Universidad Internacional de La Rioja (UNIR)

Escuela de Ingeniería

Máster en Análisis y Visualización de Datos Masivos

Detección de patrones académicos en asignaturas con altas tasas de suspensos

Trabajo Fin de Máster

Presentado por: Landacay Jaramillo, Katty Juliana

Director/a: Cubo Javier, PhD

Ciudad: Loja, Ecuador
Fecha: 14 de octubre del 2017

Contenido

Índice de tablas

Tabla de ilustraciones

RESUMEN

El suspenso o la reprobación de asignaturas, una de las causas de la desmotivación estudiantil, es también una de las razones por la que los estudiantes abandonan o se cambian de carrera e incluso de institución educativa. El objetivo de este proyecto de investigación es detectar o encontrar patrones relacionados a la reprobación de las asignaturas con altas tasas de reprobación, denominadas materias *filtradoras.* Esto mediante la aplicación de minería de datos y utilizando una metodología relacionada a proyectos de este tipo, que facilite y guíe este proceso, en este caso se ha seleccionado CRISP-DM.

Para este fin se han recopilado datos de variables académicas relacionadas al: *docente* que dicta la materia, al *estudiante* matriculado y a la *materia* ofertada. Se utilizan variables relacionadas al docente, debido a que este es considerado dentro del proceso de enseñanza como el facilitador o guía del aprendizaje. Los datos recolectados, están relacionados a factores académicos, no se han utilizado datos de factores socioeconómicos, ni familiares.

Los datos provienen de la modalidad a distancia de una Institución de Educación Superior, de este conjunto de datos, se han seleccionado sólo las materias filtradoras, de tipo *troncal o de carrera* (materias relacionadas completamente a la especialización de la carrera) y que han tenido, en los últimos tres semestres o periodos académicos, altos niveles de reprobación.

Se pretende aplicar varios tipos de algoritmos para la detección de patrones, así mismo se realizará una validación de los resultados que estos generen, con dos propósitos principales, el primero, seleccionar el o los algoritmos que se podrían utilizar en problemas similares de minería de datos educativos, y el segundo, para filtrar los mejores patrones que pueda ayudar en la toma de decisiones de la institución con respecto a la reprobación académica.

La contribución que se desea brindar al desarrollar el presente trabajo, y que lo diferencia de otros proyectos investigativos, es el problema seleccionado, l*a reprobación académica como causa para la deserción académica*, por lo cual se pretende a través de esta investigación, generar conocimiento, que pueda ser utilizado por la institución o instituciones académicas, que les permita detectar, de forma temprana, el riesgo de reprobación de un estudiante, con la finalidad de disminuir los altos índices de reprobación, y minimizar la existencia de materias filtradoras, aumentando así las tasas de continuidad académica.

Palabras Clave: Minería de datos, Comparación de algoritmos, Reprobación académica, Educación Superior a Distancia.

ABSTRACT

Suspicion or reprobation of subjects, one of the causes of student demotivation, is also one of the reasons why students abandon or change their careers and even of an educational institution. The objective of this research project is to detect or find patterns related to the reprobation of subjects with high failure rates, called filtering materials. This in the application of data mining and the use of a methodology related to projects of this type, which facilitate and guide this process, in this case CRISP-DM has been selected.

To this end, data have been collected on academic variables related to: teacher who dictates the subject, the student enrolled and the subject offered. Variables related to the teacher are used, because it is considered within the teaching process as the facilitator or guide of learning. The data collected are related to academic factors, no socioeconomic and family factors data have been used.

The data come from the distance modality of an Institution of Higher Education, from this data set, only the filtering materials, of type trunk or of race (subjects related completely to the specialization of the race) have been selected , in the last three semesters or academic periods, high levels of reprobation.

It is intended to apply several types of algorithms for the detection of patterns, as well as a validation of the results that they generate, with two main purposes, the first, to select the algorithm that could be used in similar data mining problems Para to help students make institutional decisions about academic failure.

The contribution that it is wanted to offer in developing the present work, and that differentiates it from other research projects, is the problem selected, academic failure as a cause for academic desertion, which is why through this research it is intended to generate knowledge, which can be used by the institution or academic institutions, which allows them to detect, at an early stage, the risk of student disapproval, in order to reduce the high rates of reprobation, and to minimize the existence of filtering materials, thus increasing the rates of academic continuity.

Keywords: Data Mining, Algorithm Comparison, Academic Reprobation, Higher Distance Education.

1 CAPÍTULO I: INTRODUCCIÓN

El valor que se le está dando a los datos, que se poseen dentro de las Instituciones educativas, ha ido incrementando en los últimos años. Las instituciones académicas desean explotar con mayor continuidad sus datos, desean tomar decisiones académicas en base a sus datos, y no a tendencias mundiales, experiencias personales o por casos fortuitos. Además, mejorar la calidad de los datos ingresados en los sistemas informáticos que poseen, también ha tomado mayor exigencia, puesto que se han dado cuenta, que este factor está relacionado a los resultados que desean conocer para la toma de decisiones. Esto se ve reflejando en el incremento de la investigación y la participación mundial en eventos referidos a la minería de datos educativos, como: *International Conference on Educational Data Mining* e *International Conference on Learning Analytics and Knowledge.*

La minería de datos es una ciencia muy utilizada en diversos entornos, incluido el académico, facilita realizar análisis de pequeñas o grandes cantidad de datos, descubriendo en ellos, conocimiento nuevo y comprensible. Es importante considerar que la implementación y uso de minería de datos en la educación no tienen tantos años, como el uso de esta ciencia en otras áreas (*industriales, económicas, bancarias,* entre otras). Pero su aplicación es cada vez más valiosa, puesto que ha permitido realizar personalizaciones de contenidos, de plataformas, de horarios, de material de aprendizaje, mayor enfoque en métricas del aprendizaje, entre otros; según se mencionada en el informe Horizon 2017, compartido por (Dirección de Innovación Educativa DIE-UNAH, 2017).

Si se analiza desde un punto de vista consecuente, y en base a lo mencionado en el libro "*La Responsabilidad Social de las Universidades: Implicaciones para la América Latina y el Caribe",* del autor (Aponte Hernández, 2015), el minar los datos educativos debe realizarse con la finalidad de, mejorar el aprendizaje y rendimiento académico de los estudiantes, hacia la generación de mejores profesionales, que promuevan una sociedad más productiva. Así mismo, una educación de calidad permitirá disminuir las *tasas de desempleo, pobreza, analfabetismo, prostitución, robo, drogadicción, entre otras*, que están relacionadas, de cierto modo, a la falta de formación académica de una sociedad.

Nuestra sociedad es cada más competitiva, por lo cual, es mayor la exigencia de una preparación académica eficiente.

Ante lo mencionado, si una persona ingresa a una institución educativa y por temas académicos deserta de sus estudios, es para la institución una responsabilidad con la sociedad, puesto que, esta persona tendrá mayores dificultades en su desenvolvimiento

profesional, que una persona que si finaliza sus estudios. Por ello las instituciones deben estar en constante innovación, para conocer como impulsar en sus estudiantes la culminación de sus estudios, sin descuidar la calidad educativa, con visión al desarrollo de una sociedad sostenible, más humana y progresista.

1.1 Justificación

Las instituciones educativas conocedoras de la estrecha relación entre la deserción de estudiantes y la reprobación de materias, necesitan detectar si existen características o factores comunes entre los estudiantes que reprueban las denominadas materias *filtradoras*[1].

Uno de los objetivos principales de la mayoría de entidades educativas, es lograr que los estudiantes que ingresan a su alma mater, puedan titularse en la carrera que han seleccionado, y lo puedan lograr dentro del plazo establecido, evidenciando así que existe una alta calidad en el proceso de aprendizaje que esta oferta.

Entonces *¿Qué deberían hacer las instituciones educativas para lograr identificar los factores que puedan estar reprobación de sus estudiantes? ¿Cómo y de dónde se pueden obtener esta información? E*xisten muchas metodologías y herramientas para la explotación de datos, y procesos de minería de datos, la finalidad es obtener información y conocimiento, de los datos existentes, aprovechando así uno de los activos más valiosos de las organizaciones.

Con el fin de generar información que permita disminuir la reprobación continua en este tipo de materias, se propone en el presente trabajo, detectar los factores académicos, que están provocando bajo rendimiento académico y por lo tanto el alto índice de reprobación.

Para ello se utilizará la metodología *CRISP-DM*, especializada en proyectos de minería de datos, así mismo se aplicaran dos técnicas de minería de datos: *asociación y clasificación*. Se seleccionarán dos algoritmos por cada técnica de minería, con el fin de sugerir en base al problema planteado, y al conjunto de variables determinadas, los algoritmos que faciliten la detección de patrones, y por lo tanto de factores académicos, que pueden estar causando la reprobación de los estudiantes en las *materias filtradoras.*

La detección de patrones académicos relacionados a la alta reprobación de este tipo de materias, busca brindar información que ayude a los docentes, coordinadores de titulaciones,

[1] **Materias filtradoras:** son aquellas que poseen un alto índice de reprobación, esto indica que, muchos de los estudiantes de la carrera suelen reprobarla, al menos una vez durante el transcurso de su vida académica. Suelen también provocar que algunos estudiantes no puedan continuar la carrera, por tener más de tres reprobaciones de la misma materia.

y otras autoridades, a detectar tempranamente, a los estudiantes que necesitarían de un acompañamiento personal y apoyo adicional durante el transcurso de la materia.

Al minimizar la tasa de reprobación estudiantil, se pretende lograr una existencia mínima o total de materias filtradoras, así mismo disminuir el retiro o cambio del estudiante de la carrera o de la institución, causadas por la reprobación.

1.2 Planteamiento del problema

Con la finalidad de disminuir la deserción estudiantil causada por la reprobación académica, se desea detectar patrones académicos, así como los factores que estos contienen, y que al presentarse pueden aumentar los riesgos de reprobación de un estudiante, cuando este se matricula en una materia *filtradora*.

Estos patrones y factores deben facilitar la identificación temprana de los estudiantes con riesgo de reprobación, mediante la creación de alertas dirigidas a los docentes, coordinadores de titulación y autoridades de la institución.

Se utilizan datos académicos de los estudiantes, por varias razones, entre ellas la poca actualización de los datos socioeconómicos y familiares de los estudiantes, así como la baja cantidad de estudiantes que poseen esta información registrada en algún sistema académico.

1.3 Estructura de la memoria

Se inicia el proyecto de investigación con la definición de términos y conceptos que son fundamentales dentro del estudio a realizar, y que faciliten la comprensión de los capítulos posteriores.

Luego es indispensable indicar el contexto en el que se va a aplicar el estudio, así como la descripción de la metodología seleccionada para realizarlo, lo cual se desarrollará en los capítulos II y III, respectivamente.

En los capítulos IV, V se presenta el desarrollo y aplicación, a detalle, de cada paso de la comparativa a realizarse. En estos capítulos también se incluye los resultados del proceso de validación.

En capítulo VI se presenta el análisis y una breve discusión sobre los resultados obtenidos

Finalmente, en el capítulo VII, se presentan las conclusiones que resultan del desarrollo del proyecto investigativo, así como los trabajos a futuro que se puede realizar con este conocimiento.

2 CAPÍTULO II: CONTEXTO Y ESTADO DEL ARTE

En este capítulo se presentan conceptos que deben ser comprendidos antes del desarrollo del presente proyecto de investigación. Así mismo una breve descripción de lo que representa la educación a distancia, en vista de que es la modalidad de estudio en la que se enfoca esta investigación.

Así mismo, se incluye en este capítulo algunos trabajos similares, que han sido desarrollados en otras instituciones educativas, se indica también si existen proyectos similares dentro de la institución donde se aplica este estudio, y cuáles son las contribuciones de este proyecto.

2.1 Educación a distancia

La educación a distancia busca facilitar el acceso a toda persona a una educación de calidad sin importar su ubicación, ni la presencia física de un docente: por lo que depende en gran medida de los *materiales didácticos, tecnologías de información y sistemas académicos de interacción* que posea la institución, así como de la calidad de tiempo que el estudiante y el docente dediquen al estudio como a la enseñanza.

Según un artículo de la UNESCO desarrollado por (Maya Betancourt, 1993) menciona que Educación a Distancia es: *“una modalidad educativa que permite el acto educativo mediante diferentes métodos, técnicas, estrategias y medios, en una situación en que alumnos y profesores se encuentran separados físicamente y sólo se relacionan de manera presencial ocasionalmente, según sea la distancia, el número de alumnos, tipo de conocimientos que se imparte, etc.”.*

En este mismo artículo se resaltan a los principales componentes de la educación a distancia, mencionando a: *estudiante, docente, institución educativa, el programa (titulación o carrera), los materiales y la tecnología de educación a distancia*.

Por lo antes mencionado, el estudiante como el docente, deben comprender que este tipo de modalidad, conlleva a promover un alto nivel de comunicación, que aunque sea virtual se debe tratar de que sea continua, utilizando los materiales y tecnologías que la institución facilite para este objetivo.

2.2 Minería de datos en la educación

Para profundizar sobre el concepto de minería de datos en la educación, se debe tener claro el concepto de *minería de datos (MD) o data mining (DM),* ciencia que puede aplicarse en cualquier ámbito u entorno.

Se resalta a continuación tres definiciones muy interesantes sobre la *MD*:

- Los autores (Pérez César, 2007) la definen como: "*un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos.".*
- Otra aportación interesante es la de (Orallo, Quintana, & Ramírez, 2004), quienes mencionan en su libro que *MD*, *"es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos… La tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos…*". En otro párrafo de su libro estos autores indican que *"el uso de patrones descubiertos deberían ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio para la organización".*
- Así mismo el autor (Claudio Palma, 2009), la describe con una definición un poco más profunda, enunciándola como un "*conjunto de metodologías estadísticas y computacionales que, junto a un enfoque desde las ciencias de la conducta, permite el análisis de datos y la elaboración de modelos matemáticos descriptivos y predictivos de la conducta del consumidor".*

La *MD* es por tanto, una ciencia que extrae información y conocimiento nuevo y útil, desde los datos que una institución u organización posee, sean estas de cualquier ámbito, y los datos sean de cualquier tamaño y tipo.

Con el concepto de *MD* claro, se puede determinar que la minería de datos educativos (*EDM*) es la aplicación o uso de la *MD* dentro del entorno académico, con la finalidad de poseer información que permita mejorar los procesos de enseñanza-aprendizaje. Para su aplicación se suelen utilizar los datos generados en los sistemas académicos o entornos virtuales de aprendizaje.

La *EDM* a diferencia de la *MD*, se enfoca en brindar información a los tres actores principales del proceso de enseñanza-aprendizaje, que según (Monsalvea, Aponteb, & Hoyos, 2013) son: *estudiante, docente e institución (Ilustración 1).*

El propósito es ayudarles a mejorar o corregir la forma en que realizan la actividad asignada, para que obtengan los resultados que esperan. Se habla incluso de llegar a una personalización de los sistemas y de los procesos educativos para que, el aprendizaje sea más efectivo y con mejores resultados para sus actores.

Ilustración 1 Actores en la minería de datos educativos, tomado de (Monsalvea, Aponteb, & Hoyos, 2013)

Observe la Ilustración 2 en donde se muestra como interactúa el *EDM* con estos 3 actores, y como cada uno de ellos reciben beneficios de su implementación.

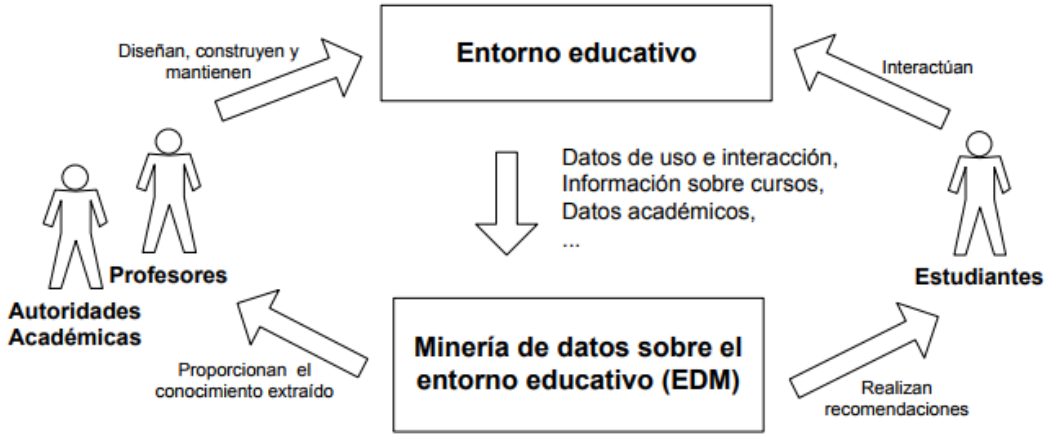


Ilustración 2 EDM y los principales actores del proceso de enseñanza-aprendizaje, tomada de (Romero, 2013)

Por ejemplo al:

- **Estudiante,** le favorece informándole sobre las actividades que puede realizar para mejorar su rendimiento académico, en base a los resultados de otros estudiantes que las han realizado y han tenido éxito. Así mismo indicarle que, si no realiza o cumple algunas actividades pueden llevarlo a no obtener un resultado positivo.
- **Docente**, le ayuda indicándole si las actividades realizadas han sido suficientes para conseguir un buen rendimiento de los estudiantes, o debe mejorar la estructura de la asignatura, mejorar planificación de las actividades o cambiar la metodología de enseñanza.

Detectar, en base a patrones, a aquellos estudiantes que pueden tener un mayor grado de dificultad, en la comprensión de los temas de la asignatura, con la finalidad de darle un mayor seguimiento e incluso una personalización del estudio.

- **Institución**, le apoya mostrándole que, los docentes con ciertas características no están dando buenos resultados dentro de la asignatura asignada. Así la institución buscará realizar capacitaciones para estos docentes, revisar los contenidos de las asignaturas, entre otros. Además puede ayudarle a detectar que cambios se deben realizar en la estructura de la malla curricular de alguna carrera.

2.3 ¿Qué son los patrones? ¿Cómo se detectan?

Los patrones son considerados como el resultado que se genera al aplicar un proceso de minería de datos. Se los define como las tendencias o variaciones de comportamiento encontradas en los datos utilizados para el análisis.

Para que un patrón sea fiable depende en gran medida de la calidad de los datos usados para su detección.

La evaluación de los patrones resultantes, suele realizarse en la fase de validación, que utiliza la minería de datos, y que se la menciona en algunas metodologías de aplicación de la *MD*.

Un patrón puede ser evaluado utilizando un análisis de sus características como: *¿Es útil?; ¿Es novedoso?, ¿Es entendible?; ¿Es efectivo para predecir?; ¿Las medidas objetivas de soporte y confianza son las adecuadas?,* entre otras.

Los patrones, según (Ballesteros A., 2013) pueden ser: *grafos, reglas de asociación, clasificaciones, una red neuronal, clustering, entre otros.*

2.3.1 Algoritmos para la detección de patrones

La *MD* y por lo tanto la *EDM* suele utilizarse tanto para la **predicción** como para la **descripción**. La primera, como su nombre lo indica, sirve para que, en base a datos generados por la institución, se pronostiquen valores desconocidos o futuros. Mientas que al aplicar la minería para la descripción se detectan o encuentran patrones que describen a los datos analizados y que sean de fácil interpretación para el ser humano.

Cada una de estas tareas posee técnicas y algoritmos de *MD* que se pueden aplicar para lograr su objetivo. En la Tabla 1 se resumen la correspondencia entre técnicas, algoritmos y tareas de *MD*.

Tabla 1 Tabla de correspondencia entre técnicas, algoritmos y tareas de minería de datos. Tomada de (Ordoñez Briceño, 2013)

Tareas / Técnicas y algoritmos	Predictivas		Descriptivas		
	Clasificación	Regresión	Agrupamiento	Asociación	Correlación
Redes neuronales	x	x	x		
Árboles de decisión (*ID.3, C4.5, C5.0*)	x				
Árboles de decisión (*CART*)	x	x			
Árboles de decisión y sistemas de reglas (*CN2*)	x			x	
Redes de *Kohonen*			x		
Modelización Estadística (Regresión lineal, regresión logarítmica)		x			x
Modelización estadística (Regresión logística)	x			x	
Métodos basados en casos y en vecindad (*K-means*)			x		
Reglas de asociación y dependencia (A priori)				x	
Métodos bayesianos (*Naive Bayes*)	x				
Métodos basados en casos y en vecindad (vecinos más próximos)	x	x	x		
Métodos basados en casos y en vecindad (*Two-step*, *COBWED*)			x		
Algoritmos genéticos y evolutivos	x	x	x	x	x
Máquinas de vectores de soporte	x	x	x		

En esta investigación se pretende aplicar dos técnicas de minería de datos, una de tipo **predictiva** y otra de tipo **descriptiva**, en la Tabla 1 se las resalta de color amarillo.

A continuación se presenta una breve descripción de cada una de estas técnicas y algoritmos a utilizarse.

2.3.1.1 Técnicas de asociación (descriptiva): reglas de asociación – algoritmo Apriori Y FP-Growth

Las reglas de asociación son utilizadas para descubrir objetos o elementos que se asocian, correlacionan o relacionan entre sí, y que se encuentran dentro de base de datos relacionales. Se las suele describir de forma general con la expresión *X -> Y,* en donde *X* e *Y* son conjuntos disjuntos de ítems. Ejemplo:

X {item1, item2} -> Y {item3}

X {computador, impresora} -> Y {toner}

Para evaluar y seleccionar las mejores reglas se utilizan medidas, entre las más conocidas son los umbrales mínimos de *soporte (minsup)* y *confianza (minconf)*. Estos valores son configurados por el usuario.

El algoritmo Apriori:

- Dentro de la técnica de reglas de asociación, este algoritmo se caracteriza por ser uno de los básicos y sencillos, con un buen nivel de eficiencia en sus resultados. Además los resultados que produce son fáciles de interpretar.
- Su principal característica es la denominada propiedad o principio ***Apriori***, que se define como *"Si un ítemset es frecuente, también lo son todos sus subconjuntos"*, lo que facilita encontrar de forma eficiente los ítemsets más frecuentes. *Ejemplo*: si *{item1, item2}* es un conjunto frecuente, entonces tanto *{ item1}* y *{ item2}* deberían ser frecuentes.
- Una desventaja de este algoritmo suele darse cuando se desea analizar una gran cantidad de transacciones, que posean muchos *ítems*, debido a la generación de los *itemsets frecuentes* que provocan un nivel computacional costoso, puesto que cada ítem es un candidato a ser ítem frecuente.

El algoritmo FP-Growth:

- Creado con la finalidad de solventar las limitaciones del algoritmo Apriori.
- Primero crea y almacena en un estructura de árboles los ítems frecuentes, para que luego de estos se puedan obtener las reglas de asociación. Logrando así una reducción de los costes computacionales.

2.3.1.2 Técnicas de clasificación (predictiva): árboles de decisión – algoritmo ID3 O C4.5 (J48) y CHAID

Las técnicas de clasificación tienen como objetivo catalogar un dato u objeto, detectando de entre varias clases, la clase específica a la que pertenece.

Estas técnicas utilizan la premisa de que todas las clases detectadas son disjuntas, es decir diferentes entre sí, por lo tanto un objeto sólo puede pertenecer a una clase, ya que no se pueden cumplir para varias clases las mismas condiciones.

El algoritmo ID3 O C4.5:

- Elige al atributo que mejor clasifica a los datos, utilizando como criterio de selección al atributo que posee el valor mayor de la *ganancia de información* (es el valor que indica que cantidad de información se clasifica con la división y sin la división de los datos)
- Se pueden utilizar atributos tipo cualitativo como cuantitativo

El algoritmo CHAID

- Al igual que el algoritmo ID3 divide e a la población en dos o más grupos distintos basados en las categorías de la variable dependiente, analizando todos los valores de la variable dependiente a través del Chi-cuadrado

2.4 Trabajos similares y contribución de la investigación

En la Tabla 2 se presentan algunos proyectos de investigación, que han utilizado la metodología *CRISP-DM*, y que están relacionados al ámbito educativo. Estos proyectos se han enfocado en analizar grupos de factores: *académicos, socioeconómicos, familiares, entre otros*; con la finalidad de predecir la deserción estudiantil.

Tabla 2 Resumen de trabajos similares. Elaboración propia.

	PROYECTO DE INVESTIGACIÓN 1 (Sotomonte Castro, Rodríguez Rodríguez, Montenegro Marín, Gaona García, & Castellanos, 2016)	PROYECTO DE INVESTIGACIÓN 2 (Timarán Pereira & Jiménez Toledo, CONGRESO IBEROAMERICANO DE CIENCIA, TECNOLOGÍA, INNOVACIÓN Y EDUCACIÓN, 2014)	PROYECTO DE INVESTIGACIÓN 3 (Hernández, 2015)	PROYECTO DE INVESTIGACIÓN 4 (Ordoñez Briceño, 2013)
Tema de investigación	Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos	Detección de Patrones de Deserción Estudiantil en Programas de Pregrado de Instituciones de Educación Superior con *CRISP-DM*	Modelo de minería de datos para identificación de patrones que influyen en el aprovechamiento académico	Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL
Técnica de minería de datos utilizada	*Árboles de decisión (algoritmo J48)*	*Algoritmos de clasificación (J48), de asociación (Apriori) y de agrupamiento (K-means)*	*Árboles de decisión (J48), Red neuronal, Clustering k-mediana*	*Clústers, Árboles de decisión (J48)*
Metodología para proyectos de minería utilizada	*CRISP – DM*	*CRISP – DM*	*CRISP – DM*	*CRISP – DM*
Herramienta para minería de datos	*WEKA*	*WEKA*	*Microsoft SQL Server Business Intelligence Development Studio y Server Analysis Services*	*WEKA*
Año de publicación de la investigación	2016	2014	2015	2013
Objetivo planteado que se resaltan	Causas de deserción de un estudiante	Detectar patrones de deserción estudiantil	Predecir si el estudiante deserta o se titula	Variable a predecir, *Deserto*: S o N
Conclusiones que se resaltan	Se indica que el número de asignaturas en las que se matricula el estudiante por semestre tiene influencia en si este decide desertar o no. Se indica que el factor socioeconómico promueve la deserción estudiantil.	Tener un promedio de notas bajo, el tener materias perdidas en los primeros semestres de la carrera y un puntaje promedio de ICFES[2] bajo.	La causa de deserción es el factor de inteligencia emocional.	El puntaje que obtienen los estudiantes en los trabajos y evaluaciones influye en la deserción estudiantil.
Diferencias con relación al presente proyecto de investigación	* Modalidad de estudio. * Variables o factores analizados. * Cantidad de datos. * Herramienta utilizada. * Objetivo: Determinar si el estudiante reprueba o no una materia filtradora.	* Variables o factores analizados. * Cantidad de datos. * Herramienta utilizada. * Objetivo: Determinar si el estudiante reprueba o no una materia filtradora.	* Variables o factores analizados. * Cantidad de datos. * Herramienta utilizada. * Objetivo: Determinar si el estudiante reprueba o no una materia filtradora.	* Variables o factores analizados. * Cantidad de datos. * Herramienta utilizada. * Objetivo: Determinar si el estudiante reprueba o no una materia filtradora.

[2] **ICFES**: Instituto Colombiano para el Fomento de la Educación Superior

De los trabajos mencionados en la Tabla 2, que son similares al presente proyecto, se resaltan los siguientes puntos:

- Analizan factores relacionados a temas socioeconómicos y familiares de los estudiantes, lo que se diferencia de esta investigación en la que se utilizarán solo factores académicos.
- Se enfocan en determinar si un estudiante se *retira o deserta*, pero no se enfoca en determinar si un estudiante reprueba o no una materia, como se lo plantea en el presente proyecto. Y en especial no se enfocan en materias que tienen altos índices de reprobación continua (materias filtradoras).

Las **contribuciones** que se desea brindar con el desarrollo de este proyecto investigativo son:

- Ser la primera investigación, usando datos oficiales y centralizados, y que sean parte del proyecto de minería institucional, que ha iniciado la universidad hace dos años, y que está relacionado a la reprobación académica de materias filtradoras, cuyos resultados puedan ser implementados y comprendidos con facilidad.
- Los resultados de esta investigación, se implementarán dentro del sistema de indicadores centralizado de la institución, en donde el docente y autoridades podrán revisar las alertas de riesgos de sus estudiantes y cursos ofertados, por semestre.
- Brindar información sobre qué técnicas y algoritmos se pueden utilizar en el ámbito educativo, al que pertenece la institución, y en problema de análisis similares al que se enfoca en este proyecto. Así mismo poder utilizar los resultados de esta investigación, como insumo o referencia para otras investigaciones, pudiéndose aumentar el número de campos utilizados o el tamaño de los datos.
- Detectar factores que faciliten determinar, de forma temprana, si un estudiante tiene riesgo de reprobar o no una materia, específicamente en estudiantes matriculados en materias filtradoras.

