UNIVERSIDAD LAICA "ELOY ALFARO" DE MANABÍ

FACULTAD DE CIENCIAS INFORMÁTICAS

TRABAJO DE TITULACIÓN MODALIDAD PROYECTO INTEGRADOR
PREVIO A LA OBTENCIÓN DEL TÍTULO DE:
INGENIERO EN SISTEMAS

TEMA:

"IMPLEMENTACIÓN DE ESTRUCTURAS DE MINERÍA DE DATOS PARA EVALUAR EL NIVEL DE TASA DE RETENCIÓN ESTUDIANTIL DE LA FACULTAD DE CIENCIAS INFORMÁTICAS"

Autora:

Srta. Katherine Ibeth Vélez Molina

Director:

Ing. Jorge Pincay Ponce, Mg

Manta – Manabí – Ecuador

Periodo Académico 2017 – 2018(1)

CERTIFICACIÓN

En calidad de docente tutor(a) de la Facultad de Ciencias informáticas de la Universidad Laica "Eloy Alfaro" de Manabí, certifico:

Haber dirigido y revisado el trabajo de titulación, cumpliendo el total de 164 horas, bajo la modalidad de Proyecto Integrador, cuyo tema del proyecto es **"IMPLEMENTACIÓN DE ESTRUCTURAS DE MINERÍA DE DATOS PARA EVALUAR EL NIVEL DE TASA DE RETENCIÓN ESTUDIANTIL DE LA FACULTAD DE CIENCIAS INFORMÁTICAS",** el mismo que ha sido desarrollado de acuerdo a los lineamientos internos de la modalidad en mención y en apego al cumplimiento de los requisitos exigidos por el Reglamento de Régimen Académico, por tal motivo CERTIFICO, que el mencionado proyecto reúne los méritos académicos, científicos y formales, suficientes para ser sometido a la evaluación del tribunal de titulación que designe la autoridad competente.

La autoría del tema desarrollado, corresponde la señora **VÉLEZ MOLINA KATHERINE IBETH**, estudiante de la carrera de Ingeniería en Sistemas, período académico 2017-2018(1), quien se encuentra apto para la sustentación de su trabajo de titulación.

Particular que certifico para los fines consiguientes, salvo disposición de Ley en contrario.

Lugar, 20 de febrero de 2018.

Lo certifico,

Ing. Jorge Pincay Ponce
Docente Tutor(a)
Área: Desarrollo de Software

TRABAJO DE TITULACIÓN MODALIDAD PROYECTO INTEGRADOR, PREVIO A LA OBTENCIÓN DEL TÍTULO DE: INGENIERO EN SISTEMAS

"IMPLEMENTACIÓN DE ESTRUCTURAS DE MINERÍA DE DATOS PARA EVALUAR EL NIVEL DE TASA DE RETENCIÓN ESTUDIANTIL DE LA FACULTAD DE CIENCIAS INFORMÁTICAS"

Tribunal examinador que declara APROBADO el Grado de INGENIERO EN SISTEMAS, de la señorita: KATHERINE IBETH VÉLEZ MOLINA

Lic. Rubén Darío Basurto Alcívar, Mg.

Ing. Fabricio Javier Rivadeneira Zambrano, Mg.

A.S. Oscar Armando González López, Mg.

Manta, 28 de Febrero del 2018

DECLARACIÓN EXPRESA

La responsabilidad del contenido de este Trabajo de Titulación de cuatro capítulos “IMPLEMENTACIÓN DE ESTRUCTURAS DE MINERÍA DE DATOS PARA EVALUAR EL NIVEL DE TASA DE RETENCIÓN ESTUDIANTIL DE LA FACULTAD DE CIENCIAS INFORMÁTICAS”, me corresponde exclusivamente y los derechos patrimoniales de la misma a la Universidad Laica “Eloy Alfaro” de Manabí.

Katherine Ibeth Vélez Molina

C.I. # 131645927-8

DEDICATORIA

La concepción de este proyecto está dedicada a quienes siempre me han apoyado y han confiado en que pueda cumplir con todo aquello que me proponga.

AGRADECIMIENTO

Agradezco a mis padres y a mi esposo e hija que me apoyaron y motivaron mi formación académica, a mi tutor le doy gracias por su paciencia y tiempo invertido en este proyecto, por último agradezco a esta institución, por permitirme prepararme y formarme para un futuro competitivo.

TABLA DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE GRÁFICOS E ILUSTRACIONES

TITULO

"IMPLEMENTACIÓN DE ESTRUCTURAS DE MINERÍA DE DATOS PARA EVALUAR EL NIVEL DE TASA DE RETENCIÓN ESTUDIANTIL DE l FACULTAD DE CIENCIAS INFORMÁTICAS"

RESUMEN

La baja tasa de retención estudiantil en las instituciones de educación superior del Ecuador se ha convertido en un problema social que ha afectado a varias instituciones y el propósito de reducir el número de estudiantes desertores es algo que tienen muy presente las autoridades. Pueden llegar a existir diversos factores y variables que afecten este fenómeno social.

El siguiente proyecto de titulación se centra en implementar una estructura de minería de datos que permitirá evaluar el índice de tasa de retención estudiantil de la Facultad de Ciencias Informáticas de la Universidad Laica Eloy Alfaro de Manabí. Para la correcta valoración se identificarán distintos factores los cuales son: retiro por asignatura, movilidad interna y externa, reingreso, pérdida de carrera y el uso de estructuras de minería de datos (árbol de decisión, redes neuronales, clúster).

INTRODUCCIÓN

En esta investigación se identificarán distintos elementos los cuales son: movilidad interna y externa, reingreso, pérdida de carrera y el uso de estructuras de minería de datos, con el fin de evaluar la tasa de retención estudiantil, para lo cual se toma como lugar de investigación la “FACULTAD DE CIENCIAS INFORMÁTICAS - ULEAM”.

La estructura de este documento se constituye por cuatro capítulos: el capítulo 1, trata sobre el marco teórico de la investigación, donde se encuentran los antecedentes legales y bibliográficos de investigaciones relacionados a la problemática. En el capítulo 2, se realiza el diagnostico o estudio de campo, definiendo los métodos y tipos de investigación, al igual que las herramientas de recolección de datos y el plan de muestreo a aplicarse. En el capítulo 3, se describe el diseño de la propuesta y se determinan los recursos necesarios para realizarla, adicionalmente se establece el estudio de factibilidad. En el capítulo 4, se muestra la evaluación de los resultados de las pruebas realizadas. Por último se redactan las conclusiones y recomendaciones pertinentes.

PLANTEAMIENTO DE PROBLEMA

I. Ubicación y Contextualización

En esta última década, las universidades de Ecuador han venido presentando el problema de baja tasa de retención estudiantil, según estudios realizados por la Secretaria de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT), de los estudiantes que ingresan a las universidades públicas ecuatorianas, el 26 % de ellos terminan desertando. Aunque los esfuerzos de autoridades mediante sus normas facilite conseguir un cupo para el ingreso realizando nuevamente el Examen Nacional para la Educación Superior (ENES) o tomar otro camino como lo efectúa la Ley de Educación Superior que permite a las universidades regular los cambios de alumnos, si quieren seguir en el mismo centro de educación, siguiendo un proceso de homologación de materias, si estas son compatibles, el alumno podrá realizar el traspaso y seguir continuando con sus estudios universitarios y habiendo otros acciones como la planteada anteriormente para evitar la deserción de los estudiantes, al menos la mitad de las personas que ingresaron a un centro de educación no continuaron con sus estudios. Lo que demuestra que son otros factores los que influyen sobre esta elección del estudiante (El Comercio, 2016).

La Universidad Laica Eloy Alfaro de Manabí no está exenta de los resultados de este estudio.

II. Génesis del Problema

Pese a que las autoridades cuentan con esta información en términos porcentuales, no ha sido posible saber con precisión qué motivos originan una baja tasa de retención estudiantil.

El diario El COMERCIO menciona posibles factores entre ellos una inadecuada orientación para elegir una carrera o ante la falta de recursos económicos las personas optan por trabajar. Más si se convierten en

padres o madres de familia de forma temprana. En el 2010 ya se tuvo información que daba cuenta de esta realidad. Entonces, el promedio de jóvenes de entre 18 y 24 años que no estudiaba ni trabajaba en el país era 19,5%, pero en el 2014 subió a 25,4%", con base en las cifras del Instituto Nacional de Estadística y Censos (INEC) (El Comercio, 2016). Podemos decir que la consecuencia de esta problemática se da cuando cursan los primeros semestres, pero son algunos los factores que afectan de forma interna o externa a los estudiantes y los llevan a tomar la decisión de abandonar sus estudios universitarios. Si bien se ha mantenido en un margen numérico cercano al 26%, no existe un conocimiento exacto sobre los causales de este problema para tomar medidas pertinentes.

En este proyecto se realiza una investigación adentrada en el ámbito académico pues es con la información que contamos para poder manejar el modelado de datos.

III. Estado Actual del Problema

La Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (Unesco) indica que el abandono de los estudios universitarios, antes de llegar a su finalización llega a un 40%. Durante la inauguración de la Sexta Conferencia Latinoamericana sobre el Abandono en la Educación Superior, el catedrático Ulises Orestes advirtió que la deserción repercute negativamente en el avance económico y social de los países, especialmente, en los que se encuentran en vías de desarrollo. Uno de ellos es Ecuador.(EL Telegrafo, 2016)

Actualmente no se ha implementado una herramienta que pueda manejar este gran volumen de información que permita conocer la tasa de retención estudiantil, ya sea porque las autoridades poseen limitaciones para el tratamiento de información suficiente para dicho análisis o por la falta de aplicación de herramientas TICS, y así poder determinar qué causas académicas influyen en la retención estudiantil.

Diagrama Causa-Efecto del Problema

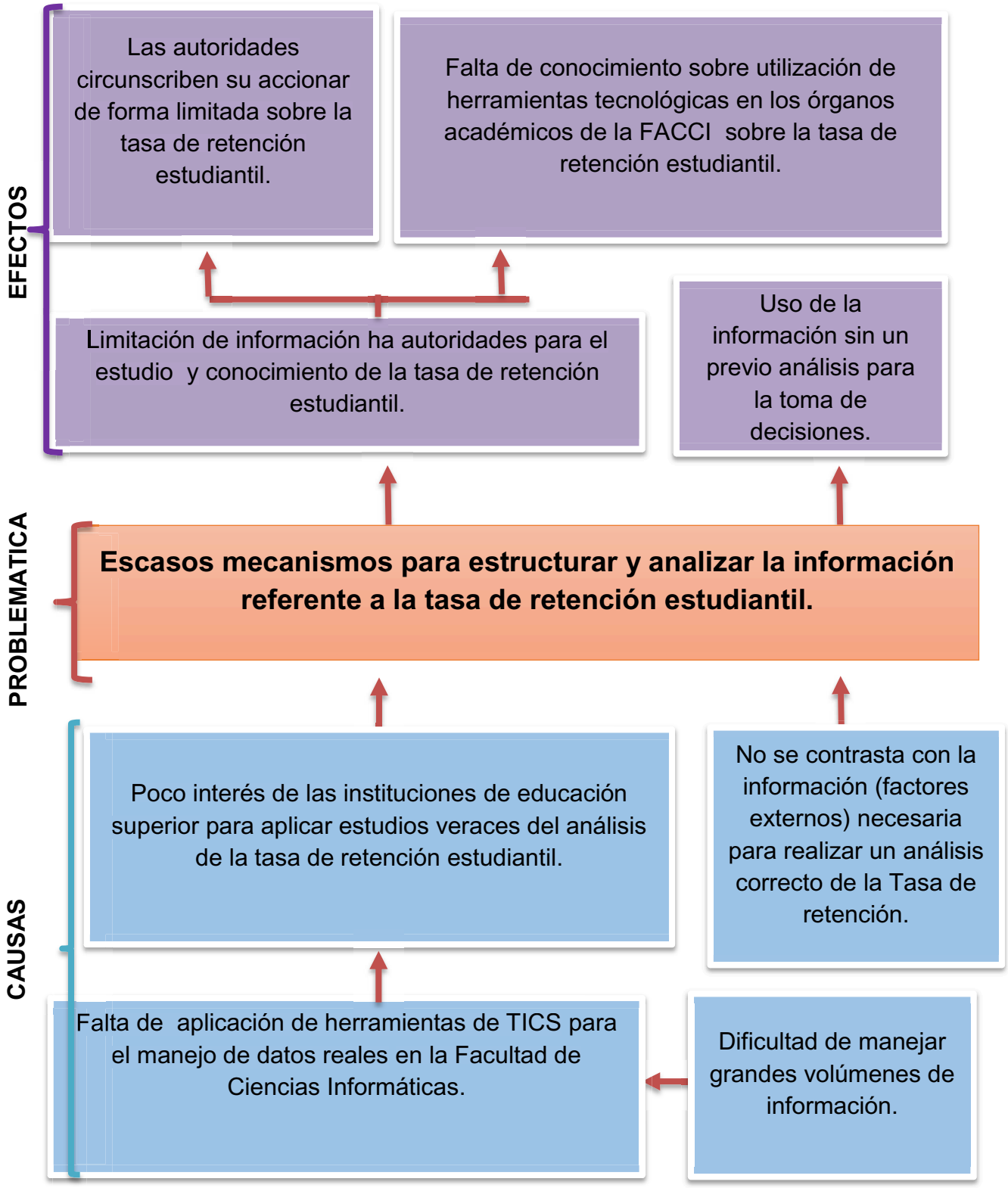


Ilustración 1: Diagrama Causa- Efecto

Formulación del problema

¿Qué factores académicos intervienen en la decisión de los estudiantes que resuelven abandonar sus estudios universitarios haciendo difícil la tarea de retenerlos en las IES?

Delimitación del problema

1) Delimitación del contenido

   Campo: Sistemas Expertos, Inteligencia de negocios
   Área: Social
   Aspecto: Estructura de minería de datos mediante flujo de trabajo en WEKA que permitirá deducir los patrones y tendencias que existen en los datos.

2) Delimitación Espacial

   La implementación de estructuras de minería de datos se orienta de acuerdo a la metodología de aplicación y al problema de retención de estudiantes y los factores académicos que inciden en este en la Facultad de Ciencias Informáticas - ULEAM.

3) Delimitación Temporal

   Esta investigación se llevará acabo entre junio 2017 hasta enero del 2018.

OBJETIVOS

Objetivo General

Implementar estructuras de minería de datos que identifiquen factores que influyan en el entorno académico para evaluar la tasa de retención estudiantil de la facultad de ciencias informáticas.

Objetivos Específicos

- Distinguir mediante investigación bibliográfica los factores que afectan la tasa de retención estudiantil.
- Determinar mediante investigación bibliográfica las estructuras de minería de datos adecuadas para estudiar y evaluar la tasa de retención estudiantil.
- Aplicar estructuras de minería de datos a la información que se recopile sobre estudiantes matriculados y calificaciones en el periodo comprendido entre los años 2011 y 2016(1).
- Identificar regularidades que causan la deserción estudiantil, mediante el uso de las estructuras de minería de datos adecuadas.

JUSTIFICACIÓN

La tasa de retención estudiantil en la carrera de Ingeniería de Sistemas de la Universidad Laica "Eloy Alfaro de Manabí" Manta, es uno de los problemas que perjudica a la eficiencia del sistema educativo. Problema que es de interés estudiantil, docente y administrativo, por lo que se crea la necesidad de investigar este suceso, lo cual permitirá visualizar algunos elementos que pudieron haber sido las causas del abandono de los estudiantes.

La importancia de la implementación de este proyecto de tesis radica en que se podrá realizar estudios para el análisis de la tasa de retención estudiantil y los factores académicos que generaron la falta de retención de estudiantes de la Facultad de Ciencias Informáticas y brindarán a las autoridades una perspectiva adecuada, para que puedan aplicar soluciones al respecto. Esta es la razón por la cual se decidió investigar sobre el suceso observado en los estudiantes de primer, segundo y tercer nivel dentro de los períodos 2011 – 2016(1).

También se pretende con el presente trabajo la aplicación de estructuras de minería de datos para un mejor análisis del problema, plantear una visión acerca de nuestra realidad académica dentro de la educación formal de la educación superior.

CAPITULO I

MARCO TEORICO DE LA INVESTIGACIÓN

1.1. INTRODUCCIÓN

La implementación de una estructura de minería de datos permite evaluar al fenómeno provocado por los alumnos que abandonan sus carreras antes de concluirlas. A pesar de los avances alcanzados en cuanto al acceso a la educación y los esfuerzos realizados para retener alumnos la Facultad de Ciencias Informáticas de la ULEAM presenta este problema, pues aunque aún no existe un análisis respecto a esto, este estudio demuestra mediante la evaluación de ciertos parámetros en qué grado se encuentra.

El análisis de la información que los estudiantes proporcionan a la facultad permite crear un modelo de análisis para obtener patrones de comportamiento de un estudiante desertor, esta información se basa en: movilidad interna y externa, reingreso y pérdida de carrera.

En este capítulo se trata a detalle los posibles factores académicos que afectan la tasa de retención estudiantil, las estructuras de minería de datos adecuada para la evaluación de la información recopilada sobre los estudiantes. Además del análisis del software WEKA el cual soporta tareas de minería de datos como: pre procesamiento de datos, clustering, clasificación, regresión, visualización, y selección.

1.2. FACTORES QUE AFECTAN LA TASA DE RETENCIÓN ESTUDIANTIL

Estudios recientes efectuados en Ecuador, demuestran que retención estudiantil es afectada por la deserción estudiantil, llegando a convertirse en un inconveniente social que se genera con mucha frecuencia en diversas Instituciones de Educación Superior, incluso a nivel mundial. Entre otros factores por la falta de apoyo económico, poco aprovechamiento de las clases, falta de interés por seguir realmente con las carreras universitarias y más bien tomarlas por un compromiso social y en muchos casos con la familia y no así consigo mismo o con su proyecto de vida. (Bazantes, Carpio, & Gutiérrez, 2017).

De acuerdo a varios estudios existen otros factores intra y extra sujetos que abarcar desde la misión – visión de la unidad académica, modelos pedagógicos, cultura universitaria, perfil profesional y ocupacional de los programas (Moreira et al., 2017).

Los factores que impiden que un estudiante logre sus metas académicas se observan sobre todo en los tres primeros niveles y en edades entre los 18 y 20 años y uno que otro entre los 21 a 26 años. La vida de muchos de ellos no solo se centra en estudiar; algunos deben trabajar para aportar en casa aunque esto les cueste dejar sus estudios, otros cuidan de sus hermanos menores y estudian, otros ven la forma de ahorrar porque no alcanza el dinero en la familia, muchos, sobre todo aquellos que vienen de otras provincias viven con lo justo; y hay otra población que estudia lo que no le gusta porque por ley están donde su puntaje los ubicó, o porque no pueden acceder a universidades particulares por lo costoso que resulta, o que iniciaron una carrera pensando en los réditos económicos y al ver la competencia sienten frustración y desesperanza.

Sin embargo, son los estudiantes de nivel socioeconómico bajo, los que verdaderamente presentan estas dificultades y aunque constituya el 28% de esta investigación, son representantes de los problemas que debemos ayudar a subsanar. (Marcillo, Blanco, Espinoza, Quinchiguano, & Andrade, 2017)

1.3. ANTECEDENTES DE INVESTIGACIÓN RELACIONADAS AL TEMA

Se han realizado varios estudios relacionados con la retención y deserción de estudiantes y la aplicación de estructuras, modelos y técnicas de minería de datos para evaluar o predecir el abandono de estos en instituciones de educación nivel básico, medio y superior. Entre los principales se tienen:

Tema 1: 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.

En el 2013 se presentó en la Universidad Técnica Particular de Loja sede Loja, en la carrera de Ingeniería en Sistema Informáticos y Computación por Ordoñez Briceño Karla Fernanda. Se aplicó técnicas de minería de datos para crear un modelo predictivo. Para la creación del modelo se tomaron en cuenta las siguientes variables:

- Información personal.
- Información académica del estudiante.
- Nivel de interacción de entorno virtual de estudiantes.
- Nivel de interacción de entorno virtual de docentes por asignatura.

CRISP-DM fue la metodología utilizada para la creación del modelo y la implementación de algoritmos de Inteligencia Artificial se realizó en la herramienta de procesamiento de datos WEKA.

Técnicas de minería de datos aplicada: Se utilizaron los arboles de decisión implementando la tarea de clasificación bajo algoritmo J48, Clustering bajo algoritmo Simple-Kmeans y reglas de asociación implementando el algoritmo A priori.

```
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.35 (666 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 15

Size of set of large itemsets L(3): 4

Best rules found:

 1. ESTADO_APROBACION=REPROBADO 1170 ==> SUPLETORIO=SI 1170    conf:(1)
 2. EDAD=16a26 ESTADO_APROBACION=REPROBADO 771 ==> SUPLETORIO=SI 771    conf:(1)
 3. ESTADO_APROBACION=REPROBADO NIVEL_INTER_PROF=Alto 746 ==> SUPLETORIO=SI 746    conf:(1)
 4. PRESENT_TODAS_LAS_EVAL=NO 731 ==> SUPLETORIO=SI 731    conf:(1)
 5. ASISTIO_SUPLETORIO=NO 709 ==> ESTADO_APROBACION=REPROBADO 709    conf:(1)
 6. ASISTIO_SUPLETORIO=NO 709 ==> SUPLETORIO=SI 709    conf:(1)
 7. SUPLETORIO=SI ASISTIO_SUPLETORIO=NO 709 ==> ESTADO_APROBACION=REPROBADO 709    conf:(1)
 8. ESTADO_APROBACION=REPROBADO ASISTIO_SUPLETORIO=NO 709 ==> SUPLETORIO=SI 709    conf:(1)
 9. ASISTIO_SUPLETORIO=NO 709 ==> ESTADO_APROBACION=REPROBADO SUPLETORIO=SI 709    conf:(1)
10. ASISTIO_SUPLETORIO=SI 705 ==> SUPLETORIO=SI 705    conf:(1)
```

Ilustración 2: Resultados de reglas de asociación – Carrera de Informática

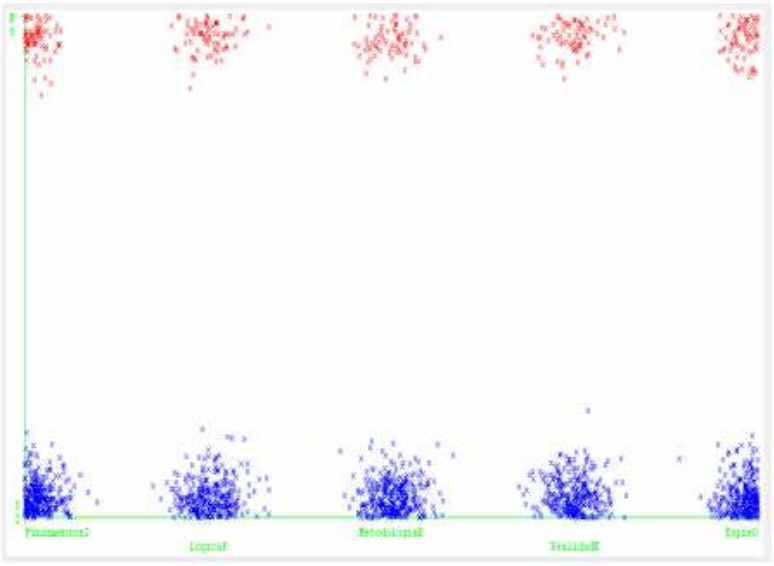


Ilustración 3: Distribución de la deserción por curso – Carrera de Informática

Tema 2: "Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes que pertenecen al Colegio Fiscomisional "San Francisco" de la Ciudad de Ibarra".

En el 2014 se presentó en la Universidad Regional Autónoma De Los Andes "UNIANDES" sede Ibarra, en la carrera de Sistema por Córdova Galarza Janeth Carolina. Los datos obtenidos corresponden a una muestra de estudiantes que estuvieron en los 6 cursos del Colegio San Francisco correspondiente al período académico 2012 – 2013, conjuntamente con la información obtenida de las encuestas realizadas a los estudiantes que cursan el periodo académico 2013- 2104. Para la predicción de la deserción se tomaron en cuenta las siguientes variables:

- Información personal.
- Información académica del estudiante.
- Información de directivos y personal docente.

CRISM-DM fue la metodología utilizada para la creación del modelo y la implementación de algoritmos de Inteligencia Artificial se realizó en la herramienta de procesamiento de datos WEKA.

Técnicas de minería de datos aplicada: Se utilizaron los arboles de decisión implementando la tarea de clasificación bajo algoritmo J48, Clustering bajo algoritmo Simple-Kmeans y reglas de asociación implementando el algoritmo A priori.

Ilustración 4: Porcentaje de deserción

PARALELO	BASCQ804	BASCQ804	BASCQ803	BASCQ802	BASCQ805
BASCQ804	40 (20%)	29 (30%)	0 (0%)	5 (12%)	6 (13%)
BASCQ805	40 (20%)	12 (12%)	1 (4%)	0 (0%)	27 (61%)
BASCQ802	40 (20%)	11 (11%)	4 (19%)	25 (64%)	0 (0%)
BASCQ803	40 (20%)	16 (16%)	14 (66%)	5 (12%)	5 (11%)
BASCQ801	40 (20%)	28 (29%)	2 (9%)	4 (10%)	6 (13%)
EDAD	13.45 +/-0.5087	13.3654 +/-0.5103	13.7619 +/-0.4364	13.6462 +/-0.3655	13.0909 +/-0.2908
GENERO	M	M	M	M	M
M	200 (100%)	96 (100%)	21 (100%)	39 (100%)	44 (100%)
TRABAJA	no	no	no	no	no
no	199 (99%)	95 (98%)	21 (100%)	39 (100%)	44 (100%)
si	1 (0%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)
ESTADO CIVIL	SOLTERO	SOLTERO	SOLTERO	SOLTERO	SOLTERO
SOLTERO	200 (100%)	96 (100%)	21 (100%)	39 (100%)	44 (100%)
DESERTOR	NO	NO	NO	NO	NO
NO	196 (98%)	93 (96%)	21 (100%)	39 (100%)	43 (97%)
SI	4 (2%)	3 (3%)	0 (0%)	0 (0%)	1 (2%)
PROMEDIO_FINAL	B	B	B	B	B
C	73 (36%)	36 (37%)	6 (28%)	12 (30%)	19 (43%)
B	97 (48%)	40 (41%)	13 (61%)	21 (53%)	23 (52%)
A	26 (13%)	17 (17%)	2 (9%)	6 (15%)	1 (2%)
E	2 (1%)	2 (2%)	0 (0%)	0 (0%)	0 (0%)
D	2 (1%)	1 (1%)	0 (0%)	0 (0%)	1 (2%)
SUPLETORIO	NO	NO	NO	SI	SI
SI	62 (31%)	0 (0%)	1 (4%)	28 (71%)	33 (75%)
NO	136 (69%)	96 (100%)	20 (95%)	11 (28%)	11 (25%)
ACOSO_ESCOLAR	NO	NO	NO	NO	NO
NO	184 (92%)	86 (89%)	19 (90%)	37 (94%)	42 (95%)
SI	16 (8%)	10 (10%)	2 (9%)	2 (5%)	2 (4%)
PAGO_MATRICULA	contado	contado	credito	contado	contado
contado	181 (90%)	95 (98%)	7 (33%)	38 (97%)	41 (93%)
credito	19 (9%)	1 (1%)	14 (66%)	1 (2%)	3 (6%)
APROBADO	SI	SI	SI	SI	SI
SI	192 (96%)	90 (93%)	21 (100%)	39 (100%)	42 (95%)
NO	8 (4%)	6 (6%)	0 (0%)	0 (0%)	2 (4%)

Ilustración 5: Resultados – Simple K-means- Octavos

Tema 3: "Las técnicas de predicción y su incidencia en la detección de patrones de deserción estudiantil en la carrera de docencia en informática de la Facultad de Ciencias Humanas y de la Educación de la Universidad Técnica de Ambato."

En el 2016 se presentó en la Universidad Técnica de Ambato sede Ambato, en la carrera de Ingeniería en Sistemas, Electrónica e Industrial por la Ing. Blanca Rocio Cuji Chacha. Se utilizó variables cualitativas para la determinación de las causas y cuantitativas para los datos numéricos como la edad, promedio, los datos estudiados fueron tomados a partir del año 2006. Para la creación del modelo predictivo se tomaron en cuenta los siguientes indicadores:

- Actividad Académica de los estudiantes
- Estudiantes graduados
- Estudiantes egresados (Es desertor)
- Estudiantes inactivos por deserción voluntaria
- Deserción por reglamento (pérdidas)

Utilización de la encuesta para la recolección de datos.

Técnicas de minería de datos aplicada: Se utilizaron los arboles de decisión.

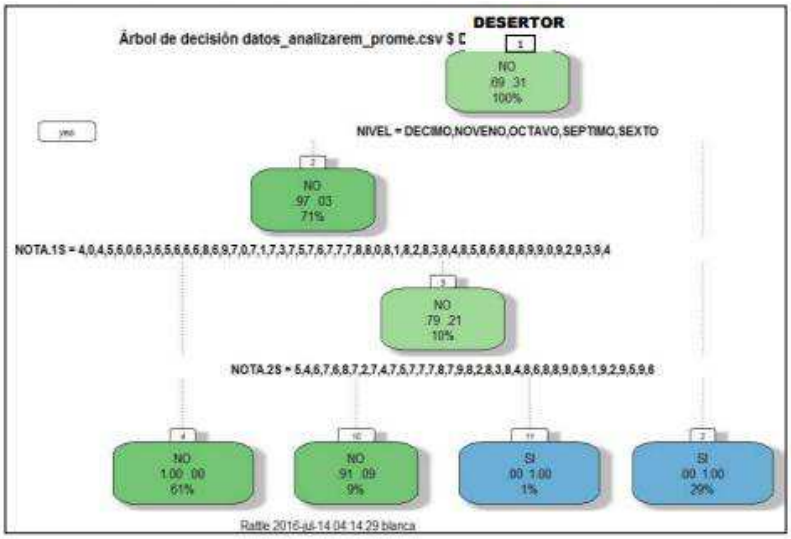


Ilustración 6: Árbol de decisión – datos históricos a partir del 2016.

1.4. DEFINICIONES CONCEPTUALES

1.4.1. Deserción estudiantil

La deserción es un fenómeno presente en todo sistema educativo, relacionado con los procesos de selección, rendimiento académico y de la propia eficiencia del sistema en general, es decir, el resultado de la combinación y efecto de distintas variables (Díaz Peralta, 2008). En este sentido, la deserción de estudiantes universitarios vinculado al desempeño académico de los mismos, es un tema que preocupa desde hace varios años.

Se han realizado estudios con el objeto de aportar información que contribuya a determinar cuáles son las causas. (Martínez-Padilla & Pérez-González, 2008) identificaron que las variables relacionadas con la trayectoria académica que mayor efecto tiene en la estimación del desempeño corresponden al promedio general alcanzado en la enseñanza media, el rendimiento académico y la cantidad de materias que fueron reprobadas durante su permanencia en la universidad; determinando así el grado de éxito y fracaso de los estudiantes mexicanos para el examen nacional de egreso de la licenciatura en ingeniería.

Otros estudios de (Soria-Barreto & Zúñiga-Jara, 2014), determinaron que las principales variables que resultaron estadísticamente determinantes en el éxito de los estudiantes fueron, en el orden de importancia, las calificaciones obtenidas en la enseñanza media, el puntaje obtenido en la prueba de aptitud académica de matemáticas, y el número de años de desfase entre el año de egreso de la enseñanza media y el año de ingreso a la universidad. Además, (Díaz, 2009) concluye que los estudiantes de ingeniería, presentan altos riesgos de deserción entre el primer y tercer semestre, siendo máximo en este último semestre, para luego descender y permanecer a tasas más estables.

Factores determinantes de la deserción universitaria:

Individuales	Académicos	Institucionales	Socioeconómicos
– Edad, género, estado civil – Entorno familiar – Calamidad y problemas de salud – Integración social – Incompatibilidad horaria en actividades extra académicas.	– Orientación profesional – Rendimiento académico – Calidad del programa – Métodos de estudio – Calificación en examen admisión – Insatisfacción en el programa u otros factores académicos – Numero de materias	– Normalidad académica – Tipo de colegio – Becas y forma de financiamiento – Recursos universitarios – Orden público – Entorno político – Relaciones con los profesores y otros estudiantes	– Estrato – Trabajo del estudiante – Situación laboral de los padres – Dependencia económica – Personas a cargo – Nivel educativo de los padres – Entorno macroeconómico del país

Tabla 1: Factores determinantes de la deserción universitaria

Fuente: (Angúlo & Sergio, 2012)

1.4.1.1. Metodología sobre la deserción

Definitivamente debe distinguirse entre la deserción (no académica) o intra-sujeto, y la mortalidad (o deserción académica) o extra-sujeto. La deserción académica podrá ser entonces por razones disciplinarias o por rendimiento y la deserción no académica, por retiro “voluntario”.

Se distingue el abandono voluntario del no voluntario, encuadrando la deserción académica como la no voluntaria y la no académica, como la voluntaria. Sin embargo, la deserción no académica no siempre será tan “voluntaria”. (Angúlo & Sergio, 2012)

La deserción no es un problema del individuo; desde luego que el desertor es aquel en donde todo se concentra, pero ello no es suficiente para declararlo culpable. De cualquier forma sí se debe mirar directamente al desertor, que es donde converge todo el proceso de la deserción.

1.4.1.2. Métodos de Predicción de Deserción

Para explorar algunos métodos que permitan predecir este comportamiento y lograr los objetivos propuestos, existen varios caminos, a saber según (Angúlo & Sergio, 2012):

Modelos estadísticos, modelos de minería de datos y modelos de Inteligencia de negocios, que comprende a los sistemas de soporte a la toma de decisión, entre otros.

Existen técnicas estadísticas que se han utilizado de forma sistemática para abordar distintos problemas de modelación y predicción a partir de distintas variables. Sin embargo, el gran volumen de datos del que se dispone hoy día hace que estas técnicas tarden mucho en numerosos problemas de interés.

La necesidad de métodos eficientes y automáticos para explorar bases de datos ha motivado un rápido avance de disciplinas conocidas hoy como Inteligencia de negocios (BI) en la que los sistemas de Soporte a la toma de decisiones a través de minería de datos, permiten desarrollar métodos que

operen de forma automática a partir de un conjunto de datos para capturar distintos patrones de comportamiento que sean apropiados para resolver un problema.

Las redes probabilísticas: Son modelos apropiados para el tratamiento de problemas con incertidumbre, y utilizan técnicas estadísticas modernas de inferencia y estimación para ajustar los parámetros a los datos y obtener conclusiones en base a los modelos resultantes. Por otra parte, los sistemas de minería de datos, permiten encontrar patrones de comportamiento, en un conjunto de datos que representan la realidad codificada a partir de los sistemas transaccionales.

1.4.2. Metodologías para el proceso de extracción de conocimiento

Teniendo en cuenta que el proceso de extracción de conocimiento KDD según (Angúlo & Sergio, 2012), es un proceso no trivial, ha surgido la necesidad de una aproximación sistemática para la realización de los proyectos de Data Mining, por lo que diversas empresas y consorcios han especificado un proceso de modelado diseñado para guiar al usuario a través de una sucesión de pasos que le dirijan a obtener buenos resultados.

Así la empresa SAS propone la utilización de la metodología SEMMA. En 1999 un importante consorcio de empresas europeas, NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra) y OHRA (Holanda), unieron sus recursos para el desarrollo de la metodología de libre distribución CRISP-DM (Cross- Industry Standard Process for Data Mining). Siendo estas las principales metodologías utilizadas para la realización de proyectos de Data Mining.

Estas metodologías comparten la misma esencia estructurando el proyecto de Data Mining en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de Data Mining en un proceso iterativo e interactivo.

1.4.2.1. Metodología CRISP-DM

La metodología CRISP-DM (Chapman et al., 2000) consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos.

A nivel más general, el proceso está organizado en fases, estando cada fase a su vez estructurada en varias tareas genéricas de segundo nivel.

Las tareas genéricas se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas. Así, si en el segundo nivel se tiene la tarea genérica “limpieza de datos”, en el tercer nivel se indican las tareas que tienen que desarrollarse para un caso específico, como por ejemplo, “limpieza de datos numéricos”, o “limpieza de datos categóricos”. El cuarto nivel, recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de Data Mining específico.

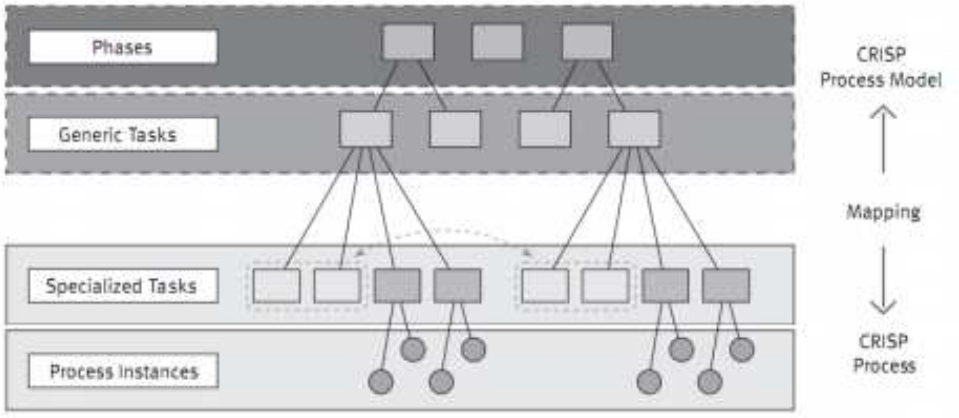


Ilustración 7: Cuatro niveles de detalle de la metodología CRISP-DM

La metodología CRISP-DM proporciona dos documentos distintos como herramienta de ayuda en el desarrollo del proyecto de Data Mining: el modelo de referencia y la guía del usuario.

El documento del modelo de referencia describe de forma general las fases, tareas generales y salidas de un proyecto de Data Mining en general. La guía del usuario proporciona información más detallada sobre la aplicación práctica del modelo de referencia a proyectos de Data Mining específicos,

proporcionando consejos y listas de comprobación sobre las tareas correspondientes a cada fase.

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Data Mining en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.

- **Fase de análisis del problema:** incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación.

- **Fase de análisis de datos:** comprende la recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis. Una vez realizado el análisis de datos, la metodología establece que se proceda a la preparación de los datos, de tal forma que puedan ser tratados por las técnicas de modelado.

- **Fase de preparación de datos:** incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

La fase de preparación de los datos, se encuentra muy relacionada con la fase de modelado, puesto que en función de la técnica de modelado que vaya a ser utilizada los datos necesitan ser procesados en diferentes formas. Por lo tanto, las fases de preparación y modelado interactúan de forma sistemática.

- **Fase de modelado:** se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas que serán utilizadas en esta fase se seleccionan en función de los siguientes criterios:

  1. Ser apropiada al problema.
  2. Disponer de datos adecuados.
  3. Cumplir los requerimientos del problema.
  4. Tiempo necesario para obtener un modelo.
  5. Conocimiento de la técnica.

  Antes de proceder al modelado de los datos, se debe establecer un diseño del método de evaluación de los modelos, que permita establecer el grado de bondad de los modelos.

  Una vez realizadas estas tareas genéricas, se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos.

- **Fase de evaluación:** se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo.

- Fase de explotación: normalmente los proyectos de Data Mining no terminan en la implantación del modelo, sino que se deben documentar y presentar los resultados de manera comprensible en orden a lograr un incremento del conocimiento. Además, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión

de los resultados (Fayyad 1996). En la imagen 7 - El proceso CRISP DM se puede ver un esquema de las diferentes fases de la metodología y las tareas generales que se deben desarrollar en cada fase.

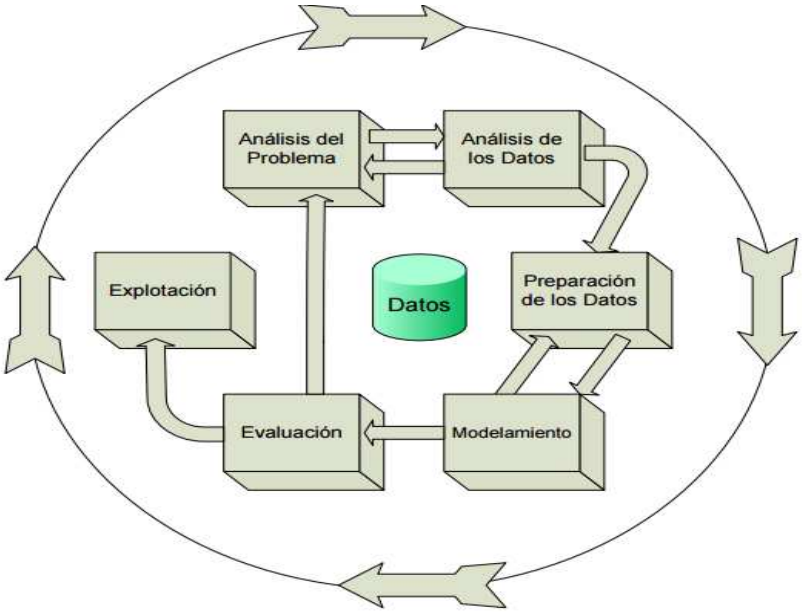


Ilustración 8: El proceso CRISP DM

1.4.2.2. Metodología MSDN

La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo. (2. Microsoft, 2016) Este proceso se puede definir mediante los seis pasos básicos siguientes:

1. Definir el problema
2. Preparar los datos
3. Explorar los datos
4. Generar modelos
5. Explorar y validar los modelos
6. Implementar y actualizar los modelos

El siguiente diagrama describe las relaciones existentes entre cada paso del proceso y las tecnologías de Microsoft SQL Server que se pueden usar para completar cada paso.

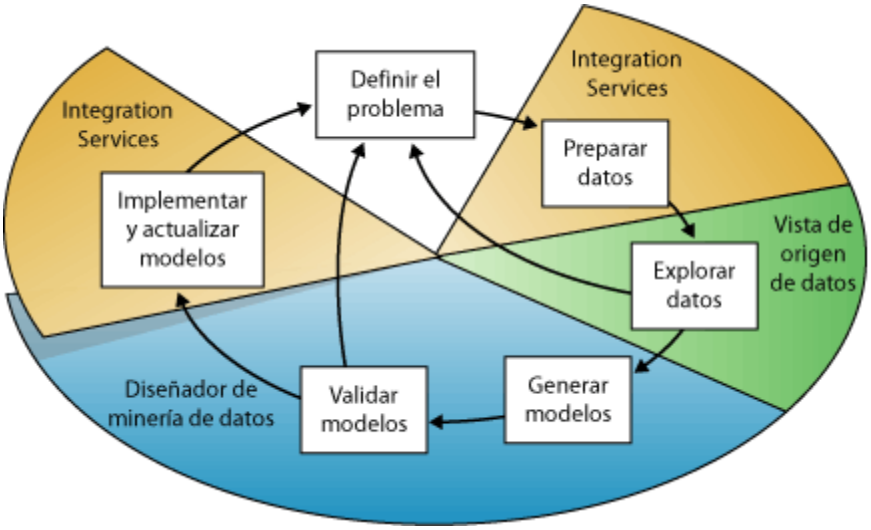


Ilustración 9: El proceso de la metodología MSDN

El proceso que se ilustra en el diagrama es cíclico, lo que significa que la creación de un modelo de minería de datos es un proceso dinámico e iterativo. La minería de datos de Microsoft SQL Server ofrece un entorno integrado para crear y trabajar con modelos de minería de datos.

- **Definir el problema**

Consiste en definir claramente el problema y considerar formas de usar los datos para proporcionar una respuesta para el mismo.

Este paso incluye analizar los requisitos empresariales, definir el ámbito del problema, definir las métricas por las que se evaluará el modelo y definir los objetivos concretos del proyecto de minería de datos. Estas tareas se traducen en preguntas como las siguientes:

1. ¿Qué está buscando? ¿Qué tipos de relaciones intenta buscar?
2. ¿Refleja el problema que está intentando resolver las directivas o procesos de la empresa?

3. ¿Desea realizar predicciones a partir del modelo de minería de datos o solamente buscar asociaciones y patrones interesantes?
4. ¿Qué resultado o atributo desea predecir?
5. ¿Qué tipo de datos tiene y qué tipo de información hay en cada columna? En caso de que haya varias tablas, ¿cómo se relacionan? ¿Necesita limpiar, agregar o procesar los datos antes de poder usarlos?
6. ¿Cómo se distribuyen los datos? ¿Los datos son estacionales? ¿Los datos representan con precisión los procesos de la empresa?

- **Preparar los datos**

Consiste en consolidar y limpiar los datos identificados en la definición del problema. Los datos pueden estar dispersos en la empresa y almacenados en formatos distintos; también pueden contener incoherencias como entradas que faltan o incorrectas.

La limpieza de datos no solamente implica quitar los datos no válidos o interpolar valores que faltan, sino también buscar las correlaciones ocultas en los datos, identificar los orígenes de datos que son más precisos y determinar qué columnas son las más adecuadas para el análisis.

Por consiguiente, antes de empezar a generar los modelos de minería de datos, debería identificar estos problemas y determinar cómo los corregirá. En la minería de datos, por lo general se trabaja con un conjunto de datos de gran tamaño y no se puede examinar la calidad de los datos de cada transacción.

Es importante tener en cuenta que los datos que se usan para la minería de datos no necesitan almacenarse en un cubo de procesamiento analítico en línea (OLAP), ni siquiera en una base de datos relacional, aunque puede usar ambos como orígenes de datos. Puede realizar

minería de datos mediante cualquier origen de datos definido como origen de datos de Analysis Services.

- **Explorar los datos**

Consiste en explorar los datos preparados. Debe conocer los datos para tomar las decisiones adecuadas al crear los modelos de minería de datos. Entre las técnicas de exploración se incluyen calcular los valores mínimos y máximos, calcular la media y las desviaciones estándar, y examinar la distribución de los datos.

Las desviaciones estándar y otros valores de distribución pueden proporcionar información útil sobre la estabilidad y exactitud de los resultados. Una desviación estándar grande puede indicar que agregar más datos podría ayudarle a mejorar el modelo. Los datos que se desvían mucho de una distribución estándar se podrían sesgar o podrían representar una imagen precisa de un problema de la vida real, pero dificultar el ajustar un modelo a los datos.

