETNIA	GENERO MATERIAS_CURSADAS MATERIAS_REP_ANT MILITAR NIVEL PARROQUIA PERIODO_TIPO PROMEDIO_ALUMNO PROV_COLEGIO SOSTENIMIENTO STATUS TIPO_ALUMNO TIPO_HORARIO TIPO_MATRICULA	28
ETAPA II	STATUS CLUSTER_CRED_MATERIA CLUSTER_CREDITOS CLUSTER_PROV_COLEGIO COLEGIO CURSO_ALUMNO DEP_ASIGNATURA EDAD ESTADO_CIVIL ETNIA GENERO	MATERIAS_CURSADAS MATERIAS_REP_ANT MILITAR NIVEL PERIODO_TIPO PROMEDIO_ALUMNO SOSTENIMIENTO TIPO_ALUMNO TIPO_HORARIO TIPO_MATRICULA	21

Observación: Las variables que no fueron tomadas en cuenta en la segunda etapa son: clsuter_curso, cluster_Edad, cluster_nivel, créditos, docente, parroquia y prov_colegio.

ETAPA I

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 28 atributos obtuvo que de 2186 instancias de prueba 1711 las clasificó correctamente, es decir, el 78.2108%. La matriz de confusión muestra que el algoritmo confundió 267 reprobados como aprobados y 208 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.4089 y el error absoluto medio sea de 0.296 como se detalla en la Figura 131.

```
=== Summary ===

Correctly Classified Instances        1711               78.2708 %
Incorrectly Classified Instances       475               21.7292 %
Kappa statistic                          0.5646
Mean absolute error                      0.296
Root mean squared error                  0.4089
Relative absolute error                 59.1825 %
Root relative squared error             81.7758 %
Total Number of Instances             2186

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,814    0,250    0,773      0,814    0,793      0,565    0,827     0,789     APROBADO
               0,750    0,186    0,794      0,750    0,771      0,565    0,827     0,798     REPROBADO
Weighted Avg.  0,783    0,219    0,783      0,783    0,782      0,565    0,827     0,794

=== Confusion Matrix ===

   a   b   <-- classified as
 911 208 |   a = APROBADO
 267 800 |   b = REPROBADO
```

Figura 131. Evaluación del Modelo 1 para el Departamento de Energía y Mecánica con árboles de decisión J48 en la ETAPA I

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 2186 instancias de prueba 1764 las clasificó correctamente, es decir, el 80.6953%. La matriz de confusión muestra que el algoritmo confundió 231 reprobados como aprobados y 191 aprobados como reprobados. El error cuadrático medio es de 0.3699 y el error absoluto medio sea de 0.268179 como se detalla en la Figura 132.

```
=== Summary ===

Correctly Classified Instances        1764               80.6953 %
Incorrectly Classified Instances       422               19.3047 %
Kappa statistic                          0.6134
Mean absolute error                      0.2681
Root mean squared error                  0.3699
Relative absolute error                 53.6164 %
Root relative squared error             73.9701 %
Total Number of Instances             2186

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,829    0,216    0,801      0,829    0,815      0,614    0,885     0,888     APROBADO
               0,784    0,171    0,814      0,784    0,798      0,614    0,885     0,869     REPROBADO
Weighted Avg.  0,807    0,194    0,807      0,807    0,807      0,614    0,885     0,879

=== Confusion Matrix ===

   a   b   <-- classified as
 928 191 |   a = APROBADO
 231 836 |   b = REPROBADO
```

Figura 132. Evaluación del Modelo 1 para el Departamento de Energía y Mecánica con Regresión Lineal en la ETAPA I

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 2186 instancias de prueba 1724 las clasificó correctamente, es decir, el 78.8655%. La matriz de confusión muestra que el algoritmo solo confundió 238 reprobados como aprobados y 224 aprobados como reprobados. El error cuadrático medio es de 0.43 y el error absoluto medio sea de 0.221 como se detalla en la Figura 133.

```
=== Summary ===

Correctly Classified Instances        1724               78.8655 %
Incorrectly Classified Instances       462               21.1345 %
Kappa statistic                          0.5769
Mean absolute error                      0.221
Root mean squared error                  0.43
Relative absolute error                 44.1947 %
Root relative squared error             85.9943 %
Total Number of Instances             2186

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,800    0,223    0,790      0,800    0,795      0,577    0,864     0,862     APROBADO
               0,777    0,200    0,787      0,777    0,782      0,577    0,864     0,853     REPROBADO
Weighted Avg.  0,789    0,212    0,789      0,789    0,789      0,577    0,864     0,857

=== Confusion Matrix ===

   a   b   <-- classified as
 895 224 |   a = APROBADO
 238 829 |   b = REPROBADO
```

Figura 133. Evaluación del Modelo 1 para el Departamento de Energía y Mecánica con clasificación lazy utilizando el algoritmo KStar en la ETAPA I

ETAPA II

ÁRBOLES DE DECISIÓN

El algoritmo J48 con 21 atributos obtuvo que de 2186 instancias de prueba 1709 las clasificó correctamente, es decir, el 78.1793%. La matriz de confusión muestra que el algoritmo confundió 241 reprobados como aprobados y 236 aprobados como reprobados. Dando como resultado que el error cuadrático medio sea de 0.4172 y el error absoluto medio sea de 0.2847 como se detalla en la Figura 134.

```
=== Summary ===

Correctly Classified Instances        1709               78.1793 %
Incorrectly Classified Instances       477               21.8207 %
Kappa statistic                          0.5633
Mean absolute error                      0.2847
Root mean squared error                  0.4172
Relative absolute error                 56.9337 %
Root relative squared error             83.4222 %
Total Number of Instances             2186

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,789    0,226    0,786      0,789    0,787      0,563    0,818     0,777     APROBADO
               0,774    0,211    0,778      0,774    0,776      0,563    0,818     0,783     REPROBADO
Weighted Avg.  0,782    0,219    0,782      0,782    0,782      0,563    0,818     0,780

=== Confusion Matrix ===

   a   b   <-- classified as
 883 236 |   a = APROBADO
 241 826 |   b = REPROBADO
```

Figura 134. Evaluación del Modelo 1 para el Departamento de Energía y Mecánica con árboles de decisión J48 en la ETAPA II

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 2186 instancias de prueba 1685 las clasificó correctamente, es decir, el 77.0814%. La matriz de confusión muestra que el algoritmo confundió 250 reprobados como aprobados y 251 aprobados como reprobados. El error cuadrático medio es de 0.3958 y el error absoluto medio sea de 0.3147 como se detalla en la Figura 135.

```
=== Summary ===

Correctly Classified Instances        1685               77.0814 %
Incorrectly Classified Instances       501               22.9186 %
Kappa statistic                          0.5414
Mean absolute error                      0.3147
Root mean squared error                  0.3958
Relative absolute error                 62.9292 %
Root relative squared error             79.1538 %
Total Number of Instances             2186

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,776    0,234    0,776      0,776    0,776      0,541    0,854     0,858     APROBADO
               0,766    0,224    0,765      0,766    0,765      0,541    0,854     0,845     REPROBADO
Weighted Avg.  0,771    0,229    0,771      0,771    0,771      0,541    0,854     0,852

=== Confusion Matrix ===

   a   b   <-- classified as
 868 251 |   a = APROBADO
 250 817 |   b = REPROBADO
```

Figura 135. Evaluación del Modelo 1 para el Departamento de Energía y Mecánica con Regresión Lineal en la ETAPA II

