ESTUDIO DE SEGUIMIENTO A EGRESADOS POR MEDIO DE TÉCNICAS DE MINERÍA DE DATOS PARA LA UNIVERSIDAD INDUSTRIAL DE SANTANDER

MAROLY MILENA MUÑOZ OSORIO
MAYRA SHIRLEY PINTO MATEUS

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERIAS FISICO-MECANICAS
ESCUELA DE ESTUDIOS INDUSTRIALES Y EMPRESARIALES
BUCARAMANGA
2013

ESTUDIO DE SEGUIMIENTO A EGRESADOS POR MEDIO DE TÉCNICAS DE MINERÍA DE DATOS PARA LA UNIVERSIDAD INDUSTRIAL DE SANTANDER

MAROLY MILENA MUÑOZ OSORIO
MAYRA SHIRLEY PINTO MATEUS

Trabajo de Grado para optar por el título de:
INGENIERO INDUSTRIAL

Director
Ph.D. HENRY LAMOS DIAZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERIAS FISICO-MECANICAS
ESCUELA DE ESTUDIOS INDUSTRIALES Y EMPRESARIALES
BUCARAMANGA
2013

DEDICATORIA

Dedicamos este proyecto a Dios por ser el inspirador para cada uno de nuestros pasos

A Fanny, Daisy, Yesid y Dionisio, por ser los guías en el sendero de cada acto

A Jair, Carlos, Andrea y Edinson por ser el incentivo para seguir adelante con este sueño
Por ellos y para ellos

AGRADECIMIENTOS

A Henry por orientar y compartir su conocimiento, por su tiempo, paciencia y dedicación.

Al grupo de investigación ÓPALO por generar conocimiento y por la disponibilidad de los recursos para la ejecución y cumplimiento de cada uno de los objetivos.

A la Vicerrectoría Académica de la UNVIERSIDAD INDUSTRIAL DE SANTANDER por apoyar e impulsar el desarrollo de este trabajo.

A Javier y Oscar por su colaboración y atentas respuestas a las diferentes inquietudes e inconvenientes surgidos durante el desarrollo de este trabajo.

TABLA DE CONTENIDO

Pág.

LISTA DE CUADROS

LISTA DE FIGURAS

RESUMEN

TÍTULO: ESTUDIO DE SEGUIMIENTO A EGRESADOS POR MEDIO DE TÉCNICAS DE MINERÍA DE DATOS PARA LA UNIVERSIDAD INDUSTRIAL DE SANTANDER.[*]

AUTOR: MAROLY MILENA MUÑOZ OSORIO, MAYRA SHIRLEY PINTO MATEUS.[**]

PALABRAS CLAVES: Minería de datos, Redes bayesianas, Árboles de decisión, Análisis de clúster, Reglas de asociación, Seguimiento a egresados.

DESCRIPCIÓN

La minería de datos nace de la necesidad de trabajar con grandes cantidades de datos sin que se pierda la escalabilidad, principal limitación de la estadística básica. La importancia de la minería de datos en estudios de seguimiento a egresados para las universidades radica en la posibilidad de obtener información no visible sobre los egresados que permite orientar las estrategias hacia ellos de una forma diferente y útil. En la presente investigación se aprecia la eficiencia de la minería de datos en la Universidad Industrial de Santander. Se emplean herramientas como árboles de decisión, redes bayesianas, análisis de clúster, reglas de asociación y cubos OLAP, con el propósito de encontrar patrones de comportamiento en los egresados y así construir perfiles que permitan a la universidad tomar decisiones favorables respecto a la oferta educativa. Como herramienta computacional se usa el software de IBM SPSS Modeler para desarrollar las técnicas anteriormente nombradas. Una vez se completa la fase de análisis y evaluación de los datos obtenidos, se logra una caracterización de los egresados elaborando estrategias de mejora en cada segmentación encontrada de la misma forma en cómo se generaron estrategias enfocadas en la captura y almacenamiento de datos, el fortalecimiento de la relación egresado – universidad y la presentación resultados concluyentes hacia la comunidad universitaria.

[*] Proyecto de grado

[**] Facultad de Ingenierías Físico-Mecánicas, Escuela de Estudios Industriales y Empresariales, Director Ph D Henry Lamos Diaz

ABSTRACT

TITLE: FOLLOW-UP STUDY OF GRADUATES THROUGH DATA MINING TECHNIQUES FOR THE INDUSTRIAL UNIVERSITY OF SANTANDER. *

AUTHOR: MILENA MUÑOZ MAROLY OSORIO, MAYRA SHIRLEY PINTO MATEUS.**

KEY WORDS: Data mining, Bayesian networks, Decision trees, Cluster analysis, Association rules, Monitoring graduates.

DESCRIPTION

Data mining originates from the need to handle large amounts of data without losing scalability, which is the main limitation of basic statistics. The importance of data mining in follow-up studies of graduates for universities lies in the possibility of obtaining information that is not visible about the graduates which allows the strategies toward them to be guided in a different and useful way. The present study shows the efficiency of data mining in the Industrial University of Santander. Tools such as decision trees, Bayesian networks, cluster analysis, association rules and OLAP cubes are used, with the purpose of finding behavior patterns in the graduates and to then build profiles that enable the university to make favorable decisions regarding what is offered educationally. As a computational tool, the IBM SPSS Modeler software is used to develop the techniques previously mentioned. Once the analysis and evaluation of the obtained data phase is completed, a characterization of the graduates is achieved, developing strategies for improvement in each segmentation which was found in the same way as strategies focused on the capture and storage of data, the strengthening of the relationship of the graduate – the university and the presentation of conclusive results to the university community.

* Work degree

** Faculty of Physico-Mechanica Engenieering, School of Industrial and Business Studies, Director Ph D Henry Lamos Diaz

INTRODUCCIÓN

El buen manejo de información permite tomar decisiones y analizar diferentes situaciones que se pueden presentar en una Institución de Educación Superior, de tal manera que se pueda incentivar el desarrollo y el uso de métodos eficientes para extraer conocimiento útil proveniente de bodegas de bases de datos que contienen información acerca de los graduados.

Las Instituciones de Educación Superior (IES) de Colombia y del mundo realizan estudios de seguimiento a graduados con el fin de determinar relaciones entre las competencias adquiridas en su formación y las requeridas por los empleadores. Naturalmente, se espera que al comprender claramente el proceso por parte de las IES, permitirá que la transición por parte de los graduados al mercado laboral sea expedita, rápida y pertinente.

Es por ello, que para la Universidad Industrial de Santander es de vital importancia supervisar constantemente la calidad de sus profesionales con el fin de realizar una mejora continua que beneficie tanto al egresado como al sector productivo del país. A partir de los datos recolectados desde el momento de egreso y durante cierta ventana de tiempo, mediante la minería de datos, se buscará encontrar patrones y relaciones entre las variables relevantes en la formación y el desarrollo de competencias para el desempeño laboral.

Una tipología adecuada de los egresados proporciona un enfoque diferente en la relación universidad - graduado, a través de características o similaridades, facilitando la construcción de estrategias de educación y relaciones más próximas que generen beneficios en las dos direcciones. La aplicación de herramientas de

minería de datos en la gestión del seguimiento a graduados es una tendencia poco usada pero que puede llegar a ser muy útil.

El objetivo del presente trabajo se enfoca en el análisis del programa de seguimiento a egresados, mediante la elaboración de perfiles de los egresados y las relaciones de dependencia e independencia entre las variables analizadas.

El estudio presenta un estado del arte sobre el seguimiento a graduados en diferentes partes del mundo, la manera en cómo se desarrolla actualmente en la Universidad Industrial de Santander y un marco teórico relacionado con el tema de minería de datos. Además, el trabajo muestra la metodología CRIPS – DM empleada en los proyectos que utilizan minería de datos en el sector educativo y finaliza con la especificación de resultados y estrategias para la construcción de los modelos conceptuales, así como sus respectivas conclusiones, recomendaciones y bibliografía.

TABLA DE CUMPLMIENTO DE OBJETIVOS

OBJETIVO	DESCRIPCIÓN	CUMPLIMIENTO
1	Encontrar las relaciones de dependencia e independencia entre las variables estudiadas por medio de los clasificadores de Redes Bayesianas.	**Capítulo 4 y 5**
2	Generar a través de los árboles de decisión un sistema de clasificación adecuado para los egresados de la Universidad Industrial de Santander.	**Capítulo 4 y 5**
3	Agrupar los datos en “grupos” naturales teniendo en cuenta la similitud de los elementos por medio de la tarea descriptiva clustering.	**Capítulo 4 y 5**
4	Encontrar correlaciones entre las variables analizadas por medio de las reglas de asociación.	**Capítulo 5**
5	Obtener patrones de los datos analizados para conocer el comportamiento general de la base de datos de egresados.	**Capítulo 5**
6	Construir un modelo de clasificación para los egresados de la Universidad Industrial de Santander.	**Capítulo 5**
7	Establecer las necesidades no conformes de cada una de las divisiones encontradas en la base de datos a fin de convertirlas en oportunidades.	**Capítulo 5**

1. GENERALIDADES DEL PROYECTO

1.1. UNIVERSIDAD INDUSTRIAL DE SANTANDER[1]

La Universidad Industrial de Santander (UIS) es una institución de educación pública de carácter oficial, del orden departamental, que está encaminada fundamentalmente a la formación del hombre, mediante la generación y difusión del saber en sus diversas ramas. Su sede principal se encuentra ubicada en la ciudadela universitaria en la carrera 27 con calle 9 de la ciudad de Bucaramanga, la facultad de salud se encuentra ubicada en inmediaciones del Hospital Universitario de Santander, cuenta también con el edificio de la Sede Bucarica ubicado en el centro de la ciudad y con la sede de Guatiguará ubicada en el Valle de Guatiguará en el municipio de Piedecuesta el cual pertenece al Área metropolitana de Bucaramanga. La UIS también cuenta con cuatro sedes regionales ubicadas en los municipios de Barbosa (Santander), Barrancabermeja, El Socorro (Santander) y Málaga (Santander).

La Universidad Industrial de Santander, fundada hace 65 años, considera a sus ex alumnos como una fuerza considerable dado que sus opiniones influyen de una manera decisiva en la organización y orientación de las instituciones, razón por la cual se debe promover, coadyuvar y estimular la creación de asociaciones, laboratorios de observación y departamentos que permitan mantener una relación entre los egresados y la universidad a fin de brindar una mejor oferta educativa acorde a las necesidades de la sociedad.

1.1.1. Misión. La Universidad Industrial de Santander es una organización que tiene como propósito la formación de personas de alta calidad ética, política y

[1] Suministrado por la Universidad Industrial de Santander

profesional; la generación y adecuación de conocimientos; la conservación y reinterpretación de la cultura y la participación activa liderando procesos de cambio por el progreso y mejor calidad de vida de la comunidad. Orientan su misión los principios democráticos, la reflexión crítica, el ejercicio libre de la cátedra, el trabajo interdisciplinario y la relación con el mundo externo.

1.1.2. Visión. Como visión general en el año 2018, la Universidad Industrial de Santander se habrá fortalecido en su carácter público, aportando al desarrollo político, cultural, social y económico del país, como resultado de un proceso de generación y adecuación de conocimiento en el cual la investigación constituye el eje articulador de sus funciones misionales. La Universidad habrá desarrollado exitosamente una política de crecimiento vertical, mediante la cual se crearán y consolidarán programas de maestría y doctorado de alta calidad, sustentados en procesos de investigación pertinente para la región y el país.

La Institución habrá contribuido al desarrollo regional, mediante la formación del talento humano, la investigación y la extensión, reflejado en el mejoramiento de la calidad de vida, la competitividad internacional y el crecimiento económico. Como parte de este proceso, se ampliará la cobertura con la creación y consolidación de programas misionales pertinentes y soportes estratégicos en su sede central y en sus sedes regionales tanto a nivel profesional como a nivel tecnológico, atendiendo a la política de formación por ciclos aprobada por el Consejo Superior.

La Universidad habrá consolidado una política de articulación global que le ha permitido incrementar de manera significativa los resultados de sus procesos misionales mediante la cooperación con instituciones educativas y de investigación de alto prestigio, empresas, entidades gubernamentales, egresados y otros entes públicos y privados nacionales e internacionales.

La Universidad habrá fortalecido en toda su organización una cultura de gestión de alta calidad de los procesos misionales, estratégicos y de apoyo.

1.1.3. Programas Académicos

Facultad de Ciencias

- Biología
- Física
- Licenciatura en Matemáticas
- Matemáticas
- Química

Facultad de Ciencias Humanas

- Derecho
- Economía
- Filosofía
- Historia
- Licenciatura en Educación Básica con Énfasis en Ciencias Naturales y Educación Ambiental
- Licenciatura en Educación Básica con Énfasis en Lengua Castellana
- Licenciatura en Español y Literatura
- Licenciatura en Inglés
- Licenciatura en Música
- Trabajo Social

Facultad de Ingenierías Fisicomecánicas

- Diseño Industrial
- Ingeniería Civil
- Ingeniería de Eléctrica

- Ingeniería de Electrónica
- Ingeniería Industrial
- Ingeniería Mecánica
- Ingeniería de Sistemas

Facultad de Ingenierías Fisicoquímicas

- Geología
- Ingeniería Metalúrgica
- Ingeniería de Petróleos
- Ingeniería Química

Facultad de Salud

- Microbiología y Bioanálisis
- Enfermería
- Fisioterapia
- Medicina
- Nutrición y Dietética

1.2. PLANTEAMIENTO DEL PROBLEMA

Desde hace varios años, se ha venido fortaleciendo a nivel internacional la tendencia de evaluación de la actividad universitaria, como una forma de rendición de cuentas a la sociedad y a los gobiernos[2], uno de los beneficios de los estudios de seguimiento a graduados es la posibilidad de analizar la relación entre las competencias adquiridas en la educación superior y las requeridas por los empleadores y así poder comprender los procesos de transición de los graduados

[2] INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY. Red GRADUA2 y la Asociación Columbus. Manual de instrumentos y recomendaciones sobre el seguimiento de egresados, 2006. 823 p.

de la educación superior al mercado laboral[3], de modo tal que se logre la vinculación regular, dinámica y permanente de los egresados de la institución con el sector productivo, propiciándose una sinergia que genere el máximo beneficio entre los entes participantes.

En Colombia, al igual que en otros países de América Latina se ha diseñado una política educativa que no tiene en cuenta en su totalidad las necesidades que demanda la sociedad. Lo anterior se puede evidenciar en una educación no estandarizada y sin planificación, en donde la población no responde a los requerimientos mundiales, ocasionando que la visión sectorial de las políticas públicas diseñadas para cumplir ese objetivo tengan poco éxito o en el peor de los casos incremente la brecha entre lo que requiere la sociedad y lo que la universidad está generando.

A medida que el mundo avanza hacia el desarrollo de una nueva sociedad del conocimiento, resulta cada vez más importante disponer de una educación superior efectiva, que garantice la creación y distribución de los conocimientos y tecnologías que demanda la sociedad. Por ello no sólo se espera que los egresados universitarios posean avanzados y especializados conocimientos y aptitudes, correspondientes a profesionales de alto nivel, sino que también sean flexibles al enfrentar retos no solo relacionados al campo específico en el que han sido entrenados.

Es claro también, que existe un gran esfuerzo por parte del Observatorio Laboral de la Educación (OLE) para incentivar a las IES (Instituciones de Educación Superior) en la realización de estudios de seguimiento a egresados, pero infortunadamente el resultado que se da en la actualidad sigue sin ser concluyente puesto que requiere de estudios continuos a través del tiempo para realizar este

[3] COLOMBIA. OBSERVATORIO LABORAL PARA LA EDUCACIÓN. Experiencias de seguimiento a graduados y articulación con el sector productivo, 2011.

tipo de análisis. Lo señalado se debe a que el estudio de seguimiento a graduados es un proceso que lleva poco tiempo de implementación en Colombia, y se encuentra en etapa de crecimiento, desarrollo y evolución.

Por consiguiente, para la correcta elaboración del seguimiento a graduados es necesario que haya interacción entre las IES (Instituciones de Educación Superior), profesores, estudiantes, egresados, empresarios, entre otros; gracias a que no todos demuestran el mismo interés por este tema, ya sea porque hay desconocimiento de lo que se pretende seguir o porque se considera que el seguimiento no contribuye con información importante para la toma de decisiones o porque existen distintos prejuicios a la hora de proporcionar información.

Se considera entonces, que la Universidad Industrial de Santander, al ser una de las universidades más destacadas a nivel nacional, debe ser partícipe del proceso de creación de una estrategia que contribuya a la consolidación de un sistema que ofrezca información completa y confiable para el contacto con los egresados y para el fortalecimiento del proceso de seguimiento realizado por el Observatorio Laboral para la Educación (OLE), esto con el fin de analizar las posibles mejoras de los procesos de formación y asegurar así una correcta restructuración de los planes de estudio

1.3. JUSTIFICACIÓN DEL PROYECTO

El seguimiento a graduados, es un tema de interés por parte de las diferentes Instituciones de Educación Superior (IES) a nivel mundial, estos estudios traen consigo abundantes beneficios como lo son el impacto que tienen los programas de pregrado sobre la sociedad, el conocimiento de la relación entre los egresados y las empresas, la retroalimentación de los programas de pregrado que permiten hallar la congruencia entre los objetivos y el perfil del egresado con las

expectativas y demanda del campo profesional, la identificación de las causas de deserción de los estudiantes y crear una herramienta eficaz para el mejoramiento e incremento de la calidad de la educación . Además se pretende crear una cultura evaluativa que ejerza controles en cuanto a las tareas académicas y la planeación de la educación sobre bases firmes.

Universidades de gran impacto nacional e internacional como la Universidad Autónoma de México, la Universidad Politécnica de Valencia, La Red Alfa Gradua2 (perteneciente a la Unión Europea), y el Instituto Tecnológico de Estudios Superiores de Monterrey realizan de manera sistemática el seguimiento a sus egresados; sin embargo a pesar del interés que ha despertado este tipo de estudios en todo el mundo, en Colombia, salvo ciertas IES, no se cuenta con una metodología estandarizada para realizar un correcto seguimiento a egresados.

En el presente estudio se pretende no solo usar técnicas estadísticas univariadas y bivariadas para la descripción del perfil de los egresados. Las técnicas de minería de datos y el análisis multivariados permiten descubrir ciertos tipos de patrones difíciles de hallar usando herramientas tradicionales de la estadística. Por lo tanto este trabajo se realiza con el propósito de ayudar a subsanar esta deficiencia proponiendo utilizar técnicas modernas de la minería de datos en la base de datos adquirida para identificar comportamientos y patrones de los egresados que contribuyan con el mejoramiento de la educación superior.

1.4. OBJETIVOS

1.4.1. Objetivo General. Utilizar técnicas de minería de datos sobre la base de datos del Observatorio Laboral para la Educación (OLE) en el seguimiento a egresados de la Universidad Industrial de Santander, con el fin de encontrar patrones de comportamiento no evidentes que generen conocimiento útil.

1.4.2. Objetivos Específicos.

- Encontrar las relaciones de dependencia e independencia entre las variables estudiadas por medio de los clasificadores de Redes Bayesianas.
- Generar a través de los árboles de decisión un sistema de clasificación adecuado para los egresados de la Universidad Industrial de Santander.
- Agrupar los datos en “grupos” naturales teniendo en cuenta la similitud de los elementos por medio de la tarea descriptiva clustering.
- Encontrar correlaciones entre las variables analizadas por medio de las reglas de asociación.
- Obtener patrones de los datos analizados para conocer el comportamiento general de la base de datos de egresados.
- Construir un modelo de clasificación para los egresados de la Universidad Industrial de Santander.
- Establecer las necesidades no conformes de cada una de las divisiones encontradas en la base de datos a fin de convertirlas en oportunidades.

1.5. ALCANCE

En la realización de este trabajo de grado se espera obtener patrones de comportamiento y una caracterización de los egresados de acuerdo a la generación de segmentos que describan por medio de características específicas la conformación de grupos sobre los cuales diseñar estrategias de mejora. El proyecto se llevará a cabo desde una perspectiva NO DIRIGIDA en la cual se parte de los datos para la obtención de patrones que, a medida que se van descubriendo, se estime si pueden ayudar a resolver algunas necesidades de la universidad en pro de fortalecer el impacto de los egresados en la sociedad.

El proyecto inicia con la consolidación, depuración y adecuación de los datos para el posterior tratamiento estadístico y de minería de datos.

Mediante programas como SPSS MODELER y SPSS PASW se aplicaran diversas técnicas de minería de datos como Redes Bayesianas, Árboles de Decisión, Reglas de asociación y Análisis de Clúster que permitan la consolidación de resultados de relaciones entre las variables, clasificación de los egresados y la asociación de los mismos en grupos o conjuntos específicos.

Por último, el entregable final es un modelo conceptual de clasificación de los egresados para la universidad que facilite la generación de estrategias de contacto que fortalezcan la relación universidad – egresado y a su vez le brinden a posibilidad a la universidad de obtener información suficiente para hacer estudios de seguimiento a egresados concluyentes.

2. SEGUIMIENTO A EGRESADOS

Estrategia evaluativa que permite conocer la situación, desempeño y desarrollo profesional de los egresados de una carrera profesional; lo anterior coincide con la definición de Barrón, et al (2003) quienes afirman que los estudios de seguimiento de egresados son todas “las propuestas metodológicas que tienen el objetivo de conocer el destino laboral, ocupacional o escolar de quienes han salido del mismo ciclo, nivel, subsistema, modalidad, institución o programa educativo (p.31).

Un estudio de seguimiento a egresados (ESE) posibilita el análisis del grado de adecuación entre la formación recibida en la universidad con las exigencias del mundo del trabajo, aportando información significativa para la toma de decisiones a nivel curricular; éste debe considerar el rendimiento académico, la pertinencia y los resultados de otros tipos de trabajo como los estudios de opinión de empleadores y especialistas, y un análisis riguroso de los planes de estudio[4].

Para la recopilación de la información se realizan preguntas principalmente sobre las siguientes áreas:

- Los antecedentes de educación superior.
- El mercado laboral.
- La situación laboral.

[4] DAMIÁN, Javier Simón. El técnico superior universitario en administración : Origen, trayectoria estudiantil y desarrollo profesional. Universidad del Papaloapan, campus Tuxtepec. Oaxaca, México, 2003. p. 26 – 28.

A continuación se presentan algunos objetivos que pueden cumplir dichos estudios para[5]:

- Evaluar la pertinencia y la calidad de los planes de estudios.
- Mejorar el diseño de los planes de estudio.
- Ayudar a los estudiantes a elegir una carrera.
- Comunicar a los ex-alumnos.
- Obtener indicadores de la calidad de la educación.
- Evaluar el nivel de satisfacción de los egresados con su formación.
- Tomar mejores decisiones de mercadeo.
- Conocer el nivel de inserción de los egresados en el mercado laboral y en sus carreras profesionales.
- Satisfacer las necesidades de los empleadores.
- Diseñar programas ad hoc de capacitación, de postgrado y de educación continua.
- Evaluar la precisión de la educación de los egresados con respecto a su trabajo. Verificar si la misión de la universidad se refleja en la realización personal de los egresados y su compromiso.

2.1. ETAPAS BÁSICAS EN EL ESTUDIO DE SEGUIMIENTO A EGRESADOS[6]

En general, la implementación de estudios de graduados y de empleadores implica el seguimiento de tres etapas:

[5] RAMOS, Teófilo. Manual de instrumentos y recomendaciones sobre el seguimiento de egresados. Red GRADUA2 : Asociación Columbus, 2006. p. 13 - 15

[6] MEXICO. SECRETARIA DE EDUCACIÓN PÚBLICA. Manual de Operación e Instructivo para la captura de datos estadísticos en el Sistema Integral de Información : SII-DGEST. México, 2012.

