UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

CENTRO UNIVERSITARIO UAEM VALLE DE MÉXICO

Análisis de los factores académicos y educativos de un estudiante para generar un clasificador de carrera utilizando el algoritmo C4.5

TESIS

Que para obtener el título de:

INGENIERO EN SISTEMAS Y COMUNICACIONES

Presenta

C. Luis Alberto García Madrid

Asesora: Dra. en C. Com. Maricela Quintana López

Atizapán de Zaragoza, Estado de México, abril 2019

RESUMEN

Se presenta el análisis de factores académicos y educativos, aplicando el algoritmo C4.5 de minería de datos, con el fin de seleccionar los más útiles y construir un clasificador para elegir una carrera. Dentro de los factores académicos, se consideraron los resultados del EXANI-II del CENEVAL y los de escolaridad; de los factores educativos se consideraron los hábitos y actividades de estudios, y las formas de aprendizaje. Los datos empleados corresponden a los estudiantes de la generación 2008 – 2013 que cursaron con éxito la carrera elegida en el centro universitario UAEM Valle de México.

La hipótesis del trabajo es que si los factores considerados del estudiante fueron suficientes para concluir la carrera elegida entonces es posible generar inductivamente un clasificador de la carrera usando el algoritmo de minería de datos C4.5, el cual genera un árbol de decisión. El clasificador generado determina si la carrera es una licenciatura o ingeniería y luego se aplica el modelo de las ingenierías o las reglas para las licenciaturas para apoyar al estudiante en esta difícil selección.

La metodología empleada está basada en el proceso de extracción del conocimiento, se aplica el algoritmo C4.5 con cada uno de los factores antes mencionados y se hace un análisis para determinar cuáles son los factores determinantes en la elección de una carrera.

El clasificador construido da como resultado un 98.06% de exactitud y determina que los factores que inciden en la separación de las clases son los relativos a los resultados del examen de nuevo ingreso EXANI-II del Centro Nacional de Evaluación (CENEVAL). Los atributos principales fueron: Razonamiento Numérico, Ciencias Sociales y Razonamiento Verbal.

ABSTRACT

The analysis of academic and educational factors, applying the C4.5 data mining algorithm is presented, in order to select the most useful and build a classifier to choose a career. Within academic factors, results of the EXANI-II from the National Evaluation Center (CENEVAL) and schooling were considered. As educational factors, studies habits, forms of learning and study activities were considered. The data used for analysis correspond to the students of the generation 2008 - 2013 who completed successfully chosen career in the university center UAEM Valley of Mexico.

The hypothesis of this work is that if the considered student factors were enough to finish the chosen career then it is possible to generate, in an inductive way, a career classifier using the data mining algorithm known as C4.5, which generates a decision tree. The generated classifier determines the type of the bachelor's degree. After that, the engineering model or the rules are applied to support the student in this difficult decision.

The methodology used is based on the knowledge extraction process, algorithm C4.5 is applied with each one of the aforementioned factors and an analysis is made to determine which are the determining factors in the choice of a career.

The built classifier model results in a 98.06% accuracy and determines that factors affecting the split of classes are those related to EXANI-II test of CENEVAL. The main attributes of the EXANI-II are: Numerical Reasoning, Social Sciences and Verbal Reasoning.

Índice

Índice de tablas

Índice de figuras

Capítulo 1. Introducción

1.1 Antecedentes

El fracaso escolar es un tema de gran preocupación en diferentes escuelas y universidades del mundo, y por ello, se han realizado investigaciones cuyo interés se centra en determinar cuáles son las causas que generan dicho fracaso. Aunque existen investigaciones de lo que ocurre en los diferentes niveles de educación, la mayoría se centra en la educación superior (Márquez, Romero, Ventura, 2012).

Actualmente en México, se enfrenta un problema en la educación superior que es el bajo desempeño, el cual puede conducir, en el peor de los casos, al abandono de los estudios o a cursarlos en un periodo de tiempo mayor, y a un bajo porcentaje de titulación respectivamente.

El problema mencionado se mide por los índices de reprobación, deserción, eficiencia terminal y titulación respectivamente. Por mencionar un dato, en el estado de México, de acuerdo con el anuario estadístico en la educación superior en el ciclo escolar 2015-2016, ingresaron 120780 estudiantes a una carrera de licenciatura, 64179 egresaron de estas carreras (53% de eficiencia terminal), y sólo 44,045 se titularon (36% titulación) (ANUIES, 2015-2016).

Particularmente, en la Universidad Autónoma del Estado de México (UAEM), las razones por las que un alumno se ve obligado a abandonar sus estudios pueden ser dos: motivos personales y baja académica; mientras que los motivos personales pueden estar refiriéndose a una gran cantidad de factores, entre ellos los sociales y los económicos; la baja académica se aplica cuando un alumno reprueba dos veces la misma unidad de aprendizaje obligatoria, o cuando acumula un total de 20 evaluaciones reprobadas.

Conscientes de esta problemática, en la UAEM, se han creado estrategias y programas de apoyo a los alumnos, que consisten en talleres, apoyo psicológico, becas institucionales, y asesorías académicas disciplinarias.

Particularmente, en la UAEM se cuenta con un Sistema de Información de Tutoría Académica y Asesoría de la UAEMex (SITAA), ver figura 1.1, el cual surge como una herramienta de apoyo para que el tutor académico del grupo determine la situación de cada uno de sus tutorados (SITAA, 2018). Este sistema le permite conocer al tutor académico, la situación social, económica y académica de cada uno de sus alumnos.

Figura 1.1 Sistema de Información para de Tutoría Académica y Asesoría de la UAEMex.

El tutor de un grupo de estudiantes, puede utilizar el sistema SITAA para revisar diversos datos de un alumno. Como se muestra en la figura 1.2, dichos datos pueden ser, los resultados obtenidos en el Examen Nacional de Ingreso a la

Educación Superior (EXANI-II), los datos del estudio socioeconómico, el seguimiento académico, y el expediente de tutorías.

Figura 1.2 Módulos de consulta del tutor del SITAA.

Particularmente, el EXANI-II es el examen que aplica el Centro Nacional de Evaluación (CENEVAL) a los alumnos que pretenden realizar sus estudios superiores (ver figura 1.3); en este examen se evalúan las áreas mostradas en la Tabla 1.1.

Tabla 1.1 Área de evaluación.

ÁREA	CLAVE
Ciencias Sociales	CS
Razonamiento numérico	RN
Matemáticas	MAT
Mundo Contemporáneo	MC
Razonamiento Verbal	RV
Ciencias Naturales	CN
Español	ESP

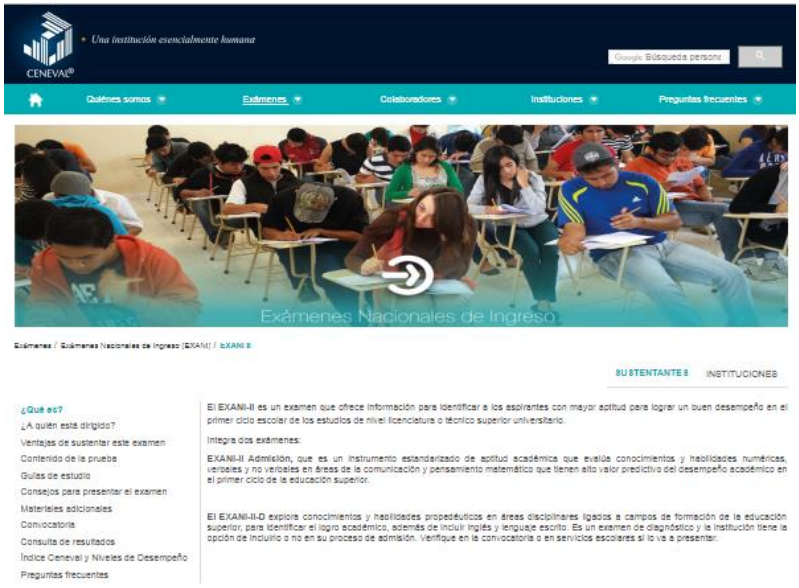
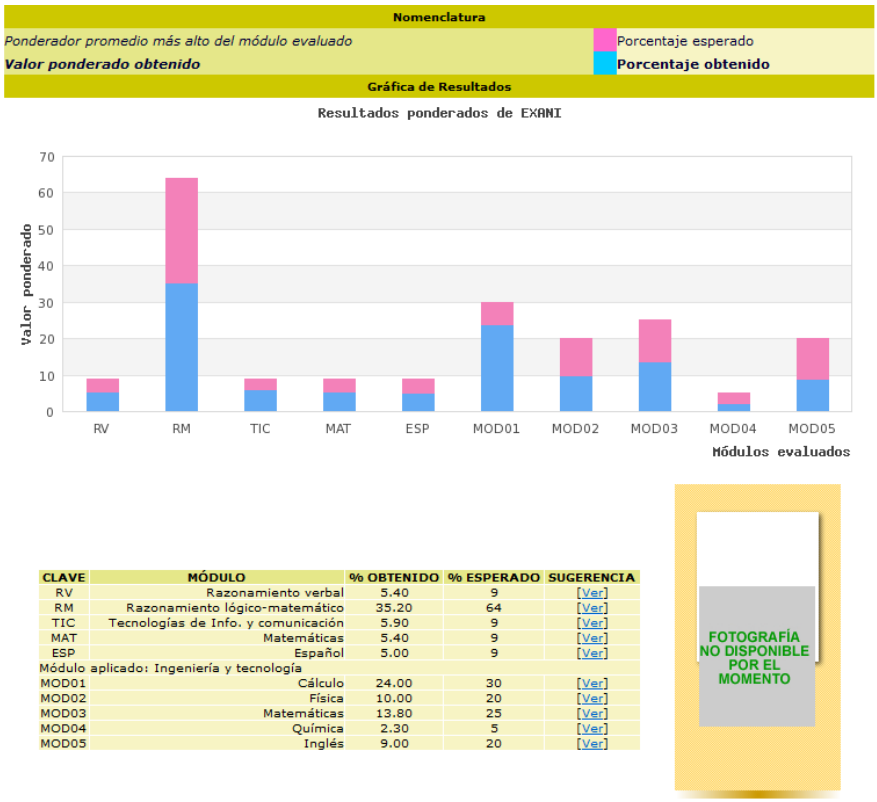


Figura 1.3 CENEVAL, EXANI-II.

Los resultados obtenidos en el examen EXANI-II pueden ser consultados por el tutor en el SITAA (ver Figura 1.4).



CLAVE	MÓDULO	% OBTENIDO	% ESPERADO	SUGERENCIA
RV	Razonamiento verbal	5.40	9	[Ver]
RM	Razonamiento lógico-matemático	35.20	64	[Ver]
TIC	Tecnologías de Info. y comunicación	5.90	9	[Ver]
MAT	Matemáticas	5.40	9	[Ver]
ESP	Español	5.00	9	[Ver]
Módulo aplicado: Ingeniería y tecnología				
MOD01	Cálculo	24.00	30	[Ver]
MOD02	Física	10.00	20	[Ver]
MOD03	Matemáticas	13.80	25	[Ver]
MOD04	Química	2.30	5	[Ver]
MOD05	Inglés	9.00	20	[Ver]

Figura 1.4 Resultados del EXANI-II.

Los datos del estudio socioeconómico se pueden observar en la figura 1.5, estos incluyen: factores generales y factores educativos. Dentro de los factores generales tenemos: Identificación del sustentante, identificación de la institución de procedencia, escolaridad, estructura familiar, situación laboral, estudio socioeconómico; mientras que en los factores educativos se encuentran: medios para mejorar la escuela, actividades culturales, actividades de los maestros, libros e idiomas, actividades de estudio, formas de aprendizaje y hábitos de estudio.

Figura 1.5 Estudio Socioeconómico.

Vale la pena mencionar, que cada uno de los grandes rubros del estudio socioeconómico contiene un cuestionario con mayor detalle. Para ejemplificar, mostramos en la figura 1.6, el cuestionario correspondiente a los datos socioeconómicos.

Por otro lado, en la parte de seguimiento académico, durante el semestre, en el sistema SITAA se incorporan las calificaciones del primer y segundo parcial, así como de la calificación final obtenida en cada una de las materias que cursa el alumno. Dichas calificaciones sirven como guía al tutor académico del grupo para apoyarlo en la toma de decisiones (figura 1.7). Respecto al expediente de tutorías, solo indica las tutorías que el estudiante ha recibido y su validación.

Factores generales -> Datos socioeconómicos
Corrobore y complete la información que proporcionó.

¿Cuál es el ingreso familiar mensual?	DE $4,001 A $5,000
¿Cuál es el ingreso personal mensual?	MENOS DE $1,000
Nivel máximo de estudios del padre	PRIMARIA
Nivel máximo de estudios de la madre	BACHILLERATO, PREPARATORIA Ó VOCACIONAL
Ocupación actual del padre	OBRERO
Ocupación actual de la madre	NO TRABAJA ACTUALMENTE

Servicios y bienes con los que cuenta en su casa	
Drenaje	DRENAJE
Agua entubada	AGUA ENTUBADA
Alumbrado público	ALUMBRADO PÚBLICO
Calles pavimentadas	CALLES PAVIMENTADAS
Recolección periódica de basura	RECOLECCIÓN PERIÓDICA DE BASURA
Calentador de gas para agua	CALENTADOR DE GAS PARA AGUA
Un cuarto propio para dormir	UN CUARTO PROPIO PARA DORMIR
Un lugar exclusivo para estudiar	0
Automóvil familiar o propio	AUTOMÓVIL FAMILIAR O PROPIO
Teléfono	TELÉFONO
Teléfono celular	TELÉFONO CELULAR
Televisión	TELEVISIÓN
Televisión por cable o por satélite	0
Videograbadora	0
Reproductor de DVD	REPRODUCTOR DE DVD
Calculadora	CALCULADORA
Computadora	COMPUTADORA
Diccionario o enciclopedia	DICCIONARIO O ENCICLOPEDIA
Suscripción a revista o periódico	0
Conexión a Internet	CONEXIÓN A INTERNET

Figura 1.6 Datos Socioeconómicos.

Sistema Inteligente para la Tutoría Académica

SEGUIMIENTO ACADÉMICO SITA

Usted está aquí: **Inicio -> Seguimiento Académico -> Calificaciones parciales**

Consulta de las evaluaciones parciales del periodo vigente

Tutor(a):	MARICELA QUINTANA LOPEZ	**Escuela:**	CENTRO UNIVERSITARIO UAEM VALLE DE MÉXICO
Programa:	Ingeniero en Sistemas y Comunicaciones	**Periodo actual:**	2014B
Alumn@:	GARCIA MADRID LUIS ALBERTO	**No. de cuenta:**	1229732

Unidades de aprendizaje en curso. Peridodo: 2014B

N.P.	Clave de acta	Periodo	Núcleo	Unidad de aprendizaje	Calificación	Créditos	Evaluacion	Indicador
1	225-230-53--01-OCT-14	2014B	Sustantivo	INVESTIGACION DE OPERACIONES	064	8	P01	A
2	225-283-53--30-SEP-14	2014B	Sustantivo	BASES DE DATOS	085	8	P01	A
3	225-284-53--26-SEP-14	2014B	Sustantivo	CIRCUITOS ELECTRICOS	090	6	P01	A
4	225-291-53--02-OCT-14	2014B	Sustantivo	PROGRAMACION ORIENTADA A OBJETOS	078	8	P01	A
5	225-301-53--29-SEP-14	2014B	Sustantivo	REDES	053	10	P01	R

NOTA: Esta es información espejo de los datos que obran en el Sistema Institucional de Control y Desempeño Escolar (SICDE), sin embargo NO es un comprobante con validez oficial.

Figura 1.7 Seguimiento Académico.

Como se mencionó, los factores que llevan a un bajo desempeño en algunas materias de la carrera, o al abandono de los estudios, puede deberse a múltiples factores: personales, escolares y socio familiares.

Uno de los posibles factores personales considerado en este trabajo, es que el estudiante no haya elegido correctamente la carrera a estudiar, esto porque son muy jóvenes al momento de decidir y, algunos aún no saben lo que quieren y/o tienen dudas respecto así terminarán la carrera o tendrán buenos resultados en esa carrera particular.

La elección de carrera representa una decisión importante, ya que, si la terminan con éxito, es algo a lo que se dedicarán toda su vida, o bien si a medio camino desertan de los estudios, habrán perdido tiempo, dinero y esfuerzo.

Como se puede percatar, el SITAA contiene información de relevancia acerca de los alumnos que han cursado o están cursando alguna carrera en el Centro Universitario Valle de México, y en la UAEM en general. Esta información puede ser utilizada para generar modelos que permitan determinar las características de los alumnos de cada carrera en particular.

Con la información almacenada en el SITAA, es posible crear clasificadores que nos permitan determinar qué influye en cada una de las carreras, qué hábitos de estudio,

qué conocimientos, qué actividades o formas de aprendizaje los caracterizan. De esta forma, se podría informar al estudiante, si los conocimientos, habilidades y actitudes que tiene, le sirven para esa carrera o si sus cualidades son mejores para otra, o bien si decide estudiar la que le llama la atención, qué actividades o formas de aprendizaje o conocimientos debe reforzar.

En este trabajo, se busca realizar un análisis de la información almacenada en el SITAA aplicando el algoritmo C4.5 para la generación de modelos que permitan a un alumno determinar la carrera a estudiar basado en sus conocimientos, hábitos y actitudes.

1.2 Planteamiento del problema

Como se ha mostrado, en el SITAA existen los datos, más no información que apoye en la toma de decisiones. Sería de gran relevancia que esos datos pudieran utilizarse para diagnosticar la carrera para la cual el alumno es apto, considerando las respuestas a su examen de ingreso EXANI y a sus hábitos de estudio.

Se sabe que pueden ser muchos los factores que pueden llevar al éxito o fracaso de un estudiante en una carrera, este trabajo se centrará en aspectos académicos y educativos.

A diferencia de aplicar un test vocacional, donde se le pregunta al estudiante que le gusta o disgusta, este trabajo se basa en utilizar el algoritmo C4.5 para analizar los factores mencionados en un grupo de estudiantes de la generación 2008 – 2013, los cuales presentaron el examen de ingreso EXANI-II del Centro Nacional de Evaluación, CENEVAL, completaron el cuestionario de factores generales y educativos y fueron aceptados para cursar una carrera, la cual concluyeron con éxito en el Centro Universitario UAEM Valle de México.

Se considera de suma importancia tener un modelo que ayude al estudiante a elegir una carrera basándose en algunos aspectos medibles de forma cuantitativa y cualitativa de su persona.

Dado que la única información que se tiene en ese momento es la recabada al momento de hacer el EXANI-II, se propone utilizar la minería de datos para realizar tareas de clasificación, en este caso particular se utilizará el algoritmo C4.5.

1.3 Justificación

Es importante contar con una herramienta que apoye a los aspirantes a elegir una carrera profesional, en ocasiones la pregunta para los estudiantes es ¿qué quieres ser? cuando a veces lo que se debe considerar son las habilidades que este tiene. Si por el contrario el aspirante está decidido sobre la carrera que quiere estudiar, sería de gran ayuda el poder indicarle, las habilidades con las que debe contar un estudiante de esa carrera.

Como se mencionó, en el SITAA existen los datos, más no información que apoye en la toma de decisiones. Se considera de suma importancia tener un clasificador que ayude al estudiante a elegir una carrera basándose en algunos aspectos académicos y educativos de su persona. Contar con este clasificador, permitiría aplicarlo a nuevos estudiantes desde el inicio de su carrera y crear estrategias y tomar acciones oportunas para lograr que cursen su carrera con éxito.

1.4 Objetivos

1.4.1 Objetivo general

Generar un clasificador que apoye la elección de carrera en el Centro Universitario UAEM Valle de México, basándose en los resultados del EXANI-II y el estudio socioeconómico, usando el algoritmo C4.5.

1.4.2 Objetivos específicos

- Preparar los datos obtenidos a ser utilizados para generar el clasificador, esto implica, recolectar e integrar los datos del EXANI-II, SITA y de Control Escolar, con el fin de seleccionar los que proporcionan más información, y transformarlos en caso de ser necesario.

- Realizar los experimentos, aplicando el algoritmo C4.5 a diferentes conjuntos de datos, con el fin de generar clasificadores que permitirán analizar los factores involucrados y medir el desempeño de cada clasificador.

- Analizar los resultados para elegir el mejor clasificador.

- Interpretar los resultados generados por el clasificador elegido.

1.5 Hipótesis

Es posible generar inductivamente, un clasificador de carrera que permita determinar lo que influye en cada una, partiendo de los hábitos de estudio, conocimientos, y formas de aprendizaje de cada estudiante; particularmente utilizando el algoritmo de minería de datos conocido como C4.5 que genera un árbol de decisión para representar el conocimiento adquirido.

1.6 Delimitación

Como se ha mencionado, se tienen como datos el estudio socioeconómico y los resultados del EXANI-II, en este trabajo son utilizados aquellos que son competencia única del estudiante, es decir, que no dependen de factores externos a él, como serían, por ejemplo, el ingreso del padre, si hay drenaje en casa o cuántos hermanos tiene.

Los experimentos para obtener el modelo se realizaron específicamente con la información correspondiente a los alumnos de la generación 2008 – 2013 del Centro Universitario UAEM Valle de México. Esto con el fin de utilizar datos de alumnos que ya hubiesen terminado sus carreras y que además hubieran contestado completamente el estudio socioeconómico.

