REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

FACULTAD DE CIENCIAS E INGENIERÍAS FÍSICAS Y FORMALES

PROGRAMA PROFESIONAL DE INGENIERÍA DE SISTEMAS

"PROPUESTA DE GESTION DE INFORMACION DE AGRUPAMIENTO (CLUSTERING), UTILIZANDO TECNICAS DE MINERIA DE DATOS PARA EGRESADOS DEL PROGRAMA PROFESIONAL DE INGENIERÍA DE SISTEMAS DE LA UNIVERSIDAD CATÓLICA DE SANTA MARÍA"

TESIS PRESENTADA POR:

León Lipe, Noé Juan

Muñoz Mazuelos, Cristhian Juan Humberto

Para Optar El Título Profesional De Ingenieros De Sistemas.

Arequipa - Perú
2013

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE
TESIS UCSM

CATÓLICA
DE SANTA MARÍA

Dedicatoria

Dar gracias a Dios por guiarnos y amarnos, con profundo amor a mis Padres Erry y Brígida, que con su ejemplo y buena voluntad me guiaron en este paso importante en mi vida, a mi hermana Melany, la cual siempre me apoyo y dio esa fuerza para salir adelante y a todas las personas que supieron confiar y valorar mi esfuerzo.

Muñoz Mazuelos, Cristhian Juan Humberto

A mis padres y hermanos, que me han ayudado y brindado su apoyo en los momentos más importantes de mi vida.

León Lipe, Noé Juan

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

INDICE

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

INDICE DE FIGURAS

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

INDICE DE TABLAS

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

RESUMEN

En la presente Tesis hemos procedido a explicar la problemática de los egresados del Programa Profesional de Ingeniería de Sistemas de la Universidad Católica de Santa María, teniendo en cuenta que al terminar la carrera universitaria los egresados tienden a desvincularse de su universidad, para lo cual hemos propuesto desarrollar una Gestión de información para los egresados.

Esta Gestión está constituida por un proceso de tratamiento de los datos, para la obtención de conocimiento, la cual se inicia con la selección de datos de las diferentes fuentes (bases de datos, hojas de cálculo, archivos planos, Data Warehouse, etc.), las cuales van a ser tratadas con las diferentes herramientas propuestas para la gestión de datos como son: Simple K-Medias, Pentaho, Weka.

Se desarrolla un proceso de extracción, transformación y carga de los datos de los egresados, continuando con el filtrado, la segmentación y agrupación de la información con características y patrones similares que nos van a permitir conocimiento para hacer análisis y obtener una buena toma de decisiones para beneficiar al Programa Profesional de Ingeniería de Sistemas de acuerdo a las necesidades tanto en el ámbito académico, laboral y tecnológico.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

ABSTRACT

In this thesis we have proceeded to explain the problem of graduates from the Professional Program in Systems Engineering from the Catholic University of Santa Maria, bearing in mind that at the end of college graduates tend to disengage from their university, for which we proposed developing an information management graduates.

This management process consists of a treatment of the data, to obtain knowledge, which begins with the selection of data from different sources (databases, spreadsheets, flat files, Data Warehouse, etc..) , which will be treated with the different tools proposed for data management such as: Simple K-Means, Pentaho, Weka.

It develops a process of extraction, transformation and loading of data from the graduates, continuing with filtering, segmentation and grouping of information with similar characteristics and patterns that knowledge will allow us to get a good analysis and decision making to benefit the Professional Program in Systems Engineering according to the needs in academic, labor and technology.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

CAPÍTULO I

PLANTEAMIENTO TEORICO

1.1. PROBLEMA.

1.1.1. Título.

"PROPUESTA DE GESTION DE INFORMACION DE AGRUPAMIENTO (CLUSTERING), UTILIZANDO TECNICAS DE MINERIA DE DATOS PARA EGRESADOS DEL PROGRAMA PROFESIONAL DE INGENIERÍA DE SISTEMAS DE LA UNIVERSIDAD CATÓLICA DE SANTA MARÍA"

1.1.2. Descripción Del Problema.

Uno de los más grandes inconvenientes que enfrentan las universidades del Perú es no saber el estado en el que se encuentran sus egresados. En la actualidad existen pocas aplicaciones que interpretan los datos almacenados de sus egresados y que están orientadas a la gestión de los mismos. Por ello es conveniente para las universidades y sus Programas Profesionales tener conocimiento de la situación actual de sus egresados para obtener beneficios mutuos y poder fortalecer las relaciones ya sea académico o profesional entre los egresados y los Programas Profesionales de las distintas universidades.

Tomamos en cuenta que los egresados se desvinculan rápidamente de la institución académica, debido a que toman diferentes rumbos, ya sea en su futuro académico o laboral, por ende no se cuenta con una verdadera

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

información de los egresados y esto nos conlleva a que no se puede realizar una verdadera categorización de dichos egresados.

La información disponible en la Web se está incrementando cada vez más volviéndose por este motivo muy compleja la organización de la misma trayendo consigo que las operaciones de segmentación y categorización sean dificultosas ante este problema nos vemos en la necesidad de estructurar la información. Cuando se realizan las consultas tradicionales estas se realizan de manera global en todo el conjunto de la base de datos, lo cual se transforma en un proceso lento y complicado una de las razones es la mala preparación de los datos.

La Minería de Datos se presenta como una tecnología de apoyo para la organización, que permite comprender y aplicar el conocimiento obtenido de la exploración, extracción y análisis de los datos almacenados para convertirlos en información útil para la toma de decisiones.

1.1.3. Área y Línea de Investigación.

- **Área**
  - Base de datos.
  - Minería de Datos (Data Mining).

- **Línea.**
  - Agrupamiento (Clustering).

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

1.1.4. Variables e indicadores.

1.1.4.1. Variable Independiente.

- K-Medias
- Weka
- Pentaho

1.1.4.2. Variable Dependiente.

Gestión de Agrupamiento (Clustering) de egresados.

- **Indicadores.**
  - Ergonómica
  - Multiplataforma
  - Dinámica

1.1.5. Tipo y Nivel de Investigación.

1.1.5.1. Tipo de Investigación.

Investigación aplicada.

1.1.5.2. Nivel de Investigación.

Explicativa porque se pretende proponer la Gestión de información de agrupamiento de los egresados en el Programa Profesional de Ingeniería de Sistemas (PPIS).

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

1.1.6. Delimitaciones.

1.1.6.1. Delimitación Espacial.

El Programa Profesional de Ingeniería de Sistemas (PPIS) de la Universidad Católica de Santa María (UCSM).

1.1.6.2. Delimitación Temporal.

Último semestre de 2012.

1.1.6.3. Delimitación Social.

Egresados del Programa Profesional de Ingeniería de Sistemas (PPIS) de la Universidad Católica de Santa María (UCSM).

1.1.6.4. Delimitación Conceptual.

Descubrimiento de la Información, Gestión de la Información.

1.2. ALCANCES Y LIMITACIONES.

1.2.1. Alcances.

- Egresados del Programa Profesional de Ingeniería de Sistemas (PPIS) de la Universidad Católica de Santa María (UCSM).

1.2.2. Limitaciones.

- Egresados del Programa Profesional de Ingeniería de Sistemas (PPIS).

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

1.3. VIABILIDAD.

1.3.1. Económica.

Los recursos serán obtenidos por financiamiento propio.

1.3.2. Técnica.

Se cuenta con la capacidad académica para resolver el problema.

1.3.3. Operativa.

Análisis de situaciones e investigaciones, en medios bibliográficos, Internet, bibliotecas y docentes del Programa Profesional de Ingeniería de Sistemas (PPIS).

1.4. OBJETIVOS.

1.4.1. Objetivo General.

Desarrollo de propuesta de gestión de información de agrupamiento (Clustering) de egresados, utilizando técnicas de minería de datos para el Programa Profesional de Ingeniería de Sistemas (PPIS) de la Universidad Católica de Santa María (UCSM).

1.4.2. Objetivos Específicos.

- Utilizar herramientas multiplataforma de minería de datos para la gestión de información de los egresados.
- Agrupar a los egresados para categorizarlos según sus patrones similares, aplicando algoritmos de agrupamiento (Clustering).

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

- Obtener grupos específicos de los egresados que nos permita tener información adecuada para la toma de decisiones.
- Beneficiar al programa profesional de ingeniería de sistemas en base al agrupamiento de la información adecuada.

1.5. HIPOTESIS.

Dado que se desconoce el agrupamiento de egresados del Programa Profesional de Ingeniería de Sistemas (PPIS), es probable que utilizando las técnicas de Minería de Datos y K-Medias estos se puedan obtener.

1.6. JUSTIFICACION DE LA INVESTIGACION

En este tiempo en que la tecnología se va desarrollando de manera abrumadora encontramos la Web semántica como una alternativa para captar información de los egresados con la finalidad de poder categorizarlos según intereses en común.

La Web semántica permite tener una mejor explotación de las redes sociales ampliando la interoperabilidad entre los sistemas informáticos.

La Minería de Datos, prepara, sondea y explora patrones de comportamiento de los datos para sacar la información oculta.

Existen varios tipos de técnicas que pueden trabajar en la gestión de la información y una de ellas es el Agrupamiento (Clustering) en la Minería de Datos, este Agrupamiento nos permite segmentar, agrupar o asociar la información. Podemos decir que la Minería de Datos es importante contar con técnicas de recuperación de información para poder obtener datos relevantes orientados al estudio de la gestión de egresados.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

CAPÍTULO II

MARCO TEÓRICO.

2.1. ESTADO DEL ARTE.

Implementación de un sistema CRM analítico orientado a la segmentación de clientes y caracterización de perfiles con técnicas de Data Mining utilizando algoritmos genéticos. Pérez Bravo Roberto y Medina Bravo Leoncio, Universidad Católica de Santa María, 2007. Esta tesis plantea el desarrollo de un sistema CRM analítico el cual está orientado a la segmentación de base de datos de clientes en grupos e indicadores homogéneos y la posterior caracterización de perfiles utilizando el algoritmo K-Means y algoritmos genéticos.

Metodología de modelamiento de Datos para Data Mining. Tello Torres Carlia, Universidad Católica de Santa María, 1999. Esta tesis propone una base de datos organizada para la toma de decisiones viendo el proceso de modelar los datos transformándolos, eliminando inconsistencias y cargándolos en una base de datos organizada para poder utilizarlos mediante la técnica de Minería de datos de esta manera se aceleran los procesos de análisis, consulta y uso de información.

Propuesta metodológica para la recuperación de información mediante Clustering y reglas de asociación para Web Mining. Álvarez Tapia Karina y Pacheco Calderón Christian, Universidad Católica de Santa María 2006. Este trabajo propone una

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

metodología que ayudara a mejorar la recuperación de información mediante la fusión de dos técnicas de Minería de Datos en primer lugar el Clustering técnica que permite corresponder cada caso a una clase utilizando medidas de similaridad en segundo lugar la técnica de reglas de asociación que permitirá encontrar tendencias para entender y explorar patrones de comportamiento de los datos.

Propuesta de un modelo de presentación empleando Web semántica para la especificación, diseño e implementación de requisitos de interfaz de usuario Velásquez Frisancho Lilian Patricia, Universidad Católica de Santa María, 2009. El propósito de este tesis es construir un modelo de presentación utilizando algunos componentes de la web semántica, para especificar requisitos de interfaz de usuario para lograr esto se muestra un conjunto de fases y patrones de los lenguajes ontológicos y para obtener la presentación se utilizó el XML.

Propuesta de optimización en la implementación de un Data Mining de operaciones a partir del análisis de metodologías existentes. Montánchez Montesinos Elizabeth y Del Carpio Béjar Christian, Universidad Católica de Santa María, 2009. El objetivo principal de esta tesis es analizar las metodologías existentes de Minería de Datos, para obtener una obtener una óptima metodología que permite implementar Minería de Datos de operaciones.

Propuesta de preparación de datos documental para consultas utilizando un algoritmo particional de Data Mining, Contreras Salas Lucia Alejandra, Universidad Católica de Santa María, 2009. El presente trabajo plantea una propuesta de segmentación de datos documental, utilizando el algoritmo Clustering particional, basándose en la técnica de K-means, para el desarrollo de este trabajo se ha utilizado la Minería de Datos en la cual mediante la segmentación de datos se darán resultados precisos en periodos cortos de tiempo.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Modelo de Integración de Sistemas de Información Empresariales con Redes Sociales a través de APIS para la Universidad Católica de Santa María, Guevara Paucar Sergio Andrés, Universidad Católica de Santa María, 2012. El objetivo principal de esta tesis es analizar sistemas de información empresariales a través de las redes sociales, perfiles, sistemas de información, etc. que permita recuperar información útil para las empresas.

2.2. INTRODUCCIÓN MINERIA DE DATOS.

La idea de la Minería de Datos no es nueva como se cree, ya desde los años sesenta los estadísticos manejaban términos como Data fishing, Minería de Datos o Data archeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido.

Las listas de discusión sobre este tema las forman investigadores de más de ochenta países y gracias a este tema de discusión es que esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios.

La Minería de Datos es una tecnología compuesta por etapas que integra varias tareas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente. [BEN, et al, 02]

Cada día las organizaciones se enfrentan a un mundo cada vez más competitivo y, por tanto, las estrategias de administración deben ser flexibles para adaptarse a las condiciones cambiantes del entorno. Lo que significa un gran reto para las organizaciones, es el poder manejar grandes volúmenes de información dentro de las mismas, para poder conocer su entorno y poder predecir su evolución.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Son muchos los motivos que nos llevan a generar información, esto nos ayuda a controlar, optimizar, administrar, examinar, investigar, planificar, predecir, someter, negociar o tomar decisiones de cualquier ámbito según el dominio en que nos desarrollemos. La información por sí misma está considerada un bien patrimonial, de esta forma, si una empresa tiene una pérdida total o parcial de información provoca muchos perjuicios. Es evidente que la información debe ser protegida, pero también explotada.

Unos de los principales factores que en la actualidad nos han permitido generar tanta información son:

- Los bajos costos de los sistemas de almacenamiento tanto temporal como permanente.
- El incremento de las velocidades de cómputo en los procesadores.
- Las mejoras en la confiabilidad y aumento de la velocidad en la transmisión de datos.
- El desarrollo de sistemas administradores de bases de datos más poderosos.

Todas estas ventajas nos han llevado a abusar del almacenamiento de la información en las bases de datos.

En la Figura 2.1 podemos mostrar la jerarquía que existe en una base de datos; entre dato, información y conocimiento. Se observa igualmente el volumen que presenta en cada nivel y el valor que los responsables de las decisiones le dan en esa jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así entre la información y el conocimiento. La Minería de Datos trabaja en el nivel superior buscando patrones, comportamientos, agrupaciones, secuencias, tendencias o asociaciones que puedan generar algún modelo que nos permita comprender mejor el dominio para ayudar en una posible toma de decisión.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Figura 2. 1 Relación entre dato, información y conocimiento.

Fuente: [BEN, et al, 02].

Tomando en consideración la importancia de extraer los conocimientos “perdidos” en los datos que almacena la organización, ha surgido desde hace un tiempo lo que se conoce como “Minería de Datos”.

El número de fuentes de información disponibles en Internet ofrece una nueva oportunidad de búsqueda y extracción de información útil a partir de esta "base de datos" dinámica y creciente. Según estadísticas se estima que cada 20 meses se duplica la cantidad de información en el mundo. Cada vez más se necesita la ayuda de computadores potentes para automatizar el proceso inductivo, para analizar de forma inteligente la cantidad de datos existentes, y extraer ese conocimiento oculto y valioso. [BEN, et al, 02]

La Minería de Datos es considerada uno de los puntos más importantes de los sistemas de bases de datos, y uno de los desarrollos más prometedores interdisciplinariamente en la industria de la información.

La Minería de Datos es parte del Proceso llamado KDD (Descubrimiento de conocimiento en bases de datos), presentamos una breve introducción de este proceso.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

2.3. KDD (DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS)

A principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a consolidar los términos de la Minería de Datos y KDD (Descubrimiento del Conocimiento en Base De Datos). A finales de los años ochenta solo existían un par de empresas dedicadas a esta tecnología. En 2002 existen más de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones.

En las empresas se están aplicando procesos de descubrimiento de conocimiento sobre las bases de datos KDD (Descubrimiento de conocimiento en bases de datos), complementándose con Sistemas de Apoyo a Decisiones (DSS) a los que corresponde la Minería de Datos como una fase fundamental para facilitar su ejecución, éstos permiten la manipulación de dichos datos, buscando obtener resultados positivos para la compañía al ser sistemas que apoyan a los administradores en la toma de decisiones.

KDD (Descubrimiento de conocimiento en bases de datos), consiste en la aplicación de una serie de pasos que inician con el tratamiento de los datos, preparándolos para la Minería de Datos hasta la evaluación y visualización de la información obtenida, en la cual se observa el conocimiento extraído de los datos que entraron inicialmente al sistema. Podemos determinar que los datos son la materia prima bruta, en el momento que el usuario les atribuye algún significado pasa a convertirse en información.

El objetivo principal del área de KDD (Descubrimiento de conocimiento en bases de datos), es el procesar automáticamente grandes cantidades de datos para encontrar conocimiento útil para un usuario y satisfacerle sus metas. “El KDD es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos” [TAN, et al, 06].

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

KDD es un proceso multidisciplinario que encierra: conocimiento, aprendizaje, bases de datos, estadística, sistemas expertos y representación gráfica, entre otros.
Las fases principales que integran el proceso de KDD son (Figura 2.2):

Selección: Selección de los datos relevantes para el análisis (son obtenidos de la base de datos).

Pre procesamiento: Limpieza de datos, estudio de la calidad de los datos y determinación de las operaciones de minería que se pueden realizar. Transformación: Conversión de datos en un modelo analítico, donde los datos se transforman y consolida en formas apropiadas para la minería.

Minería de Datos: Tratamiento automatizado de los datos seleccionados con una combinación apropiada de algoritmos para extraer conocimiento de los datos. Interpretación y evaluación: Identificación de patrones representativos del conocimiento.

Conocimiento: Aplicación del conocimiento descubierto.

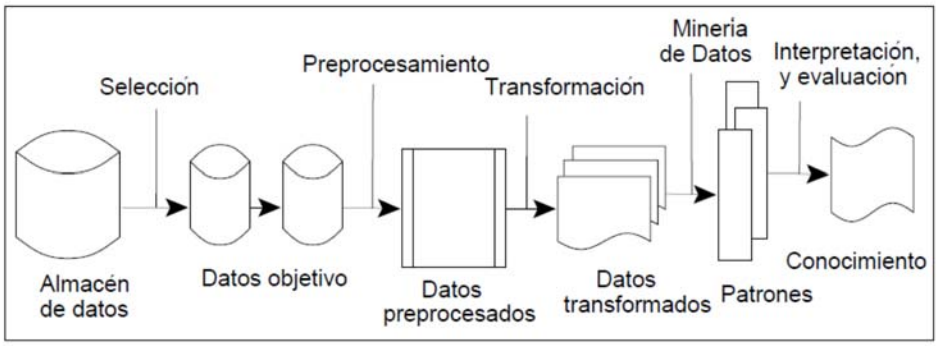


Figura 2. 2 Proceso KDD.

Fuente: [TAN, et al, 06].

A menudo se suele confundir la definición del proceso KDD con la Minería de Datos a pesar que este último es solo una fase del proceso KDD.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

En si podemos decir que la Minería de Datos es considerado uno de los puntos más importantes de los sistemas de bases de datos y uno de los desarrollos más prometedores interdisciplinariamente en la industria de la información. [HER, 06].

Lo que en verdad hace la Minería de Datos es reunir las ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Grafica, las Bases de Datos y el Procesamiento Masivo, principalmente usando como materia prima las bases de datos [BEN, et al, 02].

2.4. MINERIA DE DATOS.

2.4.1. Conceptos

Hay varias definiciones para poder conceptuar la Minería de Datos pero la esencia de estás se fundamenta en el concepto de escarbar en la información recolectada para descubrir elementos de gran utilidad y usarlos para detectar patrones de comportamiento que se repiten, o encontrar relación entre los diferentes datos. [BEN, et al, 02].

Definiendo un poco más formal la Minería de Datos podemos decir que es el conjunto de técnicas y herramientas usadas para encontrar y entender relaciones en una gran cantidad de datos y presentarlas en una forma útil y ventajosa. [FAY, et al, 96].

La Minería de Datos es un mecanismo de explotación, consistente en la búsqueda de información valiosa en grandes volúmenes de datos, con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones. [TAN, et al, 06].

La idea principal de la Minería de Datos es combinar la flexibilidad, creatividad y conocimiento general de una persona con la potencia de cálculo

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

y la capacidad de almacenamiento de un computador para una exploración de datos efectiva.

La Minería de Datos es la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión. [MOL, 00].

La Minería de Datos es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos. [FAY, et al, 96].

La Minería de Datos es conocida también como uno de los procesos núcleo del Descubrimiento de Conocimiento en Bases de datos (KDD). [CHE, 06].

Según [SAS, 02], la Minería de Datos es un conjunto de técnicas de análisis de datos que permiten:

- Extraer patrones, tendencias y regularidades para describir mejor los datos.
- Extraer patrones y tendencias para predecir comportamientos futuros.

La Minería de Datos representa la posibilidad de buscar información dentro de un conjunto de datos con la finalidad de extraer información nueva y útil que se encuentra oculta en grandes volúmenes de datos. Involucra un conjunto de técnicas de múltiples disciplinas tales como: tecnología de bases de datos, estadística, aprendizaje, reconocimiento de patrones, redes neuronales, visualización de datos, obtención de información, procesamiento de imágenes y de señales, y análisis de datos (Figura 2.3).

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

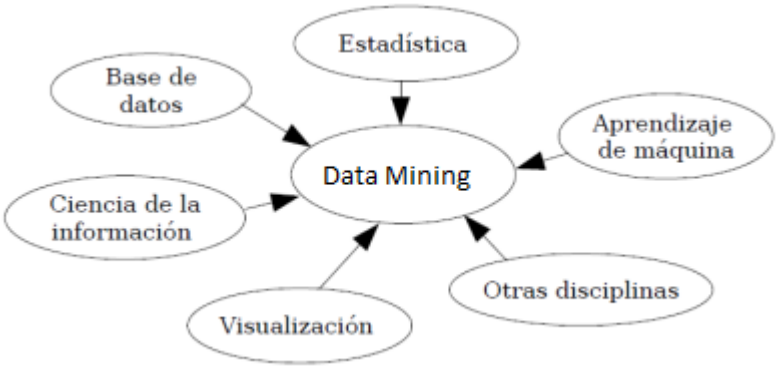


Figura 2. 3 Minería de Datos y conjunto de técnicas.

Fuente: [SAS, 02].

2.4.2. Componentes Básicos De la Minería de Datos.

Según [HER, 06]. Podemos encontrar 6 componentes básicos formando así una arquitectura de un sistema de minería clásico (Figura 2.4):

- Bases de datos, Datawarehouse o algún otro repositorio de información como por ejemplo planillas de cálculo.
- Servidores de bases de datos o de Datawarehouse que es el responsable de capturar los datos relevantes basándose en los requerimientos de minería de datos del usuario.
- Base de conocimientos que guiara la búsqueda o que evaluara el interés de los patrones resultantes. Dicho conocimiento puede incluir jerarquías de conceptos usadas para organizar valores de distintos atributos en diferentes niveles de abstracción

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

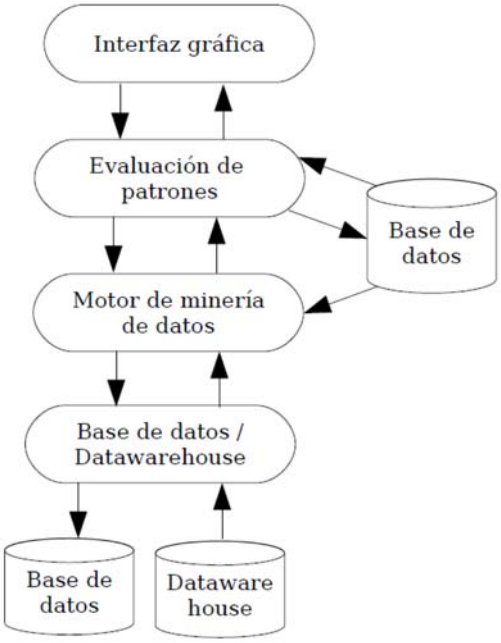


Figura 2. 4 Arquitectura típica de un sistema de minería de datos.

Fuente: [HER, 06].

- Motor de minería de datos, el cual es esencial para el sistema de minería y consiste de un conjunto de módulos funcionales que llevan a cabo distintas tareas tales como caracterización, asociación, clasificación, análisis de evolución y desviación.

- Módulo de evaluación de patrones, que generalmente utiliza medidas de interés que interactúa con los módulos de minería de datos para enfocar la búsqueda hacia patrones interesantes, así como también para filtrar o descartar patrones ya reconocidos. Alternativamente, el módulo de evaluación de patrones puede ser integrado al módulo de minería, dependiendo de la implementación del método de minería utilizado. Para una minería de datos eficiente, es altamente recomendado el incluir la evaluación de patrones de interés tanto como sea posible dentro del proceso para dirigir la búsqueda solo hacia los patrones de interés.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

- Interfaz gráfica de usuario; este módulo comunica a los usuarios y al sistema de minería de datos, permitiendo al usuario interactuar con el sistema especificando la consulta de la Minería de Datos, proveyendo información que ayude a enfocar la búsqueda y realizar exploración de datos basándose en resultados de minerías intermedias.

2.4.3. Características de la Minería de Datos.

Sus características son:

- Trabaja con grandes cantidades de datos: Volúmenes de datos tan grandes que tienen que ser analizados por técnicas automáticas. [APA, 98].

- Puede trabajar con datos incompletos y erróneos: Datos imprecisos es característica de toda corrección de datos. Las bases de datos usualmente están contaminadas por errores, no se puede asumir que los datos contenidos sean enteramente correctos. [APA, 98].

- Trabaja con datos heterogéneos almacenados. [APA, 98].

- Facilita el acceso a la información para que el usuario la analice más fácilmente.

- “Analiza” los datos.

- Este proceso consta de varias fases:
  - Preparación de Datos (selección, limpieza, y transformación), Minería de Datos, Evaluación, Difusión y Uso de Modelos.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

- Incorpora diferentes técnicas:
    - Árboles de decisión, regresión lineal, redes neuronales artificiales, máquinas de soporte vectorial, etc.
- Campos diversos:
    - Aprendizaje automático e IA (Inteligencia Artificial), estadística, bases de datos, etc.
- Aborda una tipología variada de problemas:
    - Clasificación, categorización, estimación/regresión, agrupamiento, entre otros.

2.4.4. Objetivos de la Minería de Datos.

Uno de los objetivos fundamentales de la Minería de Datos es poder predecir el valor de una variable predicativa o dependiente en función de los valores de otras variables llamadas independientes, existentes en una base de datos. Varios algoritmos realizan este tipo de tareas y pueden clasificarse como: de asociación, clasificación, Clustering, etc. [LAU, 05].

Otro objetivo principal de la Minería de Datos es la extracción de conocimiento de grandes bases de datos donde la dimensionalidad (número de variables), complejidad, o número de muestras es demasiado grande para un análisis manual. Está relacionado con campos como el análisis exploratorio de datos (exploratory Data analysis) y el descubrimiento de conocimiento en bases de datos (knowledge discovery in Databases). El objetivo de la exploración de datos es descubrir propiedades en los datos por medio de

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

medidas descriptivas (estadísticas de cada variable, entre ellas) o visualización. Se trata básicamente de llegar a una cierta "comprensión de los datos" y, de ahí, a comprender el proceso subyacente [CUA, 02].

Otros de los objetivos desde el punto de vista empresarial, es la integración de un conjunto de tareas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisiones. [BEN, et al, 02].

2.4.5. Técnicas de la Minería de Datos.

Las técnicas de la Minería de Datos son el resultado de un largo proceso de investigación y desarrollo de productos. Continuando con mejoras en el acceso a los datos, y ahora con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real.

Las técnicas de la Minería de Datos provienen de la Inteligencia artificial y de la estadística, dichas técnicas, son algoritmos sofisticados que se aplican sobre una cantidad determinada de datos para poder tomar una mejor decisión.

Las técnicas de Minería de Datos pueden dividirse en 2 categorías: predicativas (aprendizaje supervisado) y descriptivas (aprendizaje no supervisados) [MIT, et al, 03]. En la figura 2.5 se muestra la clasificación general de las categorías de minería de datos.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

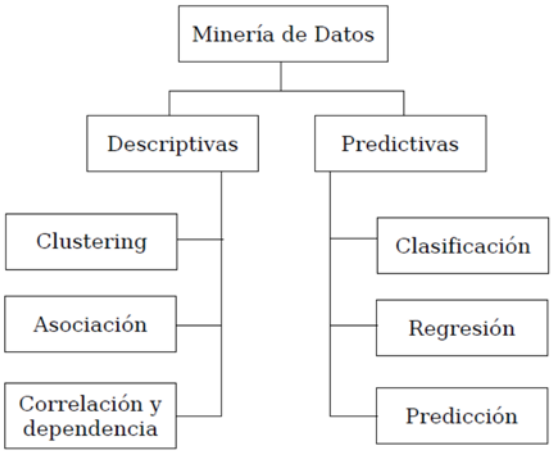


Figura 2. 5 Clasificación de las técnicas de minería de datos.

Fuente: [MIT, et al, 03].

2.4.5.1. Técnicas descriptivas.

En las tareas descriptivas, el conjunto de observaciones no tienen clases asociadas. El objetivo es derivar características (correlaciones, Clústeres, trayectorias, anomalías) que describan las relaciones entre los datos. Estas tareas son comúnmente de exploración natural y frecuentemente requieren de técnicas de post procesamiento para explicar los resultados.

Algunas de las tareas descriptivas son:

- **Clustering:** El agrupamiento/segmentación es la detección de grupos de individuos. Se diferencia de la clasificación en que en este caso, no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (Clústeres) diferenciados del resto. Ejemplo, agrupar los pacientes de un hospital a partir de su historial clínico.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

- **Asociaciones:** Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente, es relativamente alta. Ejemplo, en un supermercado se analiza si los pañales y la leche del bebe se compran conjuntamente.

- **Dependencias:** Una dependencia funcional (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro. Ejemplo: que un paciente haya sido ingresado en maternidad determina su sexo.

2.4.5.2. Tareas predictivas

En las tareas predicativas, cada observación incluye un valor de la clase a la que corresponde. El objetivo de estas tareas es predecir el valor de un atributo particular basado en los valores de otros atributos. El atributo a ser predicho se conoce como variable dependiente u objetivo (target), mientras que los atributos utilizados para realizar la predicción se llaman variables independientes o de exploración.

Algunas de las tareas predictivas son:

- **Clasificación**: Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas. Ejemplo: se sabe (por un estudio de dependencias) que los atributos edad, numero de miopías y astigmatismo han determinado los pacientes para los que su operación de cirugía ocular ha sido satisfactoria. Se pueden

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.

- **Regresión:** El objetivo es predecir los valores de una variable continua a partir de la evolución sobre otra variable continua, generalmente el tiempo. Ejemplo, se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costos, etc. a partir de los resultados de semanas, meses o años anteriores.

- **Predicción:** Encuentra una clasificación de valores faltantes o sin conocimiento previo. Se refiere tanto a la predicción de valores en los datos como a la predicción de clases utilizando la identificación de distribuciones en los datos disponibles. Ejemplo, predecir qué gobernador será elegido de acuerdo a las actuales encuestas electorales.

De estas tareas de minería de datos, aquellas que son comúnmente utilizadas son: la clasificación, el Clustering y la asociación.

A continuación se presenta una definición más formal de cada uno de ellos [TAN, et al, 06].

- **Clasificación:** Es el proceso de encontrar un conjunto de modelos (o funciones), los cuales describen y distinguen las clases definidas de los datos, con la finalidad de predecir clases de objetos cuyas clasificaciones no se han definido [HAN, et al, 00]. La clasificación requiere de un aprendizaje supervisado, es decir, se deben especificar los objetivos (clases) a los que se pretende llegar. El modelo se deriva principalmente del análisis

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

de un conjunto de datos de entrenamiento previamente clasificado.

- **Clustering:** Es una de las tareas del aprendizaje no supervisado en la que no se requiere una clasificación predefinida. El objetivo es particionar los datos obteniendo el conocimiento de acuerdo a las características de los mismos. En general, las clases de los datos no se presentan en el conjunto de datos y los objetos son agrupados basándose en el principio de maximización de similitud dentro de los Clústeres y minimización de similitud entre Clústeres diferentes [HAN, et al, 00].

- **Asociación:** Se basa en el descubrimiento de reglas de asociación demostrando condiciones en los valores de los atributos que ocurren simultáneamente de forma frecuente en un determinado conjunto de datos. La regla de asociación X) Y es interpretada como: los registros de la base de datos que satisfacen la condición X también satisfacen la condición Y [HAN, et al, 00].

2.4.6. Aplicaciones de la Minería de Datos.

Podemos mencionar algunos de los ejemplos de aplicaciones de la Minería de Datos: en mercados financieros, diagnósticos y tratamientos médicos, predicción de respuestas de clientes y la identificación automatizada de objetos de interés en grandes bases de datos de imágenes. Por ejemplo, para evaluar el riesgo crediticio se clasifican los aspirantes a un crédito como de riesgo bajo, medio o alto; en el diagnostico medico los pacientes se clasifican de acuerdo a los síntomas compadeciendo una determinada enfermedad, etc.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

La Minería de Datos es el descubrimiento automático de conocimiento utilizable procedente de los datos almacenados en el servidor y el empleo de tecnologías de reconocimiento de patrones para dar respuesta a preguntas de negocios como las siguientes:

- ¿Quiénes son los clientes más rentables de mi sitio Web?
- ¿Cómo aumentar mi cuota de mercado on-line?
- ¿Cómo optimizar mi inventario on-line?
- ¿Quiénes son los visitantes de mi sitio Web?

[BEN, et al, 02]. Respecto a los modelos inteligentes, se ha comprobado que en ellos se utilizan principalmente árboles y reglas de decisión, reglas de asociación, redes neuronales, redes bayesianas, conjuntos aproximados (rough sets), algoritmos de agrupación (Clustering), máquinas de soporte vectorial, algoritmos genéticos y lógica difusa.

- Sector Industria.
- Sector Farmacéutico y Sanitario.
- Administración Pública y Servicios.
- Sector Financiero y del Seguro.

2.4.7. Herramientas de la Minería de Datos.

Las herramientas de la Minería de Datos predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información. Los análisis prospectivos automatizados ofrecidos por un producto así van más allá de los eventos pasados provistos por herramientas retrospectivas típicas de sistemas de apoyo a la toma de decisión. Las herramientas de la Minería de Datos pueden responder a preguntas de negocios que tradicionalmente consumen

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

demasiado tiempo para poder ser resueltas y a los cuales los usuarios de esta información casi no están dispuestos a aceptar. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas. [BEN, et al, 02]

Las herramientas de la Minería de Datos se pueden clasificar en dos grandes grupos, hay dos tipos de herramientas de análisis de la Minería de Datos: descubrimiento y verificación. Ambas son necesarias para el proceso de la minería de datos. [APA, 98].

2.4.7.1. Herramientas de verificación.

Son aquellas a las que el sistema se limita a comprobar hipótesis suministradas por el usuario. No obstante, aún cuando hay muchos algoritmos interesantes nuevos en el bando del descubrimiento, las estadísticas son consideradas de valor poderoso en el bando de la verificación.

Mientras las herramientas del descubrimiento pueden ayudar a hacer descubrimientos interesantes, no pueden explicar el porqué, o si verdaderamente, el descubrimiento es válido. En la mayoría de herramientas del descubrimiento, se lleva a cabo una serie de pruebas para buscar diferencias entre grupos; a menudo estas pruebas permiten un número seguro de hallazgos incorrectos debido a su naturaleza probabilística.

Las herramientas de verificación pueden ayudar a un minero de datos a validar hallazgos desde el descubrimiento hasta tomar decisiones de negocios correctas. Aquí hay una lista de algunas herramientas de comprobación:

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

- Regresión lineal
- Regresión logística
- Análisis del discriminante
- Métodos de pronósticos
- Correlaciones.

2.4.7.2. Herramientas de descubrimiento.

Son aquellas herramientas en las que se encuentran patrones potencialmente interesantes de forma automática, incluyendo en este grupo todas las técnicas de predicción.