2.1.1. Desarrollo de concepto e instrumento. Etapa que define los objetivos del estudio y se encarga de diseñar el estudio (selección de las cohortes de graduados que serán incluidas; estrategias para rastrear a los graduados). Además incluye los conceptos técnicos para llevar a cabo el estudio, la formulación de preguntas y respuestas y la elaboración del formato de los cuestionarios.

2.1.2. Recolección de datos. Consiste en el entrenamiento del equipo de investigación, distribución y recolección de los cuestionarios y en una estrategia para asegurar la alta participación (acciones de recordatorio).

2.1.3. Análisis de los datos y elaboración del informe. Etapa que busca definir los sistemas de codificación para las respuestas a las preguntas abiertas, entrada y edición (control de calidad) de los datos, análisis de los datos, preparación del informe del estudio y el taller de socialización de resultados con estudiantes, graduados y empleadores

2.2. TEMAS PRINCIPALES EN EL SEGUIMIENTO DE EGRESADOS[7]

2.2.1. Perfil del egresado. Permite conocer la evolución profesional y personal del egresado para establecer la relación entre diversas variables relacionadas con su situación social, familiar, económica y su trabajo, estudios, entre otras.

El perfil del egresado incluye lo siguiente:

- Datos sociodemográficos. Edad, género, estado civil, lugar de nacimiento, número de hijos y las edades de estos, procedencia de los padres, fecha de nacimiento, lugar de residencia, otros.

[7] RAMOS, Teófilo. Manual de instrumentos y recomendaciones sobre el seguimiento de egresados. Red GRADUA2 : Asociación Columbus, 2006. p. 25 – 27.

- Antecedentes educativos. Educación básica, secundaria, universitaria, educación post universitaria. Se puede recopilar información sobre la carrera que estudió y graduó. Por ejemplo, si fue primera opción, qué carrera, motivaciones para seleccionar la carrera y la universidad en que estudió país y ciudad donde estudio, el turno en que estudió, tiempo que dura la carrera y tiempo real que invirtió para graduarse en dicha carrera, la calificación promedio que logró al final de la carrera, los idiomas que habla, otros.
- Otros estudios realizados. Estudios adicionales a la carrera universitaria, cambios en el nivel académico y profesional, en qué área fueron hechos los estudios posteriores, cuáles fueron los motivos por los que volvió a estudiar.
- Fuente de financiamiento de los estudios universitarios. Personas o agencias que financiaron su educación universitaria. Si tuvo beca durante todos los años o durante algunos años de los estudios universitarios. Si los pagó personalmente, o la familia contribuyó, o si recibió créditos de alguna institución financiera para pagarlos después con intereses.
- Movilidad durante la formación. Si el egresado cambió su carrera en algún momento. Si cambió de campus universitario o de universidad. Cuáles fueron los motivos. Los gastos en los que incurrió debido a los cambios.

2.2.2. Situación de los egresados en el mercado de trabajo. Las transformaciones técnico-científicas han requerido nuevos modos de organización tanto para los procesos de trabajo como para los procesos de formación. Este escenario demanda acciones constantes de seguimiento y evaluación de los egresados, ya sea para la reorganización interna de los programas, ya sea para la proposición de nuevos programas.

Para lograr estos propósitos, los estudios de seguimiento de egresados deben considerar el análisis de los siguientes factores:

- **El primer empleo**

  El acceso al primer empleo:

  ✓ Tiempo transcurrido para la obtención del primer empleo.

  ✓ Medio para la obtención del primer empleo (bolsa de trabajo, contactos personales, avisos en la prensa).

  Las características del primer empleo:

  ✓ Salario.

  ✓ Puesto desempeñado.

  ✓ Sector económico de la organización.

  ✓ Tipo de organización (pública o privada).

  ✓ Tipo de actividad (dependiente o independiente).

  ✓ Posición jerárquica en la organización.

- **La trayectoria profesional**

  ✓ Número de empleos.

  ✓ Periodos y duración de ocupación/desocupación laboral.

  ✓ Tipo de puestos desempeñados.

  ✓ Experiencia internacional.

- **Situación laboral actual.** Se deben considerar los puntos mencionados en el inciso del *"primer empleo".*

- **La coherencia entre la formación y el tipo de empleo**

  ✓ Relación del empleo con el área de estudio.

  ✓ El título le garantizó el ingreso o fue irrelevante.

  ✓ La contratación requirió preparación especial para el acceso al empleo.

  ✓ La formación le permitió al egresado responder a las demandas del empleo con relación a: desempeño de habilidades operacionales, tomas de decisiones, iniciativa necesaria en su desempeño.

2.2.3. Relación con la institución de egreso. Evalúa la satisfacción de los egresados en relación a los servicios que le ofreció la universidad, con el propósito de fortalecer la vinculación con ellos y para el mejoramiento continuo de la institución.

- **Satisfacción con la formación recibida**
  - ✓ Calidad de los docentes (nivel de conocimiento de los catedráticos, capacidades docentes o pedagógicas, vinculación de los docentes con los estudiantes).
  - ✓ Plan de Estudios. Las universidades pueden recabar de sus egresados opiniones o recomendaciones para la mejora de los planes de estudios basados en su experiencia profesional.

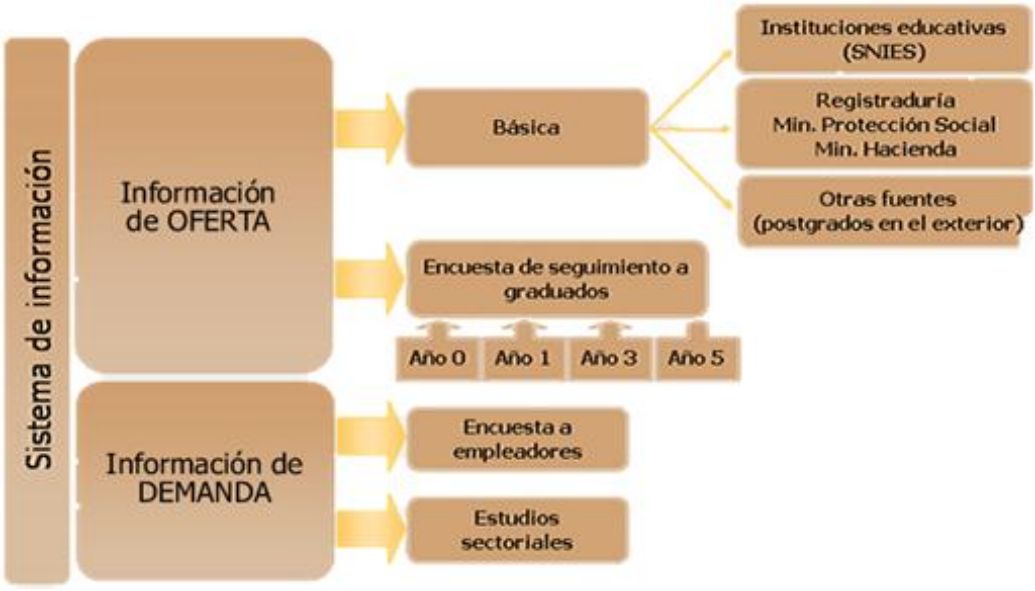
- **Satisfacción con las condiciones de estudio (servicios, infraestructura).**

2.3. INSTRUMENTO PARA LA RECOLECCIÓN DE DATOS: OBSERVATORIO LABORAL PARA LA EDUCACIÓN (OLE)

El Observatorio Laboral para la Educación es un sistema de información que brinda herramientas valiosas para analizar la pertinencia de la educación a partir del seguimiento a los graduados y su empleabilidad en el mercado laboral[8]. De esta manera, contribuye al mejoramiento de la calidad de los programas académicos ofrecidos.

[8] COLOMBIA. GRADUADOS COLOMBIA. Observatorio Laboral para la Educación. En : Visión Educativa [en línea]. [Consultado 14 Febrero 2013]. Disponible en <http://www.graduadoscolombia.edu.co>.

Figura 1. Estructura del sistema de información



Fuente. Observatorio Laboral para la Educación. Ministerio de Educación.

2.2.4. Beneficios del Sistema9

- **Beneficios del sistema para estudiantes y egresados**
  - ✓ Conocimiento sobre la oferta educativa para tomar decisiones sobre las elecciones de carreras. Las estadísticas permitirán conocer el capital humano disponible en el país y las tendencias de demanda.
  - ✓ Guía básica para elegir carrera. Adicionalmente, genera las opciones de créditos y becas que existen en el país y fuera de él para cursar una especialización, maestría o doctorado.
  - ✓ Acceso a bolsas de empleo en línea, nacional e internacional y asesoría en la Legislación Laboral para responder a todas las dudas jurídicas que surgen al ingresar al mundo del trabajo.
  - ✓ Información útil para analizar la pertinencia de la educación.

[9] COLOMBIA. MINISTERIO DE EDUCACION. Colombia APRENDE : Graduados Colombia : Beneficios del sistema de información del Observatorio Laboral para la Educación. 2007.

- **Beneficios del sistema para las instituciones de Educación Superior**
  - ✓ Consulta de información básica de los graduados de cada institución de educación superior y de la situación en el mercado laboral. Esta información proviene de los registros de graduados reportados en el SNIES, de la integración de bases de datos con los Ministerios de Hacienda y Protección Social y de la encuesta de seguimiento a graduados.
  - ✓ Información sobre iniciativas nacionales o internacionales que buscan mejorar la pertinencia de la oferta educativa con metodologías similares a la construida por el Observatorio Laboral para la Educación.
  - ✓ Conocimiento de experiencias internacionales de articulación entre la academia y el sector productivo.

- **Beneficios del sistema para el sector productivo.**
  - ✓ Conocimiento por parte de los estudiantes sobre la oferta educativa para tomar decisiones mejor informadas. Estadísticas del capital humano disponible en el país y las tendencias de demanda de las diferentes profesiones.
  - ✓ Vinculación con las estrategias lideradas por el Ministerio de Educación, las cuales buscan establecer espacios de diálogo entre el sector productivo y la academia: Alianzas estratégicas, Vínculos entre la universidad, la empresa y el Estado.
  - ✓ Información de los proyectos del Consejo Privado para la Competitividad y su relación estrecha con el Observatorio Laboral para la Educación.

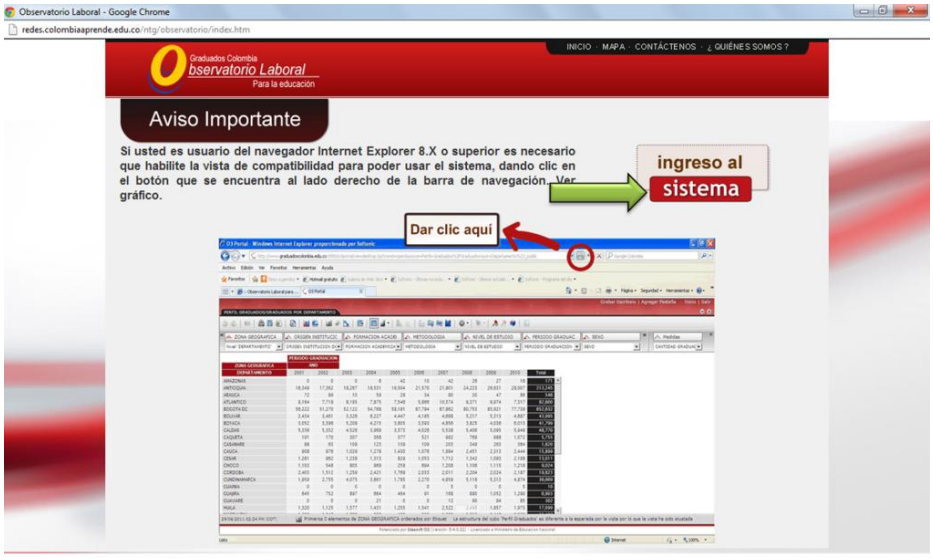
2.2.5. Captura de datos estadísticos por medio de la Plataforma del Observatorio Laboral para la Educación. El sistema de información del OLE se encuentra disponible en la dirección http://www.graduadoscolombia.edu.co. Para ingresar a la página principal del sistema se selecciona la opción “sistema de información del observatorio laboral: ingrese aquí”.

Figura 2. Portal Observatorio Laboral para la Educación.

Fuente. GRADUADOS COLOMBIA. Observatorio Laboral para la Educación.

Una vez seleccionada la opción, se direccionara la página hacia el sistema de información disponible en http://redes.colombiaaprende.edu.co/ntg/observatorio/index.htm, donde se tomara la opción “ingresar al sistema”.

Figura 3. Página principal del sistema de información del OLE.



Fuente. GRADUADOS COLOMBIA. Observatorio Laboral para la Educación.

3. PROCESO DE EXTRACCIÓN DE CONOCIMIENTO

El aumento del volumen y la variedad de información que se encuentra informatizada en bases de datos digitales y otras fuentes han crecido espectacularmente en as últimas décadas. Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido. En muchas situaciones, el método tradicional de convertir los datos en conocimiento consiste en un análisis e interpretación realizada de forma manual, pero el crecimiento exponencial de estas bodegas de datos requieren técnicas mucho más elaboradas que el simple uso de herramientas básicas de estadística, que sólo permiten generar información resumida, poco flexible y, sobre todo, poco escalable a grandes volúmenes de datos. Con el fin de poder extraer conocimiento valioso para la empresa han aparecido enfoques más refinados e inteligentes.

3.1. LA MINERÍA DE DATOS Y EL PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS (KDD)

La inteligencia de negocios se define como la habilidad corporativa para tomar decisiones. Esto se logra mediante el uso de metodologías, aplicaciones y tecnologías que permiten reunir, depurar, transformar datos, y aplicar en ellos técnicas analíticas de extracción de conocimiento (Parr 2000), los datos pueden ser estructurados para que indiquen las características de un área de interés (Stackowiak et al. 2007), generando el conocimiento sobre los problemas y oportunidades del negocio para que pueden ser corregidos y aprovechados respectivamente. (Ballard et al. 2006)

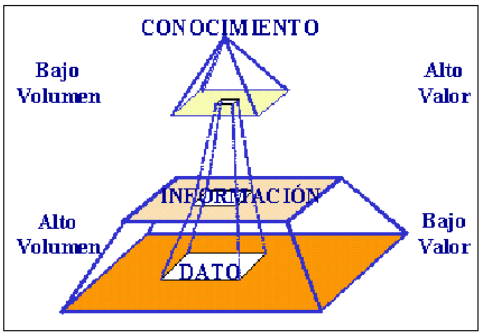
La Extracción de conocimiento está principalmente relacionado con el proceso de descubrimiento conocido como Knowledge Discovery in Databases (KDD), que se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información[10]. No es un proceso automático, es un proceso iterativo que exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información de calidad que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos.

La Minería de Datos (Data Mining) ha surgido precisamente por la necesidad de conocer, trabajar y aprovechar el aumento de los datos permitiendo a las compañías concentrarse en la información más importante de sus bases de información. De forma general, los datos son la materia prima bruta. En el momento en que se les atribuye algún significado especial pasan a convertirse en información. Una vez los especialistas elaboran o encuentran un modelo haciendo que la interpretación del confronto entre la información y ese modelo representen un valor agregado entonces nos referimos al conocimiento[11].

En la figura 5 se presenta un modelo de la relación entre datos, información y conocimiento. En el primer nivel de la pirámide se muestran los datos brutos con los cuales se desea generar nuevo conocimiento para la empresa, conocimiento que se caracteriza por tener un alto valor agregado. La minería de datos trabaja en el nivel superior buscando patrones, comportamientos, agrupaciones, secuencias, tendencias o asociaciones que puedan generar algún modelo que permita comprender mejor el dominio de los datos para ayudar en una posible toma de decisiones.

[10] ANÁLISIS DE CARACTERÍSTICAS del ambiente creativo en empresas de Manizales con técnicas KDD. Universidad Nacional de Colombia. Manizales, 2009.
[11] MOLINA, Luis Carlos. Data Mining: Torturando a los datos hasta que confiesen, 2002.

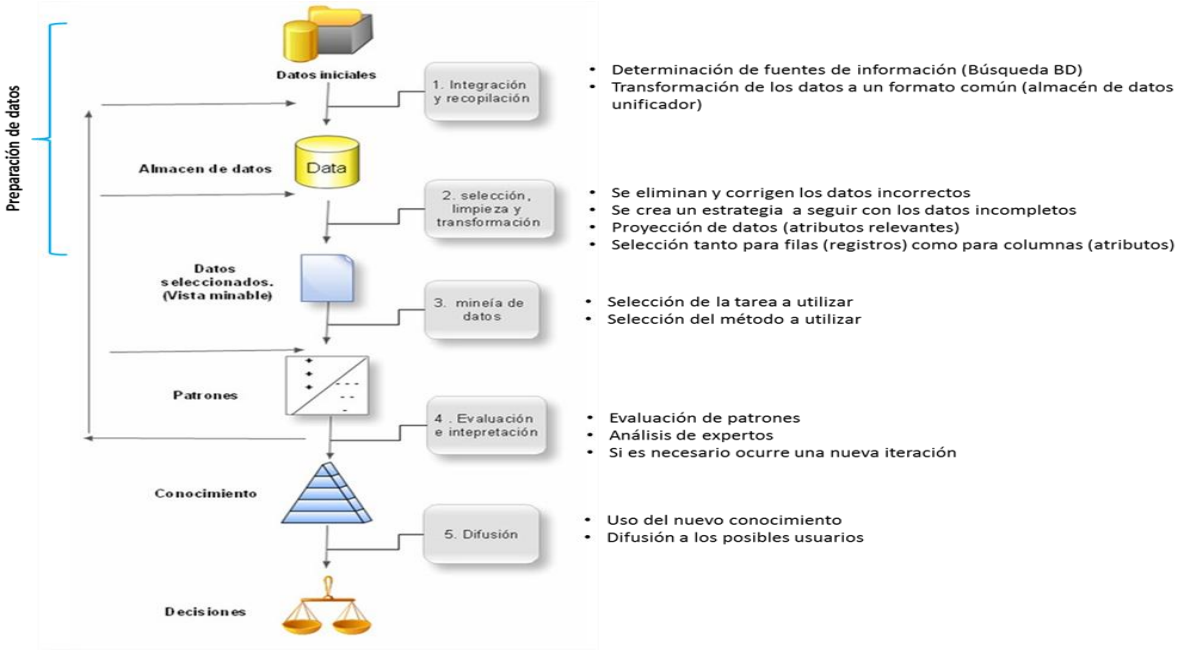
Figura 4. Relación DATO-INFORMACIÓN-CONOCIMIENTO



Fuente. Torturando a los datos hasta que confiesen. Luis Molina. 2002.

Frecuentemente se utilizan sinónimos para referirse a la MD, por ejemplo, "Descubrimiento de Conocimiento en Bases de Datos" (KDD por sus siglas en inglés), siendo esta última un proceso que consta de una serie de fases y en la cual la MD es una de ellas[12]. (Ver figura 6)

Figura 5. Fases del proceso de descubrimiento de conocimiento.



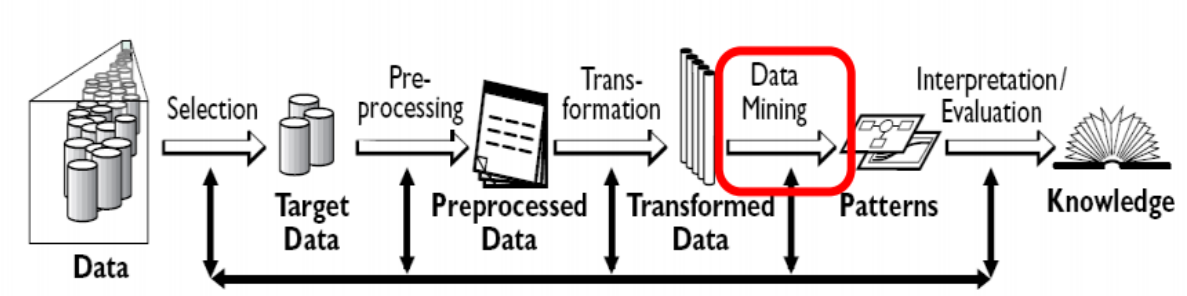
Fuente. Introducción a la minería de datos. José Hernández et al, 2004, p.20

[12] WITTEN & FRANK, Data Mining: Practical Machine Learning Tools and Techniques, citado por HERNÁNDEZ, José et al, Introducción a la minería de datos, 2004, p. 5.

El proceso incluye la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento. El conocimiento extraído debe ser válido, novedoso, potencialmente útil y en última instancia comprensible.

La diferencia puntual entre la MD y el KDD radica en que éste último es el proceso global de descubrir conocimiento útil de las bases de datos mientras que la MD es la aplicación de algoritmos específicos para extraer patrones desde los datos[13].

Figura 6. Proceso KDD – Minería de datos



Fuente. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. from data mining to knowledge discovery 1996.

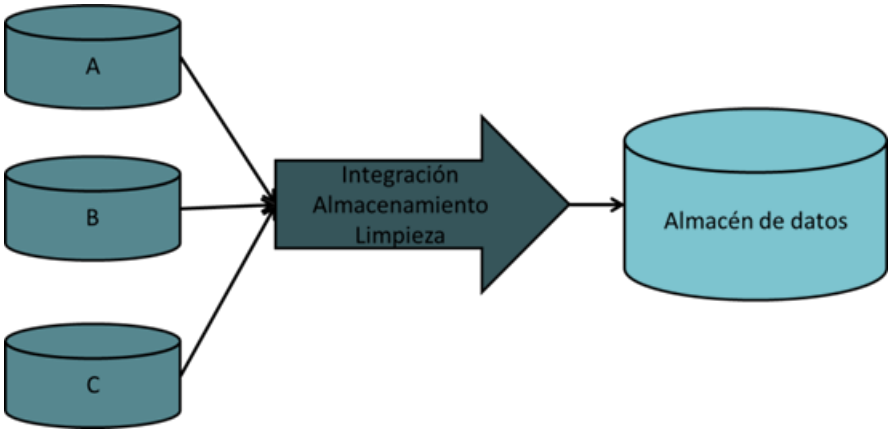
3.1.1. Fases del Proceso de Extracción de Conocimiento. El proceso KDD es iterativo e interactivo. Iterativo ya que la salida de alguna de las fases puede hacer volver a los pasos anteriores y porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. E interactivo porque el usuario debe ayudar en la preparación de los datos, validación del conocimiento extraído, etc. Las fases del KDD son[14]:

[13] HERNÁNDEZ, José et al. Introducción a la minería de datos, 2004, p. 22
[14] IBID

- **Fase de integración y recopilación:** Las bases de datos como OLPT (Procesamiento transaccional en línea) ayudan a cubrir las necesidades diarias de una organización pero son insuficientes para funciones más complejas (analizar, planificar, predecir). Lo ideal para un proceso de KDD es que los datos recolectados provengan de distintas organizaciones, departamentos de una misma entidad e incluso habrán datos que nunca hayan sido recolectados. También se deberá trabajar con bases de datos públicas conjuntamente con las bases de datos privadas.

El problema de combinar bases de datos es que usan diferentes formatos de registro, lo primero por tanto será integrar todos los datos en el almacén de datos (data warehousing) (figura 4). Se recomienda modelar los datos de acuerdo a una estructura base de datos multidimensional (cada dimensión es un atributo o conjunto de atributos) que contenga alguna medida agregada (ej: cantidad vendida de un producto en un día).