Al finalizar este capítulo se ha brindado una introducción a la minería de datos educativa, así también de los algoritmos que se utilizarán en el desarrollo del proyecto. Se ha considerado importante presentar trabajos similares y destacar la contribución de la presente investigación. Luego de esta inducción al tema investigativo, se continúa en el siguiente capítulo con el planteamiento de los objetivos que se desean cumplir, así como la selección de la metodología idónea que facilite su cumplimiento.

3 CAPÍTULO III: OBJETIVOS CONCRETOS Y METODOLOGÍA DE TRABAJO

3.1 Objetivo general

Realizar una comparativa de los resultados que se generen al utilizar diferentes técnicas de minería de datos, con la finalidad de determinar cuál sería la técnica más asertiva en sus resultados, y que se debería implementar en la Institución, considerando el tipo de problema planteado, el tipo de variables disponibles y la cantidad de datos existentes. Así mismo detectar los patrones que causan la reprobación en materias filtradoras.

3.2 Objetivos específicos

Objetivos específicos relacionados al desarrollo del proyecto de investigación:

- Seleccionar una metodología que permita guiar el proyecto de minería de datos
- Identificar las fuentes de datos y variables que se pueden obtener de ellas.
- Diseñar la estructura en donde se almacenarán y unificarán los datos recolectados
- Seleccionar algoritmos que permitan realizar detección de patrones
- Comparar y validar los patrones resultantes de cada algoritmo seleccionado, utilizando los datos reservados para la validación de resultados. Así mismo comparar los tiempos de ejecución, precisión y calidad de los resultados.

Objetivos específicos relacionados a la minería de datos a realizarse:

- Detectar las materias de carrera, denominadas *troncales* que posee un alto nivel de reprobación continua y que se las catalogará como materias *filtradoras*
- Descubrir factores académicos relacionados a la reprobación de un estudiante matriculado en una materia filtradora, considerando los datos académicos generados en los últimos 18 meses por los estudiantes, docentes y de las materias. La finalidad es que, con estos resultados se puedan crear alertas tempranas para los docentes y autoridades de la institución.

3.3 Metodología aplicada

Para el desarrollo de la investigación se ha considerado necesario, utilizar una metodología orientada a la minería de datos, que permita guiar y facilitar el desarrollo de este tipo de proyecto y que, proporcione una descripción clara sobre los procesos que se deben aplicar, sin olvidar el cumplimiento de los objetivos planteados y considerando que el ámbito es *educativo.*

3.3.1 Selección de la metodología de minería de datos

Existen varias metodologías enfocadas en proyectos de minería de datos. En la Ilustración 3 se presenta el resultado de encuentas realizadas en el 2007 y en el 2014 por (KDnuggets TM, 2014), en donde se observa que, entre las metodologías genéricas que continúan siendo las más utilizadas, se encuentran: **CRISP-DM** *(CRoss Industry Standard Process for Data Mining),* **KDD** *(*Knowledge Discovery in Databases*)* y **SEMMNA** *(Sample, Explore, Modify, Model, Assess)*, siendo la líder **CRISP-DM.**

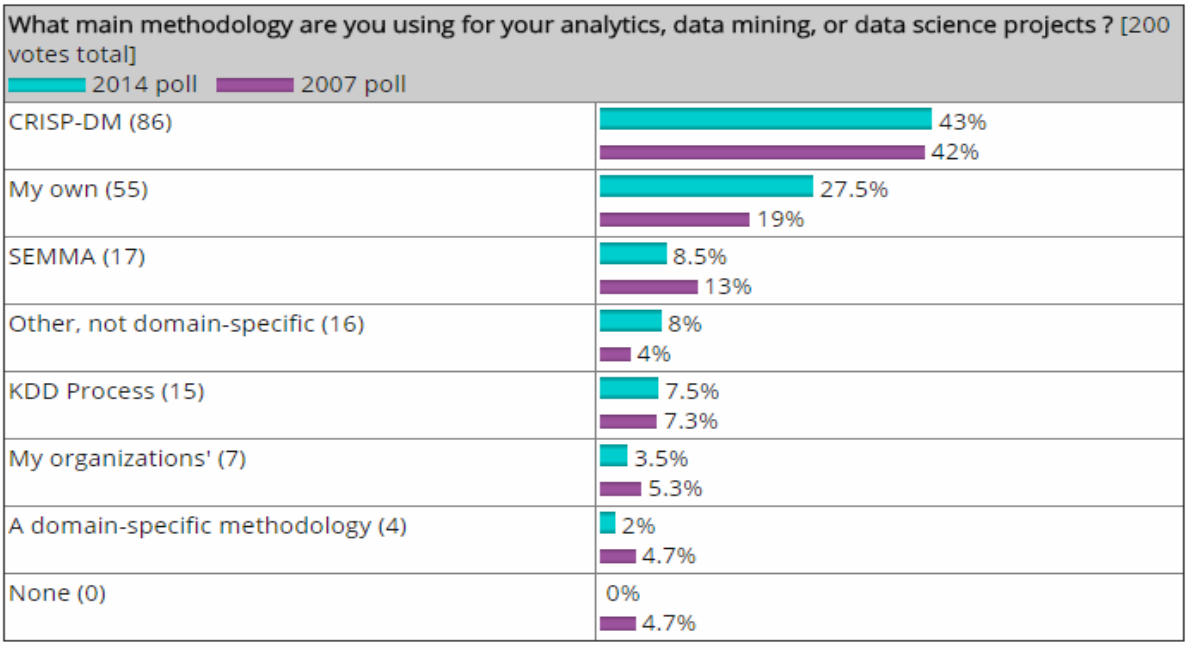


Ilustración 3 Resultados de las encuestas realizadas en el 2007 y 2014, sobre las metodologías utilizadas en proyectos de MD. Tomada de (KDnuggets TM, 2014)

En la Tabla 3 se presenta un resumen de las características utilizadas para la selección de la metodología de *MD*, que se utilizará en el presente proyecto.

Tabla 3 Características de evaluación para la selección de una metodología genérica de MD. Elaboración propia

	KDD	SEMMNA	CRISP-DM
	• **1996** • Primer modelo de MD	• **2000** • Creada por el SAS Institute • Enfocada especialmente en aspectos técnicos.	• **2000** • Creada por el grupo de empresas SPSS, NCR y Daimer Chrysler.
CARACTERÍSTICAS PARA LA SELECCIÓN DE LA METODOLOGÍA DE MD CON RESPECTO A LOS OBJETIVOS DEL PROYECTO DE INVESTIGACIÓN			
Fases que la componen.	**5 Fases** 1. Selección. 2. Pre-procesamiento. 3. Transformación. 4. Minería de datos. 5. Interpretación y evaluación.	**5 Fases** 1. Muestra (Sample). 2. Exploración (Explore). 3. Modificación o Manipulación (Modify). 4. Modelado (Model). 5. Asesoramiento o valorización (Assess).	**6 Fases** 1. Comprensión del negocio. 2. Comprensión de los datos. 3. Preparación de los datos. 4. Modelado. 5. Evaluación. 6. Despliegue.
1. Brinda un mayor detalle de las tareas y actividades que se deben ejecutar en cada etapa del proceso de MD.	**NO**	**NO**	**SI**, (Jaramillo & Paz Arias, 2015) comentan que *"Algunos modelos profundizan en mayor detalle sobre las areas y actividades a ejecutar en cada etapa del proceso de minería de datos (como CRISP-DM), mientras que otros proveen sólo una guía general del trabajo a realizar en cada fase (como el proceso KDD o SEMMA)"*
2. Comprensión del problema desde un punto de vista institucional.	**NO**	**NO,** (Ochoa, 2016) Menciona que *"SEMMA se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando. Inicia analizando los datos"*	**SI,** (Ochoa, 2016) Menciona que *"CRISP-DM empieza realizando un análisis del problema para su transformación en un problema técnico. Inicia analizando los objetivos del negocio"*
3. Análisis y comprensión del problema antes de comenzar el proceso de minería.	**SI**		
4. El proyecto recolecta datos de diversas fuentes por lo que pueden existir datos erróneos o faltantes. Por lo cual es necesaria una fase de limpieza y transformación.	**SI**	**SI**	**SI**
5. Es necesario poder aplicar cualquier tipo de modelo estadístico.	**SI**	**NO**, (Ochoa, 2016) indica que al utilizarse esta metodología se pueden utilizar únicamente los modelos incorporados en la herramienta *Enterprise Miner*	**SI**
6. Puede utilizarse cualquier herramienta que se utilice para el desarrollo del proyecto de MD.	**SI**	**NO**, Funciona con productos SAS.	**SI**
7. Debe permitir implementar la solución luego de obtener los resultados.	**SI**	**NO**	**SI**
8. Puede aplicarse proyectos de ámbito o contexto educativo.	**SI**	**SI**	**SI**, (Rojas Calvo, 2016) indica que esta metodología es *"Independiente tanto del sector de la industria y la tecnología que se utilice en el proyecto de MD"*
9. Existe proyectos de EDM que han utilizado la metodología.		**NO,** Se han buscado proyectos de EDM que hayan utilizado SEMMA como metodología de MD, pero no se han encontrado.	

			(Timarán Pereira, Hidalgo Troya, & Caicedo Zambrano, 2016) señalan que esta metodología es *"uno de los modelos más utilizados, principalmente, en los ambientes académicos e industrial y la guía de referencia más ampliamente utilizada en el desarrollo de este tipo de proyectos"*
10. Define su organización como un proceso iterativo e interactivo.	**SI,** Estas tres metodologías se componen y organizan en fases, las cuales se interrelacionan entren sí brindando información para la fase subsiguiente. Además se puede saltar una fase, así como regresar a una fase previa.		
11. Metodología abierta y gratuita.	**SI**	**NO**, atada a los productos de SAS	**SI**
Observaciones.	* Metodología base para otras metodologías.	* Metodología enfocada en el aspecto técnico.	* La más utilizada desde el 2007, encabeza el listado de las metodologías más utilizadas, según datos de (KDnuggets TM, 2014).
RESULTADO DE LA EVALUACIÓN:	De las 11 características analizadas, **KDD** cumple 9 de ellas, resultando un **81,8%** cumplimiento de las características con respecto a las necesidades del presente proyecto.	De las 11 características analizadas, **SEMMA** cumple 3 de ellas, resultando un **27.3%** cumplimiento de las características con respecto a las necesidades del presente proyecto.	De las 11 características analizadas, **CRISP-DM** cumple 11 de ellas, resultando un **100%** cumplimiento de las características con respecto a las necesidades del presente proyecto.

En base al análisis realizado y resumido en la Tabla 3, así como en los trabajos de investigación similares que se presentaron en la Tabla 2, se decide utilizar la metodología: **CRISP-DM**

3.3.1.1 Metodología CRISP-DM

La metodología CRISP-DM cuenta con esquema que va desde un modelo genérico a un modelo específico, lo que conlleva a ordenar el proceso de *MD,* desagregando cada fase por: *tareas generales, tareas específicas e instancias de proceso*, lo que facilitaría el trabajo y el desarrollo de los entregables.

Esta metodología propone 6 fases (Ilustración 4), dentro de un proceso interactivo e iterativo. El resultado de una fase previa sirve de entrada o insumo para la fase posterior. Se debe tener claro que esta metodología busca guiar, pero no exige que se cumplan todas sus fases, tareas o instancias, se sugiere realizarlas, si se desea tener mejores resultados.

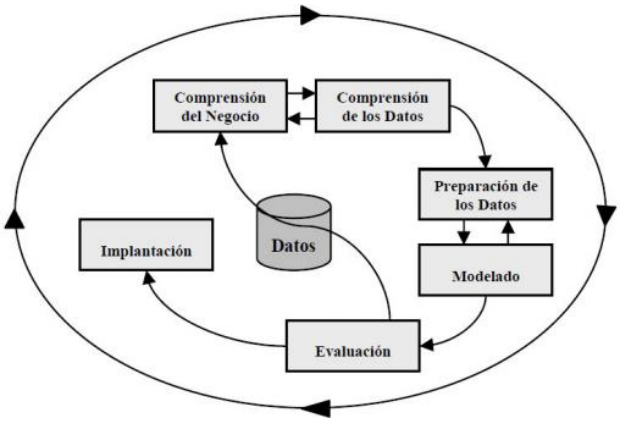


Ilustración 4 Fases del modelo propuesto por la metodología CRISP-DM. Tomada de (Gallardo Arancibia)

A continuación se presenta un breve detalle de lo que comprende cada una de las fases que involucra la metodología seleccionada.

- **Fase 1: Comprensión del negocio.-** se la considera como una de las fases más importantes, ya que tiene como principal finalidad conocer los objetivos del proyecto en beneficio de la institución, así como la definición del problema que se desea resolver aplicando MD, transformando el requerimiento del negocio en requerimientos técnicos.
- **Fase 2: Comprensión de los datos.-** su objetivo principal es la recolección de los datos, en esta fase se da una primera relación entre el problema y los datos. Junto con la fase de preparación de los datos y la fase de modelado se las considera como las fases que más tiempo y esfuerzo demandan.
- **Fase 3: Preparación de los datos.-** busca preparar los datos a través de las tareas generales que posee: *seleccionar, limpiar, estructurar, integrar y formatear los datos*. Estas tareas están enfocadas en lograr que los datos sean útiles y faciliten la aplicación de las técnicas de minería de datos que se seleccionen o apliquen en la fase de modelado, por lo tanto el resultado de esta fase afectará significativamente al proceso

y resultados de la fase de modelado. Es por ello que estas dos fases tienen un alto nivel de interacción entre sí.

- **Fase 4: Modelado.-** busca seleccionar la o las técnicas de modelado más idóneas con respecto al proyecto de MD. Para la selección de la técnica adecuada utiliza ciertos criterios de evaluación: *ser apropiada al problema, disponer de datos adecuados, cumplir los requisitos del problema, tiempo adecuado para obtener un modelo y conocimiento de la técnica.*
- **Fase 5: Evaluación.-** se enfatiza en la evaluación del modelo, en función del cumplimiento de los criterios de éxitos definidos para el problema. Para la evaluación del modelo se suelen utilizar matrices de confusión u otro tipo de herramientas. Si el modelo cumple satisfactoriamente su evaluación se puede realizar una difusión de los resultados.
- **Fase 6: Implantación.-** su principal objetivo es desarrollar o implementar el resultado de la MD en la institución, utilizando los resultados de la fase de evaluación. En esta fase se transforma, el conocimiento adquirido en acciones que pueden implementar en la organización. El alcance de este proyecto investigativo no incluye la realización de esta fase.

4 CAPÍTULO IV: PLANTEAMIENTO DE LA COMPARATIVA

En este capítulo se describe el proceso y resultados obtenidos, en cada una de las fases de la metodología seleccionada y descrita en el capítulo anterior. Para el desarrollo de este capítulo se hace uso de la guía de usuario de la metodología *CRISP-DM*, así como de algunos casos de éxitos que han utilizado esta metodología en el ámbito académico y que se han comentado anteriormente.

4.1 Fase 1: Comprensión del negocio

4.1.1 Situación actual y definición de los objetivos del negocio

La Institución en donde se desarrolla el presente proyecto es de tipo privada y pertenece al ámbito educativo. Ofrece tres modalidades de estudio: *presencial, semipresencial y a distancia.* Durante el desarrollo del proyecto no se publicarán datos de identificación de la misma, ni datos personales que puedan afectar la intimidad de los estudiantes, docentes y de la institución.

Esta institución en los últimos 3 semestres ha tenido un promedio aproximado de 37000 estudiantes por semestre. En la Ilustración 5 se presenta una gráfica del total de matriculados, agrupados por modalidad de estudios, de los últimos 3 periodos académicos. Se observa que la modalidad de estudios a Distancia, es aquella que tiene la mayor cantidad de estudiantes, con un promedio semestral de 31000 matriculados con respecto a la población total de matriculados.

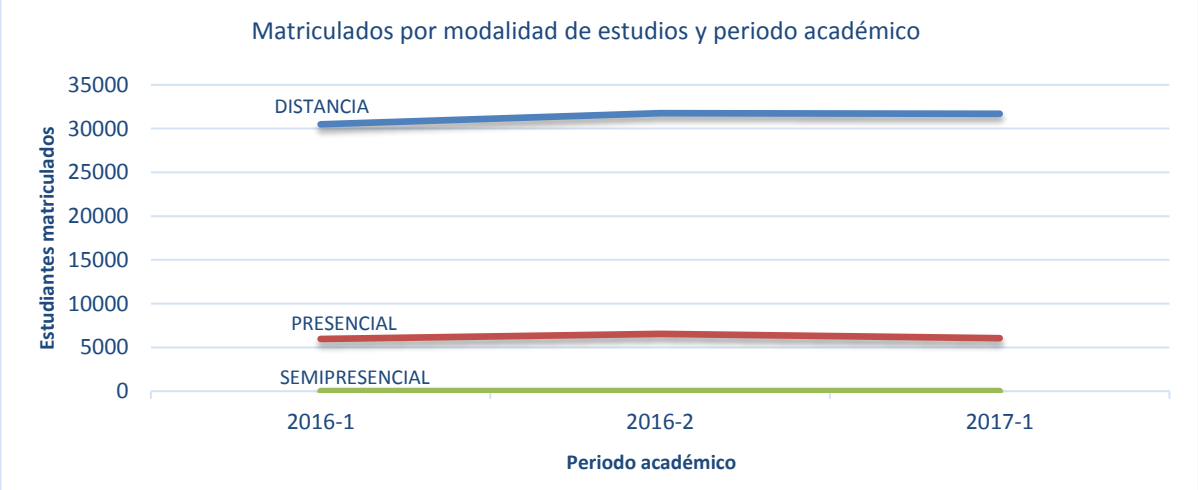


Ilustración 5 Matriculados por modalidad de estudios y periodo académico. Elaboración propia

La modalidad en la que se enfocará el presente estudio será *a Distancia,* así mismo los datos históricos a utilizar corresponden a los últimos 3 periodos académicos: 2016-1, 2016-2 y 2017-1. Se pretende detectar las causas que estén provocando la reprobación estudiantil en esta modalidad de estudios. En la Ilustración 6 se presenta la tasa de reprobación de los últimos 3 semestres de la modalidad *a Distancia*, en donde se observa que se mantiene en un valor lineal aproximado del 30% de reprobación por semestre.

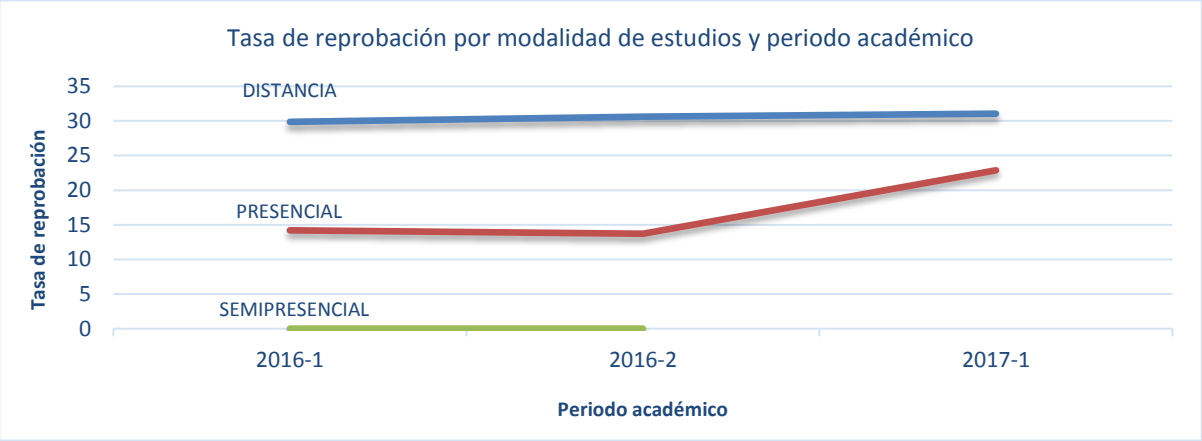


Ilustración 6 Tasa de reprobación por modalidad de estudios y periodo académico. Elaboración propia

Los efectos que causan la reprobación de materias y que preocupan a las autoridades de la institución son:

- Retrasos en el avance normal de la titulación, por lo tanto aumento del tiempo para graduarse
- Retiro de la titulación/institución
- Cambio de titulación/institución

Estos problemas afectan significativamente la tasa de graduación efectiva de la titulación y de la institución. Así mismo disminuye la continuidad académica semestral de los estudiantes.

En la Ilustración 7 se observa que, de los estudiantes que inician en el 2016-1 sólo el 50% continúan sus estudios en el semestre continuo (2016-2), y si se analiza a los que iniciaron en el 2016-1 y continúan luego de 2 periodos, es decir en el 2017-1, se obtiene que, menos del 50% siguen estudiando.

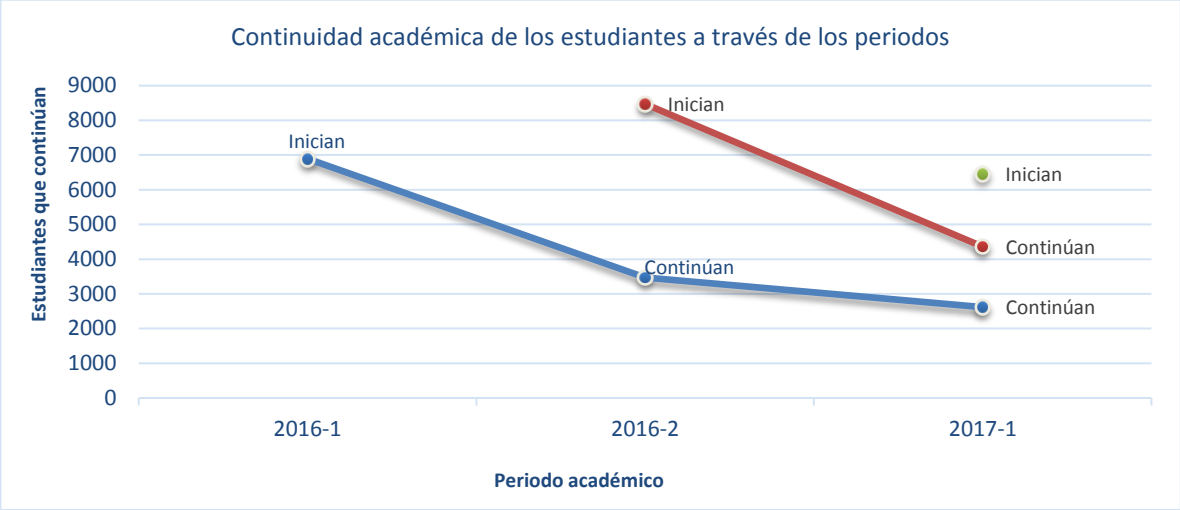


Ilustración 7 Continuidad académica de la modalidad a Distancia. Elaboración propia

Las áreas que se beneficiarán con el conocimiento que se pretende obtener de este proyecto de minería de datos son: *autoridades académicas de la institución, representantes de áreas académicas, coordinadores de titulaciones y docentes,* quienes resaltan que la falta de información impide realizar un seguimiento estudiantil a tiempo. De estas tres áreas a la que se pretende brindar más información es a los *coordinadores de titulaciones y docentes,* con la finalidad de que puedan hacer un acompañamiento y seguimiento semestral a sus estudiantes, en especial a los que tengan riesgo de reprobación. Estas áreas serán el grupo objetivo que utilizarán los resultados del presente proyecto.

Este estudio desea ofrecer información sobre los factores que están causando reprobación en materias filtradoras, utilizando datos académicos del docente, estudiante y de la materia.

Hace pocos años dentro de la institución se ha constituido un área enfocada en la gestión de datos, la cual ha iniciado varios proyectos de minería de datos institucionales. Esta área trabaja con los directivos de la institución, y ha generado algunos resultados de análisis de indicadores para toma de decisiones, pero a la fecha aún no se han desarrollado ningún proyecto relacionado al análisis de la reprobación estudiantil, siendo esta investigación la primicia en este tema.

La principal motivación para el desarrollo de este proyecto de minería de datos, es disminuir la tasa de reprobación y por lo tanto la tasa de retiro estudiantil, lo permitirá el aumentado de la continuidad académica y por lo tanto de la tasa de graduación institucional.

Se pretenden además, enfocarse en las materias filtradoras que pertenecen al grupo de créditos *troncales,* puesto que estas materias son la columna vertebral de la carrera, y su

reprobación, causa que los estudiantes no puedan tomar otras materias porque suelen ser dependientes entre sí.

A continuación se presenta un listado detallado de: *recursos, restricciones, riesgos y beneficios*, que se deben considerar para el desarrollo del proyecto de minería de datos.

4.1.1.1 Inventario de recursos para el proyecto

Los recursos que se van a utilizar para la ejecución del proyecto se resumen en la Tabla 4, se indica el *nombre, una descripción y el tipo de recurso (persona, fuente de datos, tecnológico-software, tecnológico-hardware).*

Tabla 4 Inventario de recursos para el proyecto. Elaboración propia

Recurso	Descripción y observaciones	Tipo
Estudiante investigador.	En virtud de que el presente proyecto es parte del desarrollo de una tesis, el recurso humano lo representa el estudiante, junto con su director (experto en minería), así como la colaboración de un representante académico de la modalidad seleccionada, para la presentación de resultados.	**Persona**
Director del proyecto de tesis.		
Representante académico.		
Fuente 1: **Base de datos** del Sistema académico de la institución (Oracle 11g).	De esta fuente se obtendrán datos referentes a las características personales del estudiante y su expediente académico. Así mismo el historial académico del docente y de la asignatura o materia.	**Fuente de datos**
Fuente 2: **Base de datos** del Entorno virtual de aprendizaje de la Institución (MySql).	De esta fuente se obtendrán datos referentes a la actividad académica del estudiante y del docente en este entorno.	**Fuente de datos**
Fuente 3: **Base de datos** del distributivo académico y de docencia (Oracle 11g).	De esta fuente se obtendrán las características del docente y de la asignatura.	**Fuente de datos**
Fuente 4: **Base de datos**: repositorio de datos unificados y centralizados institucionales (Oracle 11g).	Esta fuente será el repositorio centralizado de datos oficiales de la institución.	**Fuente de datos**
Herramientas que permitan extraer de forma masiva los datos de los sistemas fuentes, para colocarlos en un solo repositorio.	Actualmente la institución no cuenta con una herramienta que permita realizar un proceso de ETL (Extracción, Transformación y Carga) de datos desde varias fuentes. Para la extracción desde las fuentes se utilizará los procesos ya desarrollados por la institución, pero para la unificación de las tablas de extracción *(tablas de stage)* se analizará el uso de una herramienta que facilite este proceso.	**Tecnológico-Software**
Herramientas para realizar limpieza y transformación de los datos.	Se analizará el uso de una herramienta que facilite este proceso. Actualmente la institución no cuenta con una herramienta que permita realizar estas tareas.	**Tecnológico-Software**
Herramientas para aplicar algoritmos de minería de datos, y que permitan segmentaciones, clasificaciones, asociaciones, predicciones, etc., así como validar los resultados.		**Tecnológico-Software**

Herramientas para ofimática y análisis previos de los datos.	Ejemplo: Excel, Word.	**Tecnológico-Software**
Servidor para ubicar el repositorio central de datos.	Se utiliza el servidor de base de datos autorizado para colocar los datos en una fuente interna autorizada y centralizada.	**Tecnológico-Hardware**
Computador	En este equipo de instalará la herramienta para realizar procesos de Minería de datos. *Requerimientos mínimos:* Memoria RAM: 8G Capacidad de almacenamiento: 500 GB Procesador: i7 Sistema operativo: Windows 7	**Tecnológico-Hardware**

4.1.1.2 Restricciones

Las restricciones que se han detectado se presentan en la Tabla 5, clasificadas en 5 grupos: *seguridad, legales o éticas, de recursos y tecnológicos, económicas, y otras.*

Tabla 5 Listado de restricciones del proyecto. Elaboración propia

De seguridad	Legales o éticas
1. Acceso a los datos fuentes.- Para la recolección de los datos se tendrá acceso de lectura a las bases de datos transaccionales y a la base de datos centralizada. **2. Niveles de seguridad de la publicación de resultados.-** Se darán acceso a los resultados por el nivel de información a la que deben tener acceso. Por ejemplo cada coordinador sólo debe tener los resultados de la titulación a la que él dirige. Los directores de área sólo tendrán acceso a los resultados de las titulaciones de su área. Sólo las autoridades y representantes académicos tendrán acceso a todos los resultados, los docentes tendrán acceso solo al análisis de sus estudiantes.	**1. Privacidad de los datos.-** Se debe tener cuidado con los datos personales de los docentes y estudiantes. Se puede utilizar encriptación de datos para mayor seguridad de los mismos. No se puede hacer públicos los resultados ni listados de los estudiantes. **2. Uso de los datos.-** sólo se pueden utilizar los datos para los fines definidos y autorizados por la institución.
De recursos y tecnológicos	**Económicas**
	No aplica
Otras	
1. Tiempo para el desarrollo del proyecto.- El tiempo es corto porque el proyecto es parte de una tesis de postgrado.	