La mayoría de algoritmos de las herramientas del descubrimiento emergen de la búsqueda de la inteligencia artificial, desarrollando algoritmos de computadora para simular la actividad del cerebro humano. Para hallar patrones y tendencias en datos, las herramientas del descubrimiento pueden ayudar a las compañías a descubrir fenómenos del mercado y de este modo aumentar la base de conocimiento de la compañía.

Este conocimiento nuevo puede luego ser usado para ayudar a la compañía a ganar una ventaja competitiva y/o crecer su mercado. Aquí hay una lista de algunas herramientas del descubrimiento:

- Visualización de Datos.
- Árboles de decisión.
- Redes Neuronales.
- Algoritmos Genéticos.
- Reglas de Inducción.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

- Reglas de Asociación.

2.5. CLUSTERING.

El análisis de Clúster, o simplemente Clustering, consiste en particionar un conjunto de datos o patrones de manera que aquellos pares de patrones similares entre sí queden en el mismo grupo, mientras que los pares de patrones que no son similares queden en grupos distintos. El análisis de Clústeres suele referirse también como clasificación no supervisada, ya que en términos generales consiste en un problema de clasificación (se debe asignar una clase o etiqueta a cada patrón), sólo que en la ejecución del algoritmo no se dispone de etiquetas reales que deban ser aprendidas. El aprendizaje de las clases debe hacerse utilizando sólo los patrones mismos y medidos de similitud entre ellos. De ahí el concepto de aprendizaje no supervisado.

El Clustering es una de las tareas mentales más primitivas realizadas por los seres humanos, que la usan para manejar las inmensas cantidades de información que reciben a diario. Procesar cada pedazo de información como una entidad independiente sería imposible. Por lo tanto, los seres humanos tienden a categorizar las entidades (objetos, personas, eventos) en Clústeres como se muestra en la Figura 2.6.

El Clustering representa la división de datos en grupos de objetos similares llamados Clústeres [MIT, et al, 03].

Clustering es también conocido como clasificación no supervisada, en donde no se tienen asignación de grupos a clases ya predefinidas, sino que los grupos se van creando de acuerdo a las características de los datos.

Los grupos o Clústeres, son un conjunto de objetos que comparten características similares y juegan un papel muy importante en la manera en cómo la gente analiza y describe el mundo que los rodea. De forma natural, el humano se encarga de dividir

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

objetos en grupos (Clustering) y asignar objetos particulares a dichos grupos (clasificación).

Clustering es una de las técnicas más útiles para descubrir conocimiento oculto en un conjunto de datos. En la actualidad el análisis de Clustering en minería de datos ha jugado un rol muy importante en una amplia variedad de áreas tales como: reconocimiento de patrones, análisis de datos espaciales, procesamiento de imágenes, cómputo y multimedia, análisis médico, economía, Bioinformática y biometría principalmente [HAN, et al, 00].

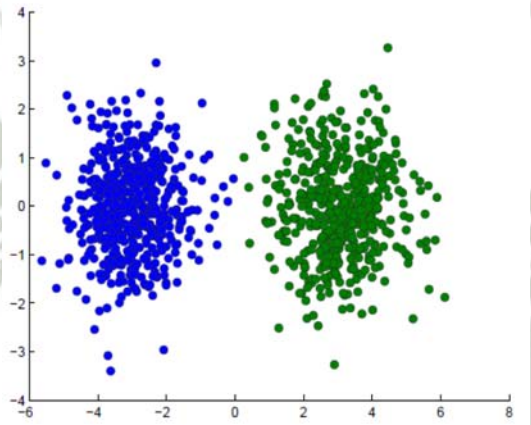


Figura 2. 6 Ejemplo de Clustering.

Fuente: [HAN, et al, 00].

2.5.1. Análisis de Clustering.

Un problema de análisis de Clustering, parte de un conjunto de casos u objetos cada uno de los cuales está caracterizado por varias variables (ver tabla 2.1). A partir de dicha información se trata de obtener grupos de objetos, de tal manera que los objetos que pertenecen a un grupo sean muy homogéneos entre si y, por otra parte, la heterogeneidad entre los distintos grupos sea muy elevada. Expresado en términos de variabilidad hablaríamos de minimizar la variabilidad dentro de los grupos para al mismo tiempo maximizar la variabilidad entre los distintos grupos.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

	C_1	...	C_i	...	C_k
O_1	X_1^1	...	X_i^1	...	X_k^1
.....					
O_j	X_1^j	...	X_i^j	...	X_k^j
.....					
O_n	X_1^N	...	X_i^N	...	X_k^N

Tabla 2. 1 Casos de partida para el análisis de Clustering.

Fuente: [HER, 06].

A partir del planteamiento de un problema de Clustering, las actividades del análisis de Clustering típicamente involucran los siguientes pasos [JAI, et al, 99]:

- Representación de patrones. Se refiere al establecimiento del número de clases, número de patrones, y el número, tipo y tamaño de las características disponibles para el algoritmo de Clustering.
- Definición de proximidad. La proximidad de los patrones es usualmente medida por una función distancia entre un par de datos.
- Clustering. La etapa de agrupamiento puede desarrollarse en un gran número de formas. Se pueden utilizar agrupamientos de Clústeres jerárquicos, particionales y otros más abarcan métodos probabilísticos o de teoría de grafos.
- Abstracción de datos. Es el proceso de extraer una representación simple y compacta del conjunto de datos.
- Verificación de resultados. Consiste en validar el análisis de Clustering realizado evaluando los resultados obtenidos.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

2.5.2. Características de los algoritmos de Clustering.

Las características deseables de la mayoría de los algoritmos de Clustering son las siguientes:

- Escalabilidad. La mayoría de los algoritmos de Clustering trabajan de manera apropiada con un número pequeño de observaciones (hasta 200 aproximadamente), mientras que se necesita una gran escalabilidad para realizar agrupamiento de datos en bases con millones de observaciones.
- Habilidad para trabajar con distintos tipos de atributos. Muchos algoritmos se han diseñado para trabajar sólo con datos numéricos, mientras que en una gran cantidad de ocasiones, es necesario trabajar con atributos asociados a tipos numéricos, binarios, discretos y alfanuméricos.
- Descubrimiento de Clústeres con formas arbitrarias. La mayoría de los algoritmos de Clustering se basan en la distancia euclidiana, lo que tiende a encontrar Clústeres todos con forma (circular) y densidad similares. Es importante diseñar algoritmos que puedan establecer Clústeres de formas arbitrarias.
- Requerimientos mínimos en el conocimiento del dominio para determinar los parámetros de entrada. La herramienta no debería solicitarle al usuario que introduzca la cantidad de clases que quiere considerar, ya que dichos parámetros en muchas ocasiones no son fáciles de determinar, y esto haría que sea difícil controlar la calidad del algoritmo.
- Habilidad para tratar con datos ruidosos. La mayoría de las Bases de Datos contienen datos con comportamiento extraño, datos faltantes, desconocidos o erróneos. Algunos algoritmos de Clustering son sensibles a tales datos y pueden derivarlos a Clústeres de baja calidad.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

- Insensibilidad al orden de las observaciones de entrada. Algunos algoritmos son sensibles al orden en que se consideran las observaciones. Por ejemplo, para un mismo conjunto de datos, dependiendo del orden en que se analicen, los Clústeres devueltos pueden ser diferentes. Es importante entonces que el algoritmo sea insensible al orden de los datos, y que el conjunto de Clústeres devuelto sea siempre el mismo.

- Alta dimensionalidad. Una Base de Datos puede contener varias dimensiones o atributos, por lo que es bueno que un algoritmo de Clustering pueda trabajar de manera eficiente y correcta no sólo en repositorios con pocos atributos, sino también en repositorios con un alto espacio dimensional, o gran cantidad de atributos.

- Clustering basado en restricciones. Es un gran desafío el agrupar los datos teniendo en cuenta no sólo el comportamiento, sino también que satisfagan ciertas restricciones. Interpretación y uso. Los usuarios esperan que los resultados del Clustering sean comprensibles, fáciles de interpretar y de utilizar.

Con estas características, se busca diseñar algoritmos más flexibles que sean capaces de manipular una gran variedad de requerimientos de acuerdo a las necesidades de los usuarios.

2.5.3. Técnicas de Clustering.

Los algoritmos de agrupación de Clustering varían entre sí por las reglas heurísticas que utilizan y el tipo de aplicación para el cual fueron diseñados. La mayoría de ellos se basa en el empleo sistemático de distancias entre vectores (objetos a agrupar) así como entre Clústeres o grupos que se van formando a lo largo del proceso de Clustering. Las características básicas por las que los algoritmos de Clustering pueden ser clasificados son en función de:

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

- El tipo de dato que manejan (numérico, categórico y/o mixto).
- El criterio utilizado para medir la similitud entre los puntos.
- Los conceptos y técnicas de Clustering empleadas (ej. lógica difusa, estadísticas).

En la literatura existen una gran cantidad de técnicas de Clustering que varían de acuerdo a la arquitectura que utilizan [JAI, et al, 99]. Una clasificación general divide los algoritmos en: Clustering particional, Clustering jerárquico, Clustering basado en densidad y Clustering basado en Grid.

Para cada una de las categorías presentadas en la Figura 2.7, existen una variedad de sub-clasificaciones que presentan algoritmos con diferentes técnicas para encontrar clústeres en los datos. Algunas de ellas son: Técnicas estadísticas, basadas en la utilización de medidas de similitud y análisis estadístico para agrupar los datos.

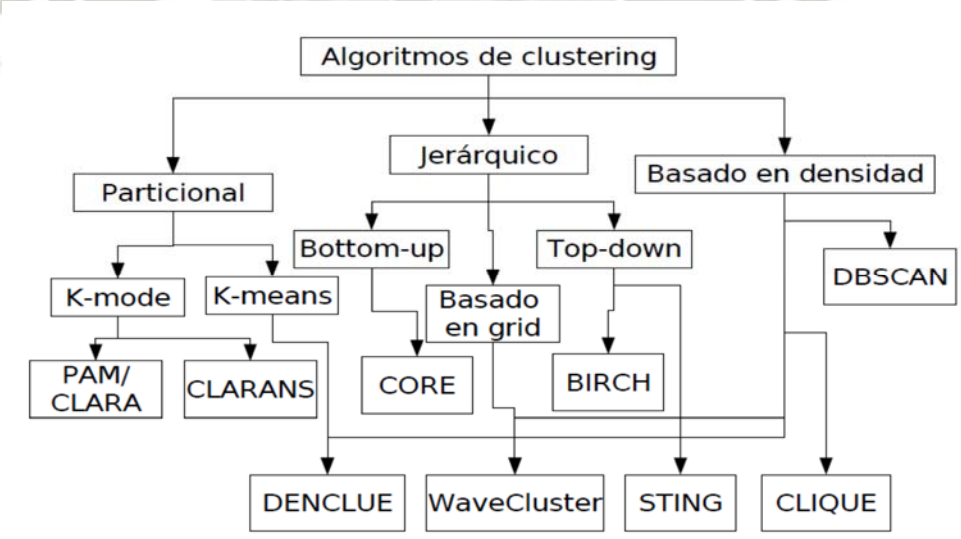


Figura 2. 7 Algoritmos básicos de Clustering.

Fuente: [HER, 06].

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

A continuación se explican brevemente, los cuatro principales tipos de algoritmos de Clustering.

2.5.3.1. Clustering particional.

Un algoritmo de Clustering particional obtiene una partición simple de los datos en vez de la obtención de la estructura del clúster tal como se produce con los dendogramas de la técnica jerárquica [JAI, et al, 99]. En la Figura 2.8 se muestra un ejemplo de Clustering particional.

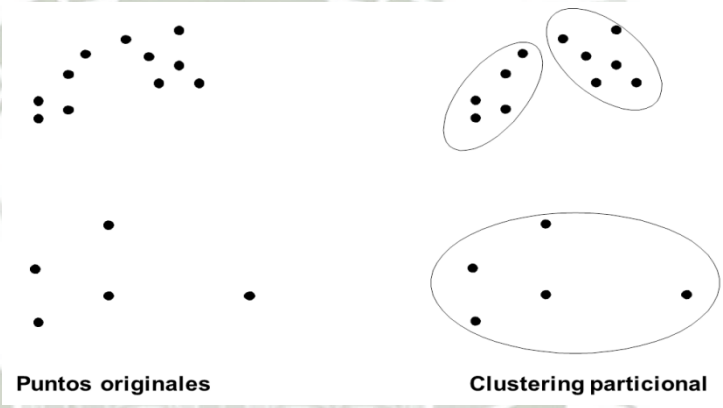


Figura 2. 8 Clustering particional.

Fuente: [JAI, et al, 99].

El Clustering particional organiza los objetos dentro de k clústeres de tal forma que sea minimizada la desviación total de cada objeto desde el centro de su clúster o desde una distribución de Clústeres. La desviación de un punto puede ser evaluada en forma diferente según el algoritmo, y es llamada generalmente función de similitud.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Los métodos particionales tienen ventajas en aplicaciones que involucran gran cantidad de datos para los cuales la construcción de un dendograma resultaría complicada.

El problema que se presenta al utilizar algoritmos particionales es la decisión del número deseado de clústeres de salida. Las técnicas particionales usualmente producen clústeres que optimizan el criterio de función definido local o globalmente. En la práctica, el algoritmo se ejecuta múltiples veces con diferentes estados de inicio y la mejor configuración que se obtenga es la que se utiliza como el Clustering de salida. Algunos algoritmos de Clustering que pertenecen a esta clasificación son:

- **PAM:** Partitioning Around Medoids [KAU, et al, 00] Para encontrar k Clústeres (agrupamientos), el modelo PAM determina un objeto representativo para cada Cluster. Este objeto representativo, llamado medoid, es el que se encuentra localizado más al centro dentro del Clúster. Una vez que los medoids han sido seleccionados, cada objeto no seleccionado es agrupado con el medoid al cual es más similar. Para encontrar los k medoids, PAM comienza con una selección arbitraria de k objetos. En cada iteración, un intercambio entre un objeto seleccionado Oi y un objeto no seleccionado Oh es realizado, si y solo sí, el intercambio resulta en un incremento de la calidad del agrupamiento (Clustering).

- **CLARA:** El algoritmo Clustering Large Applications [KAU, et al, 00] dibuja múltiples muestras de conjunto de datos aplicando PAM en cada muestra para encontrar la mediana, si la muestra es representativa las medianas de la muestra deben aproximarse a las medianas del conjunto de datos entero. Para mejorar la

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

aproximación múltiples muestras se dibujan y el mejor Clustering se retorna. El Clustering alcanzado es medido por la desemejanza promedio de todos los objetos en el conjunto de datos.

- **CLARANS:** El algoritmo Clustering Large Applications based upon RANdomized Search [NGH, et al, 94] puede ser presentado como una búsqueda en un gráfico donde cada nodo es una solución potencial, es decir, un conjunto de k medianas. Dos nodos son vecinos si sus conjuntos difieren sólo en una mediana. A cada nodo se le asigna un costo para definir la desemejanza total entre cada objeto y la mediana de su Cluster. El problema consiste en la búsqueda de un mínimo en el gráfico. En cada paso todos los vecinos del actual nodo son buscados; el vecino que corresponde al descenso más profundo en el costo es elegido como la próxima solución. Se usa PAM para la búsqueda total en el gráfico y CLARA busca en algún sub gráfico aleatorio.

- **K-MEANS:** El algoritmo k-means [JAI, et al, 03] selecciona k elementos aleatoriamente, los cuales representan el centro o media de cada Cluster. A cada objeto restante se le asigna el Cluster con el cual más se parece, basándose en una distancia entre el objeto y la media del Cluster. Después calcula la nueva media del Cluster e itera hasta no cambiar de medias.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

2.5.3.2. Algoritmos Jerárquicos.

Un método de Clustering jerárquico produce una serie de divisiones anidadas de los datos. El diagrama que representa esta clasificación jerárquica se denomina dendrograma. Un ejemplo de dendrograma se muestra en la Figura 2.9. Las líneas representan las separaciones encontradas por el algoritmo. En el nivel superior, los datos se separan en dos grandes clústeres. A medida que se desciende por el dendrograma, se obtienen particiones más finas de los datos, llegando hasta el nivel individual, donde cada patrón forma su propio Clúster. Los largos de las líneas representan las distancias entre los pares de clústeres. Los métodos jerárquicos son más apropiados que los particionales para datos con características nominales u ordinales [ZHA, 04].

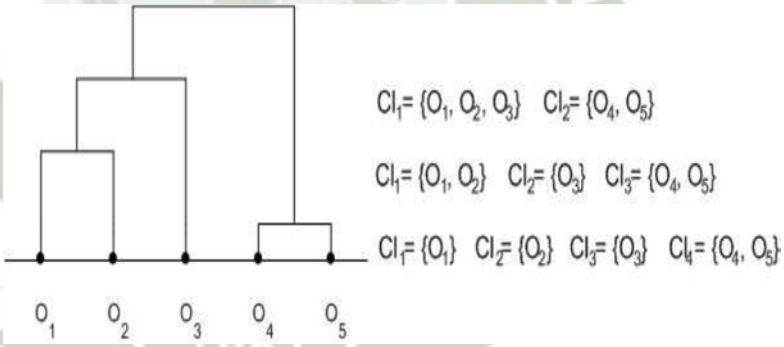


Figura 2. 9 Dendograma resultado de la clasificación ascendente jerárquica.

Fuente: [ZHA, 04].

Algunos algoritmos de Clustering que pertenecen a esta clasificación son:

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

- **Bottom-up**, también llamado aglomerativo, se comienza con cada objeto formando un grupo por separado. Los objetos o grupos se combinan sucesivamente según determinadas medidas, hasta que todos los grupos se hayan unido en uno solo, o hasta que se cumpla alguna condición de terminación.

- **Top-down**, también llamado divisivo, se comienza con todos los objetos en el mismo Clúster, y a medida que se va iterando, se dividen los grupos en subconjuntos más pequeños según determinadas medidas, hasta que cada objeto se encuentre en un clúster individual o hasta que se cumplan las condiciones de terminación.

- **CURE**, El algoritmo Clustering Use Representatives [SUD, et al, 98] también trabaja con Clúster de formas esféricas. Usa s números fijos de puntos representativos para evaluar la distancia entre los Clúster. Los puntos representativos se eligen con una primera selección de objetos bien dispersos en el Clúster y luego se contrae hacia el centro del Clúster. En cada paso, los dos Clúster con el par más cercano de puntos representativos se combinan.

- **CHAMELON**, El algoritmo Hierarchical Clustering Algorithm Using Dynamic Modeling [KAR, et al, 99] desarrollado con el objetivo de superar las limitaciones de la mayoría de los algoritmos de Clustering, como: el uso de modelo estático para los datos, el uso de diversas formas de Clúster, entre otras. CHAMELEON mide la semejanza basado en un modelo dinámico: dos Clúster se combinan sólo si la inter conectividad y proximidad entre ellos es relativamente alta para la inter

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

conectividad interna de los Clúster y proximidad de elementos dentro del Clúster.

- **ROCK**, El algoritmo Clustering Categorical Data [SUD, et al, 98] maneja datos categóricos y booleanos de tal manera que une en el mismo grupo los vecinos más cercanos según una función de similitud.

2.5.3.3. Algoritmos Basados En Grillas (Grid-Based Algorithms).

Los métodos basados en densidad suelen tener grandes problemas cuando se trabaja con bases de datos muy grandes. Para mejorar la efectividad del Clustering, un método basado en grillas usa una estructura de grilla de datos. El método divide el espacio en un número finito de celdas, formando una grilla, en donde se realizan todas las operaciones del Clustering. La mayor ventaja de este método es su veloz procesamiento del tiempo, el cual generalmente es independiente de la cantidad de objetos a procesar. Ver Figura 2.10.

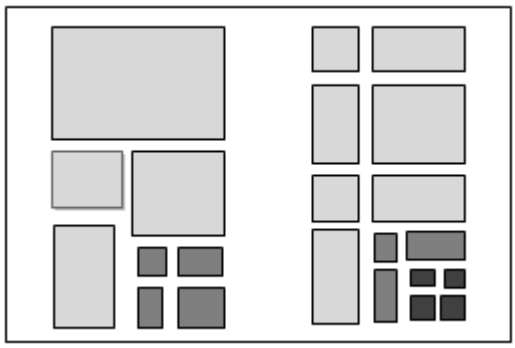


Figura 2. 10. Clustering Basado en Grid.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Algunos algoritmos de Clustering que pertenecen a esta clasificación son:

- **STING**: El algoritmo Statical Information Grid approach [WEI, et al, 97], señala que el área espacial está dividido en celdas rectangulares, donde se tiene diferentes niveles de tales celdas rectangulares correspondientes a diferentes resoluciones y estas celdas forman una estructura jerárquica. Cada celda en un nivel alto se particiona para formar un número de celdas del próximo nivel bajo. La información estadística de cada celda se calcula y almacena de antemano y se usa para responder consultas.

- **WaveCluster**: El algoritmo WaveCluster: WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases [SHE, et al, 98], está basado en métodos de rejilla y densidad e indicando para grandes bases de datos. Es una aproximación a Clustering multi-resolusión la cual aplica transformaciones wavelet a las características del espacio. Una transformación wavelet es una técnica de señal de procesamiento que descompone una señal en diferentes frecuencias de sub-bandas.

- **CLIQUE** El algoritmo Clustering In Quest, identifica automáticamente sub-espacios de un espacio de datos de alta dimensión que permite mejor Clustering que el original. Está basado en métodos de rejilla y densidad de tal forma que divide cada dimensión en el mismo número de intervalo de igual longitud, divide un espacio de datos m-dimensional en unidades rectangulares no solapadas. Una unidad es densa si la fracción de los puntos de datos total contiene en la unidad exceso de entrada de parámetros del modelo. Un Clúster es un conjunto

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

maximal de unidades densas conectadas dentro de un sub-espacio.

2.5.3.4. Algoritmos Basados En Densidad (Density-Based Algorithms).

La mayoría de los métodos de particionamiento, realizan el proceso de Clustering en base a la distancia entre dos objetos. Estos métodos pueden encontrar sólo Clústeres esféricos, y se les dificulta hallar Clústeres de diversas figuras. Otros algoritmos de Clustering han sido desarrollados en base a la noción de "densidad". Estos generalmente estiman Clústeres como regiones con gran densidad de objetos, separados de regiones de baja densidad de objetos (estos elementos aislados representan ruido). Este tipo de métodos es muy útil para filtrar ruido y encontrar Clústeres de diversas formas. En la Figura 2.12 se muestra un ejemplo de Clustering basado en densidad.

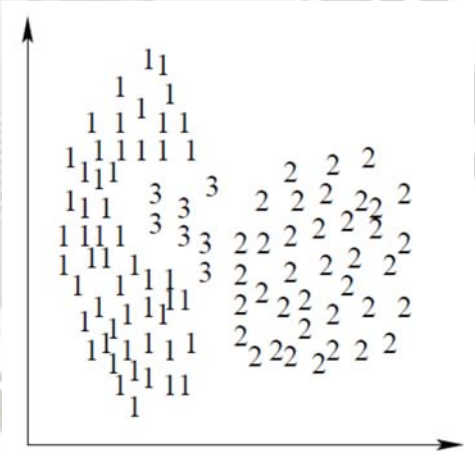


Figura 2. 11. Clustering basado en densidad.

Fuente: PROPIO.

Este tipo de métodos es muy útil para filtrar ruido y encontrar Clústers de diversas formas. La mayoría de los métodos de particionamiento, realizan el proceso de Clustering con base en la distancia entre dos objetos [JAI, et al, 99]. Estos métodos pueden

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

encontrar sólo Clústers esféricos y se les dificulta hallar clúster de formas diversas.

Algunos algoritmos de Clustering que pertenecen a esta clasificación son:

- **DBSCAN**, El algoritmo Density-Based algorithm for discovering Clústeres in large spatial Databases with Noise [EST, et al, 96] donde un grupo se define como un conjunto máximal de puntos conectados desde el punto de vista de la densidad, de esta forma descubre grupos de forma aleatoria en bases de datos con ruido.

  El algoritmo señala que se elige arbitrariamente un punto p, recupera todos los puntos densos alcanzables dentro de la distancia y el número mínimo permitido, si p es un punto central se forma un Clúster, si p es un punto borde ningún punto es alcanzable desde p y DBSCAN visita el próximo punto de la base de datos, y así continua hasta que todos los puntos han sido procesados.

- **OPTICS**, El algoritmo Ordering Points To Identify the Clustering Structure [ANK, et al, 99] ordena los puntos para identificar la estructura Clustering, agrupa los puntos por densidad de conectividad (jerarquía de Clúster), visualiza los Clúster y su jerarquía.

- **DENCLUE,** El algoritmo Desity based Clustering, está basado en fundamentos sólidos matemáticos, la idea en la que se basa es modelar la media de la densidad del espacio de datos como

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE
TESIS UCSM

CATÓLICA
DE SANTA MARÍA

la suma de funciones de influencia de todos los puntos del mismo. Permite una descripción matemática compacta de formas arbitrarias de Clúster en conjunto de datos de alta dimensión.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

CAPÍTULO III

MARCO APLICATIVO.

El presente documento tiene como objetivo la explicación de forma detallada y explícita los pasos para realizar en Minería de Datos.

En la primera sección de este capítulo se describe la definición de requerimientos y desarrollo de la aplicación, herramientas utilizadas para análisis, diseño y construcción del mismo, en la segunda sección se muestra el proceso de recuperación de datos con la metodología propuesta, en la tercera sección se detalla la implementación del aplicativo su ejecución y respectivo análisis de resultados. La integración se llevará acabo utilizando ciertas librerías y programas extras que se mencionan en los Requerimientos los cuales pueden variar según las necesidades de cada instalación.

3.1. SOFTWARE A UTILIZAR.

Se describe el software que se emplea durante el ciclo de vida de desarrollo del presente proyecto. Comenzando por el modelado de datos, la elección del repositorio usado como base de datos, extracción y explotación de los datos y finalmente la presentación de los datos trabajados. A continuación se describe el uso de cada una de las herramientas utilizadas.

3.1.1. PENTAHO.

Para el proceso de ETL se utilizara el Pentaho que es un conjunto de programas libres para generar resultados acorde a lo de inteligencia de negocios. Incluye herramientas integradas para generar informes, minería de datos, ETL, etc.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Las soluciones que Pentaho ofrece se componen fundamentalmente de una infraestructura de herramientas de análisis e informes y un motor de Workflow de procesos de negocios. Esta plataforma está especialmente diseñada para ejecutar las reglas de negocios, expresadas en forma de procesos y actividades y de presentar la información apropiada en el momento preciso (Figura 3.1).

Aunque es una plataforma de libre distribución, su modelo de ingresos parece estar orientado a los servicios: soporte, formación, consultoría y distribuciones. No obstante en la documentación de la Web oficial(http://www.pentaho.org/) de Pentaho hace referencia a algunas funcionalidades Premium que hace pensar en futuras versiones.

A parte de ser open source y sin costes de licencia, las características básicas de esta herramienta son:

- Entorno gráfico de desarrollo.
- Uso de tecnologías estándar: Java, XML, Java Script.
- Fácil de instalar y configurar.
- Multiplataforma: Windows, Macintosh, Linux.
- Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).
- PDI está formado por un conjunto de herramientas, cada una con un propósito específico.
- Soporta modo de ejecución, pre visualización con resultados parciales y repetición.
- Exploración del repositorio de datos (tablas, vistas.) y metadatos.
- Asistente para la creación de conexiones a base de datos.
- Posibilidad de verificación y análisis del impacto de las transformaciones y trabajos.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

- Admite archivos de texto, tablas de bases de datos, información del sistema, los nombres de archivos de un directorio, una entrada de XBse, archivos XML o archivos Excel como entradas de las transformaciones y trabajos.
- La salida de una transformación o trabajo puede ser una tabla de la base de datos a la cual se ha definido la conexión, un archivo de texto, un archivo XML, un archivo Excel o un archivo Access.
- Soporta muchos tipos de transformaciones básicas: mapeo de campos y valores, filtrado de filas, ordenamiento, secuencias, participación de campos, agrupación, adicción, de constantes, normalización/ des normalización de filas, uniones (join) de filas, fusión de filas y algunas operaciones matemáticas. Para operaciones más avanzadas se pueden recurrir al Scripting que usa código Java Script para definir las transformaciones.
- Ejecución de procedimientos almacenados.
- Llamadas a servicios Web mediante una URL con parámetros dinámicos, obtención de archivos vía FTP y SFTP y envío de mails.
- Las transformaciones pueden ser llamadas por otras transformaciones y por otros trabajos y los trabajos pueden ser llamados por otros trabajos. Por ello hay mecanismos para pasar la información entre transformaciones/trabajos y definir variables.
- Mecanismos de control de flujo y bucle.
- Ejecución de Shell Script y comprobación de la existencia de archivos y tablas.
- Definición del intervalo de ejecución en el planificador de trabajos.
- Importación y exportación de los metadatos de la transformación o del trabajo a archivos XML.
- Soporta Computer Clustering (Procesamiento Distribuido de Datos).
- Log minucioso con diferentes niveles de precisión.

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

Estos son las herramientas de las que se compone Kettle:

- Spoon: es la herramienta gráfica que nos permite el diseño de las transformaciones y trabajos. Incluye opciones para pre visualizar y testear los elementos desarrollados. Es la principal herramienta de trabajo de PDI y con la que construiremos y validaremos nuestros procesos ETL.
- Pan: es la herramienta que nos permite la ejecución de las transformaciones diseñadas en spoon (bien desde un fichero o desde el repositorio). Nos permite desde la línea de comandos preparar la ejecución mediante scripts.
- Kitchen: similar a Pan, pero para ejecutar los trabajos o jobs.
- Carte: es un pequeño servidor web que permite la ejecución remota de transformaciones y jobs.

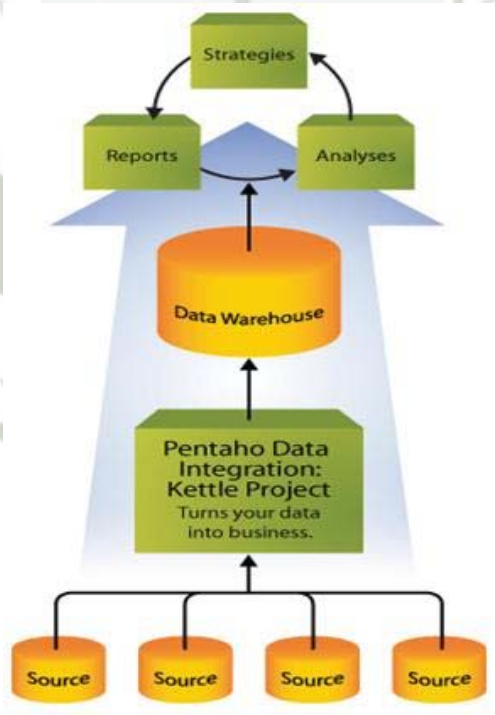


Figura 3. 1 Visualización del esquema de Pentaho Data Integration.

Fuente: [WWW3].

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

Bases de datos soportadas:

- Oracle (Nativo, ODBC, OCI).
- MySQL (Nativo, ODBC).
- AS/400 (Nativo).
- MS Access (Nativo).
- MySQL Server (Nativo, ODBC).
- IBM DB2 (Nativo, ODBC).
- PostgreSQL (Nativo, ODBC).
- Intersystems Cache (Nativo, ODBC).
- Sybase (Nativo, ODBC).
- Gupta SQL Base (Nativo, ODBC).
- Dbase III, IV y V (ODBC).
- Firebird SQL (Nativo, ODBC).
- Hypersonic (Nativo).
- maxDB – SAP DB (Nativo, ODBC).
- CA Ingres (Nativo, ODBC).
- BorlandInterbase (Nativo, ODBC).
- ExtenDB (Nativo, ODBC).
- Génerico (Nativo, ODBC).

Los requisitos generales de esta herramienta son:

- Java RuntimeEnvironment (JRE) 1.4 o superior
- Sistema operativo:
  - ✓ Microsoft Windows 95, 98,Me, 2000,XP, 2003, Vista.
  - ✓ Linux GTK:32 y 64 bits (funciona mejor en Gnome).
  - ✓ Apple OSX: funciona en con PowerPC e Intel.
  - ✓ Solaris: usando la interfaz Motif (GTK opcional).
  - ✓ AIX: usando la interfaz Motif.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

- ✓ HP-UX: usando la interfaz Motif (GTK opcional).
- ✓ FreeBSD : Funciona solo con 32 bits.

3.1.2. WEKA.

Para la segmentación y clasificación de los datos utilizaremos el WEKA (Waikato Environment for Knowledge Analysis) que es una herramienta que permite la experimentación de análisis de datos mediante la aplicación, análisis y evaluación de las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario.

Contamos con 4 diferentes interfaces en WEKA:

La primera es la Interfaz Explorer el cual nos permite realizar operaciones sobre un solo archivo de datos. En esta interfaz vemos tareas de pre-procesado, clasificación, Clustering, búsqueda de asociaciones, selección de atributos y visualización de los datos.

La segunda es la Interfaz Experimenter, nos permite fijar experimentos a gran escala, correrlos, dejarlos y analizar el funcionamiento estadístico que se ha recogido, el resultado puede ser almacenado en formato ARFF. La Interfaz Experimenter supera las limitaciones de tiempo, o que quiere decir que trabaja de manera eficiente.

La Interfaz KonwledgeFlow muestra de una manera más explícita el funcionamiento interno del programa. Se basa en situar en el panel de trabajo elementos base, a esto le llamamos un funcionamiento grafico.

La Interfaz SimpleCLI proviene de la abreviación de Simple Client. Podemos introducir comandos en una consola proporcionada por esta interfaz. Esta

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

interfaz que parece tan sencilla viene a ser complicada ya que es necesario un conocimiento completo de la aplicación. Por ende actualmente solo es útil como herramienta de ayuda a la fase de pruebas ver Figura 3.2. Los mandatos soportados en este interfaz son:

java <nombre-de-la-clase><args>

Ejecuta el método "main" de la clase dada con los argumentos especificados al ejecutar este mandato. Cada mandato es atendido en un hilo independiente por lo que es posible ejecutar varias tareas concurrentemente.

- **break:** Detiene la tarea actual.
- **kill:** "Mata" la tarea actual. Este comando es desaconsejable, solo debe usarse en el caso de que la tarea no pueda ser detenida con el mandato break.
- **cls**: (Clear Screen). Limpia el contenido de la consola.
- **exit**: Sale de la aplicación.
- **help:** <mandato> Proporciona una breve descripción de cada mandato.

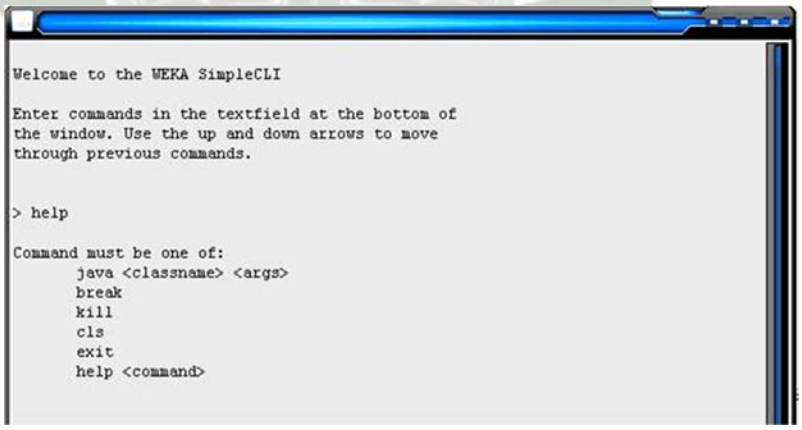


Figura 3.2 Interfaz SimpleCLI.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

WEKA se distribuye como software de libre distribución desarrollado en Java. Está constituido por una serie de paquetes de código abierto con diferentes técnicas de pre procesado, clasificación, agrupamiento, asociación, y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. Estos paquetes pueden ser integrados en cualquier proyecto de análisis de datos, e incluso pueden extenderse con contribuciones de los usuarios que desarrollen nuevos algoritmos. Con objeto de facilitar su uso por un mayor número de usuarios, WEKA además incluye una interfaz gráfica de usuario para acceder y configurar las diferentes herramientas integradas.

Los requisitos para la instalación:

- **Versiones:** Las últimas versiones estables (3.4 y 3.6) se pueden descargar del sitio web: http://www.cs.waikato.ac.nz/ml/weka/ También está disponible en la página web, la versión de desarrollo (3.7 hasta el momento), sobre la cual van corrigiendo bugs y añadiendo nuevas funcionalidades.