Figura 7. Integración en un almacén de datos



Fuente. Introducción a la minería de datos. José Hernández et al, 2004, p.21

El OLAP obtiene información agregada a partir de información detallada, además comprueba patrones y pautas hipotéticas, se trata por tanto de un proceso deductivo. Por el contrario la minería de datos es la encargada de buscar los

patrones, por la tanto es un proceso inductivo. Ambas herramientas se complementan.

- **Fase de selección, limpieza y transformación:** La calidad del conocimiento depende de la calidad de los datos minados. Los datos que resulten de este proceso es lo que se considera vista minable. Esta etapa analiza que datos son irrelevantes para la minería, además de encontrar la presencia de valores que no se ajusten al comportamiento de los datos (outliers – anómalos – ruido). Se debe tener en cuenta que en algunos casos los eventos raros pueden ser más interesantes que los regulares.

La presencia de datos faltantes o perdidos (missing values) es otro problema a resolver. No obstante lo más importante es conocer el origen del defecto para poder manejar los datos faltantes. La selección de atributos relevantes es uno de los pre-procesamientos más importantes, ya que es crucial que los atributos utilizados sean relevantes para la minería de datos. Al considerar un atributo en el proceso de minería de datos se debe generar un algoritmo, el cual aunque sea correcto puede ser inútil si la variable es irrelevante. Entre otras cosas se debe entender que el tiempo requerido para construir un modelo crece con el número de variables, por ello es importante hacer una selección según conocimientos propios.

Como en el caso de las variables, también es necesario hacer una selección de los datos que se van a procesar. Otra tarea de preparación es construir atributos nuevos aplicando alguna operación o función a los atributos originales para darles un uso más fácil con mayor poder predictivo y que tengan más relación con la técnica a utilizar (ej: numerizar atributos para técnicas numéricas); otra forma es discretizar los atributos continuos (ej: asignar a un intervalo un varo discreto).

- **Fase de minería de datos:** El objetivo es producir conocimiento que pueda utilizar el usuario. Esto se realiza construyendo un modelo (una descripción de los patrones y relaciones entre los datos) basado en los datos recopilados para ese efecto que pueden usarse para entender mejor los datos o para explicar situaciones pasados.

Decisiones a tomar:

- ✓ Tipo de tarea de minería más apropiada
- ✓ Tipo de modelo (técnica)
- ✓ Elegir el algoritmo de minería que resuelva la tarea y obtener el tipo de modelo (método). Esto puede significar hacer varias iteraciones que incluyan cambio de atributos, cambio de técnicas, parámetros, etc. a fin de encontrar el “buen modelo”.

- **Fase de evaluación e interpretación:** Aquí se evalúan los patrones y se analizan. Si es necesario se vuelve a las fases anteriores para una nueva iteración. Los patrones descubiertos tienen que tener tres cualidades: ser precisos, ser comprensibles e interesantes (novedosos), los cuales permiten medir la calidad de los mismos y según la aplicación puede interesar mejorar algún criterio en particular.

Técnicas de evaluación

Para entrenar y probar un modelo se deben emplear 2 conjuntos de datos

- ✓ Conjunto de entrenamiento (training set)
- ✓ Conjunto de prueba o test (test set)

Si no se usan conjuntos diferentes de datos la precisión del modelo será sobreestimado (optimista).

Procesos de validación

La precisión en algunos casos se obtiene dividiendo el número de total de resultados correctos sobre el número total de instancias.

- ✓ Validación simple: reserva un porcentaje de la base de datos como un conjunto de prueba y no lo usa para construir el modelo.
- ✓ Validación cruzada con n pliegues: división de los datos en n grupos. Un grupo se reserva para el conjunto de prueba y el resto n-1 se encargan se construir el modelo y se usan para predecir el resultado del grupo reservado.
- ✓ Bootstrapping: para modelos con pocos datos. Se construye un modelo con todos los datos. Después se crean numerosos conjuntos de datos llamados bootstrap samples haciendo un muestreo de datos originales con reemplazo, de esta forma los conjuntos pueden contener datos repetidos. A

- **Fase de difusión, uso y monitorización:** Una vez construido y validado el modelo puede usarse para:
  - ✓ Recomendar acciones basándose en el modelo y sus resultados
  - ✓ Aplicar el modelo a diferentes conjuntos de datos – sistemas de análisis

Es importante distribuir el conocimiento a los posibles usuarios y que pase a integrar el know – how de la organización. También se debe medir cómo evoluciona el modelo ya que los patrones pueden cambiar a fin de poder reentrenar y reconstruir el modelo si es necesario.

3.2. MINERÍA DE DATOS

En [Witten & Frank 2000] se define la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea

fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ser automático o semi-automático (asistido) y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que aporten, por tanto, algún beneficio a la organización.

3.2.1. Tipología de Tareas de Minería de Datos[15]. Es el tipo de problema a ser resuelto y cada problema tiene sus propios requisitos y características:

- **Modelos Predictivos**

  Pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo o dependientes, usando otras variables o campos de la base de datos, a las que se denominan variables independientes o predictivas. Algunas tareas de minería de datos que producen modelos predictivos son:

✓ Clasificación: Cada instancia (registro) pertenece a una clase, la cual se indica mediante el valor de un atributo que llamamos la clase de la instancia. Este atributo puede tomar diferentes valores discretos, cada uno de los cuales corresponde a una clase. El resto de los atributos de la instancia (los relevantes a la clase) se utilizan para predecir la clase. El objetivo es predecir la clase de las nuevas instancias de las que se desconoce la clase.

✓ Regresión: Consiste en aprender una función real que asigna a cada instancia un valor real. Aquí el valor a predecir es numérico. El objetivo en este caso es minimizar el error cuadrático medio entre el valor predicho y el valor real.

[15] HAN, Jiawei; KAMBER Micheline . Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers. Estados Unidos, 2000.

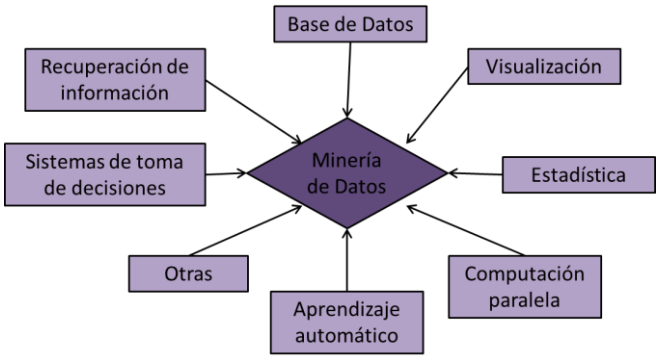
- **Modelos Descriptivos**

  Identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Algunas tareas de minería de datos que producen modelos descriptivos:

  - ✓ Agrupamiento (clustering) – segmentación: Consiste en obtener grupos naturales a través de los datos. Aquí en lugar de analizar datos etiquetados con una clase, los analiza para generar esta etiqueta.
  - ✓ Correlaciones: Se usa para examinar el grado de similitud de los valores de dos variables numéricas. Se mide mediante el coeficiente de correlación r, el cual es un valor real comprendido entre -1 y 1.
  - ✓ Reglas de asociación: Su objetivo es identificar relaciones no explícitas entre atributos categóricos. El estilo de formulación puede ser "si el atributo X toma el valor de d entonces el atributo Y toma el valor de b".

Por otra parte, la minería de datos es un campo multidiscilpinar que se ha desarrollado en paralelo o como prolongación de otras tecnologías. Por ello la investigación y los avances en la minería de datos se nutren de lo que producen estas áreas relacionadas

Figura 8. Disciplinas que contribuyen a la minería de datos.



Fuente. Minería de datos, Olmos y González. Instituto Tecnológico de Puebla. 2007.

3.2.2. Técnicas de Minería de Datos

3.2.2.1. Redes Bayesianas: Las redes bayesianas (RBs) o probabilísticas se fundamentan en la teoría de la probabilidad y combinan la potencia del teorema de Bayes con la expresividad semántica de los grafos dirigidos; las mismas permiten representar un modelo causal por medio de una representación gráfica de las independencias / dependencias entre las variables que forman parte del dominio de aplicación [Pearl, 1988]. Las RBs representan el conocimiento cualitativo del modelo mediante un grafo dirigido acíclico y no se modelan de forma cualitativa el conocimiento, sino que además expresan de forma numérica la "fuerza" de las relaciones entre variables.

Entre las características que poseen los métodos bayesianos en tareas de aprendizaje se pueden resaltar las siguientes[16]:

- Cada ejemplo observado va a modificar la probabilidad de que la hipótesis formulada sea correcta.
- Estos métodos son robustos al posible ruido presente en los ejemplos de entrenamiento y a la posibilidad de tener entre esos ejemplos de entrenamiento datos incompletos o posiblemente erróneos.
- Los métodos bayesianos permiten tener en cuenta en la predicción de la hipótesis el conocimiento a prior o conocimiento del dominio en forma de probabilidades.

Los métodos bayesianos se caracterizan por ser métodos prácticos para realizar inferencias a partir de los datos y por facilitar un marco de trabajo útil para la comprensión y análisis de otras técnicas de aprendizaje y minería de datos que no trabajan explícitamente con probabilidades.

[16] MALAGÓN, Constantino. Clasificadores bayesianos. El algoritmo de Naïve Bayes. 2003

Representación del conocimiento

Una red bayesiana representa relaciones causales en el dominio del conocimiento a través de una estructura gráfica y las tablas de probabilidad condicional entre los nodos, por lo tanto el conocimiento que representa la red está compuesto por los siguientes elementos[17]:

- Un conjunto de nodos $\{X_i\}$ que representan cada una de las variables del modelo. Cada una de ellas tiene un conjunto exhaustivo de estados $\{X_i\}$ mutuamente excluyentes.
- Un conjunto de enlaces o arcos (X_i, X_j) entre aquellos nodos que tienen una relación causal. De esta manera todas las relaciones están explícitamente representadas en el grafo.
- Una tabla de probabilidad condicional asociada a cada nodo X_i indicando la probabilidad de sus estados para cada combinación de los estados de sus padres. Si un nodo no tiene padres se indican sus probabilidades a priori.

Estructura de una red bayesiana

La estructura de una red bayesiana se puede determinar de la siguiente manera:

1. Se asigna un vértice o nodo a cada variable (X_i) y se indica de qué otros vértices es una causa directa; a ese conjunto de vértices "causa del nodo $_{Xi}$" se lo denota como el conjunto P_{ai} y se lo llamará "padres de Xi".
2. Se une cada padre con sus hijos con flechas que parten de los padres y llegan a los hijos.
3. A cada variable X_i se le asigna una matriz $P\ X_i\ P_{ai}$ que estima la probabilidad condicional de un evento $X_i = x_i$ dada una combinación de valores de los P_{ai}.

[17] FELGAER, Pablo. Optimización de redes bayesianas basado en técnicas de aprendizaje por inducción. Buenos Aires, 2005, 145 p. Tesis (Maestría en Ingeniería Informática). Universidad de Buenos Aires. Facultad de Ingeniería.

Para un conjunto de variables aleatorias $X = X_1, \ldots, X_n$ es un par $B = G, P\ \theta$ donde G es un grafo dirigido acíclico, cuyos nodos se relacionan uno a uno con las variables en el vector aleatorio X, y P es un conjunto de funciones de probabilidad local definidas por el conjunto de parámetros. Se usa P_{ai} y p_{ai} para denotar, respectivamente, a los padres y las configuraciones de los padres del nodo X_i en G.

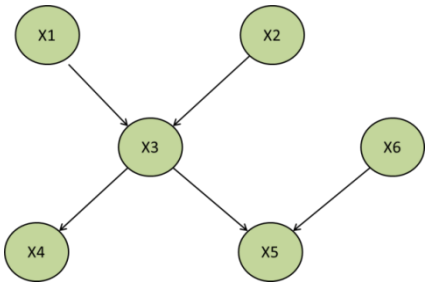
En una red bayesiana todas las relaciones de independencia condicional representadas en el grafo corresponden a relaciones de independencia en la distribución de probabilidad. Si enumeramos los nodos de una red bayesiana, X1, X2, , Xi, de manera que se cumpla que cada nodo aparece en la secuencia antes que cualquiera de sus hijos, dicha red representa el siguiente aserto de independencia:

Cada variable Xi es condicionalmente independiente de las variables del conjunto {X1, X2,…Xi-1} conocidos los valores de sus padres.

Los asertos de independencia condicional junto con las tablas de probabilidad condicional nos permiten obtener la tabla de probabilidad conjunta de todas las variables a partir de las tablas de probabilidad condicional de cada variable en función de sus padres; de esta forma, aplicando la regla de la cadena conjuntamente con la propiedad de independencia condicional se obtiene:

$$P\ X_1, X_2, \ldots, X_n = \prod_{i=1}^{n} P\ X_i\ X_1, X_2, \ldots, X_{i-1} = \prod_{i=1}^{n} P\ X_i\ P_{ai}$$

Figura 9. Ejemplo Red Bayesiana



Fuente. Autores

La regla de la cadena [Pearl, 1988] sostiene que la probabilidad conjunta puede ser calculada como:

$$P\ X_1,\ldots,X_6\ = P\ X_1\ \ P\ X_2\ \ P\ X_3\ X_1,X_2\ \ P\ X_4\ X_3\ \ P\ X_5\ X_3,X_6\ \ P(X_6)$$

El aprendizaje es una de las características que definen a los sistemas basados en inteligencia artificial porque siendo estrictos se puede afirmar que sin aprendizaje no hay inteligencia. El aprendizaje en la redes bayesianas consiste en definir la red probabilística a partir de datos almacenados en bases de datos en lugar de obtener el conocimiento del experto. Este tipo de aprendizaje ofrece la posibilidad de inducir la estructura gráfica de la red a partir de los datos observados y de definir las relaciones entre los nodos basándose también en dichos casos; según Pearl [Pearl, 1988] a estas dos fases se las puede denominar:

- Aprendizaje estructural: obtiene la estructura de la red bayesiana a partir de bases de datos, es decir, las relaciones de dependencia e independencia entre las variables involucradas.
- Aprendizaje paramétrico: dada una estructura y las bases de datos, obtiene las probabilidades a priori y condicionales requeridas.

Redes bayesianas como clasificadores[18]

La estimación de redes bayesianas, informalmente, puede describirse como: dada una muestra $D = U_1, \ldots, U_n$ encontrar una red que mejor la describa. La red aprende un modelo que incluye todas las variables (clase y atributos) del problema y luego realiza la clasificación. La metodología es encontrar una función que evalúe cada red con respecto a D y luego seleccione la mejor de acuerdo a esta función. En una red bayesiana se cumple que toda variable es independiente del resto dado su envolvente de markov, definido como:

(Padres (x) u hijos (X) u padres (hijos (X)))

Es posible seleccionar únicamente dichas variables como variables predictoras útiles para el modelo de clasificación. En general, este problema de optimización no es manejable.

3.2.2.2. Árboles De Decisión – TDIDT: La familia de los Top Down Induction Trees (TDIDT) pertenece a los métodos inductivos del aprendizaje automático que aprenden a partir de ejemplos preclasificados; en minería de datos, las mismas se utilizan para modelar las clasificaciones en los datos mediante árboles de decisión[19].

Los sistemas de aprendizaje basados en arboles de decisión son quizás e método más fácil de utilizar y de entender. Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas.

[18] HERNÁNDEZ, José et al. Introducción a la minería de datos, 2004, p. 257

[19] SERVENTE, M. y GARCÍA MARTÍNEZ, R. Algoritmos TDIDT Aplicados a la Minería Inteligente, artículo publicado en la Revista del Instituto Tecnológico de Buenos Aires. ISSN 0326-1840, 2002.

Características de los árboles de decisión

Los árboles de decisión representan una estructura de datos que organiza eficazmente los descriptores; dichos árboles son construidos de forma tal que en cada nodo se realiza una prueba sobre el valor de los descriptores.

Se puede analizar un árbol de decisión como una caja negra en función de cuyos parámetros (descriptores) se obtiene un cierto valor del clasificador; también puede analizarse como una disyunción de conjunciones donde cada camino desde la raíz hasta las hojas representa una conjunción y todos los caminos son alternativos, es decir, son disyunciones.

Existen 2 tipos de árboles de decisión:

- Árboles de clasificación: rotulan los registros y los asignan a la clase correspondiente; pueden dar una confianza de que la clasificación sea correcta. En este caso el árbol de clasificación da la probabilidad de clase, es decir, la probabilidad de que ese registro pertenezca a una clase determinada.
- Árboles de regresión: estiman el valor de una variable objetivo que toma valores numéricos. Dadas unas variables de entrada, el árbol estima el resultado de la combinación de estas y asigna un valor de salida para la variable dependiente.

La tarea para la cual los árboles de decisión se adecuan mejor es para la clasificación; la estructura de condición y ramificación de un árbol de decisión es idónea para este problema.

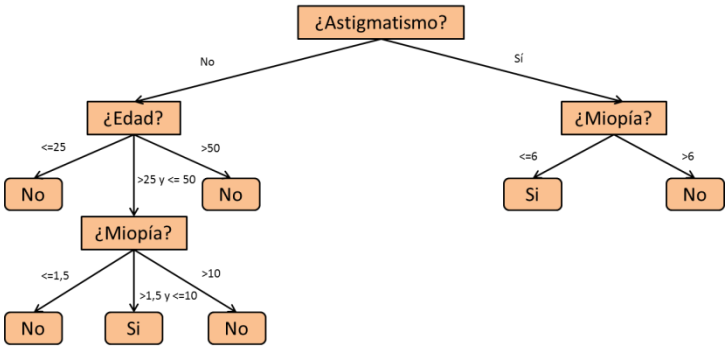
Etapas de construcción[20]

1. El árbol se escribe en forma cronológica describiendo decisiones y procesos en el orden que tienen lugar.

[20] SCHIATTINO I, SILVA C. Árboles de Clasificación y Regresión: Modelos Cart. Cienc Trab. 2008, P. 161 - 166.

2. Se asignan probabilidades a las ramas que parten de un nudo aleatorio
3. Se asignan utilidades a las ramas finales del árbol
4. Se procede desde las ramas finales hacia la base tomando valores esperados en los nudos aleatorios y maximizando en los nudos de decisión determinándose así las mejores condiciones.

Figura 10. Árbol de decisión para determinar recomendación o no de cirugía ocular



Fuente. Introducción a la minería de datos. José Hernández et al, 2004, p.282

El buen funcionamiento de los algoritmos depende de tres aspectos[21]:

- **Particiones posibles**

  Conjunto de condiciones exhaustivas y excluyentes. Cuantos más tipos de condiciones se permitan, mas posibilidades se tendrá de encontrar patrones que hay detrás de los datos. Por ejemplo, un algoritmo que permita incluir la partición ($X_i <= a*X_j^2$, $X_i > a*X_j^2$) donde Xi y Xj son atributos del problema y a es una constante, va a poder encontrar patrones cuadráticos, mientras otro algoritmo que no permita dicha partición no va a poder encontrarla. Cuanta más particiones posean los arboles de decisión serán más expresivos y probablemente más precisos.

[21] HERNÁNDEZ, José et al. Introducción a la minería de datos, 2004, p. 281

Inicialmente todos los datos están juntos (nodo raíz) y el algoritmo parte los datos utilizando cada división binaria posible en cada campo. El algoritmo busca la división que fraccione los datos en dos partes que son más puras que la original. Esta división es aplicada a cada nueva caja y continúa hasta no encontrar más divisiones útiles. Las reglas de partición en un nodo dependen exclusivamente de los atributos, por lo cual, son las mismas tanto para clasificación como para regresión

- **Criterio de selección de particiones**

  Una vez elegida una partición se continúa hacia abajo la construcción de árbol y no vuelven a plantearse las particiones ya construidas. Estos dos aspectos tienen como consecuencia que se busque un criterio que permita realizar una buena elección de la participación que parece más prometedora y que se haga sin demasiado esfuerzo computacional. Esto obliga a calcular la optimalidad de cada partición no sea muy costoso.

  La mayoría de criterios se basan por tanto en obtener medidas derivadas de las frecuencias relativas de las clases en cada uno de los hijos de la partición respecto a las frecuencias relativas de las clases del padre. La primera tarea es definir cuál de los campos independientes realiza la mejor división, la cual está definida como aquella que mejor separa los registros en grupos donde predomina una clase única. La medida utilizada para evaluar un divisor potencial es la reducción en diversidad (“medida de impureza” o “incremento de pureza”). Éste índice corresponde a la probabilidad de que el segundo registro pertenezca a una clase diferente de la primera y está dado por el nivel de entropía

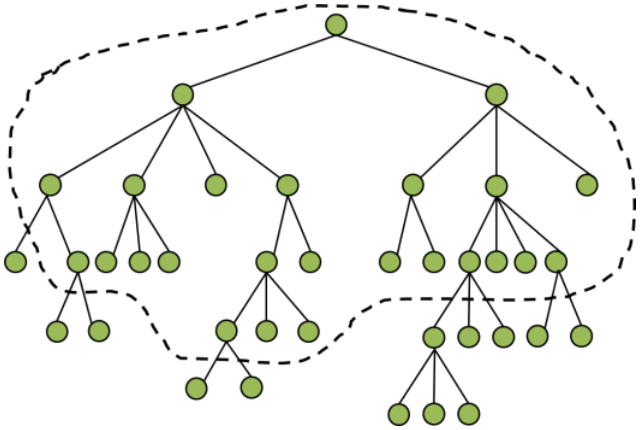
$$E\ A\ = \sum_{i=1}^{v} \frac{P_i + n_i}{P + n}\ I\ (Pi, n_i)$$

Para buscar el mejor divisor en un nodo, el algoritmo del árbol de decisión considera cada campo de entrada a su vez. En esencia cada campo es clasificado, entonces cada posible decisión es ensayada y será mejor aquella que tiene mayor disminución de diversidad. Esto se repite para todos los campos y el ganador se escoge como el divisor del nodo. de datos de entrenamiento

- **Poda y reestructuración**

  Los modelos planteados inicialmente pueden ajustarse demasiado a la evidencia suele tener como consecuencia que el modelo se comporte mal para nuevos ejemplos, ya que en la mayoría de os casos, el modelo es solamente una aproximación del concepto objetivos del aprendizaje. La manera mas frecuente de limitar este problema es modificar los algoritmos de aprendizaje de tal manera que obtengan modelos más generales. En el contexto de los árboles de decisión y conjuntos de reglas, generalizar significa eliminar condiciones de las ramas del árbol o de algunas reglas.

Figura 11. Ejemplo de Poda. Algunos nodos inferiores son eliminados



Fuente. Técnicas de aprendizaje por inducción. Universidad de Buenos Aires. 2005.

Algoritmos o sistemas de aprendizaje de árboles de decisión

- CART [Breiman et al. 1984] y derivados: son métodos “divide y vencerás” que construyen arboles binarios y se basan en el criterio de partición GINI y que srive tanto para clasificación como para regresión.

- ID3 [Quinlan 1983] [Quinlan 1985], C45 [Quinlan 1993] y derivados: son métodos “divide y vencerás” y están basados en criterios de partición derivados de la ganancia (GainRatio)

- IND [Buntine 1992], LMDT [Brodley & Utgoff 1995] y otros sistemas híbridos: incorporan características de varios sistemas o añaden otras técnicas de aprendizaje en la construcción de árboles de decisión: regresión lineal preceptores, entre otros.

3.2.2.3. Análisis De Clúster (Clustering). Los algoritmos de clustering permiten clasificar un conjunto de elementos de muestra en un determinado número de grupos basándose en las semejanzas y diferencias existentes entre los componentes de la muestra[22].

Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Un algoritmo de clustering permite extraer representantes de un conjunto de datos, que pueden ser posteriormente usados para transmisión, para eliminación de ruido o con una fase posterior de calibración, para clasificación de vectores en diferentes conjuntos.

[22] VICENTE VILLARDÓN, Jose Luis. Introducción al análisis de clúster. Universidad de Salamanca, Madrid, 2009. 345 p.

El análisis de clúster es un método que permite descubrir asociaciones y estructuras en los datos que no son evidentes a priori pero que pueden ser útiles una vez que se han encontrado. Los resultados pueden contribuir a la definición formal de un esquema de clasificación, a sugerir modelos estadísticos para describir poblaciones, asignar nuevos individuos a las clases para diagnóstico e identificación, etc.

Podemos encontrar dos tipos fundamentales de métodos de clasificación: **Jerárquicos y No Jerárquicos**. En los primeros, la clasificación resultante tiene un número creciente de clases anidadas mientras que en el segundo las clases no son anidadas. Los métodos pueden dividirse en aglomerativos y divisivos. En los primeros se parte de tantas clases como objetos tengamos que clasificar y en pasos sucesivos van obteniendo clases de objetos similares, mientras que en los segundos se parte de una única clase formada por todos los objetos que se va dividiendo en clases sucesivamente.

Pasos del Análisis de Conglomerados[23]:

- **Formulación del problema.**

Lo más importante de la formulación del problema, es la selección de las variables en las que se basará la agrupación. El conjunto de variables seleccionado debe describir la similitud entre los objetos en términos relevantes para el problema de investigación de mercados. Estas variables se seleccionan en base a investigaciones anteriores, la teoría o una consideración de las hipótesis que se prueban.

[23] JOHNSON, R., WICHERN y DEAN GONDAR, E. “Análisis Cluster, Applied Multivariate Statistical Análisis” Prentice-Hall International, Inc, 1982.

- **Selección de una medida de similitud.**

Como el conglomerado agrupa objetos similares, se necesita una medida para evaluar las diferencias y similitudes entre objetos. La Similaridad (similitud) es una medida de correspondencia o semejanza entre los objetos que van a ser agrupados. Lo más común es medir la equivalencia en términos de la distancia entre los pares de objetos. Así, los objetos con distancias reducidas entre ellos son más parecidos entre sí que aquellos con distancias mayores y se agruparán por lo tanto, dentro del mismo cluster. Los tres métodos usados en la medición de la similitud son: las medidas de correlación y las medidas de distancia (usadas cuando se tienen variables métricas) y las medidas de asociación (usadas para variables categóricas).

La función d se define sobre cierto conjunto M no vacío, como:

$$\boldsymbol{d}: \boldsymbol{M}\boldsymbol{x}\ \boldsymbol{M}\ \rightarrow \boldsymbol{R}$$

el par (x, y) →se asocia con el número d (x, y)
aquí, **R-** es el conjunto de los números reales, la función d satisface las siguientes condiciones para tres elementos cualesquiera x, y, z ∈ $\boldsymbol{M}$ se tiene

1. $d\ x,y\ \geq 0\ y\ d\ x,y\ = 0\ si\ y\ solo\ si\ x = y$
2. $d\ x,y\ = d\ y,x\ \ (simetría)$
3. $d\ x,y\ \leq d\ x,z\ + d\ z,y\ \ (Dseigualdad\ triangular)$

La primera condición exige que la distancia entre dos sujetos debe ser mayor que cero y solo valdrá cero cuando se calcula a partir del mismo sujeto, la segunda condición dice que la distancia entre los sujetos *x* y *y* es la misma que entre los sujetos *y* y *x*; la última indica que siempre debe ser menor o igual ir del sujeto al sujeto *x* que hacer el recorrido a través de *z*.

Los siguientes son ejemplos de funciones distancia, llamadas también métricas; definidas sobre el conjunto $R^n = \ x\ x = \ x_1, x_2, \dots, x_n\ ;\ x_i \in Ri = 1, \dots, n$

1. $d_p\ x, y\ = (\ \sum_{k=1}^{n} x_k - y_k\ ^p)^{1/2}$ $Distancia\ de\ Minkowsky$
2. $d_1\ x, y\ = \ \sum_{k=1}^{n} x_k - y_k\ $ $Distancia\ de\ la\ ciudad$
3. $d_2\ x, y\ = (\ \sum_{k=1}^{n} x_k - y_k\ ^2)^{1/2}$ $Distancia\ Euclídea$
4. $d_p\ x_i, x_j\ = \ \sqrt{x_i - x_j\ V^{-1}\ (x_i - x_j)}$ $Distancia\ Mahalanobis$

Formalmente la similitud se define como una función no negativa y simétrica entre dos observaciones y denotada por que satisface las siguientes propiedades:

1. $S_{ij} = 1$
2. $0\ \leq S_{ij} \leq 1$
3. $S_{ij} = S_{ji}$

La similitud es una medida adecuada para establecer el grado de similaridad entre los individuos, cuyas características han sido tomadas en escala nominal.

Cuando se elige una distancia como medida de asociación los grupos se formarán como aquellos más parecidos, es decir, que la distancia sea la mínima, generalmente es usado en aquellos datos que son medibles.

- Estandarización de dato

Como las medidas de distancia son sensibles a la diferencia de escalas o de magnitudes hechas entre variables es necesaria la estandarización de datos para evitar que las variables con una gran dispersión tengan un mayor efecto en la similaridad.

3.2.2.4. Reglas de Asociación: Las reglas de asociación es una técnica importante en la Minería de Datos y consiste en encontrar las asociaciones interesantes en forma de relaciones de implicación entre los valores de los atributos de los objetos de un conjunto de datos[24].

Minar reglas de asociación en colecciones de datos mezclados, considerando la semejanza entre objetos y partes de ellos al contar las ocurrencias de los mismos, que permitan el uso de funciones de semejanza menos restrictivas que las permitidas por el algoritmo existente.

Las reglas de asociación y dependencia se caracterizan precisamente por el hecho de que se expresan en forma de reglas tipo **"SI… ENTONCES…"** y que los algoritmos tienen como objetivo primordial la eficiencia.

Se considera que una regla es interesante si su soporte y su confianza son mayores o iguales que ciertos umbrales de mínimo soporte y mínima confianza especificados. El interés de una regla de asociación está dado por su soporte y su confianza, entendiéndose por soporte la frecuencia de aparición en la colección de la combinación de productos involucrados en la regla. Por confianza de una regla entendemos cuánto representa el soporte de la regla, del soporte del antecedente de la regla[25]. En las reglas de asociación cualquier atributo puede estar en la parte derecha de la regla.

Reglas de dependencias[26]

Una regla de dependencia se define como cualquier conjunto de variables atributo que sea dependiente. Dos eventos son independientes si $p\ A \cap B\ = p\ A\ \cdot p\ B$.

[24] ALATAS, B., E. AKIN, and A. KARCI, MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules. Applied Soft Computing, 2008. P. 646–656.
[25] KRYSZKIEWICZ, M. Fast Discovery of Representative Association Rules. In First International Conference on Rough Sets and Current Trends in Computing. Poland,1998.
[26] HERNÁNDEZ, José et al. Introducción a la minería de datos, 2004, p. 237

El método propuesto consiste en medir la significancia de las diferentes dependencias mediante el test ji-cuadrado usado en la estadística clásica para las correlaciones. Para un determinado evento, se emplea $E\ i_j\ = O_n(i_j)$ y $E\ \neg\ i_j\ = n - O_n(i_j)$. Para conjuntos de eventos se asume independencia para calcular $E\ r\ = n \cdot \frac{E\ r_1}{n} \cdot \frac{E\ r_2}{n} \cdot \ldots E(r_k\ n)$. El valro ji-cuadrado se define como

$$X^2 = \sum_{r \in R} \frac{(O\ r\ - E\ r\ )^2}{E(r)}$$

Este valor determina si los k ítems son independientes entre sí. Para saberlo, se calcula el valor, y se compara con el definido en la estadística ji-cuadrado para los grados de libertad determinados (1 para variables binarias) y para el grado de significancia requerido. Si el valor calculado es menor que el dado por la estadística, no se rechaza la hipótesis de independencia con el grado de significancia requerido.

- **Interés de dependencias**

El interés entre dos eventos x e y se define de manera similar al cálculo de la dependencia entre dos ítems:

$$I\ xy\ = \frac{p(xy)}{p\ x\ p(y)}$$

Dado un atributo, se considera un evento como cada uno de los posibles valores (o ítems) que pueda tomar. Por lo tanto, para una dependencia entre dos atributos binarios, se puede construir una tabla 2X2 que contenga el interés de cada una de las combinaciones. Un valor de interés superior a 1 significa dependencia positiva, mientras un valor inferior a 1 indica dependencia negativa.

- **Relación entre interés y correlación**

La correlación para medir la dependencia entre dos series de datos numéricos. La correlación es una medida entre 1 y -1 bastante sencilla que permite establecer el grado de similitud entre las dos series de valores.

Reglas de asociación multinivel[27]

En muchos casos es bastante difícil encontrar relaciones interesantes debido a que los datos están muy dispersos, es decir, existe una gran cantidad de atributos con respecto al pequeño número de ítems presentes en casa registro. Una medida permite subsanar este problema consiste en agrupar atributos en categorías. De esta manera el aprendizaje de reglas se realiza utilizando estas categorías y es más fácil encontrar reglas con niveles adecuados de confianza o cobertura.

Las reglas de asociación que utilizan niveles de conceptos para expresar las relaciones se denominada reglas multinivel. Para usar estas reglas se debe tener una jerarquía de conceptos que contiene un árbol de relaciones entre los atributos. Una jerarquía de conceptos, formalmente define una secuencia de relaciones entre conceptos más específicos a conceptos más generales. El nivel de agrupamiento o abstracción debe ser flexible. Un nivel excesivo de abstracción puede conducir a que se aprendan reglas poco informativas. Por el contrario un nivel demasiado concreto puede provocar que no se encuentren reglas con suficiente cobertura o confianza.

Reglas de asociación secuenciales

Este tipo de reglas expresan patrones de comportamiento secuenciales, es decir, que se dan en instantes distintos (pero cercanos) en el tiempo. El algoritmo más usado es el **AprioriAll** dividido en 5 fases: Ordenación, Selección de conjuntos de

[27] ZAKI, M.J., et al., New Algorithms for Fast Discovery of Association Rules. University of Rochester. Estados Unidos, 1997.

ítems, Transformación y renombramiento, Construcción de secuencias frecuentes, Selección de secuencias mixtas

3.3. METODOLOGÍAS PARA PROCESOS DE MINERÍA DE DATOS

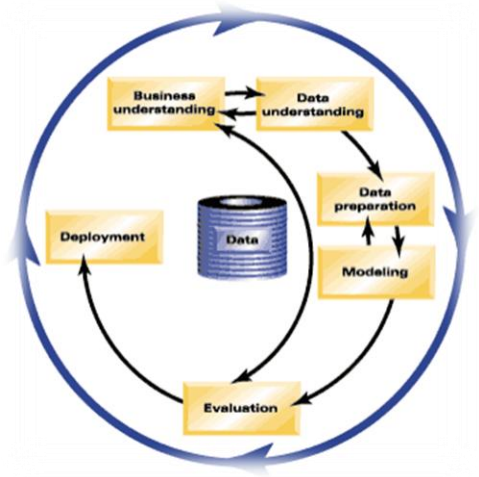
Para la aplicación de proyectos de minería de datos han surgido diferentes metodologías, que permiten ejecutar este proceso de forma óptima y ordenada. Para este caso existen dos metodologías muy usuales: la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*) y la metodología SEMMA (*Sample, Explore, Modify, Model, Assess*). SEMMA se centra en características técnicas del desarrollo del proceso y está enfocada a SAS, en tanto que CRISP-DM mantiene como foco central los objetivos empresariales del proyecto, y para fines de este proyecto será tenida en cuenta.

CRISP-DM[28]

La metodología CRISP-DM consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos. La metodología estructura el ciclo de vida de un proyecto de Data Mining en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto

[28] GALLARDO ARANCIBIA, José Alberto. Metodología para el desarrollo de proyectos de minería de datos CRISP-DM. En: EPB 603 Sistemas del Conocimiento. [en línea]. [Consultado 9 de marzo de 2013]. Disponible en: <http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf >

Figura 12. Ciclo CRISP - DM



Fuente. Manual CRISP-DM de IBM SPSS Modeler

Durante la ejecución de este proyecto de grado se llevaran a cabo las siguientes etapas:

1. **Compresión del negocio o problema:** esta fase tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar Data Mining y definir los criterios de éxito (objetivos y planteamiento del problema.

2. **Comprensión de los datos:** en esta fase se tiene un primer contacto con el problema, familiarizándose con los datos, estableciendo su calidad y planteando las primeras hipótesis. Las tareas a realizar son:
   - Recolección de datos iniciales: Esta tarea se elabora una un informe que contenga los datos existentes (para este caso son los recolectados en la base de datos del OLE)
   - Descripción de los datos: se especifican sus propiedades, volumen (número de registros), significado de cada campo y sus medidas.

- Exploración de datos: a través del uso de los conocimientos de estadística básica, se realiza un informe que contenga tablas de frecuencia, gráficos de distribución que generen una estructura general para los datos.
- Verificación de la calidad de los datos: en esta tarea se desea retirar datos que estén fuera de rango y no contribuyan a información importante.

3. **Preparación de los datos:** los datos son preparados para le técnica de modelado seleccionada, se llevan a cabo tareas como:
   - Selección de datos: se selección un subconjunto de los datos adquiridos en la fase 2, teniendo en cuenta en los criterios que se realizaron previamente
   - Limpieza de datos: se busca elevar la calidad de los datos, algunas de las técnicas a utilizar para este propósito son: normalización de los datos, desratización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc.
   - Estructuración de los datos: se generan atributos partiendo de los atributos existentes.
   - Integración de los datos: La integración de los datos, involucra la creación de nuevas estructuras, a partir de los datos seleccionados.
   - Aplicar formatos a los datos

4. **Modelado:** se selecciona la técnica para este análisis se trabajara con reglas de asociación, clustering, redes bayesianas y árboles de decisión. Se genera un modelo de prueba, se comprueba calidad y se corrigen errores, posteriormente. Después se ejecuta la técnica sobre los datos preparados para generar los modelos necesarios. . Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados

5. **Evaluación:** se evalúa el modelo, verificando el cumplimiento de los criterios de éxito del problema. Se identifican debilidades y factores a mejorar.

6. **Implementación:** con los resultados obtenidos a partir del modelo se toman decisiones con respecto al problema a resolver, es decir que acciones se podrían tomar en cuanto al programa de seguimiento a graduados de la Universidad Industrial de Santander.

   Todo este culmina con la generación de un informe final donde se pretende dejar pautados acerca de lo obtenido y los parámetros a tener en cuenta para ejecutar el programa de seguimiento a graduados de forma óptima y veraz.

4. PROCESO DE DESCUBRIMIENTO EN BASES DE DATOS

En el presente capítulo se presenta el proceso de minería de datos que se usó en la investigación. Se inicia con el reconocimiento, limpieza, integración y recopilación de los datos. En el proceso de depuración se eligieron las variables que se emplearán en el análisis según su importancia; se finaliza con una descripción del uso de las herramientas estadísticas para el tratamiento de los datos a través de ejemplos sencillos acerca de cómo se construye un modelo con las herramientas propuestas.

4.1. INTEGRACIÓN Y RECOPILACIÓN

El proceso de *Minería de Datos* inicia con la identificación de los posibles datos que se encuentren contenidos en diferentes recursos, ya sea en los servidores internos o producto de estudios externos como por ejemplo los demográficos, censos o investigaciones de mercado. Para este proyecto en particular se dispone únicamente de los datos que se encuentran en la base de datos del *OBSERVATORIO LABORAL PARA LA EDUCACIÓN (OLE)*[29], quien se encarga de recopilar la información de los egresados de las diferentes universidades del país con ayuda de las mismas. La información que se emplea es la relacionada con los egresados de la Universidad Industrial de Santander correspondiente al periodo 2010-2012.

Los datos, en este caso son capturados en conjunto por la universidad y el Observatorio Laboral para la Educación (OLE) y almacenados por el Sistema de

[29] Sistema de información que brinda herramientas valiosas para analizar la pertinencia de la educación a partir del seguimiento a los graduados y su empleabilidad en el mercado laboral.

Información del Observatorio Laboral; finalmente los datos son extraídos en diversas plantillas previamente creadas las cuales en son de tipo .xlsx.

El instrumento de medición para el Seguimiento a Graduados consta de una serie de preguntas que deben ser diligenciadas en cuatro momentos de tiempo: al momento del grado, al año de ser graduado, a los tres años y cinco años después de haber recibido el título de pregrado. La información recopilada hasta este punto por parte del OLE y la Universidad Industrial de Santander es la siguiente:

- *Momento 0* 2010-2
  2011-2
  2012-1

- *Momento 1* 2010-2
  2011-2
  2012-1
  2012-2

- *Momento 3* 2010-2
  2012-1
  2012-2

- *Momento 5* 2010-2

Para cada momento se tiene una encuesta que consta de una serie de preguntas que se repiten durante los 4 periodos y otras diferentes, esto se hace con el propósito de hacer seguimiento a egresados y para entender la evolución y el

cambio de opinión y de percepción que van teniendo los egresados conforme pasa el tiempo

Teniendo en cuenta lo anterior se obtuvieron 5 plantillas distintas. Sin embargo, para efectos de obtener un análisis con la mayor confiabilidad posible se decide trabajar únicamente con los datos del año 2010 en los diferentes momentos exceptuando el momento 5. El resto de años son considerados obsoletos dado que no tienen la cantidad de datos suficientes para iniciar un proceso de inferencia estadística.

Además se construyeron nuevas variables con el objetivo de tener una mayor cantidad de información que permitan direccionar el sentido de la investigación hacia características sobre las cuales la universidad pueda intervenir para realizar mejoras en la oferta educativa. (Ver Anexo 2: plantilla de resultados encuesta seguimiento a egresados)

En forma paralela al proceso de integración de las bases de datos se construyó el Diccionario de Datos, (ver depuración de bases de datos en el numeral 5.2.1), el cual contiene la definición para cada variable seleccionada.

4.2. SELECCIÓN, LIMPIEZA Y TRANSFORMACIÓN

En esta fase se analizan los datos y el tratamiento que se les debe realizar. Para este caso se trabaja únicamente con los datos de los egresados obtenidos en el año 2010 en los diferentes momentos.

4.2.1. Depuración de la base de datos. Para cada una de las plantillas de los momentos de las encuestas del Observatorio Laboral se realizó el proceso de depuración como sigue.

La encuesta global consta de los siguientes ítems:

- ❖ Información general
  - Nombre del programa: Nombre que tiene el programa al que pertenece el egresado.
  - Año graduación: año de egreso de la institución.

- ❖ Información personal y familiar
  - Estado civil: Situación personal en que se encuentra el egresado en relación a otra persona
  - Vivienda actual: Condición bajo la cual se encuentra la vivienda en donde reside el egresado (arriendo, propia, entre otras).
  - Educación padre: Nivel de educación más alto alcanzado por el padre del egresado.
  - Posición ocupacional del padre: Ocupación del padre del egresado.
  - Educación madre: Nivel de educación más alto alcanzado por la madre del egresado
  - Posición ocupacional de la madre.
  - Se reconoce como: Ocupación de la madre del egresado.
  - Limitaciones / discapacidades físicas: Limitación física permanente del egresado.
  - Limitación/discapacidad física que más afecta el desempeño diario: Limitación que afecta el desarrollo de algunas actividades laborales.

- ❖ Historia académica y financiación
  - Edad terminó bachillerato: Edad a la cual el egresado termino sus estudios de bachillerato.
  - Programa preferido: Áreas de interés del egresado antes de elegir la carrera.

- Factor más importante al elegir carrera: Razón más importante por la cual decidió estudiar la carrera que eligió.
- Recursos para pagar estudios: Recursos empleados por el egresado para realizar sus estudios.
- Entidades recibió beca o subsidio: Entidades de las cuales recibió alguna beca o subsidio para sustentar sus estudios.
- Entidades recibió crédito: Entidades de las cuales recibió algún crédito para sustentar sus estudios.

❖ Competencias

- Institución bilingüe: Define si la institución en donde terminó el bachillerato el egresado era bilingüe.
- Institución influyó en mejora de competencias en idiomas extranjeros: Influencia de la Universidad en la adquisición de un segundo idioma.
- Idioma ninguno: Establece si el egresado no tiene dominio de una segunda lengua.
- Idioma Inglés: Manejo de Ingles.
- Idioma Francés: Dominio de Francés.
- Idioma Italiano: Dominio de Italiano.
- Idioma Portugués: Dominio de Portugués.
- Competencia: Exponer las ideas por medios escritos.
- Competencia: Comunicarse oralmente con claridad.
- Competencia: Persuadir y convencer.
- Competencia: Símbolos de comunicación.
- Competencia: Aceptar diferencias multiculturales.
- Competencia: Utilizar herramientas informáticas básicas.
- Competencia: Aprender y mantenerse actualizado.
- Competencia: Ser creativo e innovador.
- Competencia: Buscar, analizar, administrar y compartir información.

- Competencia: Crear, investigar y adoptar tecnología.
- Competencia: Diseñar e implementar soluciones con el apoyo de tecnología.
- Competencia: Identificar, plantear y resolver problemas.
- Competencia: Capacidad de abstracción, análisis y síntesis.
- Competencia: Comprender la realidad que lo rodea.
- Competencia: Asumir cultura de convivencia.
- Competencia: Asumir responsabilidades y tomar decisiones.

❖ Plan de vida

- Pensar hacer en el largo plazo: Metas o futuros objetivos de los egresados en su plan de vida y desarrollo profesional.

❖ Situación laboral

- Actividad en que ocupa la mayor parte de su tiempo: Actividad en cuanto a situación laboral a la que se dedica actualmente el egresado.
- Alguna otra actividad remunerada: Actividad alterna a la cual el egresado dedica tiempo y que le genera ingresos económicos.
- Diligencia para conseguir un trabajo: Evalúa si en el último mes el egresado ha hecho algún trámite para obtener un trabajo.
- Desea conseguir un trabajo o instalar negocio: Interés por el egresado en establecer una ocupación fija.

❖ Graduados empleados

- Este es su primer empleo: Reconocer si el empleo actual es el primero de la trayectoria laboral del egresado.
- Canal de búsqueda: Medio que le permitió conseguir el empleo actual.
- Tipo de vinculación con la empresa: Vinculación del egresado con la empresa.

- Ocupación actual: Cargo que ocupa dentro de la empresa.
- Actividad económica: Área dentro de la cual se desempeña el egresado.
- Relación del empleo con la carrera que estudió: Verificar la coherencia entre la carrera y el cargo desempeñado actualmente.
- Ingreso laboral en el mes pasado: Ingreso económico mensual.
- Ámbito de las actividades de la empresa donde laboral: Sector económico al cual pertenece la empresa.
- Vínculo entre la institución donde estudió y la organización donde laboral: Relación existente entre las instituciones.

❖ Trabajo por cuenta propia

- Primer trabajo: Reconocer si el empleo actual desarrollado por cuenta propia es el primero de la trayectoria laboral del egresado.
- Relación actividad – carrera: Vínculo entre la actividad actual que realiza por cuenta propia con la carrera que estudió.
- Forma de trabajo que desempeña por su cuenta: Tipo de ocupación que realiza por cuenta propia.
- Ingreso mensual: Remuneración por el trabajo por cuenta propia.