4.1.1.3 Riesgos y contingencias

Los riesgos o dificultades que pueden presentarse durante el desarrollo del proyecto se listan en la Tabla 6, así mismo se plantea una solución que puede aplicarse como contingencia en caso que se produzca el riesgo.

Tabla 6 Riesgos y contingencias. Elaboración propia

Riesgo	Descripción	Nivel de criticidad	Cuando se puede dar	Contingencia
Accesos a los datos.	Demora en brindar el acceso a los datos.	Alto	Inicio del proyecto.	Revisión conjunta de los accesos con los administradores.
	Error en claves de acceso.			

4.1.1.4 Costos y beneficios

- Los datos a utilizarse son propiedad de la institución, por lo cual no se ha generado ningún costo.
- A nivel económico directamente este proyecto no generará ganancias para la institución, pero a nivel indirecto pretende ayudar a disminuir el número de estudiantes desertores por reprobación, y brindar un mejor servicio (acompañamiento) al estudiante durante su preparación académica.

4.1.2 Glosario de términos a utilizar en el proyecto

Términos **académicos**:

- ***Materias filtradoras*.-** son aquellas que poseen un alto índice de reprobación, muchos de los estudiantes de la carrera suelen reprobarla al menos una vez durante el transcurso de su vida académica. Suelen también provocar que algunos estudiantes no puedan continuar la carrera por poseer más de tres reprobaciones de esta misma materia.
- ***Sistema de estudios en créditos educativos*.-** hace referencia a un tipo de sistema educativo basado en competencias a través de créditos educativos o académicos. Define 6 tipos de asignaturas: *troncales, complementarias, genéricas, gestión productiva o prácticum, libre configuración y formación básica.*
- ***Créditos educativos o académicos*.-** es la unidad de medición de las asignaturas dentro del sistema de estudios por créditos educativos.
- ***Asignaturas troncales*.-** son un tipo de asignatura del sistema de estudios de créditos académicos, la constituyen las asignaturas cuyo contenido es propio y específico de cada titulación para la formación del estudiante en su rama profesional.
- ***Curso*.-** Es la unión de la asignatura o materia, con el paralelo y docente ofertado en un periodo académico.

Términos **técnicos**:

- ***Tabla de stage** (tabla de extracción).-* es un espacio orientado a almacenar los datos provenientes de los sistemas operacionales o de otras fuentes, con una vida temporal o no, que será el punto de partida de los procesos de depuración, transformación y carga en el DW, serán las fuentes de datos de las tablas de dimensiones.
- ***Tabla de dimensiones*.-** almacenan datos para describir procesos de negocio, contienen datos de entidades únicas del negocio.

- ***Tabla de hechos***.- contienen datos y valores que se utilizan para el análisis de alguna área del negocio, están relacionados con las dimensiones.
- ***Bases de datos relacionales***.- es un repositorio en donde sus datos se almacenan en tablas que poseen relaciones entre ellas. Base de datos relacionales más conocidas: *ORACLE, MYSQL, POSTGRESS.*
- ***PLSQL.***- lenguaje de programación utilizando en las bases de datos *ORACLE*. Sirven para manipular los datos almacenados en los objetos que existen en la base de datos.
- ***ETL***.- proceso para la obtención y explotación de datos. Incluye tres subprocesos: *Extracción, Transformación y Limpieza* de los datos.
- ***Merge tablas***.- es un sentencia que permite crear, actualizar o eliminar registros de una tabla. Se debe definir una condición para que esta sepa que operación debe realizar.

4.2 Fase 2: Comprensión de los datos

La segunda fase que propone la metodología de CRISP-DM es la definición, recolección y comprensión de los datos. En esta fase se realiza una primera inducción y contacto con los datos a utilizar.

4.2.1 Recolección de datos iniciales

Los datos a recolectarse de los estudiantes, docentes y de asignaturas deben estar relacionadas a las *materias filtradoras* de forma semestral. Para determinar una materia filtradora, esta debe cumplir:

- Más de 5 estudiantes matriculados.
- Más del 30% de tasa de reprobación en los últimos 3 semestres.
- Sea del grupo de créditos *troncales.*
- Estudiantes de modalidad a distancia de grado (tercer nivel).

Para la selección de los atributos y de los datos que se utilizan en el proyecto se han aplicado algunos criterios de evaluación, los mismos que se listan en la Tabla 7.

Tabla 7 Criterios de evaluación para la selección de atributos y datos a utilizar en el proyecto. Elaboración propia

Criterio de evaluación	Enfocado en la evaluación de
Requerido.- El atributo es necesario para cumplir con los objetivos del proyecto de minería de datos.	Selección de atributo
Tiempo de historial.- Es necesario que se tengan datos históricos del atributo, como mínimo de los últimos 3 periodos académicos.	

Número máximo de atributos.- cuantos atributos se pueden utilizar en relación con la técnica de MD que se vaya a utilizar.	
Se conoce la definición de su contenido, se comprende con exactitud a que hacen referencia los datos que contiene	
Sea mediable de forma cuantitativa o cualitativa	
Que contenga datos variables	
Los datos provengan de fuentes oficiales e institucionales	Selección de datos
Se tenga acceso autorizado a los datos para el análisis	

4.2.1.1 Informe de la recolección de datos iniciales

Se realiza la recolección de datos relacionados a los tres principales actores del aprendizaje académico: *docente, estudiante y asignatura.* En la Ilustración 8 se indica el grupo de datos a recolectar de cada uno de estos actores, así como la fuente de datos de donde serán extraídos.

Obsérvese también que todos los datos extraídos, se colocarán en una fuente centralizada de datos, denominada *fuente 4*, con la finalidad de ganar independencia de otros sistemas, procesos, restricciones, etc.

En esta fuente centralizada se realizarán las próximas fases del proceso de *MD*, ya que muchos autores sugieren que no se realicen procesos de análisis, en las mismas bases de datos en donde funcionan los sistemas transaccionales.

Ilustración 8 Grupo de datos a recolectar de cada actor que influye en el rendimiento académico y fuente de datos de donde se los obtendrá. Elaboración propia.

En la Ilustración 9 se identifican las fuentes de datos a utilizarse para la extracción de los datos que se generan en los sistemas académicos de la institución.

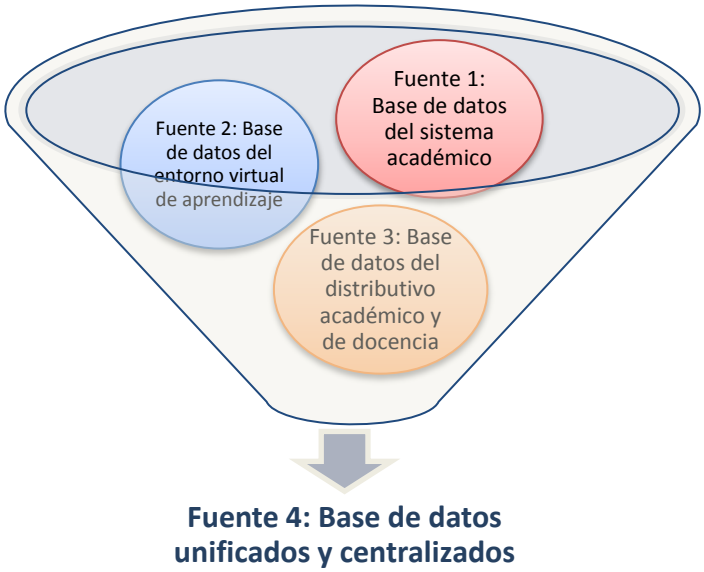


Ilustración 9 Fuentes de datos para la obtención de datos y fuente resultante. Elaboración propia

Los datos del repositorio identificado como *fuente 1,* provienen del sistema académico institucional, en donde se realizan procesos de *matrícula, ingreso de notas, postulaciones, retiros, anulaciones, gestión de expediente académico, registro de reprobaciones, entre otros.* Estos procesos están relacionados con el estudiante, docente y la asignatura.

Los datos de *fuente 2,* provienen del entorno virtual de aprendizaje o campus virtual, cuya meta es impulsar y facilitar la interacción entre estudiante y el docente dentro de la asignatura que se está cursando. De esta fuente se obtendrán las actividades propuestas por el docente, así como las actividades realizadas por el estudiante.

Finalmente de la fuente 3, se obtendrán los datos relacionados a la distribución o carga académica que ha sido asignada al docente, así mismo se obtendrán datos del docente en relación a su formación y experiencia académica del docente.

Como se ha mencionado en anteriores ocasiones, todos los datos recolectados se almacenarán en una fuente unificada y centralizada, en este caso se la ha denominado *fuente 4.* Esta fuente actualmente ya existe en la institución y facilitará los procesos de las fases de comprensión y preparación de los datos.

En la Tabla 8 se menciona, en forma resumida, información importante referente a las fuentes de datos a utilizar.

Tabla 8 Información sobre las fuentes de datos. Elaboración propia.

Fuente de datos	Tipo de fuente de datos	Método de captura	Grupo de datos a extraer
Fuente 1: **Base de datos** del sistema académico de la institución (*Oracle 11g*).	Interno	Registro de datos en los sistemas académicos de la institución, ingresados por: *secretarias de titulación, docentes y estudiantes*.	* Características del estudiante. * Expediente académico del estudiante. * Expediente académico de la asignatura.
Fuente 2: **Base de datos** del entorno virtual de aprendizaje de la institución (*MySql*).	Interno	Registro de datos en los sistemas académicos de la institución, ingresados por: *docentes y estudiantes.*	* Actividades académicas realizadas por el estudiante. * Actividades académicas propuestas y realizadas por el docente.
Fuente 3: **Base de datos** del distributivo académico y de docencia (*Oracle 11g*).	Interno	Registro de datos en los sistemas académicos de la institución, ingresados por: *secretarias de titulación.*	* Características del docente. * Expediente académico del docente.
Fuente 4: **Base de datos**: repositorio de datos unificados y centralizados institucionales (*Oracle 11g*).	Interno	Esta fuente será el repositorio centralizado de datos oficiales de la institución, los datos provienen de las fuentes 1,2 y 3.	

Antes de continuar con la descripción de los campos y los datos que van a ser extraídos, se indica el proceso realizado para extraer, cargar y transformar (uso de procesos de *ETLs*) los datos existentes en las fuentes orígenes 1, 2 y 3 y colocarlos en la fuente 4 (destino - centralización de datos), obsérvese la Ilustración 10.

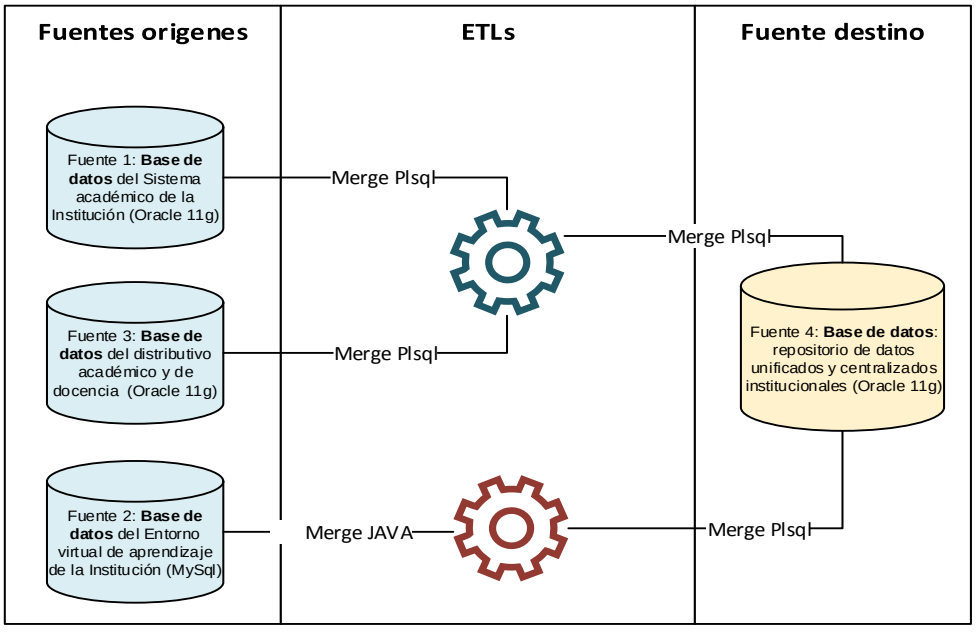


Ilustración 10 Proceso de extracción y carga de datos desde las fuentes orígenes a la fuente destino. Elaboración propia.

Los procesos de *ETLs* son programados para la fuente 1 y 3, en *PLSQL* – Oracle, en donde se crean paquetes con procesos *merge* por cada grupo de información. Si cumple la condición del *merge* se actualizan los datos (ciertos campos definidos), caso contrario se inserta un nuevo registro. También se incluye sentencias de eliminación. *Ejemplos* de paquetes de *ETLs*: *ETL estudiantes, ETL de expedientes estudiantiles, ETL para matriculas, entre otros*.

Para la fuente 2, los procesos de *ETLs* se han programado en *JAVA*, así mismo se utilizan procesos de *merge. Ejemplos*: *ETL estudiantes campus virtual, ETL de actividades, ETL docentes campus virtual, entre otros.*

Estos procesos *ETL* insertan los datos en tablas *stage*, utilizando procesos *PLSQL* para enlazarse a tablas de dimensiones que correspondan. *Ejemplo*: Enlazar la tabla stage *Actividad_Foro_Estudiante_Campus_Virtual* con la tabla dimensión *estudiante*.

Luego en paquetes diferentes se van construyendo tablas de hechos dependiendo de la parte del "*negocio*" que se desea hacer seguimiento o realizar análisis.

4.2.2 Describir los datos

En este punto se presenta una visión inicial de los datos que posee la institución, y que se pueden extraer de las fuentes de datos antes mencionadas.

Los campos catalogados como *básicos,* son aquellos que se obtienen directamente de la fuente de datos. Estos campos se utilizarán para el análisis posterior o para generar nuevos campos que serán utilizados en el análisis.

En la Ilustración 11 se presenta el esquema de base de datos de la fuente 4, que es el repositorio centralizado de datos institucionales, en donde se integran las tres fuentes de datos a utilizar.

Las tablas de *stage* a utilizarse se describe a continuación, en la Ilustración 11 se las representa de color amarillo.

- ***Oferta*.-** contiene datos semestrales, referentes a la oferta académica de la asignatura. Se obtienen campos como: *ciclo de la asignatura, área de la asignatura,* etc.
- ***Matrícula*.-** contiene datos semestrales, referentes a la matrícula académica de cada estudiante.
- ***Expediente académico del estudiante por periodo académico*.-** contiene datos académicos del estudiante por cada titulación, periodo académico y asignatura en donde

se ha matriculado. De esta tabla se extraerán datos como: *centro universitario, nota obtenida, estado obtenido,* entre otros.

- ***Actividades del estudiante en el campus virtual*.-** contiene datos de las actividades académicas realizadas por el estudiante en el campus virtual, esto por cada curso-asignatura en la que se ha matriculado en el semestre o periodo académico. Posee datos como: *total de foros en los que participó, total de cuestionarios que resolvió, total de chats a los que asistió,* entre otros.
- ***Expediente académico del docente*.-** contiene datos académicos del docente relacionados con la institución. De esta tabla se extraerán datos como: *promedio de notas del docente, tasa de reprobación del docente,* entre otros.
- ***Expediente académico del docente por periodo*.-** contiene datos académicos del docente por periodo académico. De esta tabla se extraerán datos como: *jornada del docente, área del docente, es bimodal en el periodo académico,* entre otros.
- ***Actividades del docente en el campus virtual.-*** contiene datos de las actividades académicas propuestas y realizadas por el docente en el campus virtual, esto por cada curso-asignatura que ha sido asignada al docente en el periodo académico. Posee datos como: *total de foros, total de cuestionarios, total de chats,* entre otros.

El esquema que se utiliza es tipo *copo de nieve*. Se utilizan las siguientes tablas de *dimensiones*, en la Ilustración 11 se las representa de color azul, cada una de estas tablas poseen una clave primaria, estas tablas son:

- ***Asignatura o materia***.- contiene datos básicos de la asignatura como: *nombre, código, número de unidades de aprobación, entre otros.*
- ***Áreas académicas***.- contiene datos básicos del área académica como: *nombre, código, entre otros.*
- ***Titulación***.- contiene datos básicos de cada titulación como*: nombre, código, sistema de estudios, área académica, duración, modalidad, nivel de estudios, entre otros.*
- ***Grupo de créditos*.**- contiene datos básicos del tipo de asignatura como: *nombre, código, entre otros.*
- ***Estudiante*.-** contiene los datos actuales y básicos del estudiante, como por ejemplo: *nombre, fecha de nacimiento, sexo, discapacidad, etc.*
- ***Titulación***.- contiene datos básicos de cada titulación como: *nombre, código, sistema de estudios, área académica, duración, modalidad, nivel de estudios, entre otros.*

- ***Periodo académico*.**- contiene los datos básicos de los semestres académicos en donde han existido ofertas/actividades académicas. Posee datos como: *nombre, código, año del periodo, fechas de inicio y fin, entre otros.*
- ***Docente*.**- contiene datos básicos del docente como: *nombre, código, identificación, fecha de nacimiento, sexo, nacionalidad, entre otros.*
- ***Jornada académica*.**- contiene los datos básicos de los tipos de jornadas académicas como: *nombre, código, entre otros.*
- ***Centro universitario*.**- contiene los datos básicos de los centros institucionales ubicados en todo el mundo, posee campos como: *nombre, código, ubicación, región, entre otros.*

En la Ilustración 11 se representa de color verde a las tablas de *hechos* que se van a utilizar:

- ***Ficha de la asignatura***.- contiene atributos relacionados a la asignatura-titulación como: *promedio de notas, tasa de reprobación, entre otros.*
- ***Ficha del estudiante*.-** contiene atributos relacionados al estudiante-titulación como: *promedio de notas, tasa de reprobación, promedio de materias en las que se matricula el estudiante, entre otros.*
- ***Ficha del docente.-*** contiene atributos relacionados al docente como: *promedio de notas, tasa de reprobación, número de periodos con docente en institución, entre otros.*

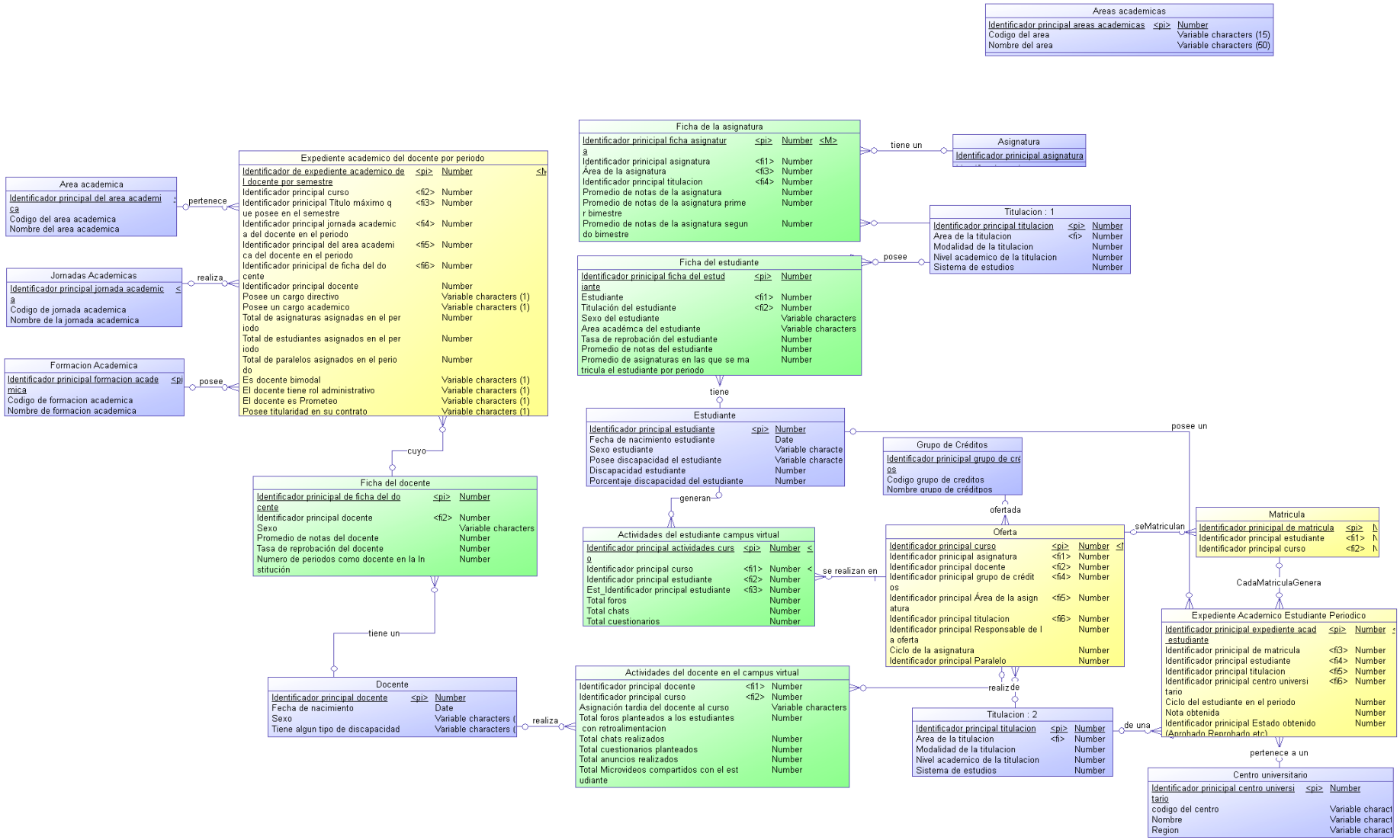


Ilustración 11 Esquema de la fuente 4 – repositorio centralizado de los datos, modelo de integración. Elaboración propia.

Los campos que se extraerán de las tablas (*stage, dimensiones, hechos*) se listan en la Tabla 9, en donde se ha colocado: *el nombre de la tabla de donde se extraerá el dato, una descripción del campo, tipo de dato, y el objetivo para el cual es extraído el dato.*

Tabla 9 Campos de donde se realizará la extracción de los datos – grupo de datos referentes a la asignatura. Elaboración propia.

Tabla de donde se extraerá los datos	Campo de donde se extraerá los datos	Descripción del campo	Tipo de campo[3]	Objetivo
Ficha de la asignatura	ASIG_PROMEDIO_NOTAS_SEM	Promedio de notas de la asignatura	NU	Campo básico para análisis
	ASIG_PROMEDIO_NOTAS_BIM1	Promedio de notas de la asignatura primer bimestre		
	ASIG_PROMEDIO_NOTAS_BIM2	Promedio de notas de la asignatura segundo bimestre		
	ASIG TASA_REPROBACION	Tasa de reprobación de la asignatura		
Oferta	ASIG_COD_TIS_RESPONSABLE _OFERT	Unidad que oferta la asignatura para los estudiantes	NO	Campo básico para generar un campo calculado
	ASIG_AREA_ACADEMICA	Código del área de la asignatura a la que esta pertenece		
	ASIG_CICLO	Ciclo de la asignatura, relacionada a la titulación del estudiante	NU	Campo básico para análisis y para calcular un nuevo campo
Estudiante	EST_FECHA_NACIMIENTO	Fecha de nacimiento	FE	Campo básico para generar un campo calculado
	EST _SEXO	Sexo	NO	Campo básico para análisis
	EST_POSEE_DISCAPACIDAD	Posee discapacidad		Campo básico para análisis
Ficha del estudiante	EST_AREA_ACADEMICA	Código del área académica a la que pertenece la titulación del estudiante		Campo básico para generar un campo calculado
	EST_COD_TIS_ESTUDIANTE	Código de la titulación del estudiante		Campo básico para generar un campo calculado
	EST _TASA_REPROBACION	Tasa de reprobación del estudiante	NU	Campo básico para análisis
	EST_PROMEDIO_NOTAS_SEM	Promedio de notas		
	EST_PROMEDIO_ASI_MAT_ES	Promedio de asignaturas en las que se matricula el estudiante por periodo		
Expediente Académico Estudiante Periódico	EST_CICLO	Ciclo del estudiante en el periodo		Campo básico para análisis y para calcular un nuevo campo
	EST_NOTA_SEM	Nota obtenida		Campo básico para análisis
	EST_ESTADO	Estado obtenido (Aprobado, Reprobado)	NO	
	EST_CENTRO_ACADEMICO	Código del centro universitario		
	EST_REGION_UBICACION	Región en donde se encuentra ubicado el estudiante		
	EST_TOT_ASIG_MATRICULA	Número de asignaturas en las que se matricula en el periodo	NU	
	EST_TOT_ASIG_REPROB_PER_ ANT	Perdió o no una asignatura en el periodo anterior	NO	
	EST_PROMEDIO_NOTAS_BIM1	Promedio de notas del primer bimestre	NU	
Actividades del estudiante	EST_FOROS	Número de foros en los que participó el estudiante		
	EST_CHATS	Número de chats en los que participó el estudiante		

[3] NO: Nominal; NU: Numérico; FE: Fecha

campus virtual	EST_CUESTIONARIOS	Número de cuestionarios en los que participó el estudiante		
Docente	DOC_FECHA_NACIMIENTO	Fecha de nacimiento	FE	Campo básico para generar un campo calculado
	DOC_SEXO	Sexo	NO	Campo básico para análisis
Ficha del docente	DOC_PROMEDIO_NOTAS_SEM	Promedio de notas del docente	NU	
	DOC_TASA_REPROBACION	Tasa de reprobación del docente		
	DOC_PERIODOS_DOCENCIA	Numero de periodos como docente en la Institución		
Expediente académico del docente por periodo	DOC_AREA	Código del área del docente a la que pertenece	NO	Campo básico para generar un campo calculado
	DOC_EDAD	Edad del docente	NU	Campo básico para análisis
	DOC_JORNADA_ACADEMICA	Código de la jornada del docente en el periodo	NO	
	DOC_TITULO_ACAD_MAX	Código del título máximo que posee		
	DOC_POSEE_CARGO_DIR	Posee un cargo directivo		
	DOC_POSEE_CARGO_ACA	Posee un cargo académico		
	DOC_TOTAL_ASIG_SEM	Total de asignaturas asignadas en el periodo	NU	
	DOC_TOTAL_EST_SEM	Total de estudiantes asignados en el periodo		
	DOC_TOTAL_PAR_SEM	Total de paralelos asignados en el periodo		
	DOC_BIMODAL	Es docente bimodal	NO	
	DOC_ADMINISTRATIVO	El docente tiene rol administrativo		
	DOC_PROMETEO	El docente es Prometeo. Docente PHD temporal		
Actividades del docente en el campus virtual	DOC_ASIGNACION_TARDIA	Asignación tardía del docente al curso. Detectar si el docente fue asignado al curso luego de iniciadas las clases		
	DOC_TOTAL_FOROS	Total foros planteados a los estudiantes con retroalimentación	NU	
	DOC_TOTAL_CHATS_ACAD	Total chats realizados		
	DOC_TOTAL_CUESTIONARIOS	Total cuestionarios planteados		
	DOC_TOTAL_ANUNCIOS	Total anuncios realizados		
	DOC_MICROVIDEOS	Total Microvideos compartidos con el estudiante		

4.3 Fase 3: Preparación de los datos

Luego de obtener los datos y tener un conocimiento previo de su contenido, es necesario realizar una preparación de los datos antes de aplicar la o las técnicas de minería de datos, lo que beneficiará los resultados del modelado que depende directamente de la calidad de los datos que se utilicen. Esta fase incluye la limpieza, construcción, integración y formateo de los datos, luego de estos pasos se realiza una selección de los atributos finales a utilizarse en la fase de modelado.

4.3.1 Construir los datos

Este paso consiste en generar nuevos campos o atributos, en base a los campos básicos, con la finalidad de facilitar el análisis y procesamiento de los datos, además de disminuir lo máximo posible la existencia de errores en los resultados que generen. Los campos calculados se listan en la Tabla 10.

