



UNIVERSIDAD NACIONAL DE CHIMBORAZO

FACULTAD DE INGENIERÍA

ESCUELA DE INGENIERÍA EN SISTEMAS Y COMPUTACIÓN

**“Trabajo de grado previo a la obtención del Título de Ingeniero en Sistemas y
Computación”**

TRABAJO DE GRADUACIÓN

Título del proyecto

**ANÁLISIS COMPARATIVO DE LAS PLATAFORMAS WEKA Y MICROSOFT
ANALYSIS SERVICES PARA OPTIMIZAR EL DESARROLLO DE MINERÍA DE
DATOS EN LA EMPRESA PRASOL “LÁCTEOS SANTILLÁN”.**

Autores:

NAY LEE MOJARRANGO PERLAZA.

JOSÉ EDUARDO CHAPALBAY OLEAS.

Director:

Ing. Margarita Aucancela.

Riobamba – Ecuador

2016

Los miembros del Tribunal de Graduación del proyecto de investigación de título: **“ANÁLISIS COMPARATIVO DE LAS PLATAFORMAS WEKA Y MICROSOFT ANALYSIS SERVICES PARA OPTIMIZAR EL DESARROLLO DE MINERÍA DE DATOS EN LA EMPRESA PRASOL LÁCTEOS SANTILLÁN”**, presentado por: José Eduardo Chapalbay Oleas, Nay Lee Mojarrango Perlaza y dirigida por: Ing. Margarita Aucancela.

Una vez escuchada la defensa oral y revisado el informe final del proyecto de investigación con fines de graduación escrito en la cual se ha constatado el cumplimiento de las observaciones realizadas, remite la presente para uso y custodia en la biblioteca de la Facultad de Ingeniería de la UNACH.

Para constancia de lo expuesto firman:

Ing. Danny Velasco
Presidente del Tribunal



Firma

Ing. Margarita Aucancela
Director del Proyecto



Firma

Ing. Ana Congacha
Miembro del Tribunal



Firma

AUTORÍA DE LA INVESTIGACIÓN

"La responsabilidad del contenido de este Proyecto de Graduación, nos corresponde exclusivamente a: Nay Lee Mojarrango Perlaza, José Eduardo Chapalbay Oleas (Autores) y del Ing. Ing. Margarita Aucancela (Director); y el patrimonio intelectual de la misma a la Universidad Nacional de Chimborazo".


Nay Lee Mojarrango Perlaza
C.I. 080318983-6


José Eduardo Chapalbay Oleas
C.I. 060505933-6

AGRADECIMIENTO

En el presente trabajo de tesis queremos agradecer a Dios por brindarnos la salud, la vida y por permitirnos hacer realidad nuestro anhelado sueño.

Expresamos nuestro eterno agradecimiento a la Universidad Nacional de Chimborazo, la cual nos abrió sus puertas para culminar con éxito una etapa más de nuestras vidas.

Un reconocimiento especial a nuestra Directora de Tesis, Ing. Margarita Aucancela por guiarnos en nuestra tesis, quien con sus conocimientos, su experiencia, su paciencia y su motivación ha logrado que con esfuerzo y dedicación terminemos nuestros estudios con éxito.

A todos los docentes de la Escuela de Ingeniería en Sistema y Computación porque han aportado con sus conocimientos y consejos para nuestra formación profesional.

Para ellos muchas gracias y que Dios les bendiga

Nay Lee Mojarrango - José Chapalbay

DEDICATORIA

*La presente tesis se la dedico ante todo a **Dios**, y a todos los **Santos**, por darme la fuerza, la fortaleza y la voluntad para afrontar las diversas dificultades y situaciones de la vida, que gracias a él las he podido superar.*

*Al apoyo incondicional de mis padres: **Nay Leopoldo Mojarrango Vásquez** y a mi madre **Maira Perlaza Borbor**, darles gracias por el gran labor que han hecho a lo largo de mi vida guiándome por el buen camino hacia el éxito, que a pesar de las adversidades siempre hemos salido adelante.*

*A mi querida hermana **Aira Lia Mojarrango Perlaza**, por siempre estar ahí conmigo apoyándome en todo lo que haga, y además servirle como ejemplo de superación, constancia y éxito, demostrándole que en esta vida todo se puede lograr si uno se lo propone.*

*A mis abuelos **Adelaida Borbor** y **Oscar Perlaza**, que aunque ya no estén con nosotros, siempre estuvieron apoyándome.*

*Finalmente a mi querida tía **Aira Perlaza** por todo el apoyo que me ha brindado a lo largo de toda mi vida.*

Nay Lee Mojarrango Perlaza

DEDICATORIA

Esta tesis se la dedico a mi Dios quién supo guiarme por el buen camino, darme fuerzas para seguir adelante y no desmayar en los problemas que se presentaban, enseñándome a encarar las adversidades sin perder nunca la dignidad ni desfallecer en el intento.

Con mucho cariño principalmente a mis padres por su apoyo, consejos, comprensión, amor, ayuda en los momentos difíciles y por ayudarme con los recursos necesarios para estudiar. Me han dado todo lo que soy como persona, mis valores, mis principios, mi carácter, mi empeño y mi perseverancia. Gracias por todo mamá y papá por darme una carrera para mi futuro y por creer en mí, aunque hemos pasado momentos difíciles siempre han estado apoyándome y brindándome todo su amor, por todo esto les agradezco de todo corazón, el que estén conmigo a mi lado.

José Eduardo Chapalbay Oleas

ÍNDICE GENERAL

AUTORÍA DE LA INVESTIGACIÓN	2
AGRADECIMIENTO	4
DEDICATORIA	5
ÍNDICE GENERAL	7
ÍNDICE DE TABLAS	10
ÍNDICE DE FIGURAS	12
RESUMEN	16
INTRODUCCIÓN	17
CAPÍTULO I	20
1.1 TÍTULO DEL PROYECTO	20
1.2 PROBLEMATIZACIÓN	20
CAPÍTULO II	24
FUNDAMENTACIÓN TEÓRICA	24
2.1 ANTECEDENTES	24
2.2 MINERÍA DE DATOS	24
2.3 MINERÍA DE DATOS COMO SOPORTE A LA TOMA DE DECISIONES ...	25
2.4 BUSINESS INTELLIGENCE	26
2.5 HERRAMIENTAS ETL	26
2.6 DATAMART	27

2.7	DATAWAREHOUSE	28
2.8	DIFERENCIA ENTRE DATAWHAREHOUSE & DATAMART.....	30
2.9	PLATAFORMA WEKA	31
2.10	MICROSOFT ANALYSIS SERVICES	36
2.11	METODOLOGÍA CRISP-DM (CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING).....	39
	CAPÍTULO III	41
	ANÁLISIS COMPARATIVO.....	41
3.1	ANÁLISIS COMPARATIVO DE HERRAMIENTAS MINERÍA DE DATOS	41
	CAPÍTULO IV	64
	DESARROLLO DE MINERÍA DE DATOS.....	64
4.1	FASE DE COMPRESIÓN DEL NEGOCIO.....	64
4.2	FASE DE COMPRESIÓN DE LOS DATOS	68
4.3	FASE DE PREPARACIÓN DE LOS DATOS	83
4.4	FASE DE MODELADO	97
4.5	FASE DE EVALUACIÓN.....	100
4.6	FASE DE IMPLEMENTACIÓN	104
	CAPÍTULO V.....	109
	METODOLOGIA.....	109
5.1	TIPO DE ESTUDIO.....	109
5.2	POBLACIÓN Y MUESTRA	109
5.3	OPERACIONALIZACIÓN DE VARIABLES.....	110

5.4	PROCEDIMIENTOS	112
5.5	PROCEDIMIENTO Y ANÁLISIS	112
	CAPÍTULO VI	113
	RESULTADOS Y DISCUSIÓN	113
6.1	RESULTADOS	113
6.2	COMPROBACIÓN DE LA HIPÓTESIS	127
6.3	DISCUSIÓN	134
	CAPÍTULO VII	137
	CONCLUSIONES Y RECOMENDACIONES	137
7.1	CONCLUSIONES	137
7.2	RECOMENDACIONES	138
	BIBLIOGRAFÍA	139
	ANEXOS	140

ÍNDICE DE TABLAS

Tabla 1: SUBFASES DE LA METODOLOGÍA CRISP-DM	53
Tabla 2: CREACIÓN DE INDICADORES DE EVALUACIÓN DE HERRAMIENTAS DE MINERÍA DE DATOS	55
Tabla 3: ANÁLISIS DE LAS HERRAMIENTAS MICROSOFT ANALYSIS SERVICES Y WEKA	59
Tabla 4: RESULTADOS DEL ESTUDIO COMPARATIVO.....	60
Tabla 5: PROMEDIO	61
Tabla 6: Plan del Proyecto.....	68
Tabla 7: Tablas útiles para la construcción del datamart.....	74
Tabla 8: Verificación Calidad de Datos BODEGA	79
Tabla 9: Verificación Calidad de Datos CLIENTE.....	80
Tabla 10: Verificación Calidad de Datos DET_FACTURA	81
Tabla 11: Verificación Calidad de Datos FACTURA.....	82
Tabla 12: Verificación Calidad de Datos PRODUCTO	82
Tabla 13: Operacionalización de Variables	111
Tabla 14: Resultados de la Sub Fase de Recolección de Datos Iniciales	113
Tabla 15: Resultados de la Sub Fase de Descripción de los datos	114
Tabla 16: Resultados de la Sub Fase de Exploración de los Datos	115
Tabla 17: Resultados de la Sub Fase de Verificación de la Calidad de los Datos.....	116
Tabla 18: Resultados de la Sub Fase de Selección de Datos	117
Tabla 19: Resultados de la Sub Fase de Limpieza de Datos	118
Tabla 20: Resultados de la Sub Fase de Estructuración de los Datos	119
Tabla 21: Resultados de la Sub Fase de Integración de los Datos	120
Tabla 22: Resultados de la Sub Fase de Formateo de los Datos	121

Tabla 23: Resultados de la Sub Fase de Selección de la Técnica de Modelado.....	122
Tabla 24: Resultados de la Sub Fase de Construcción del Modelo.....	123
Tabla 25: Resultados de la Sub Fase de Evaluación del Modelo.....	124
Tabla 26: Resultados de la Sub Fase de Informe Final.....	126
Tabla 27: Fases & Porcentaje de Tiempo.....	128
Tabla 28: Ponderación, Fases, Sub Fases y Porcentaje de Tiempo.....	129
Tabla 29: Cumplimiento de los indicadores según el análisis comparativo y en porcentaje	130
Tabla 30: Tabla que permitirá aplicar la prueba T-Student.....	131
Tabla 31: Resultado Prueba T-Student.....	133

ÍNDICE DE FIGURAS

Figura 1: Weka GUI Chooser	31
Figura 2: WEKA Simple CLI	32
Figura 3: WEKA Explorer	33
Figura 4: WEKA Experiment Environment	33
Figura 5: WEKA KnowledgeFlow Environment	34
Figura 6: Analysis Services	38
Figura 7: Metodologías de Data Mining	40
Figura 8: Resultado Puntuación Sub fases	61
Figura 9: Análisis del Porcentaje	62
Figura 10: Grafico Radial Analysis Services	63
Figura 11: Grafico Radial WEKA	63
Figura 12: ODBC's de usuario	69
Figura 13: Tabla BD Modelo Transaccional	70
Figura 14: Número registros por tablas y el número de campos por registro	71
Figura 15: Exploración de tablas seleccionadas para el datamart	75
Figura 16: Grafico Frecuencia CLIENTE	76
Figura 17: Grafico frecuencia DET_FACTURA	76
Figura 18: Grafico frecuencia FACTURA	77
Figura 19: Grafico frecuencia PRODUCTO	78
Figura 20: Datamart Ventas	83
Figura 21: Flujo del Job de Ventas	84
Figura 22: Workflow del DSA	85
Figura 23: Dataflow tabla CLIENTE	85
Figura 24: Dataflow tabla BODEGA	86

Figura 25: Dataflow tabla PRODUCTO	86
Figura 26: Dataflow tabla FACTURA	87
Figura 27: Dataflow tabla DETALLE FACTURA	87
Figura 28: Workflow Carga Dimensiones	88
Figura 29: Dataflow tabla DIM_FECHA	88
Figura 30: Dataflow tabla DIM_CLIENTE.....	89
Figura 31: Dataflow tabla DIM_BODEGA.....	89
Figura 32: Dataflow tabla DIM_PRODUCTO.....	90
Figura 33: Workflow Carga tablas de Hechos.....	90
Figura 34: Dataflow tabla FAC_VENTA.....	91
Figura 35: Dataflow tabla FAC_VENTA_DETALLE.....	91
Figura 36: Dataflow tabla FACTURA VENTA HISTORICA	92
Figura 37: Tablas del Datamart	92
Figura 38: Vistas Análisis RFM	93
Figura 39: Discretización Valores RFM.....	93
Figura 40: Vista de valores de Recencia	94
Figura 41: Vista de valores de Frecuencia.....	94
Figura 42: Vista de valor Monetario.....	94
Figura 43: Integración Datos con Análisis RFM.....	95
Figura 44: Columnas y Filas Transacciones Realizadas	96
Figura 45: Flujo Conocimiento Segmentación o Clustering	98
Figura 46: Flujo Conocimiento Asociación.....	98
Figura 47: Resultado de Error Segmentación.....	99
Figura 48: Regla N° 20 Asociacion	100
Figura 49: Parte 1-1 Clusterización	100

Figura 50: Parte 1-2 Clusterización	101
Figura 51: Porcentaje de Clusters	101
Figura 52: Asociación Productos.....	102
Figura 53: Dato Error obtenido al discretizar valor RFM_puntos.....	103
Figura 54: Ponderación Empírica de tiempo por fases en Minería de Datos-Recolección de Datos Iniciales	113
Figura 55: Ponderación Empírica de tiempo por fases en Minería de Datos- Descripción de los datos	114
Figura 56: Ponderación Empírica de tiempo por fases en Minería de Datos- Exploración de los datos	115
Figura 57: Ponderación Empírica de tiempo por fases en Minería de Datos- Verificación de la calidad de los datos	116
Figura 58: Ponderación Empírica de tiempo por fases en Minería de Datos- Selección de datos.....	117
Figura 59: Ponderación Empírica de tiempo por fases en Minería de Datos- Limpieza de datos.....	118
Figura 60: Ponderación Empírica de tiempo por fases en Minería de Datos- Estructuración de los datos	119
Figura 61: Ponderación Empírica de tiempo por fases en Minería de Datos- Integración de los datos	120
Figura 62: Ponderación Empírica de tiempo por fases en Minería de Datos- Formateo de los datos	121
Figura 63: Ponderación Empírica de tiempo por fases en Minería de Datos- Selección de la técnica de modelado	123

Figura 64: Ponderación Empírica de tiempo por fases en Minería de Datos- Construcción del modelo	124
Figura 65: Ponderación Empírica de tiempo por fases en Minería de Datos- Evaluación del modelo	125
Figura 66: Ponderación Empírica de tiempo por fases en Minería de Datos- Informe Final.....	126
Figura 67: Ponderación Empírica de tiempo de Minería de Datos.....	132
Figura 68: Ambiente de Minería (Job_Ventas)	140
Figura 69: Ambiente de Minería (QY_DSA_BODEGA)	140
Figura 70: Ambiente de Minería (QY_DSA_CLIENTE)	141

RESUMEN

El presente trabajo tiene como objetivo realizar un Análisis Comparativo de las Plataformas Weka y Microsoft Analysis Services para optimizar el desarrollo de Minería de Datos en la Empresa Prasol “Lácteos Santillán”. Este análisis es realizado a través de la creación de indicadores tomados de la conceptualización de las fases y tareas de la Metodología CRISP-DM, obteniendo como resultado que la mejor herramienta de minería de datos es Weka superando a Microsoft Analysis Services con un 7%, según los indicadores planteados enfocados a la optimización del proceso de minería con la metodología CRISP-DM, estadísticamente esta superioridad no es muy significativa.

La elaboración de la Minería de Datos se realizó aplicando la metodología CRISP-DM y utilizando la herramienta Weka, para ello se realizó un Datamart con la herramienta SAP Data Services Designer. Se genera un análisis RFM a partir de los datos dimensionales, y haciendo uso de la herramienta Weka se realiza la segmentación de clientes y asociación de productos, generando un informe enfocado al incremento de la rentabilidad de la Empresa.



UNIVERSIDAD NACIONAL DE CHIMBORAZO
FACULTAD DE INGENIERIA
CENTRO DE IDIOMAS



Msc. Janeth Caisaguano

10, de Marzo del 2016

SUMMARY

This paper aims to make a comparative analysis of the Weka Platforms and Microsoft Analysis Services to optimize the development of Data Mining Prasol Lácteos Santillán Company. This analysis was done through the creation of indicators from the conceptualization of the phases and tasks of the CRISP-DM methodology, resulting as the best tool of data mining than Weka beating Microsoft Analysis Services at 7%, according with proposed indicators focusing on optimization of mining process with CRISP-DM methodology, but this superiority is not statistically significant.

The development of data mining was conducted using the CRISP-DM methodology and using the Weka tool, for this, a Datamart was performed with the SAP Data Services Designer tool. An RFM analysis was generated from the dimensional data, and using the Weka customer segmentation tool and product association was made, generating a report focused on the Company profitability increasing.

x



INTRODUCCIÓN

Actualmente la mayoría de las empresas constan con bases de datos para varios tipos de funciones, como ventas, llevar un control, etc., pero es necesario realizar métodos y herramientas especiales para analizar grandes cantidades de datos.

Gracias a estas herramientas permiten a las empresas tomar mejores decisiones de negocios, se denominan Business Intelligence (BI), las cuales permiten el análisis multidimensional y minería de datos.

La minería de datos está orientada al descubrimiento de información, una de las características y usos más importantes es que se la utiliza para realizar campañas de marketing, ya que mediante el proceso de minería de datos (DM) se puede realizar predicciones, análisis para cada tipo de cliente que una empresa quiera llegar o clientes que las empresas saben que son rentables, este tipo de campañas se dan según la necesidad, gustos, etc., además se le ofrecerá descuentos o distintos tipos de productos, gracias a esto, las empresas mejoraran sus ventas.

En el mercado existen distintas herramientas para la realización de la minería de datos, entre ellas están Analysis Service y Weka.

Analysis Service es una herramienta de Microsoft, su licencia es pagada, y es muy utilizada en el entorno de minería de datos. Una de las ventajas es que contiene una solución ETL completa además ayuda a la toma de decisiones en tiempo real.

Por el otro Weka es una herramienta creada por la Universidad de Waikato, y se encuentra libremente disponible bajo la licencia pública general de GNU. Una de las ventajas es que proporciona interfaces para la comunicación con el usuario tal es el caso de CLI (Simple Client) y además contiene una extensa colección de técnicas para pre-procesamiento de datos y modelado.

En el caso particular de la Empresa Prasol “Lácteos Santillán”, mediante la minería de datos se ha querido mejorar las ventas, y gracias a esto adquirir más clientes para así potencializarse en el medio.

La presente investigación se compone de VII capítulos:

- En el Capítulo I inicia con un marco referencial, seguido de la problematización de la investigación.
- En el Capítulo II se sustenta teóricamente el presente trabajo con toda la información necesaria y complementaria sobre las plataformas WEKA y Analysis Services, Microsoft SQL Server, Minerías de Datos, Business Intelligence, Herramientas ETL, Datamart, Datawarehouse, Sap Data Service y Windows Server 2008.
- En el Capítulo III se plantean la metodología y procedimientos a seguirse en la investigación.
- En el capítulo IV elaboramos el análisis comparativo de las plataformas WEKA y Analysis Services, mediante la metodología CRISP-DM, la cual nos permitió cumplir con el objetivo de nuestra investigación.
- En el capítulo V desarrollamos la minería de datos mediante la metodología CRISP-DM sobre Data Mining, la cual se realizó en distintas fases y sub-fases, las fases principales son: Fase de comprensión del negocio, Fase de comprensión de los datos, Fase de preparación de los datos, Fase de modelado, Fase de evaluación y la Fase de implementación.
- En el Capítulo VI se muestran los resultados de la investigación realizada, se comprueba la hipótesis planteada en la investigación y su respectiva discusión.
- El Capítulo VII se finaliza la investigación emitiendo las conclusiones y recomendaciones.

CAPÍTULO I

1.1 TITULO DEL PROYECTO

Análisis Comparativo de las Plataformas Weka y Microsoft Analysis Services para optimizar el desarrollo de Minería de Datos en la Empresa Prasol “Lácteos Santillán”.

1.2 PROBLEMATIZACIÓN

1.2.1 IDENTIFICACIÓN Y DESCRIPCIÓN DEL PROBLEMA

Weka es una herramienta para los análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades las cuales se manifiestan a través de una colección de herramientas de visualización y algoritmos.

La versión original de Weka fue desarrollada en la plataforma TCL/TK y C en el año 1993, unos años más tarde en el año 1997 el código se decidió reformularlo para ser escrito bajo la plataforma de Java.

Microsoft Analysis Services proporciona los datos analíticos empleados en informes y aplicaciones cliente como Excel, informes de Reporting Services y otras herramientas de BI de terceros.

Use Analysis Services para crear estructuras de consulta de alto rendimiento, lógica de negocio y KPI dentro de un modelo de datos con varios fines al que puede acceder cualquier aplicación cliente que admita Analysis Services como origen de datos.

La Empresa Prasol “Lácteos Santillán”, actualmente se encuentra en crecimiento continuo, y como objetivo principal es llegar a más consumidores para que sus productos sean aprovechados por los mismos, al momento cuenta con su base de datos, y mediante la minería de datos se podrá optimizar procesos y mejorar sus ventas.

1.2.2 ANÁLISIS CRÍTICO

La creación de las bases de datos han permitido que la información de muchas entidades a nivel mundial se encuentre digitalizada, respaldada, lo que produce un efecto positivo en el desarrollo de cada una de sus actividades, estos tipos de entidades ya sean financieras, industrias, pero en la actualidad conforme con el avance tecnológico se requiere que la información este optimizada, más segura y respaldada, sin embargo posee ciertas debilidades que pueden disminuir el rendimiento de las mismas ya que puede haber duplicación de datos y falta de seguridad hacia la misma.

1.2.3 PROGNOSIS

La implementación de la plataforma Weka o Microsoft Analysis Services facilitara el desarrollo de la minería de datos en la Empresa Prasol “Lácteos Santillán” Además se manejará los datos de una manera más óptima y eficaz, con esto afectara positivamente el rendimiento de la empresa y aumentara sus ventas.

1.2.4 DELIMITACIÓN

El tema de investigación será analizado y desarrollado de acuerdo a los requerimientos de la Empresa Prasol “Lácteos Santillán”, ya que necesita optimizar sus procesos y mejorar sus ventas para poder obtener una mayor ganancia.

1.2.5 FORMULACIÓN DEL PROBLEMA

¿Cómo el análisis comparativo de las plataformas Weka y Microsoft Analysis Services, optimizara el desarrollo de minerías de datos en la Empresa Prasol “Lácteos Santillán”?

1.2.6 HIPÓTESIS

El análisis comparativo de las plataformas WEKA y MICROSOFT ANALYSIS SERVICES, determinará la mejor opción para optimizar el desarrollo de minería de datos en la empresa PRASOL “LÁCTEOS SANTILLÁN”

1.2.7 IDENTIFICACIÓN DE VARIABLES

1.2.7.1 VARIABLE DEPENDIENTE

Desarrollo de Minería de Datos

1.2.7.2 VARIABLE INDEPENDIENTE

Análisis comparativo de plataformas de minería de datos

1.2.8 OBJETIVOS

1.2.8.1 GENERAL

Realizar un Análisis Comparativo de las Plataformas Weka y Microsoft Analysis Services para optimizar el Desarrollo de Minería de Datos.

1.2.8.2 ESPECÍFICOS

- Estudiar las funcionalidades de las plataformas Weka y Microsoft Analysis Services.
- Realizar un análisis comparativo de las herramientas en base al escenario presentado del problema, para la minería de datos.
- Desarrollar una solución de minería de datos en la Empresa Prasol “Lácteos Santillán”, para mejorar las ventas de la misma.

1.2.9 JUSTIFICACIÓN

La minería de datos es un campo de las ciencias de la computación dirigido al proceso que pretende hallar patrones en grandes volúmenes de cantidad de datos. La minería de datos utiliza los métodos de inteligencia artificial y sistemas de base de datos.

Weka es un software de código abierto publicado bajo la licencia GNU (General Public License). para aprendizaje automático y minería de datos implementado en Java.

Analysis Services es un motor de datos que se usa en soluciones de ayuda a la toma de decisiones y Business Intelligence (BI).

Su objetivo se basa en la extracción de información de un conjunto de datos y transformarla en una estructura comprensible para los usuarios, lo cual posibilita organización y control dentro de grandes empresas.

La minería de datos influye en las empresas en la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisiones y puede contribuir significativamente en las aplicaciones de administración empresarial basada en la relación con el cliente. En lugar de contactar con el cliente de forma indiscriminada a través de un centro de llamadas o enviando cartas, sólo se contactará con aquellos que se perciba que tienen una mayor probabilidad de responder positivamente a una determinada oferta o promoción.

De tal manera el motivo por lo que nace la necesidad de desarrollar la minería de datos es que la empresa necesita incrementar sus ventas continuamente, y gracias a la minería de datos se llevara un control más personalizado en sus ventas así como las grandes multinacionales, llevando a un estudio de mercadeo más profundo.

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA

2.1 ANTECEDENTES

En lo que se refiere al análisis comparativo entre las herramientas Weka y Microsoft Analysis Service con respecto a la minería de datos dentro de la Empresa Prasol “Lácteos Santillán”, no existe ningún tipo de investigación o implementación de la misma que se haya realizado anteriormente.