Para el análisis de los datos, se empleó el algoritmo C4.5 dentro de la minería de datos, este algoritmo de clasificación permite obtener un modelo explicativo que captura el conocimiento en un árbol de decisión.

1.7 Metodología

La metodología empleada consistió en:

a) Recopilar la información respecto al estudio socioeconómico y los resultados del EXANI-II de los estudiantes de la generación 2008-2013.

b) Integrar la información con la situación escolar proporcionada por el departamento de control escolar con el fin de separar a los alumnos que fueron dados de baja.

c) Aplicar el algoritmo C4.5 a los resultados del EXANI-II y el estudio socioeconómico.

d) Interpretar el modelo generado por el algoritmo para identificar lo que influye en una carrera determinada.

Es conveniente aclarar que esta metodología está basada en el proceso de extracción del conocimiento, que se explicará a mayor detalle en el capítulo 2.

1.8 Publicaciones derivadas de la investigación

Modelo para la Elección de Carrera basado en el Análisis de Factores Académicos y Educativos usando Minería de Datos. Presentado en el Congreso Nacional Multidisciplinario de Educación, Ciencia y Tecnología CONAMTEC 2015 y publicado en la Revista de Tecnologías de Información, Vol. 2, No. 2, pp. 112-121. ECORFAN-Bolivia. ISSN 2410-4000 (Enero-Marzo 2015) (García y Quintana 2015).

1.9 Estructura de la tesis

La estructura de la tesis es como sigue:

- En el capítulo 2, se describirá el marco teórico correspondiente a lo que es el proceso de extracción del conocimiento.
- En el capítulo 3, se mencionan los fundamentos de la minería de datos, haciendo énfasis en los algoritmos de clasificación y se muestra el estado del arte referente a la elección de carrera usando métodos tradicionales y empleando minería de datos.
- Por otro lado, en el capítulo 4, se describe los experimentos realizados, desde la preparación de los datos hasta los resultados obtenidos.
- Finalmente se presentan las conclusiones y el trabajo futuro.

Capítulo 2. Proceso de extracción del conocimiento

El proceso de extracción o descubrimiento del conocimiento, mejor conocido como KDD por sus siglas en inglés (Knowledge Discovery from Databases), es un proceso iterativo que consiste en una serie de etapas que nos llevan a transformar los datos en conocimiento (Hernández, Ramírez, & Ferri, 2008).

En la figura 2.1 se muestran las etapas de esta metodología: integración y recopilación; selección, limpieza y transformación; minería de datos; evaluación e interpretación; y por último, difusión y uso. A continuación, se describe a detalle en qué consiste cada una de las etapas.

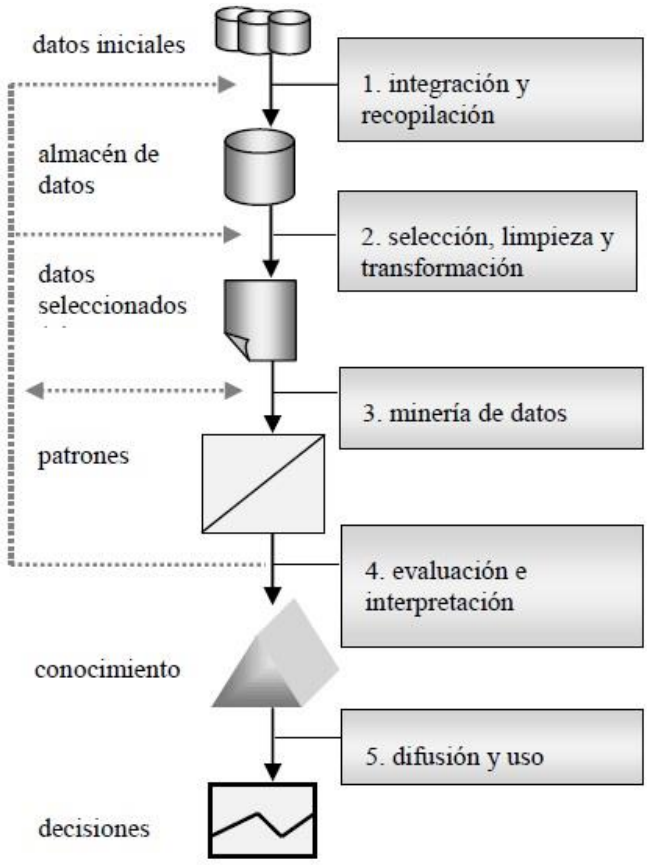


Figura 2.1 Etapas del proceso de descubrimiento de conocimiento, KDD (Hernández et al., 2008).

2.1 Integración y Recopilación

Esta etapa consiste en determinar de dónde serán obtenidos los datos, sobre todo si se trata de múltiples fuentes de almacenamiento como pueden ser archivos, bases de datos, cubos de datos o almacenes de datos.

Una vez que se tienen los datos, se procede a revisar su estructura. En el caso de que los datos tengan la misma estructura, la tarea se reduce a simplemente unirlos, de lo contrario se deberá realizar el proceso de unificación, en el cual se buscará que todos los datos tengan la misma estructura.

Para ilustrar esta fase, en el presente trabajo se cuenta con la información de tres fuentes: la información generada por el sistema de control escolar, la información del estudio socioeconómico, y los resultados del EXANI-II. Una vez que se ha reunido la información, hay que proceder a la selección, limpieza y transformación de los datos.

2.2 Selección, Limpieza y Transformación

Con el fin de que el algoritmo de minería de datos produzca conocimiento de buena calidad, es necesario identificar cuáles atributos o campos de los datos deben conservarse y cuáles no se deben tomar en cuenta. Esto también traerá como consecuencia que el algoritmo se ejecute más rápido y se requiera menos espacio de almacenamiento o bien se evite la necesidad de utilizar una máquina más potente para hacer el procesamiento.

La selección y limpieza de datos se realiza después de la etapa de integración, en ella se seleccionan los datos que se utilizarán y se eliminan aquellos que sean redundantes, presenten incoherencias, o tengan valores faltantes.

Para ilustrar lo anterior, se presenta el caso en el que se requiera generar un clasificador para determinar si se le debe otorgar una beca a un estudiante y se

cuenta con la siguiente información: número de cuenta, nombre, carrera, semestre, promedio, fecha de nacimiento, y edad. Es claro que los atributos número de cuenta y nombre no son necesarios ni deben utilizarse para construir el clasificador. Por otro lado, el atributo edad resulta innecesario, pues con la fecha de nacimiento se puede calcular.

Por otro lado, cuando los datos tienen incoherencias, valores con diferentes formatos, valores duplicados o faltantes es necesario hacer las correcciones apropiadas.

Valores con diferente formato

Ocurre cuando los valores del dominio de un atributo se especifican en diferente formato. Un ejemplo de esto son las fechas, que pueden escribirse como 19/Abril/17, 19/04/17, 17/04/19, ó 19/04/2017 (ver tabla 2.1), es necesario uniformizar las fechas para que queden en un solo formato.

Tabla 2.1 Diferentes formatos para especificar una fecha.

Matrícula	Nombre	Carrera	Semestre	Fecha de Ingreso
123456	Luis Alberto García	ISC	8	19-abr-17
123457	Luis Jesús Carvajal	ISC	8	19/04/2017
123458	Naomi García	ISC	8	17/04/2019
123459	Victoria Carvajal	ISC	8	19/04/2017

Valores duplicados

Es posible que por algún error tipográfico un registro quede duplicado, lo cual afecta a la generación del clasificador ya que se consideran 2 instancias en lugar de solo una. Un ejemplo de valores duplicados se presenta en la tabla 2.2.

Tabla 2.2 Instancias duplicadas.

Matrícula	Nombre	Carrera	Semestre	Fecha de Ingreso
123456	Luis Alberto García	ISC	8	19/04/2017
123457	Luis Jesús Carvajal	ISC	8	19/04/2017
123459	Naomi García	ISC	8	19/04/2017
123459	Naomi García	ISC	8	19/04/2017

Valores faltantes

En algunas ocasiones es posible que por algún descuido del estudiante no se llenen algunos registros provocando que el atributo pierda efectividad para lograr una predicción adecuada (ver tabla 2.3).

Tabla 2.3 Instancias incompletas.

Matrícula	Nombre	Carrera	Semestre	Fecha de Ingreso
123456		ISC	8	19/04/2017
	Luis Jesús Carvajal	ISC	8	19/04/2017
123459	Naomi García			19/04/2017
123459	Victoria Carvajal	ISC	8	

Sin embargo, se considera apropiado comentar que existen algoritmos capaces de trabajar cuando se tiene información faltante, ya sea que traten la información faltante como un valor extra del dominio del atributo, o sea porque se completa la instancia con el valor que corresponde a la moda o al promedio.

Ruido

Es cuando hay modificación de valores en las variables o son incorrectos lo cual genera una mala interpretación en los resultados.

La transformación de los datos se realiza para que el proceso de minería datos tenga mayor precisión o para que los patrones sean más simples de entender.

Los datos son normalizados para estar en un mismo rango, los atributos continuos se transforman en discontinuos para lo cual se les asigna etiquetas, esto se hace para facilitar el uso de técnicas que requieren ciertos valores específicos, pueden ser cambiados por valores numéricos por lo cual se reducen los espacios y da la posibilidad de usar técnicas numéricas o también se transforman los valores numéricos en valores nominales.

La fase de preparación de los datos que incluye a las etapas de integración y recopilación, y de selección, limpieza y transformación son de suma importancia debido a que si los datos son los apropiados, están completos y no tienen errores de duplicado o tipográficos por mencionar algunos, entonces el conocimiento que se obtenga usando la minería de datos será de mejor calidad, dicho de otra forma, en la fase de extracción y validación de patrones se obtendrá un buen resultado; El término utilizado en informática es GIGO (Garbage-In. Garbage Out) usado para indicar que si la información de entrada es mala (es basura), lo que se obtiene al final también es malo (basura).

2.3 Minería de Datos

La fase de minería de datos es la más significativa de la metodología KDD, debido a esto, en ocasiones, se utiliza el nombre de minería de datos para nombrar a todo el proceso, en esta fase se produce conocimiento que pueda ser utilizado por el usuario; esto es posible generando un modelo inductivo basado en los datos recopilados, de esta manera se consigue una descripción de los patrones y relaciones entre los datos, los cuales podrán ser utilizados para hacer predicciones.

En esta fase se identifica el tipo de tarea que se va a realizar, se elige el modelo y el algoritmo que nos ayudará a realizarla. Debido a la importancia de esta fase para la realización de este trabajo, se dedica el capítulo 3 para su estudio.

2.4 Interpretación y Evaluación

Los patrones obtenidos deben ser precisos, comprensibles, interesantes, el experto evalúa la calidad ya que el objetivo es obtener la mayor precisión y de esta manera los resultados sean útiles al usuario para el cumplimiento de sus objetivos.

La validación de los resultados dependerá de la tarea que se quiere realizar, por ejemplo, en la clasificación, usualmente el conjunto de datos se divide en 2: el conjunto de entrenamiento y el conjunto de prueba. Con el conjunto de entrenamiento se genera el clasificador, y este a su vez se utiliza para clasificar las instancias del conjunto de prueba. Con esto es posible determinar, cuántas instancias se clasificaron correctamente, y en consecuencia, permite determinar el porcentaje de error del clasificador.

2.5 Difusión y Uso

Una vez que el modelo ha sido validado, la información puede tener dos usos: uno es el de aplicar el modelo a nuevos conjuntos de datos, y el otro es el de recomendar acciones o crear estrategias basado en el análisis del modelo.

Capítulo 3. Minería de Datos

Hoy en día, utilizar una computadora para realizar transacciones es lo habitual, por lo que no es de extrañarse que, tanto en agencias gubernamentales como en empresas, escuelas, laboratorios y centros de investigación, entre otros, se generen grandes cantidades de datos.

Debido también a la baja de los costos de almacenamiento, dichos datos generados son guardados para su análisis con el fin de obtener información y conocimiento que ayude a la toma de decisiones. Sin embargo, el análisis puede resultar complejo debido a la cantidad de datos y a lo complejo de las estructuras empleadas al guardarlos. Esto se debe a que cuando se generan los datos y se almacenan, no se piensa concretamente en cuál será la utilidad que se les dará (Fayyad, 1996).

En la minería de datos, contar con datos es equivalente a tener un diamante en bruto, el cual es necesario pulir para obtener su mayor belleza y por tanto su riqueza. Pulir los datos nos lleva a obtener información y conocimiento de ellos.

La minería de datos juega un papel muy importante para encontrar dicha información o conocimiento ya que mediante los algoritmos que esta emplea es posible buscar encontrar información que se encuentra oculta y de esta manera interpretarla, de tal manera que sea de gran utilidad para beneficiar a una organización en la toma de decisiones.

Algunos autores definen la minería de datos como se muestra a continuación:

“La minería de datos se encarga de encontrar modelos inteligibles a partir de los datos, descubrir patrones cuya utilización apoye decisiones que reporten beneficios a la organización” (Hernández, Ramírez & Ferri, 2004).

“La búsqueda de nueva información, valiosa, y no trivial en grandes volúmenes de datos” (Kantardzic, 2011).

“El proceso de descubrir patrones interesantes a partir de cantidades masivas de datos” (Han, Kamber y Pei, 2006).

“La minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos” (Witten & Frank, 2000).

La minería de datos es la aplicación de una metodología mediante técnicas para descubrir conocimiento dentro de una base de datos, éste descubrimiento se puede lograr mediante procesos automáticos o manuales. Se encarga de buscar información nueva, valiosa y no trivial en grandes volúmenes de datos (Kantardzic, 2011).

Así podemos concluir que la minería de datos es un análisis inteligente de grandes bases de datos para encontrar información, potencialmente útil y comprensible, que sea utilizada como apoyo, en las diferentes disciplinas, para la toma de decisiones.

Los modelos generados, al utilizar la minería de datos, pueden ser de dos tipos descriptivos y predictivos. Los modelos descriptivos exploran propiedades de los datos para describir patrones que los explican y de esta manera puedan ser interpretados por humanos. Las reglas de asociación y el agrupamiento son tareas que generan modelos descriptivos. Por otro lado, los modelos predictivos estiman valores futuros o desconocidos a partir del valor de las variables o campos de la base de datos. Dentro de las tareas de minería de datos que producen modelos predictivos están la clasificación y la regresión. Así, dentro de las tareas que se pueden realizar en la minería de datos, se encuentran cuatro principalmente: reglas de asociación, agrupamiento, clasificación y regresión. A continuación, se describirá brevemente en qué consiste cada una, y dado que la tarea en la que se enfoca este trabajo es en la clasificación, esta será explicada con mayor detalle. Sin embargo, dada la importancia de la representación del conocimiento adquirido, primero se describirán 2 de ellos, las reglas y los árboles de decisión.

3.1 Representación del conocimiento

Existen diferentes formas de representar el conocimiento, las dos utilizadas en este trabajo tienen que ver con las reglas y los árboles de decisión.

Reglas

Para representar el conocimiento adquirido, se utilizan las reglas, las cuales tienen dos partes: el antecedente y el consecuente. En el antecedente tenemos las premisas que deben de cumplirse para poder concluir el consecuente.

ANTECEDENTE $\rightarrow$ CONSECUENTE

En el caso de reglas de clasificación, las premisas son condiciones sobre los valores de los atributos que deben cumplirse para que se asigne la clase, la cual se da en el consecuente, las reglas se leen como si condición entonces conclusión.

SI ambiente=soleado y viento=falso ENTONCES jugar

Árbol de decisión

Un árbol de decisión divide todos los registros de una base de datos en conjuntos pequeños de tal forma que los resultados de dicha división cumplan o den resultados veraces a las clases indicadas.

Un árbol de decisión está conformado por nodos que están determinados por los atributos de un objeto, después se derivan las ramas que resultan de las divisiones de los registros y finalmente están las hojas que nos indican las clases.

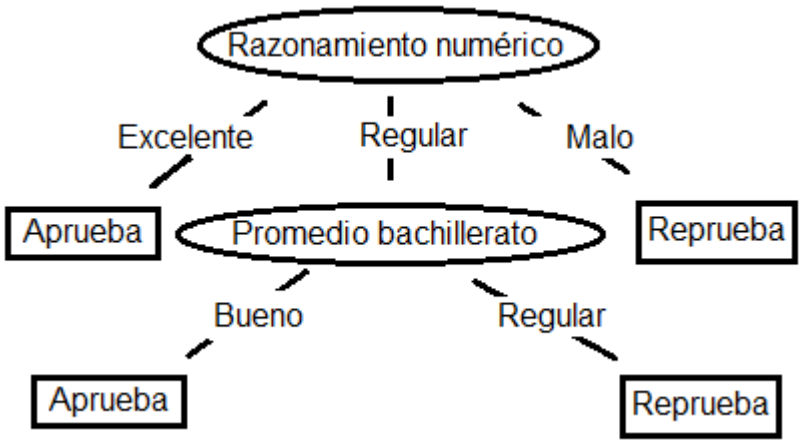


Figura 3.1 Árbol de decisión para la unidad de aprendizaje Álgebra y Geometría Analítica.

En la figura 3.1 se muestra un árbol de decisión para decidir si un alumno aprueba o reprueba la materia de Geometría. La clase solo tiene dos posibles valores: aprueba o reprueba; para decidir la clase a la que pertenece una nueva instancia, se debe recorrer el árbol, iniciando en la raíz y tomando la rama que corresponda al valor del atributo evaluado. En este ejemplo, el atributo en la raíz y primer atributo a evaluar es el razonamiento numérico, cuyo dominio es; Excelente, Regular o Malo; si el alumno tiene un desempeño excelente, podemos notar que directamente se clasifica como que aprueba, por otro lado, si el desempeño es regular (rama de en medio) es necesario evaluar otro atributo, que en este ejemplo es el promedio obtenido en el bachillerato, si es bueno aprobará la materia, pero si es regular reprobará la materia.

Un árbol de decisión tiene implícitamente una serie de reglas que se construyen al recorrer cada una de las ramas generadas y que concluyen con la clase que se especifica en la hoja.

3.2 Reglas de asociación

Las reglas de asociación son una tarea parecida a las correlaciones, identifican relaciones que no son claras entre atributos categóricos, pueden tomar muchas formas, una de ellas es “si el atributo *x* toma el valor *d* entonces el atributo *y* toma el valor *b*”. En las reglas de asociación para que exista alguna relación entre los datos, no necesariamente debe existir una causa, es decir que la relación no implica una relación causa efecto. Esta tarea es usada frecuentemente en el análisis de la cesta de la compra, se emplea para identificar los productos que son comprados de manera conjunta.

Considerando la forma de la regla, explicada anteriormente, en las reglas de asociación no hay una clase, tanto el antecedente como el consecuente puede ser una combinación de valores del dominio de cada atributo.

SI ambiente=soleado y viento=falso ENTONCES temperatura=fría y humedad=alta

En el caso de las reglas estas tienen un porcentaje de precisión dado por la cantidad de instancias predichas correctamente (consecuente correcto) entre la cantidad de instancias a las que se aplica la regla.

SI ambiente=soleado ENTONCES viento=verdadero el 50% de las veces

3.3 Agrupamiento (clustering)

Dentro de las tareas de descripción está el agrupamiento, en esta se crean grupos naturales, a partir de los datos situados en la base de datos. Esta tarea consiste en analizar los datos para generar etiquetas (ver figura 3.2) a diferencia de la clasificación que analiza datos etiquetados.

Consiste en agrupar los datos basándose en el principio de maximizar la similitud de los datos en grupo, en otras palabras, el objetivo es formar grupos de manera

que los datos de un mismo grupo sean muy similares y de la misma manera sean muy diferentes a los datos de otros grupos.

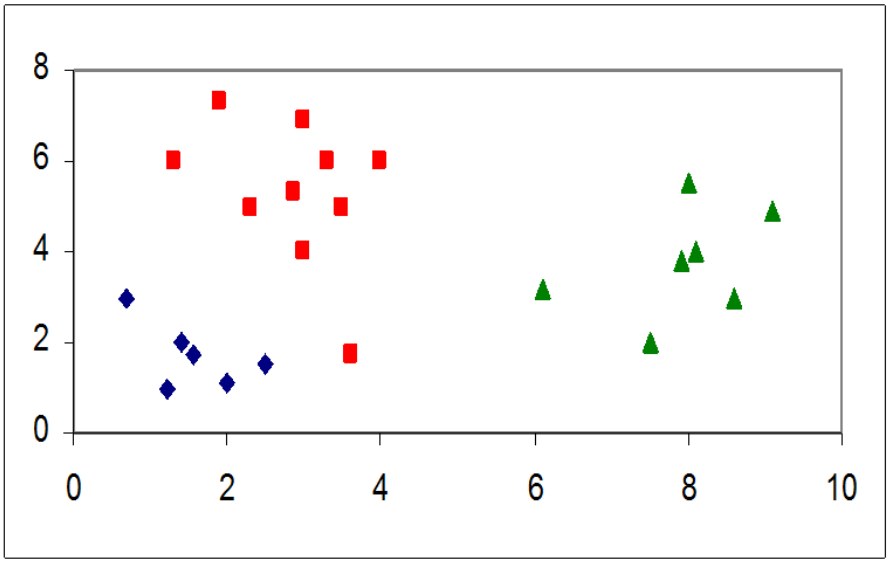


Figura 3.2 Agrupamiento.