Al explorar los datos para conocer el problema empresarial, puede decidir si el conjunto de datos contiene datos defectuosos y, a continuación, puede inventar una estrategia para corregir los problemas u obtener una descripción más profunda de los comportamientos que son típicos de su negocio.

Puede usar herramientas como Master Data Services para sondear los orígenes de datos disponibles y determinar su disponibilidad para la minería de datos. Puede usar herramientas como SQL Server Data Quality Services, o el generador de perfiles de datos de Integration Services, para analizar la distribución de los datos y solucionar problemas, como la existencia de datos incorrectos o la falta de datos.

- **Generar modelos**

Consiste en generar el modelo o modelos de minería de datos. Deberá definir qué columnas de datos desea que se usen; para ello, creará una estructura de minería de datos. La estructura de minería de datos se vincula al origen de datos, pero en realidad no contiene ningún dato hasta que se procesa. Al procesar la estructura de minería de datos, Analysis Services genera agregados y otra información estadística que se puede usar para el análisis. Cualquier modelo de minería de datos que esté basado en la estructura puede utilizar esta información. Para obtener más información sobre cómo se relacionan las estructuras de minería de datos con los modelos de minería de datos.

Antes de procesar la estructura y el modelo, un modelo de minería de datos simplemente es un contenedor que especifica las columnas que se usan para la entrada, el atributo que está prediciendo y parámetros que indican al algoritmo cómo procesar los datos. El procesamiento de un modelo a menudo se denomina entrenamiento. El entrenamiento hace referencia al proceso de aplicar un algoritmo matemático concreto a los datos de la estructura para extraer patrones.

Los patrones que encuentre en el proceso de entrenamiento dependerán de la selección de los datos de entrenamiento, el algoritmo que elija y cómo se haya configurado el algoritmo. SQL Server 2014 contiene muchos algoritmos diferentes, cada uno está preparado para un tipo diferente de tarea y crea un tipo distinto de modelo.

- **Explorar y validar los modelos**

Consiste en explorar los modelos de minería de datos que ha generado y comprobar su eficacia. Antes de implementar un modelo en un entorno de producción, es aconsejable probar si funciona correctamente. Además, al generar un modelo, normalmente se crean varios con configuraciones diferentes y se prueban todos para ver cuál ofrece los resultados mejores para su problema y sus datos.

Analysis Services proporciona herramientas que ayudan a separan los datos en conjuntos de datos de prueba para que pueda evaluar con precisión el rendimiento de todos los modelos en los mismos datos y entrenamiento. El conjunto de datos de entrenamiento se utiliza para generar el modelo y el conjunto de datos de prueba para comprobar la precisión del modelo mediante la creación de consultas de predicción.

En SQL Server 2014 Analysis Services (SSAS), estas particiones se pueden hacer automáticamente mientras se genera el modelo de minería de datos.

Puede explorar las tendencias y patrones que los algoritmos detectan mediante los visores del diseñador de minería de datos de SQL Server Data Tools.

También puede comprobar si los modelos crean predicciones correctamente mediante herramientas del diseñador como el gráfico de mejora respecto al modelo predictivo y la matriz de clasificación.

Para comprobar si el modelo es específico de sus datos o se puede usar para realizar inferencias en la población general, puede usar la técnica estadística denominada validación cruzada para crear automáticamente subconjuntos de los datos y probar el modelo con cada uno.

- **Implementar y actualizar los modelos**

Consiste en implementar los modelos que funcionan mejor en un entorno de producción. Una vez que los modelos de minería de datos se encuentran en el entorno de producción, puede llevar acabo diferentes tareas, dependiendo de sus necesidades. Las siguientes son algunas de las tareas que puede realizar:

  - Use los modelos para crear predicciones que luego podrá usar para tomar decisiones comerciales. SQL Server proporciona el lenguaje DMX, que puede usar para crear consultas de predicción y el

generador de consultas de predicción para ayudarle a generar las consultas.

- Crear consultas de contenido para recuperar estadísticas, reglas o fórmulas del modelo.

- Utilizar Integration Services para crear un paquete en el que se utilice un modelo de minería de datos para dividir de forma inteligente los datos entrantes en varias tablas.

- Crear un informe que permita a los usuarios realizar consultas directamente en un modelo de minería de datos existente.

- Actualizar los modelos después de la revisión y análisis. Cualquier actualización requiere que vuelve a procesar los modelos

- Actualizar dinámicamente los modelos, cuando entren más datos en la organización, y realizar modificaciones constantes para mejorar la efectividad de la solución debería ser parte de la estrategia de implementación.

1.4.2.3. Metodología SEMMA

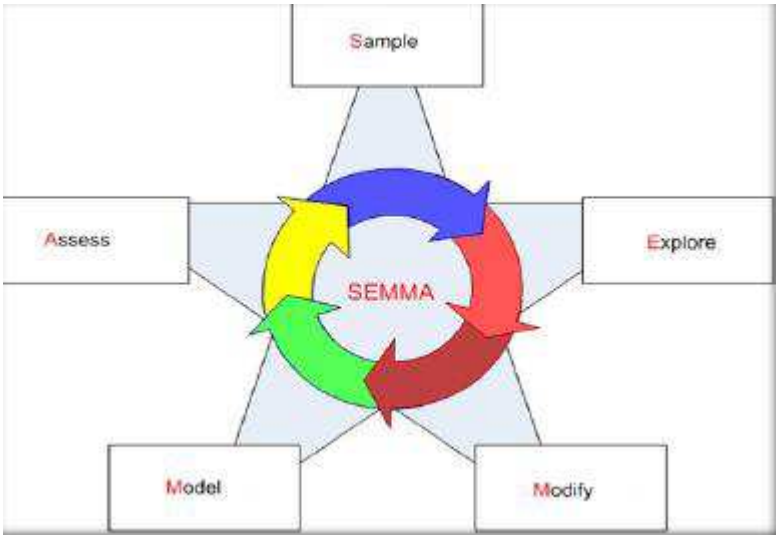


Ilustración 10: El proceso de la metodología SEMMA

Creada por el SAS Institute, se define como "el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos". El nombre de esta terminología corresponde a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess (Valoración).

Se encuentra enfocada especialmente en aspectos técnicos, excluyendo actividades de análisis y comprensión del problema que se está abordando evidenciando que el modelo está orientado especialmente a aspectos técnicos.(Moine, Haedo, & Gordillo, 2011)

1.4.2.4. Metodología KDD

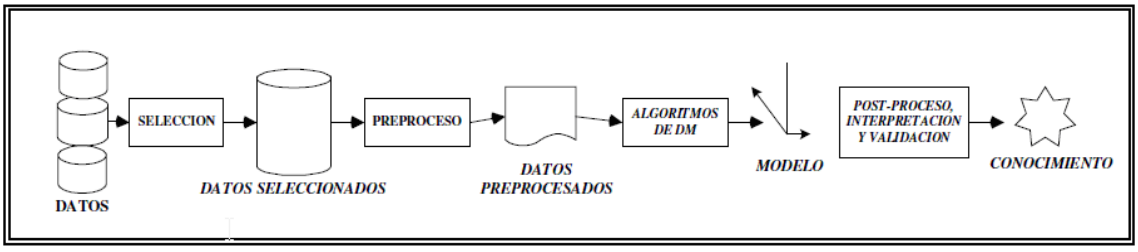


Ilustración 11: El proceso de la metodología KDD

Metodología KDD (Knowledge Discovery in Databases) constituyó el primer modelo aceptado en la comunidad científica que estableció las etapas principales de un proyecto de explotación de información, está formado por nueve etapas. Formalmente el modelo establece que la minería de datos es la etapa dentro del proceso en la cual se realiza la extracción de patrones a partir de los datos. Sin embargo actualmente, en la comunidad científica y en la literatura, el término KDD y minería de datos se utilizan indistintamente para hacer referencia al proceso completo de descubrimiento de conocimiento.(Moine et al., 2011)

1.4.2.5. Metodología CATALYST

Conocida como P3TQ (Product, Place, Price, Time, Quantity), fue propuesta por Dorian Pyle en el año 2003. Esta metodología plantea la formulación de dos modelos: el Modelo de Negocio y el Modelo de Explotación de Información.

La metodología Catalyst, en sus dos modelos, está compuesta por una serie de pasos llamados “boxes”, luego de llevar a cabo una acción, se deben evaluar los resultados y determinar cuál es el próximo paso (box) a seguir. La secuencia y la interacción entre los distintos pasos permiten una flexibilidad muy grande, y una amplia variedad de caminos posibles.

Sobresale en su fase de Modelado del Negocio, contemplando cinco puntos de partida para el proyecto, que finalmente conducirán a la definición de un conjunto de requerimientos y a una situación organizacional que deberá ser abordada desde la minería de datos.(Moine et al., 2011)

1.4.3. Minería de datos

La minería de datos es una parte importante de un proceso más amplio conocido como descubrimiento de conocimiento en bases de datos (KDD por sus iniciales en inglés). El objetivo principal de la minería de datos consiste en

extraer información oculta de un conjunto de datos. Esto puede ser alcanzado por el análisis automático o semiautomático de gran cantidad de datos, lo que permite la extracción de patrones desconocidos. Estos patrones pueden ser grupos de registros de datos (análisis clúster), inusuales registros (detección de anomalías) y dependencias entre datos (asociación de reglas). Por lo tanto, los patrones pueden ser vistos como un resumen de los datos de entrada, y se pueden utilizar para su posterior análisis. (Pérez-Palacios, Caballero, Caro, Rodríguez, & Antequera, 2014)

Hoy se dispone de grandes cantidades de información que se encuentran alojadas en bases de datos, archivos, documentos impresos, páginas web que se crean por una tarea cotidiana específica, dicha información no se analiza ni se integra con el resto de conocimiento; para lo cual existe el área de la Minería de Datos que nace de la necesidad de explicar el porqué de unos sucesos, los cuales están ocultos en datos históricos.

"La Minería de Datos es un conjunto de herramientas de diversas ciencias (Estadística, Informática, Matemáticas, Ingeniería, entre otras)" que persigue extraer conocimiento oculto o información no trivial de grandes volúmenes de datos, con la finalidad de dar soluciones a problemas específicos de una organización. (Kantardzic, 2011)

La minería de datos engloba un proceso para la obtención de conocimiento a partir de datos: selección de un conjunto, el análisis de propiedades de los datos, la transformación de conjunto de datos de entrada, seleccionar y aplicar técnica de minería de datos, seguidamente está el proceso de extracción de datos, finalmente la interpretación y evaluación de los datos.

1.4.4. Tareas de minería de datos

Se clasifican en dos grupos las tareas que se realiza en la Minería de Datos para poder extraer conocimiento oculto, estas son las predictivas que permiten predecir uno o más valores. Y el otro grupo es de las tareas descriptivas su objetivo es describir los existentes, según (Riquelme Santos, Ruiz, & Gilbert, 2006) a continuación se describen las tareas de ambos grupos:

1.4.4.1. Tareas predictivas

- Clasificación.- El objetivo de la tarea es poder clasificar un dato dentro de las clases definidas del dominio que se está modelando.
- Regresión.- El objetivo de la tarea es poder encontrar la similitud entre valores de atributos de una determinada clase de un dominio dado.

1.4.4.2. Tareas descriptivas

- Agrupamiento (clustering).- El objetivo de la presente tarea es obtener grupos o conjuntos en donde se incorpore elementos similares extraídos de las clases del dominio dado.

- Asociación.- El objetivo de la asociación es poder describir las relaciones que existen entre los valores de los atributos de un determinado ejemplo de un dominio establecido.

- Correlación.- El objetivo de la presente técnica es ver, si dos o más atributos numéricos están correlacionados linealmente o relacionados de algún otro modo mediante un análisis de varianza coeficiente de correlación lineal de los datos (Orallo et al., 2004)

1.4.5. Algoritmos de minería de datos

Un algoritmo de minería de datos es un conjunto de cálculos y reglas heurísticas que permite crear un modelo de minería de datos a partir de los datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias.

El algoritmo usa los resultados de este análisis para definir los parámetros óptimos para la creación del modelo de minería de datos.(Microsoft, 2016) A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas.

El modelo de minería de datos que crea un algoritmo a partir de los datos puede tomar diversas formas, incluyendo:

- Un conjunto de clústeres que describe cómo se relacionan los casos de un conjunto de datos.
- Un árbol de decisión que predice un resultado y que describe cómo afectan a este los distintos criterios.
- Un modelo matemático que predice las ventas.
- Un conjunto de reglas que describen cómo se agrupan los productos en una transacción, y las probabilidades de que dichos productos se adquieran juntos.

1.4.5.1. Algoritmo por tipo

Analysis Services incluye los siguientes tipos de algoritmos:

- **Algoritmos de clasificación**, que predicen una o más variables discretas, basándose en otros atributos del conjunto de datos.
- **Algoritmos de regresión**, que predicen una o más variables continuas, como las pérdidas o los beneficios, basándose en otros atributos del conjunto de datos.
- **Algoritmos de segmentación**, que dividen los datos en grupos, o clústeres, de elementos que tienen propiedades similares.
- **Algoritmos de asociación**, que buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación, que pueden usarse en un análisis de la cesta de compra.
- **Algoritmos de análisis de secuencias**, que resumen secuencias o episodios frecuentes en los datos, como un flujo de rutas web.

1.4.5.2. Algoritmo por tarea

- **Predecir un atributo discreto**
  - Algoritmo de árboles de decisión de Microsoft
  - Algoritmo Bayes naive de Microsoft
  - Algoritmo de clústeres de Microsoft
  - Algoritmo de red neuronal de Microsoft
- **Predecir un atributo continuo**
  - Algoritmo de árboles de decisión de Microsoft
  - Algoritmo de serie temporal de Microsoft
  - Algoritmo de regresión lineal de Microsoft
- **Predecir una secuencia**
  - Algoritmo de clústeres de secuencia de Microsoft

1.4.6. Algoritmo de asociación

Un modelo de asociación se compone de una serie de conjuntos de elementos y de las reglas que describen cómo estos elementos se agrupan dentro de los casos. Las reglas que el algoritmo identifica pueden utilizarse para predecir las probables compras de un cliente en el futuro, basándose en los elementos existentes en la cesta de compra actual del cliente.

El algoritmo de asociación recorre un conjunto de datos para hallar elementos que aparezcan juntos en un caso. Después, agrupa en conjuntos de elementos todos los elementos asociados que aparecen, como mínimo, en el número de casos especificado en el parámetro *MINIMUM_SUPPORT* .

Datos requeridos para los modelos de asociación

Al preparar los datos para su uso en un modelo de reglas de asociación, conviene comprender qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que estos datos se utilizan.

Los requisitos para un modelo de reglas de asociación son los siguientes:

- **Una columna de una sola clave:** cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. no se permiten las claves compuestas.
- **Una única columna de predicción**: Un modelo de asociación solo puede tener una columna de predicción. Normalmente, se trata de la columna de clave de la tabla anidada, como el campo que contiene los productos que se han comprado. Los valores deben ser discretos o discretizados.
- **Columnas de entrada:** Las columnas de entrada deben ser discretas. Los datos de entrada de un modelo de asociación suelen encontrarse en dos tablas. Por ejemplo, una tabla puede contener la información del

cliente y la otra las compras de ese cliente. Es posible incluir estos datos en el modelo mediante el uso de una tabla anidada.

Crear predicciones

Una vez procesado el modelo, puede utilizar las reglas y los conjuntos de elementos para realizar predicciones. En un modelo de asociación, una predicción indica qué elemento es probable que se produzca dada la presencia del elemento especificado, y la predicción puede incluir información como la probabilidad, el soporte o la importancia.

1.4.7. Algoritmo de clústeres

Los modelos de agrupación en clústeres identifican las relaciones en un conjunto de datos que no se podrían derivar lógicamente a través de la observación casual.

El algoritmo de clústeres se diferencia de otros algoritmos de minería de datos, como el algoritmo de árboles de decisión, en que no se tiene que designar una columna de predicción para generar un modelo de agrupación en clústeres. El algoritmo de clústeres entrena el modelo de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo.

El algoritmo de clústeres identifica primero las relaciones de un conjunto de datos y genera una serie de clústeres basándose en ellas. Un gráfico de dispersión es una forma útil de representar visualmente el modo en que el algoritmo agrupa los datos, tal como se muestra en el siguiente diagrama. El gráfico de dispersión representa todos los casos del conjunto de datos; cada caso es un punto del gráfico.

Después de definir los clústeres, el algoritmo calcula el grado de perfección con que los clústeres representan las agrupaciones de puntos y, a continuación, intenta volver a definir las agrupaciones para crear clústeres que representen mejor los datos. El algoritmo establece una iteración en este proceso hasta que ya no es posible mejorar los resultados mediante la redefinición de los clústeres.

Puede personalizar el funcionamiento del algoritmo seleccionando una técnica de agrupación en clústeres, limitando el número máximo de clústeres o cambiando la cantidad de soporte que se requiere para crear un clúster.

Pasos del análisis clúster:

- Se tiene información de n casos y k variables.
- Se describen los grupos obtenidos y se comparan unos con otros.
- Validación del análisis.
- Se crean los grupos de acuerdo a la medida de similitud.

Métodos de Agrupamiento:

- Jerárquicos: los datos se agrupan de manera arborescente.
- No jerárquicos: generar particiones a un nivel.
- Paramétricos: se asumen que las densidades condicionales de los grupos tienen cierta forma paramétrica conocida y se reduce a estimar los parámetros.
- No paramétricos: no asumen nada sobre el modo en el que se agrupan los objetos.

Crear predicciones

Una vez entrenado el modelo, los resultados se almacenan como un conjunto de patrones que se puede explorar o utilizar para realizar predicciones.

Puede crear consultas para devolver predicciones sobre si los nuevos datos se ajustan a los clústeres que se han detectado, o para obtener estadísticas descriptivas sobre los clústeres.

1.4.8. Algoritmo de árboles de decisión

Es un algoritmo de clasificación y regresión para su uso en el modelado de predicción de atributos discretos y continuos.

Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción.

Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión.

Si se define más de una columna como elemento de predicción, o si los datos de entrada contienen una tabla anidada que se haya establecido como elemento de predicción, el algoritmo genera un árbol de decisión independiente para cada columna de predicción.

Este algoritmo genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta.

Un problema común en los modelos de minería de datos es que el modelo se vuelve demasiado sensible a pequeñas diferencias en los datos de entrenamiento, en cuyo caso se dice que está sobreajustado o sobreentrenado.

Datos requeridos para los modelos de árboles de decisión

Cuando prepare los datos para su uso en un modelo de árboles de decisión, conviene que comprenda qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que estos se utilizan.

Los requisitos para un modelo de árboles de decisión son los siguientes:

- **Una columna de una sola clave:** cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
- **Una columna de predicción:** Se requiere al menos una columna de predicción. Puede incluir varios atributos de predicción en un modelo y pueden ser de tipos diferentes, numérico o discreto. Sin embargo, el incremento del número de atributos de predicción puede aumentar el tiempo de procesamiento.
- **Columnas de entrada:** Se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entrada afecta al tiempo de procesamiento.

Crear predicciones

Una vez procesado el modelo, los resultados se almacenan como un conjunto de patrones y estadísticas que se pueden usar para explorar las relaciones o para realizar predicciones.

1.4.9. Algoritmo de regresión lineal

Es una variación del algoritmo de árboles de decisión que ayuda a calcular una relación lineal entre una variable independiente y otra dependiente y, a continuación, utilizar esa relación para la predicción.

La relación toma la forma de una ecuación para la línea que mejor represente una serie de datos.

Se invoca un caso especial del algoritmo de árboles de decisión, con parámetros que restringen el comportamiento del algoritmo y requieren ciertos tipos de datos de entrada. Además, en un modelo de regresión lineal, el conjunto de datos completo se utiliza para calcular las relaciones en el paso inicial, mientras que en

un modelo de árboles de decisión estándar los datos se dividen repetidamente en árboles o subconjuntos más pequeños.

Datos requeridos para los modelos de regresión lineal

Cuando se preparan datos para utilizarse en un modelo de regresión lineal, se deben entender los requisitos del algoritmo determinado. Esto incluye saber cuántos datos se necesitan y cómo se utilizan. Los requisitos para este tipo de modelo son los siguientes:

- **Una columna de una sola clave:** cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
- **Una columna de predicción:** Se requiere al menos una columna de predicción. Se pueden incluir varios atributos de predicción en un modelo, pero deben ser tipos de datos numéricos continuos. No se puede utilizar un tipo de datos de fecha y hora como atributo de predicción aunque el almacenamiento nativo para los datos sea numérico.
- **Columnas de entrada:** Deben contener datos numéricos continuos y se les debe asignarse el tipo de datos adecuado.

Crear predicciones

Una vez procesado el modelo, los resultados se almacenan como un conjunto de estadísticas junto con la fórmula de regresión lineal, que se puede utilizar para calcular tendencias futuras.

Además de crear un modelo de regresión lineal seleccionando el algoritmo de regresión lineal, si el atributo de predicción es un tipo de datos numéricos continuo, puede crear un modelo de árbol de decisión que contenga regresiones. En este caso, el algoritmo dividirá los datos cuando encuentre puntos de

separación adecuados, pero en cambio creará una fórmula de regresión para algunas regiones de datos.

1.4.10. Algoritmo de regresión logística

La regresión logística es una técnica estadística conocida que se usa para modelar los resultados binarios.

Existen varias implementaciones de regresión logística en la investigación estadística, que utilizan diferentes técnicas de aprendizaje. El algoritmo de Regresión logística se ha implementado utilizando una variación del algoritmo de Red neuronal. Este algoritmo comparte muchas de las cualidades de las redes neurales pero es más fácil de entrenar.

Una de las ventajas de la regresión logística es que el algoritmo es muy flexible, puede tomar cualquier tipo de entrada y admite varias tareas analíticas diferentes:

- Usar datos demográficos para realizar predicciones sobre los resultados, como el riesgo de contraer una determinada enfermedad.
- Explorar y ponderar los factores que contribuyen a un resultado. Por ejemplo, buscar los factores que influyen en los clientes para volver a visitar un establecimiento.
- Clasificar los documentos, el correo electrónico u otros objetos que tengan muchos atributos.

La regresión logística es un método estadístico conocido que se usa para determinar la contribución de varios factores a un par de resultados. La implementación usa una red neuronal modificada para modelar las relaciones entre las entradas y los resultados. Se mide el efecto de cada entrada en el resultado y se ponderan las diversas entradas en el modelo acabado. El nombre regresión logística procede del hecho de que la curva de los datos se comprime mediante una transformación logística para minimizar el efecto de los valores extremos.

Datos requeridos para los modelos de regresión logística

Al preparar los datos para su uso en el entrenamiento de un modelo de regresión logística, conviene comprender qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que estos datos se utilizan.

Los requisitos para un modelo de regresión logística son los siguientes:

- **Una columna de una sola clave:** cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
- **Columnas de entrada**: cada modelo debe tener al menos una columna de entrada que contenga los valores que se utilizan como factores en el análisis. Puede tener tantas columnas de entrada como desee, pero dependiendo del número de valores existentes en cada columna, la adición de columnas adicionales podría aumentar el tiempo necesario para entrenar el modelo.
- **Al menos una columna de predicción:** el modelo debe contener al menos una columna de predicción de cualquier tipo de datos, incluidos datos numéricos continuos. Los valores de la columna de predicción también se pueden tratar como entradas del modelo, o se puede especificar que solo se utilicen para las predicciones. No se admiten tablas anidadas en las columnas de predicción, pero se pueden usar como entradas.

Crear predicciones

Una vez entrenado el modelo, puede crear consultas en el contenido del modelo para obtener los coeficientes de regresión y otros detalles, o puede usar el modelo para realizar predicciones.

1.4.11. Algoritmo Bayes naive

Es un algoritmo de clasificación basado en teoremas de Bayes para su uso en el modelado de predicción. La palabra naïve (ingenuo en inglés) del término Bayes naive proviene del hecho que el algoritmo utiliza técnicas Bayesianas pero no tiene en cuenta las dependencias que puedan existir.

Desde el punto de vista computacional, el algoritmo es menos complejo que otros algoritmos y, por tanto, resulta útil para generar rápidamente modelos de minería de datos que detectan las relaciones entre las columnas de entrada y las columnas de predicción. Puede utilizar este algoritmo para realizar la exploración inicial de los datos y, más adelante, aplicar los resultados para crear modelos de minería de datos adicionales con otros algoritmos más complejos y precisos desde el punto de vista computacional.

Datos requeridos para los modelos Bayes naive

Al preparar los datos para su uso en un modelo de entrenamiento Bayes naive, conviene comprender qué requisitos son imprescindibles para el algoritmo, incluidos el volumen de datos necesario y la forma en que estos datos se utilizan.

Los requisitos para un modelo Bayes naive son los siguientes:

- **Una columna de una sola clave :** cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
- **Columnas de entrada** en un modelo Bayes Naive, todas las columnas deben ser discretos o discretizados columnas.
    - En un modelo Bayes naive, también es importante asegurarse de que los atributos de entrada sean independientes unos de otros. Esto es particularmente importante al utilizar el modelo para la predicción.
    - El motivo es que, si utiliza dos columnas de datos que ya están estrechamente relacionadas, el efecto sería multiplicar la

influencia de esas columnas, lo que puede ocultar otros factores que influyen en el resultado.

- Al contrario, la capacidad del algoritmo de identificar las correlaciones entre las variables es útil cuando está explorando un modelo o conjunto de datos, para identificar las relaciones entre las entradas.

- **Al menos una columna de predicción:** el atributo de predicción debe contener valores discretos o discretizados.
  - Los valores de la columna predecible se pueden tratar como entradas. Este ejercicio puede ser útil si explora un nuevo conjunto de datos, para encontrar relaciones entre las columnas.

Realizar predicciones

Una vez entrenado el modelo, los resultados se almacenan como un conjunto de patrones que se puede explorar o utilizar para realizar predicciones.

Puede crear consultas para devolver las predicciones sobre cómo se relacionan los nuevos datos con el atributo de predicción, o puede recuperar estadísticas que describan las correlaciones que ha hallado el modelo.

1.4.12. Algoritmo de red neuronal

Combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción y utiliza los datos de entrenamiento para calcular las probabilidades. Posteriormente, puede usar estas probabilidades para la clasificación o la regresión, así como para predecir un resultado del atributo de predicción basándose en los atributos de entrada.

Los modelos de minería de datos construidos con el algoritmo de red neuronal de Microsoft pueden contener varias redes, en función del número de columnas que se utilizan para la entrada y la predicción, o solo para la predicción. El número de redes que contiene un único modelo de minería de datos depende

del número de estados que contienen las columnas de entrada y las columnas de predicción que utiliza el modelo.

El algoritmo de red neuronal crea una red que se compone de tres niveles de neuronas. Estos niveles son: un nivel de entrada, un nivel oculto opcional y un nivel de salida.

- **Nivel de entrada:** neuronas de entrada definen todos los valores de atributo de entrada para el modelo de minería de datos y sus probabilidades.
- **Nivel oculto:** las neuronas ocultas reciben entradas de neuronas de entrada y proporcionan salidas a las neuronas de salida. El nivel oculto es donde se asignan pesos a las distintas probabilidades de las entradas. Un peso describe la relevancia o importancia de una entrada determinada para la neurona oculta. Cuanto mayor sea el peso asignado a una entrada, más importante será el valor de dicha entrada. Los pesos pueden ser negativos, lo que significa que la entrada puede desactivar, en lugar de activar, un resultado concreto.
- **Nivel de salida:** neuronas de salida representan valores de atributo de predicción para el modelo de minería de datos.

Datos requeridos para los modelos de red neuronal

El modelo de red neuronal debe contener una columna de clave, una o más columnas de entrada y una o más columnas de predicción.

Los modelos de minería de datos que usan el algoritmo de red neuronal de Microsoft están muy influenciados por los valores que se especifican en los parámetros disponibles para el algoritmo. Los parámetros definen cómo se muestrean los datos, cómo se distribuyen o cómo se espera que estén distribuidos en cada columna, y cuándo se invoca la selección de características para limitar los valores usados en el modelo final.

Crear predicciones

Una vez procesado el modelo, puede usar la red y los pesos almacenados dentro de cada nodo para realizar predicciones. Un modelo de red neuronal admite el análisis de regresión, de asociación y de clasificación. Por lo tanto, el significado de cada predicción puede ser diferente. También puede consultar el propio modelo, revisar las correlaciones encontradas y recuperar las estadísticas relacionadas.

1.4.13. Algoritmo de clústeres de secuencia

Puede utilizar este algoritmo para explorar los datos que contiene los eventos que se pueden vincular siguiendo rutas o secuencias. El algoritmo encuentra las secuencias más comunes mediante la agrupación, o agrupación en clústeres, de las secuencias que son idénticas.

A continuación se incluyen algunos ejemplos de datos que contienen secuencias que se podrían utilizar para la minería de datos, para ofrecer una visión general de problemas comunes o escenarios empresariales:

- Rutas de clics que se crean cuando los usuarios navegan o examinan un sitio web.
- Registros que enumeran eventos que preceden a un incidente, como un disco duro erróneo o interbloqueos del servidor.
- Registros de transacciones que describen el orden en el que un cliente agrega elementos a una cesta de la compra de un comerciante en línea.
- Registros que siguen las interacciones del cliente (o paciente) a lo largo del tiempo, para predecir cancelaciones del servicio u otros malos resultados.

Este algoritmo es similar en muchas maneras al algoritmo de clústeres. Sin embargo, en lugar de encontrar clústeres de casos que contienen atributos similares, el algoritmo de clústeres de secuencia encuentra clústeres de casos que contienen rutas similares en una secuencia.

Es un algoritmo híbrido que combina técnicas de agrupación en clústeres con el análisis de cadenas de Markov para identificar los clústeres y sus secuencias. Una de las marcas distintivas del algoritmo de clústeres de secuencia es que utiliza los datos de las secuencias. Estos datos suelen representar una serie de eventos o transiciones entre los estados de un conjunto de datos, como una serie de compras de productos o los clics en web para un usuario determinado.

El algoritmo examina todas las probabilidades de transición y mide las diferencias, o las distancias, entre todas las posibles secuencias del conjunto de datos con el fin de determinar qué secuencias es mejor utilizar como entradas para la agrupación en clústeres. Después de que el algoritmo ha creado la lista de secuencias candidatas, usa la información de las secuencias como entrada para el método EM de agrupación en clústeres.

Datos requeridos para los modelos de clústeres de secuencias

Al preparar los datos para usarlos en el entrenamiento de un modelo de agrupación en clústeres de secuencia, conviene comprender qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que se usan los datos.

Los requisitos de un modelo de agrupación en clústeres de secuencia son los siguientes:

- **Una columna de clave única**: Un modelo de agrupación en clústeres de secuencia necesita una clave que identifique los registros.
- **Una columna de secuencia:** para los datos de la secuencia, el modelo debe tener una tabla anidada que contenga una columna de identificador de secuencia. El identificador de secuencia puede ser cualquier tipo de datos ordenable. Por ejemplo, puede usar el identificador de una página web, un número entero o una cadena de texto, con tal de que la columna identifique los eventos en una secuencia. Solo se admite un identificador de secuencia por cada secuencia y un tipo de secuencia en cada modelo.

- **Atributos opcionales no relacionados con la secuencia:** el algoritmo admite la incorporación de otros atributos que no tengan que ver con las secuencias. Estos atributos pueden incluir las columnas anidadas.

Crear predicciones

Una vez entrenado el modelo, los resultados se almacenan como un conjunto de patrones. Puede usar las descripciones de las secuencias más comunes en los datos para predecir el siguiente paso probable de una nueva secuencia. Sin embargo, dado que el algoritmo incluye otras columnas, puede usar el modelo resultante para identificar las relaciones entre los datos de las secuencias y las entradas que no son secuenciales. Por ejemplo, si agrega datos demográficos al modelo, puede realizar predicciones para grupos concretos de clientes. Las consultas de predicción se pueden personalizar para que devuelvan un número variable de predicciones o estadísticas descriptivas.

1.4.14. Técnicas de minería de datos.

A continuación se describen algunas técnicas de minería de datos para llevar a cabo las tareas anteriormente mencionadas:

- Modelización estadística paramétrica.
- Modelización estadística no paramétrica.
- Reglas de Asociación y Dependencia
- Métodos Bayesianos.
- Árboles de decisión y sistemas de reglas.
- Redes neuronales artificiales.
- Algoritmos de clusteing o agrupamiento

- Algoritmos de clasificación
- Algoritmos de Asociación
- Algoritmo para la Selección de atributos

De las cuales se ha seleccionado las que ofrecen los mejores resultados, para un modelo eficaz, se detallan a continuación:

1.4.14.1. Reglas de Asociación y Dependencia

Esta técnica consiste en que mediante reglas se expresan patrones de comportamiento entre los datos de las clases del dominio en función de la aparición conjunta de valores de dos o más atributos (Orallo et al., 2004). La característica principal de estas reglas es que tratan con atributos nominales es decir que puede tener un valor de un conjunto de valores establecidos, por ejemplo el atributo género (masculino, femenino).

- **A Priori**

  Se usa en minería de datos para encontrar reglas de asociación en un conjunto de datos. Este algoritmo se basa en el conocimiento previo o “a priori” de los conjuntos frecuentes, esto sirve para reducir el espacio de búsqueda y aumentar la eficiencia

1.4.14.2. Árboles de decisión y sistemas de reglas

La técnica basada en árboles de decisión es quizás el método más fácil de utilizar y de entender. (Orallo et al., 2004). Un árbol decisión es una estructura jerárquica que está formado por un conjunto de nodos, en donde cada nodo establece una condición o regla la misma que puede retornar verdadero o falso según los valores de los atributos que se desean analizar, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde el nodo raíz (superior) del árbol hasta alguno de sus nodos hojas (inferior).(Aronson, Liang, & Turban, 2005). Las tareas que utilizan este tipo de técnica son la:

clasificación, regresión y agrupamiento. Una de las ventajas de los árboles de decisión es que se puede llegar a una sola acción o decisión a tomar. (Orallo et al., 2004)

1.4.14.3. Algoritmos de clustering o agrupamiento

El presente algoritmo es utilizado para crear grupos de datos, con características similares.

- K-Means Uno de los algoritmos más utilizados para el agrupamiento de datos, es el K-Medias o KMeans, por ser uno de los más veloces y eficaces. Dicho algoritmo trabaja con un método de agrupamiento por vecindad, en el que se parte de un número determinado de prototipos de un conjunto de ejemplos a agrupar sin etiquetar.

El propósito de K-Means es ubicar a los prototipos o centros en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares (Moody, 1989).

1.4.14.4. Algoritmos de clasificación

El presente algoritmo es utilizado para clasificar un conjunto de datos, dentro de una clase especifica.

- **J48** es una implementación open source en lenguaje de programación Java del algoritmo C4.5 en la herramienta weka, el presente algoritmo permite generar un árbol de decisión, a través de los datos ingresados, seleccionando el mejor atributo que clasifique a los datos. (Wilford Ingrid, 2008). Es uno de los más utilizados en minería de datos, permite trabajar con valores

continuos para los atributos, separando los posibles resultados en las ramas respectivas.

El presente algoritmo genera un árbol de decisión de forma recursiva al considerar el criterio de la mayor proporción de ganancia de información (Hernández & Ferri, 2006) es decir, elige al atributo que mejor clasifica a los datos dentro de una categoría definida. Está formado por:

- Nodos: Nombres de los Atributos seleccionados.
- Ramas: Valores de los determinados atributos.
- Hojas: Conjuntos de datos clasificados y etiquetados con el nombre de la clase.

1.4.15. Correspondencia entre tareas, técnicas y algoritmos

En la Tabla 1.1, se muestra la correspondencia que existe entre las técnicas de minería, con las tareas y los algoritmos.

	Predictivas		Descriptiva		
Técnica (algoritmo)	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		
Métodos basados en casos y en vecindad (Two-step, COBWED)	X	X	X	X	X
Máquinas de vectores soporte	X	X	X		

Tabla 2: Correspondencia entre técnicas, algoritmos y las tareas

Fuente: (Ordoñez, 2012)

1.4.16. Modelo de minería de datos

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

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

- **Pronóstico**: cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.
- **Riesgo y probabilidad**: elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.
- **Recomendaciones**: determinación de los productos que se pueden vender juntos y generación de recomendaciones.
- **Búsqueda de secuencias**: análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.
- **Agrupación**: distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

1.4.17. Herramientas de minerías de datos

Las herramientas de minería de datos facilitan el desarrollo de los modelos para la extracción de conocimiento de un dominio establecido, dichas herramientas contienen los algoritmos específicos para la aplicación de técnicas de minería de datos, se los puede utilizar mediante una interfaz gráfica. Algunas herramientas tanto comerciales como de libres distribución, para el desarrollo de modelos de minería de datos: (Pérez C., 2007) comerciales: Intelligent Miner / DB2 Data Warehouse Edition (IBM), Enterprise Miner (SAS), DataEngine o de código libre como Weka.

1.4.17.1. Weka (Waikato environment for knowledge analysis)

Es una herramienta visual de libre distribución desarrollada por los investigadores de la Universidad de waikato en Nueva Zelanda. Sus principales característica son: Acceso de los datos desde un archivo en formato ARFF (es un archivo de texto plano); preprocesador de datos (selección, transformación de atributos); visualización del entorno y aplicación de técnicas de aprendizaje. (Holmes, Donkin, & Witten,)

El ambiente de Waikato para el análisis del conocimiento (WEKA) sobre la necesidad percibida de una mesa de trabajo unificada que permitiría a los investigadores un fácil acceso al estado de la técnica en aprendizaje automático. En el momento de inicio del proyecto en 1992, los algoritmos de aprendizaje estaban disponibles en varios idiomas, para usar en diferentes plataformas, y operado en una variedad de formatos de datos. La tarea de recolectar juntos esquemas de aprendizaje para un estudio comparativo en una colección de los conjuntos de datos fue desalentador en el mejor de los casos. Fue imaginado que WEKA no solo proporcionaría una caja de herramientas de aprendizaje y algoritmos, pero también un marco dentro del cual los investigadores podrían implementar nuevos algoritmos sin tener que preocuparse con infraestructura de soporte para la manipulación de datos y evaluación del esquema.

Hoy en día, WEKA es reconocido como un sistema histórico en minería de datos y aprendizaje automático. Ha logrado aceptación generalizada en círculos académicos y empresariales, y se ha convertido en una herramienta ampliamente utilizada para la minería de datos investigación El libro que lo acompaña es popular libro de texto para la minería de datos y se cita con frecuencia en la máquina publicaciones de aprendizaje, de haber alguno, de este éxito han sido posibles si el sistema no ha sido lanzado software de código abierto. Dar a los usuarios acceso gratuito a la fuente.

El código ha permitido que la próspera comunidad se desarrolle y facilite la creación de muchos proyectos que incorporan o extender WEKA.(Hall et al., 2009)

WEKA dispone de 4 interfaces de usuario distintos, que se pueden elegir después de lanzar la aplicación completa, son:

- **Simple CLI**: interfaz en modo texto.
- **Explorer:** interfaz gráfico básico.

  El Explorer permite visualizar y aplicar distintos algoritmos de aprendizaje a un conjunto de datos. Cada una de las tareas de minería de datos viene representada por una pestaña en la parte superior. Estas son: Preprocess: visualización y preprocesado de los datos (aplicación de filtros)
    - Classify: Aplicación de algoritmos de clasificación y regresión
    - Cluster: Agrupación
    - Associate: Asociación
    - Select Attributes: Selección de atributos
    - Visualize: Visualización de los datos por parejas de atributos
- **Experimenter:** interfaz gráfico con posibilidad de comparar el funcionamiento de diversos algoritmos de aprendizaje, sirve para aplicar varios algoritmos de aprendizaje automático sobre distintos conjuntos de datos y determinar de manera estadística cual se comporta mejor.
- **Knowledge Flow:** interfaz gráfico que permite interconectar distintos algoritmos de aprendizaje en cascada, creando una red.

1.4.17.2. Spss Clementine

Es uno de los sistemas de Minería de Datos más conocidos. Posee una herramienta visual desarrollada por ISL que tiene una arquitectura cliente /

servidor, se caracteriza por: Acceso a datos (fuentes de datos archivos ASCII); procesamiento de datos; aplicación de técnicas de aprendizaje como (redes neuronales, reglas de asociación), incorpora técnicas de evaluación de modelos visualización de resultados como (histogramas, diagramas de dispersión). (Oocities, 2004)

1.4.17.3. KEPLER

Sistema desarrollador y transformado en una KEPLER herramienta comercial distribuida por Dialogis. Posee múltiples modelos de análisis.

Sus principales herramientas de aprendizaje son:

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

Así mismo permite el pre procesado de datos, la elección de un modelo o la manipulación de la representación gráfica de los modelos obtenidos. (Oocities, 2004)

1.5. FUNDAMENTACIÓN LEGAL

Esta investigación se fundamenta en las siguientes leyes y reglamentos:

MARCO JURÍDICO

LEY ORGANICA DE EDUCACION SUPERIOR

Art. 84. Requisitos para aprobación de cursos y carreras.- Los requisitos de carácter académico y disciplinario necesarias para la aprobación de cursos y carreras, constarán en el Reglamento de Régimen Académico, en los respectivos estatutos, reglamentos y demás normas que rigen al Sistema de Educación Superior. Solamente en casos establecidos excepcionalmente en el estatuto de cada institución, un estudiante podrá matricularse hasta por tercera ocasión en una misma materia o en el mismo ciclo, curso o nivel académico. En la tercera matrícula de la materia, curso o nivel académico no existirá opción a examen de gracia o de mejoramiento.

VIGENTE

Que, el Art. 84 de la LOES, determina que los requisitos de carácter académico y disciplinario necesarios para la aprobación de cursos y carreras, constarán en el Reglamento de Régimen Académico, en los respectivos estatutos, reglamentos y demás normas que rigen el Sistema de Educación Superior;

Que, la Disposición General Primera de la Ley Orgánica de Educación Superior, dispone que todas las entidades de educación superior del país, deben adecuar su normativa interna al nuevo marco jurídico constitucional y legal, a efectos de guardar plena concordancia con ese nuevo entorno;

Que, el Consejo de Educación Superior – CES, en uso de sus atribuciones legales, expidió el Reglamento de Régimen Académico, mediante Resolución RPC-SE-13-No.051-2013, de fecha 21 de noviembre del 2013; y, posteriormente, efectuó su última reforma mediante Resolución RCP-SO-13-No.146-2014, de 09 de abril de 2014.

Que, La Disposición General Primera del Reglamento de Régimen Académico expedido por el CES, determina que, "las IES deberán asegurar, mediante normativa y políticas internas efectivas, que las relaciones entre docentes y estudiantes se desenvuelvan en términos de mutuo respeto y, en general, en condiciones adecuadas para una actividad académica de calidad. Las IES deberán vigilar, especialmente, que los derechos estudiantiles establecidos en la LOES y en sus estatutos sean respetados, de forma que no se retrase ni se distorsione arbitrariamente la formación y titulación académica y profesional, y, en particular, que se cumpla lo determinado en el artículo 5, literal a) de la LOES. La violación de este derecho estudiantil por parte del personal administrativo o académico será sancionada conforme a la normativa interna de la respectiva IES".

Que, el Consejo de Educación Superior – CES, en uso de sus atribuciones legales, expidió el Reglamento de Armonización de la Nomenclatura de Títulos Profesionales y Grados Académicos que confieren las instituciones de Educación Superior del Ecuador, mediante Resolución RPC-SO-27-No.289-20 14, de fecha 16 de julio de 2014.

Que, la Universidad Laica Eloy Alfaro de Manabí – ULEAM, dentro de su planificación, ha programado la adecuación de toda su normativa interna, al nuevo marco constitucional, legal y reglamentario vigente, con fines de prepararse permanentemente para los procesos de evaluación y acreditación dispuestos por el CEAACES;

Que, para la Universidad Laica Eloy Alfaro de Manabí – ULEAM, el Reglamento de Régimen Académico constituye una herramienta jurídico – académica fundamental para regular todo lo que conlleva el quehacer

universitario de docencia, investigación y vinculación con la sociedad, en la búsqueda de la formación profesional integral con excelencia académica, en un entorno de respeto a la diversidad, la interculturalidad, la solidaridad con el medioambiente y la identidad cultural;

Que, en el vigente Estatuto de la Universidad Laica “Eloy Alfaro” de Manabí, faculta al Consejo Universitario, expedir los reglamentos generales internos que se requieran para su funcionamiento; y, en ejercicio de sus facultades legales y estatutarias:

ESTATUTO VIGENTE DE LA ULEAM

ART-121.-LIMITACIONES DE DERECHO DE MATRICULA

Ningún/a estudiante podrá obtener matricula por tercera ocasión en una misma materia, o en el mismo ciclo, curso o nivel académico.

La tercera matricula, curso o nivel académico, será considerada de excepción y procederá únicamente, cuando el/la estudiante demuestre documentadamente que no ha reprobado por faltas o inasistencia a clases en más de tres asignaturas; enfermedades graves o catastróficas; cambio de domicilio por actividades laborales; por haber emigrado del país y que no tiene registrada ni asistencia, ni calificaciones de algún examen parcial o final, en uno de los dos años anteriores.

El/la estudiante solicitara la tercera matricula al Consejo de Facultad, Extensión o Escuela Integrada , que resolverá previo informe favorable de la Secretaria de la Unidad Académica. Cumplidos estos requisitos, la Secretaria General autorizara la legalización de la misma.

Ningún/a estudiante podrá matricularse en dos Unidades Académicas en un mismo año lectivo, salvo el caso de los/las alumnos/as que hayan obtenido muy buenas calificaciones en el periodo académico anterior y si los horarios le permiten asistir a clases regularmente, lo cual será verificado por la Secretaria General, con las certificaciones de las Secretarias de la Unidades Académicas donde realizare sus estudios.

REGLAMETO DEL REGIMEN ACADEMICO DE LA ULEAM

CAPÍTULO V

MATRÍCULAS

Artículo 27.- Proceso de matriculación.- La matrícula es el acto de carácter académico-administrativo, mediante el cual una persona adquiere la condición de estudiante, a través del registro de las asignaturas, cursos o sus equivalentes, en un período académico determinado, conforme a lo establecido en el Título Undécimo, Capítulos I y II del Estatuto en vigencia de la ULEAM, referentes al ingreso de estudiantes, requisitos previos y a la matriculación. La condición de estudiante se mantendrá hasta el inicio del nuevo periodo académico ordinario o hasta su titulación.

