UNIVERSIDAD NACIONAL DE CAJAMARCA

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

“IMPACTO DEL MODELO DE MINERÍA DE DATOS EN EL PRONÓSTICO DE VENTAS DE LA EMPRESA CELLSERVICE EIRL. EN EL PERIODO 2012-2016”

TESIS PARA OBTAR EL TÍTULO DE INGENIERO DE SISTEMAS

AUTOR:

JUAN BAUTISTA QUISPE DURAN

Bachiller

ASESOR:

ING. JAIMEAMADOR MEZA HUAMÁN

Cajamarca – Perú, Abril 2018

AGRADECIMIENTO

A Dios y a todos los maestros que me impartieron sus conocimientos y me orientaron a tener un pensamiento crítico durante mi formación académica. En particular a mi asesor y a los ingenieros de la Escuela Académico Profesional de ingeniería de Sistemas de la UNC.

DEDICATORIA

A mi familia, qué, a pesar de los pasajes de la vida, buenos o malos, creen siempre en mí. En especial a mi madre “Elena” que me apoyo en todo, siempre y a mi hermano menor “Kevin”, que partió de este mundo un poquito antes de terminar el presente trabajo, que seguro me estará dando fuerzas en todos los retos de mi vida.

CONTENIDO

CAPÍTULO I. INTRODUCCIÓN 1

CAPÍTULO II. MARCO TEÓRICO 3

2.1. Antecedentes teóricos de la investigación 3

2.1.1. Antecedentes Internacionales 3

2.1.2. Antecedentes Nacionales 5

2.2. Bases Teóricas 7

2.2.1. Descubrimiento del Conocimiento en Bases de Datos (KDD) 7

2.2.2. Minería de Datos 7

2.2.3. Disciplinas que involucran a la minería de datos 8

2.2.4. Metodologías de Minería de Datos 9

2.2.5. Técnicas de Minería de Datos 20

2.2.6. Estructura de minería de datos 28

2.2.7. Pronóstico de ventas 30

2.2.8. Herramientas para el proceso de minería de datos utilizados 36

2.3. Definición de términos básicos 41

CAPÍTULO III. MATERIALES Y MÉTODOS 43

3.1. Procedimiento 43

3.1.1. Metodología del modelo de Minería de datos 43

Fase I. Comprensión del negocio 43

a. Determinación de los objetivos del negocio 45

b. Evaluación de la situación 45

c. Determinar los objetivos de la minería de datos 47

d. Definición del plan del proyecto 48

Fase II. Entendimiento de los datos 48

e. Recolección de datos iniciales 48

f. Descripción de los datos 50

g. Exploración de los datos 59
h. Verificar la calidad de los datos 64
Fase III. Preparación de los datos 65
i. Seleccionar los datos 65
j. Limpieza de datos 66
k. Construcción de los datos 67
l. Integración de los datos 68
m. Formateo de datos 89
n. Selección de la Técnica de modelado 90
o. Diseño de las pruebas del modelo 101
p. Construcción del modelo 103
q. Evaluación del modelo 105
r. Evaluación de los resultados 111
s. Revisión del proceso 116
t. Determinación de las próximas etapas 116
u. Planificación de la implementación 116
v. Planificar el monitoreo y el mantenimiento 116
w. Creación de un reporte final 116
x. Revisión del proyecto 119
3.2. Tratamiento y análisis de datos y presentación de resultados 119
CAPÍTULO IV. ANÁLISIS Y DISCUSIÓN DE RESULTADOS 130
4.1. Análisis de resultados 130
4.1.1 Comprobación de la hipótesis 130
a. Formulación de la Hipótesis 131
b. Ubicación de la región critica 131
c. Determinación y ubicación de valor esperado en la región crítica 131
d. Aceptación o rechazo de la hipótesis 133

4.2. Discusión de resultados 133

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES 136

REFERENCIAS BIBLIOGRÁFICAS 139

Índice de tablas

Tabla 1: Fase 1 (Entendimiento del Negocio) - Metodología CRISP-DM 14
Tabla 2: Fase 2 (Entendimiento de los datos) - Metodología CRISP-DM 15
Tabla 3: Fase 3 (Preparación de los datos) - Metodología CRISP-DM 16
Tabla 4: Fase 4 (Modelado) - Metodología CRISP-DM 17
Tabla 5: Fase 5 (Modelado) - Metodología CRISP-DM..... 19
Tabla 6: Fase 6 (Implementación) - Metodología CRISP-DM 19
Tabla 7: Ejemplo de problema de clasificación 23
Tabla 8: Comparación de otras herramientas de minería de datos con Weka..... 38
Tabla 9: Descripción de algoritmos de clasificación usados para generar el modelo 40
Tabla 10: Recursos requeridos para el análisis de ventas y pronósticos..... 44
Tabla 11: Matriz de procesos y dimensiones para diseño del Data Mart 46
Tabla 12: Tabla de hechos del datamart..... 55
Tabla 13: Tabla dimensión cliente del Data Mart 56
Tabla 14: Tabla dimensión producto del dataMart 57
Tabla 15: Tabla dimensión sucursal del datamart..... 57
Tabla 16: Tabla dimensión periodo del datamart 58
Tabla 17: Tabla dimensión plan del data mart..... 58
Tabla 18: Datos seleccionados para la estructura del modelo de minería 65
Tabla 19: Comparación de algoritmos de clasificación 102
Tabla 20: Optimización del modelo con validación cruzada.....104
Tabla 21 : Optimización del modelo con porcentaje de división.....104
Tabla 22: Representación del modelo optimizado generado con WEKA 107
Tabla 23: Comparación del pronóstico del modelo y las ventas futuras.....120
Tabla 24: Comparación de recursos usados para el pronóstico, antes y después de aplicar el modelo 124
Tabla 25: Resultado de conformidad con el modelo de minería.....125
Tabla 26: Criterios de evaluación de la encuesta del Pre y Post -Test 126
Tabla 27: Resultado del Pre - Test.....127
Tabla 28: Resultado del Post - Test 128
Tabla 29: Resultado de diferencia de medias del Pre y Post - Test.....129

Índice de figuras

Fig. 1: Pasos que componen el proceso KDD 7
Fig. 2: Esquema resumido del proceso KDD 10
Fig. 3: Metodología SEMMA 10
Fig. 4: Cadena de valor empresarial 11
Fig. 5: Metodología CRISP-DM 12
Fig. 6: Técnicas de la Minería de Datos 21
Fig. 7: Ejemplo de clasificación con ID3 26
Fig. 8: Estructura de la red de neuronas 27
Fig. 9: Ejemplo de estructura de Minería de Datos de Microsoft 28
Fig. 10: Vista de la base datos del sistema transaccional de ventas 49
Fig. 11: Otros datos de clientes en formato de MS Excel 49
Fig. 12: tabla “activacion” de la base de datos relacional 50
Fig. 13: Tabla “cliente” de la base de datos relacional 51
Fig. 14: Tabla “detalleventa” de la base de datos relacional 51
Fig. 15: Tabla “distrito” de la base de datos relacional 52
Fig. 16: Tabla tienda de la base de datos relacional 52
Fig. 17: Tabla “mdproducto” de la base de datos relacional 52
Fig. 18: Tabla “venta” de la base de datos relacional 53
Fig. 19: Tabla “mdcliente” de la base de datos relacional 53
Fig. 20: Tabla “ubigeo” de la base de datos relacional 54
Fig. 21: Esquema del modelo dimensional del Data Mart 55
Fig. 22: Diseño dimensional del Data Mart de ventas 59
Fig. 23: cantidad de productos vendidos según la edad de los clientes 60
Fig. 24: cantidad de productos vendidos según el género de los clientes 60
Fig. 25: cantidad de productos vendidos según distrito del cliente 61
Fig. 26: Porcentaje de ventas por tipo de cliente 61
Fig. 27: cantidad de productos vendidos según la marca 62
Fig. 28: cantidad de equipos celulares vendidos por modelo y marca 63
Fig. 29: cantidad de equipos celulares vendidos por color 63

Fig. 30: porcentaje de equipos celulares vendidos por tipo de plan 64
Fig. 31: Diagrama de base de datos del data mart de ventas 69
Fig. 32: Interface de desarrollo de la herramienta “Pentaho Data Integraction” . 70
Fig. 33: Herramientas de pentaho usadas para las transformaciones 71
Fig. 34: Herramientas de Pentaho usadas para los trabajos o Jobs 71
Fig. 35: Configuración- extracción de datos para poblar la dimensión de los clientes 72
Fig. 36: Configuración - Insertar datos en la dimensión clientes 73
Fig. 37: Diagrama- ETL para la dimensión de los clientes 73
Fig. 38: Configuración - extracción de datos para poblar la dimensión de los clientes. 74
Fig. 39: Configuración - Insertar datos en la dimensión de los clientes. 75
Fig. 40: Diagrama- ETL para la dimensión de los productos. 75
Fig. 41: Configuración - extracción de datos para poblar la dimensión de las sucursales. 76
Fig. 42: Configuración - Insertar datos en la dimensión de los clientes. 77
Fig. 43: Diagrama- ETL para la dimensión de las sucursales. 77
Fig. 44: Configuración - extracción de datos para poblar la dimensión periodo. 78
Fig. 45: Configuración - Insertar datos en la dimensión de los clientes. 79
Fig. 46: Diagrama- ETL para la dimensión periodo 79
Fig. 47: Configuración - extracción de datos para poblar los hechos. 80
Fig. 48: Configuración - Insertar datos en la tabla de hechos 81
Fig. 49: Diagrama- ETL para la dimensión periodo 81
Fig. 50: Configuración - Limpiar dimensiones y hechos. 82
Fig. 51: Herramienta Script SQL para limpiar dimensiones y hechos. 82
Fig. 52: Representación de un paquete de transformaciones en pentaho. 82
Fig. 53: Configuración de los paquetes “job” de pentaho 83
Fig. 54: Diagrama ETL general para poblar el data mart. 84
Fig. 55: Espacio de trabajo de SchemaWorkbench. 85
Fig. 56: Diseño de las dimensiones del cubo en Schema Workbench 86
Fig. 57: Diseño del cubo en SchemaWorkbench. 87
Fig. 58: Publicación el archivo XML del cubo. 87
Fig. 59: Espacio trabajo de la herramienta OLAP JRubik. 88
Fig. 60: Ejemplo de la exploración del cubo OLAp con JRubik. 89

Fig. 61: Extracción de datos del dataMart para minería (usando consulta SQL) 90
Fig. 62: Herramienta de minería de datos WEKA..... 91
Fig. 63: Exploración de datos con WEKA..... 92
Fig. 64: Descripción del atributo “cliente” con WEKA..... 93
Fig. 65: Descripción del atributo “genero” con WEKA 94
Fig. 66: Descripción del atributo “proximo” con WEKA..... 95
Fig. 67: Descripción del atributo “quincena” con WEKA..... 96
Fig. 68: Descripción del atributo “dia” con WEKA..... 97
Fig. 69: Descripción del atributo “marca” con WEKA..... 98
Fig. 70: Descripción del atributo “Tecnología” con WEKA..... 99
Fig. 71: Características del atributo “color” con WEKA.....100
Fig. 72: Características del atributo clase “modalidad” con WEKA 101
Fig. 73: diseño de pruebas del modelo de minería.....102
Fig. 74: Ejemplo de lectura del árbol de decisión del algoritmo J48.....106
Fig. 75: Gráfica de nivel de conformidad con el modelo de minería.....126
Fig. 76: Ubicación del valor esperado en la región crítica1 132
Fig. 77: Ubicación del valor esperado en la región crítica 2 132

RESUMEN

EL presente trabajo se desarrolló en la empresa “Cell Service” de la ciudad de Cajamarca, dedicada a la venta al por mayor y menor de equipos celulares y chips. Esta empresa desde el año 2012 ha almacenado su información en una base de datos transaccional de su sistema de ventas y hojas de cálculo de MS Excel. La empresa ha decidido mejorar el análisis de sus ventas aplicando pronósticos, que le puedan ayudar a mejorar sus estrategias de ventas, pues los pronósticos que realizan actualmente son basados en reportes de hojas de cálculo y en la opinión del personal que toma decisiones en el área de ventas; pero se ha presentado inconvenientes al momento de integrar toda la información almacenada para hacer cálculos y reportes masivos que le permitan realizar pronósticos de sus ventas. Debido a esto se ha planteado elaborar un modelo de minería de datos y cuyo objetivo es evaluar el impacto que tiene en el análisis de ventas aplicando los pronósticos.

Para el proceso de minería se utilizaron: La metodología CRISP-DM; la técnica del árbol de decisión, por ser la de mejor desempeño; Schema Workbench para la construcción del data Mart de ventas; Pentaho Data Integration para poblar el data Mart y la herramienta WEKA para la construcción del modelo de minería. El modelo generado fue evaluado con datos de ventas de equipos celulares, comprendidos en el período del año 2012 y 2016, en total de 11610 registros. El impacto del modelo fue demostrado aplicando un test y post test a las 4 personas encargadas de tomar decisiones en el área de ventas, en la que se pudo verificar que, efectivamente, el modelo tiene un impacto de mejora, en el análisis de ventas aplicando pronósticos, en base a los resultados generados por el modelo.

Palabras Claves: Modelo de Minería de Datos, técnica de minería de datos, Arboles de decisión, Análisis de ventas, pronóstico de ventas, software libre.

ABSTRACT

In the present work was developed in the company "Cell Service" of the city of Cajamarca, dedicated to wholesale and retail of cell phones and chips. This company since 2012 has stored its information in a transactional database of its sales system and MS Excel spreadsheets. The company has decided to improve the analysis of its sales by applying forecasts, which can help it improve its sales strategies, since the forecasts that are currently made are based on reports of spreadsheets and the opinion of the personnel who make decisions in the area of sales; But there have been problems when integrating all the information stored to make calculations and massive reports that allow you to make forecasts of your sales. Due to this, it has been proposed to elaborate a data mining model and whose objective is to evaluate the impact that it has in the analysis of sales applying the forecasts.

The following were used for the mining process: The CRISP-DM methodology, the decision tree technique, because it is the best performing; Schema Workbench for the construction of sales data mart; Pentaho Data Integration to populate the Data Mart; the WEKA tool for the construction of the mining model. The generated model was evaluated with sales data of cellular equipment, included in the period of the year 2012 and 2016, in total of 11610 records. The impact of the model was demonstrated by applying a test and post test to the 4 people in charge of making decisions in the sales area, in which it was possible to verify that, effectively, the model has an improvement impact, in the analysis of sales applying forecasts, based on the results generated by the model.

Key Words

Data Mining Model, data mining technique, Decision trees, Sales analysis, sales forecast, free software.

CAPÍTULO I. INTRODUCCIÓN

Los avances tecnológicos en la informatización de la información han dado a las empresas la facilidad de almacenar grandes cantidades de datos, de sus diferentes áreas y de diferentes formas, desde archivos planos, base de datos transaccionales, data mart, data warehouse, etc. A todo esto, llega el momento en que surge la necesidad de explorarlos, con el fin de encontrar información valiosa y novedosa, que les ayude para tomar decisiones en base a los datos almacenados. Una de las formas más comunes en que las empresas recurren para analizar sus datos, es, entre otras, por ejemplo, las reuniones mensuales, bimestrales, trimestrales, etc., donde se discuten informes, reportes, tendencias y finalmente tomar decisiones a futuro; sin embargo se basa en un análisis e interpretación manual, por lo que esta forma de análisis de conjunto de datos es lento costoso y sumamente subjetivo [*1*]. Justamente, la Minería de Datos, que consiste en extraer conocimiento a partir de datos almacenados, apoyándose de las ciencias de la computación y de la estadística, se ha convertido en indispensable al momento de analizar comportamientos de los datos de ventas, clientes, etc., el cual genera un modelo previamente entrenado y validado con datos históricos que sirve pronosticar datos futuros y de esa manera ayudar en los planes de marketing de las empresas.

El presente trabajo, se desarrolló en la empresa “CELLSERVICE E.I.R.L”, del departamento de Cajamarca - Perú, que tiene también sucursales en otras provincias del departamento. La empresa es actualmente una agencia de la empresa de telefonía “CLARO” y se dedica a la venta de todo tipo y marca de equipos celulares, chips y también accesorios para celulares. El problema de la empresa de cara a la alta competencia en el mercado, consiste en que con los reportes que se obtiene del sistema transaccional no es suficiente y que además la forma de organización de los datos no permite analizar la información en su totalidad, ya que se desea incluir más criterios de análisis y, además, que estos procedimientos de análisis sean más rápidos, debido que la empresa desea mejorar su planes de marketing, mediante el análisis de sus ventas que almacenada en su base datos, y que le permita anticiparse a las formas de venta de sus productos y necesidades de sus clientes. Por lo que el presente trabajo ha

consistido en elaborar y proponer un modelo de minería de datos, basado pronóstico de comportamiento de ventas, y mejorar los procesos de análisis ventas que ayuden en los procesos de marketing. Con el modelo, se pretende contestar una interrogante como: "¿Cuál es el impacto del modelo de Minería de Datos, en el pronóstico de ventas de la empresa CELL SERVICE E.I.R.L, en el periodo 2012-2016?", lo que sirvió para demostrar la hipótesis formulada: "El Modelo de Minería de Datos tiene un impacto en el Pronóstico de Ventas de la Empresa CELL SERVICE E.I.R.L, en el periodo 2012 – 2016". Con un modelo de minería de datos, los que toman decisiones en el área de ventas, obtienen patrones de comportamiento de las ventas, que les ayudara al momento de tomar decisiones estratégicas para promover las ventas, que les ayude reducir también la inversión en campañas de marketing y, además, de horas de trabajo analizando manualmente los datos. El modelo propuesto, ha tenido como fuente de datos un data Mart de las ventas previamente elaborado, se usó registros de ventas desde el año 2012 hasta el 2016 que fueron en total de 11610 registros, la técnica de minería usada por el modelo fue los árboles de decisión y se ha usado la metodología de minería de datos CRISP –DM por ser la más usada actualmente. Para evaluar el impacto del modelo, se aplicó una encuesta, antes y después de aplicar el modelo, a las personas encargadas de la toma de decisiones del área de ventas, conformada por 4 personas, así mismo se realizó una comparativa de los datos de ventas de los meses posteriores al periodo de análisis. Se cumplió el objetivo principal que fue evaluar el impacto del modelo, corroborado con resultados de las encuestas aplicadas y comparando los datos de ventas del pronóstico del modelo con las ventas futuras, lográndose ver un impacto positivo del modelo al momento de realizar pronósticos, y, además, se cumplieron con los objetivos específicos: analizar la problemática de la empresa, comprender la base datos transaccionales, preparar los datos para la minería, construir el data Mart, implementar y evaluar el modelo. El contenido del presente documento comprende: Las bases teóricas, mencionados en el Capítulo II. Marco Teórico, el desarrollo del proceso de la metodología CRISP-DM, que se describen en el capítulo III. Materiales y Métodos. La presentación y análisis de resultados encontrados, que se describe el capítulo IV. Análisis y discusión de resultados. Finalmente, se presentan conclusiones y recomendaciones del trabajo realizado en el capítulo V. Conclusiones y Recomendaciones.

CAPÍTULO II. MARCO TEÓRICO

2.1. Antecedentes teóricos de la investigación

2.1.1. Antecedentes Internacionales

En la Tesis: “ANÁLISIS PARA LA PREDICCIÓN DE VENTAS UTILIZANDO MINERIA DE DATOS EN ALMACENES DE VENTAS DE GRANDES SUPERFICIES” [*2*], tiene como objetivo principal, utilizar la Minería de Datos haciendo uso de plataformas de cómo RapidMiner, para aplicar un modelo de predicción de ventas sobre un conjunto de datos seleccionados sobre una gran superficie, con el fin de encontrar relaciones entre dos o más productos. En este trabajo, la investigación parte de la necesidad de encontrar comportamientos en los inventarios, los cuales a simple vista serian tediosos, por no decir imposibles de visualizar, tales como tendencias de compras, gustos o hábitos de las mismas, con el fin de establecer políticas para la fidelidad de los mismos y el crecimiento de las ventas. El objetivo del análisis fue encontrar patrones de comportamiento en las compras de los clientes, y poder descubrir al mismo tiempo que productos son adquiridos con frecuencia interesantes para llegar a contribuir con el incremento de las ventas. También, se concluye que proceso de Minería de Datos, es una tarea que no es fácil de realizar, ya que se necesita hacer una gran cantidad de pruebas para llegar a obtener algún resultado, quizás no el mejor o esperado, pero uno que, si pueda dar cierta claridad, en los movimientos que se pueden dar en grandes superficies de ventas, como en este caso.

En el Seminario “ANÁLISIS DE PATRONES DE COMPRA DE TIENDAS RETAIL UTILIZANDO BUSINESS INTELLIGENCE” [*3*],concluye que, el análisis de los distintos modelos de trabajo de Data Mining permite ordenar las distintas posibilidades de análisis a disposición, por otro lado, la puesta en práctica del modelo de reglas de asociación en una base de datos real y desconocida permite a los investigadores encontrar información relevante respecto al negocio, además, con el uso del software de Data Mining “Rapid Miner”, permitió a los investigadores acercarse a los modelos de investigación de manera sencilla. Finalmente se indica que, en la investigación se tomaron medidas bajo un

contexto de marketing (análisis de consumidor), pero, el procedimiento puede ser similar y puede entregar gamas de respuestas para diferentes disciplinas como finanzas, recursos humanos y en realidad para la organización en general.

En la Tesis: “ALGORITMOS DE MINERÍA DE DATOS EN LA RECOLECCIÓN DE INTELIGENCIA” [*4*], el objetivo general de la investigación, es seleccionar algoritmos predictivos y de agrupación que ayuden a la toma de decisiones en la seguridad pública. Y Se concluye que, resulta valioso a los analistas delictivos aplicar los 5 algoritmos de minería de datos presentados en esta tesis. En el caso de los algoritmos predictivos, los arboles de decisión son importantes, dado que permiten visualizar los atributos que incluyen en el comportamiento delictivo de manera rápida, apoyándoles en la generación de estrategias para reducir los delitos en la comunidad. Además de predecir en que regiones se comenten más delitos. En el desarrollo de la tesis, por cuestiones de la confidencialidad de los datos y el acceso a dicha información falto integrar más bases de datos que aportaran y nutrieran el análisis delictivo, tales como: el clima, la geografía, eventos asociados, estaciones del año, hora, día de la semana, tipo de víctimas, etcétera. Por lo tanto, se concluye que mientras mayor sea la información disponible, mayores son las opciones que tiene un analista delictivo para llevar a cabo un análisis más detallado, ya sea por tipo de delito, zona, víctima o motivo. Este hecho no demerita lo investigado en la tesis, los algoritmos presentados y el análisis realizado mantienen su validez. Por último, se concluye cuán importante fue utilizar metodología CRISP-DM, que permitió mantener un orden al momento de llevar a cabo el análisis, dando pauta para análisis futuros.

En el trabajo de fin de Titulación: "APLICACIÓN DE TECNICAS DE MINERÍA DE DATOS PARA PREDECIR LA DESERCIÓN DE LOS ESTUDIANTES DE PRIMER CICLO DE LA MODALIDAD ABIERTA Y A DISTANCIA DE LA UTPL" [*5*], se ha obtenido un modelo de minería de datos aplicando la metodología CRISM- DM, que con la ayuda del análisis de la información, que los diferentes estudiantes proporcionan a la base de datos del sistema académico (Syllabus) y el entorno virtual de aprendizaje (Eva) de la Universidad, se ha podido obtener patrones de comportamiento, para con ello conocer cuáles son las posibles causas por las que un alumno que cursa las asignaturas del primer ciclo de la

Modalidad Abierta y a Distancia de la Universidad Técnica Particular de Loja, decide abandonar sus estudios universitarios. El presente modelo permitirá a la institución educativa obtener beneficios económicos, ya podrá determinar las estrategias necesarias para que un estudiante deserte de la carrera.

2.1.2. Antecedentes Nacionales

En la tesis: "PROPUESTA DE MODELO DE DETECCIÓN DE FRAUDES DE ENERGÍA ELÉCTRICA EN CLIENTES RESIDENCIALES DE LIMA METROPOLITANA APLICANDO MINERÍA DE DATOS" [*6*],se propone un modelo para predecir potenciales situaciones de fraudes de energía eléctrica en clientes residenciales basado en aprender el comportamiento de clientes que anteriormente hurtaron para ello se aplica el proceso Minería de Datos basado para analizar, extraer y almacenar información de la base de datos, las cual contiene el historia del consumos de energía. El modelo se propone para apoyar a las empresas de distribución eléctricas en especial a los técnicos eléctricos a examinar y verificar, acertadamente, de manera rápida y oportuna los resultados obtenidos y contribuya, de esta forma, en la toma de decisiones. Para la creación del modelo se utilizaron las redes neuronales por ser la de mejor desempeño en la detección. El modelo creado fue evaluado con datos de una empresa distribuidora de electricidad para el período 2009 y 2010. Las herramientas utilizadas para la creación del modelo fueron el Sql Server Management Studio (DatabaseEngine para la base de datos y creación de procedimientos, Análisis Services para la creación de Estructuras y modelos) y como herramienta interactiva y de fácil entendimiento para el usuario el Complemento de Minería de Datos para Microsoft Excel. Además, se concluye que la preparación de los datos para crear los modelos y obtener los patrones del comportamiento tomó el 60% de esfuerzo de todo el proyecto y se tuvo que regresar a una etapa anterior para mejorar las pruebas. Finalmente, entre sus recomendaciones se menciona que, si se desea desarrollar un nuevo modelo minería de datos, es recomendable apoyarse en expertos del negocio para determinar el conjunto de entrenamiento y prueba para un mejor resultado.

