UNIDAD ACADÉMICA:

DEPARTAMENTO DE INVESTIGACIÓN Y POSTGRADOS

TEMA:

DESCUBRIMIENTO DE CONOCIMIENTO EN BASE DE DATOS PARA LA TOMA DE DECISIONES EN LA UNIDAD DE NIVELACIÓN Y ADMISIÓN DE LA ESPOCH

Proyecto de Investigación y Desarrollo previo a la obtención del título de

Magister en Gerencia Informática

Línea de Investigación, Innovación y Desarrollo principal:

Sistemas de Información y/o Nuevas Tecnologías de la Información y Comunicación y sus aplicaciones

Caracterización técnica del trabajo:

Aplicación

Autora:

María Isabel Uvidia Fassler

Director:

Diego Fernando Ávila Pesantez, Ing. MsC

Ambato – Ecuador

Febrero 2016

Descubrimiento de Conocimiento en Base de Datos para la toma de decisiones en la Unidad de Nivelación y Admisión de la ESPOCH

Informe de Trabajo de Titulación
presentado ante la
Pontificia Universidad Católica del
Ecuador Sede Ambato

por

María Isabel Uvidia Fassler

En cumplimiento
parcial de los requisitos
para el Grado de Magister
en Gerencia Informática

Departamento de Investigación y Postgrados
Febrero 2016

Descubrimiento de Conocimiento en Base de Datos para la toma de decisiones en la Unidad de Nivelación y Admisión de la ESPOCH

Aprobado por:

Varna Hernández, PhD
Presidente del Comité Calificador
Director DIP

Galo López Sevilla, Mg
Miembro Calificador

Diego Fernando Ávila Pesantez, MsC
Miembro Calificador
Director de Proyecto

Dr. Hugo Altamirano Villaroel
Secretario General

Enrique Xavier Garcés Freire, Mg
Miembro Calificador

Fecha de aprobación:
Febrero 2016

Ficha Técnica

Programa: Magister en Gerencia Informática

Tema: Descubrimiento de Conocimiento en Base de Datos para la toma de decisiones en la Unidad de Nivelación Y Admisión de la ESPOCH

Tipo de trabajo: Proyecto de Investigación y Desarrollo

Clasificación técnica del trabajo: Aplicación

Autor: María Isabel Uvidia Fassler

Director: Diego Fernando Ávila Pesantez, Ing. MsC.

Líneas de Investigación, Innovación y Desarrollo

Principal: Sistemas de Información y/o Nuevas Tecnologías de la Información y Comunicación y sus aplicaciones

Resumen Ejecutivo

El Descubrimiento de Conocimiento en Base de datos es un conjunto de fases o pasos que permiten extraer información de datos que necesitan ser organizados y apoyan la toma de decisiones. Para este trabajo se desarrollaron sus fases acopladas a la Metodología HEFESTO versión 2.0. (Metodología propia para la construcción del *Data Warehouse*) pudiendo cumplir su aplicación en la Unidad de Nivelación y Admisión de la Escuela Superior Politécnica de Chimborazo (ESPOCH).

Este conjunto de pasos propuesto constó de 9 fases, 4 clasificadas dentro del subproceso *Data Warehouse* y las otras 5 dentro del subproceso *Data Mining*. El subproceso *Data Warehouse* inicia desde aprender el dominio de la aplicación, es decir, conocer el área de los datos con los que se trabaja, además de las necesidades o requerimientos de información, selección, limpieza de datos para permitir información consistente, siguiendo con el diseño del Data Warehouse (DW) donde se guarda la información mediante procesos de Extracción, Transformación y Carga (ETL) hasta obtener información preparada para el siguiente subproceso que es *Data Mining*, donde mediante la selección de técnicas y análisis de información, se puede obtener patrones que una vez observados y examinados se convierten en conocimiento para la toma de decisiones.

Esta toma de decisiones fue complementada mediante la creación de reportes *Business Intelligence* que reflejan el proceso académico de admisión y nivelación desarrollado por la ESPOCH durante estos años.

Declaración de Originalidad y Responsabilidad

Yo, María Isabel Uvidia Fassler, portador de la cédula de ciudadanía No. 060370559-1, declaro que los resultados obtenidos en el proyecto de titulación y presentados en el informe final, previo a la obtención del título de Magister en Gerencia Informática, son absolutamente originales y personales. En tal virtud, declaro que el contenido, las conclusiones y los efectos legales y académicos que se desprenden del trabajo propuesto, y luego de la redacción de este documento, son y serán de mi sola y exclusiva responsabilidad legal y académica.

María Isabel Uvidia Fassler
C.C. 060370559-1

Con todo mi amor para quienes siempre me han guiado, motivado y acompañado en mis pasos y decisiones, Fanny y John, pilar fundamental en mi vida.

A mis hermanos Johana y John por su apoyo incondicional y solidario.

A mis amigos que me dan alas para conseguir más logros.

Reconocimientos

Al Ing. Diego Ávila Pesantez, por su apoyo, motivación, amistad y atención precisa a lo largo de mi desarrollo académico durante la Maestría, que ahora se ve culminado con este proyecto de Investigación y Desarrollo.

Al Ing. Santiago Cisneros Barahona, Coordinador General de la Unidad de Nivelación y Admisión – ESPOCH, por su compromiso institucional al permitir la aplicación de este trabajo. Además, por su amistad, apoyo y confianza.

A la Ing. Ivonne Rodríguez Flores, por su apoyo incondicional a lo largo de mi vida profesional, por su generosidad de conocimientos, amistad, sinceridad y estima.

Resumen

El descubrimiento de conocimiento en base de datos es aplicado en este trabajo mediante el acoplamiento de sus fases a la metodología HEFESTO versión 2.0. (metodología propia para la construcción del Data Warehouse) en la Unidad de Nivelación y Admisión de la ESPOCH. Dichos datos fueron generados desde el año 2012, donde la ESPOCH ha trabajado con el Sistema Nacional de Nivelación y Admisión (SNNA - SENESCYT), provocando gran cantidad de información que no ha sido procesada y que no aporta a la toma de decisiones. En base a esta urgente necesidad, se propone el proceso que consta de nueve fases, cuatro clasificadas dentro del subproceso Data Warehouse y las otras cinco dentro del subproceso Data Mining. El subproceso Data Warehouse inició desde aprender el dominio de la aplicación, es decir, conocer el área de los datos con los que se trabaja, además de las necesidades o requerimientos de información, selección, limpieza de datos para permitir información consistente, siguiendo con el diseño del DW donde se almacena la información mediante procesos de Extracción, Transformación y Carga (ETL) hasta obtener información preparada para el siguiente subproceso que es Data Mining, donde mediante la selección de técnicas y análisis de información, se obtuvieron patrones que una vez observados y examinados generaron conocimiento para la toma de decisiones. Además, se crearon reportes ad-hoc Business Intelligence que reflejan los procesos académicos, de admisión y nivelación desarrollados por la ESPOCH durante estos años.

Abstract

Knowledge discovery in databases was applied in this work at the Standard Admissions Center in ESPOCH, by adapting each of their processes to the HEFESTO version 2.0. methodology (a particular methodology to build Data Warehouse). Such data was generated since 2012, where ESPOCH has been working with the Standard Admission National System (SNNA - SENESCYT), which has caused a great deal of unprocessed information that does not help decision-making. Based on this urgent matter, a process that includes nine steps, four of them were classified within the Data Warehouse subprocess and the other five steps in the Data Mining subprocess. The Data Warehouse subprocess started on how to master the application, in other words, to get the information that they are working with, besides the needs or information requirements, selection and data cleansing to get consistent information, following the DWH design, where the information is stored by the Extract, Transform and Load (ETL) process to finally obtain the information required by the next process, which is Data Mining. This process chooses some of the techniques and analyze the information ant in this way, the patterns were obtained. Such patterns were observed and examined to provide useful information to make decisions.

Furthermore, Ad-hoc Business Intelligence reports were created, which show academic, admission and leveling processes that were developed at ESPOCH during these years.

Key words: data warehouse, data mining, knowledge discovery in databases

Tabla de Contenidos

Ficha Técnica iii

Líneas de Investigación, Innovación y Desarrollo iii

Resumen Ejecutivo iii

Declaración de Originalidad y Responsabilidad iv

Dedicatoria v

Reconocimientos vi

Resumen vii

Abstract viii

Tabla de Contenidos ix

Lista de Tablas xiii

Lista de Figuras xiv

CAPÍTULOS

1. Introducción 1

1.1. Presentación del trabajo 1

1.3. Descripción del documento 3

2. Planteamiento de la Propuesta de Trabajo 4

2.1. Información técnica básica 4

Líneas de Investigación, Innovación y Desarrollo 4

2.2. Descripción del problema 4

2.3. Preguntas básicas 4

2.4. Formulación de meta 5

2.5. Objetivos 5

2.6. Delimitación funcional 6

3. Marco Teórico 7

3.1. Definiciones y conceptos 7

3.1.1. Data Warehousing 7

3.1.2. Data Warehouse 7

3.1.3. OLTP 9
3.1.4. ETL 9
3.1.5. Limpieza de datos 10
3.1.6. OLAP 10
3.1.7. Data Mining 11
3.1.8. Business Intelligence 12
3.1.9. Conocimiento 13
3.1.10. Metadatos 14
3.1.11. ROLAP 15
3.1.12. MOLAP 15
3.1.13. HOLAP 16
3.2. Estado del Arte 16
3.2.1. Análisis comparativo de metodologías para la gestión de proyectos de minería de datos 16
3.2.2. Descubrimiento de Conocimiento en Base de Datos 18
3.2.3. Fases de Descubrimiento de Conocimiento en Base de Datos 19
3.2.3.1. Aprender el dominio de la aplicación 19
3.2.3.2. Selección y creación de un conjunto de datos sobre la que se realizará el descubrimiento 20
3.2.3.3. Pre procesamiento y limpieza 21
3.2.3.4. Transformación de datos 21
3.2.3.5. Elección de la tarea de minería de datos apropiada 22
3.2.3.6. Elección del algoritmo de minería de datos 22
3.2.3.7. Empleando el algoritmo de minería de datos 22
3.2.3.8. Evaluación 23
3.2.3.9. Usando el conocimiento descubierto 23
3.2.4. HEFESTO versión 2.0. Metodología propia para la Construcción de un Data Warehouse 23
3.2.4.1. Análisis de Requerimientos 24
3.2.4.2. Análisis de los OLTP 25

3.2.4.3. Modelo Lógico del DW 26
3.2.4.4. Integración de Datos 29
3.2.5. Data Mining… 30
3.2.6. Técnicas Data Mining 30
3.2.6.1. Redes Neuronales 31
3.2.6.2. Árboles de Decisión 31
3.2.6.3. Redes bayesianas 31
3.2.6.4. Técnicas de Clustering 31
3.2.6.5. Optimización de Secuencia Mínima (SMOreg) 32
3.2.6.6. Regresión Lineal 32
3.2.7. La Minería de Datos, entre la Estadística y la Inteligencia Artificial 32
4. Metodología 35
4.1. Diagnóstico 35
4.2. Método(s) aplicado(s) 35
4.2.1. De Investigación 35
4.2.2. De Desarrollo 35
4.2.2.1. Aprender el dominio de la aplicación 37
4.2.2.2. Selección y creación de un conjunto de datos sobre la que se realizará el descubrimiento 38
4.2.2.3. Pre procesamiento y limpieza 39
4.2.2.4. Transformación de datos 40
4.2.2.5. Elección de la tarea de minería de datos apropiada 40
4.2.2.6. Elección del algoritmo de minería de datos 40
4.2.2.7. Empleando el algoritmo de minería de datos 41
4.2.2.8. Evaluación 41
4.2.2.9. Usando el conocimiento descubierto 41
4.3. Materiales y herramientas 41
5. Resultados 43

5.1. Análisis de resultados 43

5.1.1. Aprender el dominio de la aplicación 43

5.1.2. Selección y creación de un conjunto de datos sobre la que se realizará el descubrimiento 53

5.1.3. Pre procesamiento y limpieza 63

5.1.4. Transformación de datos 66

5.1.5. Elección de la tarea de minería de datos apropiada 68

5.1.6. Elección del algoritmo de minería de datos 68

5.1.7. Empleando el algoritmo de minería de datos 69

5.1.8. Evaluación 79

5.1.9. Usando el conocimiento descubierto 82

5.1.10. Reportes Business Intelligence 83

5.2. Evaluación preliminar 86

6. Conclusiones y Recomendaciones 87

6.1. Conclusiones 87

6.2. Recomendaciones 88

APÉNDICES

Apéndice A: Tareas de Limpieza de Datos 89

Apéndice B: ETLs (Extracción, Transformación y Carga) 91

Apéndice C: Reportes Business Intelligence 94

Referencias 110

Lista de Tablas

1. ROLAP vs MOLAP 15
2. Fases del proceso de Minería de Datos de cada Modelo 17
3. Comparación de la probabilidad de acierto según 4 métodos de predicción 33
4. Matriz Aprender el dominio de la aplicación 38
5. Matriz Selección y Creación DW 39
6. Herramientas tecnológicas utilizadas 42
7. Matriz Aprender el dominio de la aplicación – Proceso: Ingreso de estudiantes con cupo para la ESPOCH 45
8. Matriz Aprender el dominio de la aplicación – Proceso: Información académica de Nivelación 47
9. Matriz Aprender el dominio de la aplicación – Proceso: Aprobados y Reprobados 50
10. Estado de archivos OLTP 52
11. Matriz Selección y Creación DW Proceso: Ingreso de estudiantes con un cupo para la ESPOCH 54
12. Matriz Selección y Creación DW Proceso: Ingreso académica de Nivelación 57
13. Matriz Selección y Creación DW Proceso: Aprobados y Reprobados 59
14. Diseño del DW 61

Lista de Figuras

1. Data Warehouse 8
2. Evolución de Sistemas para toma de decisiones 13
3. Metáfora del embudo del Business Intelligence 14
4. Representación de Metadatos 14
5. El proceso de descubrimiento de conocimiento en bases de datos. 19
6. Pasos de la Metodología HEFESTO v2.0 24
7. Esquema en Estrella 26
8. Esquema Copo de Nieve 27
9. Esquema Constelación 27
10. Clasificación de técnicas Data Mining 30
11. Propuesta de Proceso KDD implementado en la Unidad de Nivelación y Admisión – ESPOCH 36
12. Diagrama del diseño del DW 62
13. Data Cleaner 63
14. Resultados de datos duplicados de docentes habilitados 64
15. Resultados de análisis de notas y sus intervalos 64
16. Resultados de análisis cédulas de estudiantes 65
17. Resultados de análisis de datos únicos 65
18. ETL dim_ubicacion_academica 66
19. ETL dim_estudiante 67
20. ETL de carga del DW 67
21. Elección del algoritmo de minería de datos para la Unidad de Nivelación y Admisión ESPOCH 69
22. Atributo Anio 70
23. Atributo Área 70
24. Atributo Género 71
25. Algoritmos de Clasificación 71
26. Resultado del algoritmo Árbol de Decisión 73
27. Resultado del algoritmo Redes Bayesianas 74
28. Resultado del algoritmo Redes Bayesianas – Nodo Género 74
29. Resultado del algoritmo Redes Bayesianas – Nodo Anio 75
30. Resultado del algoritmo Redes Bayesianas – Nodo Anio 75
31. Atributo Anio 75

32. Atributo Estudiantes 76
33. Algoritmos de Regresión 76
34. Resultado del algoritmo Regresión Lineal 77
35. Resultado del algoritmo Optimización de Secuencia Mínima 79
36. Reporte de Ubicación Geográfica 85
37. Reporte de Postulaciones 85
38. Creación de Tareas para encontrar datos duplicados 89
39. Análisis de datos de estudiantes 90
40. Transformación de datos 90
41. Conexión a base de datos 91
42. Creación de la transformación 92
43. Sentencia SQL para cargar datos de la fuente 92
44. Sentencia SQL para cargar datos de la fuente 93
45. Configuraciones de la tabla destino 93
46. SpagoBI Studio 4.0 94
47. Creación Proyecto SpagoBI 95
48. Crear nuevo modelo SpagoBI 95
49. Conexión a la base de datos 96
50. Selección de tablas Modelo Físico 96
51. Selección de tablas para el cubo dimensional 96
52. Obtención de las tablas para el Modelo de Negocios 97
53. Cubo dimensional 98
54. Creación del Cubo dimensional OLAP 98
55. Niveles de Jerarquía 99
56. Asignación de jerarquías 99
57. Relaciones de las dimensiones con el Cubo 100
58. Relación de todas las dimensiones 100
59. Asignación de Métricas 101
60. Asignación de operaciones a las métricas 101
61. Servidor SpagoBI 102
62. Panel de trabajo SpagoBI 102
63. Almacenamiento del Modelo de Negocios en el servidor SpagoBI 103
64. Almacenamiento 103
65. SpagoBI 104
66. Consultas 104

67. Visualización de consultas realizadas con éxito 105
68. Reportes con gráficas 105
69. Parámetros del reporte 106
70. Configuraciones de los encabezados del reporte 106
71. Pre visualización 107
72. Guardar reporte 107
73. Espacio de trabajo 108
74. Exportar reportes 108
75. Modificar reportes 109

Capítulo 1

Introducción

Cada día se genera una gran cantidad de datos en las instituciones públicas y privadas, datos que son almacenados en distintos repositorios y de los cuales no se obtienen resultados debido a que no se conocen procesos que puedan extraer información útil y consistente (Fernández, Duarte, Hernández, & Sánchez, 2010). La información veraz está al principio y al final de toda acción gerencial por tal motivo es importante para la toma de decisiones contar con información relevante, sólida y oportuna, para cumplir objetivos y aplicar mejoras y acciones.

"Knowledge Discovery in Databases (KDD) tiene la tarea de inferir información válida a partir de estos datos", "Sin embargo, la calidad de estos modelos está íntimamente relacionada con la de los datos almacenados, dependiendo de factores como son la existencia de datos erróneos o el gran tamaño de la base de datos" "Por lo tanto urge, como fase inicial del proceso, el preprocesamiento de los datos iniciales con el fin de elevar la calidad de los mismos" (Fernández et al., 2010). Además, una vez preprocesados los datos, *Data Warehouse* es el siguiente paso en transformar a un sistema de base de datos, donde su principal propósito es el almacenamiento confiable, siendo primordial su uso en el apoyo a la toma de decisiones (Bradley, Fayyad, & Mangasarian, 1999).

Mediante el análisis de las fases del proceso de Descubrimiento de Conocimiento en Base de Datos, se acopló a este HEFESTO versión 2.0., que es una metodología propia de construcción de *Data Warehouse*, siendo 9 fases las que se aplicaron a la Unidad de Nivelación y Admisión – ESPOCH. Las 4 primeras corresponden al subproceso *Data Warehouse*, asegurando de esta forma datos consistentes para el siguiente subproceso de 5 fases que es *Data Mining*; donde mediante la aplicación de técnicas se pueden observar patrones que una vez analizados permiten obtener conocimiento y tomar decisiones.

Para complementar este proceso de descubrimiento de conocimiento se crearon reportes de *Business Intelligence* amigables y personalizables.

1.1. Presentación del trabajo

Maimon y Rokach en el año 2010 mencionan que KDD (*Knowledge Discovery in Databases*: KDD, Descubrimiento de Conocimiento en Base de datos) es el proceso organizado para identificar

patrones válidos y útiles que permiten entender y analizar datos. La gran cantidad de datos permiten que KDD a parte de su gran importancia sea una necesidad. Dentro de las fases de KDD, *Data Mining* es la aplicación específica de algoritmos para extraer patrones de los datos, siendo esta fase particular dentro de KDD.

Para lograr la toma de decisiones efectiva se aplica el proceso KDD, que es el proceso para descubrir el conocimiento e información útil, mediante la aplicación de sus fases (Maimon & Rokach, 2010). Para aplicar este proceso dentro de la Unidad de Nivelación y Admisión de la ESPOCH, se creó una propuesta que acopla a las fases KDD la metodología HEFESTO versión 2.0. (Metodología propia para la creación de un *Data Warehouse*), siendo 9 fases: 4 fases dentro del subproceso *Data Warehouse* y las otras 5 dentro del subproceso *Data Mining*.

El subproceso *Data Warehouse* inicia desde aprender el dominio de la aplicación, es decir, conocer el área de los datos con los que se trabaja, además de las necesidades o requerimientos de información, selección, limpieza de datos para permitir información consistente, siguiendo con el diseño del DW donde se guarda la información mediante procesos de Extracción, Transformación y Carga (ETL) hasta obtener información preparada para el siguiente subproceso que es *Data Mining*, donde mediante la selección de técnicas y análisis de información, se puede obtener patrones que una vez observados y examinados se convierten en conocimiento para la toma de decisiones, siendo importante partir desde el conocimiento de la realidad, es decir conocer tendencias de postulaciones por área y años, hasta llegar a conocer predicciones mediante la comparación de técnicas de minería de datos y análisis de parámetros estadísticos que aseguren la confiabilidad en la información.

Así el valor real de los datos reside en la información que se puede extraer de ellos, información que ayude a tomar decisiones o mejorar nuestra comprensión de los fenómenos que nos rodean. Hoy, más que nunca, los métodos analíticos avanzados son el arma secreta de muchos negocios exitosos. Empleando métodos analíticos avanzados para la explotación de datos, los negocios incrementan sus ganancias, maximizan la eficiencia operativa, reducen costos y mejoran la satisfacción del cliente (Asencios, 2004).

Complementando el proceso KDD, se crearon reportes Ad-hoc de *Business Intelligence* que dan a conocer: Lugares de procedencia, postulaciones de las carreras de los estudiantes, evolución de los procesos de Nivelación de cada carrera, cantidad de estudiantes hombres y mujeres; permitiendo tomar decisiones académicas dentro de la ESPOCH, considerando que "la toma de decisiones es la selección de un curso de acción de entre varias alternativas, es un punto muy importante debido a que las consecuencias o beneficios percutirán a la organización representada" (Castillo & Chairez, 2004).

La aplicación de todo el proceso KDD se lo hizo mediante herramientas de software libre.

1.3. Descripción del documento

A continuación, se muestra una visión global del trabajo mediante la explicación de lo que en cada capítulo se presenta. El capítulo 1 introduce al proyecto en cuanto a su aplicación del proceso de Descubrimiento de Conocimiento en Base de Datos. En el capítulo 2 se presenta la propuesta del trabajo que contempla la información técnica del mismo, referenciando el tipo de trabajo y su clasificación, la descripción del problema y las preguntas básicas que ayudan a su análisis; además se conoce la delimitación funcional que determina el alcance del trabajo de investigación y desarrollo. El capítulo 3 hace referencia al marco teórico; en la sección 3.1 se abordan las definiciones y conceptos importantes que involucra el proceso KDD, en tanto que en la sección 3.2 se abordan temas referentes a la implementación del proceso KDD, como la descripción de sus fases y la Metodología HEFESTO versión 2.0. Además, se analizan proyectos que implementaron Descubrimiento de Conocimiento en Base de Datos y *Data Mining* que permiten sustentar este trabajo propuesto. En el capítulo 4 se presenta la metodología del trabajo, siendo lo más importante la propuesta realizada del proceso de Descubrimiento de Conocimiento en Base de Datos acoplado a la Metodología HEFESTO y la descripción de cada una de sus fases. El capítulo 5 muestra los resultados de las 9 fases de KDD aplicado en la Unidad de Nivelación y Admisión – ESPOCH; además del análisis de información obtenido mediante *Data Mining* y reportes de *Business Intelligence*. En el capítulo 6 se muestran las conclusiones y recomendaciones obtenidas posterior a la aplicación del proceso de Descubrimiento de Conocimiento.

En la sección Apéndices A, B, C, se muestra en detalle las tareas realizadas en las fases KDD.

Capítulo 2

Planteamiento de la Propuesta de Trabajo

2.1. Información técnica básica

Tema: Descubrimiento de Conocimiento en Base de Datos para la toma de decisiones en la Unidad de Nivelación Y Admisión de la ESPOCH

Tipo de trabajo: Proyecto de Investigación y Desarrollo

Clasificación técnica del trabajo: Aplicación

Líneas de Investigación, Innovación y Desarrollo

Principal: Sistemas de Información y/o Nuevas Tecnologías de la Información y Comunicación y sus aplicaciones

2.2. Descripción del problema

El principal problema de la Unidad de Nivelación y Admisión de la ESPOCH es que se está generando desde el 2012 gran cantidad de datos que no están procesados, ni almacenados en base de datos, haciendo que únicamente se manejen archivos planos, de los cuales no se pueden obtener reportes, ni información para la toma de decisiones. Haciendo que la Institución no pueda analizar ni tomar decisiones de la demanda que ésta posee de cada una de las carreras, los principales lugares del país de donde provienen los estudiantes, desenvolvimiento académico, fortalezas académicas y oportunidades de mejora.

Se evidencia entonces que es de gran importancia el análisis de la información y la extracción de conocimiento, para realizar la adecuada toma de decisiones a nivel institucional mediante la aplicación del proceso KDD, Descubrimiento de Conocimiento en Base de datos, que permitirá identificar patrones (Juan, Moine, Gordillo, Ana, & Haedo, 2011) válidos, útiles y comprensibles, además estructuras que conviertan los datos en información útil para alcanzar un trabajo organizado y permitir la implementación de acciones pertinentes.

2.3. Preguntas básicas

¿Cómo aparece el problema que se pretende solucionar? Se genera gran cantidad de datos en la Unidad de Nivelación y Admisión de la Escuela Superior Politécnica de Chimborazo

desde el 2012, ya que maneja los procesos de admisión y nivelación de los estudiantes que obtuvieron un cupo para el ingreso en el Sistema de Educación Superior Ecuatoriano.

¿Por qué se origina? Por la falta de un proceso para el Descubrimiento de Conocimiento en Base de datos que maneje herramientas que permitan analizar y tomar decisiones a nivel institucional.

¿Qué lo origina? La falta de sistemas y herramientas tecnológicas que manejen la gran cantidad de datos de los procesos de admisión y nivelación en la Unidad de Nivelación y Admisión de la Escuela Superior Politécnica de Chimborazo.

¿Cuándo se origina? Al inicio y al final del periodo académico de los Cursos de Nivelación de la Escuela Superior Politécnica de Chimborazo.

2.4. Formulación de meta

Disponer de información para la toma de decisiones (en el tiempo o momento requerido) en la Unidad de Nivelación y Admisión de la ESPOCH, mediante un proceso para el Descubrimiento de Conocimiento en Base de datos.

2.5. Objetivos

Objetivo General

Implementar el proceso de Descubrimiento de Conocimiento en Base de datos para la toma de decisiones en la Unidad de Nivelación y Admisión de la ESPOCH.

Objetivos Específicos

1. Fundamentar los referentes teóricos y metodológicos sobre el Descubrimiento de Conocimiento en Base de datos para la toma de decisiones.
2. Diseñar un proceso de Descubrimiento de Conocimiento en Base de datos adecuado a los requerimientos de la Unidad de Nivelación y Admisión de la ESPOCH.
3. Crear un *Data Warehouse opensource* que contenga información consistente y adecuada para la toma de decisiones de los procesos de admisión y nivelación.
4. Analizar técnicas de *Data Mining* que sean apropiadas para los datos de la Unidad de Nivelación y Admisión de la ESPOCH.
5. Generar reportes de *Business Intelligence* que sirvan de análisis de los datos de la Unidad de Nivelación y Admisión de la ESPOCH.

2.6. Delimitación funcional

Pregunta 1. ¿Qué será capaz de hacer el producto final del proyecto de titulación?

- Se implementarán las fases de Descubrimiento de Conocimiento en Base de datos para la toma de decisiones.
- Se realizará el análisis de la información para permitir que ésta sea consistente y adecuada para la toma de decisiones.
- Se elaborará *un Data Warehouse* que contenga la información académica relevante para la toma de decisiones, como calificaciones de estudiantes por materia, periodo académico y carrera, histórico de postulaciones de las diferentes carreras de la institución.
- Se generarán reportes *Business Intelligence* primordiales para la toma de decisiones, donde se podrán analizar lugares de procedencia de los estudiantes, cantidad de estudiantes por carrera, sexo y periodo académico, estudiantes aprobados, reprobados por periodo por carreras. Además, información histórica de postulaciones de los estudiantes por periodo académico y carrera.
- Se seleccionarán dos técnicas de *Data Mining* que permitan identificar los patrones de los datos de la Unidad de Nivelación y Admisión permitiendo conocer el estado de la información.

Pregunta 2. ¿Qué no será capaz de hacer el producto final del proyecto de titulación?
No aplica.

Capítulo 3

Marco Teórico

3.1. Definiciones y conceptos

3.1.1. Data Warehousing

Como lo menciona Bernabeu (2010), el *Data Warehousing* posibilita la extracción de datos de sistemas operacionales y fuentes externas, permite la integración y homogeneización de los datos de toda la empresa, provee información que ha sido transformada y sumarizada, para que ayude en el proceso de toma de decisiones estratégicas y tácticas.

El *Data Warehousing*, convierte entonces los datos operacionales de la empresa en una herramienta competitiva, debido a que pondrá a disposición de los usuarios indicados la información pertinente, correcta e integrada, en el momento que se necesita.

Para que este pueda cumplir con sus objetivos, es necesario que la información que se extrae, transforma y consolida, sea almacenada de manera centralizada en una base de datos con estructura multidimensional denominada *Data Warehouse* (DW) (Bernabeu, 2010).

3.1.2. Data Warehouse

Una de las definiciones más famosas sobre *Data Warehouse* (DW), es la de William Harvey Inmon, quien define: "Un Data Warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia" (Inmon, 1992).

Además, respecto al diseño del DW, la visión de Kimball se basa en que son los procesos de negocio los que deben de marcar la forma en la que se diseña. Admite un punto de partida en la que ya existen datos más o menos organizados en *data marts*, y que estos son la base del futuro *data Warehouse* (Kimball, 1996).

Para Kimball lo más importante es que el cálculo de los datos que servirá para la toma de decisiones sea rápido, por lo que estructura los datos del DW siguiendo patrones dimensionales. Esto suele mejorar mucho su rendimiento a la hora de realizar consultas y además organiza los datos de una forma más intuitiva y natural para los usuarios. Por todo esto se considera a la arquitectura de Kimball como una aproximación *bottom-up* del problema, ya que se parte de los

datos y procesos existentes y se modela el DW para que se adapte a ellos, tomando como premisas la eficiencia en tiempo y la representación natural de datos a costa de la normalización.

Figura 1: Data Warehouse

Fuente: HEFESTO, (2010)

Las principales características de DW (Bernabeu, 2010) son:

- Transforma datos orientados a las aplicaciones en información orientada a la toma de decisiones.
- Integra y consolida diferentes fuentes de datos (internas y/o externas) y departamentos empresariales, que anteriormente formaban islas, en una única plataforma sólida y centralizada.
- Provee la capacidad de analizar y explotar las diferentes áreas de trabajo y de realizar un análisis inmediato de las mismas.
- Permite reaccionar rápidamente a los cambios del mercado.
- Aumenta la competitividad en el mercado.
- Elimina la producción y el procesamiento de datos que no son utilizados ni necesarios, producto de aplicaciones mal diseñadas o ya no utilizadas.
- Mejora la entrega de información, es decir, información completa, correcta, consistente, oportuna y accesible. Información que los usuarios necesitan, en el momento adecuado y en el formato apropiado.
- Logra un impacto positivo sobre los procesos de toma de decisiones. Cuando los usuarios tienen acceso a una mejor calidad de información, la empresa puede lograr por sí misma: aprovechar el enorme valor potencial de sus recursos de información y transformarlo en valor verdadero; eliminar los retardos de los procesos que resultan de información incorrecta, inconsistente y/o inexistente; integrar y optimizar procesos a través del uso compartido e integrado de las fuentes de información; permitir a los usuarios adquirir mayor confianza acerca de sus propias decisiones y de las del resto, y lograr así, un mayor entendimiento de los impactos ocasionados.

- Aumento de la eficiencia de los encargados de tomar decisiones.
- Los usuarios pueden acceder directamente a la información en línea, lo que contribuye a su capacidad para operar con mayor efectividad en las tareas rutinarias o no. Además, pueden tener a su disposición una gran cantidad de valiosa información multidimensional, presentada coherentemente como fuente única, confiable y disponible en sus estaciones de trabajo. Así mismo, los usuarios tienen la facilidad de contar con herramientas que les son familiares para manipular y evaluar la información obtenida en el DW, tales como: hojas de cálculo, procesadores de texto, software de análisis de datos, software de análisis estadístico, reportes, tableros, etc.
- Permite la toma de decisiones estratégicas y tácticas (Bernabeu, 2010).

3.1.3. OLTP

OLTP (*On Line Transaction Processing*), representa toda aquella información transaccional que genera la empresa en su accionar diario, además, de las fuentes externas con las que puede llegar a disponer. Estas fuentes de información, son de características muy disímiles entre sí, en formato, procedencia, función, etc.

Entre los OLTP más habituales que pueden existir en cualquier organización se encuentran:

- Archivos de textos.
- Hipertextos.
- Hojas de cálculos.
- Informes semanales, mensuales, anuales, etc.
- Bases de datos transaccionales.

Para poder extraer los datos desde los OLTP, para luego manipularlos, integrarlos y transformarlos, para posteriormente cargar los resultados obtenidos en el DW, es necesario contar con técnicas de Integración de Datos (Bernabeu, 2010).

3.1.4. ETL

Los procesos ETL (Extracción, Transformación y Carga) son sólo una de las muchas técnicas de la Integración de Datos. Mediante este proceso se extraen los datos indicados manteniéndolos en un almacenamiento intermedio; se transforman los datos realizando análisis para verificar que sean correctos y válidos y por último se cargan los datos en el DW (Bernabeu, 2010).

3.1.5. Limpieza de datos

El objetivo principal de la limpieza de datos es realizar distintos tipos de acciones contra el mayor número de datos erróneos, inconsistentes e irrelevantes.

Un punto importante que se debe tener en cuenta al elegir alguna acción, es el de identificar el por qué de la anomalía, para luego actuar en consecuencia, con el fin de evitar que se repitan, agregándole de esta manera más valor a los datos de la organización (Bernabeu, 2010).

3.1.6. OLAP

El procesamiento analítico en línea OLAP (*On Line Analytic Processing*), es la componente más poderosa del *Data Warehousing*, ya que es el motor de consultas especializado del depósito de datos.

Las herramientas OLAP, son una tecnología de software para análisis en línea, administración y ejecución de consultas, que permiten inferir información del comportamiento del negocio. Su principal objetivo es el de brindar rápidas respuestas a complejas preguntas, para interpretar la situación del negocio y tomar decisiones. Cabe destacar que lo que es realmente interesante en OLAP, no es la ejecución de simples consultas tradicionales, sino la posibilidad de utilizar operadores tales como *drill-up, drill-down*, etc, para explotar profundamente la información.

Además:

- Permite recolectar y organizar la información analítica necesaria para los usuarios y disponer de ella en diversos formatos, tales como tablas, gráficos, reportes, tableros de control, etc.
- Soporta análisis complejos de grandes volúmenes de datos.
- Complementa las actividades de otras herramientas que requieran procesamiento analítico en línea.
- Presenta a los usuarios una visión multidimensional de los datos (matricial) para cada tema de interés del negocio.
- Es transparente al tipo de tecnología que soporta el DW, ya sea ROLAP, MOLAP u HOLAP.
- No tiene limitaciones con respecto al número máximo de dimensiones permitidas.
- Permite a los usuarios, analizar la información basándose en más criterios que un análisis de forma tradicional.

- Al contar con muestras grandes, se pueden explorar mejor los datos en busca de respuestas.
- Permiten realizar agregaciones y combinaciones de los datos de maneras complejas y específicas, con el fin de realizar análisis más estratégicos (Bernabeu, 2010).

3.1.7. Data Mining

Esta herramienta constituye una poderosa tecnología con un gran potencial que ayuda y brinda soporte a los usuarios, con el fin de permitirles analizar y extraer conocimientos ocultos y predecibles a partir de los datos almacenados en un DW o en un OLTP, siendo una mejor fuente un DW por todas las ventajas que aporta.

Implementar *Data Mining* permite analizar factores de influencia en determinados procesos, predecir o estimar variables o comportamientos futuros, segmentar o agrupar ítems similares, además de obtener secuencias de eventos que provocan comportamientos específicos, siendo su principal ventaja inferir en comportamientos, modelos, relaciones y estimaciones de los datos, para poder desarrollar predicciones sobre los mismos, sin la necesidad de contar con patrones o reglas preestablecidas, permitiendo tomar decisiones proactivas y basadas en un conocimiento acabado de la información.

Además, brinda la posibilidad de dar respuesta a preguntas complicadas sobre los temas de interés, como por ejemplo ¿Qué está pasando?, ¿Por qué? y ¿Qué pasaría sí?, estos cuestionamientos aplicados a una empresa podrían ser: ¿Cuál de los productos de tal marca y clase serán más vendidos en la zona norte en el próximo semestre? y ¿por qué? Además, se podrán ver los resultados en forma de reportes tabulares, matriciales, gráficos, tableros, etc.