Artículo 28.- Tipos de matrícula.- Dentro del Sistema de Educación Superior, se establecen los siguientes tipos de matrícula:

a. Matrículas ordinarias.- Se realizan en el plazo de 15 días anteriores al inicio de cada semestre.

b. Matrículas extraordinarias.- Se realizan en el plazo máximo de 15 días posteriores a la culminación del período de matrículas ordinarias.

c. Matrícula especial.- Es aquella que, en casos individuales excepcionales, otorga el Consejo Universitario de la ULEAM, para quienes, por circunstancias de caso fortuito o fuerza mayor debidamente documentadas, no se hayan matriculado de manera ordinaria o extraordinaria. Esta matrícula se podrá realizar hasta dentro de los 15 días posteriores a la culminación del período de matrícula extraordinaria.

En cualquier caso, los períodos de matriculación ordinario, extraordinario y especial deben concluir hasta fines de los meses de abril para el primer período académico y octubre para el segundo período académico.

Para los programas de posgrado, los períodos de matriculación ordinario y extraordinario serán establecidos por la Dirección de Postgrados de la ULEAM y aprobados por el Consejo Universitario.

Se considera como inicio de la carrera o programa la fecha de la matriculación de la primera cohorte de los mismos.

Art. 29.- Se consideran estudiantes regulares de la ULEAM, quienes se encuentren matriculados en al menos el 60% de las asignaturas, cursos o sus equivalentes, de la malla curricular, por cada periodo académico ordinario. Dependiendo de la carrera, y de lo establecido en este Reglamento, podrán existir estudiantes con dedicación a tiempo parcial, siempre que se hayan matriculado en, al menos, el 60% de las asignaturas, cursos o sus equivalentes, porcentaje calculado en números enteros, con la aproximación habitual, de 0.5 en adelante sube al inmediato superior.

Art. 30.- Proceso de matriculación: Para las carreras de grado de la ULEAM se observarán las siguientes normas generales:

1. Los estudiantes se matricularán por asignaturas, respetando las secuencias establecidas en las correspondientes mallas curriculares.

2. La matriculación se realizará en línea (“on line”), utilizando la plataforma informática de la ULEAM; para lo cual, todas las asignaturas aprobadas por cada estudiante deberán estar debidamente registradas en la base de datos académica, y el estudiante solo podrá tomar las asignaturas que la malla lo permita, de acuerdo a sus secuencias, incluso en diferentes niveles de la carrera y en diferentes horarios cuando existan, siempre que los horarios de las asignaturas que tome, no se superpongan.

3. Todos los documentos exigidos en el proceso de matriculación, deberán ser escaneados y subidos a la misma plataforma informática de la ULEAM;

incluidos los procedimientos legales de firma electrónica, cuando sea necesario. Sin perjuicio de que, posteriormente, estos documentos deben ser entregados impresos en las Secretarías de las Unidades Académicas para su verificación. En determinadas circunstancias se podrá exigir que esta documentación se presente notariada.

4. Los procedimientos de admisión en el primer semestre de cualquier carrera, son los determinados en el Sistema Nacional de Admisión y Nivelación, y en las correspondientes disposiciones del Estatuto en vigencia de la ULEAM. En cualquier caso, el Departamento de Admisión y Nivelación de la ULEAM, remitirá de manera electrónica la lista de estudiantes que pueden matricularse directamente en el primer nivel de las diferentes carreras y de aquellos que deben realizar previamente el curso de nivelación.

PLAN NACIONAL BUEN VIVIR

OBJETIVO 5: Planificando el Futuro

5.1.1. Uno de los grandes retos del Buen Vivir es mejorar la calidad de vida de los ecuatorianos. Para lograrlo, la diversificación productiva y el crecimiento de la economía deben dirigirse al cumplimiento progresivo de los derechos en educación, salud, empleo y vivienda, la reducción de la inequidad social, y la ampliación de las capacidades humanas en un entorno participativo y de creciente cohesión social, con respeto a la diversidad cultural.

1.6. CONCLUSIONES RELACIONADAS AL MARCO TEÓRICO EN REFERENCIA AL TEMA DE INVESTIGACIÓN

Una vez obtenido el conocimiento sobre las técnicas, tareas y herramientas de la minería de datos, se concluyó que posee importantes ventajas para poder descubrir patrones de comportamiento de un estudiante desertor, ya que brinda un alto valor agregado para el análisis y la generación del nuevo conocimiento. Por tal razón se aplicó esta metodología para el desarrollo del modelo predictivo ya propuesto.

Como parte de la investigación se revisaron proyectos relacionados a la temática de esta tesis, lo que permitió conocer estructuras y modelos que se realizaron con datos de estas investigaciones dando como resultado patrones o tendencias que estos generaron.

CAPITULO II

DIAGNÓSTICO O ESTUDIO DE CAMPO

2.1. INTRODUCCIÓN

Una vez examinado el marco teórico de investigación, es necesario expresar que en este capítulo estuvo presente la metodología de investigación, herramientas y técnicas que se utilizaron en la recolección de los datos que fueron necesarios y de interés en el proyecto.

Las tareas principales que se realizaron en este capítulo fueron las siguientes:

Revisión de la bibliografía relacionada a los tipos de investigación, para su aplicación adecuada a este proyecto y establecer la población que fue objeto de estudio en función de los objetivos.

Recolección de datos y organización de la información que se obtuvo para la implementación del modelo, fue obtenido de archivos en secretaria de la institución.

Indicar los métodos y técnicas para elaborar el plan de recolección de los datos. De igual forma se seleccionó la muestra probabilística para emplear una encuesta y análisis documental, que sirvió de base para las mediciones de los objetivos. Se muestra el resultado de la investigación mediante barras estadísticas y finalmente se hace un análisis de esta información.

La investigación se realiza para llegar a un análisis de la situación actual en cuanto a la retención de estudiante que posee la facultad y como estos estudiantes y docentes interactúan dentro de este ámbito y así aplicar herramientas tecnológicas que nos permitan modelar estos datos.

2.2. TIPO DE INVESTIGACIÓN

Para llevar a cabo el presente proyecto integrador, se utilizó un tipo de investigación Mixta que combina los siguientes métodos según (Soto, 2015):

2.2.1 Investigación Aplicada

Las investigaciones aplicadas son la respuesta efectiva y fundamentada a un problema detectado y analizado. La investigación aplicada concentra su atención en las posibilidades fácticas de llevar a la práctica las teorías generales, y destina sus esfuerzos a resolver los problemas y necesidades que se plantean los hombres en sociedad en un corto, mediano o largo plazo. Es decir, se interesa fundamentalmente por la propuesta de solución en un contexto físico-social específico.

Por lo tanto se implementó un modelado de datos para el respectivo análisis de factores que afectan la deserción de estudiantes y que papel realizan las autoridades en cuanto a este problema social.

2.3. METODOS DE INVESTIGACIÓN

2.3.1. Método Analítico – Sintético

El **método analítico-sintétic**o es una combinación de dos formas de investigación que son utilizadas para desarrollar trabajos formales que requieren de un esquema para lograr los objetivos planteados. (Romero)

Se utilizó el método de investigación, porque la información recopilada a través de los instrumentos de recolección de datos, debió ser sometida a un análisis previo para concretar las ideas principales del estudio. Así mismo, se analizaron varias herramientas informáticas que permiten elaborar los modelos, para el desarrollo de este estudio se hizo uso del método de investigación sintético, en virtud de que la información recopilada debió ser sintetizada y analizada para su comprensión cabal; este método investigativo permitió extraer la información más

significativa de las encuestas aplicadas y la información de muestra que proporciono la Facultad de Ciencias Informáticas.

2.4. HERRAMIENTAS DE RECOLECCIÓN DE DATOS

Para realizar el presente proyecto integrador, se utilizó las siguientes herramientas para la recolección de datos.

2.4.1. Encuesta

Fue necesario la realización de una encuesta para los docentes que integran la Facultad de Ciencias Informáticas con el fin de analizar el estado de conocimiento actual del nivel de deserción y retención de sus estudiantes y la interacción entre ellos.

Una encuesta es una técnica de investigación que se efectúa mediante la elaboración de cuestionarios o entrevistas a una población (grupal o individual) con el propósito de recabar información de diferentes variantes de la realidad o para sugerir una hipótesis. Tomando en cuenta la importancia de esta herramienta se procede a realizar una encuesta vía online utilizando la tecnología de formulario de Google para encuestar a los docentes sobre la importancia del análisis de este problema social, para así aportar la validez de la problemática del proyecto.

Las Encuestas desarrolladas fueron de tipo opción múltiple, las cuales fueron elaborados con preguntas específicas de gran importancia para la elaboración y resolución de la problemática, las personas encuestadas tendrán que elegir entre dos opciones las cuales son sí o no, de esta manera tenemos mejor recolección cuantificada de las respuestas obtenidas.(Quispe Parí & Sánchez Mamani, 2011)

2.4.2. El análisis documental

El análisis documental es una operación sistémica e intelectual que recoge y estudia información de varios tipos de documentos. El calificativo de intelectual es debido a que el investigador realiza un

proceso de interpretación y análisis de la información de todos los documentos para luego resumirlos.

Esta herramienta fue empleada en esta investigación, ya que, basándose en información de proyectos e investigaciones anteriormente realizadas, documentos de internet y demás textos afines a esta temática, se pudo realizar un análisis de toda esta información y lograr obtener lo de mayor relevancia para el proyecto.

El análisis documental de la información necesaria para este proyecto se desarrolló en tres acciones que son:

- Buscar documentos existentes y disponibles sobre la temática en cuestión.
- Seleccionar aquellos documentos con mayor relevancia para el propósito de la investigación.
- Realizar una lectura minuciosa sobre el contenido de aquellos documentos que fueron seleccionados, con la finalidad de extraer ciertos elementos y asignarlos en notas secundarias.

2.5. FUENTES DE INFORMACIÓN DE DATOS

2.5.1. Fuentes Primarias

La fuente de información primaria, pertenece a la Secretaria de la Facultad de Ciencias Informáticas - ULEAM y es la que bridó la información y permitió realizar la investigación del presente proyecto integrador.

Se consideró la información que se recolectó mediante la aplicación de técnicas como son: análisis documental y encuestas dirigidas especialmente a los docentes de la institución.

2.5.2 Fuentes Secundarias

Las fuentes de información secundaria, son las que se necesitaron para conocer que estructura de minería de datos es la más adecuada para modelar la información obtenida, entre ellas tenemos a:

- **Libros**

  Esta fuente de información permitió obtener las bases teóricas para el modelado de datos, utilizando libros de fuentes confiables para llevar a cabo esta ejecución.

- **Internet**

  Mediante el uso de fuentes confiables de internet se logró conseguir información valiosa para la implementación del modelo mediante estructuras de mimería de datos, y las debidas configuraciones con la herramienta de apoyo weka.

- **Docente Guía de la Facultad Ciencias Informáticas FACCI.**

  El Ing. Jorge Pincay, fue el docente guía en el proceso de la realización teórica de este proyecto integrador.

2.6. INSTRUMENTAL OPERACIONAL

2.6.1. Estructura y características de los documentos de recolección de datos

Una encuesta es una forma de recolección de datos, en la cual el investigador por medio de un cuestionario de autoría propia busca obtener datos reales. Por lo tanto, se aplicó esta herramienta para obtener los datos realizando un conjunto de preguntas ordenadas que fueron dirigidas a una muestra representativa de la población de docentes pertenecientes a la "FACULTAD DE CIENCIAS INFORMATICAS - ULEAM". Se realizó la selección de preguntas que estaban acorde con la naturaleza de la investigación, y así darle más importancia a la problemática. (Ver anexo 1).

En la encuesta tratada dentro de esta investigación, se hizo uso de preguntas cerradas de tipo dicotómicas y opción múltiple, que fueron previamente elaboradas, así las personas encuestadas pudieron elegir como respuesta una

de las opciones. Esta forma de encuestar dio como resultado mayor facilidad de cuantificación con respecto a las respuestas dadas.

2.7. ESTRATEGIA OPERACIONAL PARA LA RECOLECCIÓN Y TABULACIÓN DE DATOS

2.7.1. Plan de recolección

Para elaborar el plan de recolección de datos que se utilizó en el proyecto se llevaron a cabo los siguientes procesos:

- Determinar quién será el encargado de aplicar las herramientas de recolección de datos: en este caso la persona encargada del proyecto, ya que conoce la problemática y busca la solución de la misma, para obtener los resultados deseados.
- Determinar cuándo se recolectará la información: se realizó las encuestas físicamente en la mañana ya que es más factible encontrar docentes en la institución, algunos factores influyentes que afecto esta recolección fue la disponibilidad de tiempo de los encuestados, dando como resultado una participación menos activa.
- Determinar dónde se aplicarán las herramientas: En el caso de esta investigación se realizó en la "FACULTAD DE CIENCIAS INFORMÁTICAS - ULEAM".
- Asegurar la correcta recolección de la información: Aunque se disponga de herramientas de calidad, la manera en cómo se recopila la información puede afectar en la calidad de los mismos, para realizar una recolección segura de la información la autora de la investigación estuvo presente cuando el encuestado registró las respuestas.

2.7.2. Tabulación y Análisis e interpretación de los datos

Se realizó el siguiente plan:

- Presentación de los cuadros estadísticos y sus gráficos correspondientes.
- Análisis e interpretación de los datos que presentan los cuadros, resaltando los datos más importantes.
- Relacionar dichos resultados con la teoría y los procedimientos investigados en el proyecto.

2.8. PLAN DE MUESTREO

2.8.1. Técnica de muestreo

Se realiza la técnica de muestreo simple que consiste en escoger un cierto grupo de la población (docente) para poder realizar las encuestas y así validar el porcentaje de respuestas para la elaboración y resolución de la problemática.

Se escoge el muestreo simple por ser la técnica, en que cada miembro de la población tiene la misma probabilidad de ser seleccionado como sujeto.

2.8.2. Tamaño de la muestra

El número de docente perteneciente a la Facultad de Ciencias Informáticas tiempo completo, medio tiempo, tiempo parcial y por contrato da un total de 37

Se obtiene el valor de la muestra de docentes, basados en información contenida en la paginan web de la Facultad de Ciencias Informáticas y así poder aplicar la formula estadística de muestreo.

$$n = \frac{Z^2.P.Q.N}{Z^2.P.Q + N.e^2}$$

Donde:

n = Tamaño de la muestra

N = Universo

E = Margen de error admisible.- En este caso se trabajara con el 5 %

P = Probabilidad de ocurrencia

Q = Probabilidad de no ocurrencia

Z = Confiabilidad 95%

Remplazamos valores:

$$n = \frac{(1.96)2\ (0.5)\ (0.5)\ (37)}{(1.96)2\ (0.5)\ (0.5) + (37)\ (0.05)2}$$

$$n = \frac{35.53}{0.9604 + 0.0925}$$

$$n = \frac{35.53}{1.0529}$$

$$n = 33.75 = 34$$

2.9. PRESENTACIÓN Y ANÁLISIS DE LOS RESULTADOS

2.9.1. Presentación y Descripción de los resultados obtenidos

Encuesta dirigida a los docentes que integran la Facultad de Ciencias Informáticas de la ULEAM

1.- ¿Cree usted que los posibles factores que provoquen la deserción estudiantil en la institución sean los siguientes?:

- ☐ individuales
- ☐ Académicos
- ☐ Institucionales
- ☐ Otros

Tabla 3: factores de deserción estudiantil

Alternativas	Frecuencia	Porcentaje
Individuales	5	15%
Académicas	15	44%
Institucionales	10	29%
Otros	4	12%
Total	34	100%

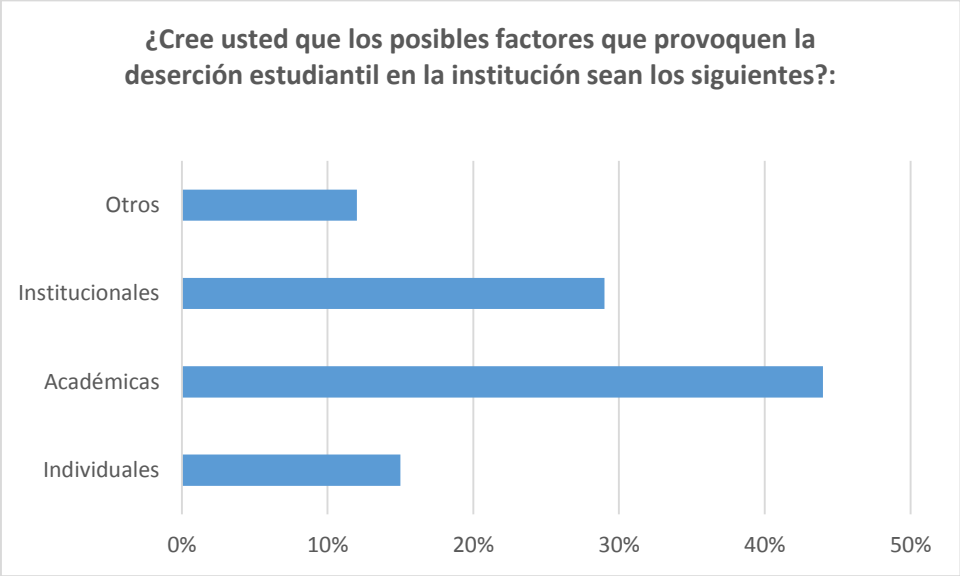


Ilustración 12: factores de deserción estudiantil

Considerando la información tabulada, se establece, que el 44% de los docentes considera que el factor académico es el que provoca la deserción estudiantil en su institución y el 12% mantiene que son otros los factores que provocan este problema entre ellos el económico y social.

2.- ¿Considera importante conocer cuáles son las razones académicas por las que el estudiante de esta facultad decide abandonar sus estudios?

- ☐ Si
- ☐ No

Tabla 4: Importancia de conocer las razones académicas de abandono de estudios

Alternativas	Frecuencia	Porcentaje
Si	34	100%
No	0	0%
Total	34	100%

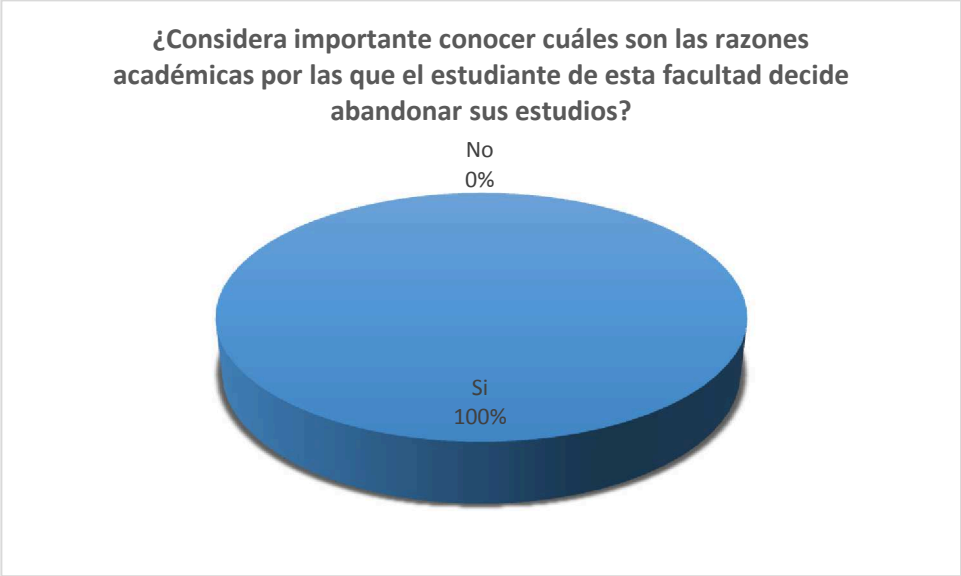


Ilustración 13: Importancia de conocer las razones académicas de abandono de estudio

Considerando la información tabulada, se establece, que el 100% de los docentes considera que SI es importante conocer las razones académicas por las que el estudiante de esta facultad decide abandonar sus estudios.

3.- ¿Cree usted que la baja tasa de retención estudiantil en el ámbito académico se da por las siguientes razones?:

- ☐ Rendimiento académico
- ☐ Métodos de estudio
- ☐ Orientación profesional
- ☐ Calidad del programa de estudio
- ☐ Otros

Tabla 5: Posibles razones de baja retención estudiantil en ámbito académico

Alternativas	Frecuencia	Porcentaje
Rendimiento académico	14	41%
Métodos de estudio	5	15%
Orientación profesional	10	29%
Calidad del programa de estudio	3	9%
Otros	2	6%
Total	34	100%

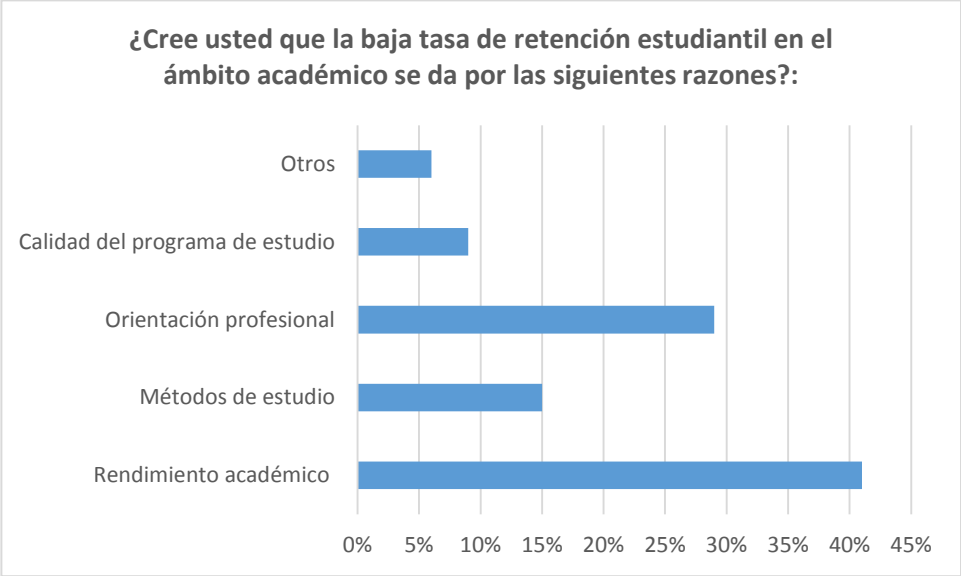


Ilustración 14: Posibles razones de baja retención estudiantil en ámbito académico

Considerando la información tabulada, se establece, que el 41% de los docentes considera que la posible razón de la baja retención estudiantil en el ámbito académico es el rendimiento del estudiante y el 6% mantiene que son otros los factores que provocan este problema.

4.- ¿Cree usted que la baja tasa de retención estudiantil en el ámbito institucional se da las siguientes razones?:

- ☐ Norma académica
- ☐ Modelos pedagógicos
- ☐ Recursos universitarios
- ☐ Perfil profesional de la carrera
- ☐ Relaciones con los profesores y otros estudiantes
- ☐ Otros

Tabla 6: Posibles razones de baja retención estudiantil en ámbito institucional

Alternativas	Frecuencia	Porcentaje
Norma académica	4	12%
Modelos pedagógicos	7	20%
Recursos universitarios	7	21%
Perfil profesional de la carrera	10	29%
Relaciones con los profesores y otros estudiantes	5	15%
Otros	1	3%
Total	34	100%

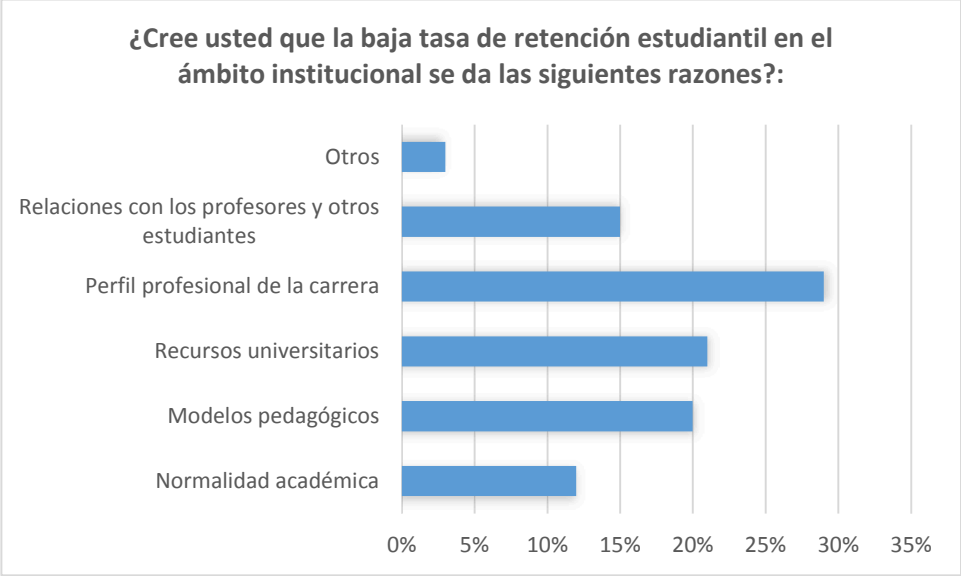


Ilustración 15: Posibles razones de baja retención estudiantil en ámbito institucional

Considerando la información tabulada, se establece, que el 29% de los docentes considera que la posible razón de la baja retención estudiantil en el ámbito institucional es el perfil profesional de la carrera y el 3% mantiene que son otros los factores que provocan este problema.

5.- ¿Cuáles según usted son las consecuencias que trae la deserción en la sociedad y para la institución?

- ☐ Trabajo mal remunerado
- ☐ Delincuencia
- ☐ Discriminación
- ☐ Otros

Tabla 7: Consecuencias de la deserción en la sociedad y la institución

Alternativas	Frecuencia	Porcentaje
Trabajo mal remunerado	18	53%
Delincuencia	10	29%
Discriminación	5	15%
Otros	1	3%
Total	34	100%

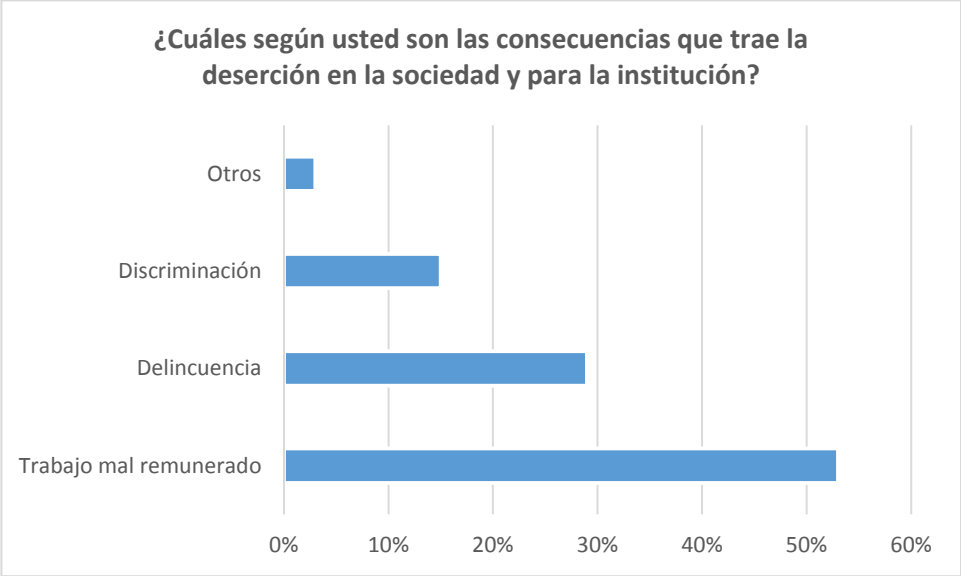


Ilustración 16: Consecuencias de la deserción en la sociedad y la institución

Considerando la información tabulada, se establece, que el 53% de los docentes considera que la posible consecuencia que trae la deserción en la sociedad es el trabajo mal remunerado y el 3% mantiene que son otros los factores que provocan estas consecuencias.

6.- ¿Considera importante este tipo de estudios sobre la deserción para la toma de decisiones en la institución?

- ☐ Si
- ☐ No

Tabla 8: Importancia del estudio de la deserción

Alternativas	Frecuencia	Porcentaje
Si	34	100%
No	0	0%
Total	34	100%

Ilustración 17: Importancia del estudio de la deserción

Considerando la información tabulada, se establece, que el 100% de los docentes considera que es muy importante realizar este tipo de estudios sobre la deserción.

7.- ¿Qué variables externas cree usted que causan deserción en su institución?

- ☐ Económicas
- ☐ Sociales
- ☐ Familiares
- ☐ Medicas

Tabla 9: Variables externas que causan deserción

Alternativas	Frecuencia	Porcentaje
Económicas	18	53%
Sociales	9	26%

Familiares	5	15%
Medicas	2	6%
Total	34	100%

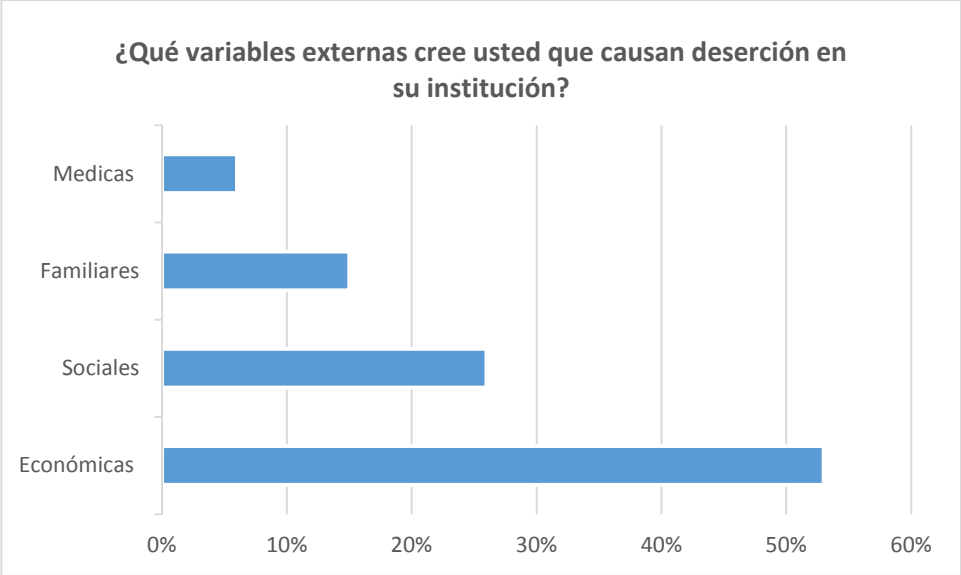


Ilustración 18: Variables externas que causan deserción

Considerando la información tabulada, se establece, que el 53% de los docentes considera que la variable externa predominante es la económica y el 6% mantiene que es medica aquella variable que afecta a los estudiantes.

8.- ¿Qué hace usted en el aula para evitar la deserción estudiantil?

- ☐ Motivar la asistencia
- ☐ Motivar el aprendizaje
- ☐ Utilizar técnicas innovadoras de estudio
- ☐ Otras

Tabla 10: Evitar deserción estudiantil

Alternativas	Frecuencia	Porcentaje
Motivar la asistencia	5	17%
Motivar el aprendizaje	9	36%
Utilizar técnicas innovadoras de estudio	11	30%

Otras	5	17%
Total	34	100%

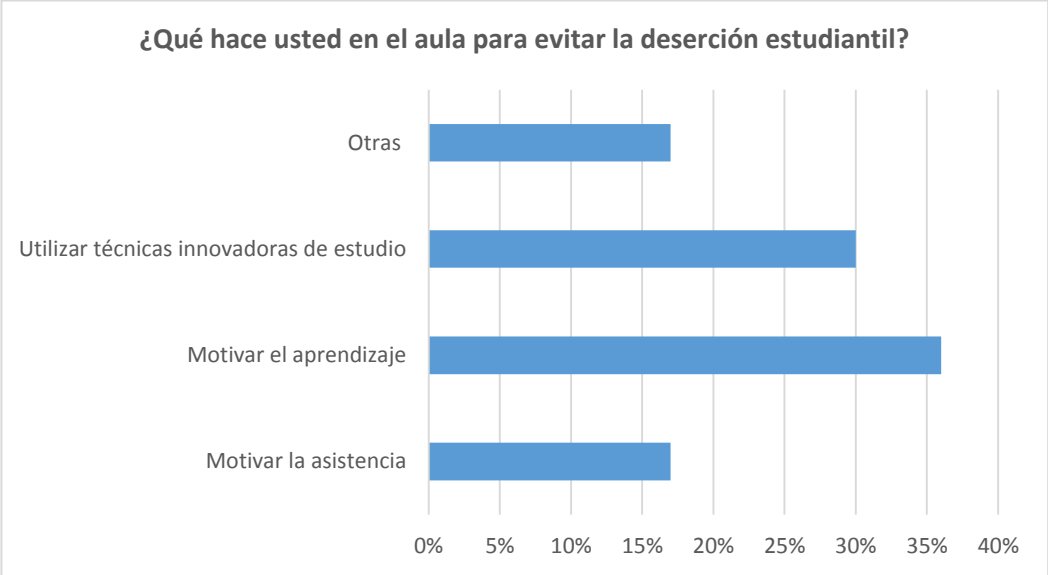


Ilustración 19: Evitar deserción estudiantil

Considerando la información tabulada, se establece, que el 36% de los docentes considera que utiliza técnicas innovadoras de estudio para evitar la deserción estudiantil y el 17% mantiene que motivar la asistencia es una opción aceptable, además de otras formas de retener al estudiante.

9.- ¿Considera importante tener datos estadísticos de los estudiantes que desertan?

- ☐ Si
- ☐ No

Tabla 11: Importancia de contar con datos estadísticos de estudiantes desertores

Alternativas	Frecuencia	Porcentaje
Si	34	100%
No	0	0%
Total	34	100%

Ilustración 20: Importancia de contar con datos estadísticos de estudiantes desertores

Considerando la información tabulada, se establece, que el 100% de los docentes considera que es muy importante contar con datos estadísticos de estudiantes desertores.

Encuesta al estudiante

1.- ¿Qué situaciones cree usted que causarían el abandono de sus estudios o si por el momento no se encuentra estudiando que lo provoco?

- ☐ Falta de recursos económicos.
- ☐ Vivir lejos de donde estudias.
- ☐ Poco aprovechamiento de las clases.
- ☐ Falta de interés por seguir realmente con las carreras universitarias.
- ☐ Porque por ley están donde su puntaje del ENES los ubicó.
- ☐ Porque no pueden acceder a universidades particulares por lo costoso que resulta.

- ☐ Por enfermedad suya o de un familiar.
- ☐ Por oportunidad de trabajo.
- ☐ Porque el perfil profesional y ocupacional de su carrera no son de su agrado.
- ☐ Porque reprueba las asignaturas constantemente.
- ☐ Falta de interacción de calidad con profesores y orientadores.
- ☐ Ambiente poco motivante en clases.
- ☐ Falta de apoyo familiar.
- ☐ Se convirtió en madre o padre de familia.
- ☐ Por motivo de viaje.
- ☐ otros.

Tabla 12: Situaciones que causan abandono de estudios

Alternativas	Frecuencia	Porcentaje
Falta de recursos económicos.	20	15%
Vivir lejos de donde estudias.	10	10%
Poco aprovechamiento de las clases.	5	5%
Falta de interés por seguir realmente con las carreras universitarias.	0	0%
Porque por ley están donde su puntaje del ENES los ubicó.	15	20%
Porque no pueden acceder a universidades particulares por lo costoso que resulta.	10	10%
Por enfermedad suya o de un familiar.	0	0%
Por oportunidad de trabajo.	25	25%
Porque el perfil profesional y ocupacional de su carrera no son de su agrado.	6	6%
Porque reprueba las asignaturas constantemente.	3	3%
Falta de interacción de calidad con profesores y orientadores.	0	0%
Ambiente poco motivante en clases.	1	1%
Falta de apoyo familiar.	5	5%
Se convirtió en madre o padre de familia.	0	0%
Por motivo de viaje.	0	0%
Otros.	0	0%
Total	100	100%

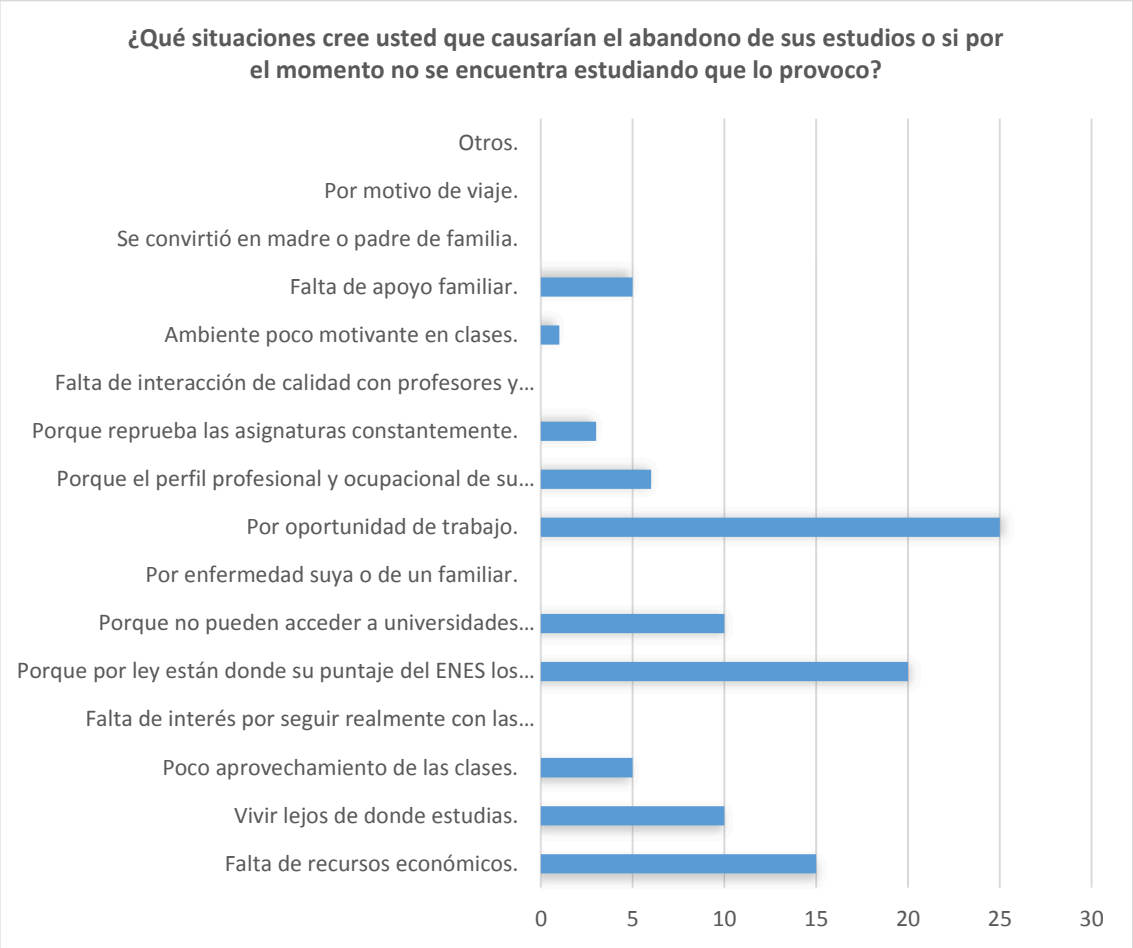


Ilustración 21: Situaciones que causan abandono de estudios

Considerando la información tabulada, se establece, que el 25% de los estudiantes abandonan sus estudios por una oportunidad de trabajo, seguido de que la ley los ubica según su puntaje del ENES con un 20%.

2.9.2. Informe final del análisis de los resultados

De acuerdo a las encuestas realizadas a docentes y estudiantes, de modo principal, se deduce lo siguiente:

- El 44% de los docentes creen que los posibles factores de deserción en la institución son académicos.
- El 100% de los docentes considera que es impórtate conocer cuáles son las razones académicas por las que sus estudiantes deciden abandonar sus estudios.

- El 41% de los docentes cree que una razón académica para el abandono de estudios es el rendimiento de sus estudiantes.
- El 29% de los docentes cree que perfil profesional de la carrera es una razón de la que exista baja retención estudiantil.
- El 53% de los docentes considera que la consecuencia de mayor impacto para la sociedad que trae la deserción es un trabajo mal remunerado.
- El 100% de los docentes considera importante estudios sobre la deserción para la toma de daciones en la institución.
- El 53% cree que una de las variables externas que causan deserción son las económicas.
- El 35% de los docentes motiva el aprendizaje de los estudiantes para evitar este problema social.
- El 100% de los docentes considera importante tener datos estadísticos sobre estudiantes desertores.
- El 25% de los estudiantes considera que una de las situaciones que los llevan a dejar sus estudios son las oportunidades de trabajo que se les presentan ya que le sigue como otro factor la falta e recursos económicos.

CAPITULO III

DISEÑO DE LA PROPUESTA

3.1. INTRODUCCIÓN

El abandono de los estudios universitarios, es un inconveniente social que se genera en diversas Instituciones de Educación Superior.

La baja tasa de retención, tasas de repetición y deserción elevadas entre otros, son elementos de incidencia en este tema de la deserción.

Las más importantes interrogantes que se dan al modelar datos es que tipo de técnica o estructura utilizar, ¿qué técnicas son más viables?, ¿cuáles nos ofrecen mejores características para el modelado?

En este capítulo se podrá observar varios puntos fundamentales en la implementación de técnicas de minería de datos para llegar a crear un modelo con la información que se tiene de los estudiantes de la facultad en cuanto a este problema social que es la deserción, como también su respectivo estudio de viabilidad, tomando en cuenta el mejor conjunto de técnicas para aplicarlas, también se analiza los recursos humanos y físicos necesarios para el desarrollo de la investigación y desarrollo de la propuesta, siendo estos todos los factores que determinan si un proyecto es viable o no.

3.2. DESCRIPCIÓN DE LA PROPUESTA

La propuesta que se plantea a continuación tiene como finalidad dar a conocer mediante el análisis realizado con técnicas de minería de datos a la información que proporcionó la institución sobre los estudiantes de primer a tercer nivel del periodo comprendido entre 2011 - 2016, que factores académicos son los que provocan este problema social identificado en relación a las dificultades que se les presentan para el posterior abandono de la carrera, por lo cual se propuso la construcción de un modelo de datos.

La metodología que se utilizó para implementación del modelo es MSDN (Microsoft Developer Network) que consta de seis fases las cuales están compuestas por actividades o una secuencia de pasos ordenados estándar donde se tiene en cuenta las técnicas y herramientas de minería de datos, que permitieron cumplir con los objetivos del trabajo de titulación.

ID	Actividades	JUNIO				JULIO				AGOSTO				SEPTIEMBRE				OCTUBRE				NOVIEMBRE				DICIEMBRE				ENERO			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Reunión y planificación inicial																																
2	Recolección de la información de movilidad de estudiantes																																
3	Investigación sobre técnicas de minería de datos																																
4	Estudio de Factibilidad																																
5	Realizar Documentación																																
6	Recolección de información de Estudiantes matriculados y notas																																
7	Encuesta sobre la deserción docentes y estudiantes																																
8	Fase I Definir El problema																																
9	Fase II Preparar los datos																																
10	Fase III Explorar los datos																																
11	Fase IV Generar, Explorar y validar Modelos																																
12	Evaluación de Resultados																																
13	Realización de conclusiones y recomendaciones																																

Ilustración 22: Cronograma de Actividades

3.2.1. OBJETIVOS

3.2.1.1. Objetivo General

Implementar estructuras de minería de datos que identifiquen factores que influyan en el entorno académico para evaluar la tasa de retención estudiantil de la facultad de ciencias informáticas.

3.2.1.2. Objetivos Específicos

- Prepara los datos y seleccionar el algoritmo de minería de datos adecuado para el estudio y modelado de la información a procesar.
- Explorar los datos para comprobar que no existan posibles errores que puedan afectar el resultado de los modelos.
- Aplicar estructuras de minería de datos a la información que se recopile sobre estudiantes matriculados y sus calificaciones para generar y validar los modelos.
- Identificar patrones o regularidades que causan la deserción estudiantil, así como las eventuales tendencias de esta problemática, mediante la implementación de las estructuras de minería de datos adecuadas.

3.2.2. ALCANCES

La propuesta descrita tiene el siguiente alcance:

- Maneja los datos de los periodos comprendidos entre 2011 – 2016 referentes a la información ligada a datos personales y calificaciones de los estudiantes de primer a tercer nivel, esto es: cedula, sexo, edad, nacionalidad, nivel, periodo, año, materia, etc.
- Existen variables que afectan a la deserción, como materia perdida, semestre perdido y la movilidad que resultarían claves.
- Es importante resaltar la existencia de una serie de factores que, pueden estar involucradas a la deserción pero que por motivos de falta de coherencia y de su posible no existencia, no fueron tomados en cuenta y se han obviado del modelo. Entre dichos factores destacamos: socioeconómicos, familiares, médicos, etc.

- La vista de los datos de la que se dispone es parcial, y se la emplea con fines demostrativos, no se muestra información de todos los estudiantes sino de aquellos que solo pertenecen al campus Manta, mas no de los demás campus o extensiones pertenecientes a la Facultad de Ciencias Informáticas.

3.3.3. DETERMINACION DE RECURSOS

3.3.3.1. Humanos

Personas que formaron parte directa e indirectamente en el desarrollo del proyecto

RECURSOS HUMANOS	FUNCION
Autora Proyecto Integrador	Vélez Molina Katherine Ibeth
Director de Proyecto Integrador	Ing. Jorge Pincay
Autoridad de Facultad	Lic. Dolores Muñoz
Docentes	Integran la facultad
Estudiantes	Integran la facultad

Tabla 13: Recursos Humanos del proyecto

3.3.3.2. Tecnológicos

- Computadora
- Impresora
- Recomendable Microsoft Windows 7 x86, x64 o superior (se detallan en la sección de factibilidad técnica)
- Waikato Environment for Knowledge Analysis WEKA
- Servicios de internet

3.3.3.3. Económicos / Presupuesto

Gasto en impresiones y materiales para la presentación del proyecto.

3.3.4. FACTIBILIDAD

3.3.4.1. Factibilidad técnica

Elección de la herramienta

A continuación, se presenta la ponderación de algunas plataformas de software para el aprendizaje automático. Con el fin de valorar cuantitativamente cuál de ellas es la mejorar aplicación, se describe en la siguiente tabla. Se utiliza la siguiente escala: Poco aceptable (1), Aceptable (2), Muy aceptable (3).