3.4 Regresión

En la minería de datos, la regresión es similar a la clasificación, la principal diferencia es que, en lugar de predecir una clase, lo que se predice es un valor numérico. Para crear un modelo de regresión, se asigna un valor numérico a cada instancia, a partir de ahí se construye una fórmula de regresión lineal. Esta fórmula se aplica a nuevas instancias para predecir un valor futuro.

SI ambiente=soleado y humedad=baja ENTONCES temperatura=40

3.5 Clasificación

La tarea de clasificación es una de las más utilizadas, usualmente se utiliza un conjunto de entrenamiento en el cual, las instancias ya tienen asignada la clase. Posteriormente, dependiendo del algoritmo empleado, se genera un modelo que puede considerar solo algunos atributos o todos. Cuando se tiene una nueva instancia, se consideran los valores de sus atributos para determinar la clase a la que pertenece. La evaluación del modelo se mide en términos de la cantidad de instancias que se clasifican correctamente, como proporción del total de instancias evaluadas.

Los algoritmos de clasificación que pueden utilizarse son varios, entre estos, se encuentran, ID3, C4.5, C5, Naive Bayes y CART, los cuales han tenido un buen desempeño en diferentes tareas (Wu & Kumar, 2010).

En el caso del algoritmo ID3 y sus variantes C4.5 (conocido como J48 en la herramienta WEKA) y C5, así como CART, utilizan solo un subconjunto de los atributos para realizar la clasificación, a diferencia de Naive Bayes que los utiliza todos.

También, la manera de representar la salida del algoritmo adquiere relevancia para poder examinar el resultado, de ahí que se prefieran estos algoritmos que capturan el conocimiento en un árbol de decisión. Algunos de los algoritmos que generan árboles de decisión son ID3, C4.5 y CART, están construidos con un enfoque de arriba hacia abajo (top-down), usando la técnica de divide y vencerás. Comienzan con un conjunto de datos, que se va dividiendo y por lo tanto, van reduciendo su tamaño conforme se desarrolla el árbol.

Los algoritmos difieren principalmente, en el método de selección del atributo, el cual especifica un procedimiento heurístico para decidir cuál es el mejor atributo, cuál atributo provee más información para separar las clases. Algunas de las

medidas utilizadas son la entropía o información, la ganancia de la información, la información de la partición y el índice Gini (Han, Kamber & Pei, 2011).

Algoritmo ID3

El algoritmo ID3 fue desarrollado por J. Ross Quinlan entre la década de 1980 y principios de 1990. Utiliza la estrategia de "divide y vencerás" generando árboles de decisión que se construyen de arriba hacia abajo partiendo de un nodo raíz y generando nodos intermedios hasta llegar a las hojas (Hernández, Ramírez & Ferri, 2004).

Este algoritmo es supervisado, lo que significa que para generar el clasificador utiliza un conjunto de entrenamiento, en el cual las instancias ya tienen la clase asignada, con el procedimiento que se mostrará a continuación, se realiza una selección de los atributos hasta que la clase esté bien definida. Es posible que algunos atributos no aparezcan en el clasificador final.

Para la elección de los atributos se utiliza la entropía de Shannon (Wu & Kumar 2010), la cual es una medida de la cantidad de información que se requiere para decidir la clase de la instancia (ver Ecuación 1).

$$Info(D) = -\sum_{i=1}^{m} \frac{n_i}{n} \log_2\left(\frac{n_i}{n}\right) \quad (1)$$

donde:

D es el sistema analizado

m es el número de clases

n_i es el número de instancias con la clase i

n es el número total de instancias

La función logarítmica base 2 se utiliza debido a que la información está codificada en bits. Específica el número mínimo de bits de información necesarios para determinar la clasificación de un elemento arbitrario de D.

Para cada valor del dominio de un atributo, se utiliza la ecuación 1 para calcular la información de ese valor. Posteriormente, es posible calcular la entropía del atributo, ver ecuación 2:

$$Info_A\,(D) = \sum_{j=1}^{v} \frac{Dj}{D}\; x\; Info(Dj) \qquad (2)$$

donde:

A es un atributo del conjunto D

v son los valores del dominio del atributo A

D_j subconjnto de instancias $j-$ ésimo valor del atributo A

$Info(D_j)$ es la entropía del subconjunto de instancias $j-$ ésimo valor del atributo A

Este procedimiento se realiza con cada uno de los atributos que se tienen, incluyendo también el no considerar ningún atributo, lo que se conoce como información del sistema.

Después, se calcula la ganancia de la información mediante la ecuación 3.

$$Gain(A) = Info(D) - Info_A(D) \qquad (3)$$

donde:

$Info\,(D) =$ Entropia del conjunto D ó del sistema

$Info\,_A(D) =$ Entropia del atributo A

Es posible entonces seleccionar al mejor atributo, siendo este el que proporcione la mayor ganancia de información, con esto el conjunto original de instancias se divide de acuerdo con los valores del dominio del atributo ganador.

El proceso se realiza iterativamente, con cada uno de los nuevos conjuntos generados. A manera de ilustrar el proceso se presenta el siguiente ejemplo.

Ejemplo:

Los datos son los resultados obtenidos en las áreas evaluadas por el EXANI-II:

- razonamiento verbal (RV),
- razonamiento numérico (RN),
- mundo contemporáneo (MC),
- ciencias naturales (CN),
- ciencias sociales (CS),
- matemáticas (MAT) y
- español (ESP).

El dominio de los atributos es:

- Bajo (BA),
- Bien (BI) y
- Excelente (EX).

Por otro lado, la clase a aprender corresponde a dos carreras de ingeniería: ingeniería en sistemas y comunicaciones (LSC) e ingeniería industrial (IIS). Una muestra de los datos se presenta en la tabla 3.1.

Para comenzar la construcción del árbol de decisión se debe determinar el atributo raíz, para ello se tiene que calcular la información del sistema, sin considerar a ningún atributo, posteriormente la información de cada atributo y su ganancia. El atributo ganador será el nodo raíz.

Entropía del sistema: 48 instancias de la clase IIN y 95 de la clase LSC, total 143

$$-\left(\frac{48}{143}\right)\ \log 2\left(\frac{48}{143}\right) - \left(\frac{95}{143}\right)\log 2\left(\frac{95}{143}\right) = 0.9205$$

Tabla 3.1 Datos de los resultados del EXANI-II asociados a las carreras IIN, y LSC.

RV	RN	MC	CN	CS	MAT	ESP	CLASE
BIEN	BIEN	EXCELENTE	EXCELENTE	BAJO	BAJO	BAJO	LSC
EXCELENTE	EXCELENTE	BAJO	BIEN	BIEN	EXCELENTE	EXCELENTE	IIN
BIEN	BIEN	BAJO	BAJO	BIEN	BIEN	BIEN	IIN
BAJO	BIEN	BIEN	BAJO	BAJO	BAJO	BAJO	LSC
BIEN	EXCELENTE	BAJO	BIEN	EXCELENTE	EXCELENTE	BIEN	IIN
BIEN	EXCELENTE	EXCELENTE	EXCELENTE	BIEN	BIEN	BAJO	LSC
BAJO	BIEN	BIEN	BAJO	EXCELENTE	BIEN	BAJO	IIN
BIEN	BIEN	EXCELENTE	EXCELENTE	BIEN	BIEN	BAJO	LSC
BIEN	BIEN	EXCELENTE	BAJO	BAJO	BAJO	BAJO	LSC
BAJO	BAJO	BAJO	BIEN	BIEN	BAJO	BAJO	LSC
BIEN	EXCELENTE	BIEN	BIEN	EXCELENTE	EXCELENTE	EXCELENTE	IIN
BAJO	BAJO	BAJO	BAJO	BIEN	BAJO	BAJO	IIN
BIEN	BIEN	BIEN	BAJO	BIEN	BIEN	BIEN	IIN
BIEN	BIEN	BIEN	BIEN	BIEN	BIEN	BAJO	LSC
BIEN	BIEN	BIEN	BAJO	BIEN	BIEN	BAJO	IIN
BAJO	EXCELENTE	BIEN	BIEN	BIEN	BIEN	BIEN	LSC
BIEN	EXCELENTE	BIEN	BIEN	BIEN	BAJO	BAJO	LSC
BIEN	BIEN	BIEN	BAJO	BAJO	BAJO	BAJO	LSC
BAJO	BIEN	BIEN	BAJO	BAJO	BAJO	BAJO	LSC

Se procede a calcular la información de cada atributo, iniciando con el razonamiento verbal, cuyos datos se muestran en la tabla 3.2. Los cálculos se presentan a continuación.

Tabla 3.2 Datos sumarizados correspondientes al razonamiento verbal.

Atributo	Razonamiento Verbal					
Valor	BAJO		BIEN		EXCELENTE	
Clase	IIN	LSC	IIN	LSC	IIN	LSC
subtotal	11	33	34	62	3	0
total	44		96		3	

Razonamiento Verbal, Valor Bajo

$-\left(\frac{11}{44}\right) log_2 \left(\frac{11}{44}\right) - \left(\frac{33}{44}\right) log_2 \left(\frac{33}{44}\right)$= 0.5 + 0.3112 = 0.8112

Razonamiento Verbal, Valor Bien

$-\left(\frac{34}{96}\right) log_2 \left(\frac{34}{96}\right) - \left(\frac{62}{96}\right) log_2 \left(\frac{62}{96}\right)$ = 0.5303 + 0.4073 = 0.9376

Razonamiento Verbal, Valor Excelente

$-\left(\frac{3}{3}\right) log_2 \left(\frac{3}{3}\right) - \left(\frac{0}{3}\right) log_2 \left(\frac{0}{3}\right)$ = 0

Entropía de Razonamiento Verbal

$\left(\frac{44}{143}\right)(0.8112) + \left(\frac{96}{143}\right)(0.9376) + \left(\frac{3}{143}\right)(0)$ = 0.8790

Ganancia del atributo razonamiento verbal

$0.9205 - 0.8790 = 0.0415$

Se presentan en la tabla 3.3, los resultados de cada uno de los atributos restantes.

Tabla 3.3 Distribución de los valores del dominio de cada atributo por clase.

Valor del Dominio	BAJO			BIEN			EXCELENTE		
CLASE	IIN	LSC	Total	IIN	LSC	Total	IIN	LSC	Total
Razonamiento Verbal	11	33	44	34	62	96	3	0	3
Razonamiento Numérico	3	10	13	31	61	92	14	24	38
Mundo Contemporáneo	20	7	27	28	52	80	0	36	36
Ciencias Naturales	32	16	48	15	53	68	1	26	27
Ciencias Sociales	10	43	53	30	52	82	8	0	8
Matemáticas	12	60	72	21	35	56	15	0	15
Español	10	84	94	33	11	44	5	0	5

En la tabla 3.4 se presentan los valores de información calculados para cada valor del dominio y para el atributo, además de la ganancia que proporciona el utilizar ese atributo. En ella se puede observar que el atributo que proporciona la mayor ganancia es español, que se encuentra en el último renglón de la tabla.

El atributo español, será la raíz del árbol y el conjunto de datos, las 143 instancias se dividen en 3 ramas, con 94, 44 y 5 instancias. Dado que las últimas instancias son de una sola clase, esa parte del árbol ya está terminada y solo resta aplicar de nuevo el procedimiento para las otras dos ramas (ver figura 3.2).

Tabla 3.4 Información y Ganancia de cada atributo.

Atributo	Información de cada valor del dominio			Información del Atributo	Ganancia del atributo
	BAJO	BIEN	EXCELENTE		
Razonamiento Verbal	0.8112	0.9376	0	0.879	0.0415
Razonamiento Numérico	0.7792	0.9219	0.9494	0.9162	0.0043
Mundo Contemporáneo	0.8255	0.934	0	0.6783	0.2422
Ciencias Naturales	9183	0.7611	0.2284	0.718	0.2025
Ciencias Sociales	0.6985	0.9474	0	0.7972	0.1233
Matemáticas	0.6499	0.9543	0	0.7009	0.2196
Español	0.4887	0.8112	0	0.5705	0.35

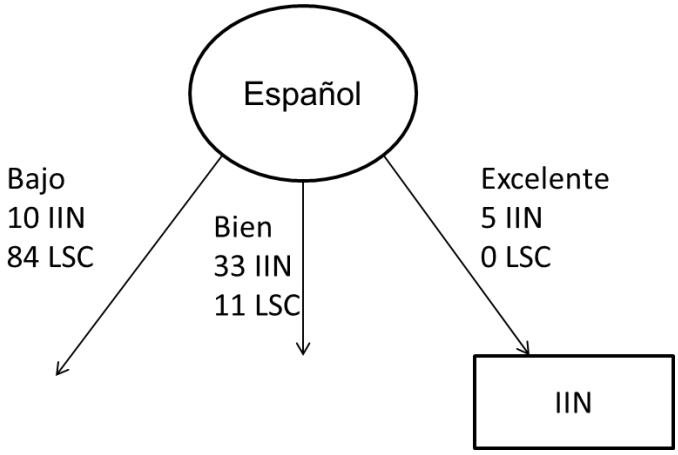


Figura 3.2 Árbol con el atributo ganador y en la rama Excelente, la clase ya queda definida.

Como se puede observar, en el proceso anterior, los datos son utilizados para determinar el atributo que será evaluado en la raíz del árbol, y los datos se dividen de acuerdo con el número de valores en el dominio del atributo un atributo es seleccionado para dividir los datos.

Se crea una rama para cada valor del atributo, el correspondiente subconjunto de la muestra que tiene el valor del atributo especificado por la rama se mueve al nodo hijo recién creado. El algoritmo se aplica a cada nodo hasta que todas las instancias en el nodo sean de una clase. El árbol final se presenta en la figura 3.3.

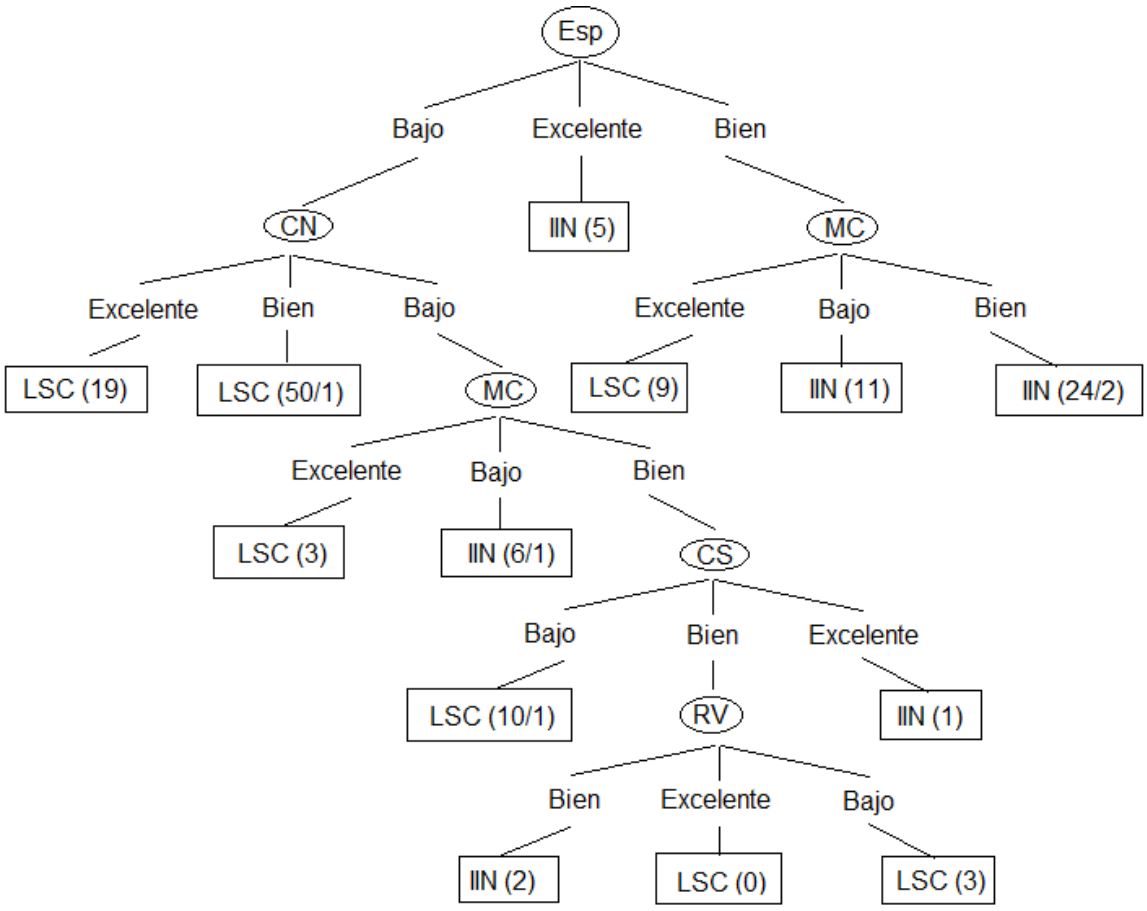


Figura 3.3 Árbol de decisión generado.

Si se recorre el árbol de decisión, desde la raíz hasta cada una de las hojas, se tienen varias rutas, cada ruta representa una regla de clasificación. Las 13 reglas que se generan a partir del árbol en la figura 3.3 se presentan a continuación.

Reglas de clasificación generadas:

1. Si (Español = Bajo) y (Ciencias Naturales = Excelente) entonces “Ingeniero en Sistemas y Comunicaciones”.
2. Si (Español = Bajo) y (Ciencias Naturales = Bien) entonces “Ingeniero en Sistemas y Comunicaciones”.
3. Si (Español = Bajo) y (Ciencias Naturales= Bajo) y (Mundo Contemporáneo = Excelente) entonces “Ingeniero en Sistemas y Comunicaciones).
4. Si (Español = Bajo) y (Ciencias Naturales = Bajo) y (Mundo Contemporáneo = Bajo) entonces “Ingeniero Industrial”.
5. Si (Español = Bajo) y Ciencias Naturales = bajo) y (Mundo Contemporáneo = Bien) y (Ciencias Sociales = Bajo) entonces “Ingeniero en Sistemas y Comunicaciones).
6. Si (Español = Bajo) y Ciencias Naturales = bajo) y (Mundo Contemporáneo = Bien) y (Ciencias Sociales = Bien) y (Razonamiento Verbal = Bien) entonces “Ingeniero Industrial”.
7. Si (Español = Bajo) y Ciencias Naturales = bajo) y (Mundo Contemporáneo = Bien) y (Ciencias Sociales = Bien) y (Razonamiento Verbal = Excelente) Entonces “Ingeniero en Sistemas y Comunicaciones”.
8. Si (Español = Bajo) y Ciencias Naturales = bajo) y (Mundo Contemporáneo = Bien) y (Ciencias Sociales = Bien) y (Razonamiento Verbal = Bajo= Entonces “Ingeniero en Sistemas y Comunicaciones”.
9. Si (Español = Bajo) y Ciencias Naturales = bajo) y (Mundo Contemporáneo = Bien) y (Ciencias Sociales = Excelente) Entonces “Ingeniero Industrial”
10. Si (español = Excelente) Entonces “ingeniero Industrial”.
11. Si (Español = Bajo) y (Mundo Contemporáneo = Excelente) Entonces “Ingeniero en Sistemas y Comunicaciones”.
12. Si (Español = Bajo) y (Mundo Contemporáneo = Bajo) Entonces “Ingeniero Industrial”.
13. Si (Español = Bajo) y (Mundo Contemporáneo = Bien) Entonces “Ingeniero Industrial”.

Las ventajas de este algoritmo son que hace una clasificación utilizando siempre los atributos que proveen la mayor información para la separación de las clases, disminuyendo con esto la cantidad de atributos que se tienen que evaluar para determinar la clase. Por otro lado, debido a la forma en que se construye, ningún atributo es evaluado dos veces en la misma rama, y en el proceso, la estructura que se forma puede ser analizada.

Dentro de las desventajas que tiene el algoritmo ID3 están que, puede favorecer a los atributos cuyo dominio sea más grande; también debido a que no se utiliza la poda del árbol, los árboles que se pueden generar son muy extensos y las reglas exhaustivas. Por otro lado, ID3 no está diseñado para trabajar cuando los valores del dominio de un atributo son numéricos.

Algoritmo C4.5

Una extensión del algoritmo ID3 es su sucesor, el algoritmo C4.5, desarrollado también por J. Ross Quinlan, que extiende el dominio de la clasificación de los atributos categóricos a los numéricos. El algoritmo básicamente elige el atributo que proporciona el máximo grado de discriminación entre las clases a nivel local (Wu & Kumar, 2010).

El Algoritmo C4.5 funciona, en esencia, de la misma manera que el algoritmo ID3, las principales diferencias que constituyen ventajas para C4.5 son, el poder trabajar con datos numéricos además de los categóricos, el considerar la información de la partición para no solo emplear la ganancia, sino el radio de ganancia, y el realizar la poda del árbol, la cual tiene como función eliminar ramas del árbol que no incide en los resultados para obtener un nodo final, debido a este proceso se obtiene un árbol más pequeño y con mayor precisión.