Entonces, se puede definir *Data Mining* como una técnica para descubrir patrones y relaciones entre abundantes cantidades de datos, que a simple vista o que mediante otros tipos de análisis no se pueden deducir, ya que tradicionalmente consumiría demasiado tiempo o estaría fuera de las expectativas (Bernabeu, 2010).

Los sistemas *Data Mining* se desarrollan bajo lenguajes de última generación basados en Inteligencia Artificial y utilizan métodos matemáticos tales como:

- Redes Neuronales.
- Sistemas Expertos.
- Programación Genética.
- Árboles de Decisión.

Soporta además, sofisticadas operaciones de análisis como los sistemas *Scoring*, aplicaciones de Detección de Desviación y Detección de Fraude. Es muy importante tener en cuenta que en las herramientas OLAP y en los reportes y consultas, el análisis parte de una pregunta o hipótesis generada por los usuarios, en cambio *Data Mining* permite generar estas hipótesis.

Generalmente las herramientas de *Data Mining* se integran con plataformas de hardware y software existentes (como DW) para incrementar el valor de las fuentes de datos establecidas y para que puedan ser integradas con nuevos productos y sistemas en línea (como OLAP). En adición a esto, hacer minería de datos sobre un depósito de datos permite entre otras ventajas contar con los beneficios de los procesos ETL y de las técnicas de limpieza de datos, tan necesarios en este tipo de análisis (Bernabeu, 2010).

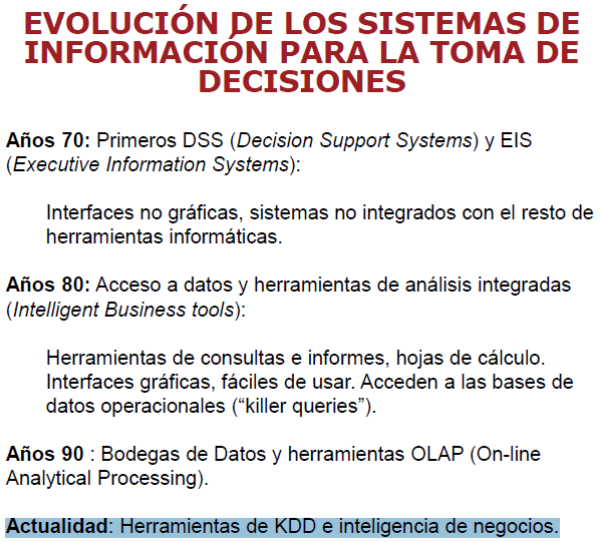
3.1.8. Business Intelligence

Según Bernabeu (2010), se puede describir *Business Intelligence* (BI), como un concepto que integra por un lado el almacenamiento y por el otro el procesamiento de grandes cantidades de datos, con el principal objetivo de transformarlos en conocimiento y en decisiones en tiempo real, a través de un sencillo análisis y exploración.

Este conocimiento debe ser oportuno, relevante, útil y debe estar adaptado al contexto de la organización. Existe una frase muy popular acerca de BI, que dice: "Inteligencia de Negocios es el proceso de convertir datos en conocimiento y el conocimiento en acción, para la toma de decisiones" (Bernabeu, 2010). BI hace hincapié en los procesos de recolectar y utilizar efectivamente la información, con el fin de mejorar la forma de operar de una organización, brindando a sus usuarios, el acceso a la información clave que necesitan para llevar a cabo sus tareas habituales y más precisamente, para poder tomar decisiones oportunas basadas en datos correctos y certeros.

Al contar con la información exacta y en tiempo real, es posible, además, identificar y corregir situaciones antes de que se conviertan en problemas y en potenciales pérdidas de control de la empresa, pudiendo conseguir nuevas oportunidades o readaptarse frente a la ocurrencia de sucesos inesperados. La Inteligencia de Negocios tiene sus raíces en los Sistemas de Información Ejecutiva (*Executive Information Systems* – EIS) y en los Sistemas para la Toma de Decisiones (*Decision Support Systems* – DSS), pero ha evolucionado y se ha transformado en todo un conjunto de tecnologías capaces de satisfacer a una gran gama de usuarios junto a sus necesidades específicas en cuanto al análisis de información (Bernabeu, 2010).

Figura 2: Evolución de Sistemas para toma de decisiones



Fuente: Claudia Jiménez R, (2015)

3.1.9. Conocimiento

De acuerdo a la definición de Harlan Cleveland (1985) el conocimiento es la "información organizada, interiorizada por uno, integrada con todo lo que se conoce, desde la experiencia, el estudio o la intuición y por lo tanto es útil para guiar la vida y el trabajo", mientras que para Alejandro Pavez (2000), este es "las creencias cognitivas, confirmadas, experimentadas y contextuadas del conocedor sobre el objeto a conocer, las cuales estarán condicionadas por el entorno, y serán potenciadas y sistematizadas por las capacidades de dicho conocedor, las que establecen las bases para la acción objetiva y la generación de valor".

Acorde a estas definiciones, el conocimiento con una base en la experiencia, el entorno e información permite que sea el apoyo a la toma de decisiones, además el elemento más importante que una vez adquirido permite repotencializar, planificar y generar acciones de perfeccionamiento del área donde se aplique. Lo transcendental del conocimiento es que este sea real, es decir, obtenido de fuentes válidas y consistentes que aseguren un conocimiento adecuado.

Figura 3: Metáfora del embudo del Business Intelligence

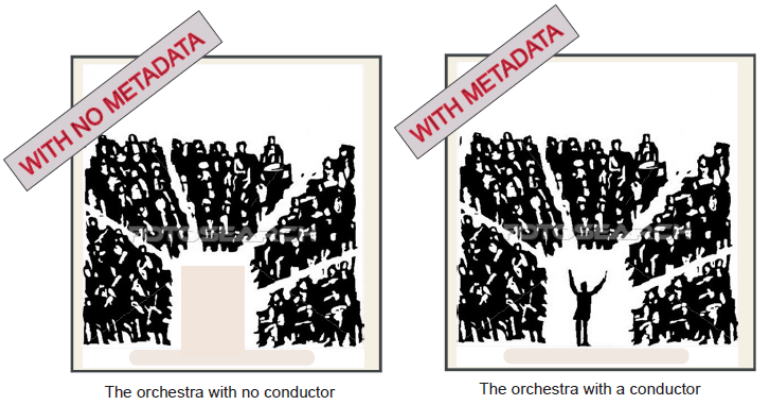
Fuente: http://www.slideshare.net/escenaenelmar/gestion-del-conocimiento-presentation-591517.

3.1.10. Metadatos

Metadato puede definirse como datos en datos, así como datos acerca de datos. porque se añade información para refinar los datos y aumentar el nivel de detalle, con el objetivo de exponer patrones y tendencias.

Como lo menciona Inmon (2010), los metadatos son el pegamento que une los diferentes componentes operativos de DW. En cierto sentido, los metadatos forman una canal que contiene los componentes de DW y permite que estos trabajen en cooperación y coordinación.

Figura 4: Representación de Metadatos



Fuente: Data Warehousing, (2010)

La figura muestra como los metadatos pueden hacer las veces de conductores, ya que estos no dirigen una orquesta, pero si el entorno de DW con sus componentes (Inmon, W, 2010).

3.1.11. ROLAP

ROLAP (*Relational On Line Analytic Processing*) cuenta con todos los beneficios de una SGBD Relacional a los cuales se les provee extensiones y herramientas para poder utilizarlo como un Sistema Gestor de DW. En los sistemas ROLAP, los cubos multidimensionales se generan dinámicamente al instante de realizar las diferentes consultas, haciendo de esta manera el manejo de cubos transparente los usuarios como lo menciona (Bernabeu, 2010).

La principal desventaja de los sistemas ROLAP, es que los datos de los cubos se deben calcular cada vez que se ejecuta una consulta sobre ellos. Esto provoca que ROLAP no sea muy eficiente en cuanto a la rapidez de respuesta ante las consultas de los usuarios.

3.1.12. MOLAP

MOLAP (*Multidimentional On Line Analytic Processing*) es almacenar físicamente los datos en estructuras multidimensionales de manera que la representación externa y la interna coincidan. Para ello, se dispone de estructuras de almacenamiento específicas (*Arrays*) y técnicas de compactación de datos que favorecen el rendimiento del DW.

MOLAP requiere que en una instancia previa se generen y calculen los cubos multidimensionales, para que luego puedan ser consultados.

Tabla 1: ROLAP vs MOLAP

ROLAP	MOLAP
Brinda mucha flexibilidad, ya que los cubos son generados dinámicamente al momento de ejecutar las consultas.	Cada vez que se requiere o es necesario realizar cambios sobre algún cubo, se debe tener que recalcularlo totalmente, para que se reflejen las modificaciones llevadas a cabo. Provocando de esta manera una disminución importante en cuanto a flexibilidad.
Los datos de los cubos se deben calcular cada vez que se ejecuta una consulta sobre ellos. Esto provoca que ROLAP no sea muy eficiente en cuanto a la rapidez de respuesta ante las consultas de l@s usuari@s.	Las consultas son respondidas con mucha rapidez, ya que los mismos no deben ser calculados en tiempo de ejecución, obteniendo de esta manera una muy buena performance.

Fuente: HEFESTO, (2010)

3.1.13. HOLAP

HOLAP (Hybrid On Line Analytic Processing) constituye un sistema híbrido entre MOLAP y ROLAP, que combina estas dos implementaciones para almacenar algunos datos en un motor relacional y otros en una base de datos multidimensional. Los datos agregados y precalculados se almacenan en estructuras multidimensionales y los de menor nivel de detalle en estructuras relacionales. Es decir, se utilizará ROLAP para navegar y explorar los datos, y se empleará MOLAP para la realización de tableros. Como contrapartida, hay que realizar un buen análisis para identificar los diferentes tipos de datos (Bernabeu, 2010).

3.2. Estado del Arte

3.2.1. Análisis comparativo de metodologías para la gestión de proyectos de minería de datos

Como lo menciona Moine, Haedo, & Gordillo (2011), la sistematización del proceso de minería de datos es un punto importante para la planificación y ejecución de este tipo de proyecto. Algunas organizaciones implementan el modelo KDD, mientras que otras aplican un estándar más específico como CRISP-DM. Si la organización ha adquirido productos de la empresa SAS, tiene a su disposición una metodología especialmente desarrollada para los mismos, la metodología SEMMA. Por otro lado, la metodología Catalyst (conocida como P3TQ) está ganando cada vez mayor popularidad debido a su completitud y flexibilidad para adaptarse en distintos escenarios. En el presente trabajo se realiza un análisis comparativo de las diferentes metodologías vigentes para minería de datos, evaluando no sólo la estructura del proceso, sino también aspectos importantes para la gestión del proyecto.

Tabla 2: Fases del proceso de Minería de Datos de cada Modelo

Fases	KDD	CRISP – DM	SEMMA	CATALYST
Análisis y comprensión del negocio	Comprensión del dominio de aplicación	Comprensión del negocio		Modelado del negocio
Selección y preparación de los datos	Crear el conjunto de datos Limpieza y pre-procesamiento de los datos Reducción y proyección de los datos	Entendimiento de los datos Preparación de los datos	Muestreo Comprensión Modificación	Preparación de los datos
Modelado	Determinar la tarea de minería Determinar el algoritmo de minería Minería de datos	Modelado	Modelado	Selección de herramientas y modelado inicial
Evaluación	Interpretación	Evaluación	Valoración	Refinamiento del modelo
Implementación	Utilización del nuevo conocimiento	Despliegue		Comunicación

Fuente: Análisis comparativo de metodologías para la gestión de proyectos de minería de datos, (2011)

En este análisis los tres autores, demuestran que se ha evidenciado la existencia de dos tipos de modelos para llevar a cabo el proceso de minería de datos. Por un lado, se encuentran aquellos que están más cercanos a un modelo de proceso, ya que sólo proponen las fases generales para el proceso de minería de datos y no incorporan actividades para la gestión del proyecto. Estos modelos son KDD y SEMMA, los cuales no llegan a ser una metodología propiamente dicha y dejan a criterio del equipo de trabajo la definición de las actividades a realizar en cada etapa del proyecto. Particularmente SEMMA excluye dos etapas importantes del proceso como son el análisis del negocio y la difusión del nuevo conocimiento, evidenciando que el modelo está orientado especialmente a aspectos técnicos.

Por otro lado, los modelos CRISP-DM y Catalyst podrían ser considerados una metodología, por el nivel de detalle con el que describen las tareas en cada fase del proceso, y porque incorporan actividades para la gestión del proyecto (como gestión del tiempo, costo, riesgo). En este aspecto, ninguno de los dos modelos incorpora actividades para el control y monitoreo del plan de trabajo. La metodología Catalyst sobresale en su fase de Modelado del Negocio (MII), contemplando cinco puntos de partida para el proyecto, que finalmente conducirán a la definición de un conjunto de requerimientos y a una situación organizacional que deberá ser abordada desde la minería de datos. Si hablamos entonces de metodologías para la gestión de un proyecto de minería de datos, los modelos a tener en cuenta deberían ser CRISP-DM y Catalyst, los

cuales se encuentran en un nivel similar de completitud en la mayoría todos los aspectos evaluados (Moine, Haedo, & Gordillo, 2011).

Análisis del artículo. De acuerdo a lo expuesto por los autores Moine, Haedo y Gordillo (2011) sobre los modelos o metodologías de KDD: CRISP-DM, SEMMA y Catalyst, éstas presentan a breves rasgos los pasos que se deben realizar dentro de cada fase, siendo CRISP-DM la más usada pero no tan específica para una fácil implementación; SEMMA tiene más enfoques técnicos y excluye el análisis del negocio, parte importante para determinar la muestra de datos con los que se va a trabajar; Catalyst (P3TQ) no llega a explicar en un alto nivel de detalle las tareas a realizarse.

Lo que referencia y admite la aplicación y desarrollo de las fases KDD dentro de diferentes negocios donde cada área de aplicación puede especificar las actividades y tareas propias a realizarse a cabalidad asegurando el manejo adecuado de los datos y las técnicas de minería de datos que presenten patrones de análisis para la adecuada y eficiente toma de decisiones cumpliendo con el objetivo del Descubrimiento de Conocimiento en base de datos sin la aplicación de una metodología definida.

Dicho en otras palabras, el proceso KDD puede ser desarrollado mediante el uso de herramientas propias que permitan cumplir con el objetivo de cada fase, como lo hace a continuación el siguiente proyecto de investigación y desarrollo acoplado a las necesidades de la Unidad de Nivelación y Admisión – ESPOCH.

3.2.2. Descubrimiento de Conocimiento en Base de Datos

El Descubrimiento de Conocimiento en Bases de Datos (KDD, *Knowledge Discovery in Databases*) es un análisis automático, exploratorio y de modelado de los repositorios de datos de gran tamaño. KDD es el proceso organizado de identificación válida, novedosa, útil y que genera patrones comprensibles de los conjuntos de datos grandes y complejos. *Data Mining* (DM) es el núcleo del proceso de KDD, que implica la inferencia de algoritmos que exploran los datos, desarrollan el modelo y descubre patrones. El modelo se utiliza para la comprensión de los fenómenos de los datos, análisis y la predicción. Además, la accesibilidad y la abundancia de datos de hoy en día hace del Descubrimiento de Conocimiento y Minería de datos una cuestión de considerable importancia y necesidad.

Según Maimon y Rokach (2010), el proceso de descubrimiento de conocimiento es iterativo e interactivo y tiene nueve pasos. Este proceso tiene muchos aspectos artísticos en el sentido de que no se puede presentar una sola fórmula o hacer una taxonomía completa para las decisiones correctas para cada paso y aplicación. Por lo tanto, se requiere profundamente entender

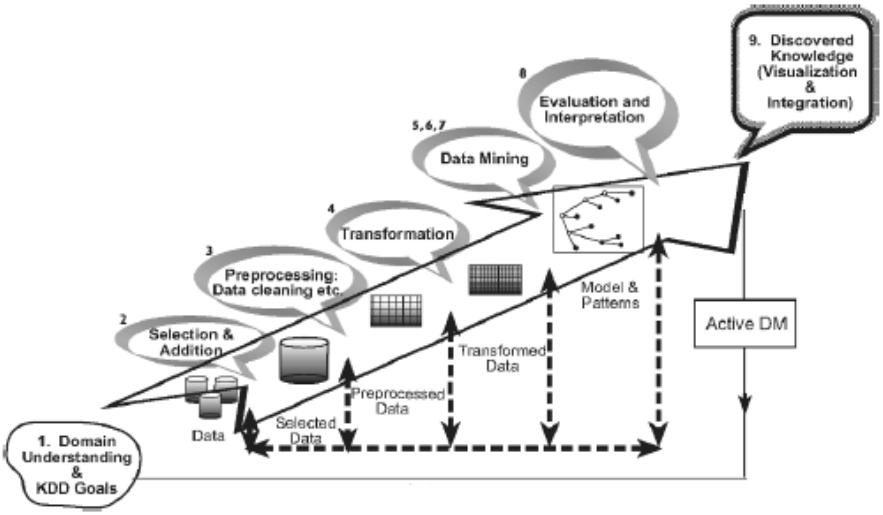
el proceso y las diferentes necesidades y posibilidades de cada paso. La taxonomía para los métodos de minería de datos es importante en este proceso.

El proceso comienza con la determinación de los objetivos de KDD, y termina con la implementación del conocimiento descubierto. Como resultado, los cambios tendrían que ser realizados en el dominio de la aplicación (por ejemplo, ofreciendo diferentes características para los usuarios de teléfonos móviles con el fin de reducir problemas). Esto cierra el círculo, y se miden entonces los efectos en los nuevos repositorios de datos, y el proceso KDD se pone en marcha de nuevo (Maimon & Rokach, 2010).

Fayyad (1996) menciona que "El descubrimiento de conocimiento en bases de datos es el proceso no trivial de identificar patrones en datos que sean válidos, novedosos, potencialmente útiles y, por último, comprensibles"(Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Además "Data Mining se refiere al acto de extraer patrones o modelos a partir de los datos" (Fayyad et al., 1996).

A continuación, se presenta una descripción del proceso de KDD de nueve pasos, comenzando con una etapa de gestión:

Figura 5: El proceso de descubrimiento de conocimiento en bases de datos.



Fuente: Data Mining and Knowledge Discovery Handbook, (2010)

3.2.3. Fases de Descubrimiento de Conocimiento en Base de Datos

3.2.3.1. Aprender el dominio de la aplicación

De acuerdo a Maimon y Rokach (2010), este es el paso inicial, en el que se prepara la escena para la comprensión de lo que debe hacerse con las decisiones (sobre la transformación, algoritmos, representación, etc.). En un proyecto KDD es necesario entender y definir los objetivos del usuario final y el medio ambiente en el que el proceso de descubrimiento de conocimiento se

llevará a cabo (incluyendo el conocimiento previo relevante). A medida que el proceso KDD avanza, puede haber incluso una revisión y puesta a punto de este paso.

De acuerdo a Saldaña y Flores (2005) aprender el dominio de la aplicación implica adquirir conocimiento del área de estudio del sistema y la meta a obtener. Esta etapa puede ser descompuesta en tres áreas:

a) Aprendizaje del tema: Se debe conocer el proceso detrás de la generación de la información para poder formular las preguntas correctas, seleccionar las variables relevantes a cada pregunta, interpretar los resultados y sugerir el curso de acción después de concluido el análisis.

b) Recolección de datos: Se debe conocer donde se encuentran los datos correctos, cómo fueron obtenidos los datos de varias fuentes, cómo se pueden combinar estos datos y el grado de confianza de cada fuente.

c) Experiencia en análisis de datos: Se debe tener conocimientos adecuados en el uso de la estadística (Reyes Saldaña & García Flores, 2005).

3.2.3.2. Selección y creación de un conjunto de datos sobre la que se realizará el descubrimiento

Teniendo definidos los objetivos, de acuerdo a los autores Maimon y Rokach (2010) los datos que se utilizarán para el descubrimiento de conocimiento deben estar determinados, esto incluye conocer los datos con los que se dispone, obteniendo datos necesarios adicionales, y luego la integración de todos los datos para el descubrimiento de conocimiento en un conjunto de datos, incluyendo los atributos que serán considerados para el proceso. Este proceso es muy importante debido a que la minería de datos aprende y descubre a partir de los datos disponibles. Esta es la base de pruebas para la construcción de los modelos. Si algunos atributos importantes faltan, entonces todo el estudio puede fallar.

Desde el éxito del proceso, es bueno tener en cuenta el mayor número posible de atributos en este punto. Por otro lado, recoger, organizar y operar datos complejos de repositorios es caro, y hay un equilibrio con la oportunidad para la mejor comprensión de los fenómenos. Este equilibrio representa un aspecto en el que el aspecto interactivo e iterativo del KDD está teniendo lugar, se inicia con el mejor conjunto de datos disponibles, se expande y se observa después el efecto en términos de descubrimiento de conocimiento y de modelado (Maimon & Rokach, 2010).

3.2.3.3. Pre procesamiento y limpieza

En esta etapa, la fiabilidad de los datos se ve reforzada al incluirse la limpieza de datos, tales como el manejo de los valores perdidos y la eliminación de ruido o valores atípicos. Puede tratarse de métodos estadísticos complejos, o el uso de algoritmos específicos para minería de datos en este contexto. Por ejemplo, si se sospecha que cierto atributo no es lo suficientemente confiable o tiene demasiados datos que faltan, entonces este atributo podría convertirse en el objetivo de un algoritmo supervisado de minería de datos y un modelo de predicción sería útil para desarrollar los datos que faltan y luego se pueden predecir (Maimon & Rokach, 2010).

Otras tareas del pre procesamiento no tan evidentes de acuerdo a Saldaña y Flores (2005) son:

a) Derivar nuevos atributos: Crear campos explícitos con relaciones entre los atributos conocidos pueden hacer el análisis más sencillo.

b) Agrupación: Donde hay relaciones uno a muchos en bases de datos, se puede convertir estas relaciones de uno a uno y agregar el campo de conteo o suma, que contabilice todos los registros de la relación (Reyes Saldaña & García Flores, 2005).

La extensión a los que se debe prestar atención en este nivel depende de muchos factores. En cualquier caso, el estudio de estos aspectos es importante, al igual que una visión propia con respecto a los sistemas de información de la empresa.

3.2.3.4. Transformación de datos

Como lo menciona Maimon y Rokach (2010), en esta etapa, se prepara y desarrolla la generación de datos apropiados para la minería de datos. Los métodos incluyen la reducción de dimensión (tales como la selección de características y la extracción y registro de la muestra), y la transformación de atributos (tales como discretización de atributos numéricos y transformación funcional). Este paso es a menudo crucial para el éxito de todo el proyecto KDD, pero usualmente es el proyecto específico. Por ejemplo, en los exámenes médicos, el cociente de atributos a menudo puede ser el factor más importante, y no cada uno por sí mismo. En comercialización, es posible que se tenga que considerar efectos más allá de nuestro control, así como los esfuerzos y cuestiones temporales (tales como el estudio del efecto de la acumulación de la publicidad). Sin embargo, incluso si no se utiliza la transformación desde el principio, puede darse un efecto sorprendente que insinúa la transformación necesaria (en la siguiente iteración). Así, el proceso KDD refleja sobre sí mismo y conduce a una comprensión de la transformación necesaria (como un

conocimiento conciso de un experto en un campo determinado en relación con los principales indicadores clave).

Después de haber completado los cuatro pasos, los siguientes cuatro pasos están relacionados con la parte de minería de datos, donde la atención se centra en los aspectos algorítmicos empleados para cada proyecto (Maimon & Rokach, 2010).

3.2.3.5. Elección de la tarea de minería de datos apropiada

Ahora se debe decidir sobre qué tipo de minería de datos utilizar, por ejemplo, clasificación, regresión, o agrupación. Esto depende principalmente de los objetivos KDD, y también en los pasos anteriores.

Hay dos objetivos principales en la minería de datos: la predicción y la descripción. Predicción se refiere a menudo como la supervisión de minería de datos, mientras que descriptiva de minería de datos incluye los aspectos sin supervisión y visualización de minería de datos. La mayoría de los datos técnicos de minería se basan en el aprendizaje inductivo, donde se construye un modelo explícito o implícitamente por generalizar a partir de un número suficiente de ejemplos de entrenamiento. El supuesto subyacente del enfoque inductivo es que el entrenado modelo es aplicable a los casos futuros. La estrategia también tiene en cuenta el nivel de meta-aprendizaje para el conjunto particular de datos disponibles (Maimon & Rokach, 2010).

3.2.3.6. Elección del algoritmo de minería de datos

Maimon y Rokach (2010) mencionan que es importante tener la estrategia, para poder decidir sobre las tácticas. Esta etapa incluye seleccionar el método específico que se utilizará para la búsqueda de patrones (incluyendo múltiples inductores). Por ejemplo, en la consideración de la precisión frente a la comprensibilidad, el primero es mejor con redes neuronales, mientras que este último es mejor con árboles de decisión. Para cada estrategia de meta-aprendizaje hay varias posibilidades de cómo se puede conseguir resultados.

El Meta-aprendizaje presenta las causas de un algoritmo de minería de datos en donde se conoce si este tiene o no éxito en un problema particular. Por lo tanto, este enfoque intenta comprender las condiciones en virtud de cual algoritmo de minería de datos es el más apropiado. Cada algoritmo tiene parámetros y tácticas de aprendizaje (como diez veces la validación cruzada u otra división para la formación y las pruebas).

3.2.3.7. Empleando el algoritmo de minería de datos

Por último, Maimon y Rokach (2010), señalan que se realiza la aplicación de los algoritmos de Minería de Datos. En este paso es posible que se necesite emplear el algoritmo varias veces

hasta obtener un resultado satisfactorio, por ejemplo, optimizando los parámetros de control del algoritmo, tales como el número mínimo de instancias en una sola hoja de un árbol de decisión.

3.2.3.8. Evaluación

En esta etapa se evalúa e interpreta los patrones extraídos (reglas, fiabilidad etc.), con respecto a los objetivos definidos en el primer paso. Aquí se considera los pasos de pre procesamiento con respecto a su efecto sobre los resultados del algoritmo de minería de datos (por ejemplo, la adición de características en el paso 4, repetición desde ese punto). Además, este paso se centra en la comprensibilidad y la utilidad del modelo inducido, el conocimiento descubierto también está documentado para su posterior uso.

El último paso es el uso y la retroalimentación general sobre los patrones y resultados de la detección obtenido por la minería de datos (Maimon & Rokach, 2010).

3.2.3.9. Usando el conocimiento descubierto

Es el momento para incorporar el conocimiento en otro sistema para la acción futura de acuerdo a Maimon y Rokach (2010). El conocimiento se vuelve activo en el sentido de que se puede realizar cambios en el sistema y medir los efectos. Actualmente el éxito de este paso determina la eficacia de todo el proceso KDD.

Hay muchos desafíos en este paso, como perder el "laboratorio de condiciones "en el que se ha trabajado. Por ejemplo, el conocimiento se descubrió desde cierto escenario estadístico (por lo general muestra) de los datos, pero ahora los datos se convierten en dinámicos, las estructuras de datos pueden cambiar (ciertos atributos se convierten en no disponibles), y el dominio de datos puede ser modificado (tal como, un atributo puede tener un valor que no se suponía antes).

3.2.4. HEFESTO versión 2.0. Metodología propia para la Construcción de un Data Warehouse

La metodología HEFESTO versión 2.0. como lo define Bernabeu (2010) comienza recolectando las necesidades de información de los usuarios obteniendo las preguntas claves del negocio. Luego, se deben identificar los indicadores resultantes de los interrogativos y sus respectivas perspectivas de análisis, mediante las cuales se construirá el modelo conceptual de datos del DW.

Después, se analizan los OLTP para determinar cómo se construyen los indicadores, señalar las correspondencias con los datos fuentes y para seleccionar los campos de estudio de cada perspectiva. Una vez hecho esto, se construye el modelo lógico del depósito, en donde se

define cuál es el tipo de esquema que se debe implementar. Seguidamente, se crean las tablas de dimensiones y las tablas de hechos, para luego efectuar sus respectivas uniones. Por último, utilizando técnicas de limpieza y calidad de datos, procesos ETL, etc, se definen políticas y estrategias para la Carga Inicial del DW y su respectiva actualización.

De acuerdo a Bernabeu (2010) los pasos definidos para la Metodología HEFESTO v2.0 se describen a continuación:

Figura 6: Pasos de la Metodología HEFESTO v2.0.

Fuente: HEFESTO, 2010

3.2.4.1. Análisis de Requerimientos

El primer paso es identificar los requerimientos de los usuarios a través de preguntas que expliciten los objetivos de la organización. Luego, se analizan estas preguntas a fin de identificar cuáles son los indicadores y perspectivas que se toman en cuenta para la construcción del DW. Finalmente se confecciona un modelo conceptual en donde se visualiza el resultado obtenido en este primer paso.

a) Identificar preguntas: El objetivo principal de esta fase, es la de obtener e identificar las necesidades de información clave de alto nivel, que es esencial para llevar a cabo las metas y estrategias de la empresa, y que facilite una eficaz y eficiente toma de decisiones.

La idea central es, que se formulen preguntas complejas sobre el negocio, que incluyan variables de análisis que se consideren relevantes, ya que son estas las que permiten estudiar la información desde diferentes perspectivas.

b) Identificar indicadores y perspectivas: Se debe tener en cuenta que los indicadores, para que sean realmente efectivos son, en general, valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: saldos, promedios, cantidades, sumatorias, fórmulas, etc. En cambio, las perspectivas se refieren a los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder a las preguntas planteadas, por ejemplo: clientes, proveedores, sucursales, países, productos, rubros, etc. Cabe destacar, que el tiempo es muy comúnmente una perspectiva.

c) Modelo conceptual: En esta etapa, se construye un modelo conceptual a partir de los indicadores y perspectivas obtenidas en el paso anterior. A través de este modelo, se puede observar con claridad cuáles son los alcances del proyecto, para luego poder trabajar sobre ellos, además al poseer un alto nivel de definición de los datos, permite que pueda ser presentado ante los usuarios y explicado con facilidad (Bernabeu, 2010).

3.2.4.2. Análisis de los OLTP

Bernabeu (2010) menciona que se analizan las fuentes OLTP para determinar cómo van a ser calculados los indicadores y para establecer las respectivas correspondencias entre el modelo conceptual creado en el paso anterior y las fuentes de datos. Luego, se deben definir los campos que están en cada perspectiva. Finalmente, se amplía el modelo conceptual con la información obtenida en este paso.

a) Conformar indicadores: En este paso se deben explicitar como se calculan los indicadores, definiendo los siguientes conceptos para cada uno de ellos:

- Hechos que lo componen
- Función de sumarización que se utiliza para la agregación

b) Establecer correspondencias: El objetivo de este paso, es el de examinar los OLTP disponibles que contengan la información requerida, como así también sus características, para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos. La idea es, que todos los elementos del modelo conceptual estén correspondidos en los OLTP.

c) Nivel de granularidad: Una vez que se han establecido las relaciones con los OLTP, se deben seleccionar los campos que contiene cada perspectiva, ya que a través de estos se examina y filtran los indicadores. Al momento de seleccionar los campos que

integran cada perspectiva, debe prestarse mucha atención, ya que esta acción determina la granularidad de la información encontrada en el DW.

d) Modelo Conceptual ampliado: En este paso, y con el fin de graficar los resultados obtenidos en los pasos anteriores, se debe ampliar el modelo conceptual, colocando bajo cada perspectiva los campos seleccionados y bajo cada indicador su respectiva fórmula de cálculo (Bernabeu, 2010).

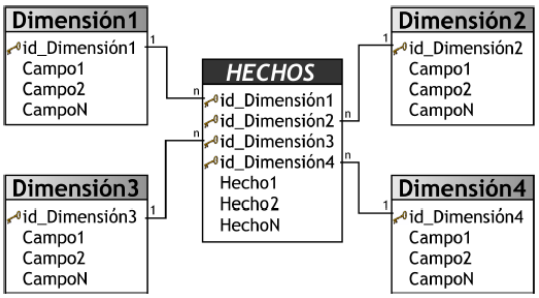
3.2.4.3. Modelo Lógico del DW

A continuación, se confecciona el modelo lógico de la estructura del DW, teniendo como base el modelo conceptual que ya ha sido creado. Para ello, primero se define el tipo de modelo que se va a utilizar y luego se llevan a cabo las acciones propias al caso, para diseñar las tablas de dimensiones y de hechos. Finalmente, se realizan las uniones pertinentes entre estas tablas.

a) Tipo de Modelo Lógico del DW: Se debe seleccionar el tipo de esquema que se va a utilizar para contener la estructura del depósito de datos, que se adapte mejor a los requerimientos y necesidades de los usuarios. Es muy importante definir objetivamente si se va a emplear un esquema en estrella, constelación o copo de nieve, ya que esta decisión afecta considerablemente la elaboración del modelo lógico.

- **Esquema en estrella:** Consta de una tabla de hechos central y de varias tablas de dimensiones relacionadas a esta, a través de sus respectivas claves.

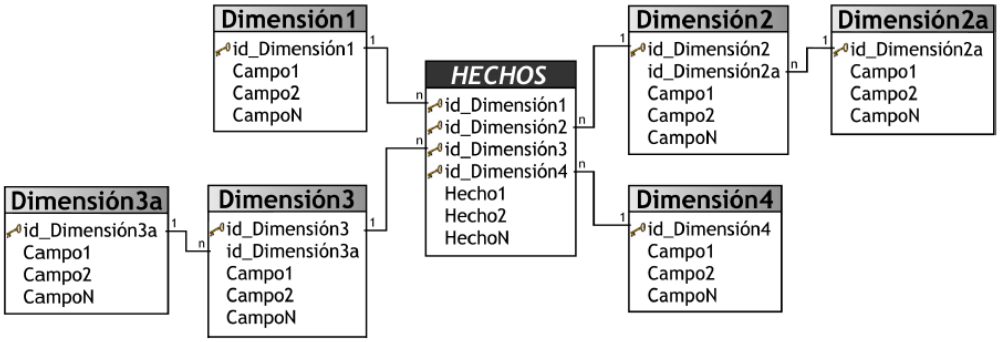
Figura 7: Esquema en Estrella



Fuente: HEFESTO, (2010)

- **Esquema Copo de Nieve:** Este esquema representa una extensión del modelo en estrella cuando las tablas de dimensiones se organizan en jerarquías de dimensiones.

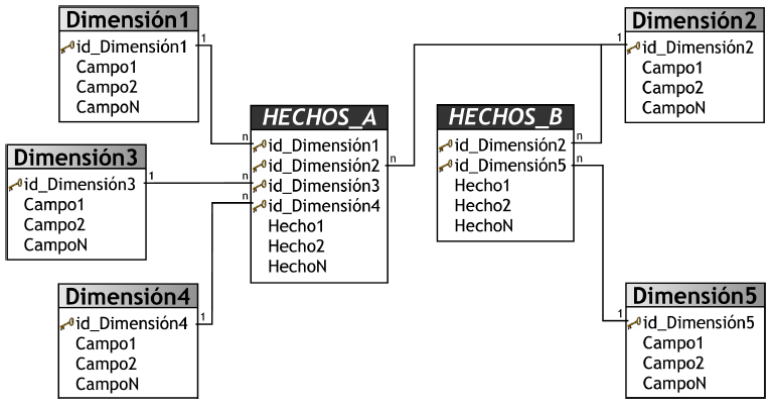
Figura 8: Esquema Copo de Nieve



Fuente: HEFESTO, (2010)

- **Esquema Constelación:** Este modelo está compuesto por una serie de esquemas en estrella, y tal como se puede apreciar en la siguiente figura, está formado por una tabla de hechos principal ("HECHOS_A") y por una o más tablas de hechos auxiliares ("HECHOS_B"), las cuales pueden ser sumarizaciones de la principal. Dichas tablas yacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones.

Figura 9: Esquema Constelación



Fuente: HEFESTO, (2010)

b) Tablas de dimensiones: En este paso se deben diseñar las tablas de dimensiones que forman parte del DW. Para los tres tipos de esquemas, cada perspectiva definida en el modelo conceptual constituye una tabla de dimensión. Para ello se debe tomar cada perspectiva con sus campos relacionados y realizar el siguiente proceso:

- Elegir un nombre que identifique la tabla de dimensión.
- Añadir un campo que represente su clave principal.

- Definir los nombres de los campos si es que no son lo suficientemente intuitivos.

c) Tablas de hechos: En este paso, se definen las tablas de hechos, que son las que contienen los hechos a través de los cuales se construyen los indicadores de estudio.

Para los esquemas en estrella y copo de nieve, se realiza lo siguiente:

- Asignar un nombre a la tabla de hechos que represente la información analizada, área de investigación, negocio enfocado, etc.
- Definir su clave primaria, que se compone de la combinación de las claves primarias de cada tabla de dimensión relacionada.
- Crear tantos campos de hechos como indicadores se hayan definido en el modelo conceptual y se les asigna los mismos nombres que estos. En caso que se prefiera, podrán ser nombrados de cualquier otro modo.