Software	Ventajas	Desventajas	Tipo de Licencia	Calif.
Weka	– Posee una amplia gama de modelados y algoritmos. – Posee una interfaz muy amigable. – Es multiplataforma. – Proporciona acceso a bases de datos vía SQL gracias a la conexión JDBC (Java Database Connectivity)	– Suele presentar ciertas complicaciones por el uso combinado de modelados. – No incluye modelados de secuencias. – Todas las técnicas de Weka se fundamentan en la asunción de que los datos están disponibles en un fichero plano (flat file) o una relación.	Libre	3
SPSS Clementine	– Acceso a datos (fuentes de datos archivos ASCII) – Procesamiento de datos – Aplicación de técnicas de aprendizaje – Incorpora técnicas de evaluación de modelos visualización de resultados como (histogramas, diagramas de dispersión).	– Si el usuario no tiene experiencia previa utilizando SPSS o si sus conocimientos de estadísticas no están actualizados es difícil discernir que opciones seleccionar. – Posee una gran cantidad de información en forma automática que distrae al usuario.	Libre	2
Kepler	– Herramienta comercial distribuida por Dialogis. – Posee múltiples modelos de análisis. – Permite el pre procesado de datos, la elección de un modelo o la manipulación de la representación gráfica de los modelos obtenidos	– La mayoría de los reportes de resultados contiene un nivel excesivo de información que muchas veces confunde al usuario.	Libre	2
Rapidminer	– Es multiplataforma. – Desarrollado bajo el lenguaje de Java. – Incluye gráficos para una mejor apreciación de resultados.	– No incluye modelados de secuencias.	Libre	2
Pentaho	– Incorpora tecnologías como Java, XML, JavaScript. – Es multiplataforma. – Fácil configuración e instalación.	– Falta de documentación fiable. – Herramientas derivadas.	Privativo	2

Tabla 14: Entornos para análisis del conocimiento a través de modelos de minería de datos

Fuente: Investigación

Elaboración: Autora

De las herramientas revisadas y sus características importantes para dar solución al problema planteado, se optó por WEKA, porque maneja a detalle un buen número de algoritmos de minería de datos, es una herramienta completa y de licencia libre.

Para tener una mejor comprensión acerca de este punto, se ha segregado los requerimientos en las siguientes categorías.

- **Requerimientos de hardware**

Los requerimientos mínimos de hardware que debe tener el equipo donde se instalará la aplicación informática son bajos, Los detalles técnicos del hardware donde es posible implementar WEKA son los siguientes:

- Procesador Pentium 233 MHz o superior.
- Mínimo 64 MB de RAM.
- Mínimo 1,5 GB de espacio disponible en disco duro

Para esta propuesta se usa un computador con las siguientes características:

- Procesador: AMD Quad-Core E2 -3800 CPU 0250GHz 250 GHz
- Memoria instalada (RAM): 8 GB
- Tipo de sistema: Sistema operativo Windows 8 de 64 bits, procesador x64.

- **Requerimientos de software**

Weka al ser multiplataforma es compatible con la mayoría de programas, se tiene la siguiente lista de compatibilidad / requerimientos de WEKA: Microsoft Windows 98SE, Microsoft Windows Me, Microsoft Windows NT, Microsoft Windows 2000, Microsoft Windows XP, x86, x64, Microsoft Windows Vista, x86, x64, Microsoft Windows 7, x86, x64, Microsoft Windows 8, x86, x64, Microsoft Windows 10, x86, x64 (Se ha usado para la presente implementación), Mac OS X, Linux y el respectivo JRE (Java Runtime Environment).

Con el análisis de los requerimientos técnicos y la información proporcionada por la Facultad d Ciencias Informáticas; se concluye que la implementación de las estructuras de minerías de datos para la creación del modelo con WEKA es factible técnicamente por los siguientes fundamentos:

- Los recursos de software y hardware son aptos para la instalación y funcionamiento de la aplicación informática WEKA.
- Las herramientas de desarrollo escogidas son adecuadas para el diseño.

Dentro de los algoritmos elegidos para los modelos se tienen los siguientes:

Se utilizaron los arboles de decisión implementando la tarea de clasificación bajo algoritmo J48, Algoritmos bayesianos como Naive Bayes, reglas como el algoritmo JRip (RIPPER), Clustering bajo algoritmo Simple-Kmeans y Farthest-First.

3.3.4.2. Factibilidad Operativa

La factibilidad operativa se basa en los puntos operativos considerados básicos e imprescindibles, para dar de manera directa satisfacción a los usuarios de los modelos de implementados WEKA. Esta aplicación permitirá hacer una efectiva detección de patrones, regularidades o tendencias de los datos de estudiantes y sus calificaciones.

- Los datos de los estudiantes matriculados que semestralmente se almacenan en el sistema permitirán reunir información importante para que los directivos tomen las decisiones correctas. Pese que en esta propuesta se emplearan datos de hace 6 años y con fines de demostrativos básicamente.

- Se pueden clasificar a los estudiantes por edad, nivel, estado académico, tramite de movilidad, entre otras.

- Cubre las expectativas de los directivos de la facultad para obtener información de apoyo a la toma de decisiones.

- WEKA presenta una interfaz gráfica llamativa y amigable, pero requiere de ciertos fundamentos de lenguaje SQL para la estructuración de archivos planos o consultas sobre los cuales se ejecuten los análisis de búsqueda de conocimiento. El personal de la facultad los posee.

Los antecedentes descritos anteriormente permiten concluir que el entorno para análisis del conocimiento WEKA, configurado para analizar los datos sobre la retención de estudiantes de la Facultad de Ciencias Informáticas, es factible desde el punto de vista operativo.

3.3.4.3. Factibilidad Económica

Para determinar la factibilidad económica, no fue necesario describir los costos que intervienen en el diseño, pues hay señalar que la implementación de la solución del modelo en WEKA, que es un software ha sido desarrollado bajo licencia GPL no genero gasto al ser gratuita.

Parte de la factibilidad económica es realizar una estimación de costo de cada una de las actividades propias se considera el costo de materiales de oficina, transporte, impresiones, entre otros. Resultado que se puede observar en la siguiente tabla:

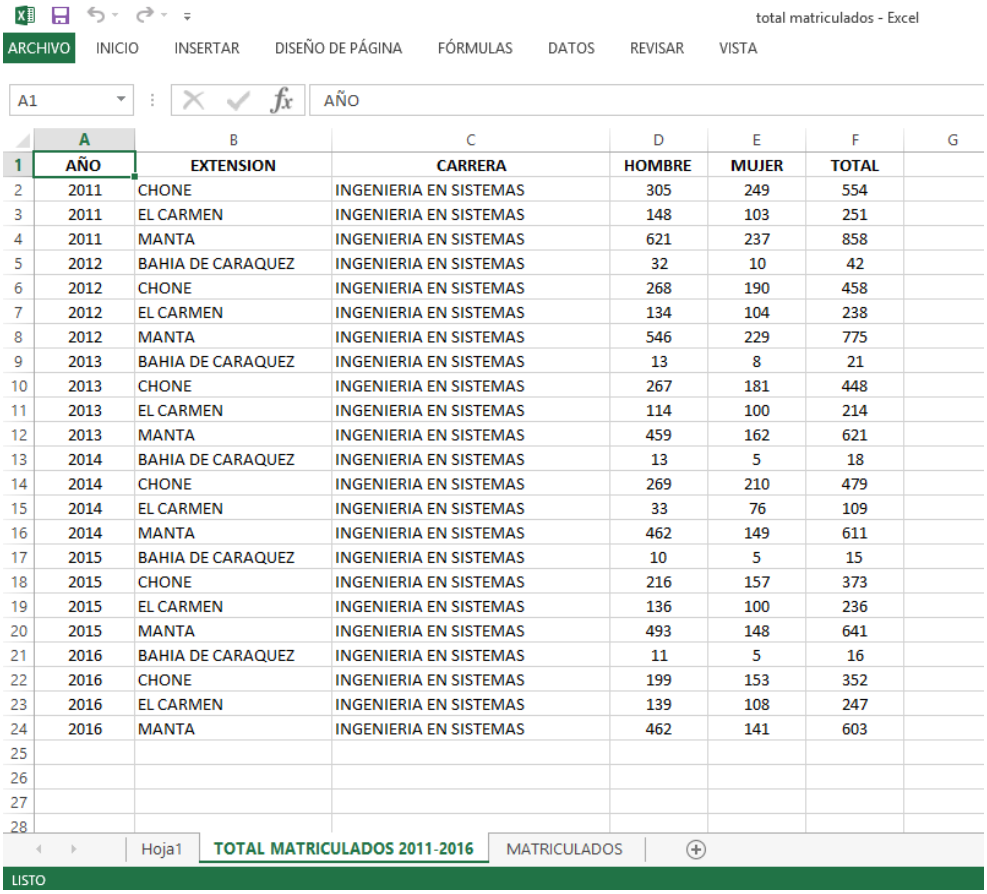
RECURSOS	DETALLE	CANT.	MEDIDA	P.U.	TOTAL
Materiales Suministros	Cartuchos de tinta	1	unidad	20	20
	Empastados del proyecto integrador	1	unidad	12	12
	Anillados	2	unidad	3	6
	transporte	1		25	25
	Materiales de oficina	1		10	10
Técnicos	Internet	3	meses	20	60
Otros	Imprevistos	1		25	25
TOTAL					**158**

Tabla 15: Factibilidad económica

3.4. ETAPAS DE LA PROPUESTA

3.4.3. FASE I: Definir el problema

La Universidad Laica Eloy Alfaro de Manabí es una Institución de Educación Superior que consta de varias facultades entre ellas la FACULTAD DE CIENCIAS INFORMATICAS. El primer paso del proceso de minería de datos, consiste en definir claramente el problema y considerar formas de usar los datos para proporcionar una respuesta para el mismo. Como se detalló en la descripción de la propuesta, se trabaja sobre los datos de los años 2011-2016(1) referentes a la información ligada a estudiantes, calificaciones, movilidad esto es: cedula, nombre, nivel, tipo de movilidad, etc. Para fines de éste estudio se tiene en consideración a la siguiente tabla:



total matriculados - Excel

ARCHIVO INICIO INSERTAR DISEÑO DE PÁGINA FÓRMULAS DATOS REVISAR VISTA

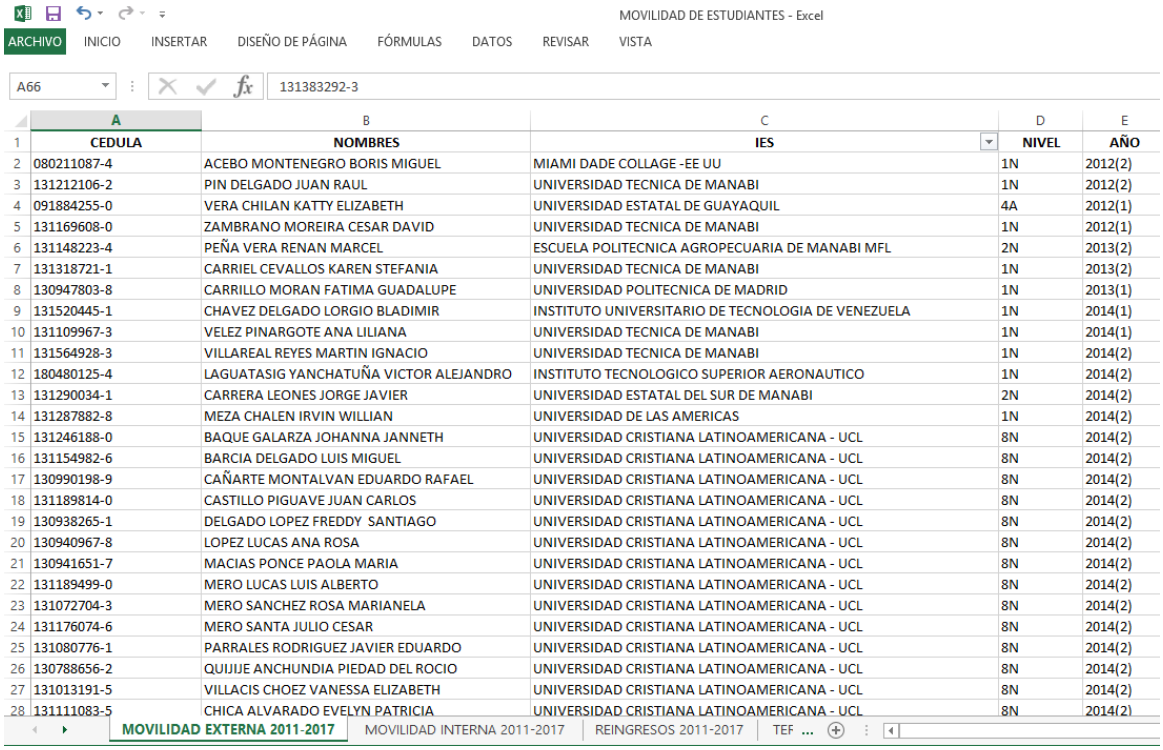
A1 | AÑO

	A	B	C	D	E	F	G
1	AÑO	EXTENSION	CARRERA	HOMBRE	MUJER	TOTAL	
2	2011	CHONE	INGENIERIA EN SISTEMAS	305	249	554	
3	2011	EL CARMEN	INGENIERIA EN SISTEMAS	148	103	251	
4	2011	MANTA	INGENIERIA EN SISTEMAS	621	237	858	
5	2012	BAHIA DE CARAQUEZ	INGENIERIA EN SISTEMAS	32	10	42	
6	2012	CHONE	INGENIERIA EN SISTEMAS	268	190	458	
7	2012	EL CARMEN	INGENIERIA EN SISTEMAS	134	104	238	
8	2012	MANTA	INGENIERIA EN SISTEMAS	546	229	775	
9	2013	BAHIA DE CARAQUEZ	INGENIERIA EN SISTEMAS	13	8	21	
10	2013	CHONE	INGENIERIA EN SISTEMAS	267	181	448	
11	2013	EL CARMEN	INGENIERIA EN SISTEMAS	114	100	214	
12	2013	MANTA	INGENIERIA EN SISTEMAS	459	162	621	
13	2014	BAHIA DE CARAQUEZ	INGENIERIA EN SISTEMAS	13	5	18	
14	2014	CHONE	INGENIERIA EN SISTEMAS	269	210	479	
15	2014	EL CARMEN	INGENIERIA EN SISTEMAS	33	76	109	
16	2014	MANTA	INGENIERIA EN SISTEMAS	462	149	611	
17	2015	BAHIA DE CARAQUEZ	INGENIERIA EN SISTEMAS	10	5	15	
18	2015	CHONE	INGENIERIA EN SISTEMAS	216	157	373	
19	2015	EL CARMEN	INGENIERIA EN SISTEMAS	136	100	236	
20	2015	MANTA	INGENIERIA EN SISTEMAS	493	148	641	
21	2016	BAHIA DE CARAQUEZ	INGENIERIA EN SISTEMAS	11	5	16	
22	2016	CHONE	INGENIERIA EN SISTEMAS	199	153	352	
23	2016	EL CARMEN	INGENIERIA EN SISTEMAS	139	108	247	
24	2016	MANTA	INGENIERIA EN SISTEMAS	462	141	603	
25							
26							
27							
28							

Hoja1 | TOTAL MATRICULADOS 2011-2016 | MATRICULADOS

LISTO

Ilustración 23: Tabla total matriculados FACCI 2011-2016



MOVILIDAD DE ESTUDIANTES - Excel

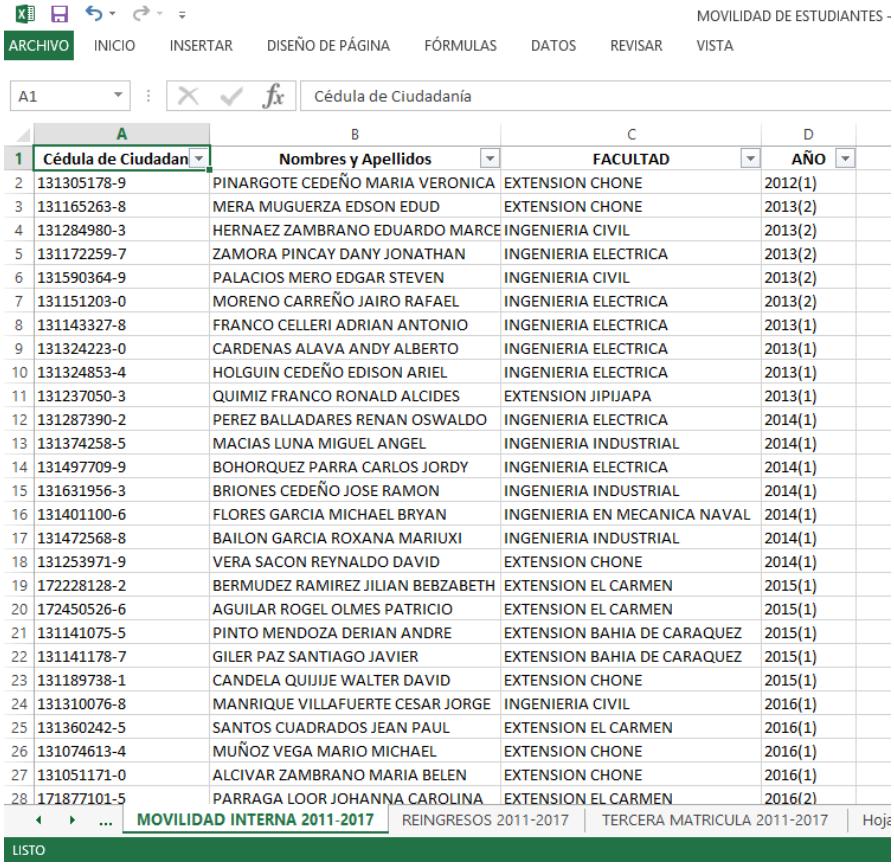
ARCHIVO | INICIO | INSERTAR | DISEÑO DE PÁGINA | FÓRMULAS | DATOS | REVISAR | VISTA

A66 | 131383292-3

	A	B	C	D	E
1	CEDULA	NOMBRES	IES	NIVEL	AÑO
2	080211087-4	ACEBO MONTENEGRO BORIS MIGUEL	MIAMI DADE COLLAGE -EE UU	1N	2012(2)
3	131212106-2	PIN DELGADO JUAN RAUL	UNIVERSIDAD TECNICA DE MANABI	1N	2012(2)
4	091884255-0	VERA CHILAN KATTY ELIZABETH	UNIVERSIDAD ESTATAL DE GUAYAQUIL	4A	2012(1)
5	131169608-0	ZAMBRANO MOREIRA CESAR DAVID	UNIVERSIDAD TECNICA DE MANABI	1N	2012(1)
6	131148223-4	PEÑA VERA RENAN MARCEL	ESCUELA POLITECNICA AGROPECUARIA DE MANABI MFL	2N	2013(2)
7	131318721-1	CARRIEL CEVALLOS KAREN STEFANIA	UNIVERSIDAD TECNICA DE MANABI	1N	2013(2)
8	130947803-8	CARRILLO MORAN FATIMA GUADALUPE	UNIVERSIDAD POLITECNICA DE MADRID	1N	2013(1)
9	131520445-1	CHAVEZ DELGADO LORGIO BLADIMIR	INSTITUTO UNIVERSITARIO DE TECNOLOGIA DE VENEZUELA	1N	2014(1)
10	131109967-3	VELEZ PINARGOTE ANA LILIANA	UNIVERSIDAD TECNICA DE MANABI	1N	2014(1)
11	131564928-3	VILLAREAL REYES MARTIN IGNACIO	UNIVERSIDAD TECNICA DE MANABI	1N	2014(2)
12	180480125-4	LAGUATASIG YANCHATUÑA VICTOR ALEJANDRO	INSTITUTO TECNOLOGICO SUPERIOR AERONAUTICO	1N	2014(2)
13	131290034-1	CARRERA LEONES JORGE JAVIER	UNIVERSIDAD ESTATAL DEL SUR DE MANABI	2N	2014(2)
14	131287882-8	MEZA CHALEN IRVIN WILLIAN	UNIVERSIDAD DE LAS AMERICAS	1N	2014(2)
15	131246188-0	BAQUE GALARZA JOHANNA JANNETH	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
16	131154982-6	BARCIA DELGADO LUIS MIGUEL	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
17	130990198-9	CAÑARTE MONTALVAN EDUARDO RAFAEL	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
18	131189814-0	CASTILLO PIGUAVE JUAN CARLOS	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
19	130938265-1	DELGADO LOPEZ FREDDY SANTIAGO	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
20	130940967-8	LOPEZ LUCAS ANA ROSA	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
21	130941651-7	MACIAS PONCE PAOLA MARIA	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
22	131189499-0	MERO LUCAS LUIS ALBERTO	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
23	131072704-3	MERO SANCHEZ ROSA MARIANELA	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
24	131176074-6	MERO SANTA JULIO CESAR	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
25	131080776-1	PARRALES RODRIGUEZ JAVIER EDUARDO	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
26	130788656-2	QUIJIJE ANCHUNDIA PIEDAD DEL ROCIO	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
27	131013191-5	VILLACIS CHOEZ VANESSA ELIZABETH	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)
28	131111083-5	CHICA ALVARADO EVELYN PATRICIA	UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	8N	2014(2)

MOVILIDAD EXTERNA 2011-2017 | MOVILIDAD INTERNA 2011-2017 | REINGRESOS 2011-2017 | TEF ...

Ilustración 24: Tabla Movilidad Externa FACCI 2011-2016



MOVILIDAD DE ESTUDIANTES -

ARCHIVO | INICIO | INSERTAR | DISEÑO DE PÁGINA | FÓRMULAS | DATOS | REVISAR | VISTA

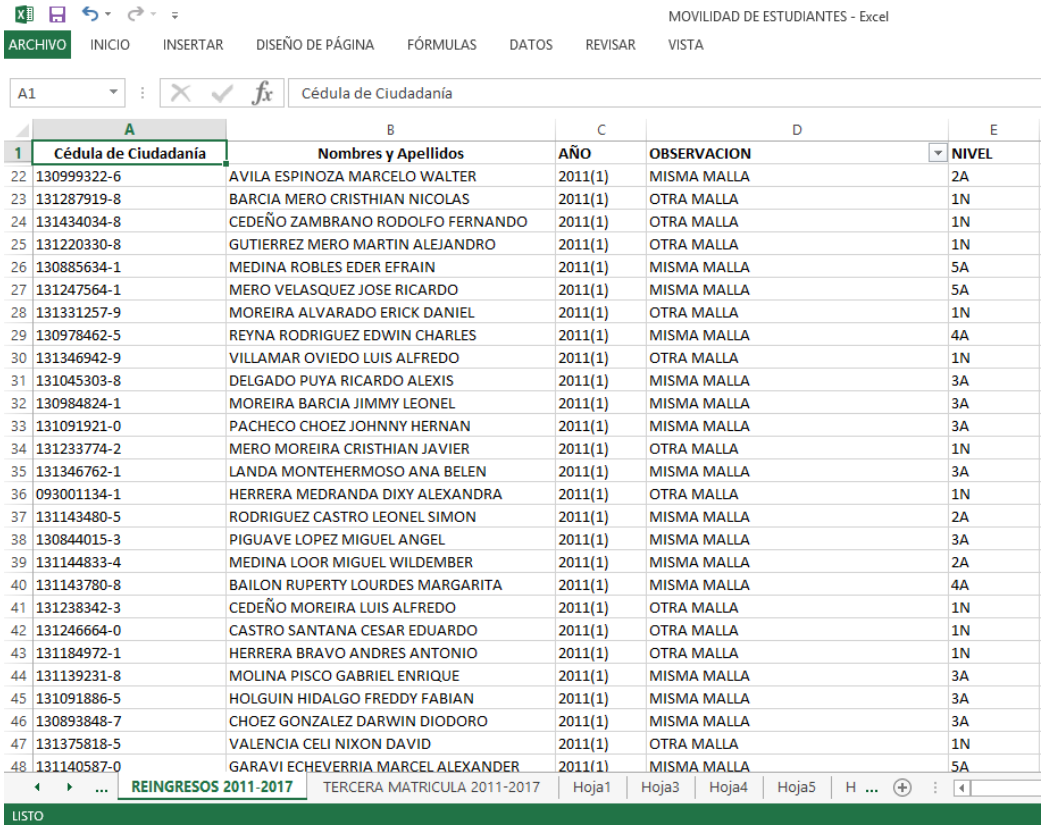
A1 | Cédula de Ciudadanía

	A	B	C	D
1	Cédula de Ciudadan	Nombres y Apellidos	FACULTAD	AÑO
2	131305178-9	PINARGOTE CEDEÑO MARIA VERONICA	EXTENSION CHONE	2012(1)
3	131165263-8	MERA MUGUERZA EDSON EDUD	EXTENSION CHONE	2013(2)
4	131284980-3	HERNAEZ ZAMBRANO EDUARDO MARCE	INGENIERIA CIVIL	2013(2)
5	131172259-7	ZAMORA PINCAY DANY JONATHAN	INGENIERIA ELECTRICA	2013(2)
6	131590364-9	PALACIOS MERO EDGAR STEVEN	INGENIERIA CIVIL	2013(2)
7	131151203-0	MORENO CARREÑO JAIRO RAFAEL	INGENIERIA ELECTRICA	2013(2)
8	131143327-8	FRANCO CELLERI ADRIAN ANTONIO	INGENIERIA ELECTRICA	2013(1)
9	131324223-0	CARDENAS ALAVA ANDY ALBERTO	INGENIERIA ELECTRICA	2013(1)
10	131324853-4	HOLGUIN CEDEÑO EDISON ARIEL	INGENIERIA ELECTRICA	2013(1)
11	131237050-3	QUIMIZ FRANCO RONALD ALCIDES	EXTENSION JIPIJAPA	2013(1)
12	131287390-2	PEREZ BALLADARES RENAN OSWALDO	INGENIERIA ELECTRICA	2014(1)
13	131374258-5	MACIAS LUNA MIGUEL ANGEL	INGENIERIA INDUSTRIAL	2014(1)
14	131497709-9	BOHORQUEZ PARRA CARLOS JORDY	INGENIERIA ELECTRICA	2014(1)
15	131631956-3	BRIONES CEDEÑO JOSE RAMON	INGENIERIA INDUSTRIAL	2014(1)
16	131401100-6	FLORES GARCIA MICHAEL BRYAN	INGENIERIA EN MECANICA NAVAL	2014(1)
17	131472568-8	BAILON GARCIA ROXANA MARIUXI	INGENIERIA INDUSTRIAL	2014(1)
18	131253971-9	VERA SACON REYNALDO DAVID	EXTENSION CHONE	2014(1)
19	172228128-2	BERMUDEZ RAMIREZ JILIAN BEBZABETH	EXTENSION EL CARMEN	2015(1)
20	172450526-6	AGUILAR ROGEL OLMES PATRICIO	EXTENSION EL CARMEN	2015(1)
21	131141075-5	PINTO MENDOZA DERIAN ANDRE	EXTENSION BAHIA DE CARAQUEZ	2015(1)
22	131141178-7	GILER PAZ SANTIAGO JAVIER	EXTENSION BAHIA DE CARAQUEZ	2015(1)
23	131189738-1	CANDELA QUIJIJE WALTER DAVID	EXTENSION CHONE	2015(1)
24	131310076-8	MANRIQUE VILLAFUERTE CESAR JORGE	INGENIERIA CIVIL	2016(1)
25	131360242-5	SANTOS CUADRADOS JEAN PAUL	EXTENSION EL CARMEN	2016(1)
26	131074613-4	MUÑOZ VEGA MARIO MICHAEL	EXTENSION CHONE	2016(1)
27	131051171-0	ALCIVAR ZAMBRANO MARIA BELEN	EXTENSION CHONE	2016(1)
28	171877101-5	PARRAGA LOOR JOHANNA CAROLINA	EXTENSION EL CARMEN	2016(2)

... MOVILIDAD INTERNA 2011-2017 | REINGRESOS 2011-2017 | TERCERA MATRICULA 2011-2017 | Hoja

LISTO

Ilustración 25: Tabla Movilidad Interna FACCI 2011-2016



MOVILIDAD DE ESTUDIANTES - Excel

ARCHIVO | INICIO | INSERTAR | DISEÑO DE PÁGINA | FÓRMULAS | DATOS | REVISAR | VISTA

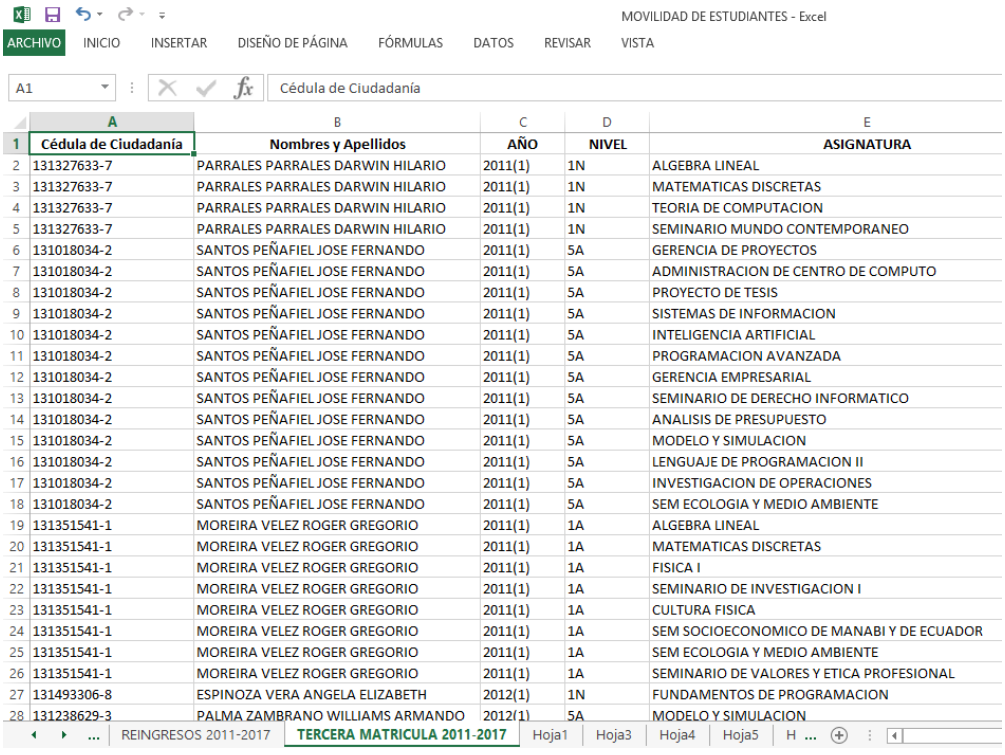
A1 | Cédula de Ciudadanía

	A	B	C	D	E
1	Cédula de Ciudadanía	Nombres y Apellidos	AÑO	OBSERVACION	NIVEL
22	130999322-6	AVILA ESPINOZA MARCELO WALTER	2011(1)	MISMA MALLA	2A
23	131287919-8	BARCIA MERO CRISTHIAN NICOLAS	2011(1)	OTRA MALLA	1N
24	131434034-8	CEDEÑO ZAMBRANO RODOLFO FERNANDO	2011(1)	OTRA MALLA	1N
25	131220330-8	GUTIERREZ MERO MARTIN ALEJANDRO	2011(1)	OTRA MALLA	1N
26	130885634-1	MEDINA ROBLES EDER EFRAIN	2011(1)	MISMA MALLA	5A
27	131247564-1	MERO VELASQUEZ JOSE RICARDO	2011(1)	MISMA MALLA	5A
28	131331257-9	MOREIRA ALVARADO ERICK DANIEL	2011(1)	OTRA MALLA	1N
29	130978462-5	REYNA RODRIGUEZ EDWIN CHARLES	2011(1)	MISMA MALLA	4A
30	131346942-9	VILLAMAR OVIEDO LUIS ALFREDO	2011(1)	OTRA MALLA	1N
31	131045303-8	DELGADO PUYA RICARDO ALEXIS	2011(1)	MISMA MALLA	3A
32	130984824-1	MOREIRA BARCIA JIMMY LEONEL	2011(1)	MISMA MALLA	3A
33	131091921-0	PACHECO CHOEZ JOHNNY HERNAN	2011(1)	MISMA MALLA	3A
34	131233774-2	MERO MOREIRA CRISTHIAN JAVIER	2011(1)	OTRA MALLA	1N
35	131346762-1	LANDA MONTEHERMOSO ANA BELEN	2011(1)	MISMA MALLA	3A
36	093001134-1	HERRERA MEDRANDA DIXY ALEXANDRA	2011(1)	OTRA MALLA	1N
37	131143480-5	RODRIGUEZ CASTRO LEONEL SIMON	2011(1)	MISMA MALLA	2A
38	130844015-3	PIGUAVE LOPEZ MIGUEL ANGEL	2011(1)	MISMA MALLA	3A
39	131144833-4	MEDINA LOOR MIGUEL WILDEMBER	2011(1)	MISMA MALLA	2A
40	131143780-8	BAILON RUPERTY LOURDES MARGARITA	2011(1)	MISMA MALLA	4A
41	131238342-3	CEDEÑO MOREIRA LUIS ALFREDO	2011(1)	OTRA MALLA	1N
42	131246664-0	CASTRO SANTANA CESAR EDUARDO	2011(1)	OTRA MALLA	1N
43	131184972-1	HERRERA BRAVO ANDRES ANTONIO	2011(1)	OTRA MALLA	1N
44	131139231-8	MOLINA PISCO GABRIEL ENRIQUE	2011(1)	MISMA MALLA	3A
45	131091886-5	HOLGUIN HIDALGO FREDDY FABIAN	2011(1)	MISMA MALLA	3A
46	130893848-7	CHOEZ GONZALEZ DARWIN DIODORO	2011(1)	MISMA MALLA	3A
47	131375818-5	VALENCIA CELI NIXON DAVID	2011(1)	OTRA MALLA	1N
48	131140587-0	GARAVI ECHEVERRIA MARCEL ALEXANDER	2011(1)	MISMA MALLA	5A

REINGRESOS 2011-2017 | TERCERA MATRICULA 2011-2017 | Hoja1 | Hoja3 | Hoja4 | Hoja5

LISTO

Ilustración 26: Tabla Reingresos FACCI 2011-2016



MOVILIDAD DE ESTUDIANTES - Excel

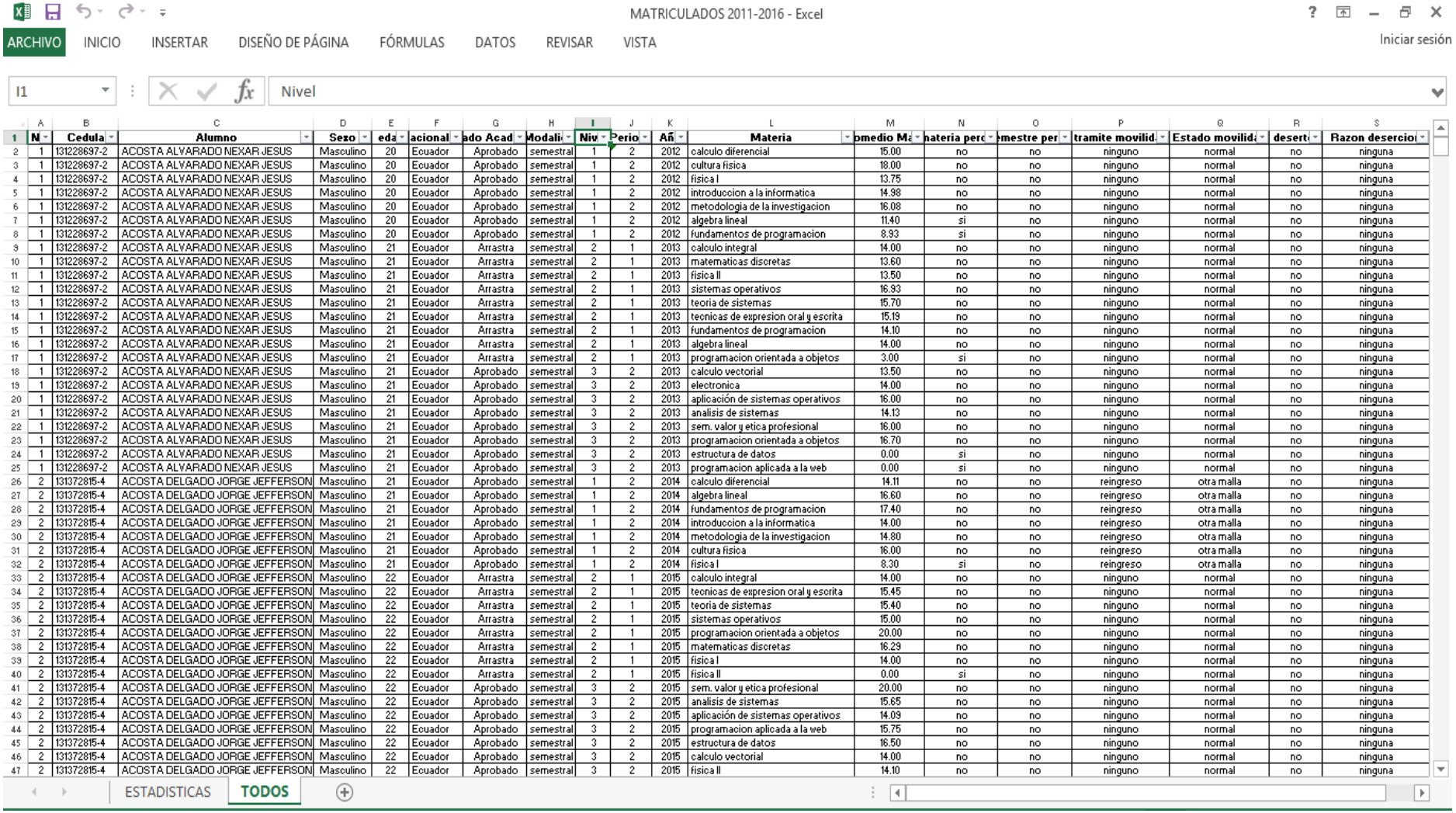
ARCHIVO | INICIO | INSERTAR | DISEÑO DE PÁGINA | FÓRMULAS | DATOS | REVISAR | VISTA

A1 | Cédula de Ciudadanía

	A	B	C	D	E
1	Cédula de Ciudadanía	Nombres y Apellidos	AÑO	NIVEL	ASIGNATURA
2	131327633-7	PARRALES PARRALES DARWIN HILARIO	2011(1)	1N	ALGEBRA LINEAL
3	131327633-7	PARRALES PARRALES DARWIN HILARIO	2011(1)	1N	MATEMATICAS DISCRETAS
4	131327633-7	PARRALES PARRALES DARWIN HILARIO	2011(1)	1N	TEORIA DE COMPUTACION
5	131327633-7	PARRALES PARRALES DARWIN HILARIO	2011(1)	1N	SEMINARIO MUNDO CONTEMPORANEO
6	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	GERENCIA DE PROYECTOS
7	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	ADMINISTRACION DE CENTRO DE COMPUTO
8	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	PROYECTO DE TESIS
9	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	SISTEMAS DE INFORMACION
10	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	INTELIGENCIA ARTIFICIAL
11	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	PROGRAMACION AVANZADA
12	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	GERENCIA EMPRESARIAL
13	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	SEMINARIO DE DERECHO INFORMATICO
14	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	ANALISIS DE PRESUPUESTO
15	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	MODELO Y SIMULACION
16	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	LENGUAJE DE PROGRAMACION II
17	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	INVESTIGACION DE OPERACIONES
18	131018034-2	SANTOS PEÑAFIEL JOSE FERNANDO	2011(1)	5A	SEM ECOLOGIA Y MEDIO AMBIENTE
19	131351541-1	MOREIRA VELEZ ROGER GREGORIO	2011(1)	1A	ALGEBRA LINEAL
20	131351541-1	MOREIRA VELEZ ROGER GREGORIO	2011(1)	1A	MATEMATICAS DISCRETAS
21	131351541-1	MOREIRA VELEZ ROGER GREGORIO	2011(1)	1A	FISICA I
22	131351541-1	MOREIRA VELEZ ROGER GREGORIO	2011(1)	1A	SEMINARIO DE INVESTIGACION I
23	131351541-1	MOREIRA VELEZ ROGER GREGORIO	2011(1)	1A	CULTURA FISICA
24	131351541-1	MOREIRA VELEZ ROGER GREGORIO	2011(1)	1A	SEM SOCIOECONOMICO DE MANABI Y DE ECUADOR
25	131351541-1	MOREIRA VELEZ ROGER GREGORIO	2011(1)	1A	SEM ECOLOGIA Y MEDIO AMBIENTE
26	131351541-1	MOREIRA VELEZ ROGER GREGORIO	2011(1)	1A	SEMINARIO DE VALORES Y ETICA PROFESIONAL
27	131493306-8	ESPINOZA VERA ANGELA ELIZABETH	2012(1)	1N	FUNDAMENTOS DE PROGRAMACION
28	131238629-3	PALMA ZAMBRANO WILLIAMS ARMANDO	2012(1)	5A	MODELO Y SIMULACION

REINGRESOS 2011-2017 | TERCERA MATRICULA 2011-2017 | Hoja1 | Hoja3 | Hoja4 | Hoja5

Ilustración 27: Tabla Tercera Matricula FACCI 2011- 2016



N	Cedula	Alumno	Sexo	eda	acional	ado Acad	Modali	Niv	Perio	Añ	Materia	omedio Ma	nateria per	emestre per	tramite movilid	Estado movilid	deserti	Razon deserci
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	20	Ecuador	Aprobado	semestral	1	2	2012	calculo diferencial	15.00	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	20	Ecuador	Aprobado	semestral	1	2	2012	cultura fisica	18.00	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	20	Ecuador	Aprobado	semestral	1	2	2012	fisica I	13.75	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	20	Ecuador	Aprobado	semestral	1	2	2012	introduccion a la informatica	14.98	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	20	Ecuador	Aprobado	semestral	1	2	2012	metodologia de la investigacion	16.08	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	20	Ecuador	Aprobado	semestral	1	2	2012	algebra lineal	11.40	si	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	20	Ecuador	Aprobado	semestral	1	2	2012	fundamentos de programacion	8.93	si	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Arrastra	semestral	2	1	2013	calculo integral	14.00	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Arrastra	semestral	2	1	2013	matematicas discretas	13.60	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Arrastra	semestral	2	1	2013	fisica II	13.50	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Arrastra	semestral	2	1	2013	sistemas operativos	16.93	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Arrastra	semestral	2	1	2013	teoria de sistemas	15.70	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Arrastra	semestral	2	1	2013	tecnicas de expresion oral y escrita	15.19	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Arrastra	semestral	2	1	2013	fundamentos de programacion	14.10	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Arrastra	semestral	2	1	2013	algebra lineal	14.00	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Arrastra	semestral	2	1	2013	programacion orientada a objetos	3.00	si	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Aprobado	semestral	3	2	2013	calculo vectorial	13.50	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Aprobado	semestral	3	2	2013	electronica	14.00	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Aprobado	semestral	3	2	2013	aplicación de sistemas operativos	16.00	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Aprobado	semestral	3	2	2013	analisis de sistemas	14.13	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Aprobado	semestral	3	2	2013	sem. valor y etica profesional	16.00	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Aprobado	semestral	3	2	2013	programacion orientada a objetos	16.70	no	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Aprobado	semestral	3	2	2013	estructura de datos	0.00	si	no	ninguno	normal	no	ninguna
1	131228697-2	ACOSTA ALVARADO NEXAR JESUS	Masculino	21	Ecuador	Aprobado	semestral	3	2	2013	programacion aplicada a la web	0.00	si	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	21	Ecuador	Aprobado	semestral	1	2	2014	calculo diferencial	14.11	no	no	reingreso	otra malla	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	21	Ecuador	Aprobado	semestral	1	2	2014	algebra lineal	16.60	no	no	reingreso	otra malla	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	21	Ecuador	Aprobado	semestral	1	2	2014	fundamentos de programacion	17.40	no	no	reingreso	otra malla	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	21	Ecuador	Aprobado	semestral	1	2	2014	introduccion a la informatica	14.00	no	no	reingreso	otra malla	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	21	Ecuador	Aprobado	semestral	1	2	2014	metodologia de la investigacion	14.80	no	no	reingreso	otra malla	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	21	Ecuador	Aprobado	semestral	1	2	2014	cultura fisica	16.00	no	no	reingreso	otra malla	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	21	Ecuador	Aprobado	semestral	1	2	2014	fisica I	8.30	si	no	reingreso	otra malla	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Arrastra	semestral	2	1	2015	calculo integral	14.00	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Arrastra	semestral	2	1	2015	tecnicas de expresion oral y escrita	15.45	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Arrastra	semestral	2	1	2015	teoria de sistemas	15.40	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Arrastra	semestral	2	1	2015	sistemas operativos	15.00	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Arrastra	semestral	2	1	2015	programacion orientada a objetos	20.00	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Arrastra	semestral	2	1	2015	matematicas discretas	16.29	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Arrastra	semestral	2	1	2015	fisica I	14.00	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Arrastra	semestral	2	1	2015	fisica II	0.00	si	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Aprobado	semestral	3	2	2015	sem. valor y etica profesional	20.00	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Aprobado	semestral	3	2	2015	analisis de sistemas	15.65	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Aprobado	semestral	3	2	2015	aplicación de sistemas operativos	14.09	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Aprobado	semestral	3	2	2015	programacion aplicada a la web	15.75	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Aprobado	semestral	3	2	2015	estructura de datos	16.50	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Aprobado	semestral	3	2	2015	calculo vectorial	14.00	no	no	ninguno	normal	no	ninguna
2	131372815-4	ACOSTA DELGADO JORGE JEFFERSON	Masculino	22	Ecuador	Aprobado	semestral	3	2	2015	fisica II	14.10	no	no	ninguno	normal	no	ninguna

Ilustración 28: Tabla de Estudiantes matriculados y sus calificaciones FACCI 2011-2016

Conocido los datos con los que se cuenta, de qué tipo son, se ha determinado que estos datos se necesitan limpiar, no es necesario agregar nuevos datos (se cuenta con 16840 filas). Si es necesario convertirlos a tipos texto (nominal) o decimal (numérico) que son los tipos de datos que soporta WEKA.

