Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), enero-febrero 2024,
Volumen 8, Número 1.

https://doi.org/10.37811/cl_rcm.v8i1

DETERMINACIÓN DE PERFILES PROFESIONALES MEDIANTE TÉCNICAS DE MINERÍA DE DATOS

DETERMINATION OF PROFESSIONAL PROFILES USING DATA MINING TECHNIQUES

María José Rodríguez Ojeda
Universidad Nacional de Loja, Ecuador

DOI: https://doi.org/10.37811/cl_rcm.v8i1.9783

Determinación de Perfiles Profesionales Mediante Técnicas de Minería de Datos

María José Rodríguez Ojeda[1]
mjrodriguezo@unl.edu.ec
https://orcid.org/0009-0009-4061-7842
Universidad Nacional de Loja
Ecuador

RESUMEN

Este artículo presenta la determinación de perfiles profesionales mediante técnicas de minería de datos enfocadas al sector educativo, estudio que busca servir de pauta en la toma de decisiones a nivel académico, profesional y como referente de investigación. Para ello se ha realizado el estudio de casos de éxito, recopilando herramientas como Rapid Miner que contiene los algoritmos y técnicas de minería de datos que permitan descubrir patrones o información oculta, seleccionadas en base a fuentes bibliográficas confiables e investigaciones continuas. El proceso tiene como base el análisis de las variables con enfoque cualitativo y cuantitativo más influyentes en el problema, que han determinado las fuentes de datos a utilizar en la busqueda de respuestas. En el desarrollo del proceso de minería, se aplicó la metodología CRISP-DM, la cual permite organizar el estudio mediante fases, sub fases y tareas la exploración de los datos, obteniendo modelos de datos que han sido evaluados y aplicados en un contexto real para seleccionar el más óptimo.

Palabras clave: minería de datos, perfil profesional, modelos, validación cruzada, fuentes de datos

[1] Autor principal
Correspondencia: mjrodriguezo@unl.edu.ec

Determination of professional profiles using data mining techniques

ABSTRACT

This paper presents the determination of professional profiles using data mining techniques focused on the educational sector, a study that seeks to serve as a pause in decision-making at an academic and professional level and as a research reference. For this purpose, a study of successful cases has been carried out, compiling tools such as Rapid Miner that contains the algorithms and data mining techniques that allow the discovery of patterns or hidden information, selected based on reliable bibliographic sources and continuous research. The process is based on the analysis of the variables with a qualitative and quantitative approach most influential in the problem, which have determined the data sources to be used in the search for answers. In the development of the mining process, the CRISP-DM methodology was applied, which allows the study to be organized through phases, subphases and data exploration tasks, obtaining data models that have been evaluated and applied in a real context to select the most optimal.

***Keywords*:** data mining;, professional profile, models, cross-validation, data sources

Artículo recibido 27 diciembre 2023
Aceptado para publicación: 30 enero 2024

INTRODUCCIÓN

La Universidad Nacional de Loja se ha caracterizado por impulsar la investigación, hacia la búsqueda de alternativas de solución a las problemáticas más preocupantes de hoy en día y de esta manera aportar al progreso académico-profesional. En el sector educativo, se ha podido evidenciar la realidad que enfrentan los egresados y profesionales, que año a año egresan y se titulan con el deseo de desarrollarse de forma plena en el ejercicio de su profesión.

Se puede evidenciar que las aptitudes adquiridas por un estudiante a lo largo de su formación académica en el aspecto cualitativo como: habilidades, capacidades, intereses y cuantitativamente el conocimiento reflejado en su record académico son algunos de los factores que determinan su perfil profesional. En vista del panorama presentado, la minería de datos surge como una alternativa de solución respaldándonos en el éxito de su aplicación en el ámbito educativo (Cobo, 2011; Eckert, 2013; Alvarado, 2017). Es por ello que mediante la aplicación de técnicas adecuadas se buscará determinar el perfil profesional de cada estudiante, el cual servirá como pauta en la toma de decisiones a nivel académico y profesional.

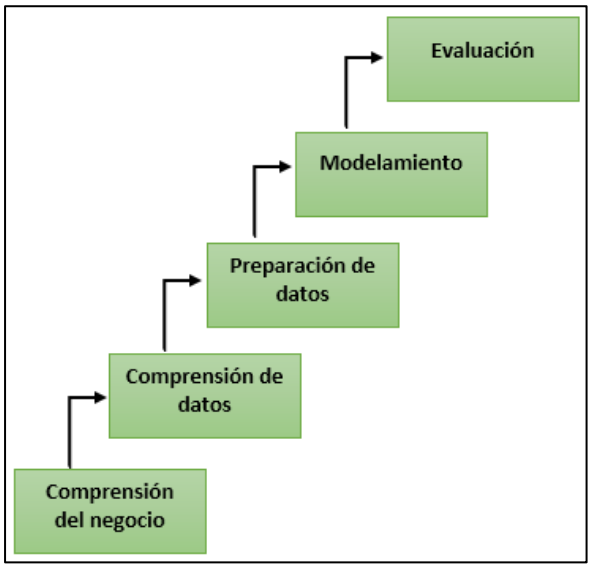
Para el desarrollo del presente trabajo se ha realizado el estudio de casos de éxito (Romero, 2005; García, 2008; Alcover, 2007, Marquez, 2015), recopilación de técnicas y herramientas de minería de datos (Astorga, 2013; Corso, 2012; Cubero, s.f.), con la finalidad de crear el escenario adecuado para solucionar el problema planteado y lograr la meta de minería. Las técnicas de minería de datos seleccionadas en base al análisis realizado son técnicas de clasificación en base a los árboles de decisión y reglas de inducción. Para el proceso de minería se ha construido dos estructuras de minería de datos, para datos agrupados y no agrupados con el fin de realizar un sin número de pruebas y encontrar los mejores resultados.