Ponemos nuevamente el ejemplo del algoritmo ID3 para saber en qué carrera se tendrá mejor desempeño ahora bajo el algoritmo C4.5 utilizando valores numéricos en los atributos que constituyen el sistema.

Tabla 3.5 ejemplo del algoritmo ID3.

RV	RN	MC	CN	CS	MAT	ESP	CARRERA
3	23	8	17	2	2	3	LSC
5	25	3	11	4	4	7	IIN
4	23	3	7	3	3	6	IIN
1	15	5	6	0	1	2	LSC
4	25	1	13	5	5	4	IIN
4	27	8	17	3	3	3	LSC
2	21	4	6	5	3	3	IIN
3	17	7	19	3	3	3	LSC
3	23	7	8	1	2	3	LSC
1	11	3	12	3	2	2	LSC
4	29	6	11	6	5	8	IIN
2	8	1	5	3	1	3	IIN
3	23	5	8	3	3	6	IIN
3	19	6	12	3	3	3	LSC
4	19	4	6	4	3	3	IIN
2	27	6	14	3	4	4	LSC
3	25	6	16	3	1	2	LSC
3	19	4	8	2	2	2	LSC
1	15	4	6	1	1	1	LSC

Las ventajas de este algoritmo son que genera árboles más pequeños y como consecuencia sus resultados son más fáciles de entender, por otro lado al realizar la poda, se reduce el error que induce la rama cortada. Además de que se pueden utilizar atributos con valores categóricos y nominales.

A continuación, se presenta la poda de árbol en la figura 3.4.

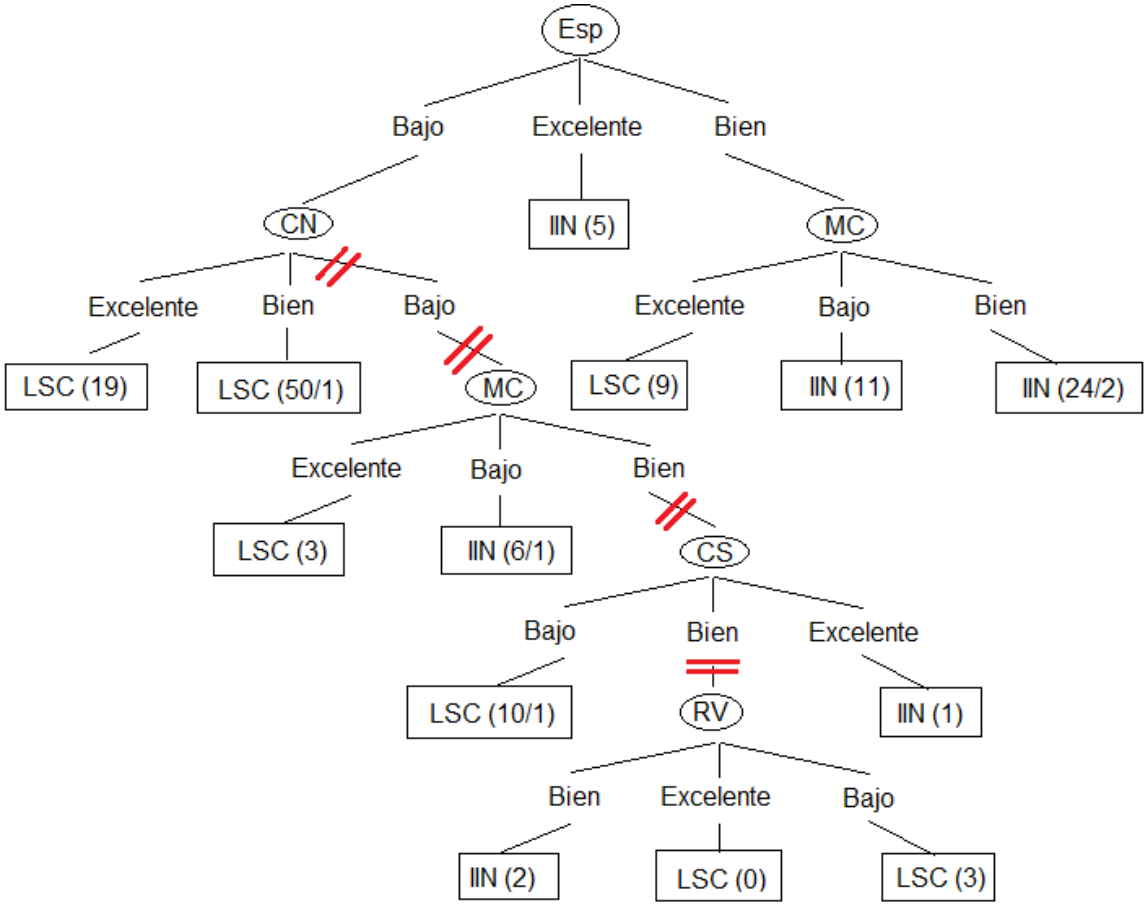


Figura 3.4 Poda del árbol.

En este trabajo se utilizó la implementación del algoritmo C4.5 del software WEKA, que se describe a continuación, llamado J48. La ventaja del algoritmo es que genera un árbol de decisión que puede ser examinado e interpretado, donde los nodos son los atributos elegidos, las ramas son los posibles valores del atributo y las hojas son las clases [Hernández, 2004].

3.6 WEKA – Waikato Environment for Knowledge Analysis

Existen diferentes herramientas para realizar minería de datos, entre ellas se encuentran RapidMiner, R, Clementine y Weka, estas incluyen algoritmos de clasificación, predicción, agrupamiento y reglas de asociación. En este trabajo en particular, se utilizó la herramienta Weka.

Weka es un software libre de minería de datos, gracias a su licencia publica General de GNU se desarrolló en la universidad de Waikato, Nueva Zelanda, el logo de Weka se puede observar en la figura 3.5.

Weka incluye una colección de algoritmos de aprendizaje en técnicas de Minería de Datos, estos algoritmos están programados en código java, los cuales pueden ser modificados y compartidos de manera libre, dichos algoritmos se pueden aplicar directamente a un conjunto de datos ya sea de tipo nominal o numérico, las herramientas que contiene Weka para hacer la tarea de minería de datos son: clasificación, regresión, agrupamiento, reglas de asociación, y la visualización de resultados, ya que dependiendo de la tarea que se vaya a realizar se debe seleccionar la herramienta adecuada (Weka, 2018).

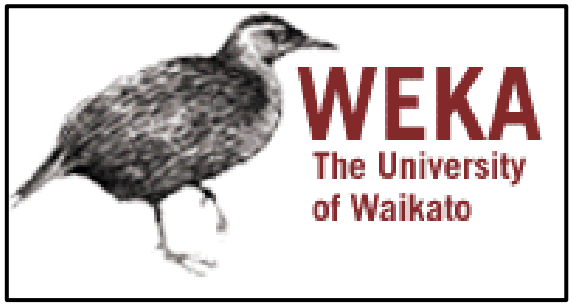


Figura 3.5 Logo Oficial de WEKA.

Para proporcionar los datos a la herramienta, los archivos deben estar en formato ARFF o bien en formato CSV.

Archivos ARFF

El formato de los archivos ARFF se muestra en la figura 3.6, el archivo se divide en 2 partes: la información de la relación y sus atributos, y los datos.

Respecto a la información se coloca el nombre de los datos precedido de @relation, y para cada atributo se coloca la palabra @attribute. En caso de que el atributo sea numérico, se indica la palabra numeric, de lo contrario, se indican los valores del dominio del atributo entre llaves. El último atributo que se coloca es la clase que se quiere aprender.

Una vez descrita la configuración de los datos, se separa esta sección de los datos en sí con la palabra @data. Los datos de cada una de las instancias, se colocan en el orden en que los atributos se especificaron.

El símbolo de porcentaje se utiliza para poner comentarios.

```
%ARFF file for the weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
%
% 14 instances
%
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Figura 3.6 Formato ARFF (Witten y Frank 2000).

A continuación, se muestra el proceso para cargar el archivo que será analizado por la herramienta Weka, así como le selección del algoritmo que se empleará y como se muestran los resultados.

Cargar datos en el sistema

Para abrir el archivo que tenemos guardado tenemos que abrir el explorador de archivos seleccionando la opción open file y de esta manera se puede seleccionar la ruta de ubicación en donde se encuentre guardado, ver figura 3.7.

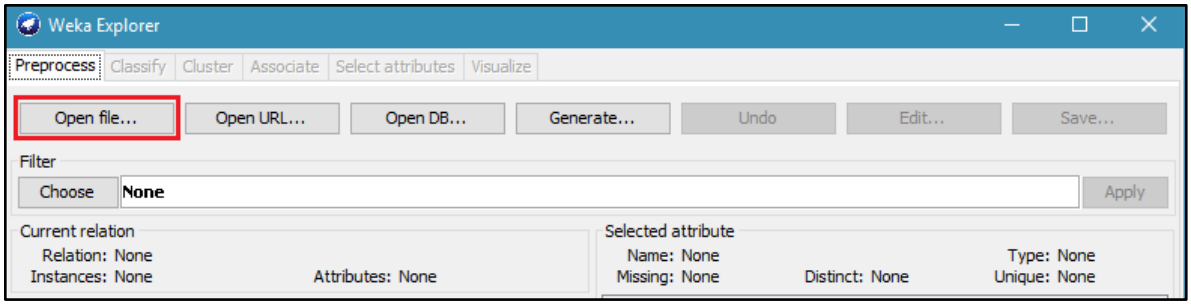


Figura 3.6 Abrir archivo en Weka.

Por omisión, se presentan los archivos con extensión ARFF, sin embargo, también puede leer archivos separados por comas (CSV).

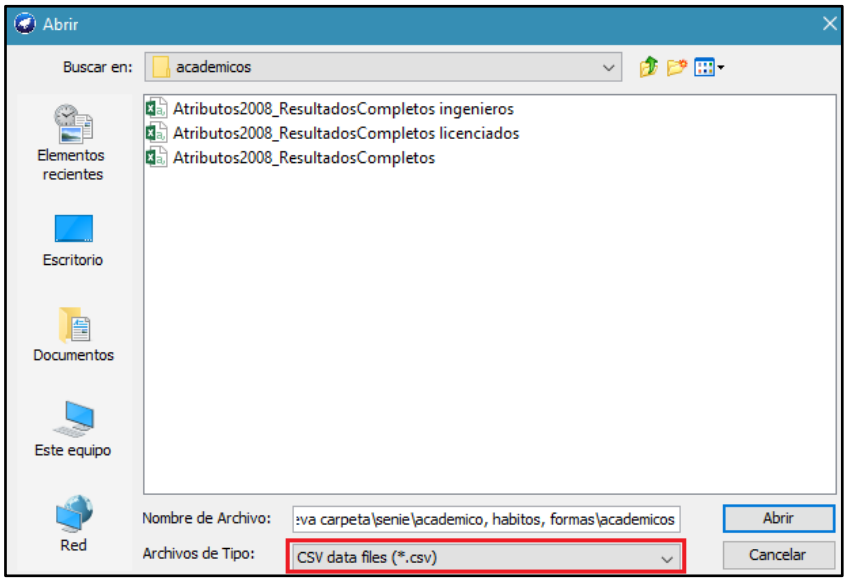


Figura 3.7 Seleccionar archivo en Weka.

Una vez que se carga el archivo, la herramienta muestra todos los atributos, asi como la clase y el tipo de dato, a continuación se selecciona la opción classify para aplicar las configuraciones en Weka, ver figura 3.9.

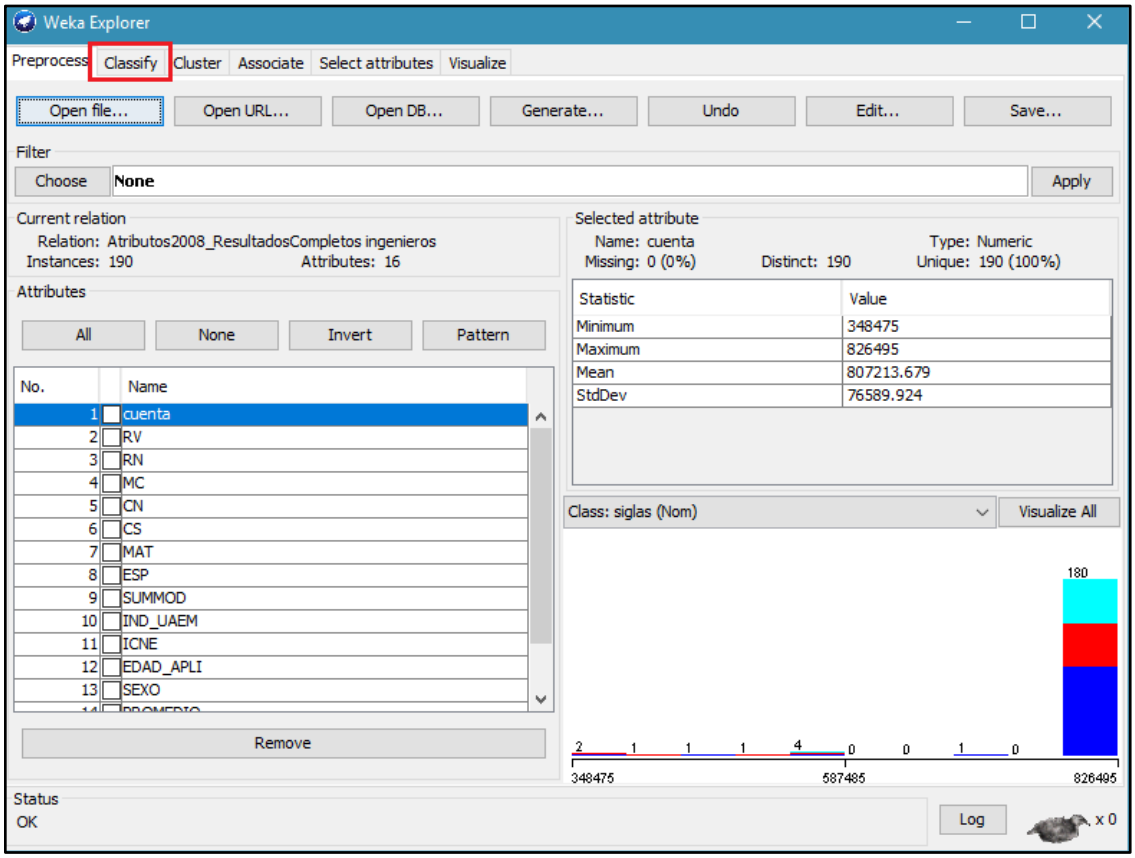


Figura 3.8 Configuraciones en Weka.

Se selecciona la opción Choose para elegir el algoritmo que se va a utilizar en la clasificación, ver figura 3.10.

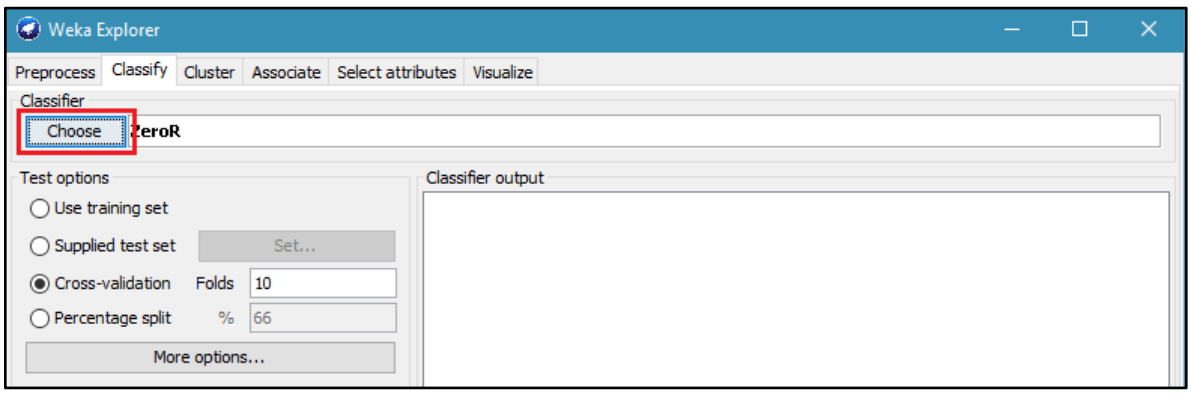


Figura 3.9 Algoritmos en Weka.

La herramienta Weka contiene diversos algoritmos que se pueden emplear dependiendo del caso de estudio, para este caso particular se selecciona el algoritmo J48, que se encuentra dentro de los algoritmos de árboles, ver figura 3.11.

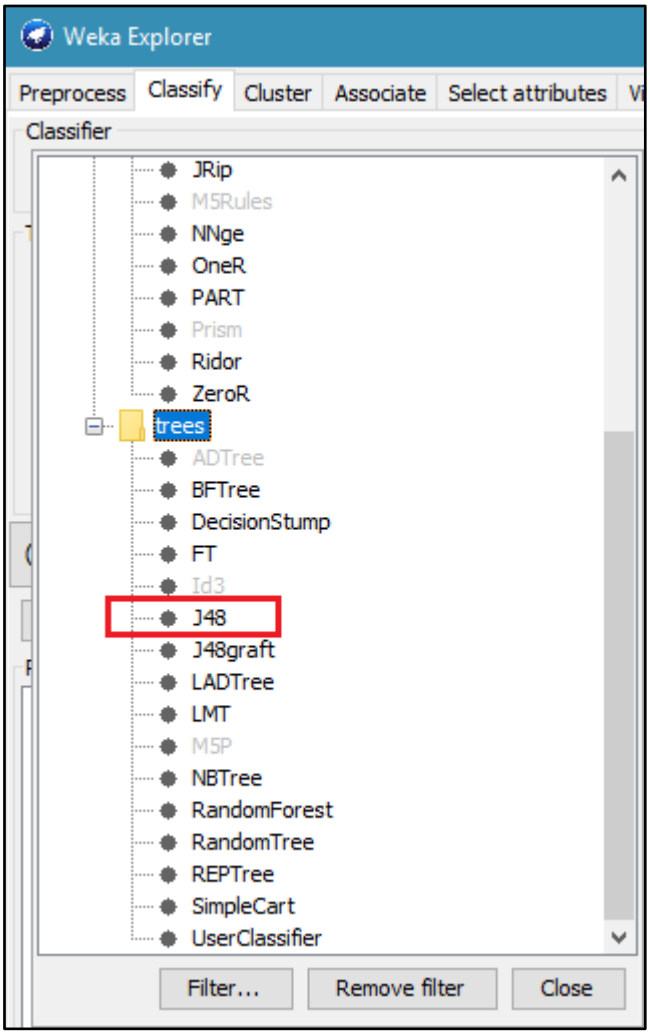


Figura 3.10 Algoritmo J48 en Weka.

Dentro de Weka también podemos configurar el minObj, esta opción controla el número mínimo de elementos en una hoja del árbol, entre más grande sea el parámetro de minObj el árbol generado será más simple y pequeño lo que conlleva a que no aparezcan todas las instancias, entre más pequeño sea el minobj el árbol es más extenso, logrando un resultado más preciso y complejo de analizar, casi generando una regla para cada instancia, ver figura 3.12.

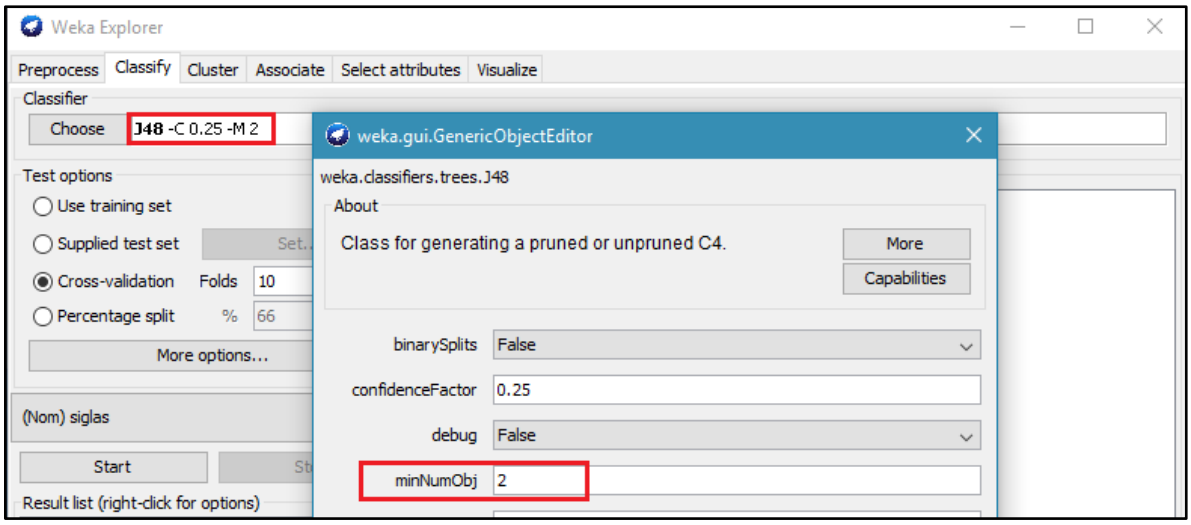


Figura 3.11 Configuración de minObj en Weka.

Una vez configurado Weka, al haber seleccionado el algoritmo requerido y el minObj, se activa la opción Start para que el algoritmo se ejecute sobre el archivo seleccionado, ver figura 3.13.

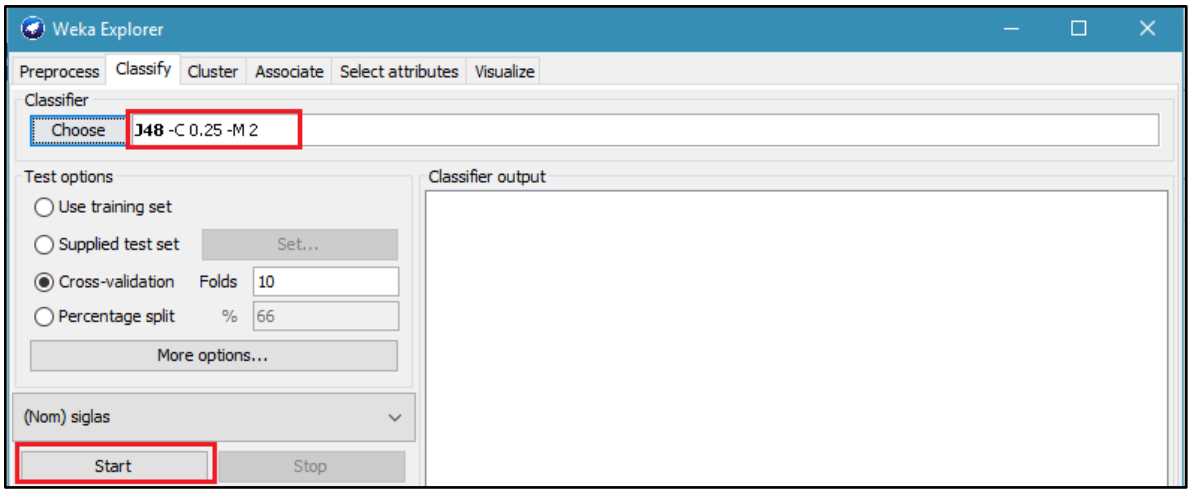


Figura 3.12 Ejecutar algoritmo en Weka.

El algoritmo analiza el archivo seleccionado y muestra los resultados de la clasificación dando a conocer el porcentaje de error y de acierto, así como el número de instancias clasificadas correcta o incorrectamente, ver figura 3.14.

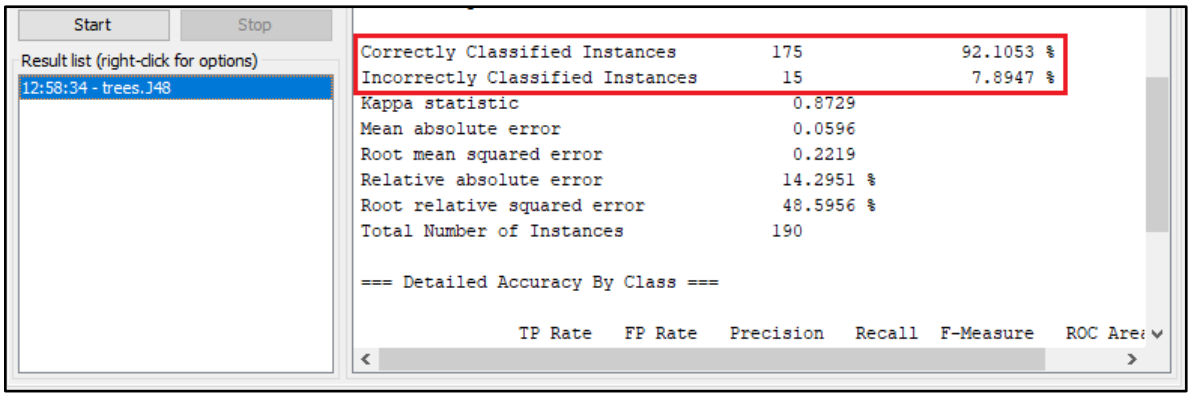


Figura 3.13 Resultados del algoritmo en Weka.

Dentro de Weka se puede consultar el árbol de decisión en forma de texto, que muestra el número de nodos y el tamaño del árbol generado por el algoritmo, ver figura 3.15.

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

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

MAT <= 4
|   MC <= 4
|   |   ESP <= 3
|   |   |   CN <= 7
|   |   |   |   MAT <= 1
|   |   |   |   |   MC <= 2: IIN (2.0)
|   |   |   |   |   MC > 2: LSC (2.0)
|   |   |   |   MAT > 1: IIN (6.0)
|   |   |   CN > 7
|   |   |   |   MAT <= 2: LSC (13.0)
|   |   |   |   MAT > 2: ICO (6.0/1.0)
|   |   ESP > 3: IIN (25.0)
|   MC > 4: LSC (80.0/1.0)
MAT > 4
|   RN <= 17: ICO (43.0/1.0)
|   RN > 17: IIN (13.0)

Number of Leaves  :     9

Size of the tree :      17
```