Dado estos antecedentes, ahora especificamos las predicciones y clasificaciones que se desean hacer:

- MATERIA PÉRDIDA: Se predice el estado de la materia pérdida en que se generan desertores a partir de los datos reales que se tienen. Con respecto al estado se debe considerar que la mayoría ocurren cuando el estado de la materia perdida es “SI”. Para que el modelo de árbol sea capaz de identificar el perfil de carga asociado a cada estado se eligieron los atributos semestre perdido, año y desertor.
- ESTADO ACADÉMICO: Se predice el estado académico en que se generan desertores a partir de los datos reales con los que se cuenta. Con respecto al estado académico se debe considerar que la mayoría ocurren cuando este es “REPITE”, es decir aquellos estudiantes que repiten el semestre. Este atributo tiene la siguientes variables “APROBADO”,”ARRASTRA” y “REPITE”.
- PROMEDIO MATERIA: Se predice el promedio de la materia en que se generan desertores a partir de los datos reales con los que se cuenta. Este comprende del 0 al 20, la materia está perdida cuando esta sea menor a 13.5, si esta es mayor o igual la materia no se encuentra perdida.
- TRÁMITE DE MOVILIDAD: Se predice que tipo de trámite en que más se generan desertores a partir de los datos reales con los que se cuenta. En el algoritmo JRip las reglas muestran que respecto al tipo de trámite se debe considerar que la mayoría ocurren cuando este es “NINGUNO”.

3.4.4. FASE II: Preparar datos

WEKA sólo reconoce dos tipos de datos, nominal o texto y numérico o decimal, por lo que es necesario convertirlos a estos tipos de datos.

MATRICULADOS.arff

Relation: MATRICULADOS 2011-2016 niveles

No.	No. Numeric	Cedula Nominal	Alumno Nominal	Sexo Nominal	edad Numeric	Nacionalidad Nominal	Estado Academico Nominal	Modalidad Nominal	Nivel Nominal	Periodo Numeric	Año Numeric	Materia Nominal	Promedio Materia Numeric	materia perdida Nominal	semestre perdido Nominal	tramite movilidad Nominal	Estado movilidad Nominal	desertor Nominal	**Razon desercion** Nominal
1	1.0	13122...	ACOS...	Mascul...	20.0	Ecuador	Aprobado	semestral	1.0	2.0	2012.0	calculo...	15.0	no	no	ninguno	normal	no	ninguna
2	1.0	13122...	ACOS...	Mascul...	20.0	Ecuador	Aprobado	semestral	1.0	2.0	2012.0	cultur...	18.0	no	no	ninguno	normal	no	ninguna
3	1.0	13122...	ACOS...	Mascul...	20.0	Ecuador	Aprobado	semestral	1.0	2.0	2012.0	fisica I	13.75	no	no	ninguno	normal	no	ninguna
4	1.0	13122...	ACOS...	Mascul...	20.0	Ecuador	Aprobado	semestral	1.0	2.0	2012.0	introd...	14.98	no	no	ninguno	normal	no	ninguna
5	1.0	13122...	ACOS...	Mascul...	20.0	Ecuador	Aprobado	semestral	1.0	2.0	2012.0	metod...	16.08	no	no	ninguno	normal	no	ninguna
6	1.0	13122...	ACOS...	Mascul...	20.0	Ecuador	Aprobado	semestral	1.0	2.0	2012.0	algebr...	11.4	si	no	ninguno	normal	no	ninguna
7	1.0	13122...	ACOS...	Mascul...	20.0	Ecuador	Aprobado	semestral	1.0	2.0	2012.0	funda...	8.93	si	no	ninguno	normal	no	ninguna
8	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Arrastra	semestral	2.0	1.0	2013.0	calculo...	14.0	no	no	ninguno	normal	no	ninguna
9	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Arrastra	semestral	2.0	1.0	2013.0	matem...	13.6	no	no	ninguno	normal	no	ninguna
10	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Arrastra	semestral	2.0	1.0	2013.0	fisica II	13.5	no	no	ninguno	normal	no	ninguna
11	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Arrastra	semestral	2.0	1.0	2013.0	sistem...	16.93	no	no	ninguno	normal	no	ninguna
12	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Arrastra	semestral	2.0	1.0	2013.0	teoria ...	15.7	no	no	ninguno	normal	no	ninguna
13	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Arrastra	semestral	2.0	1.0	2013.0	tecnic...	15.19	no	no	ninguno	normal	no	ninguna
14	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Arrastra	semestral	2.0	1.0	2013.0	funda...	14.1	no	no	ninguno	normal	no	ninguna
15	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Arrastra	semestral	2.0	1.0	2013.0	algebr...	14.0	no	no	ninguno	normal	no	ninguna
16	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Arrastra	semestral	2.0	1.0	2013.0	progra...	3.0	si	no	ninguno	normal	no	ninguna
17	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	3.0	2.0	2013.0	calculo...	13.5	no	no	ninguno	normal	no	ninguna
18	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	3.0	2.0	2013.0	electr...	14.0	no	no	ninguno	normal	no	ninguna
19	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	3.0	2.0	2013.0	aplicac...	16.0	no	no	ninguno	normal	no	ninguna
20	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	3.0	2.0	2013.0	analisi...	14.13	no	no	ninguno	normal	no	ninguna
21	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	3.0	2.0	2013.0	sem. v...	16.0	no	no	ninguno	normal	no	ninguna
22	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	3.0	2.0	2013.0	progra...	16.7	no	no	ninguno	normal	no	ninguna
23	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	3.0	2.0	2013.0	estruc...	0.0	si	no	ninguno	normal	no	ninguna
24	1.0	13122...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	3.0	2.0	2013.0	progra...	0.0	si	no	ninguno	normal	no	ninguna
25	2.0	13137...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	1.0	2.0	2014.0	calculo...	14.11	no	no	reingreso	otra malla	no	ninguna
26	2.0	13137...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	1.0	2.0	2014.0	algebr...	16.6	no	no	reingreso	otra malla	no	ninguna
27	2.0	13137...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	1.0	2.0	2014.0	funda...	17.4	no	no	reingreso	otra malla	no	ninguna
28	2.0	13137...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	1.0	2.0	2014.0	introd...	14.0	no	no	reingreso	otra malla	no	ninguna
29	2.0	13137...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	1.0	2.0	2014.0	metod...	14.8	no	no	reingreso	otra malla	no	ninguna
30	2.0	13137...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	1.0	2.0	2014.0	cultur...	16.0	no	no	reingreso	otra malla	no	ninguna
31	2.0	13137...	ACOS...	Mascul...	21.0	Ecuador	Aprobado	semestral	1.0	2.0	2014.0	fisica I	8.3	si	no	reingreso	otra malla	no	ninguna
32	2.0	13137...	ACOS...	Mascul...	22.0	Ecuador	Arrastra	semestral	2.0	1.0	2015.0	calculo...	14.0	no	no	ninguno	normal	no	ninguna
33	2.0	13137...	ACOS...	Mascul...	22.0	Ecuador	Arrastra	semestral	2.0	1.0	2015.0	tecnic...	15.45	no	no	ninguno	normal	no	ninguna
34	2.0	13137...	ACOS...	Mascul...	22.0	Ecuador	Arrastra	semestral	2.0	1.0	2015.0	teoria ...	15.4	no	no	ninguno	normal	no	ninguna
35	2.0	13137...	ACOS...	Mascul...	22.0	Ecuador	Arrastra	semestral	2.0	1.0	2015.0	sistem...	15.0	no	no	ninguno	normal	no	ninguna
36	2.0	13137...	ACOS...	Mascul...	22.0	Ecuador	Arrastra	semestral	2.0	1.0	2015.0	progra...	20.0	no	no	ninguno	normal	no	ninguna
37	2.0	13137...	ACOS...	Mascul...	22.0	Ecuador	Arrastra	semestral	2.0	1.0	2015.0	matem...	16.29	no	no	ninguno	normal	no	ninguna
38	2.0	13137...	ACOS...	Mascul...	22.0	Ecuador	Arrastra	semestral	2.0	1.0	2015.0	fisica I	14.0	no	no	ninguno	normal	no	ninguna

Ilustración 29: Archivo ARFF matriculados FACCI 2011-2016

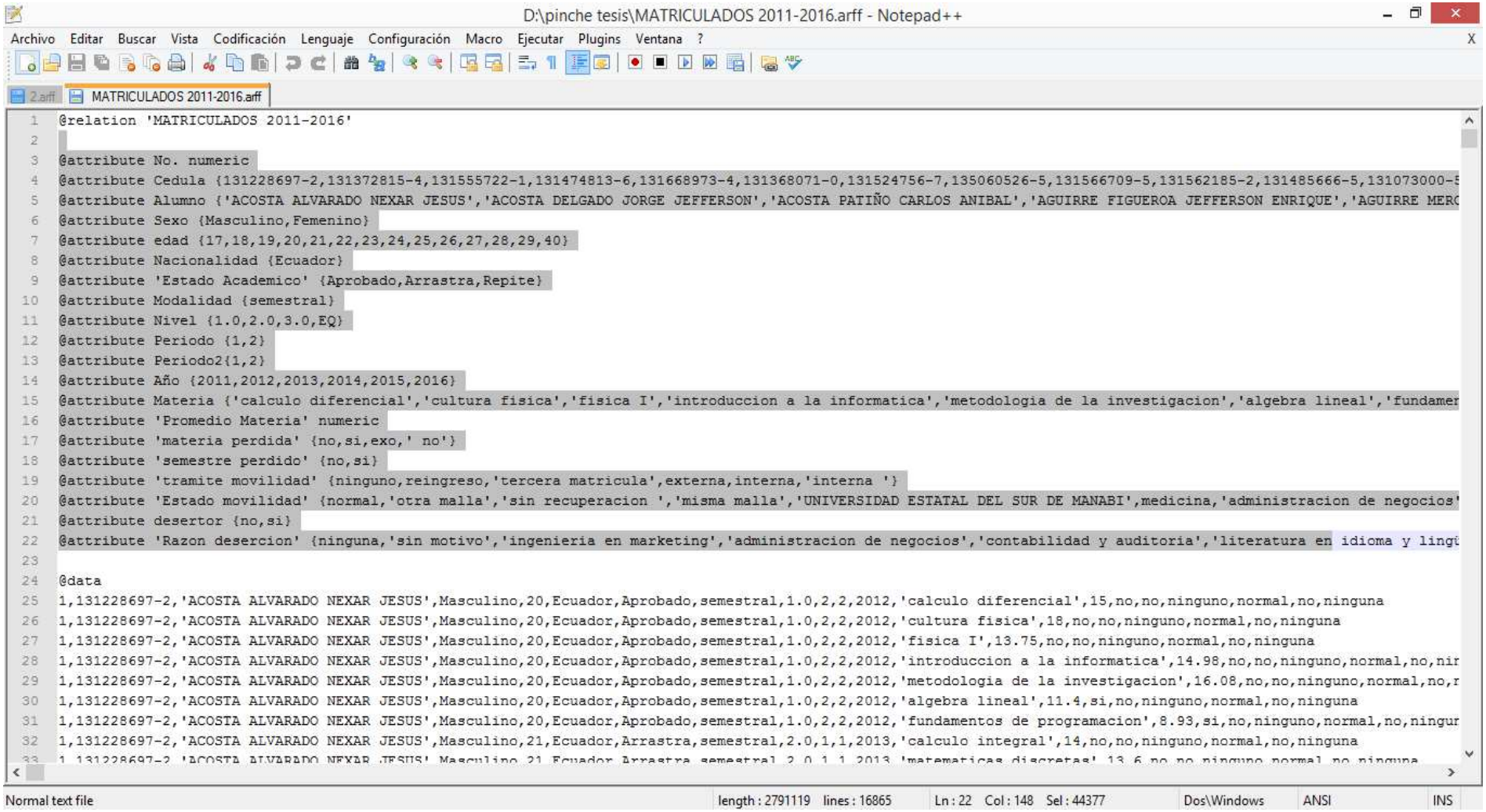
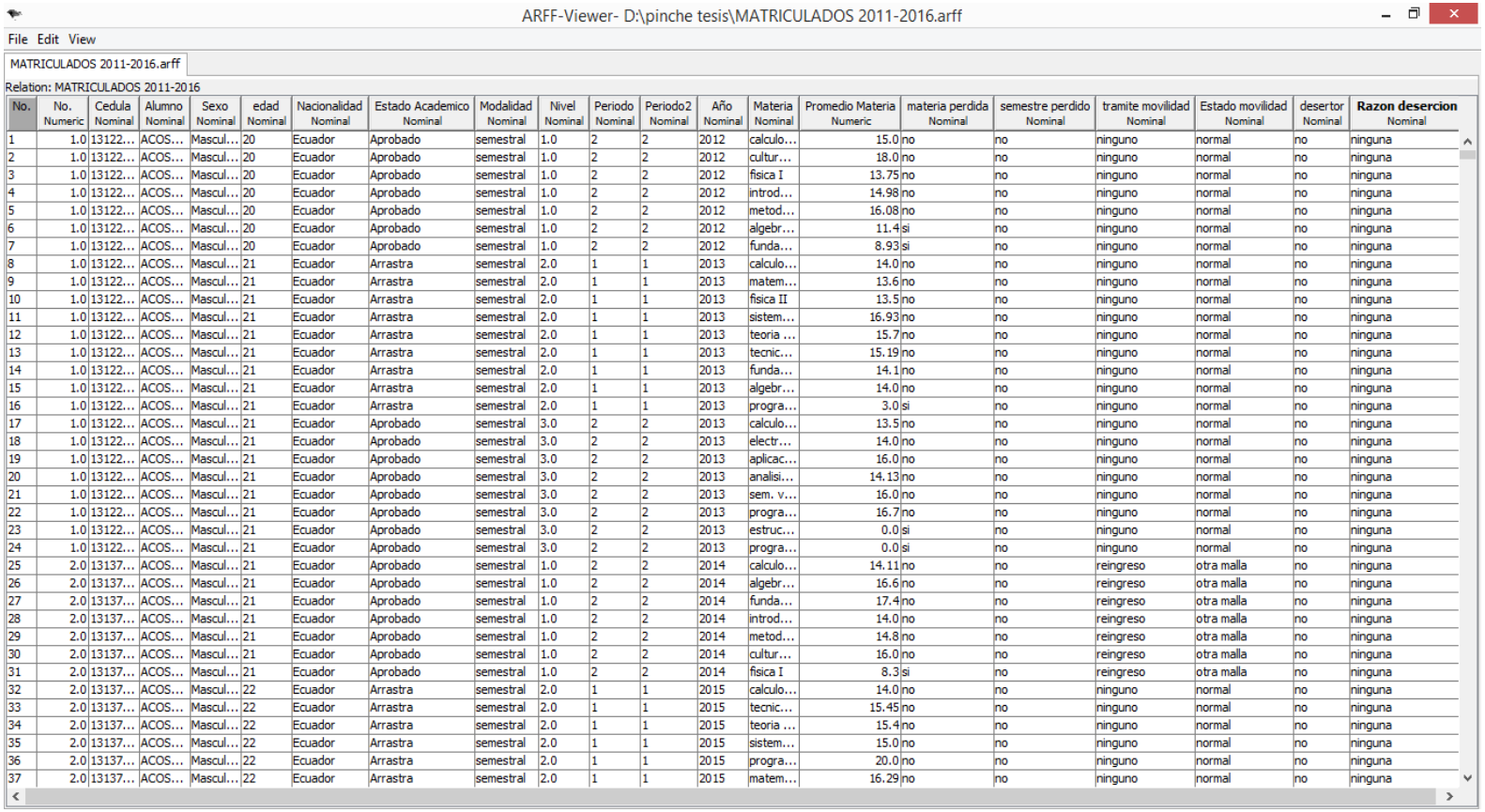


Ilustración 30: Conversión de datos a tipos numéricos y nominales



ARFF-Viewer- D:\pinche tesis\MATRICULADOS 2011-2016.arff

File Edit View

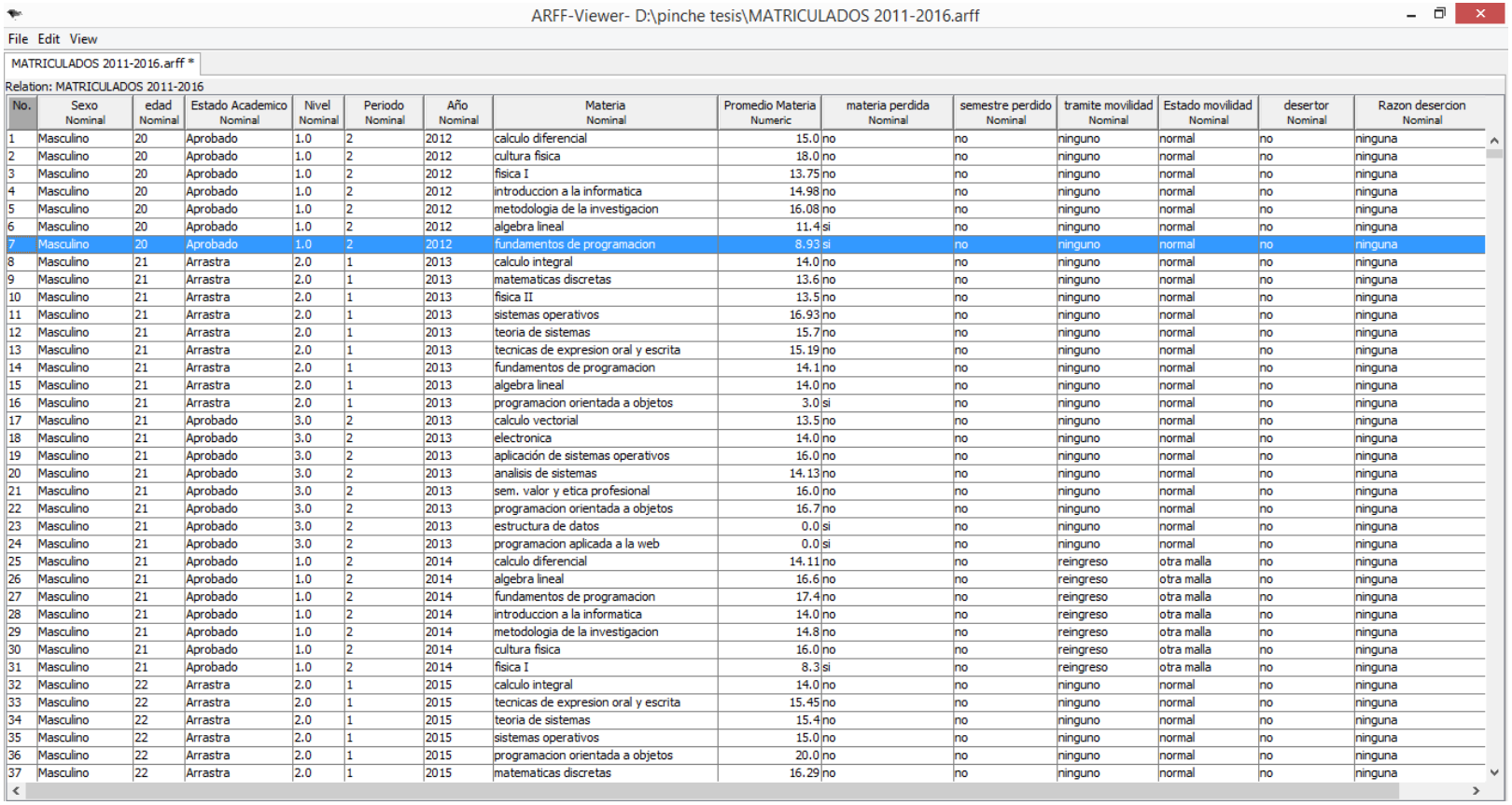
MATRICULADOS 2011-2016.arff

Relation: MATRICULADOS 2011-2016

No.	No. Numeric	Cedula Nominal	Alumno Nominal	Sexo Nominal	edad Nominal	Nacionalidad Nominal	Estado Academico Nominal	Modalidad Nominal	Nivel Nominal	Periodo Nominal	Periodo2 Nominal	Año Nominal	Materia Nominal	Promedio Materia Numeric	materia perdida Nominal	semestre perdido Nominal	tramite movilidad Nominal	Estado movilidad Nominal	desertor Nominal	Razon desercion Nominal
1	1.0	13122...	ACOS...	Mascul...	20	Ecuador	Aprobado	semestral	1.0	2	2	2012	calculo...	15.0	no	no	ninguno	normal	no	ninguna
2	1.0	13122...	ACOS...	Mascul...	20	Ecuador	Aprobado	semestral	1.0	2	2	2012	cultur...	18.0	no	no	ninguno	normal	no	ninguna
3	1.0	13122...	ACOS...	Mascul...	20	Ecuador	Aprobado	semestral	1.0	2	2	2012	fisica I	13.75	no	no	ninguno	normal	no	ninguna
4	1.0	13122...	ACOS...	Mascul...	20	Ecuador	Aprobado	semestral	1.0	2	2	2012	introd...	14.98	no	no	ninguno	normal	no	ninguna
5	1.0	13122...	ACOS...	Mascul...	20	Ecuador	Aprobado	semestral	1.0	2	2	2012	metod...	16.08	no	no	ninguno	normal	no	ninguna
6	1.0	13122...	ACOS...	Mascul...	20	Ecuador	Aprobado	semestral	1.0	2	2	2012	algebr...	11.4	si	no	ninguno	normal	no	ninguna
7	1.0	13122...	ACOS...	Mascul...	20	Ecuador	Aprobado	semestral	1.0	2	2	2012	funda...	8.93	si	no	ninguno	normal	no	ninguna
8	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Arrastra	semestral	2.0	1	1	2013	calculo...	14.0	no	no	ninguno	normal	no	ninguna
9	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Arrastra	semestral	2.0	1	1	2013	matem...	13.6	no	no	ninguno	normal	no	ninguna
10	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Arrastra	semestral	2.0	1	1	2013	fisica II	13.5	no	no	ninguno	normal	no	ninguna
11	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Arrastra	semestral	2.0	1	1	2013	sistem...	16.93	no	no	ninguno	normal	no	ninguna
12	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Arrastra	semestral	2.0	1	1	2013	teoria ...	15.7	no	no	ninguno	normal	no	ninguna
13	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Arrastra	semestral	2.0	1	1	2013	tecnic...	15.19	no	no	ninguno	normal	no	ninguna
14	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Arrastra	semestral	2.0	1	1	2013	funda...	14.1	no	no	ninguno	normal	no	ninguna
15	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Arrastra	semestral	2.0	1	1	2013	algebr...	14.0	no	no	ninguno	normal	no	ninguna
16	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Arrastra	semestral	2.0	1	1	2013	progra...	3.0	si	no	ninguno	normal	no	ninguna
17	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	3.0	2	2	2013	calculo...	13.5	no	no	ninguno	normal	no	ninguna
18	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	3.0	2	2	2013	electr...	14.0	no	no	ninguno	normal	no	ninguna
19	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	3.0	2	2	2013	aplicac...	16.0	no	no	ninguno	normal	no	ninguna
20	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	3.0	2	2	2013	analisi...	14.13	no	no	ninguno	normal	no	ninguna
21	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	3.0	2	2	2013	sem. v...	16.0	no	no	ninguno	normal	no	ninguna
22	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	3.0	2	2	2013	progra...	16.7	no	no	ninguno	normal	no	ninguna
23	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	3.0	2	2	2013	estruc...	0.0	si	no	ninguno	normal	no	ninguna
24	1.0	13122...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	3.0	2	2	2013	progra...	0.0	si	no	ninguno	normal	no	ninguna
25	2.0	13137...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	1.0	2	2	2014	calculo...	14.11	no	no	reingreso	otra malla	no	ninguna
26	2.0	13137...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	1.0	2	2	2014	algebr...	16.6	no	no	reingreso	otra malla	no	ninguna
27	2.0	13137...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	1.0	2	2	2014	funda...	17.4	no	no	reingreso	otra malla	no	ninguna
28	2.0	13137...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	1.0	2	2	2014	introd...	14.0	no	no	reingreso	otra malla	no	ninguna
29	2.0	13137...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	1.0	2	2	2014	metod...	14.8	no	no	reingreso	otra malla	no	ninguna
30	2.0	13137...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	1.0	2	2	2014	cultur...	16.0	no	no	reingreso	otra malla	no	ninguna
31	2.0	13137...	ACOS...	Mascul...	21	Ecuador	Aprobado	semestral	1.0	2	2	2014	fisica I	8.3	si	no	reingreso	otra malla	no	ninguna
32	2.0	13137...	ACOS...	Mascul...	22	Ecuador	Arrastra	semestral	2.0	1	1	2015	calculo...	14.0	no	no	ninguno	normal	no	ninguna
33	2.0	13137...	ACOS...	Mascul...	22	Ecuador	Arrastra	semestral	2.0	1	1	2015	tecnic...	15.45	no	no	ninguno	normal	no	ninguna
34	2.0	13137...	ACOS...	Mascul...	22	Ecuador	Arrastra	semestral	2.0	1	1	2015	teoria ...	15.4	no	no	ninguno	normal	no	ninguna
35	2.0	13137...	ACOS...	Mascul...	22	Ecuador	Arrastra	semestral	2.0	1	1	2015	sistem...	15.0	no	no	ninguno	normal	no	ninguna
36	2.0	13137...	ACOS...	Mascul...	22	Ecuador	Arrastra	semestral	2.0	1	1	2015	progra...	20.0	no	no	ninguno	normal	no	ninguna
37	2.0	13137...	ACOS...	Mascul...	22	Ecuador	Arrastra	semestral	2.0	1	1	2015	matem...	16.29	no	no	ninguno	normal	no	ninguna

Ilustración 31: Archivo ARFF matriculados 2011-2016 con su respectivo tipo de dato.

En un proceso de data mining, los datos pueden estar dispersos en la empresa y almacenados en formatos distintos; en el caso de los estudiantes matriculados estos han quedado del siguiente modo en un único archivo de extensión arff. Note como junto al nombre del atributo se muestra su tipo, que puede ser numérico o nominal como se ha dicho hasta ahora.



ARFF-Viewer- D:\pinche tesis\MATRICULADOS 2011-2016.arff

File Edit View

MATRICULADOS 2011-2016.arff *

Relation: MATRICULADOS 2011-2016

No.	Sexo Nominal	edad Nominal	Estado Academico Nominal	Nivel Nominal	Periodo Nominal	Año Nominal	Materia Nominal	Promedio Materia Numeric	materia perdida Nominal	semestre perdido Nominal	tramite movilidad Nominal	Estado movilidad Nominal	desertor Nominal	Razon desercion Nominal
1	Masculino	20	Aprobado	1.0	2	2012	calculo diferencial	15.0	no	no	ninguno	normal	no	ninguna
2	Masculino	20	Aprobado	1.0	2	2012	cultura fisica	18.0	no	no	ninguno	normal	no	ninguna
3	Masculino	20	Aprobado	1.0	2	2012	fisica I	13.75	no	no	ninguno	normal	no	ninguna
4	Masculino	20	Aprobado	1.0	2	2012	introduccion a la informatica	14.98	no	no	ninguno	normal	no	ninguna
5	Masculino	20	Aprobado	1.0	2	2012	metodologia de la investigacion	16.08	no	no	ninguno	normal	no	ninguna
6	Masculino	20	Aprobado	1.0	2	2012	algebra lineal	11.4	si	no	ninguno	normal	no	ninguna
7	Masculino	20	Aprobado	1.0	2	2012	fundamentos de programacion	8.93	si	no	ninguno	normal	no	ninguna
8	Masculino	21	Arrastra	2.0	1	2013	calculo integral	14.0	no	no	ninguno	normal	no	ninguna
9	Masculino	21	Arrastra	2.0	1	2013	matematicas discretas	13.6	no	no	ninguno	normal	no	ninguna
10	Masculino	21	Arrastra	2.0	1	2013	fisica II	13.5	no	no	ninguno	normal	no	ninguna
11	Masculino	21	Arrastra	2.0	1	2013	sistemas operativos	16.93	no	no	ninguno	normal	no	ninguna
12	Masculino	21	Arrastra	2.0	1	2013	teoria de sistemas	15.7	no	no	ninguno	normal	no	ninguna
13	Masculino	21	Arrastra	2.0	1	2013	tecnicas de expresion oral y escrita	15.19	no	no	ninguno	normal	no	ninguna
14	Masculino	21	Arrastra	2.0	1	2013	fundamentos de programacion	14.1	no	no	ninguno	normal	no	ninguna
15	Masculino	21	Arrastra	2.0	1	2013	algebra lineal	14.0	no	no	ninguno	normal	no	ninguna
16	Masculino	21	Arrastra	2.0	1	2013	programacion orientada a objetos	3.0	si	no	ninguno	normal	no	ninguna
17	Masculino	21	Aprobado	3.0	2	2013	calculo vectorial	13.5	no	no	ninguno	normal	no	ninguna
18	Masculino	21	Aprobado	3.0	2	2013	electronica	14.0	no	no	ninguno	normal	no	ninguna
19	Masculino	21	Aprobado	3.0	2	2013	aplicación de sistemas operativos	16.0	no	no	ninguno	normal	no	ninguna
20	Masculino	21	Aprobado	3.0	2	2013	analisis de sistemas	14.13	no	no	ninguno	normal	no	ninguna
21	Masculino	21	Aprobado	3.0	2	2013	sem. valor y etica profesional	16.0	no	no	ninguno	normal	no	ninguna
22	Masculino	21	Aprobado	3.0	2	2013	programacion orientada a objetos	16.7	no	no	ninguno	normal	no	ninguna
23	Masculino	21	Aprobado	3.0	2	2013	estructura de datos	0.0	si	no	ninguno	normal	no	ninguna
24	Masculino	21	Aprobado	3.0	2	2013	programacion aplicada a la web	0.0	si	no	ninguno	normal	no	ninguna
25	Masculino	21	Aprobado	1.0	2	2014	calculo diferencial	14.11	no	no	reingreso	otra malla	no	ninguna
26	Masculino	21	Aprobado	1.0	2	2014	algebra lineal	16.6	no	no	reingreso	otra malla	no	ninguna
27	Masculino	21	Aprobado	1.0	2	2014	fundamentos de programacion	17.4	no	no	reingreso	otra malla	no	ninguna
28	Masculino	21	Aprobado	1.0	2	2014	introduccion a la informatica	14.0	no	no	reingreso	otra malla	no	ninguna
29	Masculino	21	Aprobado	1.0	2	2014	metodologia de la investigacion	14.8	no	no	reingreso	otra malla	no	ninguna
30	Masculino	21	Aprobado	1.0	2	2014	cultura fisica	16.0	no	no	reingreso	otra malla	no	ninguna
31	Masculino	21	Aprobado	1.0	2	2014	fisica I	8.3	si	no	reingreso	otra malla	no	ninguna
32	Masculino	22	Arrastra	2.0	1	2015	calculo integral	14.0	no	no	ninguno	normal	no	ninguna
33	Masculino	22	Arrastra	2.0	1	2015	tecnicas de expresion oral y escrita	15.45	no	no	ninguno	normal	no	ninguna
34	Masculino	22	Arrastra	2.0	1	2015	teoria de sistemas	15.4	no	no	ninguno	normal	no	ninguna
35	Masculino	22	Arrastra	2.0	1	2015	sistemas operativos	15.0	no	no	ninguno	normal	no	ninguna
36	Masculino	22	Arrastra	2.0	1	2015	programacion orientada a objetos	20.0	no	no	ninguno	normal	no	ninguna
37	Masculino	22	Arrastra	2.0	1	2015	matematicas discretas	16.29	no	no	ninguno	normal	no	ninguna

Ilustración 32: Vista parcial de 16840 registros devueltos por WEKA

La limpieza de datos no solamente implica quitar los datos no válidos o interpolar valores que faltan, sino también buscar las correlaciones ocultas en los datos, identificar los orígenes de datos que son más precisos y determinar qué columnas son las más adecuadas para el análisis.

3.4.5. FASE III: Explorar datos

Los datos de los que se dispone en primera instancia son los que se muestran en la siguiente ilustración:

Ilustración 33: Datos y tipos disponibles en el archivo arff

Las decisiones se extraen principalmente de los atributos, promedio de materia, materia perdida, semestre perdido, tramite de movilidad, estado de movilidad y desertor.

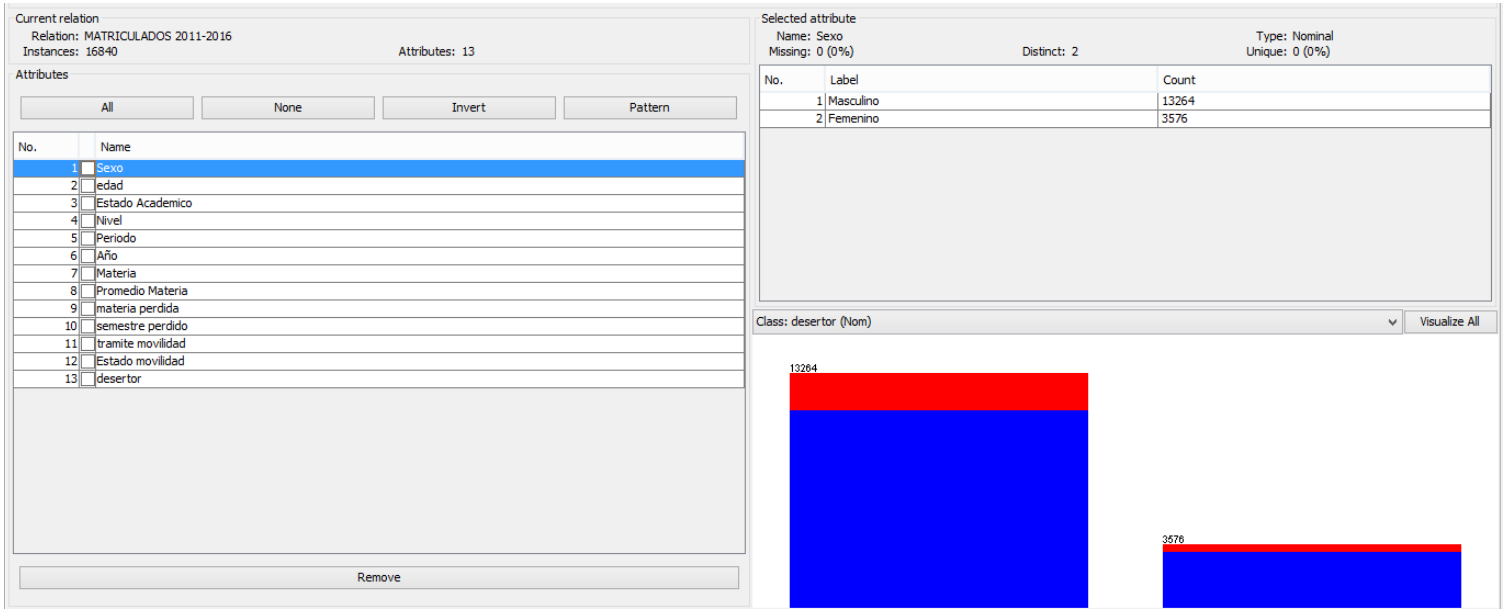


Ilustración 34: Estadísticas de sexo, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías: masculino y femenino.

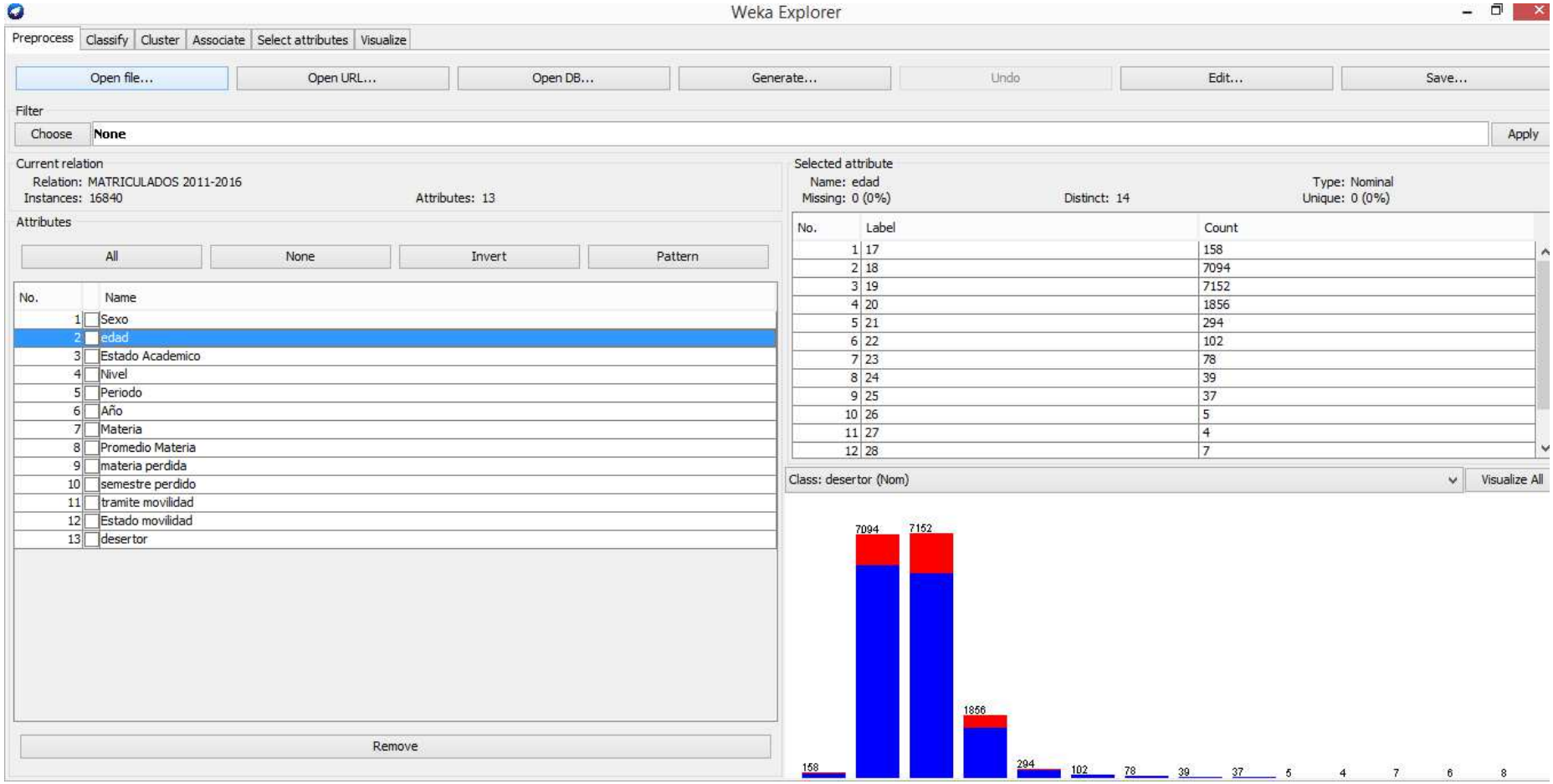


Ilustración 35: Estadísticas edad (de estudiantes del periodo 2011-2016 modalidad semestral), mismo que es relevante para las predicciones y clasificaciones buscadas, pues se registran 14 números de años de edad distintos entre los 16840 registros.

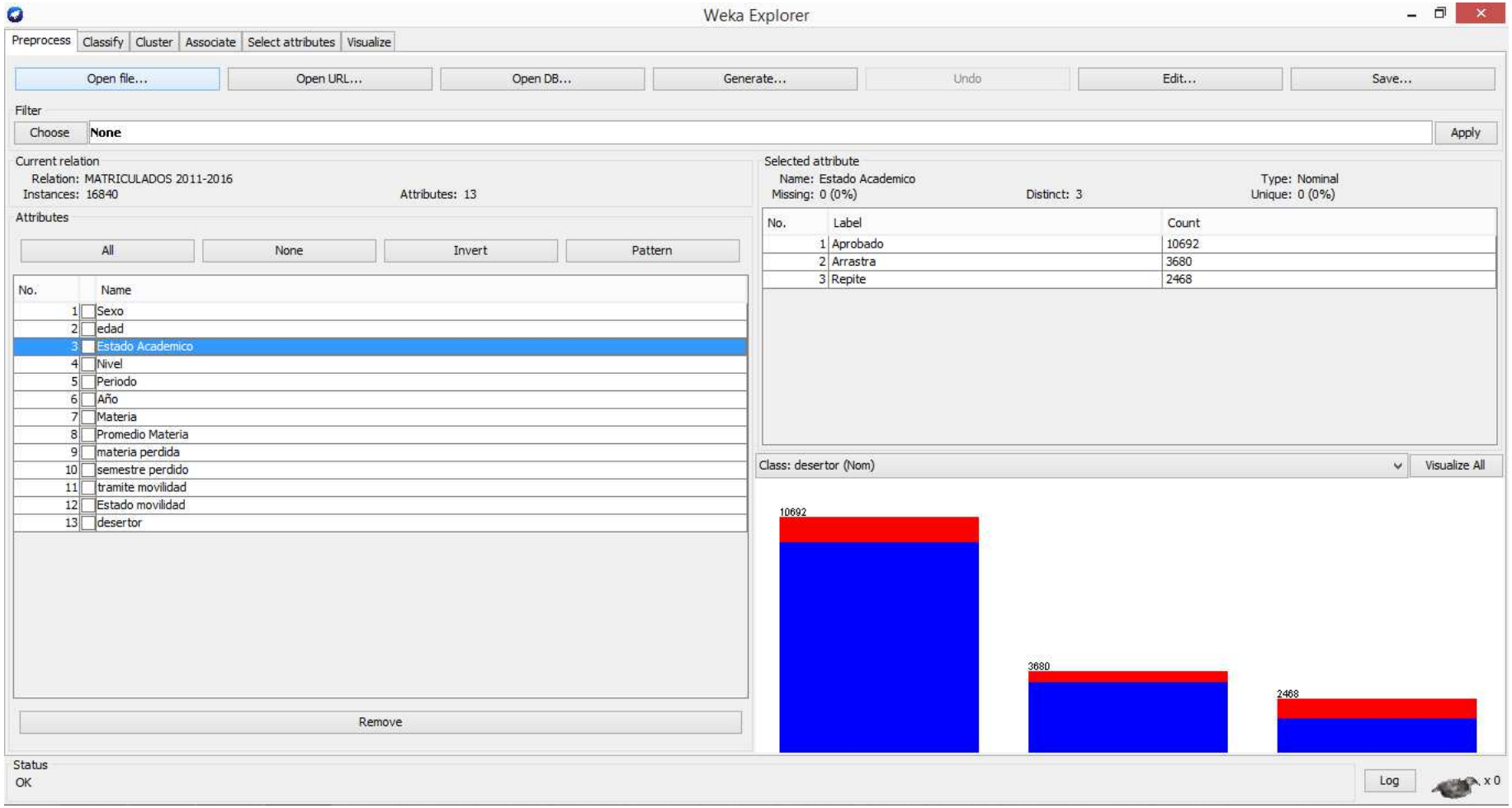


Ilustración 36: Estadísticas de Estado académico, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías: aprobado, arrastra y repite.

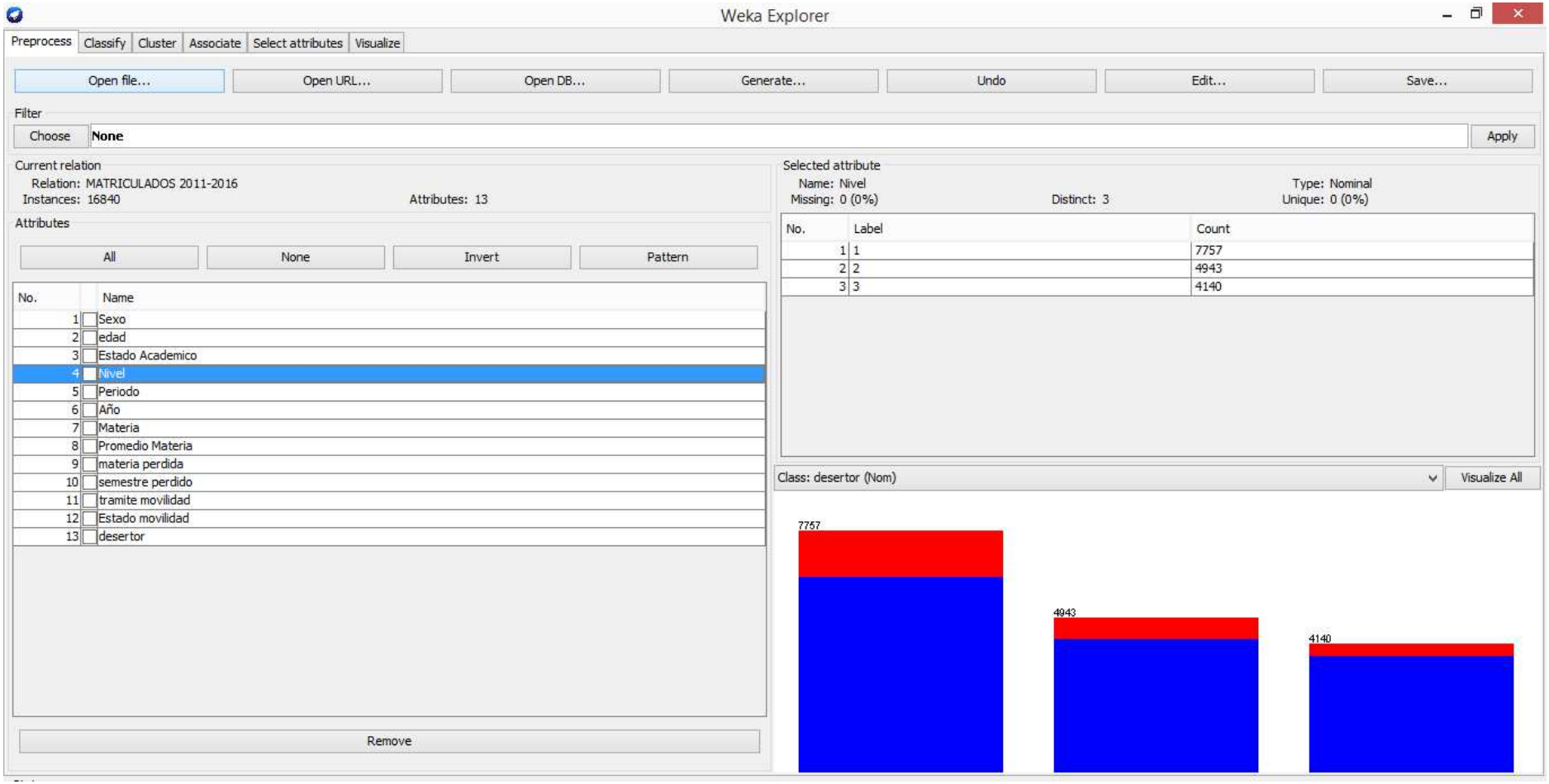


Ilustración 37: Estadísticas de Nivel, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías: primer nivel (1), segundo nivel (2) y tercer nivel (3).