2.2 MINERÍA DE DATOS¹

La minería de datos es el proceso que tiene como propósito descubrir, extraer y almacenar información relevante de amplias bases de datos, a través de programas de búsquedas e identificación de patrones y relaciones globales, tendencias, desviaciones y otros indicadores aparentemente caóticos que tienen una explicación que pueden descubrirse mediante diversas técnicas de esta herramienta..

El objetivo fundamental es aprovechar el valor de la información localizada y usar los patrones preestablecidos para que los directivos tengan un mejor conocimiento de su negocio y puedan tomar decisiones más confiables.

2.2.1 TÉCNICAS DE EXTRACCIÓN DE CONOCIMIENTO²

2.2.1.1 TÉCNICAS DESCRIPTIVAS

- Segmentación de Datos
- Clasificación

¹ LARFHETA, M. I. Á. Minería de datos: Concepto.

² Román, J. V., García, R. M. C., Rueda, J. J. G. (23/12/2011). 07 Minería de Datos. Obtenido el 10/03/2016, desde el sitio Web de OCW - UC3M: <http://ocw.uc3m.es/ingenieria-telematica/inteligencia-en-redes-de-comunicaciones/material-de-clase-1/07-mineria-de-datos>.

- Análisis de Asociaciones

2.2.1.2 TÉCNICAS PREDICTIVAS

- Análisis de patrones secuenciales
- Análisis de similitud en series temporales
- Predicción

2.3 MINERÍA DE DATOS COMO SOPORTE A LA TOMA DE DECISIONES³

El uso de la minería de datos como soporte a decisiones en los negocios es más que aplicar redes neuronales o árboles de decisión sobre los datos por un lado está el descubrimiento del conocimiento en la base de datos y por otro lado están las técnicas estadísticas como el reconocimiento de patrones y algoritmos de aprendizaje entre otros

Los datos tal y como se almacenan en las bases de datos no suelen proporcionar beneficios directos , el valor está en la información que podamos extraer de ellos , que es la información que nos ayuda en la toma de decisiones o mejorar la comprensión del entorno que nos rodea, como puede ser la comprobación de que todo va bien , analizar diferentes aspectos de la evolución de la empresa , comparar información en diferentes periodos de tiempo , comparar resultados con previsiones, para ello se tienen que definir medidas cualitativas para los patrones obtenidos como son la precisión , utilidad y beneficio obtenido.

³ Aular, Y. J. M., & Pereira, R. T. (2007). Minería de datos como soporte a la toma de decisiones empresariales. *Opción*, 23(52).

2.4 BUSINESS INTELLIGENCE

Se denomina inteligencia empresarial, inteligencia de negocios o BI al conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa. El Data Mining proporciona un enorme valor a las organizaciones.⁴

- Gran cantidad de data disponible: las organizaciones llegaron a implementar sistemas transaccionales (ventas, almacenes, producción, personal, contabilidad, etc.) y estos en el tiempo han ido almacenando información. Aunado a la baja de los costos de almacenamiento han acumulado grandes volúmenes de datos.
- Alto nivel de competencia: la competencia actualmente es alta como resultado de marketing moderno y canales de distribución como internet y comunicaciones, así como la participación de corporaciones nacionales y extranjeras en el mercado.
- Tecnología Lista: el DM anteriormente era mayormente una solución de laboratorio, ahora ya es una tecnología madura y está lista para ser aplicada en las organizaciones.⁵

2.5 HERRAMIENTAS ETL⁶

El Sistema E.T.L (Extracción - Transformación - Carga) o E.T.T (Extracción - Transformación -Transporte).

ETL es el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas necesarias para mover datos desde

⁴ <http://repositorio.espe.edu.ec/bitstream/21000/6305/1/T-ESPE-047033.pdf>

⁵ <http://f1-preview.runhosting.com/premiunnet.com/DataMining.pdf>

⁶ https://docs.google.com/presentation/d/1D9e8nt0C_rMwhdCHiimXrg9FDsTMOq8XYeREsUMAc/embed?slide=id.i0

múltiples fuentes a un almacén de datos, reformatearlos, limpiarlos y cargarlos en otra base de datos, data mart o bodega de datos. ETL forma parte de la Inteligencia Empresarial (Business Intelligence), también llamado “Gestión de los Datos” (Data Management).

2.6 DATAMART⁷

Un Datamart es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. Un datamart puede ser alimentado desde los datos de un datawarehouse, o integrar por sí mismo un compendio de distintas fuentes de información.

Por tanto, para crear el datamart de un área funcional de la empresa es preciso encontrar la estructura óptima para el análisis de su información, estructura que puede estar montada sobre una base de datos OLTP, como el propio datawarehouse, o sobre una base de datos OLAP. La designación de una u otra dependerá de los datos, los requisitos y las características específicas de cada departamento. De esta forma se pueden plantear dos tipos de datamarts:

2.6.1 DATAMART OLAP

Se basan en los populares cubos OLAP, que se construyen agregando, según los requisitos de cada área o departamento, las dimensiones y los indicadores necesarios de cada cubo relacional. El modo de creación, explotación y mantenimiento de los cubos OLAP es muy heterogéneo, en función de la herramienta final que se utilice.

⁷ http://www.sinnexus.com/business_intelligence/datamart.aspx

2.6.2 DATAMART OLTP

Pueden basarse en un simple extracto del datawarehouse, no obstante, lo común es introducir mejoras en su rendimiento (las agregaciones y los filtrados suelen ser las operaciones más usuales) aprovechando las características particulares de cada área de la empresa. Las estructuras más comunes en este sentido son las tablas report, que vienen a ser fact-tables reducidas (que agregan las dimensiones oportunas), y las vistas materializadas, que se construyen con la misma estructura que las anteriores, pero con el objetivo de explotar la reescritura de Queries (aunque sólo es posibles en algunos SGBD avanzados, como Oracle).

Los datamarts que están dotados con estas estructuras óptimas de análisis presentan las siguientes ventajas:

- Poco volumen de datos
- Mayor rapidez de consulta
- Consultas SQL y/o MDX sencillas
- Validación directa de la información
- Facilidad para la historización de los datos

2.7 DATAWAREHOUSE⁸

Un Datawarehouse es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta. La creación de un datawarehouse representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Business Intelligence.

⁸ http://www.sinnexus.com/business_intelligence/datawarehouse.aspx

La ventaja principal de este tipo de bases de datos radica en las estructuras en las que se almacena la información (modelos de tablas en estrella, en copo de nieve, cubos relacionales... etc.). Este tipo de persistencia de la información es homogénea y fiable, y permite la consulta y el tratamiento jerarquizado de la misma (siempre en un entorno diferente a los sistemas operacionales).

El término Datawarehouse fue acuñado por primera vez por Bill Inmon, y se traduce literalmente como almacén de datos. No obstante, y como cabe suponer, es mucho más que eso. Según definió el propio Bill Inmon, un datawarehouse se caracteriza por ser:

- **Integrado:** Los datos almacenados en el datawarehouse deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.
- **Temático:** Sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, todos los datos sobre clientes pueden ser consolidados en una única tabla del datawarehouse. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.
- **Histórico:** El tiempo es parte implícita de la información contenida en un datawarehouse. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el datawarehouse sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el datawarehouse se carga con los

distintos valores que toma una variable en el tiempo para permitir comparaciones.

- **No volátil:** El almacén de información de un datawarehouse existe para ser leído, pero no modificado. La información es por tanto permanente, significando la actualización del datawarehouse la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

2.8 DIFERENCIA ENTRE DATAWAREHOUSE & DATAMART⁹

DataWharehouse:

- Posee múltiples áreas temáticas
- Sostiene información muy detallada
- Trabaja para integrar todas las fuentes de datos
- No utiliza necesariamente un modelo tridimensional, pero alimenta modelos dimensionales.

Data Mart:

- A menudo ocupa sólo una zona- sujeta por ejemplo, Finanzas, o Ventas
- Pueden contener datos más resumidos (aunque muchos de ellos tienen todo detalle)
- Se concentra en la integración de la información de una determinada disciplina o conjunto de sistemas de código
- Se construye centrado en un modelo tridimensional mediante un esquema en estrella.

⁹ <http://www.datamartist.com/data-warehouse-vs-data-mart>

2.9 PLATAFORMA WEKA¹⁰

Weka se trata de un acrónimo derivado de Waikato Environment for Knowledge Analysis – Entorno para Análisis del Conocimiento de la Universidad de Waikato. Esto es porque fue esta universidad la que inició el desarrollo de Weka en 1993, no obstante, su desarrollo original fue hecho en TCL/TK y C, para en 1997 pasar a reescribirse todo el código fuente del entorno en Java, una plataforma más universal, y añadir las implementaciones de diferentes algoritmos de modelado.



Figura 1: Weka GUI Chooser

Fuente: (Programa WEKA, 2010)

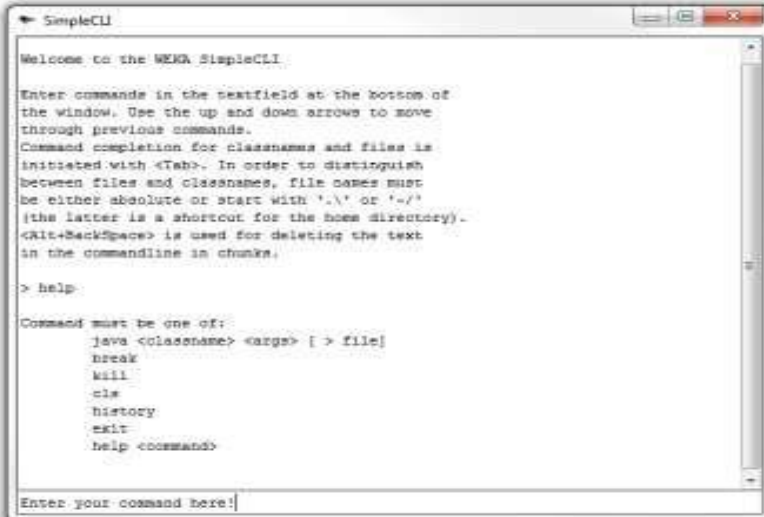
Weka está compuesta por una serie de herramientas gráficas de visualización y diferentes algoritmos para el análisis de datos y modelado predictivo. Su interfaz gráfica de usuario nos facilita el acceso a sus múltiples funcionalidades.

Esta potente herramienta de minería de datos se encuentra libremente disponible bajo la licencia pública general de GNU, además, al estar implementada en Java como ya hemos comentado, puede ejecutarse prácticamente bajo cualquier entorno.

¹⁰<http://hdl.handle.net/10251/10097>

La interfaz gráfica de Weka cuenta con 4 formas de acceso a las diferentes funcionalidades de la aplicación.

- Simple CLI (Simple command-line interface), que no es más que el acceso a través de consola de comandos a todas las opciones de Weka.



```
SimpleCLI
Welcome to the WEKA SimpleCLI.

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with './' or './'
(the latter is a shortcut for the home directory).
<Alt+Backspace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
  java <classname> <args> [ > file]
  break
  kill
  cls
  history
  exit
  help <command>

Enter your command here:|
```

Figura 2: WEKA Simple CLI

Fuente: (Programa WEKA, 2010)

- Explorer, es la opción más intuitiva para el usuario, pues dispone de varios paneles que dan acceso a las principales características del programa

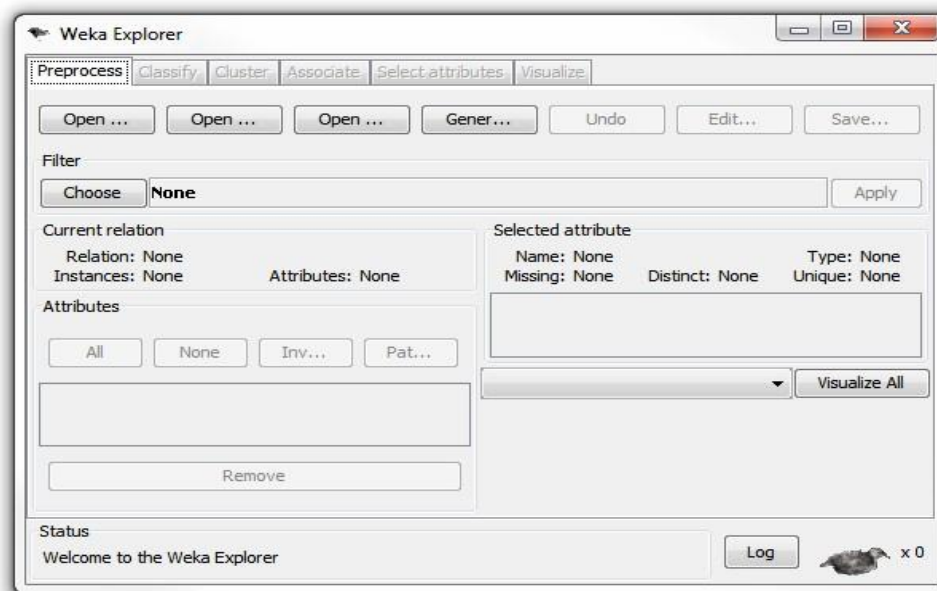


Figura 3: WEKA Explorer

Fuente: (Programa WEKA, 2010)

- Experimenter, permite la comparación sistemática de una ejecución de los algoritmos predictivos de Weka sobre una colección de conjuntos de datos.

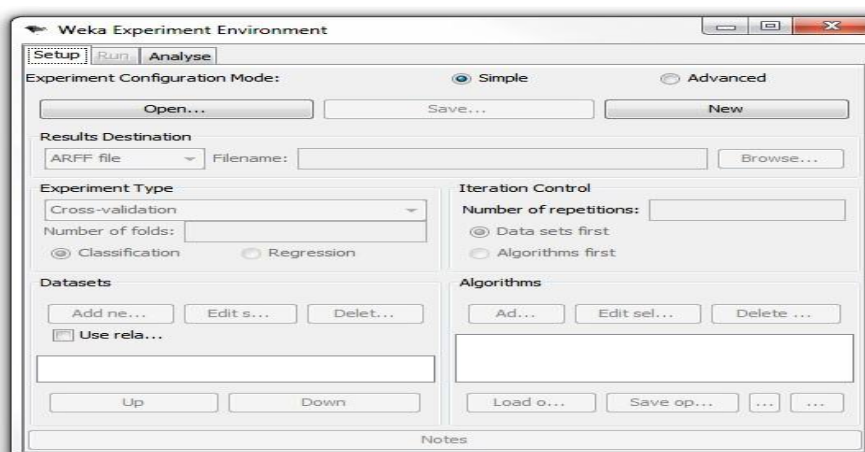


Figura 4: WEKA Experiment Environment

Fuente: (Programa WEKA, 2010)

- Knowledge Flow, soporta esencialmente las mismas opciones que la interfaz Explorer, pero esta permite “arrastrar y soltar”. Ofrece soporte para el aprendizaje incremental.



Figura 5: WEKA KnowledgeFlow Environment

Fuente: (Programa WEKA, 2010)

2.9.1 PUNTOS FUERTES DE WEKA¹¹

Weka soporta varias tareas estándar de minería de datos, especialmente, pre procesamiento de datos, clustering, clasificación, regresión, visualización, y selección. Todas las técnicas de Weka se fundamentan en la asunción de que los datos están disponibles en un fichero plano (flat file) o una relación, en la que cada registro de

¹¹<http://www.itsciudadserdan.edu.mx/Articulos%20Investigacion/WEKA%20COMO%20HERRAMIENTA%20DE%20DATA%20MINING/WEKA%20COMO%20HERRAMIENTA%20DE%20DATA%20MINING.pdf>

datos está descrito por un número fijo de atributos (normalmente numéricos o nominales, aunque también se soportan otros tipos). Weka también proporciona acceso a bases de datos vía SQL gracias a la conexión JDBC (Java Database Connectivity) y puede procesar el resultado devuelto por una consulta hecha a la base de datos. No puede realizar minería de datos multi-relacional, pero existen aplicaciones que pueden convertir una colección de tablas relacionadas de una base de datos en una única tabla que ya puede ser procesada con Weka.⁷

2.9.2 VENTAJAS & DESVENTAJAS¹²

2.9.2.1 VENTAJAS

- Weka proporciona interfaces para la comunicación con el usuario, tal es el caso de CLI (Simple Client), esta interfaz proporciona una consola para poder introducir mandatos, posee una apariencia muy simple pero nos permite realizar tareas complejas ya que permite realizar cualquier operación soportada por Weka de forma directa; no obstante, es muy complicada de manejar ya que es necesario un conocimiento completo de la aplicación.
- Nos permite ubicar patrones de comportamiento de la información a procesar de tal manera que es de gran ayuda en la toma de decisiones.
- Está disponible libremente bajo la licencia pública general de GNU.
- Es muy portable porque está completamente implementado en Java y puede correr en casi cualquier plataforma.
- Contiene una extensa colección de técnicas para pre-procesamiento de datos y modelado.

¹²<http://documents.mx/documents/weka55cf9ab9550346d033a31732.html>

2.9.2.2 DESVENTAJAS

- Existe poca documentación sobre el uso de Weka dirigida al usuario.
- Un área importante que actualmente no cubren los algoritmos incluidos en Weka es el modelado de secuencia.

2.10 MICROSOFT ANALYSIS SERVICES¹³

Microsoft Analysis Services proporciona una plataforma integrada para las soluciones que incorporan la minería de datos. Puede usar datos relacionales o de cubo para crear soluciones de Business Intelligence con análisis predictivos.

Microsoft SQL Server 2014 Analysis Services (SSAS) ofrece funciones de procesamiento analítico en línea (OLAP) y minería de datos para aplicaciones de Business Intelligence. Analysis Services admite OLAP y permite diseñar, crear y administrar estructuras multidimensionales que contienen datos agregados desde otros orígenes de datos, como bases de datos relacionales. En el caso de las aplicaciones de minería de datos, Analysis Services permite diseñar, crear y visualizar modelos de minería de datos que se construyen a partir de otros orígenes de datos mediante el uso de una gran variedad de algoritmos de minería de datos estándar del sector.

2.10.1 TIPOS DE ALGORITMOS¹⁴

Analysis Services incluye los siguientes tipos de algoritmos:

- Algoritmos de clasificación, que predicen una o más variables discretas, basándose en otros atributos del conjunto de datos.
- Algoritmos de regresión, que predicen una o más variables continuas, como las pérdidas o los beneficios, basándose en otros atributos del conjunto de datos.

¹³<http://msdn.microsoft.com/es-es/library/bb510516.aspx>

¹⁴[https://msdn.microsoft.com/es-es/library/ms175595\(v=sql.120\).aspx](https://msdn.microsoft.com/es-es/library/ms175595(v=sql.120).aspx)

- Algoritmos de segmentación, que dividen los datos en grupos, o clústeres, de elementos que tienen propiedades similares.
- Algoritmos de asociación, que buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación, que pueden usarse en un análisis de la cesta de compra.
- Algoritmos de análisis de secuencias, que resumen secuencias o episodios frecuentes en los datos, como un flujo de rutas web.

2.10.2 CARACTERÍSTICAS Y TAREAS DE MICROSOFT ANALYSIS SERVICES¹⁵

La documentación básica de Analysis Services está organizada por modelado y modo de servidor para que pueda centrarse únicamente en las herramientas, tareas y características disponibles en el modo que tiene instalado. Las tareas de administración de servidores que abarcan varios modos se encuentran en instancias de servidor y son las siguientes:

- Comparar soluciones tabulares y multidimensionales (SSAS)
- Administración de una instancia de Analysis Services
- Modelado tabular (SSAS tabular)
- Modelado multidimensional (SSAS)
- Minería de datos (SSAS)
- PowerPivot para SharePoint (SSAS)

¹⁵<http://msdn.microsoft.com/es-es/library/bb510516.aspx>



Figura 6: Analysis Services

Fuente: (Microsoft developer network, 2014)

2.10.3 VENTAJAS & DESVENTAJAS

2.10.3.1 VENTAJAS¹⁶

- Se puede contar con la información consolidada en un solo visor, es decir toda la información que se necesite analizar se puede procesar para formar un "Cubo"
- Simplifica el proceso de creación de soluciones complejas con diversas capacidades de modelado.
- Utiliza el modelo semántico de BI para proporcionar un punto de vista de negocio consolidado de datos tabulares y multidimensionales.
- Utiliza almacenamiento en caché automático para proporcionar un rendimiento excelente de consulta.
- Disfruta de una solución de copia de seguridad escalable.
- Solución ETL completa

¹⁶ <http://librozilla.com/doc/78570/presentación---asteriscus.com>

- Ayuda a la decisión en tiempo real: informes, Data Mining
- Mejoras en escalabilidad y disponibilidad

2.10.3.2 DESVENTAJAS

- El nivel de dificultad al cruzar datos de varios CUBOS OLAP es muy alto
- Los EXCEL SERVICES no están al mismo nivel de rendimiento del resto de la herramienta.
- Gran cantidad de memoria RAM de uso, para la instalación y utilización del software.
- La relación precio con herramientas OpenSource.

2.11 METODOLOGÍA CRISP-DM (CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING)

Es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Data Mining.

Los orígenes de CRISP-DM, se remontan hacia el año 1999 cuando un importante consorcio de empresas europeas tales como NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra), OHRA (Holanda), Teradata, SPSS, y Daimler-Chrysler, proponen a partir de diferentes versiones de KDD (Knowledge Discovery in Databases) [Reinartz, 1995], [Adraans, 1996], [Brachman, 1996], [Fayyad, 1996], el desarrollo de una guía de referencia de libre distribución denominada CRISP-DM (Cross Industry Standard Process for Data Mining).

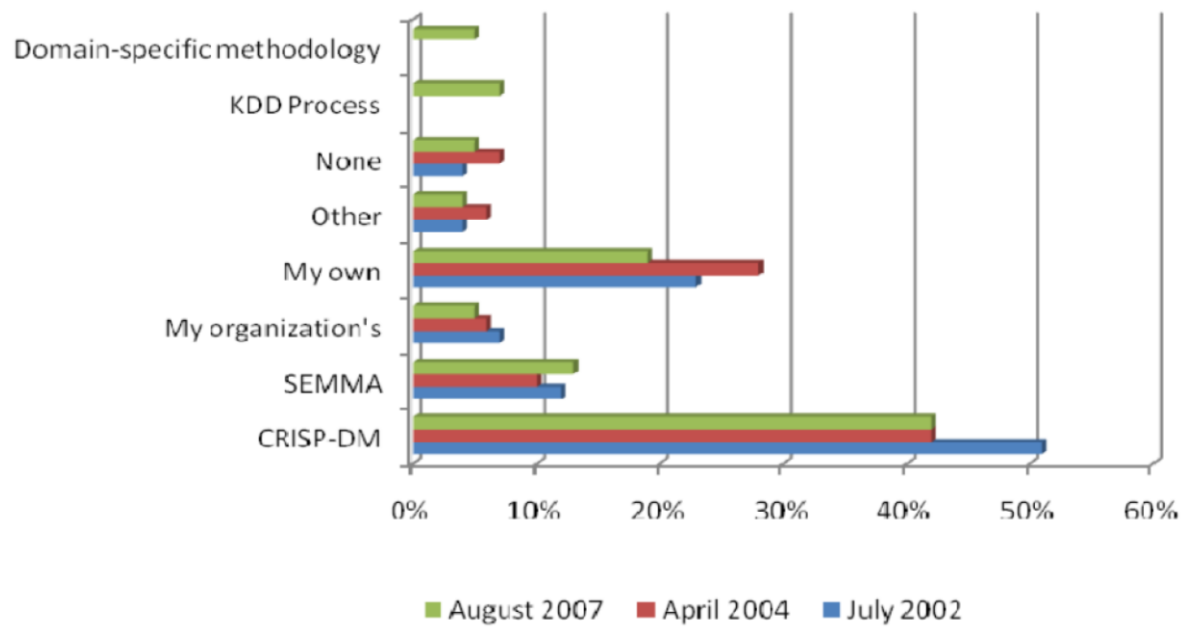


Figura 7: Metodologías de Data Mining

Fuente (Arancibia, J. A. G., 2010)

CAPÍTULO III

ANÁLISIS COMPARATIVO

3.1 ANÁLISIS COMPARATIVO DE HERRAMIENTAS MINERÍA DE DATOS

El presente análisis comparativo de las herramientas de minería de datos Microsoft Analysis Services y Weka se realiza en base de la conceptualización de las fases y sub fases de la metodología para proyectos de minería de datos CRISP-DM, con el objetivo de poder identificar si se optimiza o no el proceso de DM.

3.1.1 PROCEDIMIENTO PARA EL ANÁLISIS DE HERRAMIENTAS DE MINERÍA DE DATOS

Se propone el siguiente procedimiento para el análisis de las herramientas de minería de datos: Microsoft Analysis Services y Weka:

- 1) Conceptualización de las fases y sub fases de la Metodología CRISP-DM.
- 2) Identificación de las sub fases de la Metodología CRISP-DM que apliquen a la utilización de herramientas de software para minería de datos.
- 3) Creación de indicadores a partir de los conceptos de la sub fases de la Metodología CRISP-DM.
- 4) Análisis de las herramientas Microsoft Analysis Services y Weka para entender su funcionalidad y así poder comparar de acuerdo a los indicadores que se ha obtenido de la conceptualización de la Metodología CRISP-DM.
- 5) Análisis de resultados del estudio comparativo y elección de la mejor herramienta.

3.1.1.1 CONCEPTUALIZACIÓN DE LAS FASES Y SUBFASES DE LA METODOLOGÍA CRISP-DM

Es necesario definir a cada una de las sub fases con el objetivo de encontrar indicadores que permitan evaluar a las herramientas. La conceptualización de las Sub Fases de la Metodología CRISP-DM está basado en el trabajo de investigación “Metodología para la definición de requisitos en proyectos de Data Mining (ER-DM)” de José Alberto Gallardo Arancibia, esta se describe a continuación:

3.1.1.1.1. FASE DE COMPRENSIÓN DEL NEGOCIO O PROBLEMA

La primera fase de la guía de referencia CRISP-DM, denominada fase de comprensión del negocio o problema, es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto.