Figura 3.14 Texto del árbol de decisión.

Como parte del resultado, se presenta la matriz de confusión, la cual muestra como fueron clasificadas las instancias de cada clase, ver figura 3.16.

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 89  6  0 |  a = LSC
  4 40  4 |  b = IIN
  3  0 44 |  c = ICO
```

Figura 3.15 Matriz de confusión.

Para poder visualizar el árbol resultado de la clasificación se da clic derecho en el algoritmo y se selecciona la opción Visualize tree, ver figura 3.17.

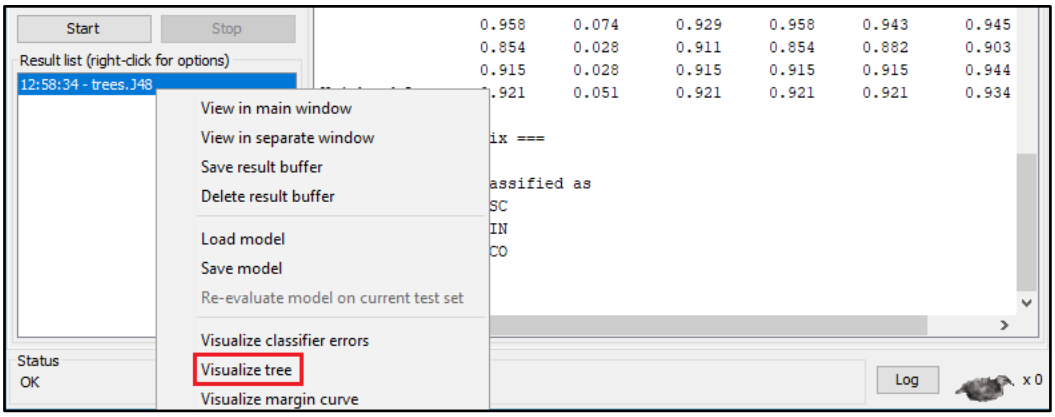


Figura 3.16 visualizar el árbol en Weka.

Weka genera el resultado visual mediante un árbol, ver figura 3.18.

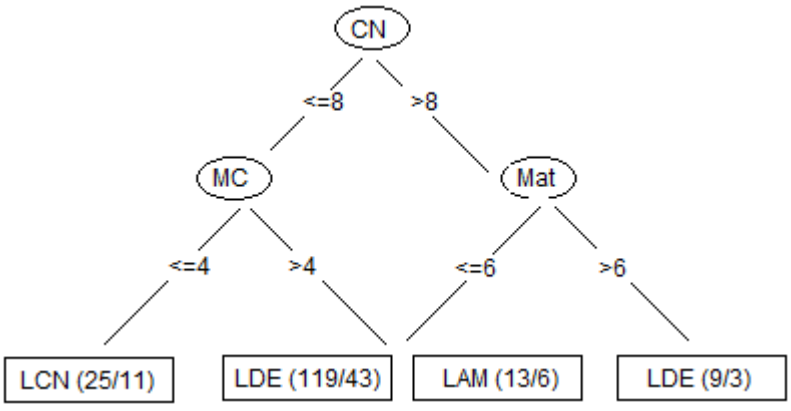


Figura 3.17 Árbol generado en Weka.

3.7 Estado del arte

El análisis de la base de datos de los alumnos que concluyeron sus estudios universitarios ha sido aplicado por diversas instituciones con el fin de evitar una mala decisión por parte del estudiante al momento de elegir una carrera. Los algoritmos o técnicas de minería de datos utilizados, entre otros, han sido los de clasificación, regresión lineal, agrupamiento y reglas de asociación. A continuación, se presentan algunos de estos trabajos.

En (Hernández y Quintana 2013) se utilizaron árboles de decisión para determinar qué inteligencia múltiple era la que proporcionaba más información en la separación de las clases, generaron un modelo clasificador usando el algoritmo C4.5 para determinar el desempeño de un estudiante en una unidad de aprendizaje en particular basándose en las inteligencias múltiples de individuo. Utilizaron el software Weka, y obtuvieron con el modelo generado un porcentaje de acierto de 89% y un 11% de error.

En (Winston y Lawrence 2008) realizaron un sistema experto basado en la prueba de personalidad de Myers-Briggs para determinar en qué carrera un estudiante se desempeñaría mejor de acuerdo con su personalidad, la interfaz se generó mediante Visual Studio y MS Access, el sistema se probó en un grupo de 104 profesionales, los cuales el 71% coincidían con su tipo de personalidad.

En (Hendahewa, Dissanayake, Samaraweera, Wijayawickrama, Ruwanpathirana, y Karunananda, 2006) utilizaron el método basado en reglas, conocido como sistema de producción para identificar los factores que deben cumplir los estudiantes al momento de cursar una trayectoria académica profesional tomando en cuenta su rendimiento del examen, las preferencias de los estudiantes y habilidades, se descubrió que el modelo tiene una capacidad de aproximadamente 70% de precisión para predecir el rendimiento.

En (Mundra, Soni, Sharma, Kumar, y Chauhan, 2014) se analizaron las habilidades de un grupo de estudiantes, así como su tipo de personalidad mediante el modelo tipo Myers Briggs y se determinó qué habilidades académicas y habilidades personales debe cumplir un estudiante para cursar una carrera particular mediante el método estadístico de aprendizaje basado en instancias, el algoritmo de aprendizaje perezoso basado en instancias que requiere menos tiempo de cálculo (lazy learning algorithm).

En (Ezenkwu, Johnson y jerame 2017) se utilizó una técnica basada en casos que razona a partir de la experiencia, se analizaron 1000 casos diferentes con 10 funciones y una salida, contemplando a su vez las materias de las trayectorias académicas, para elegir las más apropiadas para que el estudiante tenga un buen desempeño, se utilizaron 800 casos del conjunto de datos para la base de casos, se desarrolló una aplicación MATLAB basada en la GUI para recomendar carreras a nuevos estudiantes, 200 casos se usaron para la prueba, con la distancia euclidiana, como la métrica de similitud, y 10 como el parámetro k, el algoritmo del sistema logró un error de clasificación del 0% en el caso de prueba.

En (Ayman y Ahmar 2012) se usaron reglas de prototipos con una base de datos orientada a objetos, en donde cada especialidad universitaria es un objeto, con el fin de crear una herramienta que apoye en la decisión de una especialidad. Se compararon los valores de las habilidades y preferencias de los estudiantes con los valores de las habilidades y atributos requeridos; una habilidad es suficiente si su valor es igual o mayor que su valor para el mayor, de lo contrario, la habilidad no es suficiente si la preferencia del estudiante a una especialidad es satisfactoria. Para asignar un valor a una habilidad o atributo se hizo una prueba para comparar los valores principales con los valores de los estudiantes, asignaron el valor de satisfactorio, insatisfactorio o desconocido (es desconocido cuando no se tiene respuesta) los autores no mencionan porcentajes de error o acierto sobre los resultados, sin embargo, mencionan que son satisfactorios.

En (Alao, Ibam 2017) se utilizó el algoritmo de encadenamiento directo para el análisis de las asignaturas, cual es el interés profesional, y el cociente de inteligencia el cual evalúa sobre razonamiento verbal, razonamiento numérico y lógica para la recomendación de una carrera, el sistema se implementó y evaluó utilizando 200 estudiantes, los resultados muestran que la carrera recomendada por el sistema es 95% exacto y relevante, 70% satisfactorio y 80% adecuado para la información sobre orientación profesional por parte del sistema.

En (Aquino, Jara 2016) se implementan técnicas de inteligencia artificial para poder generar recomendaciones vocacionales, las muestras se dividieron en dos grupos 60% de entrenamiento de la red y 40% de prueba, El modelo de RNA utilizado fue el Perceptrón, (MLP o Multilayer Perceptron), esta red se implementó mediante cálculos, la red neuronal que obtuvo mejor resultado fue aquella cuyos datos de entrada fueron los resultados de un solo test; se hizo una comparación de los resultados de salida con las recomendaciones que dio el especialista obteniendo un total de 25 (80.6%) de coincidencia y un 6 (19.4%) que no coincidió.

Una de las principales diferencias de los trabajos antes mencionados en relación al trabajo que se propone es que en este se busca determinar qué es lo que influye en cada una de las carreras, identificando los hábitos, conocimientos, actividades formas de aprendizaje que las caracterizan, mediante el uso de diferentes algoritmos de minería de datos que representan un conocimiento explicativo en un árbol de decisión, para determinar los factores académicos y educativos aspectos medibles cuantitativamente y cualitativamente de un estudiante y así aplicarlo a nuevos estudiantes para crear estrategias y tomar acciones oportunas.

A continuación, se presentan el proceso de extracción del conocimiento para generar el clasificador de estudiantes en una carrera particular, basándose en los resultados del EXANI y el estudio socioeconómico qué estos realizan.

Capítulo 4. Preparación de los datos y diseño de los experimentos

La metodología empleada consistió en recopilar la información de los estudiantes de la generación 2008-2013, realizar la minería de datos con cada uno de los factores propuestos, realizar un análisis para determinar cuáles son los más relevantes para la tarea, considerando la precisión del algoritmo de minería utilizado y la representación del conocimiento adquirido.

Como se mencionó, la metodología utilizada está basada en el proceso de extracción del conocimiento, mejor conocido, por sus siglas en inglés, como KDD (Knowledge Discovery in Databases), el cual se muestra, considerando los datos que son empleados en la figura 4.1.

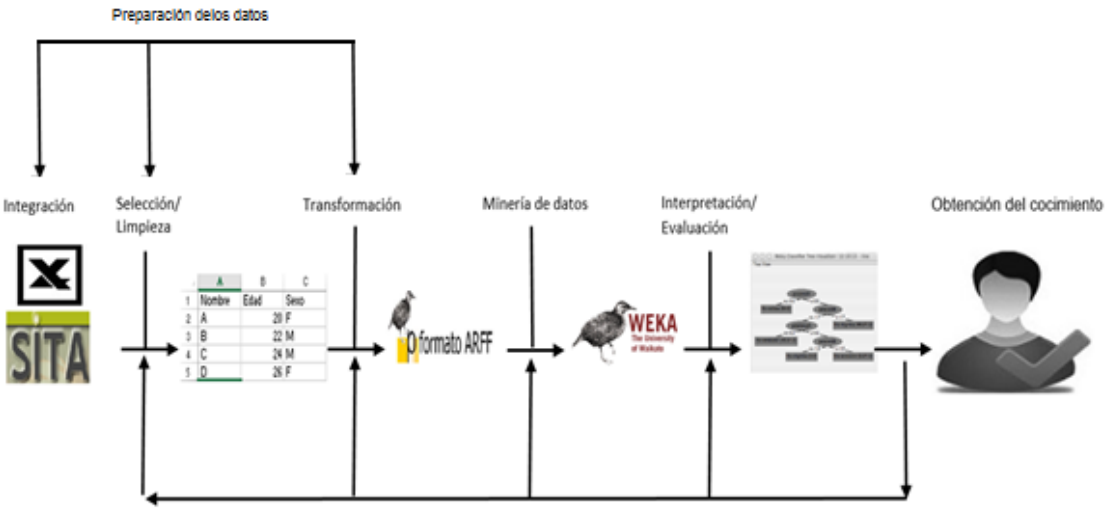


Figura 4.1 Proceso de extracción del conocimiento.

4.1 Preparación de los datos

La preparación de los datos incluye las primeras dos etapas del KDD, a continuación, se describirá lo realizado en estas fases.

Los datos corresponden a los estudiantes de la generación 2008 – 2013 que fueron aceptados en el Centro Universitario UAEM Valle de México (CUUAEMVM) y cursaron con éxito, alguna de las 10 carreras que se ofertan (ver tabla 5.1).

Tabla 4.1 Oferta educativa del CUUAEMVM.

Carrera	Siglas	Área
Actuaría	LAC	Licenciatura
Administración	LAM	
Contaduría	LCN	
Derecho	LDE	
Economía	LEC	
Informática Administrativa	LIA	
Relaciones Económicas Internacionales	REI	
Sistemas y Comunicaciones	LSC	Ingeniería
Industrial	IIN	
Computación	ICO	

De los datos del estudio socioeconómico proporcionado por el SITA, se seleccionaron los resultados del EXANI-II y la información de factores generales, específicamente los referentes a la escolaridad, y de los factores educativos los referentes a los hábitos de estudio, las actividades de estudio y las formas de aprendizaje (ver tabla 4.2).

Tabla 4.2 Factores del estudio socioeconómico considerados en esta tesis.

EXANI-II	Factores Generales	Factores Educativos
Razonamiento Verbal	Sustentante	Mejora de la Institución
Razonamiento Numérico	Institución	Actividades Culturales
Mundo Contemporáneo	**Escolaridad**	Actividades de Maestros
Ciencias Naturales	Estructura Familiar	Libros e Idiomas
Ciencias Sociales	Situación Laboral	**Hábitos de estudio**
Matemáticas	Datos socioeconómicos	**Actividades de Estudios**
Español		**Formas de Aprendizaje**

El total de atributos del estudio socioeconómico del SITA y el EXANI es de 170, sin embargo, para este trabajo solo se consideraron 52, aquellos que son considerados competencia única del estudiante y son: los 7 resultados del EXANI-II (ver tabla 4.3); de los factores generales, solo los 2 correspondientes al rubro de Escolaridad (ver tabla 4.4), mientras que en los factores educativos solo se consideraron los 9 hábitos de estudio, las 12 actividades de estudios, y las 22 formas de aprendizaje. Los 43 atributos de los factores educativos se muestran en las tablas 4.5, 4.6 y 4.7 respectivamente.

Tabla 4.3 Conocimientos evaluados en el EXANI-II.

ÁREA	CLAVE
1. Ciencias Sociales	CS
2. Razonamiento numérico	RN
3. Matemáticas	MAT
4. Mundo Contemporáneo	MC
5. Razonamiento Verbal	RV
6. Ciencias Naturales	CN
7. Español	ESP

Tabla 4.4 Factores generales, rubro escolaridad.

Promedio bachillerato
Exámenes extraordinarios

Tabla 4.5 Factores educativos, rubro actividades de estudio.

Me reúno con mis compañeros para preparar un examen
Me reúno con mis compañeros para elaborar una tarea o un trabajo en grupo
Al iniciar, identifico lo que necesito estudiar y elaboro un plan de trabajo
Reviso qué es lo que recuerdo de lo que estudié
Identifico los conceptos que aún no he comprendido
Al no entender algo busco información que me aclare dudas
Estudio principalmente con monografías
Estudio principalmente con mis apuntes de clase
Estudio principalmente con el libro de texto de la asignatura
Estudio principalmente con los apuntes de mis compañeros
Utilizo enciclopedias, diccionarios o atlas
Uso computadora o Internet para trabajos escolares

Tabla 4.6 Factores educativos, formas de aprendizaje.

Aprendo más cuando trabajo con otros compañeros
Es de gran ayuda que todos aporten ideas cuando trabajo en grupo
Estudio para asegurar económicamente mi futuro
Estudio para obtener un buen trabajo
Estudio para aprender más
Estudio para vivir mejor