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 2186 instancias de prueba 1703 las clasificó correctamente, es decir, el 77.9048%. La matriz de confusión muestra que el algoritmo solo confundió 254 reprobados como aprobados y 229 aprobados como reprobados. El error cuadrático medio es de 0.4216 y el error absoluto medio sea de 0.2455 como se detalla en la Figura 136.

```
=== Summary ===

Correctly Classified Instances        1703               77.9048 %
Incorrectly Classified Instances       483               22.0952 %
Kappa statistic                          0.5576
Mean absolute error                      0.2455
Root mean squared error                  0.4216
Relative absolute error                 49.0978 %
Root relative squared error             84.3157 %
Total Number of Instances             2186

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,795    0,238    0,778      0,795    0,787      0,558    0,847     0,845     APROBADO
                 0,762    0,205    0,780      0,762    0,771      0,558    0,847     0,828     REPROBADO
Weighted Avg.    0,779    0,222    0,779      0,779    0,779      0,558    0,847     0,837

=== Confusion Matrix ===

   a   b   <-- classified as
 890 229 |   a = APROBADO
 254 813 |   b = REPROBADO
```

Figura 136. Evaluación del Modelo 1 para el Departamento de Energía y Mecánica con utilizando el algoritmo KStar en la ETAPA II

COMPARACIÓN DE ALGORITMOS PARA EL MODELO 1

Los algoritmos J48 y Clasificación Kstar son los que presentan mejores resultados (ver Figura 137), sin embargo, el algoritmo de Regresión Lineal presenta mejores resultados para el Departamento de la Tierra y de la Construcción. Además se observa que hubo un ligero incremento de instancias correctamente clasificadas entre la Etapa I y II exceptuando para el departamento de Energía y Mecánica donde decremento.

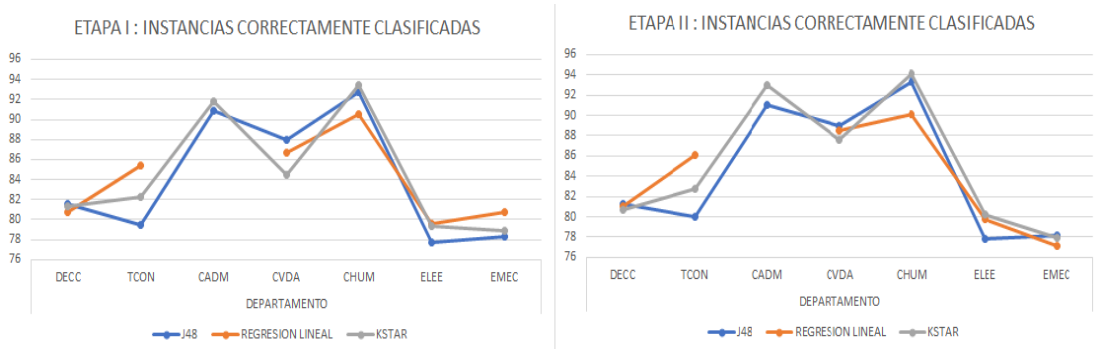

Figura 137. Instancias correctamente Clasificadas por algoritmos en cada etapa del Modelo 1

El error cuadrático medio se mantiene relativamente constante entre las dos etapas, pero disminuye significativamente para los Departamentos de Ciencias Económicas Administrativas y de Comercio y para Ciencias de la Vida (ver Figura 138), mientras que incrementa para el departamento departamento de Energía y Mecánica.

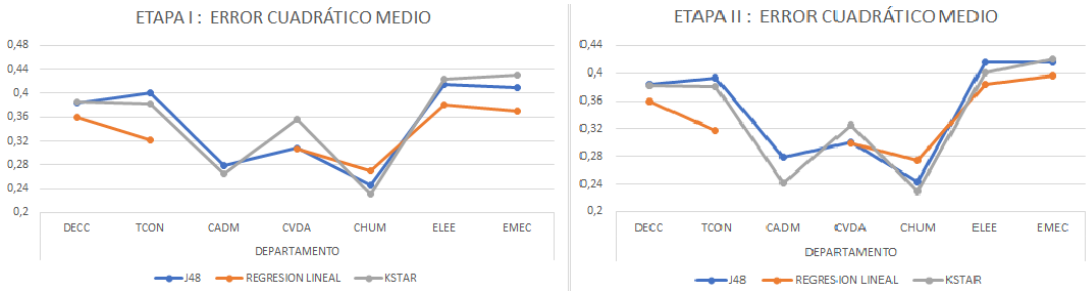

Figura 138. Error cuadrático medio por algoritmos en cada etapa del Modelo 1

El error absoluto medio que mayor incremento tiene es el algoritmo de Regresión Lineal dicha información está relacionada con los resultados arrojados en la Figura 138. De igual manera que en la figura anterior el error disminuye en la Etapa V para el algoritmo J48, siendo el algoritmo de Kstar el que menos error presenta (ver

Figura 139) en las Etapas III – IV.

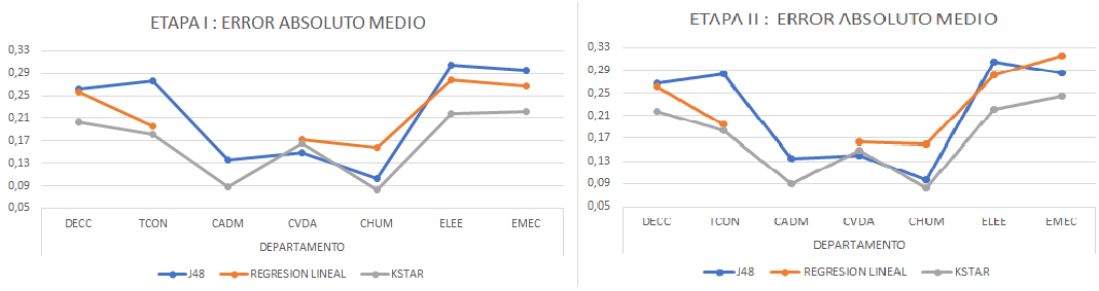

Figura 139. Error absoluto medio por algoritmos en cada etapa del Modelo 1

Considerando los resultados obtenidos en las figuras anteriores se selecciona la Etapa I para el Departamento de Energía y Mecánica y las variables de las Etapas II para los demás departamentos.

MODELO 2: Para predecir si un alumno desertará de la universidad.

Tabla 50.
Variables por etapas del modelo 2

ETAPAS	ATRIBUTOS		N°
ETAPA I	DESERCION CLUSTER_CURSO CLUSTER_MATERIAS_CURSADAS CLUSTER_MATERIAS_REP_ANT CLUSTER_NIVEL CLUSTER_PROMEDIO CLUSTER_PROV_COLEGIO CLUSTER_PROV_NAC CREDITOS EDAD	ESTADO_CIVIL GENERO MATERIAS_CURSADAS MILITAR NIVEL PERIODO_TIPO REGIMENESCOLAR SOSTENIMIENTO TIPO_ALUMNO	19
ETAPA II	DESERCION CLUSTER_CURSO CLUSTER_MATERIAS_CURSADAS CLUSTER_MATERIAS_REP_ANT CLUSTER_NIVEL CLUSTER_PROMEDIO CLUSTER_PROV_COLEGIO CLUSTER_PROV_NAC	EDAD ESTADO_CIVIL GENERO NIVEL PERIODO_TIPO REGIMENESCOLAR SOSTENIMIENTO TIPO_ALUMNO	17

	CREDITOS		

Observación: Las variables que no fueron tomadas en cuenta en la segunda etapa son: materias_cursadas y militar.

