VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGÌSTER EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TEMA: DATA MINING Y ANÁLISIS DE DATOS APLICADO AL MINISTERIO DEL CONOCIMIENTO Y TALENTO HUMANO EN ECUADOR DEL 1993 A 2015

AUTOR: GUAYASAMÍN GUANGA, CÉSAR ADRIÁN

DIRECTOR: ING. CAMPAÑA ORTEGA, EDUARDO MAURICIO

SANGOLQUÍ

2018

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**DATA MINING Y ANÁLISIS DE DATOS APLICADO AL MINISTERIO DEL CONOCIMIENTO Y TALENTO HUMANO EN ECUADOR DEL 1993 A 2015**", fue realizado por el señor ***CÉSAR ADRIÁN GUAYASAMÍN GUANGA***, el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 16 de mayo del 2018

ING. MAURICIO CAMPAÑA

C.C.: 1708856701

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORÍA DE RESPONSABILIDAD

Yo, ***CÉSAR ADRIÁN GUAYASAMÍN GUANGA***, con cédula de ciudadanía n° 1715897680, declaro que el contenido, ideas y criterios del trabajo de titulación : "**DATA MINING Y ANÁLISIS DE DATOS APLICADO AL MINISTERIO DEL CONOCIMIENTO Y TALENTO HUMANO EN ECUADOR DEL 1993 A 2015**" es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 16 de mayo del 2018

CÉSAR ADRIÁN GUAYASAMÍN GUANGA

C.C.: 1715897680

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN

Yo, ***CÉSAR ADRIÁN GUAYASAMÍN GUANGA***, con C. C. n°. 1715897680 autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"DATA MINING Y ANÁLISIS DE DATOS APLICADO AL MINISTERIO DEL CONOCIMIENTO Y TALENTO HUMANO EN ECUADOR DEL 1993 A 2015"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 16 de mayo del 2018

CÉSAR ADRIÁN GUAYASAMÍN GUANGA

C.C.: 1715897680

DEDICATORIA

Deseo dedicar esta tesis principalmente a Dios Padre, Dios Hijo y Dios Espíritu Santo, por estar presente a lo largo de mi vida.

A mi familia: padres Aníbal y Emma, mi hermana Mireya, por su apoyo incondicional durante todos estos años.

A mis amigos y amigas quienes me han acompañado a lo largo de este camino.

AGRADECIMIENTO

A Dios Padre, Dios Hijo y Dios Espíritu Santo, por estar presente a lo largo de mi vida.

A mi familia: padres Aníbal y Emma, mi hermana Mireya, por su apoyo incondicional durante todos estos años.

A la Universidad de las Fuerzas Amadas ESPE, por las facilidades brindadas para desarrollar el presente estudio.

A mis profesores, cuyos conocimientos han servido de base para llegar a culminar mi proyecto de grado.

A mis amigos y amigas quienes me han acompañado y ayudado a lo largo de este camino.

ÍNDICE DE CONTENIDO

CERTIFICADO DEL DIRECTOR ..... **I**

AUTORÍA DE RESPONSABILIDAD ..... **II**

AUTORIZACIÓN ..... **III**

DEDICATORIA ..... **IV**

AGRADECIMIENTO ..... **V**

ÍNDICE DE CONTENIDO ..... **VI**

ÍNDICE DE TABLAS ..... **IX**

ÍNDICE DE FIGURAS ..... **X**

RESUMEN ..... **XI**

ABSTRACT ..... **XII**

CAPÍTULO 1 ..... **1**

INTRODUCCIÓN E INFORMACIÓN GENERAL **1**

- **1. Introducción** **1**
- **2. Problemática** **3**
- **3. Justificación** **4**
- **4. Objetivos** **5**
- **5. Alcance** **6**
- **6. Sector de Conocimiento y Talento Humano en el Ecuador** **6**

CAPÍTULO 2 ..... **9**

MARCO TEÓRICO **9**

1. Definiciones **9**

2. Características de la minería de datos **11**

3. Beneficios de la minería de datos **12**

4. El proceso de Descubrimiento de Conocimiento en bases de Datos **13**

5. Técnicas de Análisis **13**
5.1 Análisis 13
5.1 Modelos de minería de datos 15
5.2 Agrupamiento o clustering. 16
5.3 Reglas de asociación 18

CAPÍTULO 3 ..... **20**

METODOLOGÍA **20**

1. Determinación de la metodología de desarrollo a aplicar **20**

2. Comparación cualitativa de metodologías minería de datos **21**

3. Fases de la metodología CRIPS-DM **24**

4. Herramientas de minería de datos **27**

5. Comparación cualitativa de herramientas de minería de datos **28**

CAPÍTULO 4 ..... **31**

MINERÍA **31**

1. Exploración y selección de datos **31**

2. Depuración y procesamiento **33**

3. Minería de datos **35**
3.1 Creación de clúster 36
3.2 Análisis de los clústers 38

4. Reconocimiento de patrones de comportamiento **47**

5. Resultados y análisis **50**

5.1 Patrones encontrados para el clúster 1 50
5.2 Patrones encontrados para el clúster 2 54
5.3 Patrones encontrados para el clúster 3 58
5.4 Patrones encontrados para el clúster 4 62
5.5 Patrones encontrados para el clúster 5 66

CONCLUSIONES Y RECOMENDACIONES 72

1. Conclusiones 72

2. Recomendaciones 74

BIBLIOGRAFÍA 75

ÍNDICE DE TABLAS

Tabla 1 *Comparación de las fases de las metodologías* 22
Tabla 2 *Cuadro comparativo de metodologías* 24
Tabla 3 *Cuadro comparativo de herramientas de minería de datos* 29
Tabla 4 *Clúster 1, patrón 1* 52
Tabla 5 *Clúster 1, patrón 2* 52
Tabla 6 *Clúster 1, patrón 3* 53
Tabla 7 *Clúster 1, patrón 4* 53
Tabla 8 *Clúster 2, patrón 1* 56
Tabla 9 *Clúster 2, patrón 2* 56
Tabla 10 *Clúster 2, patrón 3* 57
Tabla 11 *Clúster 2, patrón 4* 57
Tabla 12 *Clúster 3, patrón 1* 60
Tabla 13 *Clúster 3, patrón 2* 60
Tabla 14 *Clúster 3, patrón 3* 61
Tabla 15 *Clúster 3, patrón 4* 61
Tabla 16 *Clúster 3, patrón 5* 62
Tabla 17 *Clúster 4, patrón 1* 64
Tabla 18 *Clúster 4, patrón* **2** 65
Tabla 19 *Clúster 4, patrón 3* 65
Tabla 20 *Clúster 4, patrón 4* 66
Tabla 21 *Clúster 4, patrón 5* 66
Tabla 22 *Clúster 5, patrón 1* 69
Tabla 23 *Clúster 5, patrón 2* 69
Tabla 24 *Clúster 5, patrón 3* 70
Tabla 25 *Clúster 5, patrón 4* 70
Tabla 26 *Clúster 5, patrón 5* 71

ÍNDICE DE FIGURAS

Figura 1. Entidades coordinadas 8
Figura 2. Etapas del proceso KDD 13
Figura 3. Modelo predictivo. 15
Figura 4. Modelo descriptivo. 16
Figura 5. Complejidad computacional 18
Figura 6. Top de las metodologías usadas en proyectos de minería de datos 23
Figura 7. Ciclo de vida de CRISP–DM 25
Figura 8. Top de las herramientas de minería de datos usadas en el 2015 29
Figura 9. Proceso ETL 34
Figura 10. Codo de Jambú 37
Figura 11. Distribución de grupos 37
Figura 12. Comparación de clústers 40
Figura 13. Remuneraciones por grupos 41
Figura 14. Actividades económicas por grupos 42
Figura 15. Áreas de estudio por grupos 43
Figura 16. Sectores por grupos 44
Figura 17. Rangos de edad por grupos 45
Figura 18. Áreas y actividades económicas del grupo 4 46
Figura 19. Áreas y actividades económicas del grupo 1 46
Figura 20. Áreas y actividades económicas del grupo 5 47
Figura 21. Reglas generadas para el clúster 1 51
Figura 22. Reglas generadas para el clúster 2 55
Figura 23. Reglas generadas para el clúster 3 59
Figura 24. Reglas generadas para el clúster 4 63
Figura 25. Reglas generadas para el clúster 5 68

RESUMEN

Este estudio propone el análisis y minería de datos de los bachilleres y profesionales del Ecuador; y su relación con las actividades económicas, aplicado al Sector del Conocimiento y Talento Humano del Ecuador desde el año 1993 al 2015". Dicho estudio se realizó mediante la recopilación de datos del repositorio central de la entidad rectora, misma que se creó con el objetivo de desarrollar políticas públicas, que mejoren la gestión de conocimiento y la matriz productiva del país. Para el desarrollo de la investigación se utilizó la metodología CRISP-DM, por sus ventajas y flexibilidad a la hora de elaborar el diseño y construcción. Para la implementación del estudio se utilizó las herramientas de software libre R-Studio, Pentaho Data Integration y Postgresql. En la minería de datos se usó el algoritmo de clustering k-means y el algoritmo de asociación a priori. Entre los principales hallazgos y resultados está la generación dinámica de cinco clústers que muestran la relación entre el sector la educación y las actividades económicas, a los que se les aplicó las reglas de asociación que generó nuevo conocimiento útil para la toma de decisiones, que está descrito en el capítulo cuatro.

PALABRAS CLAVE

- **MINERÍA DE DATOS**
- **CRISP-DM**
- **R-STUDIO**
- **ASOCIACIÓN**
- **CLUSTERING**

ABSTRACT

This study proposes the analysis and mining of data of high school graduates and professionals of Ecuador; and its relationship with economic activities, applied to the Knowledge and Human Talent Sector of Ecuador from 1993 to 2015, this study was carried out through the collection of data from the central repository of the governing entity, which was created with the objective of to develop public policies that improve knowledge management and the productive matrix of the country. For the development of the research, the CRISP-DM methodology was used, due to its advantages and flexibility when it comes to design and construction. The free software tools R-Studio, Pentaho Data Integration and Postgresql were used to implement the study. In data mining, the k-means clustering algorithm and the a priori association algorithm were used. Among the main findings and results is the dynamic generation of five clusters that show the relationship between the education sector and economic activities, to which the rules of association that generated new knowledge useful for decision-making, which is described in chapter four.

KEY WORDS

- **DATA MINING**
- **CRISP-DM**
- **R-STUDIO**
- **ASSOCIATION**
- **CLUSTERING**

CAPÍTULO 1

INTRODUCCIÓN E INFORMACIÓN GENERAL

En el capítulo uno se muestra información del Sector del Conocimiento y Talento Humano, que es parte de la fase de comprensión del negocio de la metodología CRISP-DM. Esto da una idea y sirve de ayuda para posteriormente aplicar las técnicas de minería de datos.

1. Introducción

Contar con información oportuna en la actualidad, se ha convertido en la base para tomar decisiones importantes, que marcaran la vida de las personas e instituciones y con ello de la sociedad. Sobre todo, en las decisiones que vienen desde el ámbito educativo y económico, cuya relación podría definir que un pueblo, región o país esté dentro de las sociedades desarrolladas del futuro.

Con el propósito de impulsar la educación en el Ecuador, los Gobiernos Nacionales preocupados por la enseñanza de sus ciudadanos, crearon instituciones de control y coordinación para fortalecer a las entidades del sector de conocimiento y talento humano, con el objetivo de dar seguimiento al cumplimiento de las políticas públicas, lo que permitió consolidar la información de las bases de datos de las diferentes entidades coordinadas, en un repositorio central, el mismo que es usado para la toma de decisiones y aplicación de las políticas públicas.

En el País se han realizado estudios sobre la educación. Así por ejemplo el trabajo de tesis en el que se analiza la relación existente entre la desigualdad en la distribución de los ingresos y la educación en el Ecuador en el periodo 2000-2014 (Izurieta Ballesteros, 2016)., y la tesis que analiza el impacto en la calidad de la educación superior como consecuencia de la aplicación de la normativa vigente 2010 – 2015, en la renovación del personal académico de las Universidades y Escuelas Politécnicas (IES) (Pazmiño Lucio, 2016).

Toda esta información histórica, ha sido útil para explicar el pasado, entender el presente y tener una idea del futuro de la realidad ecuatoriana, en el ámbito educativo y productivo. Para ampliar esta visión es necesario explotar los datos y detectar patrones que a simple vista no están presentes. Esto ayuda a mejorar la toma de decisiones acertadas, para atender a los sectores de la sociedad, que hasta el momento no se los había tomado en cuenta.

El estudio desarrolló un proceso de descubrir conocimiento, para lo cual se utilizó minería de datos, donde se relacionó la formación, media y superior, con las actividades económicas desempeñadas, para luego aplicar reglas de asociación, que encontraron patrones u ocurrencias en los datos analizados. Este proceso investigativo fue totalmente transparente para el usuario final, sin necesidad de que este sea experto en minería de datos o que pierda tiempo en la interpretación de los resultados (Hasperué, 2012).

Para extraer el conocimiento se utilizó la metodología CRIPS-DM y herramientas de software libre como R., todo este proceso estuvo enmarcado dentro del KDD (proceso de extracción de conocimiento), lo que ayudo a entender el cómo se obtuvieron los resultados.

2. Problemática

Las instituciones del Sector del Conocimiento y Talento Humano en Ecuador que comprenden, entidades coordinadoras y adscritas, preocupadas por mejorar las políticas públicas, han reunido bases de datos, con el objetivo de analizarlas y ser un apoyo para la toma de decisiones. Como ejercicio de esta consolidación de información se realizó la publicación “Prospectiva 2035” (MCCTH, 2017)., que es la proyección hasta el año 2035, del estado del sector educativo y su relación con el cambio de la matriz productiva. Bajo este contexto, en el presente estudio de investigación se desarrolló un proceso que permitió descubrir nuevo conocimiento, oculto no visible a simple vista y que está presente en las bases de datos antes señaladas, en los que:

- Utilizó minería de datos para buscar patrones u ocurrencias a fin de obtener información, acerca de la relación entre el trabajo y la formación estudiantil de bachilleres y profesionales;
- Aportó con conocimiento para la toma de decisiones, que ofreció respuestas a preguntas complejas;

- Aportó con nuevo conocimiento por medio de la metodología CRIPS-DM, la misma que permitió llevar a cabo el proceso de minería de datos de manera sistemática y que a su vez ayudó a entender como se lo obtuvo.

3. Justificación

Una educación de calidad es la base para mejorar la vida de las personas y alcanzar un desarrollo sostenible; preocupados por ello la Organización de Naciones Unidas (ONU), lo ha incluido como punto importante dentro de los "Objetivos del Desarrollo Sustentable" del 2015 (ONU, 2017). En el que se indica que la educación debe ir atada al servicio de la ciudadanía, con el objetivo de erradicar la pobreza para mejorar la salud, promover la industria, el desarrollo científico, conservar la naturaleza, etc.; al tener un amplio abanico de posibilidades, los gobiernos se vieron en la necesidad de tener información consolidada de manera eficiente y eficaz. En el Ecuador, la carta magna vigente desde el año 2008 establece en sus artículos:

> Art. 26.- La educación es un derecho de las personas a lo largo de su vida y un deber ineludible e inexcusable del Estado. Constituye un área prioritaria de la política pública y de la inversión estatal, garantía de la igualdad e inclusión social y condición indispensable para el buen vivir y (Constituyente, 2008).
>
> Art. 27. […] La educación es indispensable para el conocimiento, el ejercicio de los derechos y la construcción de un país soberano, y constituye un eje estratégico para el desarrollo nacional. (Constituyente, 2008).

Como parte de las políticas del país, establecidas en el Plan Nacional del Buen Vivir (2013-1017) en el “Objetivo 4: Fortalecer las capacidades y potencialidades de la ciudadanía” (Senplades, 2013)., se crearon ministerios coordinadores y entidades adscritas dentro del sector del de Conocimiento y Talento Humano, con el objetivo de dar seguimiento a las políticas públicas implementadas.

En base a expuesto anteriormente, el presente estudio es un aporte para encontrar patrones u ocurrencias, entre el trabajo y la formación estudiantil, lo que permitió descubrir información útil que a simple vista no se visualizaba, esto se logró aplicando técnicas de minería de datos, que permitieron identificar comportamientos recurrentes, que generaron nuevo conocimiento de valor para la toma de decisiones.

4. Objetivos

Analizar y diseñar una solución de minería de datos para el Sector del Conocimiento y Talento Humano del Ecuador, basado en la metodología CRIPS-DM, que permita descubrir patrones de comportamiento o nuevo conocimiento útil, para la toma de decisiones acertadas.

Objetivos específicos

- Implementación de un proceso que permita realizar la minería de datos, que sirva de insumo para la creación de políticas educativas;
- Analizar la composición social; es decir el nivel académico de los profesionales y bachilleres que conforman el sector productivo ecuatoriano, con sus actividades económicas;

- Generar nuevo conocimiento que sirva para la toma de decisiones en los sectores: educativos y, económico.

5. Alcance

En el presente estudio se realizó el análisis y minería de datos a las fuentes de información que corresponden al Sector del Conocimiento y Talento Humano del Ecuador, en el periodo del año 1993 al 2015, se tomó como pilar principal la base de datos de los bachilleres del periodo mencionado. Se aplicaron algoritmos y técnicas de clasificación y asociación, para obtener el nuevo conocimiento. La metodología aplicada durante todo el proceso de minería de datos fue CRIPS-DM versión dos, con la herramienta de software libre R - Studio.

6. Sector de Conocimiento y Talento Humano en el Ecuador

El Ministerio Coordinador de Conocimiento y Talento Humano (MCCTH), fue creado el ocho de abril de 2011, mediante decreto ejecutivo de número 726, con sede en Quito - Ecuador, y ubicado entre las avenidas Patria y diez de agosto en el antiguo edificio Banco de Préstamos.

El Ministerio (MCCTH) pertenece al Consejo Sectorial del Conocimiento y Talento Humano, que es un espacio en donde se coordinan las políticas públicas, en conjunto con las instituciones adscritas. Las instituciones miembros trabajan en la gestión del conocimiento y en el cambio de la matriz productiva.

De la página web institucional http://www.conocimiento.gob.ec, fue de donde se extrajo la información correspondiente a la misión y objetivos, que se detalla a continuación:

> "Misión: Concertar, coordinar y evaluar la formulación, ejecución, control y seguimiento de políticas públicas, programas y proyectos, a través del apoyo y fortalecimiento a la gestión institucional de las entidades del Sector de Conocimiento y Talento Humano para consolidar la Sociedad de Conocimiento justa y solidaria (MCCTH, conocimiento.gob.ec, 2017).
>
> Objetivos Estratégicos: i) Incrementar el fortalecimiento a la gestión de las entidades que forman parte del Sector de Conocimiento y Talento Humano con la finalidad de propender al mejoramiento en la provisión de servicios de calidad con calidez, cobertura y pertinencia cultural y/o territorial; ii) Incrementar mecanismos de coordinación sectorial e intersectorial que propicien sinergias y complementariedades para la generación, aplicación y circulación del saber y el conocimiento con la producción nacional; iii) Incrementar la efectividad de las políticas públicas, de la planificación institucional y de la inversión pública; iv) Incrementar la efectividad del control, seguimiento y evaluación de la gestión en las instituciones coordinadas; v) Incrementar la calidad de la información intersectorial" (MCCTH, conocimiento.gob.ec, 2017).