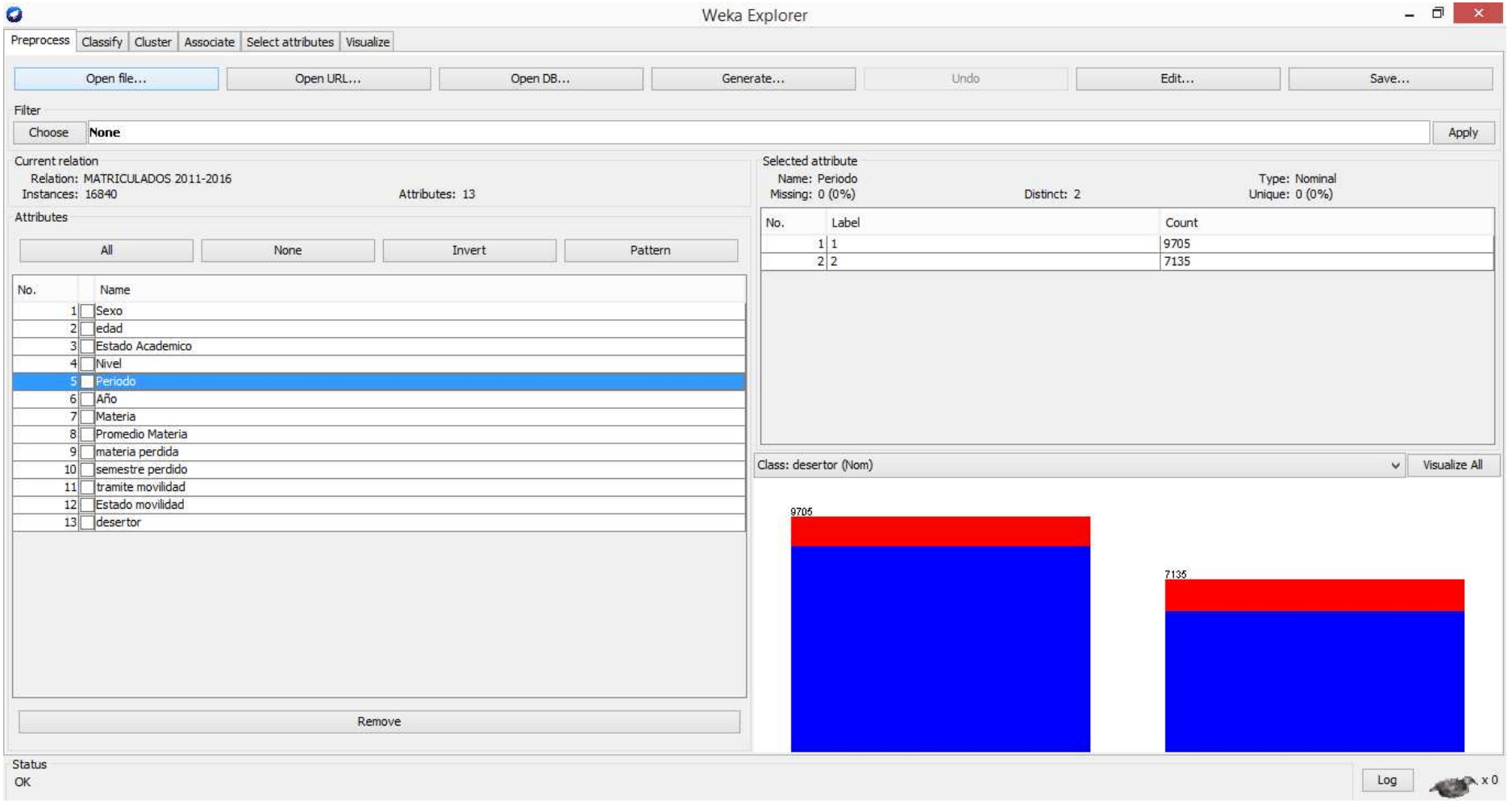


Ilustración 38: Estadísticas de Periodo, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías: periodo semestral (1 y 2).

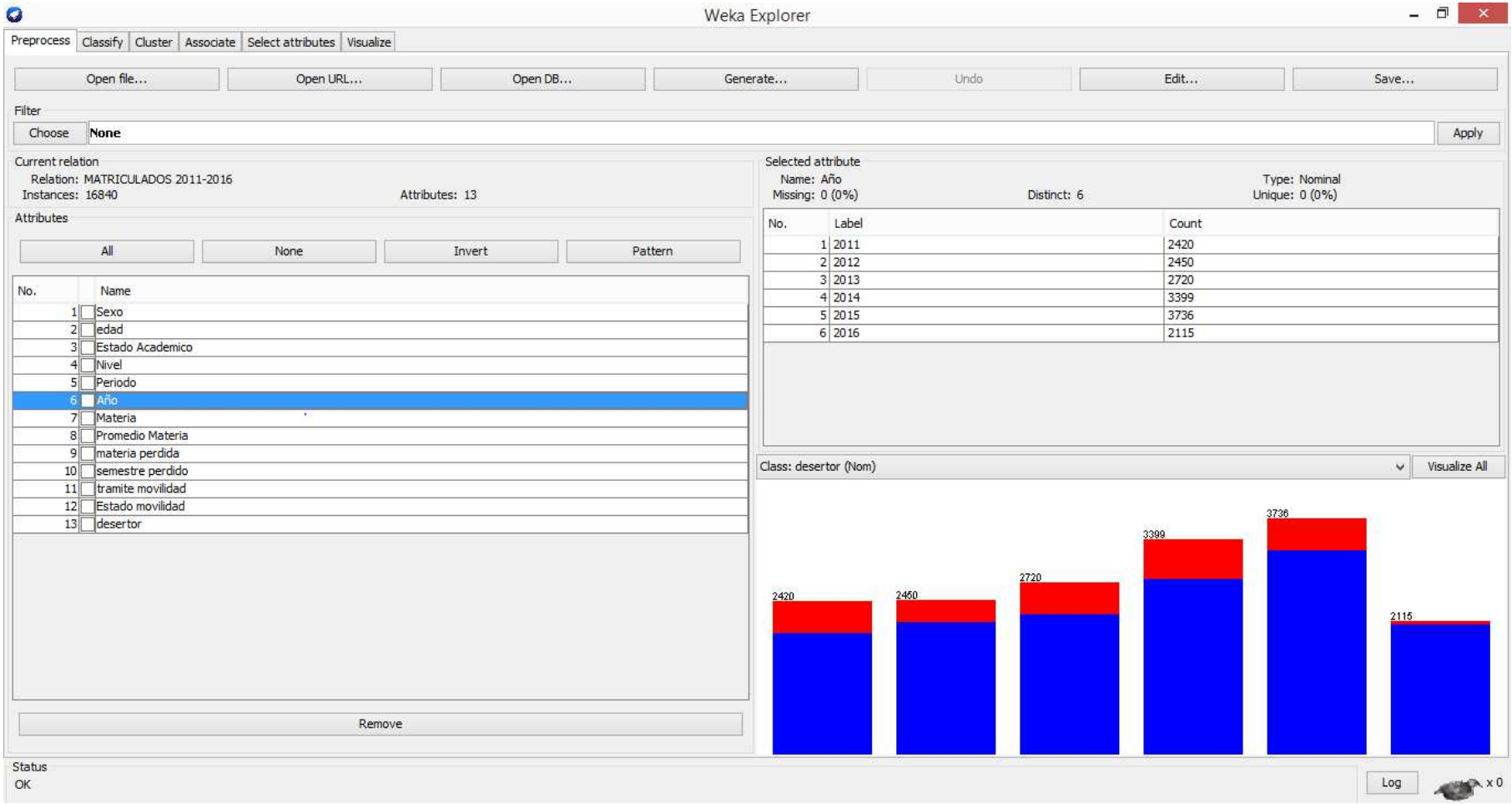


Ilustración 39: Estadísticas de Año, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías comprendidas entre el año 2011 – 2016(1).

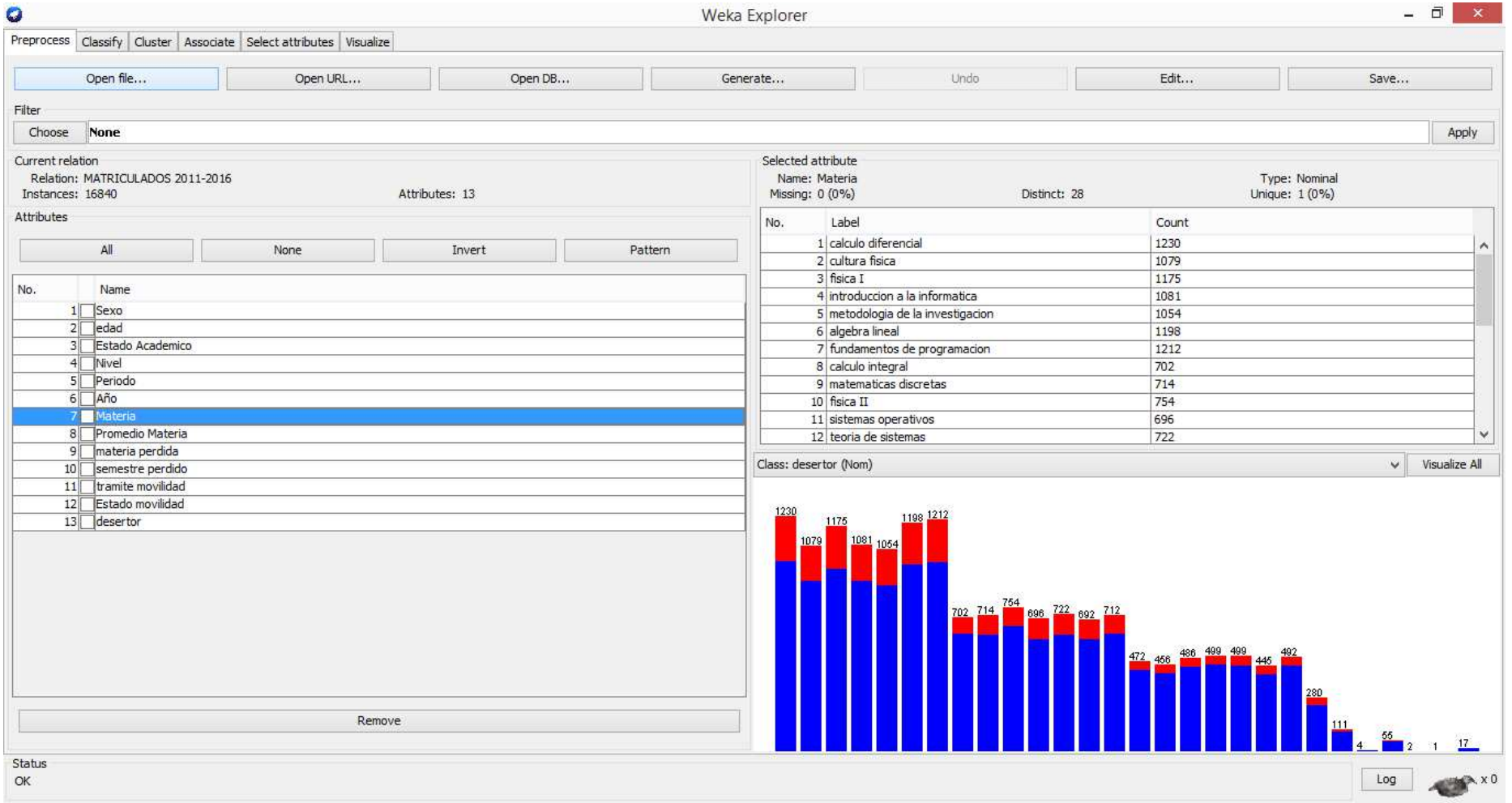


Ilustración 40: Estadísticas de Materia, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías de las materias comprendidas de primer a tercer nivel correspondiente a la malla curricular de la facultad.

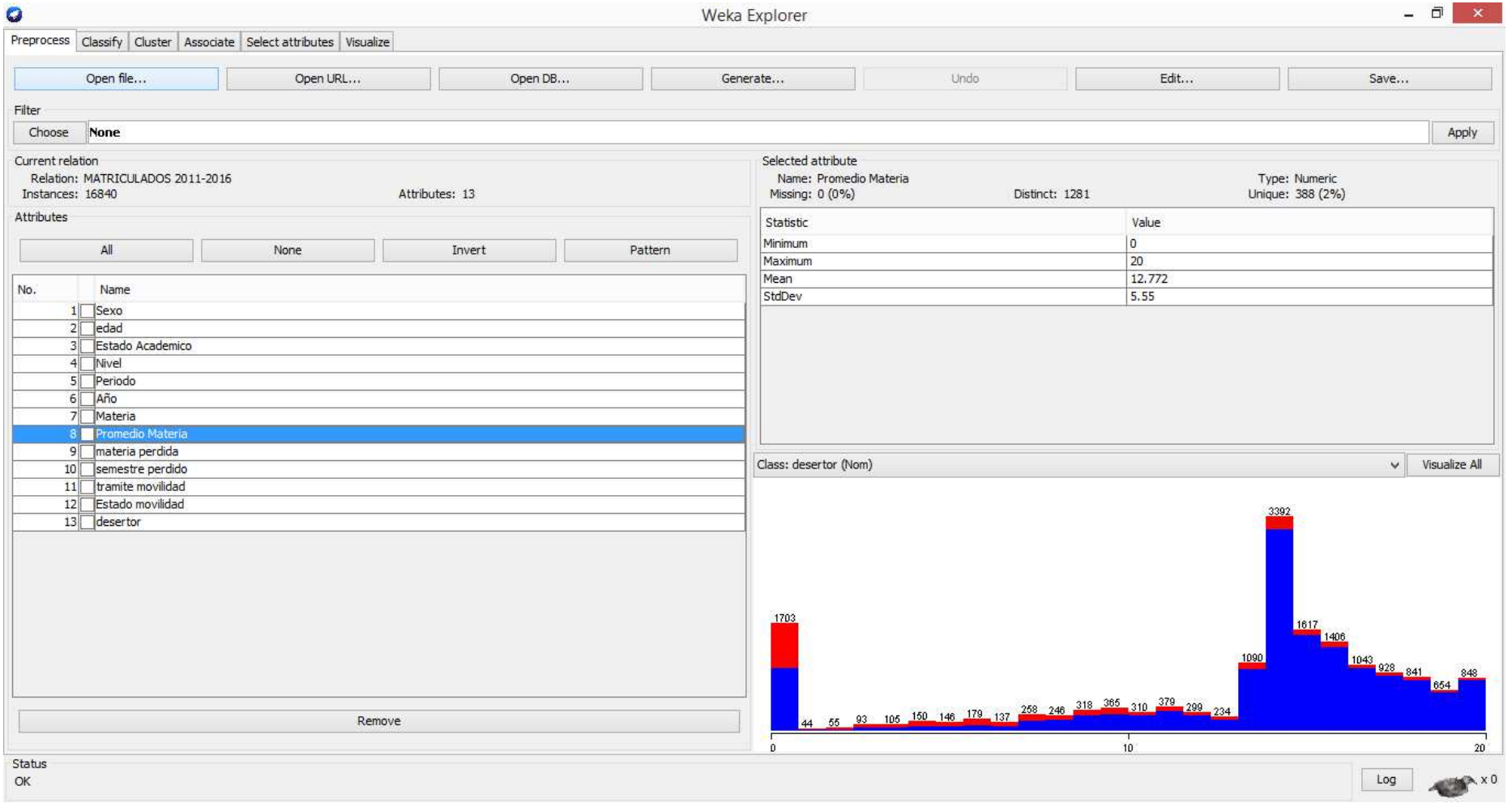


Ilustración 41: Estadísticas de Promedio materia, mismo que es relevante para las predicciones y clasificaciones buscadas, pues la desviación estándar (StdDev) es baja respecto del promedio (Mean).

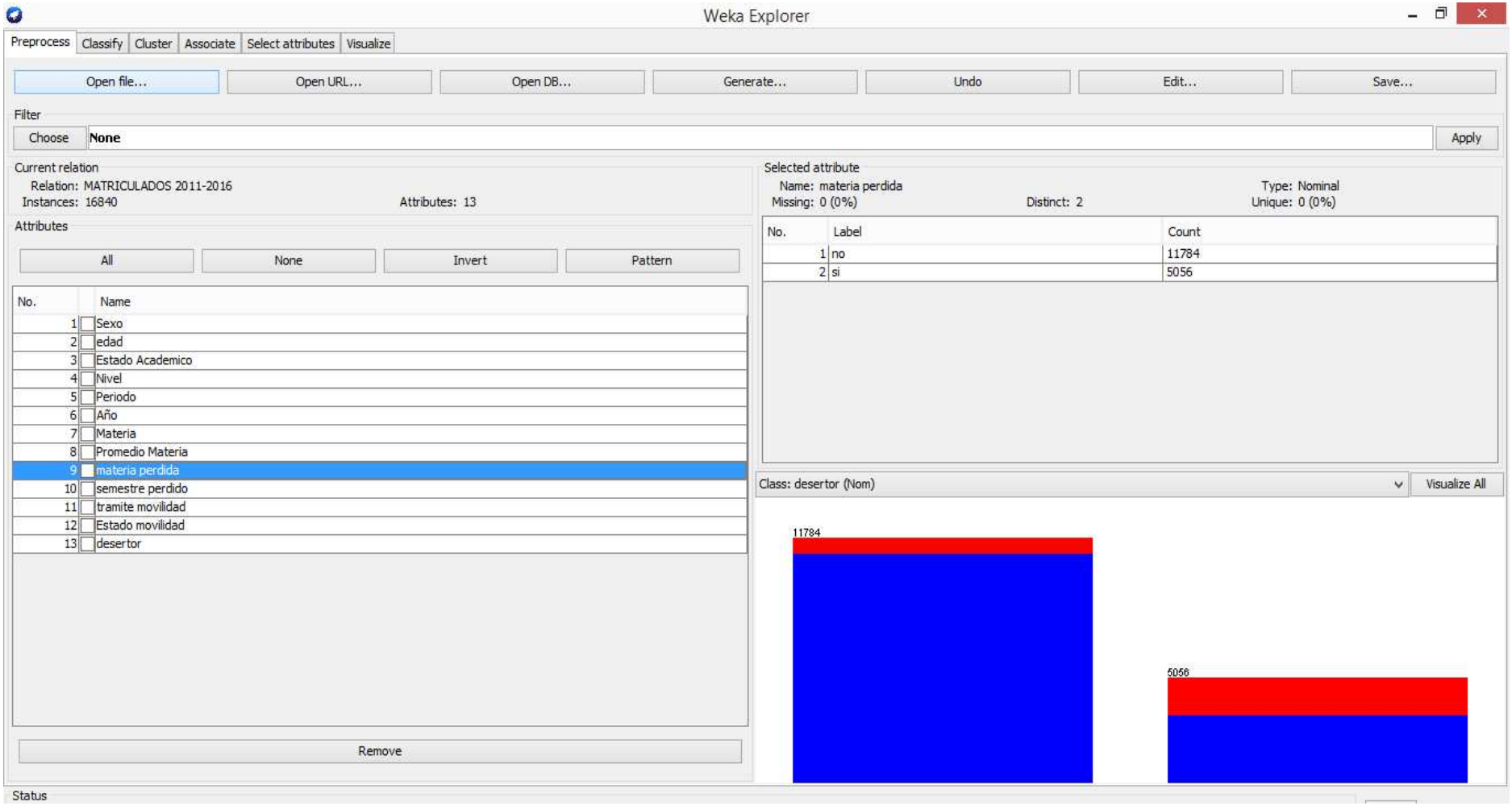


Ilustración 42: Estadísticas de materia perdida, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías: sí y no.

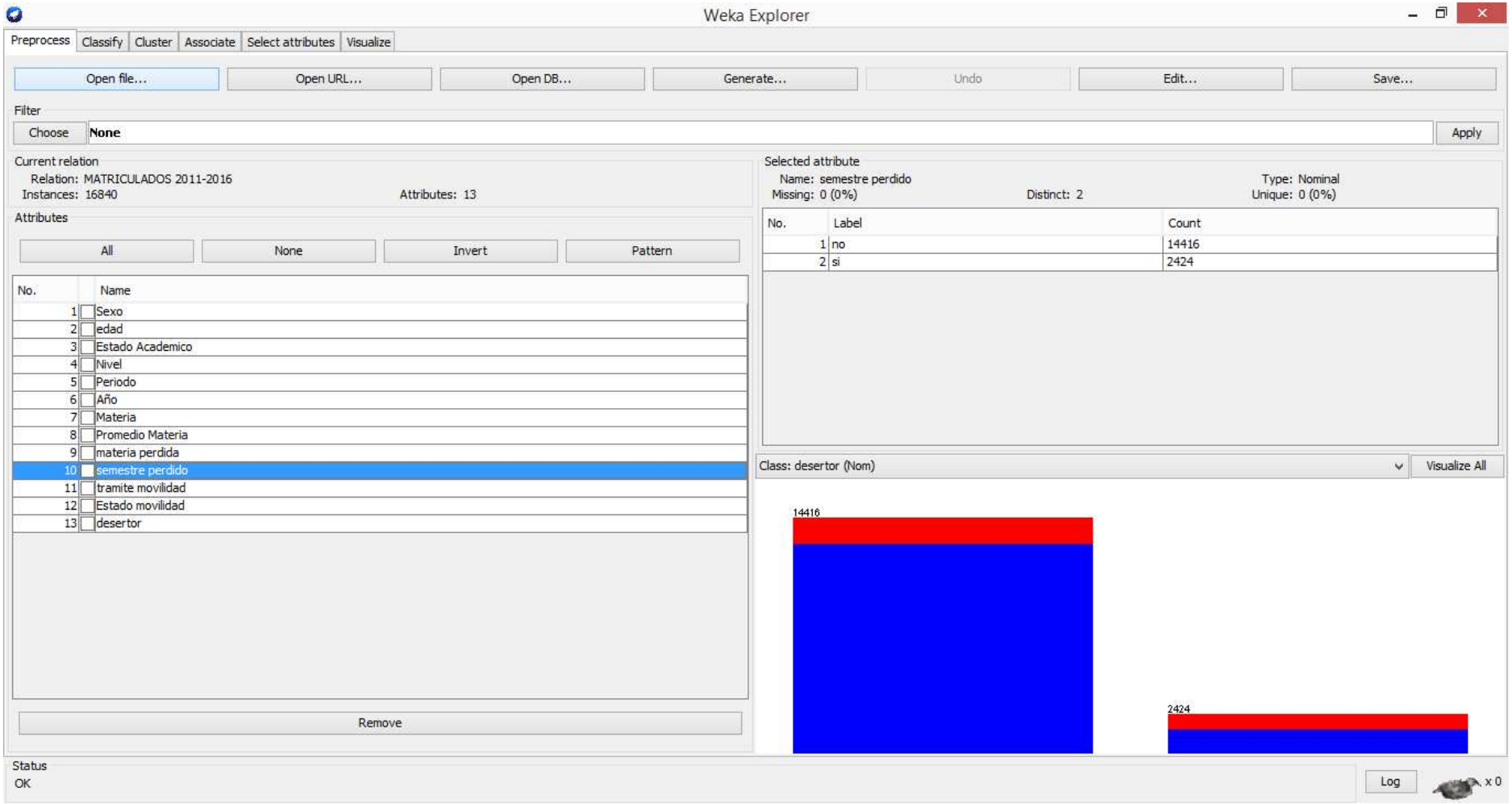


Ilustración 43: Estadísticas de semestre perdido, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías: sí y no.

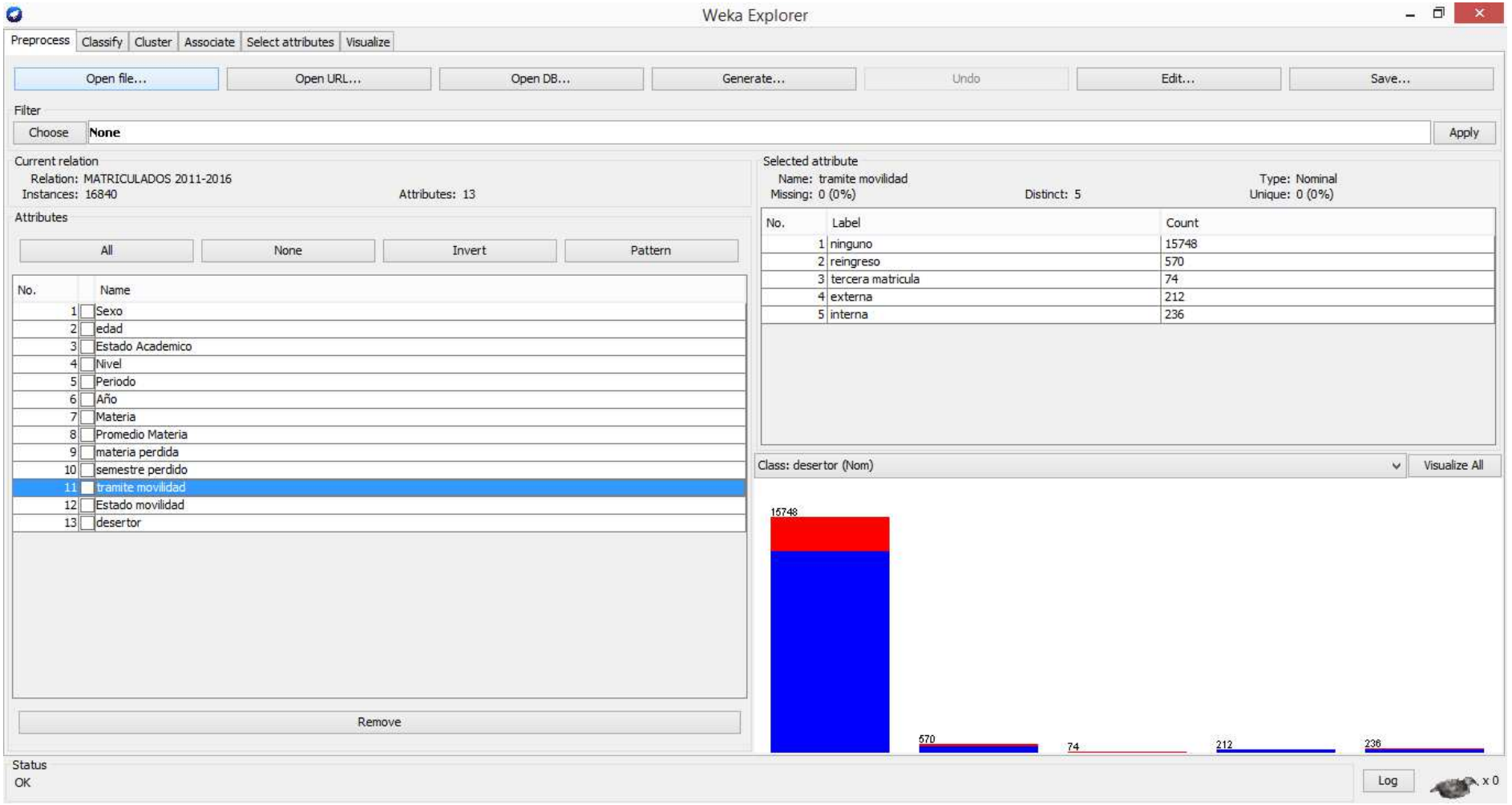


Ilustración 44: Estadísticas de trámite de movilidad, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías: ninguno, reingreso, tercera matricula, interna y externa.

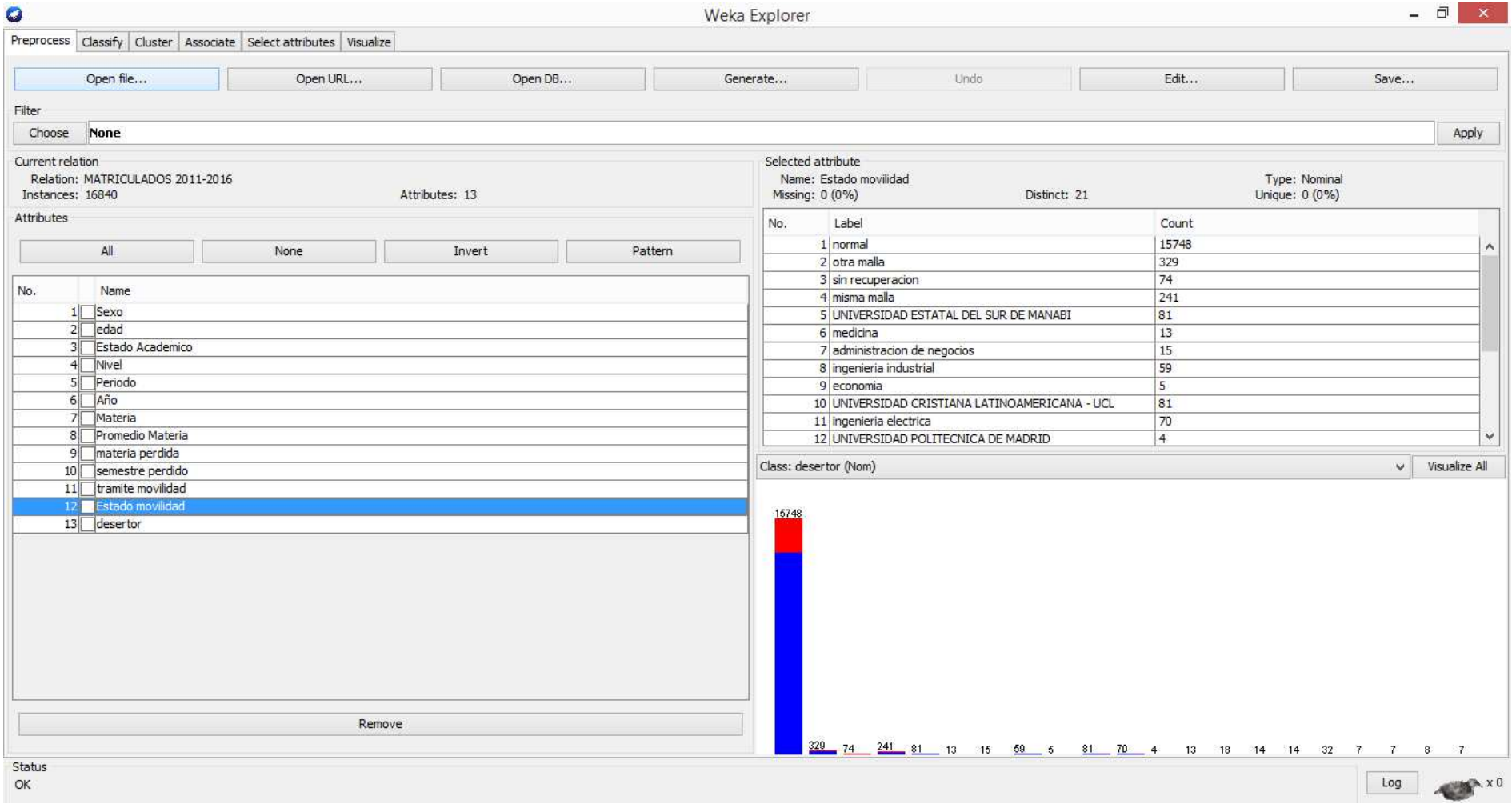


Ilustración 45: Estadísticas de estado de movilidad, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías relacionadas con el trámite de movilidad pues se observa de donde y hacia donde se dirigen los estudiantes.

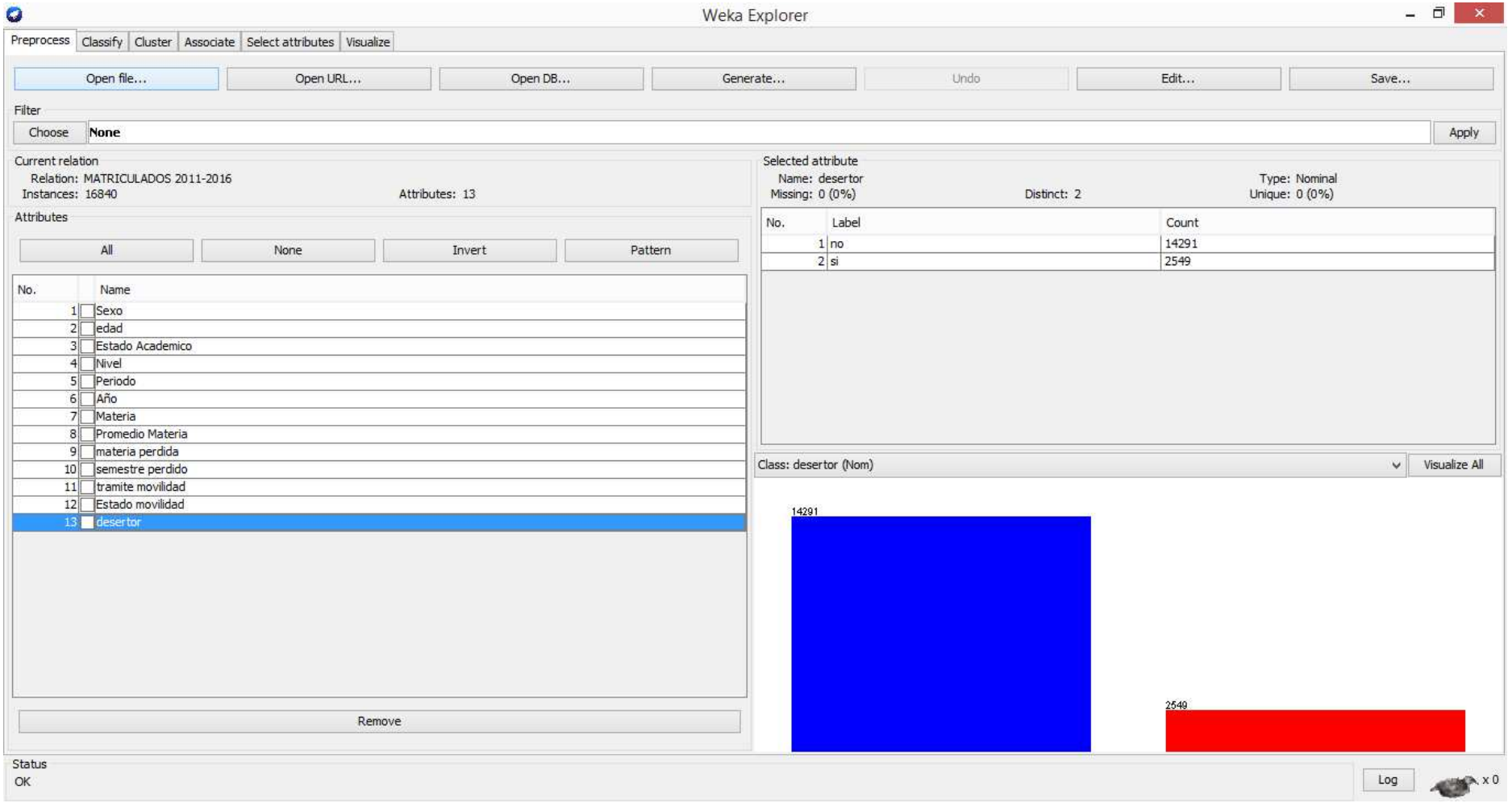


Ilustración 46: Estadísticas de desertor, mismo que es relevante para las predicciones y clasificaciones buscadas, pues refleja las categorías del estudiante dentro del entorno académico demostrando quienes es desertor o no.

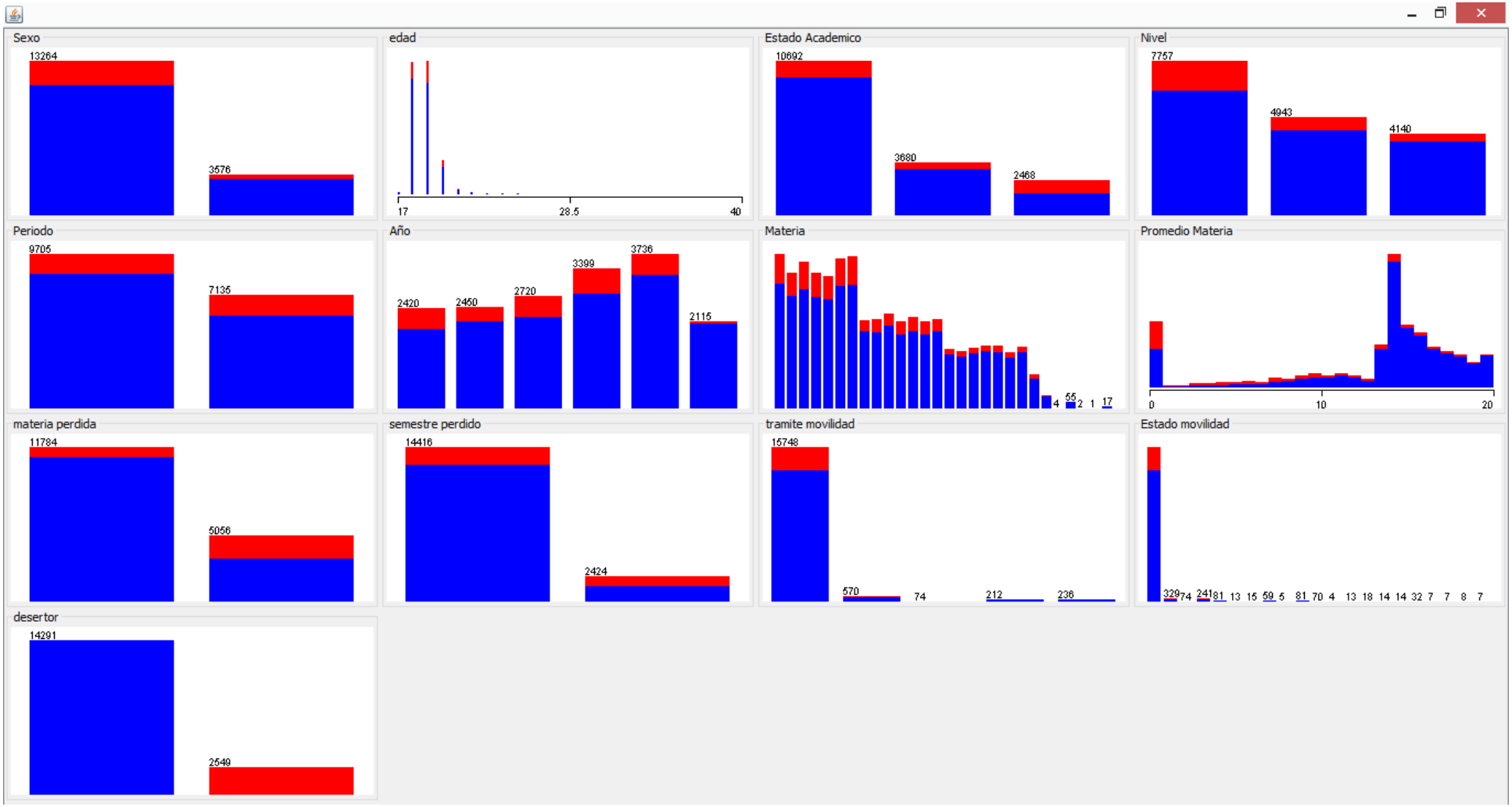


Ilustración 47: Resumen del perfil estadístico de todos los atributos empleando WEKA

Como se ha visto, entre las técnicas de exploración se incluyen calcular los valores mínimos y máximos, calcular la media y las desviaciones estándar, y examinar la distribución de los datos. Por ejemplo, al revisar el máximo, el mínimo y los valores de la media se podrían determinar que los datos no son representativos de los abonados o procesos de negocio, y que por consiguiente debe obtener más datos equilibrados o revisar las suposiciones que son la base de sus expectativas. Las desviaciones estándar y otros valores de distribución pueden proporcionar información útil sobre la estabilidad y exactitud de los resultados.

Una desviación estándar grande puede indicar que agregar más datos podría ayudarle a mejorar el modelo. Los datos que se desvían mucho de una distribución estándar se podrían sesgar o podrían representar una imagen precisa de un problema de la vida real, pero dificultar el ajustar un modelo a los datos.

Al explorar los datos para conocer el problema empresarial, puede decidir si el conjunto de datos contiene datos defectuosos y, a continuación, puede inventar una estrategia para corregir los problemas u obtener una descripción más profunda de los comportamientos que son típicos de su negocio.

3.4.6. FASE IV: Generar, Explorar y Validar los modelos

3.4.6.1. Modelo de árbol de decisión para identificar a quienes desertan, en función del atributo “promedio de materia”.

Para este ejercicio inicialmente se tiene:

- El número de instancias o registros es 16840.
- Se tienen trece atributos: sexo, edad, estado académico, nivel, periodo, año, materia, promedio materia, materia perdida, semestre perdido, trámite de movilidad, estado de movilidad y desertor.
- Se usa un conjunto de datos de entrenamiento (Use training set).

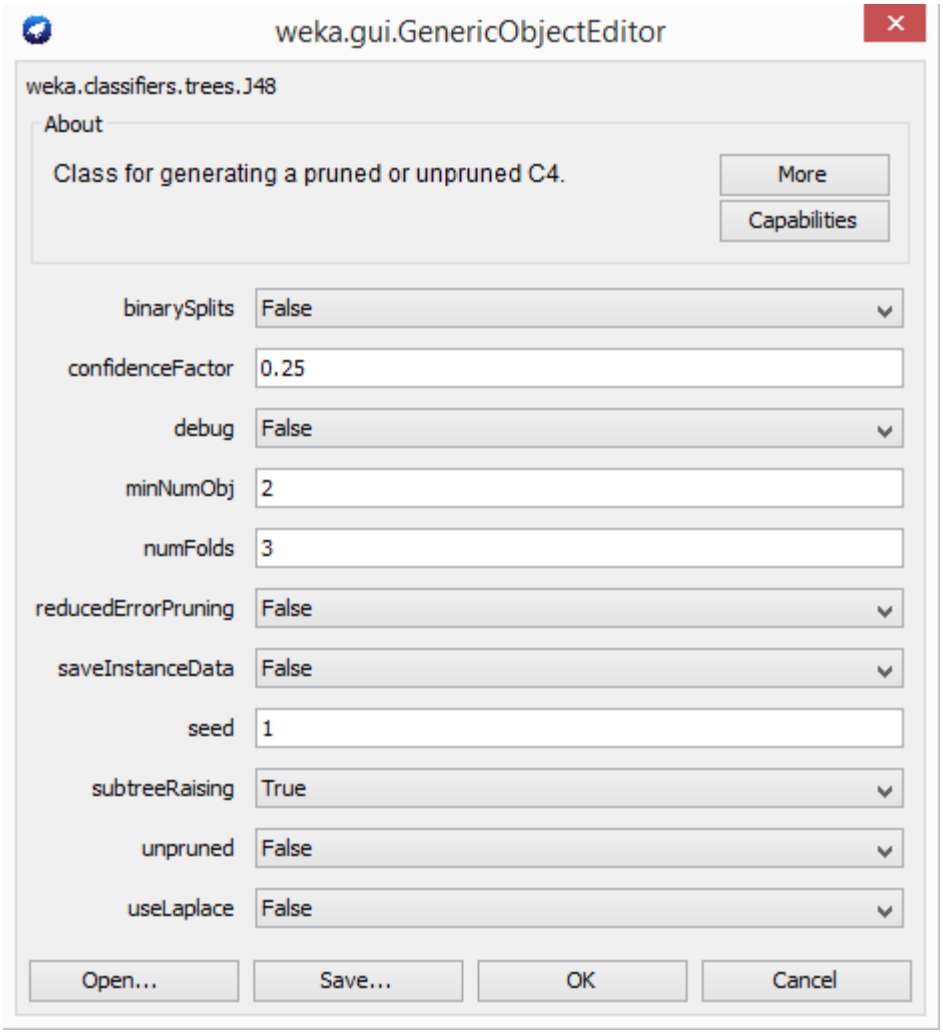


Ilustración 48: Configuración de las propiedades.