❖ Crear empresa

- Interés por crear empresa: Interés del egresado por tener una empresa propia.
- Dificultad en la creación de una empresa: Principal dificulta que ha tenido el egresado para crear una empresa.

❖ Propietarios/socios de empresa

- Primer trabajo: Reconocer si ser propietario o socio es la primera actividad en la trayectoria laboral del egresado.

- Relación actividad – carrera: Vínculo entre la actividad actual que realiza por cuenta propia con la carrera que estudió.
- Ingreso mensual: Remuneración por el trabajo por cuenta propia.

❖ Aspectos generales

- Utilidad en el trabajo de conocimientos y destrezas aprendidas en su carrera: Coherencia entre los conocimientos adquiridos y los empleados para desarrollar la actividad laboral.
- Trabajo actual contribuye a su crecimiento personal: Realización del trabajo genera en el egresado un sentido de cumplimiento con sus objetivos del plan de vida.
- Satisfacción con el trabajo actual: Grado de satisfacción que siente el egresado por su labor desempeñada.
- Nivel de estudio requerido para el trabajo actual: Nivel de estudio que considera el egresado que se debe tener para realizar su trabajo.
- Mejor aprovechamiento de las habilidades si realizara otra actividad: Si el egresado considera que debería estar en otro trabajo para desarrollar mejor sus competencias profesionales.
- Apreciación salarial del egresado: Si el egresado considera que debería ganar mejores ingresos.

❖ Buscando empleo

- Busca trabajo: Establece si está buscando trabajo por primera vez o si había trabajado antes.
- Meses buscando trabajo desde que se graduó: Tiempo que ha transcurrido desde que se graduó hasta que consiguió trabajo.
- Principal dificultad para conseguir el trabajo que busca: Establecer cuál es el mayor inconveniente para obtener un trabajo.

- Canal de búsqueda de empleo más efectivo: Medio por el cual el egresado está buscando empleo.

❖ Identidad Institución

- Volver a estudiar en la IES: Volvería el egresado a estudiar en la institución de la que se graduó.
- Principal razón para querer volver a esta institución: Razón principal para volver a elegir a la institución.
- Principal razón para no querer volver a esta institución: Razón principal para volver a elegir a la institución.
- Recomendaría a un bachiller seleccionar el programa que estudió: Intención del egresado en recomendar al IES a otra persona.

❖ Satisfacción con los recursos ofrecidos por la institución

Personal docente: Califica cada una de las características del personal docente según el grado de satisfacción en lo que respecta a:

- Relaciones interpersonales
- Formación académica
- Fundamentación teórica
- Disponibilidad de tiempo
- Procesos de aprendizaje
- Trabajo de campo/pruebas experimentales

Apoyo a estudiantes: Califica características del apoyo que brinda la universidad a los estudiantes según el grado de satisfacción en lo que tiene que ver con:

- Posibilidad intercambios
- Gestión de prácticas empresariales
- Gestión para identificar oportunidades de empleo

- Apoyo para desarrollar investigaciones
- Apoyo a seminarios de actualización
- Asistencia médica/psicológica
- Asistencia espiritual

Gestión administrativa: Califica cada una de las características de la gestión administrativa según el grado de satisfacción en referencia a:

- Agilidad trámites administrativos
- Atención del personal administrativo

Recursos físicos: Califica cada uno de los recursos físicos ofrecidos por la universidad según el grado de satisfacción según:

- Salones de clase
- Laboratorios y talleres
- Espacios para estudiar
- Ayudas audiovisuales
- Aulas de informática

Campos Eliminados

Se eliminan los siguientes campos de la base de datos por no proporcionar mayor información acerca de los egresados debido a que corresponden a información netamente personal, presentan poca variabilidad de las respuestas obtenidas o porque no hubo un nivel de contestación mínimo por parte de los egresados para que la información fuera confiable.

- Fecha de diligenciamiento
- Tipo de documento del graduado
- Número de documento del graduado

- Determina si la encuesta ya ha finalizado
- Tipo de instrumento - Encuesta
- Código IES
- Semestre graduación
- Primer nombre
- Segundo nombre
- Primer apellido
- Segundo apellido
- Código del país residencia
- Nombre del país residencia
- Teléfono fijo
- Teléfono móvil
- E-mail uno
- E-mail dos
- Número de hijos
- Observaciones parte A
- Idiomas - habla
- Idiomas - escucha
- Idiomas - lectura
- Idiomas – escritura
- Observaciones parte C
- Horas a la semana dedicadas a este empleo
- Otra dificultad en la creación de una empresa
- Otra dificultad para conseguir el trabajo que busca
- Observaciones parte E
- Otra razón para querer volver a esta institución
- Otra razón para no querer volver a esta institución
- Observaciones parte F
- Observaciones parte G

- Nombre del acudiente
- Relación del acudiente
- País en que reside el acudiente
- Municipio en que reside el acudiente
- Teléfono fijo del acudiente
- Teléfono celular del acudiente
- Observaciones parte H
- Código del municipio residencia
- Nombre del municipio residencia
- Código del programa
- Tiempo entre graduación y matrícula IES
- Razón de no ingreso a IES después de graduarse
- Idioma Mandarín: Dominio de Mandarín.
- Idioma Alemán: Dominio de Alemán.
- Idioma Japonés: Dominio de Japonés.
- Idioma Árabe: Dominio de Árabe.
- Disponibilidad para trabajar la semana pasada
- Actividad del egresado
- No hizo diligencias durante el último mes para trabajar
- Actividad económica
- Considera que será fácil conseguir el empleo que busca.

4.3. MINERÍA DE DATOS

En este capítulo se busca por medio de las herramientas de Minería de Datos generar un conocimiento nuevo útil para la UNIVERSIDAD INDUSTRIAL DE SANTANDER en lo que tiene que ver con la mejora de la oferta educativa para el desarrollo de competencias y la satisfacción con la formación recibida que le

permita a la universidad tomar decisiones sobre sus egresados y estudiantes. Las técnicas empleadas se definieron teniendo en cuenta las necesidades de la universidad y están dadas por el interés que hay en la clasificación de los egresados, la obtención de relaciones entre las variables, la construcción de grupos con comportamiento homogéneo y la obtención de patrones de comportamiento de los diferentes grupos de datos a fin de satisfacer las necesidades que estos puedan presentar.

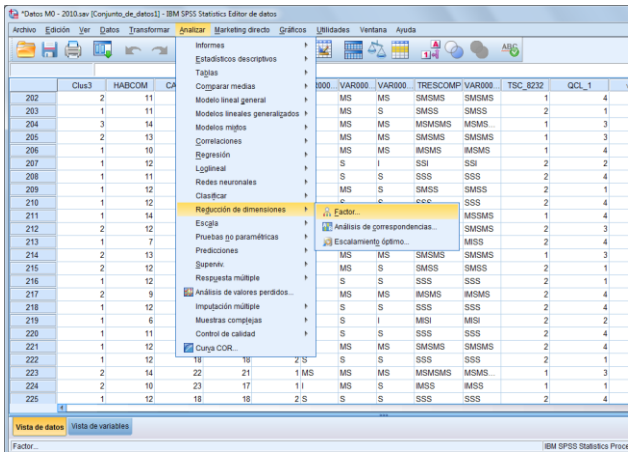
En las siguientes secciones se presentan la forma como deben ser usadas las herramientas discutidas en la investigación haciendo uso de los programas SPSS Statistics y SPSS Modeler, los cuales permiten un ahorro importante de tiempo que trasladan la atención de las tareas mecánicas de cálculo a las tareas conceptuales (interpretación de resultados, análisis crítico), aunque exigen un esfuerzo para su aprendizaje[30].

4.3.1. Análisis de componentes principales (PCA). El método de componentes principales nos permitió reducir cierto conjunto de variables a un conjunto más pequeño y de esta manera intentar describir a los graduados a través de estas nuevas variables. La construcción de este análisis en SPSS Statistics se ejecutó de la siguiente manera.

Inicialmente se busca el archivo que contiene los datos de los egresados, una vez se tengan los datos en el SPSS Statistics se sigue la ruta Analizar - Reducción de dimensiones - Factor.

[30] ÁLVAREZ, GARCÍA, GIL, MARTÍNEZ, ROMERO Y RODRÍGUEZ. Ventajas e inconvenientes del uso de la informática en el análisis de datos, 2002

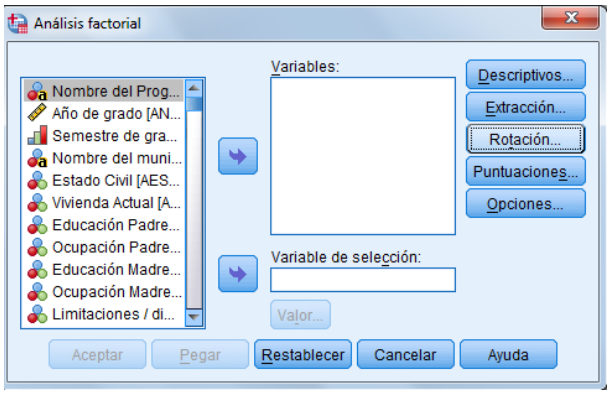
Figura 13. Ruta para realizar un PCA



Fuente. SPSS PASW

Posteriormente, se completan los espacios de variables a incluir y las características del análisis, en la cual se resalta la importancia de configurar la opción de *extracción:* basado en un auto valor y *rotación:* varimáx.

Figura 14. Definir variables del modelo



Fuente. SPSS PASW

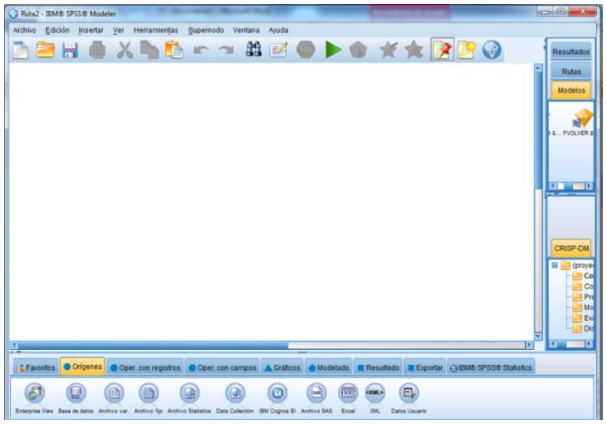
Dados los valores que arroje cada variable para cada componente se definirá la pertenencia a un grupo en específico que permita reducir la cantidad de variables. Para este caso los scrptis empleados y que contienen el conjunto de instrucciones para llevar a cabo el análisis de PCA son:

```
FACTOR
  /VARIABLES COMESC COMORAL PERCONV SIMBCOMU MULTICUL TOOLINF APREUPD
CREATIVO COMPINFO ADOPTECN APOTECNO RSOPBLEM ANALISIS COMPREAL CONVCIA
INFESPC
  /MISSING LISTWISE
  /ANALYSIS COMESC COMORAL PERCONV SIMBCOMU MULTICUL TOOLINF APREUPD
CREATIVO COMPINFO ADOPTECN APOTECNO RSOPBLEM ANALISIS COMPREAL CONVCIA
INFESPC
  /PRINT INITIAL EXTRACTION ROTATION
  /CRITERIA FACTORS(3) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION.
```

4.3.2. Redes Bayesianas en SPSS Modeler. El proceso de construcción del modelo se describe a continuación. El primer paso para la construcción de la ruta (.str) consiste en la definición de los nodos sobre el lienzo. En la pestaña orígenes de la barra inferior de herramientas se define el formato en el que se encuentra el archivo que contiene los datos de estudio, éste puede estar en diversos formatos como .sav (SPSS) o .xlsx (Excel) y otros.

Luego de seleccionar el nodo de acuerdo tipo de archivo que se va a cargar, el nodo seleccionado es arrastrado hasta el lienzo.

Figura 15. Lienzo de SPSS Modeler



Fuente. SPSS Modeler

Para cargar la base de datos, se hace doble clic sobre el nodo del lienzo; para configurarlo se despliega un nuevo cuadro en el que se busca el archivo con la base de datos y se definen los aspectos que allí se requieren. En la pestaña Tipos se define el papel para la variable objetivo o de salida.

Luego de esto, se hace la selección de un nodo Tipo en el cual se define el papel de cada una de las variables para el modelo de red bayesiana a través de la pestaña operaciones con campos contenida en la barra inferior de herramientas. Al igual que con el nodo origen, el nodo Tipo es arrastrado hasta el lienzo y conectado con el nodo origen, y después se personaliza para definir el papel de las variables.

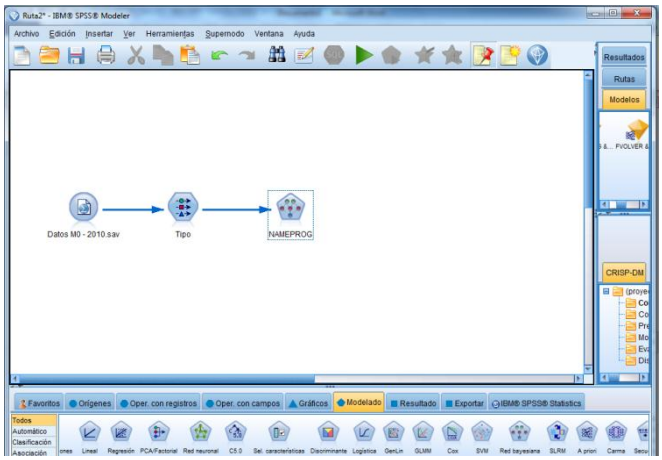
Una vez definido el papel de cada variable se ingresa al lienzo un nodo de Red Bayesiana mediante el la pestaña Modelado de la barra inferior de herramientas.

Figura 16. Paleta de nodos de modelado SPSS Modeler

Fuente. SPSS Modeler

Al igual que los nodos anteriores, se lleva este al lienzo y se conecta al nodo Tipo. Para definir los parámetros de la red se hace doble clic sobre el nodo red bayesiana del lienzo y aparecerá un nuevo cuadro.

Figura 17. Conexión nodos



Fuente. SPSS Modeler

Una vez definidos cada uno de los criterios para cada nodo, se ejecuta la ruta la cual produce un nugget.

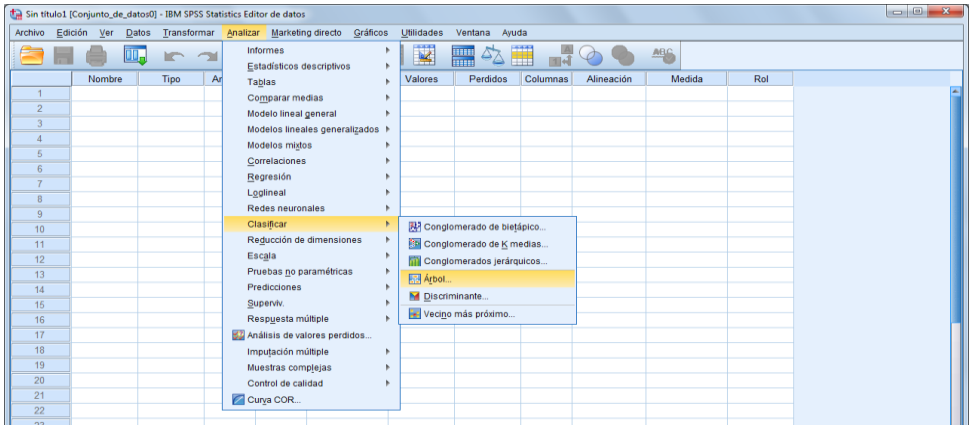
Al dar doble clic sobre el nugget el software despliega un nuevo cuadro en donde muestra la red bayesiana que construyó y las probabilidades para cada variable (nodo de la red bayesiana)

4.3.3. Árboles de decisión en SPSS. La construcción de árboles de decisión mediante el software SPSS Statisctics o SPSS Modeler tiene el proceso que se describe a continuación. En el caso de SPSS Modeler el proceso es el mismo que en las redes bayesianas con diferencia en el nodo Modelado en el cual se selecciona la técnica para la construcción del modelo CART. A continuación se describe el proceso en SPSS Statistics.

Luego de definir las variables y atributos necesarios para el análisis y la posterior depuración de los datos se procede a importar la base de datos desde el formato Excel (.xlsx) al software SPSS Statistics; para hacerlo se debe seguir la ruta Archivo, Abrir, Datos y seleccionar el archivo de datos sobre el cual se construirá

el árbol de decisión. Luego de tener los datos cargados en SPSS Statistics se sigue la ruta Analizar-Clasificar-Árbol.

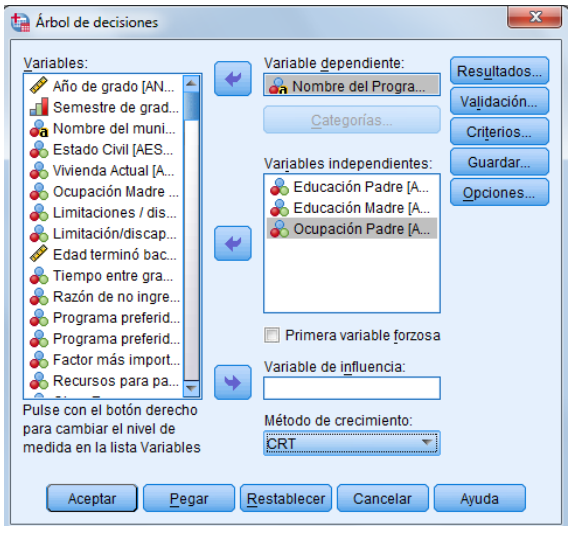
Figura 18. Ruta para crear el árbol de decisiones



Fuente. SPSS PASW

Una vez realizado el paso anterior, se escogen las variables que integraran el árbol de decisión.

Figura 19. Ventana para definir variables y criterios del árbol



Fuente. SPSS PASW

En el campo de variable dependiente se introduce la variable de salida del modelo, es decir, la variable que se define como nodo raíz o nodo padre; en el campo de variables independientes se introducirán aquellas variables que influirán sobre la variable dependiente.El campo variable de influencia se utiliza si se quiere que la creación de nodos esté influenciada por algún criterio. En el Método de crecimiento definimos CRT que hace alusión a los árboles de clasificación y regresión (CART).

Luego de este paso, se abre otra ventana, en ésta se selecciona CRT y se define la medida de impureza. Para definir el árbol que mejor clasifica a los egresados de la Universidad Industrial de Santander bajo algún criterio, definido con anterioridad, se realizan iteraciones buscando el que mejor se adapte al criterio definido.

La selección de las variables que se tomarán en cuenta para el árbol se hace mediante la elección de variables categóricas, variables escala o la combinación de estos tipos de variables. En este paso es importante definir la variable dependiente (variable del nodo raíz) y las variables independientes. También se define el método de crecimiento del árbol que para este caso será CRT.

En esta parte, es muy importante tener en cuenta que si la variable de salida es de tipo escala (continua) el árbol que se construye es de regresión, mientras que si la variable de salida es discreta el árbol será de clasificación. De acuerdo a las variables que se ingresen, el árbol seleccionará aquella variable independiente que mejor ganancia de información produzca, el proceso se realiza para cada variable independiente. Una vez determinada la primera variable que dividirá el árbol, el proceso será iterativo hasta que se encuentre la mejor partición.

Por otra parte, es posible seguir los scripts y así tener de forma inmediata el conjunto instrucciones en archivo de texto que deben ser interpretados línea a

línea en tiempo real para su ejecución, las cuales deben ser convertidos a un archivo binario ejecutable para correrlos.

```
GET
  FILE='C:\Users\Public\Documents\Trabajo de grado\ENCUESTAS\Encuestas en SPSS
Estatistics\Datos M0 - 2010.sav'.
DATASET NAME Conjunto_de_datos1 WINDOW=FRONT.
* Árbol de decisiones.
TREE ESTUACT [n] BY BRECUR1 [n] AEDUPAD [n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES NODEDEFS=YES
SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[1 2 3 4 5 6]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE=CRT MAXSURROGATES=AUTO PRUNE=NONE
  /GROWTHLIMIT MAXDEPTH=AUTO MINPARENTSIZE=100 MINCHILDSIZE=50
  /VALIDATION TYPE=NONE OUTPUT=BOTHSAMPLES
  /CRT IMPURITY=GINI MINIMPROVEMENT=0.0001
  /COSTS EQUAL
  /PRIORS FROMDATA ADJUST=NO
  /MISSING NOMINALMISSING=MISSING.
```

4.3.4. Análisis de clúster. El análisis de clúster puede ser generado por medio de las dos herramientas descritas en las técnicas anteriores y únicamente varían en la parte de modelado. Por ejemplo, la construcción de éste análisis en SPSS Modeler se realiza de forma similar a las redes bayesianas excepto porque varia el nodo de modelado en el cual se selecciona K-Medias; por su parte, en SPSS Statistics la variación se encuentra en la ruta que sigue para encontrar la técnica de modelado que sería Analizar-Clasificar- Conglomerado K-Medias. De forma similar, se seleccionan las variables interés para la construcción del modelo y se definen las propiedades para cada una de ellas. A continuación se muestra el script que contiene las instrucciones para la realización del clúster.

```
TWOSTEP CLUSTER
  /CATEGORICAL VARIABLES=COMORAL PERCONV SIMBCOMU MULTICUL TOOLINF
APREUPD CREATIVO COMPINFO ADOPTECN APOTECNO RSOPBLEM ANALISIS COMPREAL
CONVCIA INFESPC COMESC
  /DISTANCE LIKELIHOOD
  /NUMCLUSTERS FIXED=4
  /HANDLENOISE 0
  /MEMALLOCATE 64
  /CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
  /VIEWMODEL DISPLAY=YES.
```

Luego de ejecutar el modelo se crearán distintos grupos dentro del conjunto de datos inicial los cuales se caracterizan porque los datos al interior de cada grupo (conglomerado) son similares en tanto, entre los distintos grupos los datos se caracterizan por una alta heterogeneidad.

4.3.5. Reglas de Asociación. Las reglas de asociación se generan mediante una de las técnicas descritas en anteriormente y únicamente varían en la parte de modelado. La construcción de éste análisis en SPSS Modeler se realiza de forma similar a las redes bayesianas y al análisis de cluster con la diferencia de que el nodo de modelado es *A priori*. De forma similar, se seleccionan las variables interés para la construcción del modelo y se definen las propiedades para cada una de ellas.

5. RESULTADOS

Los resultados del presente trabajo de grado permitirán a las autoridades académicas tener un modelo de seguimiento a graduados en lo que respecta a la calidad y pertinencia de los profesionales entregados por la UIS a la sociedad Colombiana, además ayuda a la toma de decisiones a corto, mediano y largo plazo. En este capítulo se presentan los análisis que deben llevar a cabo con el objetivo de encontrar patrones, perfiles, modelos de pronósticos de los graduados.

5.1. MINERÍA DE DATOS Y LEVANTAMIENTO DE INFORMACIÓN

El cruce de variables y los cubos OLAP para consolidar información así como diferentes iteraciones en la aplicación de herramientas de análisis de componentes principales (PCA), redes bayesianas, árboles de decisión y análisis de clúster permitieron la construcción de perfiles de los egresados de la Universidad Industrial de Santander.

En primer lugar, es importante destacar que el proceso de reducción de dimensiones debe ser el punto de partida para los análisis siguientes dada su relevancia en encontrar patrones de comportamiento iniciales que puedan generar clasificaciones inmediatas sobre los datos de los egresados en cada uno de los momentos. Inicialmente se llevó a cabo un PCA (por sus siglas en inglés) con el objetivo de perfilar a los graduados a través de un nuevo conjunto de variables denominadas “componentes”. La reducción de dimensiones se realizó con el conjunto de variables que conformaban el concepto de *satisfacción de*

competencias adquiridas, las cuales se relacionan con las percepciones de calidad académica recibida

El análisis se lleva a cabo para las bases de datos: momento 0, momento 1,3 y 5. En la figura se muestra el análisis para los graduados al momento de egreso.