- **Sistemas Operativos:** Windows (arquitecturas x86 y x64), Mac OSX y distribuciones Linux.

- **Requerimientos:** Se debe contener el JRE (Java Runtime Environment) sobre la cual se ejecuta.

3.2. METODOLOGIA

La Metodología que aplicaremos para desarrollar nuestro proyecto es CRISP-DM. Esta metodología se formó en el año 1997 gracias al apoyo de la comisión europea con el objetivo de lograr una herramienta para la industria a fin de que se adapte a la

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

diversidad de industrias. CRISP-DM es un modelo basado en situaciones que ocurren en las empresas.

Percibe diferentes objetivos entre los principales encontramos:

- Basándonos en un proceso jerárquico nos permite aprender nuevas técnicas para aplicar y comprender de mejor manera la Minería de Datos.
- Persigue el cumplimiento de objetivos desde un punto de vista empresarial, por ende da preferencia a la comprensión del negocio.
- Desarrollar proyectos de Minería de Datos mediante un proceso estandarizado.
- Minimiza los costos que implica un proyecto de minería de datos en las empresas.

3.2.1. EL MODELO CRISP-DM.

La metodología CRISP-DM provee una representación completa del ciclo de vida de un proyecto de Minería de Datos, que se divide en seis fases, sus tareas y relaciones entre ellas ver Figuras 3.3 y 3.4.

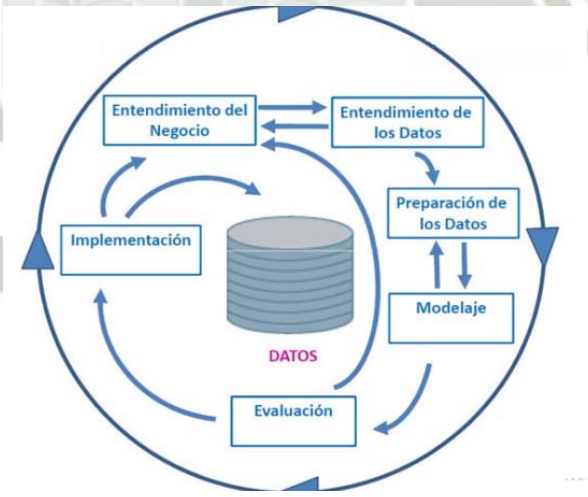


Figura 3. 3 Ciclo de Vida del Proyecto de la Metodología CRISP-DM

Fuente:[WWW4].

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

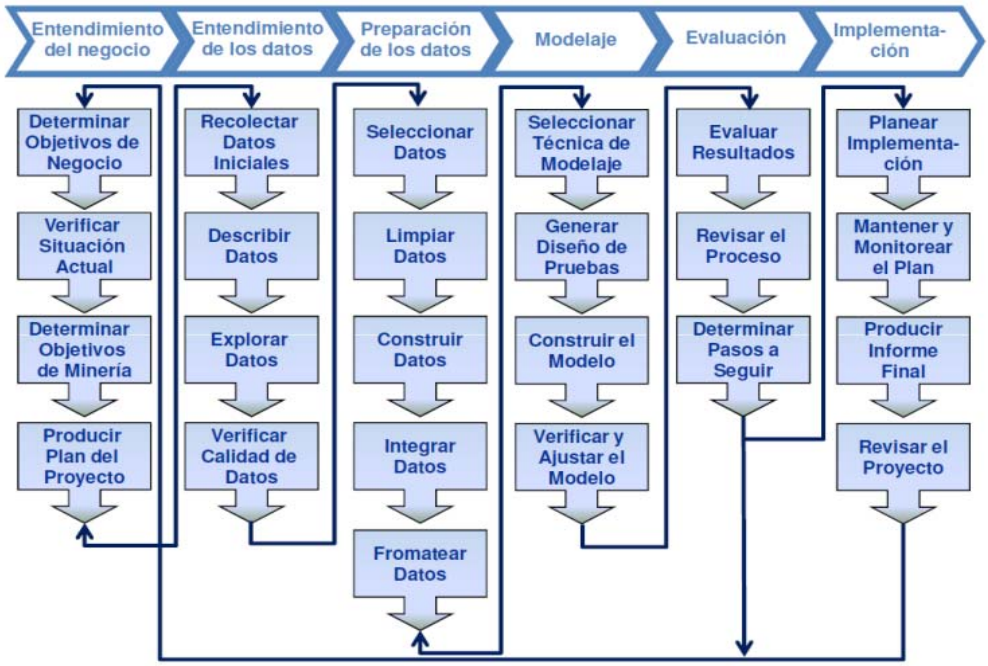


Figura 3. 4 Descripción del Proyecto con CRISP-DM.

Fuente: [WWW4].

3.2.1.1. FASE 1: CONOCIMIENTO DEL NEGOCIO.

El Programa profesional de Ingeniería de Sistemas (PPIS), dispone de una página en Facebook llamado PPIS Sistemas UCSM donde se encuentra parte de la información de los egresados, alumnos, profesores, etc.

Los datos utilizados pertenecen al período desde que se creo la página hasta el mes de julio del presente año, tomando la información personal, académica y laboral de los integrantes de esta página en Facebook.

El origen de la Base de Datos se encuentra en archivos TXT, en ellos encontramos datos personales, académicos y laborales de los egresados del Programa Profesional de Ingeniería de Sistemas

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

(PPIS). Nuestro esquema de trabajo se basa en la siguiente estructura. Ver Figura 3.5.

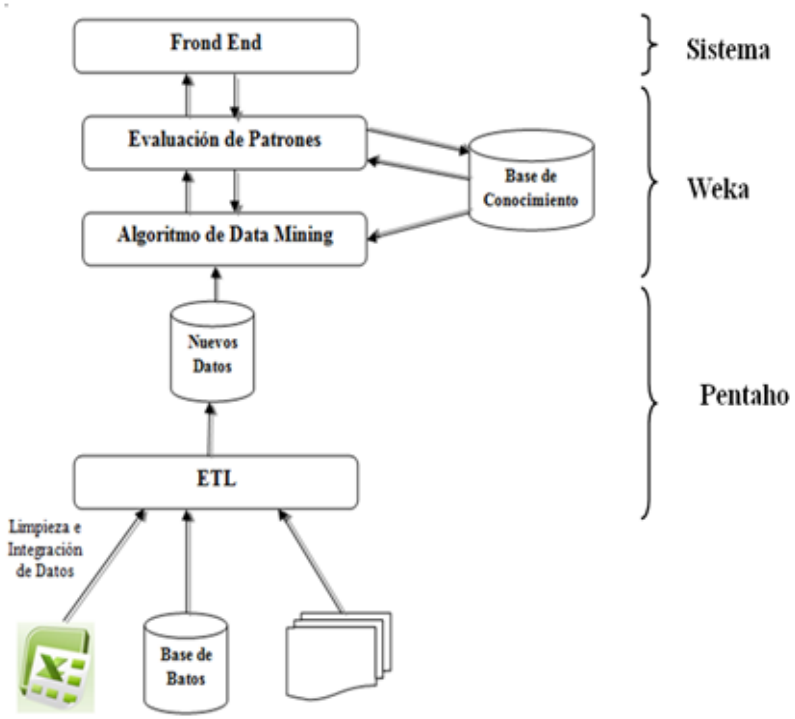


Figura 3. 5 Estructura del Proyecto.

Fuente: [PROPIO].

3.2.1.1.1. Casos de Uso del Proceso ETL.

Para Interpretar el manejo inicial de información que permita a las organizaciones mover datos desde múltiples fuentes, reformatearlos, limpiarlos y cargarlos en otra base de datos para analizar, para apoyar un proceso de negocio.

En el detalle de los casos de uso, se puede apreciar con más claridad cuál es la funcionalidad básica de cada caso. Ver Figura 3.6.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

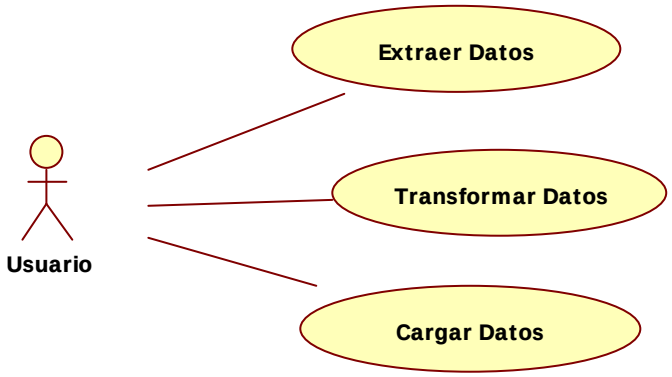


Figura 3. 6 Caso de Uso ETL.

Fuente: PROPIO.

3.2.1.1.2. Detalle Casos de Uso.

Caso de Uso: Extraer Datos	
Breve explicación: La primera parte del proceso consiste en extraer los datos desde los sistemas de origen(Fuentes de almacenamiento de datos).	
Precondiciones: • Las Bases de Datos a extraer se pueden encontrar en distintos formatos o fuentes de almacenamiento. • Si los datos a extraer son muchos, el sistema de origen se podría ralentizar e incluso colapsar, provocando que éste no pueda utilizarse con normalidad para su uso cotidiano. Por esta razón, en sistemas grandes las operaciones de extracción suelen programarse en horarios o días donde este impacto sea nulo o mínimo.	
Acciones del actor	**Acciones del Sistema**
1. Elegir Bases de Datos válidas para el sistema.	
2. Especificar momento y tiempo estimado de extracción.	
3. Verificar que el tipo de conexión sea la adecuada para el tipo de fuente	
	4. Reconoce el tipo de conexión.
	5. Determina el tamaño de las fuentes de Datos.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

	6. Recorre el conjunto de registros, verificando la validez de los datos(Campos nulos vacios).
	7. Pone a disposición las herramientas disponibles para el tipo de Archivo o Base de Datos extraída.
8. Datos extraídos listos para ser manipulados.	
Post Condiciones: • Verificar funcionalidad de las herramientas. • Establecer La estructura de la transformación.	

Tabla 3. 1 Extraer Datos.

Fuente: PROPIO.

Caso de Uso: Transformar Datos.	
Breve explicación: La fase de transformación aplica una serie de reglas o funciones sobre los datos extraídos para convertirlos en datos que serán cargados.	
Precondiciones: • Verificar funcionalidad de las herramientas para optimizar la transformación de datos. • Establecer La estructura de la transformación.	
Acciones del actor	**Acciones del Sistema**
1. Identificar columnas o campos a utilizar (Datos relevantes que enriquezcan la transformación).	
2. Traducir códigos (por ejemplo, si la fuente almacena una "H" para Hombre y "M" para Mujer pero el destino tiene que guardar "1" para Hombre y "2" para Mujer)	
	3. Identifica los tipos de datos de los campos (int, float, String, etc).

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

	4. Obtiene nuevos valores calculados (por ejemplo, total_venta = cantidad * precio).
	5. Une datos de múltiples fuentes.
	6. Calcula totales de múltiples filas de datos (por ejemplo, ventas totales de cada región).
	7. Genera campos clave en el destino.
	8. Divide una columna en varias
	9. Realiza operaciones, simples o complejas, de validación de datos.
10. Datos Transformados listos para ser cargados en una nueva fuente.	
Post Condiciones: • Revisar y validar los datos a cargarse en una nueva fuente. • Revisar el tipo de conexión de la fuente destino.	

Tabla 3. 2 Transformar Datos.

Fuente: PROPIO.

Caso de Uso: Cargar Datos.	
Breve explicación: La fase de carga es el momento en el cual los datos de la fase anterior (transformación) son cargados en el destino(nueva fuente de datos).	
Precondiciones: • Revisar y validar los datos a cargarse en una nueva fuente.	
Acciones del actor	**Acciones del Sistema**
1. Verificar Datos a Cargar.	
	2. Aplica las restricciones.
	3. Carga la nueva fuente de datos.
4. Datos Limpios Listos para el análisis.	
Post Condiciones: • Formato de archivos fuentes. • Análisis de datos (atributos como tipos, formatos, etc.).	

Tabla 3. 3 Cargar Datos.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

3.2.1.1.3. Casos de uso: Minería de Datos.

Para interpretar cuál es el objetivo de nuestra aplicación, en la Figura 3.7 se observa un diagrama de casos de uso. Este diagrama indica las interacciones básicas entre el usuario y el sistema. El usuario, en nuestro caso, es la persona encargada de realizar el proceso de Minería de Datos. Las acciones que realiza son "Preparar Datos", "Procesar Clustering" y "Analizar Resultados". En el detalle de los casos de uso, se puede apreciar con más claridad cuál es la funcionalidad básica de cada caso.

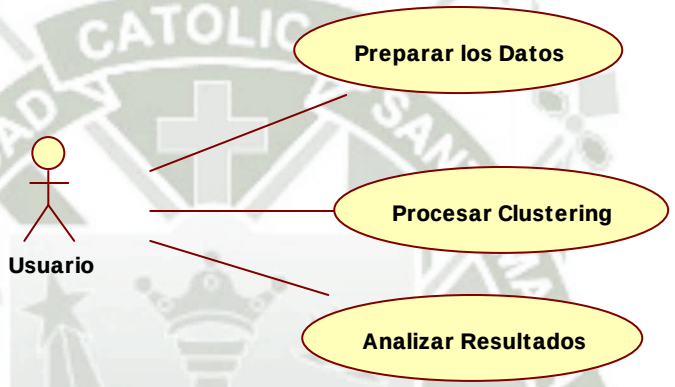


Figura 3. 7 Caso de Uso Minería de Datos.

Fuente: PROPIO.

3.2.1.1.4. Detalle de Casos de Uso.

Caso de Uso: Preparar Datos
Breve explicación: Toma los datos de una base de datos, especificados en un formato determinado, y los asigna a unidades según sus rangos de validez.
Precondiciones: • El archivo de entrada debe cumplir con el formato de los archivos arff. • Además, el archivo debe poseer: ▪ Un atributo que sea un ID único, que identifique al registro (puede haber más de un registro para el mismo ID). ▪ Uno o más atributos (numéricos o categóricos) que se compararán para realizar el Clustering.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Acciones del actor	**Acciones del Sistema**
1. Cargar un archivo válido al sistema.	
2. Especificar parámetros: crono, cantidad de unidades.	
3. Iniciar Filtrado	
	4. Determinar la diferencia entre las fechas.
	5. Recorre el conjunto de registros, verificando dentro de que intervalo es válido (comparando rangos del intervalo).
6. Guardar el Nuevo archivo.	
Post Condiciones: Archivo arff con la siguiente estructura: • ID • Atributos explícitos	

Tabla 3. 4 Preparar Datos.

Fuente: PROPIO.

Caso de Uso: Procesar Clustering	
Breve explicación: Toma los datos pre-procesados y los agrupa según similitudes en sus atributos.	
Precondiciones: • El archivo de entrada debe cumplir con el formato de los archivos arff. • Además, el archivo debe poseer: ▪ Un atributo que sea un ID único, que identifique al registro. ▪ Uno o más atributos (numéricos o categóricos) que se compararán para realizar el Clustering	
Acciones del Actor	**Acciones del Sistema**
1. Cargar un archivo válido al sistema.	
2. Especificar parámetros: Cantidad de clústers.	
3. Ejecutar proceso de Clustering.	

T

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

	4. Transformar los atributos nominales en numéricos utilizando el método de la Documentación.
	5. Elegir al azar tantos centroides como Clústeres se deseen para cada proceso.
	6. Dividir los individuos según método k-d-Median.
	7. Recalcular los centroides de los nuevos Clústeres.
	8. Repetir pasos 6 al 8 hasta que no haya más cambios y los Clústeres sean definitivos.
	9. Calcular cuántos individuos pasaron de un clúster a otro y determinar el orden de los Clústeres en cada tiempo.
	10. Armar archivo de salida con información sobre cada registro.
	11. Mostrar en pantalla la distribución de centroides.
Post Condiciones: • Archivo arff con la siguiente estructura: o Detalle de los atributos explícitos de cada registro. o Número de clúster a los que pertenece ese registro. o Flag que indica si ese registro es un centroide o no (y en caso de qué sea de qué clúster).	

Tabla 3.5 Procesar Clustering.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

<table>
<tr><td colspan="2">Caso de Uso: Analizar Resultados.</td></tr>
<tr><td colspan="2">Breve explicación: Toma el archivo de salida del proceso de Clustering y brinda diferentes gráficos para analizar los resultados obtenidos.</td></tr>
<tr><td colspan="2">Precondiciones:
Archivo arff con la siguiente estructura:
Detalle de los atributos explícitos de cada registro.
Número de clúster a los que pertenece ese registro.
Flag que indica si ese registro es un centroide o no(y en caso de qué sea, de qué clúster).</td></tr>
<tr><td>Acciones del actor</td><td>Acciones del Sistema</td></tr>
<tr><td>1. Cargar archivo de resultados *.arff</td><td></td></tr>
<tr><td></td><td>2. Leer información del archivo y configurar opciones del programa. Por default están todas deshabilitadas.*
Si tiene 2 atributos, habilitar Gráficos 2D y Gráficos 2D Tiempo*
Si tiene 3 atributos, habilitar Gráficos 3D y Gráficos 3D Tiempo*
Las opciones: Análisis por Atributo, Estadísticas de Individuos e Individuos Perdidos habilitarlas siempre.</td></tr>
<tr><td>3. Seleccionar alguna de las opciones disponibles.</td><td></td></tr>
<tr><td></td><td>4. Mostrar en pantalla el gráfico correspondiente:
Gráficos 2 D: distribución de las 2 variables en el plano, diferenciando los distintos Clústeres.
Gráficos 2D Tiempo: distribución de un mismo clúster a través de los distintos tiempos.
Gráficos 3 D: distribución de las 3 variables en el hiperplano, diferenciando los distintos Clústeres.
Gráficos 3D Tiempo: distribución de un mismo clúster a través de los distintos tiempos.
Análisis por atributo: trayectoria que sigue el centroide de cada clúster a través de los distintos tiempos.
Estadísticas de Individuos: Gráfico de barras que indica cuántos individuos pasaron entre un clúster y otro al cambiar de tiempo.</td></tr>
</table>

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

	• Individuos perdidos: Gráfico de barras que indica cuántos individuos no tienen información en un determinado tiempo.

Tabla 3. 6 Analizar Resultados.

Fuente: PROPIO.

3.2.1.2. FASE 2: CONOCIMIENTO DE LOS DATOS.

Se seleccionaron las bases de datos: AmigosFacebook, siendo este un aproximado del 22% de los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS), AmigosEstudios, AmigosTrabajos. La fuente de la cual se extraerán los datos se encuentra en el formato TXT, las fuentes de datos tienen los siguientes nombres:

- AmigosFacebook: En este archivo se encuentran los datos del usuario(Ver Figura 3.8) referidos a sus datos personales de su perfil de usuario, el cual está constituido por los siguientes campos como se muestra en la Tabla 3.7:

Campo	Significado
Idusu	ID
Nomusu	Nombre
Apeusu	Apellidos
Username	Usuario
Genusu	Sexo
Fdnusu	Fecha de Nacimiento
Maiusu	Email Usuario
Ciuact	Ciudad Actual
Linkus	Link
Polusu	Ideología Política
Fraus	Acerca de mi
Relusu	Creencias Religiosas
Tipo	Tipo
Pagweb	Página Web

Tabla 3. 7 Descripción del archivo AmigosFacebook.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Fuente: PROPIO.

```
'idusu','nomusu','apeusu','username','genusu','fdnusu','maiusu','ciunat','linkus','polusu','fraus','relusu','tipo','pagweb'
'22015287','Jean Pierre Valencia','Gomez','jean.valencia','M','04/14/1981','','Perú, Arequipa, Peru','http://www.facebook.com/jean.valencia','','','','',
'105901467','Pedro A.','Briceño','pbriceno15','M','','','','','http://www.facebook.com/pbriceno15','','','','','',
'501376269','Jose Antonio','Córdova Torres','wendellcito','M','05/18/1987','','Arequipa','http://www.facebook.com/wendellcito','','','','','',
'501635236','Dewi Mirza','Cano Torres','dewimirza.canotorres','F','06/23','','Perú, Arequipa, Peru','http://www.facebook.com/dewimirza.canotorres','','',
'501771041','Bettox','Gallegos','bettox','M','06/23','','','','http://www.facebook.com/bettox','','','','','',
'501818439','Miguel','Ramos','Superbos','M','07/30','','Arequipa','http://www.facebook.com/Superbos','','Whatever Makes you Happy , Whatever you want.','
'502244850','Jose Miguel','Choque','josemiguel.choque','M','09/07/1983','','Camaná','http://www.facebook.com/josemiguel.choque','','','','','',
```

Figura 3. 8 Contenido del archivo AmigosFacebook

Fuente: PROPIO.

- AmigosEstudios: En este archivo se encuentran los datos relacionados con los estudios del usuario como se muestra en la Figura 3.9 El archivo está constituido por los siguientes campos los cuales almacenan el contenido de cada usuario que se describen en la Tabla 3.2

Campo	Significado
Idusu	ID
Tipest	Colegio
Nomest	Universidad
Aniest	Año Postgrado
Conest	Concentracion
Graest	Grado

Tabla 3. 8 Descripción del archivo AmigosEstudios.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA


```
'idusu','tipest','nomest','aniest','conest','graest'
'22015287','High School','Colegio De La Salle Arequipa','','',''
'22015287','College','Universidad Catolica de Santa Maria - Arequipa','','Systems Engineering',''
'22015287','College','Universidad Catolica de Santa Maria','','',''
'22015287','Graduate School','DePaul University','2008','Software Engineering, Project Management',''
'501376269','College','Catholic University of Santa María','','Systems Engineering',''
'501635236','High School','Santa Ursula Arequipa','1997','',''
'501635236','College','UCSM','','Ing. Sistemas',''
'501635236','College','Catholic University of Santa María','','',''
```

Figura 3. 9 Contenido del archivo AmigosEstudios

Fuente: PROPIO.

- AmigosTrabajos: En este archivo contiene los datos referidos al trabajo del usuario como se muestra en la Figura 3.10. y las características del archivo están constituidos por los siguientes Tabla 3.3

Campo	Significado
Idusu	ID
Emptra	Empleador
Puetra	Cargo
Fecini	Fecha Inicio
Fecfin	Fecha Fin

Tabla 3. 9 Descripción del archivo AmigosTrabajos.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA


```
'idusu','emptra','puetra','fecini','fecfin'
'22015287','Depaul University','Software Engineer','2006-05-01','2006-09-01'
'22015287','Banco de Credito BCP','Software Engineer','2003-06-01','2005-12-01'
'504731451','Grupo Comet :: integrating people with technology','','2011-06-01','2011-09-01'
'507769506','KeeperTech S.A.','Software Developer','2011-07-01','2012-04-01'
'507769506','Microdata','Programmer Tesista','2010-07-01','2011-08-01'
'507769506','Banco de Credito BCP','Practicante','2010-01-01','2010-03-01'
'507769506','Global Systems Consulting','Programmer','2009-07-01','2009-10-01'
'516578878','MINSUR S.A.','','2010-01-01','2010-03-01'
'516578878','tenifluidos sac','','2008-12-01','2009-03-01'
'524461734','FONDESURCO','','2010-12-01','2011-12-01'
'524461734','Snack drive inn - El tacho','Mesera, cajera y hasta barman jaja','2007-04-01','2008-09-01'
'524821940','Google Summer of Code 2010','Student developer for Ubuntu','2010-05-01','2010-08-01'
'524821940','Haikulogic','Intern','2009-03-01','2009-07-01'
```

Figura 3. 10 Contenido del archivo AmigosTrabajos.

Fuente: PROPIO.

. 3.2.1.3. FASE 3: PREPARACIÓN DE LOS DATOS.

Con la herramienta de Pentaho se estudiaron los atributos, sus valores y el comportamiento de los mismos. Fue necesario integrar varios orígenes de datos ya que toda la información necesaria se encontraba en diferentes tablas.

Los datos personales de los egresados, estudiantes, profesores y otros integrantes de la pagina se encontraban en desorden, para lo cual se hizo una limpieza de datos logrando una salida con campos específicos, y se obtuvo las siguientes tablas de salidas llamadas: CiudadActual, EstadoLaboral, Especialidad y Carrera. Como se muestra en a Figura 3.11. Para este proyecto se seleccionó el 90% de los datos existentes.

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

Figura 3. 11 Estructura de Datos de Salida.

Fuente: PROPIO.

3.2.1.3.1. Proceso ETL.

En el proceso ETL (Figura 3.12) Procedemos a extraer los datos relevantes desde las Fuentes de Datos que se encuentran en el formato TXT. Se integran y transforman los datos, para evitar inconsistencias. Se cargan los datos para almacenarlos en nuevas fuentes de salida TXT, MySQL y Excel.

Para el proceso ETL(Extracción, Transformación y Carga) de datos utilizaremos el Pentaho la cual es una herramienta de integración de datos que nos permite unir los archivos TXT en una sola fuente de salida con los datos ya transformados según los requerimientos del usuario.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Figura 3. 12 Proceso ETL.

Fuente: PROPIO.

- **Extracción:** La extracción se realiza basándose en las necesidades y requisitos de los usuarios del Programa Profesional de Ingeniería de Sistemas(PPIS), se exploran las fuentes de datos que hemos obtenido referentes a los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS), y se extrae la información que se considere relevante de los archivos AmigosEstudios, AmigosTrabajos y AmigosFacebook. Un requerimiento importante que se debe exigir a la tarea de extracción es que ésta cause un impacto mínimo en el sistema origen. Si los datos a extraer son muchos, el sistema de origen se podría ralentizar e incluso colapsar, provocando que éste no pueda utilizarse con normalidad para su uso cotidiano. Nuestras fuentes de datos se encuentra en el formato TXT de las cuales vamos a extraer la información requerida por el usuario que por comodidad lo hemos almacenado en la siguiente dirección C:\ppis. La extracción de los datos realizada en el Pentaho como se muestran en la Figura 3.13.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

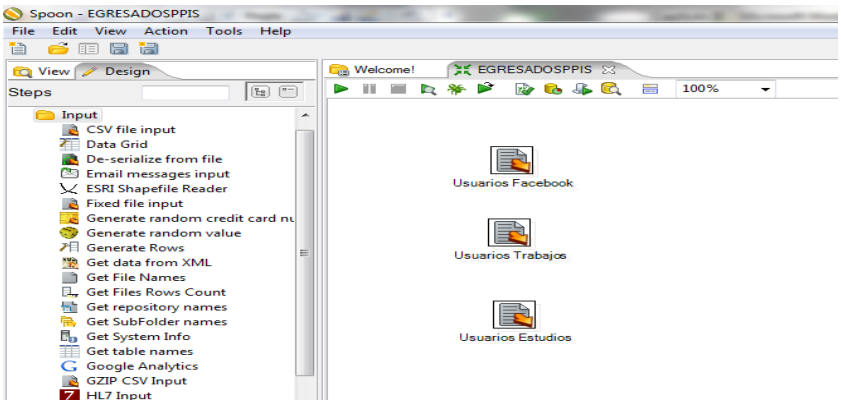


Figura 3. 13 Extracción de datos.

Fuente: PROPIO.

Configuración de la extracción de datos en el Spoon:

- En la pestaña File configuramos el step name con los nombres AmigosFacebook, AmigosTrabajos y AmigosEstudios, en File or Directory buscamos las direcciones (C:\ppis\AmigosFacebook.txt, C:\ppis\AmigosTrabajos.txt, C:\ppis\AmigosEstudios.txt) donde se encuentran nuestras fuentes de datos para así añadirlas y que se muestren en Selected files. Ver Figura 3.13.1, Figura 3.13.2 y Figura 3.13.3.

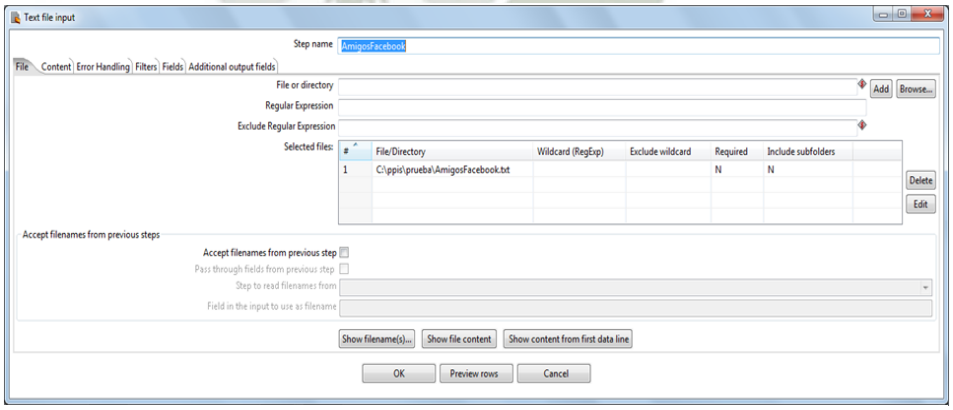


Figura 3.13. 1 Configuración de extracción de la fuente AmigosFacebook.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

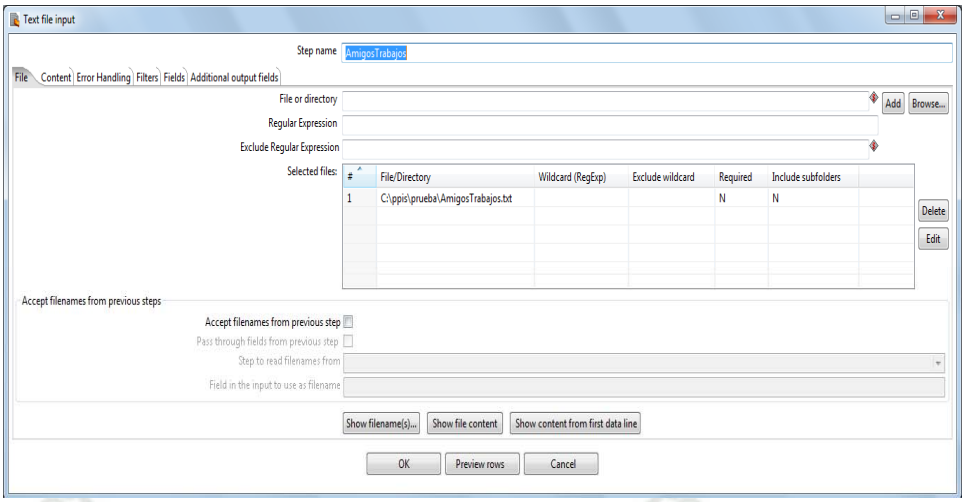


Figura 3.13. 2 Configuración de extracción de la fuente AmigosTrabajos.

Fuente: PROPIO.

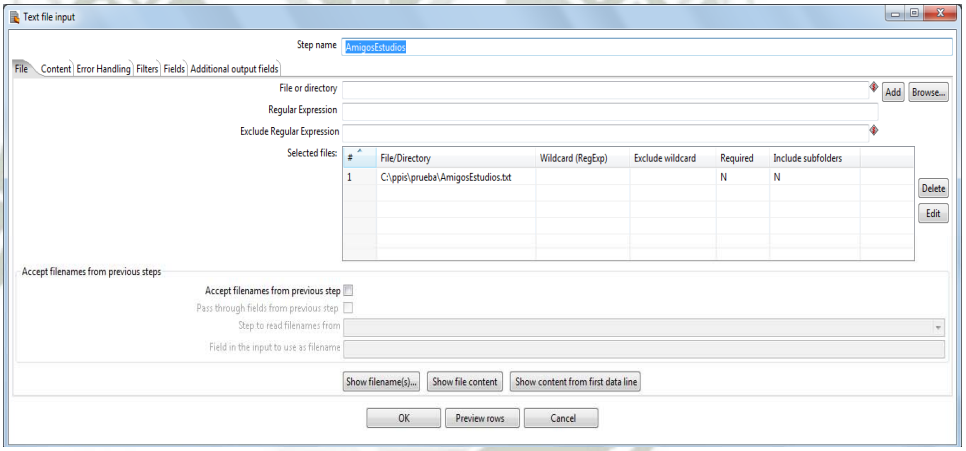


Figura 3.13. 3 Configuración de extracción de la fuente AmigosEstudios.

Fuente: PROPIO.

- En la pestaña Content configuramos los siguientes datos de la siguiente manera:

  - En Filetype conservamos por defecto el formato ya existente que en este caso es csv. el cual acepta la lectura de formato TXT.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

- Separator: conservamos el símbolo de separación predeterminada el cual es (;).
- Enclousure: conservamos el símbolo que indica espacios en blanco el cual es (¨).

Estas son las principales características que se deben tomar en cuenta para esta configuración. Ya que los demás datos son jalados por defecto según las características de la fuente de datos. Como se muestra en la Figura 3.12.4. Este paso es repetitivo en las tres fuentes de datos extraídas para el proyecto.

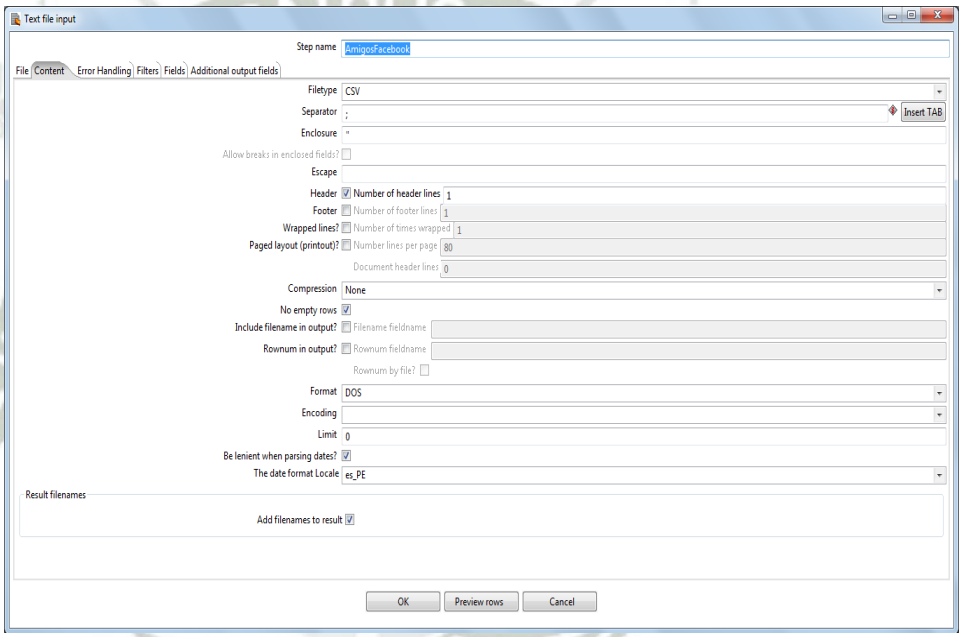


Figura 3.13. 4 Configuración del contenido de la fuente de datos.

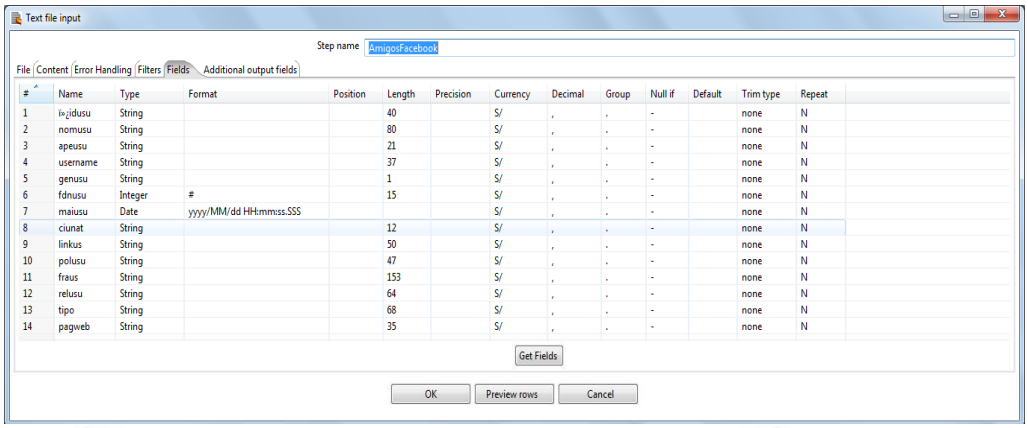
Fuente: PROPIO.