En la tesis: "IMPLANTACIÓN DE UN SISTEMA DE VENTAS QUE EMPLEA UNA HERRAMIENTA DE DATA MINING" [*7*],tiene como objetivo exponer el flujo de procesos o serie de pasos que se realiza en un proceso de implantación de un ERP y en un proceso algorítmico de Data Mining; se realiza porque la empresa a la que se aplica ambos conjuntos de procesos necesita ordenar su información en el área de ventas y obtener información que beneficie a la empresa respecto a cómo se comportan sus clientes cuando compran en un periodo de tiempo. Para que el objetivo final del proyecto se cumpla, se usaron herramientas de software, herramientas de planificación y de organización. Esta investigación concluye en que el producto final permite que la empresa beneficiada, pueda analizar y comprender porque sus clientes se comportan en sus ventas de formas distintas. El algoritmo usado para el proceso algorítmico de Data Mining es uno de los más robustos; sin embargo, se pudo haber usado otros y obteniendo diferentes resultados; es decir, cada algoritmo es usado en un escenario distinto, y la forma para escoger el adecuado es, en muchas ocasiones, la experiencia de la persona encargada del modelado.

En el Trabajo de investigación “DESARROLLO DE UN DATAMART PARA MEJORAR LA TOMA DE DECISIONES EN EL ÁREA DE VENTAS DE LA CORPORACIÓN FURUKAWA” [*8*], el objetivo de la investigaciones establecer de qué manera el desarrollo de un DataMart influye en la toma de decisiones en el área de ventas de la Corporación Furukawa. Para ello de plantean Identificar los requerimientos de análisis de información para las áreas de Ventas, elaborar un modelo de base de datos multidimensional que permita el análisis y explotación de la información identificada, construir el dataMart para mostrar la información que se necesita para poder tomar decisiones estratégicas en el área de ventas.
Las necesidades de información del Área de Ventas de la Corporación Furukawa fueron identificadas satisfactoriamente. Esto contribuyó a identificar requerimientos claros y precisos que fueron utilizados para la construcción del modelo multidimensional. El DataMart cubrió las necesidades de los usuarios estratégicos logrando así que la gerencia de ventas tenga ahora una herramienta con el cual hacer su análisis de ventas. Se concluyó que las necesidades de información del Área de Ventas de la Corporación Furukawa fueron identificadas satisfactoriamente.

2.2. Bases Teóricas

2.2.1. Descubrimiento del Conocimiento en Bases de Datos (KDD)

Es un proceso de descubrimiento del conocimiento de la base de datos. Desarrolla técnicas y métodos para dar sentido a los datos. Este proceso se centra en la aplicación de métodos específicos de minería de datos para el descubrimiento y extracción de patrones [*1*].Este proceso es muy útil a la hora de buscar información que permita tomar decisiones, en base a la información que se tiene almacenada. Debido a que esta información es demasiado grande, pues las empresas han tenido la necesidad de hacer uso de tecnologías para el tratamiento de grandes volúmenes de datos, como pueden ser aplicaciones estadísticas, sistema de gestión de bases de datos relacionales (RDBMS), entre otros. En los procesos de negocios, las áreas principales de aplicación del proceso KDD son: Marketing, Finanzas, detección de fraudes, telecomunicaciones, internet. En la siguiente figura Fig.1 se puede observar el proceso los pasos que componen el KDD.

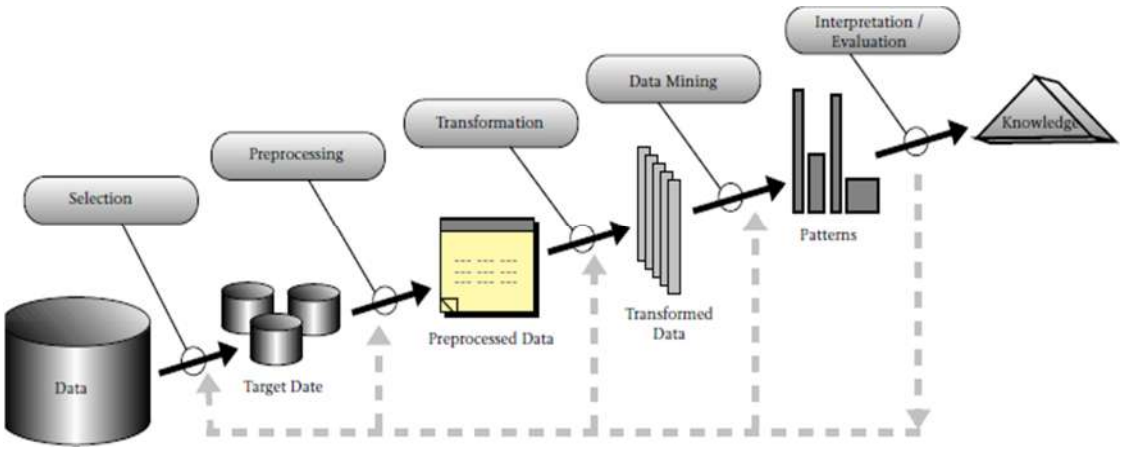


Fig. 1: Pasos que componen el proceso KDD [*1*]

2.2.2. Minería de Datos

La minería de datos o conocida en el idioma inglés “Data Mining” consiste en “la extracción de conocimiento desde una gran cantidad de datos a través del uso de tecnología (algoritmos computacionales)”. Debido a que no existe la cantidad suficiente de personas disponibles para analizar dicha cantidad de información, es

que se recurre a la tecnología a modo de lograr automatizar la extracción de conocimiento en los datos [2]. Es un proceso que reúne un conjunto de Herramientas de diversas Ciencias (Estadística, Informática, Matemáticas, Ingeniería, entre otras) que permite extraer conocimiento oculto o información no trivial en grandes volúmenes de datos, con la finalidad de dar soluciones a problemas específicos a empresas determinadas [4].

2.2.3. Disciplinas que involucran a la minería de datos

A continuación, se mencionan algunas disciplinas relacionadas al concepto de Minería de Datos. Según [*1*], la "Minería de Datos" y el descubrimiento del conocimiento en las base de datos están relacionados con las máquinas de aprendizaje, estadísticas, bases de datos y extracción del conocimiento.

- **Extracción de información:** Consiste en el primer paso, ya que Data mining (minería de datos) trabaja sobre una base de datos la cual se debe constituir a través de la extracción de datos.

- **Estadística:** Se define Data Mining como "Estadísticas en gran escala, velocidad y simplicidad", es por esto que la estadística usada en Data Mining difiere de la estadística tradicional, esencialmente en que la primera se basa en el uso de la inferencia en donde se usan pequeñas muestras para determinar el comportamiento de una variable, en este caso se usa una gran cantidad de datos para extraer conclusiones, por lo que no se usa el concepto de inferencia estadística. De hecho, el gran problema del Data Mining consiste en el "overfiting", que es el momento donde un modelo se acerca tanto a la realidad, que inclusive modela las características aleatorias de la variable en cuestión.

- **Sistemas de Bases de Datos:** La manera de estructurar los datos es relevante al momento de trabajarlos, por eso es relevante el uso y manejo de esta disciplina tanto en el procesamiento y especialmente en el pre-procesamiento de los datos, los datos deben ser claros, ordenados y acordes a los objetivos de investigación propuestos en el negocio. Se

considera que para el uso de Data Mining en negocios, la disciplina más importante de las mencionadas es la búsqueda de información.

2.2.4. Metodologías de Minería de Datos

Se puede decir que el término metodología se define como el grupo de procedimientos, empleados para el logro de un objetivo. Las metodologías permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial. Ayudan a las organizaciones a entender el proceso de descubrimiento de conocimiento y proveen una guía para la planificación y ejecución de los proyectos [*9*]. A continuación, describimos las metodologías más usadas en los procesos de minería de datos.

- **Knowdlege Discovery in Databases (KDD)**

  El Descubrimiento de Conocimiento en Bases de Datos (KDD KnowdlegeDiscovery in Databases) [*10*], constituye el primer modelo que define el descubrimiento de conocimiento en bases de datos como un "proceso", compuesto por distintas etapas y fases que van desde la preparación de los datos hasta la interpretación y difusión de los resultados. En el año 1996, Fayyad define a KDD como el "proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia entendibles en los datos". El término proceso se refiere a la secuencia iterativa de etapas o fases que lo componen. Los patrones deberían ser válidos para nuevos datos, novedosos en el sentido que deberían aportar nuevo conocimiento al dominio de aplicación y potencialmente útiles para el usuario final o tomador de decisiones.

  KDD es un proceso iterativo e interactivo. Iterativo ya que la salida de alguna de las fases puede retroceder a pasos anteriores y porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. Es interactivo porque el usuario, o más generalmente un experto en el dominio del problema, debe ayudar a la preparación de los datos y validación del conocimiento extraído. El modelo de proceso KDD se

resume en las siguientes cinco fases: Selección de los datos, Pre-procesamiento de los datos, Transformación de los datos y reducción de la dimensionalidad, Minería de datos, Interpretación y evaluación del nuevo conocimiento en el dominio de aplicación. A continuación: Fig.2, se presenta un esquema resumido del proceso KDD.

Fig. 2: Esquema resumido del proceso KDD [*10*]

- **SEMMA**

Es una metodología, creada por SAS Institute [*10*], fue propuesta especialmente para trabajar con el software SAS Enterprise Miner. Si bien en la comunidad científica se conoce a SEMMA como una metodología, en el sitio de la empresa SAS se aclara que éste no es el objetivo de la misma, sino más bien la propuesta de una organización lógica de las tareas más importantes del proceso de minería de datos. SEMMA establece un conjunto de cinco fases para llevar a cabo el proceso de minería: Sample (Muestreo), Explore (Exploración), Modify(Modificación), Model (Modelado) y Assess (Evaluación). Está especialmente enfocada al desarrollo del modelo de minería, y quedan fuera de su alcance otros aspectos del proyecto como el conocimiento del problema en estudio ola planificación de la implementación SAS Enterprise Miner. En la figura Fig.3 se puede observar el proceso de la metodología SEMMA.

Fig. 3: Metodología SEMMA [*10*]

- **Catalyst**

En el año 2003, Dorian Pyle propone en su libro “Business modelling and data mining” [*10*] una metodología para el proceso de extracción de conocimiento en bases de datos llamada “Catalyst”. Pyle recomienda que el proceso de minería de datos siempre debiera colaborar con una situación organizacional, como un problema u oportunidad. Recomienda no trabajar directamente con los datos, sino establecer de antemano la problemática que se aborda, el personal involucrado y las expectativas y necesidades de los usuarios. Para proyectos donde el problema u oportunidad de negocio no está definido, se recomienda comenzar analizando las relaciones P3TQ – Product(Producto), Place (Lugar), Price (Precio), Time (Tiempo) y Quantity(Cantidad) - que existen en la cadena de valor organizacional. Las relaciones P3TQ se refieren a tener el producto correcto, en el lugar adecuado, en el momento adecuado, en la cantidad correcta y con el precio correcto. La cadena de valor empresarial, es un modelo teórico popularizado por Michael Porter, que define las actividades de la empresa que van añadiendo valor al producto a medida que éste pasa por cada una de ellas.

Fig. 4: Cadena de valor empresarial [*10*]

- **CRISP-DM**

CRISP-DM, un acrónimo de Cross Industry Standard Process para Data Mining, es un modelo de proceso de minería de datos que incluye enfoques de uso común que utilizan las organizaciones de análisis de datos para abordar problemas empresariales relacionados con la minería de datos [*11*]. Fue presentada en el año 1999 por las empresas SPSS, Daimer Chrysler y NCR. Es una metodología abierta, no está ligada a ningún producto comercial, y fue construida en base a la experiencia de sus creadores, es decir desde un enfoque práctico [*10*]. El modelo de referencia presenta un resumen de las fases y tareas a llevarse cabo en cada una (junto con sus salidas). Es decir, describe "que" debería hacerse en un proyecto de minería de datos. La guía de usuario proporciona sugerencias para la ejecución de cada tarea del modelo de referencia [*10*]. En la siguiente figura Fig.5 se muestra el esquema de la metodología CRISP-DM.

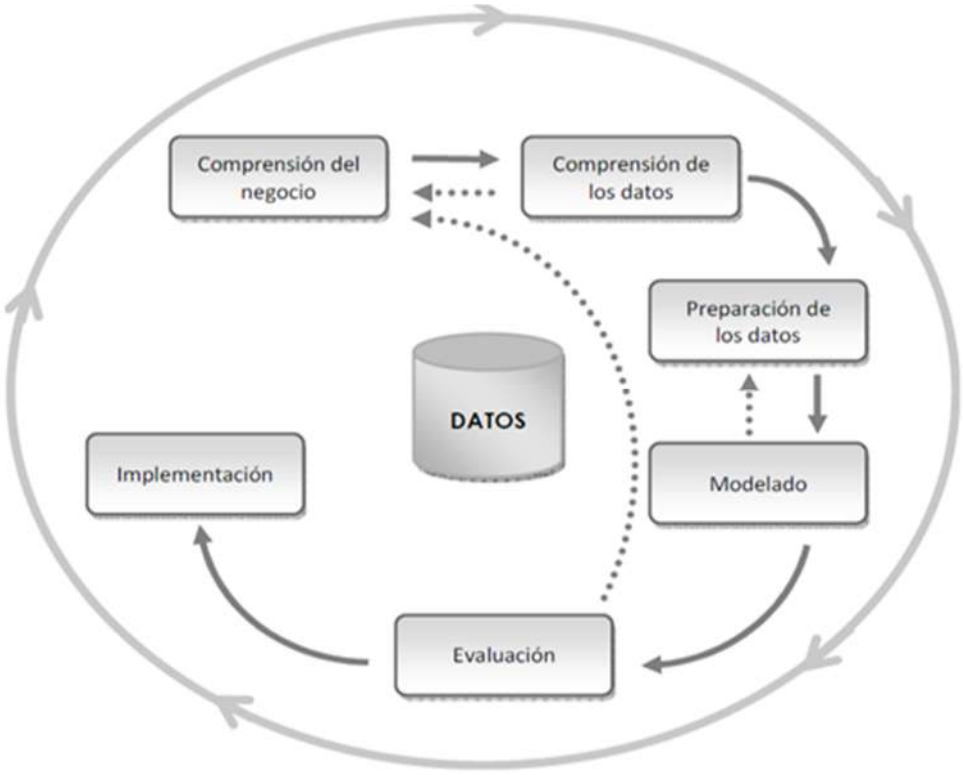


Fig. 5: Metodología CRISP-DM [*10*]

CRISP-DM propone en su nivel más alto, seis fases para el proceso de minería de datos: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación e implementación. La sucesión de fases, no es necesariamente rígida. Cada fase se descompone en un conjunto de tareas genéricas (o generales) de segundo nivel. Estas tareas son genéricas ya que tratan de abarcar la mayoría de las situaciones posibles en minería de datos. A partir del tercer nivel de abstracción, se realiza un "mapeo" de las tareas genéricas definidas en el modelo a situaciones específicas. De esta forma, las tareas genéricas se traducen en tareas específicas para casos y proyectos concretos. En el cuarto nivel, encontramos las instancias de proceso, donde se describen las acciones, decisiones y resultados de un proyecto particular de minería de datos. En las siguientes líneas, de la presente sección, se describe la metodología CRISP-DM según [*10*].

Analizando el nivel más alto de abstracción del modelo, las seis fases que componen el proceso de minería de datos son:

- ✓ **Comprensión del negocio:** en esta fase se determinan los objetivos y requerimientos del proyecto desde una perspectiva del negocio, definiendo el problema de minería y el plan de trabajo.

- ✓ **Comprensión de los datos:** fase que consiste en la recolección de datos que se utilizarán en el proyecto y la familiarización con los mismos. En esta etapa es posible el surgimiento de las primeras hipótesis acerca dela información que podría estar oculta.

- ✓ **Preparación de los datos:** comprende aquellas actividades de tratamiento de los datos para construir la vista o conjunto de datos final sobre el cual se aplicarán las técnicas de minería.

- ✓ **Modelado:** en esta etapa se aplican las diversas técnicas y algoritmos de minería sobre el conjunto de datos para obtener la información oculta y los patrones implícitos en ellos.

- ✓ **Evaluación:** fase en la que se analizan los patrones obtenidos en función de los objetivos organizacionales. En esta etapa se debería determinar si se ha omitido algún objetivo importante del negocio y si el nuevo conocimiento será implementado, es decir, si se pasará a la próxima etapa.

- ✓ **Implementación:** consiste en la comunicación e implementación del nuevo conocimiento, el cual debe ser representado de forma entendible para el usuario.

A continuación, se muestra por medio de tablas, las tareas que comprenden cada una de fases de la metodología CRISP-DM.

Tabla 1: Fase 1 (Entendimiento del Negocio) - Metodología CRISP-DM [*10*]

Fase 1: Entendimiento del Negocio	
Tarea 1.1: Determinar los objetivos del negocio	
Descripción	**Salidas**
Entender y establecer cuáles son los objetivos que el cliente pretende alcanzar, desde una perspectiva del negocio.	• Background, con la información que se conoce sobre la situación actual de la organización, incluyendo una descripción general del problema y la solución actual para el mismo (si es que existe). • Objetivos del negocio, identificando los objetivos principales del cliente. • Criterios de éxito, describiendo los resultados esperados desde una perspectiva de negocio.
Tarea 1.2: Evaluar la situación	
Descripción	**Salidas**
Profundizar en la evaluación de la situación actual del negocio. Analizar con mayor profundidad las restricciones y factores que se deben tener en cuenta para el proyecto.	• Inventario de recursos, donde deberán incluirse los recursos disponibles (como los recursos humanos, fuentes de datos, hardware y software). • Lista de requerimientos del proyecto, supuestos y restricciones que se han detectado. • Riesgos y planes de contingencia. Consiste en la identificación de los potenciales riesgos para el proyecto y la planificación de las acciones reactivas que se llevarán a cabo (planes de contingencia).

	• Glosario con terminología relevante para el proyecto. En el mismo deberá incluirse un glosario de terminología del negocio y otro de minería de datos. • Análisis costo-beneficio del proyecto.
Tarea 1.3: Determinar los objetivos de la minería de datos	
Descripción	**Salidas**
Los objetivos del negocio se describen en términos organizacionales, en cambio los objetivos de minería de datos describen los objetivos del proyecto en "términos técnicos". Es decir, si un objetivo de negocio es aumentar el volumen de ventas, el objetivo de minería de datos podría ser el "agrupamiento" de los clientes para la promoción de nuevas campañas publicitarias.	• Objetivos de minería de datos, describiendo los resultados previstos del proyecto que permiten el logro de los objetivos de negocio. • Definir un criterio de éxito para el proyecto de minería. Especificar las condiciones bajo las cuales se aceptarán los resultados obtenidos.
Tarea 1.4: Crear un plan para el proyecto de minería de datos.	
Descripción	**Salidas**
Crear una planificación para el proyecto de minería, el cual debe ser consistente con los objetivos planteados.	• Plan de proyecto: listar las tareas que deben ser ejecutadas, duraciones y recursos necesarios, así como sus entradas y salidas. El plan del proyecto es un documento dinámico, que debe ser revisado y ajustado al final de cada fase. • Evaluación inicial de técnicas y herramientas de minería que podrían ser utilizadas en el proyecto.

Tabla 2: Fase 2 (Entendimiento de los datos) - Metodología CRISP-DM [*10*]

Fase 2: Entendimiento de los datos	
Tarea 2.1: Recolectar los datos iniciales	
Descripción	**Salidas**
Recolectar todos los datos necesarios especificados en la lista de recursos del proyecto.	• Reporte de recolección inicial de datos, donde se detalla la forma en la que han sido obtenidos los conjuntos de datos y los problemas que han surgido en el proceso.
Tarea 2.2: Describir los datos	
Descripción	**Salidas**

Describir en líneas generales los datos recolectados.	• Descripción de los datos, incluyendo el formato de los mismos y su tamaño (como cantidad de registros y variables).
Tarea 2.3: Explorar los datos	
Descripción	**Salidas**
Realizar una exploración de los datos, observando la distribución y comportamiento de las variables con mayor relevancia. En esta fase es conveniente el uso de técnicas simples de análisis estadístico.	• Reporte inicial de exploración de datos, donde se expongan los resultados del análisis y las hipótesis iniciales con su impacto en el proyecto.
Tarea 2.4: Verificar la calidad de los datos	
Descripción	**Salidas**
Examinar la calidad de los datos, incluyendo un análisis de su completitud, de potenciales errores en los mismos y de los datos ausentes.	• Reporte de calidad de los datos, donde se documente el análisis de calidad efectuado y se propongan potenciales soluciones a los problemas encontrados.

Tabla 3: Fase 3 (Preparación de los datos) - Metodología CRISP-DM [*10*]

Fase 3: Preparación de los datos	
Tarea 3.1: Seleccionar los datos	
Descripción	**Salidas**
Seleccionar los datos que serán utilizados para el análisis. En esta etapa se debe seleccionar con qué atributos (columnas) y con qué observaciones (filas o registros) se trabajará. La selección debe estar justificada.	• Justificación de la selección. Documento donde se justifiquen las causas por las cuales se incluyeron y excluyeron los datos.
Tarea 3.2: Limpieza de datos	
Descripción	**Salidas**
Es una etapa que tiene por objetivo mejorar la calidad de los datos. En ella se deberán tomar decisiones sobre los problemas de calidad encontrados en los mismos, como datos ausentes o datos anómalos.	• Reporte de limpieza de datos, donde se incluyan las decisiones tomadas sobre los problemas de calidad de los datos (reportados en la fase “2.4 Verificar la calidad de los datos”)

Tarea 3.3: Construcción de los datos	
Descripción	**Salidas**
En esta fase se lleva a cabo la construcción de nuevos datos, derivados de los disponibles, que son importantes para el análisis. Estos nuevos datos pueden ser, por ejemplo, atributos calculados o atributos transformados.	• Atributos derivados. Estos atributos se calculan a partir de otros atributos del mismo registro. Por ejemplo: edad_cliente = fecha_venta – fecha_nacimiento. • Registros creados. Estos nuevos registros se crean cuando son necesarios en la fase posterior de modelado.
Tarea 3.4: Integrar los datos	
Descripción	**Salidas**
Consiste en la integración de datos provenientes de diferentes tablas o registros.	• Datos combinados. Resultan de integrar la información de dos o más tablas que tienen diferente información de las mismas observaciones. Por ejemplo, la integración de los datos personales y los datos de las atenciones efectuadas a un paciente en un centro de salud. En esta fase se incluye el cálculo de agregaciones, donde se calculan nuevos datos resumiendo información de diferentes tablas y registros. Siguiendo con el ejemplo del centro de salud, podríamos integrar en un solo registro los datos personales del paciente, el total de atenciones efectuadas, y el promedio anual de consultas médicas realizadas.
Tarea 3.5: Formatear los datos	
Descripción	**Salidas**
Esta etapa se refiere al cambio que debe realizarse en el formato de los datos (pero no en su significado) por los requisitos de las técnicas de modelado elegidas. Por ejemplo, el formato de las fechas o el ordenamiento del set de datos.	• Conjunto de datos reformateados.

Tabla 4: Fase 4 (Modelado) - Metodología CRISP-DM [*10*]

Fase 4: Modelado	
Tarea 4.1: Seleccionar la técnica de modelado	
Descripción	**Salidas**
Consiste en seleccionar qué técnica de minería de datos será utilizada. Por	• Técnica de modelado. Documentar la técnica de modelado con la que se

ejemplo, en un caso donde se ha definido un problema de agrupamiento (clustering), se puede decidir utilizar el algoritmo k-medias. Si se ha optado por el uso de múltiples técnicas, se debería repetir estatarea para cada una.	trabajará. • Supuestos del modelo. Algunas técnicas asumen supuestos sobre el conjunto de datos, como por ejemplo distribución normal de una variable. Documentar todos los supuestos realizados.

Tarea 4.2: Diseñar las pruebas del modelo	
Descripción	**Salidas**
Una vez construidos los modelos, necesitaremos un mecanismo para determinar su calidad y validez. Por ejemplo, en problemas de agrupamiento se puede utilizar el coeficiente de silueta para evaluar la robustez de los grupos encontrados y en problemas de clasificación la tasa de error para estimar la capacidad del clasificador. En esta fase se dividirá el conjunto de datos en un grupo para entrenar el modelo (training) y otro para probarlo (test).	• Diseño de los test. Determinar y documentar de qué forma se entrenarán y evaluarán los modelos generados. Incluir las decisiones tomadas sobre los datos que se utilizarán para entrenamiento y prueba.

Tarea 4.3: Construir el modelo	
Descripción	**Salidas**
Aplicar la técnica seleccionada sobre el conjunto de datos para generar uno o más modelos. En esta fase el modelo será evaluado con distintos valores de parámetros. Por ejemplo, en un algoritmo de agrupamiento kmedias, se podrían generar distintos modelos para diferentes valores de “k” o grupos.	• Parámetros seleccionados. Listar los parámetros que se han proporcionado al modelo, justificando la elección de los mismos. • Modelos producidos por las herramientas de minería. • Descripción de los modelos.