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     MATRICULADOS 2011-2016
Instances:    16840
Attributes:   13
              Sexo
              edad
              Estado Academico
              Nivel
              Periodo
              Año
              Materia
              Promedio Materia
              materia perdida
              semestre perdido
              tramite movilidad
              Estado movilidad
              desertor
Test mode:evaluate on training data

J48 pruned tree
------------------

Promedio Materia <= 13.42
|   semestre perdido = no
|   |   Estado movilidad = normal
|   |   |   Promedio Materia <= 7.6
|   |   |   |   Estado Academico = Aprobado
|   |   |   |   |   Año = 2011
|   |   |   |   |   |   Sexo = Masculino
|   |   |   |   |   |   |   Materia = calculo diferencial: no (55.0/24.0)
|   |   |   |   |   |   |   Materia = cultura fisica: si (9.0/3.0)
|   |   |   |   |   |   |   Materia = fisica I: si (37.0/16.0)
|   |   |   |   |   |   |   Materia = introduccion a la informatica: si (26.0/11.0)
|   |   |   |   |   |   |   Materia = metodologia de la investigacion: si (12.0/5.0)
|   |   |   |   |   |   |   Materia = algebra lineal
|   |   |   |   |   |   |   |   Promedio Materia <= 6.9: si (37.0/16.0)
|   |   |   |   |   |   |   |   Promedio Materia > 6.9
|   |   |   |   |   |   |   |   |   Promedio Materia <= 7.55: no (14.0)
|   |   |   |   |   |   |   |   |   Promedio Materia > 7.55
|   |   |   |   |   |   |   |   |   |   edad <= 18: no (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   edad > 18: si (2.0)
|   |   |   |   |   |   |   Materia = fundamentos de programacion
|   |   |   |   |   |   |   |   Promedio Materia <= 2.55: si (11.0/2.0)
|   |   |   |   |   |   |   |   Promedio Materia > 2.55: no (33.0/12.0)
|   |   |   |   |   |   |   Materia = calculo integral: no (0.0)
|   |   |   |   |   |   |   Materia = matematicas discretas: no (0.0)
|   |   |   |   |   |   |   Materia = fisica II: no (2.0)
|   |   |   |   |   |   |   Materia = sistemas operativos: no (1.0)
|   |   |   |   |   |   |   Materia = teoria de sistemas: no (0.0)
|   |   |   |   |   |   |   Materia = tecnicas de expresion oral y escrita: no (1.0)
|   |   |   |   |   |   |   Materia = programacion orientada a objetos: no (1.0)
```

```
|   |   |   |   |   |   |   Materia = calculo vectorial: no (0.0)
|   |   |   |   |   |   |   Materia = electronica: no (1.0)
|   |   |   |   |   |   |   Materia = aplicación de sistemas operativos: no (1.0)
|   |   |   |   |   |   |   Materia = analisis de sistemas: no (1.0)
|   |   |   |   |   |   |   Materia = sem. valor y etica profesional: no (1.0)
|   |   |   |   |   |   |   Materia = estructura de datos : no (1.0)
|   |   |   |   |   |   |   Materia = programacion aplicada a la web: no (1.0)
|   |   |   |   |   |   |   Materia = ingles I: no (0.0)
|   |   |   |   |   |   |   Materia = ingles II: no (0.0)
|   |   |   |   |   |   |   Materia = fisica I : no (0.0)
|   |   |   |   |   |   |   Materia = ingles III: no (0.0)
|   |   |   |   |   |   |   Materia = cultura fisica : no (0.0)
|   |   |   |   |   |   |   Materia = ingles I : no (0.0)
|   |   |   |   |   |   |   Materia = estructura de datos: no (0.0)
|   |   |   |   |   |   |   Sexo = Femenino: no (52.0/16.0)
|   |   |   |   |   |   Año = 2012: si (143.0/44.0)
|   |   |   |   |   |   Año = 2013
|   |   |   |   |   |   |   Sexo = Masculino
|   |   |   |   |   |   |   |   edad <= 18: no (58.0/13.0)
|   |   |   |   |   |   |   |   edad > 18
|   |   |   |   |   |   |   |   |   edad <= 19: si (60.0/23.0)
|   |   |   |   |   |   |   |   |   edad > 19: no (8.0)
|   |   |   |   |   |   |   Sexo = Femenino: no (15.0)
|   |   |   |   |   |   Año = 2014
|   |   |   |   |   |   |   Nivel = 1
|   |   |   |   |   |   |   |   materia perdida = no: si (4.0)
|   |   |   |   |   |   |   |   materia perdida = si
|   |   |   |   |   |   |   |   |   Promedio Materia <= 1.25: si (39.0/9.0)
|   |   |   |   |   |   |   |   |   Promedio Materia > 1.25
|   |   |   |   |   |   |   |   |   |   Sexo = Masculino: no (39.0/12.0)
|   |   |   |   |   |   |   |   |   |   Sexo = Femenino: si (4.0)
|   |   |   |   |   |   |   Nivel = 2: si (19.0/3.0)
|   |   |   |   |   |   |   Nivel = 3
|   |   |   |   |   |   |   |   edad <= 20: no (42.0)
|   |   |   |   |   |   |   |   edad > 20
|   |   |   |   |   |   |   |   |   edad <= 22: si (6.0)
|   |   |   |   |   |   |   |   |   edad > 22: no (2.0)
|   |   |   |   |   Año = 2015
|   |   |   |   |   |   Sexo = Masculino
|   |   |   |   |   |   |   Nivel = 1
|   |   |   |   |   |   |   |   Periodo = 1: no (82.0/18.0)
|   |   |   |   |   |   |   |   Periodo = 2: si (46.0/13.0)
|   |   |   |   |   |   |   Nivel = 2
|   |   |   |   |   |   |   |   edad <= 18: no (2.0)
|   |   |   |   |   |   |   |   edad > 18: si (4.0/1.0)
|   |   |   |   |   |   |   Nivel = 3
|   |   |   |   |   |   |   |   edad <= 22: no (40.0)
|   |   |   |   |   |   |   |   edad > 22
|   |   |   |   |   |   |   |   |   Promedio Materia <= 1.5: no (6.0/1.0)
|   |   |   |   |   |   |   |   |   Promedio Materia > 1.5: si (2.0)
|   |   |   |   |   |   Sexo = Femenino
|   |   |   |   |   |   |   edad <= 18: si (20.0/2.0)
|   |   |   |   |   |   |   edad > 18
|   |   |   |   |   |   |   |   edad <= 19: no (8.0)
|   |   |   |   |   |   |   |   edad > 19: si (7.0)
```

```
|   |   |   |   |   Año = 2016: no (75.0/13.0)
|   |   |   |   Estado Academico = Arrastra
|   |   |   |   |   Promedio Materia <= 0.5
|   |   |   |   |   |   Año = 2011: si (24.0/10.0)
|   |   |   |   |   |   Año = 2012: no (27.0/2.0)
|   |   |   |   |   |   Año = 2013: no (176.0/26.0)
|   |   |   |   |   |   Año = 2014: no (168.0/53.0)
|   |   |   |   |   |   Año = 2015
|   |   |   |   |   |   |   Periodo = 1: no (82.0/1.0)
|   |   |   |   |   |   |   Periodo = 2
|   |   |   |   |   |   |   |   Nivel = 1: no (0.0)
|   |   |   |   |   |   |   |   Nivel = 2
|   |   |   |   |   |   |   |   |   edad <= 18: no (21.0/2.0)
|   |   |   |   |   |   |   |   |   edad > 18
|   |   |   |   |   |   |   |   |   |   Sexo = Masculino: si (22.0/4.0)
|   |   |   |   |   |   |   |   |   |   Sexo = Femenino: no (4.0)
|   |   |   |   |   |   |   |   Nivel = 3: no (43.0/2.0)
|   |   |   |   |   |   Año = 2016: no (94.0/2.0)
|   |   |   |   |   Promedio Materia > 0.5
|   |   |   |   |   |   Año = 2011: no (14.0/6.0)
|   |   |   |   |   |   Año = 2012: no (20.0/7.0)
|   |   |   |   |   |   Año = 2013
|   |   |   |   |   |   |   Periodo = 1
|   |   |   |   |   |   |   |   Nivel = 1: no (0.0)
|   |   |   |   |   |   |   |   Nivel = 2
|   |   |   |   |   |   |   |   |   Promedio Materia <= 5.6: no (6.0)
|   |   |   |   |   |   |   |   |   Promedio Materia > 5.6: si (11.0/3.0)
|   |   |   |   |   |   |   |   Nivel = 3: no (6.0)
|   |   |   |   |   |   |   Periodo = 2: si (6.0/1.0)
|   |   |   |   |   |   Año = 2014: si (55.0/14.0)
|   |   |   |   |   |   Año = 2015
|   |   |   |   |   |   |   Periodo = 1: no (16.0)
|   |   |   |   |   |   |   Periodo = 2
|   |   |   |   |   |   |   |   Sexo = Masculino: si (33.0/13.0)
|   |   |   |   |   |   |   |   Sexo = Femenino: no (6.0)
|   |   |   |   |   |   Año = 2016: no (14.0/2.0)
|   |   |   |   Estado Academico = Repite: no (5.0)
|   |   |   Promedio Materia > 7.6
|   |   |   |   materia perdida = no: si (8.0/1.0)
|   |   |   |   materia perdida = si: no (1751.0/344.0)
|   |   Estado movilidad = otra malla
|   |   |   Estado Academico = Aprobado
|   |   |   |   edad <= 22: no (14.0)
|   |   |   |   edad > 22
|   |   |   |   |   Sexo = Masculino: si (5.0)
|   |   |   |   |   Sexo = Femenino: no (2.0)
|   |   |   Estado Academico = Arrastra: si (2.0)
|   |   |   Estado Academico = Repite: si (7.0)
|   |   Estado movilidad = sin recuperacion : si (5.0/1.0)
```

```
|   |   Estado movilidad = misma malla
|   |   |   Año = 2011: no (0.0)
|   |   |   Año = 2012: no (0.0)
|   |   |   Año = 2013: no (0.0)
|   |   |   Año = 2014: no (5.0)
|   |   |   Año = 2015
|   |   |   |   edad <= 20: no (4.0)
|   |   |   |   edad > 20: si (8.0/1.0)
|   |   |   Año = 2016: no (9.0)
|   |   Estado movilidad = UNIVERSIDAD ESTATAL DEL SUR DE MANABI: no (34.0)
|   |   Estado movilidad = medicina: no (2.0)
|   |   Estado movilidad = administracion de negocios: no (2.0)
|   |   Estado movilidad = ingenieria industrial: no (15.0)
|   |   Estado movilidad = economia: no (2.0)
|   |   Estado movilidad = UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL: no (26.0/2.0)
|   |   Estado movilidad = ingenieria electrica
|   |   |   Año = 2011: no (0.0)
|   |   |   Año = 2012: no (0.0)
|   |   |   Año = 2013: no (10.0)
|   |   |   Año = 2014: no (11.0)
|   |   |   Año = 2015: si (6.0)
|   |   |   Año = 2016: no (0.0)
|   |   Estado movilidad = UNIVERSIDAD POLITECNICA DE MADRID: no (0.0)
|   |   Estado movilidad = ingenieria en mecanica naval: no (1.0)
|   |   Estado movilidad = ingenieria civil: si (14.0)
|   |   Estado movilidad = UNIVERSIDAD FUERZAS ARMADAS(ESPE): no (0.0)
|   |   Estado movilidad = ingenieria en marketing: si (10.0)
|   |   Estado movilidad = UNIVERSIDAD TECNICA DE MANABI: no (3.0)
|   |   Estado movilidad = trabajo social: no (0.0)
|   |   Estado movilidad = hoteleria y turismo: no (3.0)
|   |   Estado movilidad = secretariado ejecutivo: no (0.0)
|   |   Estado movilidad = derecho: no (0.0)
|   semestre perdido = si
|   |   Periodo = 1
|   |   |   Estado Academico = Aprobado
|   |   |   |   Año = 2011: no (2.0)
|   |   |   |   Año = 2012: si (13.0)
|   |   |   |   Año = 2013: no (7.0)
|   |   |   |   Año = 2014: si (4.0)
|   |   |   |   Año = 2015: si (0.0)
|   |   |   |   Año = 2016: si (0.0)
|   |   |   Estado Academico = Arrastra: si (8.0)
|   |   |   Estado Academico = Repite
|   |   |   |   Año = 2011
|   |   |   |   |   tramite movilidad = ninguno: si (6.0)
|   |   |   |   |   tramite movilidad = reingreso
|   |   |   |   |   |   edad <= 18
|   |   |   |   |   |   |   Promedio Materia <= 4.13: si (19.0/8.0)
|   |   |   |   |   |   |   Promedio Materia > 4.13: no (24.0/9.0)
|   |   |   |   |   |   edad > 18: no (32.0/1.0)
|   |   |   |   |   tramite movilidad = tercera matricula: no (0.0)
|   |   |   |   |   tramite movilidad = externa: no (0.0)
|   |   |   |   |   tramite movilidad = interna: no (0.0)
```

```
|   |   |   |   Año = 2012
|   |   |   |   |   tramite movilidad = ninguno: si (25.0/3.0)
|   |   |   |   |   tramite movilidad = reingreso: no (5.0)
|   |   |   |   |   tramite movilidad = tercera matricula: si (1.0)
|   |   |   |   |   tramite movilidad = externa: si (0.0)
|   |   |   |   |   tramite movilidad = interna: si (0.0)
|   |   |   |   Año = 2013
|   |   |   |   |   Sexo = Masculino
|   |   |   |   |   |   Promedio Materia <= 12.05: si (46.0/7.0)
|   |   |   |   |   |   Promedio Materia > 12.05: no (3.0/1.0)
|   |   |   |   |   Sexo = Femenino: no (7.0)
|   |   |   |   Año = 2014
|   |   |   |   |   tramite movilidad = ninguno
|   |   |   |   |   |   Promedio Materia <= 11.1
|   |   |   |   |   |   |   edad <= 19
|   |   |   |   |   |   |   |   Sexo = Masculino: no (47.0/13.0)
|   |   |   |   |   |   |   |   Sexo = Femenino: si (3.0)
|   |   |   |   |   |   |   edad > 19
|   |   |   |   |   |   |   |   Sexo = Masculino: si (27.0/5.0)
|   |   |   |   |   |   |   |   Sexo = Femenino: no (14.0/5.0)
|   |   |   |   |   |   Promedio Materia > 11.1: si (6.0/1.0)
|   |   |   |   |   tramite movilidad = reingreso: si (7.0)
|   |   |   |   |   tramite movilidad = tercera matricula: si (0.0)
|   |   |   |   |   tramite movilidad = externa: si (0.0)
|   |   |   |   |   tramite movilidad = interna: si (0.0)
|   |   |   |   Año = 2015
|   |   |   |   |   edad <= 18: si (7.0)
|   |   |   |   |   edad > 18
|   |   |   |   |   |   tramite movilidad = ninguno
|   |   |   |   |   |   |   edad <= 20: no (47.0/7.0)
|   |   |   |   |   |   |   edad > 20: si (8.0/1.0)
|   |   |   |   |   |   tramite movilidad = reingreso: no (17.0)
|   |   |   |   |   |   tramite movilidad = tercera matricula: si (3.0/1.0)
|   |   |   |   |   |   tramite movilidad = externa: no (0.0)
|   |   |   |   |   |   tramite movilidad = interna: no (0.0)
|   |   |   |   Año = 2016
|   |   |   |   |   Nivel = 1
|   |   |   |   |   |   Sexo = Masculino: no (58.0/8.0)
|   |   |   |   |   |   Sexo = Femenino
|   |   |   |   |   |   |   edad <= 18: no (2.0)
|   |   |   |   |   |   |   edad > 18: si (5.0)
|   |   |   |   |   Nivel = 2: no (36.0)
|   |   |   |   |   Nivel = 3: no (31.0)
|   |   Periodo = 2
|   |   |   Promedio Materia <= 7.62
|   |   |   |   Año = 2011: si (129.0/10.0)
|   |   |   |   Año = 2012: si (23.0/5.0)
|   |   |   |   Año = 2013: si (128.0/19.0)
|   |   |   |   Año = 2014
|   |   |   |   |   Nivel = 1: si (101.0/7.0)
|   |   |   |   |   Nivel = 2: si (31.0/1.0)
|   |   |   |   |   Nivel = 3: no (5.0)
```

```
|   |   |   |   Año = 2015
|   |   |   |   |   edad <= 19
|   |   |   |   |   |   Nivel = 1
|   |   |   |   |   |   |   edad <= 18: no (32.0/13.0)
|   |   |   |   |   |   |   edad > 18: si (39.0/4.0)
|   |   |   |   |   |   Nivel = 2: si (0.0)
|   |   |   |   |   |   Nivel = 3: no (11.0)
|   |   |   |   |   edad > 19: si (27.0)
|   |   |   |   Año = 2016: si (0.0)
|   |   |   Promedio Materia > 7.62
|   |   |   |   Materia = calculo diferencial
|   |   |   |   |   Año = 2011
|   |   |   |   |   |   Promedio Materia <= 9.3: si (3.0)
|   |   |   |   |   |   Promedio Materia > 9.3
|   |   |   |   |   |   |   Promedio Materia <= 10.31: no (2.0)
|   |   |   |   |   |   |   Promedio Materia > 10.31: si (3.0/1.0)
|   |   |   |   |   Año = 2012: no (0.0)
|   |   |   |   |   Año = 2013: no (0.0)
|   |   |   |   |   Año = 2014: no (4.0)
|   |   |   |   |   Año = 2015
|   |   |   |   |   |   edad <= 18: no (6.0)
|   |   |   |   |   |   edad > 18: si (8.0/1.0)
|   |   |   |   |   Año = 2016: no (0.0)
|   |   |   |   Materia = cultura fisica: no (1.0)
|   |   |   |   Materia = fisica I
|   |   |   |   |   Sexo = Masculino
|   |   |   |   |   |   Promedio Materia <= 9.3: no (3.0)
|   |   |   |   |   |   Promedio Materia > 9.3: si (8.0/2.0)
|   |   |   |   |   Sexo = Femenino: si (2.0)
|   |   |   |   Materia = introduccion a la informatica: si (12.0/2.0)
|   |   |   |   Materia = metodologia de la investigacion: si (2.0)
|   |   |   |   Materia = algebra lineal
|   |   |   |   |   Año = 2011: si (4.0)
|   |   |   |   |   Año = 2012: si (3.0)
|   |   |   |   |   Año = 2013: si (1.0)
|   |   |   |   |   Año = 2014: no (3.0)
|   |   |   |   |   Año = 2015
|   |   |   |   |   |   edad <= 18: no (2.0)
|   |   |   |   |   |   edad > 18: si (5.0)
|   |   |   |   |   Año = 2016: si (0.0)
|   |   |   |   Materia = fundamentos de programacion
|   |   |   |   |   Sexo = Masculino: si (12.0/1.0)
|   |   |   |   |   Sexo = Femenino
|   |   |   |   |   |   Promedio Materia <= 11.15: si (3.0)
|   |   |   |   |   |   Promedio Materia > 11.15: no (2.0)
|   |   |   |   Materia = calculo integral
|   |   |   |   |   edad <= 19: no (4.0)
|   |   |   |   |   edad > 19: si (3.0/1.0)
|   |   |   |   Materia = matematicas discretas
|   |   |   |   |   edad <= 18: no (9.0)
|   |   |   |   |   edad > 18
```

```
|   |   |   |   |   |   Sexo = Masculino
|   |   |   |   |   |   |   Promedio Materia <= 10.25: si (4.0/1.0)
|   |   |   |   |   |   |   Promedio Materia > 10.25: no (5.0/1.0)
|   |   |   |   |   |   Sexo = Femenino: si (2.0)
|   |   |   |   Materia = fisica II: si (9.0/4.0)
|   |   |   |   Materia = sistemas operativos
|   |   |   |   |   Promedio Materia <= 10.79: no (2.0)
|   |   |   |   |   Promedio Materia > 10.79: si (3.0)
|   |   |   |   Materia = teoria de sistemas: si (9.0/2.0)
|   |   |   |   Materia = tecnicas de expresion oral y escrita: si (1.0)
|   |   |   |   Materia = programacion orientada a objetos
|   |   |   |   |   edad <= 20: no (6.0)
|   |   |   |   |   edad > 20: si (2.0)
|   |   |   |   Materia = calculo vectorial: no (2.0/1.0)
|   |   |   |   Materia = electronica: no (8.0/1.0)
|   |   |   |   Materia = aplicación de sistemas operativos: no (4.0/2.0)
|   |   |   |   Materia = analisis de sistemas: si (4.0)
|   |   |   |   Materia = sem. valor y etica profesional: si (1.0)
|   |   |   |   Materia = estructura de datos : no (7.0)
|   |   |   |   Materia = programacion aplicada a la web: no (4.0)
|   |   |   |   Materia = ingles I: si (1.0)
|   |   |   |   Materia = ingles II: si (0.0)
|   |   |   |   Materia = fisica I : si (0.0)
|   |   |   |   Materia = ingles III: si (0.0)
|   |   |   |   Materia = cultura fisica : si (0.0)
|   |   |   |   Materia = ingles I : si (0.0)
|   |   |   |   Materia = estructura de datos: no (1.0)
Promedio Materia > 13.42
|   semestre perdido = no: no (10571.0/547.0)
|   semestre perdido = si
|   |   Año = 2011
|   |   |   edad <= 19
|   |   |   |   Nivel = 1
|   |   |   |   |   Periodo = 1: no (52.0/8.0)
|   |   |   |   |   Periodo = 2
|   |   |   |   |   |   edad <= 18: no (54.0/9.0)
|   |   |   |   |   |   edad > 18: si (31.0/9.0)
|   |   |   |   Nivel = 2: no (66.0/2.0)
|   |   |   |   Nivel = 3: no (0.0)
|   |   |   edad > 19: no (46.0)
|   |   Año = 2012
```

```
|   |   |   tramite movilidad = ninguno
|   |   |   |   Periodo = 1
|   |   |   |   |   Sexo = Masculino: si (10.0)
|   |   |   |   |   Sexo = Femenino
|   |   |   |   |   |   edad <= 19: si (4.0)
|   |   |   |   |   |   edad > 19
|   |   |   |   |   |   |   Promedio Materia <= 15.4: si (2.0)
|   |   |   |   |   |   |   Promedio Materia > 15.4: no (8.0/1.0)
|   |   |   |   Periodo = 2
|   |   |   |   |   edad <= 18
|   |   |   |   |   |   Sexo = Masculino: no (4.0/1.0)
|   |   |   |   |   |   Sexo = Femenino: si (2.0)
|   |   |   tramite movilidad = reingreso: no (15.0)
|   |   |   tramite movilidad = tercera matricula: no (0.0)
|   |   |   tramite movilidad = externa: no (0.0)
|   |   |   tramite movilidad = interna: no (0.0)
|   |   Año = 2013
|   |   |   Estado Academico = Aprobado: no (0.0)
|   |   |   Estado Academico = Arrastra: si (2.0)
|   |   |   Estado Academico = Repite: no (186.0/24.0)
|   |   Año = 2014
|   |   |   Estado Academico = Aprobado: si (3.0)
|   |   |   Estado Academico = Arrastra: no (0.0)
|   |   |   Estado Academico = Repite
|   |   |   |   Periodo = 1
|   |   |   |   |   Nivel = 1
|   |   |   |   |   |   Promedio Materia <= 17.5
|   |   |   |   |   |   |   Sexo = Masculino: no (29.0/8.0)
|   |   |   |   |   |   |   Sexo = Femenino
|   |   |   |   |   |   |   |   edad <= 19: si (4.0)
|   |   |   |   |   |   |   |   edad > 19: no (2.0)
|   |   |   |   |   |   Promedio Materia > 17.5: si (4.0)
|   |   |   |   |   Nivel = 2: si (16.0/4.0)
|   |   |   |   |   Nivel = 3: no (24.0/6.0)
|   |   |   |   Periodo = 2: no (112.0/16.0)
|   |   Año = 2015
|   |   |   Periodo = 1: no (168.0/9.0)
|   |   |   Periodo = 2
|   |   |   |   Nivel = 1
|   |   |   |   |   edad <= 18: no (44.0/6.0)
|   |   |   |   |   edad > 18
|   |   |   |   |   |   Promedio Materia <= 14.46: si (10.0/3.0)
|   |   |   |   |   |   Promedio Materia > 14.46: no (22.0/9.0)
|   |   |   |   Nivel = 2: si (8.0)
|   |   |   |   Nivel = 3: no (33.0/1.0)
|   |   Año = 2016: no (214.0/8.0)

Number of Leaves  :     262

Size of the tree :      378
```

Ilustración 49: Vista general del listado de reglas de predicción del algoritmo J48 del árbol de decisión.

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       15269               90.671  %
Incorrectly Classified Instances      1571                9.329  %
Kappa statistic                          0.5692
Mean absolute error                      0.1537
Root mean squared error                  0.2772
Relative absolute error                 59.8236 %
Root relative squared error             77.3501 %
Total Number of Instances            16840

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.979     0.499      0.917     0.979     0.947      0.842    no
                 0.501     0.021      0.811     0.501     0.619      0.842    si
Weighted Avg.    0.907     0.427      0.901     0.907     0.897      0.842
```

Ilustración 50: En general el algoritmo J48 generó 15269 instancias correctas a partir de los 16840 registros con un error medio cuadrático de 0,277. Apenas 1571 registros se clasificaron incorrectamente.

```
=== Confusion Matrix ===

     a     b   <-- classified as
 13993   298 |     a = no
  1273  1276 |     b = si
```

Ilustración 51: La matriz de confusión muestra que los datos se están clasificando de una manera bastante aceptable, por ejemplo, en a (fila) se registraron 13993 no desertores de un total de 14291 estudiantes, de los cuales el modelo ha clasificado correctamente como a (a=no) a 13993 no desertores e incorrectamente clasificó 298 casos.

Respecto a la exploración del modelo se determina que éste funciona bastante bien, con un tamaño del árbol de 390 y 275 niveles, pues el error absoluto del modelo es apenas del 0,1537 después de que se revisó cada una de las 16840 instancias o registros.

Respecto a la implementación del modelo, corresponde al personal de secretaria, actualizarlo dinámicamente, cuando entren más datos en la organización, y realizar modificaciones constantes para mejorar la efectividad de la solución debería ser parte de la estrategia de implementación.

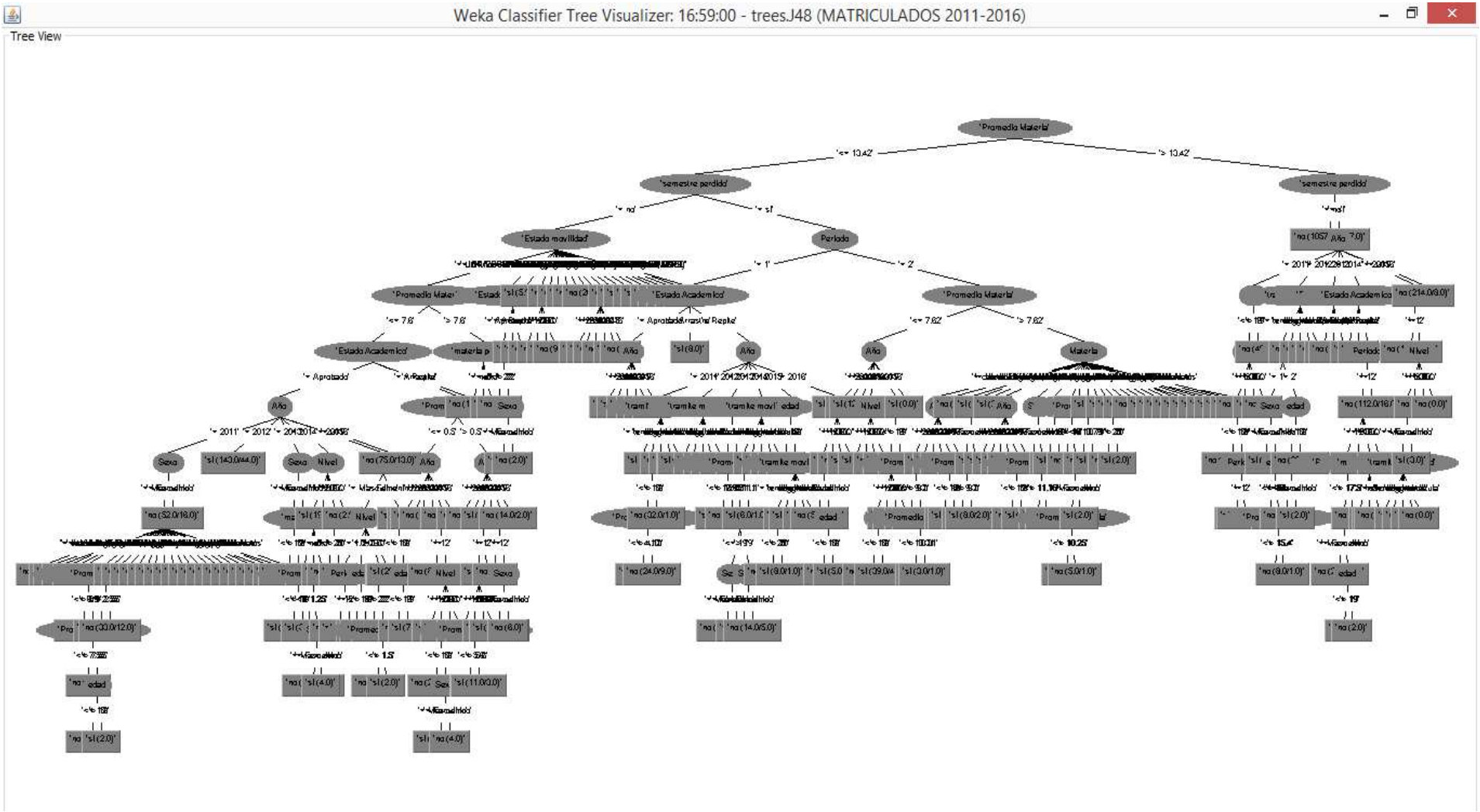


Ilustración 52: Visualización del modelo de predicción del algoritmo J48 de árbol de decisión

Validaciones del árbol de decisión

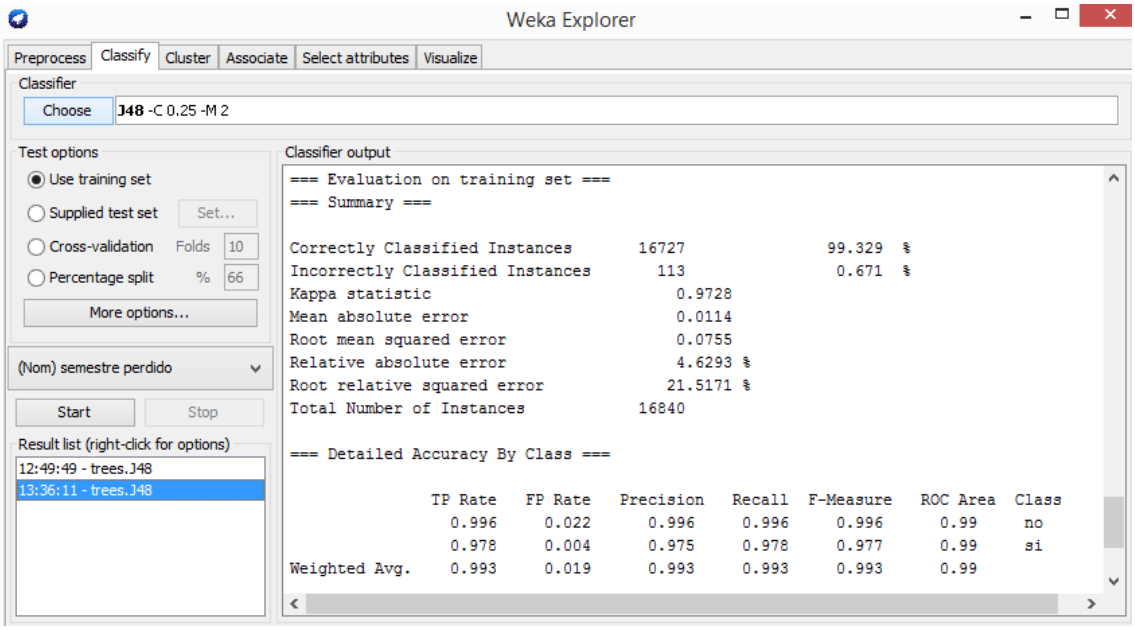
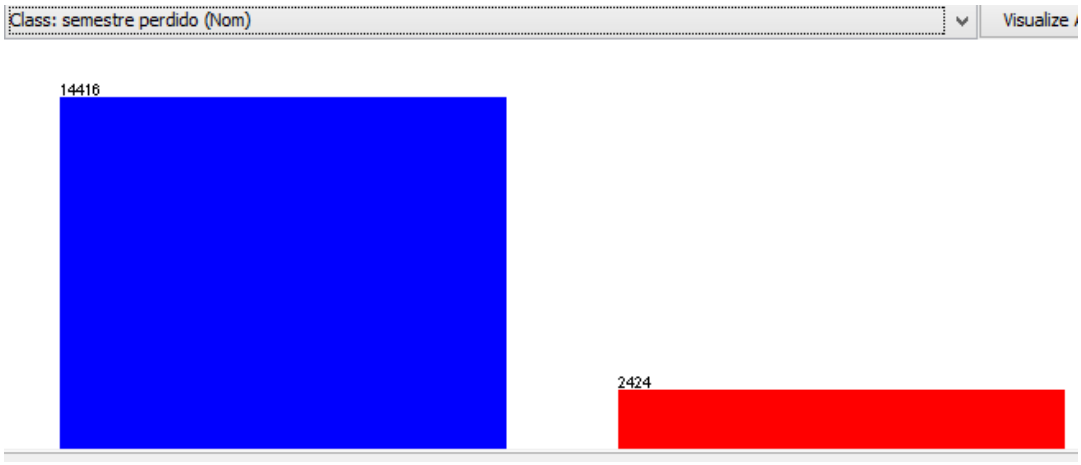


Ilustración 53: En general el algoritmo J48 generó 16727 instancias correctas a partir de los 16840 registros con un error medio cuadrático de 0,075. Apenas 113 registros se clasificaron incorrectamente.




```
=== Confusion Matrix ===

     a     b   <-- classified as
 14356    60 |     a = no
    53  2371 |     b = si
```

Ilustración 54: La matriz de confusión reporta datos alentadores, por ejemplo, para la fila "a" que tiene un total de 14416 estudiantes que no han perdido el semestre, registra que 14356 registros se clasificaron correctamente versus 60 que se clasificaron incorrectamente.

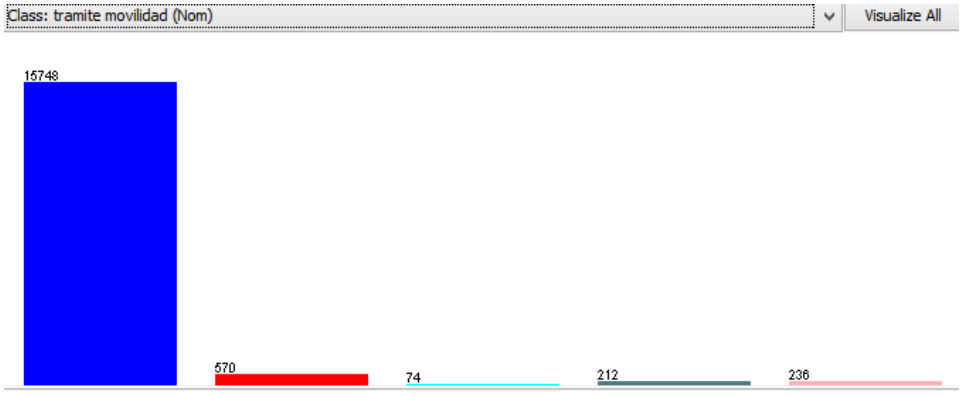
```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       16840               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances            16840

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        ninguno
                 1         0          1         1         1          1        reingreso
                 1         0          1         1         1          1        tercera matricula
                 1         0          1         1         1          1        externa
                 1         0          1         1         1          1        interna
Weighted Avg.    1         0          1         1         1          1
```

Ilustración 55: El algoritmo J48 generó 16840 instancias correctas a partir de los 16840 registros con un error medio cuadrático de 0. Ningún registro se clasifico incorrectamente.




```
=== Confusion Matrix ===

    a     b     c     d     e    <-- classified as
15748     0     0     0     0 |     a = ninguno
    0   570     0     0     0 |     b = reingreso
    0     0    74     0     0 |     c = tercera matricula
    0     0     0   212     0 |     d = externa
    0     0     0     0   236 |     e = interna
```

Ilustración 56: La matriz de confusión muestra que los datos están clasificados de una maneara aceptable, por ejemplo, en tercera matricula se registraron 74 trámites de los cuales el modelo ha clasificado correctamente todos, es decir no hay errores.

3.4.6.2. Modelo de Clasificación Naive Bayes para identificar estudiantes desertores.

Para este ejercicio inicialmente se tiene:

- El número de instancias o registros es 16840.
- Se tienen trece atributos: sexo, edad, estado académico, nivel, periodo, año, materia, promedio materia, materia perdida, semestre perdido, trámite de movilidad, estado de movilidad y desertor.
- Se usa validación cruzada (cross-validation).

Naive Bayes Calcula la probabilidad de cada estado de cada columna de entrada, dado cada posible estado de la columna de predicción.

```
Naive Bayes Classifier

                                                    Class
Attribute                                             no      si
                                                   (0.85)  (0.15)
=================================================================
Sexo
  Masculino                                       11137.0  2129.0
  Femenino                                         3156.0   422.0
  [total]                                         14293.0  2551.0

edad
  mean                                            18.7831 18.9097
  std. dev.                                        1.2441  1.0211
  weight sum                                        14291    2549
  precision                                        1.7692  1.7692

Estado Academico
  Aprobado                                         9533.0  1161.0
  Arrastra                                         3205.0   477.0
  Repite                                           1556.0   914.0
  [total]                                         14294.0  2552.0

Nivel
  1                                                6285.0  1474.0
  2                                                4270.0   675.0
  3                                                3739.0   403.0
  [total]                                         14294.0  2552.0

Periodo
  1                                                8492.0  1215.0
  2                                                5801.0  1336.0
  [total]                                         14293.0  2551.0
```

Año		
2011	1914.0	508.0
2012	2100.0	352.0
2013	2221.0	501.0
2014	2776.0	625.0
2015	3229.0	509.0
2016	2057.0	60.0
[total]	14297.0	2555.0
Materia		
calculo diferencial	998.0	234.0
cultura fisica	894.0	187.0
fisica I	954.0	223.0
introduccion a la informatica	891.0	192.0
metodologia de la investigacion	868.0	188.0
algebra lineal	979.0	221.0
fundamentos de programacion	988.0	226.0
calculo integral	615.0	89.0
matematicas discretas	610.0	106.0
fisica II	659.0	97.0
sistemas operativos	590.0	108.0
teoria de sistemas	613.0	111.0
tecnicas de expresion oral y escrita	587.0	107.0
programacion orientada a objetos	617.0	97.0
calculo vectorial	428.0	46.0
electronica	410.0	48.0
aplicación de sistemas operativos	441.0	47.0
analisis de sistemas	454.0	47.0
sem. valor y etica profesional	451.0	50.0
estructura de datos	402.0	45.0
programacion aplicada a la web	448.0	46.0
ingles I	242.0	40.0
ingles II	102.0	11.0
fisica I	4.0	2.0
ingles III	53.0	4.0
cultura fisica	3.0	1.0
ingles I	2.0	1.0
estructura de datos	16.0	3.0
[total]	14319.0	2577.0
Promedio Materia		
mean	13.671	7.7325
std. dev.	4.8793	6.3288
weight sum	14291	2549
precision	0.0156	0.0156
materia perdida		
no	11031.0	755.0
si	3262.0	1796.0
[total]	14293.0	2551.0

tramite movilidad		
ninguno	13434.0	2316.0
reingreso	432.0	140.0
tercera matricula	31.0	45.0
externa	204.0	10.0
interna	195.0	43.0
[total]	14296.0	2554.0
Estado movilidad		
normal	13434.0	2316.0
otra malla	258.0	73.0
sin recuperacion	31.0	45.0
misma malla	175.0	68.0
UNIVERSIDAD ESTATAL DEL SUR DE MANABI	82.0	1.0
medicina	14.0	1.0
administracion de negocios	16.0	1.0
ingenieria industrial	60.0	1.0
economia	6.0	1.0
UNIVERSIDAD CRISTIANA LATINOAMERICANA - UCL	80.0	3.0
ingenieria electrica	57.0	15.0
UNIVERSIDAD POLITECNICA DE MADRID	5.0	1.0
ingenieria en mecanica naval	14.0	1.0
ingenieria civil	5.0	15.0
UNIVERSIDAD FUERZAS ARMADAS(ESPE)	15.0	1.0
ingenieria en marketing	1.0	15.0
UNIVERSIDAD TECNICA DE MANABI	26.0	8.0
trabajo social	8.0	1.0
hoteleria y turismo	8.0	1.0
secretariado ejecutivo	9.0	1.0
derecho	8.0	1.0
[total]	14312.0	2570.0

Ilustración 57: vista general del modelo de clasificación Naive Bayes en WEKA.

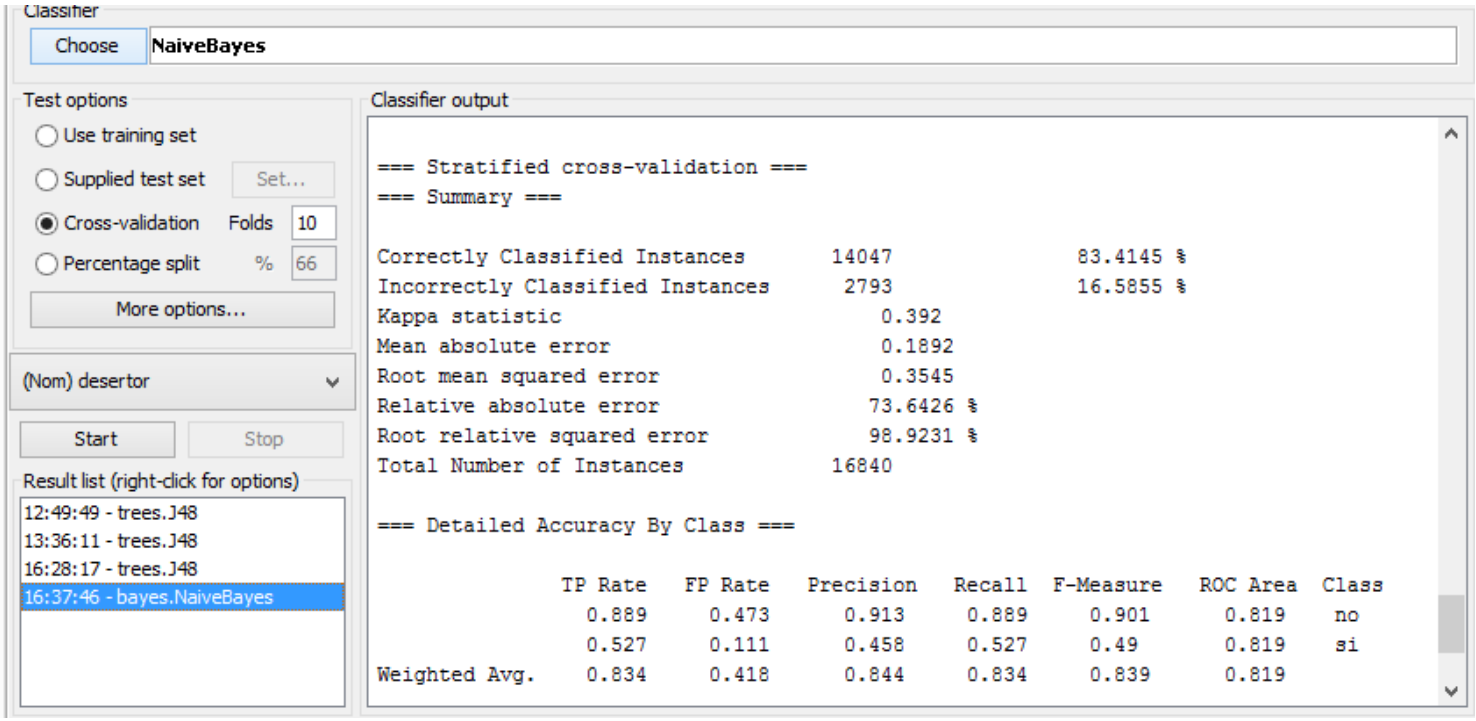


Ilustración 58: El algoritmo de clasificación Naive Bayes generó 14047 instancias correctas a partir de los 16840 registros con un error medio cuadrático de 0.3545. Apenas 2793 registros se clasificaron incorrectamente.

```
=== Confusion Matrix ===

     a     b   <-- classified as
 12703  1588 |     a = no
  1205  1344 |     b = si
```

Ilustración 59: La matriz de confusión muestra que los datos se están clasificando de una manera bastante aceptable, por ejemplo, en a (fila) se registraron 14291 estudiantes no desertores, de los cuales el modelo ha clasificado correctamente como a (a=no) a 12703 no desertores e incorrectamente clasificó 1588 casos.

Validación de Naive Bayes

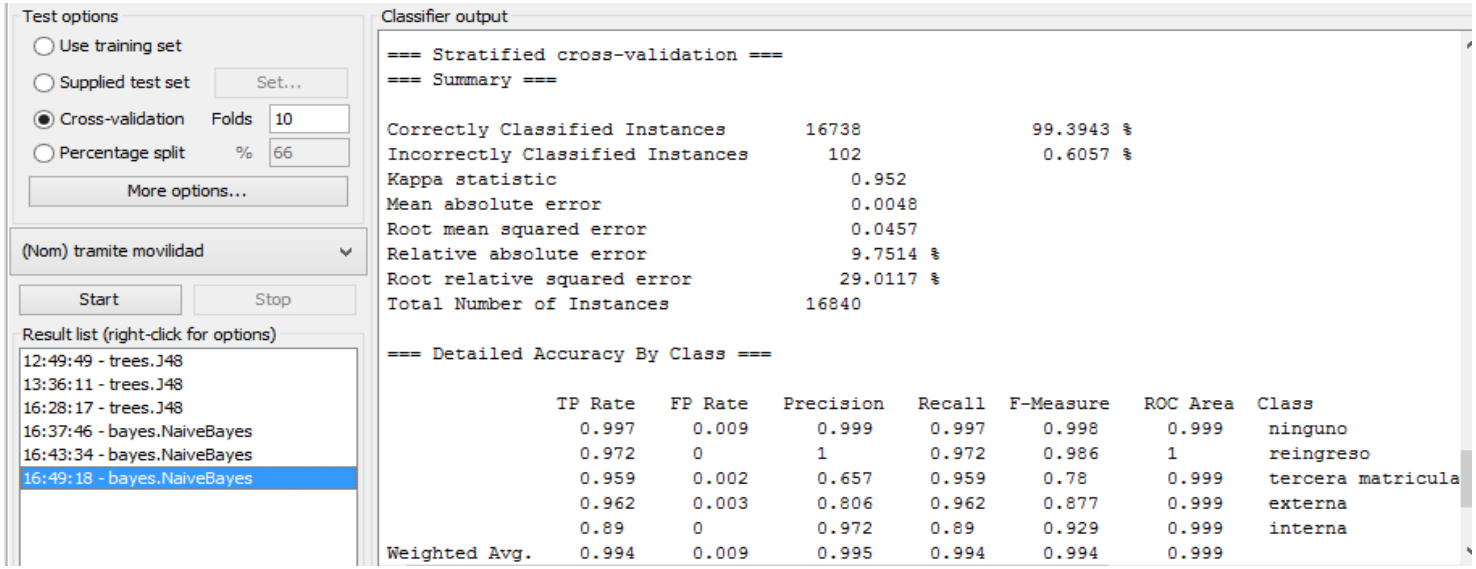


Ilustración 60: El algoritmo de clasificación Naive Bayes generó 16738 instancias correctas acerca del atributo trámite de movilidad a partir de los 16840 registros con un error medio cuadrático de 0.0457. Apenas 102 registros se clasificaron incorrectamente.

```
=== Confusion Matrix ===

    a     b     c     d     e   <-- classified as
15699     0    30    14     5 |     a = ninguno
    0   554     0    16     0 |     b = reingreso
    3     0    71     0     0 |     c = tercera matricula
    7     0     0   204     1 |     d = externa
    0     0     7    19   210 |     e = interna
```

Ilustración 61: Pese a que hay errores en la clasificación dado del volumen de datos que se analiza, la matriz de confusión reporta datos alentadores, por ejemplo, para el trámite de movilidad reingreso (fila f=2), 554 registros se clasificaron correctamente versus 16 que se clasificaron incorrectamente en otros trámites.