Tabla 10 Campos calculados o construidos. Elaboración propia.

Nombre asignado al campo calculado	Descripción	Tabla de donde se extrae el campo básico	Campos básicos	Fórmula aplicada para obtención
EST_EDAD	Edad del estudiante en el periodo	Estudiante	A: Año de nacimiento	Edad = B - A
		Periodo	B: Año del periodo	
RELACION_AREAS_EST_Y_ASG	Relación entre el área del estudiante y el área de la asignatura	Asignatura	A: Área de la asignatura	Si A = B entonces 'S' caso contrario 'N'
		Titulación	B: Área de la titulación	
RELACION_AREAS_EST_Y_TIT_OFE	Relación entre la titulación del estudiante y la titulación que ofertó la asignatura	Oferta	Responsable de la oferta (solo titulaciones que ofertan asignaturas)	
		Expediente académico estudiante	Titulación del estudiante	
EST_NUM_ASIGNATURAS_MAT	Número de materias en las que se matriculó el estudiante en el semestre.	Oferta	Identificador único del semestre	Contabilizar el número de cursos que registra el estudiante en sus matrícula por semestre, agrupado por estudiante
			Identificador único del curso	
		Matricula	Identificador único del estudiante Estado de matrícula=*Aceptada*	
DOC_EDAD	Edad del docente en el periodo	Docente	A: Año de nacimiento	Edad = B - A
		Periodo	B: Año del periodo	
DOC_PERIODOS_DICTA_ASG	Numero de periodos en los que ha dictado la asignatura	Oferta-Matricula	Identificador de la Asignatura	Contabilizan los periodos en que el docente ha sido asignado a la asignatura, agrupados por docente y asignatura
			Identificador del Docente	
			Identificador del Periodo	
RELACION_AREAS_DOC_Y_ASG	Relación entre el área del docente y el área de la asignatura	Oferta-Matricula	A: Área de la asignatura	Si A = B entonces 'S' caso contrario 'N'
		Expediente académico del docente por periodo	B: Área del docente	
TOT_MATRICULADOS_CURSO_ASG	Total de matriculados en el curso	Oferta	Identificador único del semestre	Contabilizan los estudiantes matriculados agrupados por curso y por semestre
			Identificador único del curso	
		Matricula	Identificador único del estudiante	

4.3.2 Formatear datos

En esta paso se desea transformar los valores de ciertos atributos o campos, una forma es mediante la creación de intervalos que agrupen los valores.

Al tener como objetivo la comparación de los resultados que arrojen los algoritmos seleccionados, es necesario tener campos de la misma información pero con diferentes tipos de datos, y así poder utilizar la variable en todos los algoritmos y no afectar los resultados. Por ejemplo el campo ***Ciclo del estudiante N*** es una variable numérica, se agregará una

nueva columna denominada ***Ciclo del estudiante V*** que será la transformación de la variable numérica en una cualitativa. En la Tabla 11 se listan todos los campos transformados.

Tabla 11 Campos formateados. Elaboración propia.

Campo utilizado	Formateo realizado	Valor formateado	Nombre del campo transformado
Ciclo o módulo	* Se transforma de numérico a texto codificado * Se aplica intervalos	0-1 **BAS:** Básico; 2-4 **INI:** Inicial; 5-7 **MED:** Intermedio; 8-10 **SUP:** Superior	**Su terminación será en _V**
Edad del estudiante		Menor a 17: EDE1; 17 a 19: EDE2; 20 a 22: EDE3; 23 a 25: EDE4; 24 a 28: EDE5; 29 a 31: EDE6; 32 a 34: EDE7; 35 a 37: EDE8; 38 a 40: EDE9; Mayor a 40: EDE10	
Edad del docente		Menor a 28 **EDD1**; 28 a 32 **EDD2**; 33 a 37 **EDD3**; 38 a 42 **EDD4**; 43 a 47 **EDD5**; 48 a 52 **EDD6**; 53 a 57 **EDD7**; 58 a 62 **EDD8**; Mayor a 62 **EDD9**	
Promedio de notas semestral		<28 **INS**: Insuficiente; 28-31 **BAJ**: Bajo; 32-35 **ACE**: Aceptable; 36-39 **SOB**: Sobresaliente; 40 **EXC**: Excelente	
Promedio de notas bimestral		<14 **INS**: Insuficiente; 14-15 **BAJ**: Bajo; 16-17 **ACE**: Aceptable; 18-19 **SOB**: Sobresaliente; 20 **EXC**: Excelente	
Tasa de reprobación		<11 **BAJ**: Baja; 11-30 **ACE**: Aceptable; 31-50 **ALT**: Alta; >50 **MUY**: Muy alta	
Área académica	* Se transforma de texto a numérico	**1:** Administrativa; **2:** Biológica; **3:** Técnica; **4:** Humanística	**Su terminación será en _N**
Sexo de la persona		**1:** Femenino; **2:** Masculino	
Estado obtenido		**1**: Aprobado; **2**: Reprobado	
Jornada académica del docente		**1:** Tiempo completo; **2**: Medio tiempo; **3:** Tiempo parcial	
Título máximo del docente		**1:** Tercer nivel**; 2:** Especialista; **3:** Maestría; **4:** PhD	

4.3.3 Limpieza de los datos

A los siguientes campos se ha aplicado una limpieza de datos:

***Campo: EST_EDAD*.-** se ha detectado, como se observa en la Ilustración 12, que existen estudiantes cuya edad está fuera del rango normal (17 a 80 años). En este caso, al ser pocos los estudiantes detectados con valores fuera de rango u anormales, se realiza una corrección manual de las fechas de nacimiento, utilizando los datos del expediente físico.

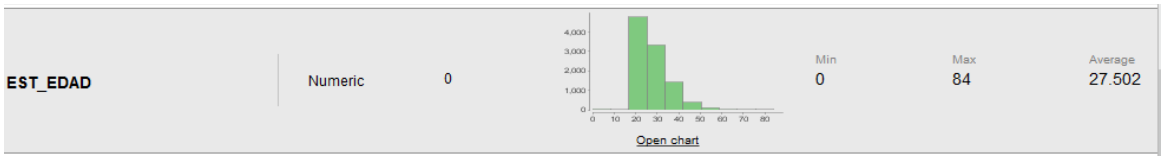


Ilustración 12 Datos antes de aplicar la limpieza - Gráfica generada con la herramienta Rapidminer 7.6. Elaboración propia.

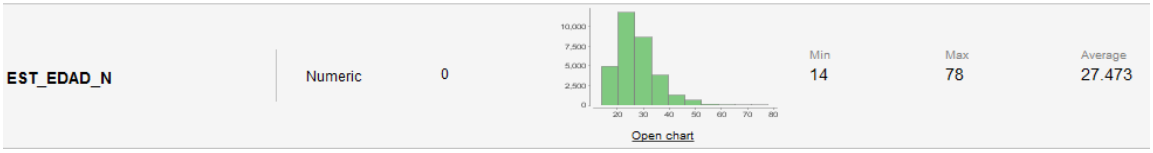


Ilustración 13 Datos después de aplicar la limpieza - Gráfica generada con la herramienta Rapidminer 7.6. Elaboración propia.

Los datos que se obtienen luego de realizar la limpieza se presentan en la Ilustración 13. Se observa que sólo queda un estudiante fuera del rango normal establecido (17 a 80 años). Este dato ha sido verificado y el estudiante posee realmente 14 años en el periodo de su matrícula.

Campo: EST_POSEE_DISCAPACIDAD.- se ha detectado, como se observa en la Ilustración 14, que la mayoría de los estudiantes no poseen ningún tipo de discapacidad, que puede estar afectando la aprobación de las asignaturas, razón por la cual se decide no utilizar este campo en los análisis posteriores.

Ilustración 14 Gráfica generada con la herramienta Rapidminer 7.6. Elaboración propia.

4.3.4 Integrar y seleccionar los atributos

Luego de realizar los pasos de recolección, descripción, exploración, limpieza, integración y formateo de los atributos iniciales, se han seleccionado y construido una *vista compacta* que contendrán los registros a utilizar, así como los atributos finales seleccionados para realizar el análisis del proyecto de minería de datos.

Obsérvese la Tabla 12 en donde se presenta un resumen de cada campo, con su tipo, valor mínimo y máximo, así como el catálogo de los valores que contendrá.

Tabla 12 Campos o atributos seleccionados. Elaboración propia

Contenido del campo[4]	Campo de donde se extraerá los datos	Tipo de campo[5]	Valor MIN.	Valor MAX.	Catálogo de valores	Campos a utilizar en técnicas de		
						Agrupación	Clasificación	Asociación
Asignatura								
Promedio de notas de la asignatura	ASIG_PROMEDIO_NOTAS_SEM_N	NU	30,81	34,68	Rango de valores: 0 - 40	x	x	x
	ASIG_PROMEDIO_NOTAS_SEM_V	NO			<28 **INS**: Insuficiente; 28-31 **BAJ**: Bajo; 32-35 **ACE**: Aceptable; 36-39 **SOB**: Sobresaliente; 40 **EXC**: Excelente			
Promedio de notas de la asignatura primer bimestre	ASIG_PROMEDIO_NOTAS_BIM1_V	NU	13,90	16,56	Rango de valores: 0 - 20	x	x	x
	ASIG_PROMEDIO_NOTAS_BIM1_N	NO			<14 **INS**: Insuficiente; 14-15 **BAJ**: Bajo; 16-17 **ACE**: Aceptable; 18-19 **SOB**: Sobresaliente; 20 **EXC**: Excelente			
Promedio de notas de la asignatura segundo bimestre	ASIG_PROMEDIO_NOTAS_BIM2_N	NU	15,05	16,77	Rango de valores: 0 - 20	x		
	ASIG_PROMEDIO_NOTAS_BIM2_V	NO			<14 **INS**: Insuficiente; 14-15 **BAJ**: Bajo; 16-17 **ACE**: Aceptable; 18-19 **SOB**: Sobresaliente; 20 **EXC**: Excelente			
Tasa de reprobación de la asignatura	ASIG TASA_REPROBACION_N	NU	15,58	72,07	Rango de valores: 0 - 100	x	x	x
	ASIG TASA_REPROBACION_V	NO			<11 **BAJ**: Baja; 11-30 **ACE**: Aceptable; 31-50 **ALT**: Alta; >50 **MUY**: Muy alta			
Unidad que oferta la asignatura para los estudiantes	ASIG_COD_TIT_RESPONSABLE_OFERT	NO			PRE_DIS_ABO,,PRE_DIS_LIC_CON,PRE_DIS_PSI,PRE_DIS_INF,PRE_DIS_ADM_EMP,PRE_DIS_ECO,PRE_DIS_GES_AMB,PRE_DIS_EDU_INF,PRE_DIS_COM,PRE_DIS_EDU_BAS,PRE_DIS_ING_ADM_HOT,PRE_DIS_ING,PRE_DIS_ADM_BAN,PRE_DIS_LEN_LIT,PRE_DIS_QUI,PRE_DIS_FIS			
Área a la que pertenece el responsable de ofertar la asignatura en el periodo	ASIG_AREA_RESPONSABLE_OFERT	NO			**AR1:** Administrativa; **AR2:** Biológica; **AR3:** Técnica; **AR4:** Humanística			
Área académica a la que pertenece la asignatura	ASIG_AREA_ACADEMICA_V	NO						
	ASIG_AREA_ACADEMICA_N	NU			**1:** Administrativa; **2:** Biológica; **3:** Técnica; **4:** Humanística	x		
Ciclo de la asignatura, relacionada a la titulación del estudiante	ASIG_CICLO_N	NU	1	8	Rango de valores: 1 – 8	x	x	x
	ASIG_CICLO_V	NO			0-1 **BAS:** Básico; 2-4 **INI:** Inicial; 5-7 **MED:** Intermedio; 8-10 **SUP:** Superior			
Estudiante								
Edad del estudiante	EST_EDAD_N	NU	14	78	Rango de valores: 14 – 78	x		
	EST_EDAD_V	NO			Menor a 17: EDE1; 17 a 19: EDE2; 20 a 22: EDE3; 23 a 25: EDE4; 24 a 28: EDE5; 29 a 31: EDE6; 32 a 34: EDE7; 35 a 37: EDE8; 38 a 40: EDE9; Mayor a 40: EDE10			
Sexo del estudiante	EST _SEXO_V	NO			**F:** Femenino; **M:** Masculino			
	EST _SEXO_N	NU			**1:** Femenino; **2:** Masculino	x		
Código del área académica a la que pertenece la titulación del estudiante	EST_AREA_ACADEMICA_V	NO			**AR1:** Administrativa; **AR2:** Biológica; **AR3:** Técnica; **AR4:** Humanística			
	EST_AREA_ACADEMICA_N	NU			**1:** Administrativa; **2:** Biológica; **3:** Técnica; **4:** Humanística	x		
Código de la titulación del estudiante	EST_COD_TIT_ESTUDIANTE	NO			PRE_DIS_ABO,PRE_DIS_PSI,PRE_DIS_INF,PRE_DIS_LIC_CON,PRE_DIS_ADM_EMP,PRE_DIS_ECO,PRE_DIS_GES_AMB,PRE_DIS_EDU_INF,PRE_DIS_ING_CON,PRE_DIS_COM,PRE_DIS_EDU_BAS,PRE_DIS_ASI_GER_REL,PRE_DIS_ADM_BAN,PRE_DIS_ADM_GES_PUB,PRE_DIS_ING_ADM_HOT,PRE_DIS_ING,PRE_DIS_LEN_LIT,PRE_DIS_QUI,PRE_DIS_FIS			
Tasa de reprobación del estudiante	EST _TASA_REPROBACION_N	NU	0	100	Rango de valores: 0 - 100	x	x	x
	EST _TASA_REPROBACION_V	NO			<11 **BAJ**: Baja; 11-30 **ACE**: Aceptable; 31-50 **ALT**: Alta; >50 **MUY**: Muy alta			

[4] **Anaranjado**: Campos que contienen datos del periodo o semestre; **Celeste**: campos que contienen datos acumulados de la entidad asignatura, docente o estudiante

[5] **NO**: Nominal; NU: Numérico;

Promedio de notas semestral del estudiante	EST_PROMEDIO_NOTAS_SEM_N	NU	0	40	Rango de valores: 0 – 40	x	x	x
	EST_PROMEDIO_NOTAS_SEM_V	NO			<28 **INS**: Insuficiente; 28-31 **BAJ**: Bajo; 32-35 **ACE**: Aceptable; 36-39 **SOB**: Sobresaliente; 40 **EXC**: Excelente			
Promedio de notas del primer bimestre	EST_PROMEDIO_NOTAS_BIM1_N	NU	0	21,63		x	x	x
	EST_PROMEDIO_NOTAS_BIM1_V	NO			<14 **INS**: Insuficiente; 14-15 **BAJ**: Bajo; 16-17 **ACE**: Aceptable; 18-19 **SOB**: Sobresaliente; 20 **EXC**: Excelente			
Promedio de asignaturas en las que se matricula el estudiante por periodo	EST_PROMEDIO_ASI_MAT_ES_N	NU	1	9	Rango de valores: 0 - 9	x		
	EST_PROMEDIO_ASI_MAT_ES_V	NO			1-2 **MIN**: Mínimo; 3-4 **NOB**: Normal Bajo; 5-6 **NOM**: Normal Medio; 7-8 **NOA**: Normal Alto; >8 **SON**: Sobrepasa normal			
Ciclo del estudiante en el periodo	EST_CICLO_N	NU	1	10	Rango de valores: 1 - 10	x		
	EST_CICLO_V	NO			0-1 **BAS:** Básico; 2-4 **INI:** Inicial; 5-7 **MED:** Intermedio; 8-10 **SUP:** Superior			
Nota primer bimestre obtenida en el periodo y en la asignatura	EST_NOTA_BIM1_N	NU	0	20	Rango de valores: 0 - 20	x	x	x
	EST_NOTA_BIM1_V	NO			<14 **INS**: Insuficiente; 14-15 **BAJ**: Bajo; 16-17 **ACE**: Aceptable; 18-19 **SOB**: Sobresaliente; 20 **EXC**: Excelente			
Nota semestral obtenida en el periodo y asignatura	EST_NOTA_SEM	NU	0	40	Rango de valores: 0 - 40	x		
		NO			<28 **INS**: Insuficiente; 28-31 **BAJ**: Bajo; 32-35 **ACE**: Aceptable; 36-39 **SOB**: Sobresaliente; 40 **EXC**: Excelente			
Estado obtenido (Aprobado, Reprobado, etc.) en el periodo y en la asignatura (CAMPO ***CLASE***)	EST_ESTADO_V	NO			**AP**: Aprobado; **RE**: Reprobado		x	x
	EST_ESTADO_N	NU	1	2	**1**: Aprobado; **2**: Reprobado	x		
Código del centro universitario	EST_CENTRO_ACADEMICO	NO			ALA1,ALU2,AMB3,AZO5,BAH6,BAL8,BOL105,CAL10,CAN11,CAR13,CAT14,CAY15,CEL16,CHO18,CMI51,CMP114,COC19,CQC12,CUE20,DAU21,DUR22,ECH90,EMS24,GUA28,GUD27, GUL29,GUR30,GUY32,GYE31,GYEC17,HUA33,IBA34,ISA35,JOY37,LAG39,LAT40,LCO38,LII41,LOJ43,MAA45,MAC44,MAD100,MAD101,MAD136,MAD137,MAD48, MAD97,MAD98,MAD99,MAT49,MCH46,MEN50,MLA47,UL53,NYK52,OTA54,PAS56,PAU57,PED58,PIS60,POR61,POV62,PUY64,QTO119,QTO67,QTOSR73,QTOV91,QUE65,QUI66,RIO68,ROM103,ROM69,SAM106,SAR81,SCR71,SDO82,SGA72, SHU83,SIS77,SLI70,SLO74,SMB75,STC76,STD79,STO80,STR78,SUC84,TEN86,TUL87,TUM88,TUR89,VIN92,YAN93,ZAM94			
Región en donde se encuentra ubicado el estudiante	EST_REGION_UBICACION	NO			S: Sierra; O: Oriente; I: Insular; E: Extranjero; C: Costa			
Número de asignaturas en las que se matricula en el periodo. Normalmente un estudiante debe matricularse en máximo 8 asignaturas por semestre.	EST_TOT_ASIG_MATRICULA_N	NU	1	10		x	x	x
	EST_TOT_ASIG_MATRICULA_V	NO			1-2 **MIN**: Mínimo; 3-4 **MEM**: Medio mínimo; 5-6 **MED**: Medio máximo; 7-8 **MAX**: Máximo; >8 **SOB**: Sobrepasa máximo			
Número de asignaturas reprobadas en periodos anteriores	EST_TOT_ASIG_REPROB_PER_ANT	NU	0	30		x	x	x
Perdió o no una asignatura en periodos anteriores	EST_PERDIO_ASIG_PER_ANT_N	NU			1: Si; 0: No	x		
	EST_PERDIO_ASIG_PER_ANT_V	NO			S: Si; N: No			
Número de foros en los que participó el estudiante	EST_FOROS	NU	0	6		x	x	x
Número de chats en los que participó el estudiante	EST_CHATS	NU	0	4		x	x	x
Número de cuestionarios en los que participó el estudiante	EST_CUESTIONARIOS	NU	0	8		x	x	x
Docente								
Sexo del docente	DOC_SEXO_V	NO			**F:** Femenino; **M:** Masculino			
	DOC_SEXO_N	NU			**1:** Femenino; **2:** Masculino	x	x	x
Promedio de notas semestral del docente	DOC_PROMEDIO_NOTAS_SEM_N	NU	18,59	33,69	Rango de valores: 0 - 40	x	x	x
	DOC_PROMEDIO_NOTAS_SEM_V	NO			<28 **INS**: Insuficiente; 28-31 **BAJ**: Bajo; 32-35 **ACE**: Aceptable; 36-39 **SOB**: Sobresaliente; 40 **EXC**: Excelente			
Tasa de reprobación del docente	DOC_TASA_REPROBACION_N	NU	11,46	74,89	Rango de valores: 0 - 100	x	x	x
	DOC_TASA_REPROBACION_V	NO			<11 **BAJ**: Baja; 11-30 **ACE**: Aceptable; 31-50 **ALT**: Alta; >50 **MUY**: Muy alta			

Numero de periodos como docente en la institución	DOC_PERIODOS_DOCENCIA	NU	1	14		x	x	x
Numero de periodos que imparte la asignatura dentro de la institución	DOC_PERIODOS_DICTA_ASG	NU	1	14		x	x	x
Código del área del docente a la que pertenece	DOC_AREA_V	NO			**AR1:** Administrativa;**AR2:** Biológica; **AR3:** Técnica; **AR4:** Humanística			
	DOC_AREA_N	NU			**1:** Administrativa; **2:** Biológica; **3:** Técnica; **4:** Humanística	x		
Edad del docente	DOC_EDAD_N	NU	25	63		x		
	DOC_EDAD_V	NO			Menor a 28 **EDD1**; 28 a 32 **EDD2**; 33 a 37 **EDD3**; 38 a 42 **EDD4**; 43 a 47 **EDD5**; 48 a 52 **EDD6**; 53 a 57 **EDD7**; 58 a 62 **EDD8**; Mayor a 62 **EDD9**			
Código de la jornada del docente en el periodo	DOC_JORNADA_ACADEMICA_V	NO			**TC**: Tiempo completo; **MT**: Medio tiempo; **TP**: Tiempo parcial		x	x
	DOC_JORNADA_ACADEMICA_N	NU			**1:** Tiempo completo; **2**: Medio tiempo; **3:** Tiempo parcial	x		
Código del título máximo que posee	DOC_TITULO_ACAD_MAX_V	NO			**TEN**: Tercer nivel; **ESP**: Especialista; **MAE**: Maestría; **PHD**: PhD		x	x
	DOC_TITULO_ACAD_MA_N	NU			**1:** Tercer nivel**; 2:** Especialista; **3:** Maestría; **4:** PhD	x		
Posee un cargo directivo	DOC_POSEE_CARGO_DIR_V	NO			S: Si; N: No			
	DOC_POSEE_CARGO_DIR_N	NU			1: Si; 0: No	x	x	x
Posee un cargo académico	DOC_POSEE_CARGO_ACA_V	NO			S: Si; N: No			
	DOC_POSEE_CARGO_ACA_N	NU			1: Si; 0: No	x	x	x
Total de asignaturas asignadas en el periodo	DOC_TOTAL_ASIG_SEM	NU	1	6		x	x	x
Total de estudiantes asignados en el periodo	DOC_TOTAL_EST_SEM	NU	41	953		x	x	x
Total de paralelos asignados en el periodo	DOC_TOTAL_PAR_SEM	NU	1	20		x	x	x
Es docente bimodal	DOC_BIMODAL_V	NO			S: Si; N: No			
	DOC_BIMODAL_N	NU			1: Si; 0: No	x	x	x
El docente tiene rol administrativo	DOC_ADMINISTRATIVO_V	NO			S: Si; N: No			
	DOC_ADMINISTRATIVO_N	NU			1: Si; 0: No	x	x	x
El docente es Prometeo. Docente PHD temporal	DOC_PROMETEO_V	NO			S: Si; N: No			
	DOC_PROMETEO_N	NU			1: Si; 0: No	x	x	x
Asignación tardía del docente al curso. Detectar si el docente fue asignado al curso luego de iniciadas las clases	DOC_ASIGNACION_TARDIA_V	NO			S: Si; N: No			
	DOC_ASIGNACION_TARDIA_N	NU			1: Si; 0: No	x	x	x
Total foros planteados a los estudiantes con retroalimentación	DOC_TOTAL_FOROS	NU	0	8		x	x	x
Total chats realizados	DOC_TOTAL_CHATS_ACAD	NU	0	22		x	x	x
Total cuestionarios planteados	DOC_TOTAL_CUESTIONARIOS	NU	0	10		x	x	x
Total anuncios realizados	DOC_TOTAL_ANUNCIOS	NU	0	19		x	x	x
Total Microvideos compartidos con el estudiante	DOC_MICROVIDEOS	NU	0	14		x	x	x
Trasversales								
Indica si existe o no relación entre el área del estudiante con el área de la asignatura	RELACION_AREAS_EST_Y_ASG_V	NO			S: Si; N: No		x	x
	RELACION_AREAS_EST_Y_ASG_N	NU			1: Si; 0: No	x		
Indica si existe o no relación entre el área del estudiante con el área de la titulación que oferto	RELACION_AREAS_EST_Y_TIT_OFE_V	NO			S: Si; N: No		x	x
	RELACION_AREAS_EST_Y_TIT_OFE_N	NU			1: Si; 0: No	x		
Indica si existe o no relación entre el área del docente con el área de la asignatura	RELACION_AREAS_DOC_Y_ASG_V	NO			S: Si; N: No		x	x
	RELACION_AREAS_DOC_Y_ASG_N	NU			1: Si; 0: No	x		
Indica si existe o no relación entre la titulación del estudiante y la titulación que oferta la asignatura en el periodo	RELACION_TIT_EST_Y_TIT_OFE_V	NO			S: Si; N: No		x	x
	RELACION_TIT_EST_Y_TIT_OFE_N	NU			1: Si; 0: No	x		
Total de matriculados en el periodo en la asignatura-curso (*asignatura + docente + paralelo + titulación*)	TOT_MATRICULADOS_CURSO_ASG	NU	1	82		x		

4.3.5 Explorar los datos

Algunos autores recomiendan realizar una exploración con los atributos iniciales, con la finalidad de ir refinando los atributos y detectar aquellos que contengan datos que puede que no aporten en este proyecto.

Así también, este paso, es muy útil para encontrar posibles errores en los procesos de *extracción, transformación o limpieza* de los datos, que estén causando inconsistencias en los mismos. Este proceso se puede realizar varias veces hasta obtener datos de alta calidad y que puedan ser utilizados.

El total de materias filtradoras detectadas en base a los requisitos mencionados en la sección 4.2.1 RECOLECCIÓN DE DATOS INICIALES, y que van a ser analizadas en este proyecto es de ***35***, obsérvese la Ilustración 15 en donde el área 1 y 3 poseen el mayor número de materias filtradoras, y el área 4 es la de menor número.

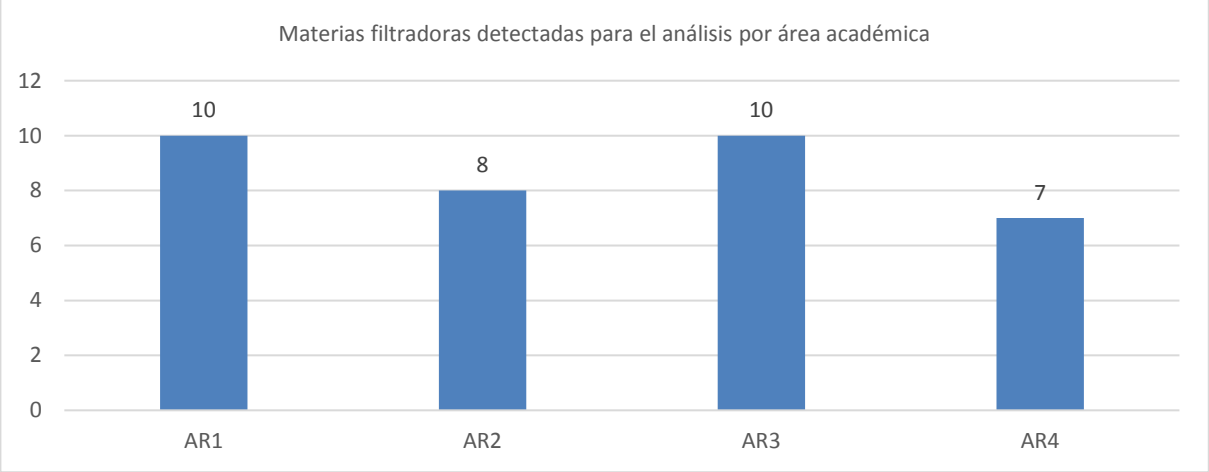


Ilustración 15 Materias filtradoras detectadas para el análisis por área académica. Elaboración propia.

Las 35 materias filtradoras obtenidas son ofertadas a estudiantes de diversas titulaciones, por lo que en cada titulación la materia o asignatura tiene su propio ciclo. Por ejemplo, la asignatura *ASIG_000866* en la titulación *PRE_DIS_ADM_GES_PUB* se encuentra en el ciclo 4, pero en la titulación *PRE_DIS_LIC_CON* pertenece al ciclo 2.