Para los esquemas constelación se realiza lo siguiente:

- Las tablas de hechos se deben confeccionar teniendo en cuenta el análisis de las preguntas realizadas por los usuarios en pasos anteriores y sus respectivos indicadores y perspectivas.
- Cada tabla de hechos debe poseer un nombre que la identifique, contener sus hechos correspondientes y su clave debe estar formada por la combinación de las claves de las tablas de dimensiones relacionadas.
- Al diseñar las tablas de hechos, se deberá tener en cuenta:
    - Caso 1: Si en dos o más preguntas de negocio figuran los mismos indicadores, pero con diferentes perspectivas de análisis, existirán tantas tablas de hechos como preguntas cumplan esta condición.
    - Caso 2: Si en dos o más preguntas de negocio figuran diferentes indicadores con diferentes perspectivas de análisis, existirán tantas tablas de hechos como preguntas cumplan esta condición.
    - Caso 3: Si el conjunto de preguntas de negocio cumple con las condiciones de los dos puntos anteriores se deberán unificar aquellos interrogantes que posean diferentes indicadores, pero iguales perspectivas de análisis, para luego reanudar el estudio de las preguntas.

d) Uniones: Para los tres tipos de esquemas, se realizan las uniones correspondientes entre sus tablas de dimensiones y sus tablas de hechos (Bernabeu, 2010).

3.2.4.4. Integración de Datos

Una vez construido el modelo lógico, se debe proceder a poblarlo con datos, utilizando técnicas de limpieza y calidad de datos, procesos ETL, etc.; luego se definen las reglas y políticas para su respectiva actualización, así como también los procesos que se van a llevar a cabo.

a) Carga Inicial: En este paso se debe poblar el modelo de datos que se ha construido anteriormente. Para lo se lleva adelante una serie de tareas básicas, tales como limpieza de datos, calidad de datos, procesos ETL, etc. La realización de estas tareas puede contener una lógica realmente compleja en algunos casos. Afortunadamente, en la actualidad existen muchos softwares que se pueden emplear a tal fin, y que facilitan el trabajo.

Se debe evitar que el DW sea cargado con valores faltantes o anómalos, así como también se deben establecer condiciones y restricciones para asegurar que sólo se utilicen los datos de interés. Cuando se trabaja con un esquema constelación, hay que tener presente que varias tablas de dimensiones son compartidas con diferentes tablas de hechos, ya que puede darse el caso de que algunas restricciones aplicadas sobre una tabla de dimensión en particular para analizar una tabla de hechos, se puedan contraponer con otras restricciones o condiciones de análisis de otras tablas de hechos.

Primero se cargan los datos de las dimensiones y luego los de las tablas de hechos, teniendo en cuenta siempre, la correcta correspondencia entre cada elemento. En el caso en que se esté utilizando un esquema copo de nieve, cada vez que existan jerarquías de dimensiones, se debe comenzar cargando las tablas de dimensiones del nivel más general al más detallado. Concretamente, en este paso se debe registrar en detalle las acciones llevadas a cabo con los diferentes softwares. Por ejemplo, es muy común que sistemas ETL trabajen con pasos y relaciones, en donde cada paso realiza una tarea en particular del proceso ETL y cada relación indica hacia donde debe dirigirse el flujo de datos. En este caso lo que se debe hacer es explicar que hace el proceso en general y luego que hace cada paso y/o relación. Es decir, se parte de lo más general y se va a lo más específico, para obtener de esta manera una visión general y detallada de todo el proceso. Es importante tener presente, que al cargar los datos en las tablas de hechos pueden utilizarse preagregaciones, ya sea al nivel de granularidad de la misma o a otros niveles diferentes.

b) Actualización: Cuando se haya cargado en su totalidad el DW, se deben establecer sus políticas y estrategias de actualización o refresco de datos. Una vez realizado esto, se tendrán que llevar a cabo las siguientes acciones:

- Especificar las tareas de limpieza de datos, calidad de datos, procesos ETL, etc., que deberán realizarse para actualizar los datos del DW.

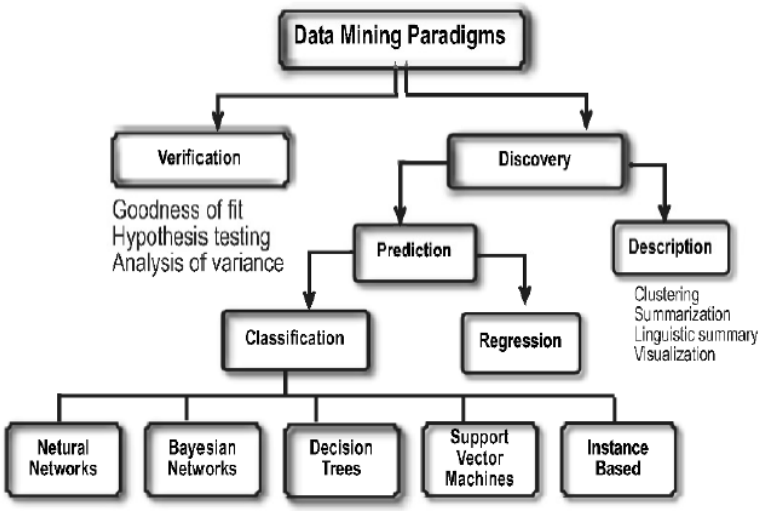
- Especificar de forma general y detallada las acciones que deberá realizar cada software (Bernabeu, 2010).

3.2.5. Data Mining

De acuerdo a Olson y Denle (2008), la minería de datos que se le llama análisis exploratorio de datos, entre otras cosas, se forma a partir de la cantidad de datos generados de distintas maneras como por: cajas registradoras, bases de datos de sistemas de las empresas, etc. Estos datos sirven para el análisis y están siendo explorados, analizados, y reutilizados. Esta búsqueda de información se está realizando a través de diferentes modelos propuestos para la predicción de las ventas, la respuesta de la comercialización y beneficios; enfoques estadísticos clásicos son fundamentales para la minería de datos.

Una variedad de modelos informáticos analíticos se ha utilizado en la minería de datos. Los tipos de modelo estándar en la minería de datos incluyen la regresión (regresión normal, para la predicción, la regresión logística para la clasificación), redes neuronales, y los árboles de decisión, siendo estas técnicas las más conocidas (Olson & Delen, 2008).

Figura 10: Clasificación de técnicas Data Mining



Fuente: Data Mining and Knowledge Discovery Handbook, (2010)

3.2.6. Técnicas Data Mining

Entre las principales técnicas de Data Mining de acuerdo a Bernabeu (2010), se encuentran:

3.2.6.1. Redes Neuronales

Se utilizan para construir modelos predictivos no lineales que aprenden a través de entrenamiento y que semejan la estructura de una red neuronal biológica. Una red neuronal es un modelo computacional con un conjunto de propiedades específicas, como la habilidad de adaptarse o aprender, generalizar u organizar la información, todo ello basado en un procesamiento eminentemente paralelo. Por ejemplo, las redes neuronales pueden emplearse para:

- Resolver problemas en dominios complejos con variables continuas y categóricas.
- Modelizar relaciones no lineales.
- Clasificar y predecir resultados (Bernabeu, 2010).

3.2.6.2. Árboles de Decisión

Son estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos, las cuales explican el comportamiento de una variable con relación a otras, y pueden traducirse fácilmente en reglas de negocio. Son utilizados con finalidad predictiva y de clasificación. Por ejemplo, los árboles de decisión pueden emplearse para:

- Optimizar respuestas de campañas.
- Identificar clientes potenciales.
- Realizar evaluación de riesgos.

3.2.6.3. Redes bayesianas

Según Aluja (2001), consiste en representar todos los posibles sucesos en los que se está interesado mediante un grafo de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones (Aluja, 2001).

3.2.6.4. Técnicas de Clustering

(Fayyad et al., 1996) hace referencia a Clustering como técnicas que parten de una medida de proximidad entre individuos y a partir de ahí, se puede buscar los grupos de individuos más parecidos entre sí, según una serie de variables mesuradas, permitiendo estos grupos realizar análisis.

3.2.6.5. Optimización de Secuencia Mínima (SMOreg)

Es el algoritmo de optimización de secuencia mínima desarrollado por John C. Esta implementación sustituye todos los valores que faltan y transforma atributos nominales en otros binarios. También normaliza todos los atributos por defecto (los coeficientes a la salida están basados en los datos normalizados, no en los originales) (González & García, 2010).

3.2.6.6. Regresión Lineal

Técnica estadística para determinar la relación entre variables. Permite predecir a partir de un muestreo de datos aleatorio. Se adapta a una amplia variedad de situaciones. La regresión ajustada con el error cuadrático medio más bajo se elige como el modelo final (González & García, 2010).

Al aplicar el análisis de funciones LinearRegression, automáticamente se genera un modelo de regresión lineal de predicción. La precisión del modelo generado depende en gran manera de la cantidad de datos que se manejen, así, la exactitud de la predicción es directamente proporcional al número de datos disponibles (Lewandowski, 2015).

3.2.7. La Minería de Datos, entre la Estadística y la Inteligencia Artificial

Aluja (2001), menciona que los datos almacenados son un tesoro para las organizaciones, es donde se guardan las interacciones pasadas con los clientes, la contabilidad de sus procesos internos, representan la memoria de la organización. Pero con tener memoria no es suficiente, hay que pasar a la acción inteligente sobre los datos para extraer la información que almacenan. Este es el objetivo de la Minería de Datos.

En el artículo se abordaos las raíces estadísticas de la minería de datos, los problemas que trata, haremos una panorámica sobre el alcance actual de la minería de datos, presentaremos un ejemplo de su aplicación en el mundo de la audiencia de televisión y, por último, daremos una visión de futuro (Aluja, 2001).

Tabla 3: Comparación de la probabilidad de acierto según 4 métodos de predicción

Problema 1	*Apren.*	*Test*
Análisis Discriminante	71.13%	**69.71%**
Redes Neuronales	71.63%	**69.12%**
Árboles de Clasificación	72.94%	**70.31%**
Regresión logística	74.18%	**71.33%**
Problema 2	*Apren.*	*Test*
Análisis Discriminante	62.18%	**61.39%**
Redes Neuronales	62.29%	**60.19%**
Árboles de Clasificación	62.70%	**61.03%**
Regresión logística	65.28%	**59.36%**

Fuente: La Minería de Datos, entre la Estadística y la Inteligencia Artificial, (2001)

Según Aluja (2001), la experiencia práctica muestra claramente la aptitud de las técnicas de minería de datos para resolver problemas empresariales. También es clara su aportación para resolver problemas científicos que impliquen el tratamiento de grandes cantidades de datos.

La minería de datos es, en realidad, una prolongación de una práctica estadística de larga tradición, la de Análisis de Datos. Existe, además, una aportación propia de técnicas específicas de Inteligencia Artificial, en particular sobre la integración de los algoritmos, la automatización del proceso y la optimización del coste.

A diferencia de la Inteligencia Artificial, que es una ciencia joven, en Estadística se viene aprendiendo de los datos desde hace más de un siglo, la diferencia consiste que ahora existe la potencia de cálculo suficiente para tratar ficheros de datos de forma masiva y automática. Esta es una realidad que cada vez será más habitual. Sin abandonar ninguno de los campos previamente abordados, la Estadística ha evolucionado de ocuparse de la contabilidad de los estados a ser la metodología científica de las ciencias experimentales, hasta ser un *problem solver* para las organizaciones modernas. Es por esta razón el énfasis dado a que los resultados sean accionables.

Por otro lado y en relación a la amplia panoplia de técnicas disponibles, conviene tener claro de que no existe la técnica más inteligente, sino formas inteligentes de utilizar una técnica y que cada uno utiliza de forma inteligente aquello que conoce. También que para la mayoría de problemas no existen diferencias significativas en los resultados obtenidos.

Por todo lo dicho, es nuestra opinión de que la minería de datos no es una moda pasajera, sino que se entronca en una vieja tradición estadística y que cada vez más debe servir para hacer más eficiente el funcionamiento de las organizaciones modernas, ayudar a resolver problemas científicos y ampliar los horizontes de la Estadística.

Análisis del artículo. De acuerdo a lo expuesto por el autor Aluja (2001), a pesar de ser una publicación antigua es importante y válido considerar las aclaraciones que realiza frente a la estadística que se ha desarrollado a lo largo de estos siglos, ya que ésta es una ciencia que no es moderna pero que tampoco sólo se ha convertido en una moda, siendo significativo la forma en cómo se realiza el análisis de los datos mediante su explotación o *Data Mining* que es la aplicación de estadística que permite analizar patrones de información para tomar decisiones que mejorarán el rendimiento de las empresas.

Sustancial y preciso el considerar que el uso de técnicas *Data Mining* va a depender de la información que se use, más no, de la calidad de la técnica, ya que también se demuestra la precisión de las diferentes técnicas que se aplicaron en este estudio, pero se debe considerar, que en realidad va a depender de la visión e intuición de la gente para lograr resolver problemas de las empresas de forma eficiente.

En el presente proyecto de investigación y desarrollo se analizarán técnicas *Data Mining* que permitan manejar la información de relevancia académica en la ESPOCH, permitiendo lograr el descubrimiento de conocimiento.

Capítulo 4

Metodología

4.1. Diagnóstico

El mecanismo empleado para la extracción de la información fue la entrevista, que se la realizó al personal del nivel estratégico de la Unidad de Nivelación y Admisión, donde los Coordinadores: General, Académico y Administrativo identificaron los siguientes problemas:

- Gran cantidad de información manejada de forma independiente en todos los años, debido a que esta información se encuentra distribuida en archivos Excel.
- No contar con reportes que permitan identificar la situación real de los estudiantes en los últimos años.
- Falta de información integrada sobre los procesos académicos de la Unidad de Nivelación y Admisión para la toma de decisiones.

4.2. Método(s) aplicado(s)

4.2.1. De Investigación

El diseño del presente trabajo de investigación y desarrollo es del tipo cuasi experimental, ya que se escogen los algoritmos de *Data Mining*, además se crea una propuesta de Descubrimiento de Conocimiento en base de Datos donde se analizará el conocimiento generado para la toma de decisiones.

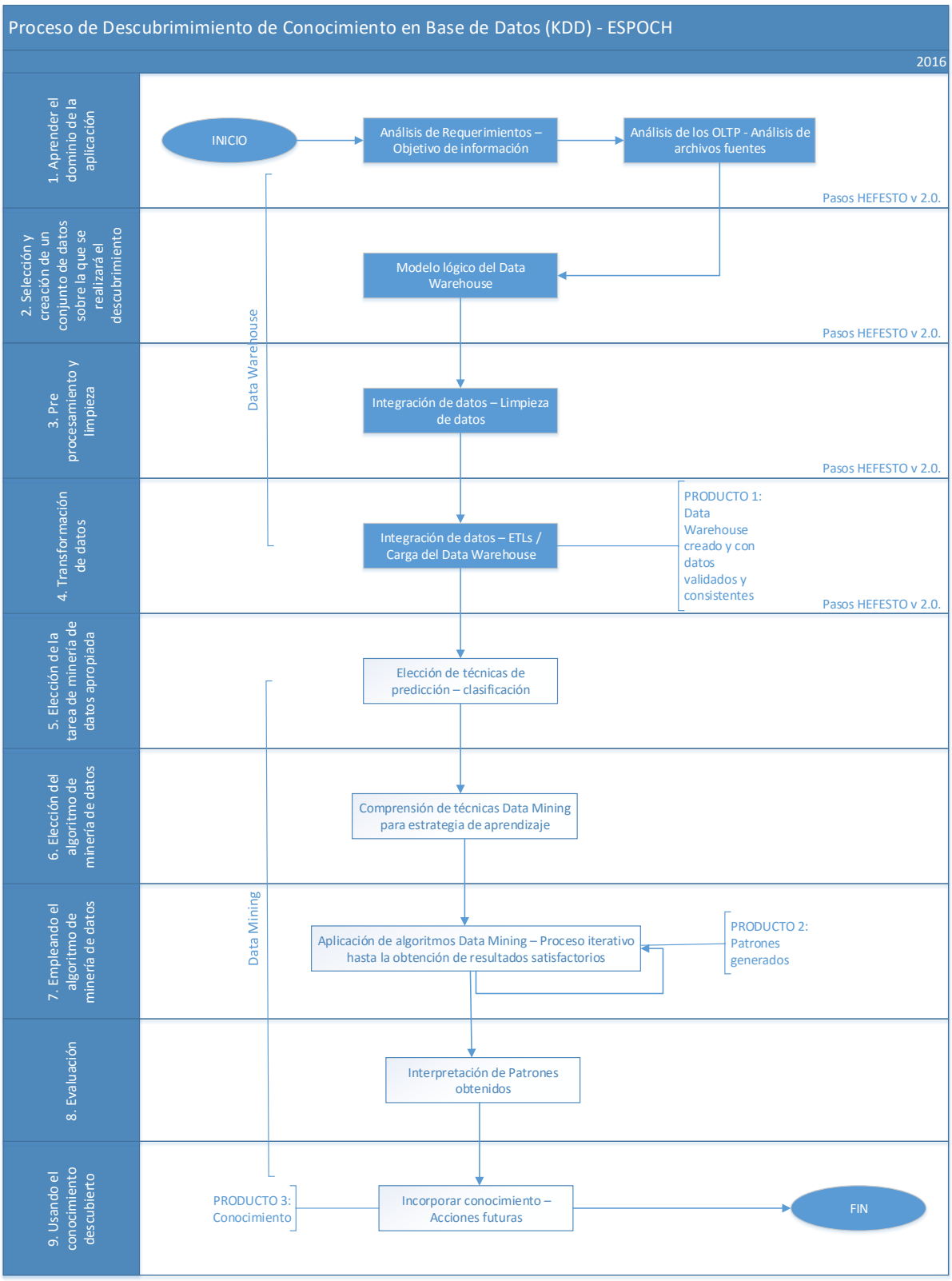
4.2.2. De Desarrollo

La aplicación del Descubrimiento de Conocimiento en base de datos en la Unidad de Nivelación y Admisión de la ESPOCH consta de 9 pasos, donde cada uno de estos refleja de forma clara los procedimientos que se deben realizar para ir cumpliendo dentro de cada paso su objetivo.

Al existir metodologías de KDD y analizarlas, el presente estudio implementa un proceso propio que se acopla a las necesidades de la ESPOCH para diseñar KDD. Además, se utiliza HEFESTO versión 2.0 que es una metodología propia para la Construcción de un *Data Warehouse* que se ajusta al proceso antes mencionado en los primeros cuatro pasos.

Producto de este análisis de metodologías y modelos, a continuación, se describe el proceso propio KDD implementado.

Figura 11: Propuesta de Proceso KDD implementado en la Unidad de Nivelación y Admisión – ESPOCH



Fuente: María Isabel Uvidia Fassler

Dentro del proceso de Descubrimiento de Conocimiento en Base de Datos se han definido dos subprocesos: *Data Warehouse* y *Data Mining*. En cada subproceso se definen las siguientes fases:

Data Warehouse:

1. Aprender el dominio de la aplicación
2. Selección y creación de un conjunto de datos sobre la que se realizará el descubrimiento
3. Pre procesamiento y limpieza
4. Transformación de datos

Data Mining:

5. Elección de la tarea de minería de datos apropiada
6. Elección del algoritmo de minería de datos
7. Empleando el algoritmo de minería de datos
8. Evaluación
9. Usando el conocimiento descubierto

Descripción de cada fase:

4.2.2.1. Aprender el dominio de la aplicación

Este es el primer paso y uno de los más importantes ya que es el punto de partida que permite definir todos los requerimientos de información que se van a tener al desarrollar Descubrimiento de Conocimiento en Base de Datos, conociendo de forma clara el área de aplicación o escena sobre la que se tomarán decisiones.

Como lo considera la metodología HEFESTO versión 2.0. para la elaboración del DW, en este primer paso de KDD se realiza:

1. **Análisis de Requerimientos:** Identifica las necesidades de información clave de alto nivel para llevar a cabo las metas y estrategias de la empresa, analizando las variables de análisis más relevantes que faciliten la adecuada toma de decisiones.
2. **Análisis de los OLTP:** Revisa los archivos fuentes de información, ya sean estos bases de datos de sistemas operacionales o archivos planos como Excel que se encuentren disponibles y contengan la información requerida para cumplir con los objetivos o necesidades de información.

Infiriendo de estos dos primeros pasos de la metodología HEFESTO y del proceso KDD, los siguientes puntos centrales a considerar en la fase: Aprender el dominio de la aplicación, son:

- El objetivo de cada requerimiento de información
- Características de los requerimientos de información
- Fuentes de datos
- Atributos de las fuentes de datos

Cuyos puntos centrales se encuentran organizados en la matriz que a continuación se propone como herramienta para esta primera fase:

Tabla 4: Matriz Aprender el dominio de la aplicación

PROCESO:

Nº	Requerimiento	Descripción	Características	Fuente	Columnas Fuente

Fuente: María Isabel Uvidia Fassler

Donde, el **proceso** se refiere al área de análisis o dominio para la aplicación de requerimientos de información que son también ingresados en las columnas **requerimiento**, **descripción** y **características** de forma desglosada de acuerdo a cada perspectiva u objeto (cliente, estudiante, docente, etc.). En las columnas **fuente** y **columnas fuente** se describe la información disponible en los archivos OLTP. Importante conocer que por cada **proceso** identificado existe una matriz.

4.2.2.2. Selección y creación de un conjunto de datos sobre la que se realizará el descubrimiento

En esta segunda fase del proceso KDD, dentro del subproceso *Data Warehouse,* una vez definidos los objetivos de información y las fuentes disponibles de datos OLTP, es importante completar la información adicional que se requiera y comenzar a trabajar en el diseño del *Data Warehouse* sobre el que se realiza el conocimiento.

Como lo determina HEFESTO en el tercer paso de su metodología, se debe realizar el Modelo Lógico de DW, empezando por su tipo de esquema, sus dimensiones, tablas de hechos (*Fact*) y las relaciones entre estas. Como propuesta dentro de la fase: Selección y creación de un conjunto de datos sobre la que se realizará el descubrimiento, se propone una matriz complementaria a la matriz: Aprender dominio de conocimiento que permite determinar toda la

información base, atributos, tipos de datos, claves principales de las tablas para continuar con el diseño y su modelo conceptual. Se muestra a continuación la matriz.

Tabla 5: Matriz Selección y Creación DW

PROCESO:

Nº	Requerimiento	Descripción	Características	Dimensiones	Atributos	Clave	Tipo de Dato	Fuente	Columnas Fuente

Fuente: María Isabel Uvidia Fassler

Donde las columnas **dimensiones** define la o el conjunto de dimensiones que son parte del diseño y que están relacionadas a la *Fact* o tabla de hecho en ese proceso. **Atributos** define los nombres de las columnas que contendrá cada tabla de Dimensión, además si este atributo es clave principal (PK) y su **tipo de dato**. Toda esta información, permite definir el modelo conceptual del DW. Su tipo de esquema se verá reflejado en el diseño y va a depender del área en donde se aplique, pudiendo tener un esquema Estrella, Copo de Nieve o Constelación.

4.2.2.3. Pre procesamiento y limpieza

Esta tercera fase del subproceso *Data Warehouse* es muy importante para asegurar la consistencia y confiabilidad de la información que se almacena en el DW, ya que mediante el uso de software exclusivo para limpieza de datos como hace referencia la metodología HEFESTO versión 2.0. en la fase Integración de Datos, se puede recabar, validar y analizar datos, además se pueden determinar valores atípicos y faltantes para una vez concluido este paso se pueda garantizar que la toma de decisiones sea adecuada y con datos reales.

Mediante las herramientas de limpieza de datos se pueden programar tareas para encontrar:

- Datos duplicados
- Tipos de datos erróneos
- Mayúsculas y minúsculas
- Caracteres especiales
- Longitudes de datos incorrectos
- Cadenas de datos incompletas, etc.

Siendo de vital importancia asegurar que los datos sean los correctos y confiables, ya que no sería útil que la empresa maneje datos que no reflejen la realidad o que emitan resultados sin tener una línea base de información consistente y adecuada para la toma de decisiones.

4.2.2.4. Transformación de datos

En esta última etapa del subproceso *Data Warehouse*, donde se consigue el primer producto que es el DW, se aplica la técnica de Integración de Información que son ETL (Extracción, Transformación y Carga) que permite cumplir la cuarta fase: Integración de Datos, de la metodología HEFESTO, donde mediante esta técnica se carga en la base de datos la información desde las fuentes, asegurando datos correctos, tipos de datos adecuados (transformación) y se los almacena en el DW que tiene el diseño adecuado para poder brindar información en tiempos de respuesta aceptables y de acuerdo a los objetivos de información que se plantearon en el primer paso.

Después de haber completado los cuatro pasos del subproceso *Data Warehouse* y obteniendo el primer producto: DW con datos que pasaron por procesos de limpieza de datos y ETL, es momento de explotar la información con los siguientes pasos de *Data Mining* que generan conocimiento y permiten la toma de decisiones.

4.2.2.5. Elección de la tarea de minería de datos apropiada

En el primer paso el subproceso de *Data Mining* y quinto paso dentro del proceso KDD, es momento de decidir sobre el tipo de minería a aplicar, ya sea clasificación, regresión, o agrupación. Esto depende principalmente de los objetivos KDD, y también en los pasos anteriores.

Dentro de KDD hay dos objetivos principales en la minería de datos: la predicción y la descripción. Predicción se refiere a la supervisión de minería de datos (técnicas probadas), mientras que descriptiva de minería de datos incluye los aspectos sin supervisión (técnicas no probadas y validadas) y visualización de minería de datos.

Estas técnicas se basan en el aprendizaje inductivo, donde se construye un modelo explícito o implícitamente por generalizar a partir de un número suficiente de ejemplos de entrenamiento. El supuesto subyacente del enfoque inductivo es que el entrenado modelo es aplicable a los casos futuros. La estrategia también tiene en cuenta el nivel de meta-aprendizaje para el conjunto particular de datos disponibles (Maimon & Rokach, 2010).

4.2.2.6. Elección del algoritmo de minería de datos

Una vez definido el objetivo de *Data Mining*, se debe escoger el algoritmo que se aplica a los datos para obtener patrones. Al escoger un algoritmo se debe considerar la precisión frente a comprensibilidad, ya que al tratarse de algoritmos estadísticos es complicado interpretar de la

forma adecuada lo que cada uno de estos muestra. Sin embargo, se deben considerar algoritmos desde los más simples hasta los más complicados, permitiendo introducir al analista en el mundo *Data Mining,* pero también permitiéndole analizar patrones con mayor porcentaje de confiabilidad y además con mucho más conocimiento que interpretar.

4.2.2.7. Empleando el algoritmo de minería de datos

En esta fase una vez definidos los objetivos de KDD y sus algoritmos, es preciso emplear cualquier herramienta de *Data Mining* para aplicar estos algoritmos que generan patrones y que en base a los resultados pueden ser ejecutados varias veces, convirtiéndose en un proceso iterativo, hasta conseguir resultados satisfactorios.

Mediante estos pasos se obtiene el segundo producto del proceso KDD, siendo los Patrones generados el resultado para en el siguiente paso del subproceso *Data Mining* analizarlos y crear conocimiento.

4.2.2.8. Evaluación

En la cuarta fase del subproceso KDD, se evalúa e interpretan los patrones obtenidos (reglas, fiabilidad etc.), con respecto a los objetivos definidos. En este paso se consideran los resultados de minería de datos, es decir, su comprensibilidad y la utilidad del modelo inducido frente al conocimiento descubierto. Considerando siempre como base de información los patrones extraídos para posteriores análisis.

4.2.2.9. Usando el conocimiento descubierto

En la última fase del subproceso *Data Mining* y el proceso de Descubrimiento de Conocimiento en Base de datos – KDD, es momento para incorporar el conocimiento para la acción futura, permitiendo que el conocimiento tome un sentido activo que permita medir efectos y aplicar soluciones, determinando el éxito del proceso KDD.

Siendo este tercer producto o producto final, el conocimiento generado, importante recurso para que una vez analizado se implementen planes de mejora y permita la adecuada toma de decisiones.

4.3. Materiales y herramientas

Para el desarrollo del proceso de Descubrimiento de Conocimiento en base de datos se utilizaron las siguientes herramientas tecnológicas:

Tabla 6: Herramientas tecnológicas utilizadas

HERRAMIENTA TECNOLÓGICA	NOMBRE
Sistema de Gestión de Base de Datos – Data Warehouse	PostgreSQL 9.3[1]
Limpieza de datos	Data Cleaner 4.5.3 [2]
Extracción, Transformación y Carga (ETLs)	Pentaho Data Integration - pdi-ce-5.4.0.1-130. Spoon [3]
Data Mining	WEKA Developer 3.7.13 [4]
Business Intelligence – Reportes	SpagoBI 4.0 [5]

Fuente: María Isabel Uvidia Fassler

[1] http://www.postgresql.org.es/
[2] http://datacleaner.org/
[3] http://community.pentaho.com/projects/data-integration/
[4] http://www.cs.waikato.ac.nz/ml/weka/
[5] http://forge.ow2.org/project/showfiles.php?group_id=204

Capítulo 5

Resultados

5.1. Análisis de resultados

El proceso de Descubrimiento de Conocimiento en base de datos constó de 9 pasos, los 4 primeros determinan la creación del *Data Warehouse* desde su planificación, determinación de objetivos, diseño, hasta conseguir tener la información adecuada y consistente después de pasar por procesos de limpieza y carga de datos; lo que se consiguió mediante la aplicación de la metodología propia para la construcción de *Data Warehouse*, HEFESTO versión 2.0. Los siguientes cinco pasos consistieron en realizar *Data Mining* hasta la generación de conocimiento.

Por tal motivo, se muestra y evidencia el trabajo desarrollado en cada paso de KDD acoplado a las necesidades de la Unidad de Nivelación y Admisión de la ESPOCH, permitiendo determinarlo a continuación:

5.1.1. Aprender el dominio de la aplicación

Este fue el primer paso que se preparó para comprender la escena sobre la que se toman decisiones. Siendo importante analizar la información con la que debe contar el DW de la Unidad de Nivelación y Admisión de la ESPOCH, en base a la metodología de desarrollo propuesta se aplicó la matriz: Aprender el dominio de la aplicación, que consolida la información de tres procesos:

- Ingreso de estudiantes con un cupo para la Escuela Superior Politécnica de Chimborazo
- Información académica de Nivelación
- Aprobados y Reprobados

Para cumplir con el objetivo de manejar toda la información académica de Nivelación y Admisión, fue necesario identificar la información base que se manejó para la obtención de indicadores y perspectivas, siendo las analizadas las siguientes:

- Información de estudiantes
- Información de docentes habilitados
- Información de ubicación académica
- Información de estados de estudiantes

- Información de ubicación geográfica
- Información de tipo de nivelación
- Información de Paralelos
- Información de Periodos
- Información de Currículos

Esta información dentro del diseño del DW se convirtió en dimensiones, y la organización de los procesos de cada matriz permitieron identificar las tablas de hechos o *fact* que tiene el DW. A continuación, se muestran las matrices de cada proceso.

Tabla 7: Matriz Aprender el dominio de la aplicación – Proceso: Ingreso de estudiantes con cupo para la ESPOCH

PROCESO: Ingreso de estudiantes con un cupo para la Escuela Superior Politécnica de Chimborazo

Nº	Requerimiento	Descripción	Características	Fuente	Columnas Fuente
1	Información de estudiantes	Información completa del estudiante de acuerdo a la información enviada por la SENESCYT	Id_Estudiante	MTN_ESTUDIANTES	Campo Incremental
			Nombres		Cédula
			Apellidos		Nombres
			Cédula de Ciudadanía		Apellidos
			Etnia		
			Género		Género
			Fecha de Nacimiento		Fecha_Nacimiento
			Discapacidad		Discapacidad
			Teléfono		Teléfono
			Teléfono móvil		Teléfono móvil
			Correo Electrónico		Correo electrónico
			Unidad Educativa		Unidad Educativa_ue
			Financiamiento		Financiamiento_ue
					Fecha_actual
					"I" Insert
2	Información de Ubicación Académica ESPOCH Información de Áreas SENESCYT	Información completa de la ubicación académica ESPOCH, Sede, Facultad, Escuela, Carrera Información de las áreas	Id_Ubicacion_Academica	Ubicación Academica ESPOCH	Campo Incremental
			Cod_Sede		Cod_Sede
			Sede		Sede
			Cod_Facultad		Cod_Facultad
			Facultad		Facultad
			Cod_Escuela		Cod_Escuela

		especificadas por la SENESCYT	Escuela		Escuela
			Cod_Carrera		Cod_Carrera
			Carrera		Carrera
			Cod_Area		Cod_Area
			Curriculo		Curriculo
			Area		Area
			Subarea_CINE		Subarea_CINE
					Fecha_actual
					"I" Insert
3	Información de Periodo	Información de todos los periodos 2012-2S, 2013-1S, 2013-2S, 2014-1S, 2014-2S, 2015-1S…	Id_Periodo	Periodo	Campo Incremental
			Cod_Periodo		Cod_Periodo
			Periodo		Periodo
			Año		Anio
			Fecha de inicio		Fecha_Inicio
			Fecha fin		Fecha_Fin
					Fecha_actual
					"I" Insert
4	Información de Tipo de Nivelación	Información del tipo de Nivelación, de Carrera, General y Especial	Id_Tipo_Nivelacion	Tipos Nivelacion	Campo Incremental
			Cod_Tipo_Nivelacion		Cod_Tipo_Nivelacion
			Nivelacion		Nivelacion
					Fecha_actual
					"I" Insert
5	Información de Ubicación Geográfica	Información de la ubicación geográfica del Ecuador	Id_Ubicacion_Geografica	Ubicación Geográfica	Campo Incremental
			Cod_Ubicacion_Geografica		Cod_Ubicacion_Geografica
			Pais		Pais
			Region		Region

			Cod_Provincia		Cod_Provincia
			Provincia		Provincia
			Canton		Canton
			Parroquia		Parroquia
					Fecha_actual
					"I" Insert
6	Información de Estado de Estudiante	Información del estado del Estudiante. Exonerado, Asignado, Matriculado, Aprobado, Reprobado	Id_Estado_Estudiante	Estados Estudiantes	Campo Incremental
			Cod_Estado_Estudiante		Cod_Tipo_Nivelacion
			Estado		Nivelacion
					Fecha_actual
					"I" Insert
7	Información del proceso de admisión y nivelación	Información de los estudiantes que obtuvieron un cupo en la ESPOCH. Ingresaron como exonerados o a nivelación con un cupo.	Id_Estudiante		Campo Incremental
			Id_Ubicacion_Academica		Id_Docente_Habilitado
			Id_Periodo		Id_Ubicacion_Academica
			Id_Tipo_Nivelacion		Id_Periodo
			Id_Ubicacion_Geografica		Id_Paralelo
			Id_Estado_Estudiante		Id_Tipo_Nivelacion
			Cod_Matricula		Cod_Matricula
			Año		Anio
			Cod_Area		Cod_Area
					Fecha_actual

Fuente: María Isabel Uvidia Fassler

Tabla 8: Matriz Aprender el dominio de la aplicación – Proceso: Información académica de Nivelación

PROCESO: Información académica de Nivelación

Nº	Requerimiento	Descripción	Características	Fuente	Columnas Fuente
8	Información de docente habilitado	Información completa del docente habilitado	Id_Docente_Habilitado	Disponibilidad_actualizada	Campo Incremental
			Cod_Docente_Habilitado		Cédula
			Nombres		Nombres
			Apellidos		Apellidos
			Género		Género
			Dirección		Dirección Postal
			Habilitacion		Habilitación
			Telefono		Teléfono
			Telefono_movil		Teléfono móvil
			Correo Electrónico		Correo electrónico
			Titulo tercer nivel		Titulo_tercer_nivel
			Titulo cuarto nivel		Titulo_cuarto_nivel
			Titulo mas alto		Titulo_mas_alto
					Fecha_actual
					"I" Insert
9	Información de Paralelo	Información de los paralelos CING-01, CING-02……	Id_Paralelo	Paralelos_Todo	Campo Incremental
			Cod_Paralelo		Cod_Paralelo
			Paralelo		Paralelo
					Fecha_actual
					"I" Insert
10	Información del Curriculo SENESCYT-SNNA	Información del Currículo SENESCYT-SNNA, área, Subárea,	Id_Curriculo	Materias	Campo Incremental
			Cod_Area		Cod_Area
			Curriculo		Curriculo
			Cod_Evaluacion		Cod_Evaluacion

		Evaluación, Materia	Evaluacion		Evaluacion
			Cod_Materia		Cod_Materia
			Módulo		Módulo
			Materia		Materia
			Horas Totales		Horas Totales
			Horas Semanales		Horas Semanales
			Cod_Tipo_Materia		Cod_Tipo_Materia
			Tipo de Materia		Tipo de Materia
			Porcentaje de Evaluacion		Porcentaje de Evaluacion
					Fecha_actual
					"I" Insert
11	Información del proceso académico		Id_Estudiante		Id_Estudiante
			Id_Docente_Habilitado		Id_Docente_Habilitado
			Id_Ubicacion_Academica		Id_Ubicacion_Academica
			Id_Periodo		Id_Periodo
			Id_Paralelo		Id_Paralelo
			Id_Tipo_Nivelacion		Id_Tipo_Nivelacion
			Id_Curriculo		Id_Curriculo
			Id_Estado_Estudiante		Id_Estado_Estudiante
			Cod_Matricula		Cod_Matricula
			Num_Matricula		Num_Matricula
			Examen_Parcial		Examen_Parcial
			Examen_Final		Examen_Final
			Proyecto_de_Aula		Proyecto_de_Aula
			Gestion_Aula_1		Gestion_Aula_1
			Gestion_Aula_2		Gestion_Aula_2

			Examen_Recuperacion		Examen_Recuperacion
			Promedio_Final_Materia		Promedio_Final_Materia
			Asistencia		Asistencia
			Estado_Materia		Estado_Materia
			Año		Anio
			Cod_Area		Cod_Area
			Curriculo		Curriculo
			Cod_Periodo		Cod_Periodo
					Fecha_actual

Fuente: María Isabel Uvidia Fassler

Tabla 9: Matriz Aprender el dominio de la aplicación – Proceso: Aprobados y Reprobados

PROCESO: Aprobados y Reprobados

Nº	Requerimiento	Descripción	Características	Fuente	Columnas Fuente
12	Información del proceso académico: Aprobados y Reprobados	Tabla resumen con estudiantes aprobados y reprobados	Id_Estudiante		Id_Estudiante
			Id_Docente_Habilitado		Id_Docente_Habilitado
			Id_Ubicacion_Academica		Id_Ubicacion_Academica
			Id_Periodo		Id_Periodo
			Id_Paralelo		Id_Paralelo

			Id_Tipo_Nivelacion		Id_Tipo_Nivelacion
			Id_Estado_Estudiante		Id_Estado_Estudiante
			Cod_Matricula		Cod_Matricula
			Num_Matricula		Num_Matricula
			Promedio_Final_Materias		Promedio_Final_Materias
			Examen_Final_Global		Examen_Final_Global
			Promedio_Final_Global		Promedio_Final_Global
			Año		Anio
			Cod_Area		Cod_Area
			Curriculo		Curriculo
			Cod_Periodo		Cod_Periodo
					Fecha_actual

Fuente: María Isabel Uvidia Fassler

Resultado 1: Al aplicar las matrices: Aprender el dominio de la aplicación, en los tres procesos analizados en la Unidad de Nivelación y Admisión, se puede observar en cada una de estas, que se tiene identificado de la forma apropiada el **requerimiento** de información con la **descripción** o detalle de lo que involucra cada requerimiento, siendo importante conocer cuáles son las **características** de cada uno de estos para identificarlas dentro de las **fuentes** de información y en el caso de no poseer información, planificar la forma en cómo se obtienen los datos adicionales. El último **requerimiento** identificado en cada matriz en cada proceso, es el resumen que va a reflejar el proceso y que está directamente relacionado con el resto de requerimientos, proporcionando de esta forma la información global del proceso, convirtiéndose a futuro en la tabla de hechos.