Tarea 4.4: Evaluar el modelo	
Descripción	**Salidas**
En esta fase, el equipo de proyecto interpreta y evalúa el modelo en función de su conocimiento del dominio, los criterios de éxito definidos para el proyecto (tarea 1.3) y las pruebas diseñadas para el modelo (tarea 4.2). Los modelos	• Evaluación de los modelos. Generar un reporte de evaluación de los modelos obtenidos, describiendo sus características y un ranking para los mismos. • Evaluación de los parámetros. En función de la evaluación anterior, revisar los parámetros y ajustar los mismos para volver a la fase de

pueden ser valorados y rankeados.	construcción del modelo (tarea 4.3). Repetir las etapas 4.3 y 4.4 hasta asegurarse de que se han encontrado los “mejores” modelos.

Tabla 5: Fase 5 (Modelado) - Metodología CRISP-DM [*10*]

Fase 5: Evaluación	
Tarea 5.1: Evaluarlos resultados	
Descripción	**Salidas**
En esta etapa se evalúa el modelo en función de los objetivos del negocio, determinando su validez de acuerdo a los intereses organizacionales. Además del modelo, puede haber surgido como parte del proceso nueva información relevante y futuras líneas de investigación.	• Evaluación de los resultados de la minería de datos con respecto a los criterios de éxito y objetivos de negocio. • Modelos evaluados y aprobados.
Tarea 5.2: Revisión del proceso	
Descripción	**Salidas**
Realizar una revisión completa del proceso efectuado en búsqueda de posibles errores u omisiones.	• Revisión del proceso, documentando un resumen del mismo. Incluir las actividades omitidas o bien aquellas que deberían ser repetidas.
Tarea 5.3: Determinar las próximas etapas	
Descripción	**Salidas**
En función de la evaluación de resultados y la revisión del proceso, se debe decidir cómo continua el proyecto: si se pasa a la próxima fase (implementación) o bien si se retorna a una fase anterior.	• Lista de posibles acciones. • Descripción de la decisión tomada.

Tabla 6: Fase 6 (Implementación) - Metodología CRISP-DM [*10*]

Fase 6: Implementación	
Tarea 6.1: Planificar la implementación	
Descripción	**Salidas**
En esta etapa se genera el plan de implementación de los resultados	• Plan de implementación, incluyendo las etapas y cómo llevarlas a cabo.

obtenidos mediante la minería de datos.	
Tarea 6.2: Planificar el monitoreo y el mantenimiento	
Descripción	**Salidas**
El monitoreo y mantenimiento es de gran importancia si los resultados de la minería formarán parte del trabajo diario del negocio y su entorno.	• Plan de mantenimiento y monitoreo.
Tarea 6.3: Crear un reporte final	
Descripción	**Salidas**
Generar un reporte final, que podría resumir el desarrollo del proyecto o bien mostrar un análisis comprensivo de los resultados obtenidos en el proceso de minería.	• Reporte final del proyecto. • Presentación final al cliente, incluyendo resultados y conclusiones.
Tarea 6.4: Revisión del proyecto	
Descripción	**Salidas**
Consiste en identificar y analizar los puntos que fueron bien realizados, los que fueron mal realizados, y los que podrían mejorarse.	• Documentación de la experiencia adquirida durante el desarrollo del proyecto.

2.2.5. Técnicas de Minería de Datos

Las técnicas de minería de datos se han convertido en una herramienta básica en el campo de la investigación como método de análisis y descubrimiento de conocimiento a partir de datos almacenados para la toma de decisiones en: procesos productivos, marketing, finanzas etc. Las técnicas de minería de datos también son conocidas como algoritmos de minería de datos que se aplican a grandes conjuntos de datos. La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos. Dichas técnicas emergentes se encuentran en continua evolución como resultado de la colaboración entre campos de investigación tales como bases de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadística, visualización, recuperación de información, y computación de altas prestaciones [*12*]. Se clasifican en dos grandes categorías: supervisadas o predictivas y no

supervisadas o descriptivas [*13*]. En la figura Fig.6 se puede observar la clasificación de las técnicas de minería de datos.

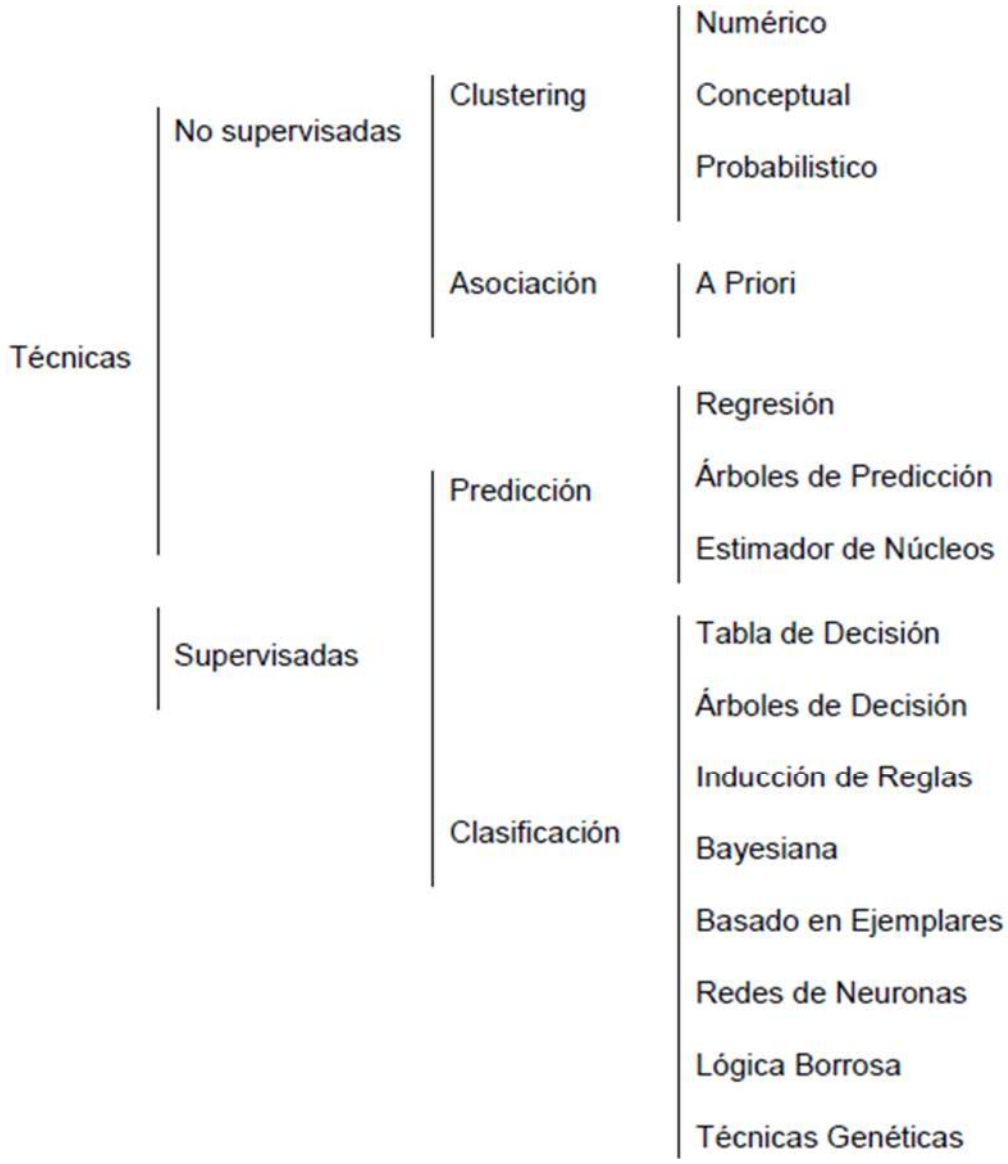


Fig. 6: Técnicas de la Minería de Datos [*13*]

Una técnica constituye el enfoque conceptual para extraer la información de los datos, y, en general, es implementada por varios algoritmos. Cada algoritmo representa, en la práctica, la manera de desarrollar una determinada técnica paso a paso, de forma que es preciso un entendimiento de alto nivel de los algoritmos para saber cuál es la técnica más apropiada para cada problema. Asimismo, es preciso entender los parámetros y las características de los algoritmos para preparar los datos a analizar. A continuación, se van a describir las técnicas más utilizadas y las que tienen mayor relevancia con la investigación [*13*].

- **Clustering**

  También llamada agrupamiento, permite la identificación de tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. Así se puede segmentar el colectivo de clientes, el conjunto de valores e índices financieros, el espectro de observaciones astronómicas, el conjunto de zonas forestales, el conjunto de empleados y de sucursales u oficinas, etc. La principal característica de esta técnica es la utilización de una medida de similitud que, en general, está basada en los atributos que describen a los objetos, y se define usualmente por proximidad en un espacio multidimensional. Para datos numéricos, suele ser preciso preparar los datos antes de realizar data mining sobre ellos, de manera que en primer lugar se someten a un proceso de estandarización.

- **Reglas de Asociación**

  Este tipo de técnicas se emplea para establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes; pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros. Son utilizadas cuando el objetivo es realizar análisis exploratorios, buscando relaciones dentro del conjunto de datos. Las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y co-ocurrencias de eventos. Debido a sus características, estas técnicas tienen una gran aplicación práctica en muchos campos como, por ejemplo, el comercial ya que son especialmente interesantes a la hora de comprender los hábitos de compra de los clientes y constituyen un pilar básico en la concepción de las ofertas y ventas cruzada, así como del "merchandising". En otros entornos como el sanitario, estas herramientas se emplean para identificar factores de riesgo en la aparición o complicación de enfermedades. Para su utilización es necesario disponer de información de cada uno de los sucesos llevados a cabo por un mismo individuo o cliente en un determinado período temporal. Por lo general esta forma de

extracción de conocimiento se fundamenta en técnicas estadísticas, como los análisis de correlación y de variación.

- **La predicción**

Es el proceso que intenta determinar los valores de una o varias variables, a partir de un conjunto de datos. La predicción de valores continuos puede planificarse por las técnicas estadísticas de regresión. Por ejemplo, para predecir el sueldo de un graduado de la universidad con 10 años de experiencia de trabajo, o las ventas potenciales de un nuevo producto dado su precio. Se pueden resolver muchos problemas por medio de la regresión lineal, por ejemplo, y puede conseguirse todavía más, aplicando las transformaciones a las variables para que un problema no lineal pueda convertirse a uno lineal.

- **La clasificación**

La clasificación es el proceso de dividir un conjunto de datos en grupos mutuamente excluyentes, de tal forma que cada miembro de un grupo esté lo más cerca posible de otros y grupos diferentes estén lo más lejos posible de otros, donde la distancia se mide con respecto a las variables especificadas, que se quieren predecir. A continuación, se presenta una tabla de datos, donde el problema es saber si se debe jugar o no de acuerdo a los datos que se tiene. En la tabla 7, se puede ver un ejemplo de un problema de clasificación.

Tabla 7: Ejemplo de problema de clasificación [*13*]

Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
1	Soleado	Alta (85)	Alta (85)	No	No
2	Soleado	Alta (80)	Alta (90)	Sí	No
3	Nublado	Alta (83)	Alta (86)	No	Sí
4	Lluvioso	Media (70)	Alta (96)	No	Sí
5	Lluvioso	Baja (68)	Normal (80)	No	Sí
6	Lluvioso	Baja (65)	Normal (70)	Sí	No

7	Nublado	Baja (64)	Normal (65)	Sí	Sí
8	Soleado	Media (72)	Alta (95)	No	No
9	Soleado	Baja (69)	Normal (70)	No	Sí
10	Lluvioso	Media (75)	Normal (80)	No	Sí
11	Soleado	Media (75)	Normal (70)	Sí	Sí
12	Nublado	Media (72)	Alta (90)	Sí	Sí
13	Nublado	Alta (81)	Normal (75)	No	Sí
14	Lluvioso	Media (71)	Alta (91)	Sí	No

El ejemplo empleado tiene dos atributos, temperatura y humedad, que pueden emplearse como simbólicos o numéricos. Entre paréntesis se presentan sus valores numéricos. Representa un sencillo problema de clasificación que consistente en que, a partir de los atributos que modelan el tiempo (vista, temperatura, humedad y viento), determinar si se puede o no jugar al tenis.

A continuación, se pasa explicar una de las técnicas más conocidas de clasificación como son los Árboles de decisión. También, se hará uso de la tabla de ejemplo anterior (Tabla 7.) para darle solución al ejemplo haciendo uso de un algoritmo de clasificación.

Árboles de decisión: El aprendizaje de árboles de decisión está englobado como una metodología del aprendizaje supervisado. La representación que se utiliza para las descripciones del concepto adquirido, es el árbol de decisión, que consiste en una representación del conocimiento relativamente simple, y que es una de las causas por la que los procedimientos utilizados en su aprendizaje son más sencillos que los de sistemas que utilizan lenguajes de representación más potentes, como redes semánticas, representaciones en lógica de primer orden etc. No obstante, la potencia expresiva de los árboles de decisión es también menor que la de esos otros sistemas. El aprendizaje de árboles de decisión suele ser más robusto frente al ruido y conceptualmente sencillo, aunque los sistemas que han resultado del perfeccionamiento y de la evolución de los más antiguos se complican con los procesos que incorporan para ganar

fiabilidad. Un árbol de decisión puede interpretarse esencialmente como una serie de reglas compactadas para su representación en forma de árbol. Dado un conjunto de ejemplos, estructurados como vectores de pares ordenados atributo-valor, de acuerdo con el formato general en el aprendizaje inductivo a partir de ejemplos, el concepto que estos sistemas adquieren durante el proceso de aprendizaje consiste en un árbol. Cada eje está etiquetado con un par atributo-valor y las hojas con una clase, de forma que la trayectoria que determinan desde la raíz los pares de un ejemplo de entrenamiento alcanzan una hoja etiquetada -normalmente- con la clase del ejemplo. Podemos citar a continuación algunas características de los problemas que se podrían resolver con esta técnica:

- Especialmente cuando los valores son disjuntos y en un número pequeño. Los sistemas actuales están preparados para tratar atributos con valores continuos, valores desconocidos e incluso valores con una distribución de probabilidad.
- Cuando el atributo que hace el papel de la clase sea de tipo discreto y con un número pequeño de valores, sin embargo, existen sistemas que adquieren como concepto aprendido funciones con valores continuos.
- Cuando las descripciones del concepto adquirido deban ser expresadas en forma normal disyuntiva.

El primer sistema que construía árboles de decisión fue CLS de Hunt, desarrollado en 1959 y depurado a lo largo de los años sesenta. Siguiendo esta misma idea, en 1979 Quinlan desarrolla el sistema ID3, que él denominaría simplemente herramienta porque la consideraba experimental. La versión definitiva, presentada por su autor Quinlan es el sistema C4.5 que expone con cierto detalle en la obra C4.5: Programsfor Machine Learning. La evolución -comercial- de ese sistema es otro denominado C5 del mismo autor, del que se puede obtener una versión de demostración restringida en cuanto a capacidades; por ejemplo, el número máximo de ejemplos de entrenamiento. El procedimiento para generar un árbol de decisión consiste, como se comentó anteriormente en seleccionar un

atributo como raíz del árbol y crear una rama con cada uno de los posibles valores de dicho atributo. Con cada rama resultante (nuevo nodo del árbol), se realiza el mismo proceso, esto es, se selecciona otro atributo y se genera una nueva rama para cada posible valor del atributo. Este procedimiento continúa hasta que los ejemplos se clasifiquen a través de uno de los caminos del árbol. El nodo final de cada camino será un nodo hoja, al que se le asignará la clase correspondiente. Así, el objetivo de los árboles de decisión es obtener reglas o relaciones que permitan clasificar a partir de los atributos.

En la siguiente figura Fig.7, podemos ver un ejemplo de la solución al problema planteado a en la tabla (Tabla 7.), haciendo uso de una técnica de clasificación, como es los Árboles de decisión, mediante el algoritmo de clasificación ID3.

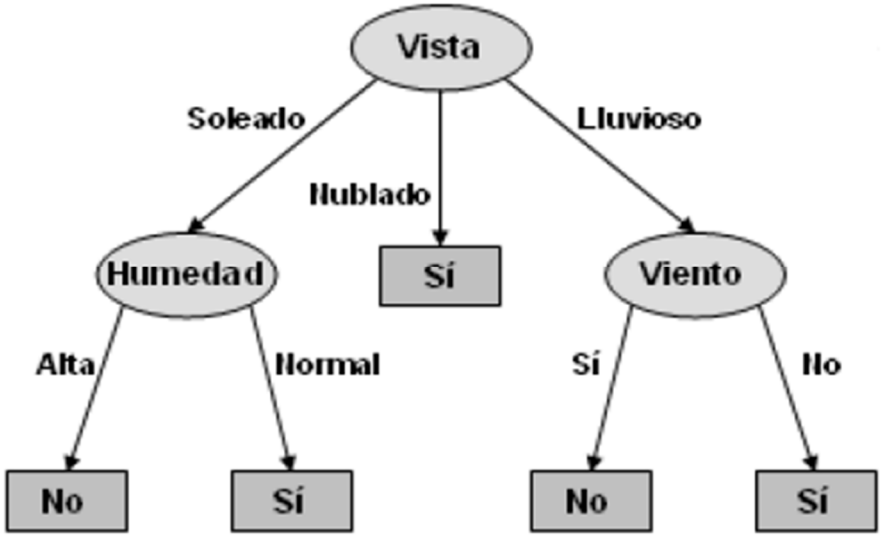


Fig. 7: Ejemplo de clasificación con ID3 [*13*]

El primer nodo del árbol se muestra cómo se llega a decidir que el mejor atributo para dicho nodo es “vista”. Se generan nodos para cada valor del atributo y, en el caso de “vista” = “Nublado” se llega a un nodo hoja ya que todos los ejemplos de entrenamiento que llegan a dicho nodo son de clase “Sí”. Sin embargo, para los otros dos casos se repite el proceso de elección con el resto de atributos y con los ejemplos de entrenamiento que se clasifican a través de ese nodo.

- **Redes de Neuronas**

Las redes de neuronas constituyen una técnica inspirada en los trabajos de investigación, iniciados en 1930, que pretendían modelar computacionalmente el aprendizaje humano llevado a cabo a través de las neuronas en el cerebro. Posteriormente se comprobó que tales modelos no eran del todo adecuados para describir el aprendizaje humano. Las redes de neuronas constituyen una nueva forma de analizar la información con una diferencia fundamental con respecto a las técnicas tradicionales: son capaces de detectar y aprender complejos patrones y características dentro de los datos. Se comportan de forma parecida a nuestro cerebro aprendiendo de la experiencia y del pasado, y aplicando tal conocimiento a la resolución de problemas nuevos. Este aprendizaje se obtiene como resultado del adiestramiento ("training") y éste permite la sencillez y la potencia de adaptación y evolución ante una realidad cambiante y muy dinámica. En aquellos casos de muy alta complejidad las redes neuronales se muestran como especialmente útiles dada la dificultad de modelado que supone para otras técnicas. Sin embargo, las redes de neuronas tienen el inconveniente de la dificultad de acceder y comprender los modelos que generan y presentan dificultades para extraer reglas de tales modelos. Otra característica es que son capaces de trabajar con datos incompletos e, incluso, contradictorios lo que, dependiendo del problema, puede resultar una ventaja o un inconveniente. En la figura Fig.8 se muestra la estructura de una red de neuronas.

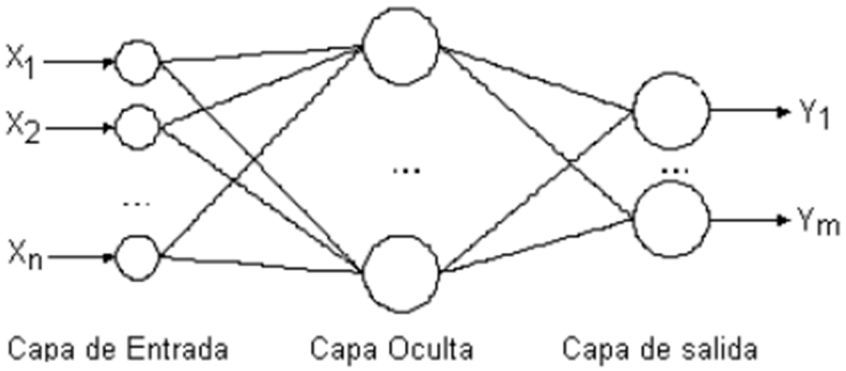


Fig. 8: Estructura de la red de neuronas [*13*]

2.2.6. Estructura de minería de datos

Todas las técnicas de minería de datos se aplican sobre conjuntos de datos que están organizadas a través de hojas de cálculo, tablas o vistas, u otro medio que permitan ser analizados. Por eso, es importante definir una estructura sobre la cual se aplicará las técnicas y sus algoritmos, para generar los modelos. A continuación, según la empresa Microsoft [*14*], se dice que la estructura de minería de datos es una estructura de datos que define el dominio de datos a partir del cual se generan los modelos de minería de datos. Una única estructura de minería de datos puede contener varios modelos de minería de datos que comparten el mismo dominio. Las unidades de creación de la estructura de minería de datos son las columnas de la estructura de minería de datos, que describen los datos que contiene el origen de datos. Estas columnas contienen información como el tipo de datos, el tipo de contenido y el modo en que se distribuyen los datos. Una estructura de minería de datos también puede contener tablas anidadas. Una tabla anidada representa una relación de uno a varios entre la entidad de un escenario y sus atributos relacionados. En la siguiente figura Fig.9 se tiene una representación de una estructura de minería de datos.

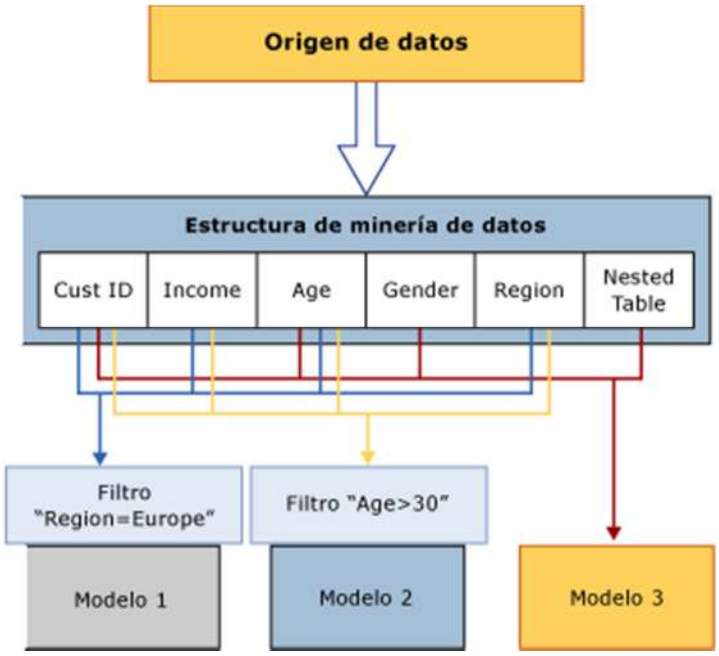


Fig. 9: Ejemplo de estructura de Minería de Datos de Microsoft [*14*]

La estructura de minería de datos del diagrama (Fig.9) está basada en un origen de datos que contiene varias tablas o vistas, combinadas en el campo CustomerID. Una tabla contiene información sobre los clientes, como la región geográfica, la edad, los ingresos y el sexo, mientras que la tabla anidada relacionada contiene varias filas de información adicional sobre cada cliente, como los productos que ha adquirido. En el diagrama, se muestra que se pueden generar varios modelos de minería de datos a partir de una misma estructura de minería de datos, y que los modelos pueden usar columnas de la estructura diferentes. En la figura Fig.9, del ejemplo se puede ver cómo están organizados los modelos, veamos a continuación:

Modelo 1: usa CustomerID, Income, Age, Region y filtra los datos de Region.
Modelo 2: usa CustomerID, Income, Age, Region y filtra los datos de Age.
Modelo 3: usa CustomerID, Age, Gender y la tabla anidada, sin filtros.

Dado que los modelos usan columnas diferentes para la entrada, y dado que dos de los modelos, además, restringen sus datos mediante la aplicación de un filtro, los modelos pueden tener resultados muy diferentes, aunque estén basados en los mismos datos. Observe que la columna CustomerID es obligatoria en todos los modelos porque es la única columna disponible que se puede usar como clave de caso. La configuración de una estructura de minería de datos consta de los pasos siguientes:

- Definir un origen de datos.
- Seleccionar las columnas de datos que se van a incluir en la estructura (no es necesario agregar todas las columnas al modelo) y definir una clave.
- Definir una clave para la estructura, incluyendo la clave de la tabla anidada, si procede.
- Especificar si los datos de origen se deben separar en un conjunto de entrenamiento y en un conjunto de prueba. Este paso es opcional.
- Procesar la estructura.

Siguiendo con las definiciones según [*14*], Microsoft ofrece dos maneras de construir estructuras de minería de datos:

- **Estructura de minería de datos relacional**

  Una estructura de minería de datos relacional puede extraer datos de orígenes dispares. Los datos sin procesar se pueden almacenar en tablas, archivos o sistemas de bases de datos relacionales, siempre y cuando los datos puedan definirse como parte de la vista del origen de datos. Por ejemplo, debe utilizar una estructura de minería de datos relacional si los datos están en Excel, en un almacén de datos de SQL Server, en la base de datos de informes de SQL Server o en los orígenes externos a los que se tiene acceso a través de los proveedores OLE DB u ODBC.