- En la pestaña Fields mediante el botón Get fields extraemos los campos de nuestras fuentes de datos: AmigosFacebook, AmigosTrabajos, AmigosEstudios. Como se muestra en la Figura 3.13.5, Figura 3.13.6 y Figura 3.13.7.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

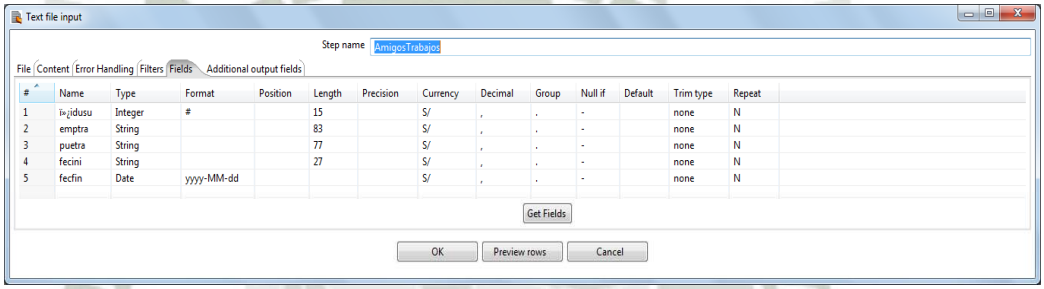
CATÓLICA DE SANTA MARÍA



#	Name	Type	Format	Position	Length	Precision	Currency	Decimal	Group	Null if	Default	Trim type	Repeat
1	ï»¿idusu	String			40		S/	,	.	-		none	N
2	nomusu	String			80		S/	,	.	-		none	N
3	apeusu	String			21		S/	,	.	-		none	N
4	username	String			37		S/	,	.	-		none	N
5	genusu	String			1		S/	,	.	-		none	N
6	fdnusu	Integer	#		15		S/	,	.	-		none	N
7	maiusu	Date	yyyy/MM/dd HH:mm:ss.SSS				S/	,	.	-		none	N
8	ciunat	String			12		S/	,	.	-		none	N
9	linkus	String			50		S/	,	.	-		none	N
10	polusu	String			47		S/	,	.	-		none	N
11	fraus	String			153		S/	,	.	-		none	N
12	relusu	String			64		S/	,	.	-		none	N
13	tipo	String			68		S/	,	.	-		none	N
14	pagweb	String			35		S/	,	.	-		none	N

Figura 3.13. 5 Contenido de los atributos de la fuente de datos AmigosFacebook.

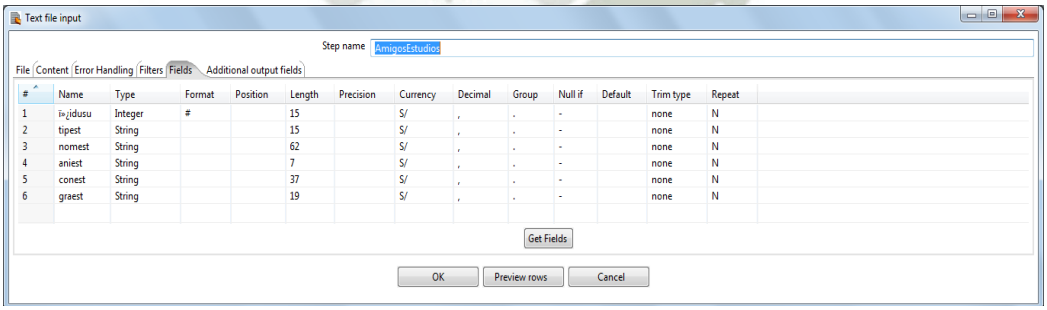
Fuente: PROPIO.



#	Name	Type	Format	Position	Length	Precision	Currency	Decimal	Group	Null if	Default	Trim type	Repeat
1	ï»¿idusu	Integer	#		15		S/	,	.	-		none	N
2	emptra	String			83		S/	,	.	-		none	N
3	puetra	String			77		S/	,	.	-		none	N
4	fecini	String			27		S/	,	.	-		none	N
5	fecfin	Date	yyyy-MM-dd				S/	,	.	-		none	N

Figura 3.13. 6 Contenido de los atributos de la fuente de datos AmigosTrabajos.

Fuente: PROPIO.



#	Name	Type	Format	Position	Length	Precision	Currency	Decimal	Group	Null if	Default	Trim type	Repeat
1	ï»¿idusu	Integer	#		15		S/	,	.	-		none	N
2	tipest	String			15		S/	,	.	-		none	N
3	nomest	String			62		S/	,	.	-		none	N
4	aniest	String			7		S/	,	.	-		none	N
5	conest	String			37		S/	,	.	-		none	N
6	graest	String			19		S/	,	.	-		none	N

Figura 3.13. 7 Contenido de los atributos de la fuente de datos AmigosEstudios.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

- Mediante el botón Previous rows podemos visualizar el contenido de las fuentes de datos: AmigosFacebook, AmigosTrabajos y AmigosEstudios. Como se muestra en la Figura 3.13.8, Figura 3.13.9 y Figura 3.13.10

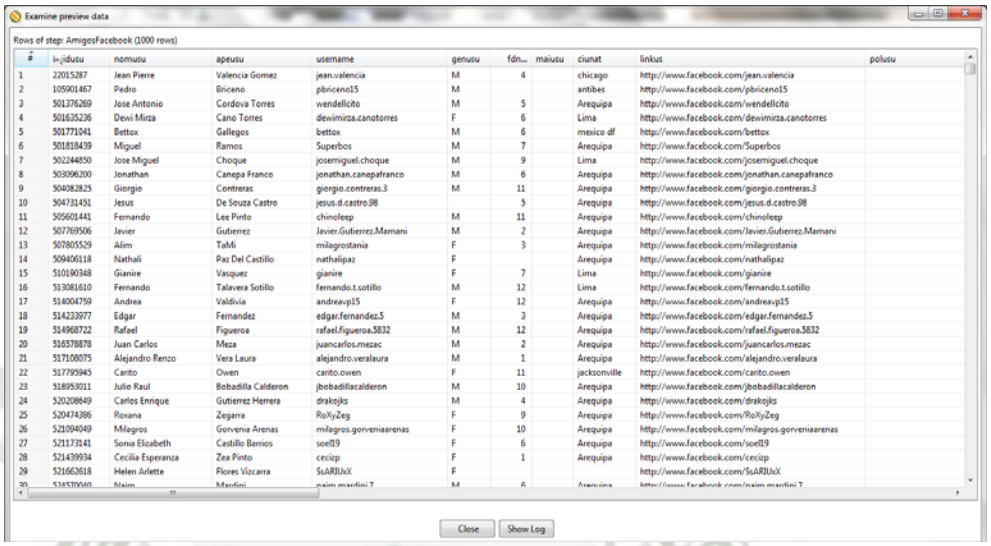


Figura 3.13. 8 Visualización de atributos de la fuente AmigosFacebook.

Fuente: PROPIO.

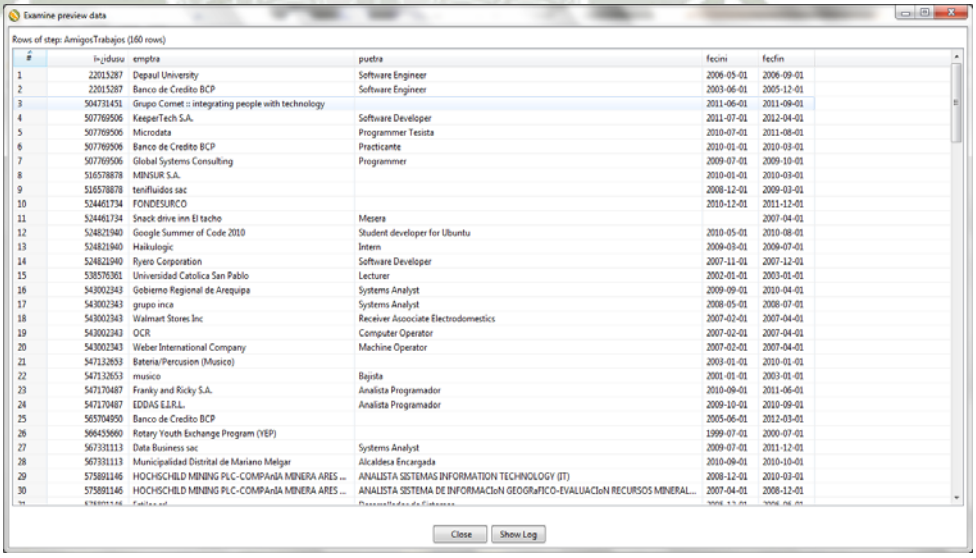


Figura 3.13. 9 Visualización de atributos de la fuente AmigosTrabajos.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

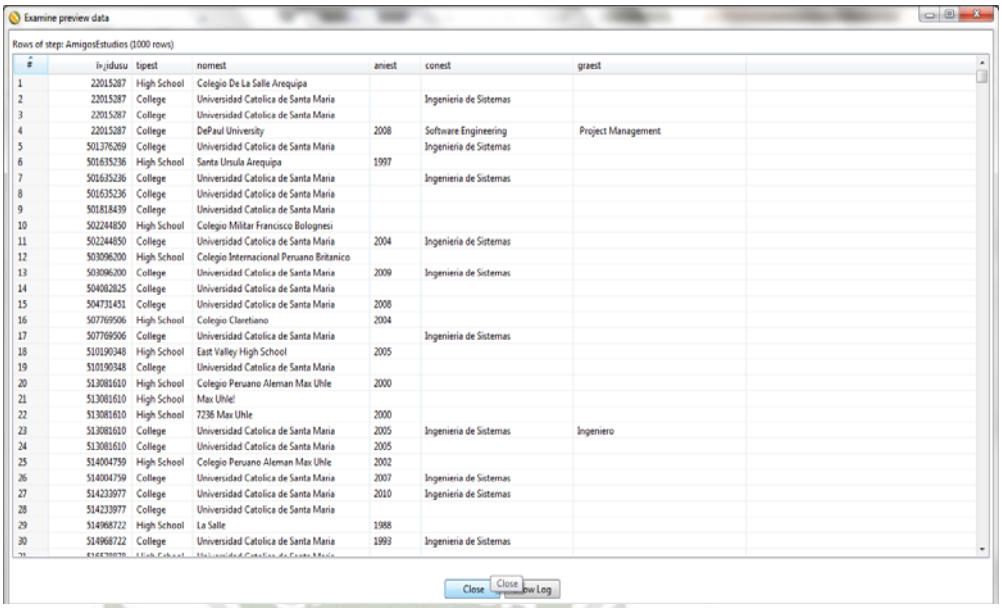


Figura 3.13. 10 Visualización de atributos de la fuente AmigosEstudios.

Fuente: PROPIO.

- **Transformación:** En esta parte del proceso nos encargamos de convertir los datos inconsistentes en un conjunto de datos compatibles y congruentes, para que puedan ser cargados en una nueva fuente de datos de salida. Estas acciones se llevan a cabo, debido a que pueden existir diferentes fuentes de información, y es vital conciliar un formato y forma única, definiendo estándares, para que todos los datos que ingresarán a la nueva fuente de datos, como se muestra en la Figura 3.14.

Figura 3. 14 Estandarización de datos.

Fuente: PROPIO.

Los casos en los que se realizara integración, son los siguientes:

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

Fuentes múltiples: El Pentaho nos permite elegir diferentes fuentes. En este caso, se debe elegir aquella fuente que se considere más fiable y apropiada para nuestro caso nuestras fuentes se encuentra con la extensión TXT:

- ✓ AmigosFacebook.txt
- ✓ AmigosEstudios.txt
- ✓ AmigosTrabajos.txt

En la Transformación elegimos los campos a integrar discriminado los datos irrelevantes a nuestro objetivo de segmentación. Las fuentes de datos de las cuales disponemos son 3 con extensión TXT con sus respectivos campos que se muestra en la Tabla 3.1, Tabla 3.2 y Tabla 3.3 de los cuales elegiremos para nuestro propósito los siguientes campos:

- ✓ AmigosFacebook: idusu, nomusu, apeusu, ciuact.
- ✓ AmigosEstudios: idusu, nomest, aniest, conest, graest
- ✓ AmigosTrabajos: idusu, emptra, puetra, fecini, fecfin

PASO 1.- Primera Transformación.

En esta transformación extraemos los datos específicos que vamos a necesitar en el desarrollo del proyecto. La estructura de la primera transformación se muestra en la Figura 3.15.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

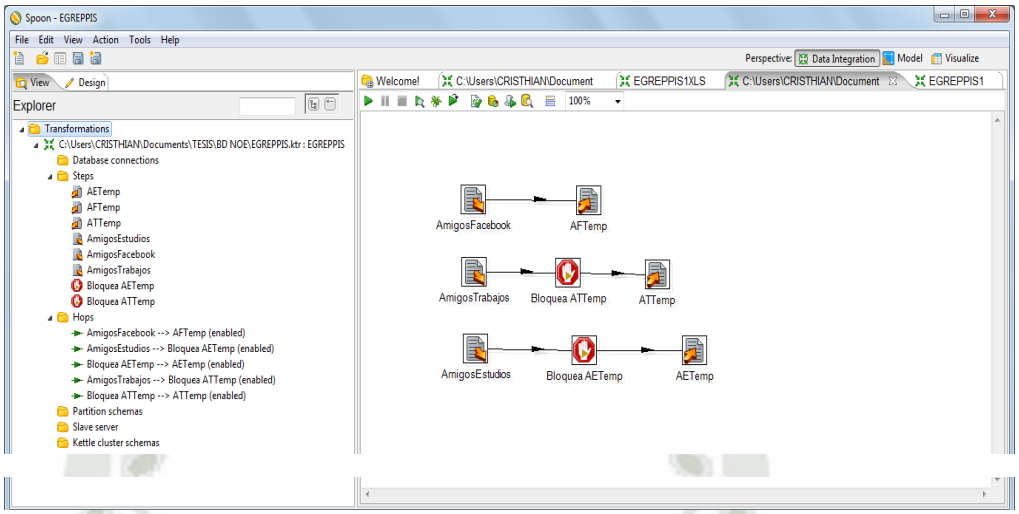


Figura 3. 15 Estructura de Primera Transformación

Fuente: PROPIO.

- Configuración de la pestaña File donde especificamos los nombres en Step name: AFTemp, ATTemp y AETemp que son fuentes de datos temporales.
- Indicamos la dirección donde se guardara la salida con los datos, mediante el botón Browse. Las direcciónes de salidas temporales serán: C:\ppis\AFTemp, C:\ppis\ATTemp, C:\ppis\AETemp. Como se muestra en la Figura 3.15.1, Figura 3.15.2 y Figura 3.15.3

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

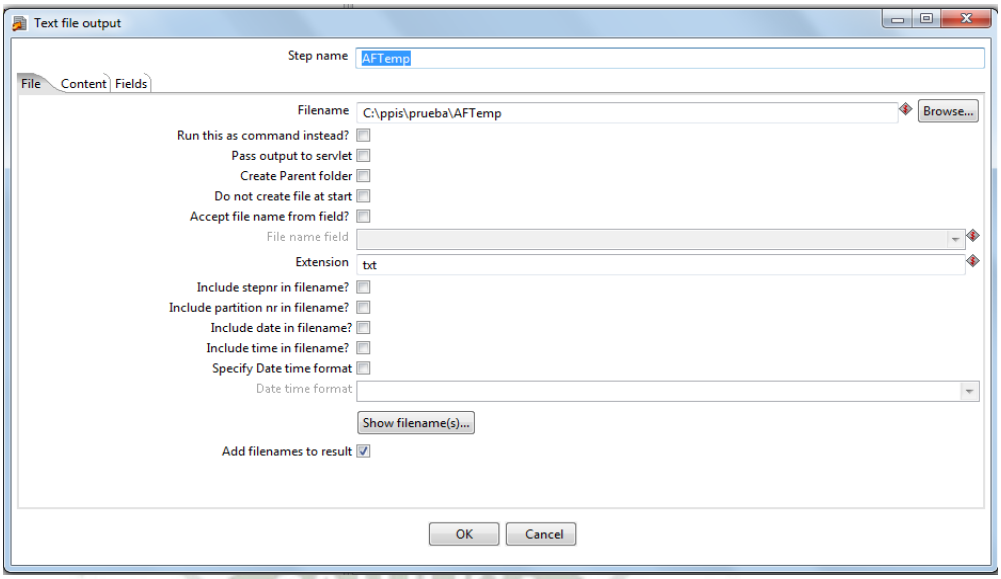


Figura 3.15. 1 Configuración de la fuente temporal AFTemp.

Fuente: PROPIO.

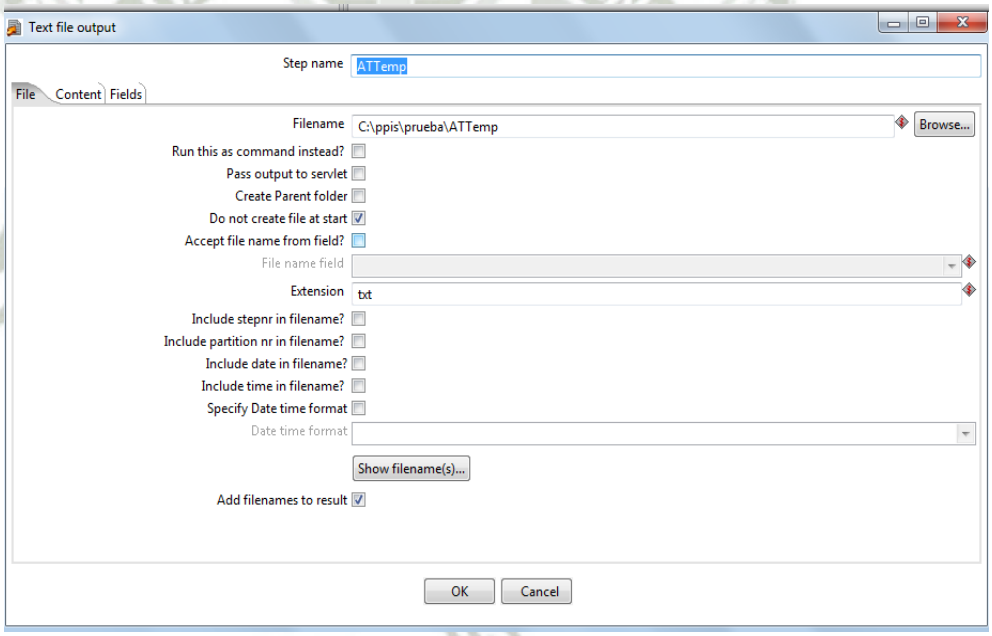


Figura 3.15. 2 Configuración de la fuente temporal ATTemp.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

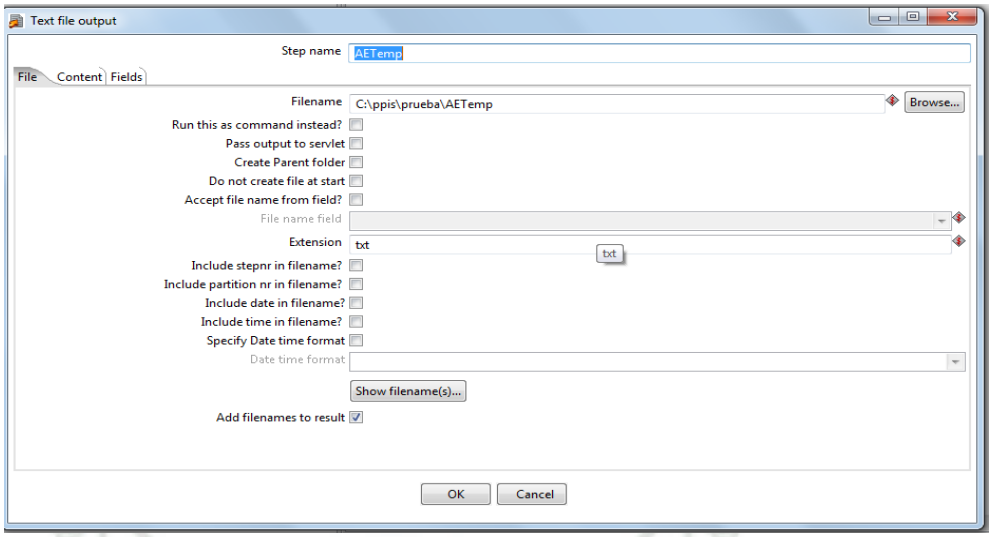


Figura 3.15. 3 Configuración de la fuente temporal AETemp.

Fuente: PROPIO.

- En la pestaña Content configuramos los siguientes datos de la siguiente manera:

  - Separator: cambiamos el símbolo de separación predeterminado (;) por (,).
  - Enclousure: conservamos el símbolo que indica espacios en blanco el cual es (¨).

Estas son las principales características que se deben tomar en cuenta para esta configuración. Ya que los demás datos son jalados por defecto según las características de la fuente de datos. Como se muestra en la Figura 3.14.4. Este paso es repetitivo en las tres fuentes de datos extraídas para el proyecto.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

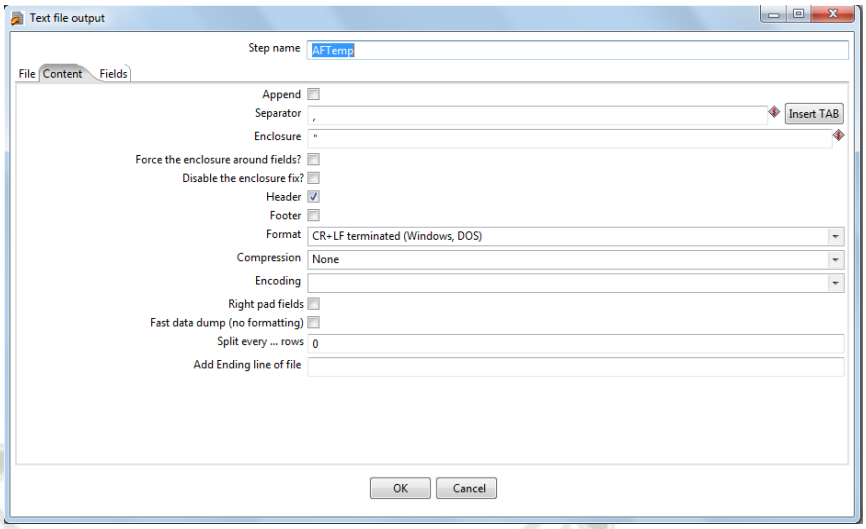


Figura 3.15. 4 Configuración del contenido de la fuente de datos.

Fuente: PROPIO.

- En la pestaña Fields mediante el botón Get fields extraemos los campos específicos de nuestras fuentes de datos originales. Como se muestra en la Figura 3.15.5, Figura 3.15.6 y Figura 3.15.7

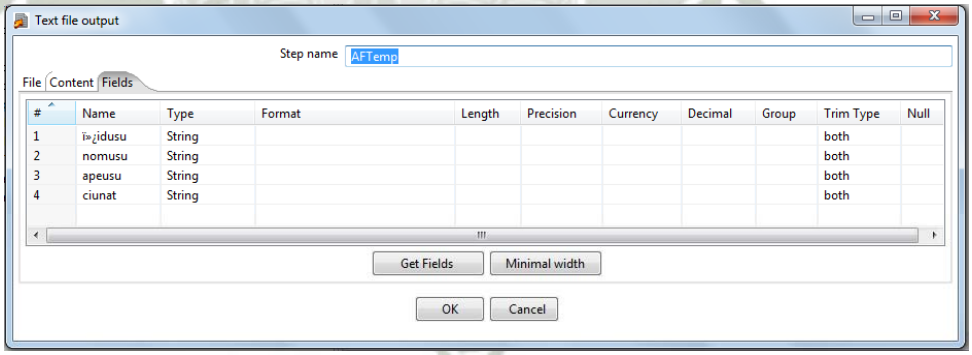


Figura 3.15. 5 Contenido de los atributos de la fuente temporal AFTemp.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

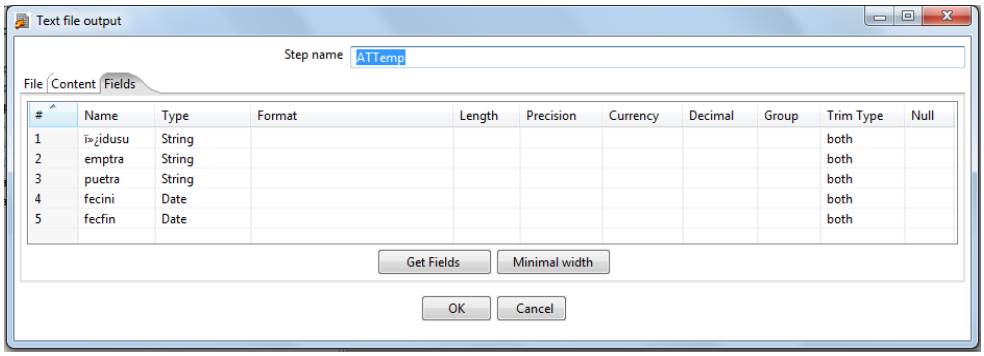


Figura 3.15. 6 Contenido de los atributos de la fuente temporal ATTemp.

Fuente: PROPIO.

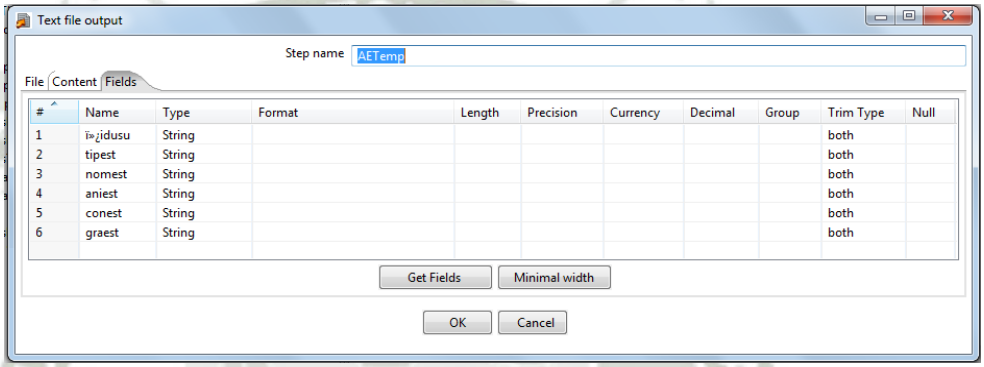


Figura 3.15. 7 Contenido de los atributos de la fuente temporal AETemp

Fuente: PROPIO.

- En la estructura de la primera transformación podemos apreciar que encontramos un paso de bloqueo, el cual nos permite detener la ejecución del proceso mientras no se hayan terminado de ejecutar los pasos anteriores. Esto lo podemos visualizar en la Figura 3.15.8

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

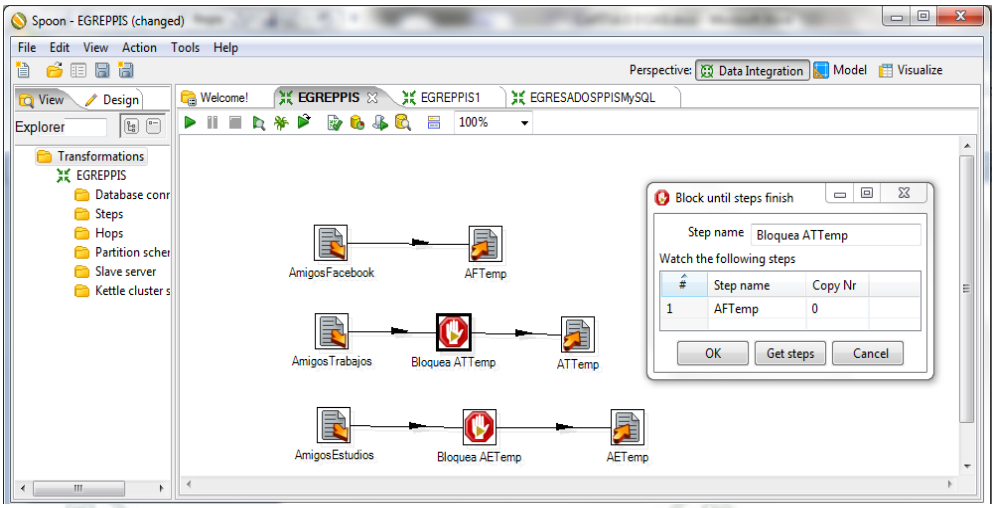
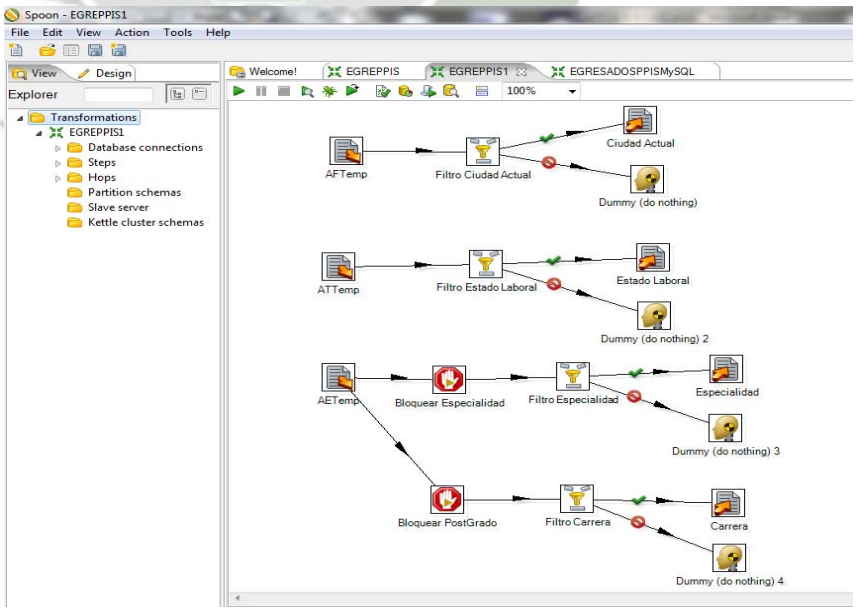


Figura 3.15. 8 Bloqueo de ejecución de procesos.

Fuente: PROPIO.

- **PASO 2.- Segunda Transformación:** En el proceso de la segunda transformación procedemos a eliminar los campos vacios, filtrar los datos, separación de campos de las fuentes de datos como se muestra en la siguiente transformación de datos Figura 3.16. Para el proceso de filtrado obtenemos los datos de las fuentes de datos temporales AFTemp, ATTemp y AETemp.



Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

Figura 3. 16 Estructura de la segunda Transformación.

Fuente: PROPIO.

- Filtro Ciudad Actual: La configuración está dada de la siguiente manera

  Step name: Filtro Ciudad Actual.

  Send 'true' data to step: Ciudad Actual.

  Send 'false' data to step: Dumny(do nothing).

  The condition: Ciudadactual IS NOT NULL, esta condición nos permite almacenar en la nueva fuente datos no nulos. Como se muestra en la Figura 3.16.1.

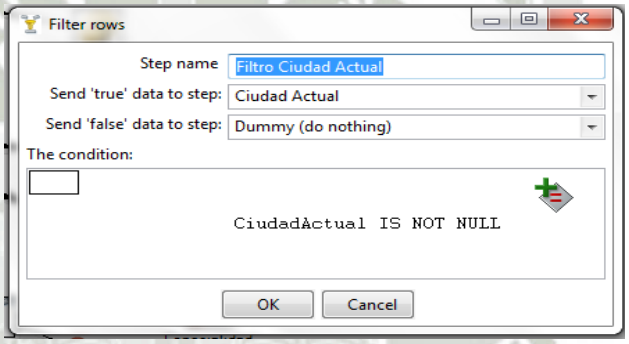


Figura 3.16. 1 Filtro Ciudad Actual.

Fuente: PROPIO.

- Filtro Estado Laboral: La configuración del Filtro Estado Laboral es el siguiente.

  Step name: Filtro Estado Laboral.

  Send 'true' data to step: Estado Laboral.

  Send 'false' data to step: Dumny(do nothing)2.

  The condition: Empleador IS NOT NULL AND Cargo IS NOT NULL, esta condición nos permite almacenar en la nueva fuente datos no nulos. Como se muestra en la Figura 3.16.2.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

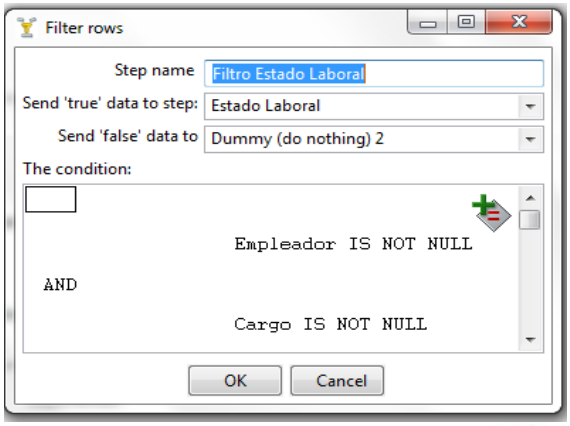


Figura 3.16. 2 Filtro Estado Laboral.

Fuente: PROPIO.

- Filtro Especialidad: La configuración del Filtro Especialidad es el siguiente.

  Step name: Filtro Especialidad.

  Send 'true' data to step: Especialidad.

  Send 'false' data to step: Dumny(do nothing)3.

  The condition: Nivel = [College] AND Nombre IS NOT NULL AND Carrera IS NOT NULL AND Especialidad IS NOT NULL, esta condición nos permite almacenar en la nueva fuente el nombre del campo College y los campos Nombre, Carrera y Especialidad no nulos. Como se muestra en la Figura 3.16.3.

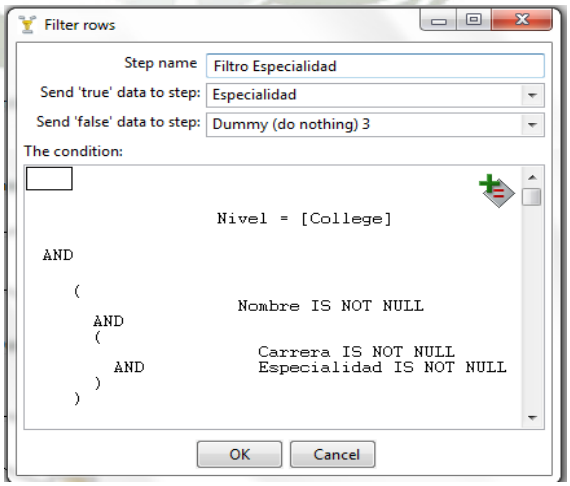


Figura 3.16. 3 Filtro Especialidad.

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

Fuente: PROPIO.

- Filtro Carrera: La configuración del Filtro Carrera es el siguiente.

  Step name: Filtro Carrera.

  Send 'true' data to step: Carrera.

  Send 'false' data to step: Dumny(do nothing)4.

  The condition: Nivel = [College] AND Nombre IS NOT NULL AND Carrera IS NOT NULL, esta condición nos permite almacenar en la nueva fuente el nombre del campo College y los campos Nombre, Carrera no nulos. Como se muestra en la Figura 3.16.4.

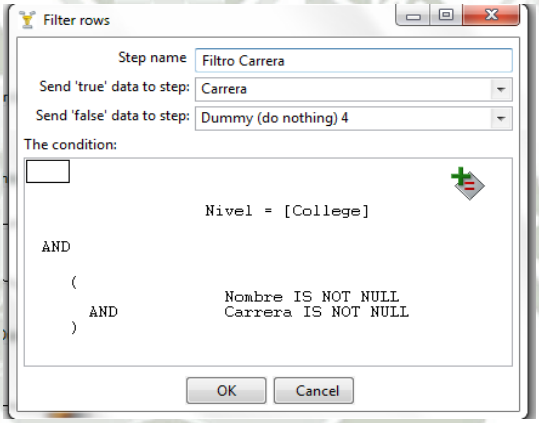


Figura 3.16. 4 Filtro Carrera.

Fuente: PROPIO.

- Dummy(Do nothing): La función de Dummy permite recibir del filtro los errores de las fuentes y no se realice ninguna acción . Ver Figura 3.16.5.

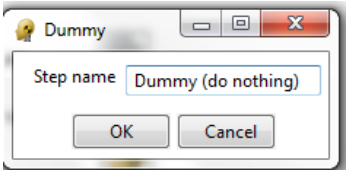


Figura 3.16. 5 Dummy

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

- Text file Output: El Text file Output para este proceso de trasformación está constituido por las siguientes salidas de datos: Estado Laboral, Ciudad Actual, Especialidad y Carrera. Como en pasos anteriores mostramos la configuración de las pestañas.
  - ✓ File con los siguientes datos:

    Step name: Nombre de la fuente de salida. Ver figura 3.16.6.