Fue posible agrupar las 16 competencias adquiridas en tres categorías debido a que presentan un comportamiento similar según el valor de cada componente en el momento 0.

Cuadro 1. Matriz de componentes rotados

	Componente		
	1	2	3
Competencia: Exponer las ideas por medios escritos	,181	,083	,698
Competencia: Comunicarse oralmente con claridad	,048	,220	,795
Competencia: Persuadir y convencer	,119	,224	,748
Competencia: Símbolos de comunicación	,237	,153	,465
Competencia: Aceptar diferencias multiculturales	,236	,587	,099
Competencia: Utilizar herramientas informáticas básicas	,673	,185	-,013
Competencia: Aprender y mantenerse actualizado	,633	,277	,189
Competencia: Ser creativo e innovador	,578	,251	,312
Competencia: Buscar, analizar, administrar y compartir información	,608	,309	,278
Competencia: Crear, investigar y adoptar tecnología	,783	,113	,171
Competencia: Diseñar e implementar soluciones con el apoyo de tecnología	,806	,092	,113
Competencia: Identificar, plantear y resolver problemas	,410	,493	,233
Competencia: Capacidad de abstracción, análisis y síntesis	,257	,510	,374
Competencia: Comprender la realidad que lo rodea	,106	,744	,250
Competencia: Asumir cultura de convivencia	,141	,823	,118
Competencia: Asumir responsabilidades y tomar decisiones	,224	,756	,169

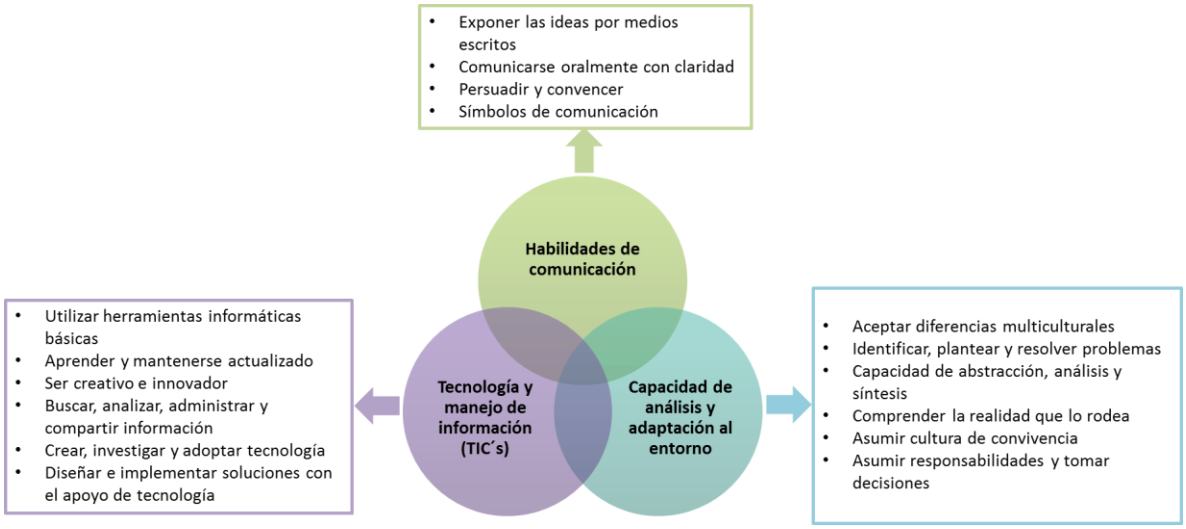
Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Varimax con Kaiser.

Fuente. SPSS PASW

De acuerdo a las competencias que fueron seleccionadas para cada grupo se definieron los nombres.

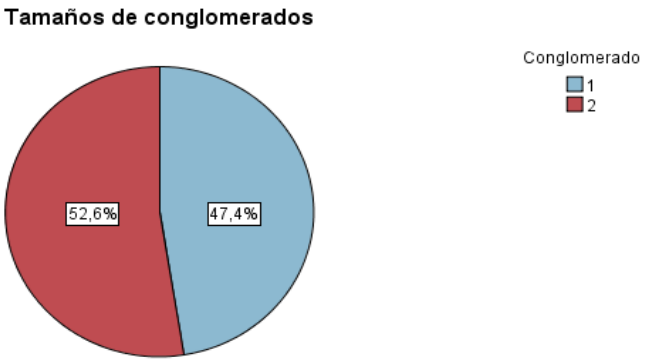
Figura 20. Grupos según competencias adquiridas



Fuente. Autor

Por otra parte, el análisis de clúster nos lleva a agrupar las nuevas competencias en diferentes segmentos con el objetivo de analizar su comportamiento. Se realizaron iteraciones buscando el número de clúster óptimo para los datos de los egresados evaluando exclusivamente aspectos relacionados con las competencias obtenidas en la reducción de dimensiones.

Figura 21. Tamaño de conglomerados



Tamaño de conglomerado más pequeño	217 (47.4%)
Tamaño de conglomerado más grande	241 (52.6%)
Cociente de tamaños: Conglomerado más grande a conglomerado más pequeño	1.11

Fuente. SPSS Modeler

Se construyeron 2 conglomerados denominados por sus características como *Insatisfechos* y *Satisfechos*. Como su nombre lo indica, los insatisfechos se muestran inconformes con 2 de las 3 competencias (TIC's y capacidad de análisis), mientras que los Satisfechos consideran que ha sido acertado el aprendizaje de las competencias.

Sin embargo, para obtener grupos con una mayor definición, se establecieron grupos como tantas combinaciones fuera posible realizar con las calificaciones dadas por los egresados en las nuevas competencias obtenidas, de esta manera se conoce más detalladamente la opinión de los egresados.

En la siguiente tabla se muestran los 8 grupos obtenidos con las diferentes combinaciones y el porcentaje de participación de los egresados para cada grupo.

Cuadro 2. Grupos según posibles combinaciones de competencias

Nombre grupo	Habilidades de comunicación	Capacidad de analisis y adaptacion al antorno	Tecnología y manejo de información (TIC´s)	Recuento	% egresados
Insatisfechos	I - MI	I -MI	I	5	1,1%
Buena capacidad de analisis	I - MI	MS - S	I - MI	10	2,2%
Buena capacidad de analisis y de manejo de TIC's	I - MI	MS - S	MS - S	55	12,0%
Satisfechos	MS - S	MS - S	MS - S	365	79,7%
Buenas habilidades de comunicación y de capacidad de analisis	MS - S	MS - S	I - MI	15	3,3%
Buenas habilidades de comunicación	S	I	I - MI	2	0,4%
Buenas habilidades de comunicación y de manejo de TIC's	S	I	MS - S	3	0,7%
Buen manejo de TIC's	I	I	S	3	0,7%
			TOTAL	458	100%

Fuente. Autor

MI= Muy Insatisfecho

I= Insatisfecho

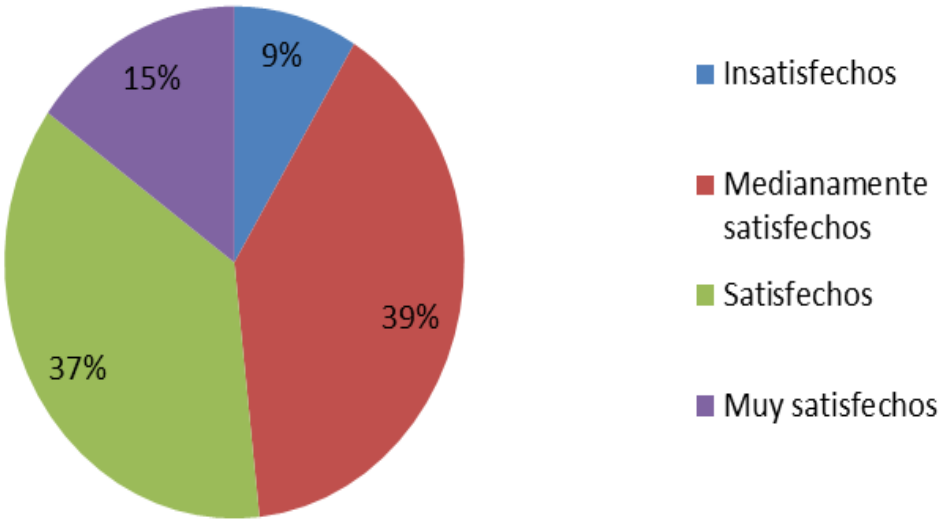
S= Satisfecho

MS= Muy Satisfecho

De acuerdo a los grupos obtenidos se pueden hallar las debilidades y fortalezas de cada uno de ellos, como se muestra en la tabla del anexo 5. Los resultados iniciales muestran que el 79% de los egresados está satisfecho con las competencias adquiridas, el segundo grupo más grande es aquel que tiene deficiencias en las habilidades de comunicación y el tercer grupo es el que tiene deficiencias en el manejo de las TIC's. Una vez realizado el análisis resulta conveniente perfilar los grupos obtenidos (Anexo 5) para resolver directamente las fallas que están generando deficiencias en el grado de satisfacción de los egresados.

Con fin de validar los resultados obtenidos se procedió a llevar a cabo el análisis de clúster usando como variables de segmentación las variables originales. En esta ocasión se encontraron cuatro grupos que pueden definirse como:

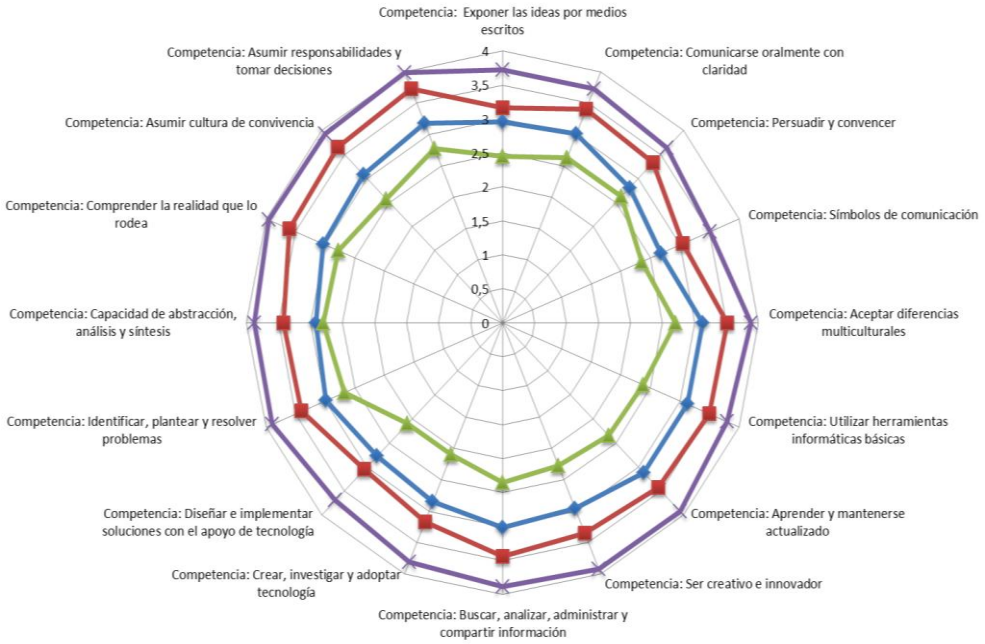
Figura 22. Grupos de competencias originales



Fuente. SPSS Modeler

En la siguiente gráfica, se observa el comportamiento de los diferentes grupos respecto a las variables analizadas, y es posible apreciar la diferencia notoria que hay entre cada grupo respecto al grado de satisfacción hacia las competencias adquiridas, pero lo que más se identifica es que cada grupo tiene una calificación promedio para todas las competencias, de ahí que los nombres de los grupos no hacen referencia a unas habilidades más o menos fuertes sino a un concepto estándar de satisfacción.

Figura 23. Comportamiento de grupos



Fuente. Autor

En el cuadro 2 se muestra el análisis comparativo de los segmentos encontrados para el *momento 0* a partir de variables de estudio que permiten establecer algunas características y comportamientos similares y en otros reflejar las diferencias para facilitar la generación de estrategias diferenciadas de servicio.

Cuadro 3. Segmentos momento 0

	Medianamente satisfechos	Satisfechos	Insatisfechos	Muy satisfechos
Nombre del programa	Ingniería Industrial, Ingeniería civil, Economía	Geología, Licenciatrua	Diseño Indsutrial, Ingeniería de Sistemas	Derecho, Trabajo social
Educacion del padre	Primaria incompleta, secundaria incompleta	Secundaria incompleta y completa, Educación universitaria completa	Educación universitaria completa, Educación de posgrado	Secundaria incompleta y completa, Educación universitaria completa
Educación de la madre	Secundaria incompleta y completa	Secundaria incompleta y completa	Secundaria incompleta y completa	Secundaria incompleta y completa, Educación de posgrado
Ocupación del padre	Trabajador independientes y de empresa particular, trabajador familiar	Trabajador independiente y del gobierno, empleador	Trabajador independientes y de empresa particular	Trabajador independientes y de empresa particular
Ocupación de la madre	Trabajador independiente, oficios del hogar	Trabajador independiente, oficios del hogar	Trabajador independientes y de empresa particular	Trabajador independiente, oficios del hogar
Factor para elegir carrera	Habilidades y destrezas, Familia	Habilidades y destrezas, Familia	Habilidades y vocación	Habilidades y vocación
Recursos para estudiar	Padrés, credito educativo	Padrés, credito educativo	Padres, recursos propios	Padres, recursos propios
Plan de vida	Estudiar un posgrado en Colombia, trabajar en Colombia	Estudiar un posgrado en Colombia, trabajar en Colombia	Estudiar un posgrado fuera de colombia, iniciar una nueva carrera	Estudiar un posgrado fuera de colombia, iniciar una nueva carrera
Actividad actual	Trabajando, buscando trabajo	Trabajando, buscando trabajo	Trabajando	Trabajando
Tipo de ocupación actual	Empleado empresa particular, gobierno	Empleado empresa particular, gobierno	Empleado empresa particular, independiente	Empleado empresa particular, independiente
Relacion de la ocupación con la carrera	Indirectamente y directamente relacionado	Nada relacionado	Directamente relacionado	Directamente relacionado
Ámbito en que se desarrolla	Local, Nacional	Local, Nacional	Nacional	Local
Interés por crear empresa	NO	NO	SI	SI
Utilidad de los conocimientos adquiridos	Muy útiles y poco útiles	Muy útiles y poco útiles	Útiles	Muy Útiles
Satisfacicon para el trabajo actual	Muy satisfecho	Muy satisfecho	Insatisfecho	Satsifecho
Nivel de estudio requerido para el trabajo actual	Universitario, Bachiller	Universitario, Bachiller	Universitario, Técnico	Universitario, Tecnológico
Considera debería estar en otro trabajo	NO	NO	SI	SI
Volvería a estudiar en la institución	SI	SI	NO	SI
Principal razon para no volver a la IES	Docentes, recursos de la universidad	Docentes	Docentes, recursos de la universidad	Recursos de la institución
Insatisfaccion Evaluación de los recursos ofrecidos por la IES	Apoyo a estudiantes	Apoyo a estudiantes	Docentes, apoyo a estudiantes	Apoyo a estudiantes, gestión admin

Fuente. Autores

En cuanto a los momentos 1 y 3 se buscó realizar igualmente la reducción de dimensiones, sin embargo no fue posible obtener grupos que se adaptaran a una descripción clara y de fácil identificación.

Respecto al análisis de clúster para los momentos 1 y 3 resulto ser efectivo, en este caso se realizó una tipología con 3 grupos para el momento 1 con el 30, 34 y 36 % de la población, los cuales presentan similitud con los grupos del momento 0, aunque con la variante de que desaparece el grupo denominado Insatisfechos a pesar de que persiste la sensación de insatisfacción en la formulación y ejecución de proyectos y en crear, investigar y adoptar tecnología. *Ver Anexo 4*.

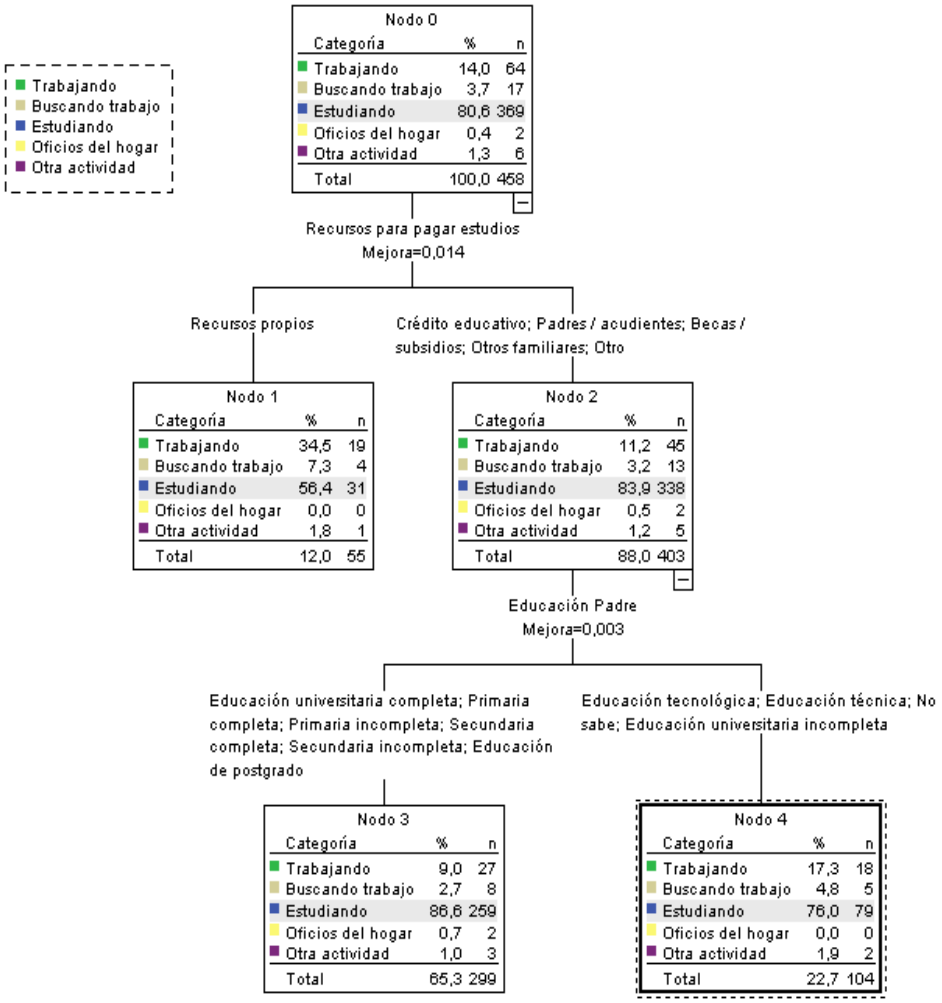
En lo que se refiere al momento 3, se obtuvieron solo 2 conglomerados cada uno con el 50% de la población, en los cuales se aprecia de nuevo la aparición de un grupo de egresados insatisfechos, e incluso también ocurre la perdida de egresados muy satisfechos, ocasionando que el promedio de los egresados demuestre una satisfacción regular por la universidad al cabo de los 3 años de haber egresado. *Ver Anexo 4.*

En cuanto a la creación de perfiles para estos momentos se pueden consultar en el *Anexo 5*. en donde se podrán apreciar algunas diferencias entre los perfiles de los grupos en cada momento.

Al emplear otras herramientas de minería de datos, como en el caso de árboles de decisión, se construyeron árboles sencillos que buscaban comprobar algunas dependencias/independencias entre variables específicas hasta árboles más elaborados. Sin embargo para los momentos 1 y 3 no fue posible la obtención de resultados favorables debido a que la cantidad de datos era muy baja para generar árboles de clasificación con probabilidades de ocurrencia confiables de los diferentes ítems de la encuesta, sin contar con que en algunos casos la variabilidad de las respuestas de los egresados era mínima.

Para el caso específico del momento 0 algunas iteraciones se presentan a continuación como parte del proceso de construcción del perfil de los egresados.

Figura 24. CART



Fuente. SPSS Modeler

El árbol clasificó a los egresados mediante la variable: *actividad actual del egresado* (*nodo origen*) empleando como variables predictoras: *recursos para pagar los estudios y educación del padre*.

La información de este árbol se puede obtener siguiendo cada una de las ramas desde el nodo origen (*actividad actual*) hasta cualquier rama del árbol; por ejemplo, la clasificación muestra que el 89,6% de los egresados que se encuentran estudiando tienen padres con estudios de primaria, secundaria, universitaria y de posgrado y financiaron sus estudios a través de crédito educativo, la ayuda del padre o acudiente, becas subsidios u otros familiares. Esta información es coherente con la clasificación por clúster encontrada para el momento 0, en donde 2 de los grupos se conforman por egresados que utilizaron recursos propios para pagar sus estudios y se encuentran trabajando, mientras que en los otros 2 grupos emplearon otros tipos de pago y los egresados se encuentran estudiando en su mayoría.

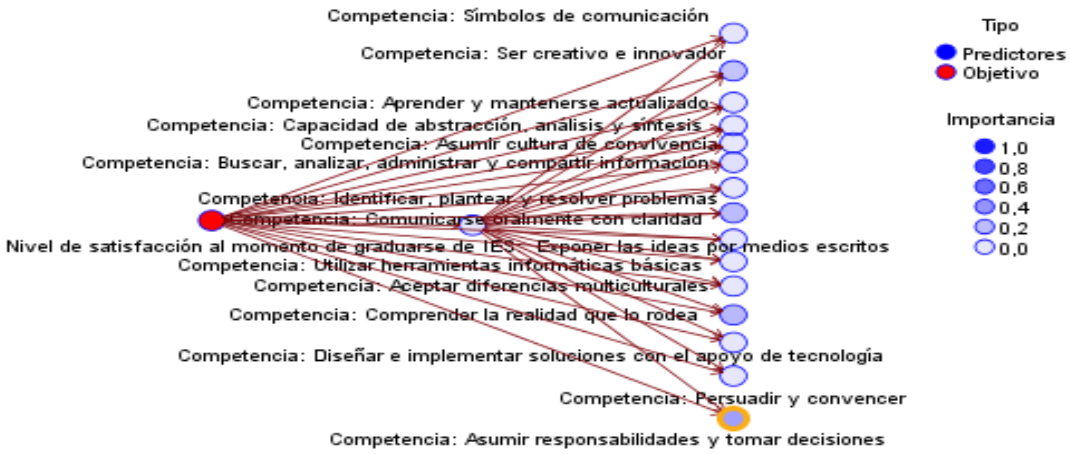
Lo anterior demuestra que en la medida en que un egresado emplea sus recursos para pagar los estudios tiene mayor interés y por tanto probabilidad en estar trabajando.

A partir de otra iteración realizada se evidenció que el 71% de los egresados que tiene interés en crear empresa forman parte de las carrera de Ingeniería Industrial, Economía, Ingeniería Civil y Trabajo social, de las cuales 58,6% propone que la principal dificultad para crear empresa proviene dudas en los negocios y falta de recursos económicos. Este árbol se encuentra en el *Anexo 6* junto con otros análisis de este tipo.