La estructura de las entidades adscritas y coordinadas se obtuvo de la intranet institucional, que está representada en la figura 1, como se muestra a continuación:

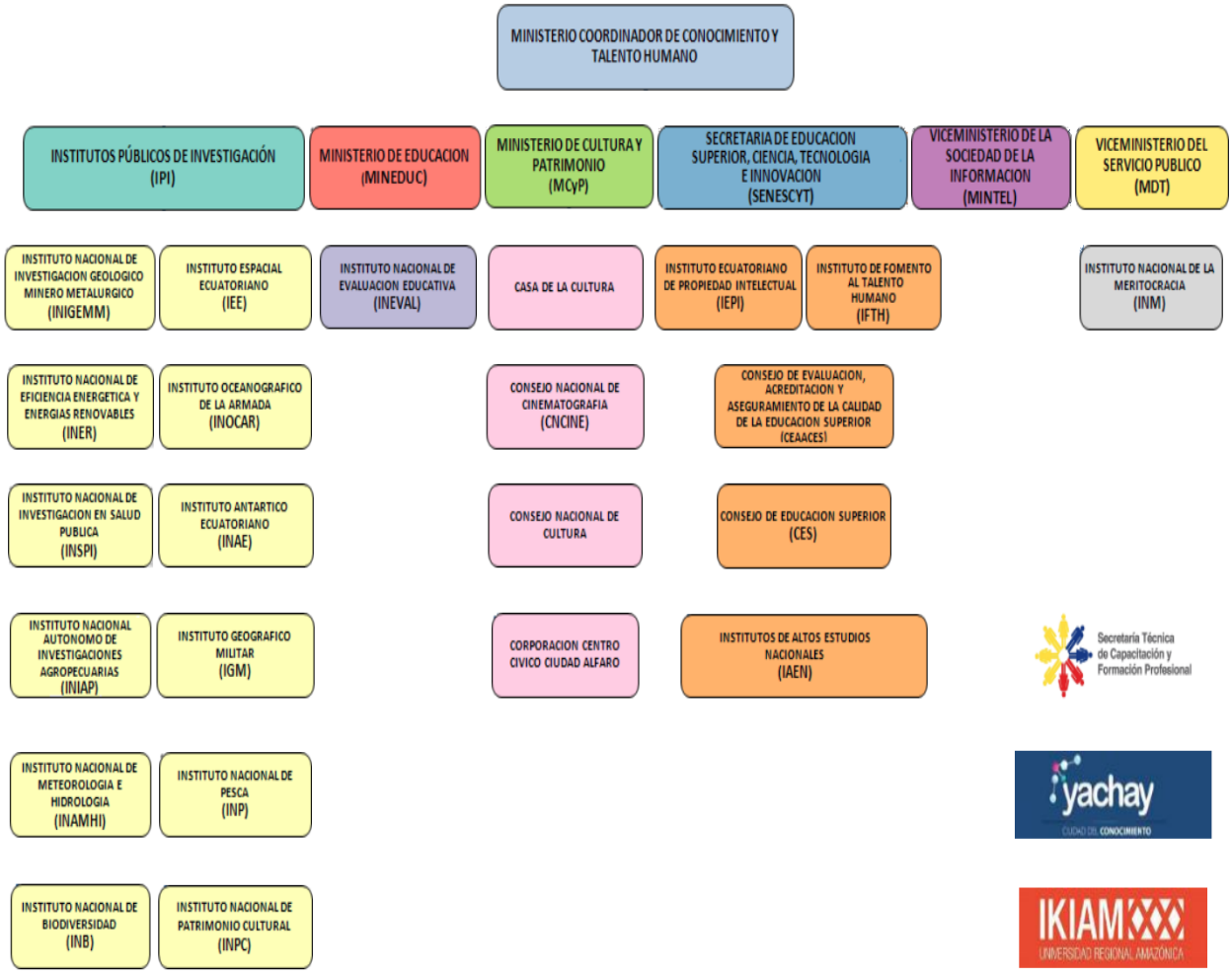

Figura 1. Entidades coordinadas
Fuente **(Subsecretaría de Gestión de Información, 2017)**

CAPÍTULO 2

MARCO TEÓRICO

En este capítulo se dan definiciones y conceptos necesarios para realizar la minería de datos. Aquí se explica los algoritmos de clustering como el de k-means y el de asociación como el A piori, utilizados en el capítulo cuatro.

1. Definiciones

Para una mejor comprensión del contexto que abarca este estudio, es necesario definir algunos conceptos que apalanquen esta investigación y que describen a continuación:

Minería de datos: Es un proceso sistemático para extraer de una gran cantidad de datos, información oculta, útil y de importancia estratégica que permite la toma de decisiones basada en conocimiento y un mejor el entendimiento de los fenómenos. Este proceso comprende las siguientes fases: La definición del problema, la preparación y selección de datos, el procesamiento de los datos, la generación y la validación de los modelos (MSDN, 2008).

CRISP-DM (CRoss-Industry Standard Processfor Data Mining) (Chapman y otros 2000).: es una metodología para minería de datos que fue presentada por el consorcio CRISP-DM, encabezado por SPSS Inc. (Estados Unidos); ésta se ha convertido en un

estándar, luego de ser liberada para su empleo y desarrollo por parte de la comunidad internacional.

Base de datos: Una base de datos es una colección de datos relacionados, diseñada, construida y rellenada con datos para un propósito específico (Shamkant, 2002).

Almacén de datos: Un Almacén de Datos (data warehouse) es una base de datos que integra datos procedentes de uno o varios sistemas de información de una organización, generalmente orientado a la toma de decisiones (Gutiérrez, 2006).

Software libre: Se basa en las libertades otorgadas a los usuarios, como son: libertad de ejecutar, copiar, distribuir, estudiar, modificar y mejorar el software. A continuación, se describen esas libertades:

- "La libertad de ejecutar el programa como lo desee, con cualquier propósito (libertad 0);
- La libertad de estudiar el funcionamiento del programa y adaptarlo a sus necesidades (libertad 1). El acceso al código fuente es un prerrequisito para esto;
- La libertad de redistribuir copias para ayudar a los demás (libertad 2).
- La libertad de mejorar el programa y de publicar las mejoras, de modo que toda la comunidad se beneficie (libertad 3). El acceso al código fuente es un prerrequisito para esto" (Free Software Foundation).

General Public Licence (GNU): es una licencia de derechos de autor usada en el mundo software libre, con el propósito de proteger los intentos de apropiación de quienes quieran incumplir, adueñarse o privar de las libertades expuestas en esta licencia (GNU, 2018). Actualmente está en su versión 3 y sus autores son: Richard Stallman y la Free Software Foundation.

2. Características de la minería de datos

Busca encontrar información oculta y no visible de forma automática o semiautomática, que a simple vista no se puede ver en los datos, donde esté conocimiento, está almacenado por años en grandes repositorios de bases de datos (Berry & Gordon, 1997). o disperso en diferentes fuentes de información;

La minería de datos busca que el nuevo conocimiento obtenido emerja para producir hipótesis (De Luca Venegas, 2006). Se diferencia del método científico donde esté, primero plantea una hipótesis y luego busca como refutar la misma.

Los mineros de datos pueden usar varias herramientas y técnicas que mejor se adapten a sus necesidades. Los algoritmos que se aplican en la minería de datos deber ser probados y calibrados cada cierto tiempo o cuando ocurran cambios importantes, para obtener mejores resultados.

La minería de datos produce los siguientes tipos de información: asociaciones, secuencias, clasificaciones, agrupamientos y pronósticos, que depende de los objetivos del estudio a realizarse.

3. Beneficios de la minería de datos

La minería de datos, puede aportar grandes ventajas, entre los investigadores y encargados de la toma de decisiones. Esto ha permitido ahorrar dinero y ha abierto nuevas oportunidades a la industria y a los profesionales. Todo esto es realizado a través de medios automatizados y análisis de datos, que permiten encontrar patrones u ocurrencias (Corporation, 1998).

Esta minería está compuesta de varias fases, entre ellas la de limpieza, donde el objetivo es tener una idea del estado de los datos y a su vez poder sugerir a los encargados de las aplicaciones, los posibles inconvenientes, causas y mejoras en la calidad, lo que evitará problemas presentes y futuros, tanto en los usuarios finales de las aplicaciones, como proyectos de inteligencia de negocio.

Las reglas de clustering permiten conocer mejor manera, el cómo están los datos agrupados, ya que toma en cuenta sus similitudes y es generado automáticamente. Lo que permite obtener las características de cada agrupo y sus relaciones entre los diferentes atributos, lo que ayuda a tener una mejor compresión del negocio.

Con las reglas de asociación se identificó a aquellos profesionales que poseen características de éxito o fracaso, y que no estén acorde a su rama de estudio. La información obtenida es de ayuda para centrar las políticas públicas y concentrar esfuerzos en áreas de estudio, que no se los había tomado en cuenta en relación con la actividad económica.

4. El proceso de Descubrimiento de Conocimiento en bases de Datos (KDD)

El Descubrimiento de Conocimiento en bases de Datos (Knowledge Discovery in Databases, KDD), es una ayuda en el proceso de minería de datos, para encontrar patrones, con ello información útil y oculta en grandes volúmenes de datos. Este proceso está compuesto de 5 fases: selección, pre procesamiento/limpieza, transformación, minería de datos e interpretación (Maria del Socorro, 2014). Nótese que la minería de datos solo en una parte del proceso y que, para su correcta aplicación, es necesario que las fases anteriores estén bien realizadas. En la figura 2 se muestra las etapas del proceso KDD.

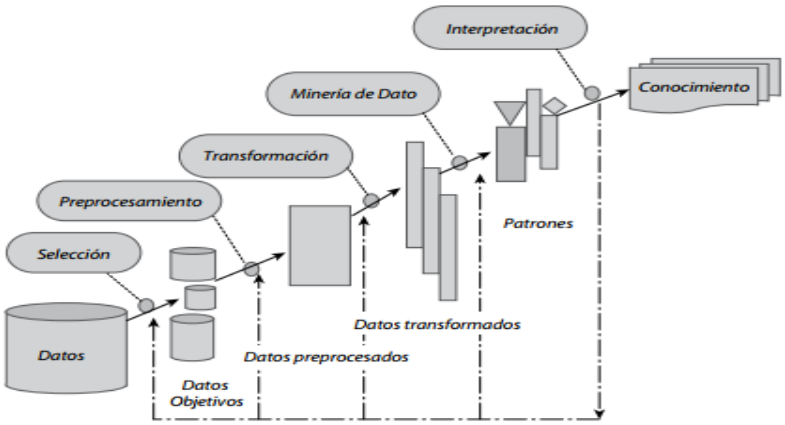

Figura 2. Etapas del proceso KDD
Fuente: **(Timarán-Pereira, 2016)**.

5. Técnicas de Análisis

5.1 Análisis

Dentro del análisis de los datos es importante conocer el tipo a los que estos corresponden, para poder aplicar correctamente los algoritmos de minería de datos. Entre los que se tiene:

a) Datos cualitativos: son los que representan variables categóricas en lugar de números (Schettini & Cortazzo, 2010). Las operaciones matemáticas no tiene sentido y se dividen en variables ordinales y nominales.
b) Datos cuantitativos: Son variables de tipo numéricas (López, Valcárce, & Barbancho)., que toman un número como valor. Se dividen en variables continuas o discretas.

Para analizar los datos y obtener nuevo conocimiento útil, ya sea este intuitivo u oculto, basado en patrones, se lo puede realizar de tres formas que se describen a continuación:

i. Mediante el uso de la estadística, que a través del tiempo ha ido acumulando métodos y análisis matemáticos, con excelentes resultados. Esta se basa principalmente en muestrear los datos (Flores, 2009) .

ii. Mediante el uso de herramientas de reportería de sistemas informáticos, que últimamente mediante la aparición de varias herramientas de Inteligencia de Negocio y OLAP (OnLine Analytical Processing) dan una perspectiva clara de los datos, a un nivel gerencial y operativo (Martínez, 2012) . Estas herramientas en lo posible tratan de acumular la mayor parte de la información en un repositorio de datos y presentan información resumida.

iii. La minería de datos, conformada por un conjunto de técnicas y algoritmos (Martínez, 2012). Buscan encontrar información no trivial, desconocida y útil, que trabaje con todos los datos; y que sirva de soporte para la toma de decisiones.

Después de revisar las formas de analizar la información, la que más se ajusta a las necesidades del presente estudio es la minería de datos, dentro de ella existen algoritmos que ayudan a obtener los resultados, definir la forma, es la tarea dentro de la minería de datos, para poder establecer modelos que sirvan de referencia para trabajos futuros.

5.1 Modelos de minería de datos

Modelos predictivos

Se refiere a descubrir patrones y para aplicarlos es necesario tener un conocimiento teórico previo (IBM, 2018). Este que servirá para constatar los resultados obtenidos antes de aceptarlos como válidos. En la figura 3 se observa la clasificación de los modelos predictivos.

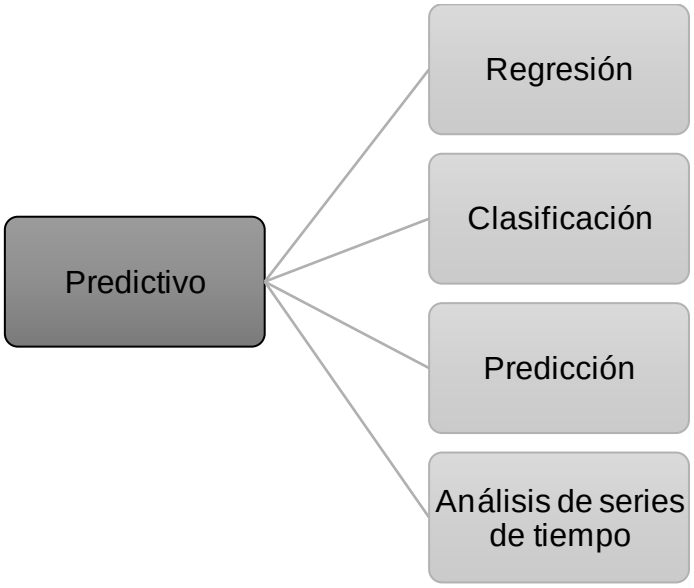

Figura 3. Modelo predictivo.

Modelos modelo descriptivos

Aquí los modelos son creados automáticamente para el reconocimiento de patrones y es por ello que las variables no tienen un papel determinado, ni se supone la existencia de variables dependientes o independientes, o de la existencia de algún modelo previo. Sirven para explorar las propiedades de los datos (Hasperué, 2012). En la figura 4, se observa la clasificación de los modelos descriptivos.

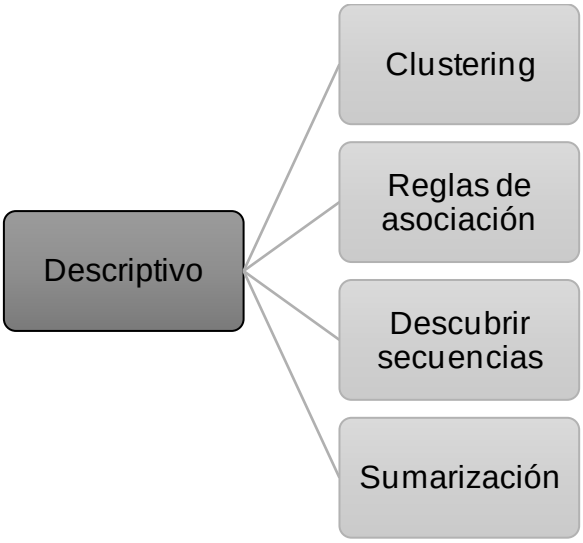

Figura 4. Modelo descriptivo.

5.2 Agrupamiento o clustering.

Técnica que agrupa diferentes datos entre sí, en donde se busca aquellos que tengan patrones o coincidencias comunes, y que a su vez tengan diferencias con los otros grupos. Es decir, internamente sean homogéneos y externamente sean heterogéneos (Timarán-Pereira, 2016). A este análisis se le conoce con el nombre de segmentación de data.

K-means

Es uno de los métodos de clustering interactivos, que segmenta la base de datos sin intervención por parte del usuario (Timarán-Pereira, 2016). Para aplicarlo es necesario que todas las variables sean de tipo cuantitativo, debido a que calcula las distancias euclidianas como medida de similitud.

Para el proceso se escogen al azar los 2 grupos y se les asignan grupos aleatoriamente. Luego se calcula el centro de gravedad y se calcula la distancia de todos contra el centro de gravedad. Posteriormente se procede a reasignar el clúster (Martínez, 2012). La ventaja de k -means es que calcula la distancia de todos contra dos, esto resuelve el problema combinatorio con matrices grandes.

Este algoritmo no es confiable ejecutarlo en bajas cantidades como por ejemplo dos, ya que podría arrojar datos aleatorios, es necesario hacer como mínimo unas 30 veces para tener mejores resultados (De La Cruz, Rivasplata, & Flores, 2012). Quizás podría parecer demasiado, pero es mejor computacionalmente, para obtener la mejor opción.

El número de clústers depende de lo que se desee hacer, así por ejemplo si se quisiera hacer una campaña para un determinado producto eso dependería de la cantidad de recursos disponibles. Si se tendría muchos clústers se complica la interpretación, mientras que en pocos no da muchos detalles, por lo que es necesario calibrar.

5.3 Reglas de asociación

Utilizado para encontrar ocurrencias en común que suceden en un determinado conjunto de datos, es decir cuando sucede un evento *X*, es posible o probable que suceda el evento *Y*.

Las reglas buscan encontrar itemsets frecuentes, cuyo soporte sea mayor a un umbral Supp(A)>=SuppMin y generar las reglas únicamente sobre los itemsets del item anterior, y la generación va a ser a partir del nivel de confianza que se generen sobre esos itermsets frecuentes (Conti & Martínez); este proceso es un problema costoso computacionalmente ya que dados *d* ítems, se tiene 2^d itemsets y *R* posibles reglas, esto se ilustra en la figura 5:

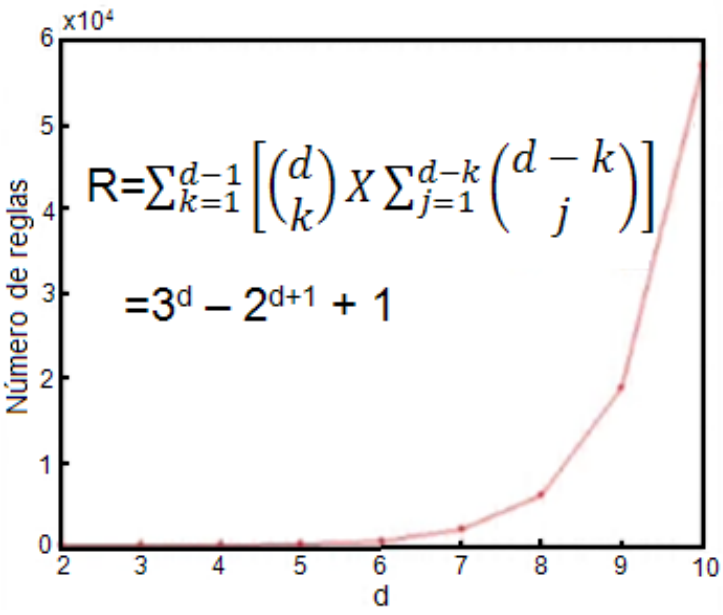

Figura 5. Complejidad computacional
Fuente: **(Berzal)**

Para un problema diminuto con 10 items: R=236 mil.

¿Cómo resolver el problema?

- Reducir la cantidad de itemsets fuertes con el uso de técnicas de poda; Ejemplo: algoritmo a priori y DirectHashing and prining;
- Reducir el número de transacciones;
- Reducir el número de comparaciones, esto va más orientado a la forma de cómo está organizada o estructurada la información.

Algoritmo a priori

Fue introducido por Rakesh Agrawal (1993), descubre las relaciones entre las variables en grandes volúmenes de datos. Se enfoca en la reducción de la cantidad itemsets (Conjunto de uno o más artículos/objetos) fuertes o frecuentes, que estén disponibles y propone candidatos para la generación de reglas (Mejia, 2002).

Si un Itemset (elemento de una transacción) es frecuente, también lo son todos los subconjuntos, y esto es porque el soporte de un itemset nunca puede ser mayor que el de cualquiera de sus subconjuntos. Esta propiedad se la conoce con el nombre de anti-monotonía (Lara, 2010). del soporte.

Si el soporte mide frecuencia, entonces la confianza mide la fortaleza de la regla. (Timarán-Pereira, 2016). Una confianza alta indica que el porcentaje de transacciones que contiene a *X* también contiene a *Y* de manera conjunta. Si se tiene un soporte bajo significa que se acepta cualquier regla, independientemente de si se presentó una vez, y si es alto significa que puede no aparecer nada (Miranda, 2015).

CAPÍTULO 3

METODOLOGÍA

En este capítulo se explica el criterio de selección de la metodología CRISP-DM y la herramienta R – Studio con su lenguaje R. Necesarios para aplicar las técnicas, procedimiento y algoritmos de minería de datos.

1. Determinación de la metodología de desarrollo a aplicar

Escoger la correcta metodología, es el primer paso para empezar los proyectos de minería y es así, que de su correcta aplicación depende en gran parte el éxito de la misma; para el presente estudio se tomó en cuenta los siguientes criterios:

- Metodología gratuita de libre acceso y que permite la estandarización de los procesos;
- Aplicación neutral en la industria;
- Independiente de un software en especial al momento de su implementación;
- Que se centre en cuestiones del negocio;
- Que sus fases de implementación sean cíclicas y flexibles;
- Que sea la más usada para proyectos de minería de datos.