Por lo tanto se detecta que, en los datos recolectados, la mayoría de estas asignaturas se agrupan en el ciclo INI (Inicial) que corresponde a los niveles o módulos 2,3 y 4; seguido por el ciclo BAS (Básico) que corresponde al nivel o módulo 1; y con bajo número de asignaturas el ciclo MED (Intermedio) que agrupa los módulos 5,6 y 7. En el ciclo SUP (Superior) se descubre poca existencia de materias filtradoras, obsérvese la Ilustración 16.

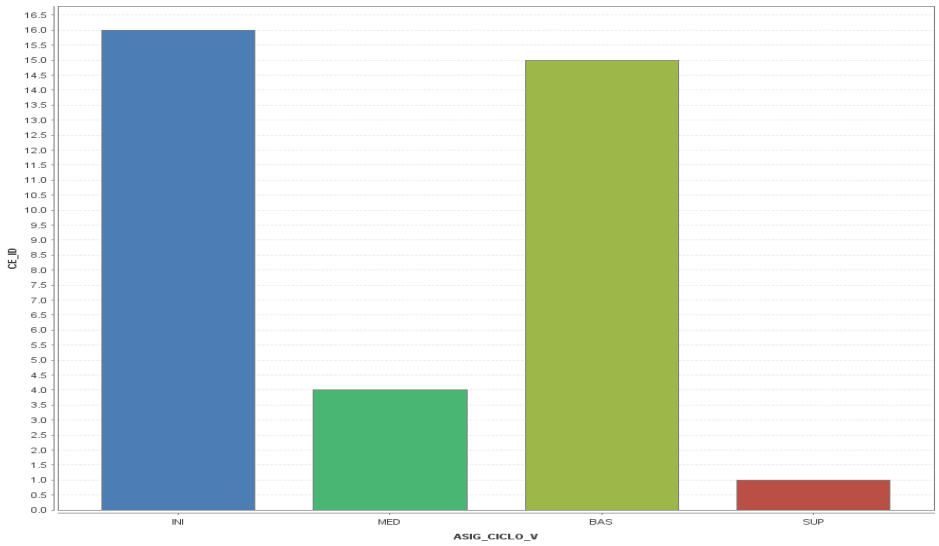


Ilustración 16 Número de materias filtrador por ciclo o módulo de la asignatura. Elaboración propia con Rapidminer

El número de registros extraídos para realizar el análisis de datos, se basa en las *35* asignaturas filtradoras detectadas, utilizando estas materias se obtienen los registros de los *3* últimos periodos académicos, de los estudiantes que se matricularon en estas asignaturas, y que las aprobaron o reprobaron. Es así que el número de registros a utilizarse es de *31321*, de los cuales *15640* son de estudiantes matriculados que aprobaron las asignaturas en alguno de los 3 periodos seleccionados, y *15681* son estudiantes que las reprobaron.

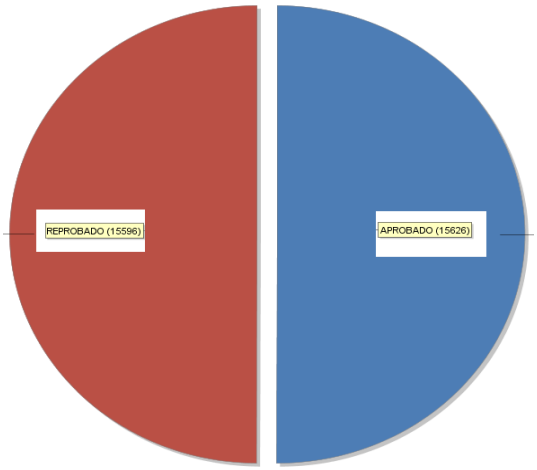


Ilustración 17 Registros por estado de aprobación y reprobación. Elaboración propia con Rapidminer

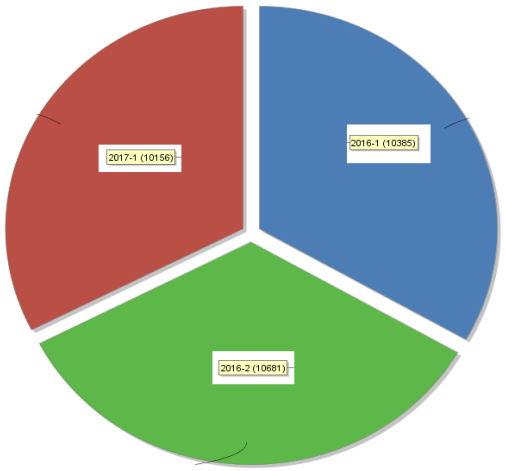


Ilustración 18 Registros extraídos por periodo académico. Elaboración propia con Rapidminer

Del total de registros extraídos, el 50% *(15660 registros)* se utilizará en la fase de entrenamiento de los algoritmos a comparar, y el otro 50% de utilizará para realizar la fase de

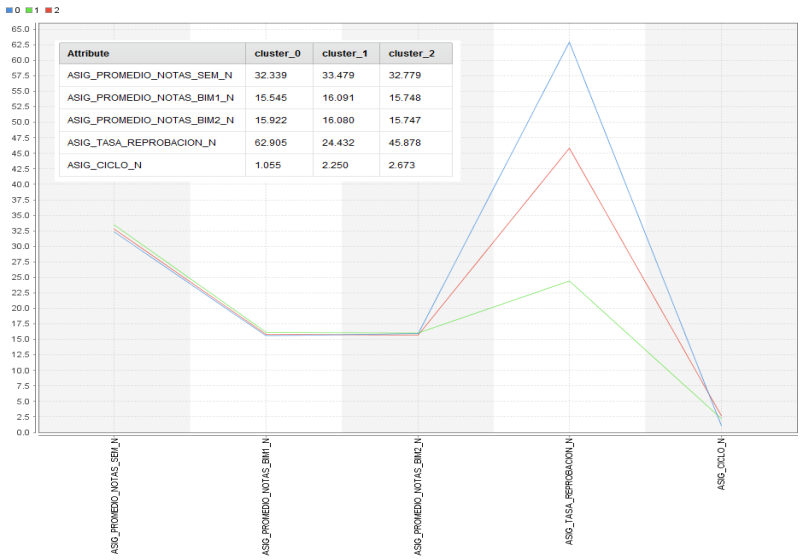
validación de los resultados que generen los algoritmos. Cada periodo académico seleccionado contiene un total aproximado de entre 10200 a 10700 registros, pudiéndose decir que cada periodo corresponde a 1/3 del total de registros extraídos, lo que indica que la cantidad de registros de cada periodo es similar, obsérvese la Ilustración 18.

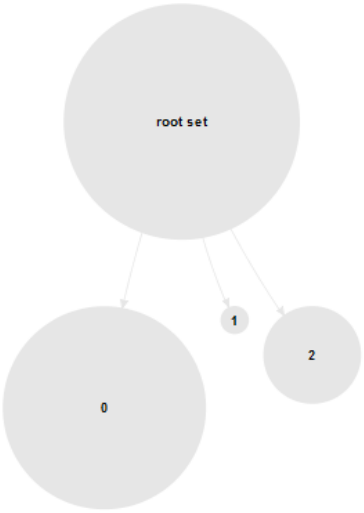
A continuación se realizarán dos exploraciones, se utilizará la técnica de minería de datos *clustering y* una comparación de varios campos de forma gráfica y manual. En ambos casos se utilizará la herramienta *Rapidminer.*

La exploración sirve para ir mejorando la selección de los campos a utilizar, con la finalidad de tener campos o atributos que tengan realmente incidencia en el problema de la reprobación.

4.3.5.1 Exploración de datos de la asignatura

Se utiliza el algoritmo *k-means,* ejecutándolo con varias valores de *k,* finalmente se ha seleccionado el valor *k=3* como apto para este conjunto de campos. Los campos numéricos que se utilizan en esta exploración contienden datos característicos de la asignatura (se han resaltado de color celeste en la Tabla 12). Los resultados obtenidos se presentan en la Ilustración 19.





A) Gráfica y tabla de cada clúster generado

B) Número de registros o transacciones distribuidas en los 3 clústers generados

Ilustración 19 Resultados clustering K-means – Datos Asignaturas, k= 3. Elaboración propia con Rapidminer.

De estos resultados se resaltan los siguientes puntos:

- *El clúster 0, es el que* agrupa la mayor parte de registros analizados. Las características de las asignaturas filtradoras que se agruparon en el *clúster 0 son:*
    - Tienen promedio de notas semestral cercano a 32.3/40 puntos
    - Tienen promedio de notas del primer y segundo bimestre cercano a 15.5/20. El promedio del primer bimestre es menor al del segundo bimestre (ASIG_PROMEDIO_NOTAS_BIM1< ASIG_PROMEDIO_NOTAS_BIM2). Lo que indica que en el segundo bimestre pueden existir contenidos menos complejos o existe mejor material de clase, etc.
    - Tienen las tasas de reprobación más altas con respecto al resto de asignaturas filtradoras.
    - Pertenecen a ciclos básicos o iniciales de las titulaciones.
- Las características de las asignaturas filtradoras que se agruparon en el *clúster 1 son:*
    - Tienen promedio de notas semestral cercano a 33.5/40 puntos
    - Tienen promedio de notas del primer y segundo bimestre cercano a 16.1/20. El promedio del primer bimestre es similar al promedio del segundo bimestre (ASIG_PROMEDIO_NOTAS_BIM1= ASIG_PROMEDIO_NOTAS_BIM2). Lo que indica que en ambos bimestres se mantiene una buena nota.
    - Las tasas de reprobación son bajas con respecto al resto de asignaturas filtradoras
    - Pertenecen a ciclos básicos o iniciales de las titulaciones.
- Las características de las asignaturas filtradoras que se agruparon en el *clúster 2 son:*
    - Tienen promedio de notas semestral cercano a 32.7/40 puntos
    - Tienen promedio de notas del primer y segundo bimestre cercano a 15.7/20. El promedio del primer bimestre es similar al promedio del segundo bimestre (ASIG_PROMEDIO_NOTAS_BIM1> ASIG_PROMEDIO_NOTAS_BIM2). Lo que indica que en el segundo bimestre, pueden existir contenidos de la materia más complejos que en el primer bimestre o menor cantidad de material, etc.
    - Las tasas de reprobación son medias altas con respecto al resto de asignaturas filtradoras.
    - Pertenecen a ciclos básicos o iniciales de las titulaciones.
- Las asignaturas filtradoras con mejores promedios de notas semestrales, tienen también un promedio de notas bimestrales mayor a 15.5 puntos, obsérvese la Ilustración 20. Se han detectado casos irregulares de lo mencionado anteriormente, obsérvese el círculo violeta de la Ilustración 20, en donde el promedio del primer bimestre es bajo en

comparación con el promedio del segundo bimestre, en donde el promedio semestral es medio alto.

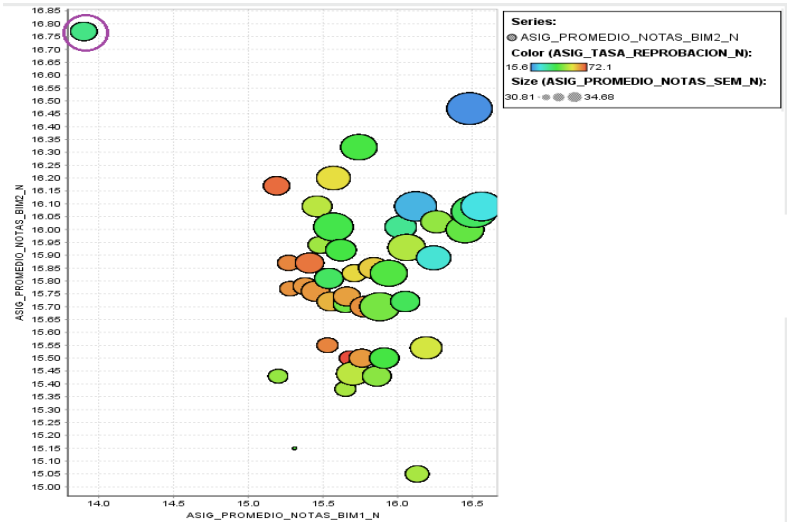


Ilustración 20 Promedio de notas del primero y segundo bimestre en escala del promedio de notas semestral de la asignatura. Elaboración propia con Rapidminer

- Las asignaturas filtradoras con mejores promedios de notas semestrales, poseen menor tasa de reprobación. En cambio las asignaturas que poseen bajos promedios de notas semestrales, tienen alta tasa de reprobación. Se detecta que existe una correlación negativa entre estos dos valores obsérvese la Ilustración 21, cuando un valor baja el otro sube:
    - La tasa de reprobación es alta y el promedio de notas semestral es bajo.
    - La tasa de reprobación es baja y el promedio de notas semestral es alto.

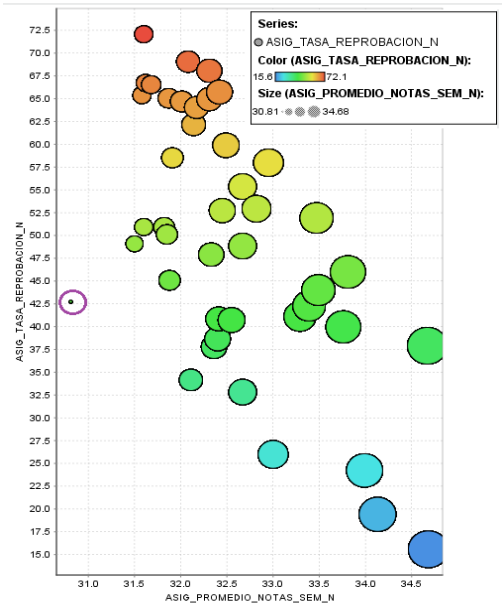


Ilustración 21 Tasa de reprobación frente al promedio de notas semestral de la asignatura-titulación. Elaboración propia con Rapidminer

- Se observa una relación entre las asignaturas con altas tasas de reprobación y el ciclo de la asignatura, detectándose que las asignaturas de ciclos básicos e iniciales tienen mayor reprobación, obsérvese la Ilustración 22.

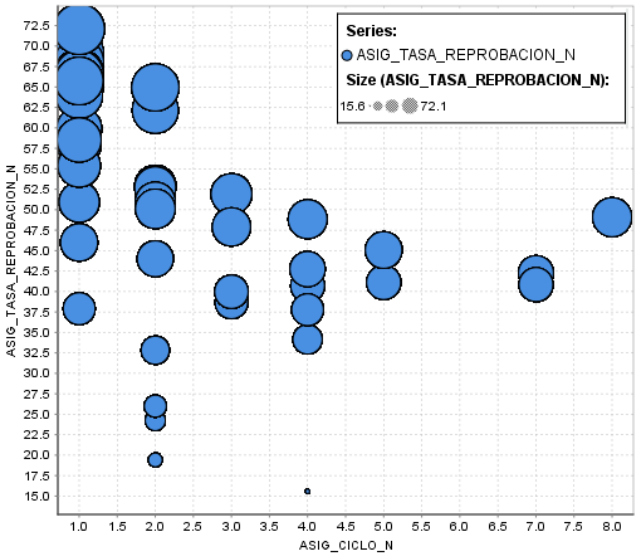


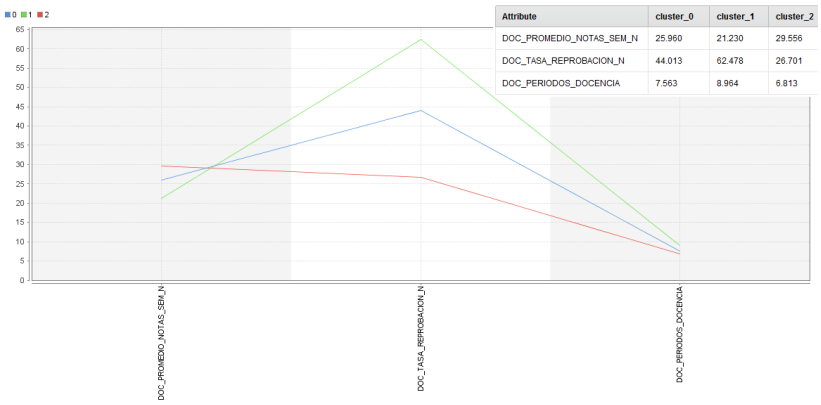
Ilustración 22 Ciclo de la asignatura frente a la tasa de reprobación de la asignatura. Elaboración propia con Rapidminer

Esta exploración permite realizar conclusiones como: *de las asignaturas filtradoras analizadas, existen aquellas que tienen un alto promedio de notas semestrales, un buen promedio de notas en sus dos bimestres, baja tasa de reprobación y pertenecen a ciclos medios altos. Existen también asignaturas que tienen bajo promedio semestral, bajo promedio de notas en el primer, alta tasa de reprobación y pertenecen a ciclos básicos e iniciales.*

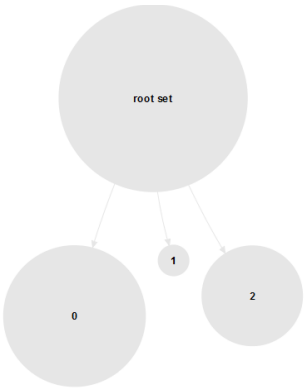
4.3.5.2 Exploración de datos del docente

Se utiliza el algoritmo *k-means,* ejecutándolo con varias valores de *k,* finalmente se ha seleccionado el valor *k=3* como apto para este conjunto de campos.

Los campos numéricos que se utilizan en esta exploración contienden datos característicos del docente (se han resaltado de color celeste en la Tabla 12). Los resultados obtenidos se presentan en la Ilustración 23.



A) Gráfica y tabla de cada clúster generado



B) Número de registros o transacciones distribuidas en los 3 clústers generados

Ilustración 23 Resultados clustering K-means – Datos Docente, k=3. Elaboración propia con Rapidminer

De estos resultados se resaltan los siguientes puntos:

- El *clúster* 0, es el que agrupa la mayor parte de registros analizados. Los docentes agrupados en este clúster tienen un bajo promedio de notas semestral <28/40, una alta de tasa de reprobación >30%, y la mitad de experiencia en semestres con respecto al resto de docentes analizados.
- La diferencia entre el *clúster* 0 y 1, es el número de semestres que el docente ha impartido clases en la institución. En el *clúster* 1 este número es mayor al del *clúster* 0.
- En el *clúster* 2 se agrupan docentes que tienen un promedio de notas semestral superior al 28/40, la tasa de reprobación más baja de los tres clústers creados. El número de semestres de experiencia docente en la institución es menor al del resto de *clústers* pero no es mucha la diferencia.

Para profundizar los puntos antes mencionados, se realiza una exploración manual de los campos: *tasa de reprobación del docente y el número de periodos en que este ha impartido clases en la institución*. En la Ilustración 24, se observa el resultado del cruce de estas dos variables, en donde la tasa de reprobación es similar entre casi todos docentes, sin verse afectada esta tasa por el número de semestres de permanencia que posean los docentes. Sin embargo hay una excepción, para los docentes que tienen menos de 8 semestres de permanencia, para ellos la tasa de reprobación es mayor que la de aquellos docentes que tienen más de 8 semestres de permanencia.

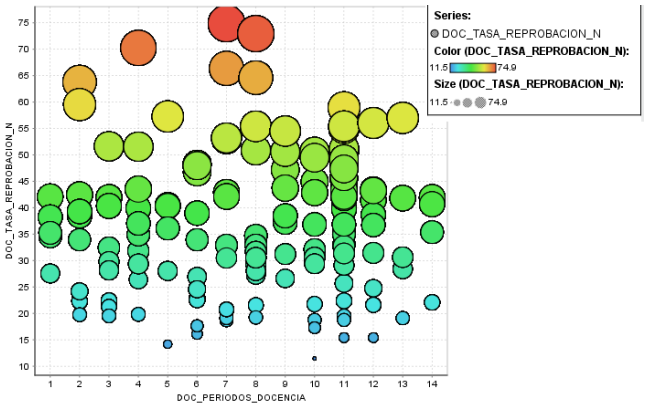


Ilustración 24 Tasa de reprobación del docente frente a los periodos de permanencia en la institución. Elaboración propia con Rapidminer.

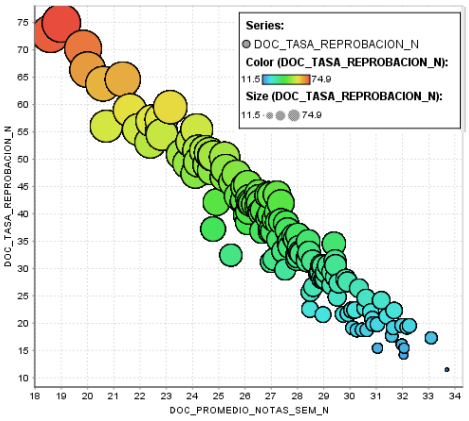


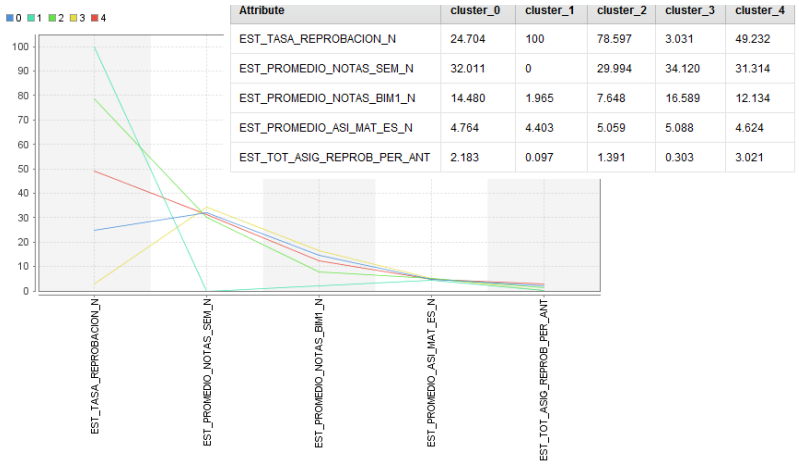
Ilustración 25 Promedio de notas semestral del docente frente a su tasa de reprobación. Elaboración propia con Rapidminer.

En la Ilustración 25 se analizan dos variables o campos del docente: *la tasa de reprobación y el promedio de notas semestral.* En esta exploración se encuentra que, existe una correlación negativa entre estos dos campos, en donde si el promedio es alto, la tasa de reprobación es baja y viceversa.

4.3.5.3 Exploración de datos del estudiante

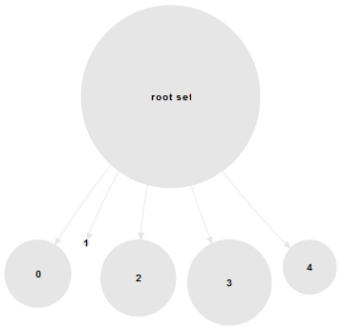
Se utiliza el algoritmo *k-means,* ejecutándolo con varias valores de *k,* finalmente se ha seleccionado el valor *k=5* como apto para este conjunto de campos.

Los campos numéricos que se utilizan en esta exploración contienden datos característicos del estudiante (se han resaltado de color celeste en la Tabla 12). Los resultados obtenidos se presentan en la Ilustración 26.



Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
EST_TASA_REPROBACION_N	24.704	100	78.597	3.031	49.232
EST_PROMEDIO_NOTAS_SEM_N	32.011	0	29.994	34.120	31.314
EST_PROMEDIO_NOTAS_BIM1_N	14.480	1.965	7.648	16.589	12.134
EST_PROMEDIO_ASI_MAT_ES_N	4.764	4.403	5.059	5.088	4.624
EST_TOT_ASIG_REPROB_PER_ANT	2.183	0.097	1.391	0.303	3.021

A) Gráfica y tabla de cada clúster generado



B) Número de registros o transacciones distribuidas en los 5 clústers generados

Ilustración 26 Resultados clustering K-means – Datos del estudiante, k= 5. Elaboración propia con Rapidminer

De este resultado se puede resaltar los siguientes puntos:

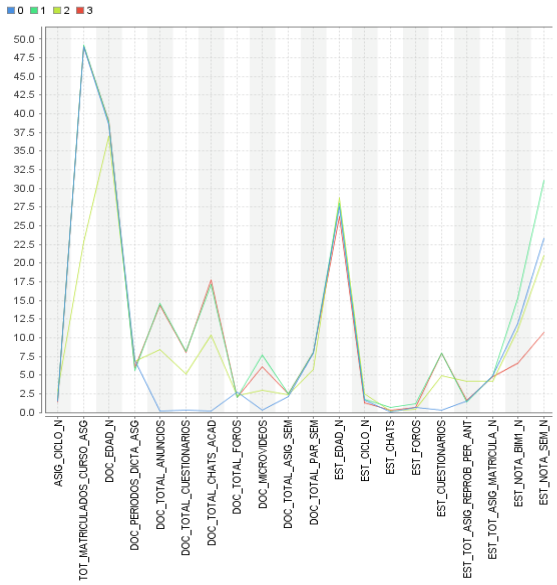
- La principal diferencia entre los *clústers* generados son, las *tasas de reprobación y los promedios de notas*. En el *clúster 3* se agrupan a los estudiantes con las más bajas tasas de reprobación de todos los estudiantes analizados, y por lo tanto poseen los más altos *promedios de notas* (*semestral y primer bimestral*). Aquí también se ubican a los estudiantes que han reprobado una o ninguna materia en periodos anteriores.
- El *clúster 2,* contiene a los estudiantes con las *tasas de reprobación* más altas,
- El campo *promedio de materias* en las que se matricula el estudiante semestralmente, no tiene mayor incidencia en la *tasa de reprobación y promedio de notas* del estudiante.

En base a los resultados obtenidos se considera, no utilizar el campo *promedio de materias en las que se matricula el estudiante semestralmente,* en la fase de modelado.

4.3.5.4 Exploración de datos semestrales del docente-asignatura-estudiante

Se utiliza el algoritmo *k-means,* ejecutándolo con varias valores de *k,* finalmente se ha seleccionado el valor *k=4* como apto para este conjunto de campos. Los campos numéricos que se utilizan en esta exploración, contienden datos obtenidos en cada semestre. Por ejemplo la nota que ha obtenido el estudiante en el semestre, cuando su docente tenía 25 años y el estudiante realizó sólo 2 foros, etc.

El resultado de esta agrupación se presenta en la Ilustración 27.



Attribute	cluster_3	cluster_2	cluster_0	cluster_1
ASIG_CICLO_N	1.403	2.831	1.513	1.612
TOT_MATRICULADOS_CURSO_A...	49.062	22.763	48.991	49.196
DOC_EDAD_N	38.956	37.010	38.498	38.844
DOC_PERIODOS_DICTA_ASG	5.976	6.803	6.964	5.682
DOC_TOTAL_ANUNCIOS	14.333	8.376	0.174	14.531
DOC_TOTAL_CUESTIONARIOS	8.089	5.119	0.265	8.190
DOC_TOTAL_CHATS_ACAD	17.670	10.389	0.213	17.135
DOC_TOTAL_FOROS	1.961	2.180	2.754	1.976
DOC_MICROVIDEOS	6.169	3.023	0.242	7.699
DOC_TOTAL_ASIG_SEM	2.488	2.348	2.068	2.469
DOC_TOTAL_PAR_SEM	8.100	5.742	7.914	8.025
EST_EDAD_N	26.136	28.684	27.579	27.999
EST_CICLO_N	1.280	2.481	1.582	1.786
EST_CHATS	0.301	0.219	0.042	0.693
EST_FOROS	0.684	0.477	0.669	1.179
EST_CUESTIONARIOS	7.902	4.914	0.253	7.908
EST_TOT_ASIG_REPROB_PER_A...	1.591	4.231	1.500	1.375
EST_TOT_ASIG_MATRICULA_N	4.783	4.231	4.957	4.925
EST_NOTA_BIM1_N	6.544	11.036	11.930	15.248
EST_NOTA_SEM_N	10.668	20.947	23.237	31.078

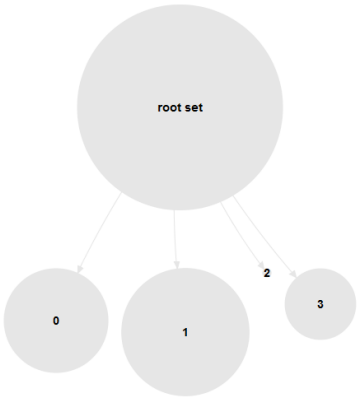


Ilustración 27 Resultados clustering K-means – Datos semestral del docente-estudiante-asignatura, k=4. Elaboración propia con Rapidminer

De este resultado se puede resaltar los siguientes puntos:

- *El clúster ,1 es el que* agrupa la mayor parte de los registros analizados. Las características que se resaltan de estos registros con respecto al total analizado es que:
    - Poseen notas finales cercadas a 31.1/40, superiores al mínimo de aprobación que es de 28 puntos, por lo cual estos estudiantes aprueban la materia.
    - Poseen notas del primer bimestre cercadas a 15.3/20, superiores al mínimo puntaje (14 puntos), que se debe obtener en el primer bimestre, para poder aprobar la asignatura.