En cuanto a las redes bayesianas se construyen diferentes redes sencillas a fin de establecer algunas relaciones entre las diferentes variables para encontrar redes complejas con nueva información. Algunas de las redes bayesianas se presentan a continuación. Para cada una de las redes construidas se identifica o define la variable de salida o de respuesta así como cada una de las variables predictoras.

En el caso de la figura 24, el modelo se construyó a partir de la estimación del nivel de satisfacción con el trabajo y soportado en las variables predictoras: 16 competencias adquiridas.

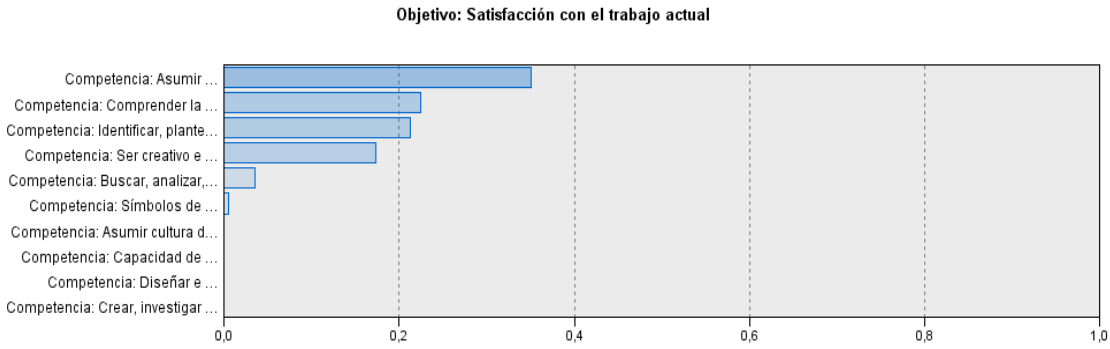
Figura 25. Red Bayesiana – Nivel de satisfacción



Fuente. SPSS Modeler

En la gráfica que se presenta a continuación de las competencias adquiridas, se observa que la más influyente para sentir satisfacción con el trabajo actual son: asumir responsabilidades y tomar decisiones, comprender a realidad que lo rodea e identificar, plantear y resolver problemas.

Figura 26. Importancia del predictor



Fuente. SPSS Modeler

Otra de las redes bayesianas obtenidas, estableció como objetivo central la variable de si le gustaría cursar otros estudios o no en esta institución, para lo cual se emplearon 7 predictores referentes a el apoyo a estudiantes.

Figura 27. Red Bayesiana – Le gustaría cursar otros estudios en la IES

Fuente. SPSS Modeler

Esta red permitió observar que el factor clave para que los egresados decidan regresar a la universidad es la asistencia espiritual recibida; una variable que resulta interesante debido a que no se esperaría que fuera la de mayor importancia a la hora de predecir el comportamiento de la variable objetivo. De acuerdo a otros análisis similares con redes bayesianas se establecieron otras conclusiones (Ver gráficos en Anexo 7):

- Para los egresados del momento 0, los recursos físicos más importantes al momento de calificar la satisfacción con la universidad son los laboratorios y talleres.

- La disponibilidad de tiempo por parte del personal docente es un factor determinante para que el egresado decida regresar a la universidad.

- Las competencias más relevantes para que los egresados decidan regresar a la universidad son:
  - ✓ Crear, investigar y adoptar tecnología
  - ✓ Comprender la realidad que lo rodea
  - ✓ Ser creativo e innovador
- Para el momento 1 el análisis según las competencias, define que las variables que más afectan la decisión de si le gustaría volver a la institución son:
  - ✓ Ser creativo e innovador
  - ✓ Utilizar herramientas informáticas básicas
  - ✓ Aceptar las diferencias multiculturales.

- Siguiendo con el momento 1, la satisfacción con el trabajo actual se encuentra determinada según los egresados crean que sus competencias están al nivel o no de lo que requiere su empleo actual y del ámbito geográfico en el cual se desarrollen, se alcanza a observar que el ingreso económico resulta importante más no el determinante.

- En cuanto al momento 3, por las características de las variables, un análisis interesante fue el que involucró como variable objetivo el ingreso laboral, el cual se ve afectado en gran parte por la satisfacción en el trabajo actual en un 78%.

- En cuanto al análisis de competencias el resultado arroja que los egresados que salieron hace 3 años determinan si vuelven a la universidad teniendo en cuenta factores como:
  - ✓ Trabajar de manera independiente sin supervisión permanente.
  - ✓ Asumir responsabilidades y tomar decisiones.
  - ✓ Ser creativo e innovador.
  - ✓ Trabajar bajo presión

- Se observa que la competencia que resulta siendo igual de importante en todos los momentos para los egresados es el ser creativo e innovador

Finalmente, otra de las técnicas de minería de datos empleada para este estudio fueron las reglas de Asociación, las cuales resultan interesantes para descubrir relaciones entre variables en grandes conjuntos de datos. Para seleccionar reglas interesantes del conjunto de todas las reglas posibles que se pueden derivar de un conjunto de datos se pueden utilizar restricciones sobre diversas medidas de "significancia" e "interés". Las restricciones más conocidas son los umbrales mínimos de "soporte" y "confianza".

El 'soporte' de un conjunto de ítems X en una base de datos D, se define como la proporción de datos en la base de datos que contiene dicho conjunto de ítems:

A partir de lo anterior se encuentran las siguientes reglas de asociación:

- Uno de los criterios que tiene en egresado para tomar la decisión de volver o no a la Universidad Industrial de Santander está dado por su nivel de satisfacción con el personal docente.

Para la regla:

$$conf \quad Insatisfacción\ con\ el\ personal\ docente \ \rightarrow \ no\ volvería\ a\ la\ institución$$
$$= \frac{sop \quad Insatisfacción\ con\ el\ personal\ docente \ \cup \ no\ volvería\ a\ la\ institución}{sop(\ Insatisfacción\ con\ el\ personal\ docente}$$
$$conf \quad Insatisfacción\ con\ el\ personal\ docente \ \rightarrow \ no\ volvería\ a\ la\ institución$$
$$= \frac{0{,}0677}{0{,}1310} = 51{,}67\%$$

A partir de la regla de asociación se encontró que el 51,67% de las reglas de la base de datos de los egresados de la Universidad Industrial de Santander que

contienen “insatisfacción con el personal docente” en el antecedente, también tienen “no volvería a la institución de educación superior” en el consecuente, en otras palabras que:

$$Insatisfacción\ con\ el\ personal\ docente \rightarrow no\ volvería\ a\ la\ institución$$

Es cierta en el 51,67% de los casos.

- En cuanto a la insatisfacción con el apoyo a estudiantes, se resalta que el 55,56% de los egresados que se encuentran muy insatisfechos con el apoyo a estudiantes no volvería a cursar otro estudios en la UIS, se puede notar que es un porcentaje alto y que se requiere por parte de la institución la evaluación de todas las actividades referentes para apoyar al estudiante, tales como gestión de prácticas, movilidad estudiantes y otras.

- La Insatisfacción con la gestión administrativa arroja que en el 42,86% de los casos, el egresado no volvería a la Universidad si se encuentra muy insatisfecho con este aspecto. La opinión del egresado frente a los trámites y atención del personal es la que genera este gran porcentaje.

- Los recursos físicos ofrecidos por la institución no son tan trascendentales para la decisión de no volver a la universidad, a partir de esto se observa que el 33,33% de los egresados que están muy insatisfechos con los recursos físicos no volvería a cursar otro estudio en la UIS.

- Existen habilidades que no están contenidas en una cátedra del programa académico de una carrera de manera formal, pero que se adquieren en el ambiente educativo de la institución. La satisfacción del egresado con estas habilidades arrojan los siguientes resultados:

- ✓ El 60% de los egresados que están muy insatisfechos con las habilidades de comunicación no volverían a la UIS
- ✓ El 50% de los casos donde el egresado siente mucha insatisfacción con la habilidades de capacidad de análisis y adaptación al entorno no regresarían a la institución
- ✓ El 45,45% donde el egresado está muy satisfecho con las habilidades en tecnología y manejo de TIC´S

Los porcentajes presentados anteriormente representan una cantidad representativa de egresados que presentan el mismo comportamiento: no volverán a la institución si sienten total insatisfacción en la adquisición de estas habilidades.

En este punto la universidad debe evaluar el ambiente que se fomenta en las aulas de clase, hay que tener en cuenta que estas habilidades se adquieren en el proceso educativo y que son importantes para el profesional en el momento de enfrentarse al mundo laboral.

Además de las técnicas estadísticas discutidas se sumó el uso de estadística básica como el cruce de variables y los cubos OLAP que permitió conocer el comportamiento de los egresados a nivel general en los distintos momentos. Algunas estadísticas para la construcción de los segmentos y observación de comportamientos se presentan a continuación.

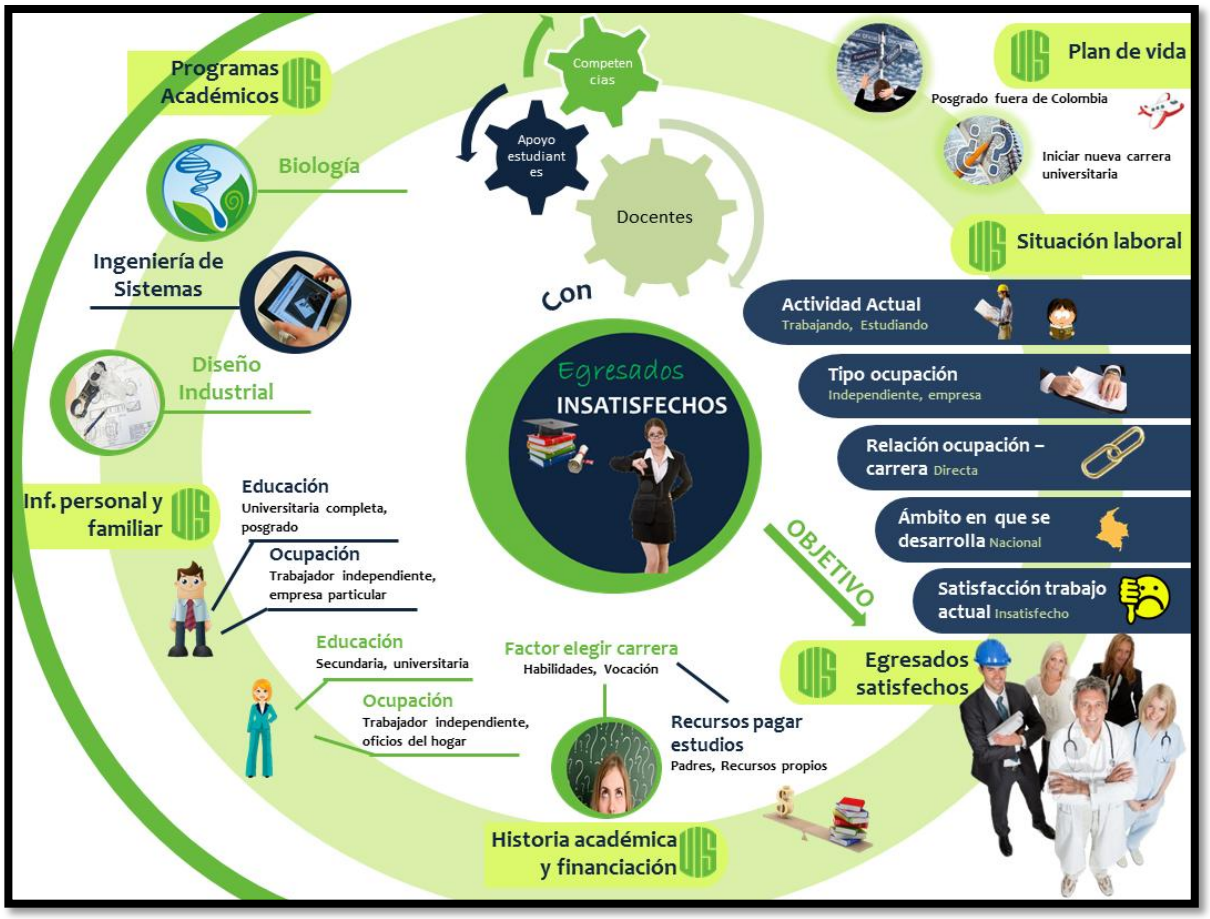
Cuadro 4. Características relevantes de cada momento.

Factores	Momento 0	Momento 1	Momento 3
Mayor participación encuestas	Licenciatura, Geología	Geología, Ing. Sistemas	Geología
Forma de vivienda	Arriendo	Arriendo	Arriendo
Competencia mas relevante		Crear, investigar y adoptar tecnología	Capacidad de abstraccion y análisis
Plan de vida	Estudiar un posgrado en Colombia	Estudiar un posgrado fuera de Colombia	Estudiar un posgrado fuera de Colombia
Actividad que realiza el egresado	Estudiando	Trabajando	Trabajando
Cargo actividad laboral	Empleado empresa particular	Empleado empresa particular	Empleado empresa particular
Tiempo al cabo del cual obtuvo el primer empleo		Ya venia trabajando	
Tipo de vinculación con la empresa	Contrato a término fijo	Contrato a término fijo	Contrato a término fijo
Carrera con mayores ingresos	-	Geología, Ing. Industrial	Geología
Mayores ingresos	-	$3.537.000 - $4.126.500	Ma de 7 smlv
Carrera con mayor interes por crear empresa	Ing. Industrial, Licenciatura, Geología	Ing. Industrial, Licenciatura, Ing Química	Ing. Industrial, Ing. Química
Tiempo de experiencia laboral de la mayoria de los egresados	0 y 6 meses	Entre 0 meses y 1 año	2 y 3 años
Nivel de estudio requerido trabajo actual	Tecnológico, Universitario	Universitario	Universitario, Mestría
Sentido de pertenencia de los egresados	Alto	Alto	Alto
Carrera con mayor sentido de pertenencia	Ing. Indutrial, Ecnomía	Biología, Economía, Ing. Química, Licenciatura	
Carrera con menor sentido de pertenencia	Ing. Sistemas, Química	Química, Ing. Sistemas	
Volveria nuevamente a estudiar en la institución	si	Si	Si
Carrera en donde hay menos interés por volver a la IES	Ing. Sistemas, Biología, Filosofía	Biología, Economía, Ing. Sistemas	
Principa razon para voler a la IES	Calidad de la formación recibida	Calidad de la formación recibida	Calidad de la formación recibida
Principal razón para no volver a la IES	Los docentes no cuenta con la preparación adecuada	Los docentes no cuenta con la preparación adecuada	Los recurso ofrecidos no son suficientes
Carrera que no volvería a la IES	-	Biología, Filosofía, Ing. Petróleros	-
Otros estudios para realizar en la IES	Especialización	Especialización	Maestría
Estudios realizados despues de graduarse		Diplomados	Maestría
Carrera con estudios depues de graduarse			Ing. Undustrial, Economía - Diplomados; Ing. Civil, Geología - Maestría

Fuente. Autores

Se resalta que la educación de las madres en el momento 0 en su mayoría llega a secundaria completa y de los padres secundaria incompleta. Sin embargo resulta interesante que son las madres las que tienen mayores estudios de posgrado y los padres mayores estudios universitarios. La mejor representación que se puede realizar sobre los segmentos es por medio de infografía que permita visualizar de forma global las características de un grupo en particular.

Figura 28. Infografía grupo Insatisfacción



Fuente. Autor

Lo realizado en esta infografía es la forma ideal de visualización para observar todos los segmentos encontrados con sus respectivas caracterizaciones, las cuales son visibles en el *Anexo 5*.

5.2. OFERTA EDUCATIVA

Como resultado de los análisis que se realizaron se identifican algunos aspectos de mejoramiento para algunos de los criterios de evaluación, los cuales fueron enfocados a las competencias adquiridas y a los recursos ofrecidos por la

universidad dado que son los factores que la universidad como ente educativo puede mejorar desde su misión.

Los elementos más relevantes relacionados con la satisfacción de los egresados hacia la universidad son:

- Desarrollo de las competencias
- Apoyo a los estudiantes
- Personal docente
- Gestión administrativa
- Recursos Físicos

Figura 29. Causas de insatisfacción de los egresados

Fuente. Autor

Actualmente, existe un 51,75% de insatisfacción de los egresados en por lo menos uno de los 5 factores nombrados anteriormente, pero el elemento sobre el cual pesa el mayor grado de insatisfacción es el *apoyo a estudiantes* con el 38,29% del total de egresados insatisfechos.

Teniendo en cuenta el análisis previo sobre los aspectos que afectan la satisfacción de los egresados se da una serie de recomendaciones para cada uno de ellos. Los criterios que permitieron evaluar los factores anteriores se encuentran en la encuesta de egresados del Observatorio Laboral. (*Anexo 1*).

5.2.1. Competencias adquiridas. Las competencias evaluadas en la encuesta son:

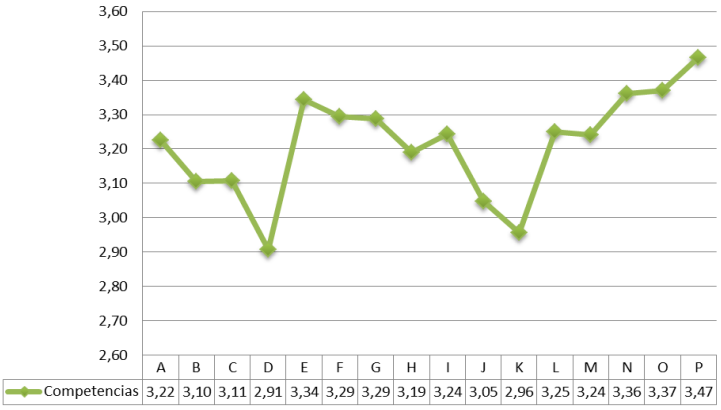
Cuadro 5: Criterios evaluados a las competencias adquiridas

A. Comunicarse oralmente con claridad
B. Exponer las ideas por medios escritos
C. Persuadir y convencer
D. Símbolos de comunicación
E. Aceptar diferencias multiculturales
F. Utilizar herramientas informáticas básicas
G. Aprender y mantenerse actualizado
H. Ser creativo e innovador
I. Buscar, analizar, administrar y compartir información
J. Crear, investigar y adoptar tecnología
K. Diseñar e implementar soluciones con el apoyo de tecnología
L. Identificar, plantear y resolver problemas
M. Capacidad de abstracción, análisis y síntesis
N. Comprender la realidad que lo rodea
O.Asumir cultura de convivencia
P. Asumir responsabilidades y tomar decisiones

Fuente. Observatorio Laboral para la Educación (OLE)

Los resultados de estos criterios se presentan en la siguiente figura:

Figura 30. Evaluación promedio de los criterios asociados a competencias



Fuente. Autor

Con respecto a lo hallado se plantean unas soluciones de mejora entorno a las competencias evaluadas.

- Se deben desarrollar de forma más efectiva las competencias relacionadas con la adopción de tecnología y el diseño de soluciones con el apoyo de la misma, de tal forma que a los estudiantes se les debe impulsar en el desarrollo de las clases el uso de software y de herramientas informáticas para agilizar el desarrollo de sus actividades. También se debe incentivar conferencias, cursos, talleres de desarrollo tecnológico y de software libre a fin de que el estudiante sienta compromiso por parte de la universidad en la introducción hacia el mundo tecnológico.

- Es de vital importancia continuar desarrollando las habilidades comunicativas, sobre todo en cuanto al reforzamiento de la escritura y la lectura como esencia para persuadir y comunicarse con los demás. Para esta situación en particular todos los docentes deberían asumir un compromiso con la exigencia en la calidad de los trabajos escritos y en las presentaciones orales realizadas por los estudiantes.

El análisis se repite para cada uno de los aspectos que siguen y en adelante se detallan las acciones para garantizar un nivel de satisfacción idóneo por parte de los egresados.

5.2.2. Personal Docente

- Los egresados consideran que el personal docente debería realizar más trabajo de carácter experimental o de campo para fortalecer el conocimiento adquirido de forma teórica. De modo que se considera necesario que la universidad empiece una mejor gestión para efectuar el contacto de los estudiantes con el mundo laboral, ya sea con convenios con empresas, y salidas o visitas empresariales.

- Por otra parte existe inconformidad con los estudiantes en cuanto a la disponibilidad de los docentes para atender inquietudes, en tanto que se propone establecer unos horarios fijos de consulta para los estudiantes tanto para los profesores planta y cátedra, horarios que deberán cumplirse estrictamente en la medida de lo posible para garantizar ese espacio de interacción con los estudiantes.

5.2.3. Apoyo a estudiantes. Este factor, al ser el que demuestra más niveles de insatisfacción por parte de los egresados, es el que debería ajustar en mayor medida, sin embargo las preocupaciones más grandes de los egresados son:

- La universidad no brinda oportunidades para gestionar empleos, ante lo cual sería importante exaltar y trabajar más fuertemente en el portal de trabajo que tiene la universidad, el cual a pesar de su funcionamiento no tiene el suficiente alcance para propiciar la participación de todos los recién egresados.

- También resulta importante fortalecer el programa de movilidad de la universidad, a fin de apoyar a los estudiantes en su interés por realizar intercambios.

- En cuanto a las prácticas empresariales se conserva aún la preocupación por parte de los egresados en tener contacto con el mundo laboral debido a que la universidad no ofrece es tipo de espacio. Por tanto realizar o comenzar a genera convenios con empresas santandereanas para la realización de prácticas o en su defecto trabajos de grado.

- Finalmente un factor decisivo que se encontró al analizar estos factores es el grado de satisfacción en la calidad de la asistencia espiritual, ya que falta acompañamiento por parte de la universidad en la búsqueda de espacios para

el desarrollo de las expresiones espirituales. Este tipo de actividades deberían coordinarse por bienestar universitario y tener una difusión con mayor alcance.

5.2.4. Gestión Administrativa

- En este factor predomina la insatisfacción en la agilidad de los trámites administrativos; para este caso se propone dar la oportunidad a los estudiantes y egresados de solicitar certificados y hacer ciertos tramites vía web a través de la plataforma de la universidad. Esto sin duda agilizaría todos los procesos internos y la sensación de demora por parte de los usuarios disminuiría.

5.2.5. Recursos Físicos

- Los laboratorios y talleres representan para los egresados la principal preocupación, por tanto es importante que la universidad empiece a gestionar recursos o convenios para mejorar estos espacios que son los que llegan a ser indispensables para probar el conocimiento teórico aprendido. Se espera que los nuevos diseños de los diferentes edificios que están siendo remodelados tengan como prioridad estos espacios, los cuales también son importantes en el momento de lograr certificaciones de calidad educativa.

5.3. ESTRATEGIAS POR SEGMENTO

A continuación se detallan algunos aspectos que permiten diferenciar el tratamiento que debe brindarse a cada segmento y otros aspectos en los cuales la Universidad Industrial de Santander debe mejorar y fortalecer para hacer más satisfactoria la oferta educativa. Existe información que puede llegar a ser igual para algunos grupos y por tanto se recomienda la verificación de cada una de las

hipótesis presentadas. Las estrategias se plantean mediante la descripción de las características más relevantes para cada segmento.

En el cuadro 3 presentada en capítulo de resultados se muestra el análisis comparativo de los segmentos a partir de variables de estudio que permiten establecer algunas características y comportamientos similares y en otros reflejar las diferencias para facilitar la generación de estrategias diferenciadas.