3.1.1.1.1.1. DETERMINAR LOS OBJETIVOS DEL NEGOCIO

Esta es la primera tarea a desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar Data Mining y definir los criterios de éxito.

3.1.1.1.1.2. EVALUACIÓN DE LA SITUACIÓN

En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de DM, se definen los requisitos del problema, tanto en términos de negocio como en términos de Data Mining.

3.1.1.1.1.3. DETERMINACIÓN DE LOS OBJETIVOS DE DM

Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de DM, como por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de DM será por ejemplo, determinar el perfil de los clientes respecto de su capacidad de endeudamiento.

3.1.1.1.1.4. PLAN DEL PROYECTO

Esta última tarea de la primera fase de CRISP-DM, tiene como meta desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada paso.

3.1.1.1.2. FASE DE COMPRENSIÓN DE LOS DATOS

Comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de DM. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos ad-hoc al proyecto de DM, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y abundantes accesos a la base de datos a objeto de realizar consultas y probablemente modificaciones, lo cual podría generar muchos problemas.

3.1.1.1.2.1. RECOLECCIÓN DE DATOS INICIALES

La primera tarea en esta segunda fase del proceso de CRISP-DM, es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo, elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.

3.1.1.1.2.2. DESCRIPCIÓN DE LOS DATOS

Después de adquiridos los datos iniciales, estos deben ser descritos. Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.

3.1.1.1.2.3. EXPLORACIÓN DE DATOS

A continuación, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución.

3.1.1.1.2.4. VERIFICACIÓN DE LA CALIDAD DE LOS DATOS

En esta tarea, se efectúan verificaciones sobre los datos, para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea en este punto, es asegurar la completitud y corrección de los datos.

3.1.1.1.3. FASE DE PREPARACIÓN DE LOS DATOS

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos.

3.1.1.1.3.1. SELECCIÓN DE DATOS

En esta etapa, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores: calidad de los datos en cuanto a completitud y corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de DM seleccionadas.

3.1.1.1.3.2. LIMPIEZA DE LOS DATOS

Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son: normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc. Estructuración de los datos. Esta tarea incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes

3.1.1.1.3.3. INTEGRACIÓN DE LOS DATOS

La integración de los datos, involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde

se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.

3.1.1.1.3.4. FORMATEO DE LOS DATOS

Esta tarea consiste principalmente, en la realización de transformaciones sintácticas de los datos sin modificar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de DM en particular, como por ejemplo la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.).

3.1.1.1.4. FASE DE MODELADO

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

Previamente al modelado de los datos, se debe determinar un método de evaluación de los modelos que permita establecer el grado de bondad de ellos. Después de concluir estas tareas genéricas, se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo, dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo.

3.1.1.1.4.1. SELECCIÓN DE LA TÉCNICA DE MODELADO

Esta tarea consiste en la selección de la técnica de DM más apropiada al tipo de problema a resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de DM existentes.

3.1.1.1.4.2. GENERACIÓN DEL PLAN DE PRUEBA

Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo. Por ejemplo, en una tarea supervisada de DM como la clasificación, es común usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.

3.1.1.1.4.3. CONSTRUCCIÓN DEL MODELO

Después de seleccionada la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

3.1.1.1.4.4. EVALUACIÓN DEL MODELO

En esta tarea, los ingenieros de DM interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos.

3.1.1.1.5. FASE DE EVALUACIÓN

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo.

3.1.1.1.5.1. EVALUACIÓN DE LOS RESULTADOS

En los pasos de evaluación anteriores, se trataron factores tales como la exactitud y generalidad del modelo generado. Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual, el modelo sea deficiente, o si es aconsejable probar el modelo, en un problema real si el tiempo y restricciones lo permiten.

3.1.1.1.5.2. PROCESO DE REVISIÓN

El proceso de revisión, se refiere a calificar al proceso entero de DM, a objeto de identificar elementos que pudieran ser mejorados.

3.1.1.1.5.3. DETERMINACIÓN DE FUTURAS FASES

Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría pasarse a la fase siguiente, en caso contrario podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. Podría ser incluso que en esta fase se decida a partir desde cero con un nuevo proyecto de DM.

3.1.1.1.6. FASE DE IMPLEMENTACIÓN

Una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso.

3.1.1.1.6.1. PLAN DE IMPLEMENTACIÓN

Para implementar el resultado de DM en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior implementación.

3.1.1.1.6.2. MONITORIZACIÓN Y MANTENIMIENTO

Si los modelos resultantes del proceso de Data Mining son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.

3.1.1.1.6.3. INFORME FINAL

Es la conclusión del proyecto de DM realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto.

3.1.1.1.6.4. REVISIÓN DEL PROYECTO

En este punto se evalúa qué fue lo correcto y qué lo incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar.

3.1.1.2 IDENTIFICACIÓN DE LAS SUBFASES DE LA METODOLOGÍA CRISP-DM QUE APLIQUEN A LA UTILIZACIÓN DE HERRAMIENTAS DE SOFTWARE PARA MINERÍA DE DATOS

La siguiente tabla proporciona información relacionada a cuál de las Sub Fases de la Metodología se pueden aplicar a la utilización de herramientas de Minería de Datos tales como las que se encuentra evaluando en este trabajo Microsoft Analysis Services y Weka.

Fases	Sub Fases	Aplica a herramienta de minería	Observaciones
Comprensión del Negocio o Problema	Determinar los objetivos del negocio	No aplica	No aplica debido a que es una actividad que debe desarrollar el personal del proyecto junto con los stakeholders.
	Evaluación de la situación	No aplica	Es una actividad de la minería de datos que no requiere una herramienta de software.
	Determinar los objetivos de la Minería de Datos	No aplica	No aplica debido a que es una actividad que debe desarrollar el personal del proyecto junto con los stakeholders.
	Realizar el plan del proyecto	No aplica	Es una tarea que deben desarrollar los miembros del proyecto de minería de datos

Fases	Sub Fases	Aplica a herramienta de minería	Observaciones
Comprensión de los Datos	Recolección de Datos iniciales	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
	Descripción de los Datos	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
	Exploración de los datos	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
	Verificación de la calidad de los datos	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
Fase de preparación de los datos	Selección de datos	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
	Limpieza de datos	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
	Estructuración de los datos	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
	Integración de los datos	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
	Formateo de los datos	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos

Fases	Sub Fases	Aplica a herramienta de minería	Observaciones
Fase de modelado	Selección de la técnica de modelado	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
	Generación del plan de prueba	No aplica	Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo. Esta tarea debe realizar el personal encargado de la minería de datos.
	Construcción del modelo	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
	Evaluación del modelo	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
Fase de evaluación	Evaluación de los resultados	No aplica	Esta tarea corresponde a personal de minería de datos debido a que es una evaluación relacionada a los objetivos del negocio
	Proceso de revisión	No aplica	El proceso de revisión, se refiere a calificar al proceso entero de DM, con el objetivo de identificar elementos que pudieran ser mejorados, esta tarea le

Fases	Sub Fases	Aplica a herramienta de minería	Observaciones
			corresponde al personal de minería de datos.
	Determinación de futuras fases	No aplica	Esta tarea corresponde a una decisión del personal de minería de datos.
Fase de implantación	Plan de implementación	No aplica	Esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación que es una tarea del personal de minería de datos.
	Monitorización y Mantenimiento	No aplica	La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente, esta tarea corresponde a personal de minería de datos
	Informe Final	Aplica	Se puede obtener indicadores enfocados a las herramientas de minería de datos
	Revisión del proyecto	No aplica	Esta actividad se basa en que requiere mejora, esta tarea corresponde a personal de minería de datos

Tabla 1: SUBFASES DE LA METODOLOGÍA CRISP-DM

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

3.1.1.3 CREACIÓN DE INDICADORES DE EVALUACIÓN DE HERRAMIENTAS DE MINERÍA DE DATOS

Se presenta los indicadores de evaluación de las herramientas de Minería de Datos, los mismos que han sido obtenidos de la conceptualización de las sub fases y del entendimiento de las mismas.

Fases	Sub Fases	Indicador
Comprensión de los Datos	Recolección de Datos iniciales	Permite la importación de Archivos de Texto (.txt, .csv)
		Permite la importación de Excel/spreadsheet
		Permite la importación de Tabla de base de datos
		Permite la importación de datos desde una URL
		Permite la importación de Otras fuentes de datos
	Descripción de los datos	Permite la visualización previa de la información
		Permite obtener el número de registros
		Permite obtener el número de Campos
		Permite la identificación de los campos
		Permite la descripción del tipo inicial de los campos
	Exploración de los datos	Permite la creación de tablas de frecuencia
		Permite la creación de gráficos de distribución estadística
		Permite obtener la cantidad de valores nulos
	Verificación de la calidad de los datos	Permite encontrar valores erróneos
Fase de preparación de los datos	Selección de datos	Permite seleccionar un subconjunto de los datos adquiridos de acuerdo a la verificación de la calidad de los datos
	Limpieza de datos	Permite la normalización de los datos
		Permite la discretización de campos numéricos
		Permite el tratamiento de valores ausentes
		Permite la reducción del volumen de datos

Fases	Sub Fases	Indicador	
	Estructuración de los datos	Permite generar nuevos atributos a partir de los ya existentes	
		Permite integrar nuevos registros	
		Permite la transformación de valores para atributos ya existentes	
	Integración de los datos	Permite la generación de nuevos campos a partir de otros existentes	
		Permite la creación de nuevos registros	
		Permite la fusión de tablas campos o nuevas tablas	
	Formateo de los datos	Permite la reordenación de los registros	
		Permite la reordenación de los campos	
	Fase de modelado	Selección de la técnica de modelado	Permite utilizar algoritmo/s de Árboles de Decisión
Permite utilizar algoritmo/s para Clustering			
Permite utilizar algoritmo/s para Series Temporales			
Permite utilizar algoritmo/s Clustering de Secuencia			
Permite utilizar algoritmo/s para Asociación			
Permite utilizar algoritmo/s para Redes Bayesianas			
Permite utilizar algoritmo/s de Redes Neuronales			
Construcción del modelo		Permite la construcción gráfica del modelo	
		Permite la ejecución del modelo con una interfaz gráfica	
		Permite la creación de un flujo gráfico del modelo	
Evaluación del modelo		Permite evaluar el performance del modelo	
		Genera gráficos de evaluación	
		Permite analizar el costo beneficio del modelo	
Fase de implantación		Informe Final	Permite crear gráficos que ilustren los resultados obtenidos

Tabla 2: CREACIÓN DE INDICADORES DE EVALUACIÓN DE HERRAMIENTAS DE MINERÍA DE DATOS

Fuente:(Nay Mojarrango & José Chapalbay, 2015)

3.1.1.4 ANÁLISIS COMPARATIVO DE LAS HERRAMIENTAS MICROSOFT ANALYSIS SERVICES Y WEKA

En esta sección se muestra una evaluación de las herramientas Microsoft Analysis Services y Weka, con indicadores enfocados al proceso de minería de datos, basados en la conceptualización de las Sub fases de la metodología CRISP-DM.

Fases	Sub Fases	Indicador	Analysis Services	Weka
Comprensión de los Datos	Recolección de Datos iniciales	Permite la importación de Archivos de Texto (.txt, .csv)	1	1
		Permite la importación de Excel/spreadsheet	1	0
		Permite la importación de Tabla de base de datos	1	1
		Permite la importación de datos desde una URL	1	1
		Permite la importación de Otras fuentes de datos	1	1
	Descripción de los datos	Permite la visualización previa de la información	1	1
		Permite obtener el número de registros	1	1
		Permite obtener el número de Campos	1	1
		Permite la identificación de los campos	1	1
		Permite la descripción del tipo inicial de los campos	1	1

Fases	Sub Fases	Indicador	Analysis Services	Weka
	Exploración de los datos	Permite la creación de tablas de frecuencia	0	1
		Permite la creación de gráficos de distribución estadística	0	1
	Verificación de la calidad de los datos	Permite obtener la cantidad de valores nulos	1	1
		Permite encontrar valores erróneos	1	1
Fase de preparación de los datos	Selección de datos	Permite seleccionar un subconjunto de los datos adquiridos de acuerdo a la verificación de la calidad de los datos	1	1
	Limpieza de datos	Permite la normalización de los datos	0	1
		Permite la discretización de campos numéricos	1	1
		Permite el tratamiento de valores ausentes	1	1
		Permite la reducción del volumen de datos	0	1
	Estructuración de los datos	Permite generar nuevos atributos a partir de los ya existentes	1	1
		Permite integrar nuevos registros	1	1

Fases	Sub Fases	Indicador	Analysis Services	Weka
	Integración de los datos	Permite la transformación de valores para atributos ya existentes	1	1
		Permite la generación de nuevos campos a partir de otros existentes	1	1
		Permite la creación de nuevos registros	0	0
		Permite la fusión de tablas campos o nuevas tablas	1	1
	Formateo de los datos	Permite la reordenación de los registros	1	1
		Permite la reordenación de los campos	1	1
Fase de modelado	Selección de la técnica de modelado	Permite utilizar algoritmo/s de Árboles de Decisión	1	1
		Permite utilizar algoritmo/s para Clustering	1	1
		Permite utilizar algoritmo/s para Series Temporales	1	1
		Permite utilizar algoritmo/s Clustering de Secuencia	1	1
		Permite utilizar algoritmo/s para Asociación	1	1
		Permite utilizar algoritmo/s para Redes Bayesianas	1	1

Fases	Sub Fases	Indicador	Analysis Services	Weka
	Construcción del modelo	Permite utilizar algoritmo/s de Redes Neuronales	1	1
		Permite la construcción gráfica del modelo	1	1
		Permite la ejecución del modelo con una interfaz gráfica	1	1
		Permite la creación de un flujo gráfico del modelo	0	1
	Evaluación del modelo	Permite evaluar el performance del modelo	1	1
		Genera gráficos de evaluación	1	1
		Permite analizar el costo beneficio del modelo	1	1
Fase de implantación	Informe Final	Permite crear gráficos que ilustren los resultados obtenidos	1	1

Tabla 3: ANÁLISIS DE LAS HERRAMIENTAS MICROSOFT ANALYSIS SERVICES Y WEKA

Fuente:(Nay Mojarrango & José Chapalbay, 2015)

Se ha realizado el análisis de las herramientas en el cuál los indicadores planteados con “0” corresponden a que no se ha podido encontrar la funcionalidad dentro de la herramienta y además no existe información relacionada a que el software de minería de datos ya sea Microsoft Analysis Services o Weka posea tal funcionalidad. Los indicadores planteados con “1” corresponden a que la herramienta permite realizar esta tarea y por lo tanto posee esta funcionalidad.

3.1.1.5 ANÁLISIS DE RESULTADOS DEL ESTUDIO COMPARATIVO

El presente Análisis de resultados permite identificar varios aspectos relacionados con el análisis comparativo realizado a las herramientas Microsoft Analysis Services y Weka, tales como la comparación de las Sub fases y Fases junto con el análisis estadístico orientado a definir la mejor herramienta que posteriormente se aplicará para el desarrollo de la minería de datos.

Se presenta el siguiente resumen enfocado a la puntuación de las Subfases:

SUBFASES	PUNTAJE ANALYSIS SERVICES	PUNTAJE WEKA
Recolección de Datos iniciales	5	4
Descripción de los datos	5	5
Exploración de los datos	0	2
Verificación de la calidad de los datos	2	2
Selección de datos	1	1
Limpieza de datos	2	4
Estructuración de los datos	3	3
Integración de los datos	2	2
Formateo de los datos	2	2
Selección de la técnica de modelado	7	7
Construcción del modelo	2	3
Evaluación del modelo	3	3
Informe Final	1	1
SUMA	35	39

Tabla 4: RESULTADOS DEL ESTUDIO COMPARATIVO

Fuente:(Nay Mojarrango & José Chapalbay, 2015)

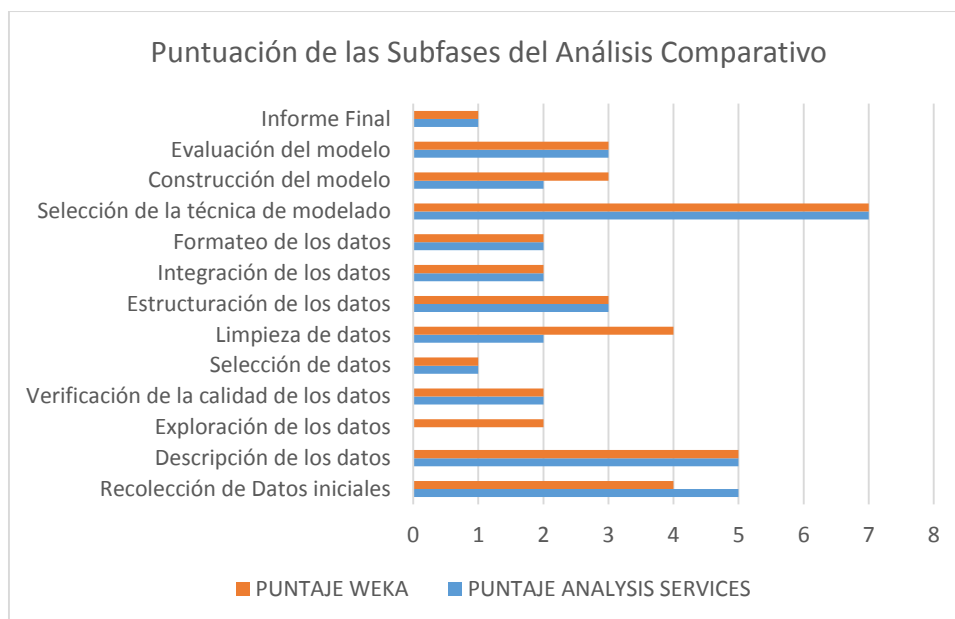


Figura 8: Resultado Puntuación Sub fases

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

En el gráfico se puede observar que Microsoft Analysis Services supera a Weka en la Recolección de Datos Iniciales, sin embargo es superada por Weka en la construcción del modelo debido a sus múltiples herramientas para la creación y ejecución de flujos y procedimientos, además en las Sub fases de limpieza de datos y significativamente superior en la Exploración de los datos.

FASE DE MINERÍA	PUNTAJE ANALYSIS SERVICES	% ANALYSIS SERVICES	PUNTAJE WEKA	% WEKA
Comprensión de los Datos	12	86%	13	93%
Fase de preparación de los datos	10	77%	12	92%
Fase de modelado	12	92%	13	100%
Fase de implantación	1	100%	1	100%
PROMEDIO	8.75	89%	9.75	96%

Tabla 5: PROMEDIO

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

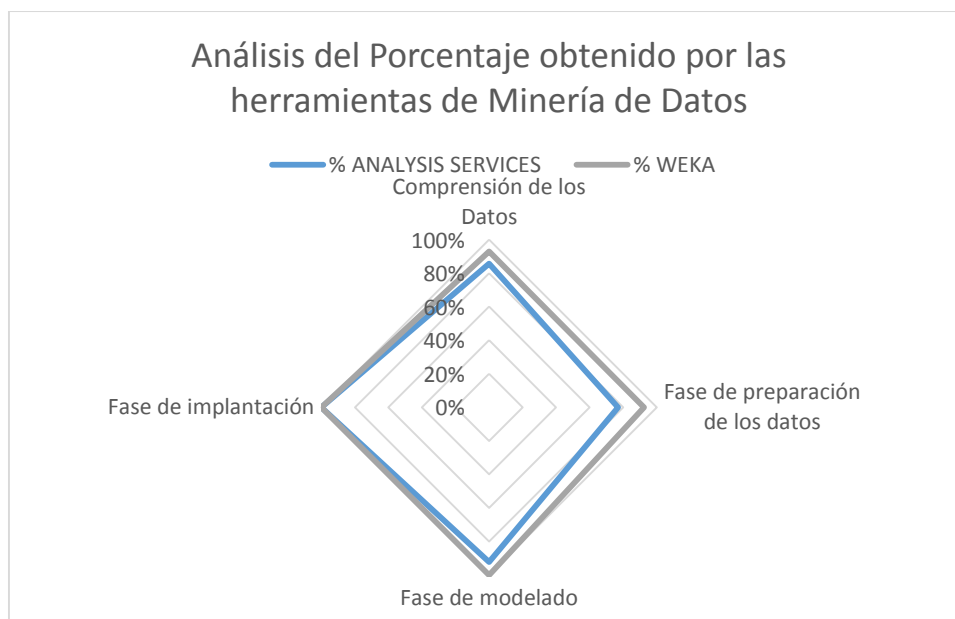


Figura 9: Análisis del Porcentaje

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

En el gráfico radial se puede observar que de acuerdo a los indicadores analizados la herramienta Weka supera a Microsoft Analysis Services en todas las fases evaluadas a excepción de la Fase de Implantación que se encuentran las dos herramientas al 100% según el indicador planteado.

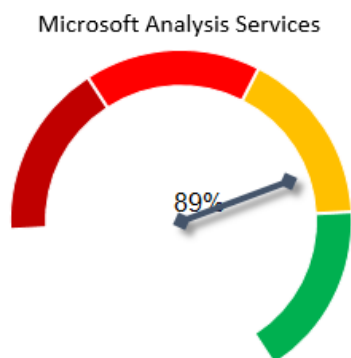


Figura 10: Grafico Radial Analysis Services

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

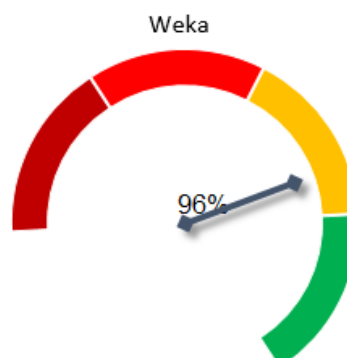


Figura 11: Grafico Radial WEKA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

En los gráficos de velocímetros se puede observar que la mejor herramienta de minería de datos es Weka superando a Microsoft Analysis Services con 7%, según los indicadores planteados enfocados a la optimización del proceso de minería con la metodología CRISP-DM.

CAPÍTULO IV

DESARROLLO DE MINERÍA DE DATOS

El presente documento se encuentra realizado bajo la metodología para minería de datos CRISP-DM, creada por el grupo de empresas SPSS, NCR y Daimler Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining. Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación. La sucesión de fases, no es necesariamente rígida. Cada fase es descompuesta en varias tareas generales de segundo nivel (Moine, 2011).

4.1 FASE DE COMPRENSIÓN DEL NEGOCIO

4.1.1 OBJETIVOS DEL NEGOCIO

La Empresa Prasol “Lácteos Santillán”, busca mejorar la rentabilidad de su empresa a través de la aplicación de información estratégica que permita a la gerencia tomar decisiones con respecto a la información generada.

4.1.2 EVALUACIÓN DE LA SITUACIÓN

La Empresa Prasol “Lácteos Santillán”, al momento cuenta con información generada por sus sistemas informáticos, posee un servidor con un sistema operativo Windows Server 2008, en donde se encuentra alojada la base de datos con el motor SQL Server 2012, esta se encuentra diseñada con un modelo transaccional, esta es una dificultad debido a que al no poseer una Datawarehouse dificulta las tareas de minería de datos.

La base de datos proporcionada por la empresa cuenta con aproximadamente 3.000 registros útiles que se considera medianamente suficiente para este proyecto.

La información almacenada por los sistemas informáticos no presenta datos demográficos, específicamente de características del cliente tales como: edad, sexo,

estado civil, sueldo, lugar de nacimiento, etc. Esto es un problema debido a que no se podrá realizar tareas de clasificación de clientes para definir posibles perfiles de compra o realizar segmentación en base a estas características.

En relación a la viabilidad del proyecto se ha decidido continuar con el mismo debido a que se podrá realizar tareas de segmentación de clientes y asociación de compra de productos lo cual permitirá incrementar la rentabilidad de la Empresa.

4.1.3 DETERMINACIÓN DE LOS OBJETIVOS DE LA MINERÍA DE DATOS

El objetivo general de este proyecto es analizar la información a través de la aplicación de técnicas de minería de datos enfocadas a descubrir patrones que permitan apoyar a la toma de decisiones que se orientan a la aplicación de inteligencia empresarial por parte de la gerencia y así incrementar la rentabilidad de la empresa. A continuación se describen los objetivos específicos para alcanzar la meta propuesta:

- 1) Determinar que técnicas de minería de datos se pueden aplicar en función de la información que posee la empresa.
- 2) Aplicar un análisis RFM con el objetivo de fidelizar a los clientes.
- 3) Establecer distintos grupos entre las personas que realizan una compra más a menudo y así generar estrategias de marketing enfocadas al incremento de las ventas.
- 4) Crear asociaciones de productos para definir una cesta de compra que permitirá crear predicciones para recomendar productos a los clientes.

4.1.4 PLAN DEL PROYECTO

Actividades	Duración (días)	Tareas a desarrollar	Técnicas a emplear
Comprensión del Negocio o Problema	10	<p>Determinar los objetivos del negocio.</p> <p>Evaluación de la situación.</p> <p>Determinar los objetivos de la Minería de Datos.</p> <p>Realizar el plan del proyecto.</p>	N/A
Comprensión de los Datos	5	<p>Recolección de Datos iniciales.</p> <p>Descripción de los datos.</p> <p>Exploración de los datos.</p> <p>Verificación de la calidad de los datos.</p>	<p>Ejecución de consultas.</p> <p>Gráficos de Frecuencia</p> <p>Resumen de Errores en los datos.</p>

Fase de preparación de los datos	8	Selección de datos. Estructuración de los datos. Integración de los datos. Formateo de los datos.	Creación de un Datawarehouse de ventas. Ejecución de consultas SQL. Creación de Vistas.
Fase de modelado	17	Selección de la técnica de modelado. Generación del plan de prueba. Construcción del modelo. Evaluación del modelo.	Análisis RFM. Segmentación de Clientes. Asociación de Productos. Performance de los modelos.
Fase de evaluación	4	Evaluación de los resultados Proceso de revisión. Determinación de futuras fases.	Generación de Gráficos estadísticos.
Fase de implantación	5	Plan de implementación.	N/A

		<p>Monitorización</p> <p>y</p> <p>Mantenimiento.</p> <p>Informe Final.</p> <p>Revisión del</p> <p>proyecto.</p>	
--	--	---	--

Tabla 6: Plan del Proyecto

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

4.2 FASE DE COMPRENSIÓN DE LOS DATOS

4.2.1 RECOLECCIÓN DE DATOS INICIALES

La Empresa ha provisto una base de datos transaccional de facturación la misma que contiene varias tablas.