ETAPA I

ÁRBOL DE DECISIÓN

El algoritmo J48 con 19 atributos obtuvo que de 701 instancias de prueba 674 las clasificó correctamente, es decir, el 96.1484%. La matriz de confusión muestra que el algoritmo confundió 20 desertores como no desertores y 7 no desertores como desertores. Dando como resultado que el error cuadrático medio sea de 0.1836 y el error absoluto medio sea de 0.0449 como se detalla en la Figura 140.

```
=== Summary ===

Correctly Classified Instances         674               96.1484 %
Incorrectly Classified Instances        27                3.8516 %
Kappa statistic                          0.9228
Mean absolute error                      0.0449
Root mean squared error                  0.1836
Relative absolute error                  8.9742 %
Root relative squared error             36.7212 %
Total Number of Instances              701

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,981    0,059    0,946      0,981    0,963      0,923    0,982     0,975     NO
               0,941    0,019    0,979      0,941    0,960      0,923    0,982     0,977     SI
Weighted Avg.  0,961    0,040    0,962      0,961    0,961      0,923    0,982     0,976

=== Confusion Matrix ===

   a   b   <-- classified as
 353   7 |   a = NO
  20 321 |   b = SI
```

Figura 140. Evaluación del Modelo 2 con árboles de decisión J48 en la ETAPA I

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 701 instancias de prueba 4279 las clasificó correctamente, es decir, el 95.1498%. La matriz de confusión muestra que el algoritmo confundió 14 desertores como no desertores y 20 no

desertores como desertores. El error cuadrático medio es de 0.1888 y el error absoluto medio sea de 0.0751 como se detalla en la Figura 141.

```
=== Summary ===

Correctly Classified Instances         667               95.1498 %
Incorrectly Classified Instances        34                4.8502 %
Kappa statistic                          0.903
Mean absolute error                      0.0751
Root mean squared error                  0.1888
Relative absolute error                 15.0122 %
Root relative squared error             37.7442 %
Total Number of Instances              701

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,944    0,041    0,960      0,944    0,952      0,903    0,992     0,993     NO
                 0,959    0,056    0,942      0,959    0,951      0,903    0,992     0,991     SI
Weighted Avg.    0,951    0,048    0,952      0,951    0,952      0,903    0,992     0,992

=== Confusion Matrix ===

   a   b   <-- classified as
 340  20 |   a = NO
  14 327 |   b = SI
```

Figura 141. Evaluación del Modelo 2 con Regresión Lineal en la ETAPA I

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 701 instancias de prueba 6764 las clasificó correctamente, es decir, el 96.8616%. La matriz de confusión muestra que el algoritmo solo confundió 124 desertores como no desertores. El error cuadrático medio es de 0.154 y el error absoluto medio sea de 0.0397 como se detalla en la Figura 142.

```
=== Summary ===

Correctly Classified Instances         679               96.8616 %
Incorrectly Classified Instances        22                3.1384 %
Kappa statistic                          0.9371
Mean absolute error                      0.0397
Root mean squared error                  0.154
Relative absolute error                  7.9472 %
Root relative squared error             30.7862 %
Total Number of Instances              701

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,986    0,050    0,954      0,986    0,970      0,938    0,998     0,998     NO
                 0,950    0,014    0,985      0,950    0,967      0,938    0,998     0,998     SI
Weighted Avg.    0,969    0,032    0,969      0,969    0,969      0,938    0,998     0,998

=== Confusion Matrix ===

   a   b   <-- classified as
 355   5 |   a = NO
  17 324 |   b = SI
```

Figura 142. Evaluación del Modelo 2 con clasificación lazy utilizando el algoritmo KStart en la ETAPA I

ETAPA II

ÁRBOL DE DECISIÓN

El algoritmo J48 con 17 atributos obtuvo que de 701 instancias de prueba 677 las clasificó correctamente, es decir, el 96.5763%. La matriz de confusión muestra que el algoritmo confundió 15 desertores como no desertores y 9 no desertores como desertores. Dando como resultado que el error cuadrático medio sea de 0.1821 y el error absoluto medio sea de 0.0459 como se detalla en la Figura 143.

```
=== Summary ===

Correctly Classified Instances         677               96.5763 %
Incorrectly Classified Instances        24                3.4237 %
Kappa statistic                          0.9314
Mean absolute error                      0.0459
Root mean squared error                  0.1821
Relative absolute error                  9.1848 %
Root relative squared error             36.4137 %
Total Number of Instances              701

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,975    0,044    0,959      0,975    0,967      0,932    0,975     0,966     NO
               0,956    0,025    0,973      0,956    0,964      0,932    0,975     0,968     SI
Weighted Avg.  0,966    0,035    0,966      0,966    0,966      0,932    0,975     0,967

=== Confusion Matrix ===

   a   b   <-- classified as
 351   9 |   a = NO
  15 326 |   b = SI
```

Figura 143. Evaluación del Modelo 2 con árboles de decisión J48 en la ETAPA II

REGRESIÓN LINEAL

El algoritmo Regresión Lineal generó un modelo donde, de 701 instancias de prueba 676 las clasificó correctamente, es decir, el 96.4337%. La matriz de confusión muestra que el algoritmo confundió 14 desertores como no desertores y 11 no desertores como desertores. El error cuadrático medio es de 0.1707 y el error absoluto medio sea de 0.0636 como se detalla en la Figura 144.

```
=== Summary ===

Correctly Classified Instances         676               96.4337 %
Incorrectly Classified Instances        25                3.5663 %
Kappa statistic                          0.9286
Mean absolute error                      0.0636
Root mean squared error                  0.1707
Relative absolute error                 12.7095 %
Root relative squared error             34.1421 %
Total Number of Instances              701

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,969    0,041    0,961      0,969    0,965      0,929    0,992     0,990     NO
               0,959    0,031    0,967      0,959    0,963      0,929    0,992     0,993     SI
Weighted Avg.  0,964    0,036    0,964      0,964    0,964      0,929    0,992     0,991

=== Confusion Matrix ===

   a   b   <-- classified as
 349  11 |   a = NO
  14 327 |   b = SI
```

Figura 144. Evaluación del Modelo 2 con regresión Lineal en la ETAPA II

CLASIFICACIÓN LAZY

El algoritmo KStar entrega como resultado que de 701 instancias de prueba 689 las clasificó correctamente, es decir, el 98.2882%. La matriz de confusión muestra que el algoritmo solo confundió 8 desertores como no desertores. El error cuadrático medio es de 0.1135 y el error absoluto medio sea de 0.0251 como se detalla en la Figura 145.

```
=== Summary ===

Correctly Classified Instances         689               98.2882 %
Incorrectly Classified Instances        12                1.7118 %
Kappa statistic                          0.9657
Mean absolute error                      0.0251
Root mean squared error                  0.1135
Relative absolute error                  5.0268 %
Root relative squared error             22.6983 %
Total Number of Instances              701

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,989    0,023    0,978      0,989    0,983      0,966    0,999     0,999     NO
                 0,977    0,011    0,988      0,977    0,982      0,966    0,999     0,999     SI
Weighted Avg.    0,983    0,017    0,983      0,983    0,983      0,966    0,999     0,999

=== Confusion Matrix ===

   a   b   <-- classified as
 356   4 |   a = NO
   8 333 |   b = SI
```

Figura 145. Evaluación del Modelo 2 con clasificación lazy utilizando el algoritmo KStart en la ETAPA II

COMPARACIÓN DE ALGORITMOS PARA EL MODELO 2

En la Figura 146 se puede observar que en la segunda etapa los tres algoritmos han incrementado el número de instancias clasificadas correctamente, además el algoritmo con mayor porcentaje de precisión es el Kstar con gran ventaja en relación a los demás algoritmos.