La organización del trabajo es la siguiente: en la Sección METODOLOGÍA, se explica la metodología utilizada para el desarrollo con todas sus fases. La Sección RESULTADOS Y DISCUSIÓN, muestra un caso de estudio realizado con la herramienta RapidMiner y aplicación de técnicas de minería de datos, detallando el proceso en base a la metodología, hasta obtener los resultados y la discusión que engloba la descripción de los resultados obtenidos. La Sección CONCLUSIONES. Finalmente se pueden encontrar en la Sección LISTA DE REFERENCIAS.

METODOLOGÍA

La metodología utilizada como modelo de referencia se denomina CRISP-DM, la cual está propiamente enfocada a proyectos de minería de datos. Está formada por varias fases, que han permitido de manera organizada detallar todo el proceso de minería de datos y cumplir con los objetivos establecidos (ver figura 1).

Figura 1. Ciclo de vida de un proyecto con CRISP-DM [13].



A continuación se ha descrito cada una de las fases, con la finalidad de comprender la estructura de esta metodología (Gallardo, 2013):

- Fase uno: Comprensión del negocio.- En esta tarea se describe los antecedentes o contexto inicial, los objetivos del negocio y los criterios de éxito.
- Fase dos: Comprensión de los datos.- Esta es la fase de la metodología donde se pretende mantener la información únicamente necesaria para realizar la minería de datos y familiarizarse con la información.
- Fase tres: Preparación de los datos.- Esta fase permite construir el conjunto de datos final, realizando tareas de selección de datos, selección de tablas, registros, y atributos a los, transformación de datos, cambios de formato, limpieza de datos, generación de variables adicionales, etc.
- Fase cuatro: Modelamiento.- En esta fase se describen las diferentes técnicas de modelado elegidas y se realiza su aplicación obteniendo los modelos y generando los resultados para su posterior evaluación.

- Fase cinco: Evaluación.- Corresponde a la fase del análisis de los resultados obtenidos en la minería, y validación de los resultados en un contexto real.

RESULTADOS Y DISCUSIÓN

A. Fase Uno: Comprensión del Negocio

La minería de datos es el proceso de exploración, análisis, extracción y refinamiento de grandes volúmenes de información de manera automatizada, con el fin de descubrir conocimiento, es decir información que ayude a la toma de decisiones (Hernández, 2006; Molina, 2006).

La meta depende del proyecto que se esté realizando, por ello en el presente proyecto al aplicar el proceso de minería de datos se busca que a partir de un conjunto de datos se descubran uno o varios modelos que determinen los perfiles profesionales mediante la aplicación de técnicas de minería de datos.

Para la elaboración del presente trabajo se ha especificado los recursos necesarios en cuanto al talento humano, hardware, software, y fuentes de datos necesarios para el desarrollo y culminación exitosa del mismo.

Objetivos del negocio

- Identificar los perfiles profesionales enfocados en la carrera de ingeniería en sistemas, a través de la formación de los estudiantes.
- Identificar los factores que determinan el perfil profesional de los estudiantes.
- Conocer los perfiles profesionales a los cuales se orienten un grupo de estudiantes.

B. Fase Dos: Comprensión de los datos.

Se ha realizado un análisis de las variables más influyentes en el problema con la finalidad de determinar las fuentes de datos a utilizar; dichas variables se las ha enfocado de dos formas: cualitativas y cuantitativas. Las cualitativas corresponden a los obtenidos de un test aplicado a la población objeto de estudio y las cuantitativas que engloban los records académicos de estudiantes egresados y graduados, estas fuentes de datos utilizadas son detalladas a continuación:

1. Datos del Sistema de Gestión Académica

Datos de los egresados de la carrera de ingeniería en sistemas respecto a las categorías académica y personal provenientes del Sistema de Gestión Académica de la institución creado en el 2008. Estos

datos se obtuvieron a través del Web Service para su posterior explotación.

2. Datos Históricos de los records académicos

Registros de los records académicos de los estudiantes egresados de la carrera de ingeniería en sistema, que se encuentran en los Libros físicos que están en poder de la secretaría del Área de la Energía, las Industrias y los Recursos Naturales no Renovables de la institución. Estos datos se han recopilado desde el año 2003, con el fin de completar la información académica de los egresados y graduados respecto de las notas de ciertos módulos que no constan en el SGA.

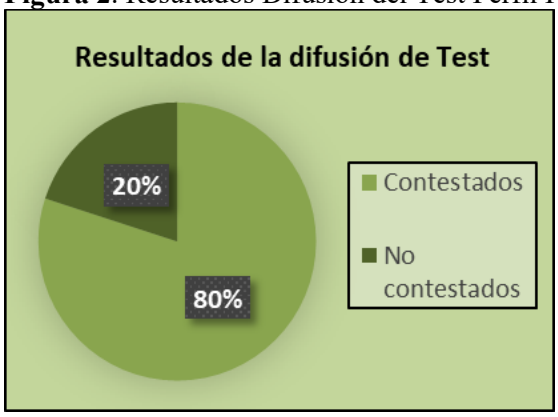
3. Test de habilidades, capacidades e intereses

El test ha sido desarrollado con el uso de la herramienta django, en base a los intereses, las capacidades, habilidades e interés de 8 perfiles planteados.