Para realizar la recolección de datos iniciales se ha creado dos ODBC's de usuario para la base de datos origen y para la base de datos en donde se cargará el datamart de ventas.

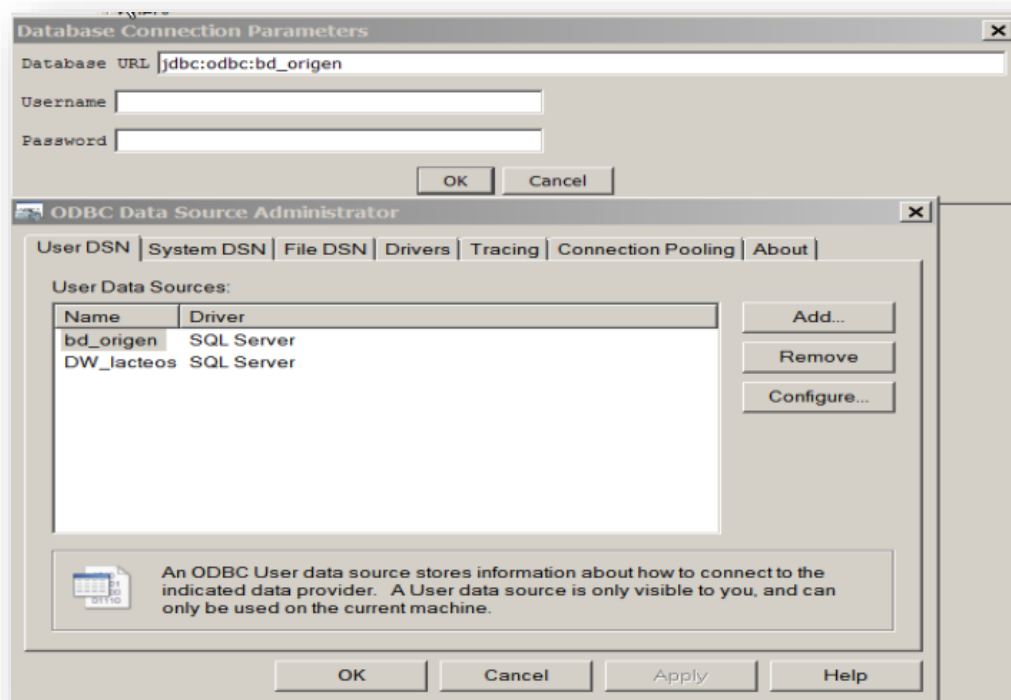


Figura 12: ODBC's de usuario

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Las tablas de la base de datos que contiene el modelo transaccional se pueden observar en el siguiente gráfico:

Row	
1	sysdiagrams
2	BODEGA
3	PRODUCTO
4	FACTURA
5	DET_FACTURA
6	CLIENTE
7	BONOS
8	CAJA
9	CAJA_CUADRE
10	CAT_ESPECIAL
11	CONFIGURACION_PV
12	CUR_DATP
13	DOCUMENTOS
14	ENCPED
15	EXCLUIR
16	EXCLUIR_HISTORICO
17	IDS
18	INIT
19	LOCAL
20	REPORTES
21	RUTAS
22	SECUENCIAS
23	TARJETAS
24	TRACAJA
25	TRANFAC
26	TRANFAC_ELIMINADAS
27	USERS

Figura 13: Tabla BD Modelo Transaccional

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Varias de estas tablas no han sido utilizadas sea la tabla completa o algunos de sus campos, sin embargo existe la posibilidad aun así de la creación de un Datamart de Ventas.

4.2.2 DESCRIPCIÓN DE LOS DATOS

A continuación se presenta el número registros por tablas y el número de campos por registro:

	TableName	RowCount
1	[dbo].[BODEGA]	3
2	[dbo].[BONOS]	0
3	[dbo].[CAJA]	5
4	[dbo].[CAJA_CUADRE]	7
5	[dbo].[CAT_ESPECIAL]	3
6	[dbo].[CLIENTE]	219
7	[dbo].[CONFIGURACION_PV]	1
8	[dbo].[CUR_DATP]	1
9	[dbo].[DET_FACTURA]	4563
10	[dbo].[DOCUMENTOS]	5
11	[dbo].[ENCPED]	0
12	[dbo].[EXCLUIR]	0
13	[dbo].[EXCLUIR_HISTORICO]	0
14	[dbo].[FACTURA]	905
15	[dbo].[IDS]	6
16	[dbo].[INIT]	1
17	[dbo].[LOCAL]	0
18	[dbo].[PRODUCTO]	62
19	[dbo].[REPORTES]	10
20	[dbo].[RUTAS]	0
21	[dbo].[SECUENCIAS]	4
22	[dbo].[sysdiagrams]	1
23	[dbo].[TARJETAS]	2
24	[dbo].[TRACAJA]	1036
25	[dbo].[TRANFAC]	3721
26	[dbo].[TRANFAC_ELIMINAD...]	1
27	[dbo].[USERS]	0

Figura 14: Número registros por tablas y el número de campos por registro

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Se observar claramente que muchas de las tablas poseen cero registros que indica que estas tablas no han sido utilizadas.

Se muestra cuáles de las tablas serán útiles para la construcción del datamart:

TABLA	SE INCLUYE EN EL DATAMART	OBSERVACIONES
BODEGA	SI	Se puede incluir en el datamart de ventas, debido a que se puede utilizar para dimensión.
BONOS	NO	No posee registros
CAJA_CUADRE	NO	No está incluido en el análisis.
CAT_ESPECIAL	NO	No está incluido en el análisis.
CLIENTE	SI	Es necesario para la segmentación de mercado.
CONFIGURACIÓN_PV	NO	Es una tabla de configuración.
CUR_DATP	NO	Es una tabla no utilizada.
DET_FACTURA	SI	Se almacena el detalle de compra de los clientes.
DOCUMENTOS	NO	No está incluido en el análisis.

ENCPED	NO	La tabla no ha sido utilizada.
EXCLUIR	NO	La tabla no ha sido utilizada.
EXCLUIR_HISTORICO	NO	La tabla no ha sido utilizada.
FACTURA	SI	Se almacena el total de compra de los clientes.
IDS	NO	Almacena información de seguridad
INIT	NO	Almacena información de seguridad
LOCAL	NO	La tabla no ha sido utilizada
PRODUCTO	SI	Almacena los productos generados en la empresa
REPORTES	NO	La tabla almacena únicamente los reportes generados
RUTAS	NO	La tabla no ha sido utilizada
SECUENCIAS	NO	La tabla almacena información de secuencias de procesos,

		no incluida en el análisis.
TARJETAS	NO	No incluida en el análisis.
TRACAJA	NO	No incluida en el análisis.
TRANFAC	NO	No incluida en el análisis.
TRANFAC_ELIMINADAS	NO	No incluida en el análisis.
USERS	NO	La tabla no ha sido utilizada.

Tabla 7: Tablas útiles para la construcción del datamart

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

4.2.3 EXPLORACIÓN DE DATOS

A continuación, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos, se realiza la exploración de las tablas seleccionadas para el datamart:

BODEGA:

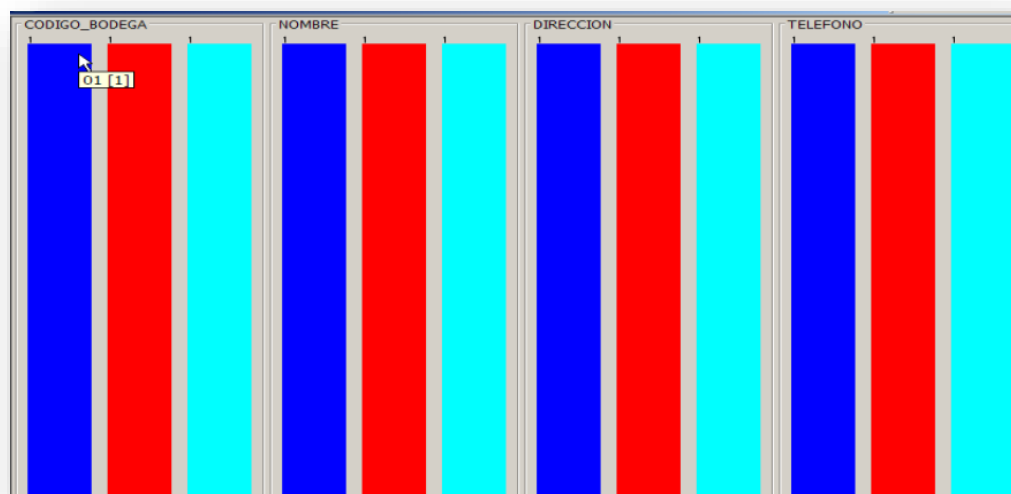


Figura 15: Exploración de tablas seleccionadas para el datamart

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Se puede observar que todos los datos son distintos.

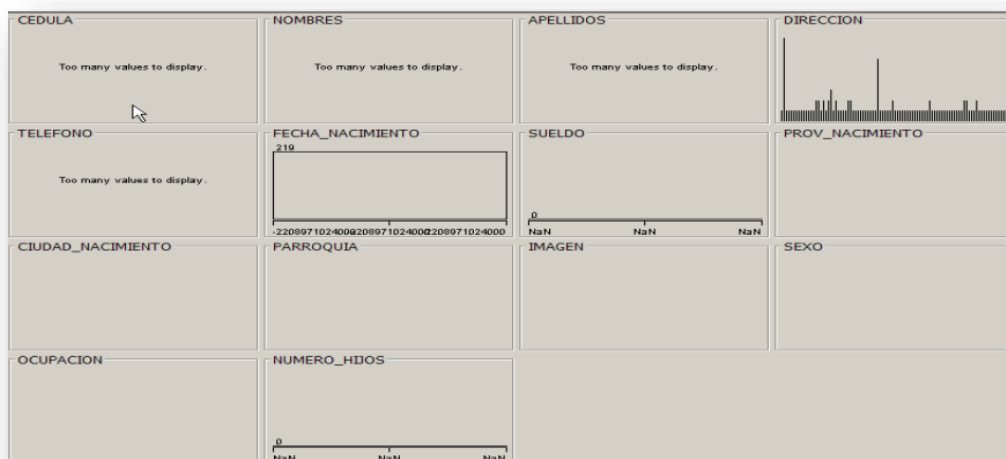
CLIENTE:

Figura 16: Grafico Frecuencia CLIENTE

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Se puede observar que muchos de los datos son distintos por lo tanto no se muestra un gráfico de frecuencias en los campos, a diferencia de dirección en el cual se puede observar un gráfico claro de frecuencia.

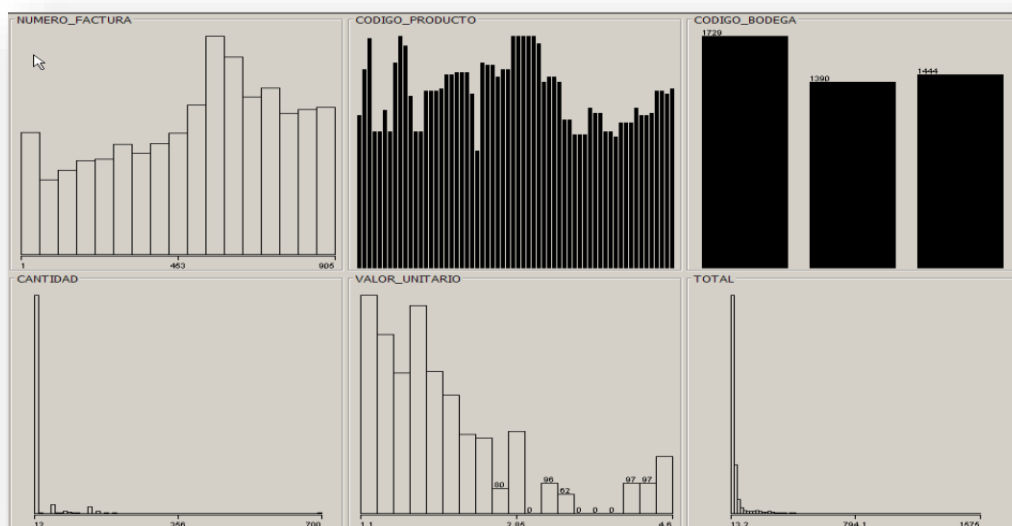
DET_FACTURA:

Figura 17: Grafico frecuencia DET_FACTURA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

En esta tabla se puede observar claras frecuencias de los datos, esto es muy útil para poder suponer que la minería de datos puede tener resultados satisfactorios.

FACTURA:

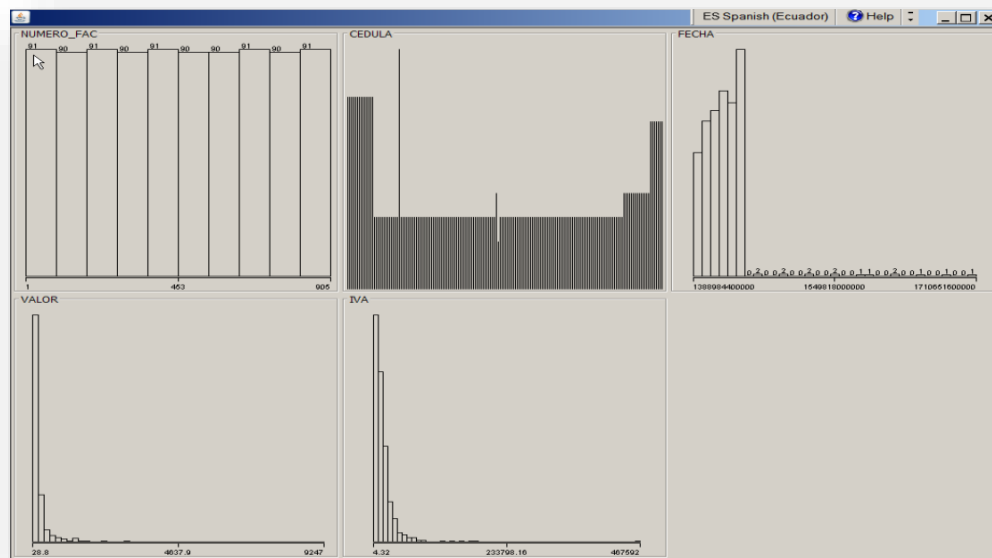


Figura 18: Grafico frecuencia FACTURA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Se puede visualizar que existe frecuencia de compra y que existen clientes que se destacan por su adquisición de productos.

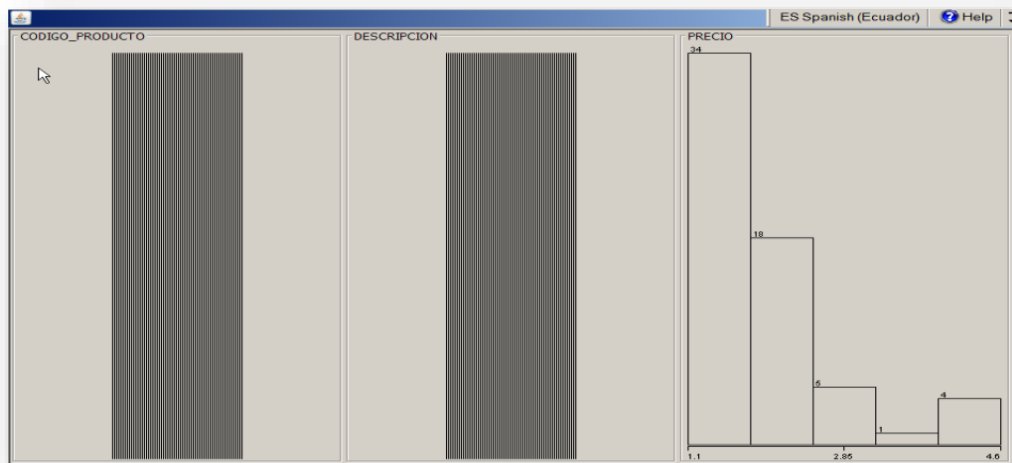
PRODUCTO:

Figura 19: Grafico frecuencia PRODUCTO

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Se puede destacar que existe una frecuencia en los rangos del precio del producto.

4.2.4 VERIFICACIÓN DE LA CALIDAD DE LOS DATOS

La verificación de la calidad de los datos se realiza con la herramienta Weka la misma ha permitido identificar los campos que poseen valores erróneos y que no son de utilidad con la finalidad de plantear de forma correcta el diseño del datamart de ventas. Los campos de la base de datos se presentan a continuación:

BODEGA:

VERIFICACIÓN DE CALIDAD	OBSERVACIONES
Selected attribute Name: CODIGO_BODEGA Type: Nominal Missing: 0 (0%) Distinct: 3 Unique: 3 (100%)	Correcto
Selected attribute Name: NOMBRE Type: Nominal Missing: 0 (0%) Distinct: 3 Unique: 3 (100%)	Correcto
Selected attribute Name: DIRECCION Type: Nominal Missing: 0 (0%) Distinct: 3 Unique: 3 (100%)	Correcto
Selected attribute Name: TELEFONO Type: Nominal Missing: 0 (0%) Distinct: 3 Unique: 3 (100%)	Correcto

Tabla 8: Verificación Calidad de Datos BODEGA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

CLIENTE:

VERIFICACIÓN DE CALIDAD	OBSERVACIONES
Selected attribute Name: CEDULA Type: Nominal Missing: 0 (0%) Distinct: 219 Unique: 219 (100%)	Correcto, este campo es necesario que ningún valor este duplicado.
Selected attribute Name: NOMBRES Type: Nominal Missing: 0 (0%) Distinct: 201 Unique: 190 (87%)	Correcto
Selected attribute Name: APELLIDOS Type: Nominal Missing: 0 (0%) Distinct: 214 Unique: 209 (95%)	Correcto
Selected attribute Name: DIRECCION Type: Nominal Missing: 0 (0%) Distinct: 177 Unique: 147 (67%)	Correcto
Selected attribute Name: TELEFONO Type: Nominal Missing: 0 (0%) Distinct: 215 Unique: 211 (96%)	Correcto

<p>Selected attribute Name: FECHA_NACIMIENTO Missing: 0 (0%) Distinct: 1 Type: Date Unique: 0 (0%)</p>	<p>A pesar de que existe un 0% de errores, este dato no se encuentra correcto, debido a que únicamente existe un valor, el dato de 0% únicos determina que es un valor por defecto y que este campo no ha sido utilizado</p>
<p>Selected attribute Name: PROV_NACIMIENTO Missing: 219 (100%) Distinct: 0 Type: Nominal Unique: 0 (0%)</p> <p>Selected attribute Name: CIUDAD_NACIMIENTO Missing: 219 (100%) Distinct: 0 Type: Nominal Unique: 0 (0%)</p> <p>Selected attribute Name: SUELDO Missing: 219 (100%) Distinct: 0 Type: Numeric Unique: 0 (0%)</p> <p>Selected attribute Name: PARROQUIA Missing: 219 (100%) Distinct: 0 Type: Nominal Unique: 0 (0%)</p> <p>Selected attribute Name: IMAGEN Missing: 219 (100%) Distinct: 0 Type: Nominal Unique: 0 (0%)</p> <p>Selected attribute Name: SEXO Missing: 219 (100%) Distinct: 0 Type: Nominal Unique: 0 (0%)</p> <p>Selected attribute Name: OCUPACION Missing: 219 (100%) Distinct: 0 Type: Nominal Unique: 0 (0%)</p> <p>Selected attribute Name: NUMERO_HDOS Missing: 219 (100%) Distinct: 0 Type: Numeric Unique: 0 (0%)</p>	<p>Atributos incorrectos en un 100%, estos campos no han sido utilizados</p>

Tabla 9: Verificación Calidad de Datos CLIENTE

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

DET_FACTURA:

VERIFICACIÓN DE CALIDAD	OBSERVACIONES
Selected attribute Name: NUMERO_FACTURA Type: Numeric Missing: 0 (0%) Distinct: 905 Unique: 0 (0%)	Correcto
Selected attribute Name: CODIGO_PRODUCTO Type: Nominal Missing: 0 (0%) Distinct: 62 Unique: 0 (0%)	Correcto
Selected attribute Name: CODIGO_BODEGA Type: Nominal Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)	Correcto
Selected attribute Name: CANTIDAD Type: Numeric Missing: 0 (0%) Distinct: 18 Unique: 0 (0%)	Correcto
Selected attribute Name: VALOR_UNITARIO Type: Numeric Missing: 0 (0%) Distinct: 33 Unique: 0 (0%)	Correcto
Selected attribute Name: TOTAL Type: Numeric Missing: 0 (0%) Distinct: 261 Unique: 81 (2%)	Correcto

Tabla 10: Verificación Calidad de Datos DET_FACTURA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

FACTURA:

VERIFICACIÓN DE CALIDAD	OBSERVACIONES
Selected attribute Name: NUMERO_FAC Type: Numeric Missing: 0 (0%) Distinct: 905 Unique: 905 (100%)	Correcto
Selected attribute Name: CEDULA Type: Nominal Missing: 0 (0%) Distinct: 219 Unique: 0 (0%)	Correcto
Selected attribute Name: FECHA Type: Date Missing: 0 (0%) Distinct: 115 Unique: 19 (2%)	Correcto
Selected attribute Name: VALOR Type: Numeric Missing: 0 (0%) Distinct: 509 Unique: 332 (37%)	Correcto
Selected attribute Name: IVA Type: Numeric Missing: 0 (0%) Distinct: 509 Unique: 332 (37%)	Correcto

Tabla 11: Verificación Calidad de Datos FACTURA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

PRODUCTO:

VERIFICACIÓN DE CALIDAD	OBSERVACIONES
Selected attribute Name: CODIGO_PRODUCTO Type: Nominal Missing: 0 (0%) Distinct: 62 Unique: 62 (100%)	Correcto
Selected attribute Name: DESCRIPCION Type: Nominal Missing: 0 (0%) Distinct: 62 Unique: 62 (100%)	Correcto
Selected attribute Name: PRECIO Type: Numeric Missing: 0 (0%) Distinct: 33 Unique: 14 (23%)	Correcto

Tabla 12: Verificación Calidad de Datos PRODUCTO

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

4.3 FASE DE PREPARACIÓN DE LOS DATOS

4.3.1 SELECCIÓN DE DATOS

Los datos seleccionados para el análisis de minería de datos previamente, son los que después del análisis de la calidad de datos se encuentran en un estado correcto, estos serán utilizados para la creación del datamart de ventas, posteriormente para los diferentes análisis se realizarán las vistas necesarias con la finalidad de que la data esté a punto para la fase de modelado. El volumen de datos es mediano es por esta razón que se utilizarán todos los datos disponibles.

El datamart de ventas tendrá el siguiente modelo, un copo de nieve debido a que se posee varias tablas de hechos relacionadas con las mismas dimensiones.