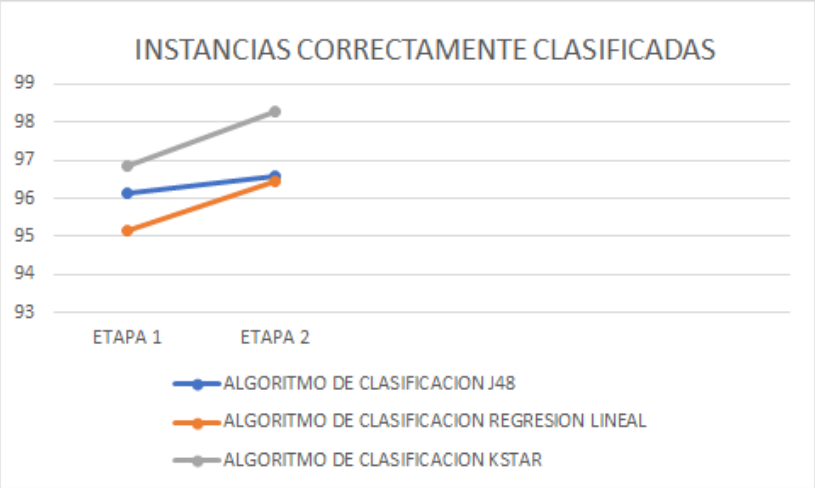

Figura 146. Instancias correctamente Clasificadas por Etapa de cada Modelo 2

La Figura 147 muestra que el algoritmo con menor porcentaje de error cuadrático medio es el KSTAR, seguido por de Regresión Lineal que en la primera etapa tuvo el más alto y que en la segunda etapa disminuye, a su vez el porcentaje de error del algoritmo J48 se mantiene en las dos etapas.

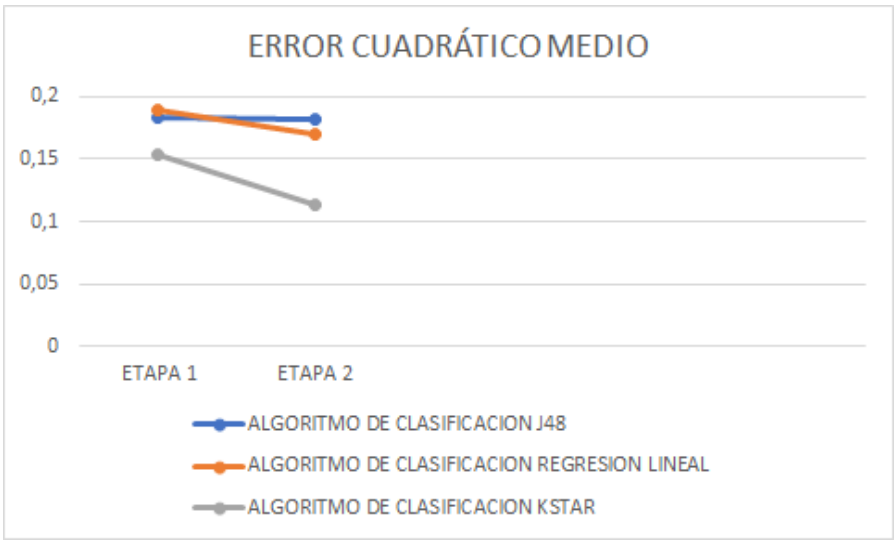

Figura 147. Error cuadrático medio por algoritmos en cada etapa del Modelo 2

Los algoritmos de Regresión Lineal y Kstar, muestran gran variación y disminución en el cambio de etapa, sin embargo, el modelo J48 durante el cambio de etapa ha incrementado en un pequeño porcentaje, pero a diferencia de la Figura anterior que estuvo en un tercer lugar con el mayor porcentaje de error ahora se sitúa en el segundo lugar con una peque diferencia con respecto al Kstar.

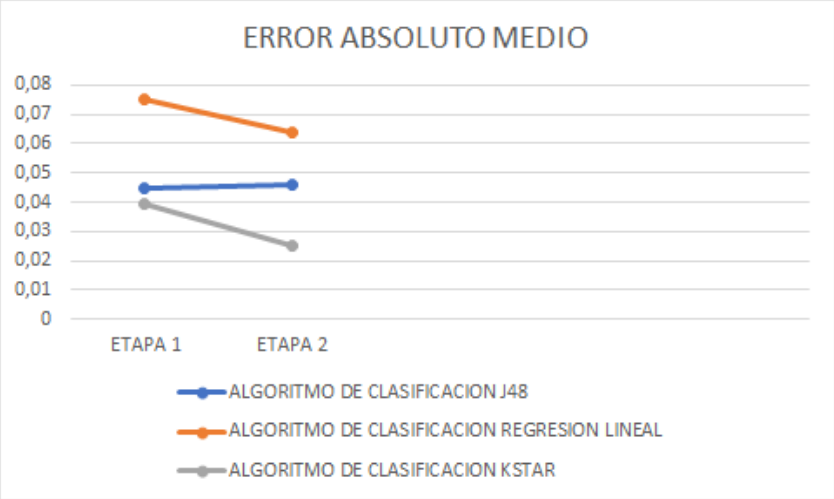

Figura 148. Error absoluto medio por algoritmos en cada etapa del Modelo 2

Considerando los resultados obtenidos en las figuras anteriores se presenta el algoritmo Kstart con mayor número de instancias clasificadas correctamente y menor porcentaje de error, sin embargo, para mantener un estándar en los modelos tanto de reprobados y deserción se opta por el uso del algoritmo J48 que es el que se sitúa en segundo lugar.

f) Evaluación

En este apartado se evalúa los modelos respecto a los criterios de éxito propuestos en los objetivos de negocio del proyecto, se analizará la factibilidad de cada algoritmo seleccionado, en esta etapa se consideran los resultados

obtenidos de la evaluación de la minería de datos de la etapa anterior.

Árboles de Decisión

El algoritmo J48 es factible porque trabaja de forma eficaz y eficiente para los dos modelos propuestos en la Etapa II, entregando un árbol factible y comprensible que permite predecir con el mínimo grado de error los alumnos que pueden reprobar una materia con un error absoluto medio para DECC el 0.2663, TCON el 0.2827, CADM el 0.1352, DVDA el 0.1409, CHUM el 0.0963, ELEE el 0.3055 y EMEC el 0.2847 0.04 en el primer modelo y 0.0459 para el segundo modelo que predice a alumnos que pueden desertar de la universidad.

Regresión Lineal

El algoritmo ViaRegression no es viable para ninguno de los dos modelos planteados ya que no nos ofrece suficientes garantías de que pueda predecir con éxito la deserción de un alumno, en comparación con los otros modelos, teniendo para DECC el 0.2613, TCON el 0.1947, CADM no ofrece ningún resultado, DVDA el 0.164, CHUM el 0.1596, ELEE el 0.2818, EMEC el 0.3147 de error absoluto medio en el primer modelo y para el modelo 2 0.0636, por lo tanto, el modelo debería ser revisado o descartado.

Clasificadores Lazy KStar

El algoritmo KStar es factible porque permite predecir con el mínimo grado de error los alumnos que pueden reprobar una materia con un error absoluto medio para DECC el 0.2187, TCON el 0.1838, CADM el 0.0899, DVDA el 0.1496, CHUM el 0.0819, ELEE el 0.2218, EMEC el 0.2455 para el primer modelo y 0.0251 para el segundo modelo que predice a alumnos que pueden desertar de la universidad, sin embargo, el costo de procesamiento es bastante elevado respecto al algoritmo J48.