Explorar los datos

Como se ha mencionado en el transcurso del desarrollo del presente trabajo, se ha realizado un test enfocado a los egresados y graduados con objeto de recabar una nueva variable de suma importancia, obteniendo un 80% de acogida por parte de la población llegando a la conclusión que la difusión ha tenido éxito (ver figura 2).

Figura 2. Resultados Difusión del Test Perfil Profesional.



El objetivo del test desarrollado obtener el perfil profesional de los egresados, estos perfiles se los ha seleccionado realizando una consulta bibliográfica de los perfiles con las respectivas características, habilidades, capacidades que puede tener un estudiante al egresar de la carrera de ingeniería en constancia en el documento del rediseño de la carrera de ingeniería en sistemas de la UNL [14] y en la documentación de algunas universidades del país y del mundo.

Finalmente para realizar un filtro de estos perfiles se ha analizado las unidades de la malla curricular

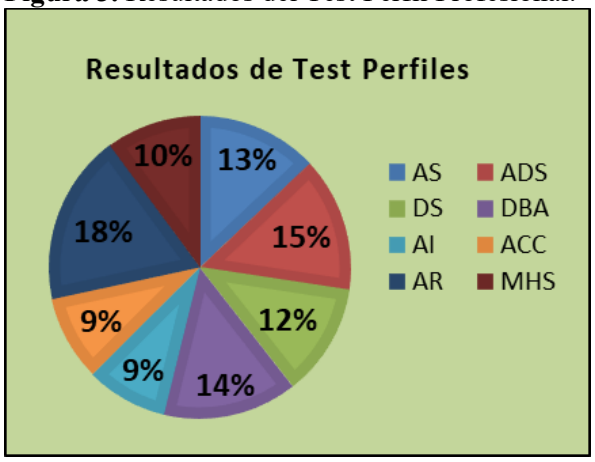
base y sus cambios a través de los diferentes periodos académicos. Por lo tanto los perfiles profesionales escogidos son 8 nombrados a continuación:

Tabla 1. Perfiles profesionales

Sigla	Perfil Profesional
AS	Analista de Sistemas de Información
ADS	Arquitecto y Diseñador de Software.
DS	Desarrollador de software
DBA	Administrador de Sistemas de Bases de Datos
AI	Auditor Informático
ACC	Administrador de Centros de computo
AR	Administrador de Redes computacionales.
MHS	Especialista en mantenimiento hardware y software.

En la figura 3 se observa de manera gráfica los resultados del test aplicado, donde el perfil profesional predominante es el perfil de 'Administrador de Redes computacionales' con el 18% lo cual no se diferencia demasiado del resto de perfiles.

Figura 3. Resultados del Test Perfil Profesional.



El perfil profesional posteriormente será tomado como nuestra variable dependiente dentro del proceso de minería de datos para realizar la predicción.

C. Fase tres: Preparación de los datos. (Selección, limpieza e integración de los datos).

Como se ha descrito anteriormente el estudio está enfocado en los egresados y graduados de la carrera de ingeniería en sistemas (2003-2013). De los cuales finalmente se cuenta con el 80% los cuales respondieron al test aplicado para determinar su perfil profesional. De este porcentaje final, se tomará un 72% con el objeto de realizar el proceso de minería y determinar las reglas que expresen en relación

con las unidades como se llega a un determinado perfil. El 28% restante se tomará para realizar la validación de las reglas obtenidas del proceso.

Construcción de Datos

En base a la meta de minería de datos y para realizar varias pruebas en búsqueda de los mejores resultados se han diseñado dos estructuras:

1. Estructura_uno: Datos no agrupados

Esta estructura está conformada por 67 variables; 66 variables correspondientes al conjunto total de unidades que existen entre todos los datos; una gran cantidad de los registros contienen los atributos de las unidades con valores nulos debido al cambio en las mallas curriculares antes mencionado y finalmente la estructura contiene la variable dependiente perfil_profesional obtenida del test aplicado clave para el proceso de predicción dentro de la minería de datos (ver Tabla II).

Tabla II. Tabla de variables utilizadas y fuente de providencia de estructura uno

Fuente	Variable
SGA y Registros de Libros Físicos	matemática, matematicas_discretas, fundamentos_basicos_computacion, calculo_diferencial, fisica_I, algebra_lineal, calculo_integral, metodologia_programacion, contabilidad_general, fisica_II, estadistica_I, programacion_I, electromagnetismo, estructura_datos_I, economía, estructura_datos_oo, estadistica_inferencial, contabilidad_costos, electronica_basica, diseno_gestion_bd, teoria_circuitos, ecuaciones_diferenciales, programacion_II, estructura_datos_II, estadistica_II, administracion_empresas, arquitectura_computadores, lenguaje_ensamblador, diseno_digital, analisis_diseno_sistemas_I, ingenieria_software_I, redes_I, proyectos_informaticos_I, teoria_telecomunicaciones, derecho_informatico, sistemas_informacion, analisis_diseno_sistemas_II, ingenieria_software_II, sistemas_operativos, redes_II, investigacion_operaciones, teoria_automatas, inteligencia_artificial, proyectos_informaticos_II, analisis_numerico, administracion_cc, auditoria_informatica, gestion_redes, sistemas_informacion_I, microprocesadores, lenguajes_formales, modelamiento_matematico, compiladores, sistemas_informacion_II, sistemas_expertos, mantenimiento_computadores, control_automatizado_asistido_c, anteproyectos_tesis, simulación, etica_profesional, legislacion_laboral, presupuestos_invesiones, diseno_asistido_computadores, aplicaciones_web, administracion_bd_mysql_uml, programacion_net.
Test Perfil Profesional	perfil_profesional