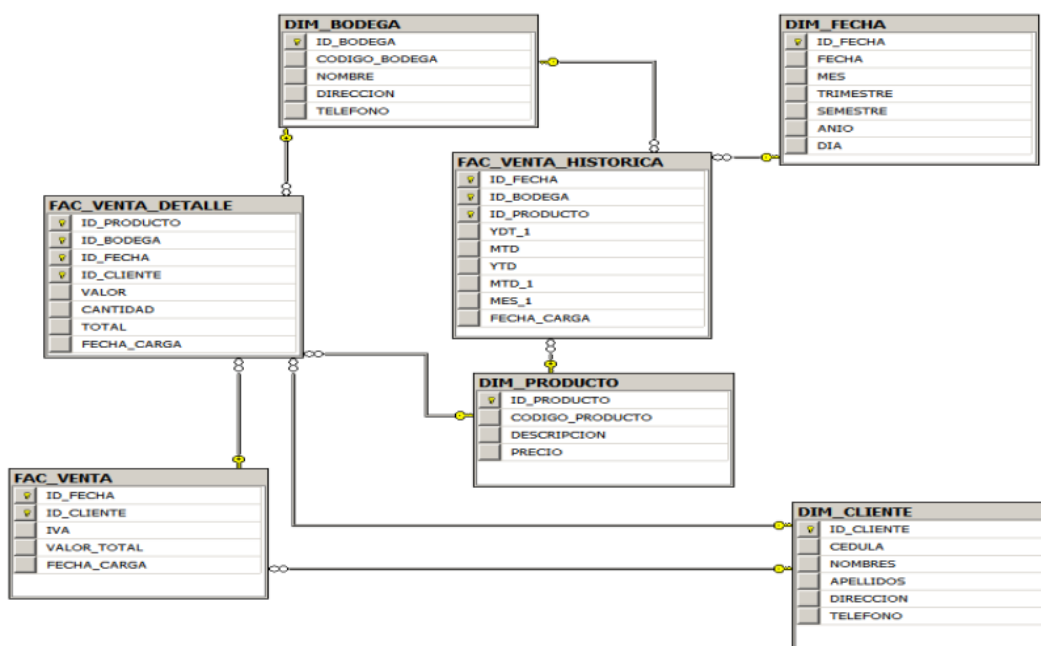


Figura 20: Datamart Ventas

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

4.3.2 LIMPIEZA DE LOS DATOS

Con la finalidad de generar el Datawarehouse en la base de datos SQL Server, se ha aplicado un ETL a través de la herramienta SAP Data Services Designer con una licencia temporal, los flujos de carga se muestran a continuación:

Flujo del Job de Ventas

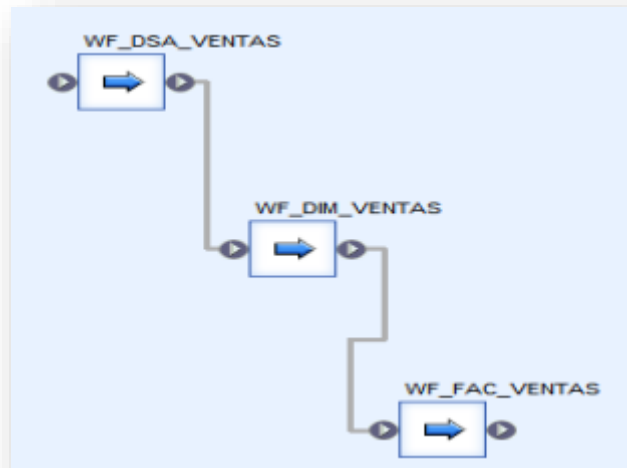


Figura 21: Flujo del Job de Ventas

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Workflow del DSA (Área de Preparación de Datos)

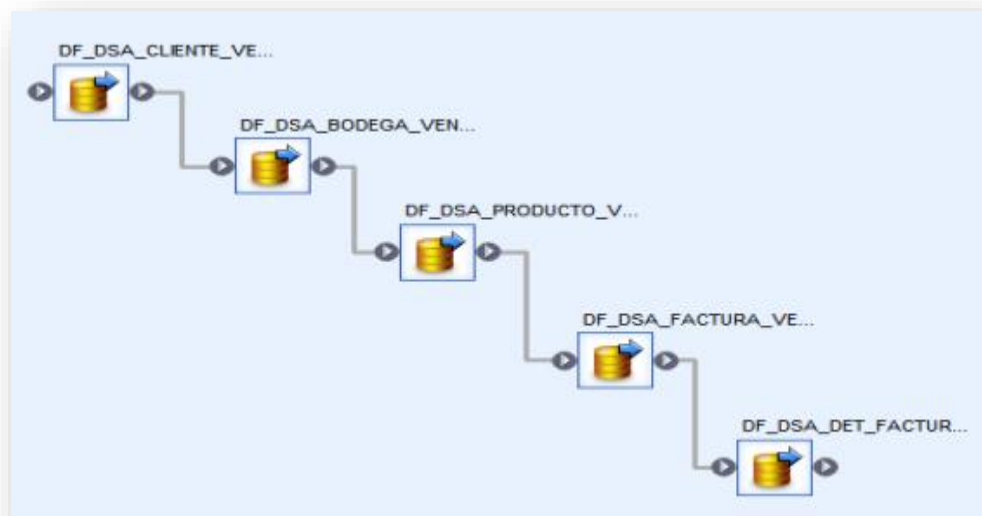


Figura 22: Workflow del DSA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow de la tabla CLIENTE



Figura 23: Dataflow tabla CLIENTE

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow de la tabla BODEGA

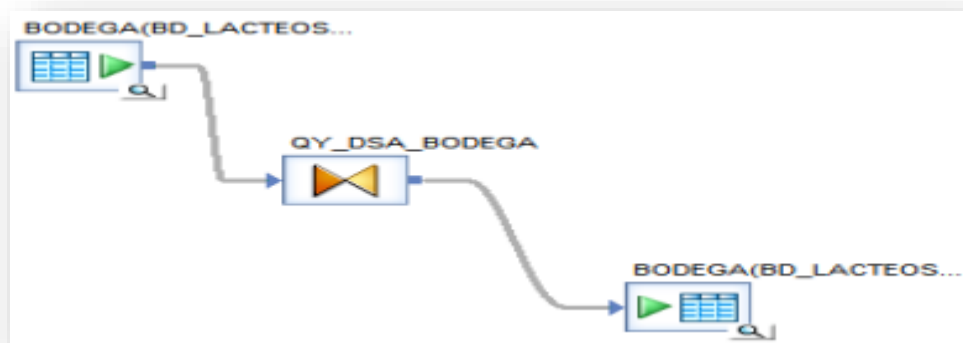


Figura 24: Dataflow tabla BODEGA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow de la tabla PRODUCTO



Figura 25: Dataflow tabla PRODUCTO

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow de la tabla FACTURA

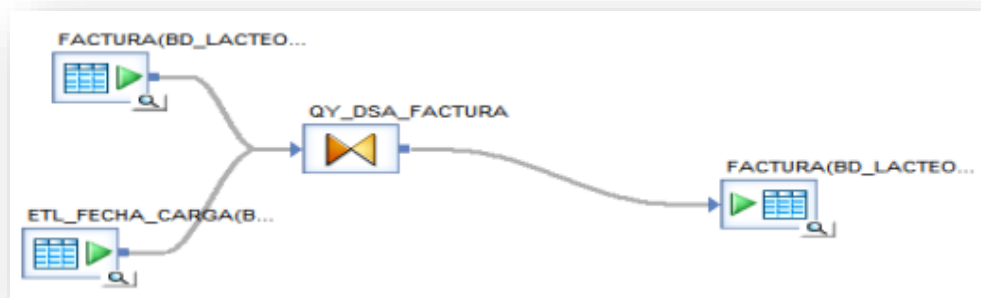


Figura 26: Dataflow tabla FACTURA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow de la tabla DETALLE FACTURA

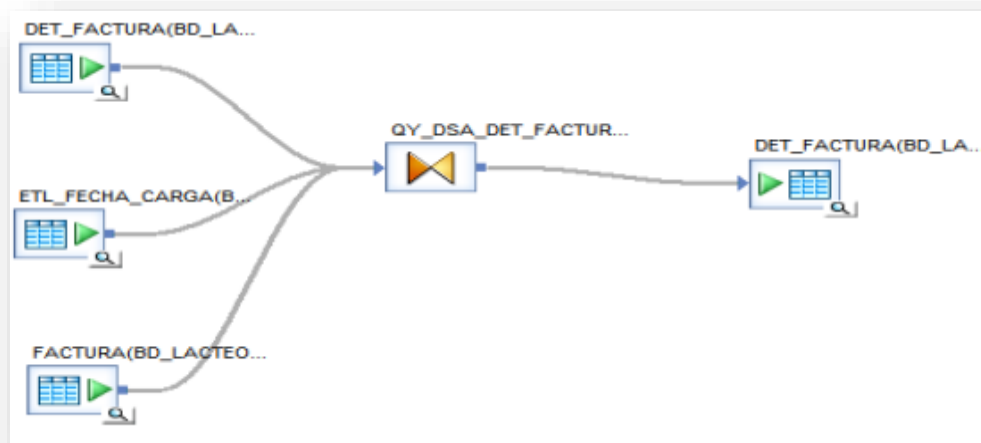


Figura 27: Dataflow tabla DETALLE FACTURA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Workflow de Carga de Dimensiones

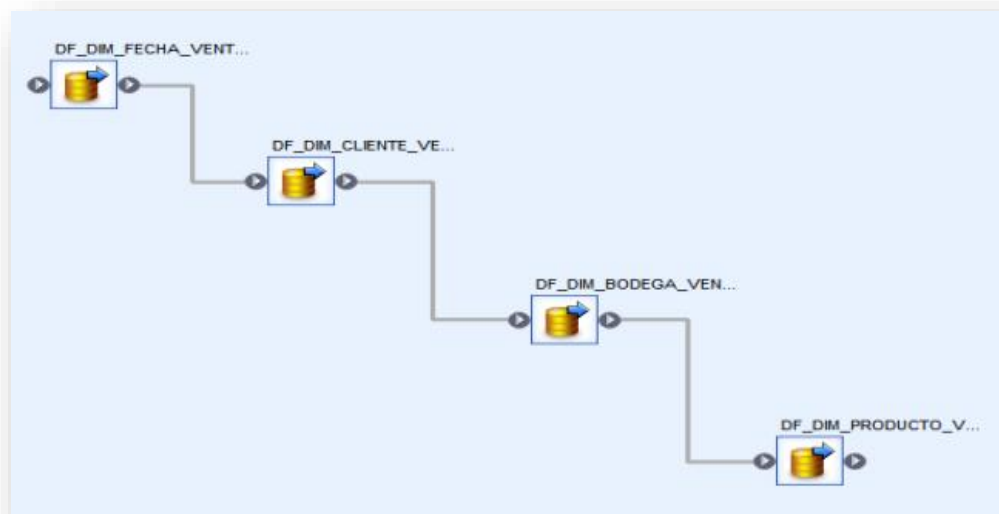


Figura 28: Workflow Carga Dimensiones

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow para la tabla DIM_FECHA

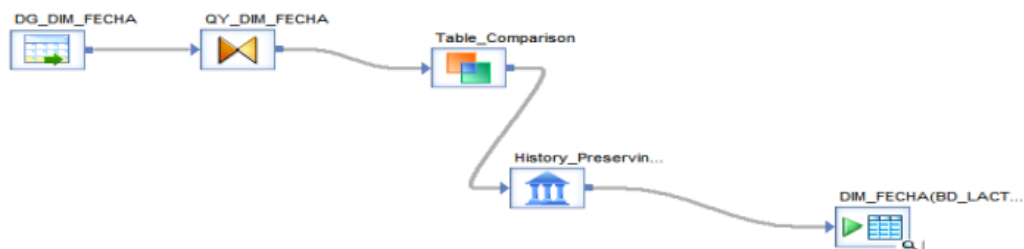


Figura 29: Dataflow tabla DIM_FECHA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow para la tabla DIM_CLIENTE

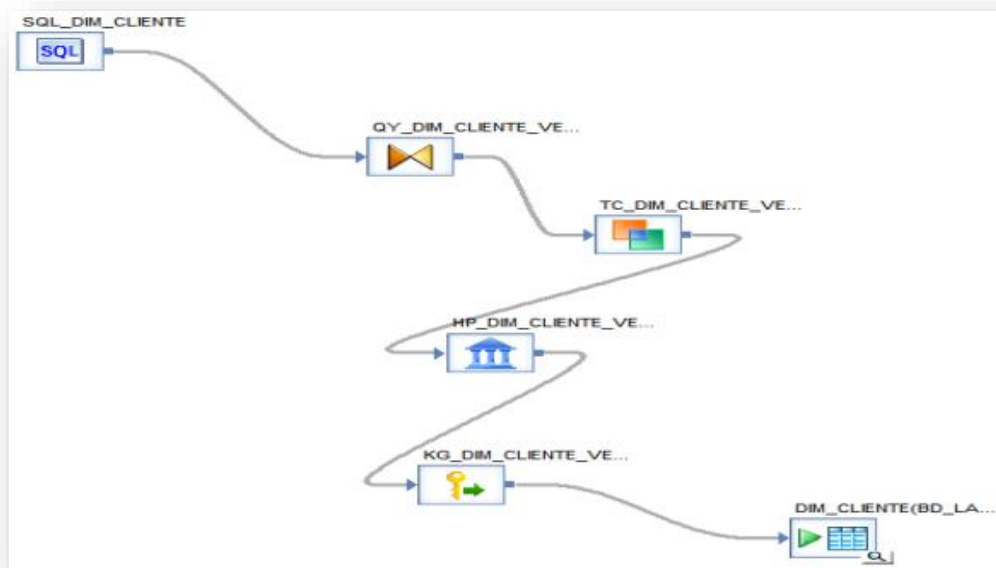


Figura 30: Dataflow tabla DIM_CLIENTE

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow para la tabla DIM_BODEGA

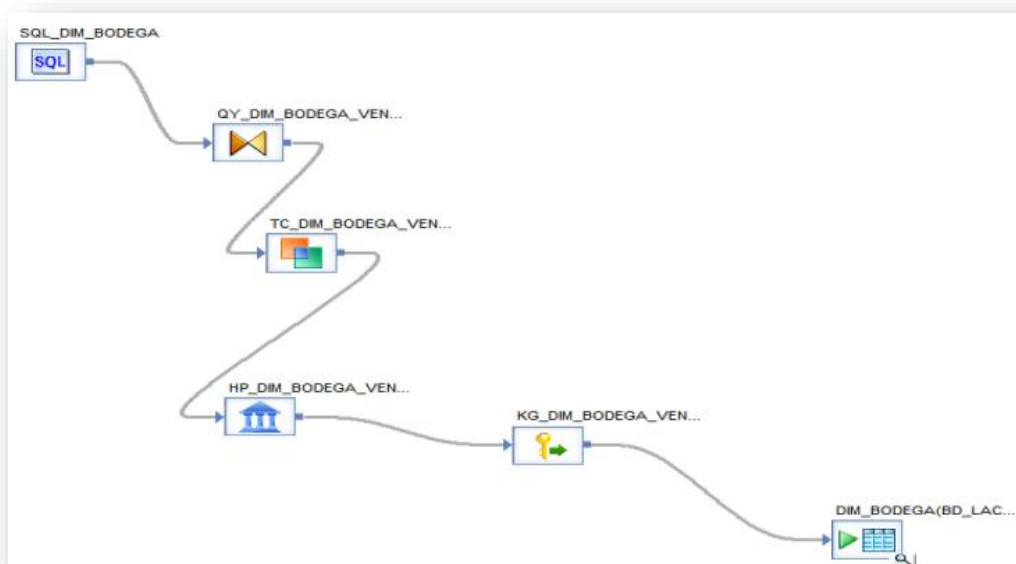


Figura 31: Dataflow tabla DIM_BODEGA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow para la tabla DIM_PRODUCTO

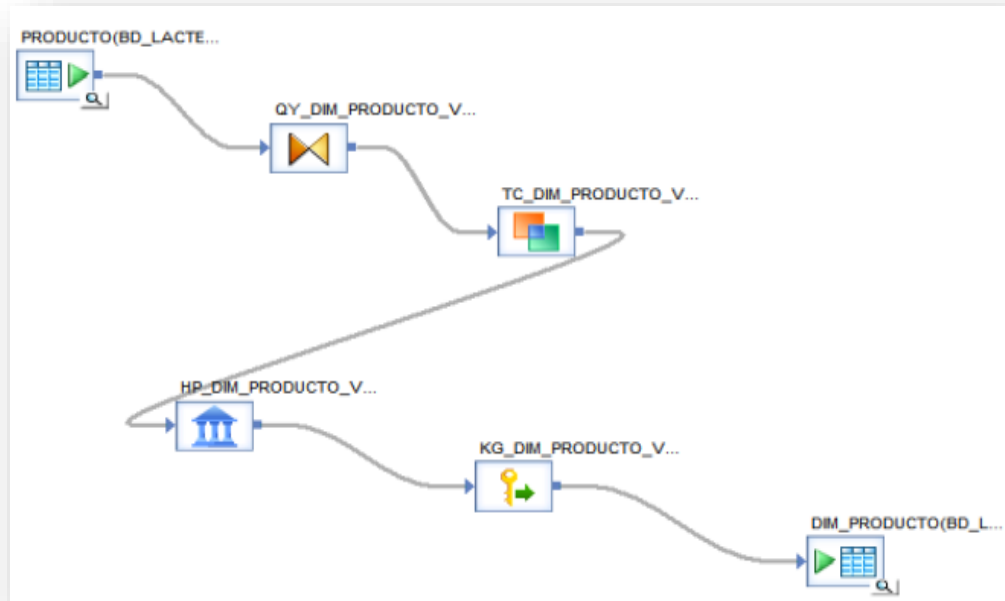


Figura 32: Dataflow tabla DIM_PRODUCTO

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Workflow de Carga de Tablas de Hechos

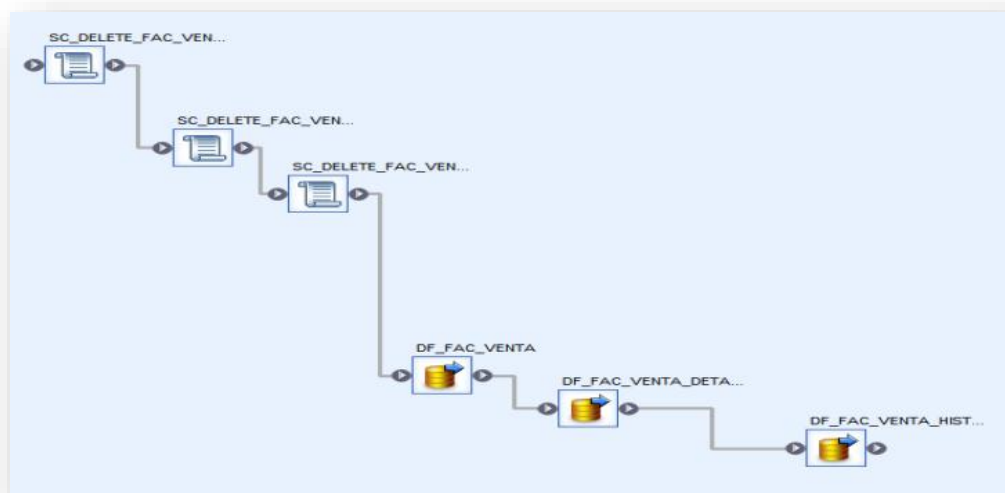


Figura 33: Workflow Carga tablas de Hechos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow para la tabla FAC_VENTA

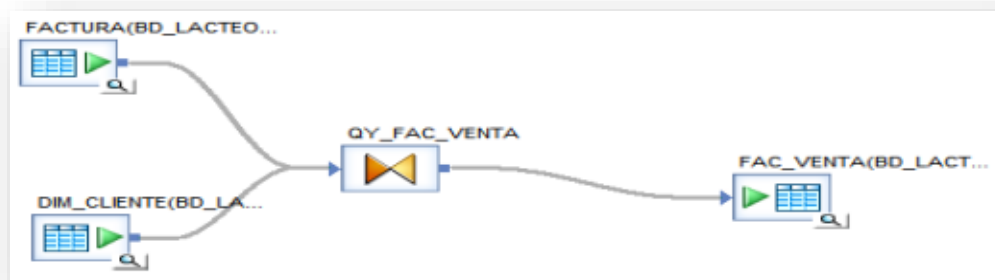


Figura 34: Dataflow tabla FAC_VENTA

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow para la tabla FAC_VENTA_DETALLE

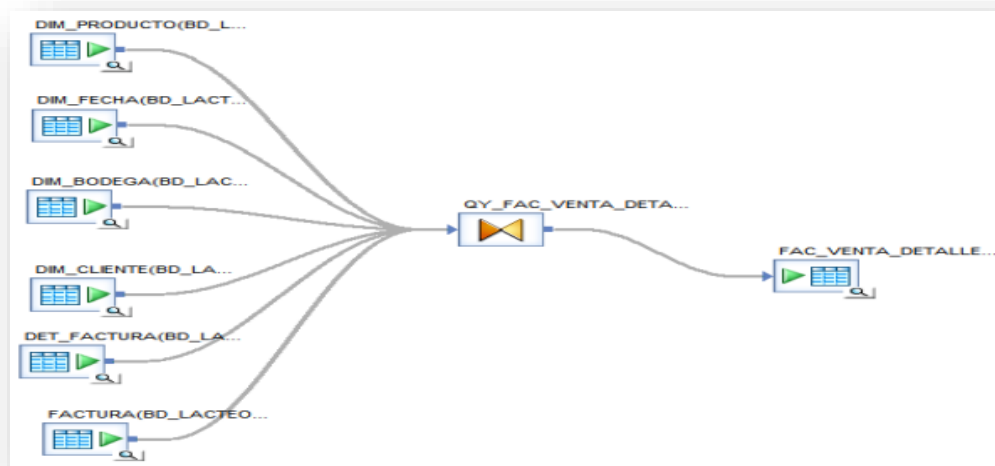


Figura 35: Dataflow tabla FAC_VENTA_DETALLE

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Dataflow de la tabla **FACTURA VENTA HISTORICA:**

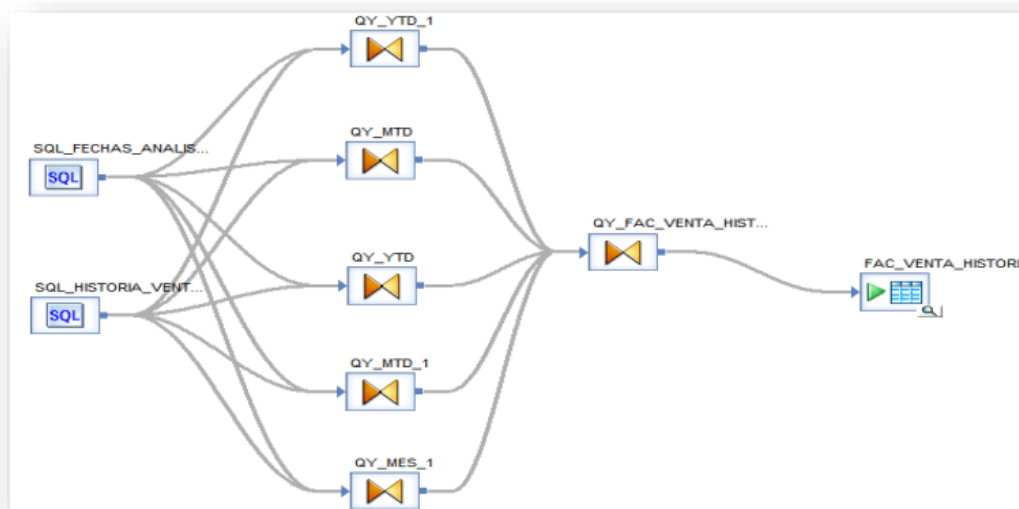


Figura 36: Dataflow tabla *FACTURA VENTA HISTORICA*

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Las tablas del datamart creado se muestran a continuación:

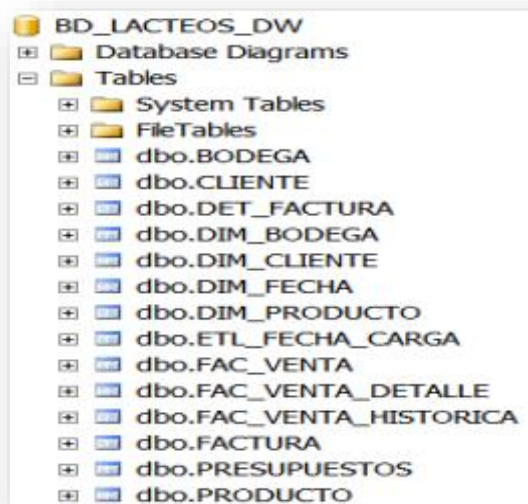


Figura 37: Tablas del Datamart

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Vistas para el Análisis RFM

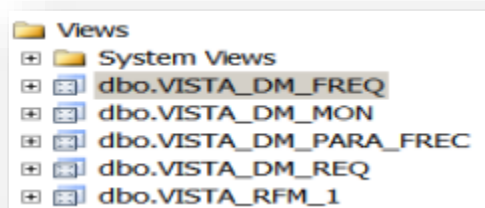


Figura 38: Vistas Análisis RFM

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Discretización de valores del RFM

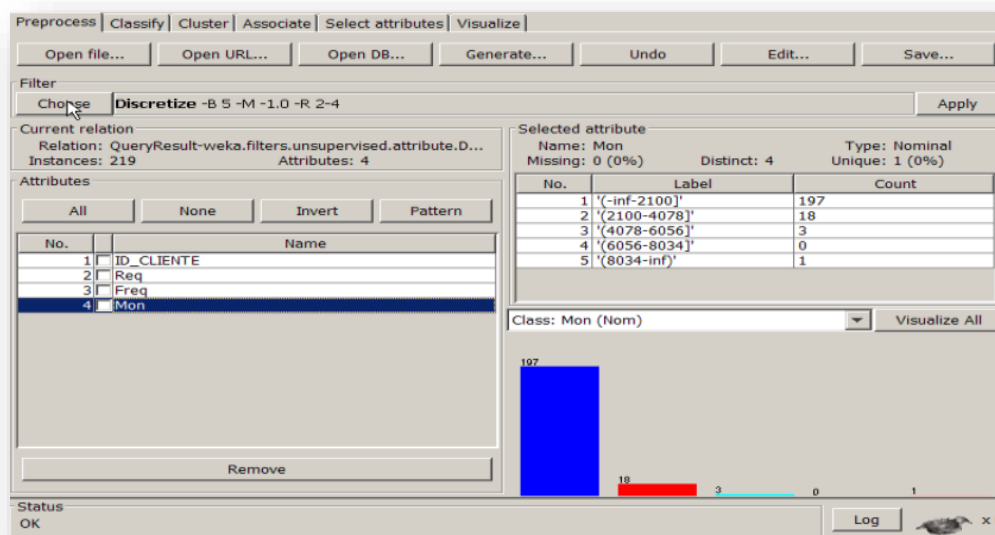


Figura 39: Discretización Valores RFM

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

A partir del datamart se ha generado una vista de valores de la Recencia, Frecuencia y Valor monetario, estos valores son discretizados a través de Weka en 5 bins.

Selected attribute		
Name: Req		Type: Nominal
Missing: 0 (0%)	Distinct: 5	Unique: 0 (0%)
No.	Label	Count
1	'(-inf-63.4]'	35
2	'(63.4-92.8]'	14
3	'(92.8-122.2]'	34
4	'(122.2-151.6]'	46
5	'(151.6-inf)'	90

Figura 40: Vista de valores de Recencia

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Selected attribute		
Name: Freq		Type: Nominal
Missing: 0 (0%)	Distinct: 5	Unique: 0 (0%)
No.	Label	Count
1	'(-inf-3.8]'	43
2	'(3.8-5.6]'	71
3	'(5.6-7.4]'	74
4	'(7.4-9.2]'	25
5	'(9.2-inf)'	6

Figura 41: Vista de valores de Frecuencia

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Selected attribute		
Name: Mon		Type: Nominal
Missing: 0 (0%)	Distinct: 4	Unique: 1 (0%)
No.	Label	Count
1	'(-inf-2100]'	197
2	'(2100-4078]'	18
3	'(4078-6056]'	3
4	'(6056-8034]'	0
5	'(8034-inf)'	1

Figura 42: Vista de valor Monetario

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Se ha realizado la discretización de los valores en la herramienta weka, sin embargo se debe considerar que en el cuarto bin del atributo Mon, no existen valores que se encuentren dentro de este rango, razón por lo cual posteriormente no se presentará el valor de 4 al reemplazar los valores por los puntos del RFM en este campo.