Finalmente como se estableció en los objetivos de negocio propuestos y

criterios de éxito, se han identificado posibles factores que inciden en el rendimiento académico en la Etapa II del modelo 1 y 2 son: Colegio, créditos de la materia, créditos tomados por el alumno en el semestre, cantidad de alumnos por curso, departamento de la asignatura, docente, edad, estado civil, etnia, genero, cantidad de materias aprobadas, cantidad de materias repetidas, si es militar o no, nivel que corresponde al número de créditos aprobados, parroquia de residencia, tipo del periodo si es en diciembre o no, promedio del alumno del semestre anterior, sostenimiento del colegio, tipo del alumno si ingreso con prueba del SENESCYT o no, el tipo de horario de la materia(mañana, tarde o noche), tipo de matrícula en la materia(primera, segunda o tercera), , tipo de sostenimiento del colegio (fiscal, particular o fiscomisional) y régimen escolar en el que estudio el colegio.

4.1.5. IMPLEMENTACIÓN

En esta última fase se explica al cliente poniendo en funcionamiento el proyecto construido en las fases de la metodología y exponer resultados obtenidos para que solos pueda entender fácilmente. Para que el cliente pueda hacer uso de los resultados obtenidos durante el proceso de minería de datos se ha diseñado e implementado un portal web de consulta, donde su proceso de desarrollo se detalla a continuación:

Diseño de los componentes

El diseño del portal web se ha optado por una arquitectura de múltiples capas que define bloques o capas estructuradas lógicamente y que detalla las responsabilidades exactas de cada capa y la forma que tienen de relacionarse entre sí como se observa en la Figura 149. Fundamentalmente en el presente proyecto se utiliza el que está constituido por tres capas que son: modelo, vista y controlador.

- **Modelo**: Se encarga de los datos, es decir de dar persistencia a la aplicación, generalmente consultando la base de datos realizando actualizaciones, consultas, búsquedas, etc.

- **Vista**: Es la representación visual de los datos, es la encargada de la interacción del usuario con la aplicación.
- **Controlador**: Se encarga de la gestión de la petición del usuario realizadas en la capa de la vista y las peticiones de la capa del modelo.

Las tres capaz son de vital importancia sin embargo al ser este un proyecto de minería de datos el núcleo de la aplicación es el modelo generado con la ayuda de la librería de WEKA.

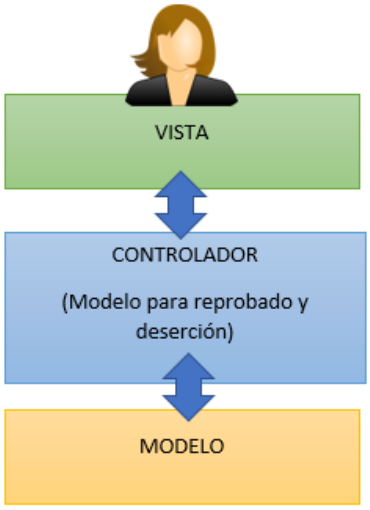

Figura 149. Diagrama de componentes del sistema

Además, para aplicar la arquitectura MVC como se mencionó en el punto anterior se ha utilizado varias herramientas las cuales son:

- Liberia de Weka para java: Que ayudará a generar el modelo de predicción
- Oracle 11G R2: Gestor de base de datos para acceso a la información academica y personal
- Netbeans v7.4: Como IDE de desarrollo en lenguaje de programación JAVA
- Glassfish v4.1: Como servidor de aplicaciones web.

Diseño Orientado a objetos

Casos de uso

El portal web tiene únicamente dos casos de usos casi similares, teniendo como actor principal al estudiante, docente u autoridad quien es el encargado de llenar los campos requeridos y de enviarlos para poder obtener los resultados de la consulta. A continuación, se los ilustra en la Figura 150.

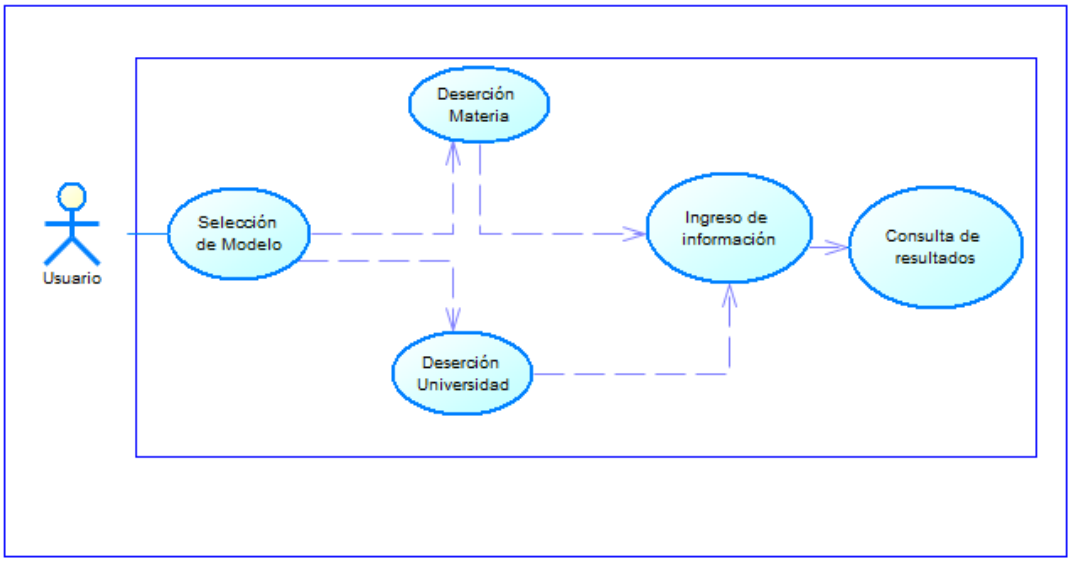

Figura 150. Caso de Uso del portal web

Tabla 51.
Descripción del caso de uso de deserción de materias

Descripción del caso de uso	
Nombre	Deserción de Materias
Descripción	Permite en base a los datos ingresados predecir si un alumno aprobará o reprobará una materia

Tabla 52.
Descripción del caso de uso de deserción de la Universidad

Descripción del caso de uso	
Nombre	Deserción de la Universidad
Descripción	Permite en base a los datos ingresados predecir si un alumno abandonará la universidad

Diagrama de Clases

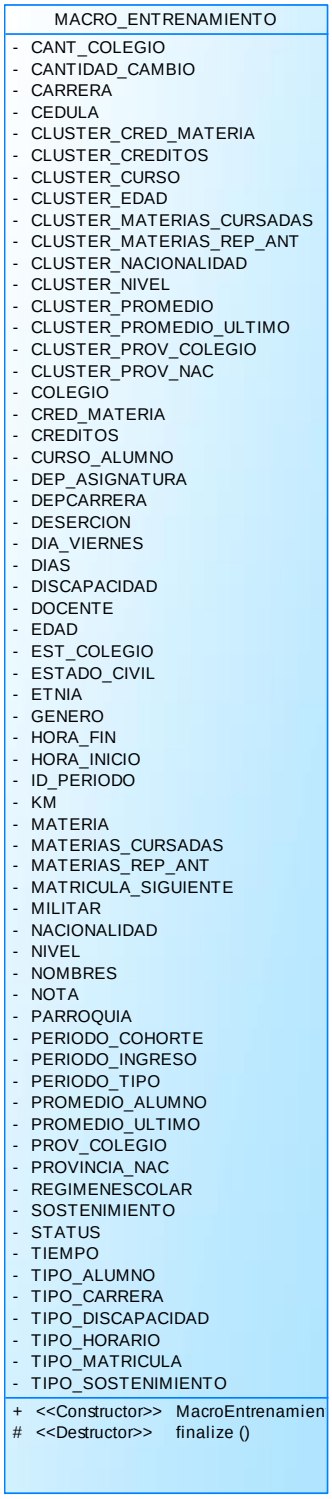

Figura 151. Diagrama de Clases

Para el presente proyecto se ha diseñado un modelo con 23 tablas sin embargo para la ejecución de algoritmo y el modelado en el proceso de minería de datos para facilitar y optimizar el proceso de extracción se ha creado una tabla maestra con el nombre de MACRO_ENTRENAMIENTO con 63 atributos que contiene una fusión de todas (ver Figura 151).