Adicionalmente se buscó que la metodología a utilizarse, permita una minería de datos confiables y repetible, para las personas interesadas en el presente estudio.

2. Comparación cualitativa de metodologías minería de datos

La elección de la mejor metodología ayudó a agilitar el proceso de la minería de datos. Es por ello que a continuación se presenta tres metodologías que se utilizan actualmente, de las cuales se seleccionó la más indicada y que se adaptó mejor a las necesidades del presente estudio y que se detallan a continuación:

i. SEMMA (Simple, Explore, Modify, Model, Assess). propuesta oficialmente 2008 por SAS (Statistical Analysis System), su nombre de debe a las cinco fases de su proceso, aunque esta metodología es independiente de la herramienta a utilizarse, está integrado con su software estadístico “SAS Enterprice”, y establece una guía al usuario en las implementaciones de soluciones de minería de datos a través de sus nodos. Esta metodología se centra únicamente en los aspectos técnicos, donde se excluye el análisis y la comprensión del problema (Flores, 2009).

ii. KDD (Knowledge Discovery Database), fue presentada en 1996 por Fayyad (Fayyad, 1996). Aunque ya existía desde los años 80, fue la que daría paso al resto de modelos. Este método es interactivo, con pasos que son cíclicos e independientes de la herramienta a manejar, por lo que pueden utilizarse en cualquier situación de minería de datos. En este proceso se combinan el descubrimiento y el análisis, para extraer patrones en forma de reglas y entrégalos al usuario (Timarán, Hernández, Caicedo, Hidalgo, & Alvarado, 2016).

iii. CRISP-DM (Cross Industry Standard Process for Data Mining), creado en el año 2000, por SPSS (Statistical Package for the Social Sciences), NCR (Systems

Engineering Copenhagen) y Daimer Chryler. Está formada por seis fases cíclicas (Flores, 2009). que profundizan el nivel de trabajo a realizarse. Actualmente es un estándar en la industria y es una metodología de libre acceso. Esta metodología es independiente de la herramienta a utilizarse, está implementada en la herramienta de software estadístico "SPPS Clementine". Esta metodología comprende no solo aspectos puramente técnicos, sino también aspectos como el análisis y la comprensión del negocio. Es financiada un consorcio de empresas europea (Maria del Socorro, 2014). y respaldada por fabricantes de herramientas de data warehouse.

Tabla 1

Comparación de las fases de las metodologías

KDD	CRISP-DM	SEMMA
(9 pasos)	(6 paso)	(5 pasos)
1. Abstracción del escenario 2. Selección de datos. 3. Limpieza y pre-procesamiento. 4. Transformación de los datos 5. Elección de tareas de Minería de Datos 6. Elección del algoritmo 7. Aplicación del algoritmo 8. Evaluación e interpretación 9. Entendimiento del conocimiento.	1. Comprensión del Negocio 2. Comprensión de los datos 3. Preparación de los datos 4. Modelado 5. Evaluación, 6. Despliegue,	1. Muestreo (Sample) 2. Exploración (Explore) 3. Modificación (Modify) 4. Modelado (Model) 5. Valoración (Assess)

En el Ecuador la metodología CRISP-DM ha sido aplicada en procesos de minería de datos, para descubrir tendencias o patrones de ataques a las redes de datos del sector publico (Macas, Fuertes, Guerrero, & Toulkeridis, 2017)

La empresa KDnuggets, realizó una encuesta a nivel mundial para saber cuál era la metodología de minería de datos más utilizada cuyos resultados se muestran la figura 6. Mientras que en la tabla 2 se muestra un cuadro comparativo de las metodologías descritas anteriormente.

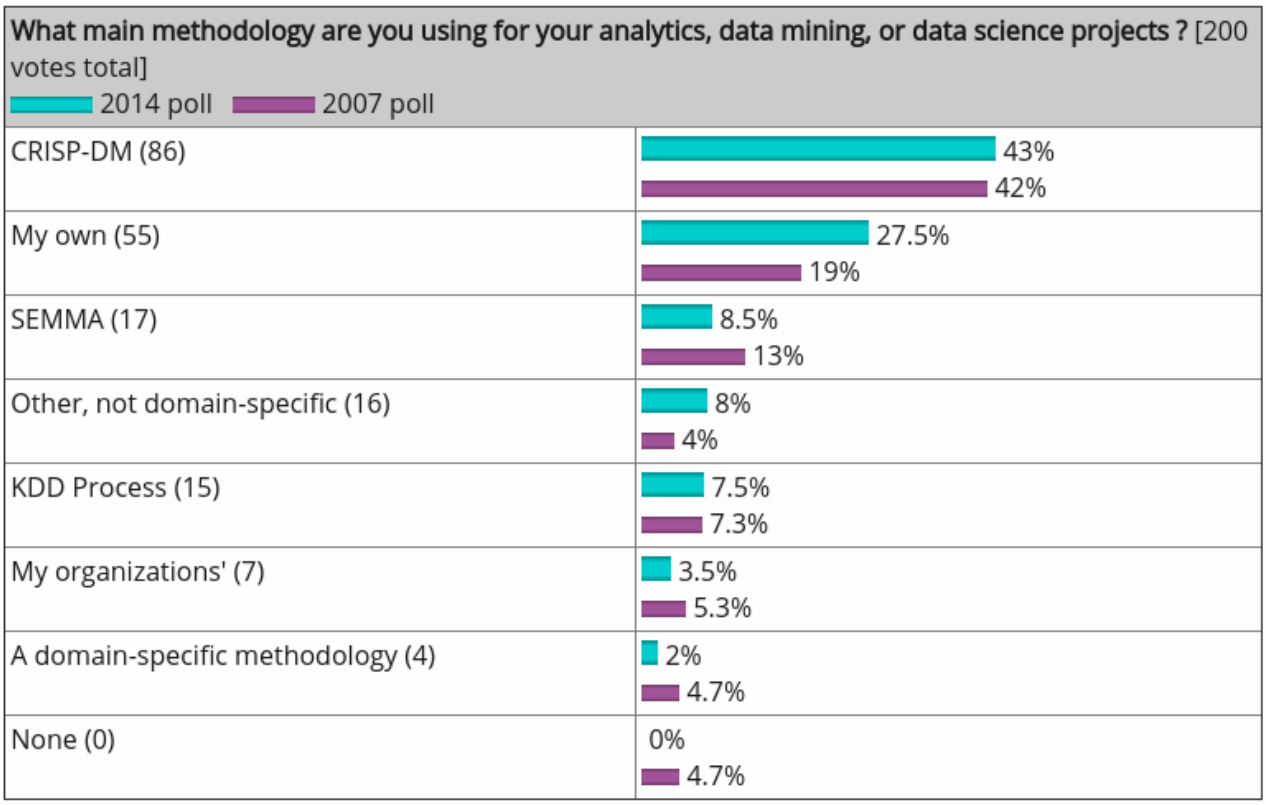

Figura 6. Top de las metodologías usadas en proyectos de minería de datos en el 2014
Fuente: (KDnuggets, KDnuggets)

Tabla 2
Cuadro comparativo de metodologías

CARACTERÍSTICAS	KDD	CRISP-DM	SEMMA
De libre acceso	Cumple	Cumple	Parcialmente
Independiente de la aplicación de la industria	Cumple	Cumple	Cumple
Neutral respecto a herramientas	Cumple	Cumple	Parcialmente
Enfocado en el negocio	Cumple	Cumple	Parcialmente
Profundiza las tareas y actividades a realizar	Parcialmente	Cumple	Parcialmente
Pose respaldo de empresas	Cumple	Cumple	Cumple
Posee financiamiento para actualización y mejoras	Cumple	Cumple	Cumple
Permite que los proyectos sean replicados	Cumple	Cumple	Cumple
Madurez en la minería de datos	Cumple	Cumple	Cumple
Fases claras, y simplificadas	Cumple	Cumple	Cumple
Costo	Cumple	Cumple	Parcialmente
Top según encuesta KDnuggets	Parcialmente	Cumple	Parcialmente

Después de revisar las metodologías y compararlas en la tabla 2, de donde se tomó en cuenta que sea una metodología abierta, gratuita, estándar en la industria, centrada en el negocio (Shafique, 2014)., más utilizada e independiente del software (De Luca Venegas, 2006). En este estudio se seleccionó para su uso la metodología CRISP-DM.

3. Fases de la metodología CRIPS-DM

Es una metodología aplicable a cualquier industria, y con cualquier tipo de datos, lo que la hace flexible para casos de estudio. Actualmente está en la versión dos, que es

la que su utilizó. Consiste de un conjunto de 6 fases que se describen en la figura 7 a continuación:

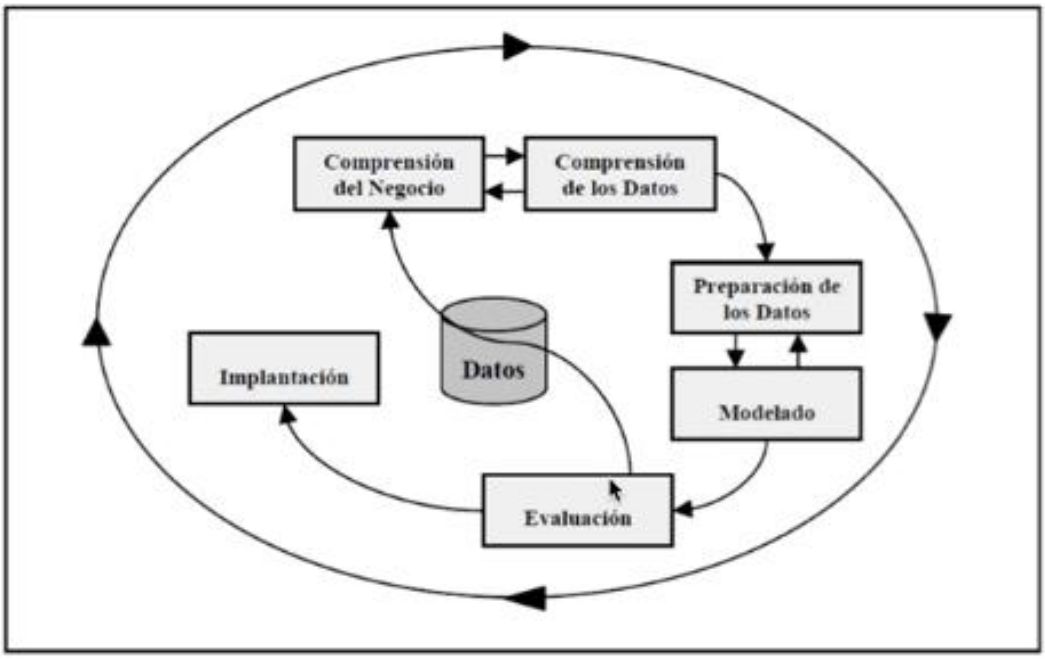

Figura 7. Ciclo de vida de CRISP–DM
Fuente: (Charman, 2000)

i. **Compresión del negocio:** Es una fase no técnica, que se enfoca en la comprensión de los objetivos del proyecto desde un enfoque institucional y que es necesaria para la correcta aplicación de minería de los datos (Timarán-Pereira, 2016);

ii. **Comprensión de los datos:** En esta fase se realiza la recolección de los datos, se los analiza y se los describe para familiarizarnos con ellos, para luego poder verificar la calidad de los mismos (Martínez, 2012). En este proceso, con un conocimiento previo de los datos se pueden empezar a formular hipótesis;

iii. **Preparación de los Datos:** Esta es una fase que requiere un considerable esfuerzo, y a su vez, es crítica, que podría incluso definir si hay posibilidad de que exista la minería, ya que se procede a la selección de los datos, formateo, limpieza e integración dejándolos listos para poder aplicar los algoritmos de minería de datos. Este es un proceso cíclico e interactivo (Hasperué, 2012). ;

iv. **Modelado:** En esta fase se aplican las diferentes técnicas de minería de datos, depende de los datos que se tenga y de las necesidades del negocio (Shafique, 2014). Aquí calibro los modelos para obtener el más óptimo y de ser necesario podría volver a la anterior fase con tal de ajustar los datos al modelo que más se adecúe;

v. **Evaluación:** En esta etapa se revisa que el modelo seleccionado cumpla con los objetivos planteados y se adapte mejor a las necesidades del negocio, esto previo a su puesta en producción (Martínez, 2012). Esta fase será necesaria realizarla cada cierto tiempo, cuando se generen nuevas condiciones en los datos;

vi. **Despliegue:** En esta parte se extrae la utilidad a los resultados obtenidos para la toma de decisiones, a través de un informe o mediante la conexión a una herramienta externa que presente el nuevo conocimiento generado (Shafique, 2014).

4. Herramientas de minería de datos

Escoger la correcta herramienta para el estudio es importante ya que ésta debe convertirse en un aporte, en vez de una carga. Para minería de datos existen soluciones open source como Weka, Rapidminer o R.

Weka: acrónimo de Waikato Environment for Knowledge Analysis (Entorno para el análisis del conocimiento de la Universidad de Waikato) (García, 2005). constituye una extensa colección de algoritmos de máquina de conocimiento implementados en Java, útiles para ser aplicados mediante las interfaces o para embeberlos dentro de cualquier aplicación. (Weka, 2018). WEKA posee la licencia GPL1 para su libre distribución y es de código abierto.

RapidMiner: Desarrollado por el departamento de inteligencia artificial de la Universidad de Dortmunt, en lenguaje java, lo que lo hace multiplataforma, distribuido bajo licencia AGPL, y disponible en el repositorio de GitHub desde el 2017, aunque existe una versión de pago que contiene mayores características y funcionalidad (RapidMiner, 2018). Posee una interfaz gráfica, puede ser programado a través de leguaje de scripts e incluye herramientas de visualización. Dispone de módulos de integración con R.

“R es un lenguaje de programación interpretado, de distribución libre, bajo Licencia GNU, y se mantiene en un ambiente para el cómputo estadístico y gráfico. Este software corre en distintas plataformas Linux, Windows, MacOS, e incluso en PlayStation 3.” (R Users Group).

En R - Studio es la consola gráfica que sirve de interfaz para el uso del Lenguaje R, que implementa comandos, funciones y librerías. Para la minería de datos existen librerías gráficas, como Rattle (Dr. Graham Williams, 2011) o FactorMineR GUI (Agrocampus Rennes University France, 2016). que son de gran ayuda cuando no se conoce bien la sintaxis del lenguaje.

Las librerías extras de R se las puede obtener de repositorios creados con la finalidad de añadir funcionalidad a la herramienta, uno de los repositorios famosos e importantes y de donde se obtuvieron las librerías es el CRAN (Santana & Farfán, 2014).

5. Comparación cualitativa de herramientas de minería de datos

Para la selección de la herramienta de minería de datos se tomó en cuenta: i) Sea el programa de estadística más usado; ii) Software libre con licencia con posibilidad de crear librerías y expandir su funcionamiento (Contreras, Molina, & Arteaga); iii) Existan paquetes que extiendan su funcionalidad; iv) Que cuente con soporte para Linux Windows y Mac; v) Que haya recibido el apoyo de empresas y que la integren en sus soluciones.

La empresa KDnuggets, realizó una encuesta a nivel mundial para saber cuál era la herramienta de minería de datos más utilizada y los resultados se muestran en la figura 8. Mientras que en la tabla 3 se muestra un cuadro comparativo de las herramientas descritas anteriormente.

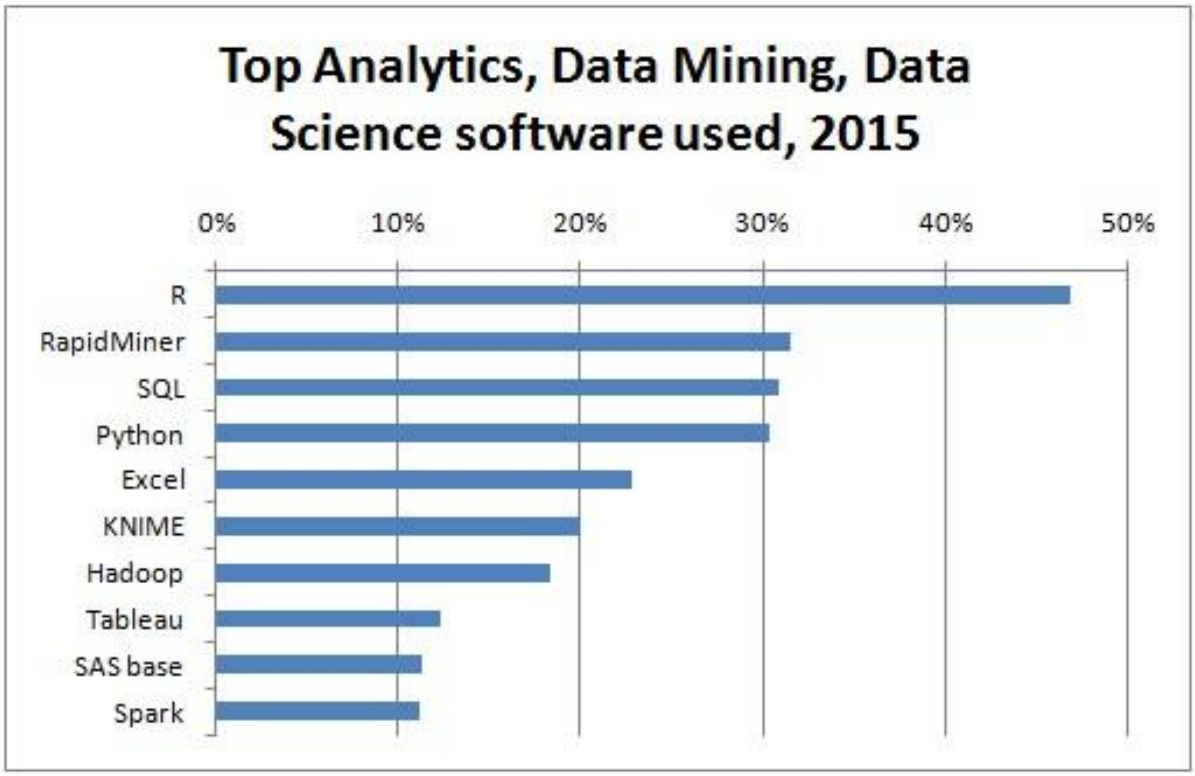

Figura 8. Top de las herramientas de minería de datos usadas en el 2015
Fuente: (KDnuggets, KDnuggets)

Tabla 3
Cuadro comparativo de herramientas de minería de datos

CARACTERÍSTICAS	WEKA	R	RAPIDMINER
De libre acceso, licencia de código abierto	Cumple	Cumple	Parcialmente
Instalación en Windows Linux y Mac	Cumple	Cumple	Cumple
Librerías que expanden su funcionalidad	Cumple	Cumple	Cumple
Línea de comandos	Parcialmente	Cumple	Parcialmente
Lenguaje de programación para estadísticos	-	Cumple	-
Apoyado por empresas	Cumple	Cumple	Cumple
Provee interfaz gráfica	Cumple	Cumple	Cumple

Después de revisar las herramientas y compararlas en la tabla 3, donde se tomó en cuenta que sea una herramienta abierta, gratuita, apoyada por empresas, con librerías que extiendan su funcionalidad, con línea de comandos (Hartl, 2012)., e independiente del sistema operativo. En este estudio se seleccionó para su uso el lenguaje R con su herramienta R- Studio.

CAPÍTULO 4

MINERÍA

En este capítulo se explicará el proceso de automatización, que toma como punto de partida obtención de los datos y cuya meta es ser una ayuda en la toma de decisiones que inicia desde la limpieza de los datos y finaliza en la obtención del conocimiento.

1. Exploración y selección de datos

El capítulo uno se detalla información del Sector del Conocimiento y Talento Humano, que es parte de la fase de comprensión del negocio, de la metodología CRISP–DM, necesario para continuar con los siguientes pasos. Para la selección de la muestra se tomó en cuenta la información disponible y que cumpla con los objetivos de la institución y del presente estudio:

La muestra se tomó de los bachilleres del Ecuador desde el año 1993 al 2015, con un total de 2´373.675 registros. Este es el universo de datos a investigar. Se analizaron las áreas de estudio o especialidades tomadas por los profesionales y su relación con las actividades económicas. De esta se tomaron los siguientes atributos:

- Especialidad: Área de estudio que tomó el bachiller;
- Fecha de grado: Fecha de la titulación que comprende, desde el periodo 1993 a 2015;
- Colegio: Institución en la que terminó sus estudios.