- **Estructura de minería de datos OLAP**

  La creación de un modelo de minería de datos basado en un cubo OLAP o en otro almacén de datos multidimensionales presenta numerosas ventajas. Una solución OLAP ya contiene enormes cantidades de datos que han sido bien organizados, limpiados y con un formato correcto.

2.2.7. Pronóstico de ventas

Un pronóstico de ventas es un cálculo de las ventas probables de la marca de un producto de una compañía durante un periodo señalado en un mercado específico, suponiendo que se sigue un plan de marketing definido [*15*].

- **Importancia del pronóstico de ventas en las empresas**

Según Stanton, Etzel y Walker [*16*], cuando se ha preparado el pronóstico de ventas, atañe a todos los departamentos de la compañía. El pronóstico de ventas es la base para decidir cuánto gastar en diversas actividades como publicidad y ventas personales. Con la base de las ventas anticipadas se planea la cantidad

necesaria de capital de trabajo, la utilización de la planta y las instalaciones de almacenaje. También dependen de éstos pronósticos el calendario de producción, la contratación de operarios fabriles y la compra de materias primas.

♦ **Alcance del pronóstico de ventas**

Es recomendable elaborar un pronóstico de ventas para cada producto (incluyendo cada uno de los items o presentaciones que tenga), línea de productos y para la empresa en su conjunto, porque de esa manera se podrá tomar decisiones más acertadas (especialmente en lo relacionado a producción, aprovisionamiento y flujo de caja) y además, se podrá realizar un mejor monitoreo y control al momento de cruzar los resultados del esfuerzo de mercadotecnia con el cumplimiento del pronóstico de ventas [*16*].

♦ **Métodos de pronósticos**

▪ **Método cuantitativo**: Son métodos estadísticos o matemáticos que se desarrollan basándose en la información histórica, ya sea de la propia empresa o del mercado en general [*17*]. A continuación de describen tipos de métodos cuantitativos:

- **Análisis de ventas históricas y la tendencia**: Consiste en pronosticar teniendo en cuenta las ventas y demanda del pasado, considerando factores del momento.

- **Análisis de los factores de mercado**: Dado que la demanda de un producto siempre se relaciona con el comportamiento de ciertos factores de mercado, se puede determinar una estimación de venta estudiando los factores relacionados con el producto.

- **Método de derivación directa:** Se trata de un estudio de los factores relacionados con un producto y las consecuencias directas de su uso y compra, determinando aspectos como desecho, recambio, rotura, moda, etc.

- **Análisis de correlación**: mide la relación directa entre dos datos o factores de mercado, se puntúa de 0 (sin relación) a 1 (relación perfecta).

- **Pruebas de mercados**: Este método consiste en realizar una prueba piloto en donde se ofrezca el nuevo producto en determinadas zonas con el fin de evaluar la respuesta del consumidor, y en base a ello pronosticar las ventas.

- **Ventas de la competencia**: Este método consiste en calcular las ventas de la competencia, y tomar éstas como referencia para pronosticar las nuestras.

- **Encuestas**: Este método consiste en obtener información a través de encuestas en donde las preguntas estarían relacionadas con la intención de compra, la frecuencia de compra y el gasto promedio.

Los métodos cuantitativos tienen como ventaja la objetividad. Los números reflejan la realidad pasada inmediata. Su limitación recae en que los pronósticos tienden a generalizar sobre la base de experiencias pasadas.

▪ **Método cualitativo:** También se les denomina Subjetivos, son pronósticos generados a partir de información que no contiene una estructura analítica bien definida. Son tipos de pronósticos que resultan útiles cuando no se cuenta con información histórica [*17*]. A continuación, se describe los tipos de métodos cualitativos:

- **Encuesta de las intenciones del comprador**: También conocido como Método de Expectativas del usuario ya que depende de las respuestas que los consumidores den en cuanto al consumo o las compras que esperan realizar del producto.

- **Participación de la fuerza de ventas**: Consiste en pronosticar las ventas con las estimaciones de la fuerza de ventas (vendedores, distribuidores, jefes de ventas, etc.). La ventaja de éste método es que abarca a las

personas que tendrán la responsabilidad de los resultados, ayuda a controlar y dirigir el esfuerzo de ventas. La desventaja es que los vendedores tienen un interés manifiesto y los resultados podrían verse alterados.

- **Juicio de los ejecutivos**: Este método consiste en hacer un sondeo interno o informal, de la opinión de los ejecutivos claves de la empresa, para saber cómo evalúan las posibilidades de las ventas. La ventaja de este método es que es fácil y rápido de realizar; y no requiere estadísticas elaboradas. Las desventajas es que la responsabilidad no la toman los vendedores, los pronósticos pueden ser más elevados de lo que realmente podrían ser.

▪ **Minería de datos como método de pronóstico:** La minería de datos (DM, Data Mining) consiste en la extracción no trivial de información que reside de manera implícita en los datos, es decir, el proceso de estudiar datos para encontrar información y relaciones previamente desconocidas; esta información es entonces aplicada para lograr objetivos específicos del negocio. En otras palabras, podemos decir que dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos. Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. Está fuertemente ligado con la supervisión de procesos industriales ya que resulta muy útil para aprovechar los datos almacenados en las bases de datos. Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación. La minería de datos, para pronosticar utiliza técnicas de la minería de datos provienen de la Inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados. Muchos nuevos algoritmos han sido desarrollados, y el horizonte del análisis de datos ha sido significativamente expandido. Una de las técnicas más representativas los arboles de decisión [*18*].

- **Árboles de decisión:** Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.

  El funcionamiento para pronósticos, de la técnica los arboles de decisión, así como los algoritmos utilizados, son detalladamente descritos en el apartado anterior 2.2.5: Técnicas de minería de datos: Clasificación: Arboles de decisión, del presente documento.

♦ **Pautas para mejorar el pronóstico de ventas**

A continuación se describen pautas muy importantes a tener en cuenta para lograr la efectividad de los pronósticos de ventas en las empresas [*18*].

- **Involucrar a altos ejecutivos**: Pronósticos fuertes requieren el soporte de altos ejecutivos debido a los recursos requeridos, estos ejecutivos deben aprobar la instalación de la tecnología, así como, la contratación de profesionales en el desarrollo de pronósticos. Además, una vez involucrados, se generan reuniones de seguimiento para entender de una mejor manera la demanda y todo lo relacionado para cubrirla.

- **Explicar los beneficios mutuos**: Dado que es común que algunas áreas se resistan a participar en el desarrollo de pronósticos, resulta necesario el explicar el beneficio específico que tendrá el área en cuestión con el desarrollo de un buen pronóstico. Esta recomendación aplica, incluso para clientes y proveedores externos, pues si se puede motivar a los clientes a compartir su información se puede mejorar el nivel de la cadena de suministros.

- **Definir claramente metas y acuerdos**: La mejor manera de ver que los pronósticos de una empresa están mejorando es cuando la cadena de

suministros se vuelve más eficiente, para esto es necesario dejar en claro cuáles son las métricas que se estarán desarrollando, esto toma especial importancia en los acuerdos logrados con externos.

- **Utilizar la mejor tecnología**: Las compañías deben utilizar la tecnología de vanguardia referente a pronósticos, que pueda asegurar el mejor tratamiento a los datos disponibles y el mejor desempeño de los pronósticos logrados, que permita el uso de datos estandarizados que puedan ser compartidos con todas las áreas de la organización. Actualmente ya existe software que apoya a esta labor, sin embargo, se sigue esperando software de simulación más avanzado, que puedan llevar a decisiones más rápidas y acertadas.

- **Enfocarse donde la utilidad es mayor**: Debido a que usualmente los recursos son limitados, las compañías suelen enfocar sus pronósticos en aquellos productos que generan mayor utilidad, lo cual viene a ser un resultado de la ley general de clasificación ABC.

- **Ligar incentivos a las metas generales de la compañía**: Para asegurar una mejor precisión de pronósticos, los incentivos y premios de empleados deben basarse en las metas generales de la compañía, más que en las metas específicas de un departamento. Cuando no se hace de esta manera, las áreas se enfrascan en sus propias metas y se contraponen en situaciones comunes de trabajo.

- **Seguir con mejora continua**: Los errores en pronósticos pueden resultar de datos incorrectos, suposiciones inadecuadas o a un modelo defectuoso, por lo que es importante llevar a cabo un análisis una vez que se ha terminado el reporte de cada periodo, para tomar acciones y corregir problemas.

A manera de resumen, se puede decir que los pronósticos se pueden realizar haciendo uso de herramientas de software o de manera manual. Se pueden hacer uso de datos históricos de ventas, o basada en

experiencia de personal del área de ventas u opiniones de expertos. Los pronósticos son de mucho beneficio para las empresas, desde varias perspectivas, pues ayuda a planificar las operaciones de ventas a futuro, expandir la línea de productos o descontinuar productos a futuro, optimizar el nivel de inventario de acuerdo al nivel de rotación, orientar las acciones de marketing de los productos reduciendo a si costos innecesarios. Sin embargo, los pronósticos podrían, también tener algunas desventajas, por ejemplo, cuando se quiere pronosticar productos nuevos, pues dichos productos no tienen un historial de datos, otro ejemplo sería cuando se quiere realizar un pronóstico de manera manual es muy costoso y depende de la organización de la información. Actualmente, el avance de la tecnología de la información, permite realizar el almacenaje de los datos de manera óptima, así como su transformación para muchos fines empresariales, lo cual es una ventaja importante al momento de hacer pronósticos de ventas con datos históricos.

2.2.8. Herramientas para el proceso de minería de datos utilizados

Las herramientas de software usadas en un proceso de minería, pueden variar según la problemática que se tenga que abordar. Se debe tener en cuenta también el tema de las licencias de uso que tiene cada herramienta.

Para el presente trabajo de investigación de ha hecho uso de las siguientes herramientas, desde luego, se indica que son todas herramientas de la categoría Open Source y son:

- **Mysql**

  MySQL es un sistema de gestión de base de datos relacional (RDBMS) de código abierto, basado en lenguaje de consulta estructurado (SQL). MySQL se ejecuta en prácticamente todas las plataformas, incluyendo Linux, UNIX y Windows [*19*].

- **Pentaho Data Integration**

  Es una herramienta que permite implementar los procesos de extracción, transformación y carga de datos. El uso de kettle permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar [*20*].

- **Schema Workbench**

  MondrianSchemaWorkbench es una interfaz de diseñador que le permite crear y probar esquemas de cubo OLAP Mondrian visualmente. El motor Mondrian procesa las solicitudes MDX con los esquemas ROLAP (OLAP relacional). Estos archivos de esquema son modelos de metadatos XML que se crean en una estructura específica utilizada por el motor Mondrian. Estos modelos XML se pueden considerar estructuras tipo cubo que utilizan tablas FACT y DIMENSION existentes que se encuentran en su RDBMS. No requiere que se construya o mantenga un cubo físico real; solo que se crea el modelo de metadatos [*21*].

- **JRubik**

  Es una aplicación cliente capaz de conectar a fuentes Olap basadas en el motor relacional Mondrian. Las consultas Olap pueden realizarse por medio del lenguaje MDX [*22*].

- **WEKA**

  Weka. Waikato Environment for Knowledge Analysis - Entorno para Análisis del Conocimiento de la Universidad de Waikato), es una Plataforma de Software para aprendizaje Automático y minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es un software libre distribuido bajo licencia GNU-GPL [*23*].

- **Comparación de Weka con otras herramientas de minería**

Tabla 8: Comparación de otras herramientas de minería de datos con Weka

Herramienta	Descripción	Código abierto	Multiplataforma
Weka	WEKA como una plataforma pública de trabajo de minería de datos, una colección de una gran cantidad de tareas de minería de datos puede soportar el algoritmo de aprendizaje automático, que incluye pre procesamiento de datos, clasificación, regresión, clustering, reglas de asociación y visualización en la nueva interfaz interactiva [*24*].	SI	SI
Orange	Orange es una suite de software de aprendizaje de aplicaciones y minería de datos basada en componentes que presenta una interfaz de programación visual amigable, potente, rápida y versátil para examinar el análisis y visualización de datos, con enlace de Python para scripts. Contiene un conjunto completo de componentes para pre procesamiento de datos y proporciona funciones de contabilidad, transición, modelado, evaluación de patrones y exploración de datos. Desarrollado por C ++ y Python, su biblioteca	SI	SI

	de gráficos se basa en un marco Qt multiplataforma [*24*].		
KNINE	KNINE (Konstanz Information Miner) es una plataforma de integración de datos, procesamiento de datos, exploración de datos y exploración de datos de fácil uso e inteligente [*24*].	SI	SI
SQL Server Analysis Service	Analysis Services proporciona una plataforma integrada para soluciones que incorporan minería de datos. Puede usar datos relacionales o de cubo para crear soluciones de inteligencia empresarial con análisis predictivos [*25*].	NO	SI
SAS Analytics y Enterprise	Desarrolla modelos para minería de datos y trabajos estadísticos. Un entorno de diagrama de flujo de proceso interactivo y auto-documentado mapea eficientemente todo el proceso de minería de datos para producir los mejores resultados. Y tiene más técnicas de modelado predictivo que cualquier otro paquete de minería de datos comercial [*26*].	NO	SI

- **Algoritmos de clasificación de weka**

A continuación, se realiza una breve descripción de los algoritmos de clasificación que fueron tomados en cuenta para construir el modelo de minería según [*27*]

Tabla 9: Descripción de algoritmos de clasificación usados para generar el modelo

Algoritmo	Descripción
REPTree	Aprende rápidamente el árbol de decisión. Construye un árbol de decisión por medio de regresión usando ganancia por la varianza de información y lo poda usando poda de errores reducidos, con ajuste posterior. Solo ordena valores para atributos numéricos una vez. Los valores faltantes se tratan dividiendo las instancias correspondientes en partes.
AttributeSelectedClassifier	La dimensionalidad de los datos de entrenamiento y prueba se reduce mediante la selección de atributos antes de pasarse a un clasificador. El tamaño del árbol generado, en algunos casos es menor que otros algoritmos.
J48	Es uno de los algoritmos de clasificación más conocido, cuyo procedimiento es seleccionar un atributo como raíz del árbol y crear una rama con cada uno de los posibles valores de dicho atributo, con cada rama resultante que es un nuevo nodo del árbol, se realiza el mismo proceso. Este procedimiento continúa hasta que los ejemplos se clasifiquen a través de uno de los caminos del árbol. El nodo final de cada camino será un nodo hoja, al que se le asignará la clase correspondiente. Su objetivo es obtener reglas o relaciones que permitan clasificar a partir de los atributos.
Jrip	Esta clase implementa reglas proposicionales básicas, Poda repetida e incrementalmente para producir

	reducción de errores (RIPPER), que fue propuesto por William W. Cohen como una versión optimizada de IREP.
LMT	Clasificador para construir 'árboles de modelo logístico', que son árboles de clasificación con funciones de regresión logística en las hojas. El algoritmo puede tratar con variables de destino binarias y de clases múltiples, atributos numéricos y nominales y valores perdidos.

2.3. Definición de términos básicos

- **Operatividad**

  Capacidad para funcionar o estar en activo [*28*]. Desde el punto de vista informático se dice que es la Capacidad para estar operativo al usuario.

- **Funcionalidad**

  Conjunto de características que hacen que algo sea práctico y utilitario [*28*].

- **Análisis de ventas**

  Permite tomar decisiones sobre las orientaciones comerciales de la empresa. Para ello es necesario contar con una información cuantitativa y cualitativa, a nivel general de las ventas de la empresa, a nivel de delegación, de vendedor [*29*].

- **Procesos de marketing**

  El proceso de marketing es el proceso mediante el cual se buscan oportunidades de negocios, se segmenta el mercado y se selecciona un mercado resultante, se analiza dicho mercado, se formulan estrategias de marketing, se diseñan planes de acción, se implementan las estrategias, y se controlan y evalúan los resultados [*30*].

- **Modelo de Minería de datos**

  Es un conjunto de datos, estadísticas y patrones que se pueden aplicar los nuevos datos para generar predicciones y deducir relaciones, que se crea mediante la aplicación de un algoritmo a los datos [*31*].

- **Pronostico de ventas**

  Un pronóstico de ventas es un cálculo de las ventas probables de la marca de un producto, de una compañía durante un periodo señalado [*15*].

- **Impacto**

  Es el efecto que determinados fenómenos tienen sobre la realidad [*32*]. De acuerdo al presente trabajo, el impacto estará representado por el cambio, en la forma de hacer pronóstico de ventas. Sera positivo si el personal que toma decisiones en el área de ventas está conforme con el modelo y los resultados que serán comparados con las ventas de periodos futuros.

CAPÍTULO III. MATERIALES Y MÉTODOS

El trabajo de investigación se desarrolló en la empresa “CELLSERVICE EIRL”, en la ciudad de Cajamarca - Perú, donde actualmente funciona como una agencia de la empresa de telefonía CLARO. El trabajo de investigación se ha desarrollado entre los años 2016 y 2017.

3.1. Procedimiento

El procedimiento esta dado de acuerdo a la metodología de minería de datos usada para construir el modelo de minería.

3.1.1. Metodología del modelo de Minería de datos

Se ha tomado como referencia la metodología de minería de datos CRISP – DM, que es un modelo de procesos para proyectos de minería de datos, después de analizar cada una de las metodologías mencionadas en el capítulo dos (CAPITULO II), en el apartado 2.2.4 (Metodología CRISP-DM). Se ha desarrollado cada una de las tareas y fases de la metodología, según el caso de la presente investigación. Es importante aclarar que uno de los objetivos de la investigación es la construcción del DataMart, por ello en algunas fases de la metodología se explica cómo se va desarrollando el proceso para la construcción del Data Mart. A continuación, se listan cada una de las tareas y fases desarrolladas de la metodología CRISP-DM, alineadas a los objetivos.

Fase I. Comprensión del negocio

La empresa “Cell Service”, actualmente es una agencia de la empresa de telefonía “CLARO”, hasta fines del año 2016 fue una agencia de la empresa “MOVISTAR”. Desde fines del año 2012 la empresa empieza hacer uso de un sistema automatizado de ventas para el registro de sus ventas y también hace uso de MSExcel para registrar algunos datos adicionales de clientes. Desde entonces ha almacenado una cantidad importante de datos de ventas. La empresa requiere aumentar sus ventas debido a la alta competencia en el

mercado, para ello ha necesitado mejorar el proceso de análisis de sus ventas aplicando pronósticos en base a la información que tiene almacenada.

La forma como se realiza los pronósticos actualmente en la empresa, es con el uso de datos registrados en hojas de cálculo Excel, donde se tienen almacenados los registros de ventas de los últimos meses. Cada jefe de tienda y jefe de ventas de campo, entregan reportes de ventas diariamente para su control interno. Para pronosticar las ventas de equipos en la ciudad de Cajamarca y en las provincias, se preparan los reportes con semanas de anticipación, después hay una reunión cada 4 meses, tanto el dueño que es a su vez el gerente, con el jefe de control interno o administrador, el jefe de ventas de tienda y de campo donde se exponen los reportes y las estimaciones de ventas del mes. El pronóstico es complementado en su mayoría por opinión de los expertos, en este caso por el jefe de ventas, dado que estas personas están en contacto directo con los ejecutivos de ventas de cada sucursal y conocen muy bien el movimiento de clientes y equipos celulares en la zona, además de la opinión del gerente comercial. En todo el proceso, donde se utiliza un tiempo considerable es cuando se tiene que preparar la información, para ello se hace uso de hojas de cálculo y también se hace uso de algunos reportes del sistema de ventas.

Tabla 10: Recursos requeridos para el análisis de ventas y pronósticos

Preparar la información	Medios de uso	Tiempo requerido
Reporte de ventas de campo	Uso de hojas de cálculo MS Excel	2 semanas
Reporte de ventas de tienda	Uso de hojas de cálculo y el sistema transaccional	1 semana
Disgregar la información, según los criterios de análisis (Local de venta, marca, generación, plan, periodo, etc.)	Uso de hojas de cálculo MS Excel	1 semana
Juntar y cruzar la información con el periodo anterior	Uso de hojas de cálculo MS Excel	5 días aprox.

a. Determinación de los objetivos del negocio

- ♦ Mejorar el proceso de análisis datos de ventas aplicando pronósticos.
- ♦ Mejorar el proceso de análisis datos de clientes.
- ♦ Mejorar el proceso de Marketing.

b. Evaluación de la situación

Para el desarrollo del trabajo, se ha contado con el apoyo del siguiente equipo humano:

- El dueño de empresa CELLSERVICE EIRL.
- El gerente de la empresa
- El administrador
- Personal de Activaciones
- Personal de Almacén
- Personal de ventas

Los datos de ventas con la que cuenta la empresa desde el año 2002, se encuentran organizados de la siguiente manera:

- Los datos antes del año 2012, se encuentran registradas en papel pre impreso (boletas de venta) y también en hojas de cálculo Excel, pero solo con datos totalizados para fines contables.
- Los datos del año 2012 hacia adelante se encuentran almacenados en una base de datos que está alojada en un servidor de hosting privado en la web. La base de datos usa el sistema MYSQL y los datos son registrados por medio de un sistema automatizado para ventas (OLTP).

Los requerimientos que se han tenido en cuenta para el proyecto de minería:

- Delimitación del alcance del proyecto, que ha está de acuerdo al del trabajo de investigación.

- Disponer de la base de datos del sistema del sistema transaccional y los datos que están en otros formatos que sean necesarios para el proyecto.
- Definir los requerimientos para la construcción del Data Mart.
- Disponer de los datos del Data Mart, construido para el área de ventas.
- Contar con una herramienta de minería de datos, para construir el modelo de minería.

Los requerimientos para la construcción del Data Mart están en función del cumplimiento de los objetivos del negocio. En este caso, se han captado los procesos importantes para el análisis de ventas y las posibles dimensiones del Data Mart. A continuación se muestra una matriz de procesos y dimensiones, que es una herramienta de diseño clave que representa los procesos empresariales principales de la organización y la dimensionalidad asociada [*33*]. Esto ha servido para definir las dimensiones que del Data Mart.

Tabla 11: Matriz de procesos y dimensiones para diseño del Data Mart

	Dimensiones					
procesos	Producto	cliente	sucursal	plan	periodo	Cliente revendedor
Analizar de ventas histórico	X	x	x	x	X	
Analizar datos de clientes		x	x	x	X	x
analizar ventas por sucursal	X	x	x	x	x	X
Analizar importes y cantidades de ventas	X		x		x	x
Proyecciones de ventas	X	x	x	x	x	
Marketing para mejorar ventas	X	x	x		x	

En la tabla anterior (Tabla 10), el símbolo “x”, representa si un proceso está relacionado con alguna de las dimensiones posibles del Data Mart.

También, como parte de la metodología, el proyecto se basa en los siguientes supuestos: Se supone que los clientes mayoristas de la empresa no influyen en la evaluación de las ventas, desde el punto de vista del comportamiento del cliente, porque los clientes mayoristas ya no reportan datos de sus clientes a la empresa.

Las restricciones del proyecto:

- Los datos de ventas desde el año 2012 hacia atrás, están en papel pre impreso (boletas manuales de venta) y también en hojas de cálculo Excel con datos totalizados sólo para fines contables. Volver a registrar las ventas de manera detallada, para ser evaluadas, implicaría mucho tiempo y mano de obra.
- La ilegibilidad de muchos datos de los clientes, captados a la hora de registrar las ventas y las activaciones.
- Limitaciones en los registros de datos de clientes, por el sistema automatizado de ventas
- Muy poco conocimiento sobre Minería de datos, por parte de las áreas administrativas y gerencia.

Riesgos y planes de contingencia del proyecto:

- En el proceso de limpieza de datos, por desconocimiento, se podría eliminar atributos que son relevantes a las ventas. Para evitar esta situación, se tendrá que consultar al personal responsable del área y del proceso de minería.

c. Determinar los objetivos de la minería de datos

Los objetivos de la minería de datos están acordes con los objetivos generales y específicos del trabajo de investigación. El objetivo general es: “Evaluar el impacto que tiene el modelo de minería de datos en el pronóstico de Ventas, de la

empresa CELL SERVICE E.I.R.L., en el periodo del año 2012-2016". Los objetivos específicos son: Analizar la problemática de la empresa, Comprender la base de datos del sistema Automático Transaccional, Preparar de los datos que serán tratados, Construir la estructura del DATA MART, Implementar y Evaluar el modelo de Minería de Datos.

d. Definición del plan del proyecto

La planificación del proyecto de minería se ha hecho según al cronograma de la investigación. Dentro de la planificación está incluida también la construcción del Data Mart. El alcance de la construcción del Data Mart estará delimitado exclusivamente al área de ventas.

Fase II. Entendimiento de los datos

En esta fase de la metodología pues vamos a detallar actividades de: recolección de datos iniciales, descripción de los datos, exploración de datos y verificación de la calidad de los datos.

e. Recolección de datos iniciales

Los datos proporcionados por la empresa se encuentran almacenadas en la base datos del sistema transaccional y en también otros en hojas de cálculo Excel.

Los datos más fáciles de disponer para el análisis, son de los periodos comprendidos entre los años 2012 y 2016, porque los datos anteriores a éstos están registrados en boletas manuales.