Diseño de la Base de Datos

Al realizar minería de datos es recomendable que se tenga una base de datos exclusiva para dicho proceso cuya estructura permita almacenar e identificar de fácilmente a cada registro de tal forma que el proceso se realice de la forma más rápida y precisa (ver Figura 152).

MATERIA	VARCHAR2 (2000 BYTE)
MATERIAS_CURSADAS	NUMBER
MATERIAS_REP_ANT	NUMBER
MATRICULA_SIGUIENTE	VARCHAR2 (2 BYTE)
MILITAR	VARCHAR2 (2000 BYTE)
NACIONALIDAD	VARCHAR2 (100 BYTE)
NIVEL	NUMBER
NOMBRES	VARCHAR2 (2000 BYTE)
NOTA	NUMBER
PARROQUIA	VARCHAR2 (300 BYTE)
PERIODO_COHORTE	VARCHAR2 (2000 BYTE)
PERIODO_INGRESO	VARCHAR2 (2000 BYTE)
PERIODO_TIPO	VARCHAR2 (20 BYTE)
PROMEDIO_ALUMNO	NUMBER
PROMEDIO_ULTIMO	NUMBER
PROV_COLEGIO	VARCHAR2 (100 BYTE)
PROVINCIA_NAC	VARCHAR2 (100 BYTE)
REGIMENESCOLAR	VARCHAR2 (2000 BYTE)
SOSTENIMIENTO	VARCHAR2 (2000 BYTE)
STATUS	VARCHAR2 (20 BYTE)
TIEMPO	VARCHAR2 (20 BYTE)
TIPO_ALUMNO	VARCHAR2 (20 BYTE)
TIPO_CARRERA	VARCHAR2 (20 BYTE)
TIPO_DISCAPACIDAD	VARCHAR2 (2000 BYTE)
TIPO_HORARIO	VARCHAR2 (20 BYTE)
TIPO_MATRICULA	VARCHAR2 (7 BYTE)
TIPO_SOSTENIMIENTO	VARCHAR2 (20 BYTE)

Figura 152. Diseño de la Base de Datos

Diseño de la Interfaz de usuario

Se diseño un portal web con el objetivo de generar consultas de forma amigable, rápida y efectiva. El portal permitirá consultas para predecir dos tipos de modelos de deserción el primero para predecir la probabilidad de un estudiante para reprobar una materia y la segunda para predecir la probabilidad de un estudiante para desertar o abandonar la universidad. A continuación, se presentan las distintas interfaces que contiene el sistema:

En la Figura 153 se observa la pantalla de Inicio del portal web, la misma que está constituida por dos opciones que representan a los modelos seleccionados, además muestra información del tipo de algoritmo de minería de datos utiliza junto con el nombre de la herramienta.

Figura 153. Pantalla de inicio del portal

La Figura 154 muestra la interfaz para la predicción de que un alumno apruebe una materia que se obtiene al seleccionar la opción "MODELO 1", dónde se deben ingresar los datos requeridos para obtener el resultado.

Figura 154. Interfaz para predicción a probación o reprobación de materias

La Figura 155 ilustra los resultados obtenidos al realizar la predicción con los

datos ingresados en la Figura 154 y adicionalmente los datos personales y académicos del alumno.

Figura 155. Resultados de la predicción de la deserción de una materia

La Figura 156 muestra la interfaz para la predicción de que un alumno apruebe una materia que se obtiene al seleccionar la opción “MODELO 2”, dónde se deben ingresar los datos requeridos para obtener el resultado.

Figura 156. Interfaz para predicción de deserción de la Universidad

La Figura 157 ilustra los resultados obtenidos al realizar la predicción con los datos ingresados en la Figura 156 y adicionalmente los datos personales y académicos del alumno.

Resultado: NO

Datos Personales

Característica	Valor
Nombres	PAGUAY FLORES, SANDRA ELIZABET
Cédula	1718325523
Departamento de la Carrera	CIENCIAS DE LA COMPUTACION
Carrera	[PRES] SISTEMAS E INFORMATICA
Tipo Alumno	SENESCYT
Edad	23
Colegio	MILITAR ABDON CALDERON
Sostenimiento del Colegio	FISCOMISIONAL
Provincia del Colegio	PICHINCHA
Género	FEMENINO
Militar	NO
Provincia de nacimiento	PICHINCHA
Estado Civil	SOLTERO
Etnia	MESTIZOS

Datos Académicos

Característica	Valor
Materias Cursadas	65
Materias Repetidas	0
Créditos actuales	30
Créditos de la Materia	4
Tipo de Matrícula de la Materia	PRIMERA
Departamento de la Asignatura	CIENCIAS DE LA COMPUTACION
Créditos Aprobados	265
Promedio semestre anterior	18.19
Tipo Horario	MEDIA MANANA-TARDE

Figura 157. Resultados de la predicción de la deserción de la universidad

CAPÍTULO IV

4.1.CONCLUSIONES

Se eligió el algoritmo de árboles de decisión J48 luego de un exhaustivo análisis por su fácil interpretación y mejor precisión en comparación con los otros algoritmos (Regresión Lineal y Kstar).

Los algoritmos de clasificación utilizados tienden a ignorar las clases con menor frecuencia, es necesario un balance de los datos creando un nuevo conjunto conformado con 50% registros de alumnos aprobados y 50% reprobados. Con la finalidad de solventar este problema se crearon modelos individuales por departamento de la carrera (7 modelos).

Uno de los principales retos de este proyecto se presentó durante la etapa de limpieza de datos, este proceso fue aplicado en varias iteraciones hasta conseguir que los datos de entrada del modelo sean aceptables y no generen ruido que afecte significativamente el resultado del modelo.

CRISPM DM es una guía sencilla y flexible que facilitó el desarrollo del proyecto ofreciendo tareas claras y puntuales para su implementación. Por otro lado la herramienta.

WEKA ofrece una gran cantidad de técnicas de minería de datos con un alto grado de usabilidad y compatible con la mayoría de plataformas.

Los posibles factores que inciden en la deserción universitaria son: cantidad de créditos tomados en el semestre, departamento de la materia, edad, género, cantidad de materias aprobadas, cantidad de materias repetidas, promedio, tipo de alumno, tipo de horario, y tipo de matrícula en la materia.

Por lo tanto, este proyecto de investigación aporta significativamente a la sociedad brindando un panorama completo de la deserción universitaria mediante modelos de minería de datos. Dichos modelos tienen un alto porcentaje de precisión; varían entre 78% y 93%, sin embargo, se podría mejorar el resultado mediante la inclusión de nuevas características del alumno como son: situación socioeconómica, situación

laboral o conductual que actualmente no se encuentran disponibles.

4.2. RECOMENDACIONES

La selección de la técnica de minería de datos debe ir acorde con los objetivos del proyecto y con el tipo de datos que se utilizarán para la construcción del modelo.