2. Estructura_dos: Datos agrupados

La estructura_dos está formada por 28 variables; 18 variables correspondientes a los grupos generados en base de 47 unidades de las 66 totales; con la finalidad de eliminar la gran cantidad de valores nulos existentes y por su relación entre sí, 9 atributos de unidades que no han sido alteradas manteniéndose de la estructura_uno y la variable dependiente denominada perfil_profesional obtenida por cada estudiante de manera personal en base al test aplicado (ver tabla III).

Tabla III. Tabla de variables utilizadas y fuente de providencia de estructura dos

Fuente	Variable
SGA y Registros de Libros Físicos	matemática, física, calculo, programación, estructura_datos, estadística, presupuestos_contabilidad, redes, proyectos_informaticos, sistemas_informacion, analisis_diseno_sistemas, ingenieria_software, arquitectura_computadores, electronica_telecomunicaciones, base_datos, lenguaje_ensamblador, derecho, teoria_automatas, inteligencia_artificial, administracion_centros_computo, auditoria_informatica, lenguajes_formales, compiladores, sistemas_expertos, anteproyectos_tesis, simulación, etica_profesional
Test Perfil Profesional	perfil_profesional

Cabe mencionar que las notas de las unidades tienen un valor comprendido entre 0 y 10, para realizar el proceso de minería se ha visto importante realizar la discretización de estos valores en las dos estructuras. (ver tabla IV):

Tabla IV. Discretización de las notas de cada unidad

Nro.	Nomenclatura	Rango
1	Regular	Menor a 7.5
2	Bueno	Entre 7.5 a 8.5
3	Excelente	Mayor a 8.5

D. Fase Cuatro: Selección de técnicas y generación de pruebas

En esta fase se ha utilizado la herramienta RapidMiner luego de una previa selección para realizar el proceso de minería y la aplicación de dos grupos de técnicas: Clasificación y las técnicas de reglas basadas en inducción, seleccionadas en base a los casos de estudio analizados. Dentro de cada grupo de técnicas se ha utilizado diferentes algoritmos descritos a continuación:

Técnicas de Clasificación

Este tipo de técnicas los algoritmos son robustos a datos con ruido, la función aprendida es representada

como un árbol, permitiendo obtener a su vez de forma visual las reglas de clasificación bajo las cuales operan los datos del experimento (Aluja, 2001; Vizcaino, 2008). La aplicación de estas técnicas has sido basadas en los algoritmos de árboles de decisión, específicamente 2 algoritmos: ID3 y CHAID.

Técnicas Basadas en Reglas de Inducción

Este tipos de algoritmos arrojan como resultados un sin número de reglas respecto al análisis de los datos, el proceso interno que realiza es la búsqueda de patrones, relaciones y características similares entre los datos. Estas reglas tienen la ventaja que son fáciles de entender (Servente, 2002; Hernandez 2003). Para la aplicación de estas técnicas se han escogido 6 algoritmos: JRip, Part, Ridor, Decisión Table, DTNB y NNge.

Diseño de pruebas

En las pruebas con el conjunto de entrenamiento se ha tomado un 72% de los datos mientras que el 28% restante será utilizado para la evaluación de los modelos. A su vez se realizará la evaluación de los modelos con el método de validación cruzada. Los resultados obtenidos al aplicar los distintos algoritmos han sido detallados en las siguientes tablas tanto para la estructura de datos no agrupados (ver tabla V), como para la estructura de datos agrupados (ver tabla VI).

Tabla V. Comparación del rendimiento de algoritmos con datos no agrupados

Clasificador	Modo de Prueba	Instancias bien clasificadas (%)	Instancias mal clasificadas (%)
CHAID	Conjunto de Entrenamiento	98.00%	2.00%
	Validación Cruzada	10.67%	89.33%
DECISION TABLE	Conjunto de Entrenamiento	33.33%	66.67%
	Validación Cruzada	10.00%	90.00%
DTNB	Conjunto de Entrenamiento	35.33%	64.67%
	Validación Cruzada	10.00%	90.00%
ID3	Conjunto de Entrenamiento	28.67%	71.33%
	Validación Cruzada	14.67%	85.33%
JRIP	Conjunto de Entrenamiento	94.00%	6.00%
	Validación Cruzada	12.00%	88.00%
PART	Conjunto de Entrenamiento	81.33%	18.67%
	Validación Cruzada	15.33%	84.67%
RIDOR	Conjunto de Entrenamiento	51.33%	48.67%
	Validación Cruzada	10.67%	89.33%
NNGE	Conjunto de Entrenamiento	100%	0.0%
	Validación Cruzada	10.67%	89.33%

La tabla V nos muestra que en pruebas de entrenamiento con datos no agrupados se puede observar que los algoritmos que presentan mejor rendimiento son: CHAID, PART, JRip y NNge; con estos resultados ya se logró tener una aproximación de los algoritmos con mejor rendimiento.