Para poder ver la cantidad de registros de datos de ventas y de clientes, se ha tenido que ingresar a la base datos del sistema transaccional, así como también revisar los archivos relacionados en Excel.

Los registros de datos de ventas, están organizadas dentro de tablas en la base datos del sistema transaccional de ventas, por este motivo se ha tenido que

escoger las tablas que serán necesarias para el análisis y construcción del modelo. En la siguiente figura (Fig. 10) se muestra el gestor de base de datos mysql, en la que se encuentra alojadas las tablas del sistema de ventas.

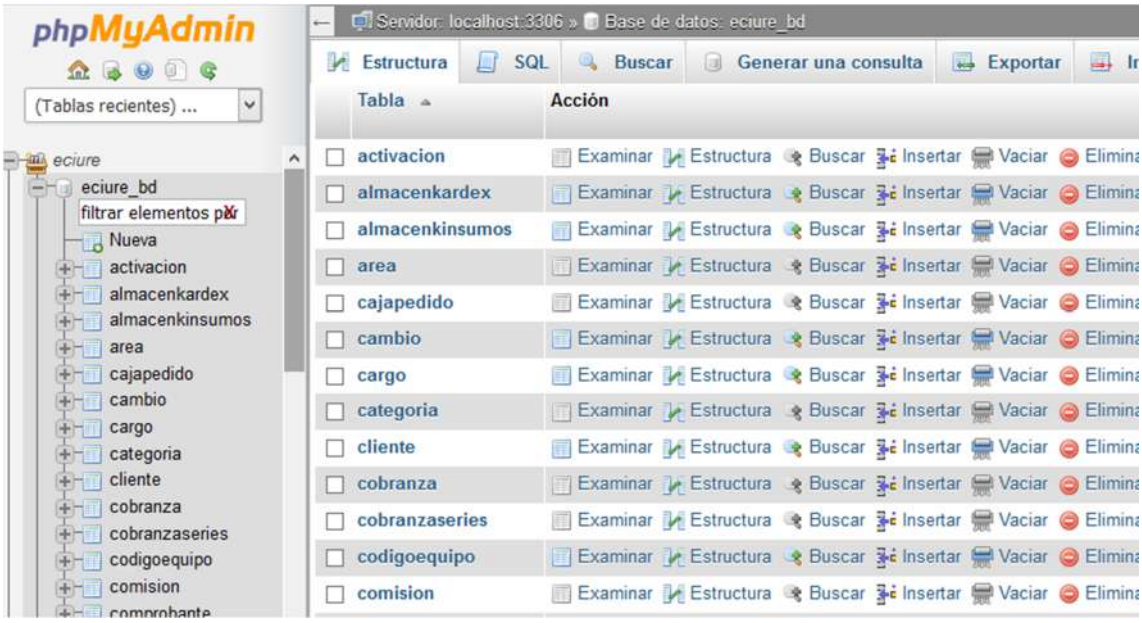


Fig. 10: Vista de la base datos del sistema transaccional de ventas

M22

	A	DNI	UBIGEO	APPAT	APMAT	NOMBRES	FECHA.NAC	GENE
1								
2	1	00010969	140103	SAAVEDRA	DEL AGUILA DE CACHAY	NILDA	1949-09-05	F
3	2	00042084	250107	CANANAHUAY	SILVA	JUAN MANUEL	1947-06-08	M
4	3	00043642	250101	LOPEZ	DE MUÑOZ	LUZ ESTELA	1958-05-28	F
5	4	00043904	250102	RODRIGUEZ	DE MACEDO	BLANCA	1964-03-25	F
6	5	00043985	250101				--	
7	6	00070378	060501	LEON	DE MIRANDA	NILDA ALBERTINA	1951-09-11	F
8	7	00103663	250101	CHUNG	RIOS	HUGO BENITO	1972-06-28	M
9	8	00113246	060101	DELGADO	SAUCEDO	HUGO ALBERTO	1970-10-04	M
10	9	00114776	250102	VARGAS	RAMIREZ	FLORINDA	1976-06-22	F
11	10	00160032	140116	CORO	CAMASI	EMILIO	1959-05-22	M
12	11	00186708	060201	SOLIS	MALPICA	PATRICIA DEL CARMEN	1975-04-15	F
13	12	00204577	060408	VILLENA	TEJADA	GLADIS ELIZABETH	1966-10-02	F
14	13	00247886	230105	RODRIGUEZ	CORNEJO	ANGEL IVAN	1974-03-22	M
15	14	00252649	060501	GALLARDO	BECERRA	JOSE MATIAS	1975-05-14	M
16	15	00322412	060701	LOZANO	VALLEJOS	AGUSTIN	1962-02-28	M
17	16	00327722	060501	PUELLES	ADRIANO	MARTIN	1974-01-08	M
18	17	00439598	110113	SALAMANCA	CHOQUE	ROBERTO	1957-08-03	M
19	18	00476676	940920	MATTA	DE LOS RIOS	FERNANDO TEODOMIRO	1961-10-31	M
20	19	00518252	220113	ALAVE	VARGAS	MARCELINO	1975-10-07	M
21	20	00799474	170301	GOMEZ	HUARACHE	NESTOR ADOLFO	1975-03-31	M
22	21	00827804	60501	LLATAS	MUÑOZ	SEGUNDO JUAN	1971-10-25	M
23	22	00829205	060101	VALQUI	SANCHEZ	SABINA	1974-10-27	F
24	23	00864678	060501	SILVA	RIMARACHIN	JOSE ANTONIO	1970-12-07	M
25	24	00890406	060701	QUEVEDO	SANCHEZ	ZOILA MERCEDES	1973-01-05	F
26	25	00952794	060501	DELGADO	CUBAS	JOSE HUMBERTO	1974-09-03	M

Fig. 11: Otros datos de clientes en formato de MS Excel

f. Descripción de los datos

Los datos se encuentran almacenados en un sistema de base datos MYSQL, alojado en un servidor UNIX - Percona Server (GPL) y administrado mediante phpMyAdmin versión 4.0.

La base de datos cuenta con 95 tablas. De acuerdo a los objetivos de la investigación, solo se tomará las tablas que intervendrán en el análisis de datos de ventas. Las tablas que se consideraron son: “activación”, “cliente”, “detalleventa”, “distrito”, “tienda”, “venta”, “mdproducto”. A continuación, se pasan a describir cada uno de los tipos de datos que corresponden a cada campo que componen cada tabla, las representaciones están tomadas de la base de datos del sistema PHPMyAdmin. Esta descripción servirá como base para crear el modelado dimensional del datamart del área de ventas.

	Campo	Tipo
☐	idactivacion	int(11)
☐	nomequipo	varchar(50)
☐	imei	varchar(30)
☐	iccid	varchar(30)
☐	ncel	varchar(10)
☐	fecha	date
☐	idtienda	int(11)
☐	iddistrito	int(11)
☐	idventa	int(11)
☐	idpersonal	int(11)
☐	plan	varchar(50)
☐	modalidad	varchar(8)

Fig. 12: tabla “activacion” de la base de datos relacional

Tabla: cliente

Examinar | Estructura | SQ

	Campo	Tipo
☐	idcliente	int(11)
☐	nombres	varchar(50)
☐	direccion	varchar(50)
☐	ruc	varchar(11)
☐	dni	varchar(8)
☐	telefono	varchar(10)
☐	referencia	varchar(200)
☐	fnac	datetime
☐	email	varchar(50)
☐	lugar	varchar(30)
☐	estado	varchar(10)
☐	tipo	varchar(14)

Fig. 13: Tabla "cliente" de la base de datos relacional

Tabla: detalleventa

Examinar | Estructura | SQ

	Campo	Tipo
☐	id	int(11)
☐	idventa	int(11)
☐	idproducto	int(11)
☐	precio	double
☐	cantidad	double
☐	uso	varchar(30)
☐	total	double
☐	descuento	double
☐	identificador	int(11)
☐	nombre	varchar(40)

Fig. 14: Tabla "detalleventa" de la base de datos relacional

Fig. 15: Tabla "distrito" de la base de datos relacional

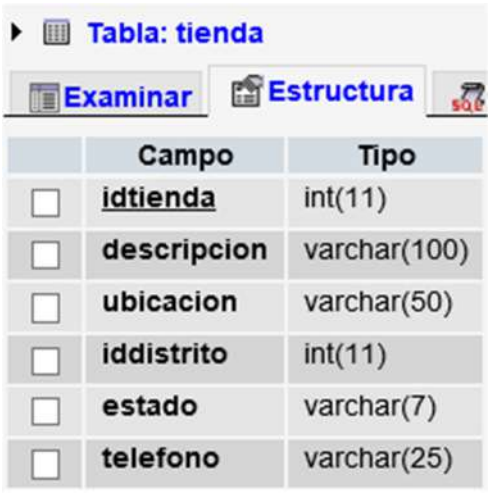


Fig. 16: Tabla tienda de la base de datos relacional

Fig. 17: Tabla "mdproducto" de la base de datos relacional

Fig. 18: Tabla "venta" de la base de datos relacional

Dado el caso que algunos datos los clientes como es: código del distrito, código de la provincia, código del departamento, edad, género, están en formato de MS Excel, se ha creado una nueva tabla "mdcliente", donde se importaran los datos de clientes.

Fig. 19: Tabla "mdcliente" de la base de datos relacional

Para obtener las descripciones de los códigos del distrito, provincia, departamento respectivamente, se ha creado una tabla “ubigeo”.

Fig. 20: Tabla “ubigeo” de la base de datos relacional

- **Descripción del modelado dimensional del DataMart de ventas**

Uno de los objetivos de la minería de datos es construir el dataMart, que servirá como fuente de datos del modelo de minería de datos.

Después de tener en claro los objetivos del negocio en el área de ventas y la matriz de requerimientos para la construcción del dataMart, se ha tenido que priorizar las posibles dimensiones según la importancia para la solución de la investigación. Se ha considerado omitir la posible dimensión “cliente revendedor”, dado que este tipo de cliente no refleja el comportamiento de un cliente final. Seguidamente, pasamos ver el modelo dimensional de alto nivel del datamart, donde H_ventas: representa a la tabla de hechos de ventas, D_cliente: representa la dimensión cliente, D_Producto: representa la dimensión producto, D_sucursal: representa la dimensión sucursal, D_Plan: representa la dimensión plan, D_Periodo: representa la dimensión periodo de la venta.

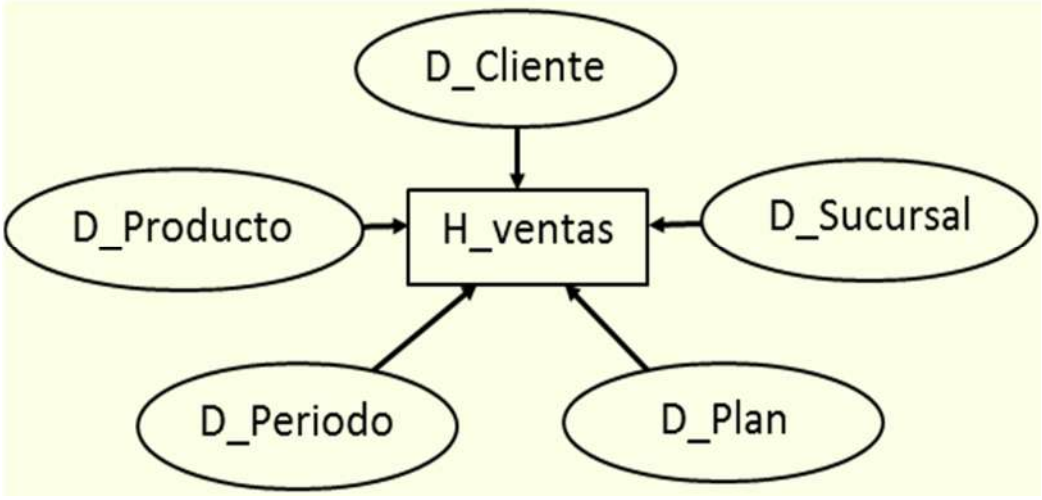


Fig. 21: Esquema del modelo dimensional del Data Mart

En esta parte, cada vez que mencionemos la palabra “base de datos transaccional”, no estaremos refiriendo a la “base de datos relacional” del sistema de ventas automatizado de la empresa.

A continuación, pasamos a describir los datos y las tablas, que se consideraron en la elaboración del Data Mart, o sea, la tabla de “Hechos” y las “dimensiones”.

1) Hechos del data Mart

Tabla 12: Tabla de hechos del datamart

Tabla:“hechos_ventas”		
Campo	**Tipo**	**Descripción**
dimcliente_id	Dato numérico, entero, identity	Clave foránea, que hace referencia a la dimensión “dimcliente”.
dimproducto_id	Dato numérico, entero, identity.	Clave foránea, que hace referencia a la dimensión “dimproducto”.
dimsucursal_id	Dato numérico, entero, identity	Clave foránea, que hace referencia a la dimensión “dimsucursal”.
dimperiodo_id	Dato numérico, entero,	Clave foránea, que hace

	identity	referencia a la dimensión “dimperiodo”.
Dimplan_id	Dato numérico, entero, identity	Clave foránea, que hace referencia a la dimensión “dimplan”.
cantidad	Dato numérico entero	Cantidad de equipos vendidos, por registro en el detalle de venta.
total	Dato numérico decimal	Subtotal de la venta, por registro en el detalle de venta.

1) Dimensión cliente

Tabla 13: Tabla dimensión cliente del Data Mart

Tabla:“dimcliente”		
Campo	**Tipo**	**Descripción**
dimcliente_id	Dato numérico, entero, identity	Clave primaria, de la tabla dimensión
idcliente	Dato numérico, entero	Identificador del registro del cliente en la tabla “cliente” de la base de datos relacional
nombres	varchar(50)	Nombre del cliente
genero	Char	Representa el género del cliente(masculino: M, femenino: F)
edad	Dato numérico, entero	Edad del cliente, valor numérico
distrito	Varchar(45)	Nombre del distrito del cliente
provincia	Varchar(45)	Nombre de la provincia del cliente
departamento	Varchar(45)	Nombre del departamento del cliente

2) Dimensión producto

Tabla 14: Tabla dimensión producto del dataMart

Tabla:"dimproducto"		
Campo	**Tipo**	**Descripción**
dimproducto_id	Dato numérico, entero, identity	Clave primaria, de la tabla dimensión "producto"
idproducto	Dato numérico, entero	Identificador del registro del producto en la tabla "producto" de la base de datos relacional
nombre	varchar(50)	Nombre del producto o equipo telefónico
generacion	varchar(3)	Describe la tecnología del equipo telefónico (2G, 3G, 4G)
tipo	varchar(7)	Tipo producto(chip, Celular)
marca	Varchar(50)	Marca del equipo telefónico (LG, SAMSUNG, ALCATEL,..)
modelo	Varchar(50)	Nombre del modelo del equipo que corresponde a una determinada marca
color	Varchar(45)	Nombre del color del equipo telefónico

3) Dimensión Sucursal

Tabla 15: Tabla dimensión sucursal del datamart

Tabla: "dimsucursal"		
Campo	**Tipo**	**Descripción**
dimsucursal_id	Dato numérico, entero, identity	Clave primaria, de la tabla dimensión "sucursal"
idtienda	Dato numérico, entero.	Identificador del registro de la tienda en la tabla "tienda" de la base de datos relacional
descripcion	varchar(45)	Nombre de la tienda o sucursal
provincia	varchar(3)	Nombre de la provincia donde se ubica la tienda

4) Dimensión Periodo

Tabla 16: Tabla dimensión periodo del datamart

Tabla:"dimperiodo"		
Campo	**Tipo**	**Descripción**
dimperiodo_id	Dato numérico, entero, identity	Clave primaria, de la tabla dimensión "periodo"
fecha	Fecha	Representa la fecha en la que se realizó la venta
año	Dato numérico, tipo entero	Año correspondiente a la venta.
mes	varchar(10)	Nombre del mes en que se realizó la venta
dia	varchar(10)	Nombre del dia en que se realizó la venta

5) Dimensión Plan

Tabla 17: Tabla dimensión plan del data mart

Tabla:"dimplan"		
Campo	**Tipo**	**Descripción**
dimplan_id	Dato numérico, entero, identity	Clave primaria, de la tabla dimensión "plan"
descripcion	varchar(45)	Nombre del tipo de plan, con lo que se vende el equipo (POSTPAGO, PREPAGO)

A continuación, se muestra en la siguiente figura Fig.22, una representación del diseño dimensional de la base de datos del dataMart de ventas.

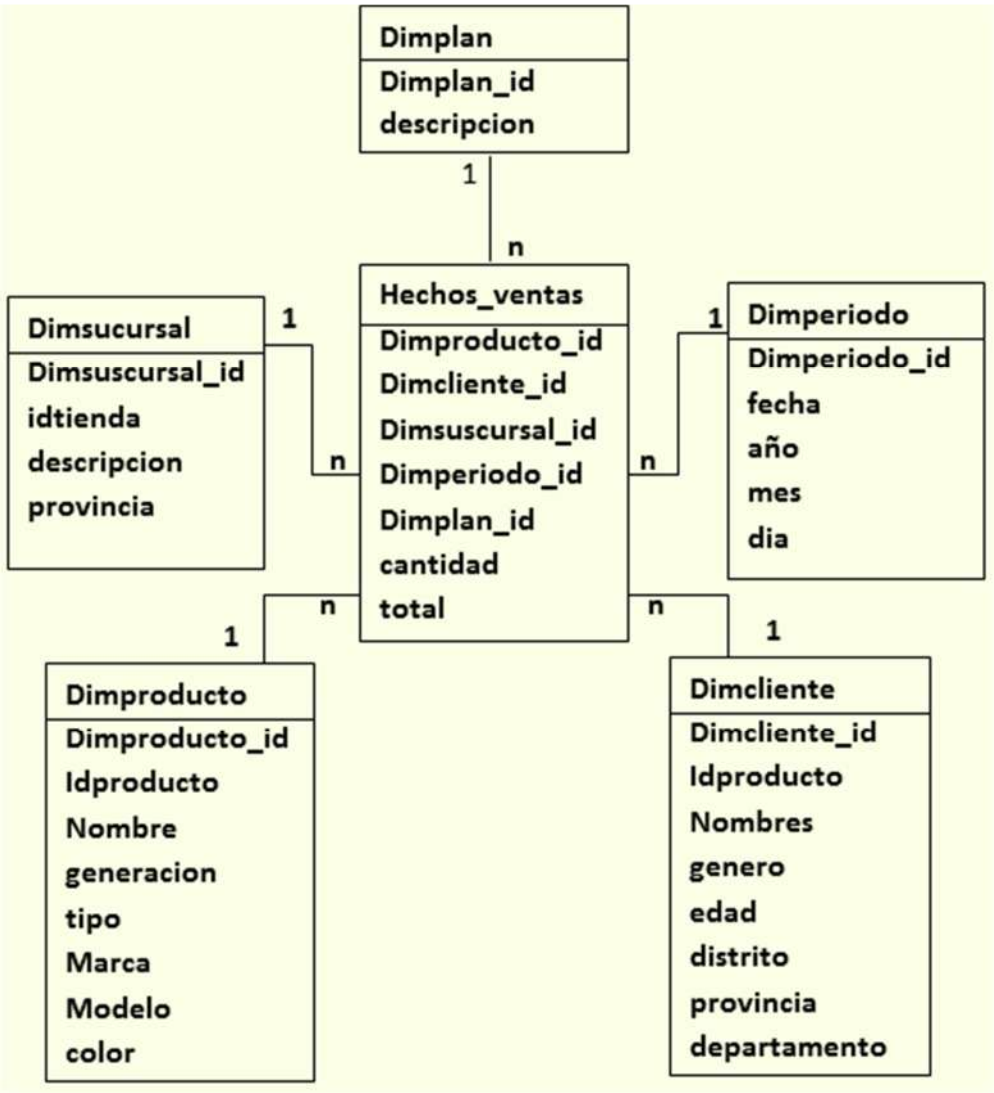


Fig. 22: Diseño dimensional del Data Mart de ventas

g. Exploración de los datos

Esta tarea, se ha realizado desde la base datos original del sistema transaccional, para ello se ha hecho uso de las tablas: “activación”, “cliente”, “detalleventa”, “distrito”, “venta”, “tienda”, “mdproducto”, “mdcliente”, mencionadas en la “tarea f: descripción de datos” de la metodología.

Para consultar los datos se ha instalado el sistema MYSQL en una computadora personal, por medio de ello se ha realizado consultas SQL y el uso de filtros y, para visualizar el comportamiento de los datos se ha hecho uso de histogramas y diagramas, de acuerdo al grado de importancia para llevar a cabo la minería de datos.

→ Ventas por edad del cliente: como se ve en la siguiente figura Fig.23, los clientes que han comprado más productos de teléfono celular, según la edad están entre los 23 y 25 años, además también los de edad entre 37 y 39, los clientes entre 46 y 48 años de edad.

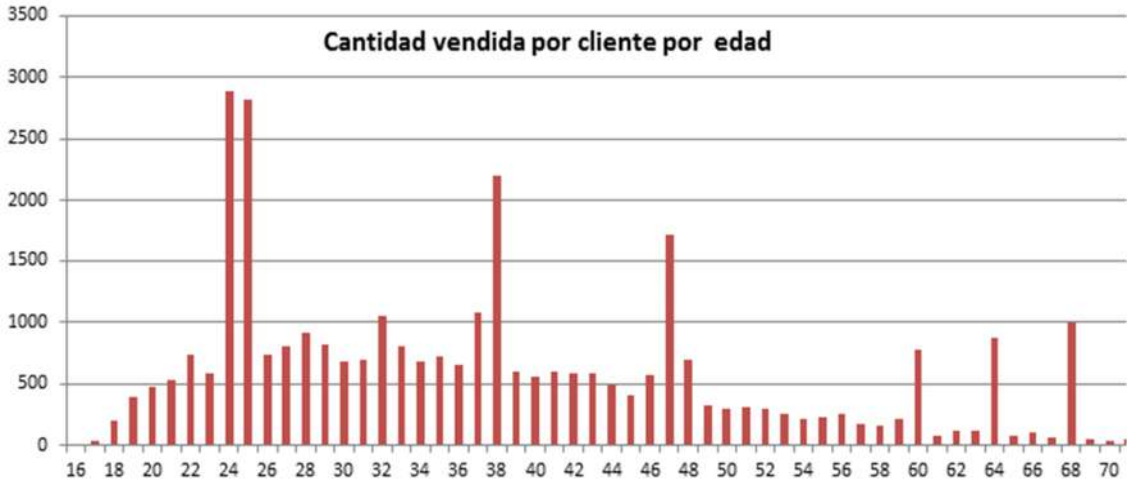


Fig. 23: cantidad de productos vendidos según la edad de los clientes

→ Ventas según género del cliente: se puede en la siguiente figura Fig.24, los clientes de género masculino (M) son los que compras más que los clientes de género femenino (F).

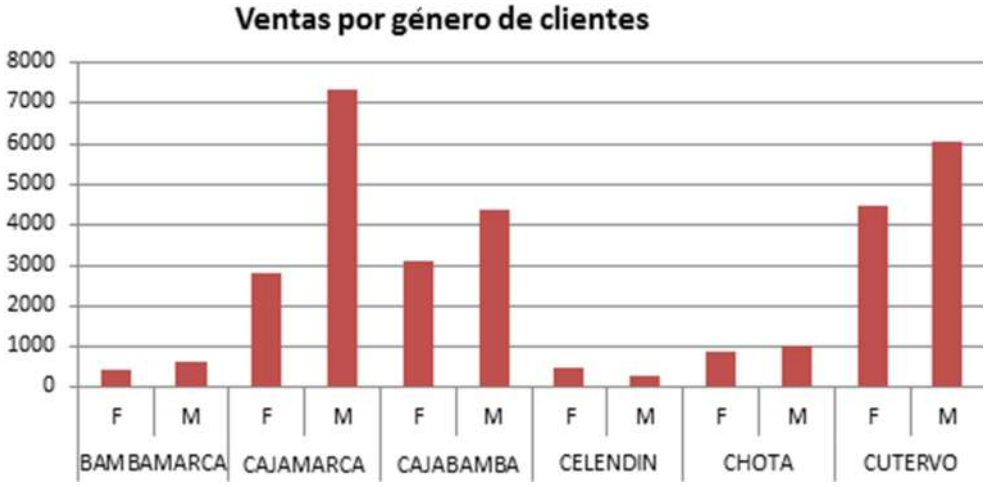


Fig. 24: cantidad de productos vendidos según el género de los clientes

→ En la figura Fig.25, se muestra las Ventas según el distrito del cliente: en la siguiente imagen se muestra las cantidades de equipos vendidos según el

distrito de procedencia del cliente, para este caso solo se consideró los distritos de la provincia de Cajamarca. Según el grafico se puede ver que los clientes cuyo distrito de procedencia es Cajamarca son los que más han comprado, debido a que los puntos de venta en la provincia de Cajamarca están ubicados en el mismo distrito de Cajamarca.

Fig. 25: cantidad de productos vendidos según distrito del cliente

→ Ventas por tipo de cliente: como se puede ver en la siguiente figura Fig.26, la empresa CELLSERVICE vende más a clientes finales.

Fig. 26: Porcentaje de ventas por tipo de cliente

→ Ventas de productos por marca: en la siguiente figura Fig.27, se puede observar el comportamiento de las ventas de productos de telefonía celular, según la marca. Es importante aclarar en este punto, que por razones de espacio solo se están mostrando algunas de las marcas registradas en la base de datos. Además, se puede observar que algunas marcas están duplicadas. No vamos a discutir sobre la calidad de los datos en este punto, pues, eso se verá en el paso siguiente de esta fase de la metodología (Identificación de la calidad de los datos).