La base de trabajo, que integra las bases de los afiliados al seguro social y trabajadores, con fecha de corte de octubre 2015 hasta enero del 2016, con un total de 4´502.023 registros, de esta base se tomaron los siguientes atributos:

- Remuneraciones: Sueldo en dólares americanos de los empleados públicos, privados y aportes voluntarios;
- Actividad económica: Actividad según el CIUU4 en la que se desempeña el trabajador;
- Fecha de ingreso: Fecha desde que ingreso a su última institución hasta la fecha de corte.

La base de profesionales, que terminaron el tercer nivel o estudios superiores, que contiene el registro de todos los profesionales del Ecuador desde 2000 al 2016, con un total de 1´869.288 registros. De esta se tomaron los siguientes atributos:

- Título: Obtenido por el profesional otorgado por una institución registrada por la entidad rectora;
- Áreas de estudio: Clasifica y agrupa a los diferentes títulos obtenidos en diferentes sectores de estudio;
- Fecha de grado: Fecha en la que obtuvo su último título profesional;
- Nombre de la institución: Institución o universidad donde culminó sus estudios de tercer nivel.

La base de Rentas Internas, cuya función es el cobro de impuestos, con corte a octubre del 2016, con un total de 4´904.645 registros. Esta base de datos es pública y

libre para la descarga. Se la puede obtener de la siguiente dirección: http://www.sri.gob.ec/web/guest/ruc;jsessionid=hCiivx4yBlYlimrqVoN VFRWs; de esta se tomaron los siguientes atributos:

- Estado: Tipo de establecimiento que comprende, si es persona natural o institución;
- Actividades económicas. Actividad según el CIUU4 en la que se desempeña el trabajador;
- Fecha de apertura: Fecha en la que registra el inicio de las actividades económicas.

La base del registro nacional de ciudadanos con corte a octubre del 2015, con un total de 16´914.661 registros, de esta se tomaron los siguientes atributos:

- Estado civil del ciudadano;
- Edad: del ciudadano expresada en años.

2. Depuración y procesamiento

Para la fase de preparación de datos se utilizó la herramienta Pentaho Data Integration, que ha sido utilizada en proyectos similares, así por ejemplo análisis de vulnerabilidades para incrementar el nivel de seguridad en un CSIRT académico (Reyes-Mena, Fuertes-Díaz, Villacís-Silva, Guzmán-Jaramillo, Pérez-Estévez, & Bernal-Barzallo, Jan. 2017.). o la implementación de un business intelligence para el análisis

de vulnerabilidades para mitigar el impacto de los incidentes en la red de investigación CEDIA (Reyes Mena, 2017).

Adicionalmente se utilizó el gestor de base de datos PostgreSQL, para el almacenamiento de los datos: fuente, transformados y limpios. Este proceso fue el más extenso, arduo, y constituye uno de los más importantes, ya que este depende la calidad de los resultados.

Para la limpieza se migraron los archivos entregados formato csv, hacia una base de datos denominada fuente, posteriormente estos datos pasaron a una base de datos intermedia donde se realizó la limpieza y transformaciones necesarias, en algunos casos se utilizaron scripts de SQL. Una vez que todo estuvo correcto se la migro hacia la base final o de producción. La figura 9 ilustra lo antes descrito.

Figura 9. Proceso ETL

Para la verificación de la calidad de los datos al revisar las tablas se identificaron problemas, como que las letras estaban en mayúsculas o minúsculas para un mismo dato, o que contenían faltas ortográficas, para esto se los estandarizo a un solo formato. Otro caso fue el de encontrar datos que poseían referencias a otros datos con su código, lo que se resolvió al colocar el valor que corresponde.

La limpieza de los datos es un trabajo complejo, que debe iniciar por ingresar datos coherentes y correctos a los sistemas, para luego proceder a los procesos de minería. Se han hecho esfuerzos por mejora la calidad de la data ya se en nombres, (Delgado, Galárraga, Fuertes, Toulkeridis, Villacís, & Castro, 2016)., fechas o número, pero mientras existe el factor humano serán necesarios los procesos ETL.

Los datos atípicos que se encontraron fueron con respecto a las remuneraciones. Aquí se encontraron dos registros que presentaban valores atípicos, los mismos que fueron dejados fuera del análisis. Para establecer los rangos salariales y el sueldo básico se los discretizó (Hasperué, 2012). y se tomó en cuenta la escala salarial fijada en el año 2010 y publicada en el Registro Oficial N° 133.

Al final de todo este proceso de limpieza de datos (ETL), se obtuvo una vista minable lista para su uso, con un universo de 2´373.675 registros, a los que se le aplicaran los algoritmos de clustering y asociación.

3. Minería de datos

Para la creación de los clústeres se tomaron los atributos de las tablas de actividades, área de estudio de bachilleres, área de estudio profesionales, el estado civil, la remuneración y el tipo de sector en donde trabaja; los que fueron puestos en forma de columna y se les asigno un valor, esto dependió del número de coincidencias. Una vez ejecutado el algoritmo y creados los clústeres, se los volvió al formato de fila original y se añadió el clúster correspondiente.

3.1 Creación de clúster

Para el proceso de minería de datos, se procedió a organizar los datos en grupos, de manera automática. Se tomó en cuenta que en su clasificación el algoritmo no consuma tiempo en el procesamiento; todo esto con el objetivo de analizar grupos que tengan características similares entre sí, para ello se utilizó el algoritmo de k-means; propuesto por Forgy (Forgy, 1965). para encontrar clústers más separados y más homogéneos.

K-means, es usado para procesar millones de datos de manera eficiente en comparación de algoritmos como clasificación jerárquica donde todos los datos se procesan todos contra todos, lo que requiere de más recursos computacionales, (Escobar, 2007). Otra técnica que se pudo utilizar es la de reducción de la dimensión (Hernández, Delgado, Rivera, & Castellanos, 2006). pero al ser millones de datos no es de mucha utilidad.

Los datos utilizados fuero cualitativos del tipo nominal (Mena, 2014). El proceso de creación de los grupos está en el Anexo 1. Para encontrar el *k* óptimo se utilizó el método experimental llamado “codo de Jambú” (Medina, Luna, Tavarez, & Narvaez, 2016). donde se usó la inercia Inter clases y se encontró donde se estabilizaban los datos.

En la figura 10 en el punto *k* número cinco es donde se forma el codo y se estabilizan los clústers, es decir, se usaron cinco clústeres en k-means. Para generar el codo de Jambú se probó de 1 hasta 30 clústers. Para la aplicación y generación de los grupos se utilizó con un total de 100 iteraciones.

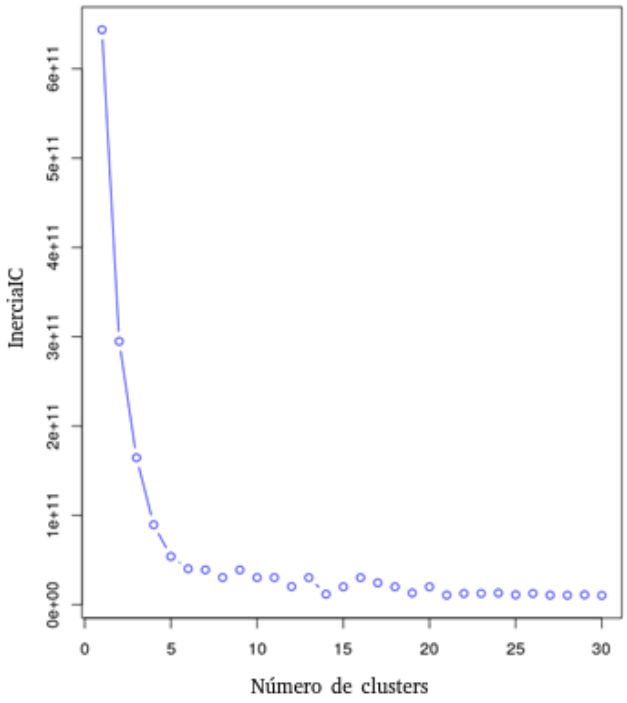

Figura 10. Codo de Jambú

Al revisar la distribución de los grupos se encontró que el grupo tres y el grupo dos, son lo más grandes al cubrir el 85,2% del total versus los grupos uno, cuatro y cinco.

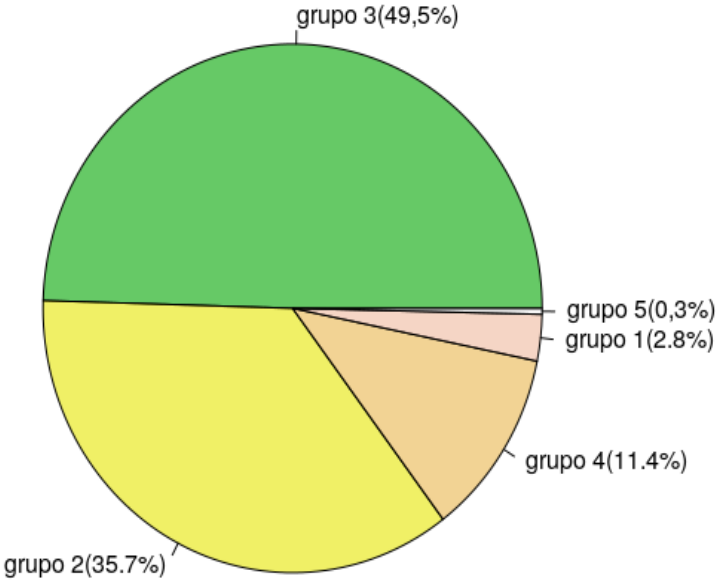

Figura 11. Distribución de grupos

3.2 Análisis de los clústers

Clúster 1 (verde): Con un total de 67.504 registros, se tiene poca presencia de bachilleres y mayor presencia de profesionales en los que se destacan las áreas de estudio de artes y humanidades, servicios, agricultura, ingeniería, ciencias y salud; en cuanto al sector al que pertenece está el público y financiero; en las actividades que se desarrollan son la enseñanza, la salud, la comunicación, electricidad, científicas y de organización.

Clúster 2 (azul): Con un total de 848.960 registros, en este grupo las remuneraciones son bajas, aquí hay una presencia media en cuanto a bachilleres a diferencia del clúster tres, hay poca presencia de la educación superior; en cuanto al trabajo está los sectores de servicio doméstico, artesanales, voluntarios, especiales y empresas públicas. En el tipo de actividad realizada predominan la reparación de vehículos, comercio al por mayor y menor, servicios administrativos, manufactureras, inmobiliarias, servicio de comidas, de transporte, construcción aportaciones voluntarias al seguro social, científico-técnicas, organizaciones y sin actividad económica.

Clúster 3 (rojo): Con un total de 1´175.712 registros, muestra que en este grupo están principalmente los bachilleres hombres, donde la remuneración es menor y se desconoce el tipo de trabajo que realizan o son independientes, en cuanto a su actividad laboral no está disponible o no registrada. Aquí se agrupa una parte de los que tienen el seguro social campesino.

Clúster 4 (amarillo): Con un total de 272.947 registros, con un nivel de remuneración superior al clúster tres y dos, hay una mayor presencia de profesionales de estudios universitarios destacándose, artes y humanidades, servicios, agricultura y educación, con poca presencia de bachilleres; los sectores en los que trabajan están el sector financiero y el sector público, en donde las actividades en las que se desarrollan son: atención a la salud, seguros, administración pública y alcantarillado.

Clúster 5 (negro): Con un total de 8.552 registros, con una mayor presencia de mujeres, casi no hay presencia de bachilleres a excepción de la especialidad de físico matemáticos; en la educación superior se encuentran los que son de áreas de ciencias sociales, agricultura, ciencias e ingeniería. Aquí se registran las mayores remuneraciones y una mayor edad. Los sectores en lo que se encuentran están el sector privado, sector público y empresas públicas, desarrollándose en áreas como financieras, manufactureras, inmobiliarias, entretenimiento, administración pública, minería y científicas.

Para la visualización y análisis se utilizó un gráfico tipo radar, generado en el modelo de minería de datos, que se ilustra en la figura 12, en la que se observan las características de los todos los clústers formados.

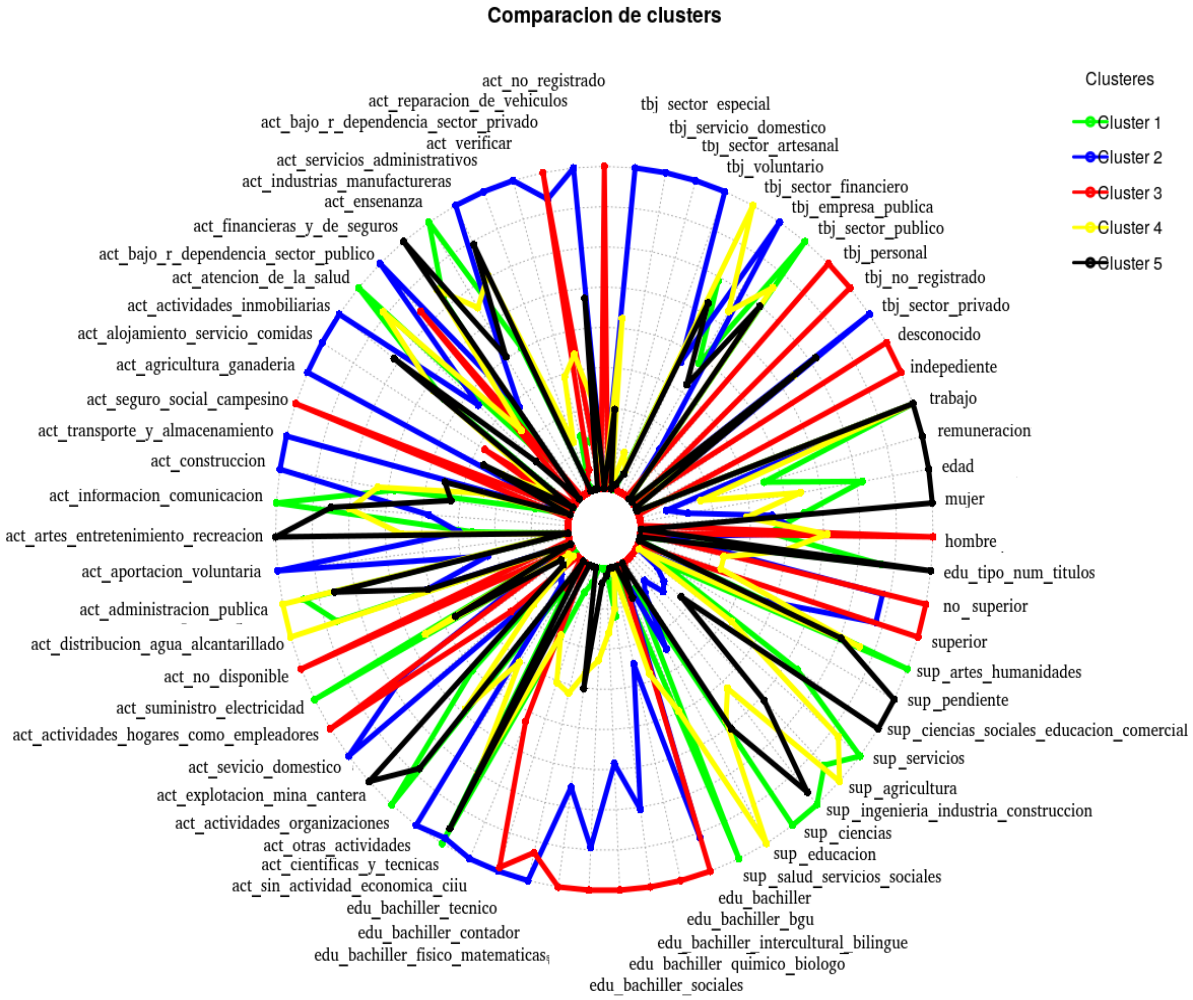

Figura 12. Comparación de clústers

Una vez que se han formado los grupos, fue necesario hacer un análisis exploratorio sobre los datos de manera inicial para conocer de mejor manera su situación, a esto se le llama minería de datos visual (Wong, 1999). Aunque la minería no se trata de buscar conocimiento que se encuentra a simple vista, esto ayudó a entender el negocio, descartar conocimiento trivial, y mejorar en los resultados.

Al comparar las proporciones que corresponden a la remuneración se encontró que el grupo tres, está compuesto por remuneración desconocida o sin datos. Esto sucede

porque no existen datos de si esa persona tiene trabajo o no; en el caso de remuneración desconocida no se tienen información de sus ingresos mientras que para el caso de sin datos, es para aquellos que tienen un negocio independiente.

En el grupo dos, ilustrado en la figura 13, existen remuneraciones menores al sueldo básico y estas aumentan de valor en el grupo uno y tres. Mientras que en el grupo cuatro, están las remuneraciones más altas.

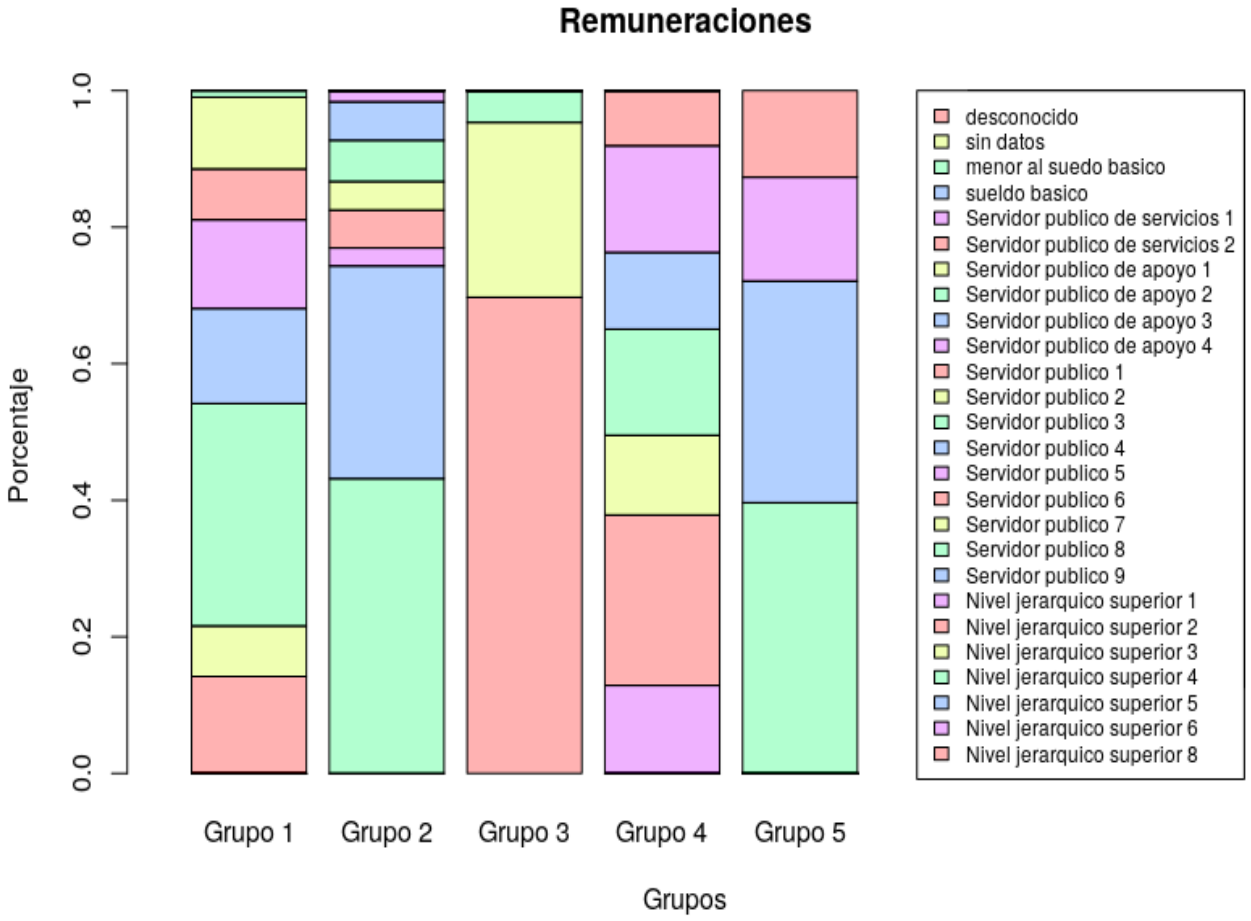

Figura 13. Remuneraciones por grupos

Al analizar los grupos que integran las actividades económicas, ilustrados en la figura 14, se evidencia que en el grupo tres están los que no tienen registro de empleo.