Tabla VI. Comparación del rendimiento de algoritmos con datos agrupados

Clasificador	Modo de Prueba	Instancias bien clasificadas (%)	Instancias mal clasificadas (%)
CHAID	Conjunto de Entrenamiento	96.67%	3.33%
	Validación Cruzada	14.00%	86.00%
DECISION TABLE	Conjunto de Entrenamiento	27.33%	72.67%
	Validación Cruzada	8.67%	91.33%
DTNB	Conjunto de Entrenamiento	24.67%	75.33%
	Validación Cruzada	9.33%	90.67%
ID3	Conjunto de Entrenamiento	87.33%	12.67%
	Validación Cruzada	14.00%	86.00%
JRIP	Conjunto de Entrenamiento	94.67%	5.33%
	Validación Cruzada	14.00%	86.00%
PART	Conjunto de Entrenamiento	82.67%	17.33%
	Validación Cruzada	8.00%	92.00%
RIDOR	Conjunto de Entrenamiento	42.67%	57.33%
	Validación Cruzada	12.67%	87.33%
NNGE	Conjunto de Entrenamiento	100.0%	0.0%
	Validación Cruzada	11.33%	88.67%

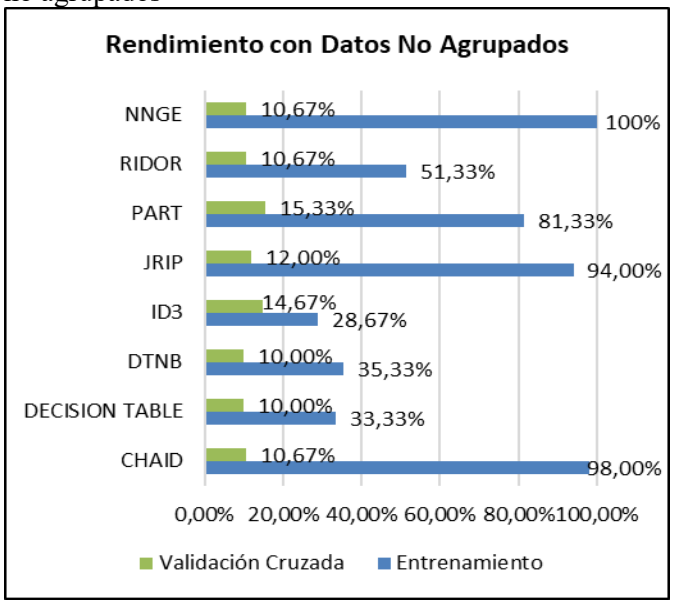
En pruebas de entrenamiento con datos agrupados se puede observar que los algoritmos que presentan mejor rendimiento son: CHAID, ID3, PART, JRip y NNge; con estos resultados ya se logró tener una aproximación de los algoritmos con mejor rendimiento.

Comparación general de la evaluación de modelos con datos agrupados y no agrupados:

En las pruebas de Validación Cruzada todos los algoritmos muestran porcentajes bajos en la clasificación debido a la poca cantidad de datos (Andrew, 2020) y a la presencia de outliers o valores atípicos que le restan calidad a los datos (Dapozo, 2000).

Podemos observar que al agrupar las unidades los porcentajes de clasificación varían muy poco entre las dos estructuras, el único algoritmo que logro tener un considerable aumento en pruebas de entrenamiento es el ID3, por lo tanto se ha notado que realizar la agrupación en los datos no es tan esencial, por ello se ha realizado la selección de los mejores algoritmos en base a datos no agrupados (ver figura 4), descartando los modelos de los datos agrupados.

Figura 4. Comparación Rendimiento en pruebas de Entrenamiento y Validación Cruzada con datos no agrupados



En la figura 4, se observa el rendimiento de los algoritmos, donde aparentemente NNge es el más óptimo con un porcentaje de rendimiento del 100%, el cual ha sido aplicado en una herramienta diferente a RapidMiner, debido a que es el único algoritmo que no soporta, esta nueva herramienta hace uso de la librería de weka.jar, y ha sido desarrollada en base al lenguaje de programación java [30]. Sin embargo a pesar del perfecto porcentaje de clasificación siendo un algoritmo robusto no se lo tomó en cuenta debido que no maneja correctamente la presencia de valores nulos, reflejado en que las reglas generadas por el algoritmo toman en cuenta los valores nulos o perdidos (Kaiser, 2014), por lo tanto son difíciles de interpretar y de utilizar.

Continuando con el análisis CHAID presenta el 98% de instancias bien clasificadas seguido de JRip con el 94%, siendo los mejores en esta característica así como en el análisis del rendimiento, lógica de reglas y medidas de error por lo que han sido seleccionados como los mejores hasta el momento.