Fig. 27: cantidad de productos vendidos según la marca

→ Ventas de productos por modelo y por marca:

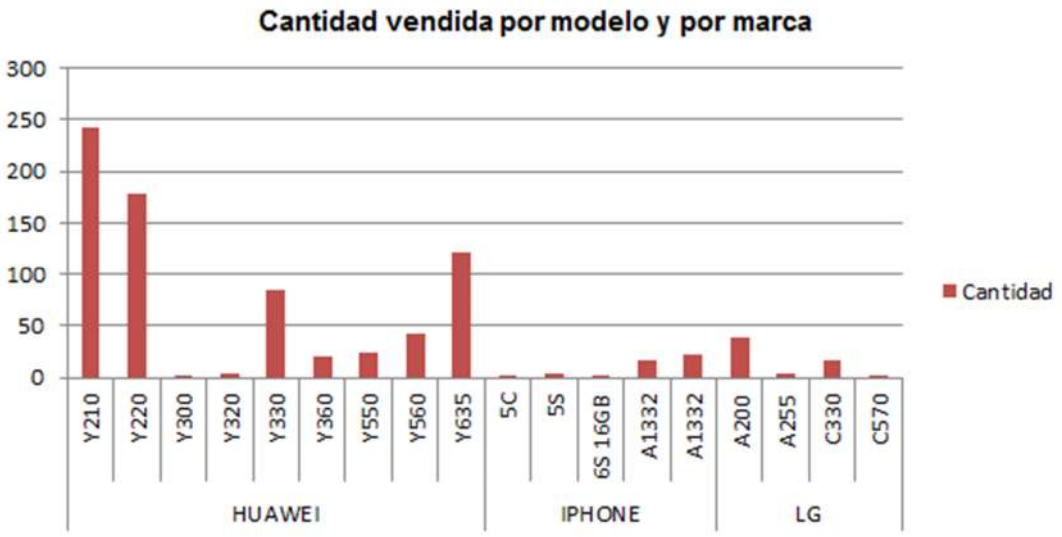


Fig. 28: cantidad de equipos celulares vendidos por modelo y marca

→ Venta de equipos celulares por color: igual que en el caso anterior, por razones de espacio solo se muestra algunos de los colores y sus respectivas cantidades vendidas.

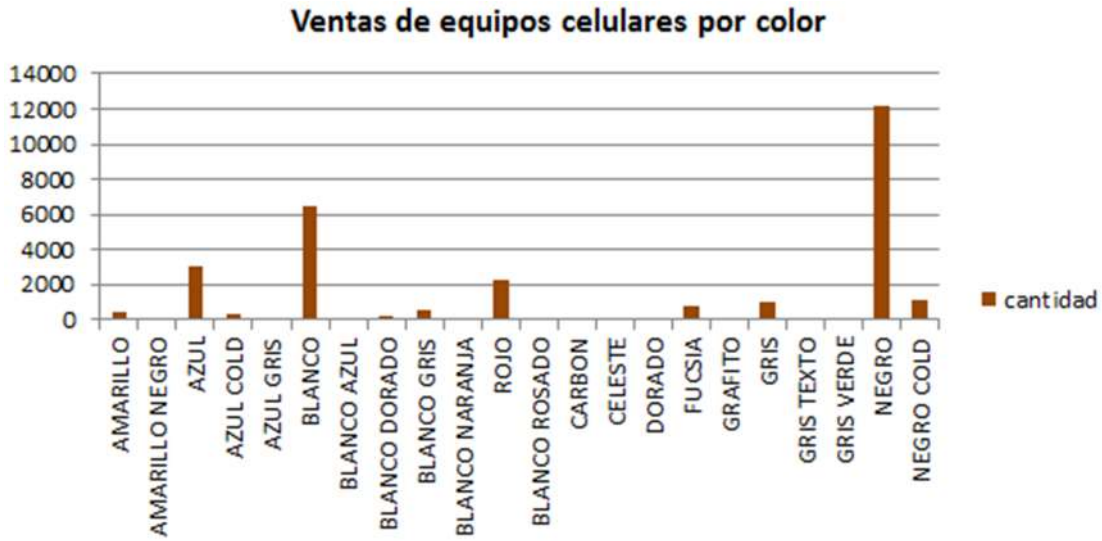


Fig. 29: cantidad de equipos celulares vendidos por color

→ Porcentaje de ventas por tipo de plan: en la siguiente figura Fig.30, se puede observar que los clientes prefieren pagar al contado la compra de su equipo celular.

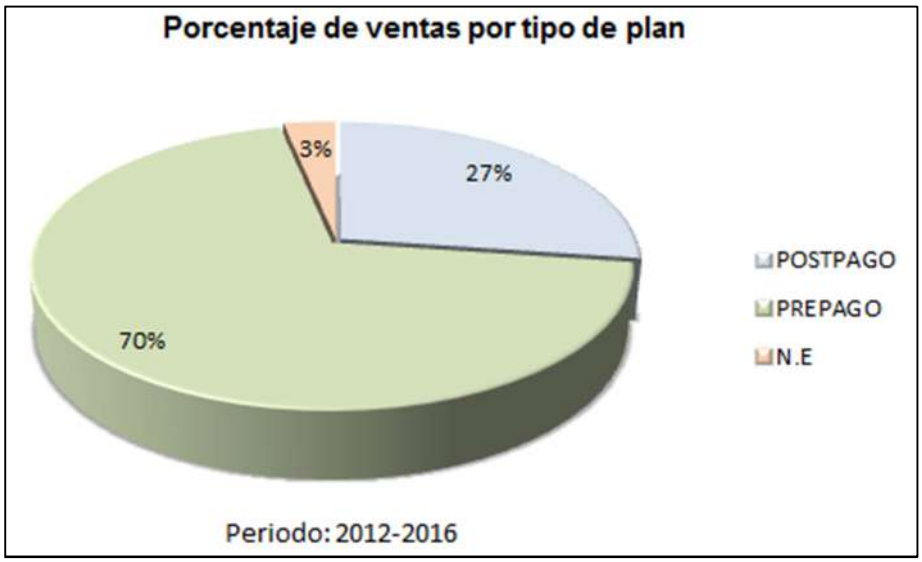


Fig. 30: porcentaje de equipos celulares vendidos por tipo de plan

h. Verificar la calidad de los datos

En esta tarea de la metodología, se ha visto como están registrados los datos en las fuentes de datos. Se encontrado que hay: datos repetidos, datos mal escritos, datos en blanco, datos con valor nulo. A continuación, se lista los casos encontrados con algunos reportes.

→ En la tabla “cliente” de la base de datos transaccional, solo se encuentra registrado los nombres, el DNI y el estado (“Activo”, “Inactivo”), los demás datos como el género, la fecha de nacimiento y los datos de ubigeo, están en archivo Excel.

→ En la tabla “cliente”, 196 registros de clientes están con nombres y DNI en blanco.

→ En la tabla “cliente”, existen registros de clientes repetidos o se refieren al mismo cliente, pero están mal escritos sus nombres

→ En la tabla “producto”, los datos de la marca del equipo celular, tiene registros repetidos o se refieren a lo mismo, pero están mal escritos.

→ En la tabla “producto”, los datos de la modelo del equipo celular, tiene registros repetidos o se refieren a lo mismo, pero están mal escritos.

→ En la tabla “producto”, los datos del color del equipo celular, tiene registros repetidos o se refieren a lo mismo, pero están mal escritos. También, se puede ver que existe datos en blanco.

→ En la tabla “venta”, 38 registros no presentan detalle de venta.

→ En la tabla “venta”, 814 registros se encuentran en anulados por errores en el ingreso de datos.

→ En la tabla “venta”, 58 registros tienen monto de S/.0.00 por errores en el ingreso de datos.

→ En la tabla “detalleventa”, hay registros duplicados de productos, por cada registro de venta en la tabla “venta”.

Fase III. Preparación de los datos

i. Seleccionar los datos

Para el presente trabajo, se ha construido la estructura de minería de datos tomando con fuente de datos el Data Mart elaborado para el área de ventas. Por tanto, el criterio de la selección de datos para la elaboración del modelo de minería, incluye los datos que se usaron para la elaboración del dataMart. A continuación, se detallan los campos de la estructura del modelo de minería.

Tabla 18: Datos seleccionados para la estructura del modelo de minería

Campo	Descripción	Tipo
“próximo”	Proximidad del cliente al punto de venta	Texto
“quincena”	Primera o segunda quincena del mes de la venta	
“cliente”	Edad del cliente (“muy joven”, “joven”, “adulto”)	Texto
“Tec”	Tecnología del equipo celular (“2G”, “3G”, “4G”,…)	Texto
“marca”	Marca del equipo celular (“SAMSUNG”, “NOKIA”,…)	Texto
“dia”	Día de la semana en que se realizó la venta	Texto
“genero”	Genero del cliente (Masculino “M”, Femenino “F”)	Texto
“color”	Color del equipo celular (“color claro”, “color oscuro”, “otro color”)	Texto
“modalidad”	Modalidad de venta (“PREPAGO”, “POSTPAGO”)	Texto

Los campos seleccionados para el modelo de minería, listados en la tabla anterior, se realizaron teniendo en cuenta los objetivos de la investigación. También se consideró los valores que toman cada uno de los campos y el grado de importancia que tendrían estos datos a la hora de probar el modelo.

j. Limpieza de datos

Tomando en cuenta punto, mencionado en el apartado h: “verificación de la calidad de datos”, los registros donde se encontraron inconsistencias fueron separados y no serán considerados en el análisis del modelo. Es importante aclarar, que los registros que presentan inconsistencias no han sido eliminados de la base de datos, solo que se han usado sentencias sql para filtrar los registros necesarios.

La limpieza de datos ha sido realizada antes del proceso ETL de carga de datos al Data Mart. Por lo que, el modelo de minería de datos ha tomado como datos de origen el Data Mart.

Para realizar el filtro de registros, se ha considerado los siguientes criterios:

- ⇨ Los registros de clientes que tienen el valor del DNI igual a nulo ("NULL”), no se consideraron, debido a que éste dato es importante para relacionar los demás datos del cliente que se encuentran en el formato MS Excel. Se procedió del mismo modo con los que tienen el valor en blanco.
- ⇨ Se procedió a contar la cantidad se nombres que registraba un cliente con el mismo DNI, de esta manera se filtraron todos los clientes duplicados
- ⇨ Se seleccionaron solo los registros de clientes, cuyo valor de la columna “tipo” de la tabla “cliente” sea diferente a “VENDEDOR”. Se consideró así, porque un cliente que es de tipo vendedor no refleja el comportamiento de un cliente final, debido a las compras que realiza es por volúmenes mayores.
- ⇨ Algunos registros de clientes, no tienen DNI y tampoco nombres, en total 196 registros fueron filtrados y sacados del análisis.

- ⇨ Los datos de clientes que están en formato MS Excel, que no tienen ubigeo, genero, edad, fueron eliminado del archivo (45 registros).
- ⇨ Los registros de ventas que tienen el valor de la columna “estado” como “ANULADO”, que representan un registro de venta no real, fueron excluidos.
- ⇨ Se ha relacionado las tablas “venta” y “detalleventa” para filtrar los registros de ventas que no tengan detalle.
- ⇨ Se han excluido del análisis los registros que tengan productos repetidos en el detalle de venta, en total 1029 registros.

k. Construcción de los datos

En esta parte de la metodología, se ha realizado algunas transformaciones y también se han calculado algunos atributos de acuerdo al objetivo de análisis y la estructura de minería definida.

En la tabla “cliente”, los campos de género, fecha de nacimiento, ubigeo (distrito, provincia, departamento), no están creadas. Debido a que la tabla “cliente” está en uso, y para no alterar su funcionamiento, se ha creído conveniente crear una nueva tabla en la base datos, con el nombre de “mdcliente”, donde se migraran los datos de año de nacimiento, ubigeo y género de los clientes, que están en formato MS Excel. Esta nueva tabla se relacionará con la tabla “cliente” por medio del atributo “dni”.

En la tabla “mdproducto”, el campo “tecnología”, ha sido creado como campo adicional. Para ello, para cada registro, se ha realizado una extracción de una subcadena que contenga la palabra “2G” o “3G” o “4G” del valor de la cadena del campo “nombre”. También, en la misma tabla “mdproducto” se han modificado los valores de los campos “marca”, “modelo”, “color” bajo los siguientes criterios: Los valores que representan el campo o columna “marca” tienen 142 valores diferentes, y muchos de ellos se refieren a la misma marca, solo que están escritos con unos caracteres adicionales o una descripción adicional. Para solucionar esto, se ha tomado como referencia una lista tipo de marcas de equipos celulares que maneja el área de ventas, y mediante una sentencia sql se

extrae el valor de la subcadena que contenga el nombre de la marca de la lista tipo, y si se encuentra el nombre de la marca en la subcadena se procede a reemplazar todo el valor del campo respectivo con el nombre de la marca de la lista tipo. De este modo los valores de la columna “marca” se ha reducido a 28 valores diferentes, que representan los nombres de las marcas exactas, lo que ha hecho más fácil el análisis de las ventas cuando se agrupan por marcas. Para la columna “color”, se ha procedido de la misma forma como se ha procedido columna “marca”, debido a que los valores de la columna color también tienen 77 valores distintos.

En la tabla “mdcliente”, el valor de la columna “edad” se ha calculado a partir de la columna “fnacimiento” y el valor de la columna “fecha” de la tabla “venta”. Para calcular el dato, se ha usado una setenciasql, donde se resta el valor de la fecha de venta “fecha” de la fecha de nacimiento del cliente “fnacimiento”.

Dado que uno de los objetivos de la investigación es construir el Data Mart, a partir del cual se elaborara el modelo de minería de datos, también se ha considerado algunos aspectos respecto trasformación de datos que se serán cargados al data Mart.

En la tabla “dimcliente” del data Mart, los campos “distrito”, “provincia”, “departamento” representan los nombres del distrito, la provincia y el departamento del cliente, respectivamente, estos datos son obtenidos al relacionar la tablas de la base datos relacional “mdcliente” y “ubigeo”.

En la tabla “dimperiodo” del Data Mart, lo campos “año”, “mes”, “dia” ha sido elaborados a partir del campo fecha de la tabla venta, haciendo uso de sentencias sql y las funciones de fecha.

I. Integración de los datos

Ha consistido en reunir los datos necesarios de las diferentes tablas de la base de datos transaccional para poblar la data Mart. Se usó herramientas de software tanto para crear el esquema del dataMart, poblar dataMart, diseñar el cubo OLAP

y explorar el cubo. Las herramientas usadas fueron MySQL, para diseñar el esquema del datamart, para poblar el data mart se usó Pentaho Data Integration, para diseñar el cubo se usó Schema Workbench, y para explorar el cubo se usó JRubik, todas estas herramientas son de código abierto “open source”.

- **Implementación del data mart:** Primeramente se ha implementado el data Mart en el mismo servidor Mysql. A continuación, se presenta el diagrama del data mart de ventas, para ello se ha hecho uso del programa Mysql Workbench.

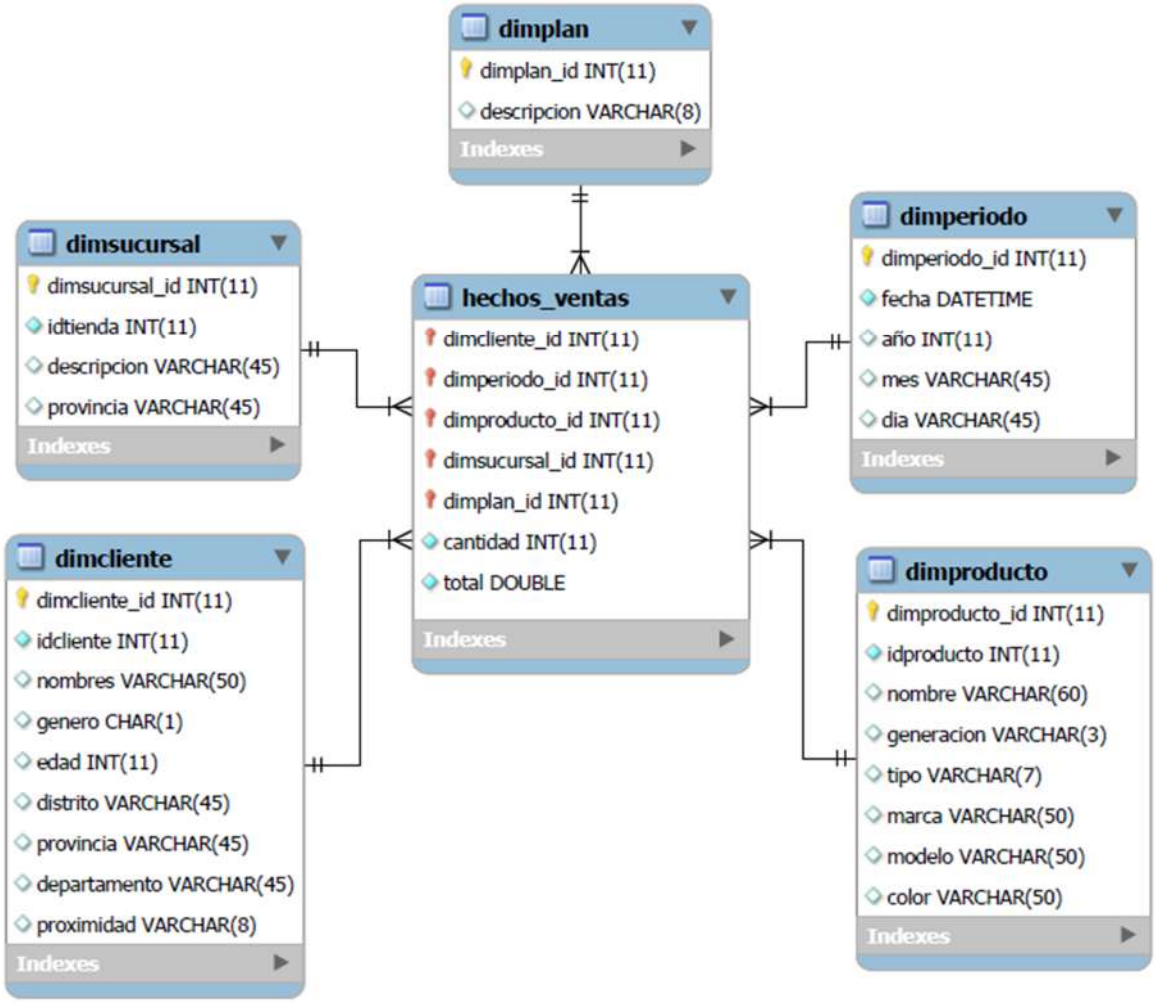


Fig. 31: Diagrama de base de datos del data mart de ventas

- **Proceso de integración con Pentaho Data Integration:** Seguidamente, se ha llevado a cabo el proceso de integración los datos de las diferentes tablas de la base de datos transaccional. Para ello se ha hecho uso de la herramienta de integración de datos “Pentaho Data Integration”.

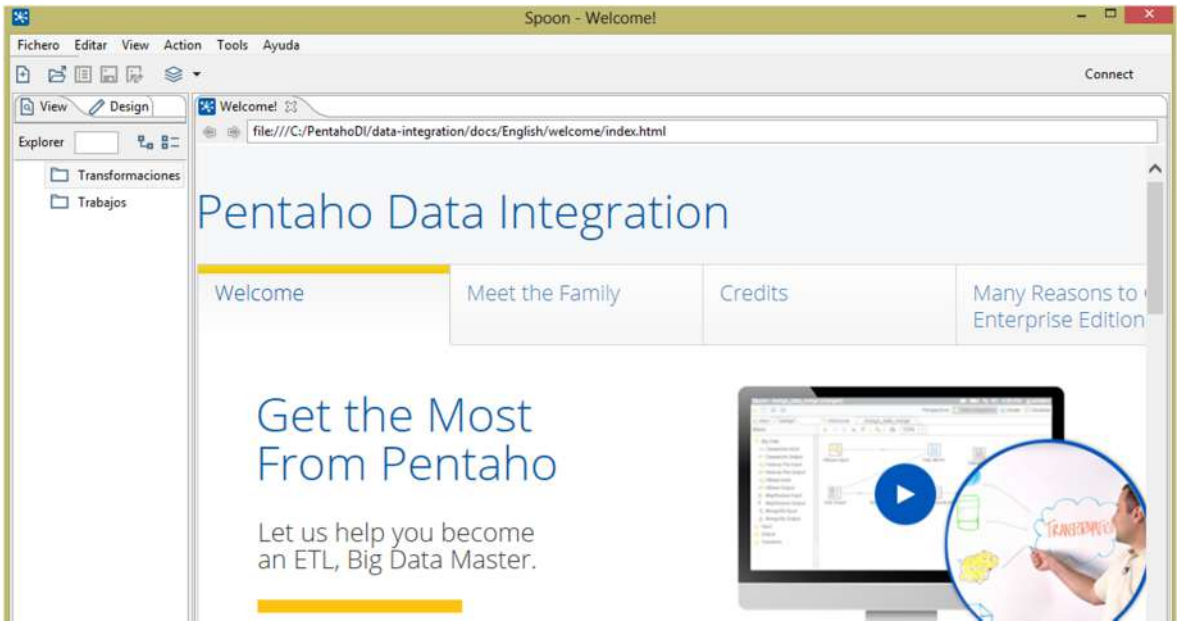


Fig. 32: Interface de desarrollo de la herramienta "Pentaho Data Integraction"

El integrador de datos de pentaho cuenta con muchos objetos que cumplen diferentes funcionalidades, que están agrupadas en 2 capas básicas: La capa de trabajo (en inglés: Job) y la capa de transformaciones. En esta etapa del proceso se ha hecho uso de ambas capas. Se ha hecho uso de la capa de trabajos, que permite ordenar las secuencias de las transformaciones y se diseñó las transformaciones que contienen la secuencia de extracción, transformación y carga de datos, conocido como el proceso ETL (en inglés: Extract, Transform and Load). Cada uno de estos pasos se realizó haciendo uso de la herramienta Pentaho, el cual es explicado detalladamente a continuación.

- **Herramientas de Pentaho usadas para la integración:** son las herramientas de diseño de pentaho usadas tanto para implementar el trabajo y las transformaciones. En la capa de transformaciones se usó como entrada de datos una tabla "Entrada Tabla" y como destino de los datos a otra tabla "Salida Tabla".

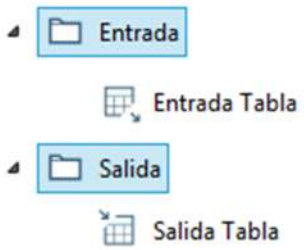


Fig. 33: Herramientas de pentaho usadas para las transformaciones

En la capa de trabajo o "job" se usó las herramientas de inicio "Start", transformación "Transformation", y las utilidades de mensajes y salida de procesos.

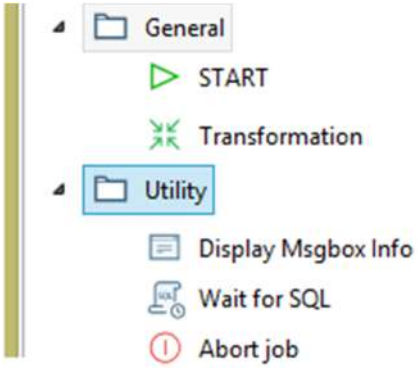


Fig. 34: Herramientas de Pentaho usadas para los trabajos o Jobs

- **Extracción, transformación y carga de datos:** Para realizar estos pasos ha sido necesario haber realizado correctamente cada una de las fases y tareas previas a ésta fase, de la metodología usada. Este paso a consistido en extraer mediante consultas SQL, cada una de los campos y registros que poblaran los hechos y las dimensiones del data mart, los mismos que fueron usadas para crear la estructura de la minería de datos. Como parte del proceso ETL, se usó la capa de transformaciones de pentaho. Para especificar la secuencia, se creó para cada dimensión y los hechos del data mart, una transformación para cada una, respectivamente. A continuación, se detallan las transformaciones realizadas.

  Para poblar la dimensión de los clientes, se usó como datos de entrada una tabla, donde se insertó una sentencia SQL para extraer los datos que conformaran la dimensión de los clientes.