Para la etapa de preparación de datos es recomendable buscar un software que facilite la limpieza con el objetivo de eliminar datos incompletos, erróneos o duplicados. Además, es necesario que ciertas variables sean discretizadas para equilibrar el balance de datos y obtener puntos de corte más intuitivos.

Es necesario que las autoridades de la Universidad de las Fuerzas Armadas – ESPE tomen en cuenta los resultados obtenidos en esta investigación con el fin de tomar decisiones y proponer estrategias que mejoren el servicio entregado a los alumnos, prevenir y disminuir el índice de deserción estudiantil.

Por otra parte, se recomienda que la Unidad de Tecnologías de la Información UTIC’s tome como referencia este proyecto de tesis para mejorar su bodega de datos (Data Warehouse) que soporten futuros proyectos de Minería de Datos y Business Intelligence permitiendo mejorar la toma de decisiones.

4.3. LÍNEAS DE TRABAJO FUTURO

Como trabajos futuros se propone la implantación del modelo de predicción en un sistema online de alerta temprana permitiendo que el proceso de entrenamiento del modelo se adapte en tiempo real, a la información de los alumnos y vaya encontrando nuevos patrones de que influyan en la deserción.

Además, se plantea generar nuevos modelos, recopilando información sobre estilos de aprendizaje, variables socioeconómicas y relaciones interpersonales de los alumnos para obtener un modelo que revele nuevos patrones que influyen en la deserción universitaria.

Evaluar nuevas tendencias de modelos de predicción, en especial modelos que

tengan mayor relación a variables subjetivas como por ejemplo lógica difusa y comparar sus resultados.

Es importante dar continuidad a este proyecto de investigación, enfocando los esfuerzos a otras problemáticas de la universidad como, por ejemplo: predecir el tiempo que tardaría en graduarse un alumno de la universidad.

REFERENCIAS BIBLIOGRÁFICAS

AGUILAR, J. & Estrada, C. (2012). *Minería de Datos (Data Mining).* [en línea]. Disponible en: http://www.slideshare.net/miriam1785/mineria-de-datos-8768313 [2016, 11 de febrero].

AMIE. (2012). *Ecuador*: *Indicadores educativos 2011-2012*. [On line] Disponible en la World Wide Web: https://educacion.gob.ec/wp-content/uploads/downloads/2013/10/Indicadores_Educativos_10-2013_DNAIE.pdf

AMIE. (2016) Listado Instituciones Educativas Distribuidas Por Zona Distrito y Circuito Disponible en: https://es.scribd.com/doc/209711949/Listado-Instituciones-Educativas-Distribuidas-Por-Zona-Distrito-y-Circuito

AMORIM, R. C.; MIRKIN, B (2012). *Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering*. Pattern Recognition 45 (3): 1061–1075.

Andrienko, N., Andrienko, G., Savinov, A. and Wettschereck, D. (2017). *Descartes and Kepler for Spatial Data Mining*. [online] Ercim.eu. Available at: https://www.ercim.eu/publication/Ercim_News/enw40/andrienko1.html [Accessed 12 Mar. 2017].

ASAMBLEA NACIONAL, (2010). *Ley Orgánica De Educación Superior*. Art 100.- La Evaluación Externa.

BAGGA SIMMI, G. (2012). *Applications of Data Mining*. International Journal for Science and Emerging Technologies with Latest Trends .

BERMEO, M (2013). *Árboles de decisión aplicados a la minería de datos.* Obtenido de: http://es.slideshare.net/Migu3lB/ejemplificacion-de-arboles-de-de

BERZAL, F. (2017). *Patrones secuenciales*. [online] Available at: http://elvex.ugr.es/idbis/dm/slides/22%20Pattern%20Mining%20-%20Sequences.pdf [Accessed 12 Mar. 2017].

CARVAJAL, L (2013). *El método deductivo de investigación*. Obtenido de: http://www.lizardo-carvajal.com/el-metodo-deductivo-de-investigacion/

CHATFIELD, C. 2003. *The analysis of time series: an introduction, CRC press*

CEAACES (2015). *Adaptación del Modelo de Evaluación Institucional de Universidades y Escuelas Politécnicas 2013 al Proceso de Evaluación, Acreditación y Recategorización de Universidades y Escuelas Politécnicas 2015*

Obtenido de:

http://www.ceaaces.gob.ec/sitio/wpcontent/uploads/2013/10/ADAPTACIO%CC%81N-DEL-MODELO-DE-EVALUACIO%CC%81N-INSTITUCIONAL-DE-UNIV.-Y-ESC.-POLITE%CC%81C.-2013-AL-PROCESO-DE-EVAL-ACREDIT-Y-RECATEG-DE-UNIVERS.-Y-ESC.-POLIT-2015PLENOFINAL-NOTIF.pdf.

Cubero J. y Berzal F. (2014) Departamento de Ciencias de la Computación e I.A. Universidad de Granada. *Sistemas Inteligentes de Gestión Guión de Prácticas de Minería de Datos*. Obtenido de:

http://elvex.ugr.es/decsai/intelligent/workbook/D1%20KNIME.pdf

DATAPRIX. (2015). *El modelo de referencia CRISP-DM.* Obtenido de http://www.dataprix.com/es/el-modelo-referencia-crisp-dm

DE LA CALLE, J. (2017). *Carga de datos – Weka*. [online] Opiniones sobre Ciencia. Available at: https://opinionessobreciencia.com/2016/03/17/carga-de-datos-weka/ [Accessed 12 Mar. 2017].

Dpto. de Matemática Aplicada (Biomatemática). Facultad de Biología. UCM. *Análisis de Regresión Lineal Simple.* Obtenido de: http://e-stadistica.bio.ucm.es/cont_mod_1.html

GALLARDO, J. (2010). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* [en línea]. http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf. [fecha de consulta: 05 /05/2013].

EL COMERCIO. (2016). *El 26% de los universitarios se retiró en los primeros años.* [On line] Disponible en la World Wide Web: http://www.elcomercio.com/actualidad/ecuador-universitarios-desercion-educacion-jovenes.html

EL COMERCIO. (2016). *Los alumnos aún desertan de las carreras universitarias.* [On line] Disponible en la World Wide Web: http://www.elcomercio.com/tendencias/alumnos-desercion-carreras-universidad-educacion.html

GALÁN Cortina, V. (2016). Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario (Bachelor's thesis).

GARCÍA, M. N. M., Quintales, L. A. M., García-Peñalvo, F. J., & Martín, M. J. P. (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. In ADIS.

GARCÍA, R. & BORRAJO, D. (2000) "*An Integrated Approach of Learning, Planning and Executing,*" Journal of Intelligent and Robotic Systems, pp. 47-78

GONZÁLEZ, L. E. (2005). *Digital Observatory for higher education in Latín América and The Caribbean.* Digital Observatory for higher education in Latín América and The Caribbean.

GONZÁLEZ, L. E. y D. URIBE(2002). *Estimaciones sobre la "repitencia" y deserción en la Educación Superior chilena. Consideraciones sobre sus implicaciones.* Revista Calidad de la Educación Superior 17, 2o Semestre. [On line] Disponible en la World Wide Web: www.cse.cl/public/Secciones/seccionpublicaciones/publicaciones_revista_calidad_detalle.aspx?idPublicacion=35 Obtenido el 18/01/08.

HEREDIA, D., Amaya, Y., & Barrientos, E. (2015). *Student Dropout Predictive Model Using Data Mining Techniques.* IEEE Latin America Transactions, 13(9), 3127-3134.

HIMMEL K. Erika. (2002) *Modelo de Análisis de la Deserción estudiantil en la Educación Superior*, Calidad en la Educación, 17, 91- 108.