4.3.3 INTEGRACIÓN DE LOS DATOS

Para continuar con el análisis RFM ha sido necesario la creación de campos adicionales, se ha reemplazado los campos Req, Freq y Mon por sus valores puntos (1, 2, 3, 4 y 5), como resultado se ha obtenido los campos R, F y M; además el valor resultante RFM_puntos.

ID_CLIENTE	Req	Freq	Mon	RFM_puntos
23	4	5	1	3.33333333
215	3	3	1	2.33333333
46	4	4	1	3
192	4	2	1	2.33333333
69	5	1	1	2.33333333
92	5	2	1	2.66666667
115	5	4	1	3.33333333
209	1	4	1	2
138	1	3	1	1.66666667
161	5	4	5	4.66666667
29	4	3	1	2.66666667
175	5	3	1	3
169	1	5	1	2.33333333
221	5	1	1	2.33333333
75	1	3	1	1.66666667
201	1	3	1	1.66666667
132	5	3	1	3
9	2	4	2	2.66666667
15	2	2	1	1.66666667

Figura 43: Integración Datos con Análisis RFM

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Weka solicita para la técnica de asociación que los datos a los cuales se aplicarán las reglas en este caso los productos se encuentren cada uno como columnas y como filas las transacciones realizadas.

TRANSACCION	Crema de Le	Crema De Le	Crema de Le	Crema de le	Crema De Le	Crema De Q	Crema Leche Dulce de Lec	Gelatina	Leche Baja e	Leche Conde
CLIENTE:100FECHA:20140224										
CLIENTE:100FECHA:20141027						1				
CLIENTE:100FECHA:20150907										
CLIENTE:101FECHA:20140505										
CLIENTE:101FECHA:20150105										
CLIENTE:101FECHA:20151026	1					1				
CLIENTE:102FECHA:20140317										
CLIENTE:102FECHA:20141117										
CLIENTE:102FECHA:20150921										
CLIENTE:103FECHA:20140512										
CLIENTE:103FECHA:20150119						1				
CLIENTE:103FECHA:20151026			1	1			1			1
CLIENTE:104FECHA:20140519										1
CLIENTE:104FECHA:20150119										
CLIENTE:104FECHA:20150518								1		1
CLIENTE:104FECHA:20151102	1		1	1			1			
CLIENTE:105FECHA:20141110								1		1
CLIENTE:105FECHA:20150921						1				
CLIENTE:106FECHA:20140519										

Figura 44: Columnas y Filas Transacciones Realizadas

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

4.3.4 FORMATEO DE LOS DATOS

Para la asociación de productos es necesario incluir un cambio de numérico a nominal, el resto de valores gracias al datamart y a la preparación de los datos se encuentran listos para la fase de modelado.

4.4 FASE DE MODELADO

4.4.1 SELECCIÓN DE LA TÉCNICA DE MODELADO

Los algoritmos que se utilizarán para la fase de modelado son los siguientes, esta selección se basa en el trabajo de investigación denominado “Data Mining Using RFM Analysis” de Derya Birant:

- Para la segmentación de clientes se utilizará el algoritmo de Simple K Means de Weka que es un algoritmo de cluster que reúne en tablas distintas atributos con semejanza, a estos los divide en la cantidad de clúster seleccionados en este caso ocho debido a que los clientes según el trabajo mencionado, se dividen en: Mejores, Valiosos, Compradores, Iniciales, Perdidos, Frecuentes, Gastadores e Inciertos.
- Para la asociación de productos es necesario aplicar el algoritmo Predictive Apriori con el objetivo de combinar las dos medidas: de soporte y confianza y proporcionar una mejor precisión del modelo.

4.4.2 GENERACIÓN DEL PLAN DE PRUEBA

Una vez construido un modelo, se deberá probar la validez del mismo, para ello se utilizará el componente Clusterer Performance Evaluator de Weka para medir el error del modelo. Para las reglas de asociación se tomarán en cuenta las primeras 20 reglas de asociación siempre y cuando la precisión sea superior al 0,95.

4.4.3 CONSTRUCCIÓN DEL MODELO

Se ha construido en la herramienta Weka los siguientes flujos de conocimiento para la segmentación y asociación.

Flujo de Conocimiento para la Segmentación o Clustering

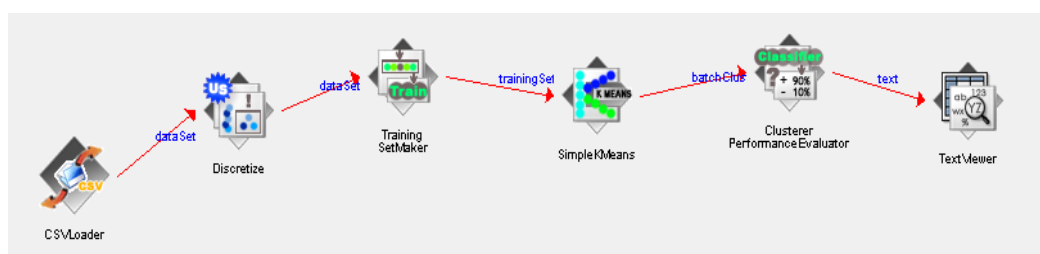


Figura 45: Flujo Conocimiento Segmentación o Clustering

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Flujo de Conocimiento para la Asociación

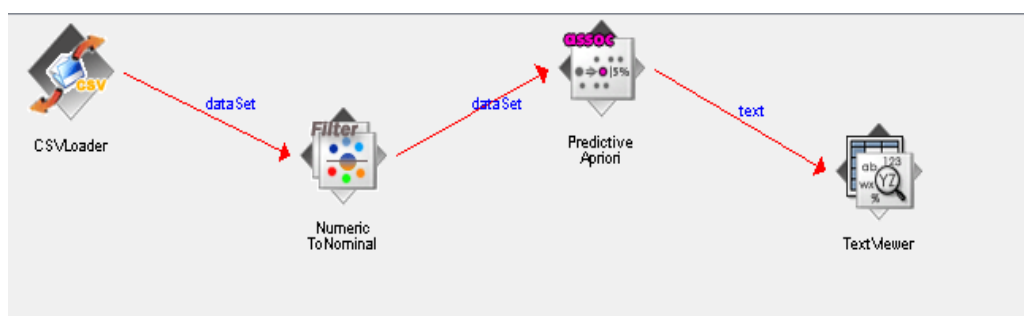


Figura 46: Flujo Conocimiento Asociación

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

4.4.4 EVALUACIÓN DEL MODELO

Como se mencionó en la sección de la Generación del Plan de Pruebas se obtiene los siguientes resultados de la evaluación.

SEGMENTACIÓN

```
kMeans
=====

Number of iterations: 2
Within cluster sum of squared errors: 0.0
Missing values globally replaced with mean/mode
```

Figura 47: Resultado de Error Segmentación

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Este resultado de error de la suma de los cuadrados que muestra como resultado 0,0 es debido a la discretización realizada a partir de los datos de la columna RFM_puntos. Posteriormente en la Fase de Evaluación del modelo se discutirá de mejor manera este resultado, sin embargo se puede mencionar que la precisión del modelo es alto.

ASOCIACIÓN

Se ha generado 20 reglas de las cuales todas superan la precisión del 95%. En el siguiente gráfico se muestra la regla N° 20 la cual es el último registro y posee una precisión del 97%, todas las restantes tienen una precisión superior a esta regla.

```
20. Crema de Leche=1 Queso Holandés=1 36 ==> Queso Mesa=1 Queso Ricotta =1 36 acc:(0.97338)
```

Figura 48: Regla N° 20 Asociación

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

4.5 FASE DE EVALUACIÓN

4.5.1 EVALUACIÓN DE LOS RESULTADOS

Los resultados generados a partir de la aplicación del modelo son los siguientes:

Para la segmentación de clientes posterior al análisis RFM se ha obtenido la clusterización de los clientes en ocho partes que se muestra a continuación:

```
kMeans
=====

Number of iterations: 2
Within cluster sum of squared errors: 0.0
Missing values globally replaced with mean/mode

Cluster centroids:
Attribute          Full Data          Cluster#
                   (219)              (57)              1              2              3
                   (35)              (71)              (38)
-----
RFM_puntos         '(2.583333-3)' '(2.166667-2.583333)' '(-inf-1.75)' '(2.583333-3)' '(1.75-2.166667)'
```

Figura 49: Parte 1-1 Clusterización

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

4	5	6	7
(3)	(2)	(12)	(1)
'(3.416667-3.833333]'	'(3.833333-4.25]'	'(3-3.416667]'	'(4.25-inf]'

Figura 50: Parte 1-2 Clusterización

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

En el gráfico se muestra los rangos del modelo a los cuales RFM_puntos ha clasificado a los clientes.

Clustered Instances	
0	57 (26%)
1	35 (16%)
2	71 (32%)
3	38 (17%)
4	3 (1%)
5	2 (1%)
6	12 (5%)
7	1 (0%)

Figura 51: Porcentaje de Clusters

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Se puede observar los porcentajes en cada uno de los clústers, se puede destacar al clúster 7 que es el que mejores clientes, posee únicamente uno el cuál no llega a representar el 1%; en contraste el clúster 2 es en donde se encuentran clasificados un mayor número con un total de 71 que corresponde al 32%.

Para la asociación se muestran las reglas obtenidas las cuales se encuentran acorde a lo esperado las mismas que poseen una precisión superior al 95%.

1. Queso Crema=1 Queso Parmesano=1 62 ==> Queso Crema con Hierbas=1 62 acc:(0.98386)
2. Queso Campesino=1 Queso Mesa=1 61 ==> Queso Holandés=1 61 acc:(0.98362)
3. Queso Crema con Hierbas=1 Queso Finesse tajadas=1 57 ==> Queso Crema=1 57 acc:(0.98258)
4. Queso Campesino=1 Queso Crema con Hierbas=1 56 ==> Queso Parmesano=1 56 acc:(0.98229)
5. Crema De Queso Flotemys=1 Queso Finesse=1 52 ==> Queso Mozzarella =1 52 acc:(0.98105)
6. Crema De Queso Flotemys=1 Queso Camembert Pierrot=1 51 ==> Queso Gudbrandsdals=1 51 acc:(0.98071)
7. Queso Camembert Pierrot=1 Queso Mascarpone Fabioli=1 51 ==> Queso Courdavault=1 51 acc:(0.98071)
8. Queso Courdavault=1 Queso Gorgonzola Telino=1 49 ==> Queso Mascarpone Fabioli=1 49 acc:(0.97999)
9. Crema de Leche=1 Crema de leche UHT Finesse=1 48 ==> Crema de Leche UHT=1 48 acc:(0.9796)
10. Crema de Leche=1 Queso Mesa=1 47 ==> Queso Ricotta =1 47 acc:(0.9792)
11. Crema De Leche Larga Vida Sancor=1 Yogurt de platanos=1 47 ==> Yogurt saborizado=1 47 acc:(0.9792)
12. Crema de Leche UHT=1 Crema Leche cantina=1 47 ==> Crema de leche UHT Finesse=1 47 acc:(0.9792)
13. Queso Campesino=1 Queso Ricotta =1 46 ==> Queso Holandés=1 46 acc:(0.97878)
14. Crema De Queso Flotemys=1 Queso Courdavault=1 43 ==> Queso Camembert Pierrot=1 43 acc:(0.97742)
15. Crema de Leche UHT=1 Queso Ricotta =1 42 ==> Crema de Leche=1 42 acc:(0.97692)
16. Dulce de Leche=1 Mantequilla de Nata Cruda=1 41 ==> Mantequilla Batida =1 41 acc:(0.9764)
17. Queso Crema con Hierbas=1 Queso Finesse=1 40 ==> Queso Crema=1 40 acc:(0.97586)
18. Crema de leche UHT Finesse=1 Leche de Chocolate=1 38 ==> Crema Leche cantina=1 38 acc:(0.97468)
19. Crema De Queso Flotemys=1 Queso Finesse tajadas=1 38 ==> Queso Finesse=1 38 acc:(0.97468)
20. Crema de Leche=1 Queso Holandés=1 36 ==> Queso Mesa=1 Queso Ricotta =1 36 acc:(0.97338)

Figura 52: Asociación Productos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Por ejemplo se puede observar que se pueden asociar productos como cuando un cliente compra Queso Crema y Parmesano también se le puede ofrecer el Queso Crema con Hierbas, y además se puede incluir esta combinación en una cesta de compra.

4.5.2 PROCESO DE REVISIÓN

Si se realiza un análisis de revisión del proyecto de Minería de Datos para la Empresa de Lácteos Santillán se puede evidenciar que se ha obtenido resultados satisfactorios, a pesar de las limitantes de la base de datos proporcionada la cual no permitió realizar una clasificación de perfiles del clientes y está será una de las recomendaciones que se realizará al Gerente de la Empresa. Sin embargo al conocer está limitante en la sección de evaluación de la situación inicial, los objetivos de la minería de datos han sido cubiertos completamente.

Relacionado a la precisión de los algoritmos particularmente en el algoritmo para la segmentación se observó un dato de error de cero, este fue obtenido al discretizar el valor de RFM_puntos debido a que el modelo respondía a un error del 0,076 como se puede observar en el siguiente gráfico.

```

kMeans
=====

Number of iterations: 2
Within cluster sum of squared errors: 0.07638095257592387
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute      Full Data      Cluster#
                (219)          0          1          2          3          4          5          6          7
=====
RFM_puntos     2.4094         2.3333     1.6095         3          2          2.6667     3.6667     4.2222     3.3333

Clustered Instances

0      57 ( 26%)
1      35 ( 16%)
2      28 ( 13%)
3      38 ( 17%)
4      43 ( 20%)
5       3 (  1%)
6       3 (  1%)
7      12 (  5%)

```

Figura 53: Dato Error obtenido al discretizar valor RFM_puntos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Este modelo fue mejorado cuando se aplicó la discretización, y además proporcionó una idea clara de cuáles clientes se encontraban en cada clúster para posteriormente la gerencia aplique campañas de marketing estratégico.

4.5.3 DETERMINACIÓN DE FUTURAS FASES

Al cumplir los objetivos de la minería de datos no existe inconveniente en continuar con la fase de implementación, que está enfocada a transformar en conocimiento lo encontrado en el proceso de minería de datos y sugerir a la gerencia aplicarlos al negocio.

4.6 FASE DE IMPLEMENTACIÓN

4.6.1 PLAN DE IMPLEMENTACIÓN

El procedimiento que se sugiere para aplicar los resultados obtenidos de la Minería de Datos en la Empresa Santillán es el siguiente:

1. En este procedimiento de minería de datos se ha creado un datamart adhoc que sirve para no interrumpir las actividades normales del sistema debido a la carga de procesamiento que se requirió, se sugiere que el datamart de ventas sea implementado posterior a la adquisición de las licencias de las herramientas respectivas.
2. El preprocesamiento se ha realizado con la herramienta Weka esta no posee la funcionalidad para realizar un RFM automáticamente por esta razón es necesario a partir de vistas de la base de datos, generar la discretización en 5 partes, mapear los valores obtenidos de la discretización a valores numéricos funcionalidad que no se ha encontrado en la herramienta Weka, es por esta razón que este procedimiento se realizó generando un archivo .CSV en el cual se aprovechó para realizar los cálculos respectivos obteniendo el campo RFM_puntos y cargando consecuentemente esta data al flujo de conocimiento para la segmentación creado.
3. Para la asociación es necesario trasladar la información de la vista realizada en SQL de tal forma que los productos se encuentren como columnas y que las transacciones se encuentren como filas generando una tabla dinámica la cual se expandirá por el número de transacciones, posteriormente se crea un archivo .CSV para la carga de la data al flujo de conocimiento de asociación creado.

Para el mapeo de valores y la generación de la tabla dinámica se ha hecho uso de la herramienta Excel.

4.6.2 MONITORIZACIÓN Y MANTENIMIENTO

Se sugiere realizar una retroalimentación del conocimiento generado por el modelo cada mes con el objetivo de garantizar que el mismo está siendo utilizado adecuadamente, y una nueva implementación de la estrategia de marketing dependiendo de los recursos de la empresa al menos cada año.

Además es necesario posterior a la implementación de las estrategias de marketing analizar cómo van creciendo las ventas o a su vez si han disminuido, con la finalidad de revisar el modelo y ajustarlo, este procedimiento se sugiere realizarlo al menos una vez al mes.

4.6.3 INFORME FINAL

El objetivo general de este proyecto fue el de analizar la información a través de la aplicación de técnicas de minería de datos enfocadas a descubrir patrones que permitan apoyar a la toma de decisiones enfocadas a la inteligencia empresarial por parte de la gerencia y así incrementar la rentabilidad de la empresa, en relación a esto a continuación se describen como se han cumplido los objetivos específicos para alcanzar la meta propuesta:

- 1) Determinar que técnicas de minería de datos se puede aplicar en función de la información que posee la empresa.
 - a. Con la base de datos proporcionada por la empresa y una vez realizada la preparación de los datos se ha realizado la segmentación de clientes y la asociación de productos, debido a que no se han presentado datos demográficos de los clientes no se ha realizado la clasificación de los perfiles de los mismos.
 - b. Se ha utilizado el algoritmo de Simple K means para segmentar a los clientes en 8 grupos vinculándolos con la tipología de los clientes.

c. Para la asociación de productos se ha utilizado el algoritmo de Predictive Apriori que proporciona una mayor precisión del modelo al funcionar dos medidas la de soporte y confianza.

2) Aplicar un análisis RFM.

a. Se ha aplicado un análisis RFM para la segmentación de clientes debido a que es una herramienta poderosa para el marketing estratégico y además porque no se han presentado datos demográficos de los clientes como se había mencionado anteriormente que hubiesen permitido segmentar los clientes sin necesidad de un análisis RFM, sin embargo este análisis presenta una mejor funcionalidad.

3) Establecer distintos grupos entre las personas que realizan una compra más a menudo y así generar estrategias de marketing enfocadas al incremento de las ventas.

a. Para cumplir este objetivo se ha utilizado el campo de RFM_puntos junto con el algoritmo para la realización de clustering, en el cuál se ha obtenido 8 segmentos que corresponden a los tipos de clientes: Mejores, Valiosos, Compradores, Iniciales, Perdidos, Frecuentes, Gastadores e Inciertos.

b. Las estrategias que se sugiere a la gerencia pueden estar enfocadas a los siguientes puntos:

- A aquellos clientes que se encuentren en el nivel de compromiso con la compañía superior 333 RFM puntos se puede realizar: Muestras de aprecio a través de campañas exclusivas, proporcionarles descuentos especiales, realizar encuestas de satisfacción, entrevistas, muestras de productos gratis, acciones que recompensen su lealtad como llamadas para eventos de la empresa, entre otros.

- A aquellos que se encuentran en la mitad, 233 a 332 RFM puntos se debe motivarles con: Descuentos atractivos dependiendo de su posición en la tabla con un porcentaje de acuerdo a su orden de puntuación, sorteos, educación para el mejor aprovechamiento de los productos y entrega de guías prácticas.
- A aquellos que se encuentran en la parte inferior 133 a 232 RFM puntos: Descuentos atractivos focalizados en el tiempo a un crecimiento de compra, encuestas de satisfacción para identificar problemas.
- A los clientes que se encuentren en un rango inferior a 133 dependiendo de los recursos para marketing de la empresa se podrá realizar cualquiera de los mencionados en el rango de 133 a 232 RFM puntos, sin embargo de acuerdo a este análisis no son prioritarios.

4) Crear asociaciones de productos para definir una cesta de compra que permitirá crear predicciones para recomendar productos a los clientes.

- a. Se ha creado veinte reglas de asociación para recomendar productos a los clientes a través del uso del algoritmo de Predictive Apriori, el mismo que proporciona asociaciones con un nivel de precisión alto. Es necesario aprovechar estas reglas para la creación de cestas compra con descuento, además sugerir en el momento de la venta estos productos asociados.

Estas técnicas de marketing sugeridas son consideradas como estratégicas para poder incrementar las ventas de la empresa y serán de responsabilidad de la gerencia, que deberá tomar la decisión en virtud al conocimiento adquirido a través de la minería

de datos realizada cuál de estas técnicas aplicarlas a corto, medio y largo plazo, dependiendo de sus recursos y de su planificación estratégica.

4.6.4 REVISIÓN DEL PROYECTO

El proyecto de minería de datos se ha realizado de forma satisfactoria para ellos se ha seguido la metodología CRISP-DM garantizando la calidad del procedimiento, se ha podido identificar sin embargo ciertos aspectos que se debe mejorar: Es sumamente importante que se incluyan los datos demográficos en el procesamiento diario de los clientes con el objetivo de poder clasificarlos y predecir cuáles son los sectores en donde se podría encontrar nuevos clientes, el campo e-mail es imprescindible debido a que con ello las sugerencias de marketing estratégico se puede realizar por este medio, además que con los mismos se podrá entablar una comunicación más estrecha.

Se sugiere implementar un datamart continuo en la empresa con el objetivo de contrastar la información histórica y facilitar la implementación continua de un proceso de minería de datos.

CAPÍTULO V

METODOLOGIA

- **Método Científico.-** Se utiliza en la realidad de los hechos de la empresa.
- **Método Inductivo.-** Se utiliza en la etapa de observación y registro de los hechos.
- **Método Bibliográfico.-** Son las fuentes más importantes de donde se obtengan información y documentación como: códigos, libros, datos, etc.
- **Método Analítico.-** Se utilizará en la comparación de las dos herramientas de nuestra investigación luego para seleccionar la mejor herramienta de acuerdo a los indicadores y sus resultados.

5.1 TIPO DE ESTUDIO

5.1.1 SEGÚN EL OBJETO DE ESTUDIO

Investigación Aplicada.

5.1.2 SEGÚN LA FUENTE DE INFORMACIÓN

Investigación Documental.

5.1.3 SEGÚN EL NIVEL DE CONOCIMIENTOS

Investigación Descriptiva.

5.1.4 SEGÚN EL MÉTODO A UTILIZAR

Investigación Comparativa

5.2 POBLACIÓN Y MUESTRA

Plataforma Weka y Analysis Services

5.3 OPERACIONALIZACIÓN DE VARIABLES

VARIABLE	TIPO DE VARIABLE	CONCEPTO	DIMENSIÓN	INDICADORES
Análisis comparativo de plataformas de minería de datos	Independiente	Análisis comparativo de las plataformas de minería de datos Weka y Microsoft Analysis Services basado en indicadores tomados de la conceptualización de las fases y tareas de la Metodología	Fases, subfases o tareas, e indicadores tomados de la minería de datos.	<ul style="list-style-type: none"> • Porcentaje de cumplimiento de las fases. • Puntuación de las subfases o tareas. • Soporte de indicadores seleccionados tomados de la conceptualización de la Metodología CRISP-DM.

		CRISP-DM que se aplican al entorno de la herramienta de software.		
Desarrollo de Minería de Datos	Dependiente	El desarrollo de la minería de datos está basado en la las Fases y Subfases de la Metodología CRISP-DM.	Procedimiento para el desarrollo de la minería en base a la metodología CRISP-DM	<ul style="list-style-type: none"> • Porcentaje de optimización por fases. • Porcentaje de optimización por Subfases o tareas.

Tabla 13: Operacionalización de Variables

Fuente:(Nay Mojarrango & José Chapalbay, 2015)

5.4 PROCEDIMIENTOS

5.4.1 TÉCNICA DE INVESTIGACIÓN

Técnica Documental.- Permite la recopilación de datos para enunciar las distintas teorías que respaldarán la investigación.

5.4.2 INSTRUMENTOS DE RECOLECCIÓN DE DATOS

- Entrevistas
- Diálogos
- Observación

5.5 PROCEDIMIENTO Y ANÁLISIS

Los datos que serán obtenidos, se cuantificaran mediante hojas electrónicas, y se podrán emitir conclusiones y recomendaciones en base a la investigación realizada.

CAPÍTULO VI

RESULTADOS Y DISCUSIÓN

6.1 RESULTADOS

6.1.1 FASE DE COMPRENSIÓN DE DATOS

6.1.1.1 RECOLECCIÓN DE DATOS INICIALES

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	100%	80%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	4,002	3,202

Tabla 14: Resultados de la Sub Fase de Recolección de Datos Iniciales

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

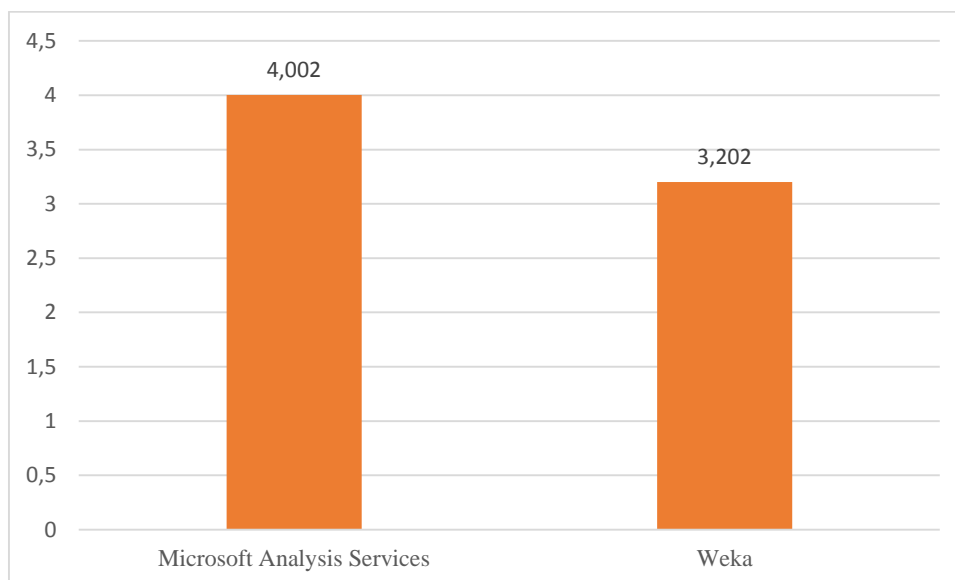


Figura 54: Ponderación Empírica de tiempo por fases en Minería de Datos-Recolección de Datos Iniciales

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Recolección de Datos Iniciales existe una superioridad de Microsoft Analysis Services cumpliendo con el indicador al 100% con relación a Weka que cumple un 80% del mismo, la ponderación empírica del tiempo corresponde a una diferencia de 0,8 entre las dos herramientas.