El sector público está presente en los grupos uno, dos, cuatro y cinco, conjuntamente con la reparación de vehículos, comercio al por mayor y menor, las actividades de salud y la industria manufacturera.

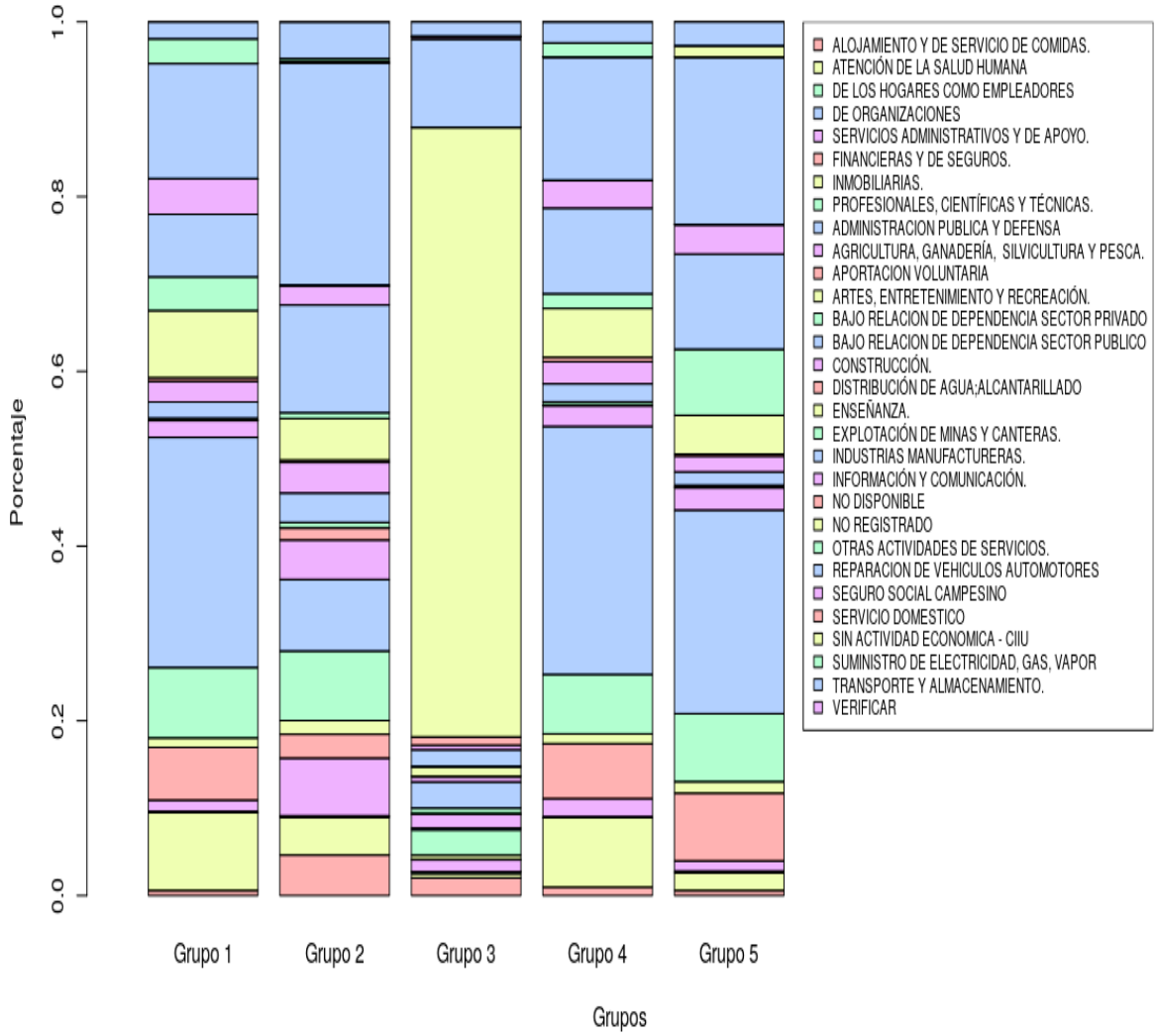

Figura 14. Actividades económicas por grupos

Al analizar las áreas de estudio ilustradas en la figura 15, en los grupos se evidencia que el grupo tres (87,38%) y el grupo dos (76,33%) está conformado por bachilleres que no tienen educación de tercer nivel; en el grupo cuatro la proporciones disminuye

(35,75%) mientras que en los grupos uno y cinco están conformados en su mayor parte por profesionales que tienen tercer nivel de estudios o un cuarto nivel.

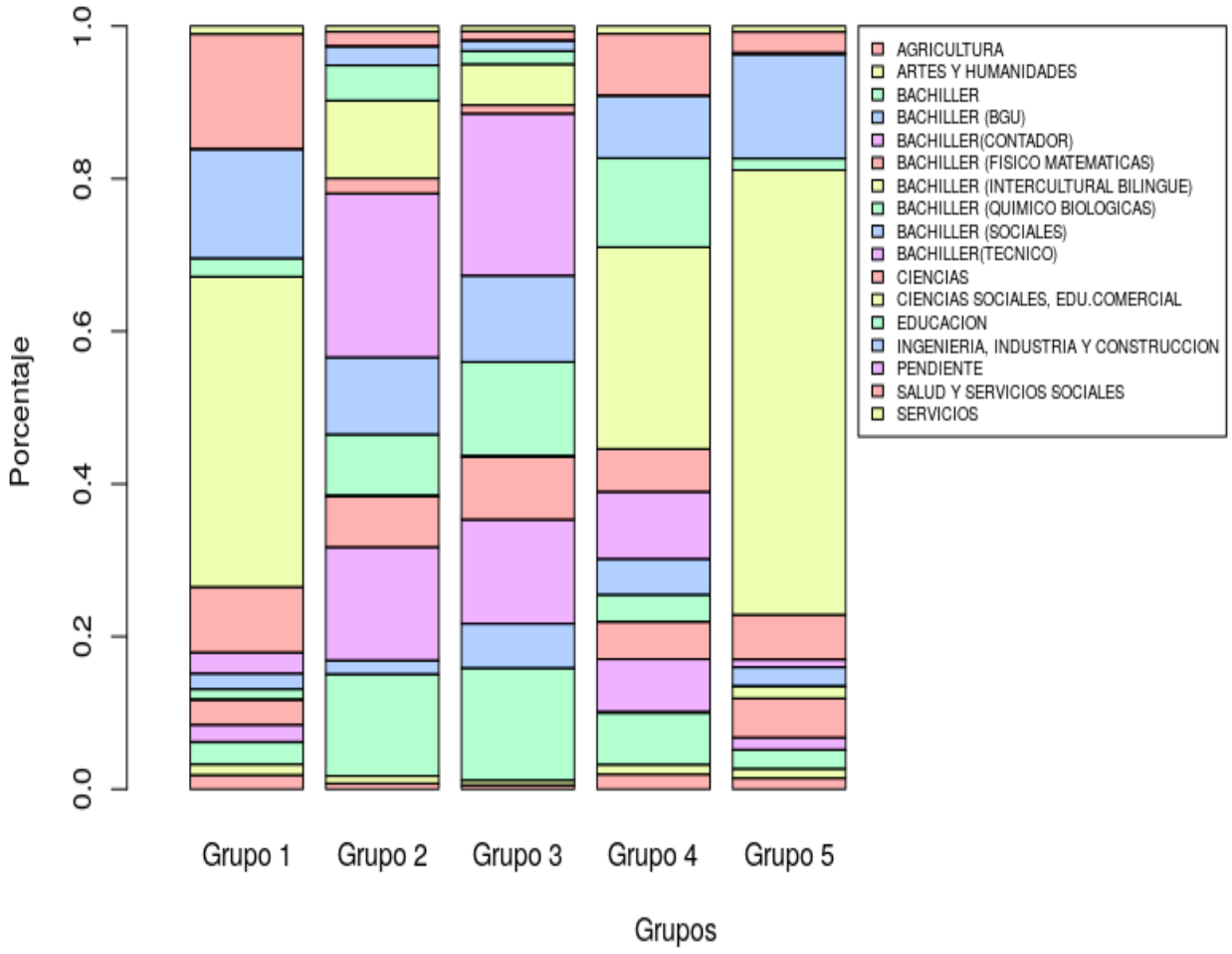

Figura 15. Áreas de estudio por grupos

Al revisar al sector de los que tiene trabajo, ilustrado en la figura 16, se tiene que grupo tres, está formado por personas de las que no se tiene información de su trabajo o son independientes; en el grupo dos en su mayor parte está formado por el sector privado (77,9%), mientras que en el resto de grupos se mantiene aproximadamente por la mitad.

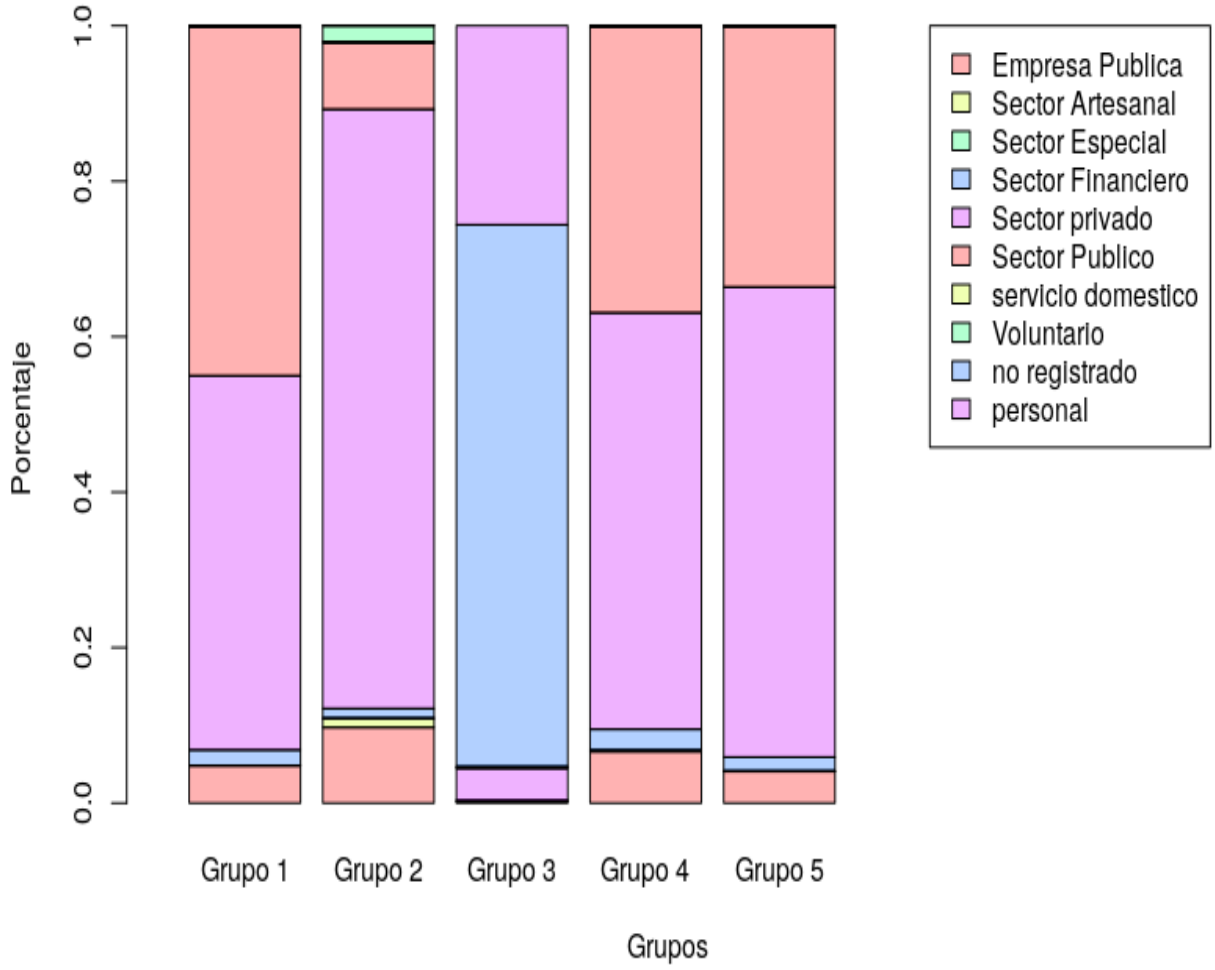

Figura 16. Sectores por grupos

En cuanto a los rangos de edad ilustrado en la figura 17, se evidencia que la población más joven domina los grupos dos y tres (entre 18 y 29 años) mientras que la población mayor a los 30 años, está en su mayor parte presente en los grupos uno, cuatro y cinco.

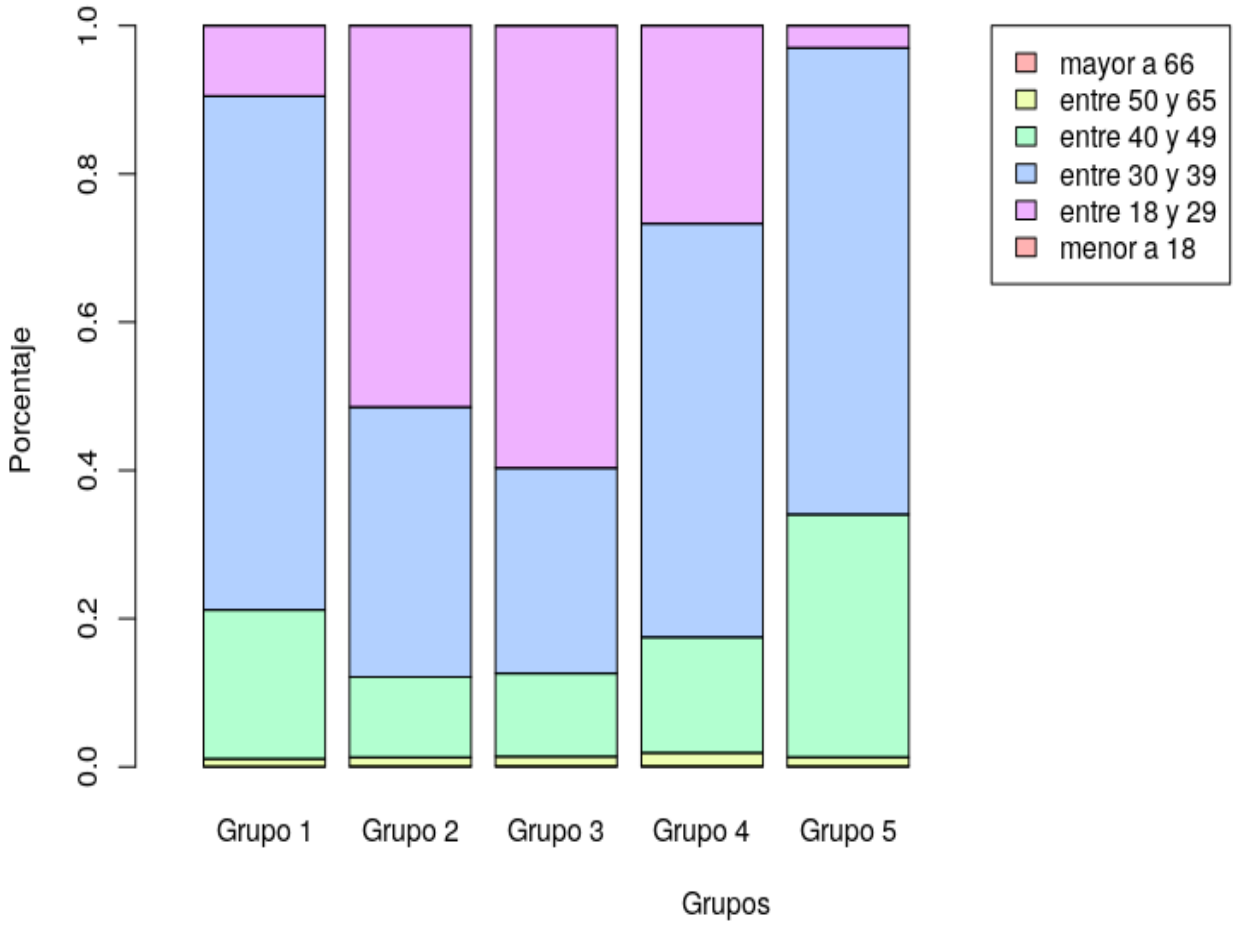

Figura 17. Rangos de edad por grupos

Al relacionar el área de estudio con la actividad económica, que se encuentra ilustrada en la figura 18, muestra que, en todos los grupos la mayor concentración de actividades está en el área de administración pública y defensa que cuyo empleador es el estado, seguido de las actividades relacionadas con reparación de vehículos, comercio al por mayor y menor; en cuanto al tercer lugar cambia en el grupo cinco con industrias manufactureras mientras que en el grupo cuatro y uno están presente las actividades relacionadas a salud.

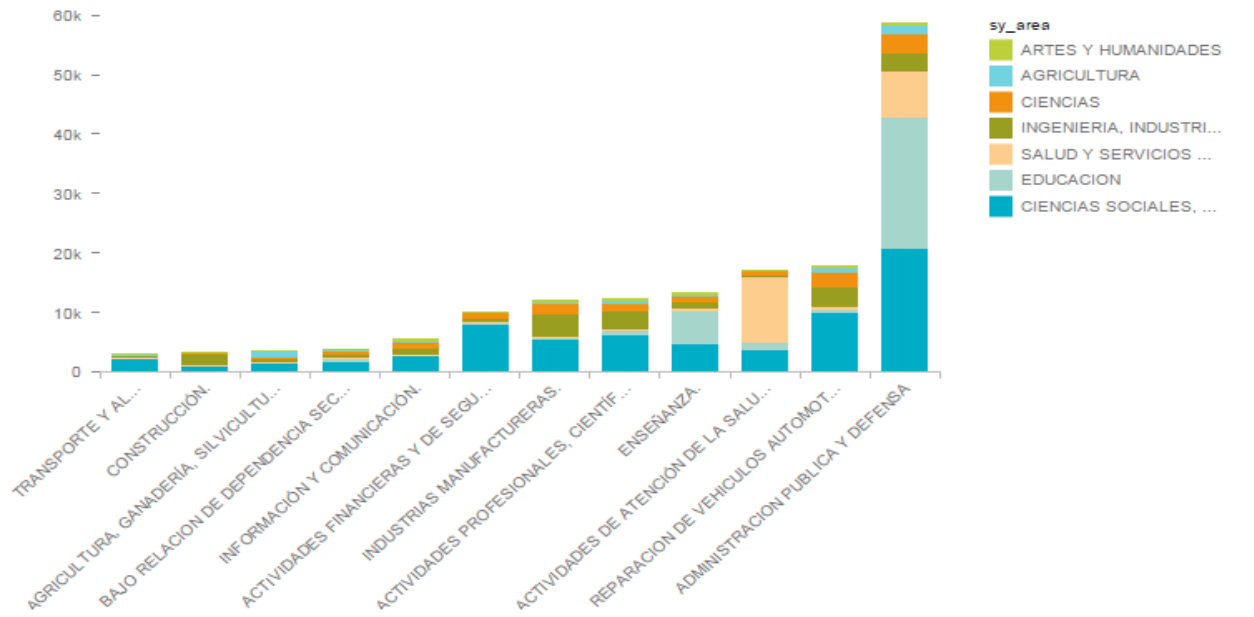

Figura 18. Áreas y actividades económicas del grupo 4

De igual manera se observa que en las figuras 17, 18 y 19, existe la presencia de profesionales en el área de ciencias sociales, en casi todas las áreas, mientras que los profesionales de las áreas de ingeniaría y ciencias, su presencia es menor.