E. Fase Cinco: Evaluación

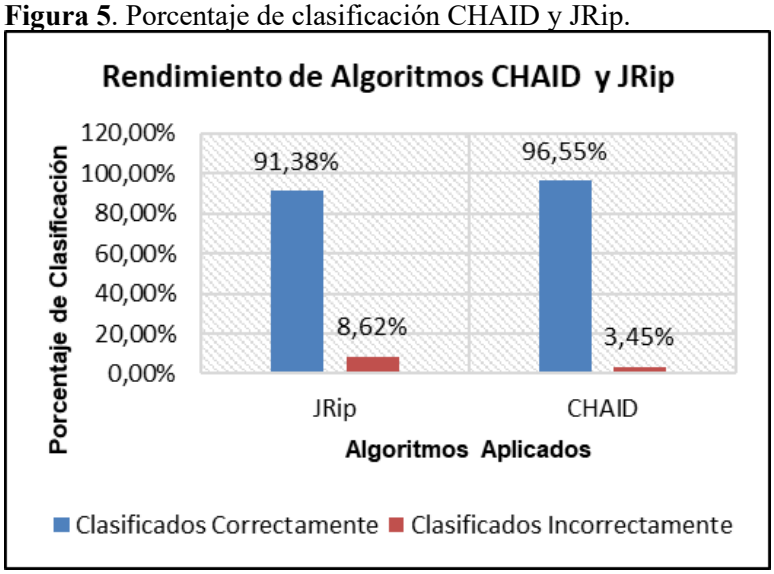
En este apartado se realizó un análisis del rendimiento con los mejores algoritmos en este caso CHAID y JRip, los datos utilizados para realizar estas pruebas corresponden al 28% restante de los recopilados inicialmente. Donde la evaluación realizada del rendimiento de los modelos, han presentado algunas variaciones del porcentaje de clasificación para cada perfil profesional, las cuales las observamos en la tabla VII.

Tabla VII. Resultados de la evaluación de los modelos generados con Chaid y JRip

PERFIL PROFESIONAL	Instancias bien clasificadas (%)	
	JRip	CHAID
Analista de Sistemas de Información	100	100
Arquitecto y Diseñador de Software.	100	100
Desarrollador de software	100	100
Administrador de Sistemas de Bases de Datos	100	86.67
Auditor Informático	100	100
Administrador de Centros de computo	100	100
Administrador de Redes computacionales.	64.29	100
Especialista en mantenimiento hardware y software.	100	100

Los resultados de clasificación mostrados en la table VII corresponden a los valores de cada matriz de confusión generada, donde se describe el porcentaje de precisión para las clases definidas como perfiles profesionales, en donde se puede observar que varían muy poco, ya que cada uno rinde el 100% en 7 perfiles, pero JRip presenta el 64.29% para el perfil Administrador de Redes computacionales y CHAID lo supera con el 86.67% en el perfil Administrador de Sistemas de Bases de Datos. A continuación realizaremos una evaluación global del porcentaje de clasificación (ver figura 5).

Figura 5. Porcentaje de clasificación CHAID y JRip.

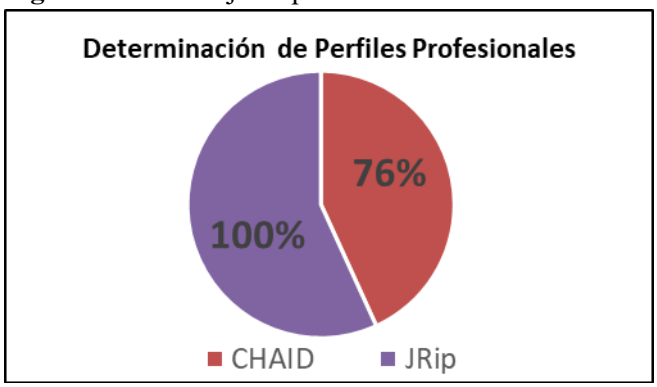


Como se puede observar en la figura 5 el algoritmo que mejor rendimiento presenta es CHAID logrando clasificar el 96.55% de las instancias mientras que el algoritmo JRip clasificó el 91.38%, siendo una leve diferencia, sin embargo en base a ello se ha seleccionado al algoritmo CHAID como el más óptimo para la predicción.

Aplicación de los modelos de minería de datos en un contexto real.

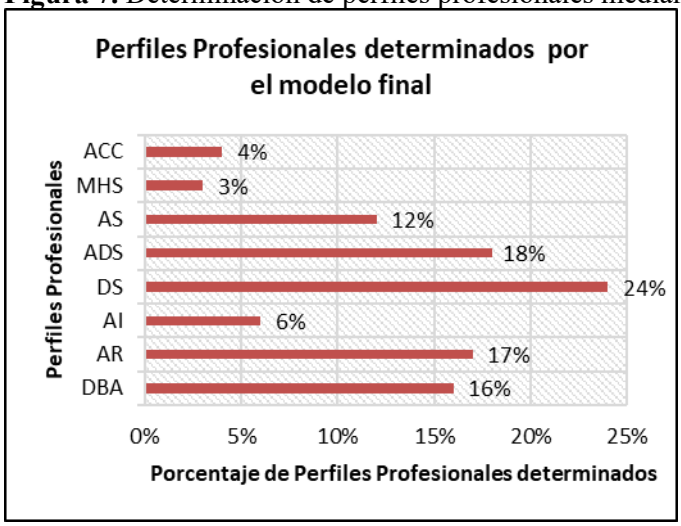
Finalmente se ha realizado la validación de los modelos generados al probarlos en un contexto real, para seleccionar el más óptimo. Los datos tomados corresponden a los últimos egresados de la carrera de ingeniería en sistemas año 2014. De esta validación se ha obtenido el porcentaje global de determinación de perfiles profesionales mediante los modelos CHAID y JRip (ver figura 6).