    Filename: La dirección donde se almacena la nueva fuente.
  - ✓ Content: Separator: (,) y Enclousure: ("). Ver figura 3.16.7.
  - ✓ Fields: Get fields para obtener los campos y Minimal width para indicar los tipos de datos mínimos. Ver figura 3.16.8.

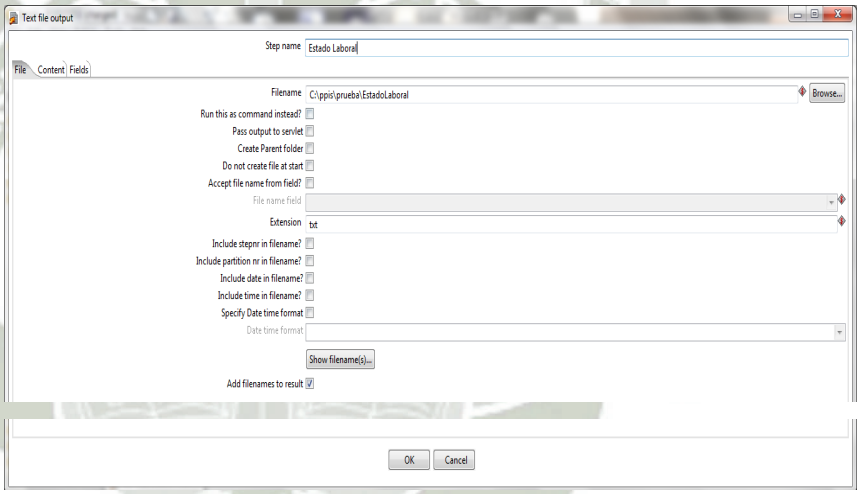
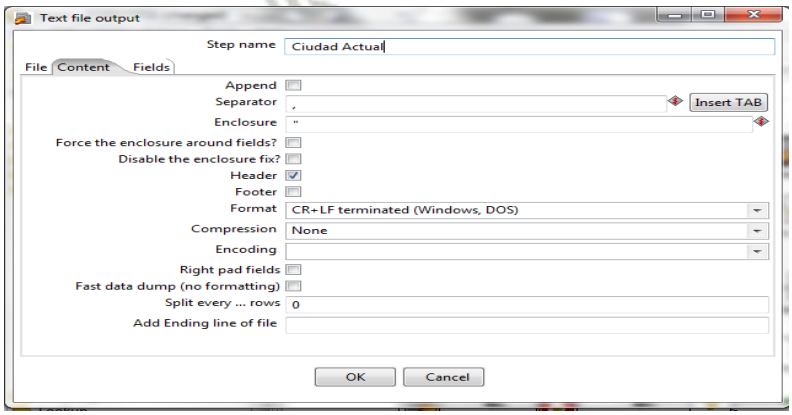


Figura 3.16. 6 Configuración de file de Text file Output.

Fuente: PROPIO.



Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Figura 3.16. 7 Configuración de Content de Text file Output.

Fuente: PROPIO.

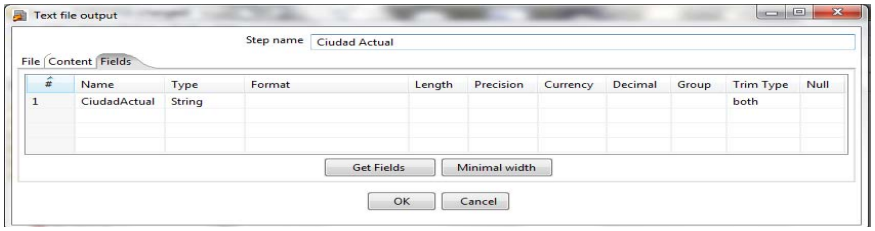


Figura 3.16. 8 Configuración de Fields de Text file Output.

Fuente: PROPIO.

PASO 3.- Tercera Transformación: En el proceso de la tercera transformación procedemos indicamos las salidas de datos en MySQL y Excel como se muestra en la Figura 3.17. La transformación de datos se realizan de las fuentes de datos Temporales: AFTemp, ATTemp y AETemp.

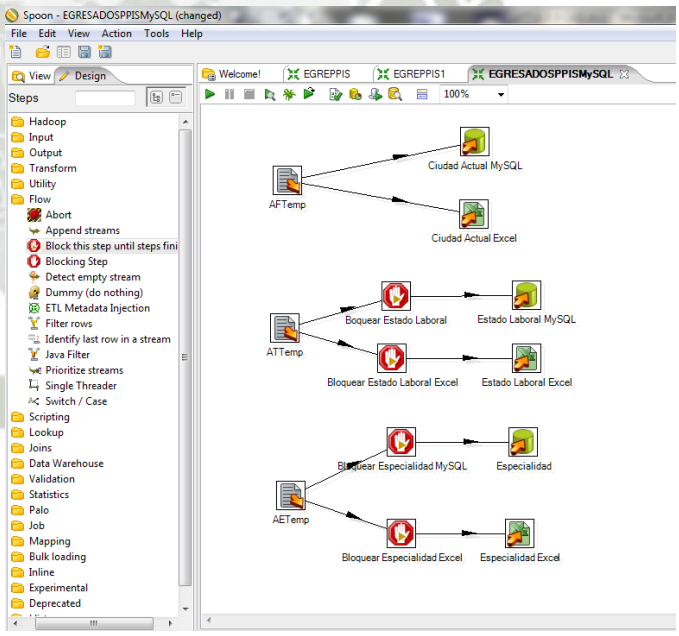


Figura 3. 17 Salida de fuentes MySQL y Excel.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

- Creamos bloqueos para completar la salida de datos en las fuentes cuando estas sean cargadas.

- La salida de datos para MySQL se configura de la siguiente manera. Ver Figura 3.17.1.
  Step name: Nombre de la tabla de salida.
  Connection: aquí se realiza la conexión con la base de datos.
  Target schema: Aquí va el esquema destino.
  Target table: aquí va el esquema de la tabla.
  Specify database fields: check on

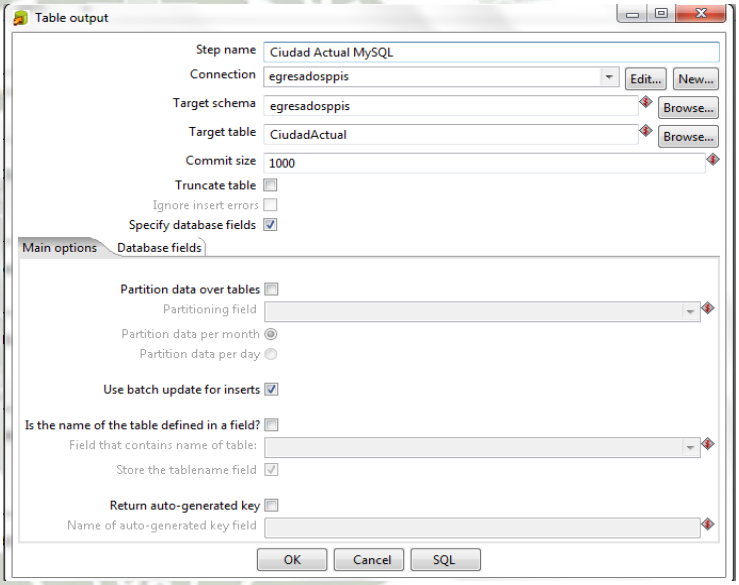


Figura 3.17. 1 Configuración Table Output.

Fuente: PROPIO.

- Antes de hacer la conexión para la salida de la nueva fuente creamos una Base de Datos en MySQL llamada: egresadosppis. Ver Figura 3.17.2

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

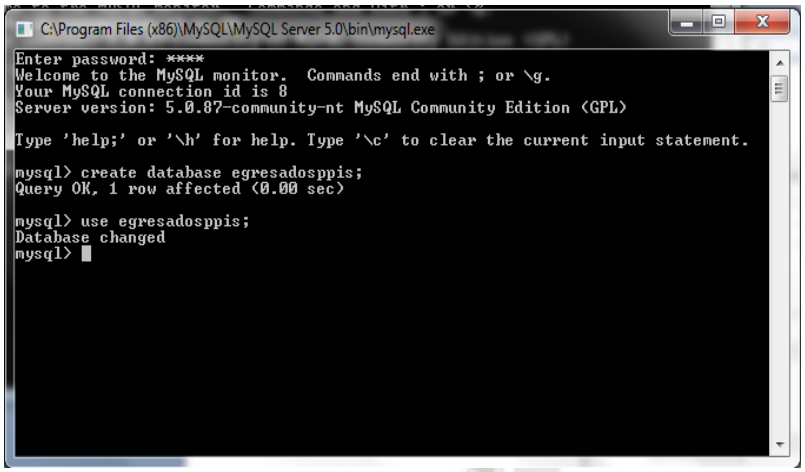


Figura 3.17. 2 Creación de Base de Datos egresadosppis.

Fuente: PROPIO.

- Configuración de la conexión con la base de datos. Ver Figura 3.17.3.

  Connection Name: egresadosppis, es el nombre de la conexión.

  Connection Type: MySQL, aquí se elije la base de datos en la que se cargaran los datos.

  Access: Native(JDBC)

  Hots Name: Localhost

  Database Name: egresadosppis

  Port Number: 3306

  User Name: root

  Password: root

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

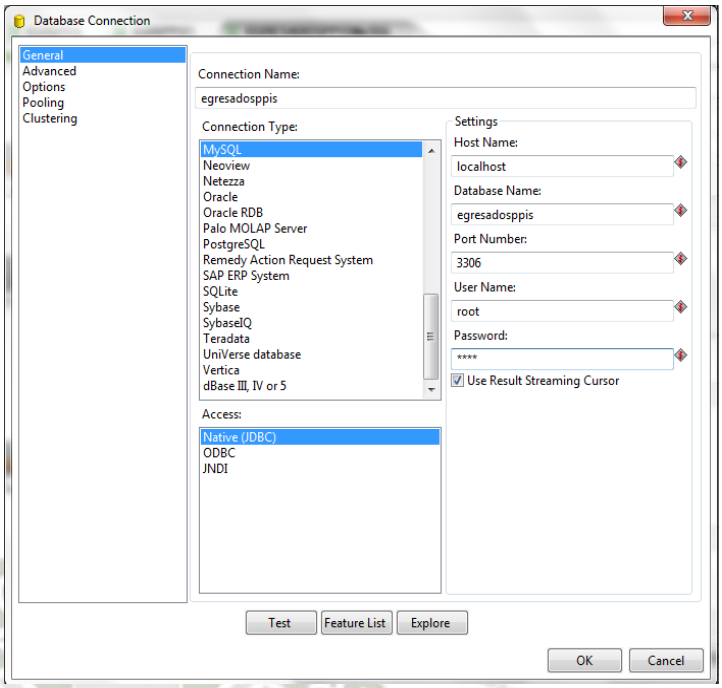


Figura 3.17. 3 Configuración de la conexión con la Base de Datos.

Fuente: PROPIO.

- Test de conexión con la base de datos. Nos indica si se realizó con éxito la conexión si existe la base de datos como los campos de las fuentes. Como se muestra en la Figura 3.17.4.

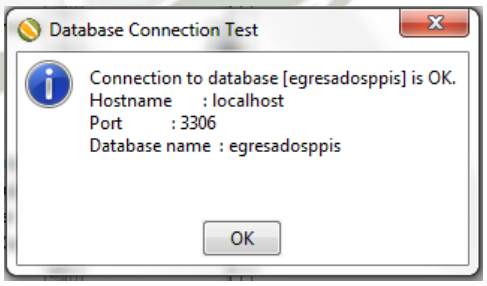


Figura 3.17. 4 Test de Conexión.

Fuente: PROPIO.

- Database Fields. Aquí mostramos los campos de la fuente con el botón Get Fields. Como se muestra en la Figura 3.17.5.

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

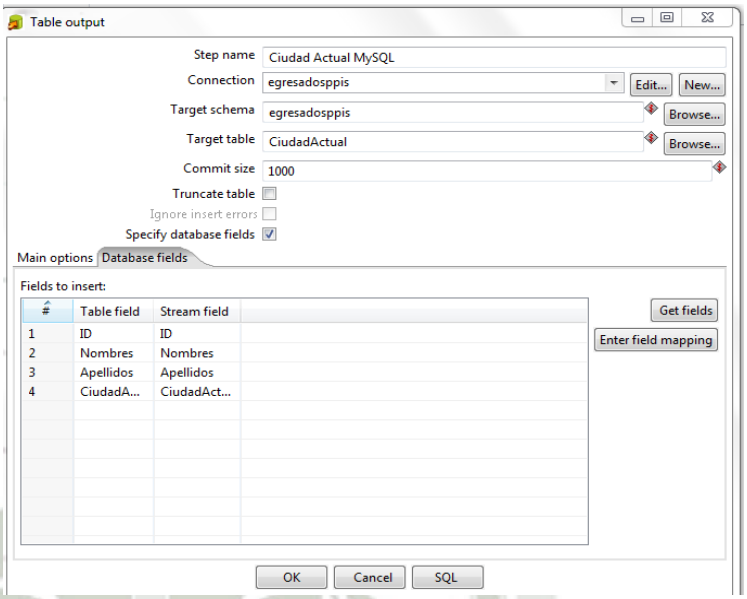


Figura 3.17. 5 Muestra de campos de las fuentes.

Fuente: PROPIO.

- Simple SQL Editor. Aquí nos genera la estructura de la tabla de la fuente de datos, este editor nos permite modificar la estructura en caso ser necesario. Como se muestra en la Figura 3.17.6.

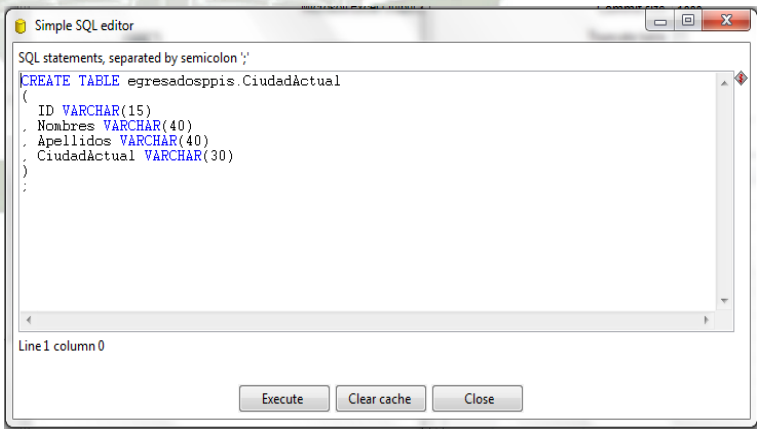


Figura 3.17. 6 Simple SQLEditor.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

- Results of the SQL statements. Aquí se muestra la estructura de la tabla a crearse en la base de datos en MySQL. Como se muestra en la Figura 3.17.7.

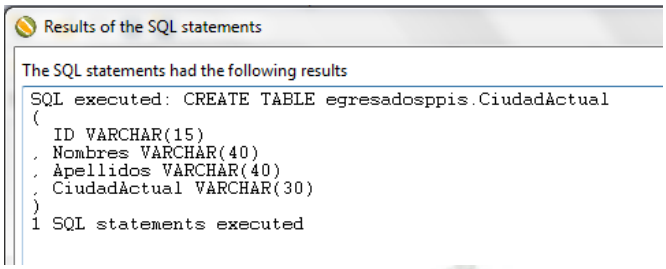


Figura 3.17.7 Results of the SQL statements.

Fuente: PROPIO.

- Para la configuración de salida de datos en el formato Excel.

  Step name: Estado Laboral Excel.

  Filename: C:\ppis\EstadoLaboral.xls

  Extension: .xls.

  Como se muestra en la Figura 3.17.8.

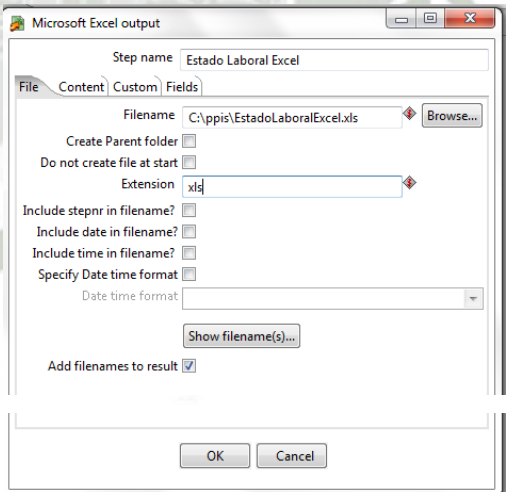


Figura 3.17. 8 Ruta de almacenamiento Excel.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

- Visualización de los datos de la fuente con el botón Get Fields, permite ver si los datos fueron configurados correctamente. Como se muestra en la Figura 3.17.9.

.

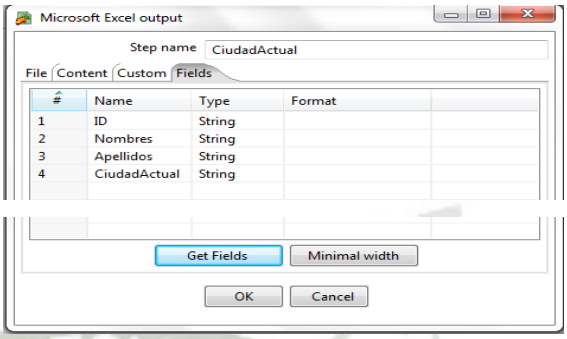


Figura 3.17. 9 Get Fields.

Fuente: PROPIO.

- **Carga**: En esta etapa del proceso se genera una nueva fuente con los datos integrados de las fuentes de datos iniciales. Se debe tener en cuenta, que los datos antes de moverse al almacén de datos, deben ser analizados con el propósito de asegurar su calidad, ya que este es un factor clave, que no debe dejarse de lado. Las fuentes de salida producto del ETL en tres formatos (Excel, MySQL y CSV).

- Una vez terminado el diseño y la estructura con sus respectivas configuraciones de datos, procedemos a ejecutar la transformación(icono verde de reproducción) y le damos clic en Launch, como se muestra en la Figura 3.18

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

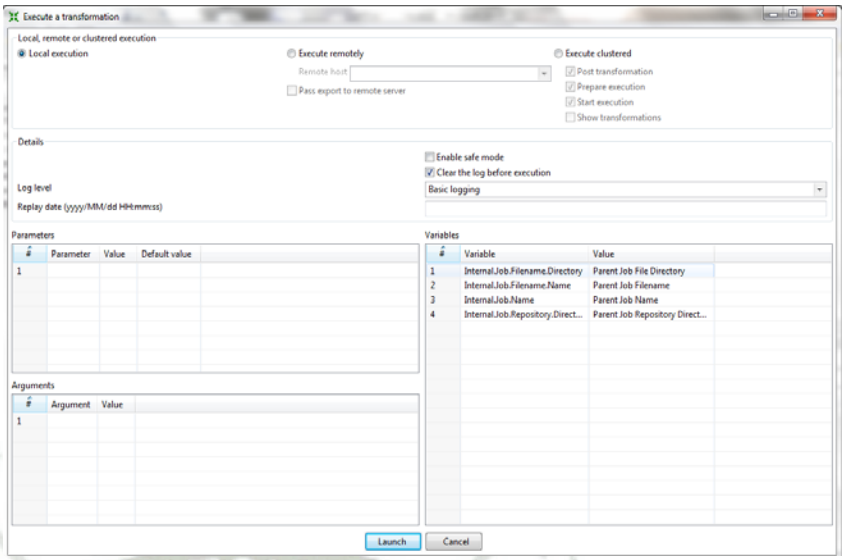


Figura 3. 18 Ejecución de la Transformación.

Fuente: PROPIO.

- Después de la ejecución de transformación, se extraen los datos de las fuentes y se crean las salidas de datos detallando en la parte inferior de la Figura 3.18.1 y Figura 3.18.2 . La descripción de los procesos como datos de extraídos, salida de datos y procesos intermedios.

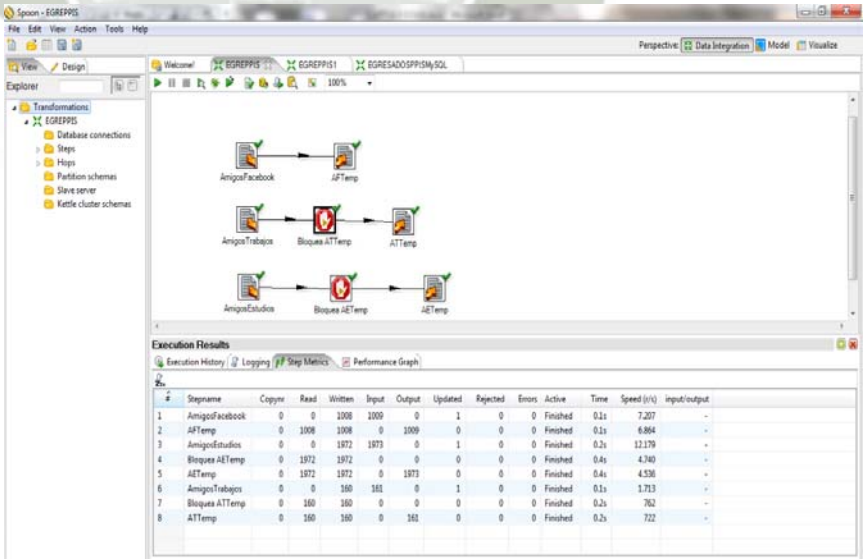
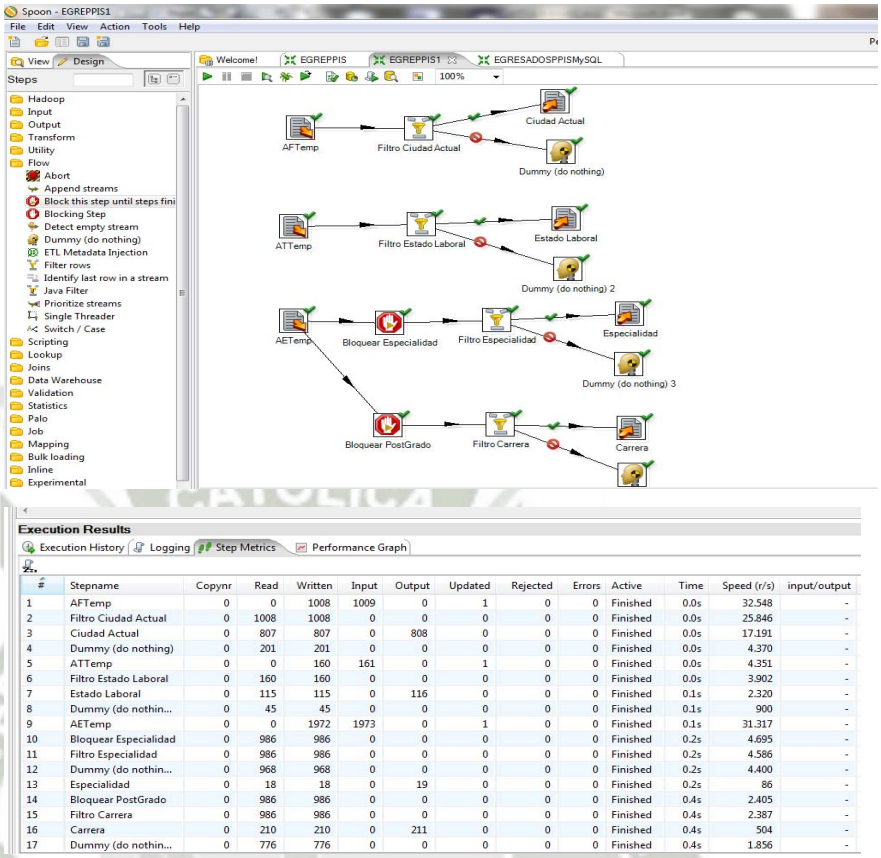


Figura 3.18. 1 Resultados de la primera transformación.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Fuente: PROPIO.



#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	AFTemp	0	0	1008	1009	0	1	0	0	Finished	0.0s	32.548	-
2	Filtro Ciudad Actual	0	1008	1008	0	0	0	0	0	Finished	0.0s	25.846	-
3	Ciudad Actual	0	807	807	0	808	0	0	0	Finished	0.0s	17.191	-
4	Dummy (do nothing)	0	201	201	0	0	0	0	0	Finished	0.0s	4.370	-
5	ATTemp	0	0	160	161	0	1	0	0	Finished	0.0s	4.351	-
6	Filtro Estado Laboral	0	160	160	0	0	0	0	0	Finished	0.0s	3.902	-
7	Estado Laboral	0	115	115	0	116	0	0	0	Finished	0.1s	2.320	-
8	Dummy (do nothin...	0	45	45	0	0	0	0	0	Finished	0.1s	900	-
9	AETemp	0	0	1972	1973	0	1	0	0	Finished	0.1s	31.317	-
10	Bloquear Especialidad	0	986	986	0	0	0	0	0	Finished	0.2s	4.695	-
11	Filtro Especialidad	0	986	986	0	0	0	0	0	Finished	0.2s	4.586	-
12	Dummy (do nothin...	0	968	968	0	0	0	0	0	Finished	0.2s	4.400	-
13	Especialidad	0	18	18	0	19	0	0	0	Finished	0.2s	86	-
14	Bloquear PostGrado	0	986	986	0	0	0	0	0	Finished	0.4s	2.405	-
15	Filtro Carrera	0	986	986	0	0	0	0	0	Finished	0.4s	2.387	-
16	Carrera	0	210	210	0	211	0	0	0	Finished	0.4s	504	-
17	Dummy (do nothin...	0	776	776	0	0	0	0	0	Finished	0.4s	1.856	-

Figura 3.18. 2 Resultados de la segunda transformación.

Fuente: PROPIO.

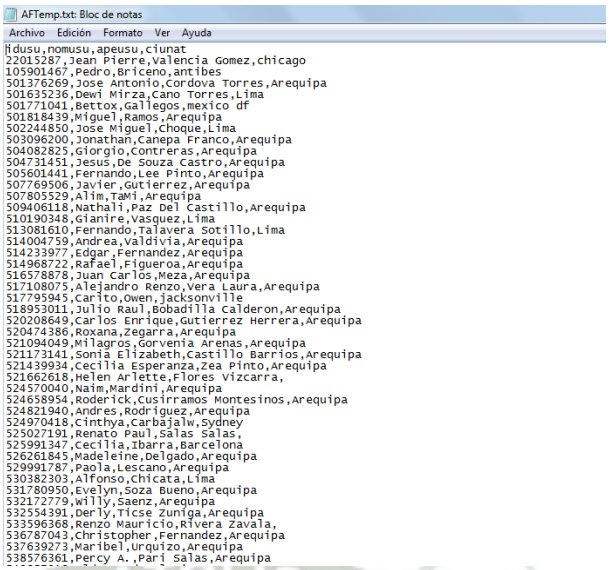
OUTPUT DE LAS TRANFORMACIONES

- OUTPUT 1.- Es la primera salida temporal con los campos siguientes: idusu(ID), nomusu(nombre de usuario), apeusu(apellido de usuario) y ciunat(cuidad actual). Ver Figura 3.19

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA



AFTemp.txt: Bloc de notas

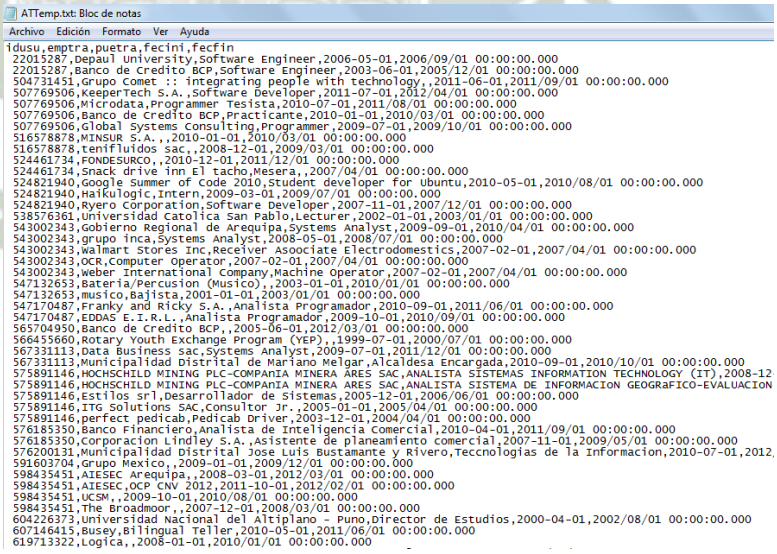
Archivo Edición Formato Ver Ayuda

```
idusu,nomusu,apeusu,ciunat
22015287,Jean Pierre,Valencia Gomez,chicago
105901467,Pedro,Briceno,antibes
501376269,Jose Antonio,Cordova Torres,Arequipa
501635236,Dewi Mirza,Cano Torres,Lima
501771041,Bettox,Gallegos,mexico df
501818439,Miguel,Ramos,Arequipa
502244850,Jose Miguel,Choque,Lima
503096200,Jonathan,Canepa Franco,Arequipa
504082825,Giorgio,Contreras,Arequipa
504731451,Jesus,De Souza Castro,Arequipa
505601441,Fernando,Lee Pinto,Arequipa
507769506,Javier,Gutierrez,Arequipa
507805529,Alim,TaMi,Arequipa
509406118,Nathali,Paz Del Castillo,Arequipa
510190348,Gianire,Vasquez,Lima
513081610,Fernando,Talavera Sotillo,Lima
514004759,Andrea,Valdivia,Arequipa
514233977,Edgar,Fernandez,Arequipa
514968722,Rafael,Figueroa,Arequipa
516578878,Juan Carlos,Meza,Arequipa
517108075,Alejandro Renzo,Vera Laura,Arequipa
517795945,Carito,Owen,jacksonville
518953011,Julio Raul,Bobadilla Calderon,Arequipa
520208649,Carlos Enrique,Gutierrez Herrera,Arequipa
520474386,Roxana,Zegarra,Arequipa
521094049,Milagros,Gorvenia Arenas,Arequipa
521173141,Sonia Elizabeth,Castillo Barrios,Arequipa
521439934,Cecilia Esperanza,Zea Pinto,Arequipa
521662618,Helen Arlette,Flores Vizcarra,
524570040,Naim,Mardini,Arequipa
524658954,Roderick,Cusirramos Montesinos,Arequipa
524821940,Andres,Rodriguez,Arequipa
524970418,Cinthya,Carbajalw,Sydney
525027191,Renato Paul,Salas Salas,
525991347,Cecilia,Ibarra,Barcelona
526261845,Madeleine,Delgado,Arequipa
529991787,Paola,Lescano,Arequipa
530382303,Alfonso,Chicata,Lima
531780950,Evelyn,Soza Bueno,Arequipa
532172779,Willy,Saenz,Arequipa
532554391,Derly,Ticse Zuniga,Arequipa
533596368,Renzo Mauricio,Rivera Zavala,
536787043,Christopher,Fernandez,Arequipa
537639273,Maribel,Urquizo,Arequipa
538576361,Percy A.,Par¦ Salas,Arequipa
```

Figura 3. 19 Salida temporal de AmigosFacebook

Fuente: PROPIO.