Obteniendo de esta forma como primer resultado el levantamiento de requerimientos de información cumpliendo con el objetivo de manejar todos los datos de la parte académica de la Unidad de Nivelación y Admisión – ESPOCH, además se identificaron las fuentes de información OLTP.

Para el manejo de la información es primordial analizar el estado de la información de los archivos OLTP, siendo la siguiente:

Tabla 10: Estado de archivos OLTP

CARACTERÍSTICA	ESTADO
TIPO DE ARCHIVO OLTP:	Archivos planos en Excel
CANTIDA DE ARCHIVOS:	1270 aproximadamente
ESTADO:	Grado medio de inconsistencias

Fuente: María Isabel Uvidia Fassler

Resultado 2: Se logró identificar que los archivos OLTP son archivos en Excel, siendo alrededor de 1270 con un nivel medio de inconsistencias, lo que se focalizará el tercer paso de limpieza de datos y muestra el trabajo arduo y organizado que se debe realizar para poder constituir la información.

5.1.2. Selección y creación de un conjunto de datos sobre la que se realizará el descubrimiento

Como segundo paso dentro del proceso de KDD fue importante analizar completamente el modelo lógico, atributos del DW y toda la información que permita la construcción de este en base a los requerimientos que se manejaron en el primer paso.

Para lograr este paso se manejó la matriz: Selección y Creación DW, para permitir como siguiente paso de esta fase la implementación del modelo lógico del DW. A continuación, se muestran las matrices por cada proceso identificado.

Tabla 11: Matriz Selección y Creación DW Proceso: Ingreso de estudiantes con un cupo para la ESPOCH

PROCESO: Ingreso de estudiantes con un cupo para la Escuela Superior Politécnica de Chimborazo

Nº	Requerimiento	Descripción	Características	Dimensiones	Atributos	Clave	Tipo de Dato	Fuente	Columnas Fuente
1	Información de estudiantes	Información completa del estudiante de acuerdo a la información enviada por la SENESCYT	Id_Estudiante	DIM_ESTUDIANTE	Id_Estudiante	PK	Serial	MTN_ESTUDIANTES	Campo Incremental
			Nombres		Cod_Estudiante		Varchar(15)		Cédula
			Apellidos		Nombres		Varchar(50)		Nombres
			Cédula de Ciudadanía		Apellidos		Varchar(50)		Apellidos
			Etnia		Etnia		Varchar(20)		
			Género		Genero		Varchar(10)		Género
			Fecha de Nacimiento		Fecha_Nacimiento		Timestamp		Fecha_Nacimiento
			Discapacidad		Discapacidad		Varchar(20)		Discapacidad
			Teléfono		Telefono		Varchar(20)		Teléfono
			Teléfono móvil		Telefono_movil		Varchar(20)		Teléfono móvil
			Correo Electrónico		Correo_Electronico		Varchar(100)		Correo electrónico
			Unidad Educativa		Unidad_Educativa		Varchar(100)		Unidad Educativa_ue
			Financiamiento		Financiamiento		Varchar(50)		Financiamiento_ue
					ETL_Fecha_Carga		Timestamp		Fecha_actual
					ETL_Operacion		Varchar(5)		"I" Insert
2	Información de Ubicación Académica ESPOCH Información de Areas SENESCTY	Información completa de la ubicación académica ESPOCH, Sede, Facultad, Escuela, Carrera Información de las áreas especificadas por la SENESCYT	Id_Ubicacion_Academica	DIM_UBICACION_ACADEMICA	Id_Ubicacion_Academica	PK	Serial	Ubicación Academica ESPOCH	Campo Incremental
			Cod_Sede		Cod_Sede		Varchar(10)		Cod_Sede
			Sede		Sede		Varchar(50)		Sede
			Cod_Facultad		Cod_Facultad		Varchar(10)		Cod_Facultad
			Facultad		Facultad		Varchar(50)		Facultad
			Cod_Escuela		Cod_Escuela		Varchar(10)		Cod_Escuela

			Escuela		Escuela		Varchar(100)		Escuela
			Cod_Carrera		Cod_Carrera		Varchar(10)		Cod_Carrera
			Carrera		Carrera		Varchar(100)		Carrera
			Cod_Area		Cod_Area		Varchar(10)		Cod_Area
			Curriculo		Curriculo		Varchar(150)		Curriculo
			Area		Area		Varchar(150)		Area
			Subarea_CINE		Subarea_CINE		Varchar(150)		Subarea_CINE
					ETL_Fecha_Carga		Timestamp		Fecha_actual
					ETL_Operacion		Varchar(5)		"I" Insert
3	Información de Periodo	Información de todos los periodos 2012-2S, 2013-1S, 2013-2S, 2014-1S, 2014-2S, 2015-1S…	Id_Periodo	DIM_PERIODO	Id_Periodo	PK	Serial	Periodo	Campo Incremental
			Cod_Periodo		Cod_Periodo		Varchar(10)		Cod_Periodo
			Periodo		Periodo		Varchar(50)		Periodo
			Año		Anio		Integer		Anio
			Fecha de inicio		Fecha_Inicio		Timestamp		Fecha_Inicio
			Fecha fin		Fecha_Fin		Timestamp		Fecha_Fin
					ETL_Fecha_Carga		Timestamp		Fecha_actual
					ETL_Operacion		Varchar(5)		"I" Insert
4	Información de Tipo de Nivelación	Información del tipo de Nivelación, de Carrera, General y Especial	Id_Tipo_Nivelacion	DIM_TIPO_NIVELACION	Id_Tipo_Nivelacion	PK	Serial	Tipos Nivelacion	Campo Incremental
			Cod_Tipo_Nivelacion		Cod_Tipo_Nivelacion		Varchar(10)		Cod_Tipo_Nivelacion
			Nivelacion		Nivelacion		Varchar(50)		Nivelacion
					ETL_Fecha_Carga		Timestamp		Fecha_actual
					ETL_Operacion		Varchar(5)		"I" Insert
5	Información de Ubicación Geográfica	Información de la ubicación geográfica del Ecuador	Id_Ubicacion_Geografica	DIM_UBICACION_GEOGRAFICA	Id_Ubicacion_Geografica	PK	Serial	Ubicación Geográfica	Campo Incremental
			Cod_Ubicacion_Geografica		Cod_Ubicacion_Geografica		Varchar(10)		Cod_Ubicacion_Geografica

			Pais		Pais		Varchar(50)		Pais
			Region		Region		Varchar(20)		Region
			Cod_Provincia		Cod_Provincia		Varchar(10)		Cod_Provincia
			Provincia		Provincia		Varchar(50)		Provincia
			Canton		Canton		Varchar(50)		Canton
			Parroquia		Parroquia		Varchar(50)		Parroquia
					ETL_Fecha_Carga		Timestamp		Fecha_actual
					ETL_Operacion		Varchar(5)		"I" Insert
6	Información de Estado de Estudiante	Información del estado del Estudiante. Exonerado, Asignado, Matriculado, Aprobado, Reprobado	Id_Estado_Estudiante	DIM_ESTADO_ESTUDIANTE	Id_Estado_Estudiante	PK	Serial	Estados Estudiantes	Campo Incremental
			Cod_Estado_Estudiante		Cod_Estado_Estudiante		Varchar(10)		Cod_Tipo_Nivelacion
			Estado		Estado		Varchar(50)		Nivelacion
					ETL_Fecha_Carga		Timestamp		Fecha_actual
					ETL_Operacion		Varchar(5)		"I" Insert
7	Información del proceso de admisión y nivelación	Información de los estudiantes que obtuvieron un cupo en la ESPOCH. Ingresaron como exonerados o a nivelación con un cupo.	Id_Estudiante	**FACT_CUPOS**	Id_Estudiante	FK	Serial		Campo Incremental
			Id_Ubicacion_Academica		Id_Ubicacion_Academica	FK	Serial		Id_Docente_Habilitado
			Id_Periodo		Id_Periodo	FK	Serial		Id_Ubicacion_Academica
			Id_Tipo_Nivelacion		Id_Tipo_Nivelacion	FK	Serial		Id_Periodo
			Id_Ubicacion_Geografica		Id_Ubicacion_Geografica	FK	Serial		Id_Paralelo
			Id_Estado_Estudiante		Id_Estado_Estudiante	FK	Serial		Id_Tipo_Nivelacion
			Cod_Matricula		Cod_Matricula		Serial		Cod_Matricula
			Año		Anio		Integer		Anio
			Cod_Area		Cod_Area		Varchar(10)		Cod_Area
					ETL_Fecha_Carga		Timestamp		Fecha_actual

Fuente: María Isabel Uvidia Fassler

Tabla 12: Matriz Selección y Creación DW Proceso: Ingreso académica de Nivelación

PROCESO : Información académica de Nivelación

Nº	Requerimiento	Descripción	Características	Dimensiones	Atributos	Clave	Tipo de Dato	Fuente	Columnas Fuente
8	Información de docente habilitado	Información completa del docente habilitado	Id_Docente_Habilitado	DIM_DOCENTE_HABILITADO	Id_Docente_Habilitado	PK	Serial	Disponibilidad_actualizada	Campo Incremental
			Cod_Docente_Habilitado		Cod_Docente_Habilitado		Varchar(15)		Cédula
			Nombres		Nombres		Varchar(50)		Nombres
			Apellidos		Apellidos		Varchar(50)		Apellidos
			Género		Genero		Varchar(10)		Género
			Dirección		Direccion		Varchar(150)		Dirección Postal
			Habilitacion		Habilitacion		Varchar(20)		Habilitación
			Telefono		Telefono		Varchar(20)		Teléfono
			Telefono_movil		Telefono_movil		Varchar(20)		Teléfono móvil
			Correo Electrónico		Correo_Electronico		Varchar(100)		Correo electrónico
			Titulo tercer nivel		Titulo_tercer_nivel		Varchar(200)		Titulo_tercer_nivel
			Titulo cuarto nivel		Titulo_cuarto_nivel		Varchar(200)		Titulo_cuarto_nivel
			Titulo mas alto		Titulo_mas_alto		Varchar(200)		Titulo_mas_alto
					ETL_Fecha_Carga		Timestamp		Fecha_actual
					ETL_Operacion		Varchar(5)		"I" Insert
9	Información de Paralelo	Información de los paralelos CING-01,	Id_Paralelo	DIM_PARALELO	Id_Paralelo	PK	Serial	Paralelos_Todo	Campo Incremental
			Cod_Paralelo		Cod_Paralelo		Varchar(15)		Cod_Paralelo
			Paralelo		Paralelo		Varchar(50)		Paralelo

		CING-02……			ETL_Fecha_Carga		Timestamp		Fecha_actual
					ETL_Operacion		Varchar(5)		"I" Insert
10	Información del Curriculo SENESCYT-SNNA	Información del Currículo SENESCYT-SNNA, área, Subárea, Evaluación, Materia	Id_Curriculo	DIM_CURRICULO	Id_Curriculo	PK	Serial	Materias	Campo Incremental
			Cod_Area		Cod_Area		Varchar(10)		Cod_Area
			Curriculo		Curriculo		Varchar(150)		Curriculo
			Cod_Evaluacion		Cod_Evaluacion		Varchar(20)		Cod_Evaluacion
			Evaluacion		Evaluacion		Varchar(100)		Evaluacion
			Cod_Materia		Cod_Materia		Varchar(10)		Cod_Materia
			Módulo		Módulo		Varchar(50)		Módulo
			Materia		Materia		Varchar(50)		Materia
			Horas Totales		Horas Totales		Integer		Horas Totales
			Horas Semanales		Horas Semanales		Integer		Horas Semanales
			Cod_Tipo_Materia		Cod_Tipo_Materia		Varchar(10)		Cod_Tipo_Materia
			Tipo de Materia		Tipo de Materia		Varchar(20)		Tipo de Materia
			Porcentaje de Evaluacion		Porcentaje de Evaluacion		Integer		Porcentaje de Evaluacion
					ETL_Fecha_Carga		Timestamp		Fecha_actual
					ETL_Operacion		Varchar(5)		"I" Insert
11	Información del proceso académico		Id_Estudiante	**FACT_ACADEMICO_MATERIA**	Id_Estudiante	FK	Serial		Id_Estudiante
			Id_Docente_Habilitado		Id_Docente_Habilitado	FK	Serial		Id_Docente_Habilitado
			Id_Ubicacion_Academica		Id_Ubicacion_Academica	FK	Serial		Id_Ubicacion_Academica
			Id_Periodo		Id_Periodo	FK	Serial		Id_Periodo
			Id_Paralelo		Id_Paralelo	FK	Serial		Id_Paralelo
			Id_Tipo_Nivelacion		Id_Tipo_Nivelacion	FK	Serial		Id_Tipo_Nivelacion
			Id_Curriculo		Id_Curriculo	FK	Serial		Id_Curriculo
			Id_Estado_Estudiante		Id_Estado_Estudiante	FK	Serial		Id_Estado_Estudiante

			Cod_Matricula		Cod_Matricula	FK	Serial		Cod_Matricula
			Num_Matricula		Num_Matricula		Smallint		Num_Matricula
			Examen_Parcial		Examen_Parcial		Numeric(5,2)		Examen_Parcial
			Examen_Final		Examen_Final		Numeric(5,2)		Examen_Final
			Proyecto_de_Aula		Proyecto_de_Aula		Numeric(5,2)		Proyecto_de_Aula
			Gestion_Aula_1		Gestion_Aula_1		Numeric(5,2)		Gestion_Aula_1
			Gestion_Aula_2		Gestion_Aula_2		Numeric(5,2)		Gestion_Aula_2
			Examen_Recuperacion		Examen_Recuperacion		Numeric(5,2)		Examen_Recuperacion
			Promedio_Final_Materia		Promedio_Final_Materia		Numeric(5,2)		Promedio_Final_Materia
			Asistencia		Asistencia		Numeric(5,2)		Asistencia
			Estado_Materia		Estado_Materia		Varchar(50)		Estado_Materia
			Año		Anio		Integer		Anio
			Cod_Area		Cod_Area		Varchar(10)		Cod_Area
			Curriculo		Curriculo		Varchar(150)		Curriculo
			Cod_Periodo		Cod_Periodo		Varchar(10)		Cod_Periodo
					ETL_Fecha_Carga		Timestamp		Fecha_actual

Fuente: María Isabel Uvidia Fassler

Tabla 13: Matriz Selección y Creación DW Proceso: Aprobados y Reprobados

PROCESO: Aprobados y Reprobados

Nº	Requerimiento	Descripción	Características	Dimensiones	Atributos	Clave	Tipo de Dato	Fuente	Columnas Fuente
12			Id_Estudiante	FACT_ACADEMICO_ESTUDIANTE	Id_Estudiante	FK	Serial		Id_Estudiante

	Información del proceso académico: Aprobados y Reprobados	Tabla resumen con estudiantes aprobados y reprobados	Id_Docente_Habilitado		Id_Docente_Habilitado	FK	Serial		Id_Docente_Habilitado
			Id_Ubicacion_Academica		Id_Ubicacion_Academica	FK	Serial		Id_Ubicacion_Academica
			Id_Periodo		Id_Periodo	FK	Serial		Id_Periodo
			Id_Paralelo		Id_Paralelo	FK	Serial		Id_Paralelo
			Id_Tipo_Nivelacion		Id_Tipo_Nivelacion	FK	Serial		Id_Tipo_Nivelacion
			Id_Estado_Estudiante		Id_Estado_Estudiante	FK	Serial		Id_Estado_Estudiante
			Cod_Matricula		Cod_Matricula	FK	Serial		Cod_Matricula
			Num_Matricula		Num_Matricula				Num_Matricula
			Promedio_Final_Materias		Promedio_Final_Materias		Smallint		Promedio_Final_Materias
			Examen_Final_Global		Examen_Final_Global		Numeric(5,2)		Examen_Final_Global
			Promedio_Final_Global		Promedio_Final_Global		Numeric(5,2)		Promedio_Final_Global
			Año		Anio		Integer		Anio
			Cod_Area		Cod_Area		Varchar(10)		Cod_Area
			Curriculo		Curriculo		Varchar(150)		Curriculo
			Cod_Periodo		Cod_Periodo		Varchar(10)		Cod_Periodo
					ETL_Fecha_Carga		Timestamp		Fecha_actual

Fuente: María Isabel Uvidia Fassler

Considerando en base al análisis de las matrices: Selección y Creación DW, la siguiente información para el diseño del DW:

Tabla 14: Diseño del DW

CARACTERÍSTICA	IMPLEMENTACIÓN
Esquema de DW	Esquema Constelación
Tablas de Dimensión	1. DIM_ESTUDIANTE 2. DIM_UBICACION_ACADEMICA 3. DIM_PERIODO 4. DIM_TIPO_NIVELACION 5. DIM_UBICACION_GEOGRAFICA 6. DIM_ESTADO_ESTUDIANTE 7. DIM_DOCENTE_HABILITADO 8. DIM_PARALELO 9. DIM_CURRICULO
Tablas de Hechos	1. FACT_CUPOS 2. FACT_ACADEMICO_MATERIA
Tabla de Hechos Agregada	1. FACT_ACADEMICO_ESTUDIANTE
Tipo de implementación DW	ROLAP

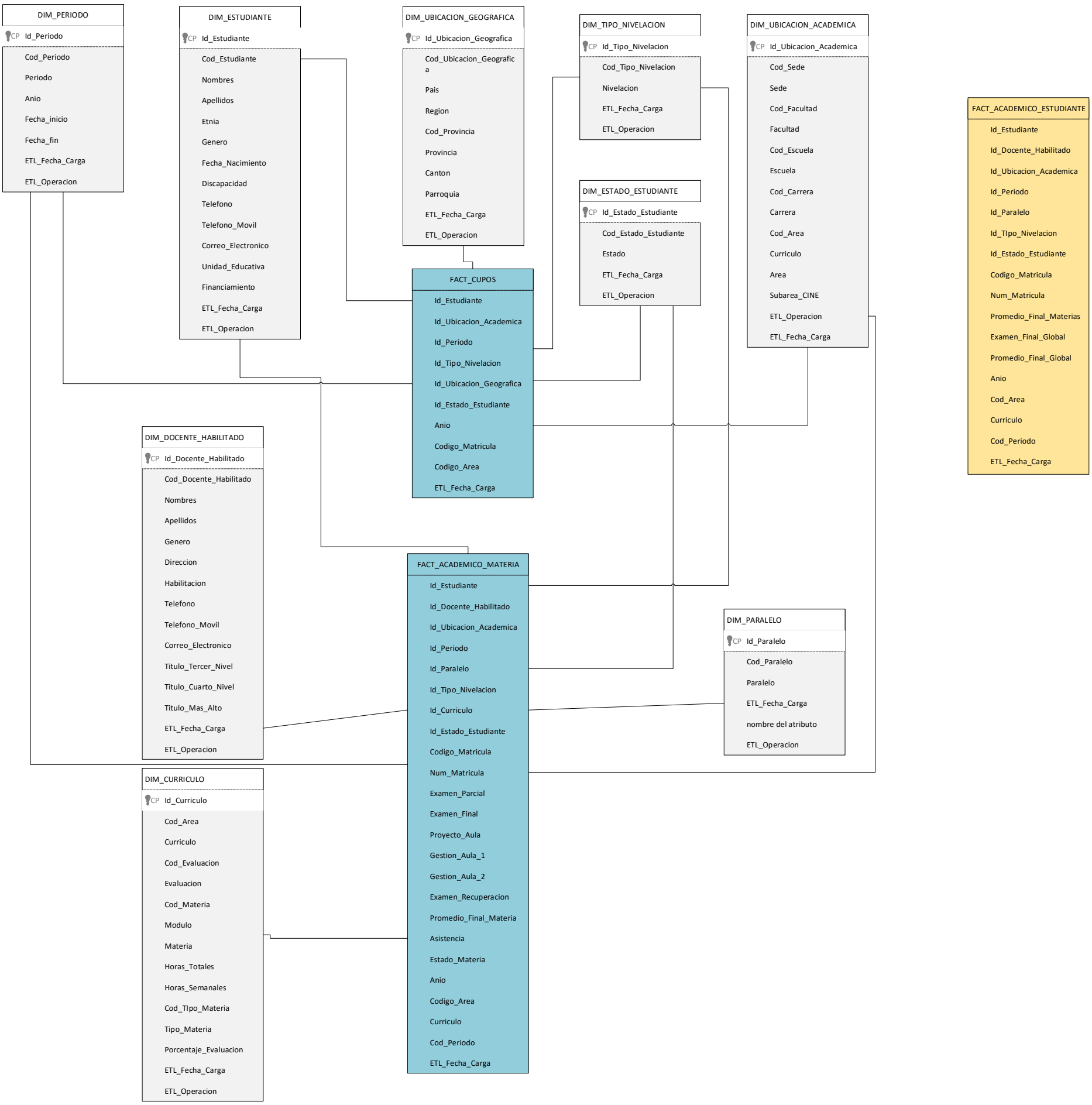
Fuente: María Isabel Uvidia Fassler

Resultado 1: Mediante la aplicación de las matrices: Selección y Creación DW, aplicado a cada proceso, se logró identificar las **dimensiones**, que se generan por cada requerimiento, además sus **atributos**, **claves principales** y **tipos de datos**. Información que muestra la relación entre las tablas para determinar las 9 dimensiones identificadas, además de las 3 tablas de hechos que son la información primordial para la toma de decisiones.

Resultado 2: Se logró identificar que el esquema de diseño de DW a implementarse es un esquema constelación, ya que es una unión entre estrellas. Además la implementación del DW es ROLAP, es decir es una implementación de Análisis en línea - Relacional.

Como siguiente resultado se obtuvo el diagrama del diseño del DW.

Figura 12: Diagrama del diseño del DW



Fuente: María Isabel Uvidia Fassler

Resultado 3: En esta fase se pudo obtener el diseño del DW implementado, cumpliendo con los requerimientos y objetivos de información, como se puede observar en el diagrama de diseño, donde las tablas se encuentran relacionadas y además se conocen los atributos y tipos de datos que se manejan.

Resultado 4: La implementación del DW se la realizó en PostgreSQL 9.3.

5.1.3.Pre procesamiento y limpieza

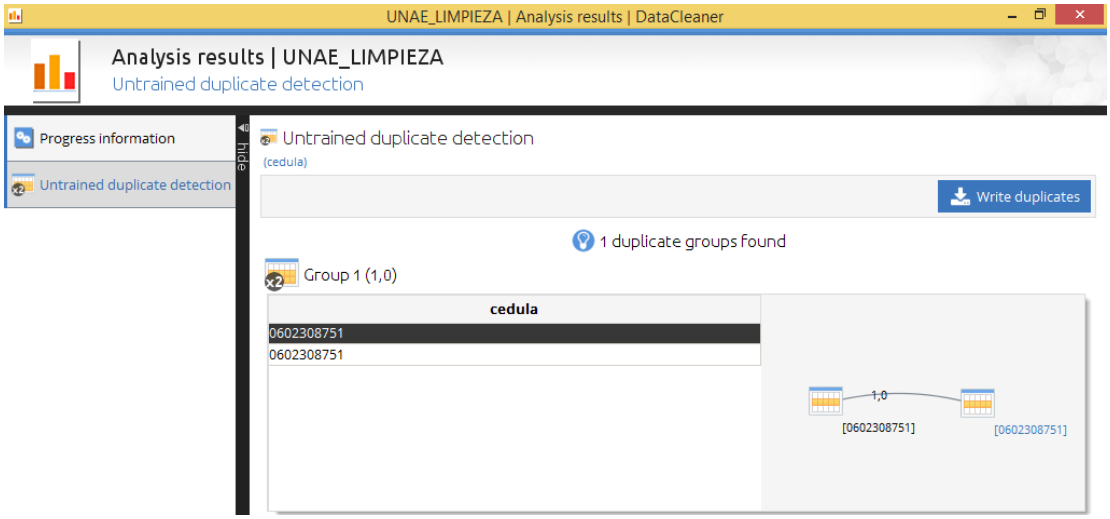
En esta etapa, la fiabilidad de los datos se ve reforzada, por lo tanto, se realizó la limpieza de datos mediante la herramienta *opensource* o de código abierto: *Data Cleaner* 4.5.3.

Figura 13: Data Cleaner

Fuente: *Data Cleaner* 4.5.3

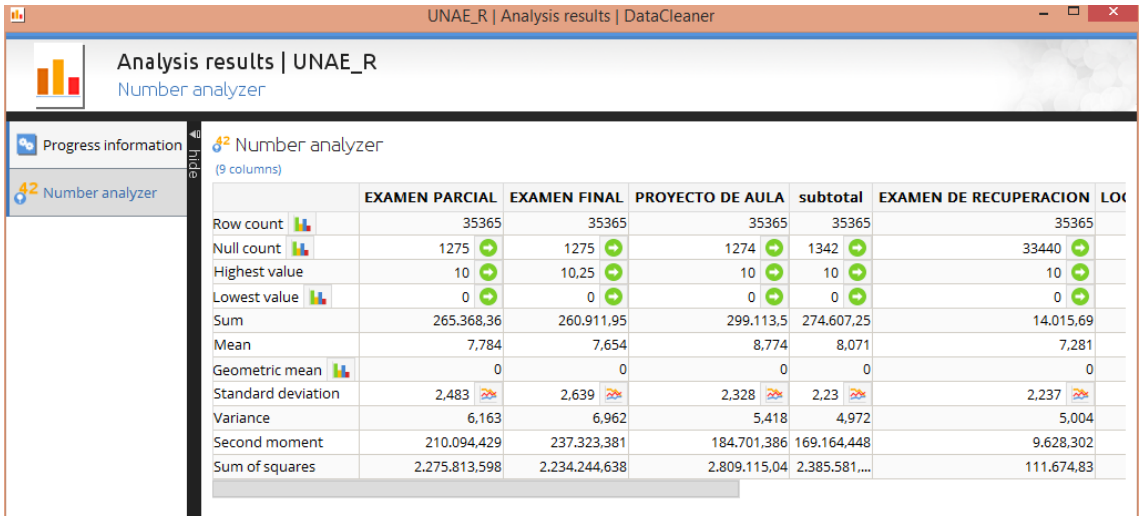
El proceso de limpieza de datos fue determinado en la primera fase de KDD, dentro del subproceso *Data Warehouse*, donde se observó un nivel medio de inconsistencias. En el Apéndice A: Tareas de limpieza de datos, se puede observar en detalle el proceso realizado. A continuación, se muestran algunos de los resultados obtenidos del sin número de tareas de limpieza realizadas.

Figura 14: Resultados de datos duplicados de docentes habilitados



Fuente: *Data Cleaner* 4.5.3

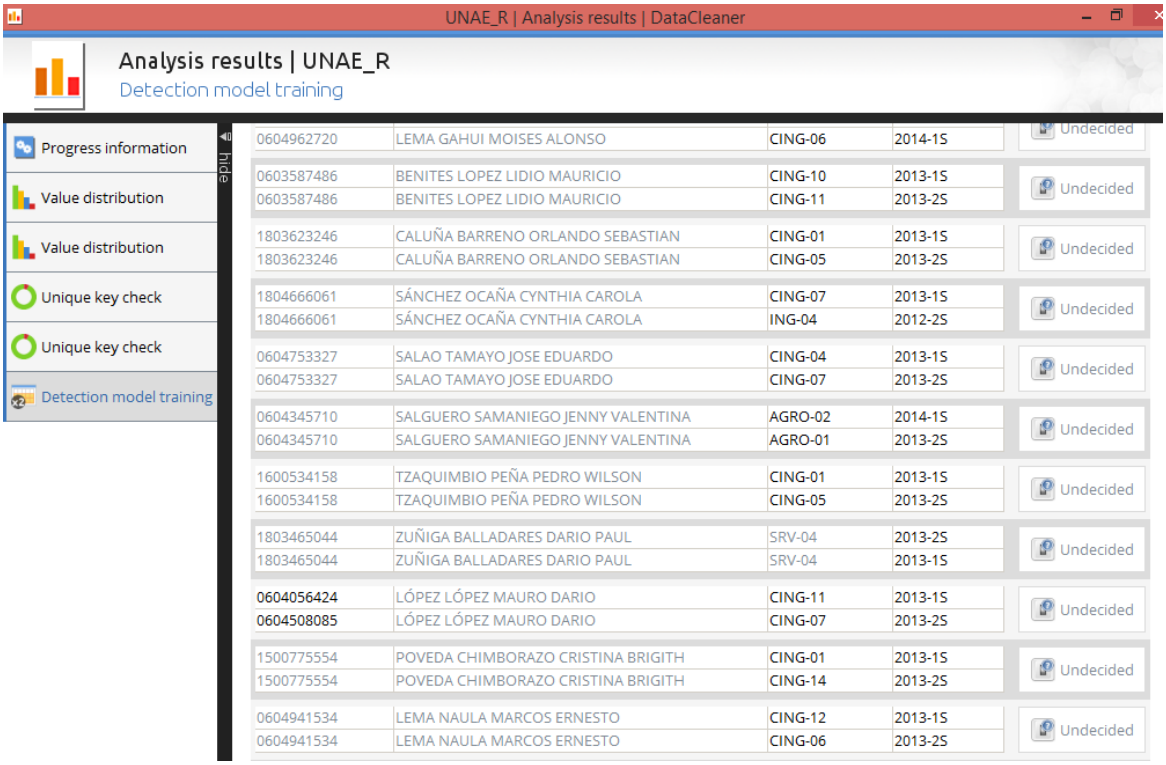
Figura 15: Resultados de análisis de notas y sus intervalos



	EXAMEN PARCIAL	EXAMEN FINAL	PROYECTO DE AULA	subtotal	EXAMEN DE RECUPERACION	LO
Row count	35365	35365	35365	35365	35365	
Null count	1275	1275	1274	1342	33440	
Highest value	10	10,25	10	10	10	
Lowest value	0	0	0	0	0	
Sum	265.368,36	260.911,95	299.113,5	274.607,25	14.015,69	
Mean	7,784	7,654	8,774	8,071	7,281	
Geometric mean	0	0	0	0	0	
Standard deviation	2,483	2,639	2,328	2,23	2,237	
Variance	6,163	6,962	5,418	4,972	5,004	
Second moment	210.094,429	237.323,381	184.701,386	169.164,448	9.628,302	
Sum of squares	2.275.813,598	2.234.244,638	2.809.115,04	2.385.581,...	111.674,83	

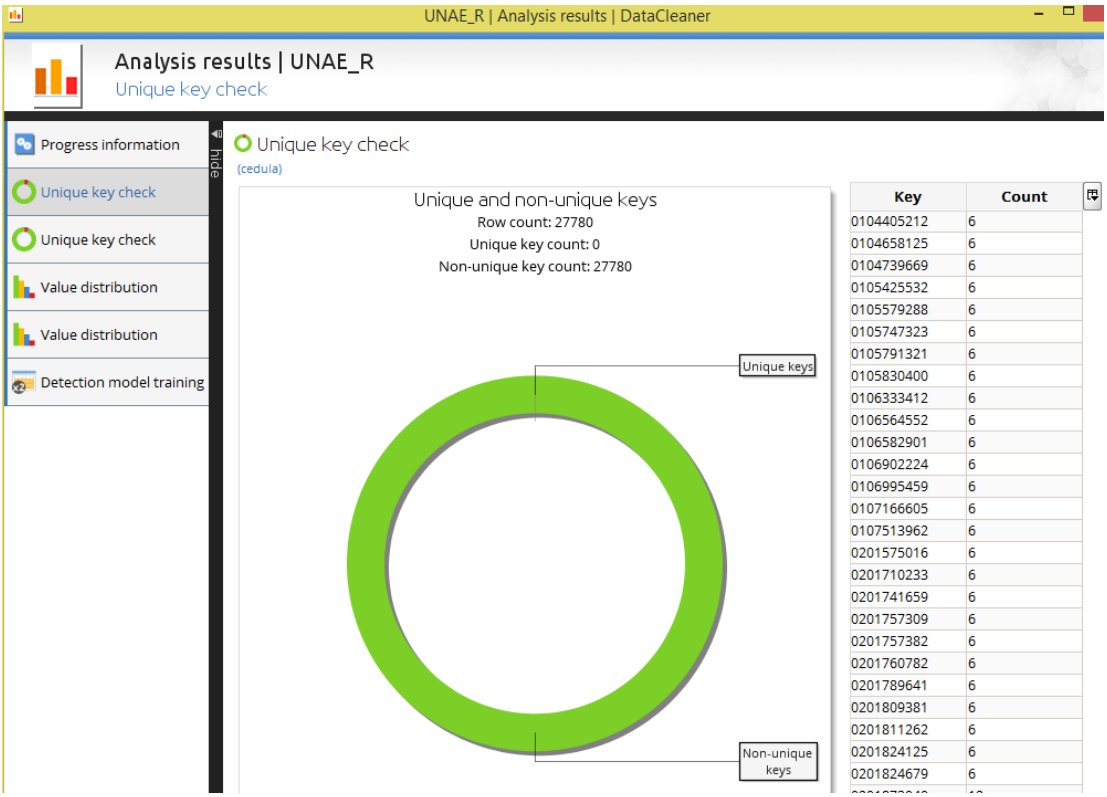
Fuente: *Data Cleaner* 4.5.3

Figura 16: Resultados de análisis cédulas de estudiantes



Fuente: *Data Cleaner* 4.5.3

Figura 17: Resultados de análisis de datos únicos



Fuente: *Data Cleaner* 4.5.3

Resultado 1: Como se pueden observar en las figuras 14, 15, 16 y 17, mediante este proceso se pudo identificar la información inconsistente dentro de cada área de información, como fue la información de los estudiantes, docentes y notas de los Cursos de Nivelación de Carrera. Siendo la duplicación de las cédulas de ciudadanía el principal problema encontrando.