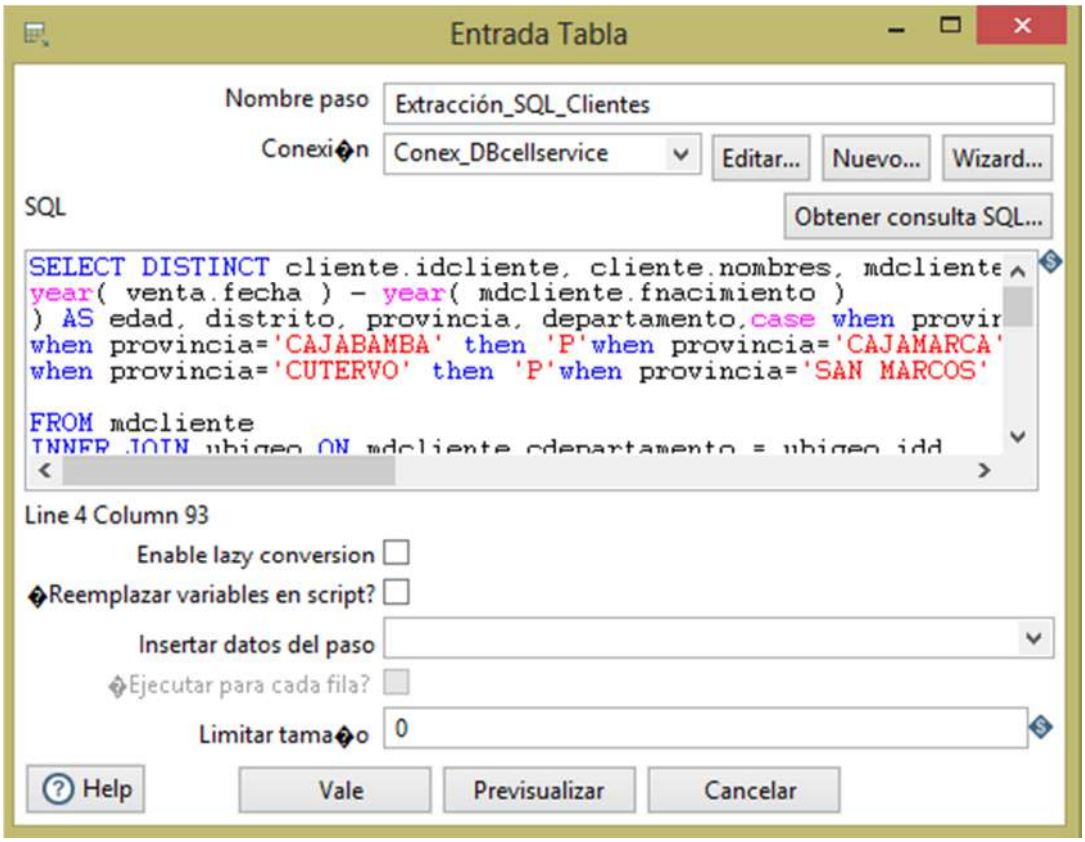


Fig. 35: Configuración- extracción de datos para poblar la dimensión de los clientes

Para cargar los datos provenientes en la dimensión de clientes, se utilizó la herramienta de tabla de salida “Salida Tabla”.

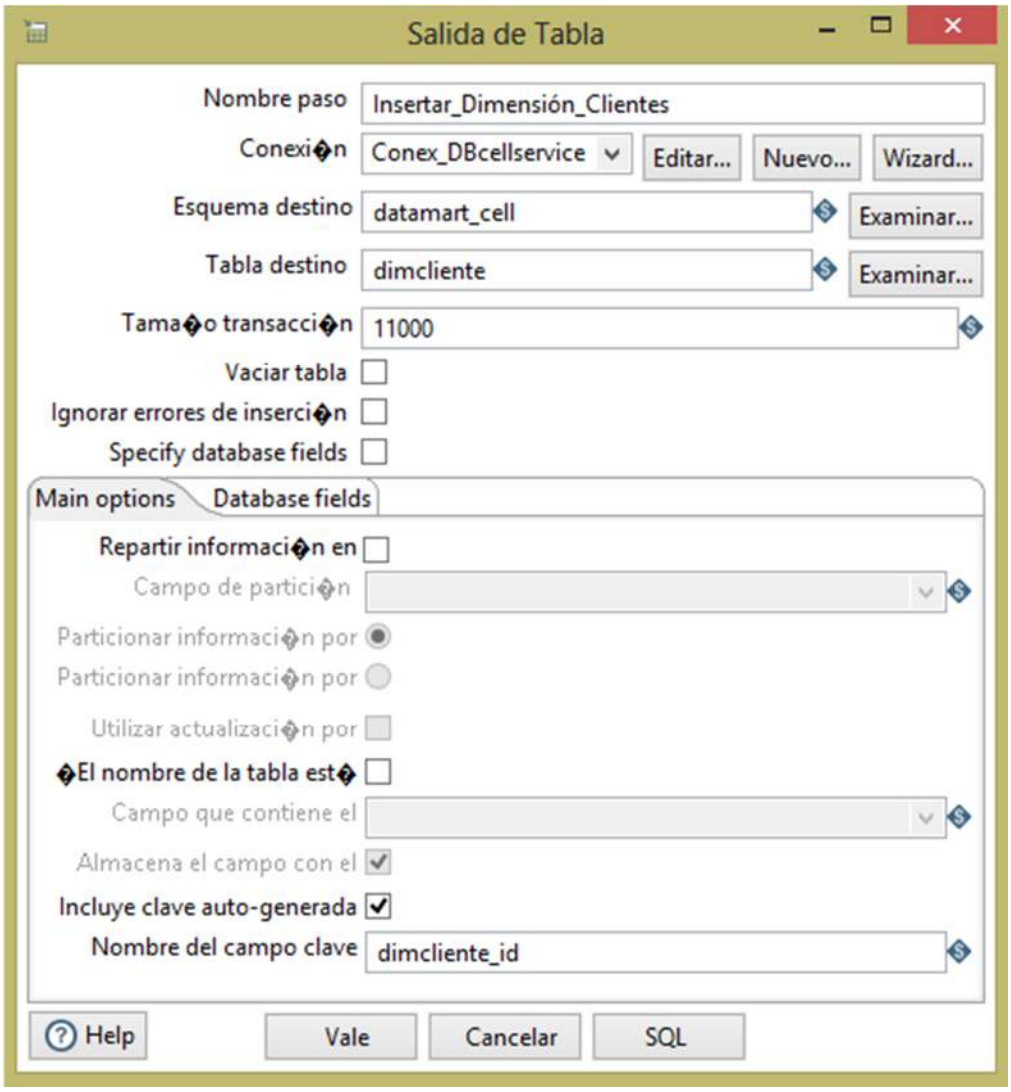


Fig. 36: Configuración - Insertar datos en la dimensión clientes

Fig. 37: Diagrama- ETL para la dimensión de los clientes

De manera análoga al caso anterior, se procedió para poblar la dimensión de los productos del data mart.

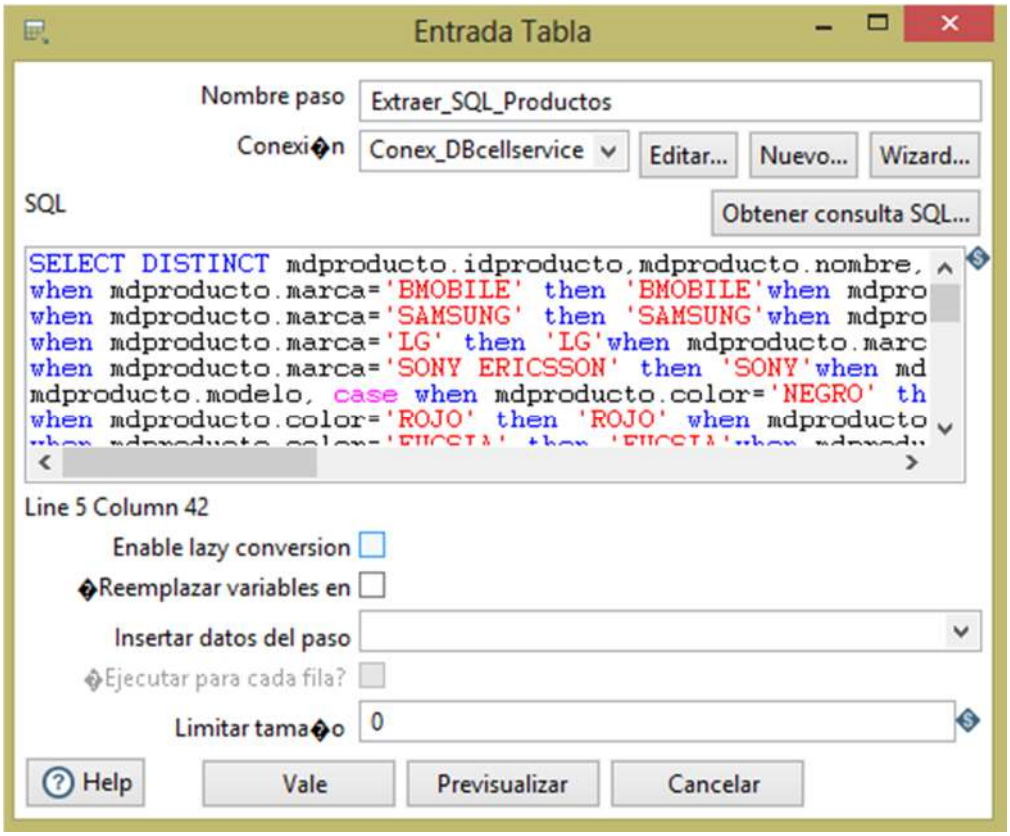


Fig. 38: Configuración - extracción de datos para poblar la dimensión de los clientes.

Para cargar los datos provenientes, en la dimensión de productos, se utilizó la herramienta de tabla de salida “Salida Tabla”.

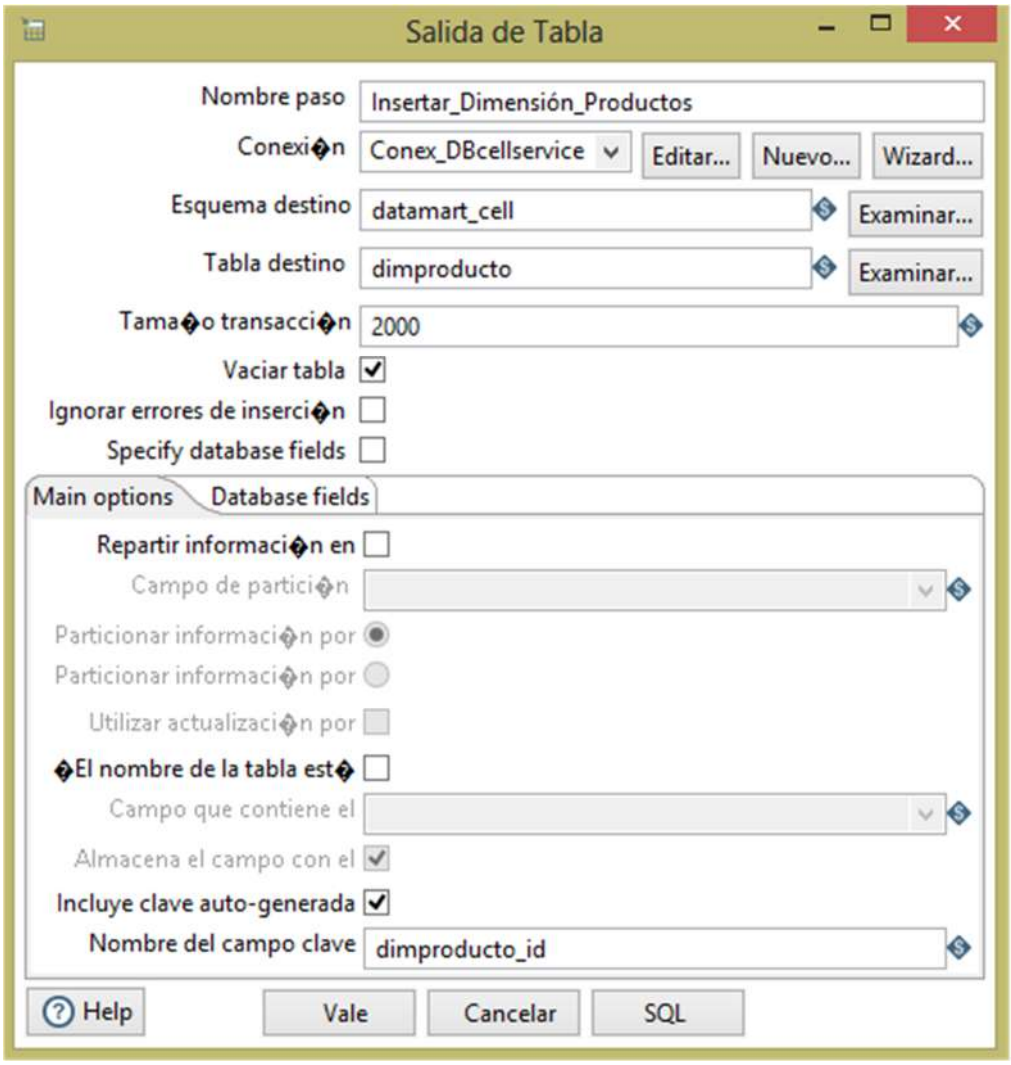


Fig. 39: Configuración - Insertar datos en la dimensión de los clientes.

Fig. 40: Diagrama- ETL para la dimensión de los productos.

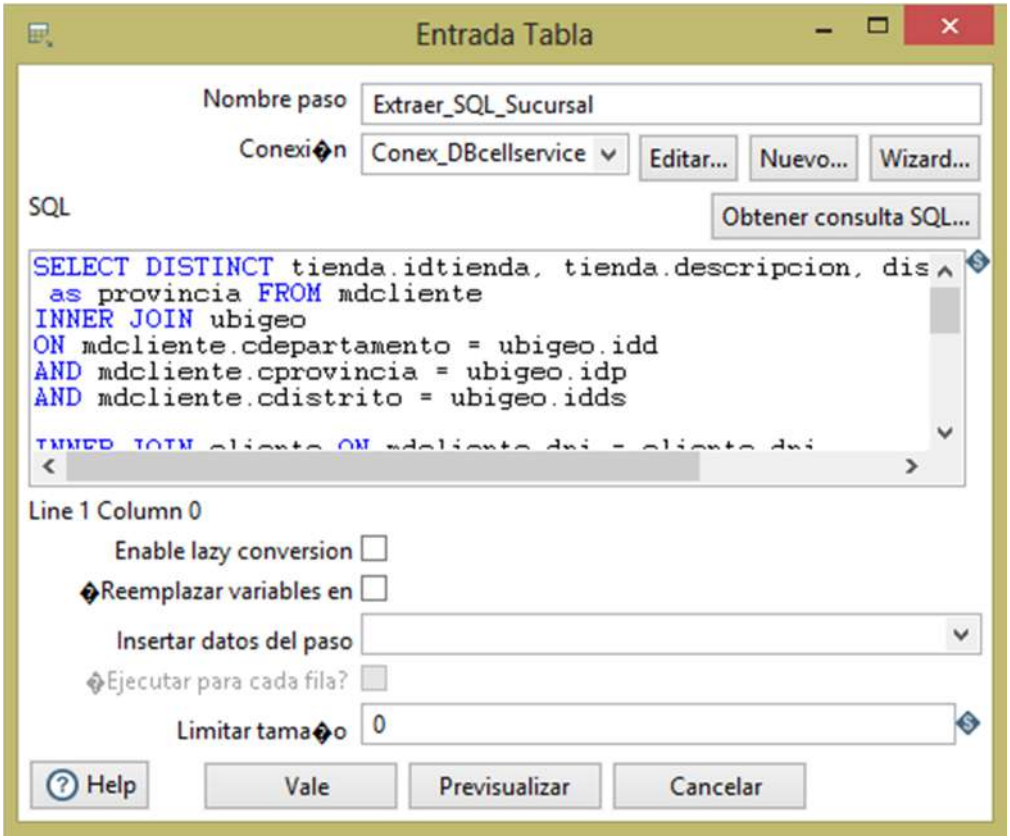


Fig. 41: Configuración - extracción de datos para poblar la dimensión de las sucursales.

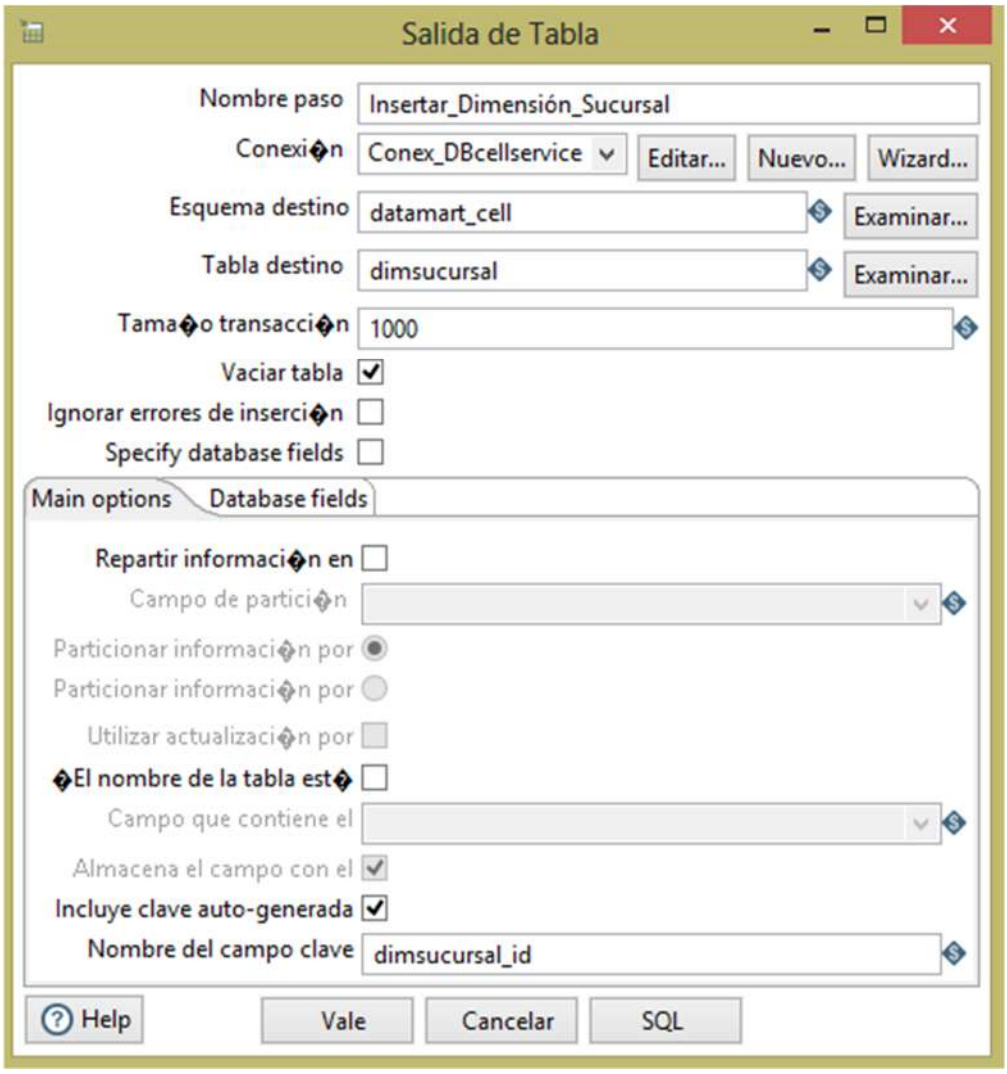


Fig. 42: Configuración - Insertar datos en la dimensión de los clientes.

Fig. 43: Diagrama- ETL para la dimensión de las sucursales.

Fig. 44: Configuración - extracción de datos para poblar la dimensión periodo.

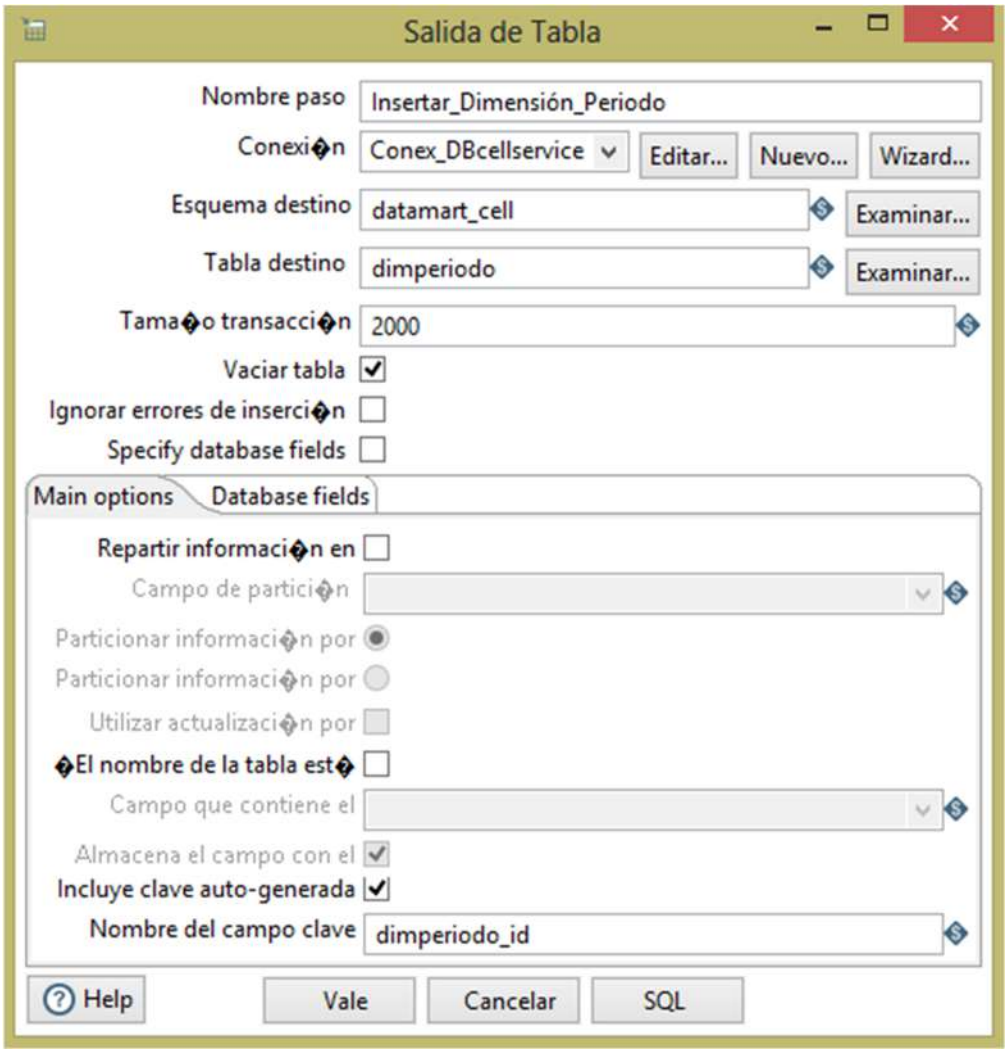


Fig. 45: Configuración - Insertar datos en la dimensión de los clientes

Fig. 46: Diagrama- ETL para la dimensión periodo

De la misma manera, como se ha procedido para las dimensiones se realizó para poblar los hechos del data mart. En la consulta SQL se hace uso de las tablas normalizadas y además de las dimensiones pobladas.

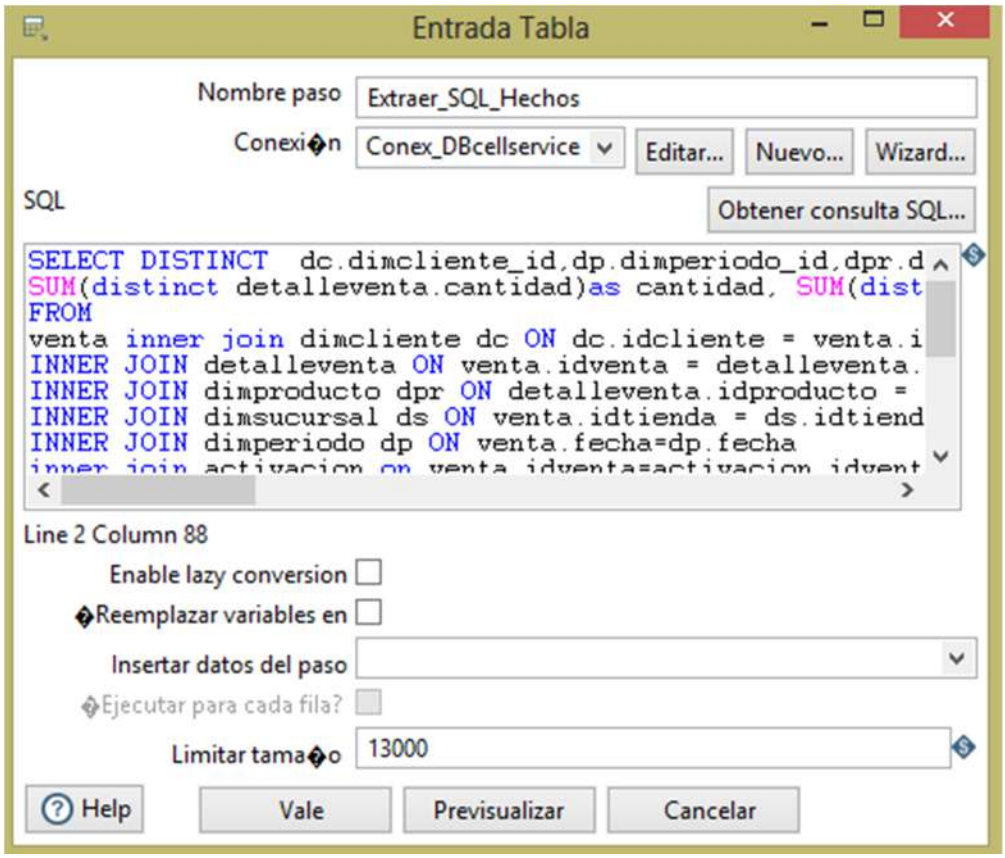


Fig. 47: Configuración - extracción de datos para poblar los hechos.