- OUTPUT 2.- Es la segunda salida temporal con los campos siguientes: idusu(ID), emptra(empleador), puetra(puesto de trabajo), fecini(fecha de inicio) y fecfin(fecha de fin). Ver Figura 3.20



ATTemp.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
idusu,emptra,puetra,fecini,fecfin
22015287,Depaul University,Software Engineer,2006-05-01,2006/09/01 00:00:00.000
22015287,Banco de Credito BCP,Software Engineer,2003-06-01,2005/12/01 00:00:00.000
504731451,Grupo Comet :: integrating people with technology,,2011-06-01,2011/09/01 00:00:00.000
507769506,KeeperTech S.A.,Software Developer,2011-07-01,2012/04/01 00:00:00.000
507769506,Microdata,Programmer Tesista,2010-07-01,2011/08/01 00:00:00.000
507769506,Banco de Credito BCP,Practicante,2010-01-01,2010/03/01 00:00:00.000
507769506,Global Systems Consulting,Programmer,2009-07-01,2009/10/01 00:00:00.000
516578878,MINSUR S.A.,,2010-01-01,2010/03/01 00:00:00.000
516578878,tenifluidos sac,,2008-12-01,2009/03/01 00:00:00.000
524461734,FONDESURCO,,2010-12-01,2011/12/01 00:00:00.000
524461734,Snack drive inn El tacho,Mesera,,2007/04/01 00:00:00.000
524821940,Google Summer of Code 2010,Student developer for Ubuntu,2010-05-01,2010/08/01 00:00:00.000
524821940,Haikulogic,Intern,2009-03-01,2009/07/01 00:00:00.000
524821940,Ryero Corporation,Software Developer,2007-11-01,2007/12/01 00:00:00.000
538576361,Universidad Catolica San Pablo,Lecturer,2002-01-01,2003/01/01 00:00:00.000
543002343,Gobierno Regional de Arequipa,Systems Analyst,2009-09-01,2010/04/01 00:00:00.000
543002343,grupo inca,Systems Analyst,2008-05-01,2008/07/01 00:00:00.000
543002343,Walmart Stores Inc,Receiver Asociate Electrodomestics,2007-02-01,2007/04/01 00:00:00.000
543002343,OCR,Computer Operator,2007-02-01,2007/04/01 00:00:00.000
543002343,Weber International Company,Machine Operator,2007-02-01,2007/04/01 00:00:00.000
547132653,Bateria/Percusion (Musico),,2003-01-01,2010/01/01 00:00:00.000
547132653,musico,Bajista,2001-01-01,2003/01/01 00:00:00.000
547170487,Franky and Ricky S.A.,Analista Programador,2010-09-01,2011/06/01 00:00:00.000
547170487,EDDAS E.I.R.L.,Analista Programador,2009-10-01,2010/09/01 00:00:00.000
565704950,Banco de Credito BCP,,2005-06-01,2012/03/01 00:00:00.000
566455660,Rotary Youth Exchange Program (YEP),,1999-07-01,2000/07/01 00:00:00.000
567331113,Data Business sac,Systems Analyst,2009-07-01,2011/12/01 00:00:00.000
567331113,Municipalidad Distrital de Mariano Melgar,Alcaldesa Encargada,2010-09-01,2010/10/01 00:00:00.000
575891146,HOCHSCHILD MINING PLC-COMPANIA MINERA ARES SAC,ANALISTA SISTEMAS INFORMATION TECHNOLOGY (IT),2008-12
575891146,HOCHSCHILD MINING PLC-COMPANIA MINERA ARES SAC,ANALISTA SISTEMA DE INFORMACION GEOGRAFICO-EVALUACION
575891146,Estilos srl,Desarrollador de Sistemas,2005-12-01,2006/06/01 00:00:00.000
575891146,ITG Solutions SAC,Consultor Jr.,2005-01-01,2005/04/01 00:00:00.000
575891146,perfect pedicab,Pedicab Driver,2003-12-01,2004/04/01 00:00:00.000
576185350,Banco Financiero,Analista de Inteligencia Comercial,2010-04-01,2011/09/01 00:00:00.000
576185350,Corporacion Lindley S.A.,Asistente de planeamiento comercial,2007-11-01,2009/05/01 00:00:00.000
576200131,Municipalidad Distrital Jose Luis Bustamante y Rivero,Tecnologias de la Informacion,2010-07-01,2012
591603704,Grupo Mexico,,2009-01-01,2009/12/01 00:00:00.000
598435451,AIESEC Arequipa,,2008-03-01,2012/03/01 00:00:00.000
598435451,AIESEC,OCP CNV 2012,2011-10-01,2012/02/01 00:00:00.000
598435451,UCSM,,2009-10-01,2010/08/01 00:00:00.000
598435451,The Broadmoor,,2007-12-01,2008/03/01 00:00:00.000
604226373,Universidad Nacional del Altiplano - Puno,Director de Estudios,2000-04-01,2002/08/01 00:00:00.000
607146415,Busey,Bilingual Teller,2010-05-01,2011/06/01 00:00:00.000
619713322,Logica,,2008-01-01,2010/01/01 00:00:00.000
```

Figura 3. 20 Salida temporal de AmigosTrabajos

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

- OUTPUT 3.- Es la tercera salida temporal con los campos siguientes: idusu(ID), tipest(nivel educativo), nomest(nombre de institución), aniest(año de estudio), conest(carrera de estudio) y graest(especialidad). Ver Figura 3.21.

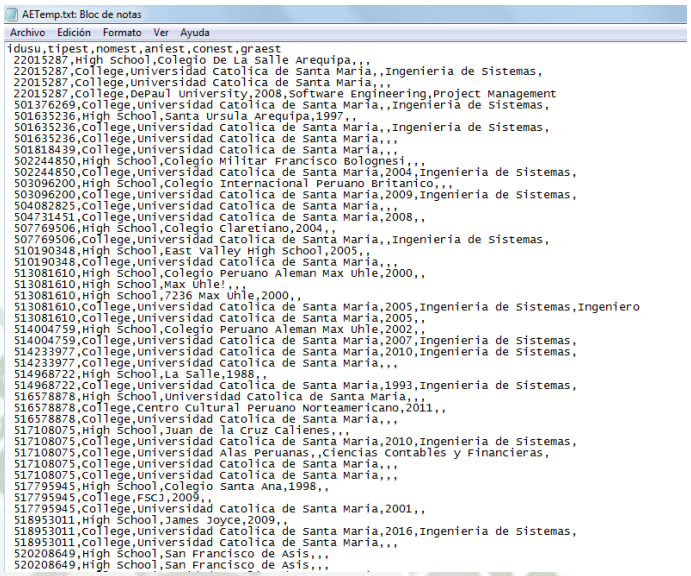


Figura 3. 21 Salida temporal de AmigosEstudios

Fuente: PROPIO.

- OUTPUT 4.- La cuarta salida nos muestra solo el campo: CiudadActual, que vamos a utilizar en nuestro siguiente paso de la metodología. Ver Figura 3.22.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

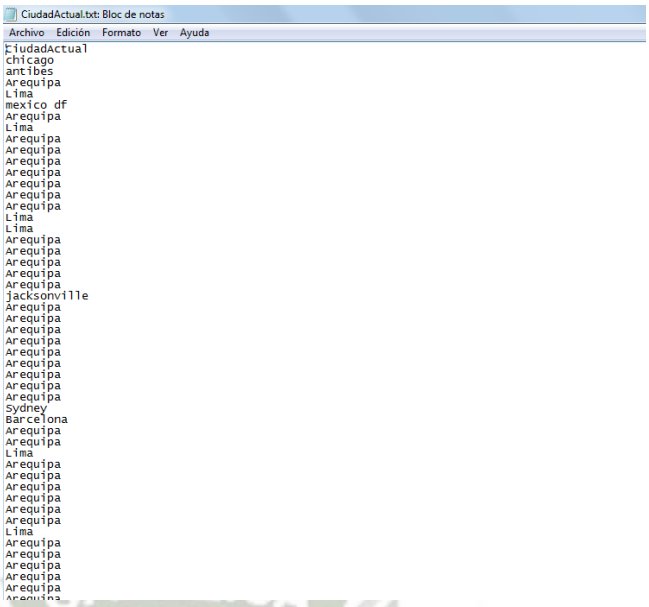


Figura 3. 22. Salida del archivo CiudadActual

Fuente: PROPIO.

- OUTPUT 5.- La quinta salida nos muestra los campos: Empleador y Cargo, que se van a utilizar en nuestro siguiente paso de la metodología. Ver Figura 3.23.

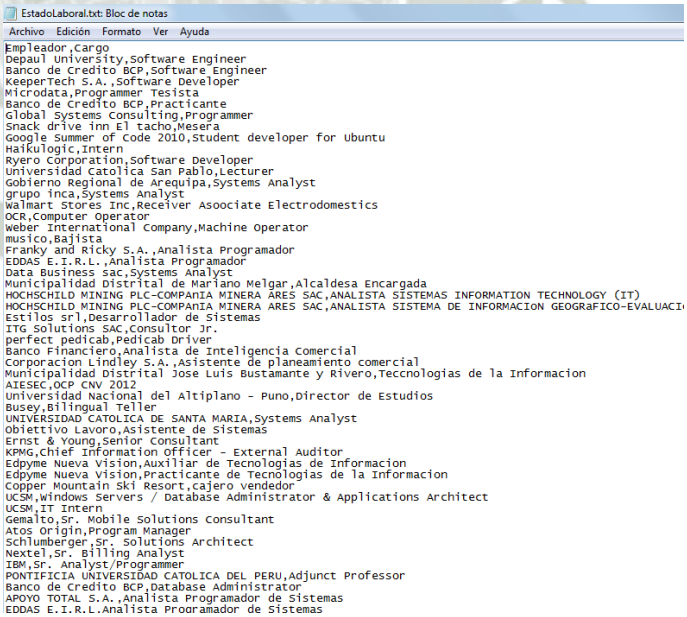


Figura 3. 23 Salida del archivo Estado Laboral.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

Fuente: PROPIO.

- OUTPUT 6.- La sexta salida nos muestra los campos: Nombre, Carrera y Especialidad, que se van a utilizar en nuestro siguiente paso de la metodología. Ver Figura 3.24.

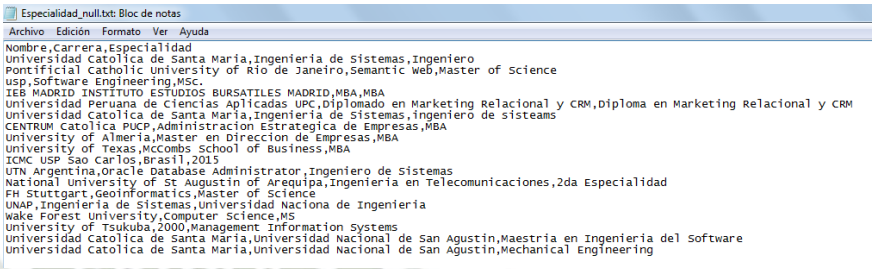


Figura 3. 24 Salida del archivo Especialidad.

Fuente: PROPIO.

- OUTPUT 7.- La séptima salida nos muestra los campos: Nombre y Carrera que se van a utilizar en nuestro siguiente paso de la metodología. Ver Figura 3.25

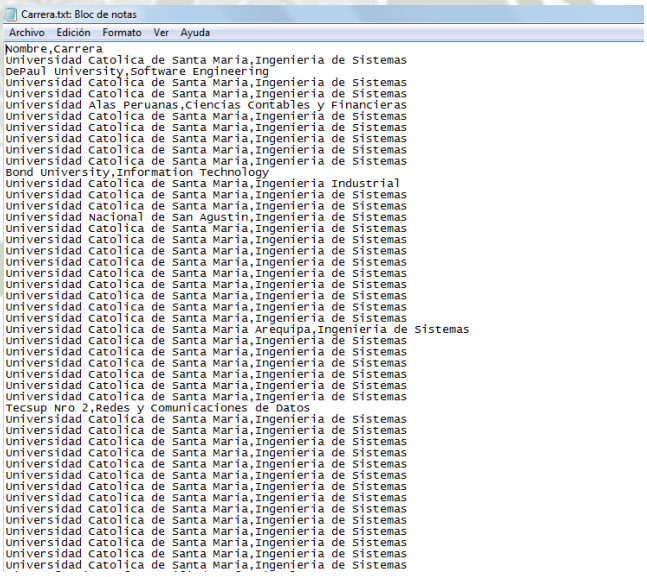


Figura 3. 25 Salida del archivo Carrera

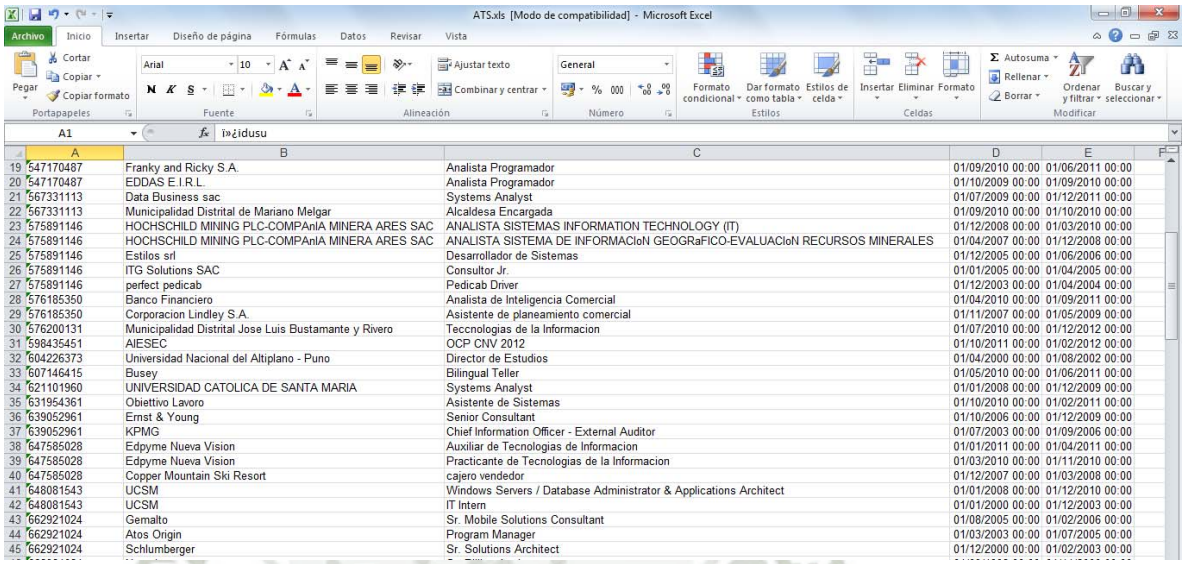
Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

- OUTPUT 8.- Se ha considerado otro tipo de salida en formato Excel (XLS) que muestran los datos producto de las transformaciones anteriormente detalladas. Ver Figuras 3.26 y 3.27

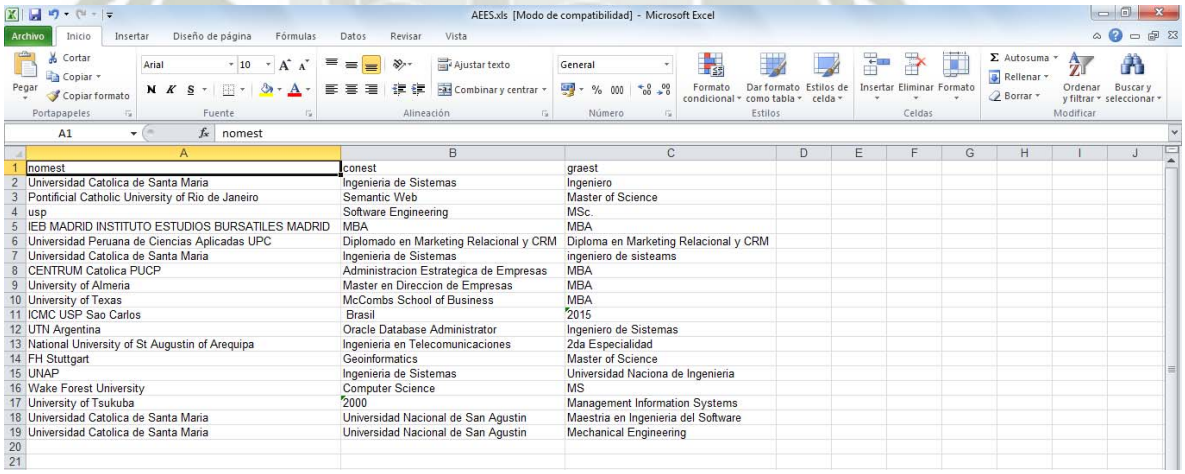


ATS.xls [Modo de compatibilidad] - Microsoft Excel

	A	B	C	D	E
19	547170487	Franky and Ricky S.A.	Analista Programador	01/09/2010 00:00	01/06/2011 00:00
20	547170487	EDDAS E.I.R.L.	Analista Programador	01/10/2009 00:00	01/09/2010 00:00
21	567331113	Data Business sac	Systems Analyst	01/07/2009 00:00	01/12/2011 00:00
22	567331113	Municipalidad Distrital de Mariano Melgar	Alcaldesa Encargada	01/09/2010 00:00	01/10/2010 00:00
23	575891146	HOCHSCHILD MINING PLC-COMPAnIA MINERA ARES SAC	ANALISTA SISTEMAS INFORMATION TECHNOLOGY (IT)	01/12/2008 00:00	01/03/2010 00:00
24	575891146	HOCHSCHILD MINING PLC-COMPAnIA MINERA ARES SAC	ANALISTA SISTEMA DE INFORMACIoN GEOGRaFICO-EVALUACIoN RECURSOS MINERALES	01/04/2007 00:00	01/12/2008 00:00
25	575891146	Estilos srl	Desarrollador de Sistemas	01/12/2005 00:00	01/06/2006 00:00
26	575891146	ITG Solutions SAC	Consultor Jr.	01/01/2005 00:00	01/04/2005 00:00
27	575891146	perfect pedicab	Pedicab Driver	01/12/2003 00:00	01/04/2004 00:00
28	576185350	Banco Financiero	Analista de Inteligencia Comercial	01/04/2010 00:00	01/09/2011 00:00
29	576185350	Corporacion Lindley S.A.	Asistente de planeamiento comercial	01/11/2007 00:00	01/05/2009 00:00
30	576200131	Municipalidad Distrital Jose Luis Bustamante y Rivero	Teccnologias de la Informacion	01/07/2010 00:00	01/12/2012 00:00
31	598435451	AIESEC	OCP CNV 2012	01/10/2011 00:00	01/02/2012 00:00
32	604226373	Universidad Nacional del Altiplano - Puno	Director de Estudios	01/04/2000 00:00	01/08/2002 00:00
33	607146415	Busey	Bilingual Teller	01/05/2010 00:00	01/06/2011 00:00
34	621101960	UNIVERSIDAD CATOLICA DE SANTA MARIA	Systems Analyst	01/01/2008 00:00	01/12/2009 00:00
35	631954361	Obiettivo Lavoro	Asistente de Sistemas	01/10/2010 00:00	01/02/2011 00:00
36	639052961	Ernst & Young	Senior Consultant	01/10/2006 00:00	01/12/2009 00:00
37	639052961	KPMG	Chief Information Officer - External Auditor	01/07/2003 00:00	01/09/2006 00:00
38	647585028	Edpyme Nueva Vision	Auxiliar de Tecnologias de Informacion	01/01/2011 00:00	01/04/2011 00:00
39	647585028	Edpyme Nueva Vision	Practicante de Tecnologias de la Informacion	01/03/2010 00:00	01/11/2010 00:00
40	647585028	Copper Mountain Ski Resort	cajero vendedor	01/12/2007 00:00	01/03/2008 00:00
41	648081543	UCSM	Windows Servers / Database Administrator & Applications Architect	01/01/2008 00:00	01/12/2010 00:00
42	648081543	UCSM	IT Intern	01/01/2000 00:00	01/12/2003 00:00
43	662921024	Gemalto	Sr. Mobile Solutions Consultant	01/08/2005 00:00	01/02/2006 00:00
44	662921024	Atos Origin	Program Manager	01/03/2003 00:00	01/07/2005 00:00
45	662921024	Schlumberger	Sr. Solutions Architect	01/12/2000 00:00	01/02/2003 00:00

Figura 3. 26 Salida del archivo Estado Laboral en Excel.

Fuente: PROPIO.



AEES.xls [Modo de compatibilidad] - Microsoft Excel

	A	B	C
1	nomest	iconest	graest
2	Universidad Catolica de Santa Maria	Ingenieria de Sistemas	Ingeniero
3	Pontificial Catholic University of Rio de Janeiro	Semantic Web	Master of Science
4	usp	Software Engineering	MSc.
5	IEB MADRID INSTITUTO ESTUDIOS BURSATILES MADRID	MBA	MBA
6	Universidad Peruana de Ciencias Aplicadas UPC	Diplomado en Marketing Relacional y CRM	Diploma en Marketing Relacional y CRM
7	Universidad Catolica de Santa Maria	Ingenieria de Sistemas	ingeniero de sisteams
8	CENTRUM Catolica PUCP	Administracion Estrategica de Empresas	MBA
9	University of Almeria	Master en Direccion de Empresas	MBA
10	University of Texas	McCombs School of Business	MBA
11	ICMC USP Sao Carlos	Brasil	2015
12	UTN Argentina	Oracle Database Administrator	Ingeniero de Sistemas
13	National University of St Augustin of Arequipa	Ingenieria en Telecomunicaciones	2da Especialidad
14	FH Stuttgart	Geoinformatics	Master of Science
15	UNAP	Ingenieria de Sistemas	Universidad Naciona de Ingenieria
16	Wake Forest University	Computer Science	MS
17	University of Tsukuba	2000	Management Information Systems
18	Universidad Catolica de Santa Maria	Universidad Nacional de San Agustin	Maestria en Ingenieria del Software
19	Universidad Catolica de Santa Maria	Universidad Nacional de San Agustin	Mechanical Engineering
20			
21			

Figura 3. 27 Salida del archivo Especialidad en Excel.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

PASO 4.- PREPARACION DE ARCHIVOS ARFF.

Una vez obtenidos los datos producto de las transformaciones en los archivos: CiudadActual.txt, EstadoLaboral.txt, y Especialidad.txt; procedemos a cambiar el formato TXT por ARFF (*Attribute-Relation File Format)*. Iniciamos cambiando cada extensión de cada archivo de TXT por ARFF.

A continuación procedemos a desarrollar la estructura de los archivos ARFF que cuentan con dos partes principales: la cabecera y los datos o instancias.

En la cabecera, cada línea inicia con símbolo de @. En la segunda parte se encuentran los datos separados por comas. Todas las líneas que comiencen con un % corresponden a comentarios. La cabecera inicia con el tag @relation indicando el nombre de la relación representada por los datos.

Los atributos se definen con el tag @attribute seguido de su nombre y tipo, uno por línea (Ver Figura 3.28). Los atributos son de diferentes tipos de acorde con los valores que puedan tomar.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

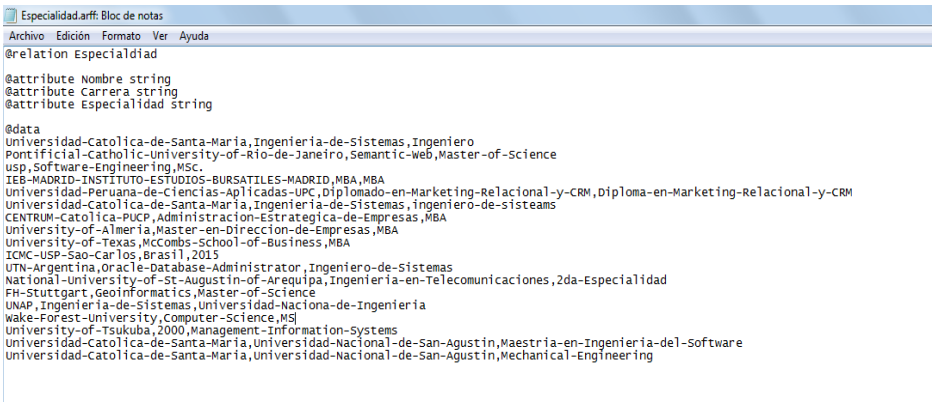


Figura 3.28 Estructura de archivos ARFF

Fuente: PROPIO.

3.2.1.4. FASE 4: MODELAJE.

Para la realización de este paso se utilizaron las técnicas de Minería de Datos utilizando la herramienta WEKA. A continuación se muestran las técnicas a utilizar por cada objetivo de la Minería:

Objetivo de minería	Técnica
Segmentar a los egresados por residencia actual, centros de trabajo, empleador, puesto en el que se desempeña, fecha de inicio y fin de trabajo, nivel de educación, postgrado, especialidad e intereses en común en el área de sistemas. Analizar y categorizar los clústeres obtenidos de acuerdo a los ítems antes mencionados como paso analítico para el próximo objetivo. Obtener patrones y segmentos tanto en el campo educativo como en la experiencia profesional, el cual nos permita una mejor gestión de los egresados en el Programa Profesional de Ingeniería de Sistemas (PPIS).	Algoritmos de Clustering: • Algoritmo Simple k-Means • Algoritmo DBSCAN

Tabla 3. 10 Objetivos y Técnicas de Minería de Datos

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

ALGORITMOS DE CLUSTERING.

En esta parte describiremos la funcionalidad de los algoritmo a utilizar, esta funcionalidad la podremos observar mejor mediante la herramienta WEKA.

Algoritmo Simple K-Means.

Es uno de los más simples y conocidos algoritmos de agrupamiento, sigue una forma fácil y simple para dividir una base de datos dada en k grupos (fijados a priori).

Básicamente este algoritmo busca formar Clústeres (grupos) los cuales serán representados por K objetos. Cada uno de estos K objetos es el valor medio de los objetos que pertenecen a dicho grupo, entonces decimos que es un algoritmo eficaz para particionar.
Inicialmente se seleccionan K objetos del conjunto de entrada. Estos K Objetos serán los centroides iniciales de los K-grupos.
Se calculan las distancias de los objetos (datos) a cada uno de los centroides. Los Datos (Objetos) se asignan a aquellos grupos cuya distancia es mínima con respecto a todos los centroides.
Se actualizan los centroides como el valor medio de todos los objetos asignados a ese grupo. Se repite el paso 2 y 3 hasta que se satisface algún criterio de convergencia.

Finalmente quedarán agrupados por Clústeres, los grupos de simulaciones según la cantidad de Clústeres que se definió en el momento de ejecutar el algoritmo. El problema del empleo de estos esquemas es que fallan cuando los puntos de un grupo están muy cerca del centroide de otro grupo, también cuando los grupos tienen diferentes tamaños y formas.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Algoritmo DBSCAN

Este algoritmo es el primero que es basado en densidad. Se definen los conceptos de punto central (puntos que tienen en su vecindad una cantidad de puntos mayor o igual que un umbral especificado), borde y ruido.

El algoritmo comienza seleccionando un punto p arbitrario, si p es un punto central, se comienza a construir un grupo y se ubican en su grupo todos los objetos denso-alcanzables desde p.
Si p no es un punto central se visita otro objeto del conjunto de datos.
El proceso continúa hasta que todos los objetos han sido procesados.

Los puntos que quedan fuera de los grupos formados se llaman puntos ruido, los puntos que no son ni ruido ni centrales se llaman puntos borde.
De esta forma DBSCAN construye grupos en los que sus puntos son: o puntos centrales o puntos borde, un grupo puede tener más de un punto central.

Antes de comenzar con la aplicación de las técnicas de WEKA a los datos de este dominio, es conveniente tomar en consideración los objetivos perseguidos en el análisis.
De esta manera, el proceso de análisis de datos (proceso KDD), permitirá dirigir la búsqueda y hacer refinamientos, con una interpretación adecuada de los resultados generados.
En nuestro caso, el principal objetivo es el poder hacer una mejor gestión sobre los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS). Algunas de las preguntas que podemos plantearnos a responder como objetivo del análisis podrían ser las siguientes:

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

- ¿En qué ciudades se encuentran el mayor número de egresados del Programa Profesional de Ingeniería de Sistemas (PPIS)?

- ¿En qué empresas laboran y que cargos ocupan los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS)?

- ¿En qué países y que especializaciones realizan los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS)?

Para poder continuar con la estructura del proyecto y la realización del mismo en el WEKA nos enfocaremos en gran parte en la Opción Explorer. Cuando seleccionamos esta opción, se crea una ventana con 6 pestañas en la parte superior, cuenta con etapas independientes, en las cuales se pueden realizar operaciones sobre los datos:

- Preprocess: Se selecciona la fuente de datos y a la vez su preparación mediante el filtrado.
- Clasify: Facilidades para aplicar esquemas de clasificación, entrenar modelos y evaluar su precisión
- Cluster: Encontramos los diferentes Algoritmos de agrupamiento que se encuentran en la herramienta.