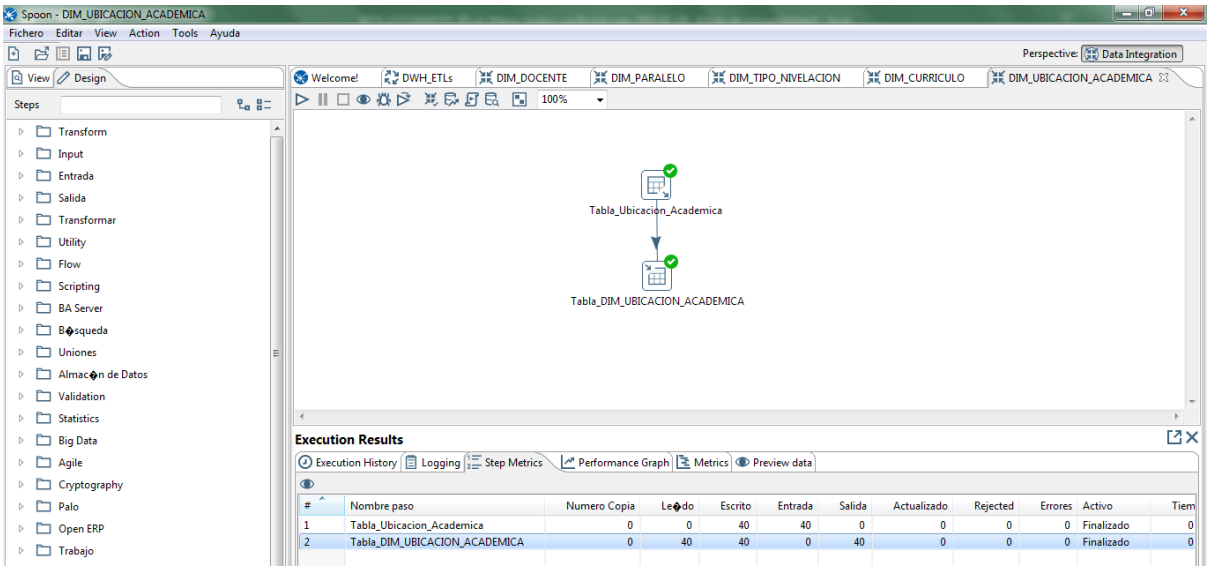
Resultado 2: Se logró y aseguró contar con información consistente y adecuada que permita tomar decisiones y obtener patrones mediante *Data Mining*, posterior a la aplicación de la herramienta *Data Cleaner* 4.5.3 dentro de la fase de Limpieza de Datos.

5.1.4. Transformación de datos

En esta etapa, se cargaron los datos académicos de la Unidad de Nivelación y Admisión ESPOCH, mediante ETL, que es una técnica de Integración de Información, que aparte de cargar la información en el *Data Warehouse,* permitió definir los tipos de datos apropiados.

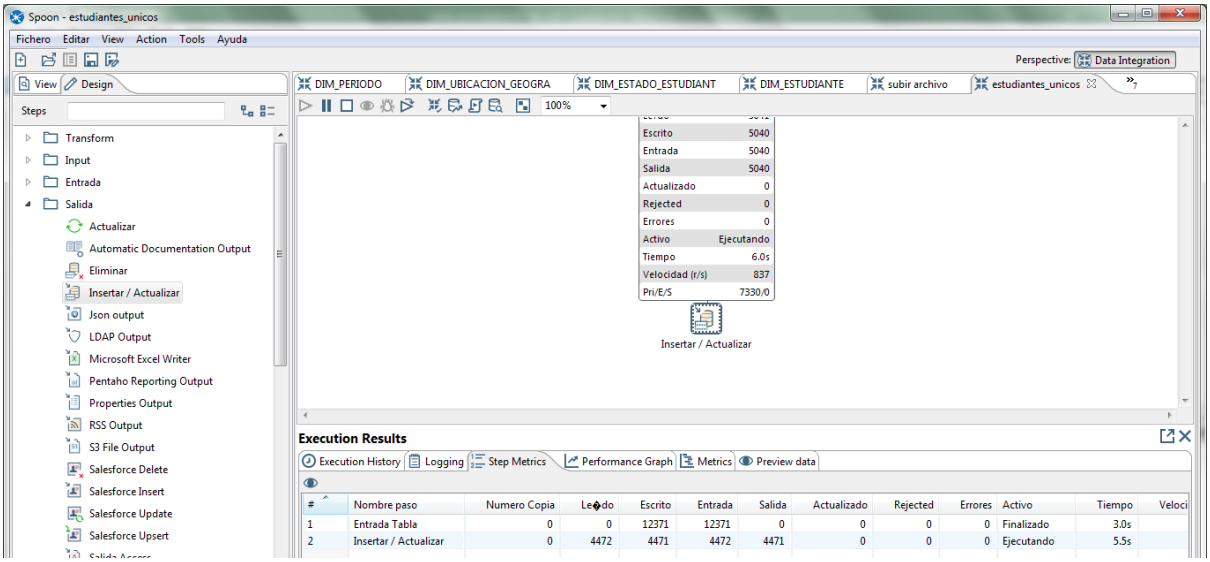
En el Apéndice B: ETL (Extracción, Transformación y Carga), se puede observar de mejor forma el proceso de elaboración de ETL. A continuación, se muestran algunas de las tareas realizadas mediante la herramienta Pentaho *Data Integration* - pdi-ce-5.4.0.1-130. Spoon.

Figura 18: ETL dim_ubicacion_academica



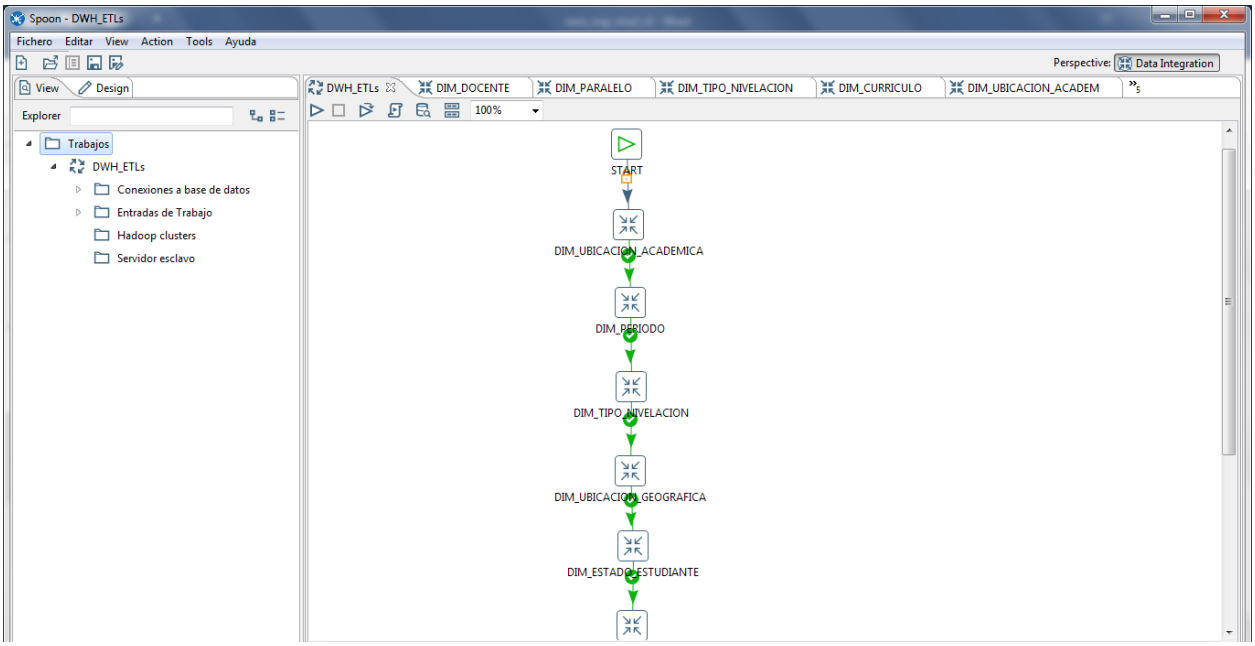
Fuente: Pentaho Data Integration - pdi-ce-5.4.0.1-130. Spoon

Figura 19: ETL dim_estudiante



Fuente: Pentaho Data Integration - pdi-ce-5.4.0.1-130. Spoon

Figura 20: ETL de carga del DW



Fuente: Pentaho Data Integration - pdi-ce-5.4.0.1-130. Spoon

Resultado 1: Mediante este procedimiento se cumplió con el Producto 1 de los pasos del Proceso de Descubrimiento de Conocimiento de Base de datos: Creación del DW y carga de datos, finalizando con el subproceso *Data Warehouse*. Obteniendo un DW de tamaño personal, es decir de menos de 1GB, con toda la información académica de la Unidad de Nivelación y Admisión de la

ESPOCH, contenida en 9 tablas de Dimensión y 3 tablas de hecho. Permitiendo continuar con los pasos de *Data Mining* para la obtención de patrones y toma de decisiones.

5.1.5. Elección de la tarea de minería de datos apropiada

Dentro del primer paso del subproceso *Data Mining* y quinto paso del proceso KDD propuesto, es importante tomar la decisión de escoger el objetivo y tipo de minería de datos, que para este estudio se lo realizó mediante la predicción, que tiene técnicas supervisadas (técnicas probadas y validadas).

Se realizaron varios análisis de técnicas, por lo que los algoritmos escogidos son los de clasificación, debido a que estos pueden explicar el comportamiento de una variable con relación a otras y se pueden obtener fácilmente reglas de negocio útiles para la toma de decisiones; y también algoritmos de regresión, para poder contar con datos predichos.

Resultado 1: Se logró escoger Data Mining de Predicción con algoritmos de clasificación y regresión.

5.1.6. Elección del algoritmo de minería de datos

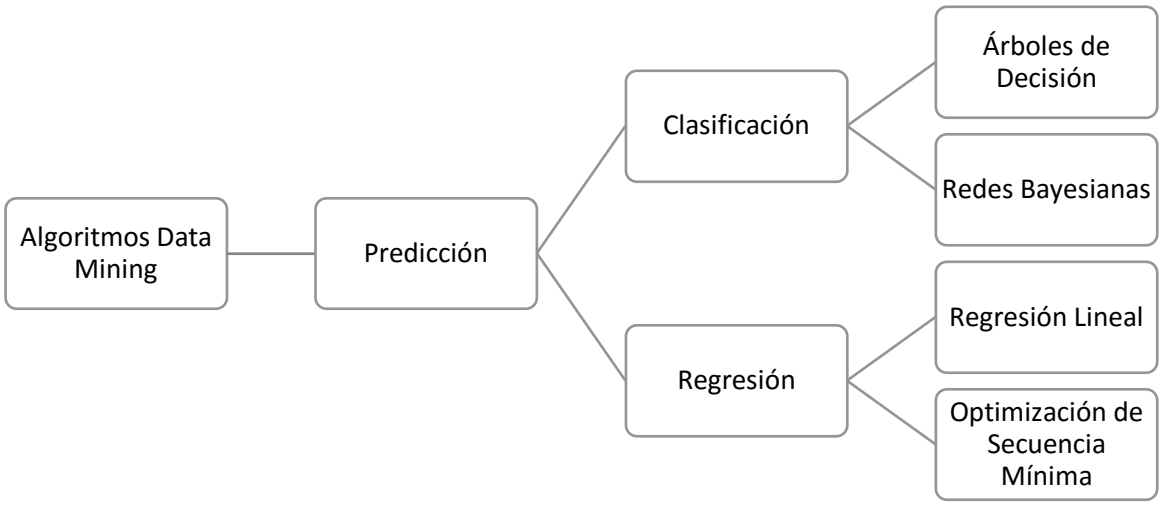
Para este segundo paso importante dentro de *Data Mining*, se analizaron los algoritmos de clasificación que existen y que se pueden aplicar. Para ello se hizo un análisis, considerando el equilibrio entre la precisión y la comprensibilidad al tratarse de algoritmos estadísticos que no son fáciles de interpretar.

En la Unidad de Nivelación y Admisión (UNAE) es importante conocer la realidad de las postulaciones de los estudiantes que obtienen un cupo para la ESPOCH y sus reglas del negocio para tomar decisiones, por tal motivo se consideró que los algoritmos a aplicarse deben mostrar esta información significante.

Los algoritmos que muestran los sucesos presentados en la información y pueden ser comprendidos son: Arboles de Decisión y Redes Bayesianas. En consecuencia, se consideraron estos dos algoritmos debido a que pueden establecer relaciones causales entre las variables y describir las reglas del negocio para identificar las estrategias a tomar dentro del proceso de selección de carreras de la ESPOCH.

Además, para la ESPOCH, también es importante conocer la predicción de la cantidad de estudiantes que serán parte de la institución en los próximos años, es por eso que el segundo análisis de técnicas *Data Mining*, dentro de los algoritmos de regresión son: Regresión Lineal y Optimización de Secuencia Mínima.

Figura 21: Elección del algoritmo de minería de datos para la Unidad de Nivelación y Admisión ESPOCH



Fuente: María Isabel Uvidia Fassler

Resultado 1: Se seleccionaron a los algoritmos: Arboles de Decisión y Redes Bayesianas como algoritmos de *Data Mining* de clasificación.

Resultado 2: Se seleccionaron a los algoritmos: Regresión Lineal y Optimización de Secuencia Mínima como algoritmos de *Data Mining* de regresión.

5.1.7. Empleando el algoritmo de minería de datos

En este paso se realiza la aplicación de los algoritmos de *Data Mining*, que generan los patrones que son el segundo producto del proceso KDD propuesto. La aplicación de *Data Mining* se lo realizó en la herramienta WEKA Developer 3.7.13. Para este análisis en el que se prueban dos algoritmos de minería de datos: clasificación y regresión, para mejor entendimiento se los clasificó en dos escenarios.

ESCENARIO A

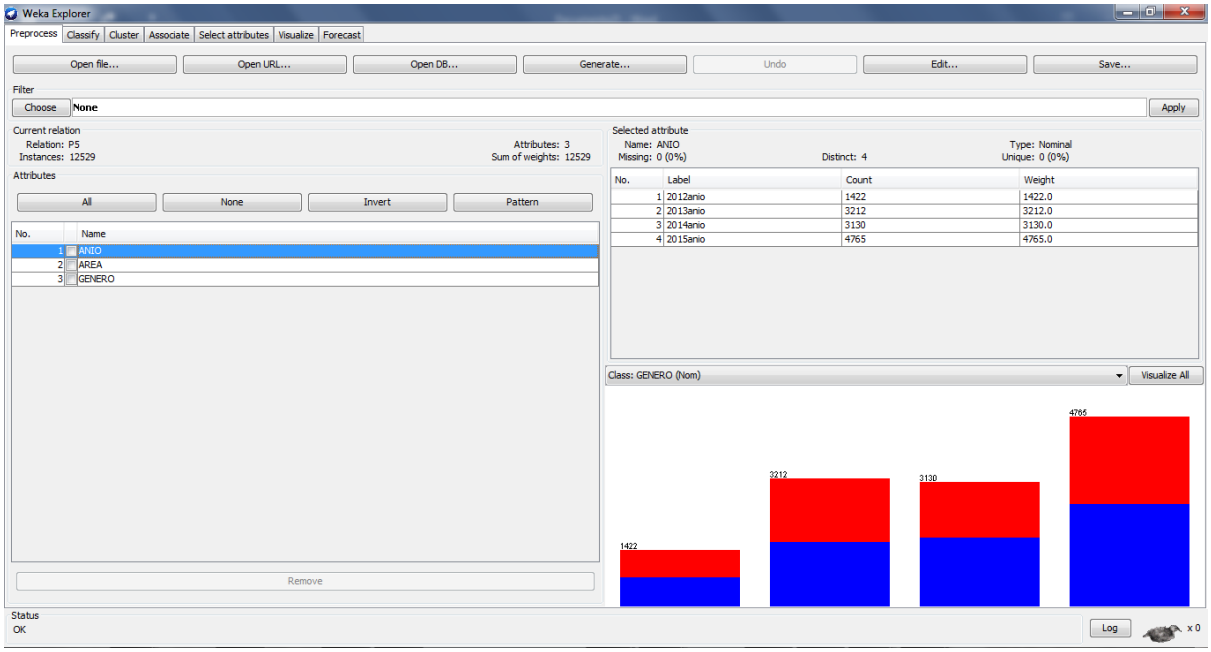
Algoritmo: CLASIFICACIÓN

Algoritmos de Clasificación: Árboles de Decisión y Redes Bayesianas

Datos analizados: Estudiantes por año, área de estudio de las carreras y género

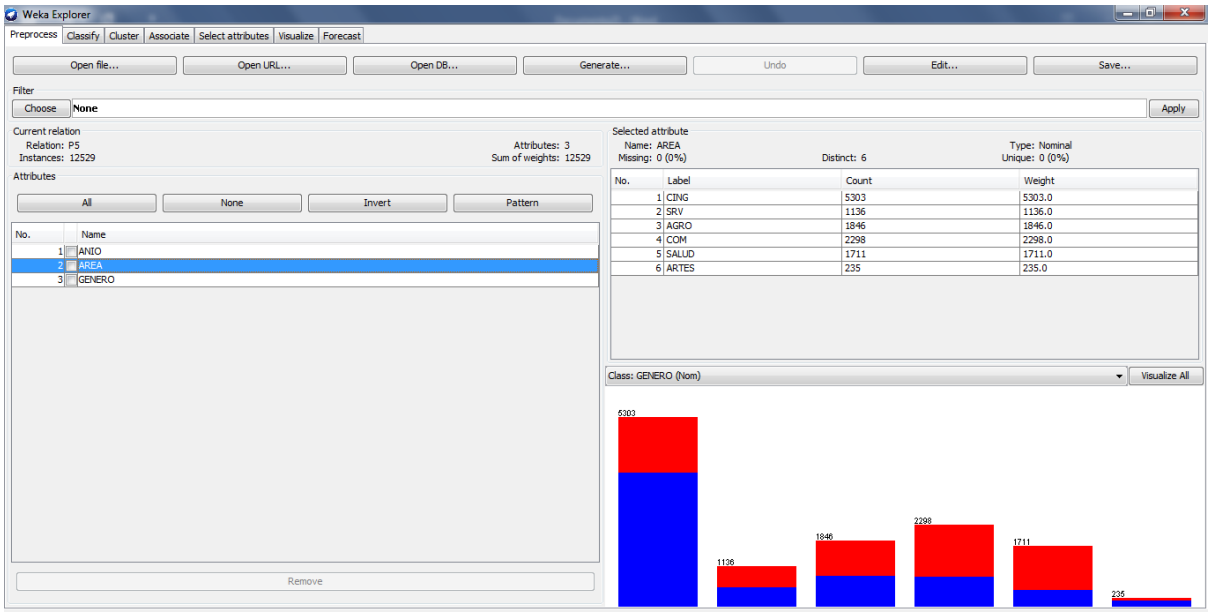
Atributos: Anio, Area, Genero

Figura 22: Atributo Anio



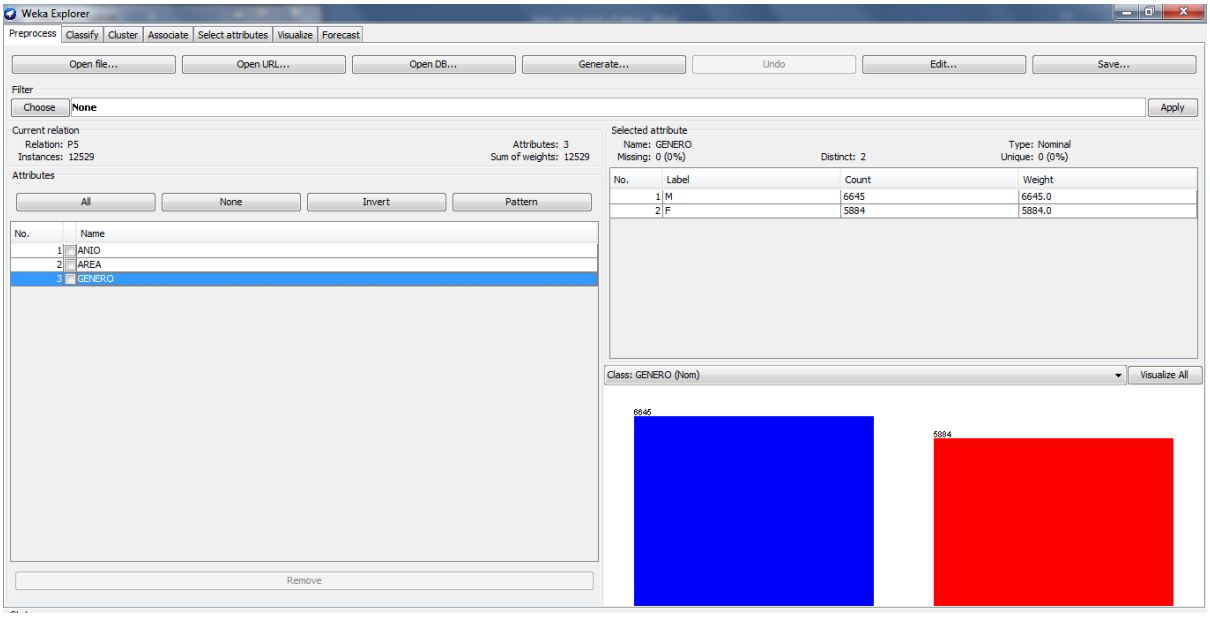
Fuente: Weka Developer 3.7.13

Figura 23: Atributo Área



Fuente: Weka Developer 3.7.13

Figura 24: Atributo Género

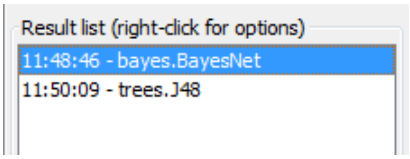


Fuente: Weka Developer 3.7.13

APLICACIÓN DE LOS ALGORITMOS DE DATA MINING:

WEKA permite la aplicación de estos algoritmos de forma intuitiva, por tal motivo después de indicar los datos que se consideran para el análisis, es necesario únicamente escoger el algoritmo, que para este estudio fueron: Árboles de Decisión y Redes Bayesianas.

Figura 25: Algoritmos de Clasificación



Fuente: Weka Developer 3.7.13

Resultado 1: Se obtuvieron los patrones generados al aplicar el algoritmo de Árboles de Decisión

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: P5
Instances: 12529
Attributes: 3
ANIO
AREA
GENERO
Test mode: 10-fold cross-validation

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

J48 pruned tree

AREA = CING: M (5303.0/1549.0)
AREA = SRV
| ANIO = 2012anio: F (198.0/98.0)
| ANIO = 2013anio: F (331.0/150.0)
| ANIO = 2014anio: M (323.0/155.0)
| ANIO = 2015anio: F (284.0/128.0)
AREA = AGRO: F (1846.0/868.0)
AREA = COM: F (2298.0/839.0)
AREA = SALUD: F (1711.0/469.0)
AREA = ARTES: M (235.0/64.0)

Number of Leaves : 9

Size of the tree : 11

Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	8160	65.1289 %
Incorrectly Classified Instances	4369	34.8711 %
Kappa statistic	0.3058	
Mean absolute error	0.4408	
Root mean squared error	0.4697	
Relative absolute error	88.4776 %	
Root relative squared error	94.1072 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	100 %	
Total Number of Instances	12529	

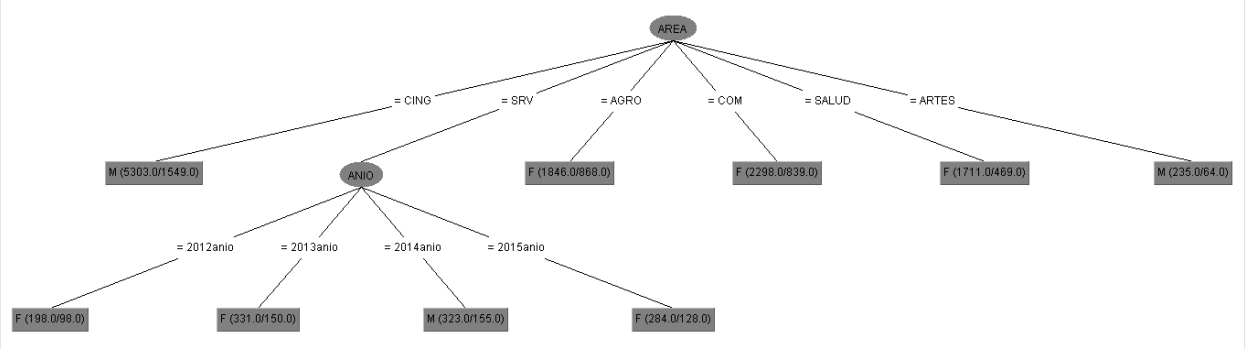
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.606	0.298	0.697	0.606	0.648	0.309	0.680	0.665	M
	0.702	0.394	0.612	0.702	0.654	0.309	0.680	0.629	F
Weighted Avg.	0.651	0.343	0.657	0.651	0.651	0.309	0.680	0.648	

=== Confusion Matrix ===

a	b	<-- classified as
4028	2617	a = M
1752	4132	b = F

Figura 26: Resultado del algoritmo Árbol de Decisión



Fuente: Weka Developer 3.7.13

Resultado 2: Se obtuvieron los patrones generados al aplicar el algoritmo de Redes Bayesianas

=== Run information ===

Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
Relation: P5
Instances: 12529
Attributes: 3
ANIO
AREA
GENERO
Test mode: 10-fold cross-validation

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

Bayes Network Classifier
not using ADTree
#attributes=3 #classindex=2
Network structure (nodes followed by parents)
ANIO(4): GENERO
AREA(6): GENERO
GENERO(2):
LogScore Bayes: -43451.04175206494
LogScore BDeu: -43484.15420239049
LogScore MDL: -43480.792713834584
LogScore ENTROPY: -43400.5884033265
LogScore AIC: -43417.5884033265

Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 8194 65.4003 %
Incorrectly Classified Instances 4335 34.5997 %
Kappa statistic 0.3118

Mean absolute error 0.4402
Root mean squared error 0.4695
Relative absolute error 88.3741 %
Root relative squared error 94.071 %
Coverage of cases (0.95 level) 100 %
Mean rel. region size (0.95 level) 100 %
Total Number of Instances 12529

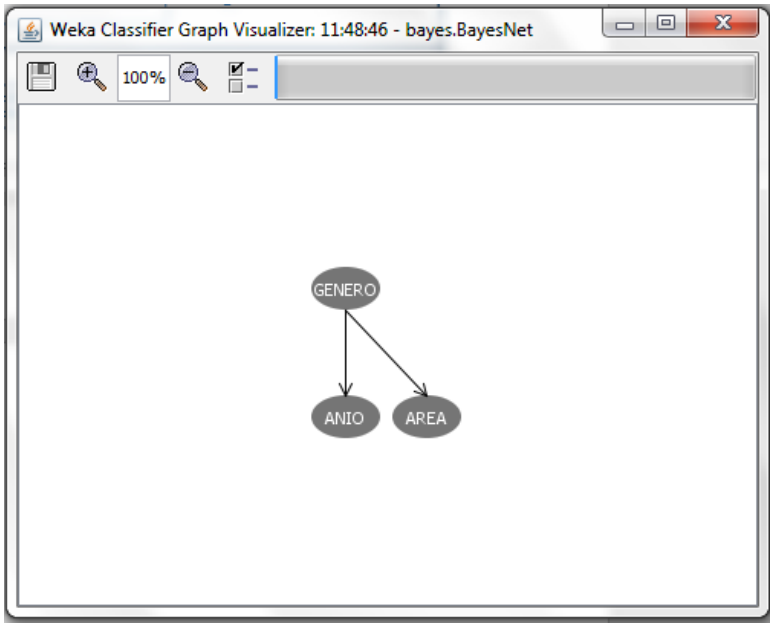
=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.602	0.288	0.703	0.602	0.649	0.315	0.685	0.672	M
	0.712	0.398	0.613	0.712	0.659	0.315	0.685	0.641	F
Weighted Avg.	0.654	0.339	0.661	0.654	0.654	0.315	0.685	0.657	

=== Confusion Matrix ===

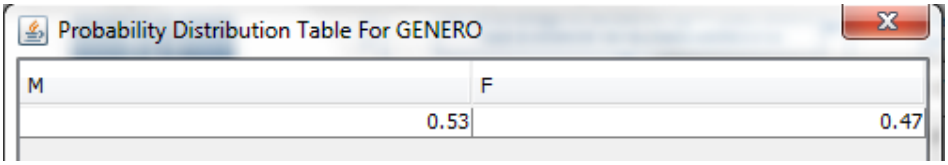
a	b	<-- classified as
4002	2643	a = M
1692	4192	b = F

Figura 27: Resultado del algoritmo Redes Bayesianas



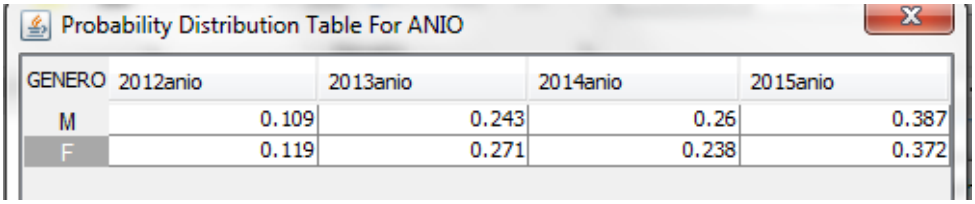
Fuente: Weka Developer 3.7.13

Figura 28: Resultado del algoritmo Redes Bayesianas – Nodo Género



Fuente: Weka Developer 3.7.13

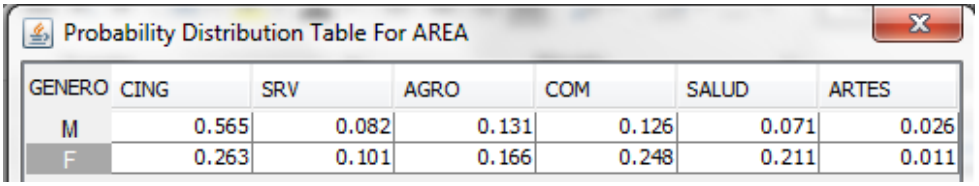
Figura 29: Resultado del algoritmo Redes Bayesianas – Nodo Anio



GENERO	2012anio	2013anio	2014anio	2015anio
M	0.109	0.243	0.26	0.387
F	0.119	0.271	0.238	0.372

Fuente: Weka Developer 3.7.13

Figura 30: Resultado del algoritmo Redes Bayesianas – Nodo Anio



GENERO	CING	SRV	AGRO	COM	SALUD	ARTES
M	0.565	0.082	0.131	0.126	0.071	0.026
F	0.263	0.101	0.166	0.248	0.211	0.011

Fuente: Weka Developer 3.7.13

ESCENARIO B

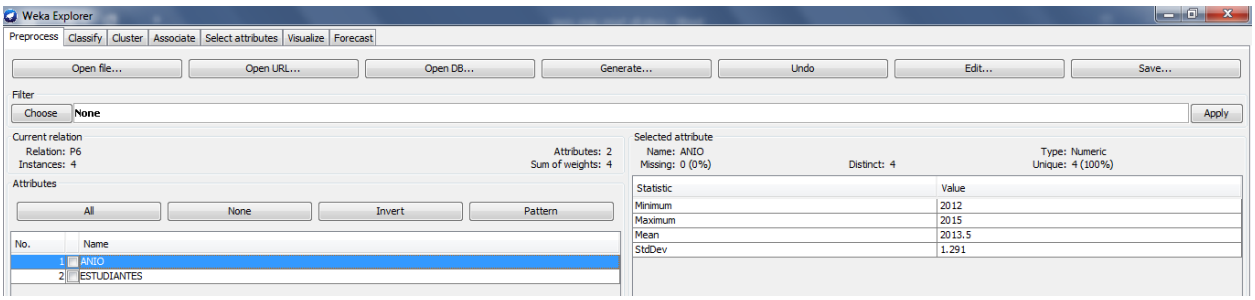
Algoritmo: REGRESIÓN

Algoritmos de Regresión: Regresión Lineal y Optimización de Secuencia Mínima

Datos analizados: Estudiantes por año

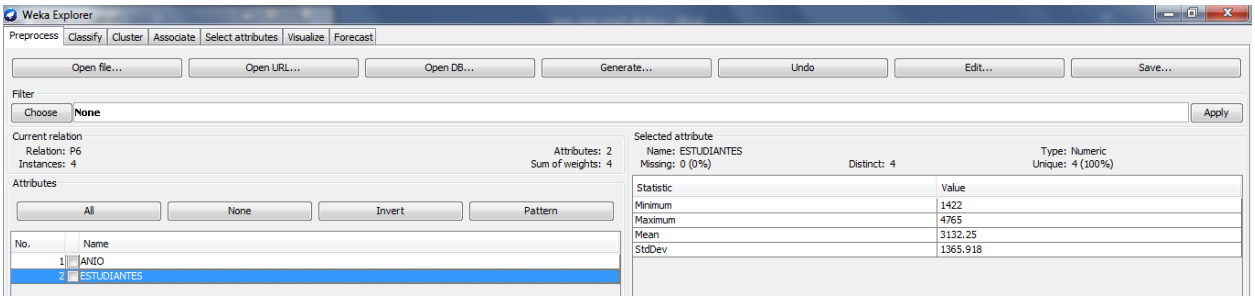
Atributos: Anio, Estudiantes

Figura 31: Atributo Anio



Fuente: Weka Developer 3.7.13

Figura 32: Atributo Estudiantes

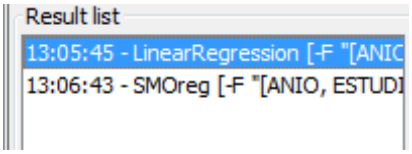


Fuente: Weka Developer 3.7.13

APLICACIÓN DE LOS ALGORITMOS DE DATA MINING:

En la opción Forecast de WEKA es necesario escoger los algorimos a aplicar y el tiempo de proyección que se desea conocer.

Figura 33: Algoritmos de Regresión



Fuente: Weka Developer 3.7.13

Resultado 3: Se obtuvieron los patrones generados al aplicar el algoritmo Regresión Lineal.