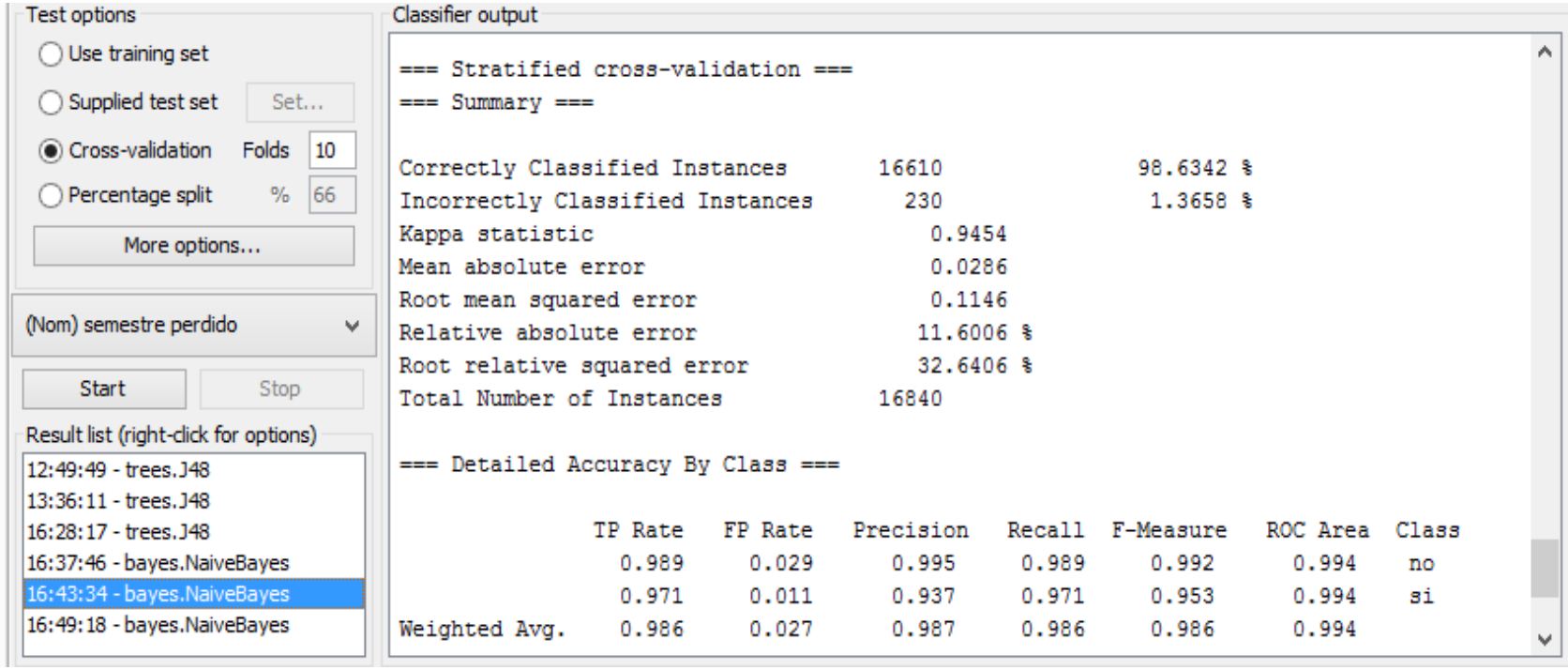


Ilustración 62: El algoritmo de clasificación Naive Bayes generó 16610 instancias correctas acerca del atributo semestre perdido a partir de los 16840 registros con un error medio cuadrático de 0.1146. Apenas 230 registros se clasificaron incorrectamente.

```
=== Confusion Matrix ===

     a     b   <-- classified as
 14257   159 |     a = no
    71  2353 |     b = si
```

Ilustración 63: La matriz de confusión reporta datos alentadores, por ejemplo, para el atributo semestre perdido en la que existen 14416 estudiantes que no han perdido el semestre, 14257 clasificaron correctamente versus 159 que se clasificaron incorrectamente.

3.4.6.3. Modelo de Reglas JRIP para la predicación de desertores

JRip (RIPPER) es uno de los algoritmos básicos y más populares. Las clases son examinadas en aumentar el tamaño y se genera un conjunto inicial de reglas para la clase usando incrementalmente reducido error JRip (RIPPER) procede al tratar todos los ejemplos de un juicio particular en los datos de entrenamiento como clase, y encontrar un conjunto de reglas que cubran a todos los miembros de esa clase.

A continuación, pasa a la siguiente clase y hace lo mismo, repitiendo esto hasta que todas las clases hayan sido cubiertas.

Para este ejercicio inicialmente se tiene:

- El número de instancias o registros es 16840.
- Se tienen nueve atributos: sexo, edad, estado académico, nivel, año, materia perdida, semestre perdido, trámite de movilidad y desertor.
- Se usa un conjunto de datos de entrenamiento (Use training set).

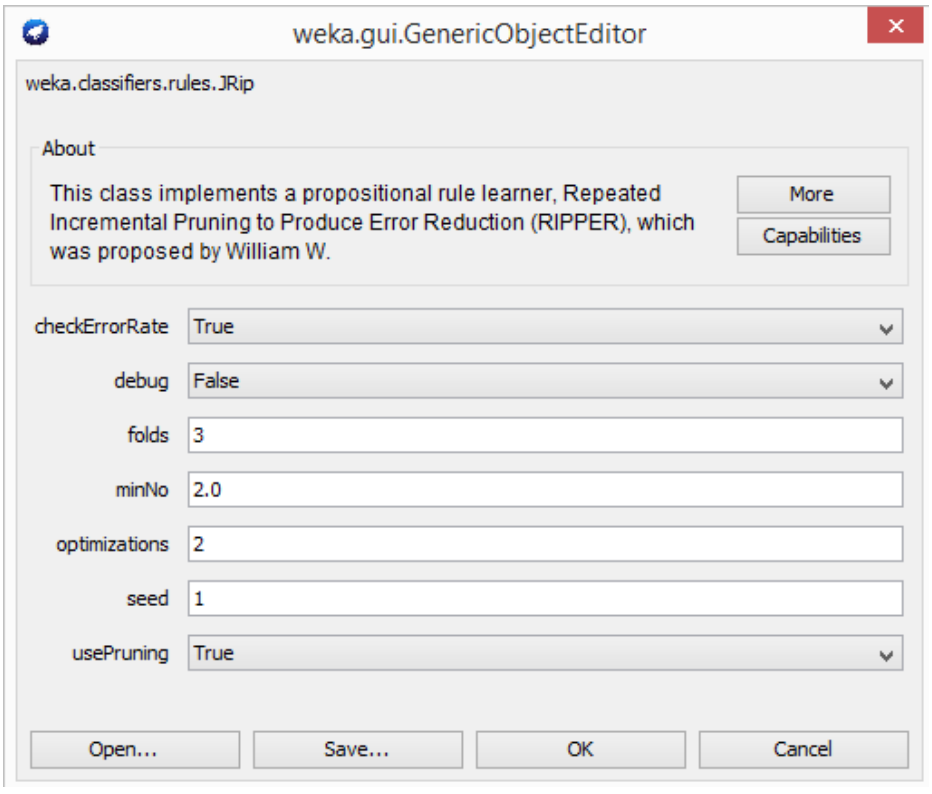


Ilustración 64: Configuración del algoritmo JRip

```
=== Run information ===

Scheme:weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation:     MATRICULADOS 2011-2016
Instances:    16840
Attributes:   9
              Sexo
              edad
              Estado Academico
              Nivel
              Año
              materia perdida
              semestre perdido
              tramite movilidad
              desertor
Test mode:evaluate on training data

=== Classifier model (full training set) ===
JRIP rules:
===========

(materia perdida = si) and (semestre perdido = si) and (Año = 2013) => desertor=si (236.0/61.0)
(materia perdida = si) and (semestre perdido = si) and (Nivel = 1.0) and (tramite movilidad = ninguno) and (Año = 2011) => desertor=si (150.0/20.0)
(materia perdida = si) and (semestre perdido = si) and (Año = 2014) => desertor=si (266.0/78.0)
(materia perdida = si) and (semestre perdido = si) and (edad = 18) => desertor=si (138.0/66.0)
(materia perdida = si) and (semestre perdido = si) and (Año = 2015) => desertor=si (190.0/79.0)
(materia perdida = si) and (Año = 2012) and (semestre perdido = si) => desertor=si (67.0/26.0)
(materia perdida = si) and (Año = 2011) and (Sexo = Masculino) and (edad = 19) and (semestre perdido = no) and (Estado Academico = Repite) => desertor=si (7.0/0.0)
(materia perdida = si) and (Nivel = 1.0) and (edad = 19) and (Año = 2012) and (tramite movilidad = ninguno) and (Sexo = Masculino) => desertor=si (86.0/27.0)
(materia perdida = si) and (Año = 2011) and (Sexo = Masculino) and (Estado Academico = Arrastra) => desertor=si (61.0/29.0)
(materia perdida = si) and (Año = 2014) and (Nivel = 2.0) and (edad = 20) => desertor=si (29.0/11.0)
(materia perdida = si) and (Estado Academico = Aprobado) and (Año = 2013) and (edad = 19) and (Nivel = 1.0) and (Sexo = Masculino) => desertor=si (37.0/12.0)
 => desertor=no (15573.0/1691.0)

Number of Rules : 12
```

Ilustración 65: Vista general de las doce reglas generadas por el algoritmo de JRip

Interpretaciones de las Reglas

En esta sección tratamos de interpretar la regla de JRip.

Regla 1 interpretada como: Si los estudiantes han perdido la materia y el semestre y se encuentran cursando en el año 2013 son posibles desertores.

Regla 2 interpretada como: Si los estudiantes han perdido la materia y el semestre, no han realizado trámite de movilidad y se encuentran cursando el primer nivel en el año 2011 son posibles desertores.

Regla 3 interpretada como: Si los estudiantes han perdido la materia y el semestre y se encuentran cursando en el año 2014 son posibles desertores.

Regla 4 interpretada como: Si los estudiantes tienen 18 y han perdido la materia y el semestre son posibles desertores.

Regla 5 interpretada como: Si los estudiantes han perdido la materia y el semestre y se encuentran cursando en el año 2015 son posibles desertores.

Regla 6 interpretada como: Si los estudiantes han perdido la materia y el semestre y se encuentran cursando en el año 2012 son posibles desertores.

Regla 7 interpretada como: Si los estudiantes de sexo masculino de 19 años han perdido la materia aunque su estado de académico sea repite, pero no el semestre y se encuentra cursando en el año 2011 son posibles desertores.

Regla 8 interpretada como: Si los estudiantes de sexo masculino de 19 años han perdido la materia, y hayan realizado ningún trámite de movilidad sea repite, pero no el semestre y se encuentra cursando en el año 2011 nivel 1 son posibles desertores.

Regla 12: Si los estudiantes no han perdido sus materia y el semestre no son desertores.

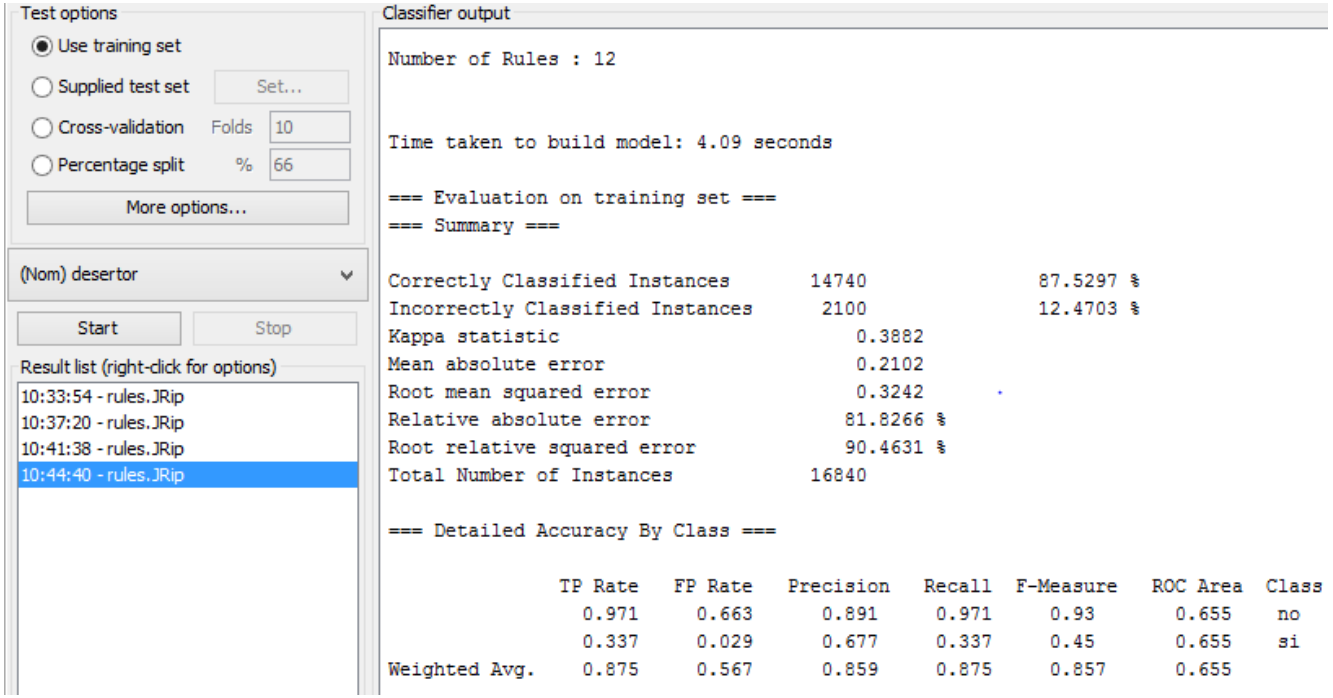


Ilustración 66: El algoritmo JRip generó 14740 instancias a partir de los 16840 registros con un error medio cuadrático de 0.3242. Apenas 2100 registros se clasificaron incorrectamente.

```
=== Confusion Matrix ===

     a     b   <-- classified as
 13882   409 |     a = no
  1691   858 |     b = si
```

Ilustración 67: La matriz de confusión muestra que los datos se están clasificando de una manera bastante aceptable, por ejemplo, en a (fila) se registraron 14291 estudiantes de los cuales el modelo ha clasificado correctamente como a (a=no) a 13882 no desertores e incorrectamente clasificó 409 casos.

3.4.6.4. Modelo Clúster utilizando el algoritmo Simple K-Means para la predicción de estudiantes desertores.

Para este ejercicio inicialmente se tiene:

- El número de instancias o registros es 16840.
- Se tienen diez atributos: sexo, nivel, año, materia, promedio materia, materia perdida, trámite de movilidad, estado de movilidad y desertor.
- Se usa un conjunto de datos de entrenamiento (Use training set)

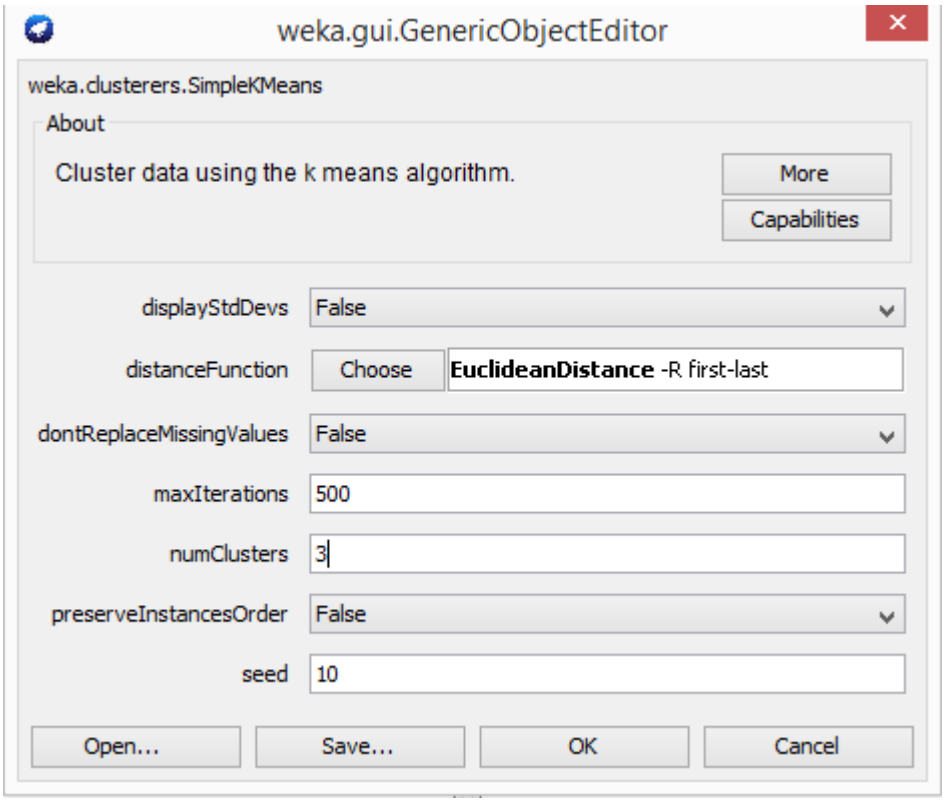


Ilustración 68: Configuración de Simple K-Means

Clusterer output

```
Within cluster sum of squared errors: 46902.31851947573 6

Missing values globally replaced with mean/mode

Cluster centroids:
                                                    Cluster#
Attribute                                Full Data          0          1          2
                                           (16840)    (11515)     (3597)     (1728)
=============================================================================================
Sexo                                     Masculino  Masculino  Masculino  Masculino
Nivel                                          1.0        1.0        1.0        1.0
Año                                           2015       2015       2014       2011
Materia                       calculo diferencial  cultura fisica  fisica I calculo diferencial
Promedio Materia                           12.7722    15.7733     6.9698     4.8519
materia perdida                                 no         no         si         si
semestre perdido                                no         no         no         si
tramite movilidad                           ninguno    ninguno    ninguno    ninguno
Estado movilidad                             normal     normal     normal     normal
desertor                                        no         no         no         si



Time taken to build model (full training data) : 0.83 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      11515 ( 68%)
1       3597 ( 21%)
2       1728 ( 10%)
```

Ilustración 69: El algoritmo k-Means muestra que el 68% de los casos más cercanos a la media son aquellos estudiantes del género masculino que si aprobaron la materia "cultura física" en el año 2015(1) cuya estado de movilidad es normal y no han desertado.

Además se muestra otro clúster donde el 21% de los casos más cercanos a la media son estudiantes de sexo masculino que no han aprobado la materia "Física I" del periodo 2014(1) y sin embargo no son desertores.

3.4.6.5. Modelo Clúster utilizando el algoritmo Farthest-First para la predicción de estudiantes desertores.

Para este ejercicio inicialmente se tiene:

- El número de instancias o registros es 16840.
- Se tienen diez atributos: sexo, nivel, año, materia, promedio materia, materia perdida, trámite de movilidad, estado de movilidad y desertor.
- Se usa un conjunto de datos de entrenamiento (Use training set)

Tiene el mismo procedimiento que K-Means también elige los centroides pero este algoritmo utiliza el punto arbitrario más alejado a la media de los valores, aquí la asignación del grupo es diferente obtenemos un enlace con un alto número de sesiones como en el cluster-0 más que en cluster-1, y así sucesivamente. El primer algoritmo más alejado necesita menos ajustes.

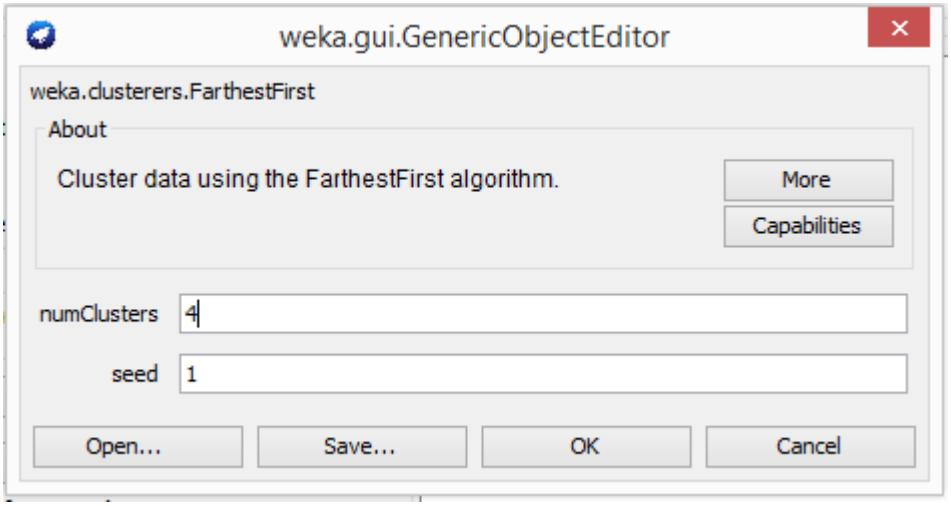


Ilustración 70: Configuración de Algoritmo Farthest-First

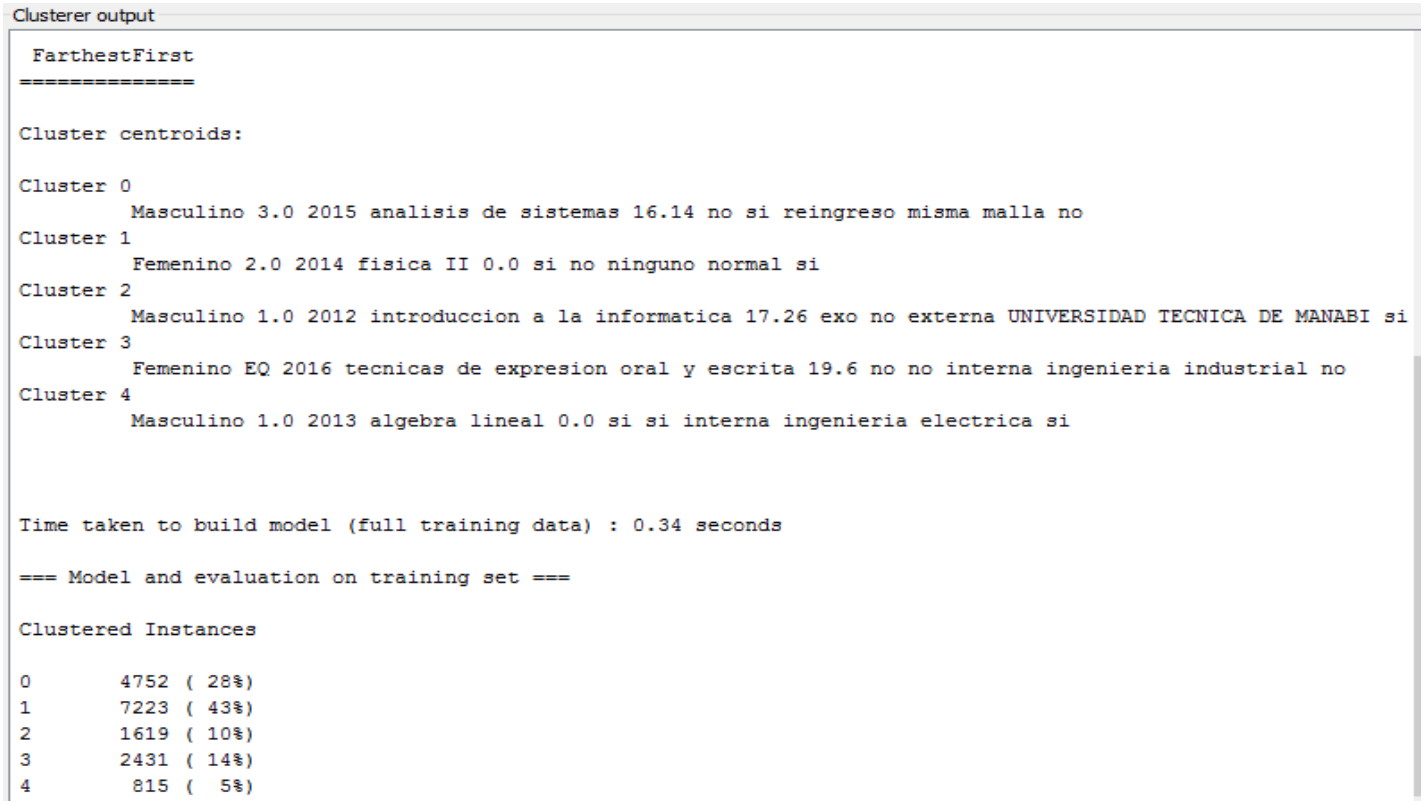


Ilustración 71: El algoritmo Farthest-First muestra que el 43% de los casos más alejados de la media son aquellos desertores de sexo femenino que han perdido la materia física II en el año 2014(2).

Además se muestra otro clúster donde el 5% de los casos más lejanos a la media son estudiantes desertores de sexo masculino provenientes de la facultad de ingeniería eléctrica que ingresan a 1N en el 2013 y que han perdido el semestre y la materia en este caso “algebra lineal”.

CAPITULO IV

EVALUACIÓN DE RESULTADOS

Los resultados de cada modelo se han ido analizando al tiempo de irlos exponiendo en el capítulo anterior, en este espacio se muestran algunos resultados erróneos producto de haber seleccionado de forma incorrecta los campos.

Como caso de ejemplo, se revisa lo que sucede al pretender predecir el nivel en la que los estudiantes desertan.

```
Attributes:   4
              Sexo
              Estado Academico
              Nivel
              desertor
```

Ilustración 72: Atributos de un listado incorrecto de reglas, empleando el algoritmo J48.

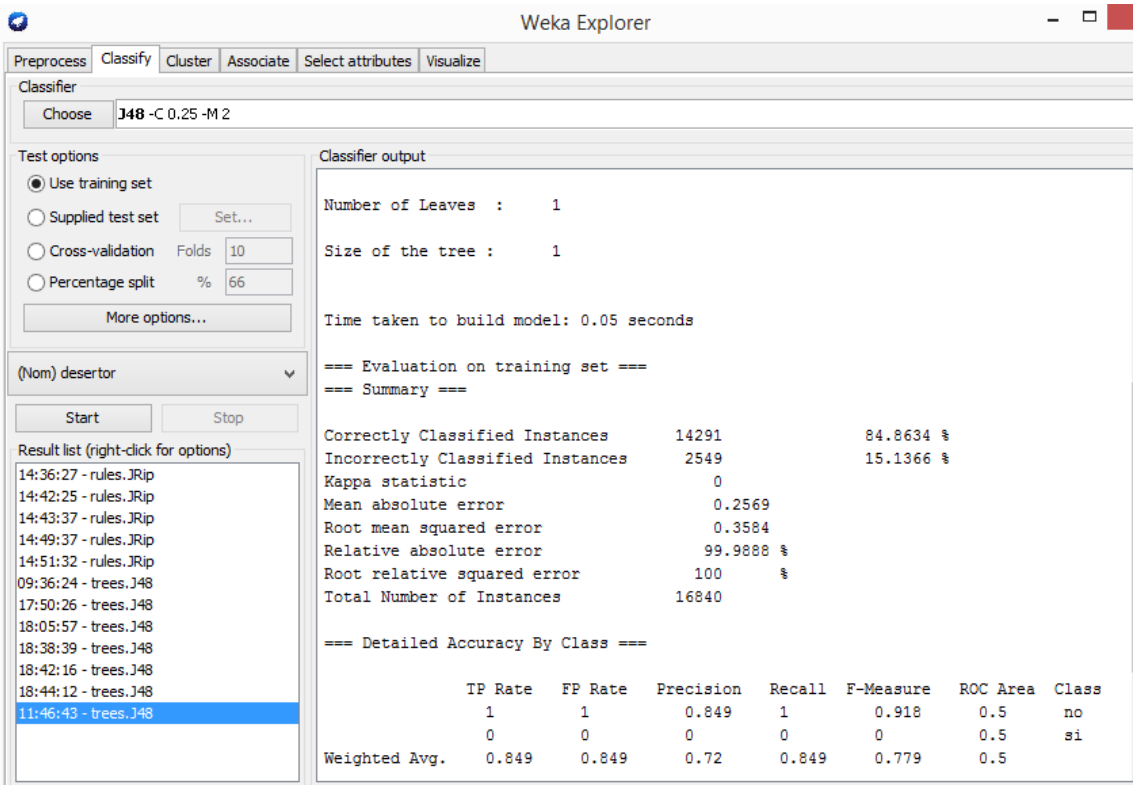


Ilustración 73: Aplicando otros atributos para la predicción WEKA registra que el 15.13 % de los registros han sido incorrectamente clasificados.

Otro caso de ejemplo, se revisa lo que sucede al pretender predecir el año en que se presentaron desertores de acuerdo al promedio.

```
Attributes:   3
              Año
              Promedio Materia
              desertor
```

Ilustración 74: Atributos de un listado incorrecto de reglas, empleando Naive Bayes.

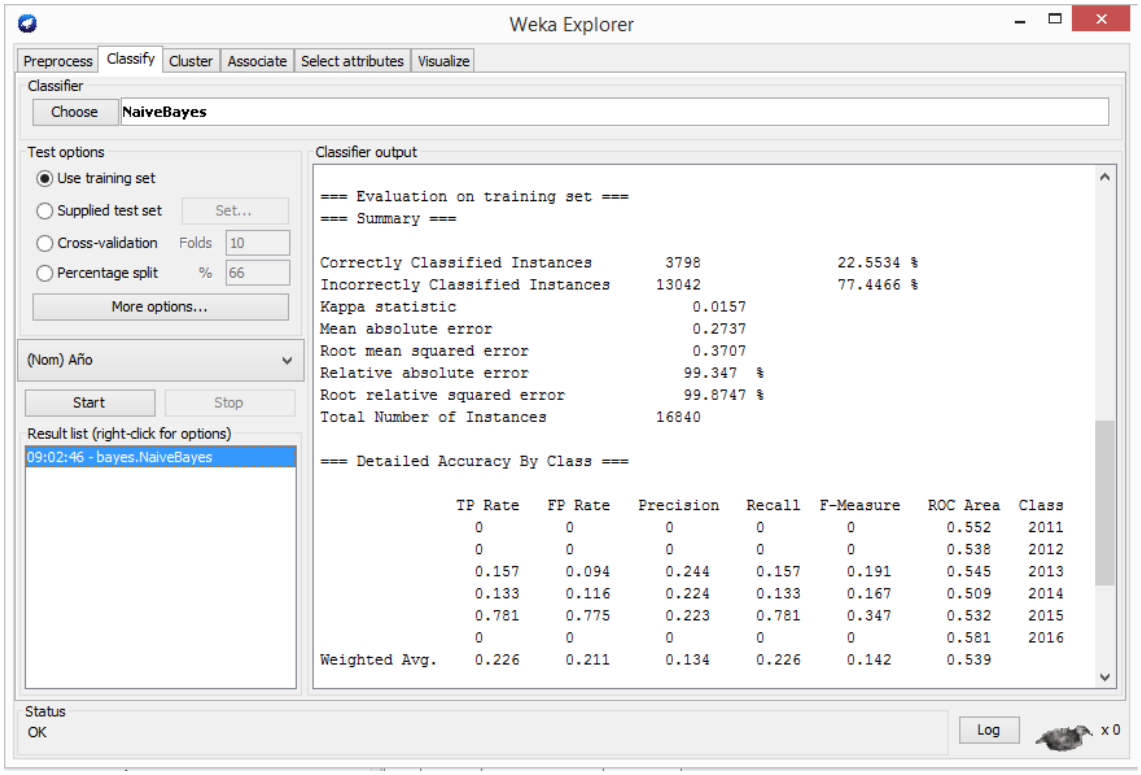


Ilustración 75: Aplicando otros atributos para la predicción WEKA registra que el 77.446 % de los registros han sido incorrectamente clasificados.

Otro caso de ejemplo, se revisa lo que sucede al pretender predecir los estudiantes desertores de acuerdo con los atributos escogidos y tomando como referencia el atributo año.

```
Attributes:   10
              Sexo
              Nivel
              Año
              Materia
              Promedio Materia
              materia perdida
              semestre perdido
              tramite movilidad
              Estado movilidad
              desertor
```

Ilustración 76: Atributos de un listado incorrecto de reglas, empleando JRip.

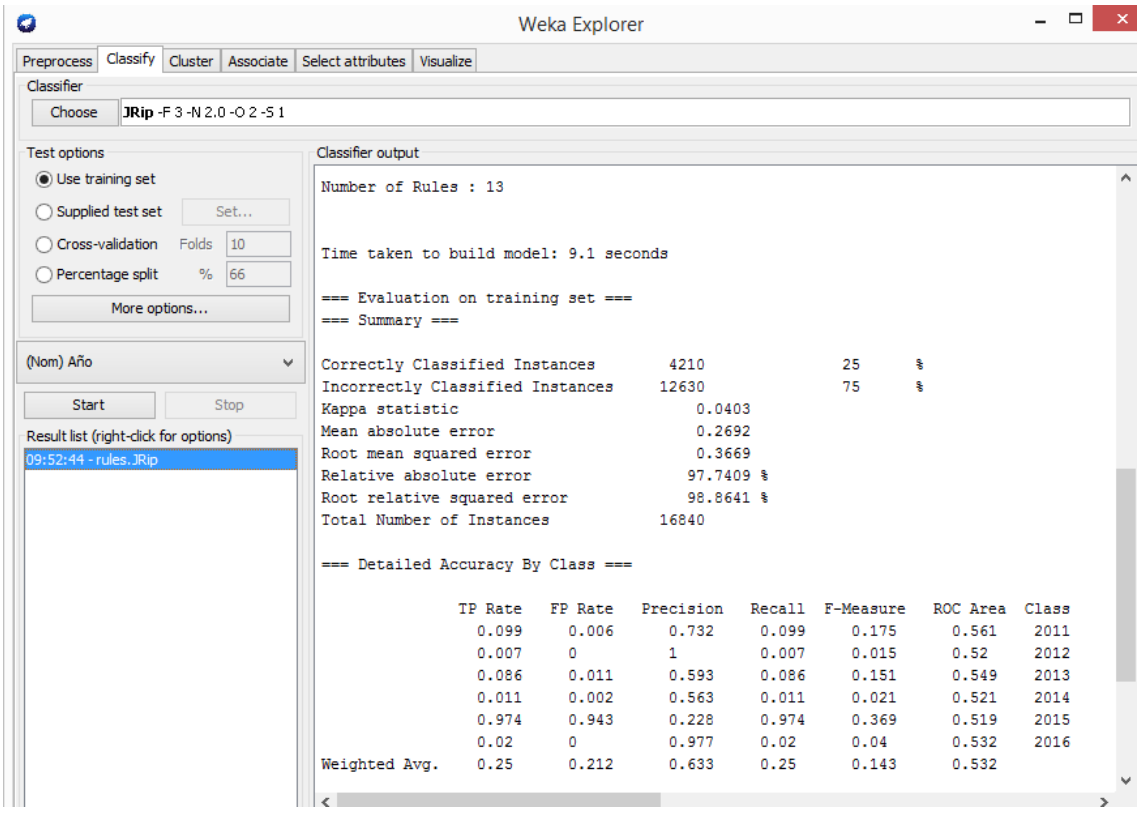


Ilustración 77: Si bien el programa ha producido 13 reglas, reconoce que el 75% de los registros ha sido incorrectamente clasificados.

Como se puede observar si los atributos no son escogidos correctamente las predicciones no serán tan aceptables, lo que puede ocasionar una mala interpretación de los datos.

CONCLUSIONES

Seguidamente una síntesis de las importantes opiniones formadas durante el proceso del presente trabajo. A partir de estas se podrá mostrar el cumplimiento de los objetivos planteados.

A continuación, se describen las principales conclusiones de este proyecto:

- Se distinguió en el marco teórico de la investigación diferentes factores que afectan la tasa de retención estudiantil entre estos la falta de apoyo económico, poco aprovechamiento de las clases, problema que se observan sobre todo en los tres primeros niveles y en edades entre los 18 y 20 años y uno que otro entre los 21 a 26 años.
- Se estudió en el marco teórico de la investigación diferentes técnicas para desarrollar modelos de predicción, basados en sistemas de soporte a las decisiones, utilizando técnicas de minería de datos, tales como Árboles de Decisión, técnicas de clasificación, como Clúster K-Means.
- Se seleccionó un subconjunto de modelos que han presentado un mejor desempeño, probando modelos para clasificar en forma automática a los alumnos con mayor riesgo de deserción, determinando que el modelo árbol de decisión, tiene un mejor comportamiento respecto a los modelos de Redes Bayesianas, Reglas de clasificación y Clústers.
- En esta investigación se ha descubierto que en general las instituciones no recolectan la suficiente información referente a la caracterización del estudiante al momento de ingresar a estudiar, que permitan establecer modelos de predicción de retención.
- Este trabajo permitió apreciar la importancia que tiene el proceso de recopilación de datos, abarcando las fases de análisis y preparación de los datos.

RECOMENDACIONES

Como se indicó desde el principio el actual proyecto ha asumido como objetivo implementar estructuras de minería de datos que identifiquen factores que influyan en el entorno académico para evaluar la tasa de retención estudiantil. Sin embargo, los datos de los estudiantes de pre grado de la Facultad de Ciencias Informáticas de la ULEAM, se puede apreciar que la información actual que recoge la institución es relevante tan solo en lo académico, pero se necesitan de más variables para la realización de estudios completos de minería de datos asociados a la deserción.

De lo ya expuesto, en este trabajo planteo que se recoja un conjunto de atributos asociados al problema de deserción universitaria que además de incluir los académicos también se agrupen más factores como institucionales y socioeconómico, se recomienda sean tomados al momento que el estudiante se matricula en la institución, con el objeto de poder crear un repositorio de registro histórico.

Tipo	Atributo	Descripción de Valores
Académicos	Nombre	Nombre del estudiante
	Edad	Rango de edad posible
	Sexo	Femenino, Masculino
	Estado Civil	Soltero, casado, divorciado, viudo
	Lugar Residencia	Ubicación de vivienda
	Fecha ingreso	Fecha valida de rango de estudio
	Tramite de movilidad	Interno, externo, reingreso, tercera matricula
	Promedio ENES	Puntaje alcanzado 100 -900
	Promedio	Promedio General 0-20
Institucionales	Colegio	Nombre del colegio
	Tipo de colegio	Técnico, científico, informático.
	Universidad	Nombre de la universidad
	Tipo de Universidad	Fiscal, Particular
Socio económicos	Estrato	Bajo, Medio, Alto
	Estado Laboral	Si, No
	Nivel Educativo de Padres	Primarios, Secundarios, Sin estudios.
	Personas a cargo	Si, No
	Desertor	Si, No

Tabla 16: Variables a considerar en estudios futuros

La recomendación antes detallada, está orientada a mejorar el proceso de toma de datos al momento que el estudiante se matricula, con el objeto de mejorar las predicciones en los futuros estudios.

BIBLIOGRAFÍA

References

Angúlo, F., & Sergio, E. (2012). Modelo para la automatización del proceso de determinación de riesgo de deserción en alumnos universitarios.

Aronson, J. E., Liang, T., & Turban, E. (2005). *Decision support systems and intelligent systems* Pearson Prentice-Hall.

Bazantes, Z. P., Carpio, M. L. R., & Gutiérrez, M. L. Á. (2017). Deserción estudiantil universitaria en ecuador y su influencia en la calidad del egresado. *Revista Magazine De Las Ciencias.ISSN 2528-8091, 1*(4), 65-70.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 step-by-step data mining guide.

Díaz Peralta, C. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. *Estudios Pedagógicos (Valdivia), 34*(2), 65-86.

Díaz, C. J. (2009). Factores de deserción estudiantil en ingeniería: Una aplicación de modelos de duración. *Información Tecnológica, 20*(5), 129-145.

El Comercio. (2016,). **El 26% de los universitarios se retiró en los primeros años**. *Diario EL COMERCIO*

EL Telegrafo. (2016, 10-NOV-2016). La deserción universitaria bordea el 40% . *Diario EL TELEGRAFO*

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18.

Holmes, G., Donkin, A., & Witten, I. H.Weka: A machine learning workbench. Paper presented at the *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference On,* 357-361.

Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms* John Wiley & Sons.

Marcillo, C., Blanco, M., Espinoza, T., Quinchiguano, D., & Andrade, K. (2017). Factores que impiden el logro de metas académicas de un estudiante centralino de pregrado. Paper presented at the *Anales, , 373*(1) 147-154.

Martínez-Padilla, J. H., & Pérez-González, J. A. (2008). Efecto de la trayectoria académica en el desempeño de estudiantes de ingeniería en evaluaciones nacionales. *Formación Universitaria, 1*(1), 3-12.

Microsoft. (2016). Algoritmos de mineria de datos. Retrieved from https://msdn.microsoft.com/es-es/library/ms174916(v=sql.120).aspx

Microsoft, 2. (2016). **Conceptos de minería de datos**. Retrieved from https://msdn.microsoft.com/es-es/library/ms174949(v=sql.120).aspx#DefiningTheProblem

Moine, J. M., Haedo, A. S., & Gordillo, S. E. (2011). Estudio comparativo de metodologías para minería de datos. Paper presented at the *XIII Workshop De Investigadores En Ciencias De La Computación,*

Moreira, E. V. G., Rodríguez, R. C., Tumbaco, S. C., Santana, Y. P., Fernández, R. L., & Benítez, L. B. C. (2017). Factores que influyen en la deserción de los estudiantes en la universidad de guayaquil. *Revista De La Facultad De Ciencias Medicas,*

Oocities. (2004). Sistemas y herramientas de minería de datos. ejemplos: Retrieved from http://www.oocities.org/es/mineria.datos/sistemas_herramientas_mineria_datos.pdf

Orallo, H., RAMIREZ, J., QUINTANA, C. R., Orallo, M. J. H., Quintana, M. J. R., & Ramírez, C. F. (2004). *Introducción a la minería de datos* Pearson Prentice Hall,.

Pérez-Palacios, T., Caballero, D., Caro, A., Rodríguez, P. G., & Antequera, T. (2014). Applying data mining and computer vision techniques to MRI to estimate quality traits in iberian hams. *Journal of Food Engineering, 131*, 82-88.

Quispe Parí, D. J., & Sánchez Mamani, G. (2011). Encuestas y entrevistas en investigación científica. *Revista De Actualización Clínica Investiga, 10*, 490-494.

Riquelme Santos, J. C., Ruiz, R., & Gilbert, K. (2006). Mineria de datos: Conceptos y tendencias. *Inteligencia Artificial: Revista Iberoamericana De Inteligencia Artificial, 10*(29), 11-18.

Romero, M. L.**Método analítico-sintético**. Retrieved from https://www.lifeder.com/metodo-analitico-sintetico/

Soria-Barreto, K., & Zúñiga-Jara, S. (2014). Aspectos determinantes del éxito académico de estudiantes universitarios. *Formación Universitaria, 7*(5), 41-50.

Soto, E. (2015). Tecnicas de estudio. Retrieved from http://emilsesoto.blogspot.com/2015/07/republica-bolivariana-de-venezuela_25.html

ANEXOS

Encuesta al personal docente FACCI

UNIVERSIDAD LAICA ELOY ALFARO DE MANABÍ

FACULTAD DE CIENCIAS INFORMÁTICAS

Dirigido a: Personal docente de la Facultad de Ciencias Informáticas - ULEAM. Manta.

Objetivo. - Implementar estructuras de minería de datos que identifiquen factores que influyan en el entorno académico para evaluar la tasa de retención estudiantil de la facultad de ciencias informáticas.

Instrucciones: Lea detenidamente cada pregunta y marque con un (✓) en la respuesta que considere conveniente.

Implementación de estructuras de minería de datos para evaluar el nivel de tasa de retención estudiantil de la Facultad de Ciencias Informáticas.

1.- Cree usted que los posibles factores que provoquen la deserción estudiantil en la institución sean los siguientes:

- ☐ individuales
- ☐ Académicos
- ☐ Institucionales
- ☐ Otros

2.- ¿Considera importante conocer cuáles son las razones académicas porque las que el estudiante de esta facultad decide abandonar sus estudios?

- ☐ Si

- ☐ No

3.- Cree usted que la baja tasa de retención estudiantil en el ámbito académico se da las siguientes razones:

- ☐ Rendimiento académico
- ☐ Métodos de estudio
- ☐ Orientación profesional
- ☐ Calidad del programa de estudio
- ☐ Otros

4.- Cree usted que la baja tasa de retención estudiantil en el ámbito institucional se da las siguientes razones:

- ☐ Normalidad académica
- ☐ Modelos pedagógicos
- ☐ Recursos universitarios
- ☐ Perfil profesional de la carrera
- ☐ Relaciones con los profesores y otros estudiantes
- ☐ Otros

5.- ¿Cuáles según usted son las consecuencias que trae la deserción en la sociedad y para la facultad?

- ☐ Trabajo mal remunerado
- ☐ Delincuencia
- ☐ Discriminación
- ☐ Otros

6.- ¿Considera importante este tipo de estudios sobre la deserción para la toma de decisiones en la institución? Escriba el porqué de su respuesta.

- ☐ Si
- ☐ No

7.- ¿Qué variables externas cree usted que causan deserción en su institución?

- ☐ Económicas
- ☐ Sociales
- ☐ Familiares
- ☐ Médicas

8.- ¿Qué hace usted en el aula para evitar la deserción estudiantil?

- ☐ Motivar la asistencia
- ☐ Motivar el aprendizaje
- ☐ Utilizar técnicas innovadoras de estudio
- ☐ Otras

9.- ¿Considera importante tener datos estadísticos de los estudiantes que desertan? Escriba el porqué de su respuesta.

- ☐ Si
- ☐ No

Encuesta al estudiante

1.- ¿Qué situaciones cree usted que causarían el abandono de sus estudios o si por el momento no se encuentra estudiando que lo provoco?

- ☐ Falta de recursos económicos.
- ☐ Vivir lejos de donde estudias.
- ☐ Poco aprovechamiento de las clases.
- ☐ Falta de interés por seguir realmente con las carreras universitarias.
- ☐ Porque por ley están donde su puntaje del ENES los ubicó.
- ☐ Porque no pueden acceder a universidades particulares por lo costoso que resulta.
- ☐ Por enfermedad suya o de un familiar.
- ☐ Por oportunidad de trabajo.
- ☐ Porque el perfil profesional y ocupacional de su carrera no son de su agrado.
- ☐ Porque reprueba las asignaturas constantemente.
- ☐ Falta de interacción de calidad con profesores y orientadores.
- ☐ Ambiente poco motivante en clases.
- ☐ Falta de apoyo familiar.
- ☐ Se convirtió en madre o padre de familia.
- ☐ Por motivo de viaje.
- ☐ otros.