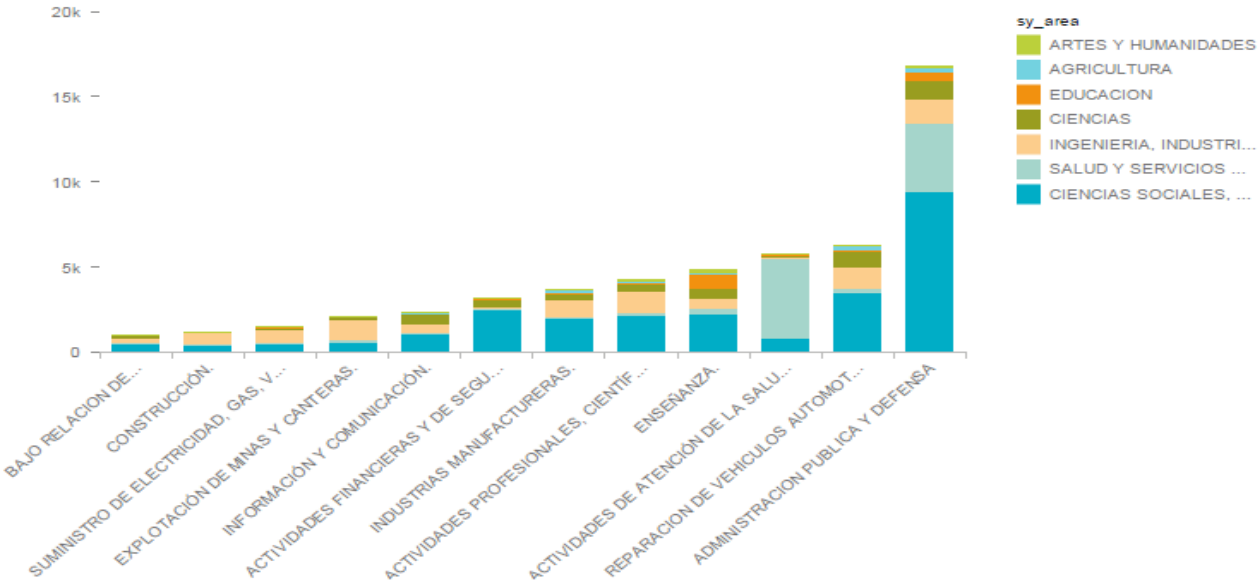

Figura 19. Áreas y actividades económicas del grupo 1

En el caso de los profesionales de la salud ilustrados en las figuras 17, 18 y 19, se observa que se concentran en los grupos uno y cuatro, y que están en la actividad económica de salud.

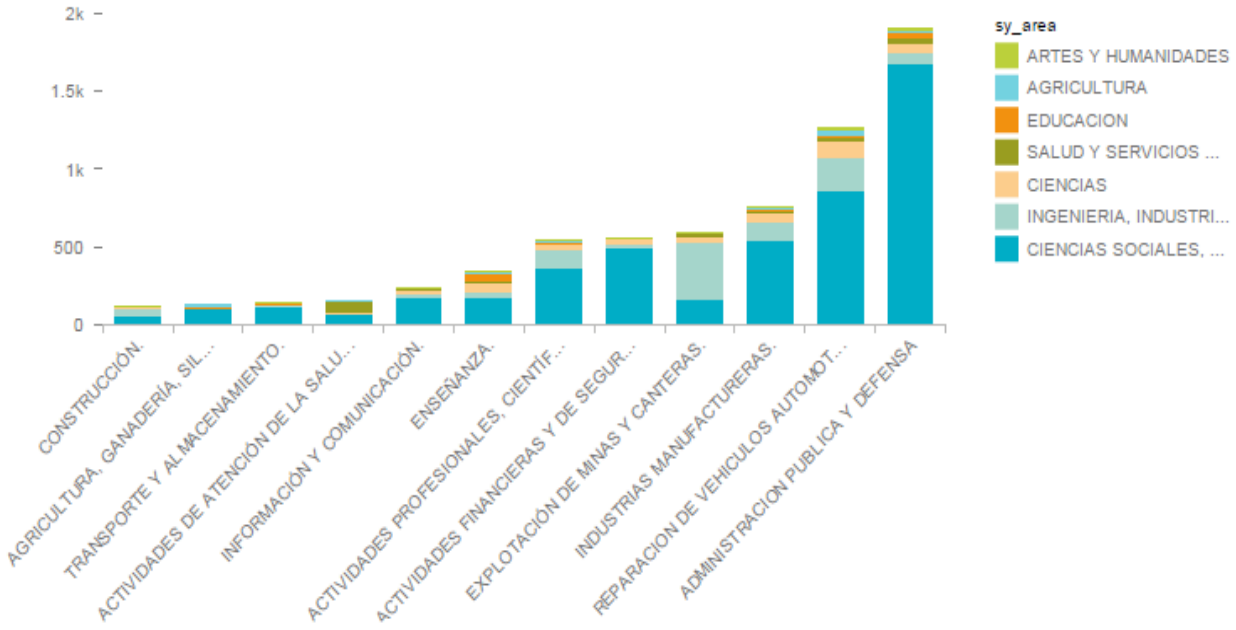

Figura 20. Áreas y actividades económicas del grupo 5

Una vez hecha la revisión previa del estado de los datos se procedió a realizar la minería mediante asociaciones para encontrar patrones o tendencias ocultas en los profesionales y su actividad económica.

4. Reconocimiento de patrones de comportamiento

El objetivo de realizar asociación, fue encontrar reglas que describan el comportamiento de los datos, conocer la ocurrencia de los mismos y ver la asociación entre sí. Para la elaboración de las reglas se utilizó el algoritmo A piori.

A partir de los datos obtenidos y una vez que estos se clasificaron automáticamente entre sí, se buscaron patrones entre las áreas de estudio y las actividades laborales, para saber qué áreas presentan nuevo conocimiento antes ignorado, a fin de establecer recomendaciones oportunas en el ámbito educativo lo que permitirá establecer programas de formación profesional orientados a satisfacer las necesidades de esa población de la cual no se tiene registros.

Para la elaboración de las reglas fueron excluidas las redundantes, ya que ellas son subconjunto de otras. Se tomó en cuenta las reglas que tengan un lift mayor a 1, lo que ayudó a que los ítems que aparezcan sean los que tengan, no solamente más frecuencia, sino que, no sean solo producto del azar. La idea es encontrar reglas fuertes (Shapiro, 1991).

Para la generación de las reglas el programa R servirá como base, conjuntamente con los paquetes arules y arulesViz. El primero para estimar las reglas de asociación con el método a priori y el segundo para visualízalas. Para todos los clústers se han utilizado las siguientes variables.

- cv_sexo : Sexo de la persona;
- cv_edad: Edad dividida por rangos;
- trb_remuneracion_esp: Remuneración dividida según escalas salariales de las instituciones públicas;
- sy_area : Área de estudio en las universidades o instituciones encargadas de la enseñanza del tercer nivel de educación;

- sy_subarea : Sub clasificación de las área de estudio de las instituciones de que imparten enseñanza de tercer nivel;
- sy_nivel_formacion: Nivel de formación, que poseen los ciudadanos ya sea bachiller o tercer nivel;
- trb_trabaja: Ocupación de las personas ya sean trabajadores, independientes o no se tengan datos;
- trb_tipo_empresa: Tipo de empresa o instituciones en la que se desenvuelve los ciudadanos;
- trb_actividad_n1: Actividad económica desempeñada por las personas clasificada según el ciu4;
- trb_cargo_hom: Cargo en su lugar de trabajo que es desempeñado por la persona.

De los grupos formados, se determinó la relación entre el área de estudio y la actividad económica. Luego se encontraron reglas que describieron el comportamiento o tendencia de un conjunto de transacciones. A partir de las reglas encontradas se pudo recomendar las decisiones a tomar en el área educativa y de trabajo.

Para aplicar el método a priori la única variable que necesitó ser transformada fue la de remuneración, que se clasificó según la escala salarial del sector público, con el sueldo básico correspondiente al año 2015.

5. Resultados y análisis

5.1 Patrones encontrados para el clúster 1

Para el clúster número 1 se obtuvieron 67.504 observaciones y 8 variables con 124 ítems, el algoritmo se aplicó con un soporte de 0,0047 y una confianza del 80%, donde se generaron 1.020 reglas, que al ser eliminadas las redundantes y escoger solo las que tienen un lift superior a 1, se obtuvieron 330. En la figura 21 se muestran las reglas generadas y las relaciones entre sí. Las variables utilizadas fueron:

- cv_sexo: Con 2 niveles, por ejemplo: "HOMBRE","MUJER",…
- cv_edad_rango: Con 5 niveles, por ejemplo: "entre 18 y 29",…
- cv_ecivil: Con 5 niveles, por ejemplo: "CASADO","DIVORCIADO",…
- trb_remuneracion_esp: Con 8 niveles, por ejemplo: "Nivel jerárquico superior 1"
- sy_nivel_formacion: Con 5 niveles, por ejemplo: "BACHILLER",…
- sy_titulo: Con 4.923 niveles, por ejemplo: "ABOGADO",…
- trb_actividad_n1: Con 226 niveles, por ejemplo:"ACTIVIDADES ADMINISTRATIVAS Y DE APOYO DE OFICINA.",…
- trb_cargo_hom: Con 625 niveles, por ejemplo: "05-LEY ORGÁNICA DE SERVICIO PÚBLICO -LOSEP",…

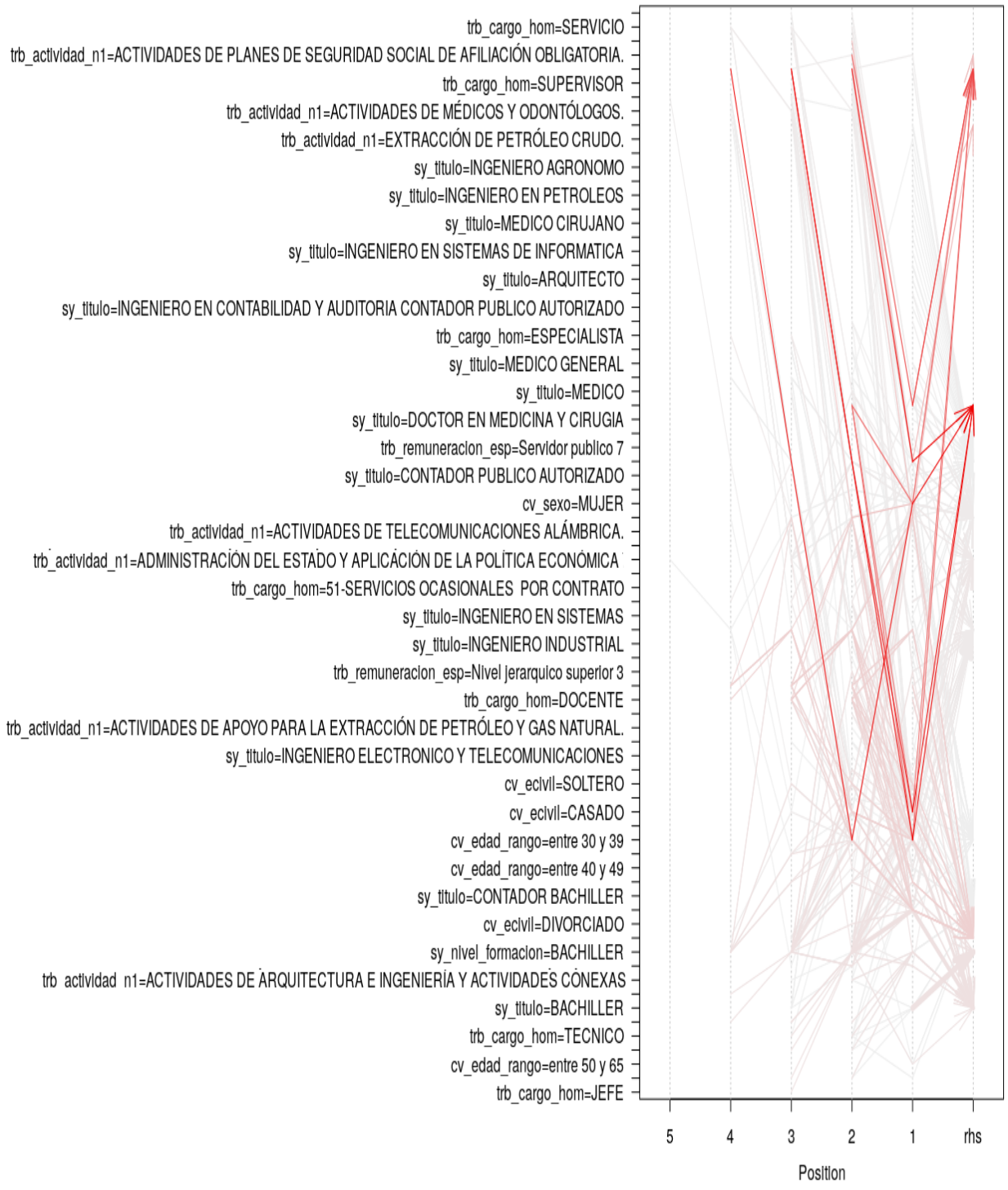

Figura 21. Reglas generadas para el clúster 1

De las reglas generadas, se seleccionó cinco, donde se tomó en cuenta que no sean redundantes, que no sean visibles a simple vista, y que tengan coherencia. Estas muestran a continuación:

i. Si es mujer con un rango de edad entre 30 y 39 años con formación de tercer nivel y trabaja con cargo de enfermera; es posible que ocurra que tenga un título de licenciada en enfermería. Los resultados se muestran en la tabla 4.

Tabla 4
Clúster 1, patrón 1

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[1] {cv_sexo=mujer, cv_ dad_rango= entre 30 y 39, sy_nivel_formación= tercer_nivel,trb_cargo_ hom=enfermero}	=>{sy_titulo=licenciado en enfermería}	0,005392273	0,9077307	78,65911 6	364

ii. Si tiene formación de tercer nivel, con cargo de secretario; es posible que ocurra que tenga el título de abogado de los tribunales de justicia de la república. Los resultados se muestran en la tabla 5.

Tabla 5
Clúster 1, patrón 2

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[10]sy_nivel_formacion=tercer_nivel,trb_cargo_hom=secretario}	=> {sy_título= Abogado de los tribunales de justicia de la república}	0,005762621	0,8206751	28,29359 2	389

iii. Si es hombre en un rango de edad de 30 y 39 años, con formación de cuarto nivel y tiene un cargo de docente; es posible que ocurra que trabaje en actividades enseñanza superior. Los resultados se muestran en la tabla 6.

Tabla 6
Clúster 1, patrón 3

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[16]{cv_sexo=hombre,cv_edad_rango=entre 30 y 9, sy_nivel_formacion=cuarto_nivel,trb_cargo_hom=docente}	=>{trb_actividad_n1= enseñanza superior.}	0,006029272	0,9644550	14,50959	407

iv. Si está casado con una remuneración equivalente al nivel jerárquico superior 3, de la escala salarial discretizada (sueldo entre 2640 y 3168 dólares americanos), que trabaja en actividades de médicas y odontológicas; es posible que ocurra que su formación sea de cuarto nivel. Como se muestra en los resultados de la tabla 7.

Tabla 7
Clúster 1, patrón 4

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[123] {cv_ecivil= casado, trb_remuneración_esp=nivel jerárquico superior 3, trb_actividad_n1=actividad de médicos y odontólogos.}	=>{sy_nivel_formación= cuarto_nivel}	0,005155250	0,9456522	3,209094	348

5.2 Patrones encontrados para el clúster 2

Para el clúster número 2 se obtuvieron 848.960 observaciones y 8 variables con 74 ítems, el algoritmo se aplicó con un soporte de 0,009 y una confianza del 80%, donde se generaron 1.244 reglas, que al ser eliminadas las redundantes y escoger solo las que tienen un lift superior a 1, se obtuvieron 354. En la figura 22 se muestran las reglas generadas y las relaciones entre sí. Las variables utilizadas fueron:

- cv_sexo: Con 2 niveles, por ejemplo: "HOMBRE","MUJER",…
- cv_edad_rango: Con 5 niveles, por ejemplo: "entre 18 y 29",…
- cv_ecivil: Con 5 niveles, por ejemplo: "CASADO","DIVORCIADO".
- trb_remuneracion_esp: Con 8 niveles, por ejemplo: "menor al sueldo básico",…
- sy_nivel_formacion: Con 5 niveles, por ejemplo: "BACHILLER",…
- sy_titulo: Con 6.228 niveles, por ejemplo: "ABOGADO",…
- trb_actividad_n1: Con 256 niveles, por ejemplo: "ACTIVIDADES ADMINISTRATIVAS Y DE APOYO DE OFICINA.",…
- trb_cargo_hom: Con 1.445 niveles, por ejemplo: "ADMINISTRADOR ALMACEN",…

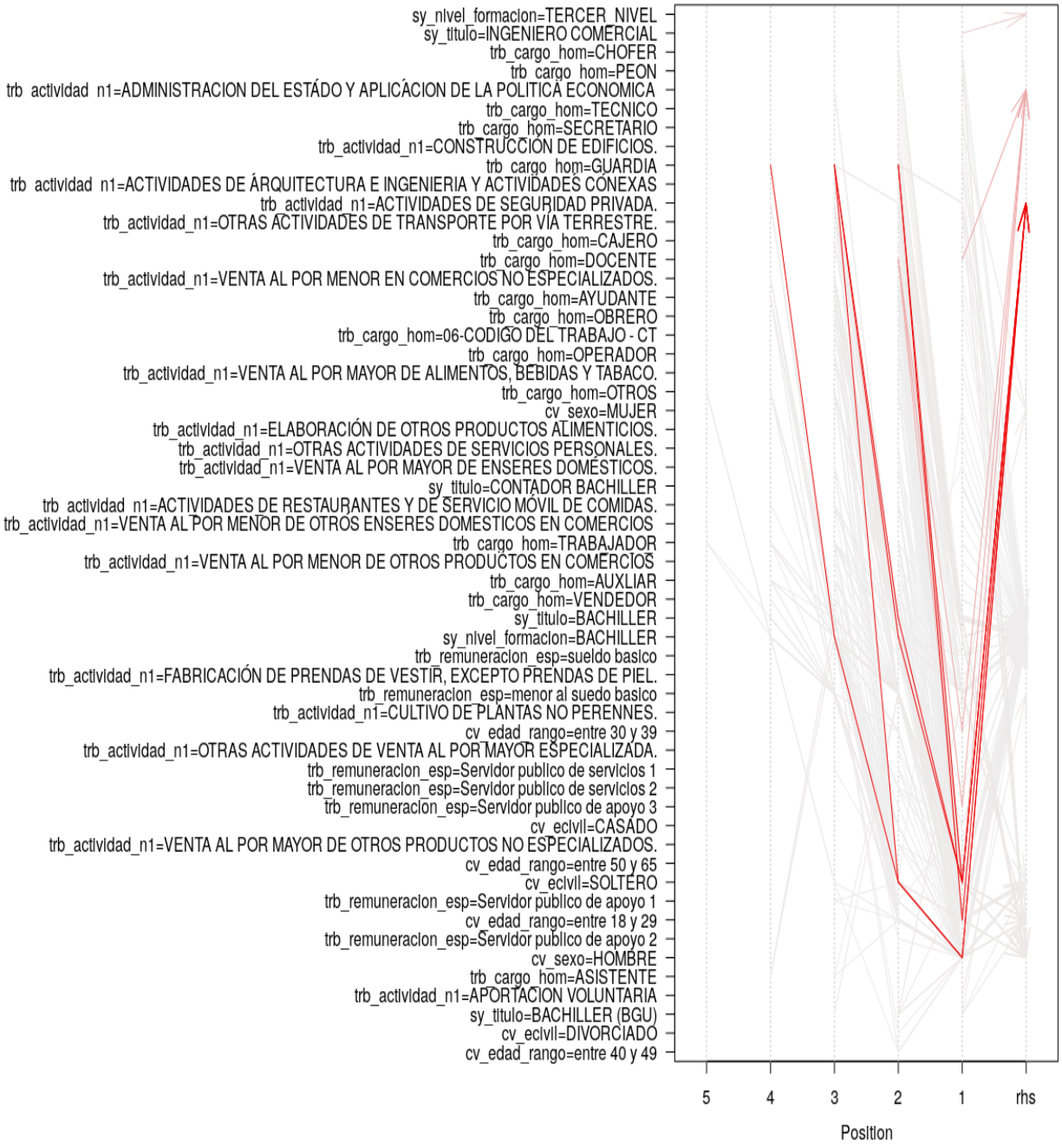

Figura 22. Reglas generadas para el clúster 2

De las reglas generadas, se seleccionó cuatro, donde se tomó en cuenta que no sean redundantes, que no sean visibles a simple vista, y que tengan coherencia. Estas muestran a continuación:

i. Si es hombre con estado civil solero, con nivel de formación bachiller sin estudios de tercer nivel y trabaja con el cargo de guardia; es posible que ocurra que labore en actividades de seguridad privada. Los resultados se muestran en la tabla 8.