6.1.1.2 DESCRIPCIÓN DE LOS DATOS

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	100%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	2,801	2,801

Tabla 15: Resultados de la Sub Fase de Descripción de los datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

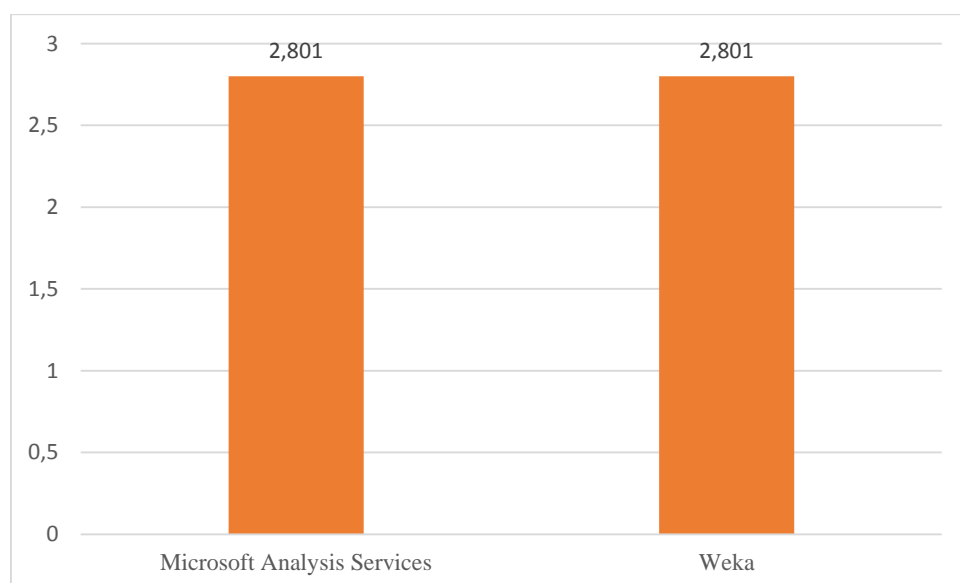


Figura 55: Ponderación Empírica de tiempo por fases en Minería de Datos- Descripción de los datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Descripción de los datos se observa que las dos herramientas de Microsoft Analysis Services y Weka cumplen con el indicador al 100%, la ponderación empírica del tiempo corresponde a 2,8 de las dos herramientas.

6.1.1.3 EXPLORACIÓN DE LOS DATOS

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	0%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	0,000	2,495

Tabla 16: Resultados de la Sub Fase de Exploración de los Datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

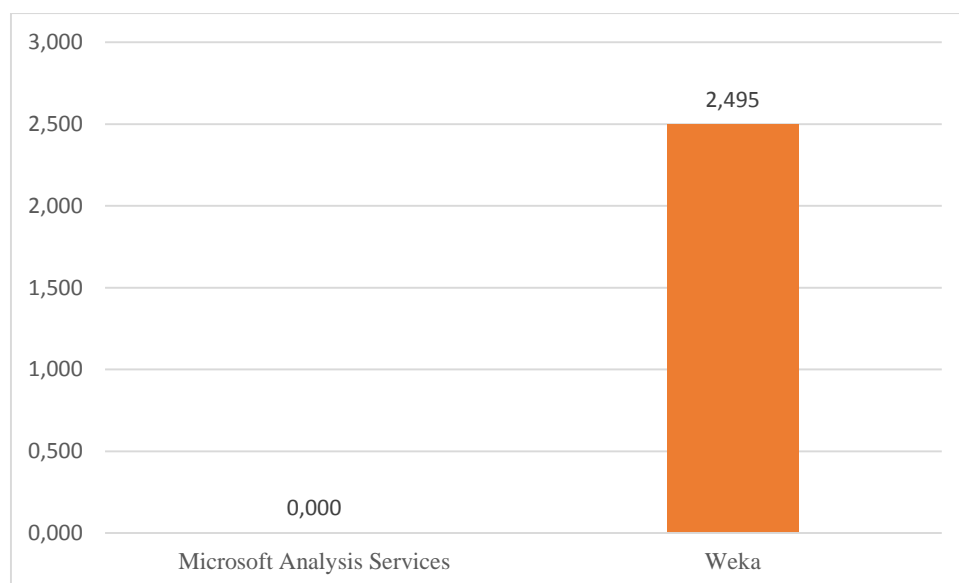


Figura 56: Ponderación Empírica de tiempo por fases en Minería de Datos- Exploración de los datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Exploración de los datos se observa que la herramienta Weka cumple con el indicador al 100% con relación a Microsoft Analysis Services que no cumple el mismo, la ponderación empírica del tiempo corresponde a una diferencia de 2,46 entre las dos herramientas.

6.1.1.4 VERIFICACIÓN DE LA CALIDAD DE LOS DATOS

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	100%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	1,536	1,536

Tabla 17: Resultados de la Sub Fase de Verificación de la Calidad de los Datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

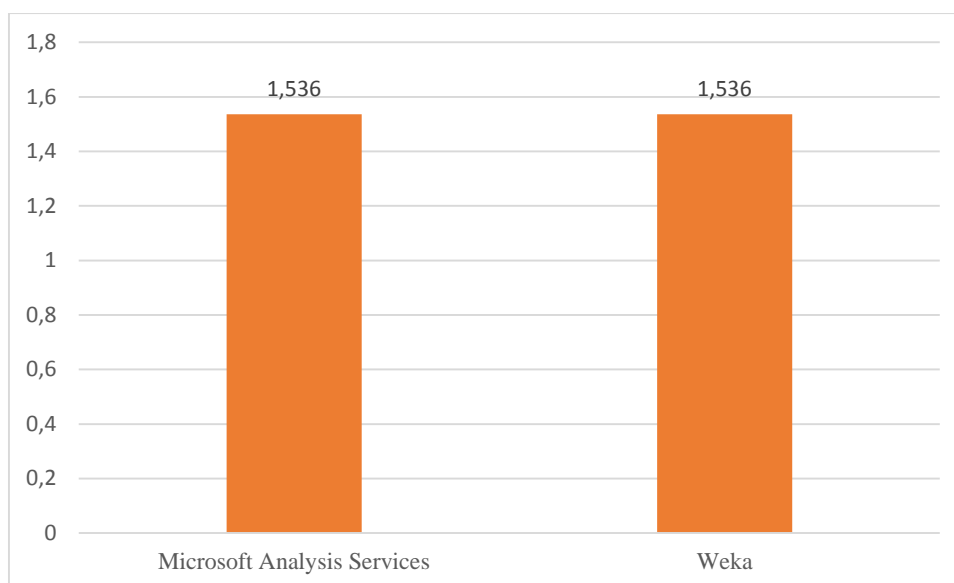


Figura 57: Ponderación Empírica de tiempo por fases en Minería de Datos- Verificación de la calidad de los datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Verificación de la calidad de los datos se observa que las dos herramientas de Microsoft Analysis Services y Weka cumplen con el indicador al 100%, la ponderación empírica del tiempo corresponde a 1,54 de las dos herramientas.

6.1.2 FASE DE PREPARACIÓN DE LOS DATOS

6.1.2.1 SELECCIÓN DE DATOS

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	100%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	6,537	6,537

Tabla 18: Resultados de la Sub Fase de Selección de Datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

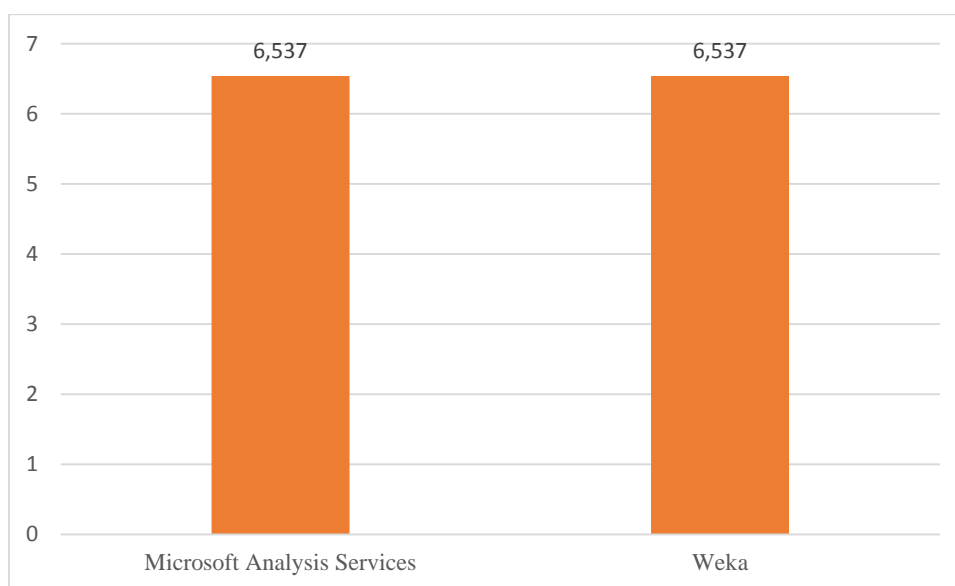


Figura 58: Ponderación Empírica de tiempo por fases en Minería de Datos- Selección de datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Selección de datos se observa que las dos herramientas de Microsoft Analysis Services y Weka cumplen con el indicador al 100%, la ponderación empírica del tiempo corresponde a 6,54 de las dos herramientas.

6.1.2.2 LIMPIEZA DE DATOS

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	50%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	0,801	1,602

Tabla 19: Resultados de la Sub Fase de Limpieza de Datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

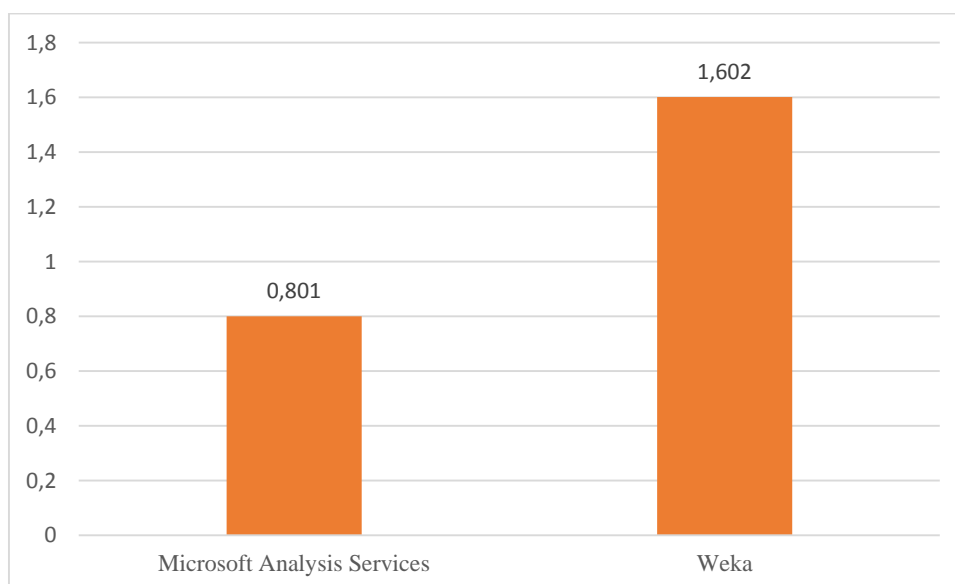


Figura 59: Ponderación Empírica de tiempo por fases en Minería de Datos- Limpieza de datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Limpieza de datos se observa que la herramienta Weka cumple con el indicador al 100% con relación a Microsoft Analysis Services que cumple con un 50%, la ponderación empírica del tiempo corresponde a una diferencia de 0,8 entre las dos herramientas.

6.1.2.3 ESTRUCTURACIÓN DE LOS DATOS

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	100%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	1,388	1,388

Tabla 20: Resultados de la Sub Fase de Estructuración de los Datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

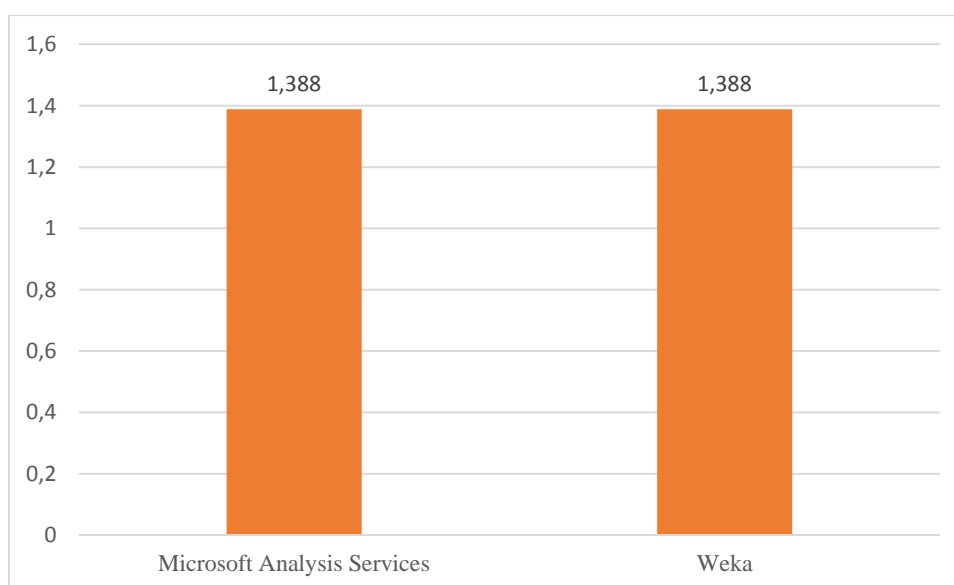


Figura 60: Ponderación Empírica de tiempo por fases en Minería de Datos- Estructuración de los datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Estructuración de los datos se observa que las dos herramientas de Microsoft Analysis Services y Weka cumplen con el indicador al 100%, la ponderación empírica del tiempo corresponde a 1,39 de las dos herramientas.

6.1.2.4 INTEGRACIÓN DE LOS DATOS

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	67%	67%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	0,454	0,454

Tabla 21: Resultados de la Sub Fase de Integración de los Datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

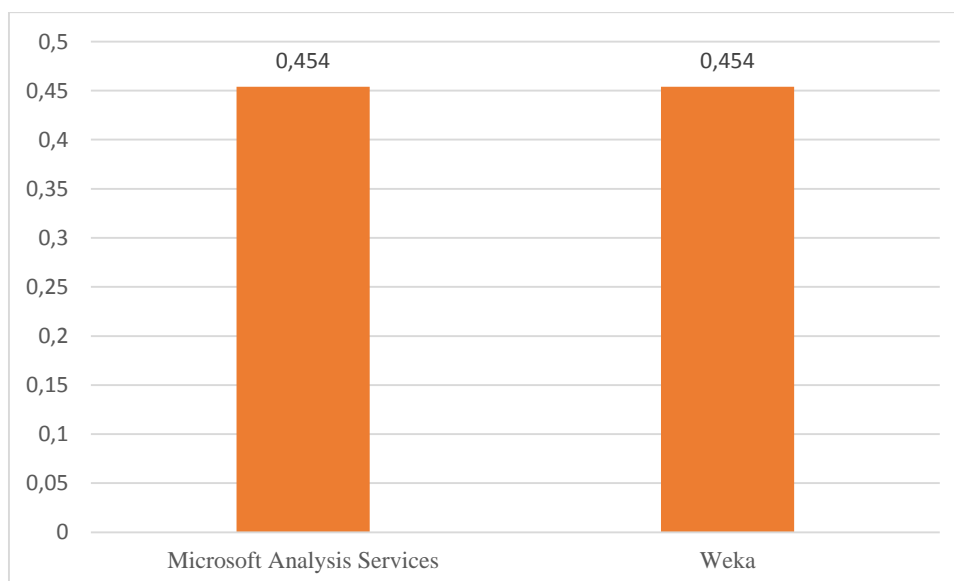


Figura 61: Ponderación Empírica de tiempo por fases en Minería de Datos- Integración de los datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Integración de los datos se observa que las dos herramientas de Microsoft Analysis Services y Weka cumplen con el indicador al 67%, la ponderación empírica del tiempo corresponde a 0,45 de las dos herramientas.

6.1.2.5 FORMATEO DE LOS DATOS

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	100%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	5,404	5,404

Tabla 22: Resultados de la Sub Fase de Formateo de los Datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

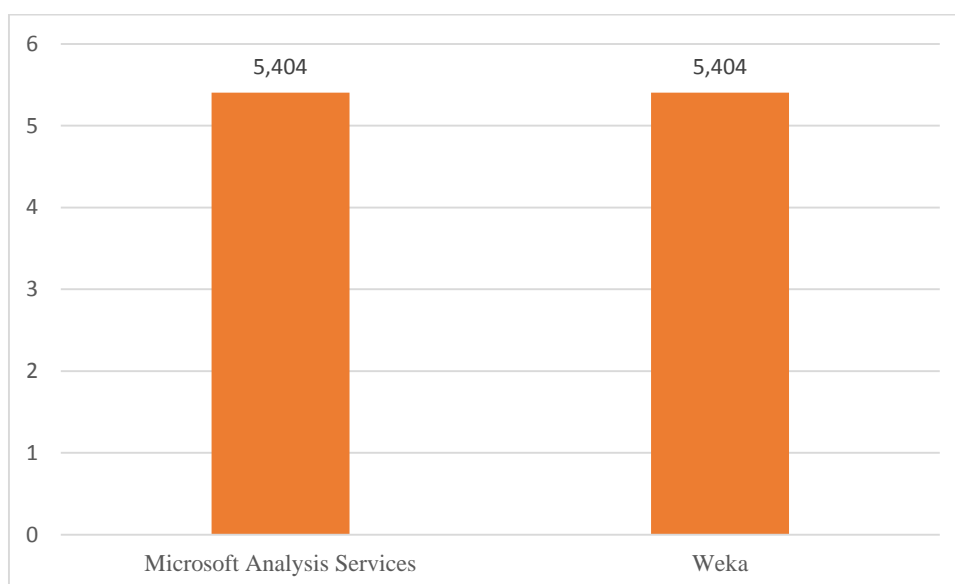


Figura 62: Ponderación Empírica de tiempo por fases en Minería de Datos- Formateo de los datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Formateo de los datos se observa que las dos herramientas de Microsoft Analysis Services y Weka cumplen con el indicador al 100%, la ponderación empírica del tiempo corresponde a 5,4 de las dos herramientas.

6.1.3 FASE DE MODELADO

6.1.3.1 SELECCIÓN DE LA TÉCNICA DE MODELADO

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	100%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	5,402	5,402

Tabla 23: Resultados de la Sub Fase de Selección de la Técnica de Modelado

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

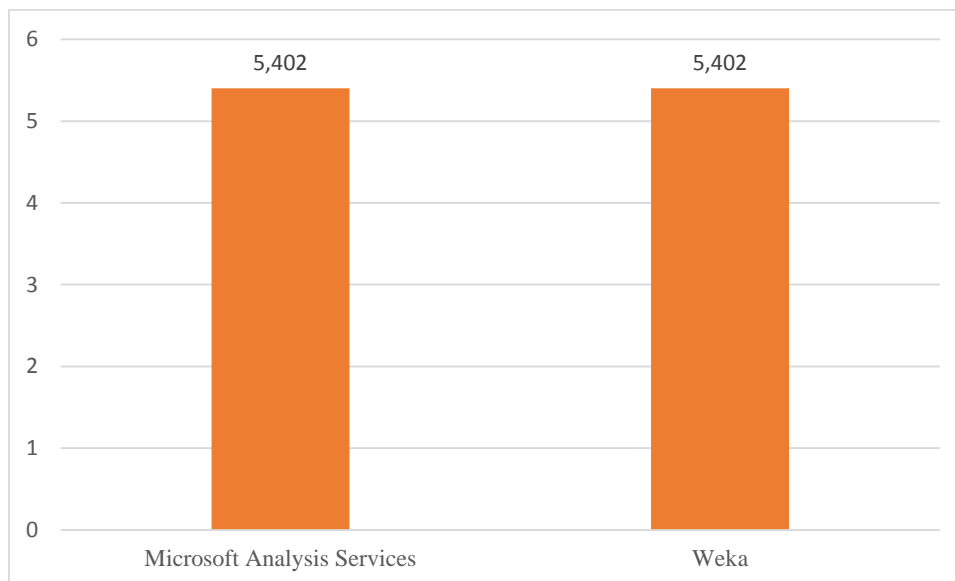


Figura 63: Ponderación Empírica de tiempo por fases en Minería de Datos- Selección de la técnica de modelado

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Selección de la técnica de modelado se observa que las dos herramientas de Microsoft Analysis Services y Weka cumplen con el indicador al 100%, la ponderación empírica del tiempo corresponde a 5,4 de las dos herramientas.

6.1.3.2 CONSTRUCCIÓN DEL MODELO

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	67%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	14,445	21,668

Tabla 24: Resultados de la Sub Fase de Construcción del Modelo

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

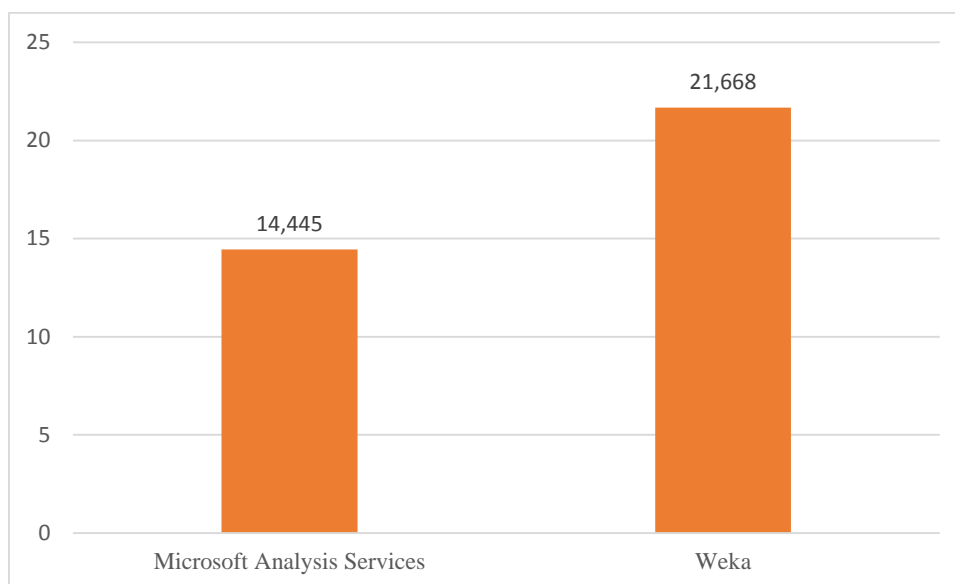


Figura 64: Ponderación Empírica de tiempo por fases en Minería de Datos- Construcción del modelo

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Construcción del modelo existe una superioridad de Weka cumpliendo con el indicador al 100% con relación a Microsoft Analysis Services que cumple un 67% del mismo, la ponderación empírica del tiempo corresponde a una diferencia de 7,22 entre las dos herramientas.

6.1.3.3 EVALUACIÓN DEL MODELO

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	100%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	4,590	4,590

Tabla 25: Resultados de la Sub Fase de Evaluación del Modelo

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

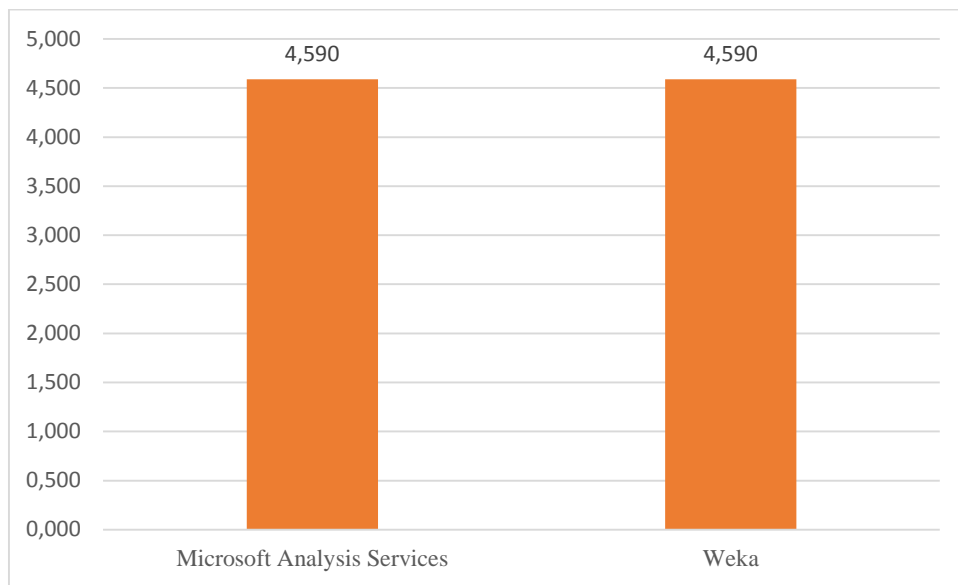


Figura 65: Ponderación Empírica de tiempo por fases en Minería de Datos- Evaluación del modelo

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Evaluación del modelose observa que las dos herramientas de Microsoft Analysis Services y Weka cumplen con el indicador al 100%, la ponderación empírica del tiempo corresponde a 4,59 de las dos herramientas.