IBM, (2017). *SPSS Modeler*. [online] Www-03.ibm.com. Available at: http://www-03.ibm.com/software/products/es/spss-modeler [Accessed 12 Mar. 2017].

LIN, S. and ANTON, C. (2017). *Evaluation of DBMiner System*. [online] CMPUT 690 Assignment 1. Available at: https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/work/group2/dbminer.html [Accessed 12 Mar. 2017].

LÓPEZ, C. P. (2007). *Minería de datos: técnicas y herramientas.* Editorial Paraninfo.

MABROUKEH, N. R.; EZEIFE, C. I. (2010). *"A taxonomy of sequential pattern mining algorithms".* ACM Computing Surveys

MOLINA, J., & García, J. (2006). *Técnicas de análisis de datos: aplicaciones prácticas utilizando Microsoft Excel y Weka*. Universidad Carlos III de Madrid España.

MARQUES, M. P. (2013). *Minería de datos. Técnicas de Segmentación*. Madrid: Createspace Independent Publishing Platform.

ORACLE, (2017). *ODM on the Cloud*. [online] Oracle.com. Available at: http://www.oracle.com/technetwork/database/enterprise-edition/odm-on-the-cloud-100619.html [Accessed 12 Mar. 2017].

PEREIRA, R. T., Romero, A. C., y Toledo, J. J. (2013). *Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil* [Application of data

mining in extracting student dropout profiles]. Ventana Informática, (28).

PRADEEP, A., Das, S., & Kizhekkethottam, J. J. (2015, February). *Students dropout factor prediction using EDM techniques.* In Soft-Computing and Networks Security (ICSNS), 2015 International Conference on (pp. 1-7). IEEE.

RODRÍGUEZ, Olmedar. (2010). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM,* 2010.

ROMÁN, M. (2013). *Factores asociados al abandono y la deserción escolar en América Latina: Una Mirada de Conjunto*. REICE. Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación.

SHALEENA, K. P., y Paul, S. (2015, Marzo). *Data mining techniques for predicting student performance.* In Engineering and Technology (ICETECH), 2015 IEEE International Conference on (pp. 1-3). IEEE.

TAMAYO, M. (2004). *El proceso de la investigación científica.* Editorial Limusa.

TEAMBETAMETRICA, (2017). *Importar una base de datos de Excel a R Studio*. [online] Betametrica.com.ec. Available at: http://www.betametrica.com.ec/2015/05/27/importar-una-base-de-datos-de-excel-a-r-studio/ [Accessed 12 Mar. 2017].

TIMARÁN R y Jiménez,J (2014, Noviembre). *Detección de Patrones de Deserción Estudiantil en Programas de Pregrado de Instituciones de Educación Superior con CRISP-DM.* Congreso Iberoamericano de Ciencia, Tecnología, Innovación y Educación. Artículo 758.

UNESCO. (2015). *Situación Educativa de América Latina y el Caribe: Hacia la educación de calidad para todos al 2015,* 2015.

UNESCO. (2009). *Indicadores de la educación: Especificaciones Técnicas.* Recuperado el 07 de febrero del 2017, de http://www.uis.unesco.org/Library/Documents/eiguide09-es.pdf

VALERO, C. S.(2009) *Minería de datos para series temporales*.

Orea, S. V., Vargas, A. S., & Alonso, M. G. (2005). *Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos*. Ene, 779(73), 33.

VALLEJOS, S. *Minería de Datos. Corrientes,* Argentina, Universidad Nacional de Noreste, 2006, págs. 11-16.

WALPOLE, R. E., Myers, R. H., & Myers, S. L. (1999). *Probabilidad y estadística para ingenieros*. Pearson Educación. WEISS, S.M. y Indurkhya, N. "Predictive Data Mining. A practical guide", Morgan Kaufmann Publishers, San Francisco,

1998.

WHATASOFTWARE, (2017). *RapidMiner vs. Datameer: Reviews of RapidMiner , Datameer Business Intelligence & Analytics Software. Compare features, Pricing | WhataSoftware*. [online] Whatasoftware.com. Available at: https://www.whatasoftware.com/compare/RapidMiner-vs-Datameer/MTc0/MTEy [Accessed 12 Mar. 2017].

WINTERS, T (2006) *Educational Data Mining: Collection and Analysis of Score Matrices for Outcomes- Based Assessment* .USA, University of California: Riverside

Zhang S., C. Zhang, and Q. Yang. *Data preparation for data mining. Applied Articial Intelligence*, 17:375381, 2003.

ANEXOS

ANEXO 1: Cronograma del proyecto

	Task Mod	Task Name	Duration	Start	Finish	Predecessors
1		Análisis taxonómico	175 days	Sun 1/1/17	Thu 7/20/17	
2		**Capítulo I: INTRODUCCIÓN**	**20 days**	**Sun 1/1/17**	**Tue 1/24/17**	
3		1.1. Antecedentes	4 days	Sun 1/1/17	Thu 1/5/17	
4		1.2. Problematica	4 days	Thu 1/5/17	Mon 1/9/17	3
5		1.3. Justificación	3 days	Mon 1/9/17	Wed 1/11/17	4
6		1.5. Objetivos	3 days	Wed 1/11/17	Fri 1/13/17	
7		1.6. Alcance	3 days	Fri 1/13/17	Mon 1/16/17	6
8		**Capítulo II: MARCO TEORICO**	**25 days**	**Mon 1/16/17**	**Mon 2/13/17**	**6**
9		2.1. Deserción Universitaria	6 days	Mon 1/16/17	Sat 1/21/17	
10		**2.2. Minería de Datos**	**19 days**	**Sat 1/21/17**	**Sat 2/11/17**	**9**
11		2.2.1 Técnicas de Minería de Datos	5 days	Sat 1/21/17	Thu 1/26/17	
12		2.2.2 Herramientas de Minería de Datos	5 days	Thu 1/26/17	Tue 1/31/17	11
13		2.2.3 Áreas de aplicación de la minerìa de datos	3 days	Tue 1/31/17	Thu 2/2/17	12
14		2.2.4 Proyectos Relacionados a la Minería de Datos en la deserción estu	4 days	Thu 2/2/17	Mon 2/6/17	13
15		2.4. Metodologías de MIneria de Datos	2 days	Mon 2/6/17	Tue 2/7/17	14,10
16		**Capítulo III: DESARROLLO DEL PROYECTO**	**120 days**	**Tue 2/7/17**	**Fri 6/23/17**	
17		3.1. Comprensión del Negocio	10 days	Tue 2/7/17	Fri 2/17/17	
18		3.2. Comprensión de los Datos	20 days	Fri 2/17/17	Sat 3/11/17	17
19		3.3. Preparación de los Datos	40 days	Sat 3/11/17	Wed 4/26/17	18
20		3.4. Modelado	20 days	Wed 4/26/17	Thu 5/18/17	19
21		3.5. Evaluación	10 days	Thu 5/18/17	Mon 5/29/17	20
22		3.6. Implantación	20 days	Mon 5/29/17	Tue 6/20/17	21
23		**Capítulo IV: CONCLUSIONES, RECOMENDACIONES Y LÍNEAS DE TRABAJO FU**	**10 days**	**Tue 6/20/17**	**Fri 6/30/17**	**16**
24		4.1. Conclusiones	4 days	Mon 6/26/17	Thu 6/29/17	
25		4.2. Recomendaciones	3 days	Thu 6/29/17	Sat 7/1/17	24
26		4.3. Líneas de Trabajo Futuro	3 days	Sat 7/1/17	Tue 7/4/17	25