Tabla 8
Clúster 2, patrón 1

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[2]sexo=hombre, ecivil=soltero,nivel_formación=bachiller, cargo_hom =guardia}	=>{trb_actividad_n1=actividades de eguridad privada.}	0,010337354	0,8063212	25,815902	8.776

ii. Si se tiene una remuneración equivalente al servidor público de apoyo 3 (sueldo entre 640 y 695 dólares americanos), que trabaja con el cargo de docente; es posible que ocurra que labore en actividades de administración del estado y aplicación de la política económica y social de la comunidad. Los resultados se muestran en la tabla 9.

Tabla 9
Clúster 2, patrón 2

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[7]Remuneración=servidor público de apoyo_3,cargo=docente	=>{actividad=administración del estado y aplicación de la política económica y social de la comunidad	0,010896862	0,9068719	12,134696	9.251

iii. Si se tiene una remuneración equivalente o que sea menor al sueldo básico (sueldo entre 1 y 366 dólares americanos) y tiene el título en bachillerato general unificado, sin estudios de tercer nivel; es posible que ocurra, que su edad este en el rango entre 18 y 29 años. Los resultados se muestran en la tabla 10.

Tabla 10
Clúster 2, patrón 3

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[17]{remuneracion=menor al suedo basico, titulo=bachiller(bgu)}	=>{cv_edad_rango=entre 18 y 29}	0,010704862	0,9695935	1,88376	9.088

iv. Si es de estado civil soltero con remuneración equivalente o que sea menor al sueldo básico (sueldo entre 1 y 366 dólares americanos) con título de bachiller, sin estudios de tercer nivel, y que tiene el cargo de asistente; es posible que ocurra que su edad este en el rango entre 18 y 20 años. Los resultados se muestran en la tabla 11.

Tabla 11
Clúster 2, patrón 4

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[60]{ecivil=soltero,remuneracion=menor al suedo basico,titulo=bachiller,cargo=asistente}	=>{cv_edad_rango=entre 18 y 29}	0,010095882	0,8314094	1,615298	8.571

5.3 Patrones encontrados para el clúster 3

Para el clúster número 3 se obtuvieron 1´75.710 observaciones y 8 variables con 35 ítems, al algoritmo se aplicó con un soporte de 0,005 y una confianza del 80%, en la cual se generaron 2916 reglas, que al ser eliminadas las redundantes y escoger solo las que tienen un lift superior a 1, se obtuvieron 340. En la figura 22 se muestran las reglas generadas y las relaciones entre sí. Las variables utilizadas fueron:

- cv_sexo: Con 2 niveles, por ejemplo: "HOMBRE","MUJER",…
- cv_edad_rango: Con 6 niveles, por ejemplo: "entre 18 y 29",…
- cv_ecivil: Con 5 niveles, por ejemplo: "CASADO","DIVORCIADO",…
- trb_remuneracion_esp: Con 3 niveles, por ejemplo: "desconocido",…
- sy_nivel_formacion: Con 5 niveles, por ejemplo: "BACHILLER",…
- sy_titulo: Con 5.983 niveles, por ejemplo: "ABOGADO",…
- trb_actividad_n1: Con 256 niveles, por ejemplo: "ACTIVIDADES ADMINISTRATIVAS Y DE APOYO DE OFICINA.",…
- trb_cargo_hom: Con 454 niveles, por ejemplo: "03-FUTBOLISTA",…

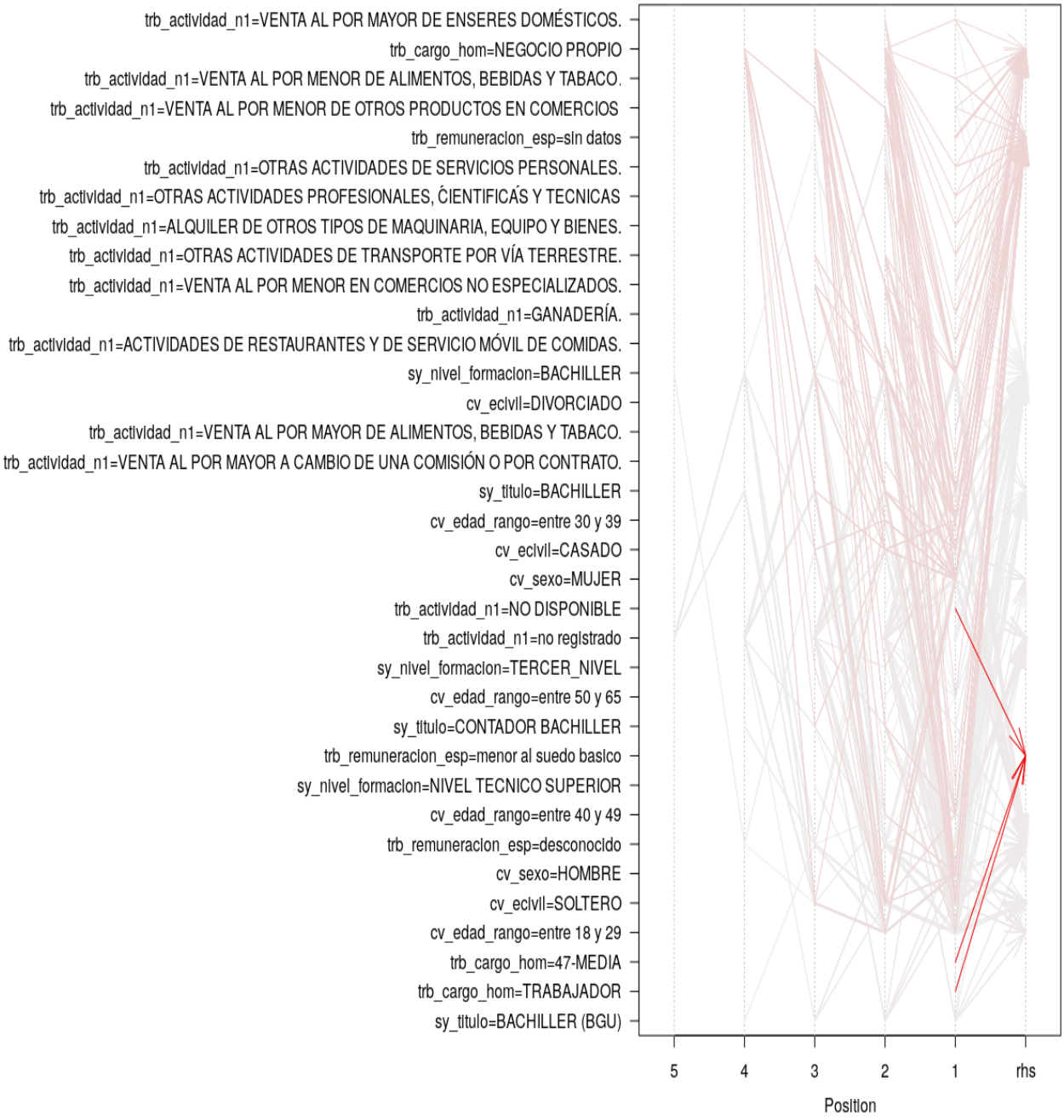

Figura 23. Reglas generadas para el clúster 3

De las reglas generadas, se seleccionó cinco, donde se tomó en cuenta que no sean redundantes, que no sean visibles a simple vista, y que tengan coherencia. Estas muestran a continuación:

i. Si es bachiller sin estudios de tercer nivel, que se desempeña en actividades de venta al por menor de alimentos, bebidas y tabacos en comercios especializados; es posible que ocurra que trabaje en su negocio propio. Los resultados se muestran en la tabla 12.

Tabla 12
Clúster 3, patrón 1

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[79]{titulo=bachiller,actividad=venta al por menor de alimentos, bebidas y tabaco en comercios especializados}	=>{cargo=negocio propio}	0,005558334	0,9703044	3,7883166	6.535

ii. Si está en el rango de edad entre 30 y 39 años, con nivel de formación bachiller, sin estudios de tercer nivel y trabaja en actividades relacionadas a la venta al por menor de otros productos en comercios especializados; es posible que ocurra que trabaje en su negocio propio. Los resultados se muestran en la tabla 13.

Tabla 13
Clúster 3, patrón 2

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[81]{cv_edad_rango=entre 30 y 39, sy_nivel_ formación =bachiller,trb_actividad_n1=venta al por menor de otros productos en comercios especializados.}	=>{trb_cargo_hom= negocio propio}	0,005871336	0,9700675	3,7873916	6.903

iii. Si es mujer con título de bachiller sin estudios de tercer nivel, que se desempeña en actividades de venta al por menor en comercio no especializados; es posible que ocurra que trabaje en su negocio propio. Los resultados se muestran en la tabla 14.

Tabla 14
Clúster 3, patrón 3

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[103]{sexo=mujer,titulo=bachiller,actividad_n1=venta al por menor en comercios no especializados}	=>{cargo_hom=negocio propio}	0,007264534	0,9606343	3,7505623	8.541

iv. Si es hombre con título de bachiller sin estudios de tercer nivel, que se desempeña en otras actividades de transporte por vía terrestre; es posible que ocurra que trabaje en negocio propio. Los resultados se ilustran en la tabla 15.

Tabla 15
Clúster 3, patrón 4

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[107]cv_sexo=hombre,título=bachiller,actividad=otras actividades de transporte por vía terrestre.}	=>{trb_cargo_hom= negocio propio}	0,007947525	0,9569848	3,7363137	9.344

Si es bachiller sin estudios de tercer nivel, que se desempeña en actividades de venta al por mayor de enseres domésticos; es posible que ocurra que trabaje en negocio propio. Los resultados se presentan en la tabla 16.

Tabla 16
Clúster 3, patrón 5

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[141]{sy_titulo=bachiller,trb_actividad_n1=venta al por mayor de enseres domésticos.}	=>{trb_cargo_hom= negocio propio}	0,00519600	0,8974585	3,5039076	6.109

5.4 Patrones encontrados para el clúster 4

Para el clúster número 4 se obtuvieron 272.947 observaciones y ocho variables con 115 ítems, el algoritmo se aplicó con un soporte de 0,0055 y una confianza del 80%, en donde se generaron 1.070 reglas, que al ser eliminadas las redundantes y escoger solo las que tienen un lift superior a 1, se obtuvieron 323. En la figura 24 se muestran las reglas generadas y las relaciones entre sí. Las variables utilizadas fueron:

- cv_sexo: Con 2 niveles, por ejemplo:"HOMBRE","MUJER",…
- cv_edad_rango: Con 5 niveles, por ejemplo:"entre 18 y 29",…
- cv_ecivil: Con 5 niveles, por ejemplo: "CASADO","DIVORCIADO".
- trb_remuneracion_esp: Con 7 niveles, por ejemplo:"Servidor público",…
- sy_nivel_formacion: Con 5 niveles, por ejemplo:"BACHILLER",…
- sy_titulo: Con 5.968 niveles, por ejemplo: "ABOGADO","ABOGADO DE LOS TRIBUNALES DE JUSTICIA DE LA REPUBLICA",…

- trb_actividad_n1: Con 245 niveles, por ejemplo: "ACTIVIDADES ADMINISTRATIVAS Y DE APOYO DE OFICINA.",…
- trb_cargo_hom: Con 1.051 niveles, por ejemplo: "ADMINISTRADOR ALMACEN",…

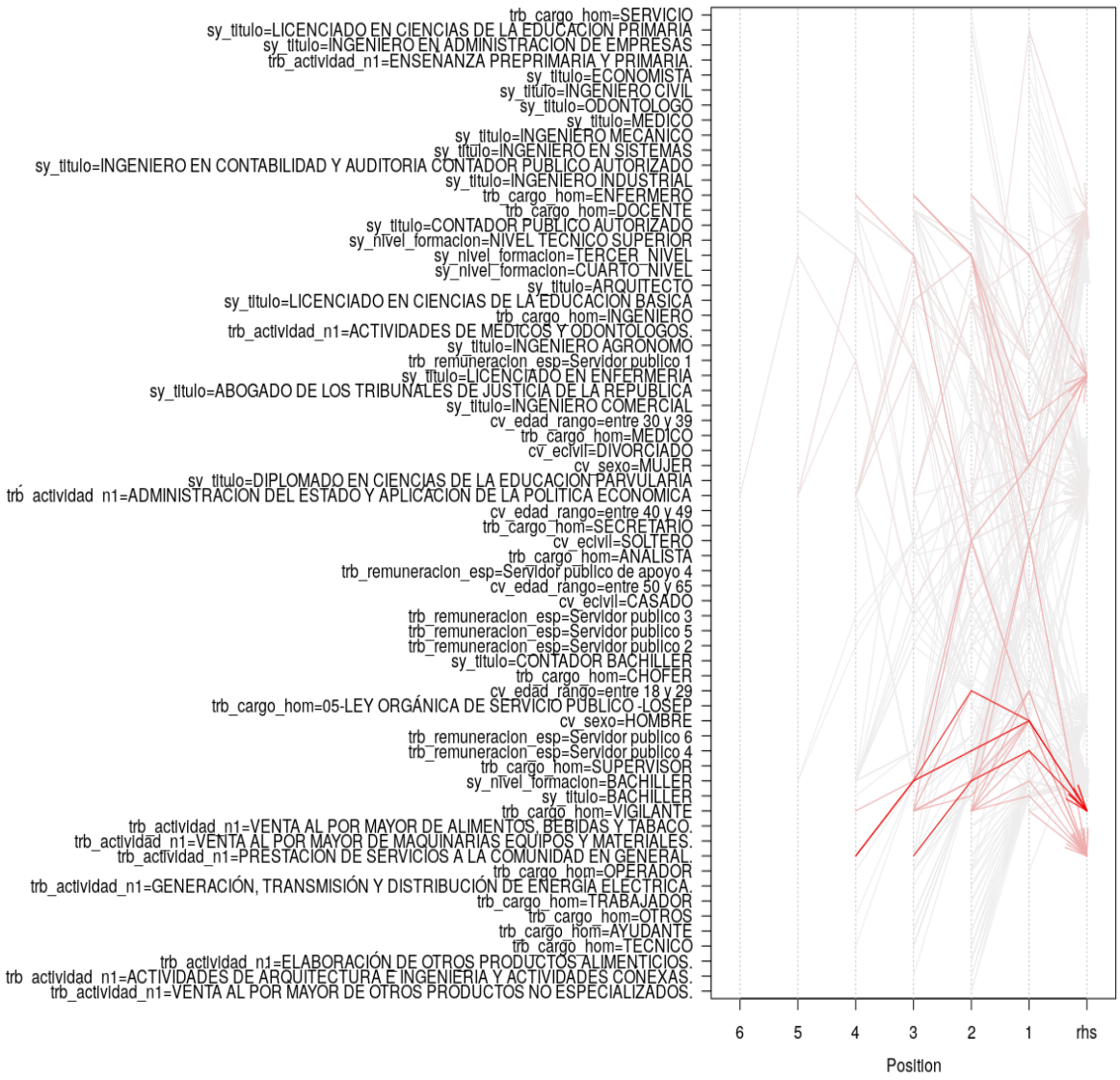

Figura 24. Reglas generadas para el clúster 4

De las reglas generadas, se seleccionó cinco, donde se tomó en cuenta que no sean redundantes, que no sean visibles a simple vista, y que tengan coherencia. Estas muestran a continuación:

i. Si es hombre con una remuneración equivalente al servidor público 4 (sueldo entre 1.030 y 1.150 dólares americanos), con nivel de formación bachiller sin estudios de tercer nivel y que trabaja en actividades de prestación de servicios a la comunidad en general; es posible que ocurra que trabaje como vigilante. Los resultados se muestran en la tabla 17.

Tabla 17
Clúster 4, patrón 1

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[1] {cv_sexo=hombre, trb_remuneracion_esp=servidor publico 4,nivel_formacion =bachiller,trb_actividad_n1=prestación de servicios a la comunidad en general.}	=>{trb_cargo _hom= vigilante}	0,006188015	0,936772 0	97,183 246	1.689

ii. Si tiene remuneración equivalente al servidor público 1 (sueldo entre 775 y 855 dólares americanos), con el título de diplomado en ciencias de la educación básica y que se desempeña en actividades de administración del estado y aplicación de la política económica y social de la comunidad; es posible que ocurra que trabaje como docente. Los resultados se muestran en la tabla 18.

Tabla 18
Clúster 4, patrón 2

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[25]{trb_remuneracion_esp=servidor público 1,sy_titulo= licenciado en ciencias de la educacionbasica, trb_actividad_n1=administración del estado y aplicación de la política económica y social de la comunidad.}	=>{trb_cargo _hom= docente}	0,008704 987	0,940990 1	10,799 328	2.376

iii. Si es mujer con título de diplomado en ciencias de la educación parvularia y que trabaja en actividades de administración del estado y aplicación de la política económica y social de la comunidad; es posible que ocurra que trabaje con una remuneración equivalente al servidor público 1 (sueldo entre 775 y 855 dólares americanos). Los resultados se muestran en la tabla 19.

Tabla 19
Clúster 4, patrón 3

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[67] {cv_sexo=mujer, sy_titulo= diplomado en ciencias de la educación parvularia, trb_actividad_n1=Administración del estado y aplicación de la política económica y social de la comunidad.}	=>{trb_remune racion_esp= servidor público 1}	0,00577 4015	0,9141531	3,66 3095	1.576

iv. Si trabaja como chofer; es posible que ocurra que tenga nivel de formación bachiller sin estudios de tercer nivel. Los resultados se muestran en la tabla 20.

Tabla 20
Clúster 4, patrón 4

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[201]{trb_cargo_hom= chofer}	=>{sy_nivel_formacion = bachiller}	0,006048793	0,924930	2,58741	1.651

v. Si es hombre, casado y que trabaja como operador; es posible que ocurra que tenga nivel de formación bachiller sin estudios de tercer nivel. Los resultados se muestran en la tabla 21.