=== Run information ===

Scheme:
LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

Lagged and derived variable options:
-F "[ANIO, ESTUDIANTES]" -L 1 -M 2 -G ANIO

Relation: P6
Instances: 4
Attributes: 2
ANIO
ESTUDIANTES

Transformed training data:

ANIO
ESTUDIANTES
Lag_ANIO-1
Lag_ANIO-2

Lag_ESTUDIANTES-1
Lag_ESTUDIANTES-2
ANIO^2
ANIO^3
ANIO*Lag_ANIO-1
ANIO*Lag_ANIO-2
ANIO*Lag_ESTUDIANTES-1
ANIO*Lag_ESTUDIANTES-2

ANIO:

Linear Regression Model

ANIO =

0.8623 * Lag_ANIO-1 +
0.0002 * ANIO^2 +
0 * ANIO^3 +
-0.0002 * ANIO*Lag_ANIO-1 +
189.0071

ESTUDIANTES:

Linear Regression Model

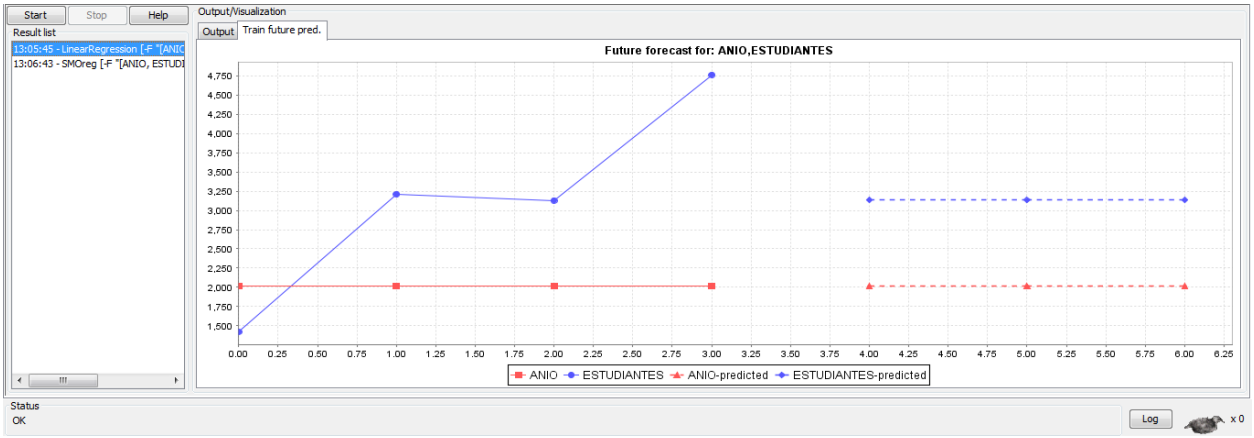
ESTUDIANTES =

+
3132.25

=== Future predictions from end of training data ===

inst#	ANIO	ESTUDIANTES
2012	2012	1422
2013	2013	3212
2014	2014	3130
2015	2015	4765
2016*	2016.0002	3132.25
2017*	2017.0006	3132.25
2018*	2018.0011	3132.25

Figura 34: Resultado del algoritmo Regresión Lineal



Fuente: Weka Developer 3.7.13

Resultado 4: Se obtuvieron los patrones generados al aplicar el algoritmo de Optimización de Secuencia Mínima.

```
=== Run information ===

Scheme:
        SMOreg -C 1.0 -N 0 -I "RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001 -W 1" -K
"PolyKernel -E 1.0 -C 250007"

Lagged and derived variable options:
        -F "[ANIO, ESTUDIANTES]" -L 1 -M 2 -G ANIO

Relation:     P6
Instances:    4
Attributes:   2
        ANIO
        ESTUDIANTES

Transformed training data:

        ANIO
        ESTUDIANTES
        Lag_ANIO-1
        Lag_ANIO-2
        Lag_ESTUDIANTES-1
        Lag_ESTUDIANTES-2
        ANIO^2
        ANIO^3
        ANIO*Lag_ANIO-1
        ANIO*Lag_ANIO-2
        ANIO*Lag_ESTUDIANTES-1
        ANIO*Lag_ESTUDIANTES-2

ANIO:
SMOreg

weights (not support vectors):
 +      0.0005 * (normalized) Lag_ANIO-1
 +      0.0003 * (normalized) Lag_ANIO-2
 +      0.0004 * (normalized) Lag_ESTUDIANTES-1
 +      0.0003 * (normalized) Lag_ESTUDIANTES-2
 +      0.4989 * (normalized) ANIO^2
 +      0.4988 * (normalized) ANIO^3
 +      0.0005 * (normalized) ANIO*Lag_ANIO-1
 +      0.0003 * (normalized) ANIO*Lag_ANIO-2
 +      0.0004 * (normalized) ANIO*Lag_ESTUDIANTES-1
 +      0.0003 * (normalized) ANIO*Lag_ESTUDIANTES-2
 -      0.001

Number of kernel evaluations: 10 (93.151% cached)

ESTUDIANTES:
SMOreg
```

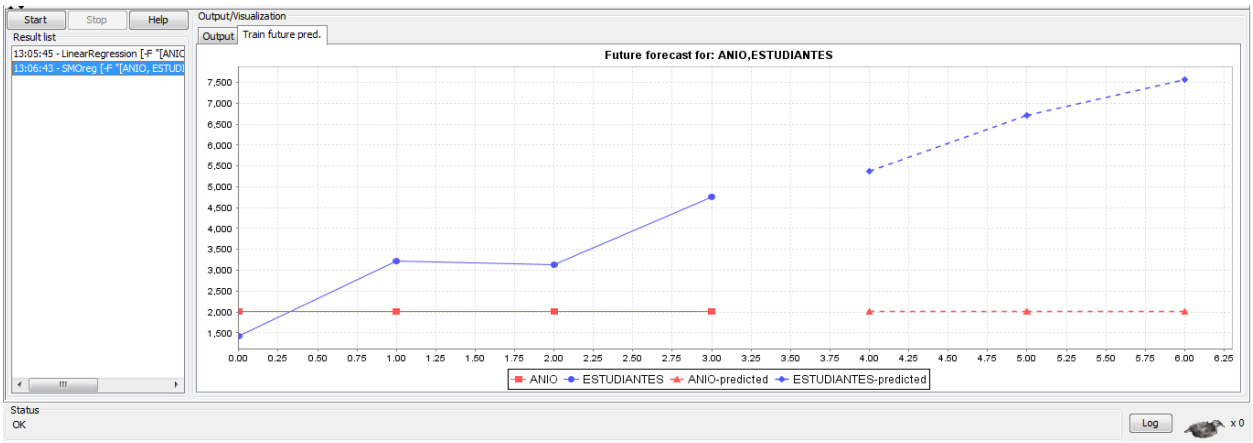
weights (not support vectors):
- 0.0554 * (normalized) Lag_ANIO-1
+ 0.0485 * (normalized) Lag_ANIO-2
- 0.106 * (normalized) Lag_ESTUDIANTES-1
+ 0.0485 * (normalized) Lag_ESTUDIANTES-2
+ 0.5107 * (normalized) ANIO^2
+ 0.5106 * (normalized) ANIO^3
- 0.0554 * (normalized) ANIO*Lag_ANIO-1
+ 0.0485 * (normalized) ANIO*Lag_ANIO-2
- 0.106 * (normalized) ANIO*Lag_ESTUDIANTES-1
+ 0.0485 * (normalized) ANIO*Lag_ESTUDIANTES-2
+ 0.0971

Number of kernel evaluations: 10 (87.952% cached)

=== Future predictions from end of training data ===

inst#	ANIO	ESTUDIANTES
2012	2012	1422
2013	2013	3212
2014	2014	3130
2015	2015	4765
2016*	2016.0029	5379.4973
2017*	2017.0081	6712.3302
2018*	2018.0142	7576.2995

Figura 35: Resultado del algoritmo Optimización de Secuencia Mínima



Fuente: Weka Developer 3.7.13

5.1.8.Evaluación

En esta octava fase del proceso KDD, se interpretan y analizan los patrones obtenidos en cada escenario, para determinar cuál es el algoritmo más apropiado dentro del tipo de *Data Mining*. A continuación, se muestran los resultados:

ESCENARIO A

ALGORITMO: Clasificación

Algoritmos de Clasificación: Árboles de Decisión y Redes Bayesianas

Para este análisis de los patrones obtenidos como resultado de cada uno de los algoritmos, se van a considerar indicadores importantes como:

- Instancias correctamente clasificadas
- Kappa: Indice que compara la coincidencia entre varios expertos con el nivel de coincidencia que se pondría por casualidad. Resultados:
  - +1: total coincidencia
  - Valores de 0: No más coincidencia que la que puede esperarse por casualidad
  - -1: total desacuerdo
- Error absoluto: Diferencia entre los valores obtenidos frente al valor exacto.
- Matriz de Confusión

Análisis:

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Correctly Classified Instances 8160 65.1289 %
Incorrectly Classified Instances 4369 34.8711 %
Kappa statistic 0.3058
Mean absolute error 0.4408
Root mean squared error 0.4697
Relative absolute error 88.4776 %
Root relative squared error 94.1072 %
Coverage of cases (0.95 level) 100 %
Mean rel. region size (0.95 level) 100 %
Total Number of Instances 12529

=== Confusion Matrix ===

a	b	<-- classified as
4028	2617	a = M
1752	4132	b = F

Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -

Correctly Classified Instances 8194 65.4003 %
Incorrectly Classified Instances 4335 34.5997 %
Kappa statistic 0.3118
Mean absolute error 0.4402
Root mean squared error 0.4695
Relative absolute error 88.3741 %
Root relative squared error 94.071 %
Coverage of cases (0.95 level) 100 %
Mean rel. region size (0.95 level) 100 %

=== Confusion Matrix ===

a	b	<-- classified as
4002	2643	a = M
1692	4192	b = F

Al realizar una comparación entre algoritmos empleados se deduce que:

- La mayor cantidad de instancias correctamente clasificadas del total de 12529 registros, lo realiza el algoritmo de Redes Bayesianas con un total de 8194, que representa el 65,4003%.
- El indice Kappa del algoritmo Redes Bayesianas se acerca más a 1, es decir, con 0,3118 es el que más coincidencia entre variables presenta, siendo este un punto aceptable.
- El error absoluto es de 0.4402 en las Redes Bayesianas que mientras menor sea es mejor ya que existe menos diferencia con el promedio exacto.
- Las diagonales de la Matriz de Confusión de las Redes Beyesianas y Árboles de Decisión, tienen casi el mismo equilibrio, por lo que en ambos casos son valores aceptables.

Resultado 1: El algoritmo de Clasificación – Redes Bayesianas, es el que mejor se acopla al manejo de información de la Unidad de Nivelación y Admisión de la ESPOCH, aunque Árboles de Decisión tiene valores muy cercanos al otro algoritmo.

ESCENARIO B

ALGORITMO: Regresión

Algoritmos de Regresión: Regresión Lineal y Optimización de Secuencia Mínima

Análisis:

Scheme:
LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

inst#	ANIO	ESTUDIANTES
2012	2012	1422
2013	2013	3212
2014	2014	3130
2015	2015	4765
2016*	2016.0002	3132.25
2017*	2017.0006	3132.25
2018*	2018.0011	3132.25

Scheme:
SMOreg -C 1.0 -N 0 -I "RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001 -W 1" -K "PolyKernel -E 1.0 -C 250007"

Number of kernel evaluations: 10 (87.952% cached)

inst#	ANIO	ESTUDIANTES
2012	2012	1422
2013	2013	3212
2014	2014	3130
2015	2015	4765
2016*	2016.0029	5379.4973
2017*	2017.0081	6712.3302
2018*	2018.0142	7576.2995

Al realizar el análisis entre los dos algoritmos de regresión, es evidente notar que la optimización de secuencia mínima presenta una proyección de datos, pero que estos no son tan aceptables debido a que presentan un crecimiento grande en la cantidad de datos, pero para tomar esta decisión también se deben considerar otras variables externas. Por lo que la predicción puede ser medianamente aceptable. Mientras que los resultados de Regresión Lineal mantienen una constante, que también podría ser considerada como una predicción válida, pero con menos probabilidad de que ocurra.

Resultado 2: Al ser algoritmos utilizados para la predicción, y mostrar un 87,952% de confiabilidad, el algoritmo de Optimización de Secuencia Mínima, es el algoritmo que se puede tomar como referencia para el crecimiento de la cantidad de estudiantes en la ESPOCH, para los siguientes 3 años.

5.1.9. Usando el conocimiento descubierto

En la última fase del Descubrimiento de Conocimiento en Base de Datos, el tercer producto es el conocimiento, por tal motivo es preciso en base a los algoritmos utilizados presentar el conocimiento descubierto en los datos de la Unidad de Nivelación y Admisión de la ESPOCH.

Conocimientos:

- La mayor cantidad de estudiantes de la ESPOCH en los años 2012, 2013, 2014 y 2015, están en el área de Ingeniería, donde de los 5303 estudiantes, 1549 fueron del género femenino.
- En el área de Servicios, en el año 2014 existió mayor cantidad de estudiantes, siendo 155 mujeres de los 323 estudiantes.
- El área Comercial en los últimos cuatro años ha generado el segundo lugar con mayor cantidad de estudiantes, siendo 2298, con tan sólo 839 hombres.
- En el área de Artes la mayoría de los estudiantes son hombres.
- El área con el tercer lugar de cantidad de estudiantes es Agricultura con un total de 1846, donde tan sólo hubo 868 hombres.
- La cantidad de estudiantes hombres en el universo global de estudiantes representa un 53%, siendo mayor, pero que muestra también que la diferencia entre estos dos géneros casi se está equiparando.
- El año en donde existió mayor cantidad de estudiantes mujeres fue el 2015 con un 37.2%

- El área con mayor cantidad de estudiantes hombres y mujeres es el área de Ingeniería, mientras que el segundo lugar con mayor cantidad de estudiantes mujeres es el área Comercial y para los estudiantes hombres el área de Agricultura.
- La ESPOCH se proyecta a tener mayor cantidad de estudiantes en los próximos 3 años.

5.1.10. Reportes Business Intelligence

Cuando se empezaron a analizar los requerimientos de información dentro de cada proceso en la primera fase de Descubrimiento de Conocimiento en Base de Datos, se definió también la necesidad de información que requiere la Unidad de Nivelación y Admisión, la cual necesita ser mostrada para realizar la adecuada la toma de decisiones. Por tal motivo, como complemento al proceso KDD, se crearon reportes de *Business Intelligence* ad-hoc, es decir, reportes que manejan filtros y que pueden tomar la apariencia que el tomador de decisiones requiera.

Esta implementación se la realizó mediante SpagoBI 4.0. que es una herramienta de código libre u opensource, que brinda las opciones para manipular la información que se requiere aparezca en los reportes.

En base a las necesidades de información en la Unidad de Nivelación y Admisión, se crearon los siguientes reportes:

- Reporte de Ubicación Geográfica
- Reporte de Postulaciones
- Reporte de Cantidad de estudiantes
- Reporte de Estudiantes Aprobados y Reprobados
- Reporte de la Evolución de las Carreras

Reporte de Ubicación Geográfica

Reporte creado para identificar los lugares de procedencia de los estudiantes que aceptaron un cupo para la ESPOCH. Este reporte permite mediante los filtros implementados escoger el año, periodo, país, provincia y estado del estudiante para realizar el análisis correspondiente.

Reporte de Postulaciones

Para conocer a los estudiantes y cuáles fueron las carreras más postuladas dentro de la ESPOCH, se creó este reporte, que permite también seleccionar el año, periodo, datos del estudiante como: estado, género y discapacidad, permitiendo conocer a fondo la situación de los

estudiantes de la institución y brindarles la opción de servicios apropiados en base a sus realidades.

Reporte de Cantidad de estudiantes

Este reporte permite identificar la cantidad de estudiantes del género femenino y masculino que se encuentran estudiando en las diferentes carreras de la ESPOCH, de acuerdo al año, periodo, área y al estado del estudiante.

Reporte de Estudiantes Aprobados y Reprobados

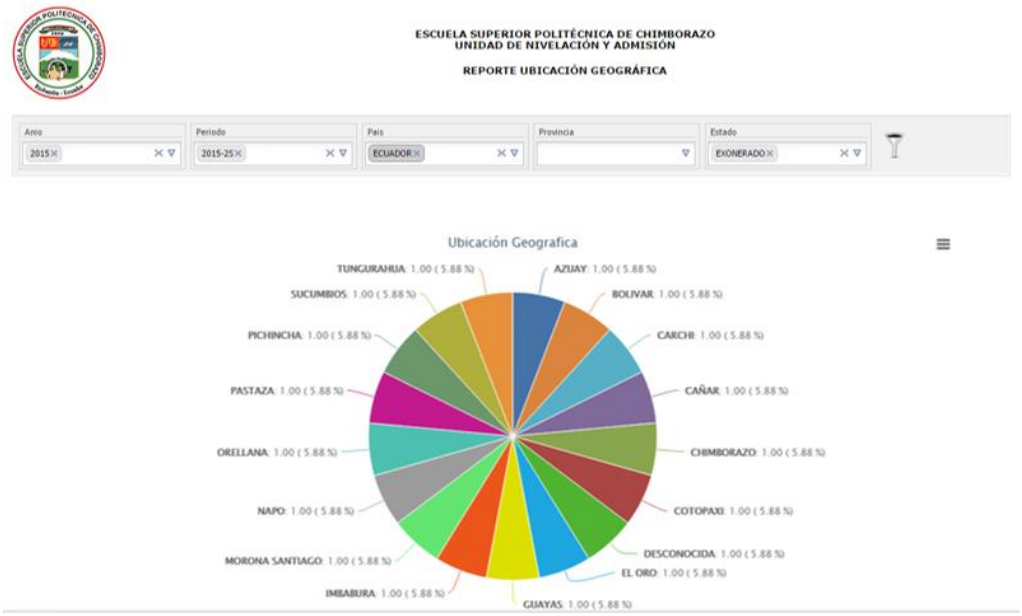
Es importante conocer la cantidad de estudiantes que aprueban o reprueban el Curso de Nivelación de Carrera, para cubrir esta necesidad se creó el reporte que, en base a años, periodos, áreas y carreras, evidencia la tendencia de los estudiantes. Pudiendo tomar estrategias para mejorar el rendimiento de los estudiantes politécnicos.

Reporte de la Evolución de las Carreras

Desde que el estudiante decide postular a una carrera es primordial analizar estas tendencias, comenzando en su estado de: admitidos, matriculados, aprobados o reprobados del Curso de Nivelación de Carrera. Por tal motivo este reporte muestra la evolución de los estudiantes en las diferentes carreras de acuerdo a años, periodos, áreas y géneros.

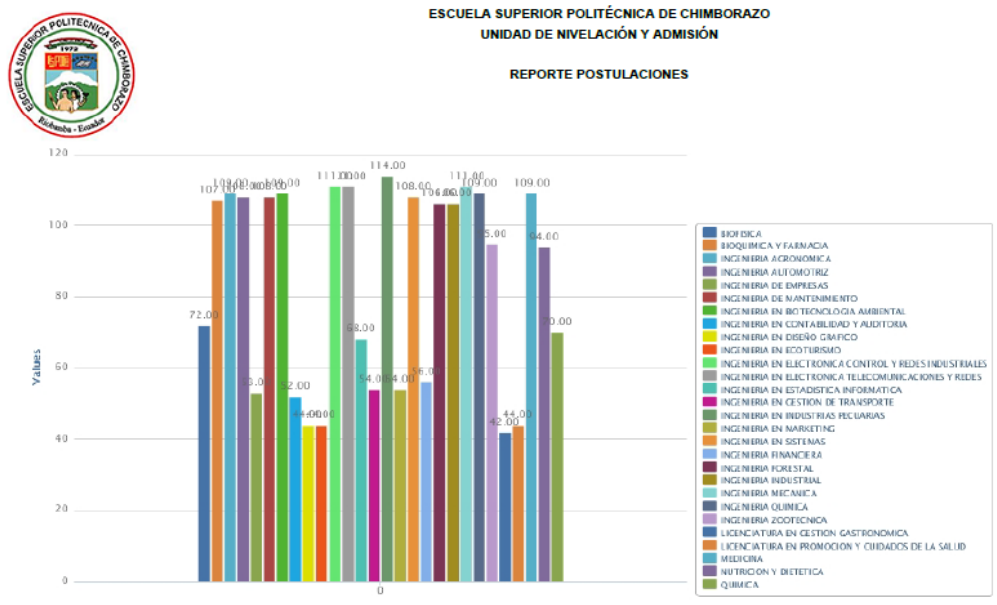
A continuación, se muestran algunos de los reportes creados:

Figura 36: Reporte de Ubicación Geográfica



Fuente: SpagoBI 4.0.

Figura 37: Reporte de Postulaciones



Fuente: SpagoBI 4.0.

En el Ápendice C: Reportes *Business Intelligence*, se puede observar de forma detallada la elaboración de los reportes ad-hoc.

5.2. Evaluación preliminar

Al aplicar el proceso de Descubrimiento de Conocimiento en Base de datos, es evidente como se solucionó la falta de información que la Unidad de Nivelación y Admisión tenía. Mediante entrevistas a los coordinadores General, Académico y Administrativo, que son el personal del nivel estratégico que toman decisiones y son usuarios directos de la información, posterior a la aplicación y explicación de los resultados obtenidos en *Data Mining* y presentación de los reportes ad-hoc de *Business Intelligence*, se obtuvieron los siguientes puntos considerables mencionados por el personal:

- El poseer información consistente de los procesos académicos de Nivelación permite tomar decisiones y plantear estrategias para que el rendimiento de los estudiantes mejore.
- El conocer información directa de los estudiantes, como los lugares de donde provienen, género y discapacidades, permite conocer de mejor forma la realidad institucional, creando programas de bienestar y vinculación que mejoren el transcurso del estudiante en la ESPOCH.
- Conocer la tendencia al momento de la selección de las carreras de la institución permiten potenciar las más escogidas y buscar estrategias para las menos postuladas.

Mediante estas entrevistas se pudo ratificar el objetivo de KDD, que es crear conocimiento para la toma de decisiones mediante información apropiada, correcta, consistente que se encuentre disponible.

Capítulo 6

Conclusiones y Recomendaciones

6.1. Conclusiones

- En el presente trabajo se analizó la teoría y las metodologías del proceso de Descubrimiento de Conocimiento en Base de Datos, además de la metodología HEFESTO versión 2.0. y los algoritmos de *Data Mining*, que permitieron plantear un proceso de KDD con HEFESTO versión 2.0. acoplado a las necesidades de la Unidad de Nivelación y Admisión ayudando a la adecuada toma de decisiones.
- El proyecto de investigación y desarrollo posee un diseño del proceso de Descubrimiento de Conocimiento en Base de datos adecuado a los requerimientos de la Unidad de Nivelación y Admisión de la ESPOCH, el cual plantea nueve fases, las cuatro primeras dentro del subproceso *Data Warehouse*, que acopló la metodología HEFESTO versión 2.0, que es una metodología propia para la creación de DW. Las otras cinco fases estuvieron dentro del subproceso *Data Mining*, con el que se analizó los algoritmos útiles para conseguir conocimiento. Las nueve fases permitieron conseguir tres productos: *Data Warehouse*, patrones de *Data Mining* y el Conocimiento, brindando a la UNAE la posibilidad de contar con información correcta, consistente y adecuada para la toma de decisiones.
- La aplicación del proceso de Descubrimiento de Conocimiento en Base de Datos, produjo como resultado de sus primeras cuatro fases, la creación de un *Data Warehouse* alojado en PostgreSQL 9.3, que consta de nueve dimensiones y tres tablas de hechos, que contienen la información consistente y adecuada para la toma de decisiones, ya que los datos pasaron por fases de limpieza y ETL; garantizando de esta forma la disponibilidad de información para la generación de reportes de análisis en tiempo real a la Unidad de Nivelación y Admisión sobre sus procesos académicos.
- Dentro de las cinco técnicas del subproceso *Data Mining* del proceso KDD, se aplicaron y analizaron algoritmos de minería de datos de clasificación y regresión; siendo el algoritmo de Redes Bayesianas y el de Optimización de Secuencia Mínima los más apropiados para los datos de la Unidad de Nivelación y Admisión de la ESPOCH, que permitieron generar conocimiento y conocer las tendencias de los estudiantes a la

hora de escoger carreras, información importante para la planificación de cupos de la oferta académica de la ESPOCH, siendo soporte para la toma de decisiones.

- Para brindar herramientas que muestren la información académica, se generaron cinco reportes de Business Intelligence ad-hoc, que permiten el análisis de los datos de la Unidad de Nivelación y Admisión de la ESPOCH, tanto de sus ubicaciones geográficas o lugares de procedencia, como sus postulaciones, tendencias de las carreras por áreas y años, cantidad de estudiantes por género, discapacidad y periodos académicos. Información útil y consistente que está disponible para el análisis que se requiera y soporte la adecuada toma de decisiones, que va acompañada de estrategias de mejora para garantizar la calidad en la educación superior.

6.2. Recomendaciones

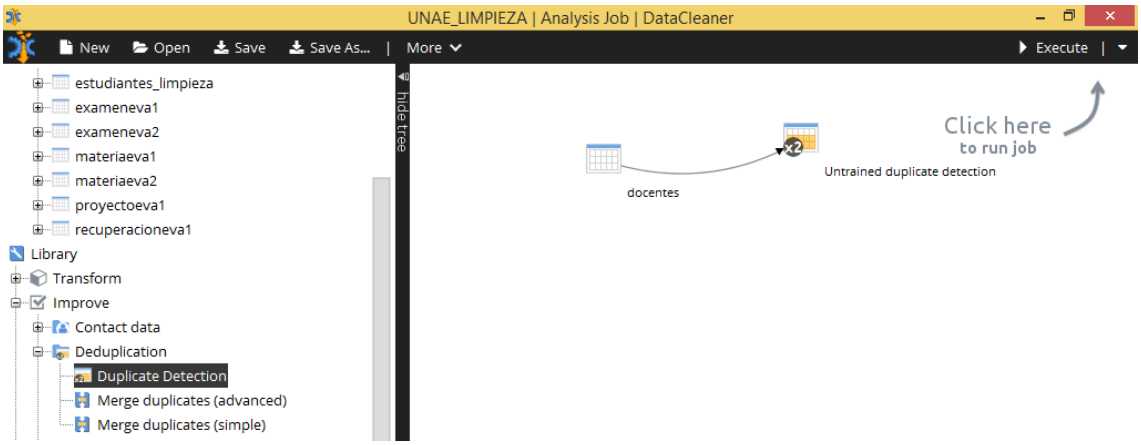
- Al momento de implementar el proceso de Descubrimiento de Conocimiento en Base de Datos, es primordial conocer el objetivo de la información, ya que de esta forma se define el alcance el *Data Warehouse*.
- Para la selección de algoritmos de *Data Mining* es importante conocer cómo funciona el algoritmo, además de cómo se debe ingresar la información para obtener patrones adecuados.
- Es importante aclarar que la calidad de información que estará en el *Data Warehouse* o que se mostrará en los reportes *Business Intelligence*, siempre va a depender de los datos fuentes con los que se dispongan y de esta consistencia de información.

Apéndice A

Tareas de Limpieza de Datos

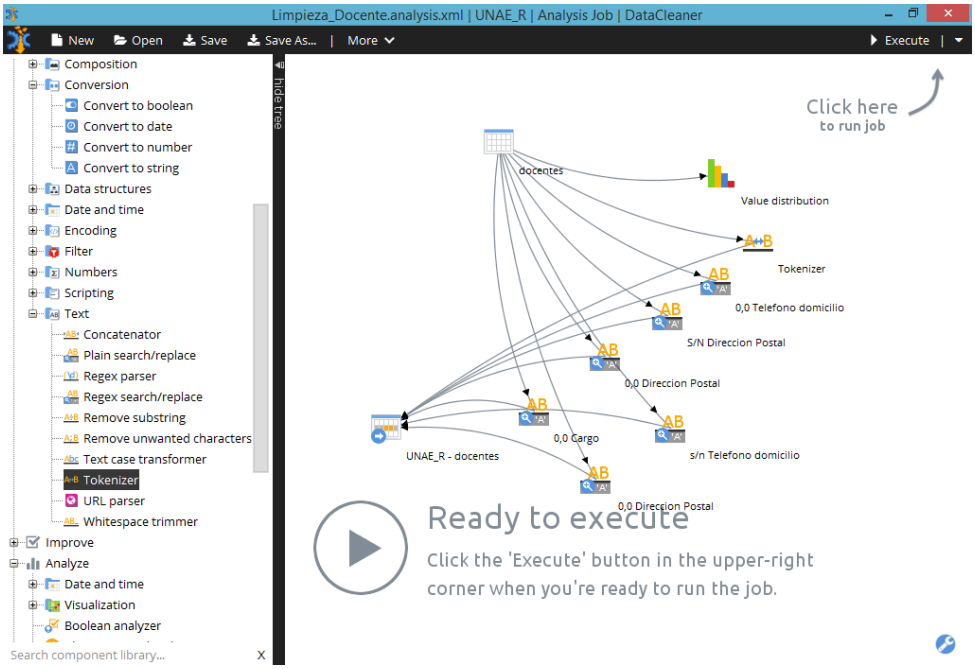
Para realizar las tareas de limpieza de datos en Pentaho es importante identificar el objetivo a cumplir. Esta herramienta ofrece un sin número de opciones, a continuación, se presentan algunas de las tareas realizadas.

Figura 38: Creación de Tareas para encontrar datos duplicados



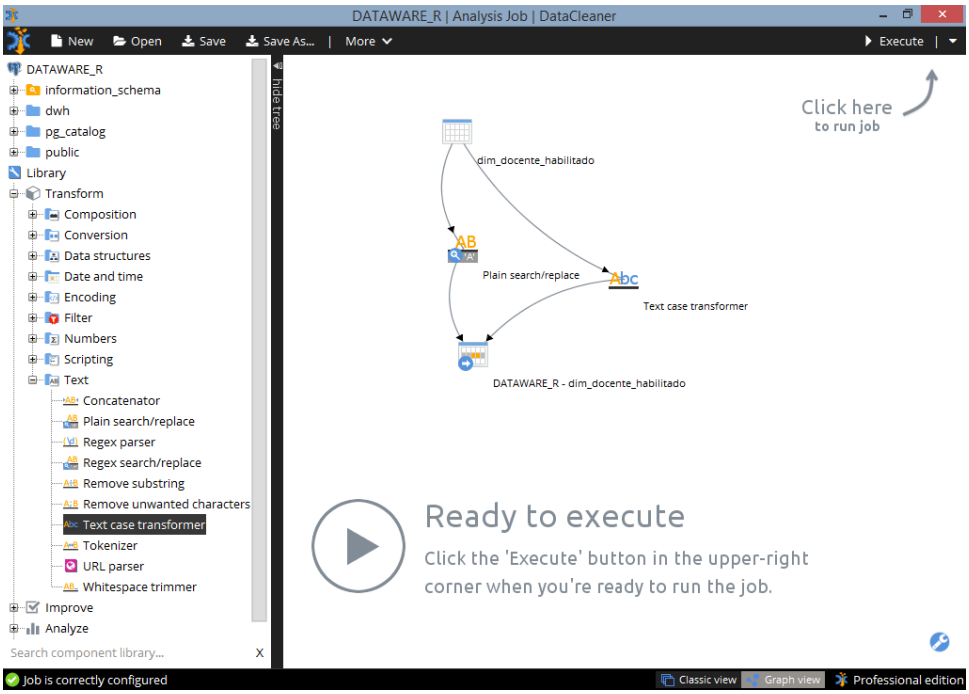
Fuente: *Data Cleaner* 4.5.3

Figura 39: Análisis de datos de estudiantes



Fuente: *Data Cleaner* 4.5.3

Figura 40: Transformación de datos



Fuente: *Data Cleaner* 4.5.3

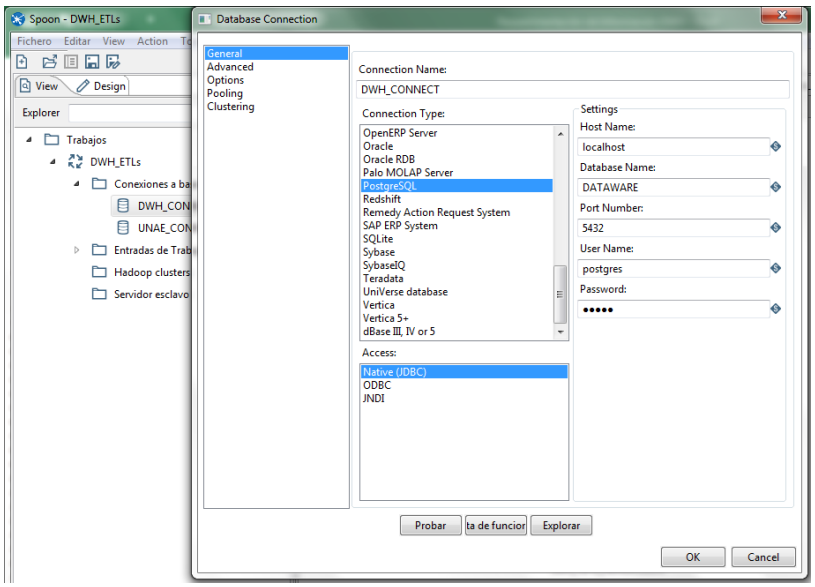
Apéndice B

ETL (Extracción, Transformación y Carga)

Para la elaboración de ETL en la herramienta Pentaho Data Integration - pdi-ce-5.4.0.1-130. Spoon, es importante considerar los siguientes pasos:

Configuración de la conexión con la base de datos.

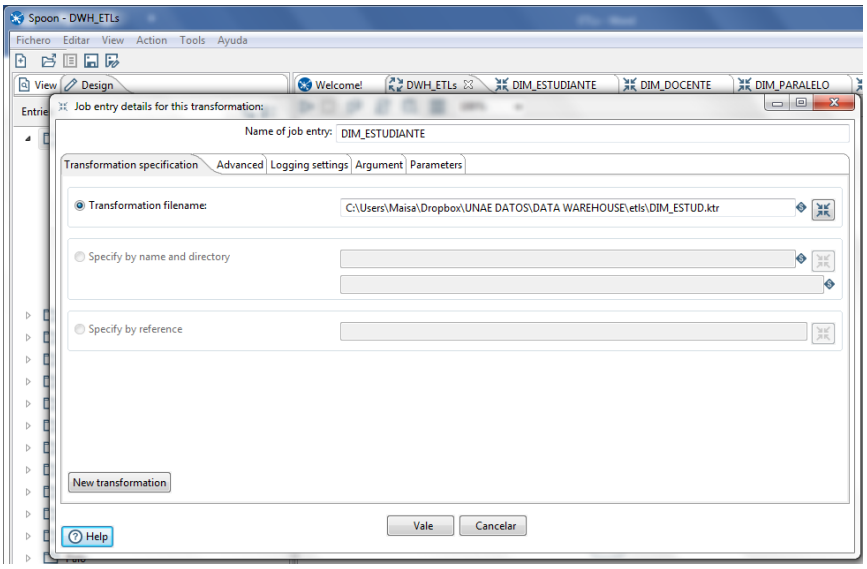
Figura 41: Conexión a base de datos



Fuente: Pentaho Data Integration - pdi-ce-5.4.0.1-130. Spoon

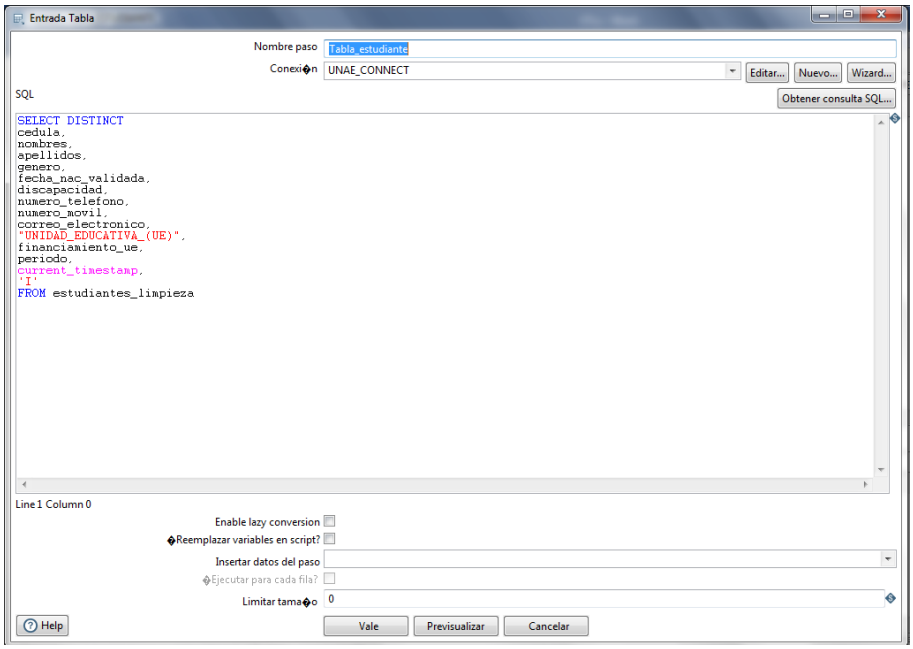
Creación de la transformación, en donde se inserta la sentencia sql con la que se extrae información de la fuente.