Figura 6. Porcentaje de predicción final de los modelos CHAID y JRip



Al observar los resultados de la figura 6 se demuestra que ha sido una sido fundamental la aplicación de los modelos en un contexto real, debido a que JRip realiza la determinación de los perfiles profesionales en un 100%, mientras que CHAID lo hace en un 76%, cambiando la perspectiva y validando el modelo generado por el algoritmo JRip perteneciente a las técnicas de reglas de inducción como el óptimo para la determinación de perfiles profesionales en la carrera de ingeniería en sistemas de la UNL. A continuación analizaremos el porcentaje de los perfiles profesionales determinados por el modelo final (ver figura 7).

Figura 7. Determinación de perfiles profesionales mediante el modelo JRip.



En la figura 7 se puede observar que el perfil profesional que más se destaca entre la población es el

perfil ‘Desarrollador de software’, seguido de ‘Arquitecto y Diseñador de software’, mientras que el perfil con menor porcentaje corresponde a ‘Especialista en mantenimiento Hardware y Software’, observando que la mayor parte de los egresados de la carrera de ingeniería en sistemas del año 2014 son desarrolladores.

CONCLUSIONES

La minería de datos hoy en día se ha convertido en una herramienta de vital importancia en el tratamiento, análisis y obtención de resultados del procesamiento de grandes cantidades de datos; para guiar la toma de decisiones tanto en las instituciones educativas como en todo tipo de organizaciones que así lo requieran.

La tarea más costosa a lo largo del proyecto, tanto en tiempo como esfuerzo, fue la recolección y armado de la Base de Datos, puesto que se obtuvieron de una fuente digital así como de una física; lo que conllevó al análisis, clasificación y limpieza de la información para luego agruparlos en una sola Base de datos.

Posterior al desarrollo del presente proyecto se puede concluir la importancia de determinar el perfil profesional de los egresados y graduados con los factores determinantes como: el récord académico que muestra el desempeño del estudiante a lo largo de la carrera y la parte cualitativa de cada individuo, ésta última se determinó mediante la aplicación de una encuesta que permitió posteriormente contrastar sus conocimientos con sus habilidades, intereses y capacidades que los hacen únicos y candidatos potenciales y competentes a diferentes Áreas y temáticas dentro del mundo laboral.

En el trabajo desarrollado se evidenció que el perfil profesional que más predomina en los últimos egresados del año 2014 es desarrollador de software cuyos conocimientos obtenidos en las aulas universitarias serán puestos en práctica en el desarrollo de su vida profesional, ésta información es muy útil para los futuros cambios que se realicen a nivel de la malla curricular y perfil de carrera.

Para el proceso de minería de datos se escogió los algoritmos ID3 y CHAID que pertenecen a las técnicas de clasificación basadas en árboles de decisión y los algoritmos JRip, PART, Ridor, Decisión Table, DTNB y NNge, pertenecientes al grupo de técnicas de reglas de inducción. Ya en el desarrollo y generación de modelos, los mejores algoritmos fueron CHAID y JRip los cuáles se hicieron con el 72% de los datos y con el 28% restante se hizo la evaluación de los mismos para verificar su validez,

donde CHAID resultó el más óptimo al clasificar el 96.55% de las instancias; mientras que JRip clasificó el 91,38%. Posterior a ello se realizó la aplicación de éstos algoritmos en un contexto real para validar y realizar sí la elección final, en donde JRip tuvo el mejor rendimiento en la predicción con el 100%; mientras que CHAID realizó la predicción del 76%, llegando a la conclusión que JRip es el modelo que se debe aplicar para la obtención de los perfiles profesionales.

REFERENCIAS BIBLIOGRAFICAS

Aluja Tómas. (2001). *La minería de datos entre la estadística y la inteligencia artificial.* Recuperado de: https://upcommons.upc.edu/bitstream/handle/2099/4162/article.pdf?sequence=4

Álvaro J. Galindo, Álvarez G. Hugo. (2017). *Minería de Datos en la Educación.* Recuperado de: https://repository.ucatolica.edu.co/server/api/core/bitstreams/bab2d285-cf09-4869-afbf-786092d38ca0/content.

Andrew Y. Ng. (2020). *Preventing "Overfitting" of Cross-Validation Data.* Recuperado de: https://dblp.org/pid/n/AndrewYNg.html.

Astorga Nathalia, Salinas Maruxa. (2013). Weka para minería de datos, 2013. Recuperado de: http://prezi.com/_gli7zt6vv0t/weka-para-mineria-de-datos-2013/ .

Cobo O. Angel, Rocha B. Rocío. (2011). *Selección de atributos predictivos del rendimiento académico de estudiantes en un modelo de B-Learning.* Recuperado de: https://www.edutec.es/revista/index.php/edutec-e/article/view/390/127

Corso Cynthia L, Gibellini Fabián. (2012). *Facultad Regional Córdoba/Universidad Tecnológica Nacional. Argentina. Uso de herramienta libre para la generación de reglas de asociación, facilitando la gestión eficiente de incidentes e inventarios.* Recuperado de: http://41jaiio.sadio.org.ar/sites/default/files/16_JSL_2012.pdf .