Tengo confianza en que puedo entender los textos más difíciles
Tengo confianza en que puedo realizar un excelentes trabajos escolares
Tengo seguridad en que domino las habilidades que me enseñaron
Aprendo rápidamente en la mayoría de las asignaturas
Soy competente en la mayoría de las asignaturas
Resuelvo bien los exámenes en la mayoría de las asignaturas
Me gusta trabajar con otros compañeros
Solamente leo cuando tengo la obligación de hacerlo
La lectura es uno de mis pasatiempos favoritos
Me gusta comentar los libros con otras personas

Tabla 4.6 Factores educativos, formas de aprendizaje (continuación).

Me cuesta trabajo terminar de leer un libro
Me gusta que me regalen libros
La lectura me parece una pérdida de tiempo
Disfruto el visitar librerías o bibliotecas
Solamente leo para obtener la información que necesito
Me cuesta trabajo sentarme a leer por mucho tiempo

Tabla 4.7 Factores educativos, hábitos de estudio.

¿Cuántas horas a la semana estudia fuera del horario escolar?
¿Cuántas horas a la semana lee sobre lo que le gusta o interesa?
¿Cuántos libros completos ha leído en el último año no cuentan los libros de texto
Libros de literatura (novela, teatro, poesía)
Libros de otros temas (ciencia, tecnología, economía, etc.)
Revistas
Periódicos
Historietas
Páginas de Internet

La información del EXANI-II y del estudio socioeconómico se relacionó con la información de control escolar que indica si el alumno egresó o se dio de baja por motivos personales o académicos. Seleccionando únicamente los datos de los alumnos que cursaron con éxito su carrera y que tuvieran la información completa tanto en el EXANI-II como en los factores considerados.

La generación con la que se trabajó es la del 2008 – 2013 quedando un total de 879 registros en el caso de los resultados del EXANI-II y 510 para el caso de los otros factores. La diferencia en el número es porque todos los alumnos, sin excepción deben hacer el EXANI-II, pero no todos llenan el cuestionario del estudio socioeconómico, o lo dejan incompleto.

4.2 Diseño de los experimentos.

El primer experimento consistió en aplicar el algoritmo C4.5 (J48 en Weka) con diferentes valores de minObj, primero para todos los factores y posteriormente para cada factor, esto para determinar qué valor de minObj nos permitía tener un mejor análisis del árbol, creando los árboles con 5, 10 y 15 minObj y a su vez saber que factor tenía mejores resultados.

Una vez concluido el primer experimento, se procedió a analizar el conocimiento capturado en el árbol generado por el algoritmo J48 cuando se utilizan los resultados del mejor factor.

El segundo experimento consistió en modificar los datos para no poner el nombre de cada licenciatura, sino solo manejar dos clases: licenciatura e ingeniería; de igual manera que el primer experimento se hizo la prueba del algoritmo con todos los factores y posteriormente para cada factor, se generaron los árboles con 5, 10 y 15 elementos, se procedió a analizar el conocimiento capturado con el minObj que nos permitirá un mejor análisis con el factor que da mejores resultados.

El tercer experimento consistió en separar los datos de los ingenieros (190) y de los licenciados (689) para aplicar el mismo algoritmo con los diferentes valores de minObj al conjunto de factores y posteriormente a cada factor primero para los ingenieros y posteriormente a los licenciados.

4.3 Resultados.

El primer experimento consistió en aplicar el algoritmo C4.5 (J48 en Weka) con diferentes valores de minObj primero a todos los factores y posteriormente a cada grupo de factores, los resultados se pueden ver en la tabla 4.8.

Tabla 4.8 Resultados del modelo clasificador con todos los factores.

	Todos los factores.	
J48	% Éxito	% Error
5	72.74	27.26
10	67.05	32.95
15	63.13	36.87

Los árboles generados con 5, 10 y 15 elementos para la prueba de cada una de las carreras y todos los factores, se diferencian en tamaño, ya que con 5 elementos se obtiene una extensión de 33 hojas y el número de nodos es de 59, a diferencia del árbol generado con 10 elementos que tiene 18 hojas y 33 nodos, por último el árbol generado para 15 elementos se generan 12 hojas y 23 nodos, los árboles generados se pueden observar en las figuras 4.2 a 4.11.

En la figura 4.2 se encuentra la raíz del árbol para 5 elementos, debido a su extensión se dividió en subárbol A1 y subárbol A2.

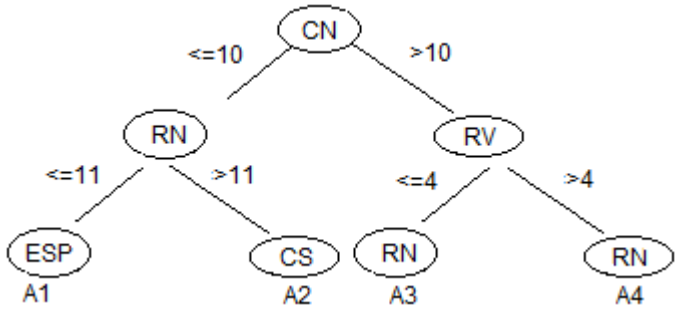


Figura 4.2 Raíz del árbol de decisión con 5 elementos para cada carrera y todos los factores.

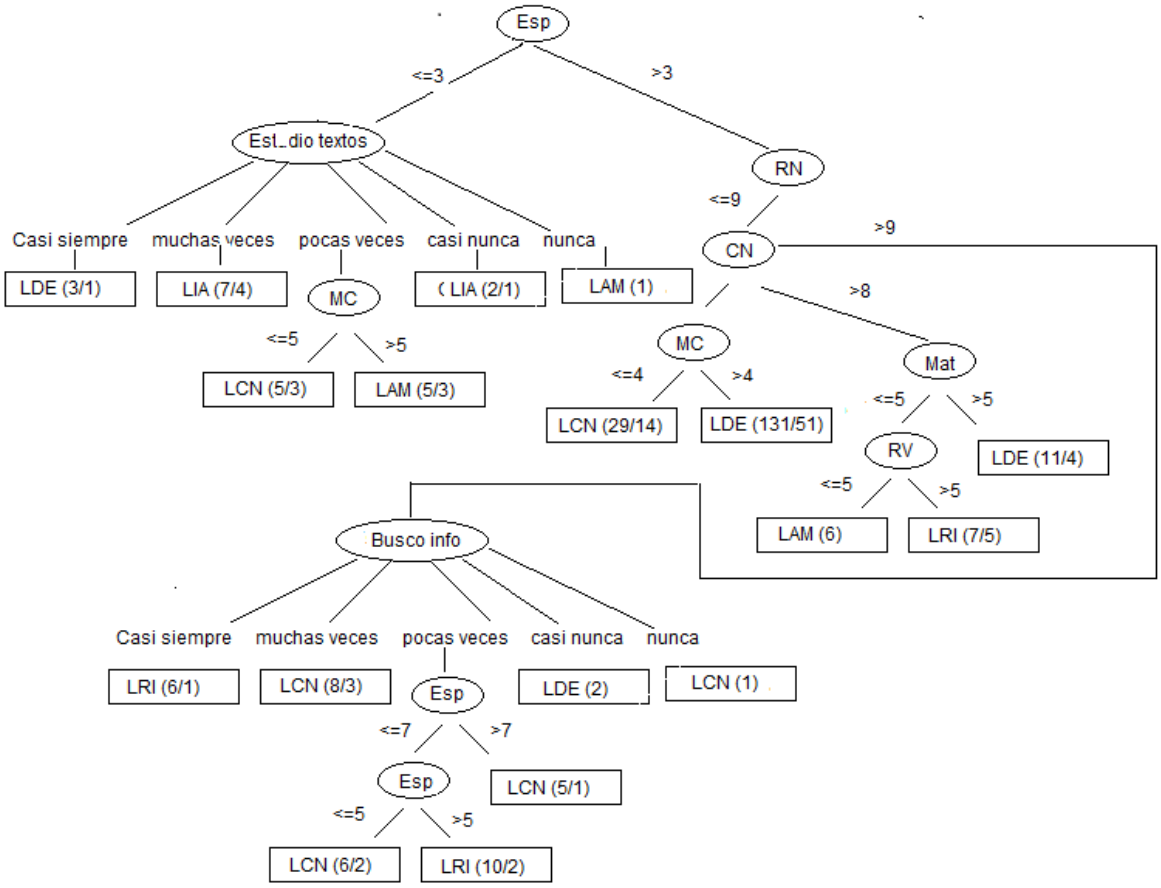


Figura 4.3 Subárbol A1 del árbol con 5 elementos para cada carrera y todos los factores.

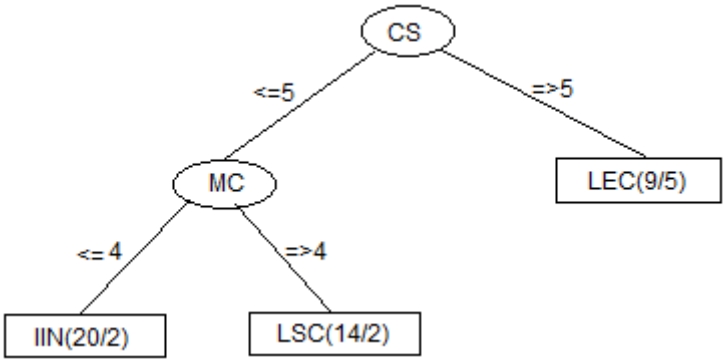


Figura 4.4 Subárbol A2 del árbol con 5 elementos para cada carrera y todos los factores.

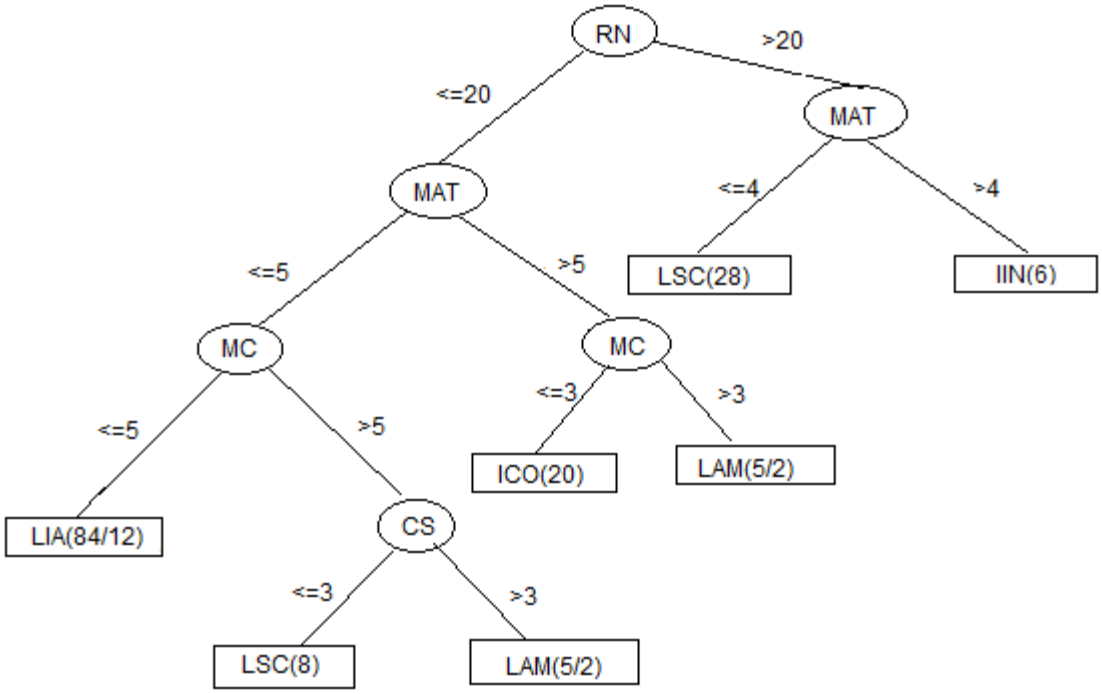


Figura 4.5 Subárbol A3 del árbol con 5 elementos para cada carrera y todos los factores.

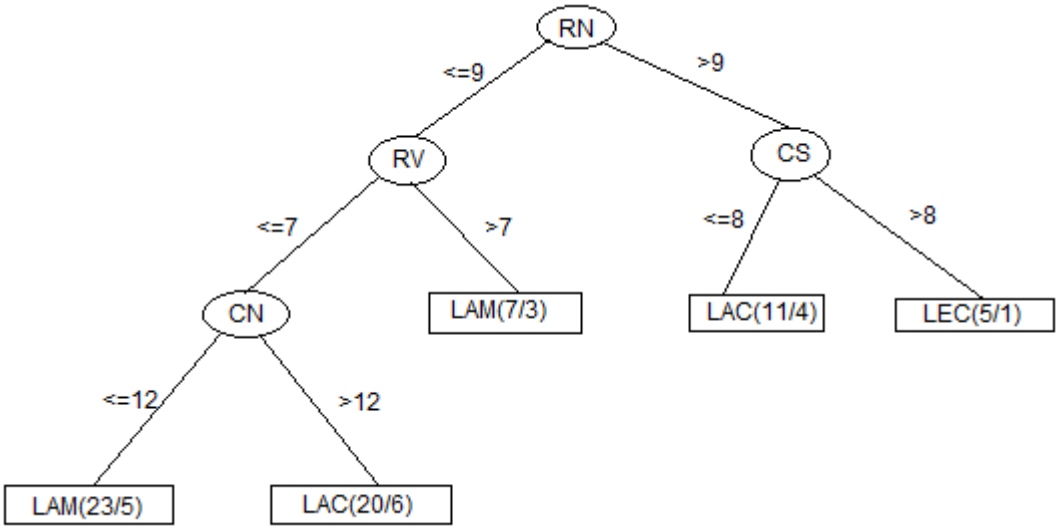


Figura 4.6 Subárbol A4 del árbol con 5 elementos para cada carrera y todos los factores.

Tabla 4. 9 Matriz de confusión para 5 objetos para cada carrera y todos los factores.

=== Confusion Matrix ===											
a	b	c	d	e	f	g	h	i	j	<-- classified as	%Acierto
37	1	4	0	2	0	0	0	0	6	a = LAM	74.0
0	15	31	0	8	1	0	0	0	1	b = LRI	26.7
4	2	91	1	9	0	0	0	0	0	c = LDE	85.0
3	0	0	76	0	0	2	0	0	1	d = LIA	92.6
1	3	19	1	31	0	4	0	0	0	e = LCN	52.5
0	0	0	4	0	48	0	0	2	0	f = LSC	88.8
4	1	1	3	4	1	8	0	0	1	g = LEC	34.7
1	0	0	5	0	0	0	20	0	0	h = ICO	76.9
0	0	1	1	0	0	0	0	24	1	i = IIN	88.8
2	1	0	2	0	0	0	0	0	21	j = LAC	80.7

En la tabla 4.9 se puede observar en la parte derecha de la matriz se encuentran cada una de las carreras asignadas a una letra, en el primer renglón se encuentra el número de alumnos clasificados en cada carrera, para Administración, LAM, 37 alumnos fueron clasificados correctamente ya que fueron clasificados como licenciados en administración y 6 alumnos fueron clasificados erróneamente en la carrera de Actuaría, y 7 en otras carreras, dando un total de 37 alumnos clasificados correctamente de 50. En la tabla se observa que las carreras con mayor error en la clasificación son las de LRI, LCN y LEC, por ejemplo, en el renglón 2, se observa que hay mucha confusión entre las carreras de Relaciones Económicas Internacionales y Derecho siendo clasificados solo 15 alumnos de manera correcta, 31 en la carrera de derecho y 10 en otras carreras; sin embargo, se puede notar en el renglón 3, que 91 licenciados en derecho (de 107), son clasificados correctamente. Se puede observar también que una de las carreras que tiene poco grado de confusión es la de los ingenieros en sistemas, teniendo un total de 48 alumnos clasificados correctamente de 54 y que la confusión es con Ingenieros industriales, y con la carrera de Licenciado en Informática administrativa que tienen más afinidad que otras como derecho actuaría, etc. hay un alto grado de confusión porque solamente 8 alumnos de economía están clasificados correctamente y 15 están clasificados erróneamente en las demás carreras. De hecho, las carreras de

ingeniería (letras f, h, i, renglones 6,8 y 9) muestran un buen resultado en la clasificación. La última columna de la tabla 4.9 muestra el porcentaje de acierto de cada carrera basado en la matriz de confusión.

A diferencia del árbol de 5 elementos que tenía 4 subárboles, el de 10 objetos tiene solo 2. La raíz de este árbol se muestra en la figura 4.7 y sus subárboles en las figuras 4.8 y 4.9.

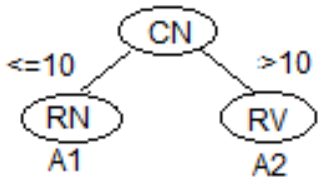


Figura 4. 7 Raíz del árbol de decisión con 10 objetos para cada carrera y todos los factores.

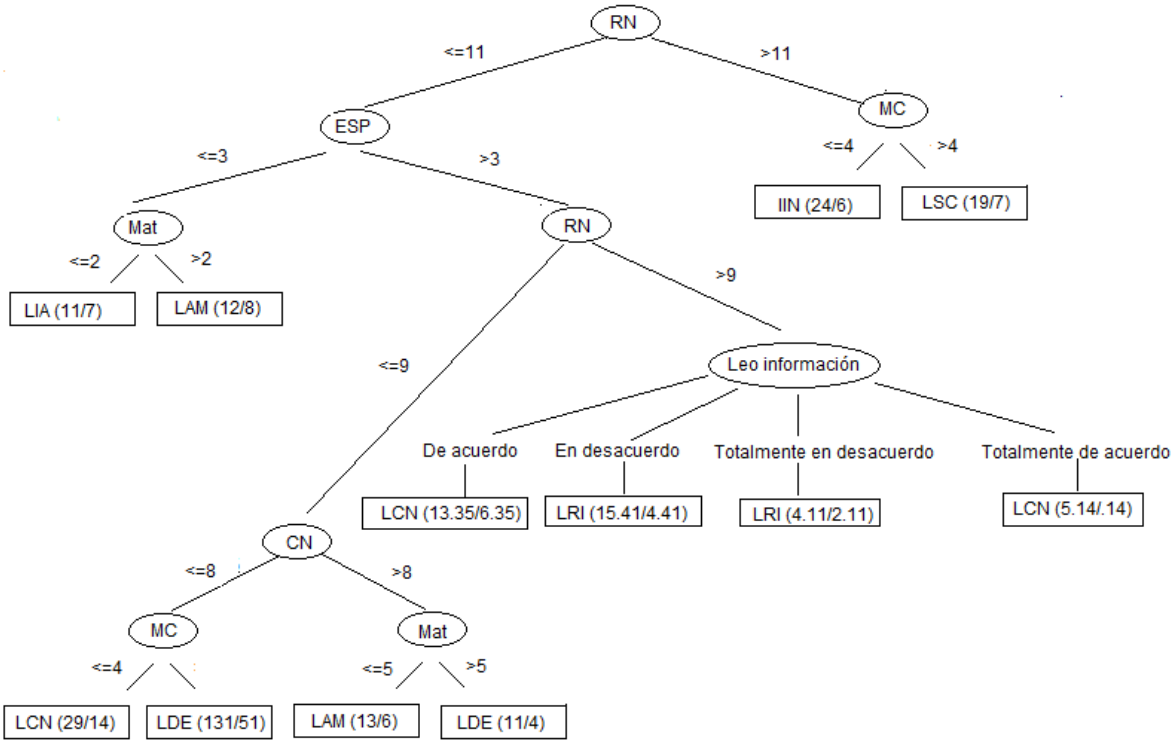


Figura 4.8 Subárbol A1 del árbol con 10 elementos para cada carrera y todos los factores.

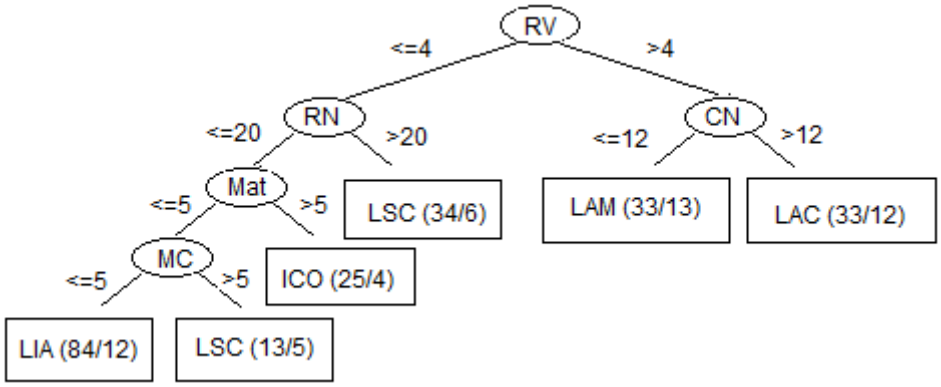


Figura 4.9 Subárbol A2 con 10 elementos para cada carrera y todos los factores.

En la tabla 4.10 se observa la matriz de confusión en donde la primer línea están los alumnos de la carrera de administración con 31 alumnos clasificados correctamente de 50, posteriormente 13 alumnos de la carrera de relaciones económicas internacionales de 56, 87 alumnos de derecho de 107, 76 alumnos de informática administrativa de 82, 27 de contaduría de 59, 48 alumnos de sistema y comunicaciones de 54, 0 alumnos de economía de 23, 21 alumnos de ingenieros en computación de 26, 18 alumnos de ingenieros industriales de 27, 21 alumnos de actuaría que fueron identificados correctamente de 26, para la carrera de relaciones económicas internacionales hubo mucha confusión porque no hubieron alumnos clasificados en dicha carrera y 23 fueron confundidos en diversas carreras, otra carrera donde existió confusión fue en la carrera de contaduría al tener 27 alumnos clasificados correctamente y 29 en diversas carreras.

Tabla 4. 10 Matriz de confusión para 5 objetos para cada carrera y todos los factores.

=== Confusion Matrix ===										
a	b	c	d	e	f	g	h	i	j	<-- classified as
31	0	4	0	1	3	0	3	0	8	| a = LAM
4	13	31	0	7	1	0	0	0	0	| b = LRI
8	0	87	1	11	0	0	0	0	0	| c = LDE
1	0	0	76	0	2	0	0	2	1	| d = LIA
3	5	19	2	27	3	0	0	0	0	| e = LCN
0	0	0	4	0	48	0	0	2	0	| f = LSC
9	1	1	3	2	3	0	0	2	2	| g = LEC
0	0	0	5	0	0	0	21	0	0	| h = ICO
0	0	0	2	0	6	0	0	18	1	| i = IIN
2	0	0	2	0	0	0	1	0	21	| j = LAC

El tamaño del árbol se reduce drásticamente en comparación con los de 5 y 10 elementos, la raíz de este árbol es idéntico al de 10 objetos por lo que en la figura 4.10 y 4.11 se mostrarán los subárboles A1 y A2.

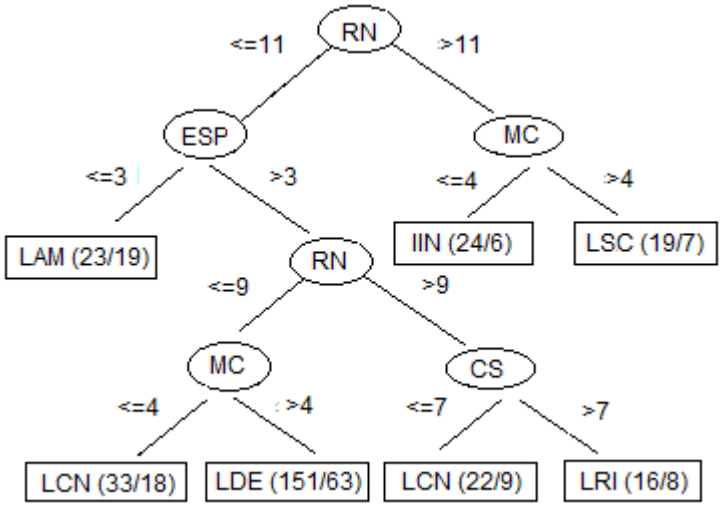


Figura 4. 10 Subárbol A1 con 15 elementos para cada carrera y todos los factores.

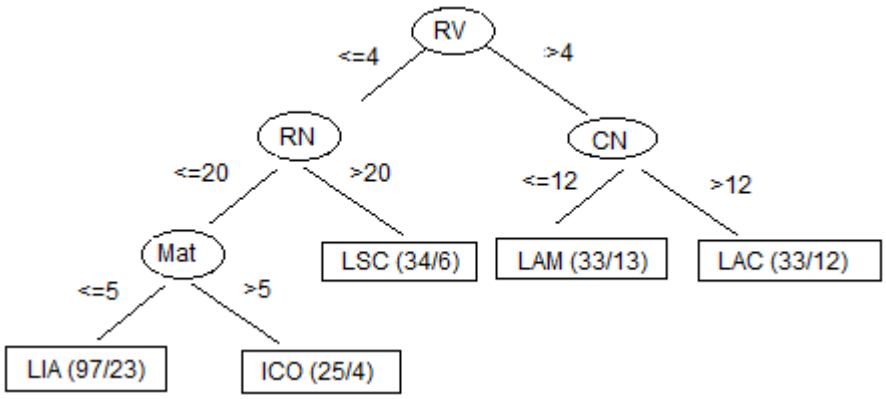


Figura 4. 11 Subárbol A2 con 15 elementos para cada carrera y todos los factores.

En la tabla 4.11 se muestra la matriz de confusión, en la cual se pueden identificar las carreras donde hubo poca confusión por parte del algoritmo y a su vez cuales fueron las carreras en donde los alumnos no fueron bien identificados. En la primera línea está la carrera de administración teniendo un total de 24 alumnos clasificados correctamente y 10 alumnos que fueron mal identificados al situarlos como licenciados en derecho, una carrera donde hubo mucha confusión fue en la carrera de relaciones económicas internacionales por que 33 alumnos fueron mal identificados en la carrera de derecho y 8 fueron clasificados de forma correcta, sin embargo para la licenciatura en derecho se tuvo 88 alumnos identificados de manera correcta y 9 alumnos mal clasificados al situarlos en la carrera de contaduría.

Tabla 4. 11 Matriz de confusión para 15 elementos con cada carrera y todos los factores.

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j   <-- classified as
 24  0 10  3  2  0  0  3  0  8 |  a = LAM
  2  8 33  0 12  1  0  0  0  0 |  b = LRI
  7  3 88  0  9  0  0  0  0  0 |  c = LDE
  5  0  0 74  0  0  0  0  2  1 |  d = LIA
  5  4 19  0 28  3  0  0  0  0 |  e = LCN
  1  0  0 11  0 40  0  0  2  0 |  f = LSC
 10  1  1  1  3  3  0  0  2  2 |  g = LEC
  0  0  0  5  0  0  0 21  0  0 |  h = ICO
  1  0  0  1  0  6  0  0 18  1 |  i = IIN
  1  0  0  2  1  0  0  1  0 21 |  j = LAC
```