La nota del primer bimestre incide en la nota semestral, por lo cual se puede definir que si un estudiante tiene una alta nota en el primer bimestre, tendrá más seguridad de que la nota semestral será también alta, obsérvese la Ilustración 28.

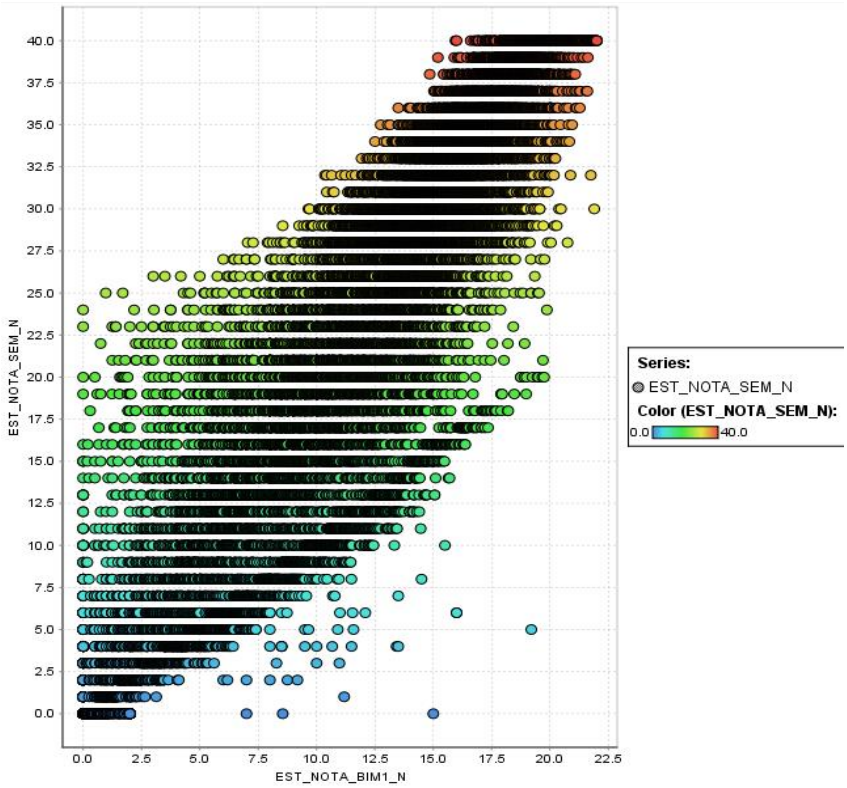
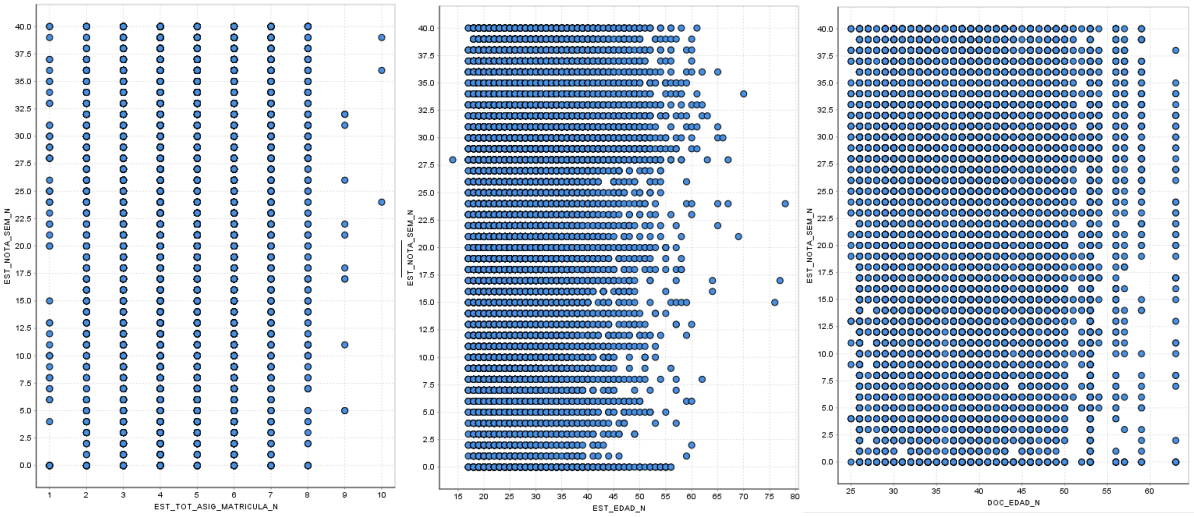


Ilustración 28 Nota del primer bimestre frente a la nota semestral. Elaboración propia con Rapidminer.

- En todos los clústers generados, se observa que el campo *total de asignaturas en las que se matricula el estudiante en el semestre (EST_TOT_ASIG_MATRICULA_N)* es similar, por lo cual se detecta que, esta variable no tiene mayor incidencia en la nota final que obtiene el estudiante. De igual forma sucede con las variables edad del docente *(DOC_EDAD_N)* y del estudiante *(EST_EDAD_N)*.
  Por lo tanto estas tres variables no serán utilizadas en la fase de modelado.



a) Total asignaturas matriculas *b) Edad del estudiante* *c) Edad del docente*

Ilustración 29 Exploración manual de las variables a, b y c, frente a la nota semestral final que ha obtenido el estudiante en algún semestre. Elaboración propia con Rapidminer

- Se observa que aunque no existe una correlación entre e*l número de asignaturas reprobadas en periodos anteriores (EST_TOT_ASIG_REPROB_PER_ANT)* y la *nota que obtiene el estudiante en el semestre (EST_NOTA_SEM_N)*, los estudiantes que tienen alto número de asignaturas reprobadas, tienen también notas semestrales bajas, menores a la nota mínima (28 puntos), obsérvese el cuadrante 2 y 4 de la Ilustración 30.

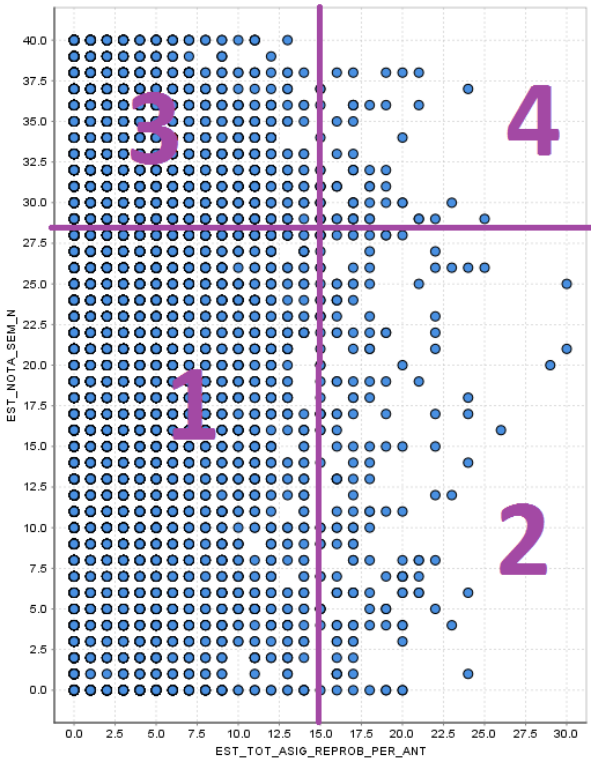


Ilustración 30 Total de asignaturas reprobadas en periodos anteriores frente a la nota obtenida en el semestre. Elaboración propia con Rapidminer

- Al comparar el ciclo del estudiante y el ciclo de la asignatura, se observa que existen pocos casos en los que un estudiante de un ciclo superior tome una asignatura de un ciclo menor. Por lo tanto tienen una correlación positiva en donde, si la asignatura es de un ciclo menor, el estudiante pertenece a un ciclo menor, y viceversa.

Se observa también que cuando un estudiante se matricula en una asignatura de mayor ciclo, que en el que se encuentra, la nota que obtiene es menor que, cuando el ciclo de la asignatura es igual al ciclo del estudiante, observe la Ilustración 31.

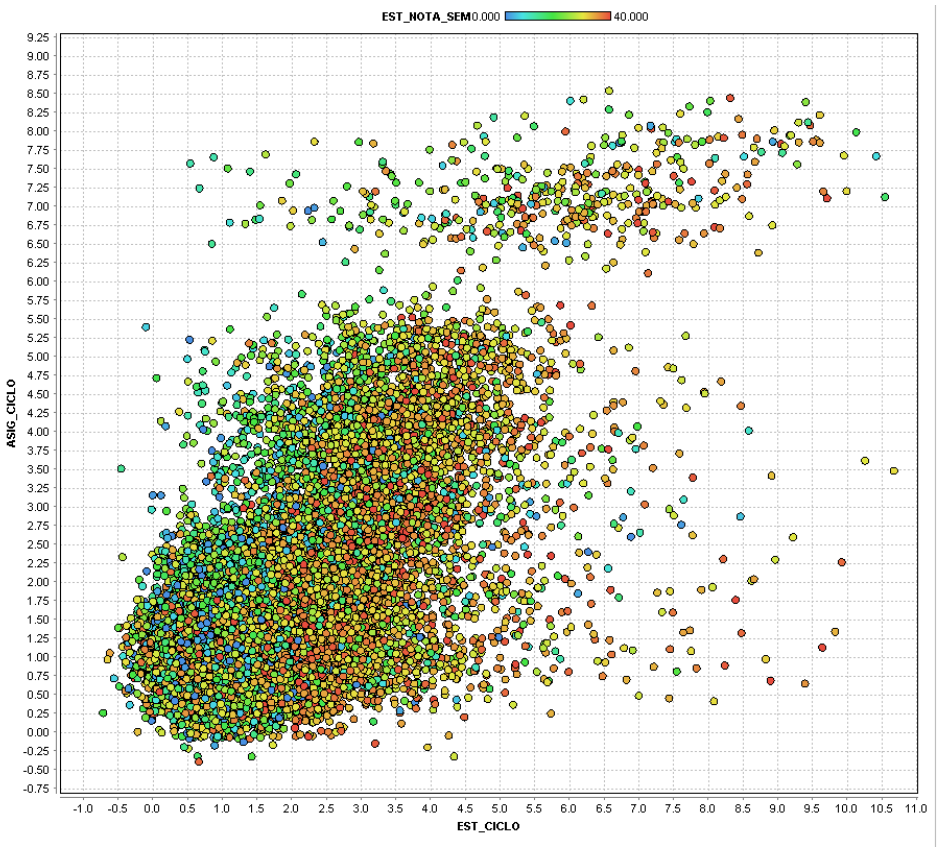


Ilustración 31 Ciclo del estudiante frente al ciclo de la asignatura, indicando la nota semestral obtenida. Elaboración propia con Rapidminer

En la Ilustración 32 se observa que, existe una correlación negativa entre el campo *tasa de reprobación del estudiante con el ciclo del estudiante*, esto es que entre mayor sea el ciclo del estudiante menor es su tasa de reprobación.

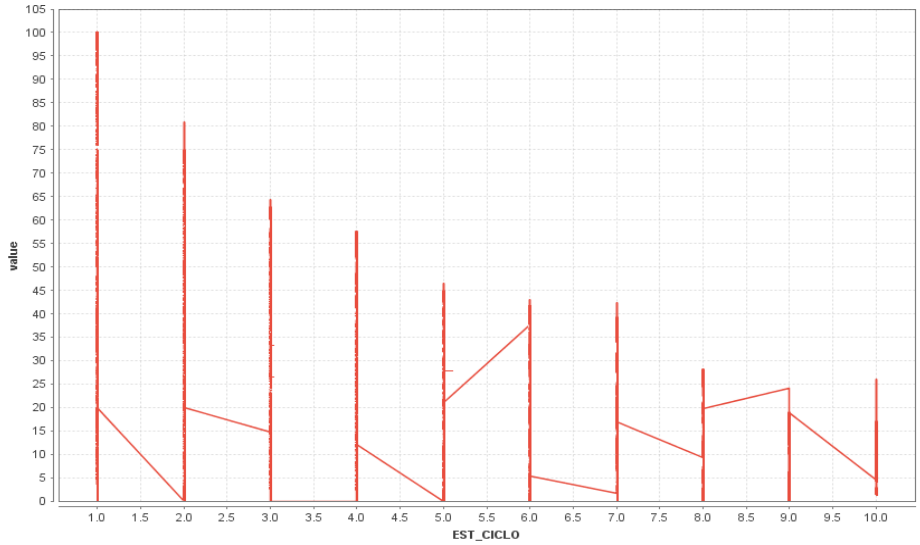


Ilustración 32 Ciclo del estudiante frente a su tasa de reprobación acumulada. Elaboración propia con Rapidminer

En este capítulo se ha logrado brindar un conocimiento sobre la situación actual de la institución, en referencia a la reprobación estudiantil, para resaltar la importancia del análisis de este problema académico.

Además en este capítulo se ha realizado, todo el proceso de *ETL* (extracción, transformación y limpieza) de los datos a utilizar para la construcción de los modelos, así también se ha realizado una exploración de los datos obtenidos, con la finalidad de:

- *Tener mayor conocimiento de los campos y datos, así como de su relación.*
- *Eliminar campos con poco o ninguna influencia en el problema de análisis.*
- *Seleccionar los campos a utilizar en la fase de modelado.*

En el capítulo siguiente, se inicia con el desarrollo y la comparación de los modelos generados con las técnicas de clasificación y asociación, utilizando los datos seleccionados en este capítulo.

5 CAPÍTULO V: DESARROLLO DE LA COMPARATIVA

5.1 Fase 4: Modelado

Esta fase comprende los pasos desde la selección de la técnica de minería de datos hasta la obtención del modelo y su validación. Los campos o variables a utilizarse en esta fase, así como la técnica de minería en donde se usaran se presentaron en la Tabla 12.

Uno de los objetivos de la investigación, es la comparación de dos técnicas de minería de datos y de sus resultados, los criterios de evaluación para seleccionar la técnica y herramienta de minería de datos se listan en la Tabla 13.

Tabla 13 Criterios de evaluación para seleccionar herramientas y técnicas de minería de datos. Elaboración propia.

Criterio de evaluación	Enfocado en la evaluación de	Asociación	Clasificación
Permita detectar patrones en los datos que se poseen	Técnica de MD	Si	Si
Permita trabajar con varias columnas o atributos	Técnica de MD	Si	Si
Pueda utilizarse para pequeñas y grandes cantidades de datos	Técnica de MD	Si	Si
Pueda ser utilizada en entornos educativos	Técnica de MD	Si	Si
Criterio de evaluación	**Enfocado en la evaluación de**	**Rapidminer**	**Weka**
Posee los algoritmos de MD que sean seleccionados en esta investigación	Herramienta de MD	Si. Además contiene una extensión de WEKA	Si
No sólo se enfoque en la fase de Modelado, sino brinde ayuda en las fases de preparación (limpieza, transformación, etc.) y comprensión (exploración) de los datos	Herramienta de MD	Si	No
Pueda ser utilizada en entornos educativos	Herramienta de MD	Si	Si

De cada técnica de *MD* se ha seleccionado los algoritmos básicos:

- Asociación: Reglas de asociación *APRIORI y FP-GROWTH*
- Clasificación: Árboles de decisión *ID3 y CHAID*

Como herramienta de MD se ha seleccionado *Rapidminer* en su versión 7.6, con licencia educativa, por el cumplimiento de los criterios de evaluación mencionados en la Tabla 13.

5.1.1 Construcción y evaluación de los modelos

A continuación se describen los procesos o flujos desarrollados para ejecutar los algoritmos seleccionados, los mismos que fueron construidos con la herramienta Rapidminer. También

se presenta el análisis de rendimiento de cada algoritmo, así como el análisis de los resultados que estos generan. La finalidad es seleccionar el o los algoritmos a utilizar, así como la parametrización más idónea para el problema analizado y que pueden ser implementados en la institución educativa.

Ambas técnicas están enfocadas en detectar de forma temprana a los estudiantes que tengan riesgos de reprobar este tipo de asignaturas. La finalidad es alertar a los docentes y autoridades sobre los posibles reprobados, para que puedan realizar cambios y tomar decisiones a tiempo, que ayuden al estudiante en su rendimiento académico, y por lo tanto a evitar su reprobación.

5.1.1.1 Reglas de asociación

El flujo desarrollado en la herramienta Rapidminer, que sirve para la obtención de reglas de asociación, se presenta en la Ilustración 33. Este flujo contiene los dos algoritmos seleccionados: *Apriori y FP-Growth.*

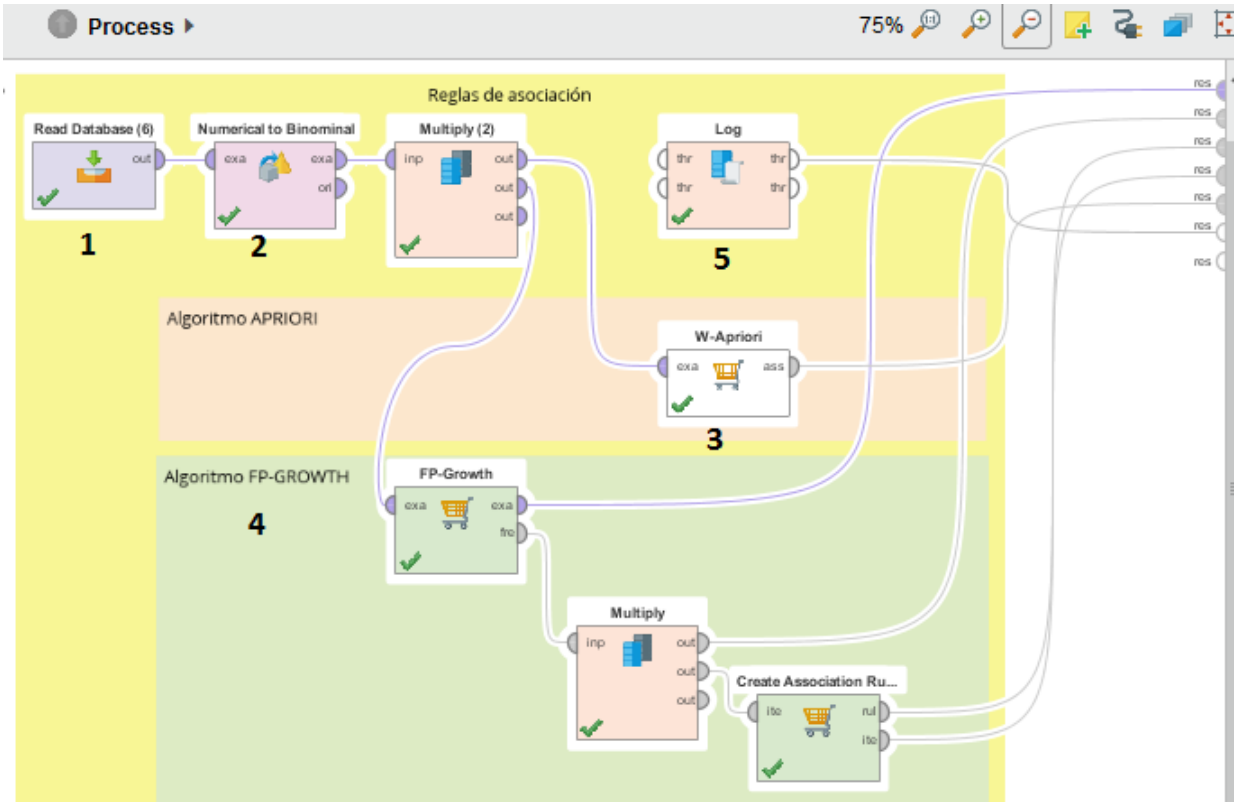


Ilustración 33 Proceso de construcción de reglas de asociación con la herramienta RapidMiner. Elaboración propia.

A continuación se resumen los pasos que forman parte del flujo o proceso desarrollado:

1. ***Read Database.-*** operador para ejecutar la sentencia *select* y obtener los datos desde la base de datos centralizada – repositorio 4.
2. ***Numerical to binominal.-*** operador que sirve para realizar la transformación de valores numéricos a binominales. Se realiza este paso, porque los algoritmos utilizados requieren que los campos sean de tipo binomial, que significa que los valores de los campos tengan sólo dos valores: *true o false.*
3. ***W-Apriori.-*** operador tipo *extensión*, permite ejecutar el algoritmo *Apriori* de WEKA. Es una de potencialidades de la herramienta Rapidminer. En este operador se deben establecer el *mínimo de soporte y de confianza* que se utilizarán para la generación de las reglas de asociación.
4. ***Fp-Growth.-*** para la obtención de los resultados que genere el algoritmo *FP-Growth*, se utilizan dos operadores. El primero operador *FP-Growth* permite obtener los ítems frecuentes, en este operador se puede configurar el valor mínimo de soporte que se utilizará para la detección de *itemsets* frecuentes. El segundo operador es *Create Association Rules*, el cual se encarga de generar las reglas de asociación, utilizando los itemsets frecuentes generados por el operador *FP-Growth*. En este se puede configurar el valor mínimo del umbral de confianza, que se desea que cumplan las reglas a generarse.
5. ***Log.-*** este operador sirve para calcular y presentar los tiempos de ejecución de los algoritmos *FP-Growth* y *Apriori*.

5.1.1.1.1 Análisis de rendimiento de FP-Growth y Apriori frente a las reglas de asociación que generan.

Para evaluar el rendimiento de *FP-Growth* y *Apriori* se obtiene el tiempo de ejecución de cada algoritmo, aplicando diversas cantidades de registros o transacciones, varios tamaños de ventana, y diferentes valores para los umbrales de soporte y confianza. En la Ilustración 34 e Ilustración 35, se presentan los tiempos de ejecución obtenidos.

De estos resultados se resalta lo siguiente:

- Con el 50% y 100% de los registros o transacciones que se poseen para el análisis de datos, y con los tamaños de ventana de 16 y 20, el algoritmo *Apriori tiene mayor tiempo de ejecución que el algoritmo FP-Growth.* Por lo cual se puede concluir que en el algoritmo *FP-Growth* es más rápido y escalable que el algoritmo *Apriori.*

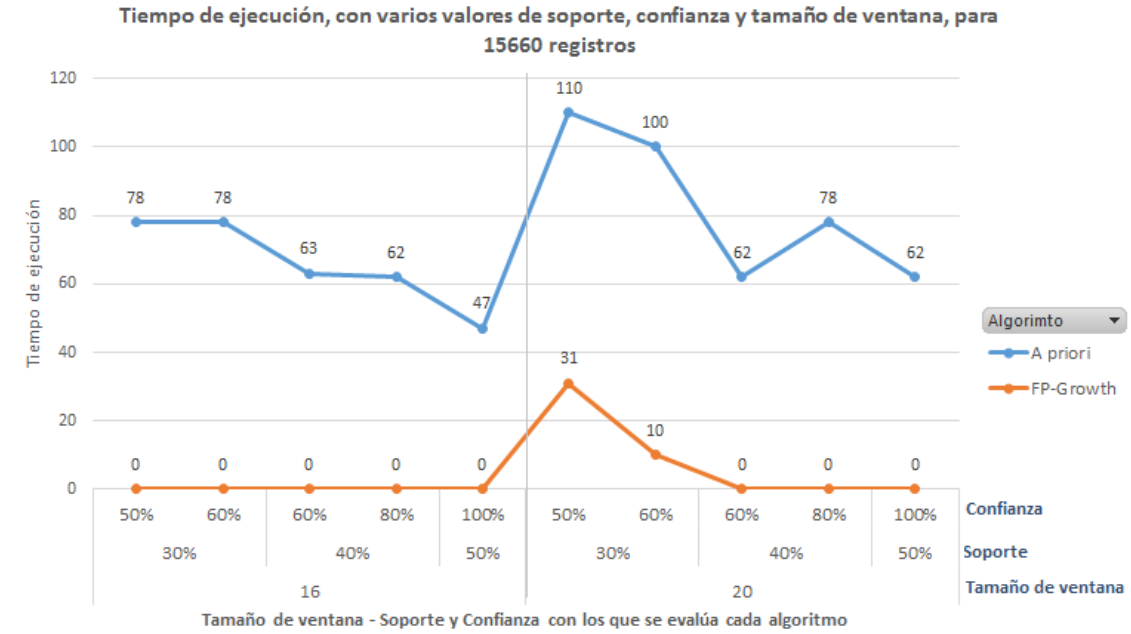


Ilustración 34 Tiempo de ejecución de FP-Growth y Apriori con 15660 registros y varios valores de soporte, confianza y tamaño de ventana. Elaboración propia.

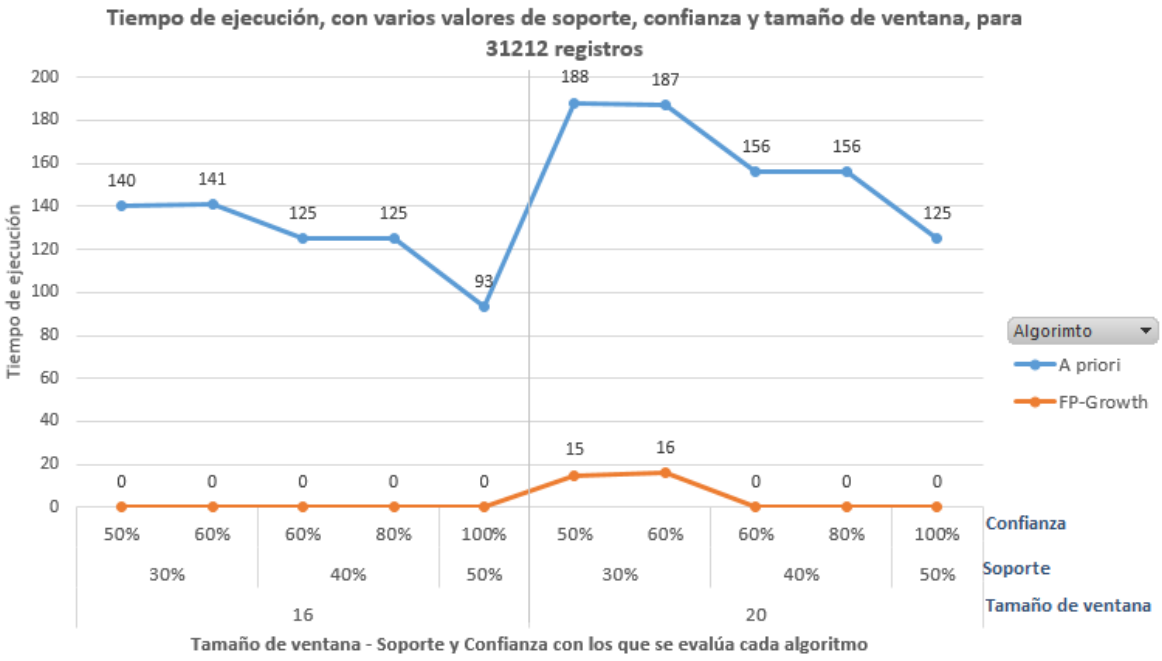


Ilustración 35 Tiempo de ejecución de FP-Growth y Apriori con 31212 registros y varios valores de soporte, confianza y tamaño de ventana. Elaboración propia.

En cuanto al número de reglas generadas por los algoritmos *Apriori* y *FP-Growth*, se obtienen los resultados que se muestran en la Ilustración 36 e Ilustración 37. De igual manera se han aplicado diversas cantidades de registros o transacciones, varios tamaños de ventana, y diferentes valores para los umbrales de soporte y confianza.

De los resultados obtenidos se puede resaltar lo siguiente:

- Considerando las variaciones del tamaño de ventana y de los umbrales de soporte y confianza, se observa que al aumentar el número de registros o transacciones, el número de reglas que genera el algoritmo *FP-Growth* disminuye. Contrariamente el algoritmo *Apriori* mantienen el mismo número de reglas sin verse afectado por la variación de estos parámetros.