Momento 0

Egresados Insatisfechos

- Dirigir las diferentes estrategias a los estudiantes y egresados de los programas de Diseño Industrial, Ingeniería de Sistemas y Biología.
- Direccionar las diferentes ofertas del portal de empleo a este grupo para lograr una mayor satisfacción con el trabajo obtenido.
- Realizar un estudio más profundo sobre la satisfacción con el personal docente para recoger las fallas evidenciadas por los estudiantes.
- Observar más detenidamente los espacios dentro de los cuales desarrollan las actividades académicas de los programas a fin de realizar mejoras factibles en estos lugares generando ambientes propicios para el desarrollo del conocimiento.
- Existe inconformidad por el apoyo hacia los estudiantes, especialmente en el apoyo a la gestión de empleos.
- Fortalecer el compromiso en las competencias adquiridas (16) por los estudiantes, especialmente en aquellas que están en el grupo de manejo de información y tecnología TIC’s.

Egresados medianamente satisfechos

- Estrategias direccionadas principalmente a los programas de Ingeniería Industrial, Ingeniería Civil, Economía.
- Hacer un estudio que permita conocer la satisfacción actual de los estudiantes con los docentes para tomar acciones correctivas si son necesarias.
- Mejorar el apoyo a estudiantes en temas relacionados con prácticas e intercambios, ya sea a través de la creación de convenios o de mayor difusión del plan de movilidad de la universidad.
- Reforzar la exigencia en las habilidades comunicativas (escritura, redacción, ortografía, capacidad de persuasión) y en el uso de TIC's para resolver problemas.
- Incentivar la creación de empresas y hacer difusión de los diferentes mecanismos financiadores de ideas nuevas.

Egresados satisfechos

- Dirigida especialmente a estudiantes de Geología, Licenciatura, Ingeniería de Petróleos, Ingeniería Química y Filosofía.
- Hacer revisión de la satisfacción de los estudiantes respecto al personal docente debido a que hay inconformidad con la disponibilidad de tiempo de los profesores.
- Mejorar el apoyo a estudiantes en cuanto a ayuda espiritual con la ayuda de bienestar universitario y desarrollar formas para generar más prácticas.
- Reforzar específicamente las competencias de redacción y exposición de ideas por medios escritos, las cuales tienen el menor grado de satisfacción y a su vez son consideradas como las más útiles. También se buscar reforzar el uso de software para la solución de problemas.

- Hacer énfasis en la creación de empresas puesto que no hay interés ni motivación en este grupo por la generación de nuevas ideas.
- Mantener contacto con los egresados a través del portal de trabajo para que los trabajos conseguidos estén directamente relacionados con la carrera.

Egresados Muy Satisfechos

- Orientado esencialmente a los programas de Derecho y Trabajo social.
- Se pretende conservar el trabajo realizado hasta el momento en las diferentes competencias aunque se sigue evidenciando la importancia de reforzar las competencias de manejo de tecnología e información.
- Hacer revisión de los espacios para estudiar y la biblioteca, puesto que existen inconformismos con estas áreas.

Momento 1

Egresados Medianamente satisfechos

- En este momento hay un cambio en cuanto a que se considera que la calidad de la universidad ha bajado, por tanto se pretende reforzar las competencias aportadas por la institución, especialmente en formulación de proyectos y creación e innovación.
- Ingeniería Industrial deja de formar parte de este grupo y se incluyen carreras como Geología, Biología, filosofía, Física, Licenciatura.

Egresados satisfechos

- Egresados pertenecientes a Ingeniería de sistemas (Nuevo integrante) e Ingeniería Química.

- Persiste la inconformidad con el personal docente, por tanto es necesario el análisis de ese factor.
- Se evidencia disminución del interés por crear empresa.

Egresados Muy Satisfechos

- Forman parte Ingeniería de petróleos e Ingeniería Industrial, las cuales no formaban parte de este grupo.
- Existe preocupación por la disminución del reconocimiento de la universidad.

Momento 3

Para este momento se mantienen las características y estrategias anteriormente nombradas, aunque la percepción de un alto nivel de satisfacción desaparece.

5.4. INDICADORES

A partir de los resultados obtenidos y de la visualización del comportamiento de las diferentes variables se generaron los siguientes indicadores:

$$\%\text{Insatisfaccion personal docente} = \frac{Cantidad\ de\ egresados\ insatisfechos\ con\ el\ PD\ en\ un\ año}{Total\ de\ egresados\ encuestados\ en\ un\ año}$$

$$\%\text{Insatisfaccion apoyo a estudiantes} = \frac{Cantidad\ de\ egresados\ insatisfechos\ con\ el\ AE\ en\ un\ año}{Total\ de\ egresados\ encuestados\ en\ un\ año}$$

%Insatisfaccion recursos físicos

$$= \frac{Cantidad\ de\ egresados\ insatisfechos\ con\ el\ RF\ en\ un\ año}{Total\ de\ egresados\ encuestados\ en\ un\ año}$$

%Insatisfaccion Gestión Admin

$$= \frac{Cantidad\ de\ egresados\ insatisfechos\ con\ el\ G.A.\ en\ un\ año}{Total\ de\ egresados\ encuestados\ en\ un\ año}$$

%Egresados desempleados

$$= \frac{Cantidad\ de\ egresados\ desempleados\ en\ un\ año\ en\ cada\ momento}{Total\ de\ egresados\ encuestados\ en\ un\ año\ en\ cada\ momento}$$

%Egresados insatisfechos

$$= \frac{Cantidad\ de\ egresados\ insatisfechos\ con\ algún\ aspecto}{Total\ de\ egresados\ encuestados\ en\ un\ año}$$

%Egresados trabajando en áreas relacionadas a su carrera

$$= \frac{Cantidad\ de\ egresados\ de\ cada\ escuelatrabajando\ en\ algo\ relacionado\ a\ su\ carrera}{Total\ de\ egresados\ encuestados\ en\ un\ año\ pertenecientes\ a\ cada\ escuela}$$

%Egresados a los que los padres les pagaron el estudio

$$= \frac{Cantidad\ de\ egresados\ \text{a los que los padres les pagaron el estudio}}{Total\ de\ egresados\ encuestados\ en\ un\ año}$$

%Egresados insatisfechos con uno de los grupos de competencias adquiridas

$$= \frac{Cantidad\ de\ egresados\ \text{insatisfechos con unp de los grupos de competencias}}{Total\ de\ egresados\ encuestados\ en\ un\ año}$$

%Egresados con salarios superiores a 5 smlv en cada escuela

$$= \frac{Cantidad\ de\ egresados\ \text{con salarios superiores a 5 smlv en cada escuela}}{Total\ de\ egresados\ encuestados\ en\ un\ año\ en\ cada\ escuela}$$

5.5. BASE DE DATOS

Durante el desarrollo del presente trabajo la más fuerte limitación encontrada fue la obtención de una base de datos o suficientemente robusta y pura para realizar el proceso de minería de datos. Básicamente el problema surge porque el seguimiento a egresados es un proceso que se encuentra en etapa desarrollo y todavía no ha alcanzado la madurez suficiente dentro de las Instituciones de Educación Superior para llevarse a cabo con toda la severidad del caso; es por ello que a pesar de la gran cantidad de variables contenidas en la base de datos, no fue posible analizarlas todas, era notoria la falta de respuestas a muchas de las preguntas, las cuales hubieran jugado un papel interesante en cuanto a la generación de información si hubiesen sido contestadas.

Se recomienda por tanto formular encuestas más amigables con un sistema de recolección que permita guardar el contenido de las preguntas aunque el egresado tenga fatiga por la contestación de la misma, además se hace necesario fijar algunas preguntas como obligatorias debido a la importancia de sus respuestas.

Finalmente y quizás el factor clave dentro de este proceso es velar por mantener el contacto con los egresados a través del tiempo, el diseño de estrategias específicas de contacto no solo permitirá una mejora en este procedimiento, sino que el sentido de pertenencia de los egresados será mucho mayor. También en necesario que este tipo de proyectos tengan la colaboración de cada una de las escuelas y cuenten con un direccionamiento del proceso por parte de una oficina de egresados o de un departamento de relaciones exteriores, que consagre gran parte de su trabajo a observar la calidad de los estudiantes de su universidad, los cuales en este caso son el producto que entrega la institución a la sociedad.

6. CONCLUSIONES

- La Universidad Industrial de Santander se encuentra en una etapa de crecimiento en el proceso de seguimiento a egresados que debe continuar fortaleciendo para que la consecución de datos permita realizar estudios con análisis más precisos y en los que se facilite el uso de técnicas y herramientas de minería de datos.

- El diseño actual de la encuesta resulta ser dispendioso para los egresados, por tanto la recolección de datos se dificulta en la medida en que la extensión de la encuesta no permite que los egresados conserven el interés de contestar todas las preguntas y se pierda gran parte de la información que debió ser recolectada.

- La carencia de un departamento o una oficina de egresados que se encargue del proceso de seguimiento dificulta la observación que necesita realizar la universidad en pro de mejorar la oferta educativa y la calidad de sus egresados.

- La escuela de Geología resulta ser la escuela más comprometida con la actualización de la base de datos de sus egresados, permitiendo mantener el contacto con los mismos y la colaboración para realizar los estudios de seguimiento.

- Se dificulta la realización de un seguimiento longitudinal de egresados, debido a que en la medida en que aumentan los años de egreso es más fácil perder contacto con el egresado, por tanto, la cantidad de egresados contactados

para el momento 0 se reduce en un 90,39% para el momento 1 y un 96,4% para el momento 3, imposibilitando la opción de este tipo de análisis.

- El 78,3% de los egresados del momento 0, tiene como actividad principal estudiar, mientras que para el momento 1 es trabajar con el 55% y para el momento 3 lo es con el 90%, mostrando que a través del tiempo los egresados logran estabilizar su situación laboral.

- Es posible agrupar las competencias adquiridas en 3 categorías debido a que su comportamiento es similar y tiene relación entre sí, estas categorías son habilidades comunicativas, manejo de tecnología e información y capacidad de análisis.

- En el momento 0, la segmentación de los egresados arrojó 4 grupos, mientras que para el momento 1 fueron 3 y para el momento 2 se obtuvieron 2.

- Los factores que registran los mayores grados de insatisfacción son, apoyo a estudiantes con el 38%, gestión administrativa con el 26% y competencias adquiridas con el 15%.

- El estudio del nivel de satisfacción demostró que la Universidad Industrial de Santander se encuentra en un nivel óptimo, sin embargo existen oportunidades de mejora específicas en cada uno de los aspectos evaluados para que la universidad mejores su calidad formativa, entre ellos el apoyo a estudiantes.

- La construcción de los perfiles de los egresados de acuerdo a las competencias adquiridas permitió la identificación de oportunidades de mejora, facilitó la generación de estrategias enfocadas a cada grupo y afianzó la necesidad de continuar con un estudio profundo del nivel de satisfacción de los egresados.

- El uso de redes bayesianas permitió la identificación de relaciones entre las variables de satisfacción como relaciones de independencia/dependencia entre aspectos familiares y recursos ofrecidos por la universidad.

- El conocimiento de los egresados juega un papel fundamental en la generación de nuevas estrategias que permitan a la universidad mejorar sus procesos educativos y tomar decisiones pertinentes que apoyen e fortalecimiento de la calidad de la educación.

7. RECOMENDACIONES

- La Universidad Industrial de Santander debe asumir un compromiso mucho más fuerte con sus egresados a través del diseño y ejecución de estrategias de contacto que permitan tener el apoyo de los egresados en la realización de estudios que midan el impacto de sus profesionales en el mundo laboral.

- Es importante definir una oficina o departamento dentro de la universidad que se encargue de mantener el contacto con los egresados y de liderar todos los estudios referentes a este tema, de modo que el proceso de seguimiento a egresados se tome con mayor severidad y diligencia.

- El diseño de una plataforma específica para la recolección de datos, que le brinde autonomía a la universidad para realizar investigaciones frente a este tema, es de vital importancia debido a que permite diseñar mecanismos para seccionar las encuestas, almacenar información conforme se va diligenciando el formato y direccionar preguntas de acuerdo al tiempo de egreso o de características específicas que tenga el encuestado.

- La renovación en el sistema de recolección de datos es ineficiente sino se realizan análisis minuciosos y constantes con el interés de encontrar información útil para tomar decisiones, es por ello que programar fechas específicas para los tipos de estudios denota una mayor seriedad por parte de la universidad en continuar con esta labor.

- Direccionar estrategias correctivas para mejorar la oferta educativa partiendo de las opiniones de los egresados, lo cual es fundamental para la generación de valor, ya que no solo se está teniendo en cuenta la opinión de los

egresados para conocer el impacto externo de la universidad sino que también se está buscando reforzar el proceso educativo desde adentro y por medio del conocimiento de las personas que reciben los servicios prestados por la universidad.

- Emplear herramientas tecnológicas y hacer uso de las diferentes redes sociales para mantener el contacto con los egresados sin duda ayudaría a que la recolección de datos fuera mucho más eficiente. Como propuesta más ambiciosa sería interesante la creación de una red social propia de la universidad, en la cual fuera posible fomentar el contacto entre egresados, docentes, estudiantes y demás personal que labore dentro de la institución.

- Ofrecer servicios específicos a los egresados por hacer parte de la comunidad UIS reforzaría el sentido de pertenencia de los mismos hacia la institución, como por ejemplo invitaciones a congresos, festivales, eventos deportivos, así como el ofrecimiento de opciones de estudio, propuestas de trabajo entre otros beneficios que resulten atractivos conociendo los interese particulares de los egresados.

- A partir de los resultados obtenidos definir estrategias mínimas para reforzar los procesos educativos de la universidad y los factores que afectan la realización de dichos procesos. Se sugiere ser constante en la realización de las encuestas e ir reformando las mismas dependiendo de las necesidades de los diferentes momentos. De este modo las estrategias diferenciadas por cada segmento también se debe ir re-definiendo teniendo como punto de partida las diseñadas por los análisis del año 2010.

- Generar un mayor compromiso por parte de las diferentes escuelas en el fortalecimiento de las relaciones universidad-egresado al igual que concientizar

sobre la relevancia de llevar a cabo estudios como el de seguimiento para realizar mejoras profundas con un enfoque definido sobre el cual

BIBLIOGRAFÍA

- ALATAS, B., E. AKIN, and A. KARCI, MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules. Applied Soft Computing, 2008. P. 646–656.

- ALEMANIA. BUNDESMINISTERIUM FÜR WIRTSCHAFTLICHE ZUSAMMENARBEIT UND ENTWICKLUNG. Estudio de seguimiento de egresados de Programas de posgrado regionales Centroamericanos. Tegucigalpa, 2005. 161 p.

- ALLEN, Jim and VAN DER VELDEN, Rolf. La transición desde la educación superior al empleo. In: Jiménez Aguilera, Juan de Dios, Sánchez Campillo, José and Montero Granados, Roberto (eds.): Educación Superior y Empleo: la Situación de los Jóvenes Titulados en Europa. La Encuesta CHEERS. Universidad de Granada. Granada, 2003. p.69-90.

- ALVAREZ, Hernán Darío. Creación y puesta en marcha de la oficina de egresados. Universidad Nacional. Bogotá, 2004.

- ANÁLISIS DE CARACTERÍSTICAS del ambiente creativo en empresas de Manizales con técnicas KDD. Universidad Nacional de Colombia. Manizales, 2009.

- BAEZ, Fabiola. Plan de desarrollo de la facultad de ingeniería 2007 – 2011. Mexico, 2007.

- BELTRÁN, Yolima. Seguimiento eficiente a graduados : propuesta de una metodología sistemática para implementar la política de egresados de la UIS. Bucaramanga, 2008.

- BRODLEY, C.E.; LANE, T. and STOUGH, T.M. Knowledge discovery and data mining. American Scientist. Vol. 86, 1999. p. 55-65.

- CANALLI, Andrea. IX Informe de AlmaLaurea sobre la condición ocupacional de los licenciados. Osservatorio Statistico de la Universidad de Bolonia, 2007. p. 93 – 101.

- COLOMBIA. DECRETO 1108. MINISTERIO DE EDUCACION NACIONAL. Bogotá, 2008.

- COLOMBIA. DIRECCION DE FOMENTO DE LA EDUCACIÓN SUPERIOR. Estudio de Seguimiento a Egresados, 2008.

- COLOMBIA. GRADUADOS COLOMBIA. Observatorio Laboral para la Educación. En : Visión Educativa [en línea]. [Consultado 14 Febrero 2013]. Disponible en <http://www.graduadoscolombia.edu.co>.

- COLOMBIA. MINISTERIO DE EDUCACION. Colombia APRENDE : Graduados Colombia : Beneficios del sistema de información del Observatorio Laboral para la Educación. 2007.

- _____ Convocatoria para apoyar proyectos que fortalezcan el proceso de seguimiento a graduados en las instituciones de educación superior, 2007. 127 p.

- COLOMBIA. OBSERVATORIO LABORAL PARA LA EDUCACIÓN. Experiencias Internacionales de Seguimientos a Graduados, 2011.

- CONGRESO NACIONAL EMPRESAS, ORGANISMOS ESTATALES, ASOCIACIONES E IES. Bogotá, 2006.

- DAMIÁN, Javier Simón. El técnico superior universitario en administración : Origen, trayectoria estudiantil y desarrollo profesional. Universidad del Papaloapan, campus Tuxtepec. Oaxaca, México, 2003. p. 22 – 35.

- ESQUEMA BÁSICO para estudios de egresados. ANUIES. Universidad Autónoma de México, México. 1998.

- ESTATUTOS UNIANDINOS. Asociación de egresados de la Universidad de los ANDES Bogotá, 2001.

- FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. and UTHURUSAMY, R. Knowledge and data mining. Cambridge (Massachussets), 1996.

- FELGAER, Pablo. Optimización de redes bayesianas basado en técnicas de aprendizaje por inducción. Buenos Aires, 2005, 145 p. Tesis (Maestría en Ingeniería Informática). Universidad de Buenos Aires. Facultad de Ingeniería.

- GALLARDO ARANCIBIA, José Alberto. Metodología para el desarrollo de proyectos de minería de datos CRISP-DM. En: EPB 603 Sistemas del Conocimiento. [en línea]. [Consultado 9 de marzo de 2013]. Disponible en: <http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf >

- GARCIA-ARACIL, Adela; GABALDÓN, Daniel; MORA, Jose-Gines and VILA, Luis E. The relationship between life goals and fields of study among young european graduates. In: Higher Education, vol. 53, no. 6. 2007. p. 843-865

- GOMEZ, Nini Johana. Seguimiento a graduados pontificia universidad bolivariana. Bucaramanga, 2008. p.87.

- HAN, Jiawei; KAMBER Micheline . Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers. Estados Unidos, 2000.

- HERNÁNDEZ, José et al. Introducción a la minería de datos, 2004, p. 22, 237, 257,281.

- INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY. Red GRADUA2 y la Asociación Columbus. Manual de instrumentos y recomendaciones sobre el seguimiento de egresados, 2006. 823 p.

- JOHNSON, R., WICHERN y DEAN GONDAR, E. “Análisis Cluster, Applied Multivariate Statistical Análisis” Prentice-Hall International, Inc, 1982.

- KRYSZKIEWICZ, M. Fast Discovery of Representative Association Rules. In First International Conference on Rough Sets and Current Trends in Computing. Poland,1998.

- LUAN, Fing. Business Analytics Software : Aplicaciones de minería de datos en la educación Superior. IBM Corporation, Estados Unidos, 2010.

- MALAGÓN, Constantino. Clasificadores bayesianos. El algoritmo de Naïve Bayes. 2003

- MARTINEZ, Patricia. Origen del observatorio laboral para la educación. Bogotá, 2005.

- MEMORIAS DEL PRIMER encuentro nacional de oficina de egresados. Medellín, 2005. p. 345.

- MEXICO. SECRETARIA DE EDUCACIÓN PÚBLICA. Manual de Operación e Instructivo para la captura de datos estadísticos en el Sistema Integral de Información : SII-DGEST. México, 2012.

- MOLINA, Luis Carlos. Data Mining: Torturando a los datos hasta que confiesen, 2002.

- MORA, José Guinés y CAROT, José Miguel. El profesional flexible en la sociedad del conocimiento. Universidad Politécnica de Valencia. España, 2010. p. 13 – 19, 243-256.

- PAGANI, Rafaella. Glosario. Tuning educational structures in Europe, Español-Inglés Inglés-Español : Agencia Nacional de Evaluación de la Calidad y Acreditación y Agencia de Calidad : Acreditación y Prospectiva de las Universidades de Madrid. Madrid. En: DAAD. [En línea] (2003). [Consultado 22 febrero 2013]. Disponible en: <http://www.uam.es/europea/glosario_convergencia_tuning.pdf.>

- PAUL, Jean-Jacques and MURDOCH, Jake. The teaching quality of higher education institutions and graduate employment: a comparison across 10 European countries. Paper presented at the seventh annual workshop of the European Research Network on Transitions in Youth. Antwerp, 7-10, 2000.

- PROGRAMA DE EGRESADOS. Universidad de la Sabana. Bogotá, 2006.

- RAMOS, Teófilo. Manual de instrumentos y recomendaciones sobre el seguimiento de egresados. Red GRADUA2 : Asociación Columbus, 2006. p. 13 – 27.

- SAAVEDRA, Gustavo Andrés. Estructura observatorio laboral para la educación. Bogotá, 2007. p. 133.

- SCHIATTINO I, SILVA C. Árboles de Clasificación y Regresión: Modelos Cart. Cienc Trab. 2008, P. 161 - 166.

- SCHOMBURG, Harald. Graduate surveys in germany as a tool to measure and improve the relevance of higher education, 2010. p. 235 – 283.

- SEGUIMIENTO EFICIENTE a graduados : propuesta de una metodología sistemática para implementar la política de egresados UIS. Grupo de Investigaciones Educativas ATENEA. Escuela de Educación. Universidad Industrial de Santander. Bucaramanga, 2008.

- SEMINARIO PROYECTO ALFA GRADUA2. Segunda reunión Monterrey. Mexico,2005. p. 384.
- SERIE ORIENTADORES Universitarios N°33 “Los antiguos alumnos y sus asociaciones”. Universidad del Norte, 2005.

- SERVENTE, M. y GARCÍA MARTÍNEZ, R. Algoritmos TDIDT Aplicados a la Minería Inteligente, artículo publicado en la Revista del Instituto Tecnológico de Buenos Aires. ISSN 0326-1840, 2002.
- TORRES, José Maximiliano. Dinámica del observatorio laboral : Para qué y quienes. Bogotá, 2007.

- VALENTI, Giovanna and VARELA, Gonzalo. Diagnóstico sobre el estado actual de los estudios de egresados. Asociación Nacional de Universidades e Instituciones de Educación Superior ANUIES, 2004. p. 10 – 35.

- VAN DER VELDEN, Rolf and ALLEN, Jim. Europe needed : flexible professionals, 2008. p. 157 – 185.

- VEITCH, William R. Identifying Characteristics of High School Dropouts: Data Mining With A Decision Tree Model. Annual Meeting of the American Educational Research Association. San Diego, CA , 2004.

- VENEGAZ, Victor Alejandro. Balance y proyecciones del programa de seguimiento a egresados. Bogotá, 2007. p. 231.

- VICENTE VILLARDÓN, Jose Luis. Introducción al análisis de clúster. Universidad de Salamanca, Madrid, 2009. 345 p.

- VILA, Luis Eduardo. Experiencias de España en el seguimiento a graduados. Colombia APRENDE. Bogotá, 2011. 148 p.

- WITTEN & FRANK, Data Mining: Practical Machine Learning Tools and Techniques, citado por HERNÁNDEZ, José et al, Introducción a la minería de datos, 2004, p. 5.

- ZAKI, M.J., et al., New Algorithms for Fast Discovery of Association Rules. University of Rochester. Estados Unidos, 1997

ANEXOS

(VER ARCHIVO ANEXO)