- Associate: Encontramos los diferentes Algoritmos de búsqueda de reglas de asociación
- Select Attributes: Se realiza una búsqueda supervisada de subconjuntos de atributos representativos.
- Visualize: Herramienta interactiva en la cual se presenta una forma gráfica en 2D.

Además de estas pestañas de selección, en la parte inferior de la ventana aparecen dos elementos comunes. Uno es el botón de “Log”,

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

que al activarlo presenta una ventana textual donde se indica la secuencia de todas las operaciones que se han llevado a cabo dentro de la opción "Explorer", sus tiempos de inicio y fin, así como los mensajes de error más frecuentes. Junto al botón de log aparece un icono de actividad (el pájaro WEKA, que se mueve cuando se está realizando alguna tarea) y un indicador de status, que indica qué tarea se está realizando en este momento dentro de la opción Explorer

3.2.1.5. FASE 5: EVALUACIÓN.

En esta fase vamos a medir diferentes patrones para poder tener una mejor gestión sobre los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS), se interactúa con el o los algoritmos de Minería de Datos para guiar la búsqueda hacia patrones interesantes. Esta fase se desarrollara en el Capitulo 4.

3.2.1.6. FASE 6: IMPLEMENTACIÓN.

En esta fase procedemos a implementar un sistema de poyo para la visualización de los datos obtenidos en los diferentes procesos de tratamiento de la información que nos van a permitir diferenciar características específicas como: carrera, especialidad, etc. Así lograremos una mejor interacción con el usuario. Ver Anexo A: Implementación del Sistema.

Los modelos y reglas obtenidas podrán ser utilizados por el Programa Profesional de Ingeniería de Sistemas (PPIS), Por diferentes Programas de la Universidad Católica de Santa María UCSM y en otras investigaciones sobre la gestión de información de los egresados. Con las relaciones y patrones encontrados se podrán trazar estrategias que permitan gestionar de una mejor

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE
TESIS UCSM

CATÓLICA
DE SANTA MARÍA

manera la información de los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS).

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

CAPITULO IV

4. EVALUACION DE RESULTADOS

Basándonos en el cumplimiento de los objetivos del negocio y no desde el punto de vista general evaluamos el modelo. Revisamos el proceso teniendo en cuenta los resultados obtenidos. Vamos a poder encontrar diferentes grupos en los cuales los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS) se van a clasificar.

4.1. EVALUACION

En relación a la variable dependiente “Propuesta de Gestión de información de Agrupamiento (Clustering) de egresados.” y tomando en cuenta los indicadores: Ergonómica**,** Multiplataforma y Dinámicase ha evaluado en referencia a ciudad actual, cargo que desempeñan, empresas empleadoras, especialidades y centros de estudios.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

PREGUNTA N°1

- ¿En qué ciudades se encuentran el mayor número de egresados del Programa Profesional de Ingeniería de Sistemas (PPIS)?

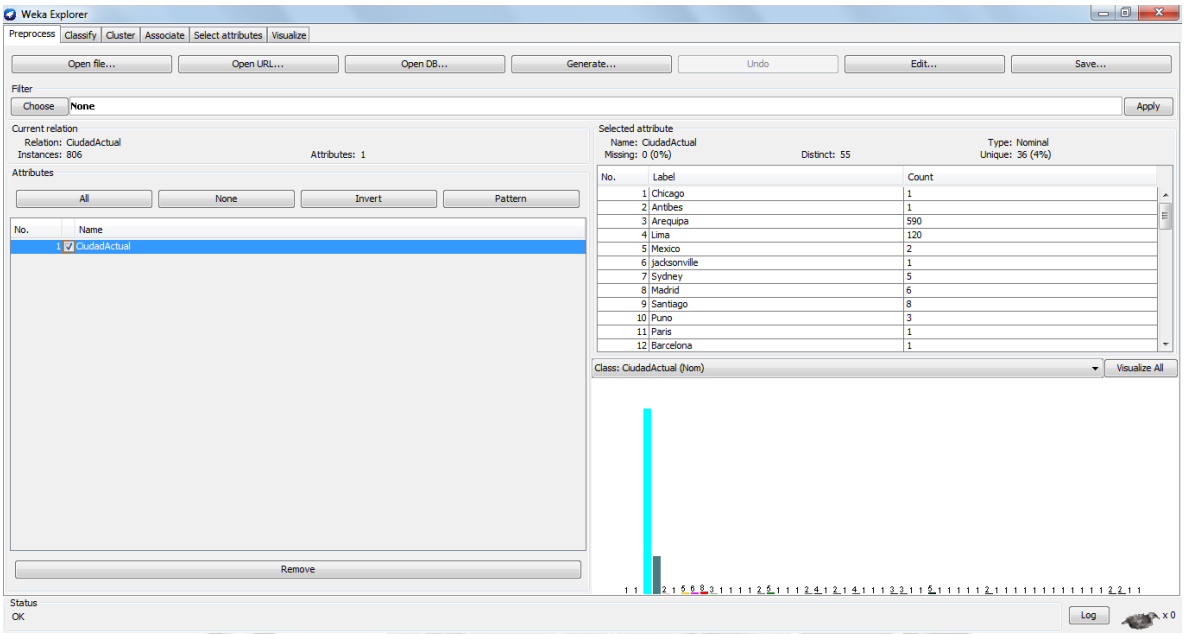


Figura 4. 1 Carga de Datos "CIUDAD ACTUAL"

Fuente: PROPIO.

DESCRIPCION 4.1.- Utilizando la data de "CiudadActual" con un total de 806 egresados podemos deducir que en la ciudad donde se encuentran la mayoría de los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS) es en Arequipa con un numero de 590, seguida de la ciudad de Lima con 120 egresados respectivamente y otras ciudades como por ejemplo Santiago de Chile, Madrid España y Sídney Australia con un número que no excede a 8 egresados.

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

4.11.ALGORITMO SIMPLE K-MEANS: CONFIGURACION Y VISUALIZACION DE RESULTADOS DE CIUDAD ACTUAL.

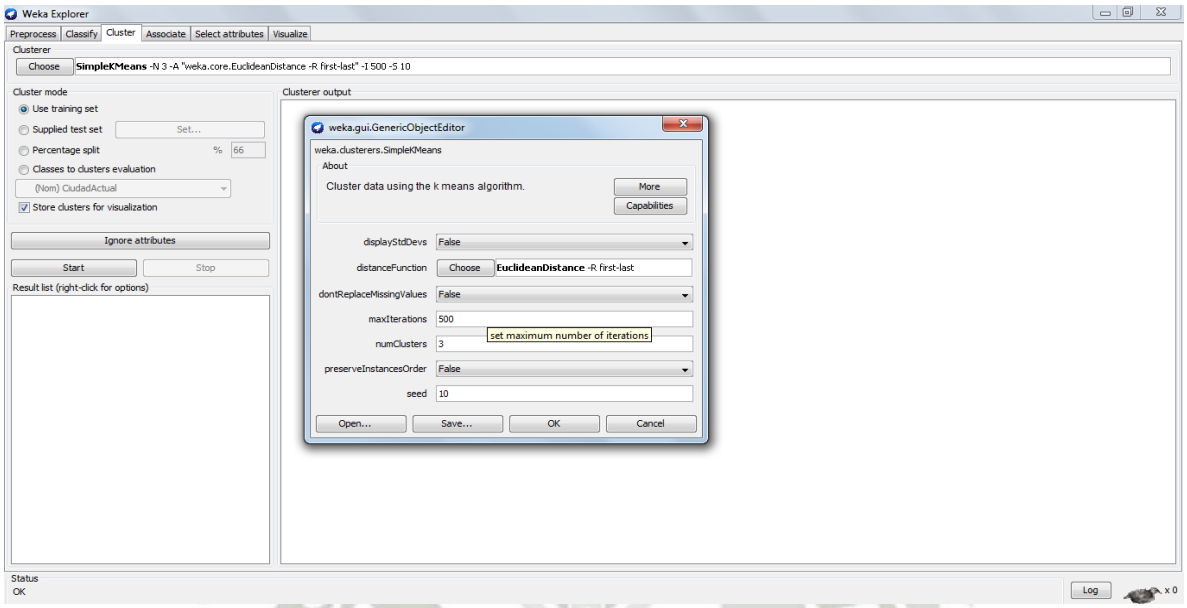


Figura 4. 2 Configuración de algoritmo Simple K-Means en Tabla Ciudad Actual.

Fuente: PROPIO.

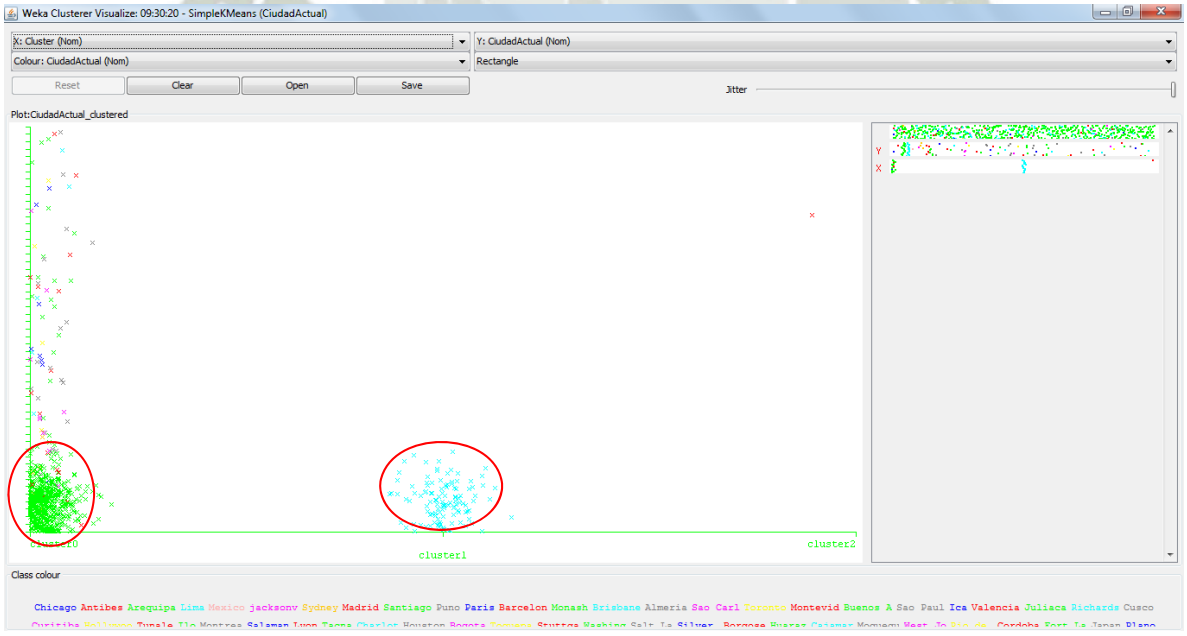


Figura 4. 3 Visualzización de Clústeres en Simple K-Means en Tabla Ciudad Actual.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

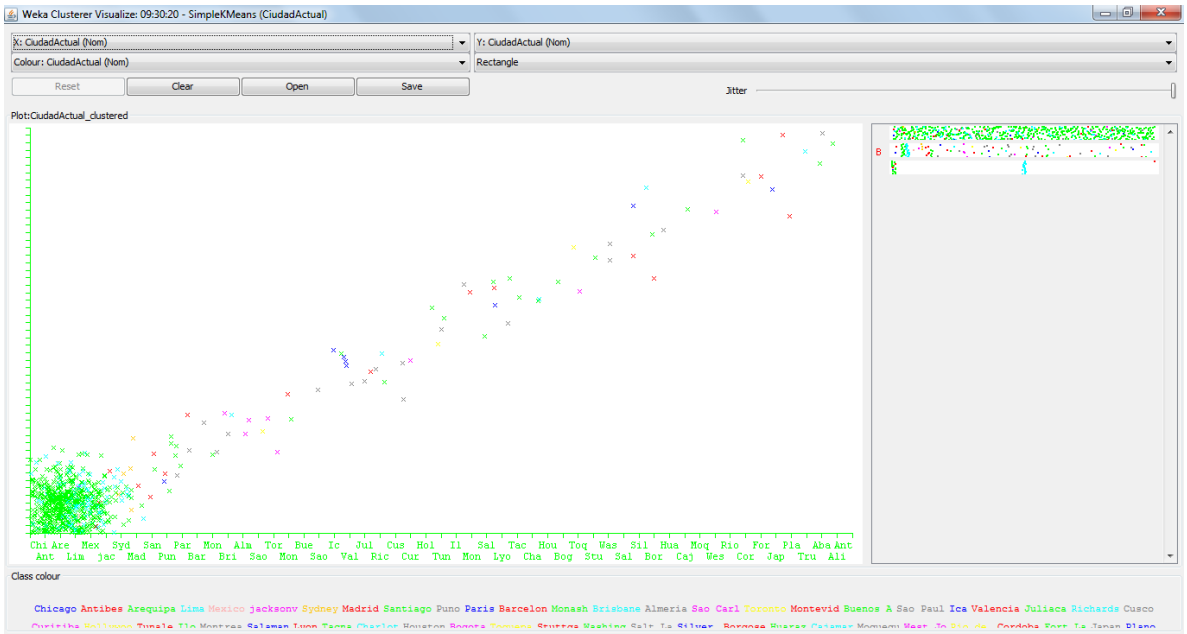


Figura 4. 4 Visualización de Ciudad Actual en Simple K-Means.

Fuente: PROPIO.

DESCRIPCION 4.2.- Teniendo en cuenta el algoritmo SIMPLE K-MEANS, comenzamos el proceso de evaluación configurando los ítems del recuadro que se observa en la Figura 4.2, siendo de mayor importancia el ítem numCluster (número de clústeres) el cual tendrá un valor de 3, ya que mientras menos clústeres sean podremos encontrar grupos más específicos, también verificamos la distancefunction (función de distancia) y escogemos la distancia: EuclideanDistance, los demás ítems los trabajamos por default. Analizando el proceso de Clustering mediante el algoritmo Simple K-Means podemos indicar que los clústeres o grupos son bien definidos e indicamos que la mayor parte de los egresados se establecen en la ciudad de Arequipa. Ver Figuras 4.3 y 4.4

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

4.1.2. ALGORITMO DBSCAN: CONFIGURACION Y VISUALIZACION DE RESULTADOS DE CIUDAD ACTUAL

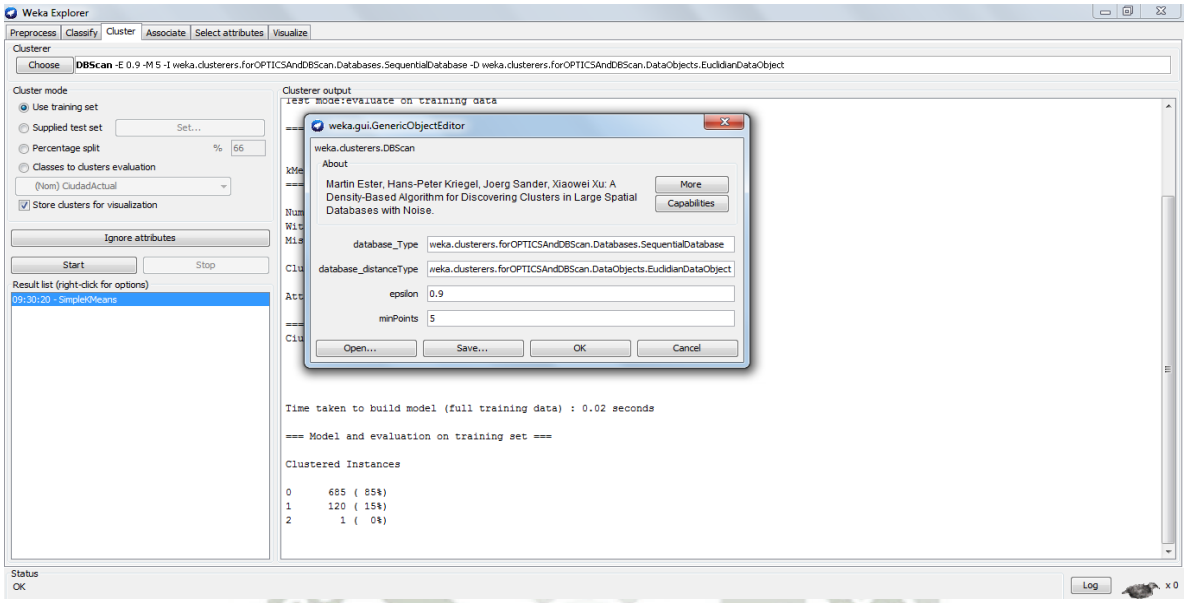


Figura 4. 5 Configuración de DBSCAN en Tabla Ciudad Actual

Fuente: PROPIO.

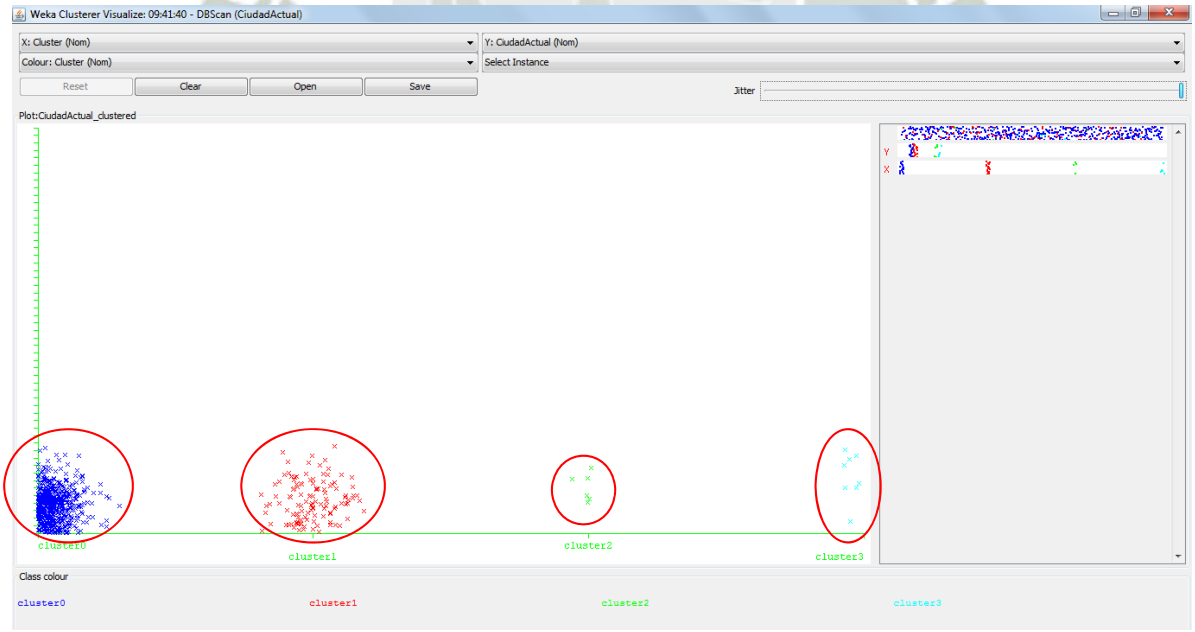


Figura 4. 6 Visualización de Clústeres en DBSCAN en Tabla Ciudad Actual

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

DESCRIPCION 4.3.- Teniendo en cuenta el algoritmo DBSCAN, comenzamos el proceso de evaluación, configurando los ítems del recuadro que se observa en la Figura 4.5, siendo de mayor importancia el ítem minPoints (mínimo de puntos) el cual tendrá un valor de 5 ya que será el mínimo de puntos que concuerdan, también verificamos la DataBase Distance Type (tipo de distancia en a base de datos) y por default nos muestra la distancia: EuclideanDistance, los demás ítems los trabajamos por default. Analizando el proceso de Clustering mediante el algoritmo DBSCAN podemos indicar que los clústeres o grupos son bien definidos e indicamos que la mayor parte de los egresados se establecen en la ciudad de Arequipa. Ver Figura 4.6

COMPARACION 4.1.- La tendencia a establecerse en la ciudad de Arequipa de los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS), la podemos verificar de forma similar en ambos Algoritmos ya sea en Simple K-Means como en DBSCAN

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

PREGUNTA N°2

- ¿En qué empresas laboran y que cargos ocupan los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS)?

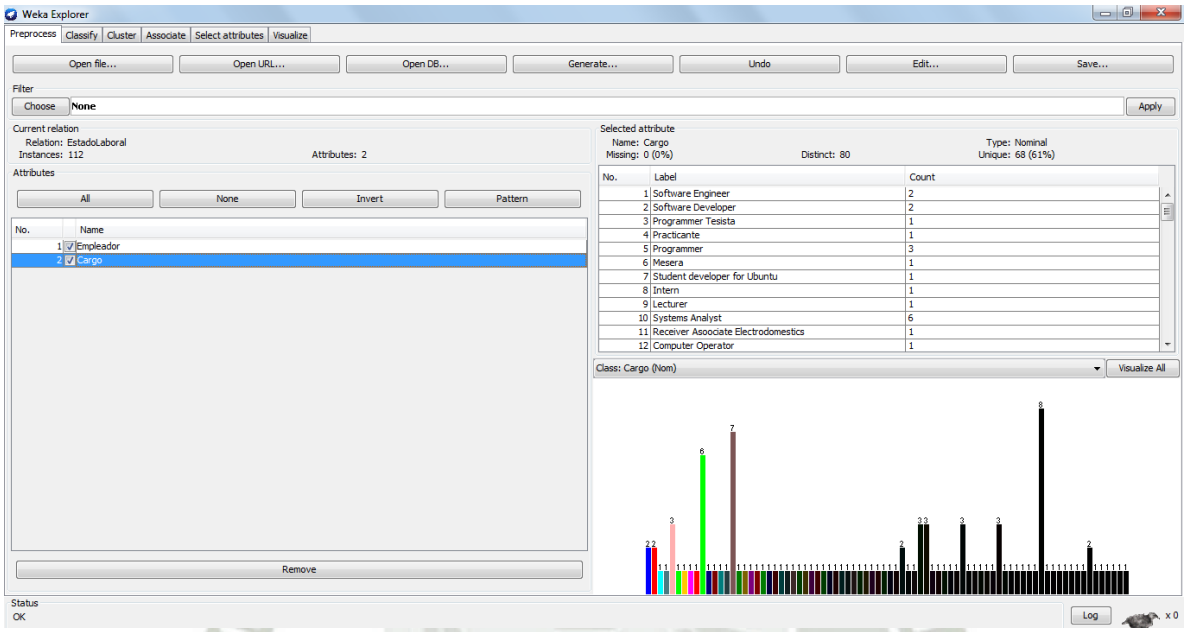


Figura 4. 7 Carga de datos de “ESTADO LABORAL”

Fuente: PROPIO.

DESCRIPCION 4.4.- Utilizando la data de “EstadoLaboral” con un total de 112 egresados que ocupan un Cargo y que cuentan con un Empleador deducimos que el Cargo en el que más se desempeñan los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS), es en Profesor, Analista Programador y System Analist, y el centro de trabajo donde se encuentran o han estado es el Banco de Crédito del Perú (BCP).

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

4.1.3. ALGORITMO SIMPLE K-MEANS: VISUALIZACION DE RESULTADOS DE ESTADO LABORAL

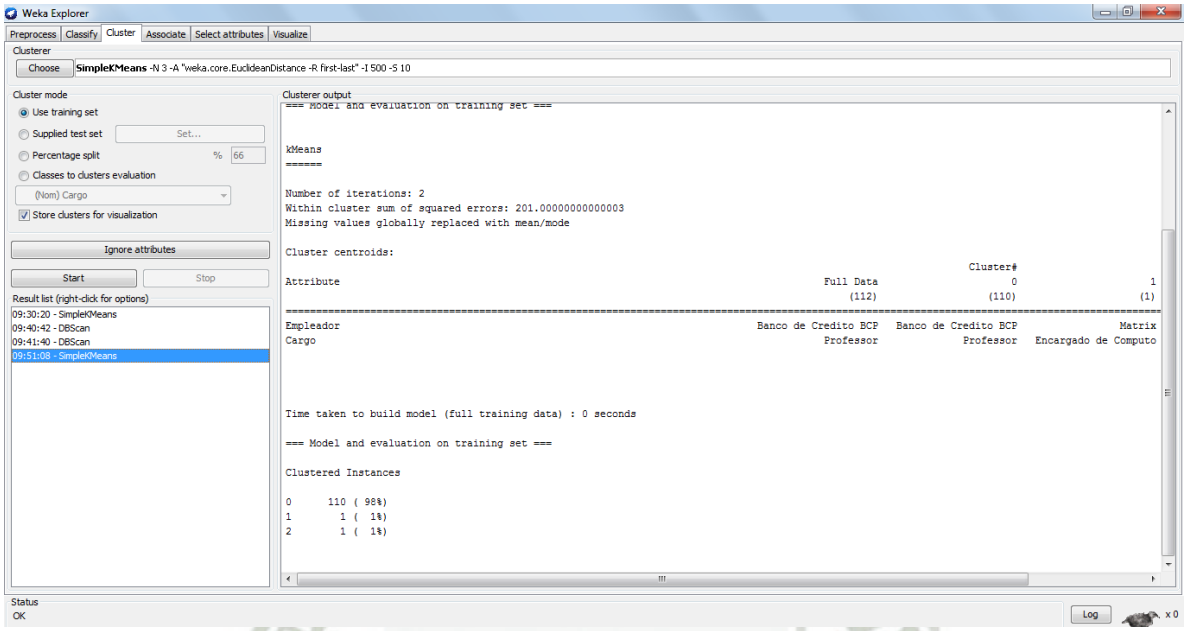


Figura 4. 8 Resultados de Tabla Estado Laboral en Simple K-Means.

Fuente: PROPIO.

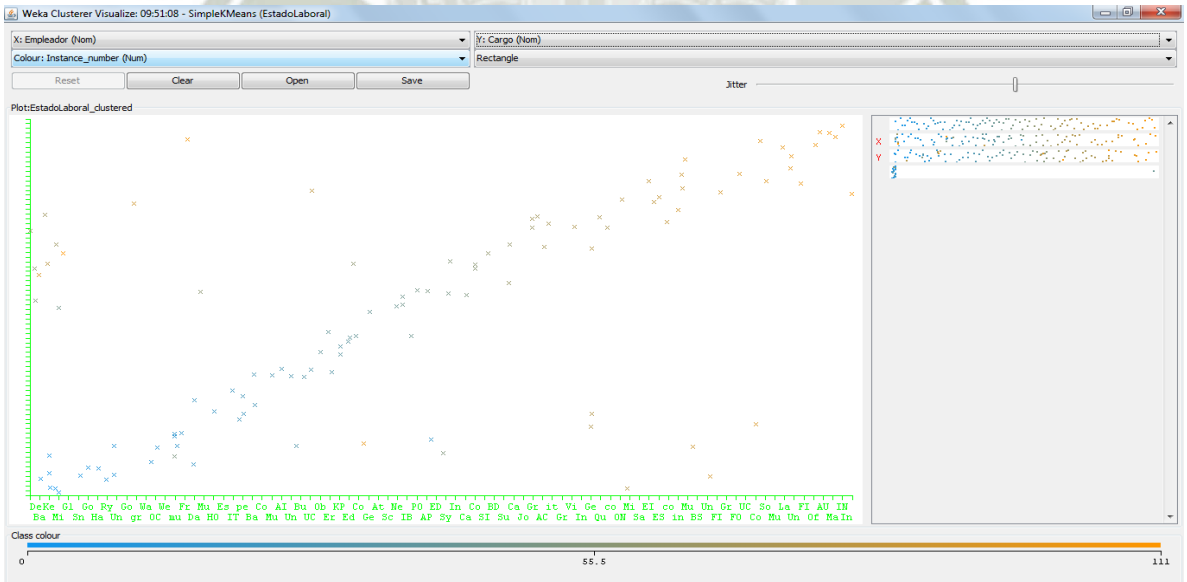


Figura 4. 9 Visualización de Clústeres en Simple K-Means en Estado Laboral.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

DESCRIPCION 4.5.- Teniendo en cuenta el algoritmo SIMPLE K-MEANS, comenzamos el proceso de evaluación configurando los ítems del recuadro que se observa en la Figura 4.2, siendo de mayor importancia el ítem numCluster (numero de clústeres) el cual tendrá un valor de 3, ya que mientras menos clústeres sean podremos encontrar grupos más específicos, también verificamos la distancefunction (función de distancia) y escogemos la distancia: EuclideanDistance, los demás ítems los trabajamos por default. Analizando el proceso de Clustering mediante el algoritmo Simple K-Means y segmentando a los egresados encontramos que un porcentaje significativo tiende a enfocarse a la Banca y Finanzas como también a involucrarse en el Plano Académico, estos grupos encontrados los podemos ver en la Figura 4.9.

4.1.4. ALGORITMO DBSCAN: VISUALIZACION DE RESULTADOS DE ESTADO LABORAL

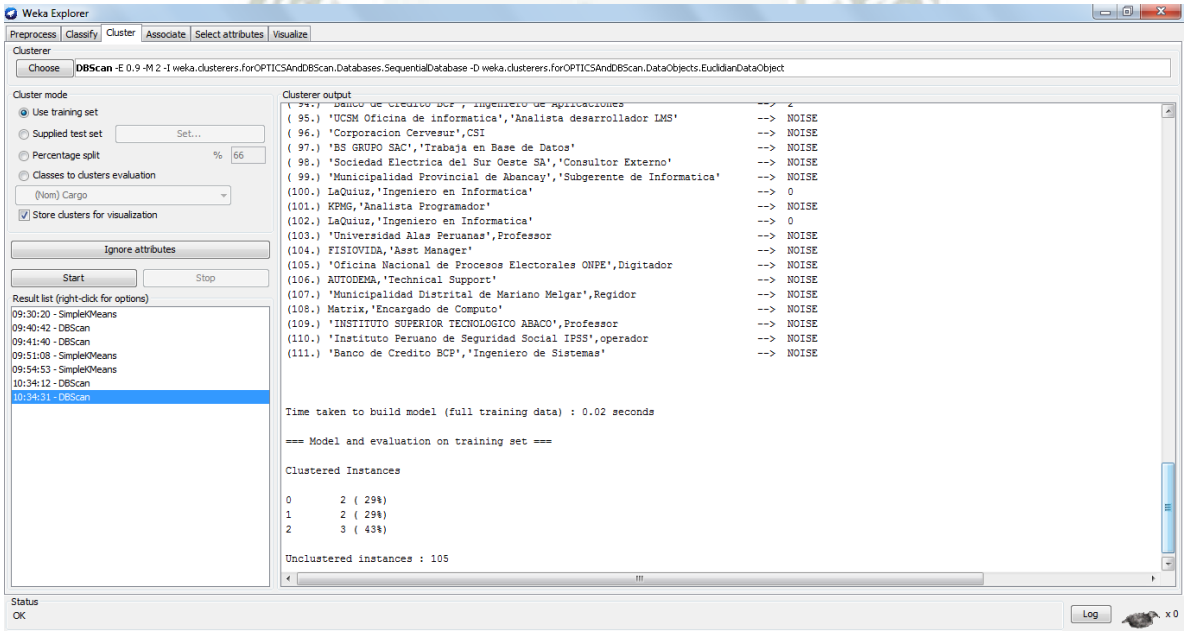


Figura 4. 10 Resultados de Tabla Estado Laboral en DBSCAN

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

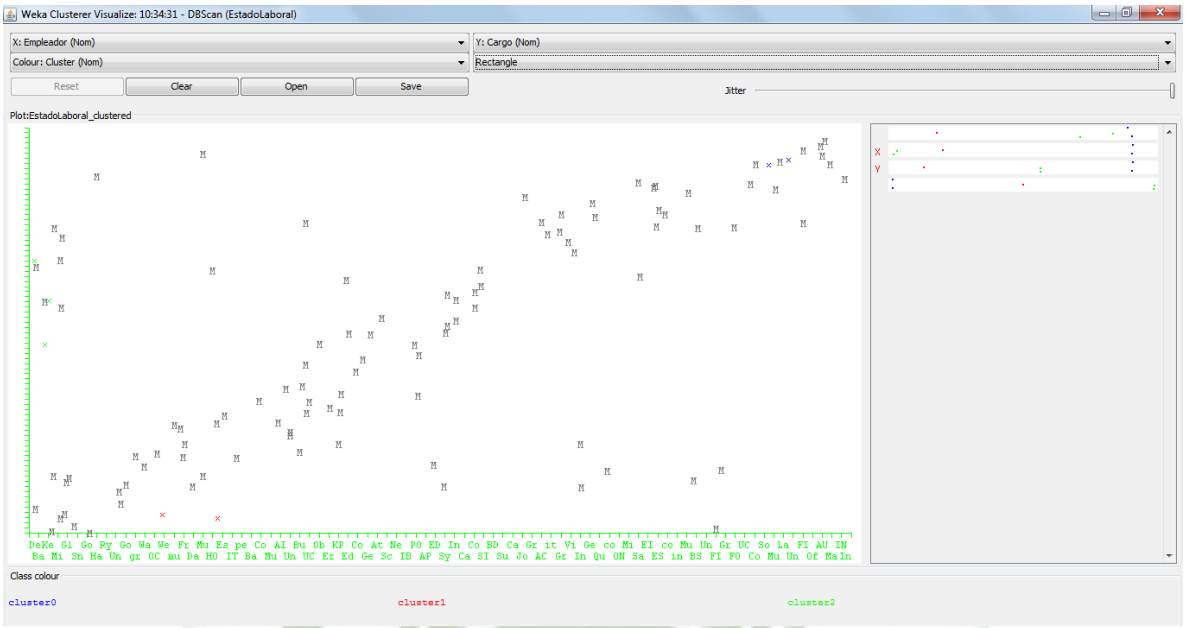


Figura 4. 11 Visualización de Clústeres en DBSCAN en Tabla Estado Laboral.

Fuente: PROPIO.

DESCRIPCION 4.6.- Teniendo en cuenta el algoritmo DBSCAN, comenzamos el proceso de evaluación configurando los ítems del recuadro que se observa en la Figura 4.5, siendo de mayor importancia el ítem minPoints (mínimo de puntos) el cual tendrá un valor de 2 ya que será el mínimo de puntos que concuerdan, también verificamos la DataBase Distance Type (tipo de distancia en a base de datos) y por default nos muestra la distancia: EuclideanDistance, los demás ítems los trabajamos por default. Analizando el proceso de Clustering mediante el algoritmo DBSCAN y dado que este algoritmo trabaja en función a densidad nos permite visualizar 3 clústeres o grupos los cuales nos muestran coincidencias entre el cargo que ocupan y la empresa donde laboran logrando así encontrar que un grupo está enfocado a la Banca y Finanzas, otro grupo se desarrolla en el área de Informática.

COMPARACION 4.2.- En ambos análisis efectuados mediante los algoritmos Simple K-Means y DBSCAN encontramos similitud en el grupo que se enfoca a la Banca y Finanzas, por el contrario también se ha descubierto dos grupos significativos, uno que se desarrolla en el área de Informática y el otro que incursiona en el Plano Académico.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

PREGUNTA N°3

- ¿En qué países y que especializaciones realizan los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS)?

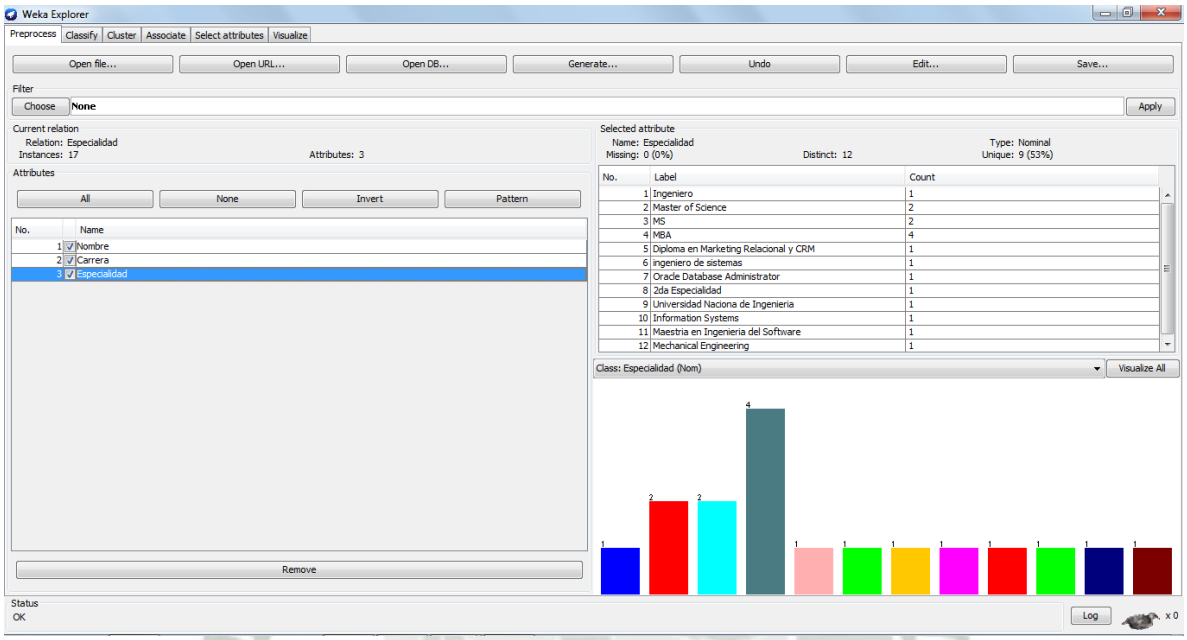


Figura 4. 12 Carga de Datos de “Especialidad”

Fuente: PROPIO.

DESCRIPCION 4.7.- Utilizando la data de “Especialidad” con un total de 17 egresados que llevan una Especialidad y en donde también se muestran que carrera han llevado y en que Centro de Estudios han realizado la especialidad, deducimos que los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS), tienden a especializarse en MBA y Master of Science.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

4.1.5. ALGORITMO SIMPLE K-MEANS: VISUALIZACION DE RESULTADOS DE ESPECIALIDAD

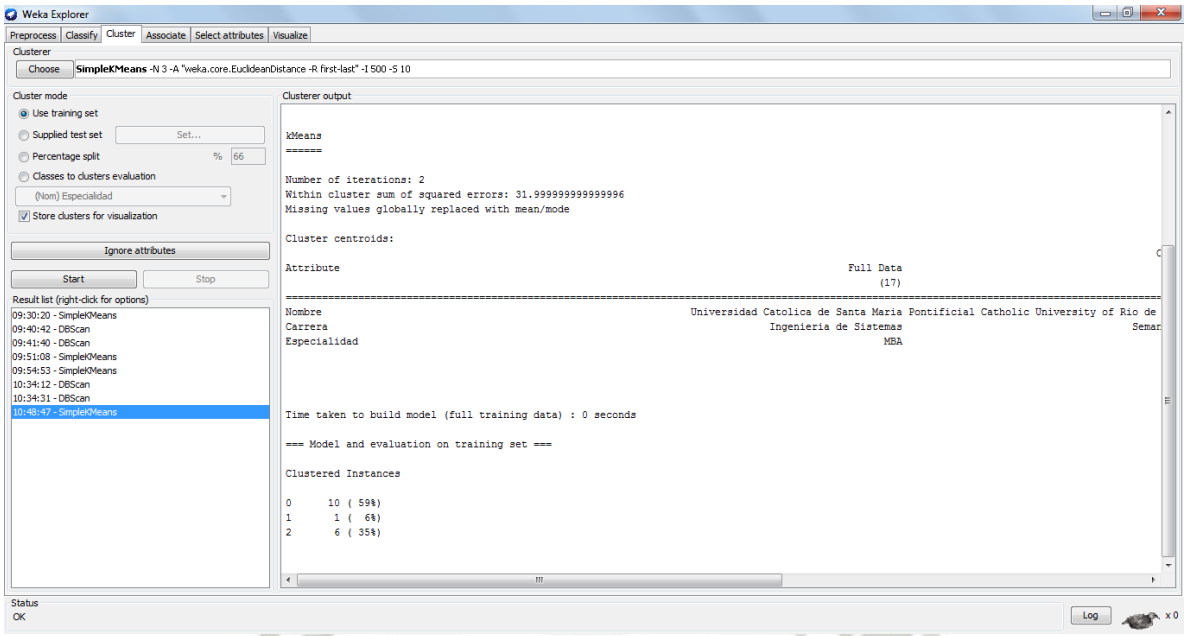


Figura 4. 13 Resultado de Tabla Especialidad en Simple K-Means (I)

Fuente: PROPIO.

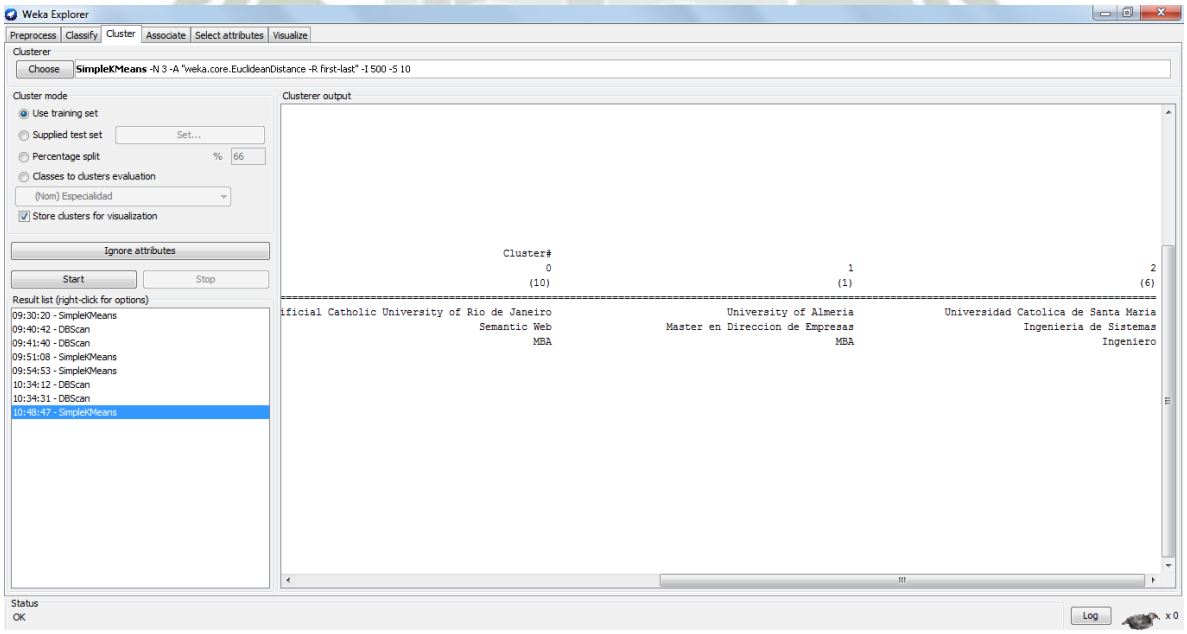


Figura 4. 14 Resultado de Tabla Especialidad en Simple K-Means (II)

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

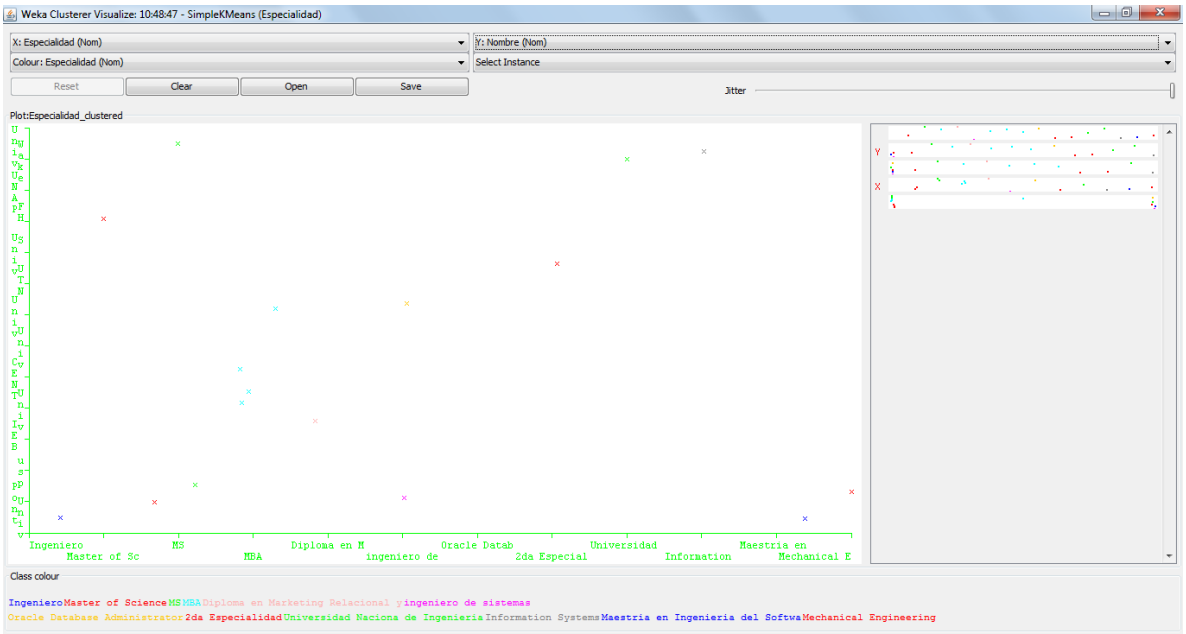


Figura 4. 15 Visualización de Clústeres en Simple K-Means en Tabla Especialidad

Fuente: PROPIO.

DESCRIPCION 4.8.- Teniendo en cuenta el algoritmo SIMPLE K-MEANS, comenzamos el proceso de evaluación configurando los ítems del recuadro que se observa en la Figura 4.2, siendo de mayor importancia el ítem numCluster (numero de clústeres) el cual tendrá un valor de 3, ya que mientras menos clústeres sean podremos encontrar grupos mas específicos, también verificamos la distancefunction (función de distancia) y escogemos la distancia: EuclideanDistance, los demás ítems los trabajamos por default.

Analizando el proceso de Clustering mediante el algoritmo Simple K-Means y segmentando a los egresados encontramos que realizan la especialidad en otros países, lo cual nos da a entender que los egresados prefieren hacer sus especializaciones en el extranjero, en este mismo estudio se deduce que se especializan mas en el área de Desarrollo Web así como aspirar a lograr el grado académico en MBA ver Figuras 4.13. 4.14 y 4.15

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

4.1.6. ALGORITMO DBSCAN: VISUALIZACION DE RESULTADOS DE ESPECIALIDAD.

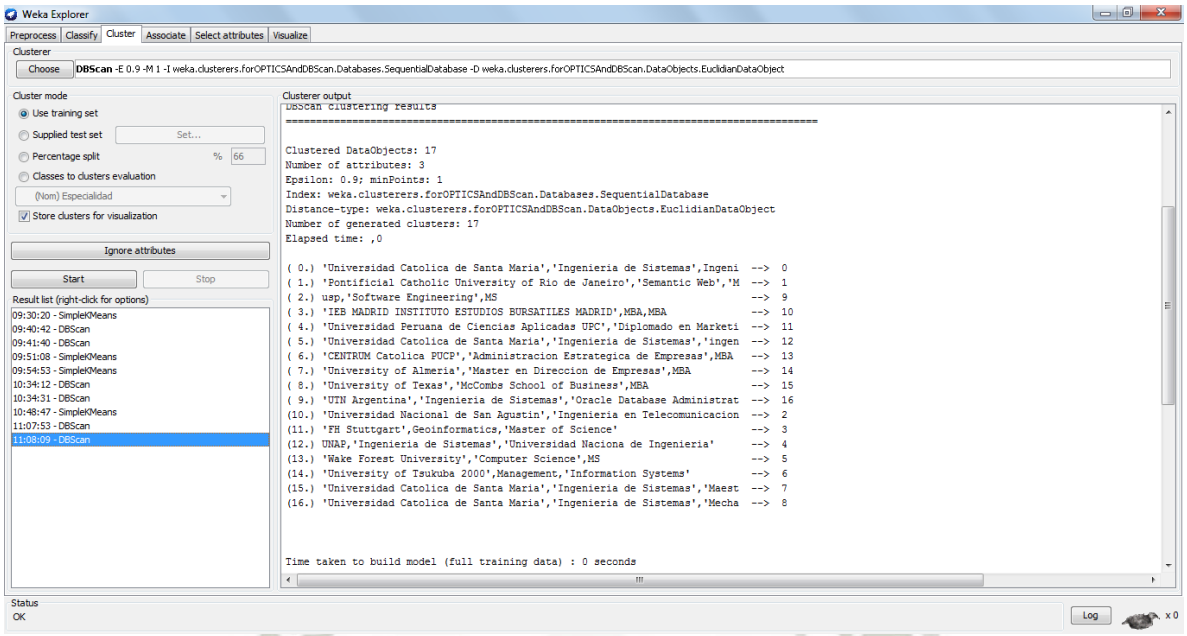


Figura 4. 16 Resultado de Tabla Especialidad en DBSCAN

Fuente: PROPIO.

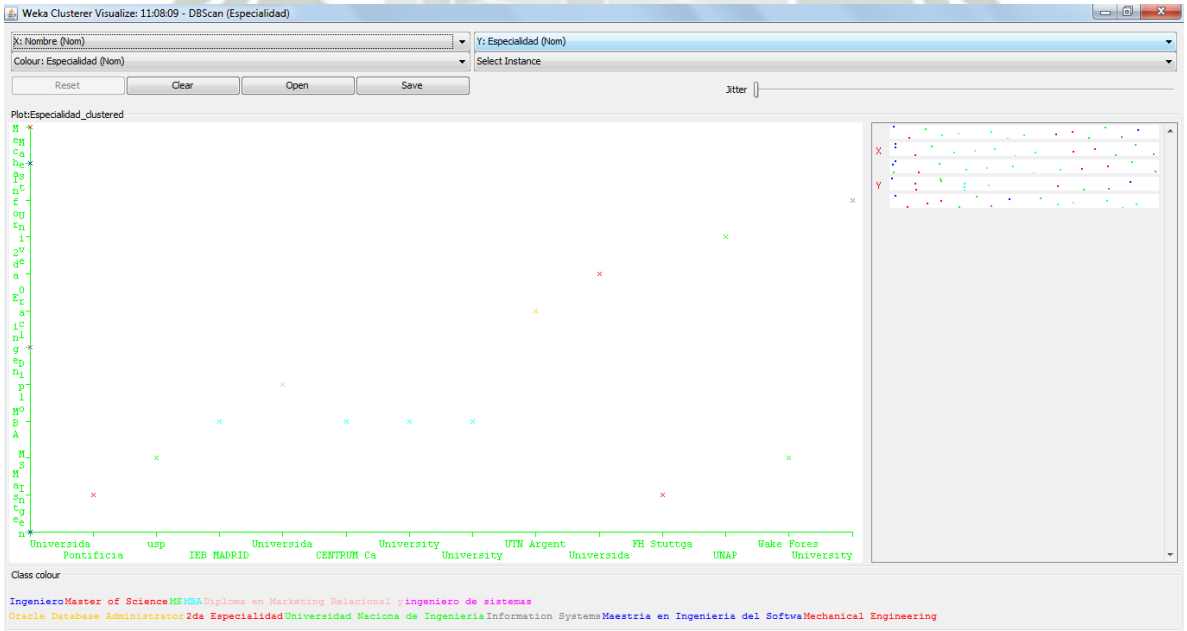


Figura 4. 17 Visualización de Clústeres en DBSCAN en Tabla Especialidad.

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

DESCRIPCION 4.9.- Teniendo en cuenta el algoritmo DBSCAN, comenzamos el proceso de evaluación configurando los ítems del recuadro que se observa en la Figura 4.5, siendo de mayor importancia el ítem minPoints (mínimo de puntos) el cual tendrá un valor de 1 ya que será el mínimo de puntos que concuerdan, también verificamos la DataBase Distance Type (tipo de distancia en a base de datos) y por default nos muestra la distancia: EuclideanDistance, los demás ítems los trabajamos por default. Analizando el proceso de Clustering mediante el algoritmo DBSCAN y dado que este algoritmo trabaja en función a densidad nos permite visualizar 16 clústeres o grupos los cuales nos muestran coincidencias entre la especialidad que estudian y la universidad donde la desarrollan. Entonces el algoritmo nos describe de forma precisa todas las especialidades que existen con sus respectivos centros de estudios, logrando obtener que la un porcentaje significativo estudia en el extranjero y que se especializan en diferentes áreas como son. Desarrollo Web e Investigación.

COMPARACION 4.3.- Revisando los grupos que se han encontrado en ambos algoritmos, vemos que existe una similitud en los egresados que estudian la especialidad en el extranjero y que también se enfocan en el área de Desarrollo Web, en el algoritmo DBSCAN encontramos un grupo que se enfoca al área de Investigación.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

4.1.7. CONFIGURACION Y EVALUACION DE DISTANCIAS: EUCLIDEAN Y MANHATTAN

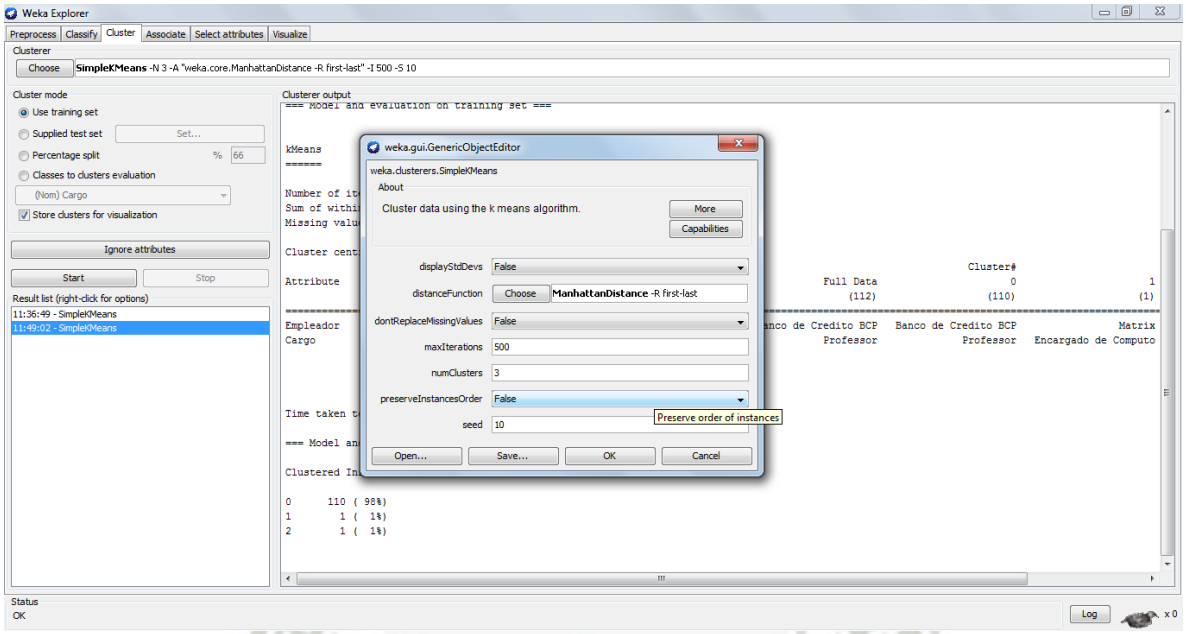


Figura 4. 18 Configuración Distancia: ManhattanDistance

Fuente: PROPIO.

COMPARACION 4.4.- Evaluando las distancia en el algoritmo Simple K-Means, podemos observar que tanto la distancia Euclidean y la distancia Manhattan obtienen resultados similares en cuanto a la distancia entre los grupos que se identificaron. Ver Figura 4.18

4.2. REPRESENTACIÓN DE LOS DATOS OBTENIDOS POR GRUPOS

CIUDAD ACTUAL

De la data CiudadActual analizada obtenemos el grupo Ciudad Actual en el cual predomina la ciudad de Arequipa con el mayor número de residentes egresados del Programa Profesional de Ingeniería de Sistemas (PPIS), seguida por la ciudad de Lima y otras ciudades con menor porcentaje de egresados que habitan en ellas. Ver Figura 4.19.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

Data:	Ciudad Actual	
Grupo 1:	Arequipa	84%
	Lima	14%
	Otros	2%

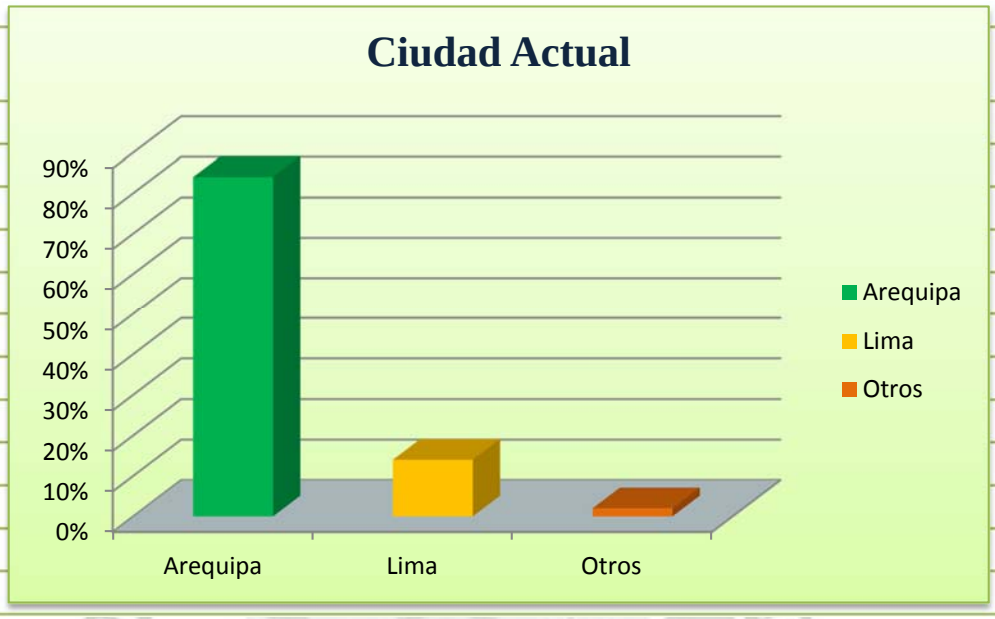


Figura 4. 19 Ciudad Actual

Fuente: PROPIO.

ESTADO LABORAL

De la data EstadoLaboral analizada obtenemos tres grupos (Ver Figura 4.20) los cuales son:

- Banca y Finanzas: en este grupo predomina por tener el mayor número de coincidencias de los egresados que laboran en entidades financieras con un 70.50%.

- Los grupos Académico y Área de Informática tienen coincidencias de un 14.75% cada grupo respectivamente.

Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

Data:	Estado Laboral	
Grupo 2:	Banca y Finanzas	70.50%
	Académico	14.75%
	Área Informática	14.75%

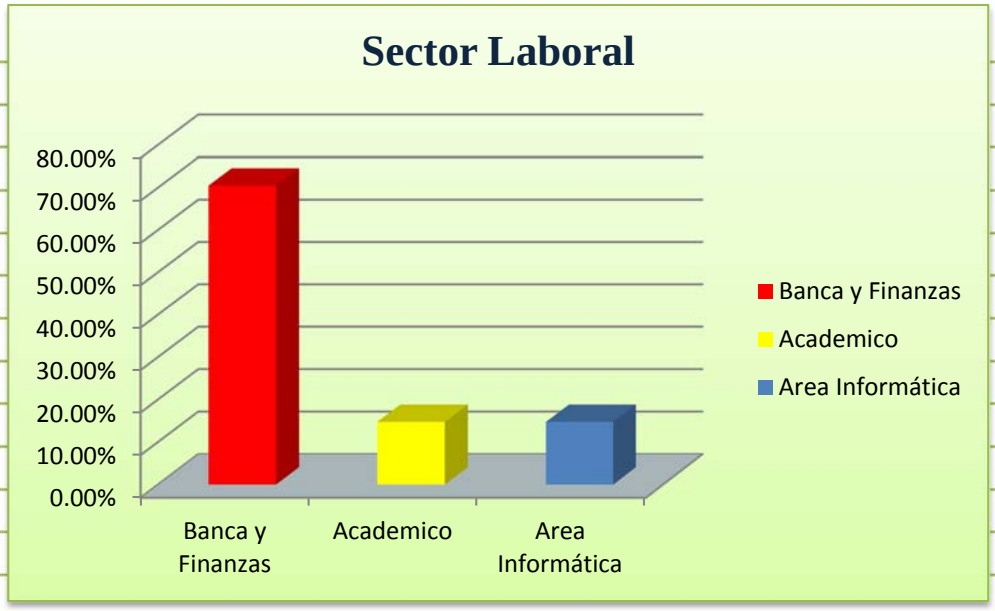


Figura 4. 20 Sector Laboral

Fuente: PROPIO.

ESPECIALIDAD.

- De la data Especialidad analizada obtenemos el grupo Lugar de Especialización en el cual predomina que los egresados tienden a realizar sus especialidades académicas de postgrado en el extranjero y con un menor índice en el país de origen. Ver Figura 4.21.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

Data:	Especialidad:	
Grupo 3:	Nacional	35%
	Extranjero	65%

Figura 4. 21 Lugar de Especialización

Fuente: PROPIO.