De la prueba anterior, se observa que el árbol que tiene mayor facilidad de análisis es el de 15 elementos ya que no es tan extenso; al notar que el porcentaje de éxito no es muy favorable, se procedió a aplicar el mismo algoritmo para cada uno de los factores realizando la siguiente prueba la cual consistió en separar cada uno de los factores y aplicar el mismo algoritmo con diferentes valores de minObj, los resultados se observan en las tablas 4.12, 4.13, 4.14 y 4.15.

Tabla 4.12 Resultados del modelo clasificador con los resultados del EXANI-II y los factores de escolaridad.

	EXANI-II - Escolaridad	
J48	% Éxito	% Error
5	71.56	28.44
10	66.07	33.93
15	63.13	36.87

Tabla 4.13 Resultados del modelo clasificador usando los resultados de Actividades de Estudio.

	Actividades de Estudio	
J48	% Éxito	% Error
5	42.15	57.85
10	33.33	66.67
15	30.98	69.02

Tabla 4.14 Resultados del modelo clasificador usando los resultados de Formas de Aprendizaje.

	Formas de Aprendizaje	
J48	% Éxito	% Error
5	43.33	56.67
10	31.56	68.44
15	29.8	70.2

Tabla 4.15 Resultados del Modelo Clasificador usando los resultados de Hábitos de Estudio.

	Hábitos de Estudio	
J48	% Éxito	% Error
5	40.78	59.22
10	32.74	67.26
15	27.25	72.75

Del resultado anterior se observa que los 4 rubros mencionados, el que mejor funciona para separar las clases son los resultados del EXANI-II-Escolaridad, seguido de las formas de aprendizaje, las actividades de estudio y los hábitos de estudio. De lo anterior, se puede concluir que los 3 últimos no se presentan como un factor determinante para la elección de carrera.

Por lo anterior, procedimos a analizar el conocimiento capturado en el árbol generado por el algoritmo J48 cuando se utilizan los resultados del EXANI-II-Escolaridad, el cual se muestra de manera simplificada en la figura 4.12, por cuestiones de espacio y legibilidad. Los subárboles se mostrarán y analizarán en figuras subsecuentes. Se muestra el árbol generado cuando se consideran un mínimo de 15 elementos que, aunque no es el que tiene la mejor precisión, si es por su extensión el que permite un mejor análisis.

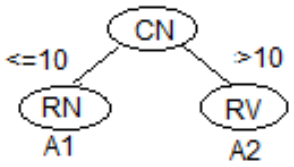


Figura 4. 12 Raíz del árbol de decisión usando los resultados del EXANI-II-Escolaridad.

La raíz del árbol nos muestra que el atributo más relevante para separar a las carreras fue el resultado obtenido en las Ciencias Naturales, seguido del Razonamiento Numérico en la rama izquierda y del Razonamiento Verbal en la rama de la derecha.

Al analizar los subárboles A1 y A2 podemos observar que existen ramas que nos llevan a solo tener ingenierías y otras a solo licenciaturas.

Particularmente, en el subárbol A1 (Ciencias Naturales <=10 y Razonamiento Numérico <=11), mostrado en la figura 4.13, encontramos cosas interesantes, la primera de ellas es que las carreras son todas licenciaturas, la segunda es que cuando el Razonamiento Numérico >11 encontramos a dos de las tres ingenierías, LSC: Ingeniería en Sistemas y Comunicaciones e IIN: Ingeniería Industrial

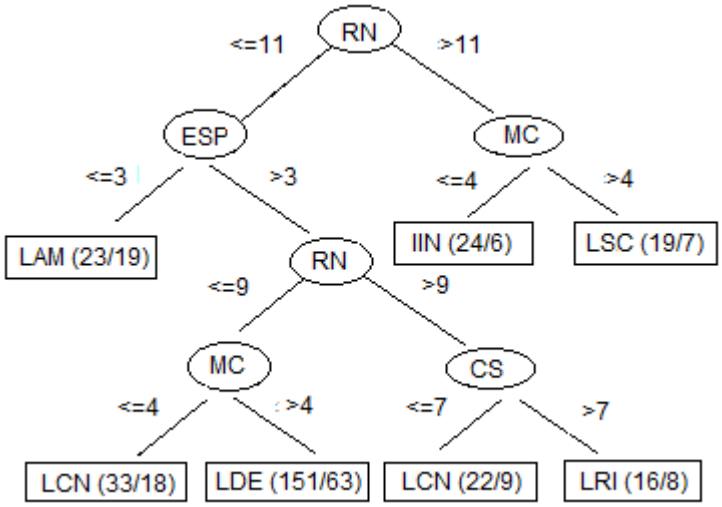


Figura 4. 13 Subárbol A1, licenciaturas e ingenierías.

En el subárbol A2 (Ciencias Naturales >10 y Razonamiento Verbal <=4), lo que notamos es que en su mayoría son carreras relacionadas con la computación, ICO: Ingeniería en Computación, LSC: Ingeniería en Sistemas, LIA: Informática Administrativa, por otro lado cuando el Razonamiento Verbal es >4, encontramos que son solo carreras de licenciatura y relacionadas con el razonamiento numérico, LAM: Administración y LAC: Actuarios (ver figura 4.14).

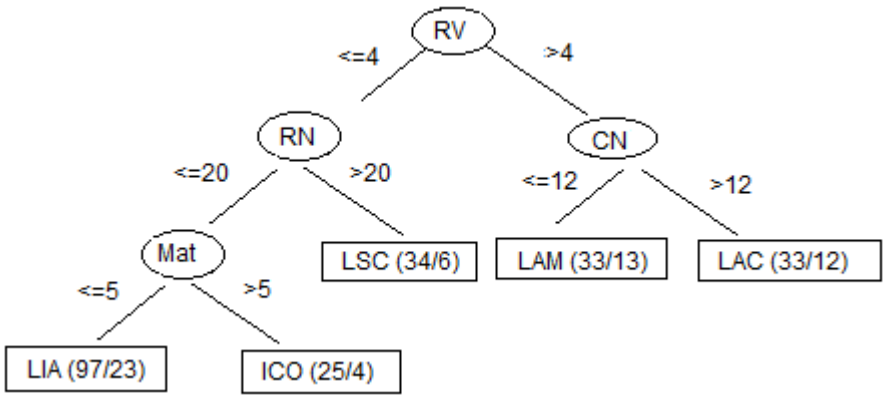


Figura 4. 14 Subárbol A2, carreras relacionadas con la computación y Razonamiento Numérico.

En la tabla 4.16 se observa la matriz de confusión del algoritmo en donde se nota que de igual manera la carreras que tienen un alto grado de exactitud al clasificar correctamente a los alumnos son la carrera de administración al tener 24 alumnos bien clasificados, la carrera de derecho con 88 alumnos bien clasificados, 74 en la carrera de informática administrativa y 40 para la carrea de sistemas, por otro lado las carreras donde existe mucha confusión son la carrera de relaciones internacionales con solo un alumno bien clasificado y 33 alumnos confundidos en la carrera de derecho, de la misma manera en las carreras de economía y actuaría hay un alto índice de confusión porque tienen alumnos en todas las carreras y no se encuentra en la carrera a la que pertenecen.

Tabla 4. 16 Matriz de confusión usando los resultados del EXANI-II-Escolaridad.

=== Confusion Matrix ===										
a	b	c	d	e	f	g	h	i	j	<-- classified as
24	0	10	3	2	0	0	3	0	8	| a = LAM
2	8	33	0	12	1	0	0	0	0	| b = LRI
7	3	88	0	9	0	0	0	0	0	| c = LDE
5	0	0	74	0	0	0	0	2	1	| d = LIA
5	4	19	0	28	3	0	0	0	0	| e = LCN
1	0	0	11	0	40	0	0	2	0	| f = LSC
10	1	1	1	3	3	0	0	2	2	| g = LEC
0	0	0	5	0	0	0	21	0	0	| h = ICO
1	0	0	1	0	6	0	0	18	1	| i = IIN
1	0	0	2	1	0	0	1	0	21	| j = LAC

Estas observaciones, nos llevaron al siguiente experimento que consistió en modificar los datos para no poner el nombre de cada licenciatura, sino solo manejar dos clases: licenciatura e ingeniería; de igual manera que el primer experimento se hizo la prueba del algoritmo con todos los factores, a continuación se muestran los resultados en la tabla 4.17.

Tabla 4.17 Resultados del modelo clasificador con todos los factores.

	Todos los factores	
J48	% Éxito	% Error
5	98.62	1.38
10	96.86	3.12
15	94.9	5.1

Los árboles generados con 5, 10 y 15 elementos para licenciatura e ingeniería con a todos los factores dan los siguientes resultados, con 5 elementos se obtiene una extensión de 11 hojas y el número de nodos es 21, a diferencia del árbol generado con 10 elementos que tiene 9 hojas y 17 nodos, por último, el árbol generado para 15 elementos se generan 5 hojas y 9 nodos, los árboles generados se pueden observar en las figuras 4.15 a 4.20.

En la figura 4.15 se encuentra la raíz del árbol de decisión para 5 elementos, debido a su extensión se dividió en subárbol A1 y subárbol A2, los cuáles se encuentran en la figura 4.16 y 4.17.

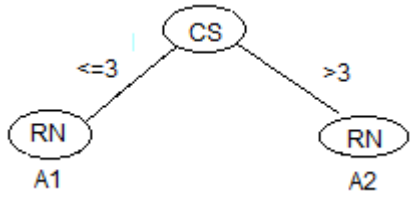


Figura 4. 15 Raíz del árbol de decisión con 5 elementos raíz para ingenieros y licenciados con todos los factores.

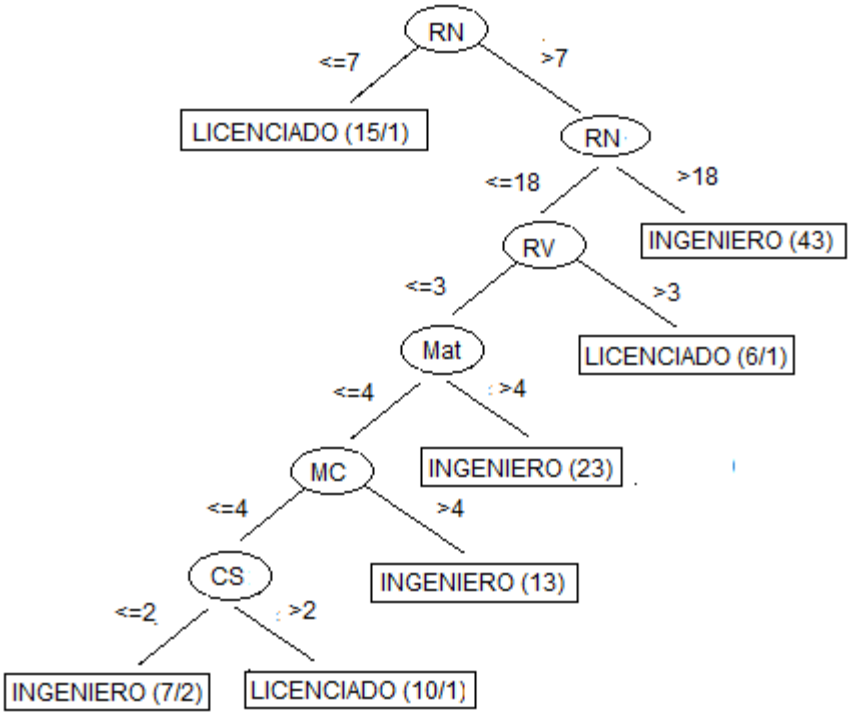


Figura 4.16 Subárbol A1 con 5 elementos para Ingenieros y Licenciados con todos los factores.

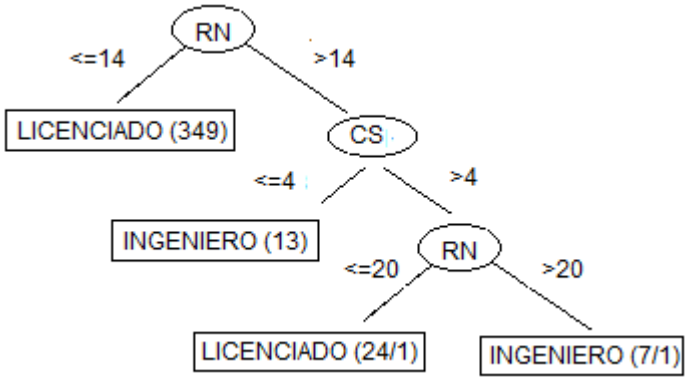


Figura 4.17 Subárbol A2 con 5 elementos para Ingenieros y Licenciados con todos los factores.

En la matriz de confusión de la tabla 4.18 se observa que 400 licenciados fueron clasificados correctamente y solo 3 fueron confundidos como ingenieros, por otra parte 103 ingenieros fueron clasificados correctamente y 4 fueron confundidos como licenciados.

Tabla 4. 18 Matriz de confusión con 5 elementos para ingenieros y licenciados con todos los factores.

=== Confusion Matrix ===	
a	b <-- classified as
103	4 | a = INGENIERO
3	400 | b = LICENCIADO

Aunque el árbol obtenido con 5 elementos no fue tan extenso, al tener 10 elementos se simplificó aún más, siendo entonces la misma raíz del árbol que la de 5 elementos y se redujeron los subárboles A1 y A2 situados en la figura 4.18 y 4.19.

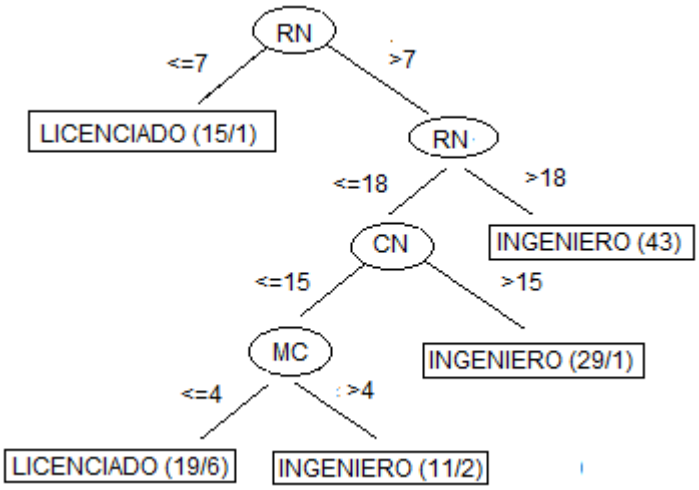


Figura 4.18 Subárbol A1 con 10 elementos para Ingenieros y Licenciados con todos los factores.

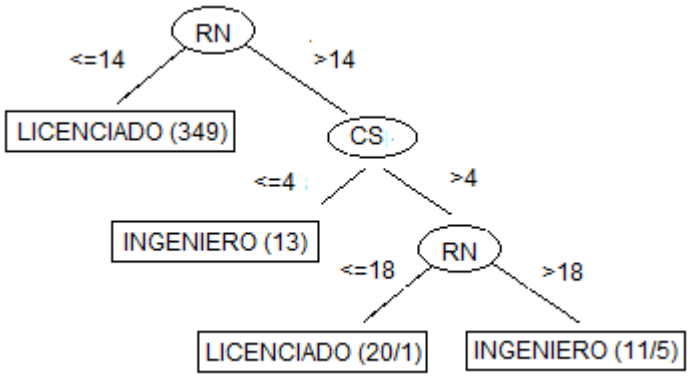


Figura 4. 19 Subárbol A2 con 10 elementos para Ingenieros y Licenciados con todos los factores.

En la Tabla 4.19 se puede observar la matriz de confusión que nos muestra a 395 licenciados bien clasificados y 8 que se confundieron como si fueran ingenieros, a 99 ingenieros bien clasificados correctamente y 8 que fueron confundidos como licenciados.

Tabla 4. 19 Matriz de confusión con 10 elementos para ingenieros y licenciados con todos los factores.

=== Confusion Matrix ===
a b <-- classified as
99 8 | a = INGENIERO
8 395 | b = LICENCIADO

Al tener 15 elementos, se reduce y se simplifica aún más que teniendo 5 y 10 elementos, logrando un cambio en la raíz del árbol el cual se puede ver en la figura 4.20.

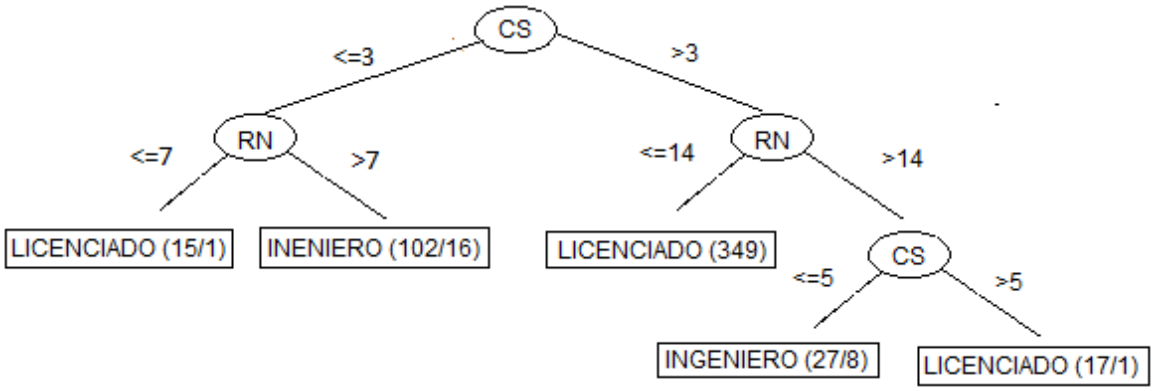


Figura 4.20 Árbol de decisión con 15 elementos para Ingenieros y Licenciados con todos los factores.