Figura 42: Creación de la transformación



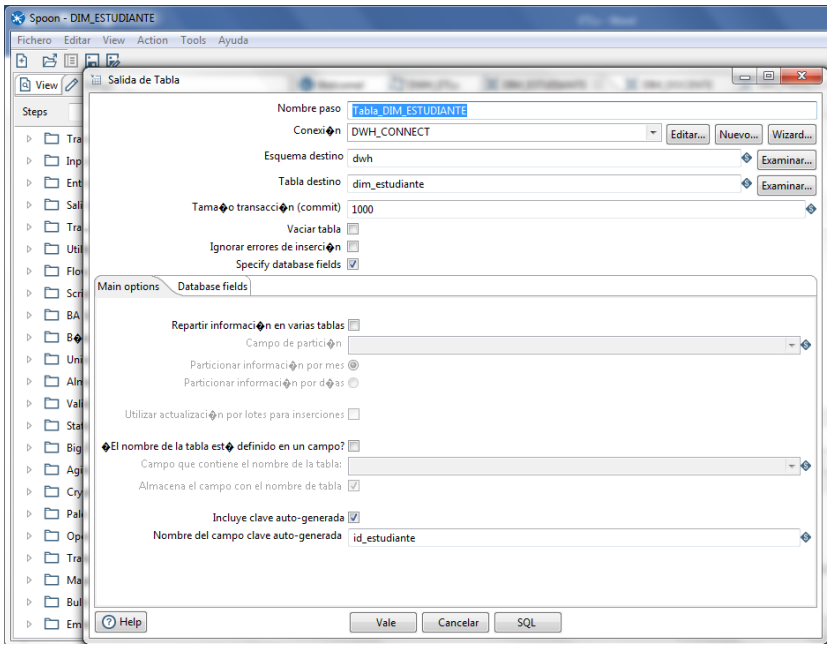
Fuente: Pentaho Data Integration - pdi-ce-5.4.0.1-130. Spoon

Figura 43: Sentencia SQL para cargar datos de la fuente



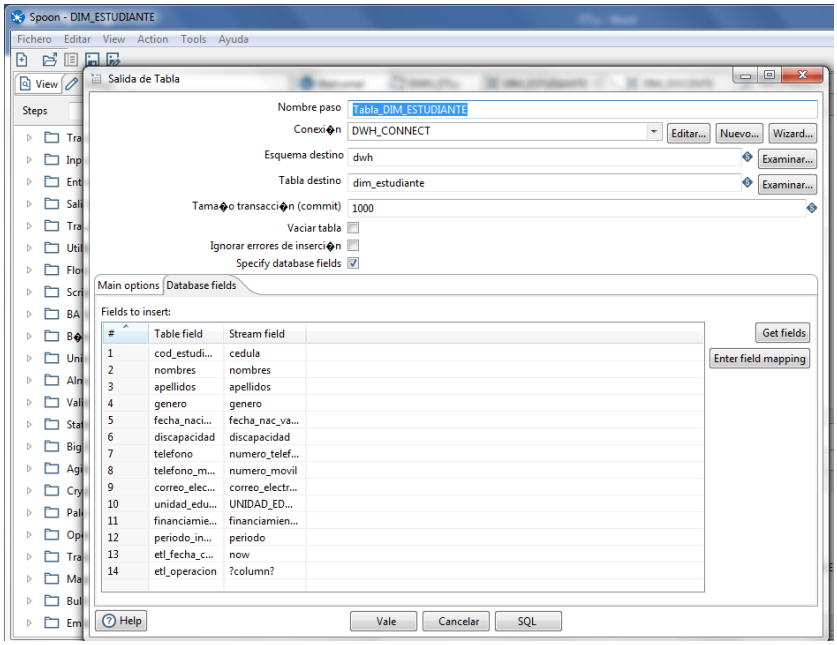
Fuente: Pentaho Data Integration - pdi-ce-5.4.0.1-130. Spoon

Figura 44: Sentencia SQL para cargar datos de la fuente



Fuente: Pentaho Data Integration - pdi-ce-5.4.0.1-130. Spoon

Figura 45: Configuraciones de la tabla destino



Fuente: Pentaho Data Integration - pdi-ce-5.4.0.1-130. Spoon

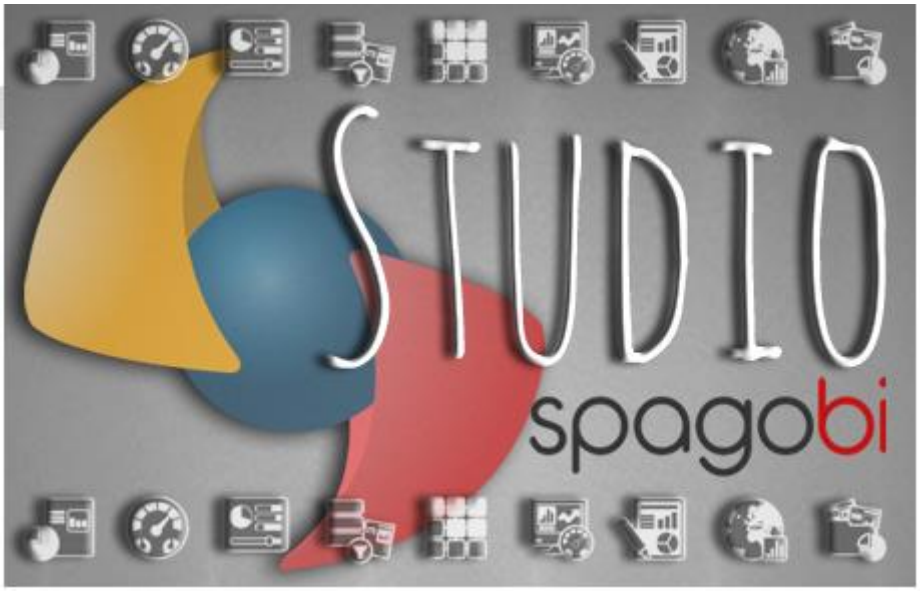
Una vez configurado la fuente y las tablas destino se debe ejecutar la tarea que carga toda la información.

Apéndice C

Reportes Business Intelligence

A continuación, se muestra el procedimiento que se utilizó para la creación de los reportes Ad-hoc en SpagoBI 4.0.

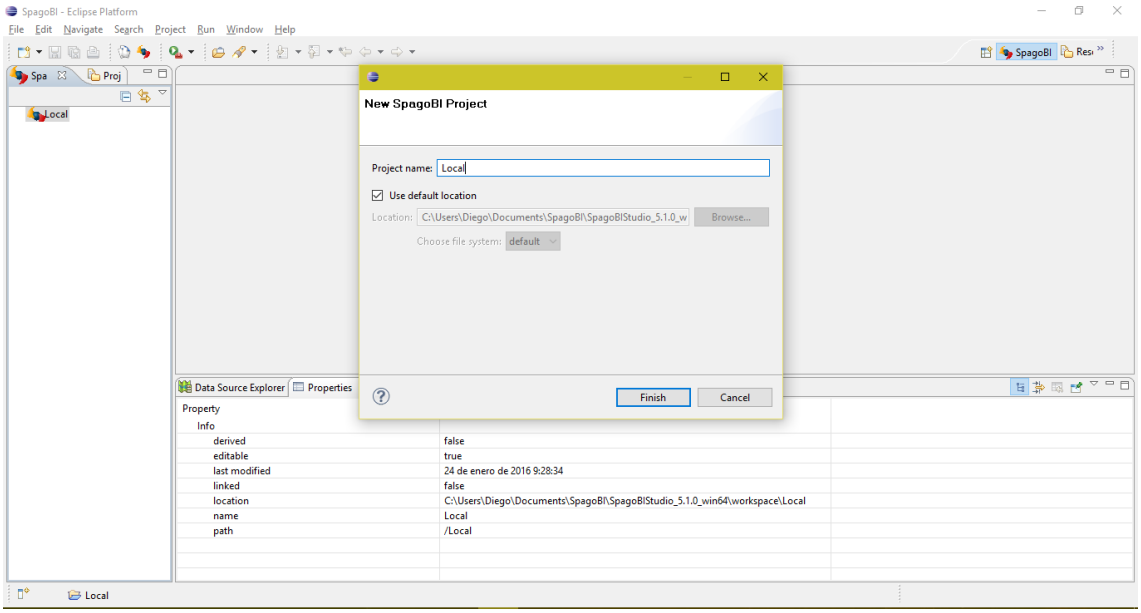
Figura 46: SpagoBI Studio 4.0.



Fuente: SpagoBI 4.0

1. El primer paso es crear un nuevo Proyecto SpagoBI

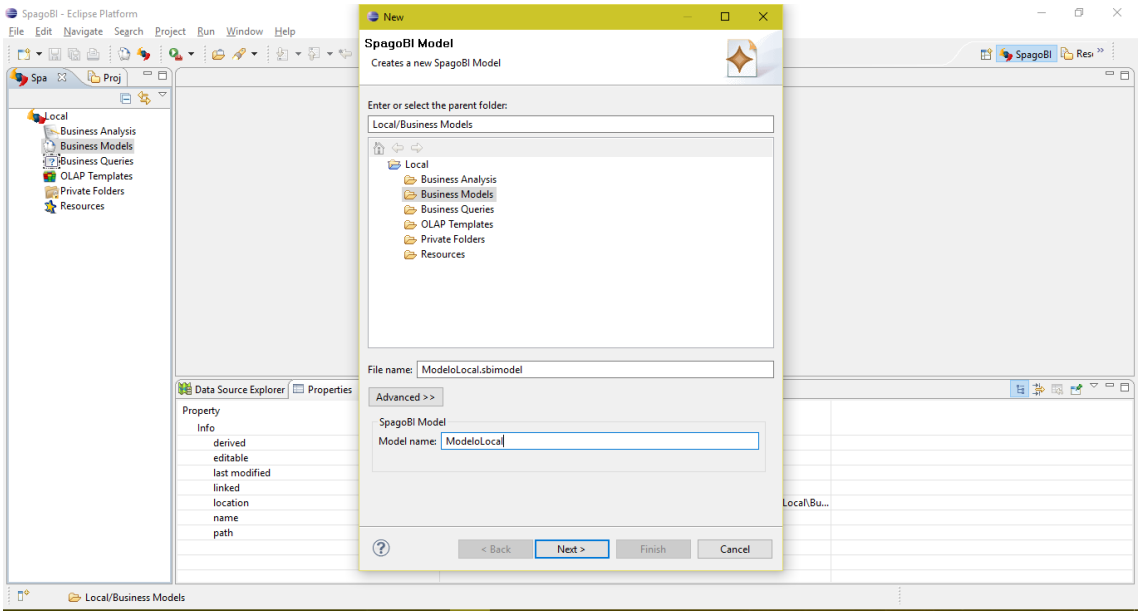
Figura 47: Creación Proyecto SpagoBI



Fuente: SpagoBI 4.0

2. El siguiente paso es crear un nuevo modelo SpagoBI

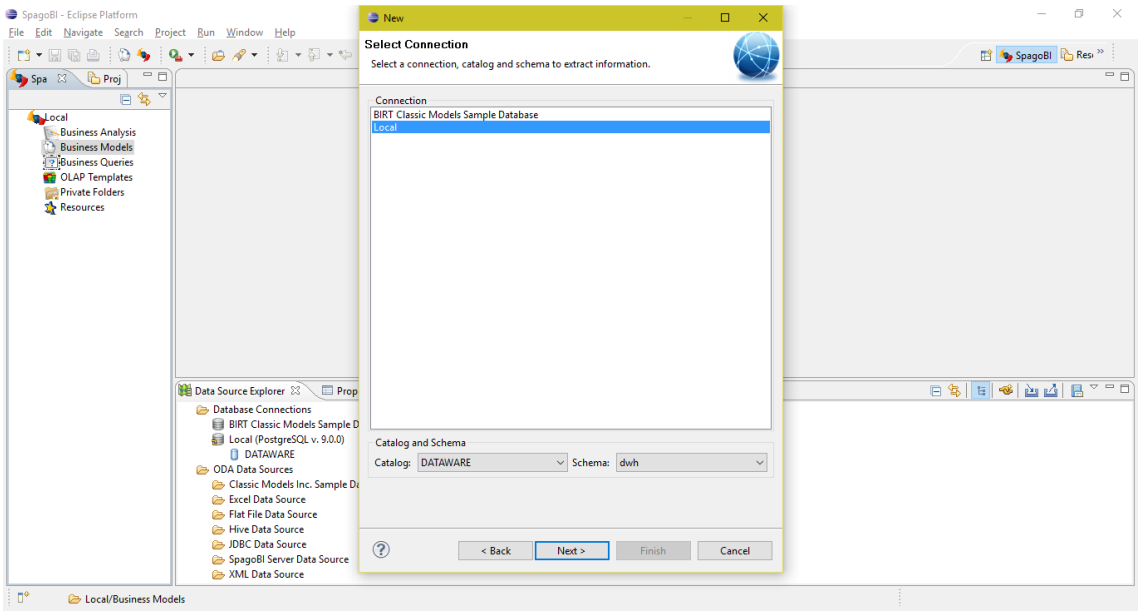
Figura 48: Crear nuevo modelo SpagoBI



Fuente: SpagoBI 4.0

3. Realizar una Conexión a nuestra Base de Datos

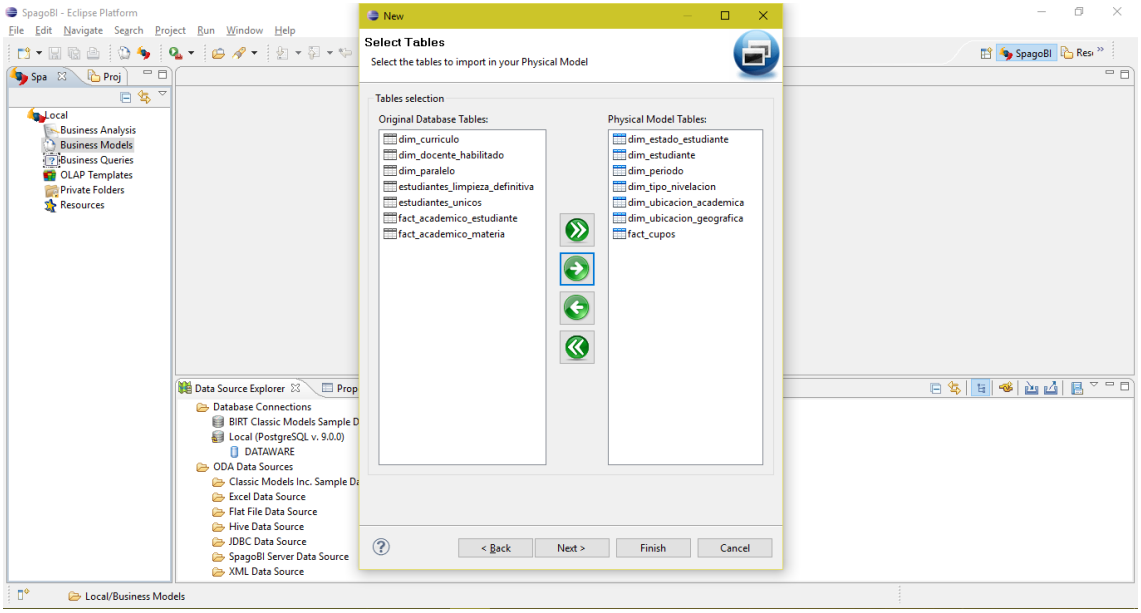
Figura 49: Conexión a la base de datos



Fuente: SpagoBI 4.0

4. Seleccionar las tablas del modelo Modelo Físico

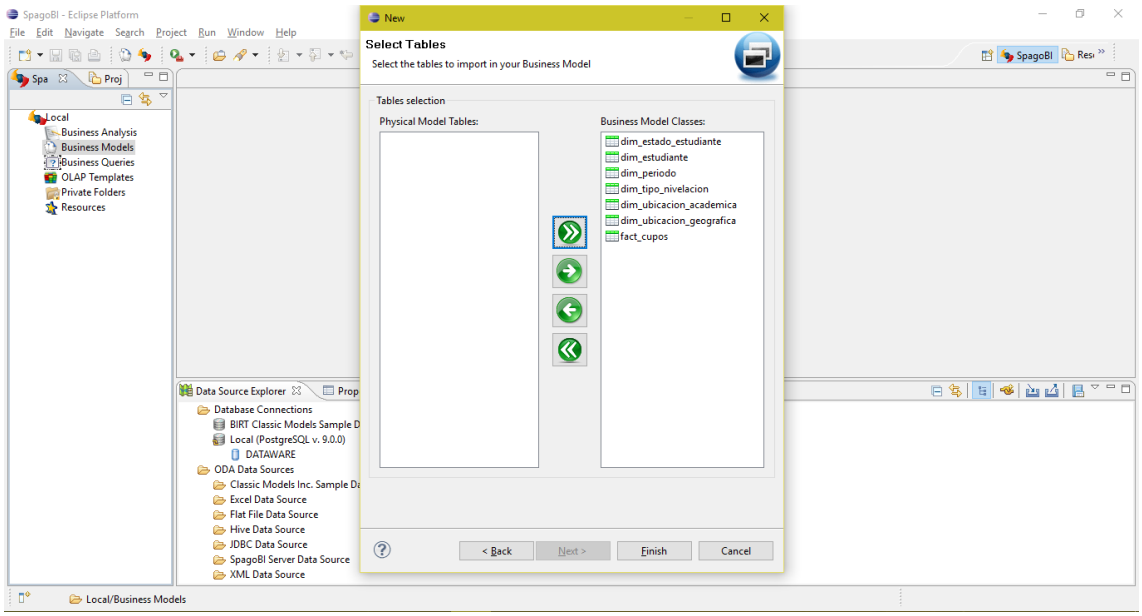
Figura 50: Selección de tablas Modelo Físico



Fuente: SpagoBI 4.0

5. Seleccionar las tablas para nuestro Cubo Dimensional

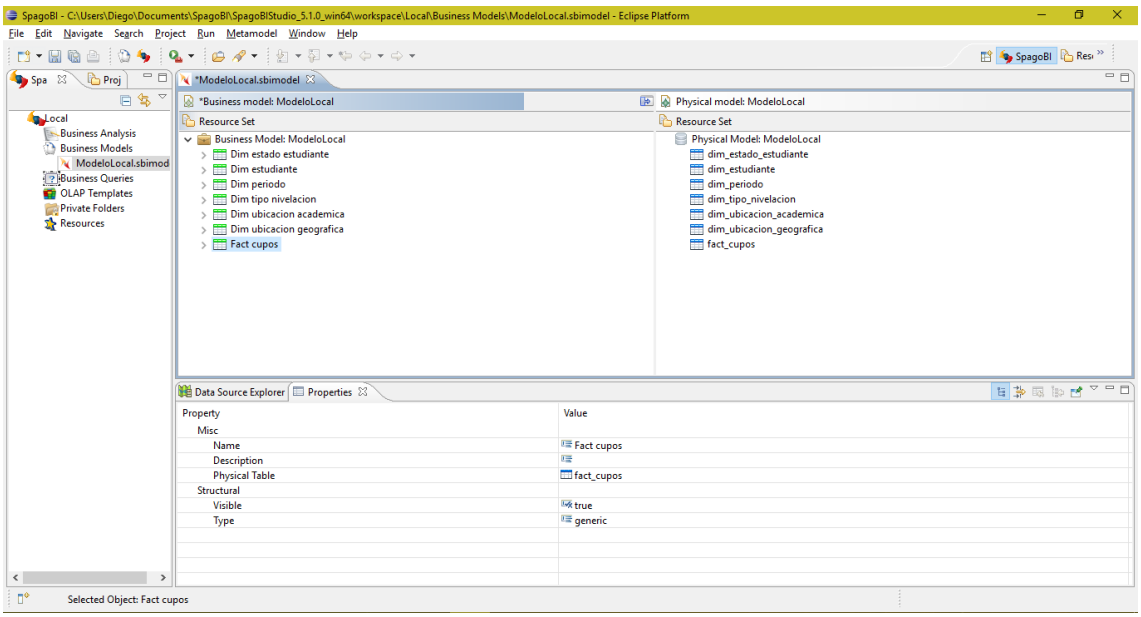
Figura 51: Selección de tablas para el cubo dimensional



Fuente: SpagoBI 4.0

6. Obtención de las tablas para el desarrollo del Modelo de Negocios

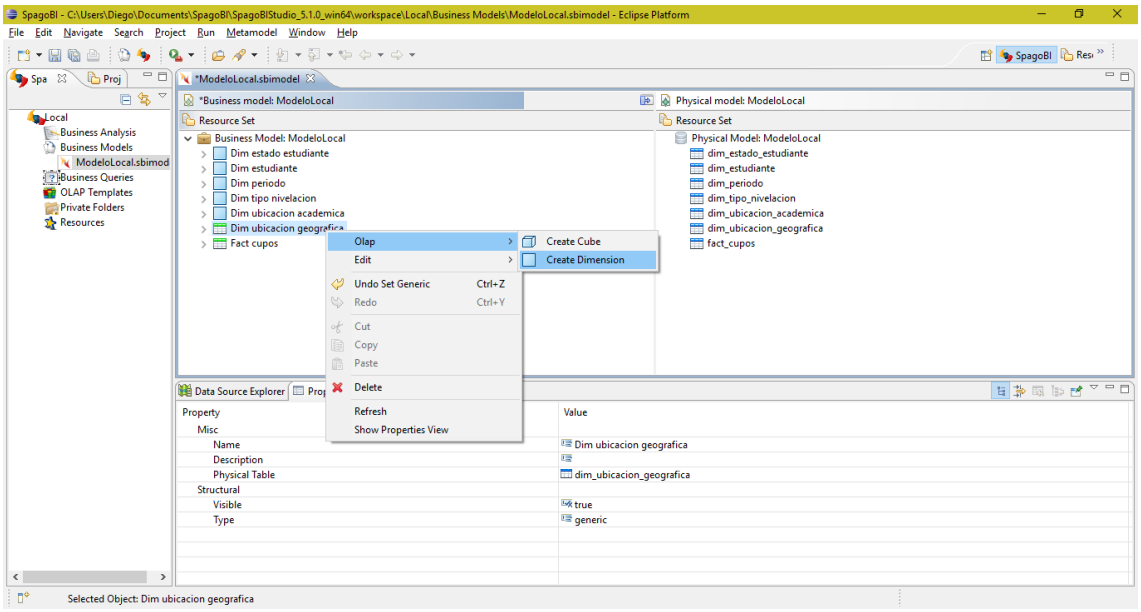
Figura 52: Obtención de las tablas para el Modelo de Negocios



Fuente: SpagoBI 4.0

7. Se crean las Dimensiones para el Cubo Dimensional

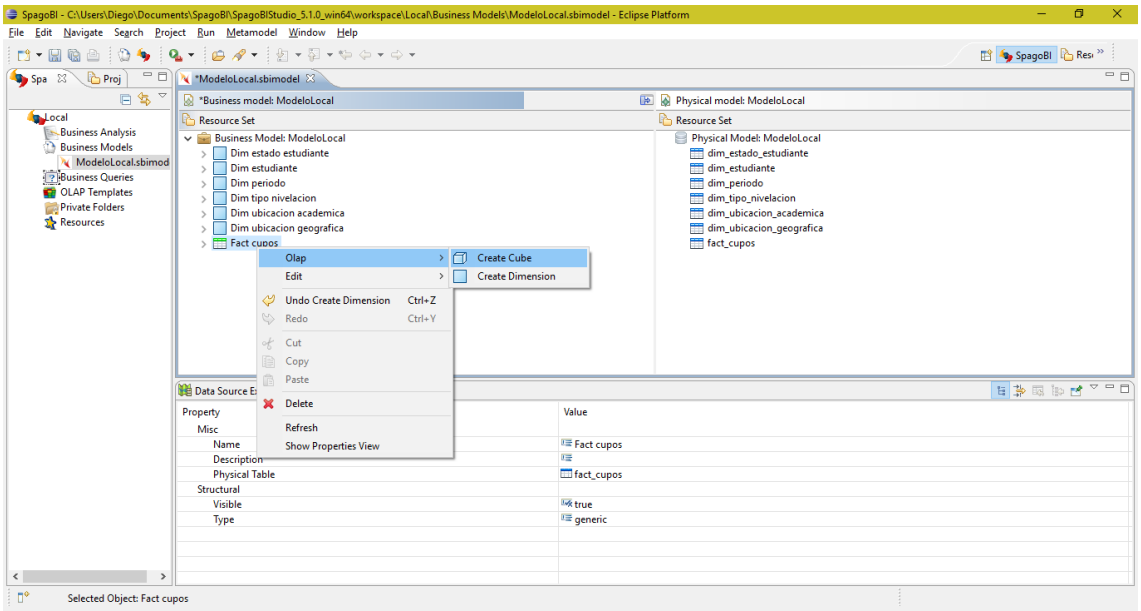
Figura 53: Cubo dimensional



Fuente: SpagoBI 4.0

8. Creamos nuestro Cubo Dimensional

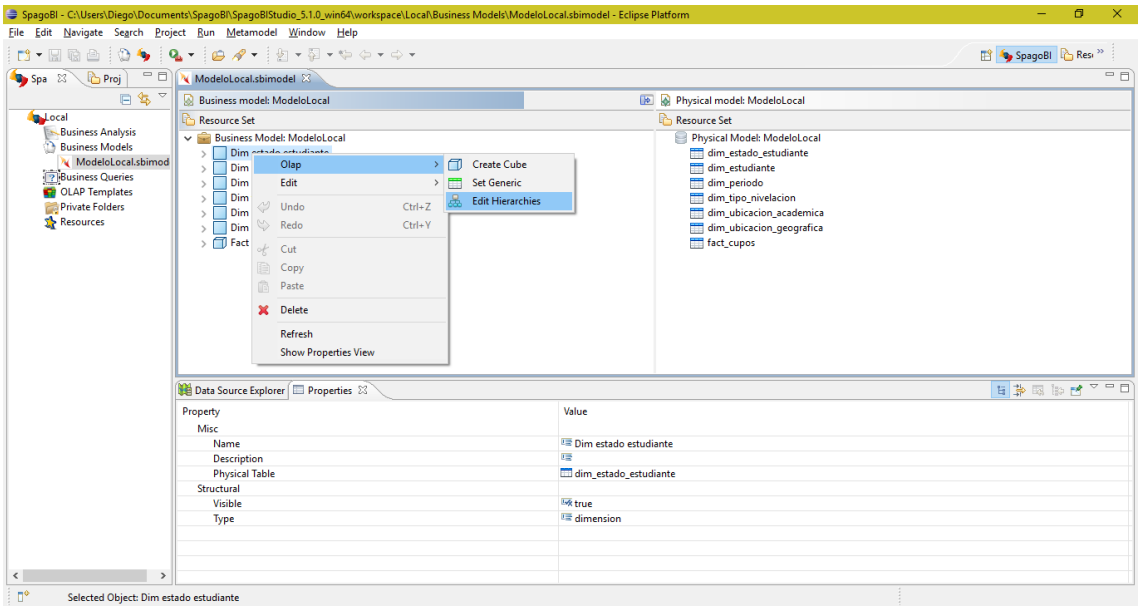
Figura 54: Creación del Cubo dimensional OLAP



Fuente: SpagoBI 4.0

9. Se asigna el nivel de Jerarquía a las Dimensiones

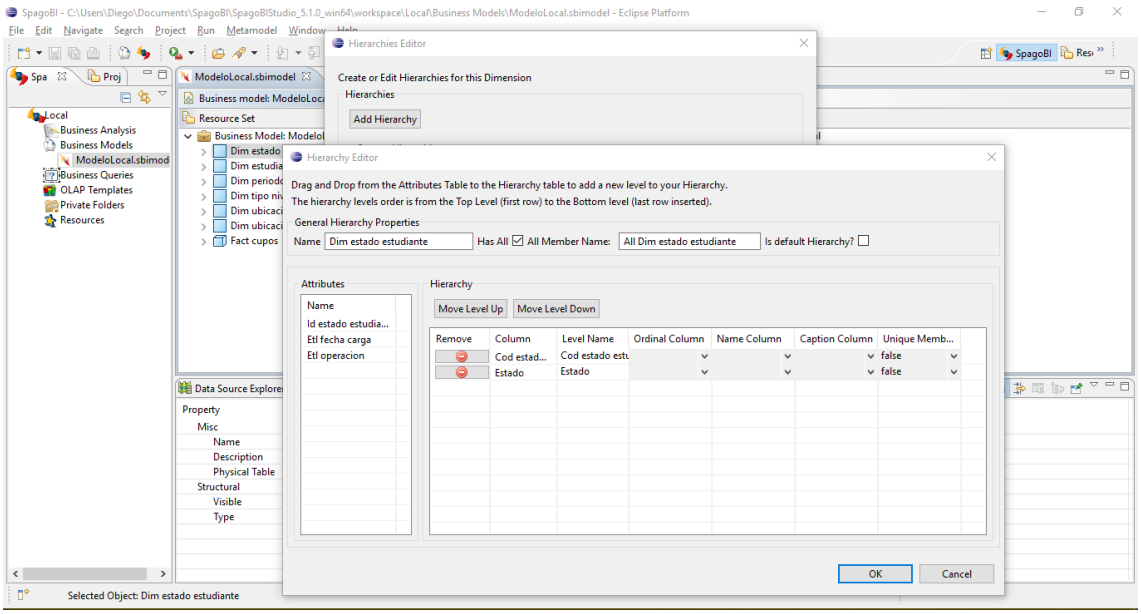
Figura 55: Niveles de Jerarquía



Fuente: SpagoBI 4.0

10. De esta forma se asigna jerarquías a todas las Dimensiones

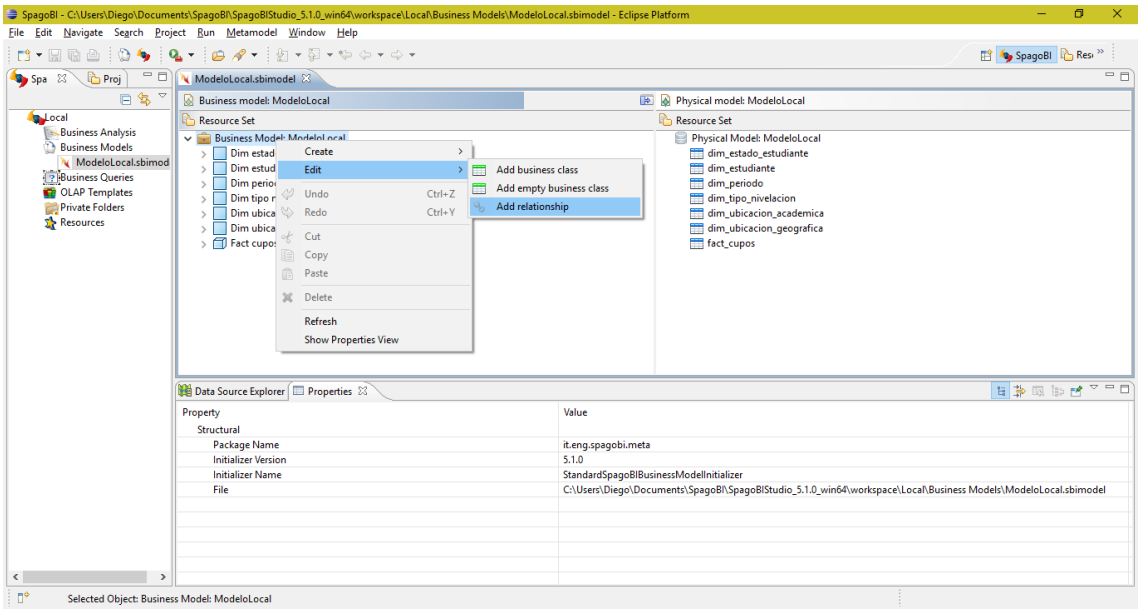
Figura 56: Asignación de jerarquías



Fuente: SpagoBI 4.0

11. Al concluir con las jerarquías, se deben asignar las relaciones de las Dimensiones con el Cubo

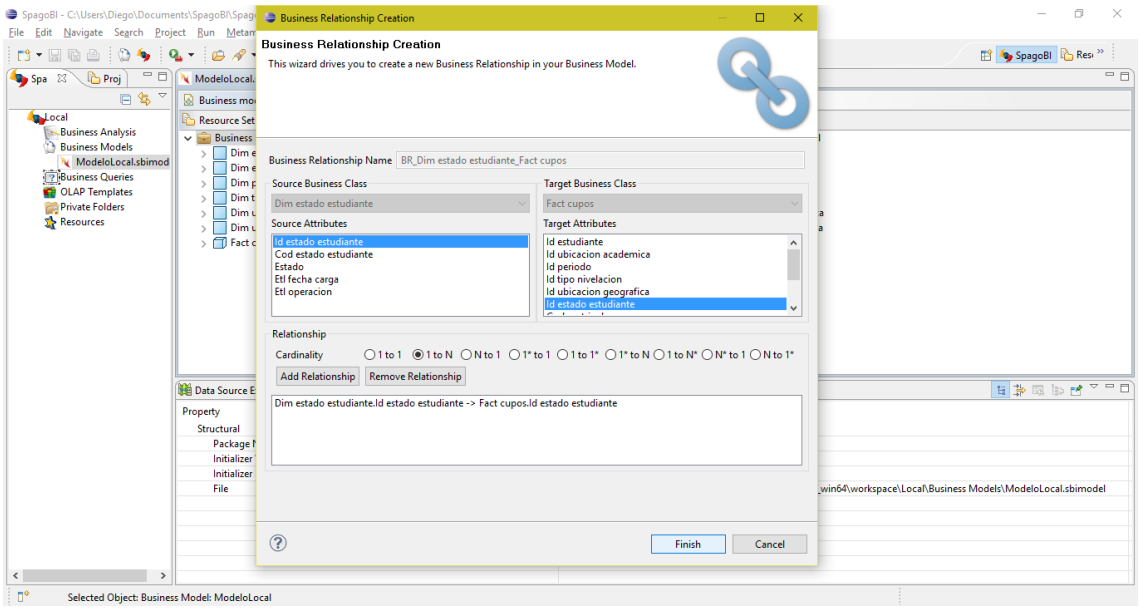
Figura 57: Relaciones de las dimensiones con el Cubo



Fuente: SpagoBI 4.0

12. Se añaden todas las relaciones que están en el proyecto

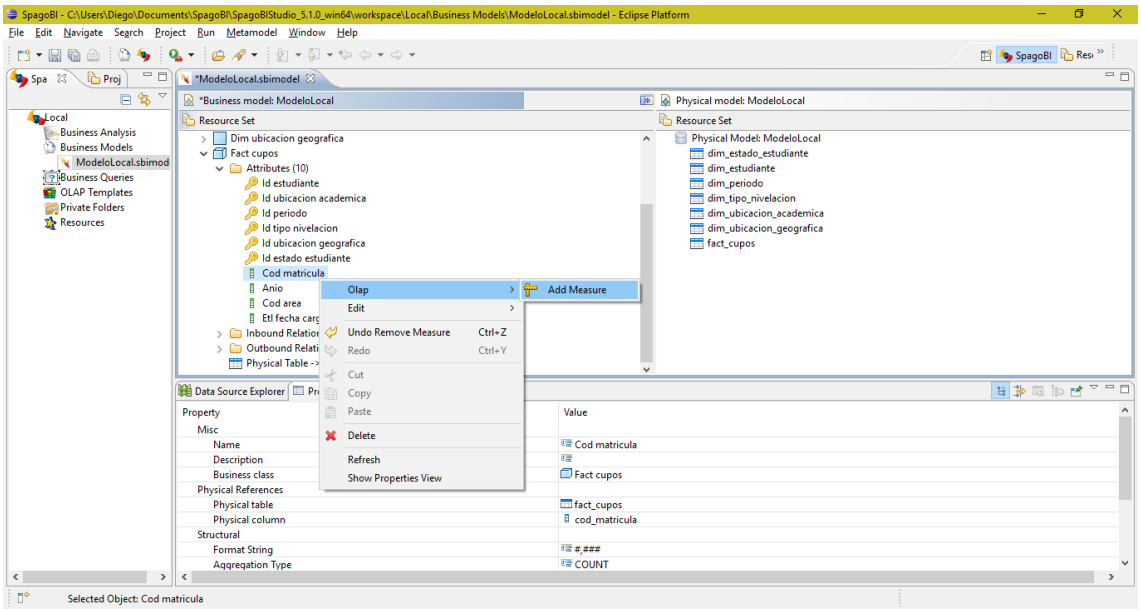
Figura 58: Relación de todas las dimensiones



Fuente: SpagoBI 4.0

13. A continuación, se debe asignar las métricas de nuestro cubo

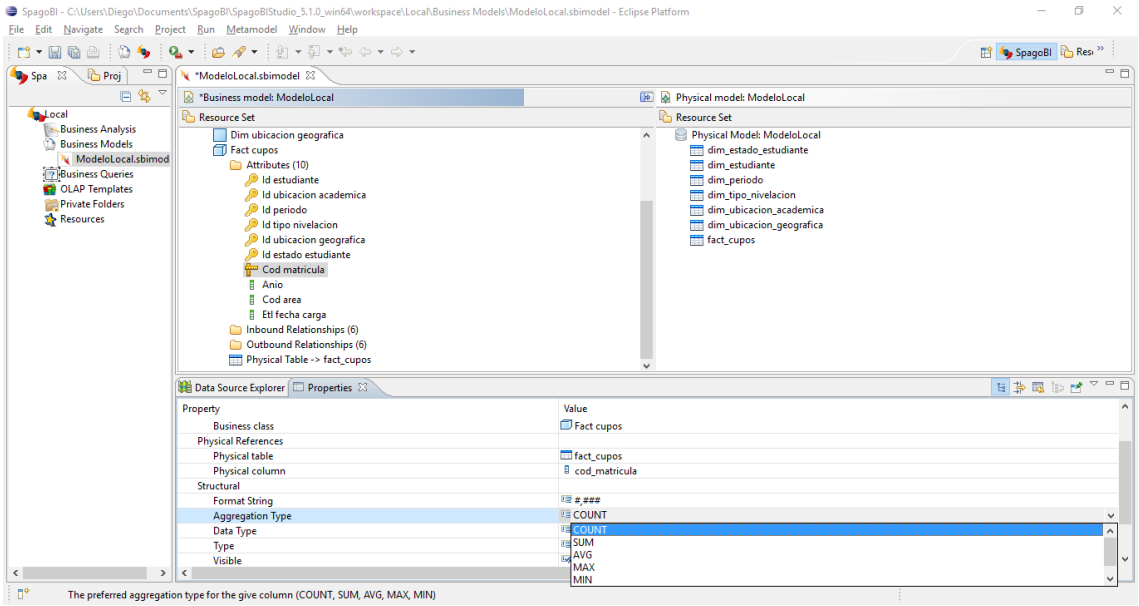
Figura 59: Asignación de Métricas



Fuente: SpagoBI 4.0

14. Se asignan las operaciones que va a realizar cada métrica.

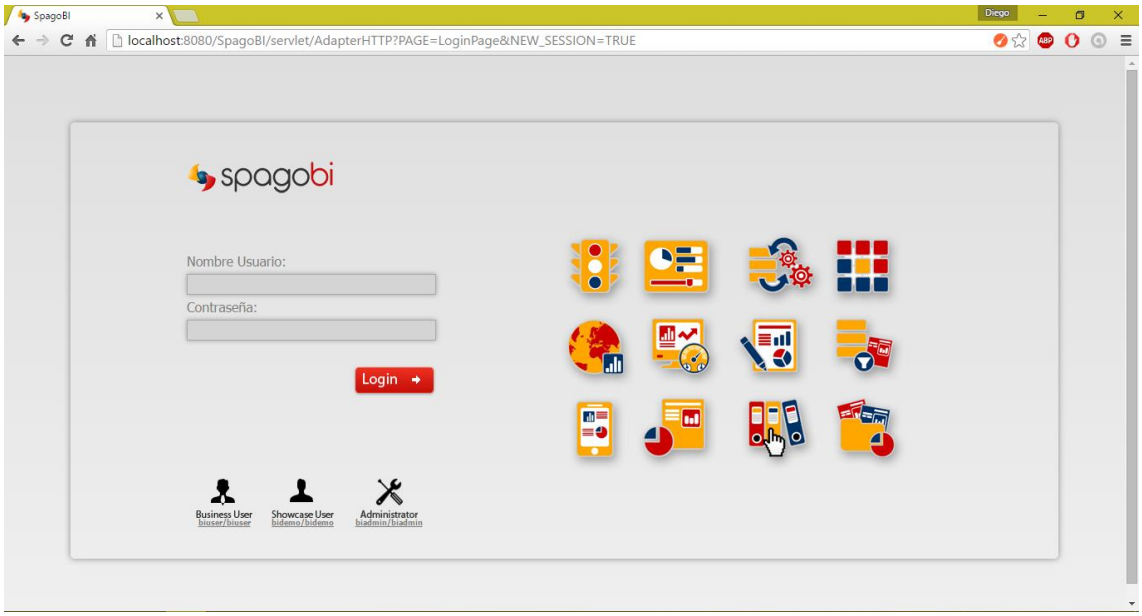
Figura 60: Asignación de operaciones a las métricas