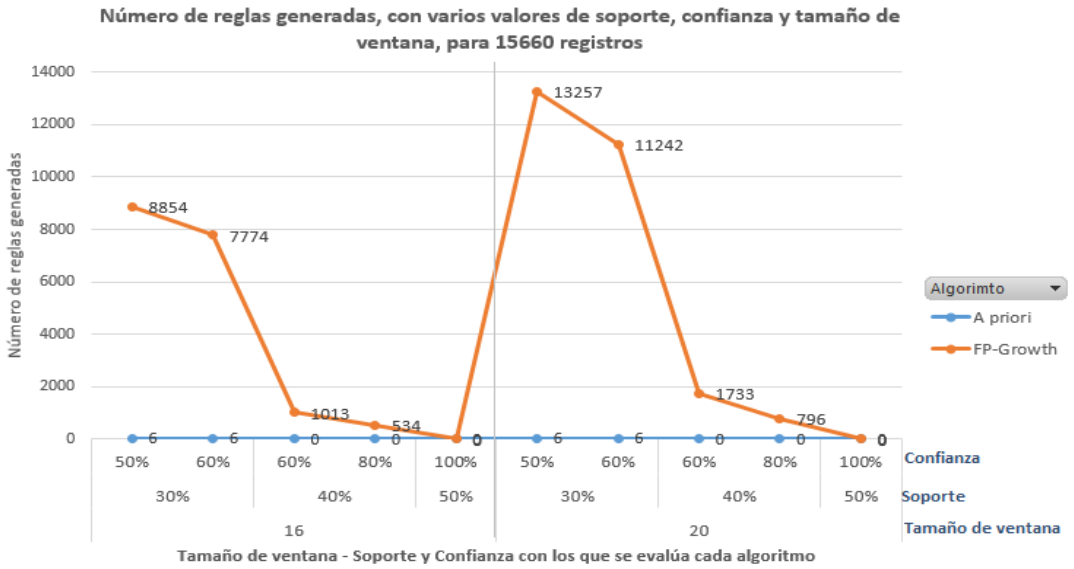


Ilustración 36 Número de reglas generadas con FP-Growth y A priori, con diversos valores de soporte, confianza y tamaño de ventana con 15660 registros. Elaboración propia, utilizando Rapidminer.

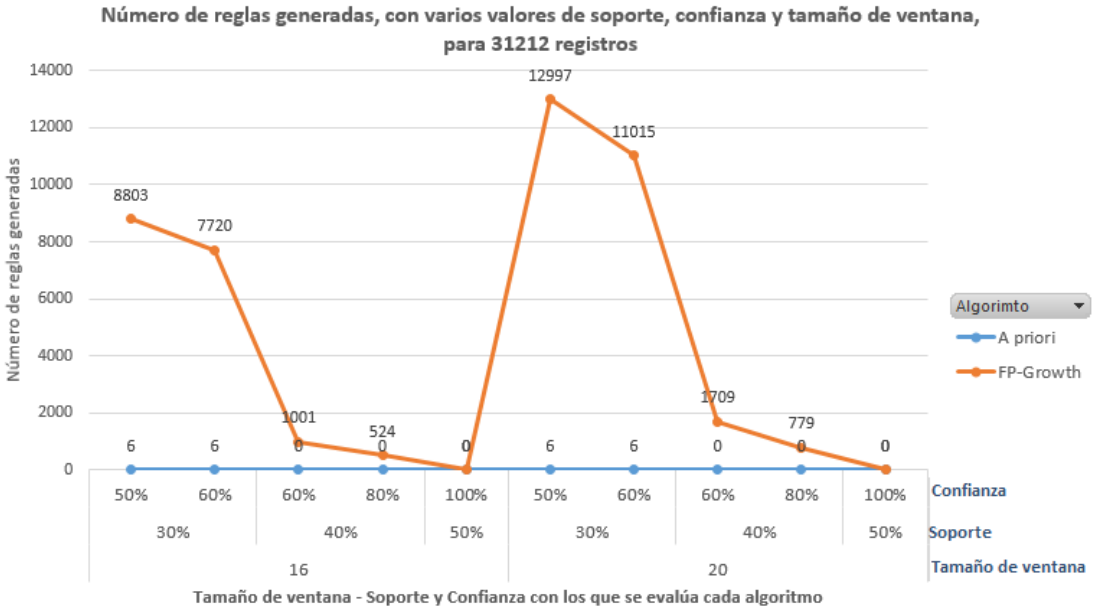


Ilustración 37 Número de reglas generadas con FP-Growth y A priori, con diversos valores de soporte, confianza y tamaño de ventana con 31212 registros. Elaboración propia, utilizando Rapidminer.

Para encontrar las mejores de reglas de asociación, así como el algoritmo a utilizar, se han definido las siguientes consideraciones:

- ***Consideración 1.-*** Reglas de campos afines. Quitar campos que son afines entre ellos. Por ejemplo se han eliminado los campos *total de foros, total de cuestionarios y total de chats académicos del docente*, debido a que el *total de foros, total de cuestionarios y total de chats académicos del estudiante,* son el resultado de estos, es quiere decir que, si un estudiante realiza cualquier actividad de estos tres tipos, es por el docente la propuso, por lo tanto no puede existir un actividad de un estudiante sino el docente no la planteó.

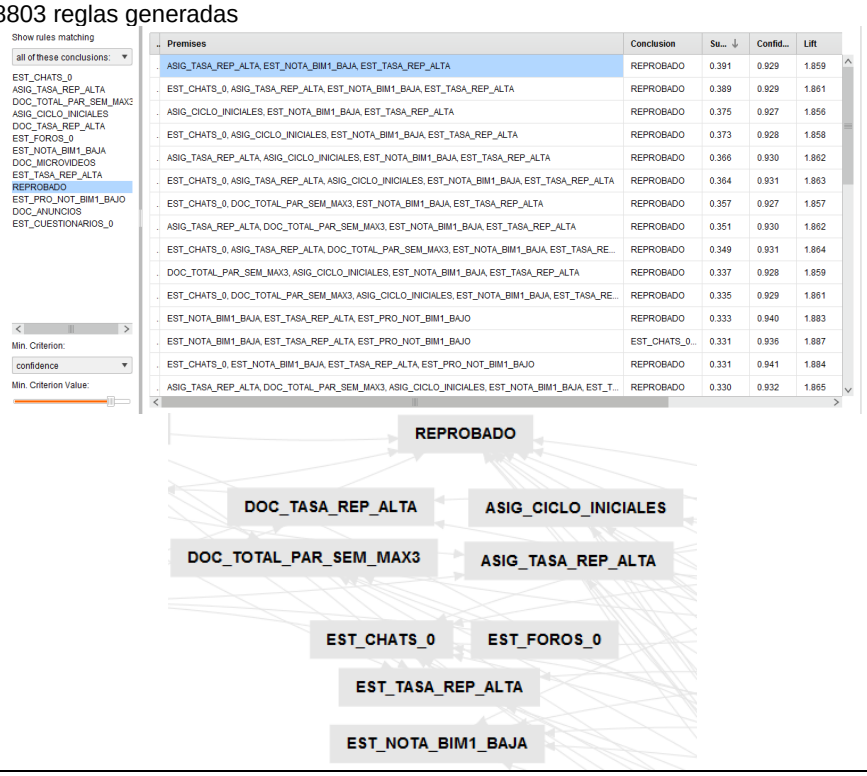
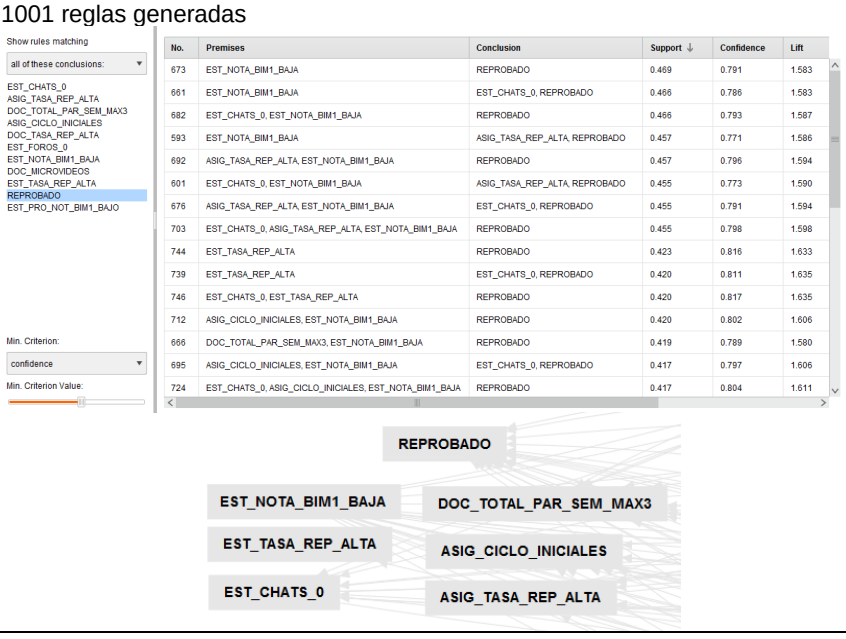
  En esta investigación se plantea evaluar si la realización o no de la actividad, por parte del estudiante le afecta o no en su rendimiento. Puesto que, si el docente no plantea ninguna actividad, el estudiante tampoco la realizará, y si el docente plantea la actividad pero el estudiante no la realiza tendría el mismo resultado: *estudiante sin actividad realizada*
  Con esta consideración se evita tener reglas como:
  {DOC_TOTAL_CUESTIONARIOS = true → EST_TOTAL_CUESTIONARIOS =true}
  {DOC_TOTAL_CUESTIONARIOS = false → EST_TOTAL_CUESTIONARIOS = false}

- ***Consideración 2.-*** Reglas de asociación que permitan crear alertas preventivas, las cuales puedan ser comprendidas con facilidad por los docentes. Estas alertas deben indicar el riesgo de reprobación de cada estudiante.
- ***Consideración 3.-*** Reglas de asociación que brinden conocimiento nuevo, que pueda ser reportado a las autoridades, e implementado en la institución.
- ***Consideración 4.-*** Algoritmo cuyo tiempo de ejecución de análisis sea el mínimo posible.
- ***Consideración 5.-*** Algoritmo permita aumentar el tamaño de ventana, y de registros o transacciones, sin que esto afecte contraproducentemente en su rendimiento y calidad de reglas generadas.

Estas consideraciones también influyen en los valores de los umbrales de soporte, confianza y el tamaño de la ventana a utilizar.

En la Tabla 14, se presentan las reglas de asociación generadas por los algoritmos *Apriori* y *FP-Growth*, con diferentes valores de soporte y confianza. Así mismo en esta tabla se han colocado dos columnas, que hacen referencia a dos de las cinco consideraciones antes mencionadas, las cuales permiten seleccionar las mejores reglas de asociación.

Tabla 14 Reglas de asociación generadas por los algoritmos Apriori y FP-Growth. Elaboración propia.

Tamaño de ventana: 16 y Cantidad de registros: 31212				Consideraciones	
Algoritmo	**Min. Soporte**	**Min. Confianza**	**Resultado** Se resaltan las reglas que tienen como consecuente al estado *reprobado = true*	**Consideración 2.- Permite alertar tempranamente**	**Consideración 3.- Conocimiento nuevo**
Apriori	**0.3**	**0.5**	Minimum support: 0.3 (4698 instances) Minimum metric <confidence>: 0.5 Number of cycles performed: 14 Generated sets of large itemsets: Size of set of large itemsets L(1): 9 Size of set of large itemsets L(2): 3 Best rules found: 1. EST_CUESTIONARIOS_0=true 5291 ==> DOC_ANUNCIOS=true 5244 conf:(0.99) 2. DOC_ANUNCIOS=true 5904 ==> EST_CUESTIONARIOS_0=true 5244 conf:(0.89) 3. EST_PRO_NOT_BIM1_BAJO=true 7292 ==> REPROBADO=true 5889 conf:(0.81) 4. EST_NOTA_BIM1_BAJA=false 6372 ==> EST_TASA_REP_ALTA=false 5019 conf:(0.79) 5. REPROBADO=true 7821 ==> EST_PRO_NOT_BIM1_BAJO=true 5889 conf:(0.75) 6. EST_TASA_REP_ALTA=false 7544 ==> EST_NOTA_BIM1_BAJA=false 5019 conf:(0.67)	**SI**, Se puede indicar al docente que estudiantes tienen bajos promedios en sus notas bimestrales	**NO**
	0.4	**0.5**	Sólo generó reglas con soporte del 30%		
FP-Growth	**0.3**	**0.5**	8803 reglas generadas  	**SI**	**SI**
	0.4	**0.5**	1001 reglas generadas  	**SI**	**SI**

En cuanto al **rendimiento**, se observa que sin importar el tamaño de registros o de la ventana y de los valores de los umbrales de soporte y confianza, *FP-Growth* es más rápido y escalable que el algoritmo *Apriori.*

En cuanto a la **cantidad** de las reglas de asociación que genera cada algoritmo, se observa que al aumentar el valor de los umbrales, el algoritmo *Apriori* no encuentra ninguna regla de asociación para el conjunto de datos utilizados, mientras que el algoritmo *FP-Growth* va reduciendo la cantidad de reglas y mejorando la calidad de las mismas.

En cuanto a la **calidad** de las reglas de asociación que genera cada algoritmo, se ha encontrado que las pocas reglas que genera el algoritmo *Apriori* no cumplen con la tercera consideración establecida, que mencionada que se desea obtener: *"reglas de asociación que brinden conocimiento nuevo".* En caso contrario el algoritmo *FP-Growth* si genera reglas que cumplan con esta consideración.

Ante los resultados obtenidos, así como a las consideraciones expuestas y analizadas anteriormente, se determina utilizar el algoritmo *FP-Growth*, con los siguientes parámetros:

- **Cantidad de registros:** 31212
- **Tamaño de ventana:** 16
- **Mínimo de soporte:** 0.4
- **Mínimo de confianza:** 0.7

Reglas de asociación seleccionadas, se presentan en la Tabla 15 sólo las reglas que tienen como consecuente al estado reprobado = true.

Tabla 15 Reglas de asociación seleccionadas generadas por el algoritmo FP-Growth, sólo se han seleccionada las reglas relacionadas a la reprobación. Elaboración propia

Premisa - antecedente				Conclusión - consecuente	Soporte	Confianza
ASIG_CICLO_INICIALES	EST_NOTA_BIM1_BAJA			REPROBADO	0,466	0,793
ASIG_TASA_REP_ALTA	EST_NOTA_BIM1_BAJA			REPROBADO	0,457	0,797
ASIG_TASA_REP_ALTA	EST_TASA_REP_ALTA			REPROBADO	0,454	0,799
ASIG_TASA_REP_ALTA	ASIG_CICLO_INICIALES	EST_NOTA_BIM1_BAJA		REPROBADO	0,422	0,817
ASIG_TASA_REP_ALTA	DOC_TOTAL_PAR_SEM_MAX3	EST_NOTA_BIM1_BAJA		REPROBADO	0,419	0,801
DOC_TOTAL_PAR_SEM_MAX3	EST_NOTA_BIM1_BAJA			REPROBADO	0,418	0,787
EST_CHATS_0	EST_NOTA_BIM1_BAJA			REPROBADO	0,417	0,804
EST_CHATS_0	ASIG_TASA_REP_ALTA	EST_NOTA_BIM1_BAJA		REPROBADO	0,415	0,79
EST_CHATS_0	EST_TASA_REP_ALTA			REPROBADO	0,413	0,827
EST_CHATS_0	ASIG_CICLO_INICIALES	EST_NOTA_BIM1_BAJA		REPROBADO	0,411	0,828
EST_CHATS_0	DOC_TOTAL_PAR_SEM_MAX3	EST_NOTA_BIM1_BAJA		REPROBADO	0,408	0,809
EST_CHATS_0	ASIG_TASA_REP_ALTA	EST_TASA_REP_ALTA		REPROBADO	0,407	0,794
EST_CHATS_0	ASIG_TASA_REP_ALTA	ASIG_CICLO_INICIALES	EST_NOTA_BIM1_BAJA	REPROBADO	0,406	0,812
EST_CHATS_0	ASIG_TASA_REP_ALTA	DOC_TOTAL_PAR_SEM_MAX3	EST_NOTA_BIM1_BAJA	REPROBADO	0,405	0,797

Estas reglas demuestran que existe una asociación entre los campos que se muestran en la Ilustración 38, con referencia a la reprobación de un estudiante.

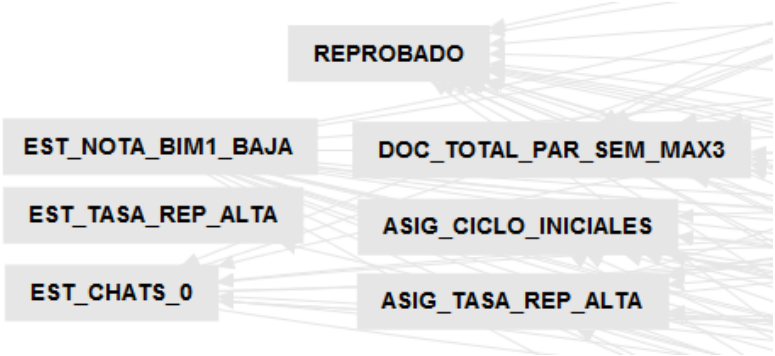


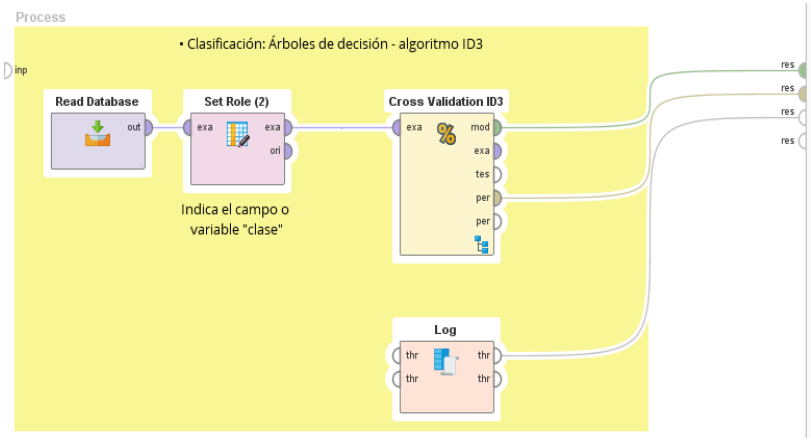
Ilustración 38 Factores detectados en las reglas de asociación. Elaboración propia, utilizando la herramienta Rapidminer

5.1.1.2 Árboles de decisión

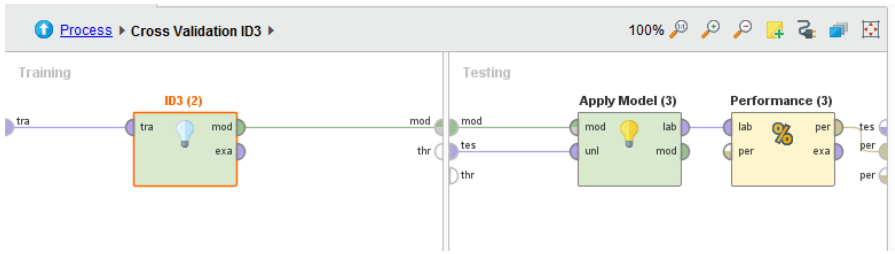
En cuanto a los algoritmos de árboles decisión, se ha construido, de igual manera en la herramienta Rapidminer, el proceso para su generación.

En este caso se han creado dos flujos separados, uno para el algoritmo ID3 y otro para el algoritmo CHAID. Cada flujo contiene un subproceso, que facilita realizar la validación del modelo utilizando validación cruzada, la cual permite evaluar la precisión de los modelos de clasificación generados. Obsérvese la Ilustración 39 e Ilustración 40.

Los campos de entrada de estos dos algoritmos no deben ser de tipo numérico.

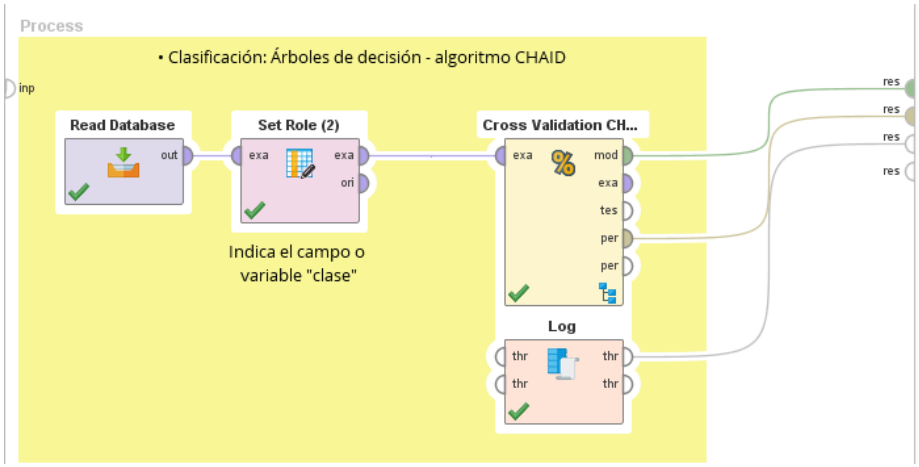


a) Proceso general algoritmo ID3

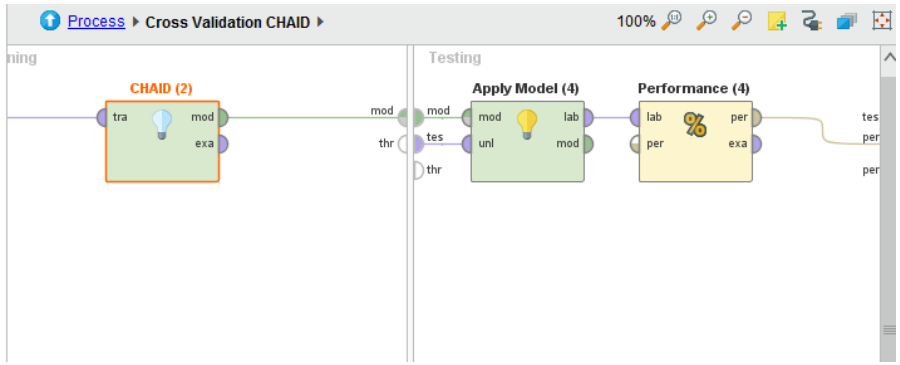


b) Subproceso algoritmo ID3

Ilustración 39 Proceso para la construcción de árboles de decisión, algoritmo ID3, con validación cruzada.



a) Proceso general algoritmo CHAID



b) Subproceso algoritmo CHAID

Ilustración 40 Proceso para la construcción de árboles de decisión, algoritmo CHAID, con validación cruzada. Elaboración propia utilizando Rapidminer

A continuación se resumen los pasos que forman parte de ambos procesos desarrollados. Como se puede observar en la Ilustración 39 e Ilustración 40, estos procesos son similares, sólo difieren uno del otro por el algoritmo que se utiliza en el flujo:

1. ***Read Database.-*** operador para ejecutar la sentencia *select* y obtener los datos desde la base de datos centralizada – repositorio 4.
2. ***Set role.***- al utilizar árboles de decisión, es requerido definir el campo que será catalogado como la *clase*, este operador es el encargado de permitir al usuario indicar que campo será.
3. ***Cross validation ID3 y CHAID.-*** método que permite obtener los índices que informaran sobre la precisión de los modelos de clasificación que se generen.
4. ***ID3.-*** operador que representa al algoritmo *ID3*, en este se puede cambiar los parámetros por defecto como *información de ganancia,* etc.
5. ***CHAID.-*** operador que representa al algoritmo *CHAID*, en este se puede cambiar los parámetros por defecto *información de ganancia*, etc.
6. ***Apply model.-*** operador que permite la construcción del modelo generado
7. ***Perfomance.-*** genera la matriz de confusión, así como otros índices que informan sobre la calidad del modelo de clasificación.
8. ***Log.-*** este operador sirve para calcular y presentar los tiempos de ejecución de los algoritmos *ID3 y CHAID*

5.1.1.2.1 Análisis de rendimiento de los algoritmos ID3 y CHAID, frente a las reglas de asociación que generan.

Para evaluar el rendimiento de estos algoritmos, se obtiene el tiempo de ejecución de cada algoritmo, utilizando diversas cantidades de registros o transacciones. En la Ilustración 41 se presentan los tiempos de ejecución obtenidos, de estos resultados se resalta lo siguiente:

- Con un tamaño de datos aproximado del 50%, 75% y 100% del total de registros que se poseen para el análisis de datos, el algoritmo *ID3 tiene mayor tiempo de ejecución que el algoritmo CHAID.* Por lo cual se puede concluir que en el algoritmo *CHAID,* es más rápido y escalable que el algoritmo *ID3.* En ambos se utiliza la misma cantidad de campos para realizar la clasificación.

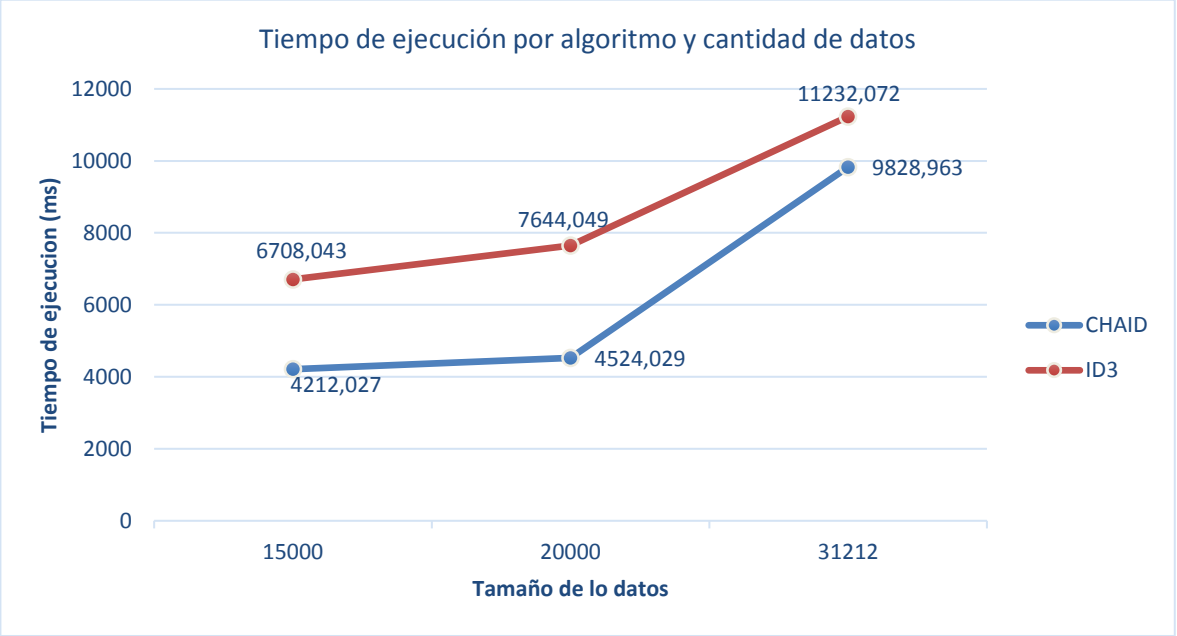


Ilustración 41 Tiempo de ejecución de ID3 y CHAID con diversos tamaños de datos. Elaboración propia.

Los árboles de decisión se componente de ramificaciones o caminos, estos son considerados como reglas de clasificación. En los resultados generados, se observó que el algoritmo *CHAID* generó menos reglas de clasificación que el algoritmo *ID3*. El número de reglas no fue afectado por la variación del tamaño de los datos utilizados.

Para evaluar estas reglas de clasificación que han generado los algoritmos, así como determinar al algoritmo más idóneo, se definen las siguientes consideraciones:

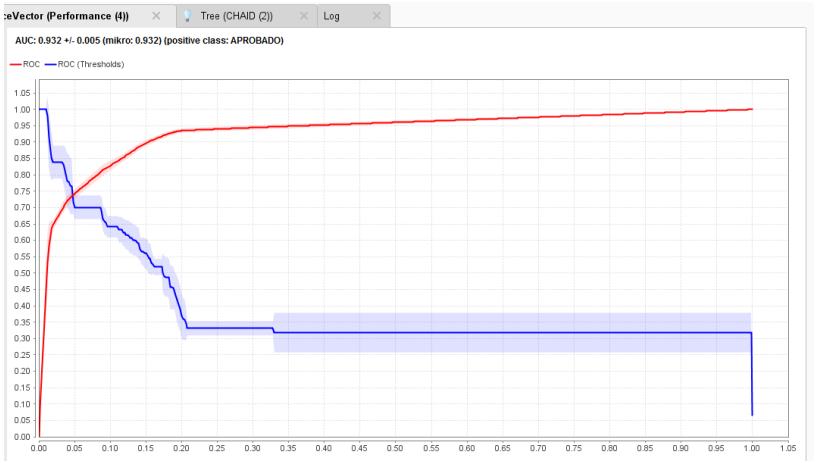
- ***Consideración 1 - Precisión de la clasificación.-*** alta calidad de las reglas del modelo de clasificación. Se deben revisar los porcentajes de casos clasificados correcta e incorrectamente. Utilizar la *matriz de confusión*, e índice de *accuracy.* Este índice es fiable, debido a que los casos a utilizar de estudiantes aprobados y reprobados están equilibrados.
- ***Consideración 2 - Interpretable.-*** Reglas de clasificación que permitan crear alertas preventivas, las cuales puedan ser comprendidas con facilidad por los docentes. Estas alertas deben indicar el riesgo de reprobación de cada estudiante.
- ***Consideración 3 – Conocimiento nuevo***.- Reglas de clasificación que brinden conocimiento nuevo, que pueda ser reportado a las autoridades, e implementado en la institución.
- ***Consideración 4 - Eficiencia.-*** Algoritmo cuyo tiempo de ejecución sea el mínimo posible.