En la tabla 4.20 se encuentra la matriz de confusión que nos indica que 379 licenciados fueron clasificados adecuadamente y 24 se confundieron como

ingenieros, se tienen a 105 ingenieros bien clasificados y 2 confundidos como licenciados.

Tabla 4. 20 Matriz de confusión con 15 elementos para ingenieros y licenciados con todos los factores.

=== Confusion Matrix ===		
a	b	<-- classified as
105	2	a = INGENIERO
24	379	b = LICENCIADO

Al realizar esta prueba, se observa que se tiene buen resultado en el porcentaje de éxito y aunque es bueno se procedió a realizar la siguiente prueba, tomando en cuenta los resultados del primer experimento en donde se observó que los que mejor daban porcentaje de éxito eran las pruebas del EXANI-II, se procedió a aplicar el algoritmo a dicho factor; los resultados se muestran en la tabla 4.21 y el árbol en la figura 4.21.

Tabla 4.21 Resultados usando EXANI-II en Licenciaturas e Ingenierías.

	EXANI-II	
J48	% Éxito	% Error
5	98.06	1.94
10	96.24	3.76
15	96.13	3.87

En el caso de las hojas del árbol de la figura 4.21, estos muestran 2 valores, el primero nos indica el número de elementos que cayeron en esa rama y el segundo cuántos son clasificados incorrectamente. Podemos observar en el árbol generado que con solo 2 atributos del EXANI-II: Razonamiento numérico (RN) y Ciencias Sociales (CS) es posible determinar si estudiar una carrera de licenciado o de ingeniero.

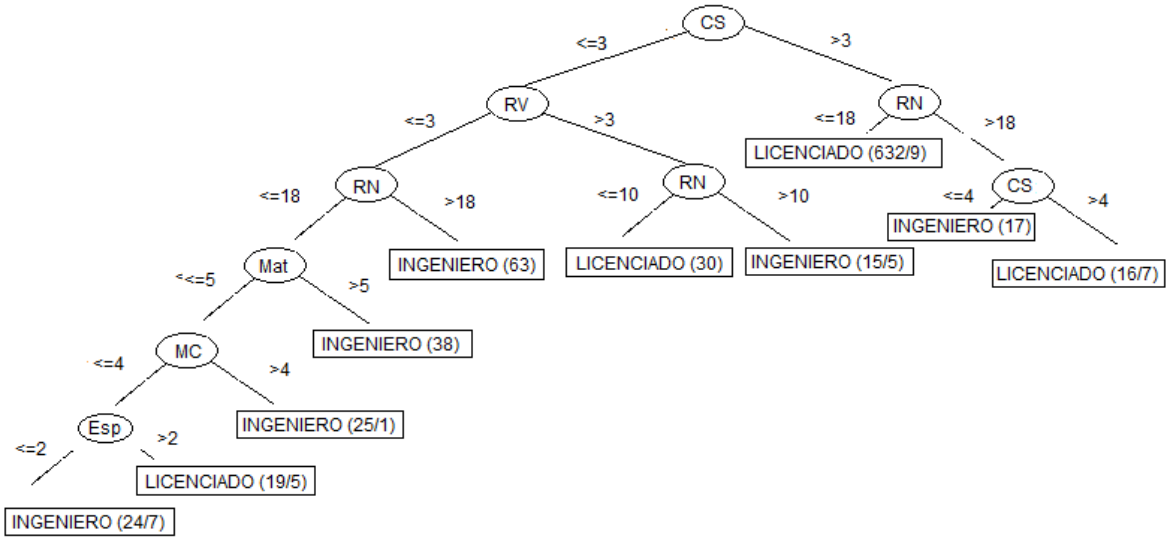


Figura 4.21 Árbol Generado usando EXANI-II en Licenciaturas e Ingenierías.

La matriz de confusión de la tabla 4.22 muestra a 676 licenciados clasificados correctamente y 13 confundidos como ingenieros, por otra parte, se muestra a 169 ingenieros clasificados correctamente y 21 confundidos como ingenieros.

Tabla 4. 22 Matriz de confusión usando EXANI-II en Licenciaturas e Ingenierías

=== Confusion Matrix ===		
a	b	<-- classified as
676	13	| a = LICENCIADO
21	169	| b = INGENIERO

Los resultados coinciden con el razonamiento común: los licenciados requieren más conocimientos de las ciencias sociales y de un razonamiento verbal, mientras que los ingenieros de un razonamiento numérico.

Posterior a este experimento, se separaron los datos de los ingenieros (190) y de los licenciados (689) para el caso del EXANI-II y 403 licenciados, 107 ingenieros para los otros factores y se realizaron de nuevo los experimentos con los diferentes factores.

Los resultados para los licenciados se muestran en las tablas 4.23 a la 4.27, donde notamos que los resultados del EXANI-II siguen siendo los mejores.

Tabla 4.23 Resultados usando Todos los factores, para licenciados.

	Todos los factores	
J48	% Éxito	% Error
5	68.48	31.52
10	65.6	34.4
15	60.54	39.46

Tabla 4.24 Resultados usando EXANI-II y Escolaridad, para licenciados.

	EXANI-II	
J48	% Éxito	% Error
5	72.85	27.15
10	65.74	34.26
15	64.44	35.56

Tabla 4.25 Resultados usando las Actividades de Estudio, para licenciados.

	Actividades de Estudio	
J48	% Éxito	% Error
5	44.66	55.34
10	39.70	60.3
15	31.51	68.49

Tabla 4.26 Resultados usando los resultados de Formas de Aprendizaje, para licenciados.

	Formas de Aprendizaje	
J48	% Éxito	% Error
5	48.38	51.62
10	41.19	58.81
15	38.46	61.54

Tabla 4.27 Resultados usando los resultados de Hábitos de Estudio, para licenciados.

	Hábitos de Estudio	
J48	% Éxito	% Error
5	45.4	54.6
10	37.22	62.78
15	26.55	73.45

De manera similar, el análisis del árbol de los licenciados se realizó segmentando el árbol en subárboles (Ver figuras 4.22 a la 4.26), con ello se generaron algunas reglas que si bien no permitirán una clasificación exacta, si reducen el abanico de opciones.

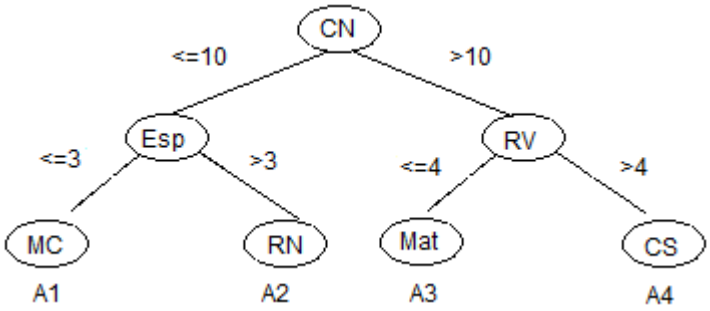


Figura 4.22 Árbol generado para las licenciaturas.

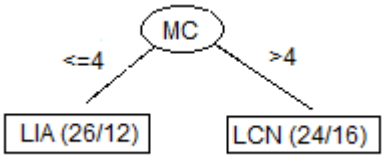


Figura 4.23 Subárbol A1 generado para las licenciaturas.

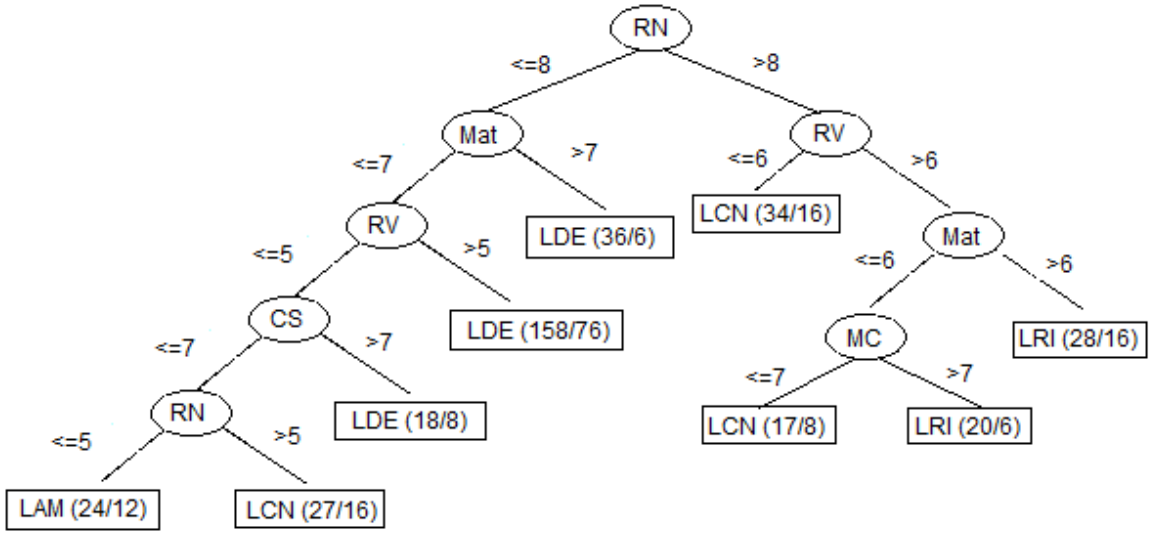


Figura 4.24 Subárbol A2 generado para las licenciaturas.

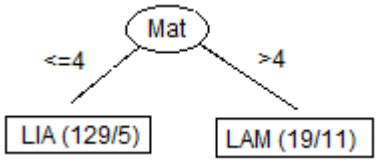


Figura 4.25 Subárbol A3 generado para las licenciaturas.

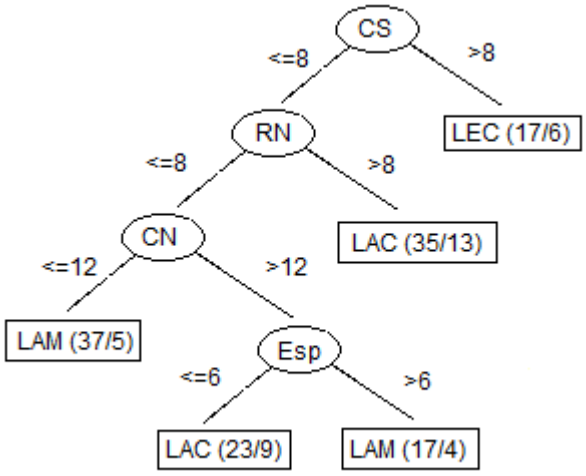


Figura 4.26 Subárbol A4 generado para las licenciaturas.

Las reglas son:

- Cuando en ciencias naturales es mayor a 10 y razonamiento verbal es mayor a 4 las carreras a elegir son: LEC, LAM o LAC.

- Cuando en ciencias naturales es mayor a 10 y razonamiento verbal es menor o igual a 4 puede ser LIA, o LAM.

- Cuando en ciencias naturales es menor o igual 10 y español es menor o igual a tres puede ser LIA o LCN.

- Cuando en ciencias naturales es menor o igual 10, español es mayor a 3 y razonamiento numérico es menor o igual a 8 puede ser LDE, LAM o LCN.

- Cuando en ciencias naturales es menor o igual 10, español es mayor a tres y razonamiento numérico es mayor a 8 puede ser LRI o LCN.

En la tabla 4.28 se muestra la matriz de confusión en donde se puede observar las licenciaturas; en la línea 1 está la carrera de informática administrativa que tiene 131 alumnos clasificados correctamente y 17 alumnos mal clasificados en otras carreras, para la carrera de actuaría hay 28 alumnos bien clasificados y 24 alumnos mal clasificados, para informática administrativa hay 62 alumnos bien clasificados y 33 alumnos que se confundieron en otras carreras, en la carrea de derecho hay una gran cantidad de alumnos clasificados de manera adecuada teniendo 114 alumnos bien clasificados y 35 que se confundieron en otras carreras para las carreras de relaciones económicas internacionales, contaduría y economía el índice de confusión fue mayor.

Tabla 4. 28 Matriz de confusión para las licenciaturas.

=== Confusion Matrix ===							
a	b	c	d	e	f	g	<-- classified as
138	7	7	0	0	2	2	| a = LIA
5	36	9	0	1	1	0	| b = LAC
1	10	65	0	8	9	2	| c = LAM
1	2	2	26	43	22	0	| d = LRI
2	1	6	9	122	8	1	| e = LDE
2	1	4	13	34	46	1	| f = LCN
6	1	4	0	4	14	11	| g = LEC

Para el caso de las ingenierías, los resultados son similares, El árbol generado resulta compacto y se presentan en la figura 4.27, y las tablas 4.29 a 4.34 condensan los resultados de los experimentos.

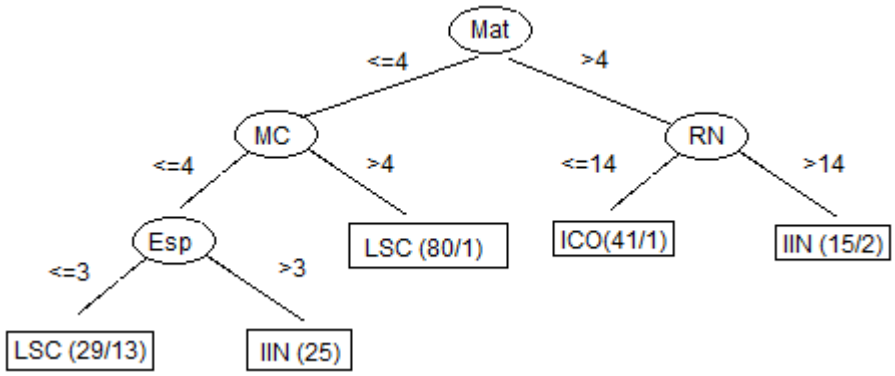


Figura 4.27 Modelo generado para las Ingenierías utilizando los resultados del EXANI-II del CENEVAL.

En la tabla 4.29 se observa la matriz de confusión que nos muestra en la línea 1 a los alumnos de la carrera de sistemas se clasifican a los 95 alumnos identificados, en la segunda línea están 38 ingenieros industriales clasificados correctamente, 9 que se confundieron como ingenieros en sistemas y 1 que fue confundido como

ingeniero en computación, por último están los ingenieros en computación que tiene a 40 alumnos clasificados correctamente, 5 que fueron confundidos como ingenieros en sistemas y 2 que fueron confundidos como ingenieros industriales.

Tabla 4. 29 Matriz de confusión para las Ingenierías con los resultados del EXANI-II del CENEVAL.

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 95  0  0 |  a = LSC
  9 38  1 |  b = IIN
  5  2 40 |  c = ICO
```

Tabla 4.30 Resultados usando todos los factores, para ingenieros.

	Todos los factores	
J48	% Éxito	% Error
5	94.39	5.61
10	91.58	8.42
15	91.58	8.42

Tabla 4.31 Resultados usando los resultados del EXANI-II y los factores de escolaridad, para ingenieros**.**

	EXANI-II	
J48	% Éxito	% Error
5	97.36	2.64
10	95.26	4.74
15	91.05	8.95

Tabla 4.32 Resultados usando los resultados de Actividades de Estudio, para ingenieros.

	Actividades de Estudio	
J48	% Éxito	% Error
5	50.46	49.54
10	50.46	49.54
15	50.46	49.54

Tabla 4.33 Resultados usando los resultados de Formas de Aprendizaje, para ingenieros.

	Formas de Aprendizaje	
J48	% Éxito	% Error
5	57.94	42.06
10	50.46	49.54
15	57.94	42.06

Tabla 4. 34 Resultados del Modelo Clasificador usando los resultados de Hábitos de Estudio, para ingenieros.

	Hábitos de Estudio	
J48	% Éxito	% Error
5	67.28	32.72
10	61.68	38.32
15	57	43

Conclusiones y Trabajo futuro.

La elección de carrera es una decisión muy importante en la vida de un estudiante, el no tomar una buena decisión conlleva a la pérdida de tiempo, dinero y esfuerzo. Con la aplicación de la minería de datos, se pudo comprobar que se pueden construir clasificadores con un 98.06% de corrección, el factor que mejor separa las clases son los resultados del EXANI-II. Los hábitos de estudio, las formas de aprendizaje y las actividades de estudio no se presentan como un factor determinante para la elección de carrera, el árbol de 15 elementos no fue el de mayor precisión, aunque si fue el mejor para ser analizado ya que no fue tan extenso.

El manejar el minObj en 15 objetos nos permitió lograr un 96.13% aceptable, con un árbol manejable para su estudio. En el análisis encontramos que con solo 3 atributos del EXANI-II: Razonamiento numérico, Ciencias Sociales y Razonamiento Verbal es posible determinar si estudiar una carrera de licenciado o de ingeniero.

Con esto, se ha comprobado al aplicar el algoritmo C4.5 de minería de datos resulta útil para el análisis de los diversos factores que pueden influir en la elección de carrera. Particularmente, sirvió para determinar que únicamente con los resultados del examen de nuevo ingreso EXANI-II del Centro Nacional de Evaluación CENEVAL es posible generar un clasificador para la elección de carrera.

El clasificador primero determina si la carrera a estudiar puede ser una licenciatura o una ingeniería. Posteriormente se puede utilizar el clasificador de las ingenierías o las reglas para las licenciaturas, y de esta forma apoyar al estudiante en esta difícil selección.

Referencias

Alao Kazeem, A., & Ibam Onwuka, E. (2017). Development of a Web-based Intelligent Career Guidance System for Pre-Tertiary Science Students in Nigeria.

Aquino, N. M. R., & Jara, E. A. M. (2016). Aplicación de Redes Neuronales Artificiales en la Orientación Vocacional. In *Memorias de Congresos UTP* (Vol. 1, No. 1, pp. 4-8).

Asociación Nacional de Universidades e Instituciones de Educación Superior (2016). *Anuario estadístico Población escolar y docente en la educación media superior y superior ciclo escolar 2015-2016.* Recuperado de http://www.anuies.mx/

Ayman, M., & Ahmar, A. (2012). A Prototype Rule-based Expert System with an Object-Oriented Database for University Undergraduate Major Selection.

Ezenkwu, C. P., Johnson, E. H., & Jerome, O. B. (2017) Automated Career Guidance Expert System Using Case-Based Reasoning.

FAYYAD, U. M. (1996). *Data Mining and Knowledge Discovery: Making Sense out of Data.* IEEE Intelligent Systems, Vol. 11, No. 5, USA, ISSN: 0885-9000.

García, M. A., Quintana, L. M. (2015). Modelo para la elección de carrera basado en el Análisis de Factores Académicos y Educativos usando Minería de Datos. Revista de Tecnologías de Información Enero-Marzo 2015, Vol 2, No. 2. Pp 112-121. ECORFAN-Bolivia. ISSN 2410-4000

Han, J., Kamber, M., y Pei, J. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.

Han, J., Kamber, M., y Pei, J. (2011). *Data Mining Concepts and Techniques*. USA: Morgan Kaufmann.

Hendahewa, C., Dissanayake, M., Samaraweera, S., Wijayawickrama, N., Ruwanpathirana, A., & Karunananda, A. S. (2006). Artificial intelligence approach to effective career guidance. *Sri Lanka Association for Artificial Intelligence*.

Hernández, J., Ramírez, M. J., y Ferri C. (2004). *Introducción a la Minería de Datos*. Prentice Hall.

Hernández, J., Ramírez, M. J., y Ferri, C. (2008). *Introducción a la Mineria de Datos.* Madrid: Pearson Education.

Hernandez, P., J., Quintana., L., M. (2013). Análisis de la Influencia de las Inteligencias Múltiples en el Desempeño Académico de un Alumno aplicando Técnicas de Minería de Datos. Research in Computing Science. Vol. 67. pp 51-60. ISSN 1870-4069.

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

Mehmed, K. (2011). *Data mining concepts, models, methods, and algortihms* Second edition, Publisher: W.l.y, ISBN: 0470890452,University of Louisville.

Mundra, A., Soni, A., Sharma, S. K., Kumar, P., & Chauhan, D. S. (2014). Decision support system for determining: right education career choice. *ICC 2014-Computer Networks and Security*, 8-17.

Sistema Integral de Tutoría Académica. (2016). Recuperado de https://www.sita.uaemex.mx/tutoria/index_ok.html

Vera, C. M., Morales, C. R., y Soto, S. V. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. *Revista Iberoamericana de Tecnologías del/da Aprendizaje/Aprendizagem*, 109

Winston, O., & Lawrence, M. (2008). Career guidance using expert system approach. *Strengthening the Role of ICT in Development*, *123*.

Witten, H. I., y Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques witch Java Implementations.* Morgan Kaufmann, 2000.

Wu, X., & Kumar, V. (Eds.). (2010). *The top ten algorithms in data mining*. CRC Press.

Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/