Fuente: SpagoBI 4.0

15. Al finalizar la creación y asignación de métricas, se guarda el proyecto y se abre el servidor SpagoBI, que muestra los reportes y es donde se aloja el Modelo de Negocios

Figura 61: Servidor SpagoBI



Fuente: SpagoBI 4.0

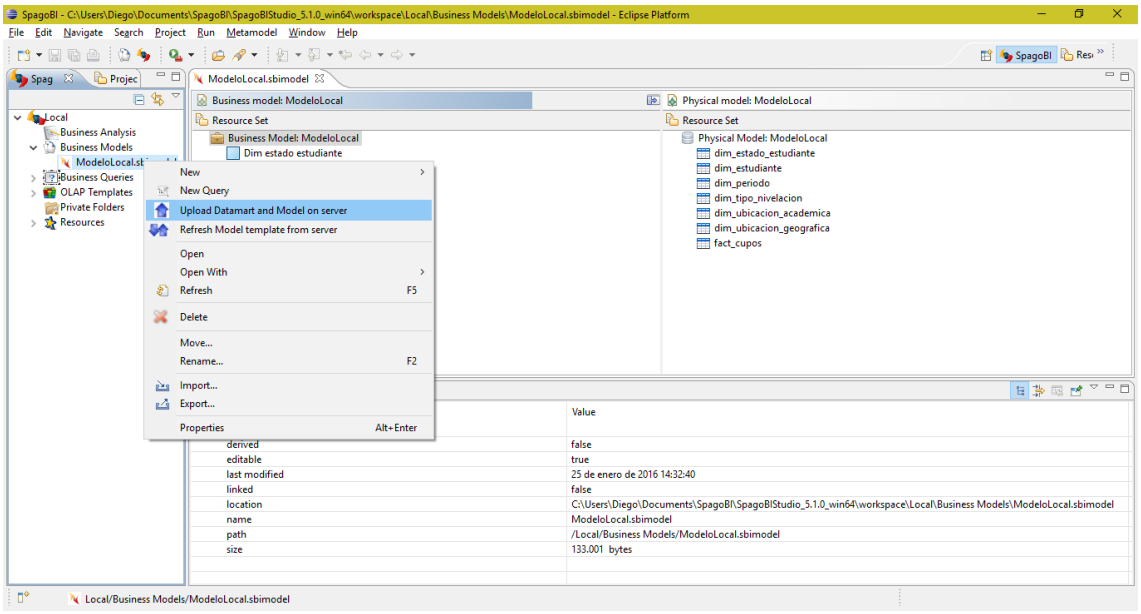
16. Se inicia como Administrador para ver el Panel de trabajo

Figura 62: Panel de trabajo SpagoBI

Fuente: SpagoBI 4.0

17. En nuestro Modelo de Negocios, se escoge la opción de almacenar nuestro proyecto al Servidor SpagoBI

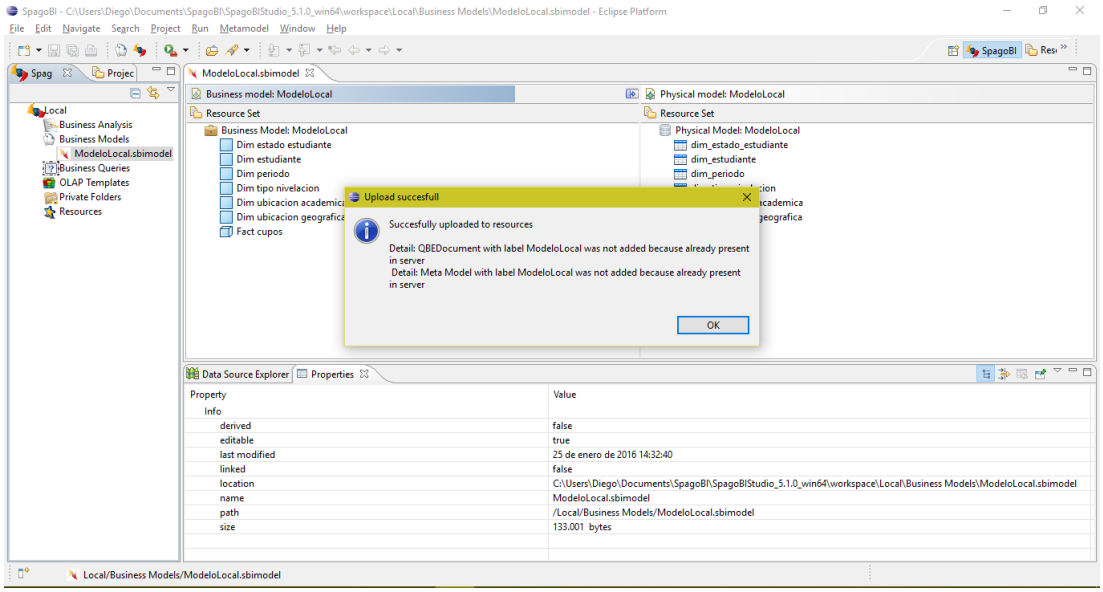
Figura 63: Almacenamiento del Modelo de Negocios en el servidor SpagoBI



Fuente: SpagoBI 4.0

18. Automáticamente se muestra el mensaje de almacenamiento correcto.

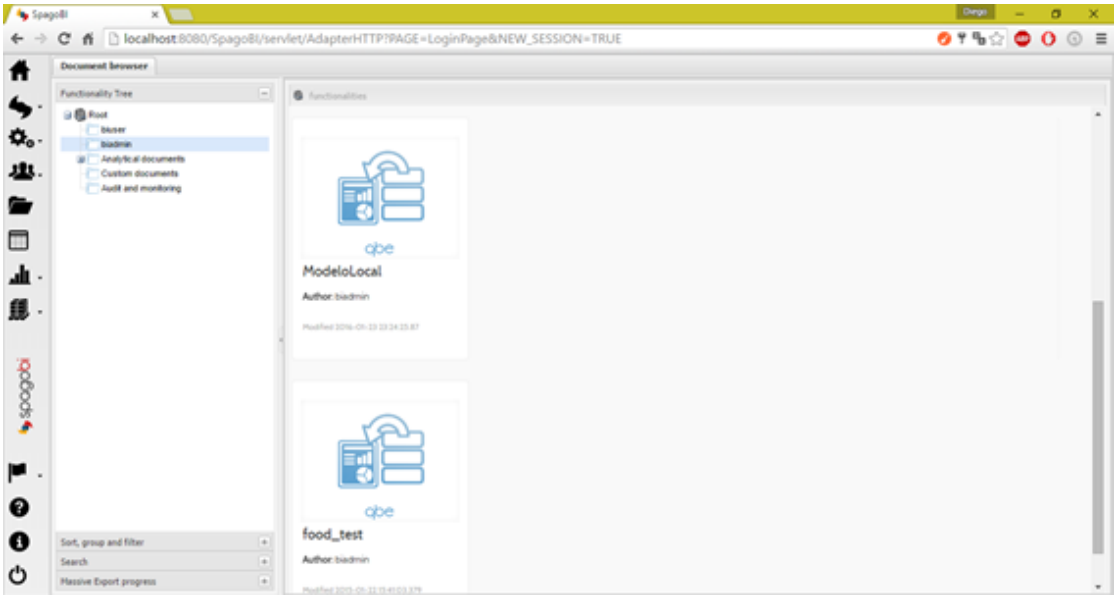
Figura 64: Almacenamiento



Fuente: SpagoBI 4.0

19. Se abre el Servidor Local con cualquier navegador y en el navegador de documentos se puede encontrar el Modelo de Negocios

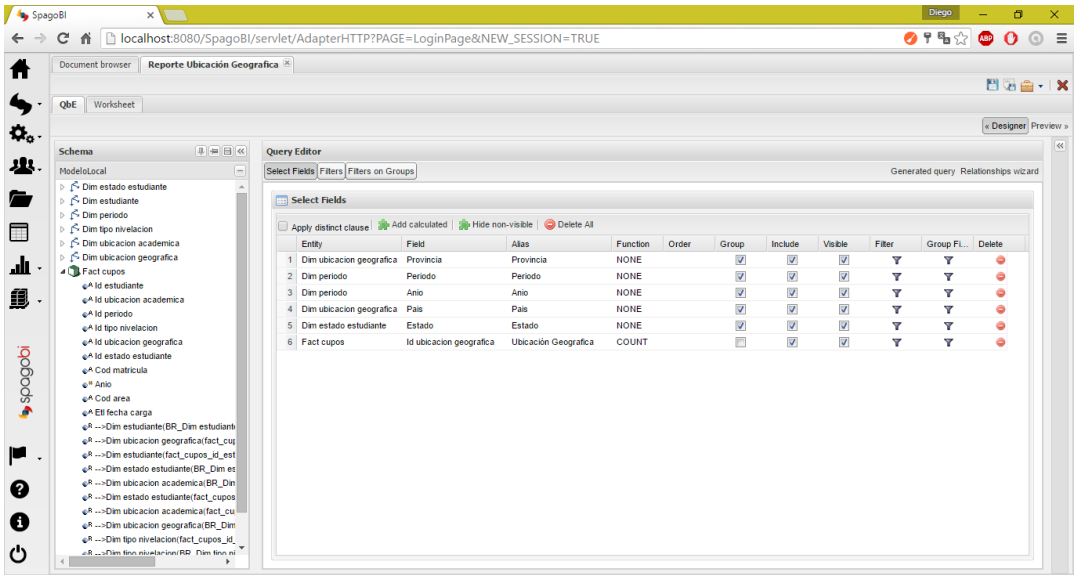
Figura 65: SpagoBI



Fuente: SpagoBI 4.0

20. Al abrir el Modelo se pueden desarrollar diferentes consultas para obtener reportes que ayuden a la toma de decisiones. Estas se realizan arrastrando los atributos al panel central y asignando una operación

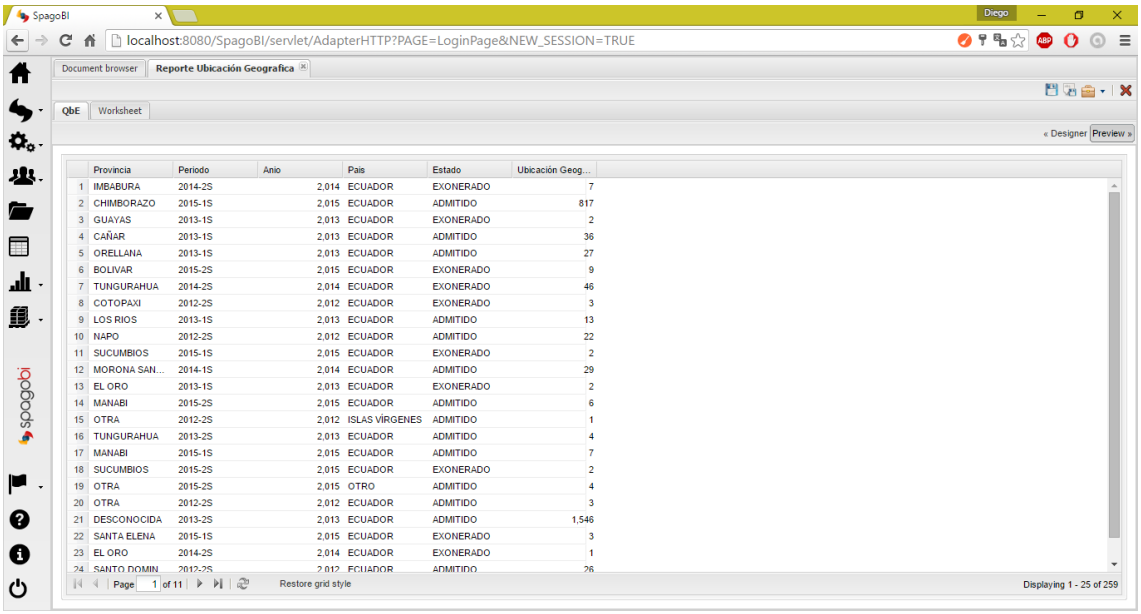
Figura 66: Consultas



Fuente: SpagoBI 4.0

21. Se puede visualizar correctamente nuestro reporte si se realizan con éxito las consultas.

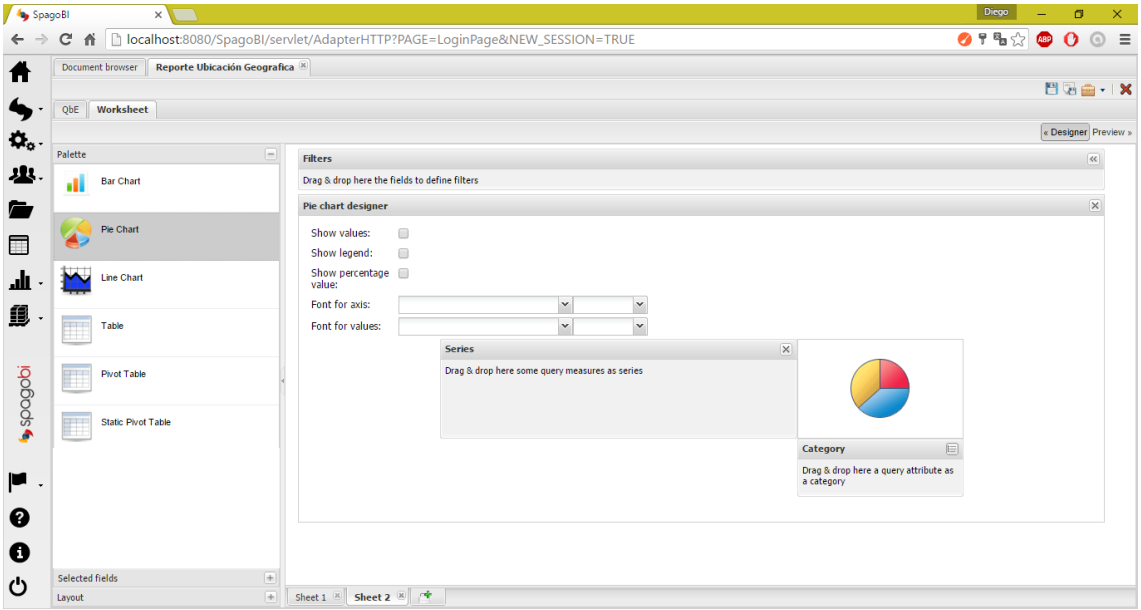
Figura 67: Visualización de consultas realizadas con éxito



Fuente: SpagoBI 4.0

22. Para realizar reportes con gráficas, se ingresa a *Worksheet* y se puede escoger los diferentes gráficos que van en los reportes.

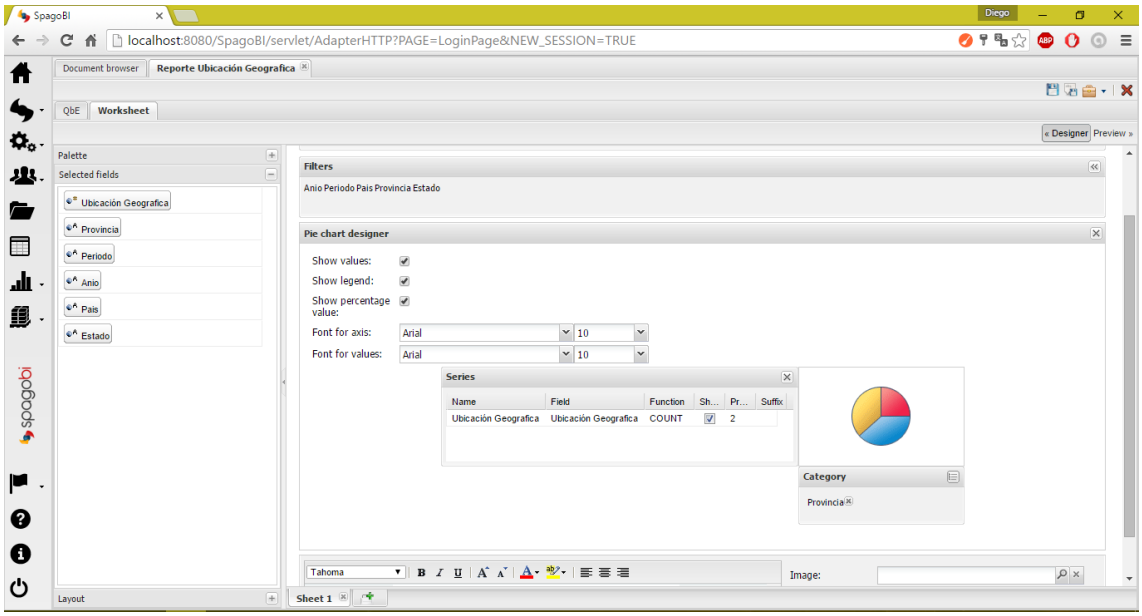
Figura 68: Reportes con gráficas



Fuente: SpagoBI 4.0

23. Se escogen los parámetros para mostrar el reporte y ayude a la toma de decisiones

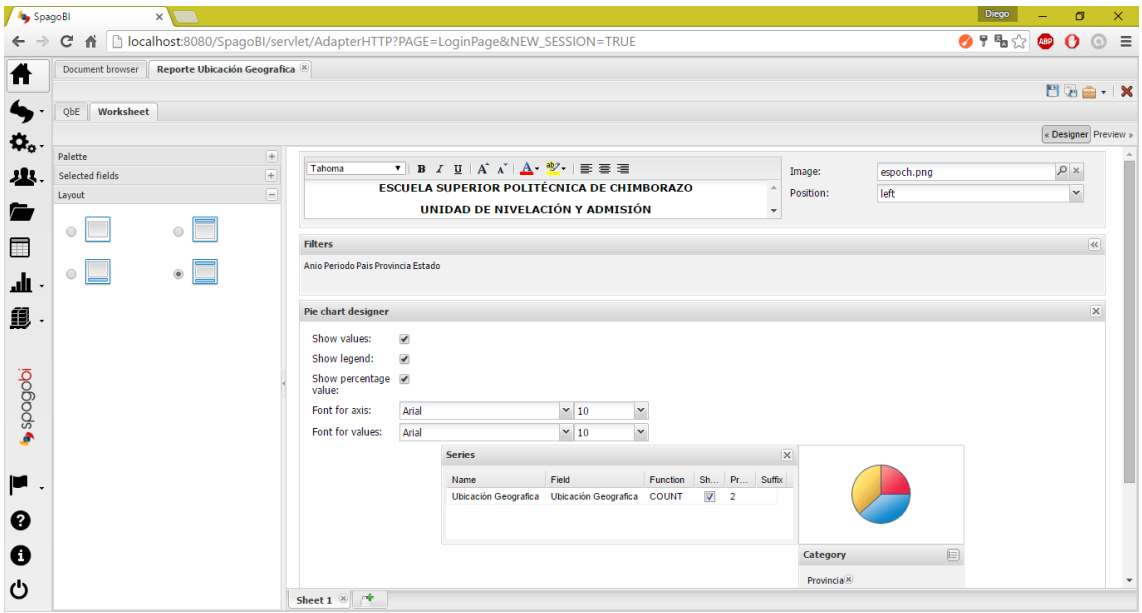
Figura 69: Parámetros del reporte



Fuente: SpagoBI 4.0

24. A continuación, se escoge el formato como se presentan los reportes donde se puede personalizar con títulos y sellos

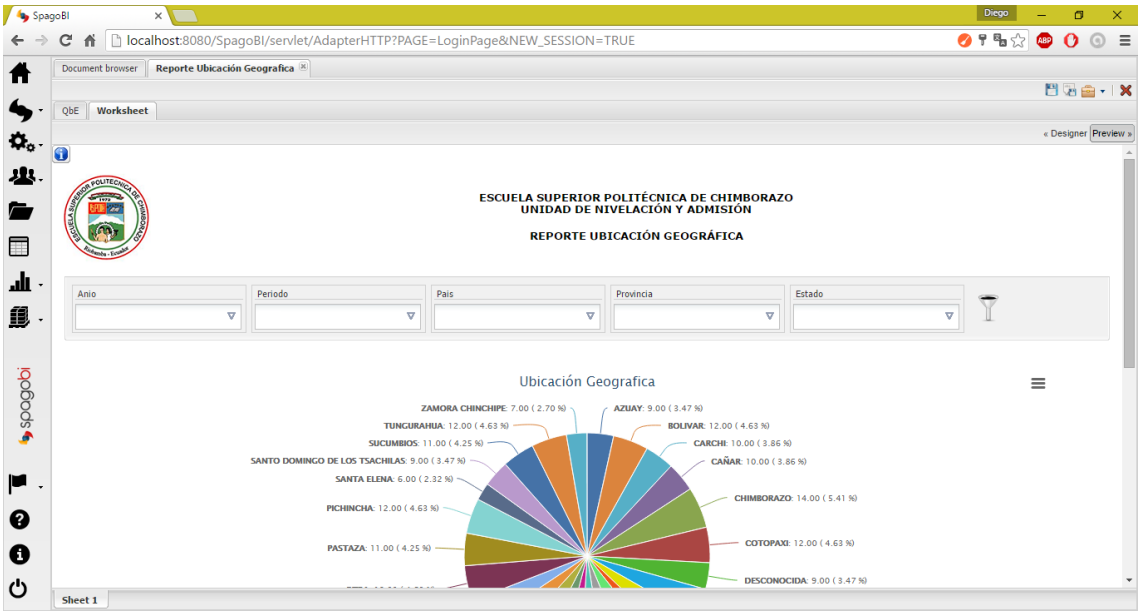
Figura 70: Configuraciones de los encabezados del reporte



Fuente: SpagoBI 4.0

25. Se selecciona la pre visualización para poder apreciar el reporte elaborado

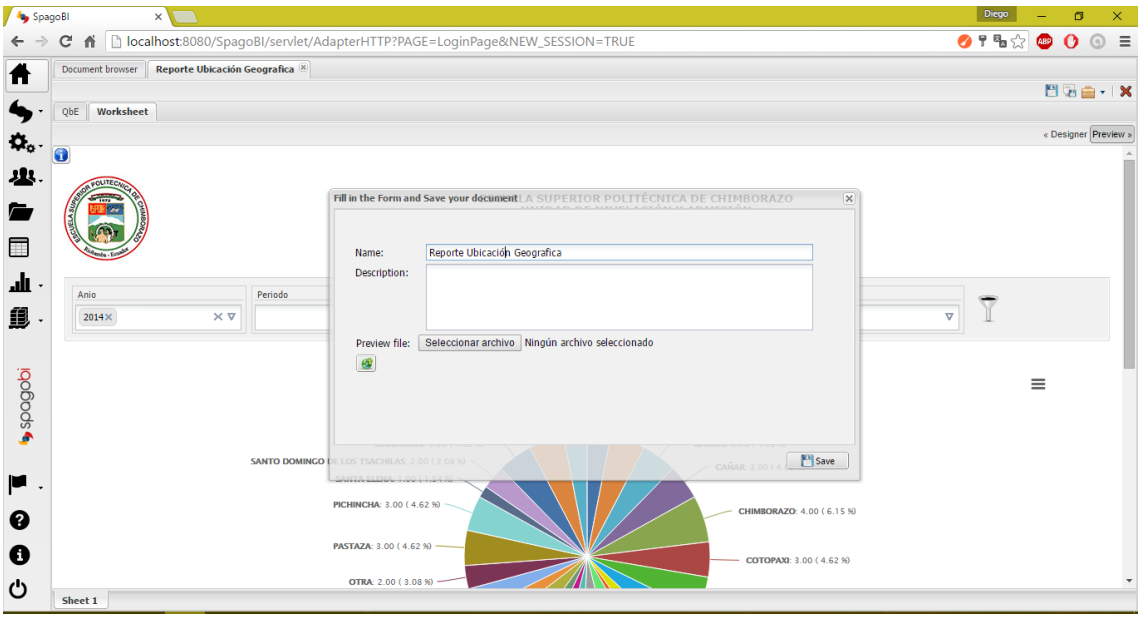
Figura 71: Pre visualización



Fuente: SpagoBI 4.0

26. Si el reporte está bien elaborado, se guarda para que se pueda visualizar en nuestro espacio de trabajo

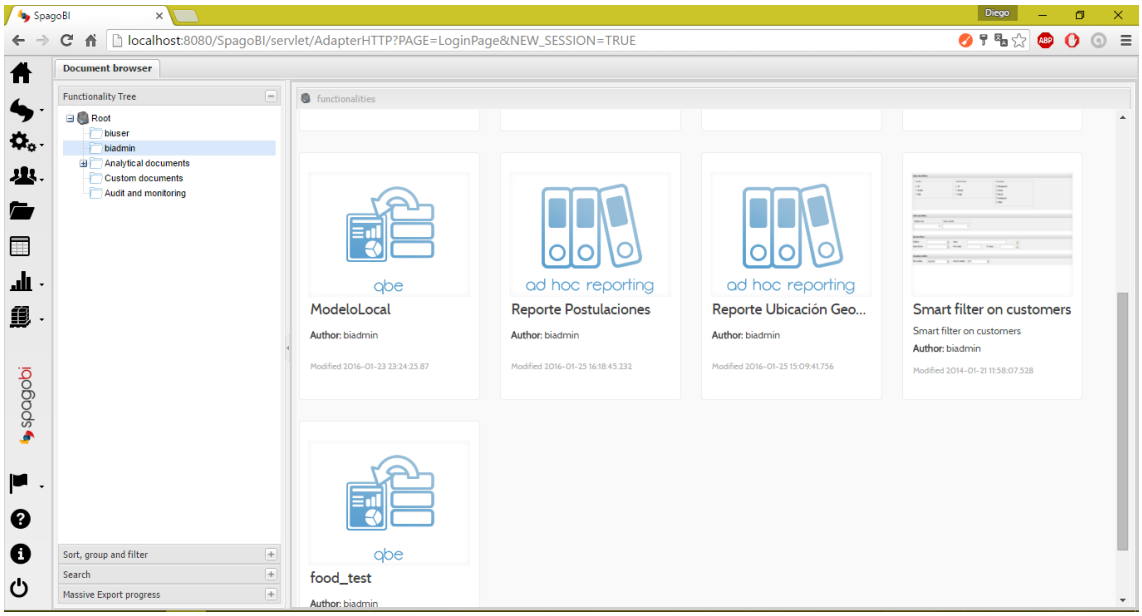
Figura 72: Guardar reporte



Fuente: SpagoBI 4.0

27. Finalmente se ingresa a nuestro espacio de trabajo para observar los diferentes reportes que se han creado y soportan la adecuada toma de decisiones en la Unidad de Nivelación y Admisión de la ESPOCH.

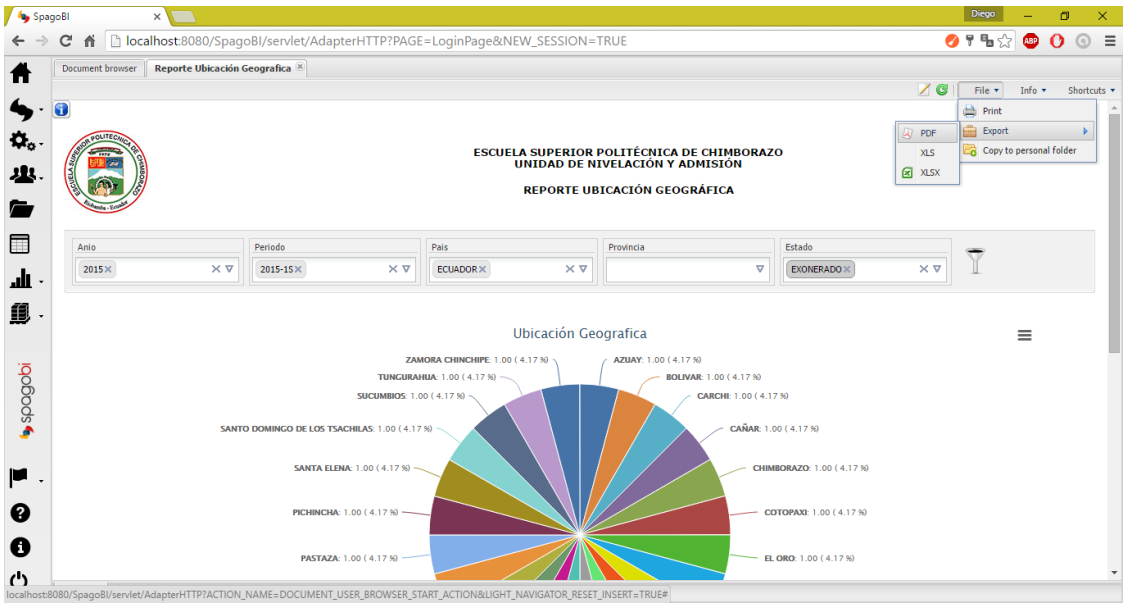
Figura 73: Espacio de trabajo



Fuente: SpagoBI 4.0

28. Adicionalmente, cada reporte tiene la posibilidad de ser exportado en formato PDF y en hojas de cálculo EXCEL

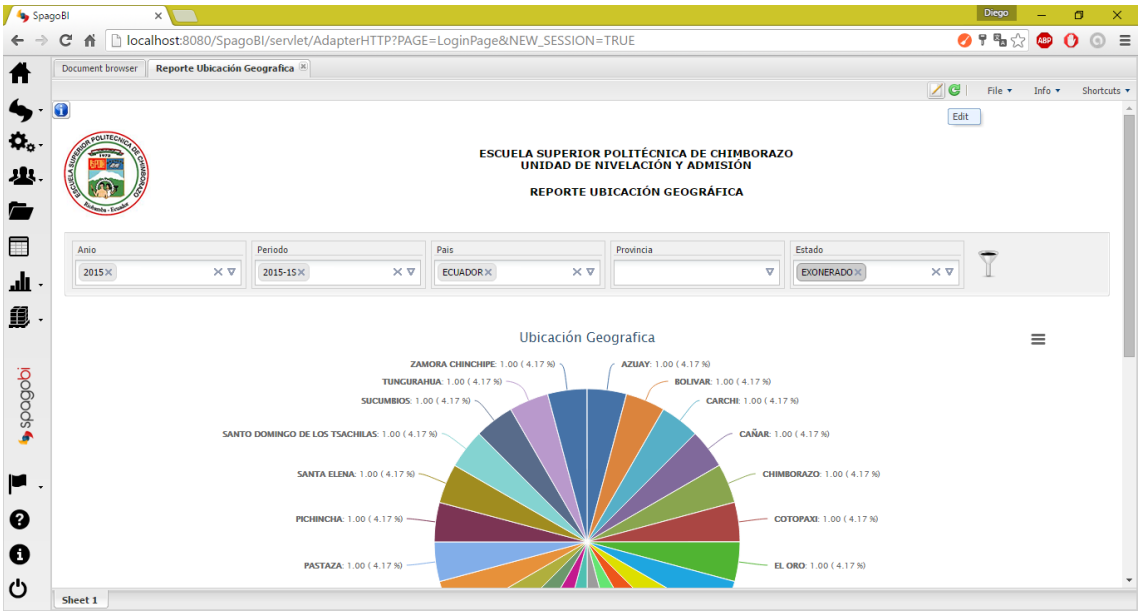
Figura 74: Exportar reportes



Fuente: SpagoBI 4.0

29. Si en el caso de querer modificar el reporte, se debe seleccionar la opción editar que permite modificar la consulta o la forma de presentación de la información

Figura 75: Modificar reportes



Fuente: SpagoBI 4.0

Referencias

Aluja, T. (2001). La Minería de Datos, entre la Estadística y la Inteligencia Artificial. *Qüestiió*, *25*(3), 479–498.

Asencios, V. V. (2004). Data Mining y el descubrimiento del conocimiento. *Industrial Data*, *7*(2), 83–86. Retrieved from http://www.redalyc.org/resumen.oa?id=81670213

Bernabeu, R. (2010). Hefesto, 146. Retrieved from http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/hefesto-metodologia-propia-para-la-construccion-un-data-warehhttp://www.dataprix.com/data-warehousing-y-metodologia-hefesto/ii-hefesto-metodologia-propia-para-la-construccion-un-data-wa

Bradley, P., Fayyad, U. M., & Mangasarian, O. (1999). Data mining: Overview and optimization opportunities. *INFORMS Journal on Computing*, *11*(January 1998), 217–238. Retrieved from http://www.ergasya.tuc.gr/Users/matsatsinis/courses/postgrad/Data Mining.pdf

Castillo, C., & Chairez, M. (2004). Toma de decisiones.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases, *17*(3), 37. http://doi.org/10.1609/aimag.v17i3.1230

Fernández, T., Duarte, A., Hernández, R., & Sánchez, Á. (2010). GRASP aplicado al problema de la selección de instancias en KDD.

González, C. B., & García, F. (2010). Práctica Final INTELIGENCIA EN COMUNICACIONES Minería de Datos : Predicción de las condiciones meteorológicas.

Inmon, W, H. (2010). Data Warehousing 2 . 0 Modeling and Metadata Strategies for Next Generation Architectures. *Architecture*, 13.

Juan, I., Moine, M., Gordillo, D. S., Ana, D., & Haedo, S. (2011). proyectos de minería de datos, 931–938.

Lewandowski, C. M. (2015). No Title No Title. *The Effects of Brief Mindfulness Intervention on Acute Pain Experience: An Examination of Individual Difference*, *1*, 1689–1699. http://doi.org/10.1017/CBO9781107415324.004

Maimon, O., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. In *Data Mining and Knowledge Discovery Handbook* (pp. 22–38). http://doi.org/10.1007/0-387-25465-x_2

Moine, M., Haedo, S., & Gordillo, D. S. (2011). Estudio comparativo de metodologías para minería

de datos. Retrieved December 3, 2015, from http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques*. http://doi.org/10.1007/978-3-540-76917-0

Reyes Saldaña, J., & García Flores, R. (2005). El Proceso de Descubrimiento de Conocimiento en Bases de datos.

Resumen Final

Descubrimiento de Conocimiento en Base de Datos para la toma de decisiones en la Unidad de Nivelación y Admisión de la ESPOCH

María Isabel Uvidia Fassler

129 páginas

Proyecto dirigido por: Diego Ávila Pesantez. Ing. MsC.

El Descubrimiento de Conocimiento en Base de datos es un conjunto de fases o pasos que permiten extraer información de datos que necesitan ser organizados y apoyan la toma de decisiones. Para este trabajo se desarrollaron sus fases acopladas a la Metodología HEFESTO versión 2.0. (Metodología propia para la construcción del *Data Warehouse*) pudiendo cumplir su aplicación en la Unidad de Nivelación y Admisión de la Escuela Superior Politécnica de Chimborazo (ESPOCH).

Este conjunto de pasos propuesto constó de 9 fases, 4 clasificadas dentro del subproceso *Data Warehouse* y las otras 5 dentro del subproceso *Data Mining*. El subproceso *Data Warehouse* inicia desde aprender el dominio de la aplicación, es decir, conocer el área de los datos con los que se trabaja, además de las necesidades o requerimientos de información, selección, limpieza de datos para permitir información consistente, siguiendo con el diseño del DW donde se guarda la información mediante procesos de Extracción, Transformación y Carga (ETL) hasta obtener información preparada para el siguiente subproceso que es *Data Mining*, donde mediante la selección de técnicas y análisis de información, se puede obtener patrones que una vez observados y examinados se convierten en conocimiento para la toma de decisiones.

Esta toma de decisiones fue complementada mediante la creación de reportes *Business Intelligence* que reflejan el proceso académico de admisión y nivelación desarrollado por la ESPOCH durante estos años.