6.1.4 FASE DE IMPLEMENTACIÓN

6.1.4.1 INFORME FINAL

	Microsoft Analysis Services	Weka
% de Cumplimiento del Indicador	100%	100%
Resultado de la Ponderación Empírica de tiempo por fases en Minería de datos	6,152	6,152

Tabla 26: Resultados de la Sub Fase de Informe Final

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

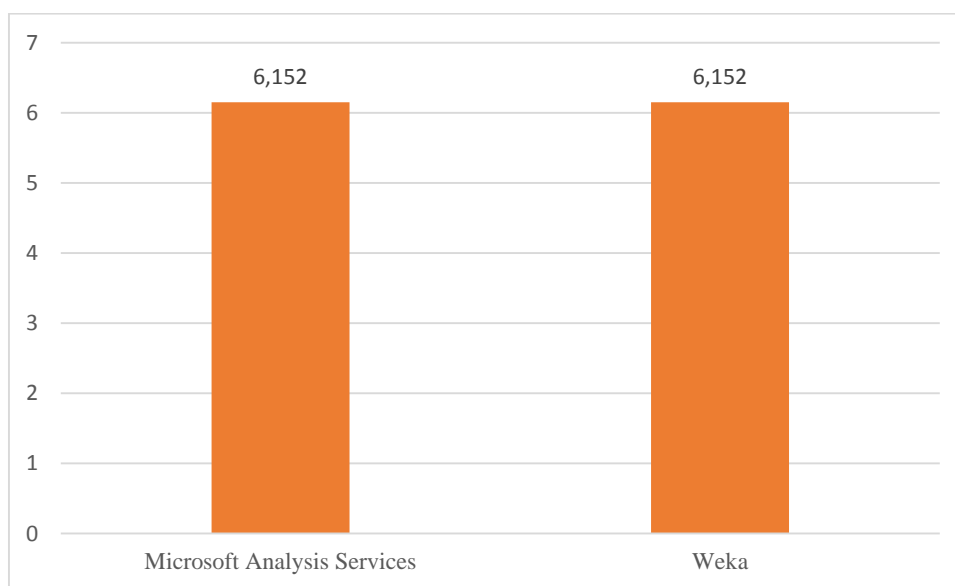


Figura 66: Ponderación Empírica de tiempo por fases en Minería de Datos- Informe Final

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

INTERPRETACIÓN:

En la Sub Fase de Informe Final se observa que las dos herramientas de Microsoft Analysis Services y Weka cumplen con el indicador al 100%, la ponderación empírica del tiempo corresponde a 6,15 de las dos herramientas.

6.2 COMPROBACIÓN DE LA HIPÓTESIS

A continuación se plantea la Hipótesis estadística de la investigación, la hipótesis Nula H_0 y la hipótesis alternativa H_1 .

- **H_0** =No hay diferencia significativa en las medidas del tiempo de la herramienta Weka y Analysis Services.
- **H_1** =Las medidas del tiempo de la herramienta Weka es mayor a las medidas de tiempo de la herramienta Microsoft Analysis Services.
- **Alfa**=0,05

Al considerar un estudio longitudinal de dos medidas, la variable de comparación es una variable numérica, y el número de parámetros es menor a 30 al tratarse de 13 indicadores de comprobación, por lo que se aplicará la prueba T student.

El siguiente análisis se realiza a través del tiempo, el porcentaje que conlleva el desarrollo de la minería de datos con la metodología CRISP-DM, está ponderado basado en el trabajo de investigación denominado “Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información” de Rodríguez, D., Pollo-Cattaneo, F., Britos, P., García-Martínez, R. es así que esta ponderación es la siguiente:

Fases	% de Tiempo
Fase de Comprensión del Negocio o Problema	20.7
Fase Comprensión de los Datos	10.9
Fase de preparación de los datos	15.61
Fase de modelado	34.41
Fase de evaluación	7.41
Fase de implantación	10.93

Tabla 27: Fases & Porcentaje de Tiempo

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Para las sub-fases es la siguiente ponderación:

Fases	Sub Fases	% de Tiempo
Comprensión del Negocio o Problema	Determinar los objetivos del negocio	35.59
	Evaluación de la situación	32.24
	Determinar los objetivos de la Minería de Datos	12.83
	Realizar el plan del proyecto	19.34
Comprensión de los Datos	Recolección de Datos iniciales	36.72
	Descripción de los Datos	25.7
	Exploración de los datos	22.89
	Verificación de la calidad de los datos	14.09
Fase de preparación de los datos	Selección de datos	41.88
	Limpieza de datos	10.26
	Estructuración de los datos	8.89
	Integración de los datos	4.36
	Formateo de los datos	34.62
Fase de modelado	Selección de la técnica de modelado	15.7
	Generación del plan de prueba	7.99
	Construcción del modelo	62.97
	Evaluación del modelo	13.34

Fases	Sub Fases	% de Tiempo
Fase de evaluación	Evaluación de los resultados	50.36
	Proceso de revisión	21.51
	Determinación de futuras fases	28.14
Fase de implantación	Plan de implementación	17.7
	Monitorización y Mantenimiento	12.82
	Informe Final	56.29
	Revisión del proyecto	13.19

Tabla 28: Ponderación, Fases, Sub Fases y Porcentaje de Tiempo

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

A continuación se muestra el cumplimiento de los indicadores según el análisis comparativo y en porcentaje el cual se obtiene dividiendo para el número de indicadores por sub-fases:

Fases	Sub Fases	Analysis Services	% de Cumplimiento del Indicador Analysis Services	Weka	% Cumplimiento del Indicador Weka
Comprensión de los Datos	Recolección de Datos iniciales	1	100	1	80
		1		0	
		1		1	
		1		1	
		1		1	
	Descripción de los datos	1	100	1	100
		1		1	
		1		1	
		1		1	
	Exploración de los datos	0	0	1	100
		0		1	
	Verificación de la calidad de los datos	1	100	1	100
1		1			
Fase de preparación de los datos	Selección de datos	1	100	1	100
	Limpieza de datos	0	50	1	100
		1		1	
		1		1	
		0		1	
	Estructuración de los datos	1	100	1	100
		1		1	
		1		1	
	Integración de los datos	1	67	1	67
		0		0	
		1		1	
Formateo de los datos	1	100	1	100	
	1		1		
Fase de modelado		1		1	

Fases	Sub Fases	Analysis Services	% de Cumplimiento del Indicador Analysis Services	Weka	% Cumplimiento del Indicador Weka
Fase de implantación	Selección de la técnica de modelado	1	100	1	100
		1		1	
		1		1	
		1		1	
		1		1	
		1		1	
	Construcción del modelo	1	67	1	100
		1		1	
		0		1	
	Evaluación del modelo	1	100	1	100
		1		1	
		1		1	
		Informe Final	1	100	1

Tabla 29: Cumplimiento de los indicadores según el análisis comparativo y en porcentaje

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

Es así que se construye la siguiente tabla que permitirá aplicar la prueba T-Student

Fases	Sub Fases	% de Cumplimiento del Indicador Analysis Services	Ponderación Empírica de tiempo por Subfases en Minería de datos Analysis Services	Ponderación Empírica de tiempo por fases en Minería de datos Analysis Services	% Cumplimiento del Indicador Weka	Ponderación Empírica de tiempo en Subfases Minería de datos Weka	Ponderación Empírica de tiempo por fases en Minería de datos Weka
Comprensión de los Datos	Recolección de Datos iniciales	100	36,720	4,002	80	29,376	3,202
	Descripción de los datos	100	25,700	2,801	100	25,700	2,801
	Exploración de los datos	0	0,000	0,000	100	22,890	2,495
	Verificación de la calidad de los datos	100	14,090	1,536	100	14,090	1,536
Fase de preparación de los datos	Selección de datos	100	41,880	6,537	100	41,880	6,537
	Limpieza de datos	50	5,130	0,801	100	10,260	1,602
	Estructuración de los datos	100	8,890	1,388	100	8,890	1,388
	Integración de los datos	67	2,907	0,454	67	2,907	0,454
	Formateo de los datos	100	34,620	5,404	100	34,620	5,404
Fase de modelado	Selección de la técnica de modelado	100	15,700	5,402	100	15,700	5,402
	Construcción del modelo	67	41,980	14,445	100	62,970	21,668
	Evaluación del modelo	100	13,340	4,590	100	13,340	4,590
Fase de implantación	Informe Final	100	56,290	6,152	100	56,290	6,152

Tabla 30: Tabla que permitirá aplicar la prueba T-Student

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

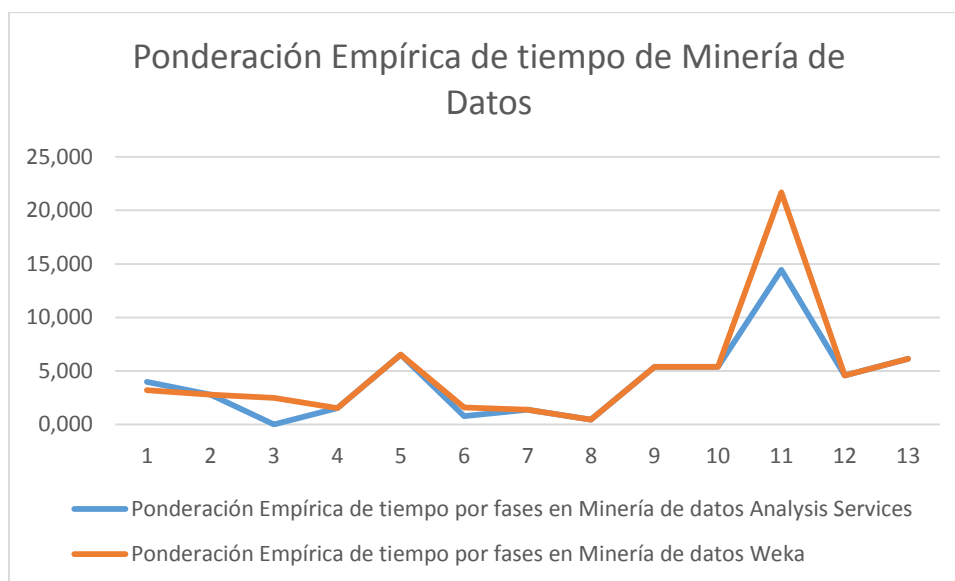


Figura 67: Ponderación Empírica de tiempo de Minería de Datos

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

El gráfico indica la superioridad de Weka en relación a Microsoft Analysis Services en la ponderación empírica del tiempo tomando en cuenta las fases en Minería de Datos.

El resultado de la prueba T-Student es el siguiente:

	<i>Ponderación Empírica de tiempo por fases en Minería de datos Analysis Services</i>	<i>Ponderación Empírica de tiempo por fases en Minería de datos Weka</i>
Media	4,116459359	4,863995205
Varianza	14,7859322	29,44793234
Observaciones	13	13
Varianza agrupada	22,11693227	
Diferencia hipotética de las medias	0	
Grados de libertad	24	
Estadístico t	-0,405253018	
P(T<=t) una cola	0,344440682	
Valor crítico de t (una cola)	1,71088208	
P(T<=t) dos colas	0,688881365	
Valor crítico de t (dos colas)	2,063898562	

Tabla 31: Resultado Prueba T-Student

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

El resultado del P valor es de 0,34 mayor a 0,05; rechazando la hipótesis alternativa y aceptando la hipótesis nula. Esto significa que a pesar de observar una superioridad de la herramienta Weka con respecto a Analysis Services, resultantes del análisis comparativo para minería de datos, estadísticamente esta superioridad no es significativa.

6.3 DISCUSIÓN

La siguiente discusión está basada en los resultados obtenidos del análisis comparativo de las plataformas Weka y Microsoft Analysis Services, que tiene la finalidad de optimizar el desarrollo de minería de datos en la Empresa Prasol “Lácteos Santillán”.

Se ha evaluado las fases para el desarrollo de la minería de datos, en las cuales se puede destacar lo siguiente:

En las fases de Comprensión de los Datos, Preparación de los datos y Modelado Weka supera a Microsoft Analysis Services, a excepción de la fase de Implantación en la que se encuentran con igual puntaje, a continuación, se explica los aspectos relevantes por cada una de las fases mencionadas:

En la fase de Comprensión de los Datos Weka supera a Microsoft Analysis Services, debido a que en la Sub Fase Exploración de los datos permite la creación de tablas de frecuencia y gráficos de distribución estadística, necesarios para una indagación preliminar de los datos, esto equivale a un aporte del 22,89% del tiempo necesario para culminar esta fase.

En la fase de Preparación de los datos Weka supera a Microsoft Analysis Services, debido a que en la Sub Fase de Limpieza de datos cumple dos de cuatro indicadores al permitir la normalización y la reducción del volumen de datos, esto equivale al 5,13% del tiempo necesario para culminar esta fase.

En la fase de Modelado Weka supera a Microsoft Analysis Services, al permitir la creación de un flujo gráfico del modelo de minería de datos afectando a la sub fase de Construcción del modelo, y a su vez en un tiempo considerable equivalente al 20,99% para culminar esta fase.

La valoración final define a Weka como la herramienta que supera a Microsoft Analysis Services con una diferencia del 7%, según los indicadores planteados enfocados a la optimización del proceso de minería de datos con la metodología CRISP-DM.

Relacionado al proceso realizado para la Minería de datos se puede manifestar que el objetivo planteado de analizar la información a través de la aplicación de técnicas de minería de datos enfocadas a descubrir patrones que permitan apoyar a la toma de decisiones fue cumplido, esto permitirá aplicar disposiciones por parte de la gerencia en búsqueda de incrementar la rentabilidad de la empresa, los aspectos relevantes se presentan a continuación:

Debido a que la base de datos proporcionada por la empresa no posee datos demográficos se realizó la segmentación de clientes y la asociación de productos. Se ha utilizado el algoritmo de Simple K means para segmentar a los clientes en 8 grupos vinculándolos con la tipología de los clientes y el algoritmo de Predictive Apriori que proporciona una mayor precisión del modelo al funcionar dos medidas la de soporte y confianza; además un análisis RFM para la segmentación de clientes debido a que es una herramienta poderosa para el marketing estratégico. Sin embargo, es importante tomar en cuenta que para un análisis de minería de datos más completo es necesario que se almacenen los datos demográficos los cuáles permitirán a través de una técnica de clasificación encontrar el perfil idóneo para la búsqueda de potenciales clientes.

Con respecto a la comprobación de la hipótesis, a pesar de la superioridad de la herramienta de minería de datos Weka evidenciado esto a través de los resultados de las Fases y Sub Fases del Análisis comparativo y de la ponderación empírica del tiempo, estadísticamente el resultado del P valor es de 0,34 el cuál es mayor a 0,05; rechazando la hipótesis alternativa y aceptando la hipótesis nula, lo cual representa que

a pesar de la superioridad de Weka con respecto a Microsoft Analysis Services, estadísticamente esta diferencia no es significativa.

CAPÍTULO VII

CONCLUSIONES Y RECOMENDACIONES

7.1 CONCLUSIONES

- Se ha realizado el estudio de las funcionalidades de las herramientas Weka y Microsoft Analysis Services, en donde se ha determinado las principales características, se examina sus principales ventajas y desventajas, obteniendo una visión general de las plataformas que permite enfocar de forma correcta a este trabajo de investigación.
- El análisis comparativo de las herramientas, permitió identificar varios aspectos, tales como: la comparación de las fases, sub fases e indicadores planteados en función de la conceptualización de la metodología CRISP-DM. Como resultado se define a Weka como la mejor herramienta superando con un 7% a Microsoft Analysis Services, y esta se aplica para el desarrollo de la minería de datos. Sin embargo, en el análisis estadístico se aplica la prueba T student, en donde se rechaza la hipótesis alternativa y se acepta la hipótesis nula. A pesar de observar la superioridad de la herramienta Weka con respecto a Microsoft Analysis Services, estadísticamente esta superioridad no es significativa.
- El objetivo planteado de analizar la información a través de la aplicación de técnicas de minería de datos fue cumplido, al no poseer datos demográficos de clientes, fue necesario aplicar un análisis RFM, y posteriormente se realizó las técnicas de segmentación de clientes y la asociación de productos, enfocadas a descubrir patrones que permitan apoyar a la toma de decisiones. Se entrega a la empresa el informe de minería de datos para que sea aplicado por los directivos con la finalidad de incrementar la rentabilidad de la misma.

7.2 RECOMENDACIONES

- Se recomienda, a los futuros investigadores enfocar el estudio de las funcionalidades de las herramientas de forma estándar, de tal forma que se realice una sustentación teórica justa, sobre todo en los ambientes donde se analizan dos herramientas como Weka y Microsoft Analysis Services que son de uso libre y propietaria respectivamente y las cuales presentan múltiples diferencias.
- En este estudio se realiza un análisis comparativo en función de la optimización en el desarrollo de la minería de datos, sin embargo, es necesario tomar en cuenta, además, aspectos como el costo monetario y de implementación de las herramientas, con la finalidad que la plataforma se adapte convenientemente a las necesidades de la organización.
- Para realizar un análisis de minería de datos más completo es imprescindible que se almacenen los datos demográficos en las bases de datos de la Empresa Prasol “Lácteos Santillán”, los cuáles permitirán a través de una técnica de clasificación encontrar el perfil idóneo para la búsqueda de potenciales clientes.

BIBLIOGRAFÍA

- Laudon, J. P. (2004). *Sistemas de información gerencial: administración de la empresa digital*. Pearson Educación.
- Microsoft developer network. (2014). *Microsoft developer network*. Obtenido de <http://msdn.microsoft.com/es-es/library/bb510516.aspx>
- Nay Mojarrango & José Chapalbay, N. M. (2015). Riobamba, Riobamba, Ecuador.
- Programa WEKA, A. C. (Diciembre de 2010). *riunet.upv.es*. Obtenido de [riunet.upv.es: http://hdl.handle.net/10251/10097](http://hdl.handle.net/10251/10097)
- Moine, J. M., Haedo, A., & Gordillo, S. (2011). Estudio comparativo de metodologías para minería de datos. In *XIII Workshop de Investigadores en Ciencias de la Computación*.
- Moine, J. M., Haedo, A., & Gordillo, S. (2011). Estudio comparativo de metodologías para minería de datos. In *XIII Workshop de Investigadores en Ciencias de la Computación*.
- Arancibia, J. A. G. (2010). Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. *Recuperado de http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM,2385037*.
- Birant, D. (2011). *Data Mining Using RFM Analysis*. INTECH Open Access Publisher.
- Rodríguez, D., Pollo Cattaneo, M. F., Britos, P. V., & García Martínez, R. (2010). Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información. In *XVI Congreso Argentino de Ciencias de la Computación*.

ANEXOS

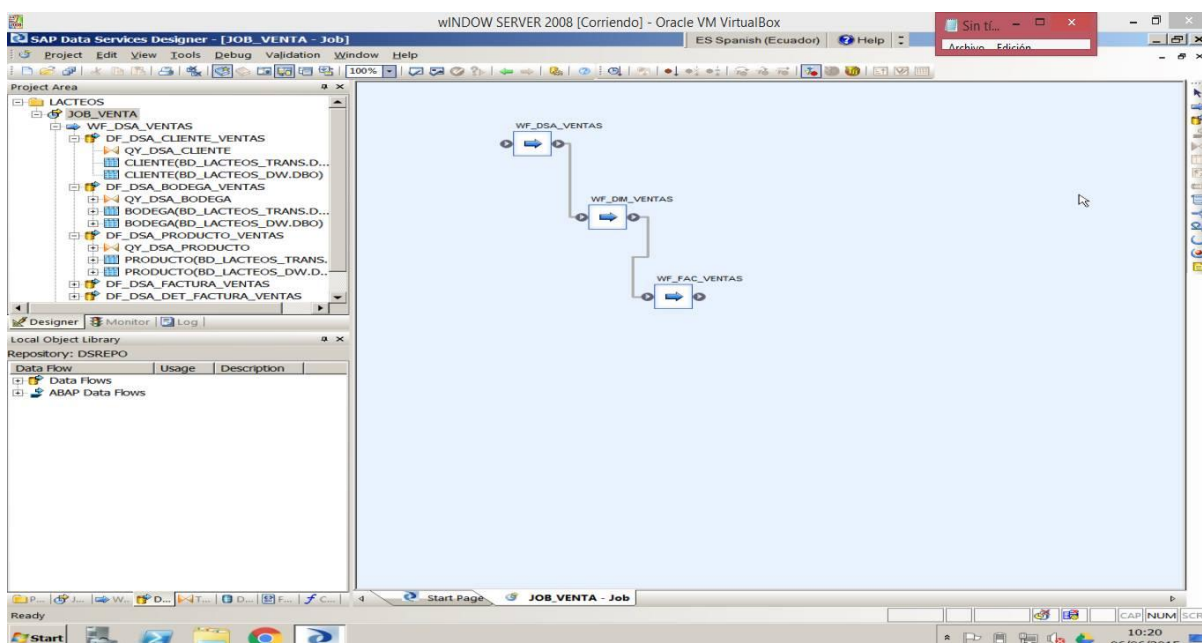


Figura 68: Ambiente de Minería (Job_Ventas)

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

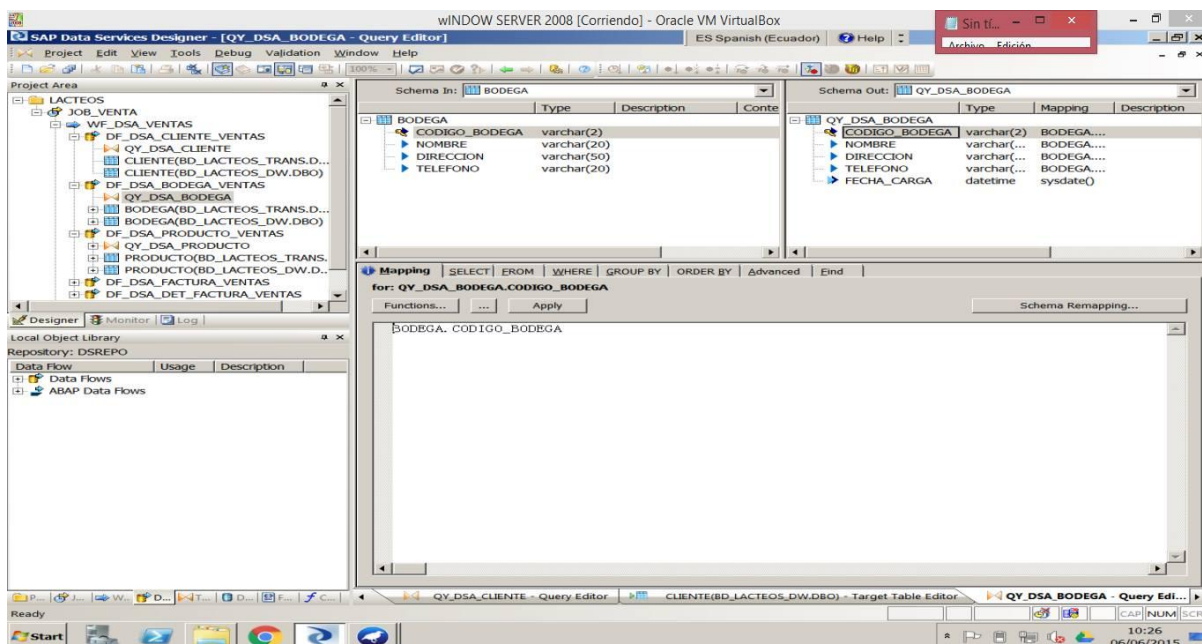


Figura 69: Ambiente de Minería (QY_DSA_BODEGA)

Fuente: (Nay Mojarrango & José Chapalbay, 2015)

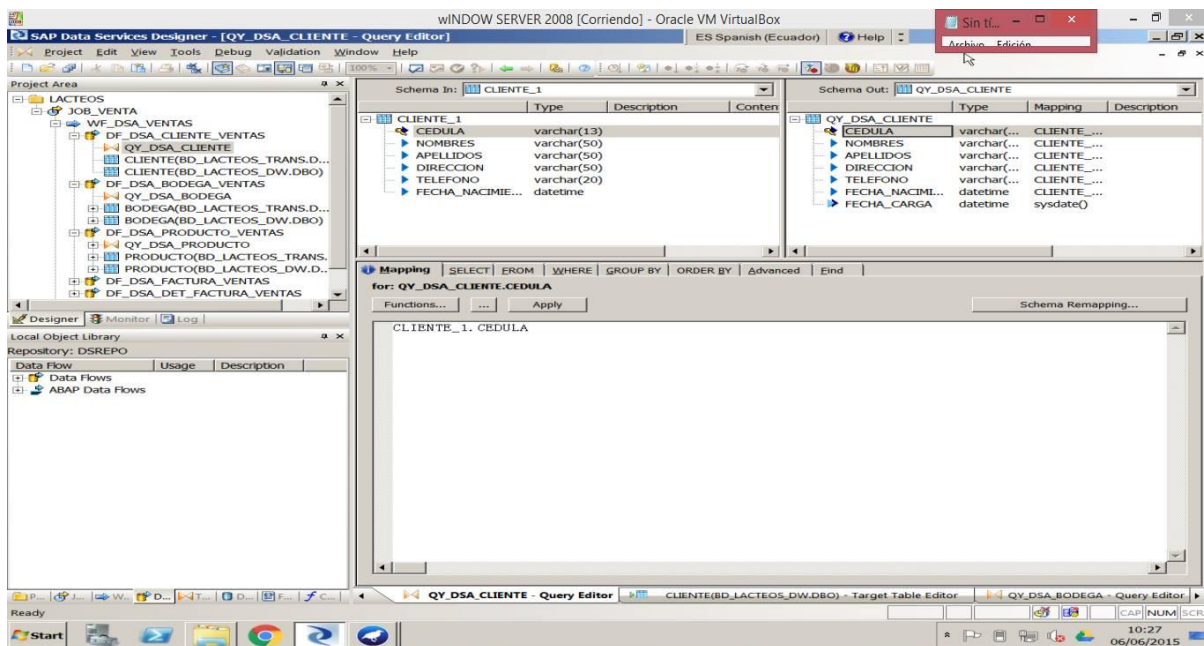


Figura 70: Ambiente de Minería (QY_DSA_CLIENTE)

Fuente: (Nay Mojarrango & José Chapalbay, 2015)