Cubero Juan C, Berzal Fernando. (s.f.). *Departamento de Ciencias de la computación, Universidad de Granada. Guión de prácticas de minería de datos, herramientas de minería de datos, KNIME.* Recuperado de: http://elvex.ugr.es/decsai/intelligent/workbook/D1%20KNIME.pdf

Dapozo Gladys, Porcel Eduardo, López María V, Bogado Verónica. (2000). *Técnicas de preprocesamiento para mejorar la calidad de los datos en un estudio de caracterización de ingresantes universitarios.* Recuperado de:

http://sedici.unlp.edu.ar/bitstream/handle/10915/20453/Documento_completo.pdf?sequence=1

Eckert Karina, Suénaga Roberto. (2013). *Aplicación de técnicas de Minería de Datos al análisis de situación y comportamiento académico de alumnos de la UGD.* Recuperado de: http://sedici.unlp.edu.ar/bitstream/handle/10915/27103/Documento_completo.pdf?sequence=1

Gallardo A. José A. (2013). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* Recuperado de: https://es.scribd.com/document/480481042/Metodologia-para-el-Desarrollo-de-Proyectos-en-Mineria-de-Datos-pdf

García S. Enrique, Romero M. Cristóbal, Ventura S. Sebastián, Castro Carlos. (2008*). Sistema recomendador colaborativo usando minería de datos distribuida para la mejora continua de cursos e-learning.* Recuperado de: http://rita.det.uvigo.es/200805/uploads/IEEE-RITA.2008.V3.N1.A3.pdf

Hernández O. José. (2006). *Minería De Datos. Otros Aspectos.* Recuperado de: https://www.redalyc.org/pdf/925/92502902.pdf

Hernández O. José. (2003). *Práctica 2 de minería de datos, profundizando en el Clementine.* Recuperado de: https://docplayer.es/10288244-Practica-2-de-mineria-de-datos-profundizando-en-el-clementine.html

Instituto Tecnológico Superior de Irapuato. Ingeniería en Sistemas Computacionales. Carretera Irapuato - Silao Km. 12.5, C.P. 36821 Irapuato, Guanajuato, México. Recuperado de: http://www.itesi.edu.mx/Oferta%20Educativa/Nivel%20Superior/IngSistemas.html.

Kaiser Jiri. (2014). *Dealing with Missing Values in Data.* Recuperado de: https://www.researchgate.net/publication/304500093_Dealing_with_Missing_Values_in_Data

Márquez V. Carlos, Romero M. Cristóbal, Ventura S. Sebastián. (2015). *Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos.* Recuperado de: https://dialnet.unirioja.es/servlet/tesis?codigo=66343

Molina L. José M, Garcia H. Jesús. (2006). *Técnicas de Análisis de Datos.* Recuperado de: https://matema.ujaen.es/jnavas/web_recursos/archivos/weka%20master%20recursos%20natural

es/apuntesAD.pdf

Paz A. Henry P. (2011). *Publicaciones Henry. Weka with Data Mining done in java.* Recuperado de: http://publicacioneshenry.wordpress.com/ .

Romero M. Cristóbal, Ventura S. Sebastián, Hervás M. Cesar. (2005). *Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web.* Recuperado de: http://www.lsi.us.es/redmidas/CEDI/papers/189.pdf .

R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch, M. Celma, C. Ferri, J. Hernández-Orallo, L. Iniesta, J. Más, M. J. Ramírez-Quintana, A. Robles, J. M. Valiente, M. J. Vicent, L. R. Zúnica. (2007). *Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos.* Recuperado de: https://www.researchgate.net/publication/337285746_Analisis_del_rendimiento_academico_en_los_estudios_de_informatica_de_la_Universidad_Politecnica_de_Valencia_aplicando_tecnicas_de_mineria_de_datos .

Servente Magdalena. (2002). *Algoritmos TDIDT aplicados a la minería de datos inteligente*. Recuperado de: https://docplayer.es/10351628-Algoritmos-tdidt-aplicados-a-la-mineria-de-datos-inteligente-tesis-de-grado-en-ingenieria-informatica.html .

Universidad Nacional de Loja. Reglamento de Rediseño de la carrera de Ingeniería en Sistemas. Perfil de Egreso del Ingeniero en Sistemas. Recuperado de: https://drive.google.com/file/d/0By4B2OXR-vftTHpZV3F0a2NiNWs/edit?usp=sharing .

Universidad Peruana de Ciencias e Informática. Ingeniería de sistemas e Informática. Recuperado de: http://www.upci.edu.pe/facultades.php?ac=ci_ing&op=ing .

Universidad Católica del Ecuador. Facultad de Ingeniería. Ingeniero en Sistemas. Recuperado de: http://www.puce.edu.ec/portal/content/Ingenier%C3%ADa%20en%20Sistemas/292;jsessionid=248B1730885DCEC54DFC19EC224DCF7F.node0?link=oln30.redirect .

Facultad de Sistemas, México. Ingeniero en Sistemas Computacionales. Recuperado de: http://www.sistemas.uadec.mx/index.php/carreras/isc .

Universidad del Valle. Cede central Cochabamba - Bolivia. Ingeniería de Sistemas Informáticos.

Recuperado de: http://www.univalle.edu/index.php/facultades/informatica/sistemas.

Universidad de las Fuerzas Armadas. (ESPE). Ingeniería de Sistemas e Informática. Recuperado de: http://www.espe.edu.ec/portal/portal/main.do?sectionCode=107 .

Universidad del Valle. Sede Central Cochabamba, Bolivia. Ingeniería de Sistemas Informáticos. Recuperado de: http://www.univalle.edu/index.php/facultades/informatica/sistemas.

Vizcaino G. Paula A. (2008). *Aplicación de técnicas de inducción de Árboles de Decisión a problemas de clasificación mediante el uso de weka (Waikato Environment For Knowledge Analysis).* Recuperado de: https://www.studocu.com/es/document/universidad-rey-juan-carlos/economia/arbol-weka/17561912