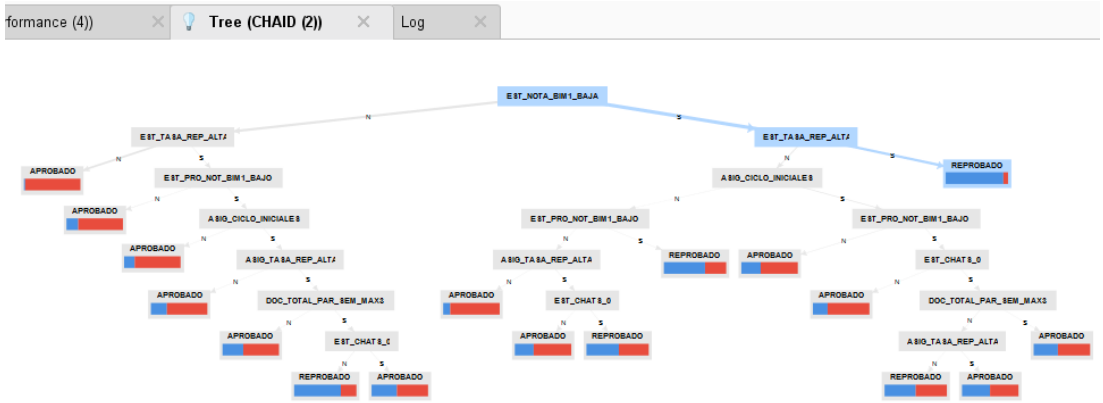
- ***Consideración 5 - Escalabilidad.-*** Algoritmo permita aumentar el tamaño de los datos registros sin que esto afecte contraproducentemente en su rendimiento y calidad de reglas generadas.

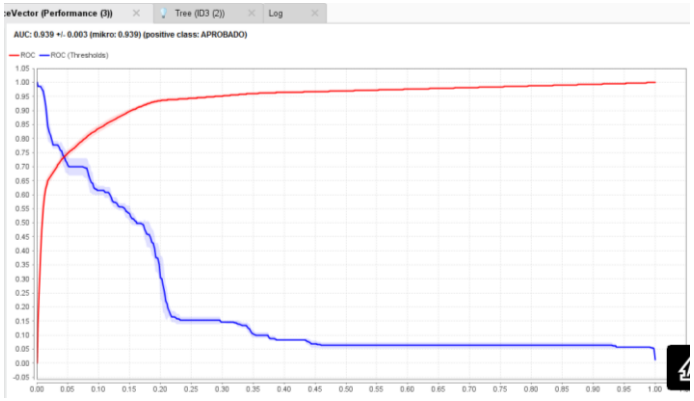
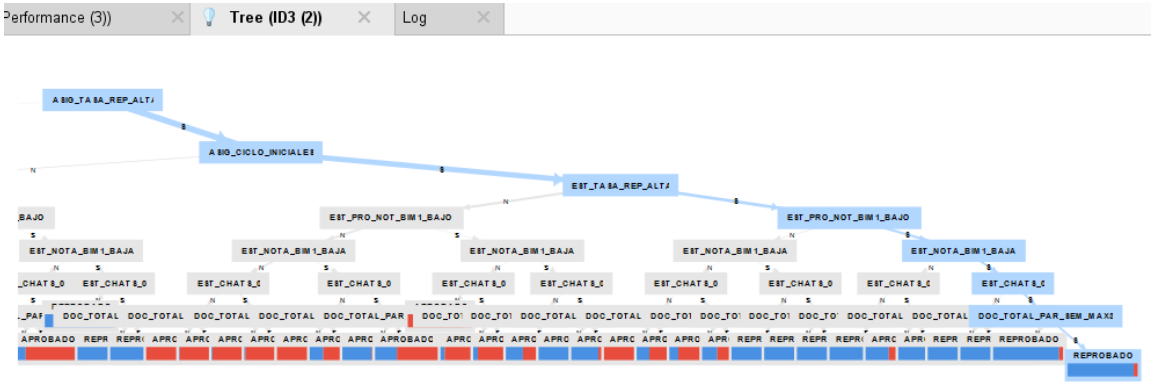
En la Tabla 16, se presentan los resultados de la ejecución de los algoritmos ID3 y CHAID. Además se presenta la matriz de confusión, la curva ROC y el árbol de decisión que se genera con el 100% de los datos que se poseen. También se ha colocado en forma de columna las consideraciones de evaluación de los resultados que se utilizarán y que fueron mencionadas anteriormente.

Los campos que se utilizaron como entrada en la ejecución de los algoritmos de clasificación, son aquellos que fueron detectados por el algoritmo *FP-Growth* y que están relacionados a las reglas de asociación seleccionadas en la sección anterior.

Tabla 16 Árboles de decisión generados por los algoritmos CHAID e ID3. Elaboración propia.

<table>
<tr><th colspan="2">Tamaño de ventana: 8 y Cantidad de registros: 31212</th><th colspan="5">Consideraciones</th></tr>
<tr><th>Algoritmo</th><th>Resultado</th><th>1) Precisión de la clasificación</th><th>2) Interpretable</th><th>3) Conocimiento nuevo</th><th>4 - Eficiencia</th><th>5) Escalabilidad</th></tr>
<tr><td>CHAID</td><td>Se observa que el porcentaje de precisión de la clasificación de este algoritmo para la clase Reprobado es del 89.33% y para la clase Aprobado es de 85.60%. Y de la precisión del algoritmo es del 87.37%
accuracy: 87.37% +/- 0.44% (mikro: 87.37%)
<table><tr><td></td><td>true REPROBADO</td><td>true APROBADO</td><td>class precision</td></tr><tr><td>pred. REPROBADO</td><td>13268</td><td>1584</td><td>89.33%</td></tr><tr><td>pred. APROBADO</td><td>2372</td><td>14097</td><td>85.60%</td></tr><tr><td>class recall</td><td>84.83%</td><td>89.90%</td><td></td></tr></table> 


</td><td>Si, el tamaño de los datos no ha afectado el porcentaje de precisión de la clasificación</td><td>Si</td><td>Si</td><td colspan="2">Si, el aumento del tamaño de datos, no afecta en el tiempo de ejecución</td></tr>
</table>

ID3	Se observa que el porcentaje de precisión de la clasificación de este algoritmo para la clase Reprobado es del 89.34% y para la clase Aprobado es de 85.60%. Y de la precisión del algoritmo es del 87.37% **accuracy: 87.37% +/- 0.44% (mikro: 87.37%)** 		true REPROBADO	true APROBADO	class precision
pred. REPROBADO	13269	1584	89.34%
pred. APROBADO	2371	14097	85.60%
class recall	84.84%	89.90%	|   	Si, el tamaño de los datos no ha afectado el porcentaje de precisión de la clasificación	Si, aunque genera muchas reglas de clasificación	Si	No, al aumentar el tamaño de datos el tiempo de ejecución incrementa.

En cuanto al **rendimiento**, se observa que sin importar el tamaño de registros *CHAID* es más rápido y escalable que el algoritmo *ID3.*

En cuanto a la **cantidad** de las reglas de clasificación que genera cada algoritmo, se observa que el algoritmo CHAID genera menos reglas que el algoritmo ID3. El algoritmo ID3 genera muchas clasificaciones, lo que dificulta su comprensión y la detección de conocimiento nuevo. Es importante recalcar que para las ejecuciones de ambos algoritmos, se estableció en *0.3* el valor de *ganancia de información.* Por lo cual los árboles de decisión generados por el algoritmo *ID3* fueron más complejos que los generados por el algoritmo *CHAID*.

En cuanto a la **calidad y precisión** ambos algoritmos son altamente precisos, lo respalda el índice de *accuracy* obtenido, la curva de *ROC* y los resultados de la matriz de confusión generados.

Ante los resultados obtenidos, así como a las consideraciones expuestas y analizadas anteriormente, se establece utilizar el algoritmo *CHAID*, con los siguientes parámetros:

- **Cantidad de registros:** 31212
- **Tamaño de ventana:** 8
- **Ganancia de información:** 0.3
- **Reglas de clasificación generadas y árbol de decisión generado**

```
EST_NOTA_BIM1_BAJA = N
|   EST_TASA_REP_ALTA = N: APROBADO {REPROBADO=222, APROBADO=9797}
|   EST_TASA_REP_ALTA = S
|   |   EST_PRO_NOT_BIM1_BAJO = N: APROBADO {REPROBADO=390, APROBADO=1480}
|   |   EST_PRO_NOT_BIM1_BAJO = S
|   |   |   ASIG_CICLO_INICIALES = N: APROBADO {REPROBADO=12, APROBADO=52}
|   |   |   ASIG_CICLO_INICIALES = S
|   |   |   |   ASIG_TASA_REP_ALTA = N: APROBADO {REPROBADO=19, APROBADO=48}
|   |   |   |   ASIG_TASA_REP_ALTA = S
|   |   |   |   |   DOC_TOTAL_PAR_SEM_MAX3 = N: APROBADO {REPROBADO=29, APROBADO=50}
|   |   |   |   |   DOC_TOTAL_PAR_SEM_MAX3 = S
|   |   |   |   |   |   EST_CHATS_0 = N: REPROBADO {REPROBADO=3, APROBADO=1}
|   |   |   |   |   |   EST_CHATS_0 = S: APROBADO {REPROBADO=286, APROBADO=351}
EST_NOTA_BIM1_BAJA = S
|   EST_TASA_REP_ALTA = N
|   |   ASIG_CICLO_INICIALES = N
|   |   |   EST_PRO_NOT_BIM1_BAJO = N
|   |   |   |   ASIG_TASA_REP_ALTA = N: APROBADO {REPROBADO=1, APROBADO=7}
|   |   |   |   ASIG_TASA_REP_ALTA = S
|   |   |   |   |   EST_CHATS_0 = N: APROBADO {REPROBADO=1, APROBADO=2}
|   |   |   |   |   EST_CHATS_0 = S: REPROBADO {REPROBADO=429, APROBADO=396}
|   |   |   EST_PRO_NOT_BIM1_BAJO = S: REPROBADO {REPROBADO=292, APROBADO=150}
|   |   ASIG_CICLO_INICIALES = S
|   |   |   EST_PRO_NOT_BIM1_BAJO = N: APROBADO {REPROBADO=778, APROBADO=1490}
|   |   |   EST_PRO_NOT_BIM1_BAJO = S
|   |   |   |   EST_CHATS_0 = N: APROBADO {REPROBADO=7, APROBADO=20}
|   |   |   |   EST_CHATS_0 = S
|   |   |   |   |   DOC_TOTAL_PAR_SEM_MAX3 = N
|   |   |   |   |   |   ASIG_TASA_REP_ALTA = N: REPROBADO {REPROBADO=1, APROBADO=1}
|   |   |   |   |   |   ASIG_TASA_REP_ALTA = S: APROBADO {REPROBADO=71, APROBADO=72}
|   |   |   |   |   DOC_TOTAL_PAR_SEM_MAX3 = S: APROBADO {REPROBADO=561, APROBADO=745}
|   EST_TASA_REP_ALTA = S: REPROBADO {REPROBADO=12538, APROBADO=1019}
```

Ilustración 42 Árbol de decisión generado con el algoritmo CHAID – formato texto. Elaboración propia utilizando Rapidminer

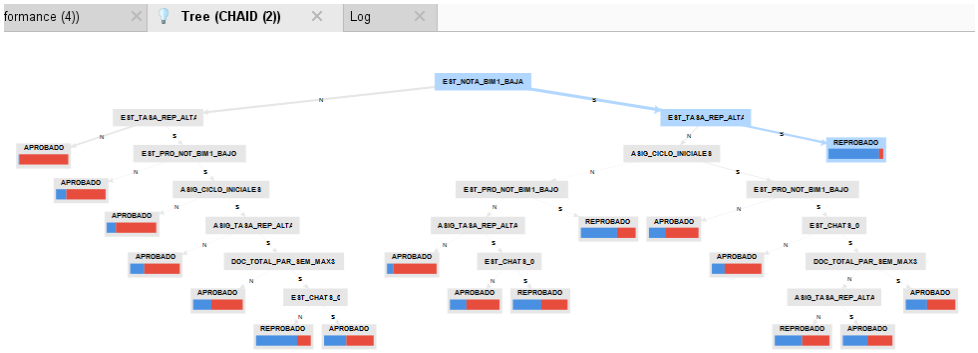


Ilustración 43 Árbol de decisión generado con el algoritmo CHAID – formato gráfico. Elaboración propia utilizando Rapidminer

Los resultados de estos dos algoritmos permitirán crear conocimiento, que se puede mostrar, por ejemplo en un reporte similar al que se presenta en la Tabla 17.

Tabla 17 Ejemplo de presentación del conocimiento obtenido con los modelos generados por las técnicas de minería de datos utilizadas. Elaboración propia.

Periodo académico	2017 - 2			**Docente:**	Docente A				
Asignatura	Asignatura A			**Paralelo**	A				
Estudiante	Factor 1 Tasa de reprobación de la asignatura	Factor 2 Ciclo de la asignatura	Factor 3 Carga académica del docente (Paralelos asignados)	Factor 4 Tasa de reprobación del estudiante	Factor 5 Chats académicos - tutorías en que ha participado el estudiante	Factor 6 Nota primer bimestre	Factor 7 Promedio de notas del primer bimestre	Tiene riesgo de reprobar	Nivel de riesgo de reprobación
Estudiante 1	NO	NO	NO	NO	NO	NO	NO	NO	NINGUNO
Estudiante 2	NO	NO	NO	SI	NO	SI	NO		ALTA
Estudiante 3	SI	SI	SI	SI	SI	SI	NO	SI	ALTA
Estudiante 4	NO	SI	NO	SI	NO	NO	NO	SI	BAJO
Estudiante 5	SI	NO	SI	NO	NO	SI	SI	SI	MEDIO
Estudiante N...	NO	NO	SI	NO	NO	SI	NO	SI	NINGUNO

En este capítulo se han presentado los modelos de datos, generados con los algoritmos y técnicas de minería de datos seleccionadas. Estos modelos han sido evaluados en cuanto a su *escalabilidad, complejidad, precisión y eficiencia*. Un resumen de los modelos generados por los algoritmos seleccionados, se presentan de forma resumida en la Tabla 18. Cabe recalcar que estos resultados están en dependencia del tamaño de datos, tamaño de ventana, umbrales o parámetros establecidos y características del computador en donde se ejecutaron los procesos.

Tabla 18 Resumen comparativo de los aspectos de evaluación utilizados. Elaboración propia.

	Apriori	FP-Growth	CHAID	ID3
Escalabilidad *(Incremento del tamaño de datos)*	Baja	Alta	Media	Baja
Complejidad *(Modelo generado)*	Baja	Baja	Media	Alta
Precisión *(Resultados)*	Baja	Media	Alta	Alta
Eficiencia *(Tiempos de ejecución)*	Baja	Alta	Media	Baja

6 CAPÍTULO VI: DISCUSIÓN Y ANÁLISIS DE RESULTADOS

En este capítulo se presenta una discusión sobre los resultados que se han obtenido con los algoritmos y técnicas de minería de datos utilizadas, y que se presentaron en el capítulo anterior.

Los campos obtenidos son de diversos tipos, por ello para poder aplicar los algoritmos de asociación seleccionados, se tuvo que realizar una transformación de valores numéricos o nominales a binominales. Y para aplicar los algoritmos de clasificación seleccionados, se tuvo que realizar discretización de los campos numéricos, debido a que estos algoritmos sólo aceptaban campos tipo nominal.

Durante la exploración de los datos, se detectaron campos que poseen correlación negativa entre ellos, si un valor aumenta el otro disminuye. Los campos detectados fueron:

- Tasa de reprobación de la asignatura con el promedio de notas semestrales de la asignatura.
- Tasa de reprobación del docente con el promedio de notas semestrales del docente.
- Tasa de reprobación del estudiante con el promedio de notas semestrales del estudiante.

Esta detección permitió disminuir el número de campos a utilizar en el modelado, puesto que se decidió utilizar sólo la tasa de reprobación de estas tres entidades.

Otras deducciones importantes que se obtuvieron durante la exploración de datos fueron:

- La nota del primer bimestre incide en la nota semestral, por lo cual se puede definir que si un estudiante tiene una alta nota en el primer bimestre, se tendrá mayor seguridad de que la nota semestral también será alta.
- El campo *total de asignaturas en las que se matricula el estudiante en el semestre,* es similar para todos los estudiantes analizados, por lo cual se decide que esta variable no tiene mayor incidencia en la nota final que obtiene el estudiante. De igual forma sucedió con las variables edad del docente y del estudiante. Por lo tanto estas tres variables no fueron utilizadas en la fase de modelado.
- Al comparar el ciclo del estudiante y el ciclo de la asignatura, se observó que existen pocos casos en los que un estudiante de un ciclo superior se matricule en una asignatura de un ciclo menor. En este caso se detecta una correlación positiva, en

donde, si la asignatura es de un ciclo menor, el estudiante pertenece a un ciclo menor, y viceversa. También se encontró que cuando un estudiante de ciclo menor toma una asignatura de mayor ciclo, la nota que obtiene es baja, en comparación con la nota que obtiene cuando el ciclo de la asignatura es igual al ciclo del estudiante.

Los tiempos de ejecución de los algoritmos, además de ser afectados por los parámetros de: *tamaño de datos, tamaño de ventana y umbrales*, también se relacionan a las características del computador, en donde se ejecutan los procesos o flujos desarrollados. Si se mejoran estas características, los tiempos de ejecución podrían disminuirse.

Se detectó que los algoritmos con mejores resultados para el problema planteado, así como para los datos utilizados, fueron *FP-Growth* y *CHAID.* Cada uno de estos pertenece a diferente técnica de minería de datos. *FP-Growth* busca relación de asociación o dependencia de datos, y el *CHAID*, busca predecir un resultado.

Visto así estos algoritmos no tendrían el mismo objetivo, pero al realizar el análisis de los modelos que estos dos algoritmos generaron, se ha encontrado que, sus resultados no son excluyentes, al contrario se complementan, y refuerzan los resultados obtenidos por el otro.

Por lo tanto si se desea mayor precisión, comprensión y disminución de la complejidad en los resultados de los árboles de decisión, se sugiere realizar una clasificación basada en resultados de asociación, como se realizó en esta investigación.

Se observó también que los algoritmos de asociación generaron resultados similares a los algoritmos de clasificación, pero los primeros lo realizaron con menos costes computacionales.

Los resultados obtenidos indican que la reprobación de un estudiante, se ve afectada por:

- La poca o nula interacción del estudiante a través del chat académico o de tutorías.
- La nota que el estudiante obtiene en su primer bimestre, si es baja determina un alto riesgo de reprobación para el estudiante.
- Las altas tasas de reprobación del estudiante y de la asignatura. Estos dos valores deben ser comunicados al docente al momento en que se le asigne la asignatura y los estudiantes, puesto que la tasa de reprobación del estudiante indica que este viene con problemas en sus estudios, y la segunda indica que la asignatura tiene un alto nivel de complejidad, por lo cual se debe considerar el uso de una metodología de enseñanza de mayor efectividad.

- Con respecto al ciclo de la asignatura, si esta es de ciclo iniciales, se debe considerar aplicar una metodología de enseñanza diferente a las asignaturas de ciclos superiores, debido a que los estudiantes que se matriculan en estas asignaturas, en su mayoría llevan poco tiempo en la institución. Además se debe evaluar si los contenidos de las asignaturas deben ser estudiados en los ciclos iniciales de la carrera, *¿necesita el estudiante de conocimientos previos para poder comprenderla?*
- *Docente con alta carga académica,* se debe analizar la asignación del número de paralelos por docente, puesto que se ha detectado que el asignar más de tres paralelos a un docente, está afectando en el rendimiento de sus estudiantes.

Es importante recalcar que la finalidad de detectar los factores antes mencionados, es prevenir la reprobación de un estudiante de forma temprana, para que se puedan tomar medidas correctivas. Estos factores son de tipo académico, no se han podido utilizar factores de tipo socioeconómico o familiar, debido a que la ficha estudiantil, que en la actualidad posee la institución, no se actualiza semestralmente por los estudiantes, y además se permiten valores vacíos.

Los factores detectados por las reglas de asociación fueron también utilizados en la ejecución de los árboles de decisión, lo que permitió descubrir, que estos campos tienen un alto porcentaje de precisión en la detección de estudiantes con alto riesgo de reprobación.

7 CONCLUSIONES Y TRABAJO FUTURO

En base a los resultados obtenidos, se pueden realizar las siguientes conclusiones:

- El uso de una metodología enfocada en proyectos de minería de datos, así como de una herramienta que pueda utilizarse en las diferentes fases que la metodología propone, facilitó el desarrollo y comprensión de los procesos realizados.
- Para disminuir la complejidad y tiempos de ejecución, así como para mejorar la precisión de los árboles de decisión, se sugiere agregar como un proceso más, la generación de reglas de asociación, puesto que se ha detectado que al aplicarlos conjuntamente, mejora la calidad de los resultados.
- De entre estas dos técnicas de minería de datos, se observa que los árboles de decisión requieren mayores requerimientos computacionales.
- Ambas técnicas de minería de datos descubrieron patrones académicos similares. Los resultados que generaron son equivalentes, pudiéndose concluir con mayor certeza, que los factores que pueden determinar el riesgo de reprobación de un estudiante son: *la no realización de chats académicos o de tutorías, la nota que el estudiante obtiene en su primer bimestre, las altas tasas de reprobación del estudiante y de la asignatura, el ciclo de la asignatura, en especial si la asignatura es de ciclos iniciales, y docentes con alta carga académica.*
- De entre todas las actividades realizadas en el campus virtual, se detecta que el chat académico semanal, tiene mayor efecto en la posible reprobación de un estudiante, por lo cual debería ser una de las actividades que más fomente el docente y la institución en sus estudiantes.

Como una mejora a los campos utilizados en esta investigación, se propone utilizar un campo llamado *porcentaje de participación en actividades del campus virtual,* el cual contenga el resultado entre, el *número de actividades en las que participa el estudiante* y el *número de actividades propuestas por el docente*.

Se ha detectado que el número de cuestionarios que realiza el estudiante no está ayudando como debería en la obtención de un buen rendimiento académico, por lo que se considera necesario que como trabajo a futuro se realice un análisis de la calidad del cuestionario *(análisis de contenido),* en donde se revise el contenido de las preguntas, existencia de retroalimentación, relación a los temas revisados, y si esta actividad profundizan en el conocimiento que se desea que, obtenga el estudiante con la asignatura.

También como un trabajo a futuro, se podría realizar un análisis similar, pero enfocado en la modalidad presencial, en donde se podrían utilizar los horarios, asistencias a clases, calidad de aulas, entre otros factores, característicos de esta modalidad de estudio.

9 REFERENCIAS

Aponte Hernández, E. (2015). *UNESCO.ORG.* Recuperado el 2017, de http://unesdoc.unesco.org/images/0024/002442/244270m.pdf

Ballesteros A., S. D. (2013). Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo. *Alejandro Ballesteros Román, Daniel Sánchez-Guzmán and Ricardo García Salcedo*.

Claudio Palma, W. (2009). *Data Mining. El arte de anticipar. 10 casos reales.* RIL Editores. Recuperado el 2017

Dirección de Innovación Educativa DIE-UNAH. (2017). Recuperado el 2017, de https://die.unah.edu.hn/assets/Uploads/Tendencias-en-la-Educacion-Superior-2017.pdf

Galán Cortina, V. (2016). *Biblioteca de la Universidad de San Carlos III de Madrid.* Recuperado el 2017, de http://hdl.handle.net/10016/22198

Gallardo Arancibia, J. (s.f.). Recuperado el 2017, de http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

Guevara Maldonado, C. B. (2012). *Repositorio de la producción académica de la Universidad Complutense de Madrid.* Recuperado el 2017, de http://eprints.ucm.es/16711/1/TFM_CESAR_BYRON_GUEVARA_MALDONADO.pdf

Hernández, J. (09 de 2015). Obtenido de http://posgrado.itlp.edu.mx/uploads/55f7167f2302e.pdf

IBM Knowledge Center. (s.f.). Recuperado el 2017, de https://www.ibm.com/support/knowledgecenter/es/SS3RA7_16.0.0/com.ibm.spss.modeler.help/clementine/nodes_associationrules.htm

Jaramillo, A., & Paz Arias, H. (2015). Recuperado el 2017, de https://www.google.com.ec/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjvwOeh6ofWAhXJ4iYKHTulCcIQFggqMAA&url=http%3A%2F

%2Fwww.rte.espol.edu.ec%2Findex.php%2Ftecnologica%2Farticle%2Fdownload%2F351%2F229&usg=AFQjCNF2lwr_8_XLiN7C1j0F6rl

Jiménez Ramírez, C. (2012). Recuperado el 2017, de https://tecaprendizajeest.wikispaces.com/file/view/reglas+de+asociaci%C3%B3n.pdf

KDnuggets TM. (2014). Obtenido de http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Macías, M. (2008). *Comisión Nacional de Seguros y Finanzas.* Recuperado el 2017, de http://www.cnsf.gob.mx/Eventos/Premios_2014/ANIVDELAREV.pdf

Maya Betancourt, A. (1993). *UNESCO.* Recuperado el 2017, de http://www.unesco.org/education/pdf/53_21.pdf

Molina López, J., & García Herrero, J. (2012). *OPEN COURSE WARE UNIVERSIDAD CARLOS III DE MADRID.* Recuperado el 2017, de http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/libroDataMiningv5.pdf

Monsalvea, J., Aponteb, F., & Hoyos, J. (2013). Aplicación de minería de datos educativos a procesos. Colombia.

Ochoa, L. (2016). Recuperado el 2017, de http://tesis.ucsm.edu.pe/repositorio/bitstream/handle/UCSM/6021/N7.0298.SE.pdf?sequence=1&isAllowed=y

Orallo, J., Quintana, M., & Ramírez, C. (2004). *Introducción a la minería de datos.* Editorial Alhambra S. A. (SP).

Ordoñez Briceño, K. (2013). *Repositorio Institucional de la UTPL (RiUTPL).* Recuperado el 2017, de http://dspace.utpl.edu.ec/bitstream/123456789/7897/1/Ordonez%20Brice%C3%B1o%20Karla-%20Informatica.pdf

Pérez César, S. D. (2007). *Minería de datos: técnicas y herramientas.* Editorial Paraninfo, 2007.

Rodríguez León, C., & García Lorenzo, M. M. (2016). *Revista Universidad y Sociedad, 8(4), 43-53.* Recuperado el 29 de 08 de 2017, de

http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202016000400005&lng=es&tlng=es.

Rojas Calvo, W. (Junio de 2016). Recuperado el 2017, de http://repositorio.autonoma.edu.co/jspui/bitstream/11182/1032/1/Solucion%20de%20BI-DM%20Proceso%20de%20Cirugia.pdf

Romero, C. (2013). Minería de Datos en Educación y Análisis del Aprendizaje.

Soft Computing and Intelligent Information Systems. (2016). Recuperado el 2017, de http://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/InteligenciaDeNegocio/Tema07-Modelos%20de%20Asociacion%20%2020015-16.pdf

Sotomonte Castro, J., Rodríguez Rodríguez, C., Montenegro Marín, C., Gaona García, P., & Castellanos, J. (2016). Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos. *Revista Científica*, 15.

Timarán Pereira, R., & Jiménez Toledo, J. (2014). *CONGRESO IBEROAMERICANO DE CIENCIA, TECNOLOGÍA, INNOVACIÓN Y EDUCACIÓN.* Recuperado el 2017, de http://www.oei.es/historico/congreso2014/contenedor.php?ref=memorias

Timarán Pereira, R., Hidalgo Troya, A., & Caicedo Zambrano, J. (2016). Recuperado el 2017, de https://www.researchgate.net/publication/311370703_Proceso_de_Descubrimiento_de_Patrones_de_Desempeno_Academico_en_la_Competencia_de_Ingles_con_CRISP-DM