Tabla 21
Clúster 4, patrón 5

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[204]{cv_sexo=hombre,cv_ecivil=casado, trb_cargo_hom=operador}	=>{sy_nivel_formacion= bachiller}	0,008697659	0,8119015	2,271229	2.374

5.5 Patrones encontrados para el clúster 5

Para el clúster número 5 se obtuvieron 8.552 observaciones y 8 variables con 97 ítems, el algoritmo se aplicó con un soporte de 0,006 y una confianza del 80%, donde se generaron 1.169 reglas, que al ser eliminadas las redundantes y escoger solo las

que tienen un lift superior a 1, se obtuvieron 370. En la figura 25 se muestran las reglas generadas y las relaciones entre sí. Las variables utilizadas fueron:

- cv_sexo: Con 2 niveles, por ejemplo: "HOMBRE","MUJER",…
- cv_edad_rango: Con 5 niveles, por ejemplo: "entre 18 y 29","30 y 39", "40 y 49",...
- cv_ecivil: Con 5 niveles, por ejemplo: "CASADO","DIVORCIADO", "SOLTERO"…
- trb_remuneracion_esp: Con 4 niveles, por ejemplo: "Nivel jerárquico superior 4",…
- sy_nivel_formacion: Con 4 niveles, por ejemplo: "BACHILLER","CUARTO_NIVEL",…
- sy_titulo: Con 1.627 niveles, por ejemplo: "ABOGADO",...
- trb_actividad_n1: Con 178 niveles, por ejemplo: "ACTIVIDADES ADMINISTRATIVAS Y DE APOYO DE OFICINA.",…
- trb_cargo_hom: Con 289 niveles, por ejemplo: "05-LEY ORGÁNICA DE SERVICIO PUBLICO -LOSEP",…

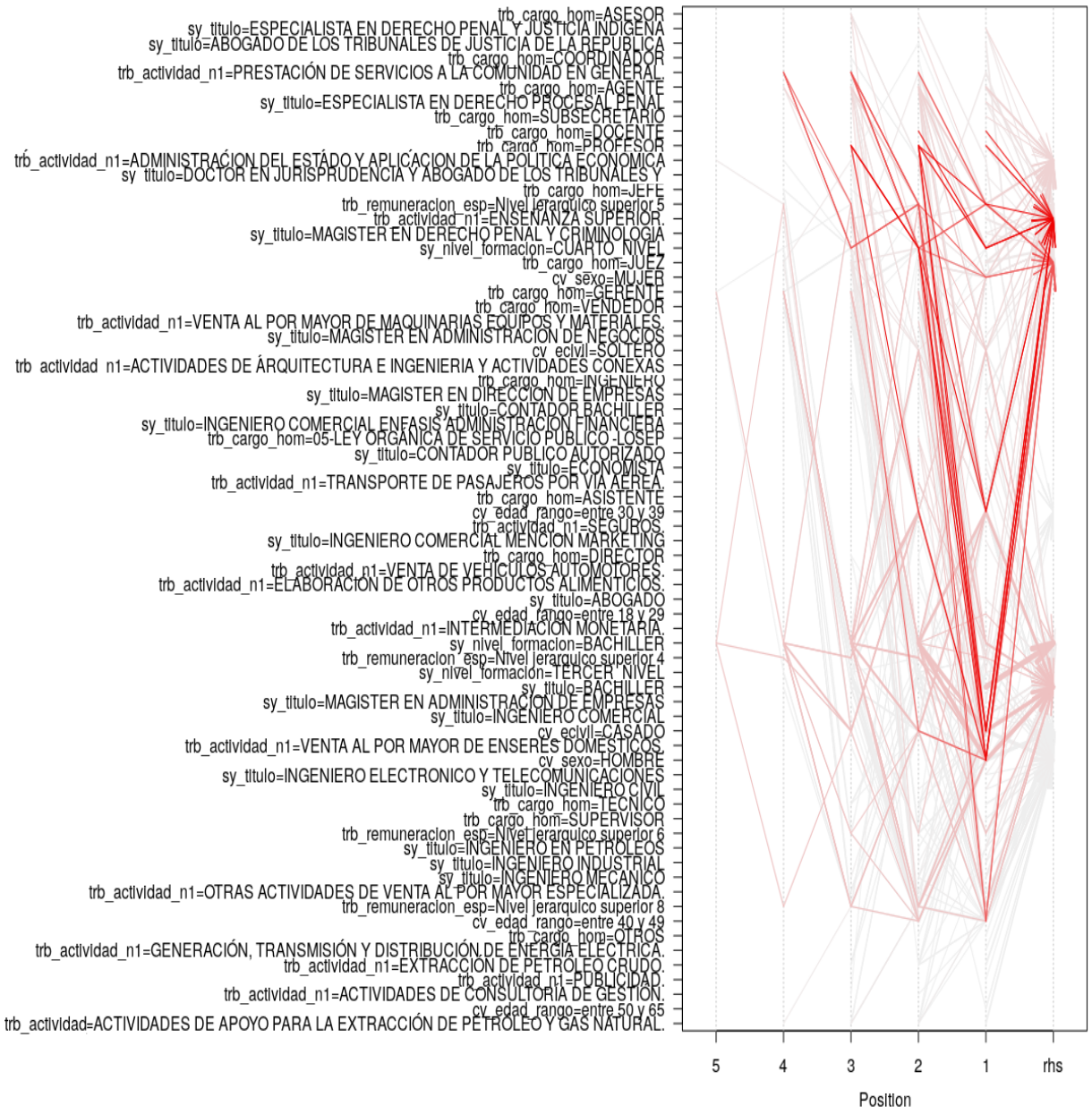

Figura 25. Reglas generadas para el clúster 5

De las reglas generadas, se seleccionó cinco, donde se tomó en cuenta que no sean redundantes, que no sean visibles a simple vista, y que tengan coherencia. Estas muestran a continuación:

i. Si tiene formación de cuarto nivel con cargo de docente; es posible que ocurra que trabaje en actividades de enseñanza superior. Los resultados se muestran en la tabla 22.

Tabla 22
Clúster 5, patrón 1

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[1]{sy_nivel_formación =cuarto_nivel,trb_cargo_hom=docente}	=>{trb_actividad_n1=enseñanza superior.}	0,007015903	0,9677419	25,0791	60

ii. Si es casado con una remuneración equivalente al nivel jerárquico superior 5 (sueldo entre 1.150 y 1.340 dólares americanos), con formación de cuarto nivel, que trabaja en actividades de prestación de servicios a la comunidad en general; es posible que ocurra que tenga el cargo de juez. Los resultados se muestran en la tabla 23.

Tabla 23
Clúster 5, patrón 2

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[13]{cv_ecivil=casado,trb_remuneracion_esp=niveljerarquico superior 5,nivel_ formación= cuarto_nivel, trb_actividad_n1 =prestación de servicios a la comunidad en general.}	=>{trb_cargo_hom= juez}	0,010991581	0,9038462	18,761389	94

iii. Si tiene una remuneración equivalente al nivel jerárquico superior 5 (sueldo entre 1.150 y 1.340 dólares americanos), con el título de especialista en derecho penal y justicia indígena; es posible que ocurra que se desempeñe en actividades de administración del estado y aplicación de la política económica y social de la comunidad. Los resultados se muestran en la tabla 24.

Tabla 24
Clúster 5, patrón 3

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[86]{trb_remuneracion_esp=nivel jerarquicosuperior 5, sy_titulo= especialista en derecho penal y justicia indigena}	=>{trb_actividad_n1= dministración del estado y aplicación de la política económica y social de la comunidad.}	0,007600561	0,9154930	4,682593	65

iv. Su está casado con una remuneración equivalente al nivel jerárquico superior 4 (sueldo entre 1.030 y 1.150 dólares americanos), con el título de ingeniero mecánico; es posible que ocurra que sea hombre. Los resultados se muestran en la tabla 25.

Tabla 25
Clúster 5, patrón 4

Lhs	Rhs	Soporte	Confianza	Lift	cuenta
[197]{cv_ecivil=casado,trb_remuneracion_esp=nivel jerárquico superior 4, sy_título=ingeniero mecánico}	=>{cv_sexo=hombre}	0,007951356	0,9855072	1,475759	68

v. Si tiene el título de ingeniero civil o ingeniero mecánico o ingeniero industrial o ingeniero electrónico y telecomunicación o ingeniero en petróleos; es posible que ocurra que sea hombre. Los resultados se muestran en la tabla 26.

Tabla 26
Clúster 5, patrón 5

Lhs	Rhs	Soporte	Confianza	Lift	Cuenta
[204]{sy_titulo=ingeniero civil}	=> {cv_sexo= hombre}	0,009354537	0,9523810	1,426153	80
[206]{sy_titulo=ingeniero mecanico}	=> {cv_sexo= hombre}	0,022333957	0,9408867	1,408941	191
[217]{sy_titulo=ingeniero industrial}	=> {cv_sexo= hombre}	0,008068288	0,9200000	1,377664	69
[223]{sy_titulo=ingeniero electronico y telecomunicaciones}	=> {cv_sexo= hombre}	0,007366698	0,9130435	1,367247	63
[295]{sy_titulo=ingeniero en petroleos}	=> {cv_sexo= hombre}	0,016487371	0,8294118	1,242012	141

CONCLUSIONES Y RECOMENDACIONES

1. Conclusiones

A través de los resultados obtenidos se presentan las siguientes conclusiones:

- A partir del análisis de las metodologías SEMMA, KDD y CRISP-DM se puede concluir que la última, es apropiada para ser aplicada en el sector del conocimiento y talento humano, por ser totalmente independiente de la herramienta de software, ordenada, de libre acceso, cumple con las necesidades del negocio; a su vez se ha probado que es flexible en todas sus fases, lo que da pautas para ser una ayuda en desarrollos futuros.
- El proceso que más tiempo consumió en el presente estudio, fue el de la limpieza de los datos, donde se procedió a la depuración, transformación, modificación, substitución, entre otros procesos para llegar a obtener una vista minable, ya que de esta dependía el éxito o fracaso del presente estudio. Este tiempo podría ser acortado para futuros proyectos de minería de datos, si en las instituciones existiera un departamento de calidad de datos, que examine y aplique correcciones en las aplicaciones y bases de datos existentes.
- Con el uso de la herramienta R - Studio, aunque posee librerías gráficas como por ejemplo Rattle y FactorMineR GUI, que ayudan en la minería de datos y son a la vez intuitivas. Alcanzan su mayor potencialidad cuando se usan los comandos propios del lenguaje, que por medio de scripts permite la automatización e integración a sistemas externos para que en el futuro pueda ser nuevamente recreado u optimizado.

- En la generación de los grupos el algoritmo k-means, tuvo que ser calibrado con el método experimental codo de Jambú, a fin obtener el *k* (número de clústers) más óptimo y de evitar obtener información redundante o incompleta.
- Antes de proceder a realizar la minería en los datos se debería hacer uso de herramientas de Inteligencia de Negocios o de reportería, lo que se conoce como minería de datos visual. Para tener una idea clara de cómo están comportándose los datos, lo que es de gran ayuda a la hora de encontrar nuevo conocimiento. Esto evita obtener conocimiento redundante o innecesario.
- El uso de reportes gráficos generados y aplicados a los grupos creados por el clúster, permitió conocer la composición de los datos y las relaciones entre el área estudio y las actividades económicas. Se generaron cinco grupos, que con el conocimiento previo del negocio permitió obtener resultados intuitivos.
- Al aplicar las reglas de asociación con sus respectivos ajustes y configuraciones a los grupos generados por el clúster, se encontraron patrones u ocurrencias en los datos que, llevados a un proceso de interpretación, generaron nuevo conocimiento útil, descrito en el capítulo cuatro, que ayudaran a la toma de decisiones oportunas.

2. Recomendaciones

- Para iniciar un proceso de minería de datos, es importante tener claro los objetivos de la institución, a fin de entender de mejor manera la información y discriminar aquella redundantes o innecesaria, ya que así es como lo recomienda la fase uno de la metodología CRIPS - DM.
- Antes de aplicar cualquier técnica o proceder a la limpieza de dados es recomendable, usar aplicaciones de tipo Inteligencia de negocios para tener una idea de los datos y a la vez tomar solo aquellos atributos que aporten mayor información para el análisis. Esto evita perder tiempo en limpiezas innecesarias y reduce la complejidad de los cálculos al aplicar los algoritmos.
- Para obtener el éxito en todo proceso de minería de datos, es necesario hacer una limpieza exhaustiva, esto a través de herramientas, algoritmos o scripts, que ayuden a crear la vista minable, ya que de ello dependerá el éxito de los proyectos.
- En el capítulo cuatro, se muestran el análisis de los cinco clústeres generados con sus respectivas reglas de asociación, que muestran el comportamiento del sector educativo con el de las actividades económicas. Se recomienda tomarlo en cuenta para la toma de decisiones.

BIBLIOGRAFÍA

De La Cruz, K., Rivasplata, J., & Flores, C. (2012). *Aplicación del modelo de clusterización basado en el algoritmo de k-means para la segmentación de la morbilidad materna en el Hospital San Bartolomé de la ciudad de Lima-2012*. Lima: Universidad Peruana Unión.

Berry, M., & Gordon, L. (1997). *Data Mining Techniques for Marketing, Sales, and Customer Support*. John Wiley & Sons, Inc.

Berzal, F. (s.f.). *Reglas de asociación*. España: Departamento de Ciencias de la Computación e IA, Universidad de Granada.

Charman, P. (2000). *Step-by-Step Data Mining Guide*. crisp - dm 1.0.

Constituyente, A. (2008). *Constitución de la República del Ecuador*. Montecristi.

Conti, D., & Martínez, J. (s.f.). Reglas de Asociación en Series Temporales:panorama referencial y tendencias. *II Congreso Español de Informática* (pág. 9). La Rioja: II Congreso Español de Informática.

Contreras, J., Molina, E., & Arteaga, P. (s.f.). *Introduccion a la programacióon estadíistica con R para profesores*. EDUC (MEC-FEDER).

Corporation, T. C. (1998). *Introduction to Data Mining and Knowledge Discovery, Two Cross Corporation*. U.S.A: Two Crows Corporation.

De Luca Venegas, M. P. (2006). *Plan para enfocar las campañas bancarias utilizando datamining*. Chile: Universidad de Chile facultad de Ciencias Físicas y Matemáticas.

Delgado, J., Galárraga, F., Fuertes, W., Toulkeridis, T., Villacís, C., & Castro, F. (2016). A proposal of an entity name recognition algorithm to integrate governmental databases. *Fidel* .

Escobar, V. (2007). *Minería Web de Uso y Perfiles de Usuario*. Granada: Editorial de la Universidad de Granada.

Fayyad, U. M. (1996). *Data Mining and Knowledge Discovery: Making Sense out of Data*. IEEE Expert.

Flores, H. (2009). *Deteccion de patrones de daños y averías en la industria automotriz*. Buenos Aires: Universidad Tecnológica Nacional.

Forgy, E. (1965). *Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classification*. Biometrics.

Free Software Foundation, G. (s.f.). *GNU*. Recuperado el 22 de 12 de 2017, de https://www.gnu.org/home/es.html

GNU. (7 de 01 de 2018). *GNU*. Recuperado el 7 de 01 de 2018, de https://www.gnu.org

Gutiérrez, M. (2006). *El rol de las bases de datos espaciales en una infraestructura de datos*. Santiago, Chile: Proceedings of the 9th Conference Global Spatial Data Infrastructure.

Hartl, F. (03 de 2012). *Thoughts on machine learning – the statistical software r*. Recuperado el 04 de 01 de 2018, de https://florianhartl.com/thoughts-on-machine-learning-the-statistical-software-r.html

Hasperué, W. (2012). *Extracción de conocimiento en grandes bases de datos utilizando estrategias adaptativas*. La Plata: Universidad Nacional de la Plata.

Hernández, A., Delgado, E., Rivera, J., & Castellanos, G. (2006). Reducción de Dimensiones para Clasificación de Datos Multidimensionales Usando Medidas de Información . *Scientia et Technica A*, 6.

Humano, M. C. (01 de 06 de 2017). *Ministerio Coordinador de Conocimiento y Talento Humano*. Obtenido de http://www.conocimiento.gob.ec/

IBM. (2018). *IBM BigInsights*. Recuperado el 04 de 01 de 2018, de https://www.ibm.com/support/knowledgecenter/es/SSPT3X_4.1.0/com.ibm.swg.im.infosphere.biginsights.product.doc/doc/bi_predictive_modeling.html

Izurieta Ballesteros, G. A. (2016). *Estudio de la igualdad de acceso a la educación general por grupos de edad y áreas geográficas en el Ecuador y su relación con la equidad de la distribución de ingresos en el período 2000-2014*. Quito - Ecuador: Escuela Politécnica Nacional.

KDnuggets. (s.f.). *KDnuggets*. Recuperado el 22 de 12 de 2017, de https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

KDnuggets. (s.f.). *KDnuggets*. Recuperado el 22 de 12 de 2017, de https://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html

Lara, J. (2010). *Marco de Descubrimiento de Conocimiento para Datos Estructuralmente Complejos con Énfasis en el Análisis de Eventos en Series Temporales.* Madrid: Universidad Politécnica de Madrid.

López, A., Valcárce, M., & Barbancho, M. (s.f.). Indicadores Cuantitativos y Cualitativos para la Evaluación de la Actividad Investigadora. *Cuadernos IRC*.

Macas, M., Fuertes, W., Guerrero, G., & Toulkeridis, T. (2017). Data Mining Model in the Discovery of Trends and Patterns of Intruder Attacks on the Data Network as a Public-Sector Innovation. *IEEE*, 8.

Maria del Socorro, B. (2014). *Algoritmos de minería de datos en la recolección de inteligencia.* México: Unidad Profesional Interdisciplinaria de Ingeniería y Ciencia Sociales y Administrativas.

Martínez, C. (2012). *Aplicación de técnicas de minería de datos para mejorar El proceso de control de gestión en entel.* Santiago de chile: Universidad de Chile.

MCCTH. (2017). *Agenda prospectiva para el Sector del Conocimiento y talento Humano.* Quito: MCCTH.

MCCTH. (2017). *conocimiento.gob.ec*. Recuperado el 01 de 05 de 2017, de Ministerio Coordinador de Conocimiento y Talento Humano: http://www.conocimiento.gob.ec/mision-vision

MCCTH. (2017). *conocimiento.gob.ec*. Recuperado el 01 de 05 de 2017, de Ministerio Coordinador de Conocimiento y Talento Humano: http://www.conocimiento.gob.ec/objetivos

MCCTH. (2017). *Conocimiento.gob.ec*. Recuperado el 01 de 05 de 2017, de Ministerio Coordinador de Conocimiento y Talento Humano: http://www.conocimiento.gob.ec/ministro

Medina, G., Luna, F., Tavarez, J., & Narvaez, R. (2016). Calibración y selección del modelo de aprendizaje no supervisado K-Medias, de una encuesta sobre factores de riesgo en el consumo de drogas entre estudiantes. *Revista de Análisis Cuantitativo y Estadístico*, 9.

Mejia, J. (2002). *Aplicaciones de reglas de asociación para web mining*. Universidad Autonoma Metropolitana.

Mena, D. (2014). *Clasificación de flujos de datos basada en similitud*. Granada: Universidad de Granada.

Miranda, N. (2015). *Reglas de asociación para líneas espectrales*. Chile: Universidad de Chile.

MSDN. (2008). *Conceptos de minería de datos (Analysis Services - Minería de datos),*. Recuperado el 01 de 01 de 2017, de http://msdn.microsoft.com/es-es/library/ms174949.aspx

ONU. (12 de 12 de 2017). *Objetivos de desarrollo sostenible* . Recuperado el 12 de 12 de 2017, de http://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/

Pazmiño Lucio, H. A. (2016). *Diagnóstico del impacto en la calidad de la educación como consecuencia de la normativa vigente 2010-2015, en la renovación del personal académico en las instituciones de educación superior de la Provincia del Pichincha*. Quito-Ecuador: Escuela Politécnica Nacional.

R Users Group, E. (s.f.). Recuperado el 23 de 12 de 2017, de http://rusersgroup.com/blog/R_potencial/

RapidMiner. (25 de 01 de 2018). *RapidMiner*. Recuperado el 25 de 01 de 2018, de https://rapidminer.com

Reyes Mena, F. (2017). *Iimplementación de una aplicación de business intelligence que permita el análisis de vulnerabilidades para mitigar el impacto de los incidentes en la red de investigación cedia. caso de estudio espe*. Sangolqui: ESPE.

Reyes-Mena, F., Fuertes-Díaz, W., Villacís-Silva, C., Guzmán-Jaramillo, C., Pérez-Estévez, E., & Bernal-Barzallo, C. (Jan. 2017.). Application of business intelligence for analyzing vulnerabilities to increase the security level in an academic CSIRT. *Rev. Fac. Ing.*, 15.

Santana, J. S., & Farfán, E. M. (2014). *El arte de programar en R: un lenguaje para la estadística*. Mexico: Instituto Mexicano de Tecnología del Agua.

Schettini, P., & Cortazzo, I. (2010). *Análisis de datos cualitativos en la investigación social*. La Plata: Universidad Nacional de la Plata.

Senplades, S. N. (2013). *Plan Nacional para el Buen Vivir 2013-2017*. Quito: Senplades.

Servicio de rentas Internas, S. (s.f.). Recuperado el 29 de 01 de 2018, de http://www.sri.gob.ec/web/guest/que-es-el-sri

Shafique, U. (2014). *A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)*. Pakistan: University of Gujrat.

Shamkant, E. R. (2002). *Fundamentos de sistemas de bases de datos*. Addison-Wesley.

Shapiro, P. (1991). *Descubrimiento, análisis y presentación de las reglas fuertes*. Cambridge: Knowledge Discovery in Databases.

Subsecretaría de Gestión de Información, M. (2017). *intranet.mccth*. Recuperado el 01 de 05 de 2017, de http://intranet.mccth/wordpress

Timarán, S., Hernández, I., Caicedo, S., Hidalgo, A., & Alvarado, J. (2016). *El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional*. Bogotá: Ediciones Universidad Cooperativa de Colombia.

Timarán-Pereira, S. R.-A.-Z.-T.-P. (2016). *El proceso de descubrimiento de conocimiento en bases de datos*. Bogotá: Ediciones Universidad Cooperativa de Colombia.

Weka. (01 de 2018). *Weka*. Recuperado el 03 de 01 de 2018, de https://www.cs.waikato.ac.nz/ml/weka/

Wong, P. (1999). *Visual data mining*. IEEE Computer Graphics and Applications.