- Otro grupo obtenido del análisis de la data Especialidad es el grupo Área de Especialización en el cual predominan los egresados que se especializan en Desarrollo Web e Investigación. Ver Figura 4.22.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

Data:	Especialidad	
Grupo 4:	Desarrollo Web	58%
	Investigación	42%

Figura 4. 22 Área de Especialización

Fuente: PROPIO.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

PRIMERA: Se llegó a agrupar a los egresados en base a la información obtenida de las distintas fuentes de datos utilizando las herramientas Pentaho y Weka que nos lleve a realizar una adecuada gestión de información de los egresados.

SEGUNDA: Para la elección del algoritmo de agrupamiento (Clustering) en minería de datos se ha tomado en cuenta distintas características de agrupamiento como tamaño de grupo, la distancia entre objetos, la similaridad de objetos; habiendo realizado una evaluación y un análisis documentario de los algoritmos que se ajustan a la propuesta presentada.

TERCERA: Podemos observar en los resultados que haya un predominio de grupos bien definidos como son: Ciudad Actual donde residen la cual cuenta con un 84% de egresados que radican en la Ciudad de Arequipa, Banca y Finanzas con un 70.50% en el Sector Laboral, en el área de especialización el grupo de Desarrollo Web con 58% y que realizan sus estudios de especialidad en el Extranjero con un 65%.

CUARTA: Según los resultados obtenidos mediante el agrupamiento de los egresados concluimos que se encontraron grupos en diferentes áreas como son la academica, laboral; los cuales son importantes y nos ayudan a poder realizar una buena toma de decisiones en base a la gestión de los egresados del PPIS.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

RECOMENDACIONES

PRIMERA: Realizar este proceso utilizando Agrupamiento Difuso para manejar la incertidumbre de los datos y enfrentar el problema de sensibilidad ante ruido y valores atípicos, y así obtener datos más precisos de los egresados respecto al campo en que se desvuelven.

SEGUNDA: Utilizar otras técnicas de Minería de Datos, para la gestión de información de egresados como son las Redes Neuronales, Arboles de decisión, etc. Que son técnicas que provienen de la inteligencia artificial y de la estadística.

TERCERA: Hacer un sistema web que permita a los egresado actualizar sus información: personal, profesional, laboral y otros a través de las redes sociales, verificando los datos para poder contar con información más exacta y real.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

REEFERENCIAS BIBLIOGRAFICAS.

1. [ALV, et al, 06], Alvares y Pacheco. Propuesta metodológica para la recuperación de información mediante Clustering y reglas de asociación para Web Mining. Universidad Católica de Santa María, 2006.

2. [ANK, et al, 99] M. Ankerst, M. Breuning, H.P. Kriegel, and J. Sander. Optics: Ordering points to identify clustering structure. In Proceedings of the ACM SIGMOD Conference, pages 49–60, Philadelphia, PA., 1999.

3. [APA, 98] C. Aparicio Mollepaza, Construcción de una Metodología para la búsqueda de patrones en Sistemas de Soporte a la Toma de Decisiones, utilizando tecnología Data Mining, Universidad Católica de Santa María, Facultad de Ciencias e Ingenierías Físicas y Formales, Arequipa, Perú, 1998.

4. [BEN, et al, 02], F. Benavides, J. Marchant, “tesis de DataMining”, Minería de datos y la extracción de información oculta y predecible de grandes bases de datos. ITS. Andrés Bello, 2002.

5. [BER, et al, 01], T. Berners-Lee, J. Hendler, O Lassila. The Semantic Web. Scientific American, May 2001.

6. [CHE et al, 06], Chen, M. S, Han, J., y Yu, P. S., Data Mining: an overview from a database perspective, 2006.

7. [CON, 09], L. Contreras. Propuesta de preparación de datos documental para consultas utilizando un algoritmo particional de Data Mining. Universidad Católica de Santa María, 2009.

8. [CUA, 02], Abel Alberto Cuadrado Vega, Supervisión de procesos complejos mediante Técnicas de Data Mining con incorporación de conocimiento previo, Gijon, Noviembre de 2002.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

9. [EST, et al, 96] M. Esther, H.P Krieguel, J. Sander, and X. Xu. A density-based algorithm for discovering Clústeres in large spatial databases. In KDD ’96: Conf. Knowledge Discovery Data Mining, pages 226–231, 1996.

10. [FAY, et al, 96], U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From Data Mining to Knowledge Discovery in Databases, AAAI Press, 1996, S.1-34.

11. [GUE,12], Guevara Paucar Sergio, Modelo de Integración de Sistemas de Información Empresariales con Redes Sociales a través de APIS para la Universidad Católica de Santa María, 2012.

12. [HAN, et al, 00], J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, Publishers Inc, 2000.

13. [HAN, et al, 00] D. J. Hand, H. Mannila, and P. Smyth. Principles of Data Mining. Addison Wesley, 2000.

14. [HER, 06], Hernández Valadez Edna, Algoritmo de Clustering basado en entropía para descubrir grupos en atributos de tipo mixto, Centro de Investigación y de estudios avanzados del Instituto Politécnico Nacional , Departamento de Ingeniería Eléctrica Sección de Computación, México, D.F., Agosto de 2006.

15. [HOR, et al, 03].I. Horrocks and P.F. Patel-Schneider. “Three theses of representation in the semantic web”. In Proc. of the Twelfth International World Wide Web Conference (WWW 2003). 2003.

16. [JAI, et al, 99], A. K. Jain, M.N. Murty, and P. J. Flynn. Data clustering: a review. ACM Computing, Surveys, 31(3):264–323

17. [JAI, et al, 03], A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Prentice Hall, 2003.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

18. [KAR, et al, 99], G. Karyapis, E. H. Han and B. Kumar, A Hierarchical Clustering Algorithm Using Dynamic Modeling, IEEE Computer, Special Issue on Data Analysis and Mining, 1999.

19. [KAU, et al, 00], L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 2000.

20. [LAU, 05], Romina Laura Bot, Data Mining utilizando Redes Neuronales, Facultad de Ingeniería de la Universidad de Buenos Aires, Ingeniería en Informática, Tutor: Dr. Juan M. Ale, Buenos Aires, Argentina, Mayo de 2005

21. [MIT, et al, 03], S. Mitra and T. Acharya Data Mining: Multimedia, Soft Computing and Bioinformatics. Wiley Inter-Science, 2003.

22. [MOL, 00], Molina, L., Torturando los datos hasta que confiesen, Departamento de Lenguajes y Sistemas Informáticos, Universidad Politécnica de Cataluña. Barcelona, España, 2000.

23. [MON, et al, 09], E. Montánchez y Del Carpio. Propuesta de optimización en la implementación de un Data Mining de operaciones a partir del análisis de metodologías existentes. Universidad Católica de Santa María, 2009.

24. [NGH, et al, 94], R. Ng and J. Han. Efficient and Effective Clustering method for Spatial data Mining, Santiago de Chile, 1994.

25. [PER, et al, 07], Pérez y Medina. Implementación de un sistema CRM analítico orientado a la segmentación de clientes y caracterización de perfiles con técnicas de Data Mining utilizando algoritmos genéticos.. Universidad Católica de Santa María, 2007.

26. [POT, 05]. Potok. Co-Chairs The Semantic Web: The Goal of Web Intelligence Proceedings of the 38th Hawaii International Conference on System Sciences.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

27. [SHE, et al, 98], G. Sheikholeslami, S. Chatterjee, and A. Zhang, WaveCluster: A Multi-Resolution Clustering Approach for very large Spatial Database, 24th International Conference on Very Large Data Bases, New York City, 1998.

28. [SUD, et al, 98] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: an efficient clustering algorithm for large databases. In SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data, pages 73–84, New York, NY, USA, 1998. ACM Press.

29. [TAN, et al, 06], P. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining. Addison Wesley, 2006.

30. [TEL, 99], Tello. Metodología de modelamiento de Datos para Data Mining. Universidad Católica de Santa María, 1999.

31. [VEL, 09], L. Velásquez. Propuesta de un modelo de presentación empleando Web semántica para la especificación, diseño e implementación de requisitos de interfaz de usuario. Universidad Católica de Santa María, 2009.

32. [WEI, et al, 97], Wei Wang, Jiong Yang, and Richard Muntz, STING: A Stadistical Information Grid Approach to Spatial Data, Departament of Computer Science University of California, Los Angeles, 1997.

33. [ZHA, 04], T. Zhang. "Minimum entropy clustering and applications to gene expression analysis," in Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE, 16-19 Aug. 2004, pp. 142–151.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

REFERENCIAS WEB

1. [SAS, 02], SAS ENTERPRISE, Implicancias del Data Mining, http://www.fceco.uner.edu.ar/extinv/publicdocent/sarangur/pdf/datamining.pdf, 2002 .
2. [WWW0] http://es.wikipedia.org/wiki/Web_semántica
3. [WWW1] Página principal de la web semántica en W3C. http://www.w3.org/2001/sw/
4. [WWW2] [http://www.daml.org].
5. [WWW3] http://www.dataprix.com/723-caracter-sticas-pentaho.
6. [WWW4] http://www.dataprix.com/la-metodolog%C3%AD-crisp-dm

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

GLOSARIO.

PPIS: Programa Profesional de Ingeniería de Sistemas

UCSM: Universidad Católica de Santa María

KDD: Descubrimiento de Conocimiento en Bases de Datos

PAM: Partitioning Around Medoids.

CLARA: Clustering Large Applications

CLARANS: Clustering Large Applications based upon RANdomized Search

CHAMELON: Hierarchical Clustering Algorithm Using Dynamic Modeling

CURE: Clustering Use Representatives

ROCK: Clustering Categorical Data

STING: Statical Information Grid approach

WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases

CLIQUE: Clustering In Quest

DBSCAN: Density-Based algorithm for discovering Clústeres in large spatial Databases with Noise

OPTICS: Ordering Points To Identify the Clustering Structure

DENCLUE: Desity based Clustering

XML: Extensible Mark-up Language

SGML: Sistema Anidado de Marcas

DTDs: Document Type Definitions

NS: Name Spaces

RDF: Resource Description Framework

OWL: Lenguaje Ontológico Web (Web Ontology Language)

DAML: DARPA Agent Mark-Up Language

OIL: Ontology Inference Layer

XKMS: Gestión de Claves XML (XML Key Management - XKMS).

SAML: SAML (Security Assertion Mark-up Language)

ETL: Extraccion, Transformacion y Carga

WWW: World Wide Web Consortium

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE
TESIS UCSM

CATÓLICA
DE SANTA MARÍA

ANEXOS

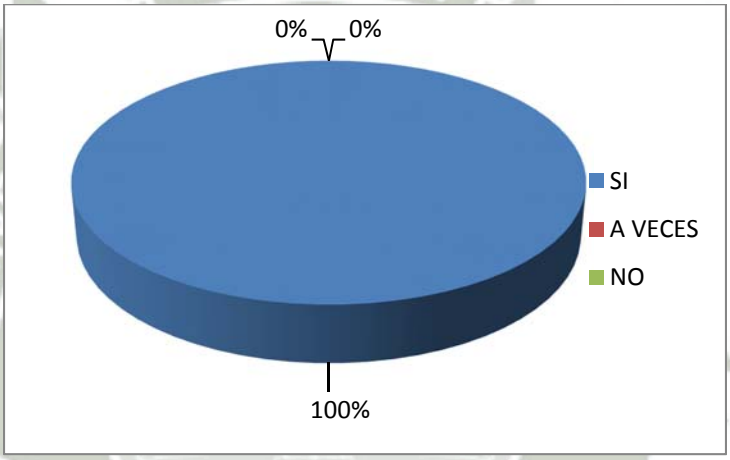
Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

ENCUESTA

1. Cree Ud. Que la gestión de los egresados es importante y beneficioso para el Programa Profesional?

Detalle	Frecuencia	Porcentaje
SI	13	100%
AVECES	0	0%
NO	0	0%
TOTAL	13	100%



Análisis e Interpretación: El 100% de los Directores de los Programas Profesionales de las Universidades de Arequipa encuestados consideran que la Gestión de egresados es importante y beneficioso para el Programa Profesional.

2. Cree que la gestión propuesta de egresados tendrá influencia en la toma de decisiones sobre el plan curricular de su Programa Profesional?

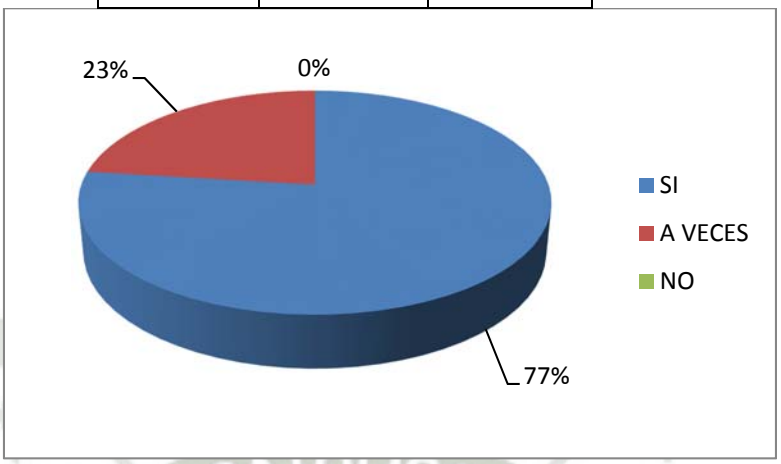
Detalle	Frecuencia	Porcentaje
SI	10	77%
A VECES	3	23%

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

NO	0	0%
TOTAL	13	100%



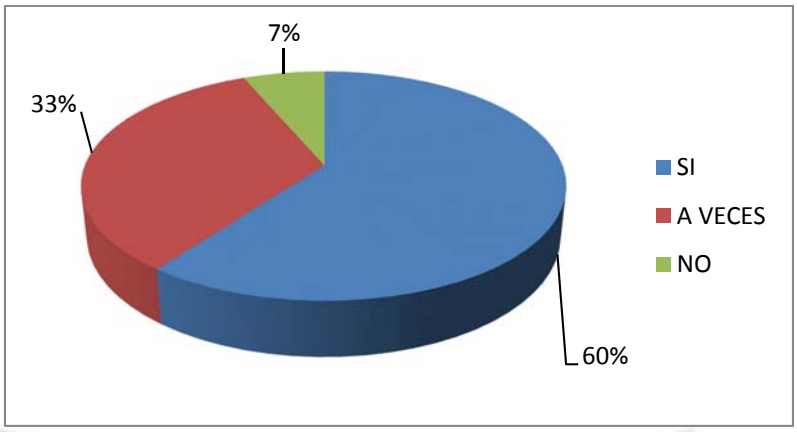
Análisis e Interpretación: El 77.00% de la población encuestada consideran que dicha gestión de egresados influiría de manera significativa en la toma de decisiones sobre el plan curricular dentro del Programa Profesional, ya que consideran que año tras año aparecen nuevas tendencias en el ámbito profesional y laboral, mientras que el 23.00% restante opinaron que a veces influiría en el plan curricular, ya que consideran que existen una estructura curricular de la cual no se puede prescindir.

3. Cree Ud. Que el Pentaho es una herramienta viable y flexible para la preparación de datos (ETL)?

Detalle	Frecuencia	Porcentaje
SI	47	60%
A VECES	26	33%
NO	5	7%
TOTAL	78	100%

Publicación autorizada con fines académicos e investigativos

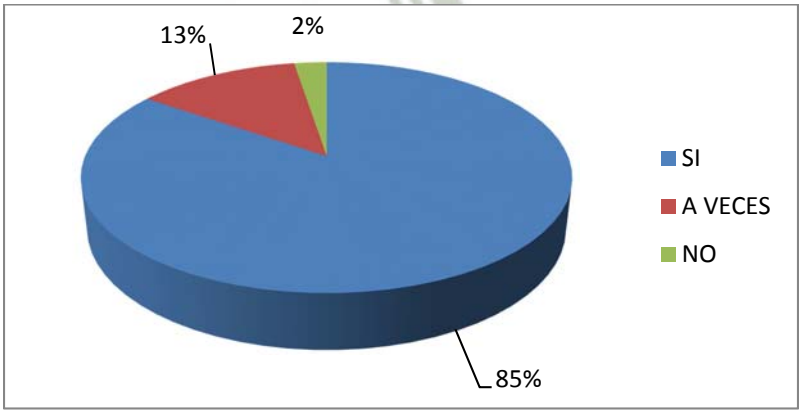
REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA



Análisis e Interpretación: El 60.00% de la población encuestada consideran que el Pentaho es una herramienta viable y flexible para la preparación de datos (ETL), mientras que el 33.00% opinaron que a veces es conveniente utilizar otro tipo de herramientas de acuerdo a los tipos de datos a trabajar y el 7.00% de la población encuestada no opina por desconocimiento de la herramienta.

4. Cree Ud. Que el WEKA es una herramienta que nos permite visualizar un adecuado análisis sobre los datos trabajados para la gestión de los egresados?

Detalle	Frecuencia	Porcentaje
SI	66	84%
A VECES	10	13%
NO	2	3%
TOTAL	78	100%



Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

Análisis e Interpretación: El 84.00% de la población encuestada consideran que el WEKA es una herramienta que permite visualizar un adecuado análisis sobre la gestión de datos, mientras que el 13.00% opinaron que a veces es conveniente utilizar otro tipo de herramientas de acuerdo a los tipos de datos a trabajar y el 3.00% de la población encuestada no opina ya que no tiene experiencia en el manejo de la herramienta.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM
CATÓLICA DE SANTA MARÍA

ANEXO A: IMPLEMENTACIÓN DEL SISTEMA.

Para visualizar los resultados obtenidos del proceso ETL, nos apoyamos en un sistema Web desarrollado en PHP y MySQL.

Pantalla Principal.

Se muestra la presentación del sistema donde se visualiza de forma general la estructura, proceso ETL y análisis. Figura. Pantalla Principal.

Figura. Pantalla Principal.

Pantalla Ciudad Actual.

En esta pantalla se visualiza la ciudad actual donde residen los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS), en la cual se listan las ciudades de acuerdo al país seleccionado. Figura Pantalla Ciudad Actual.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

Figura. Pantalla Ciudad Actual.

Pantalla Empresas Empleadoras.

En esta pantalla se visualiza la lista de empresas empleadoras y el numero de egresados que laboran en cada empresa. Figura Pantalla Empresas Empleadoras.

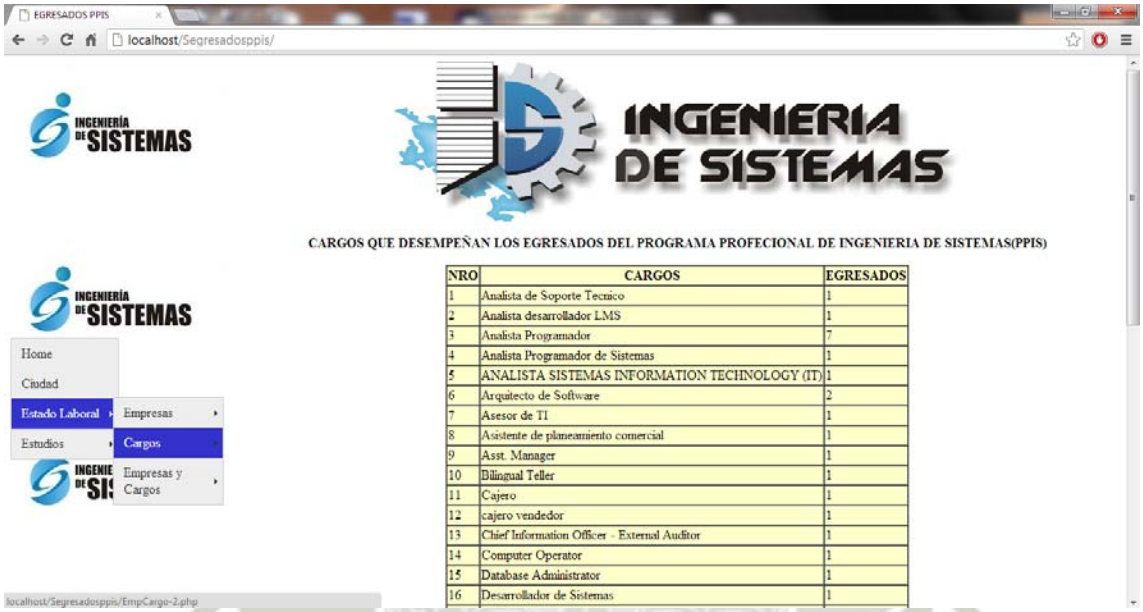
Publicación autorizada con fines académicos e investigativos

TESIS UCSM
CATÓLICA
DE SANTA MARÍA

Figura Pantalla Empresas Empleadoras.

Pantalla Cargos.

En esta pantalla se visualiza la lista de cargos que desempeñan los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS) y el número de egresados que se desempeñan en dichos cargos. Figura Pantalla Cargos.



NRO	CARGOS	EGRESADOS
1	Analista de Soporte Tecnico	1
2	Analista desarrollador LMS	1
3	Analista Programador	7
4	Analista Programador de Sistemas	1
5	ANALISTA SISTEMAS INFORMATION TECHNOLOGY (IT)	1
6	Arquitecto de Software	2
7	Asesor de TI	1
8	Asistente de planeamiento comercial	1
9	Asst. Manager	1
10	Bilingual Teller	1
11	Cajero	1
12	cajero vendedor	1
13	Chief Information Officer - External Auditor	1
14	Computer Operator	1
15	Database Administrator	1
16	Desarrollador de Sistemas	1

Figura Pantalla Cargos.

Pantalla Universidad y Carrera.

En esta pantalla se visualiza la lista de Universidades y carreras de los egresados o alumnos asociados a la red social del Programa Profesional de Ingeniería de Sistemas (PPIS). Figura Pantalla Universidad y Carrera.

Publicación autorizada con fines académicos e investigativos

REPOSITORIO DE TESIS UCSM

CATÓLICA DE SANTA MARÍA

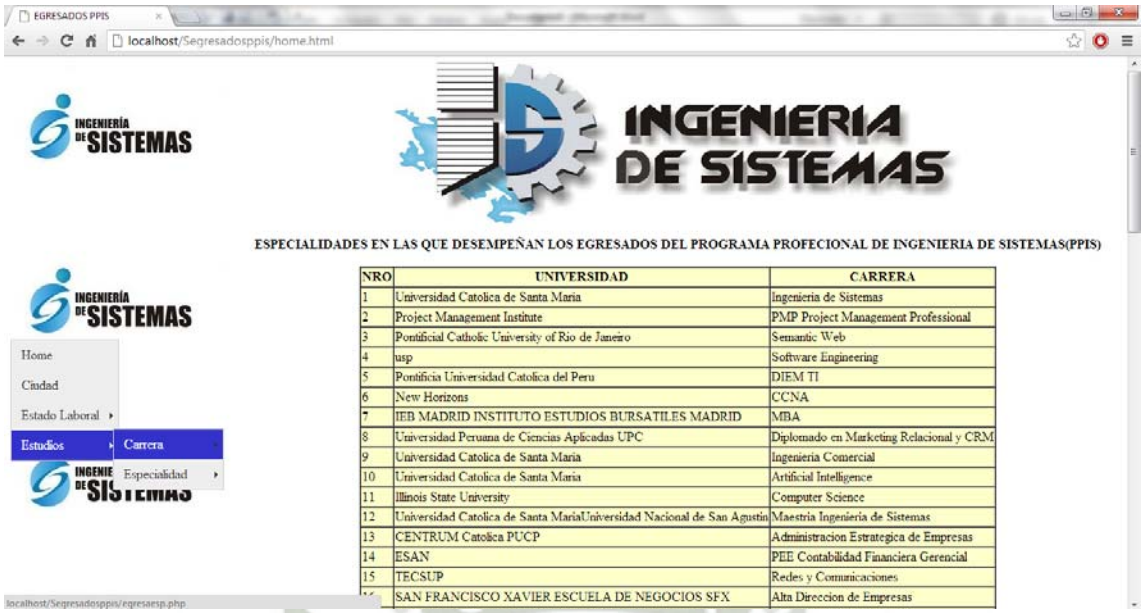


Figura Pantalla Universidad y Carrera.

Pantalla Especialidad.

En esta pantalla se visualiza Especialidad de los egresados del Programa Profesional de Ingeniería de Sistemas (PPIS), en la cual se listan las especialidades y el numero de egresados que tiene dicha especialidad. Figura. Pantalla Especialidad.

Figura. Pantalla Especialidad

Publicación autorizada con fines académicos e investigativos