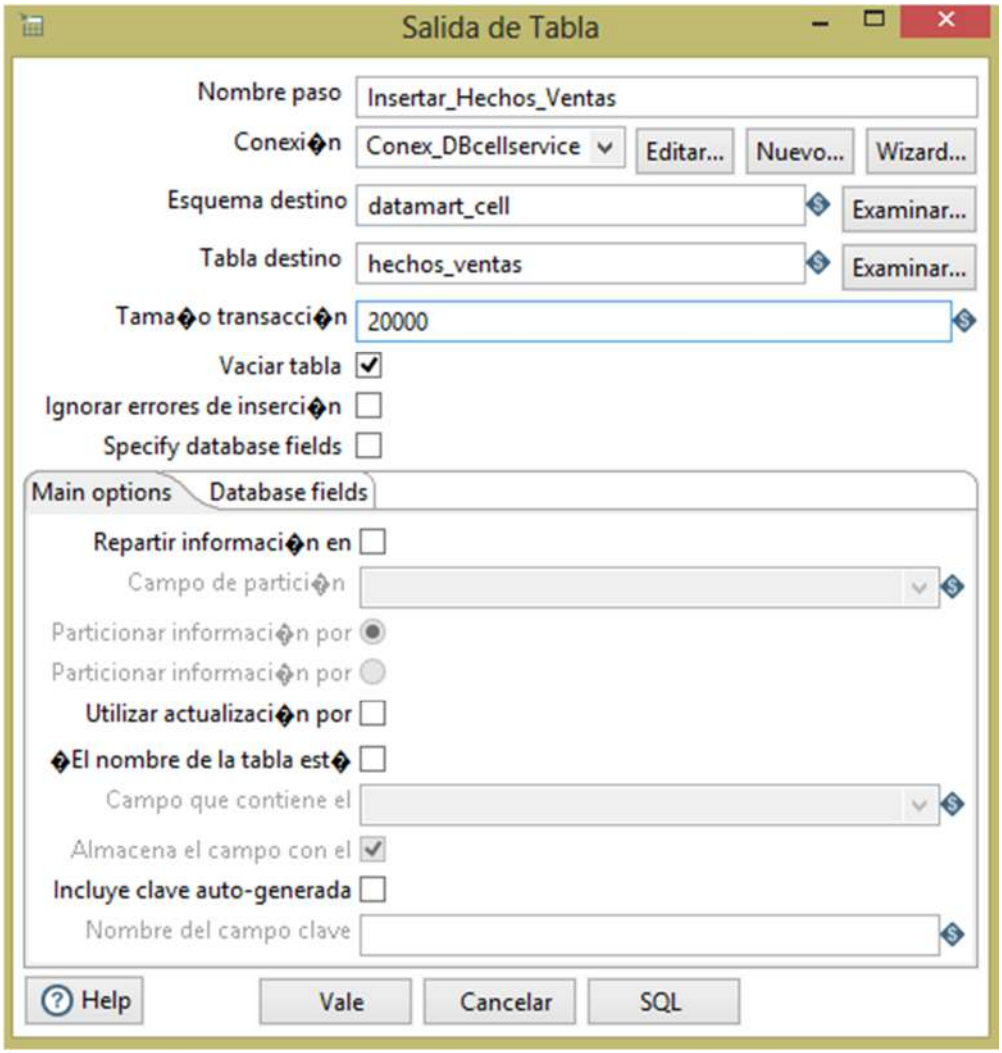


Fig. 48: Configuración - Insertar datos en la tabla de hechos

Fig. 49: Diagrama- ETL para la dimensión periodo

Finalmente, antes de proceder a ejecutar los ETLs de las dimensiones y hechos se ha creado una rutina que cosiste en limpiar las dimensiones y la tabla hechos para no duplicar los registros en la fase de prueba.

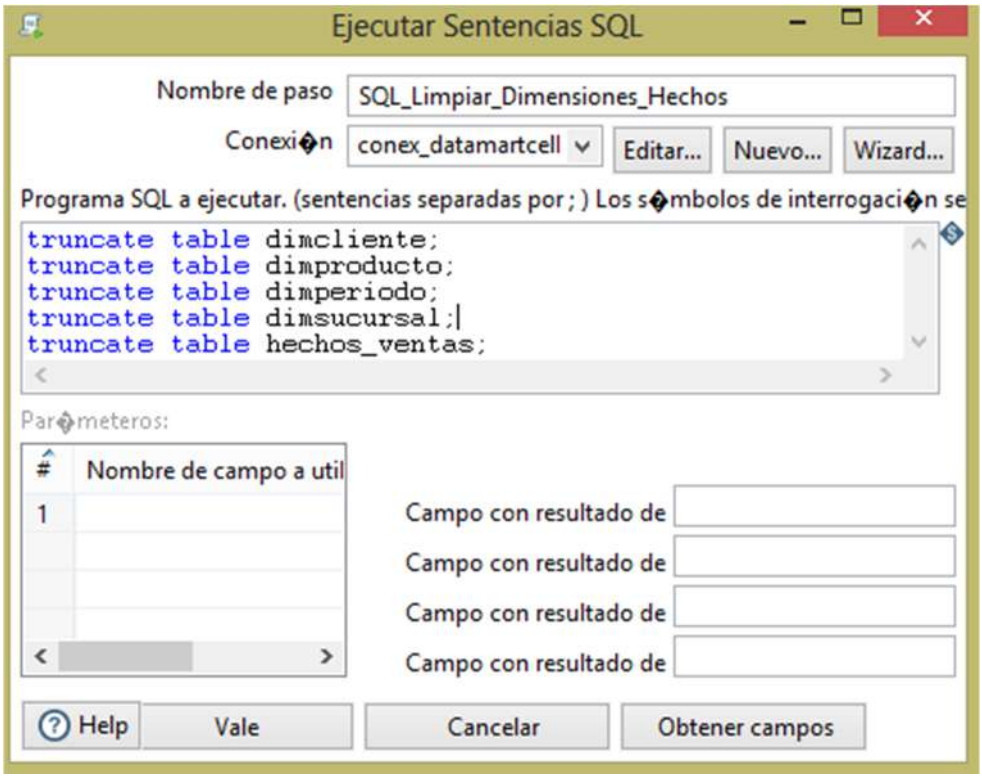


Fig. 50: Configuración - Limpiar dimensiones y hechos.

Fig. 51: Herramienta Script SQL para limpiar dimensiones y hechos.

- **Secuencia de las trasformaciones**: para establecer la secuencia en que se realizaran las transformaciones, se ha creado un trabajo o "Job", en la cual se integran todas las transformaciones que contienen los ETLs para cada una de las dimensiones y hechos. Esto ha sido posible haciendo uso de las herramientas que ofrece pentaho en la capa de trabajos.

Fig. 52: Representación de un paquete de transformaciones en pentaho.

Los paquetes de transformaciones de pentaho han permitido agrupar los ETLs y establecer las secuencias de las mismas. A cada paquete se le

asignó una transformación o ETL para cada dimensión y la tabla hechos, descritos en los pasos anteriores.

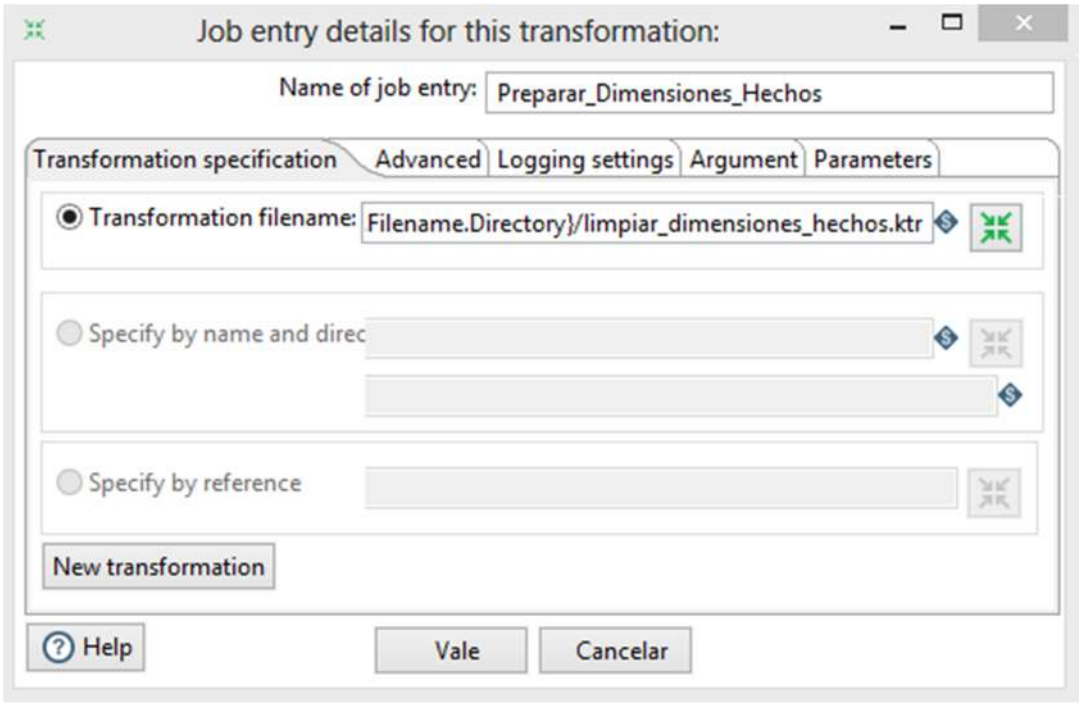


Fig. 53: Configuración de los paquetes "job" de pentaho

A continuación, se muestra el diagrama (figura 54) de ETL general usado para poblar la data Mart.

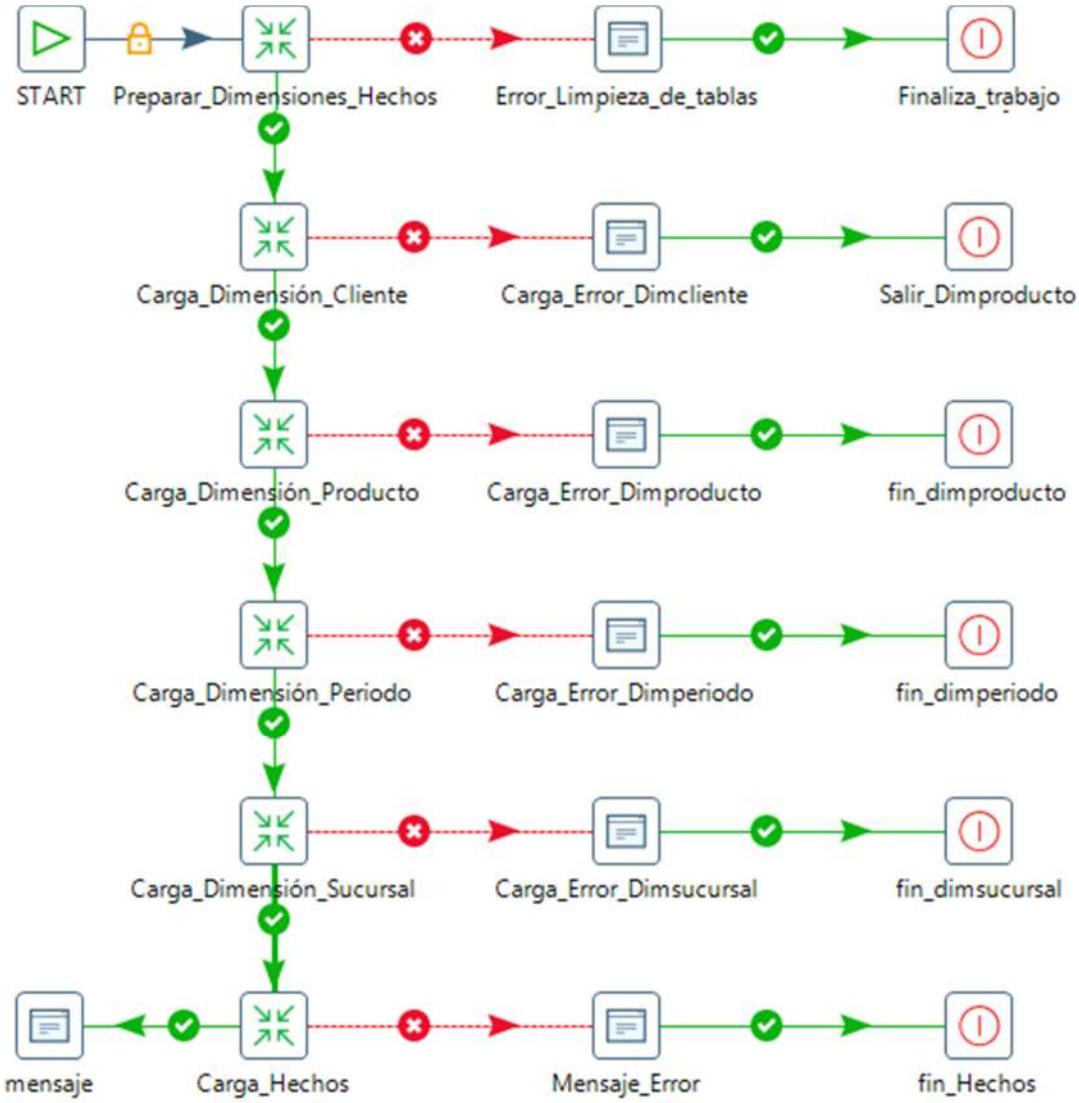


Fig. 54: Diagrama ETL general para poblar el data mart.

- **Diseño del cubo OLAP**

El diseño del cubo OLAP ha permitido visualizar los datos integrados de ventas, productos, clientes, sucursales, además, ha permitido visualizar el comportamiento de los atributos usados en el modelo de minería de datos. Para diseñar el cubo se ha hecho uso de la herramienta Pentaho Schema Workbench.

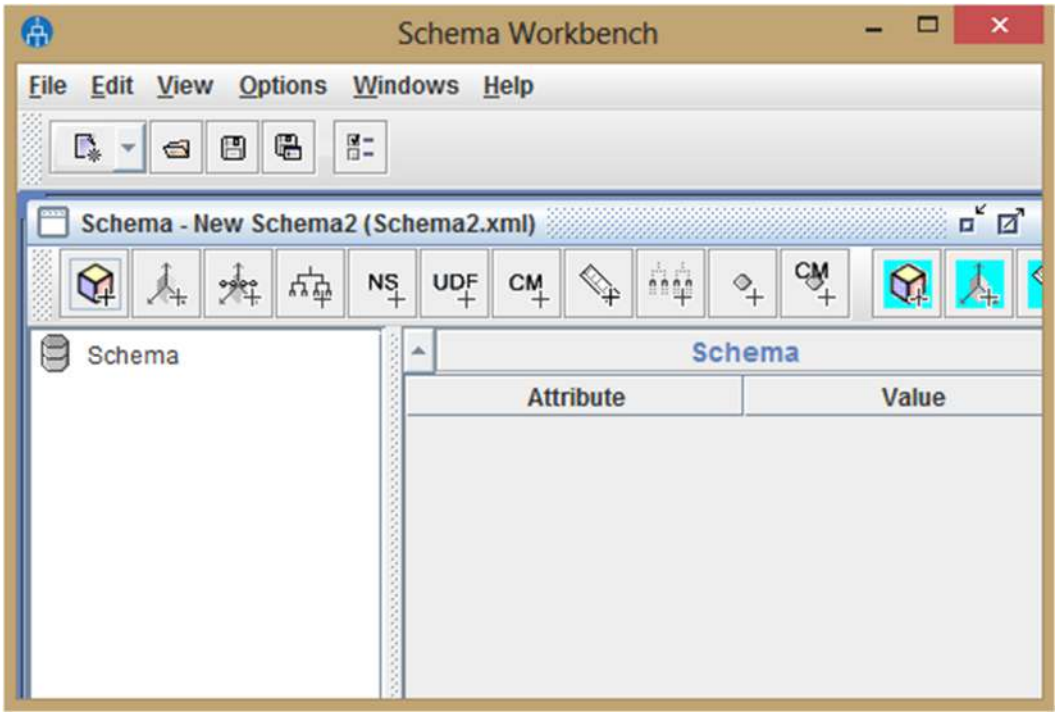


Fig. 55: Espacio de trabajo de SchemaWorkbench

Primeramente, se ha realizado una conexión con la base de datos dimensional “dataMart_cell”, que ha servido como datos de origen para diseñar el cubo. Seguidamente se ha creado las dimensiones generales que conformaran el cubo. A continuación, a manera de ejemplo, se muestra como una figura, donde se puede ver como se diseñó una dimensión, en este caso para los productos.

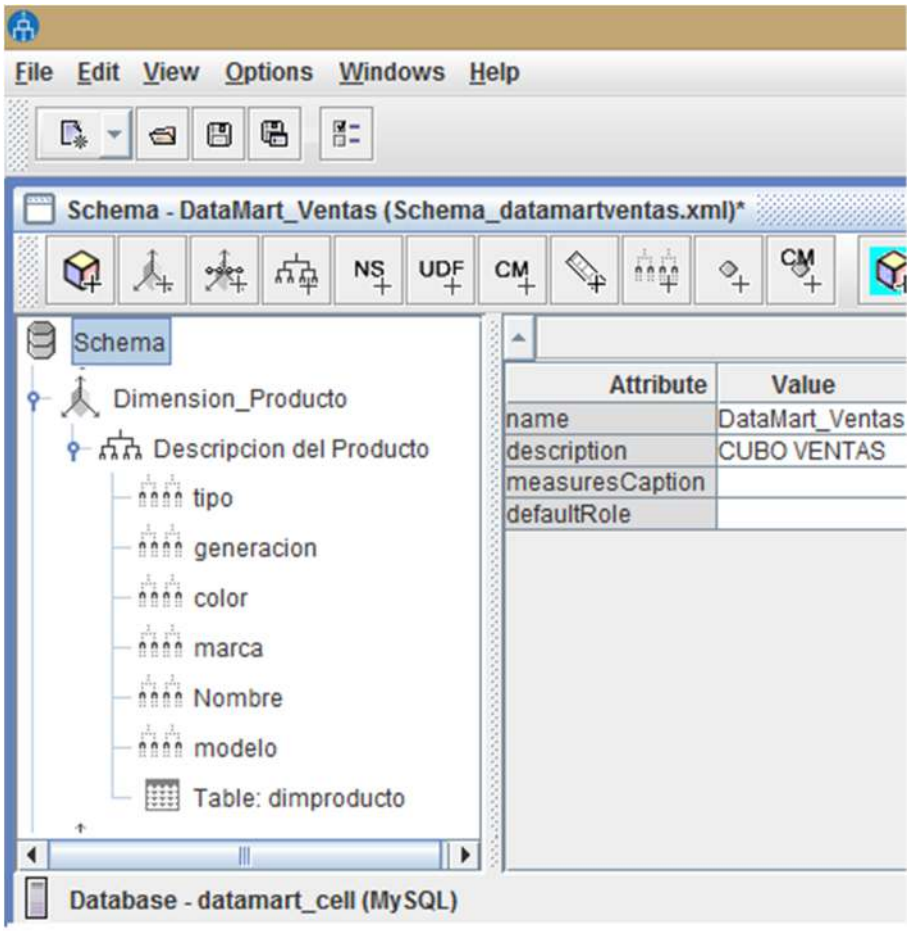


Fig. 56: Diseño de las dimensiones del cubo en Schema Workbench

Después de diseñar las dimensiones, se diseñó el cubo, con sus respectivas métricas y las dimensiones de uso, como se puede ver en la siguiente figura.

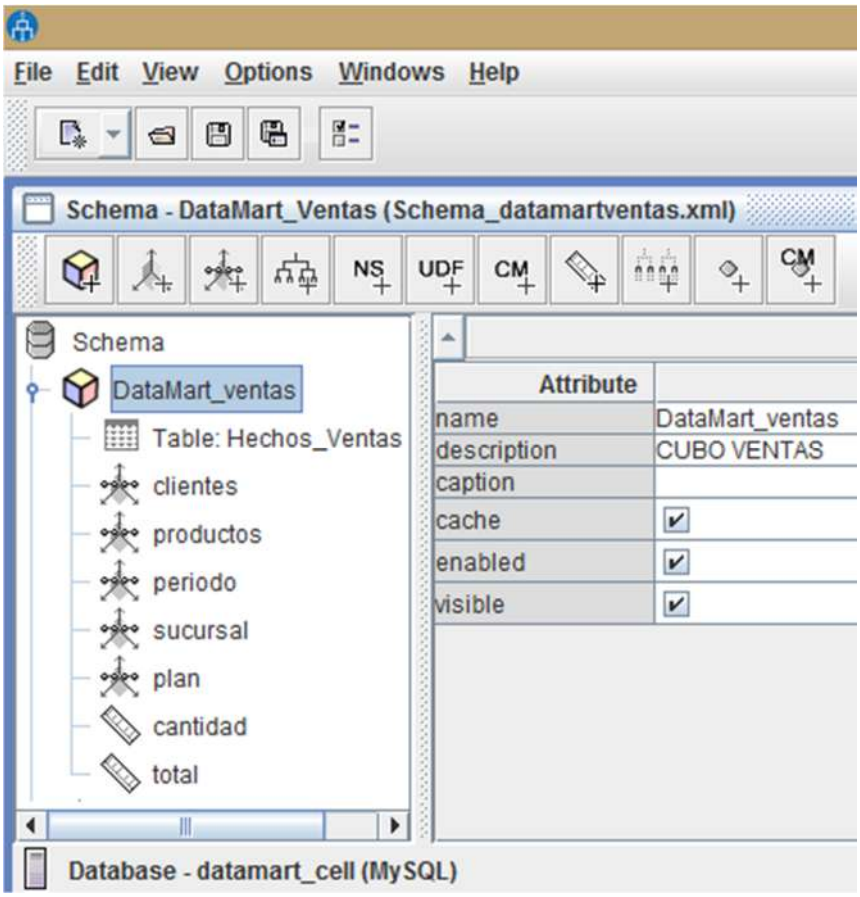


Fig. 57: Diseño del cubo en SchemaWorkbench.

Finalmente, se generó el archivo XML, que ha sido utilizado para realizar el análisis del contenido del cubo. En la siguiente figura se muestra como se ha podido publicar el cubo, en este caso se guardó en un repositorio, pero también se podría haber publicado en un servidor de pentaho.

Fig. 58: Publicación el archivo XML del cubo.

- **Exploración del Cubo OLAP**

Para analizar la información del cubo se implementó la herramienta "JRubik", que ha permitido ver la información a manera de reportes y gráficos.

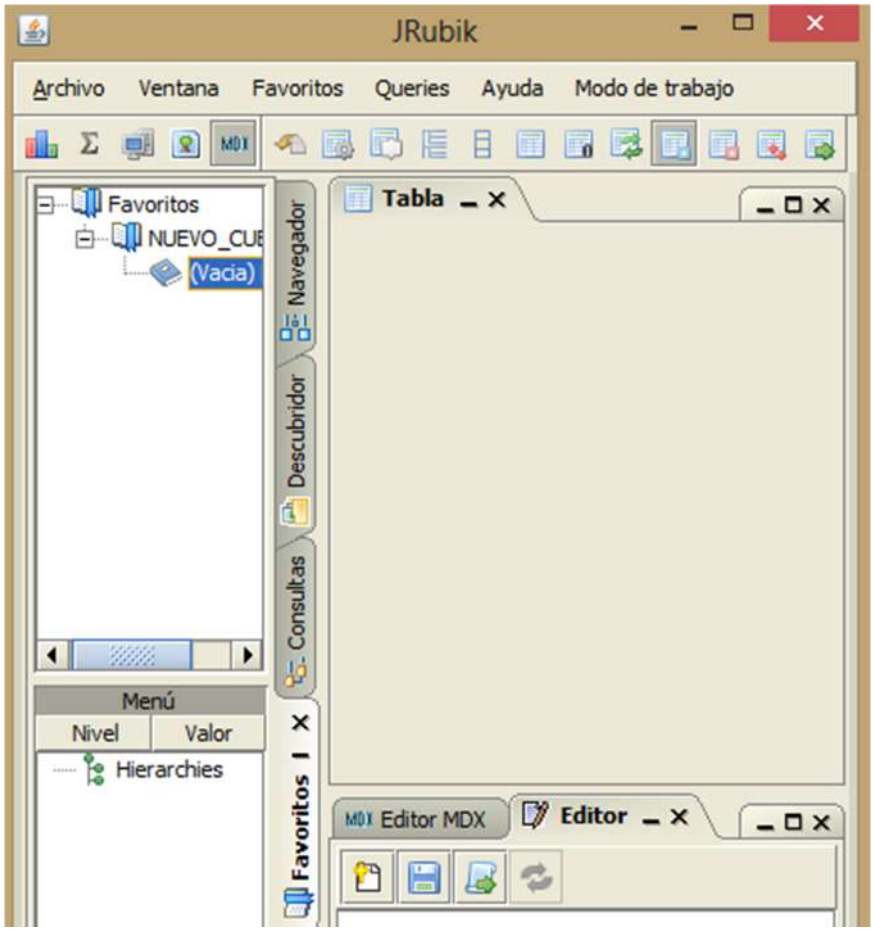


Fig. 59: Espacio trabajo de la herramienta OLAP JRubik.

Para explorar los datos del cubo, primeramente, se ha tenido que configurar la conexión con la base de datos multidimensional "dataMart_cell", seguidamente se ha enlazado con el archivo XML creado en el paso anterior "Schema_datamartventas.xml" y que tiene la estructura del cubo.

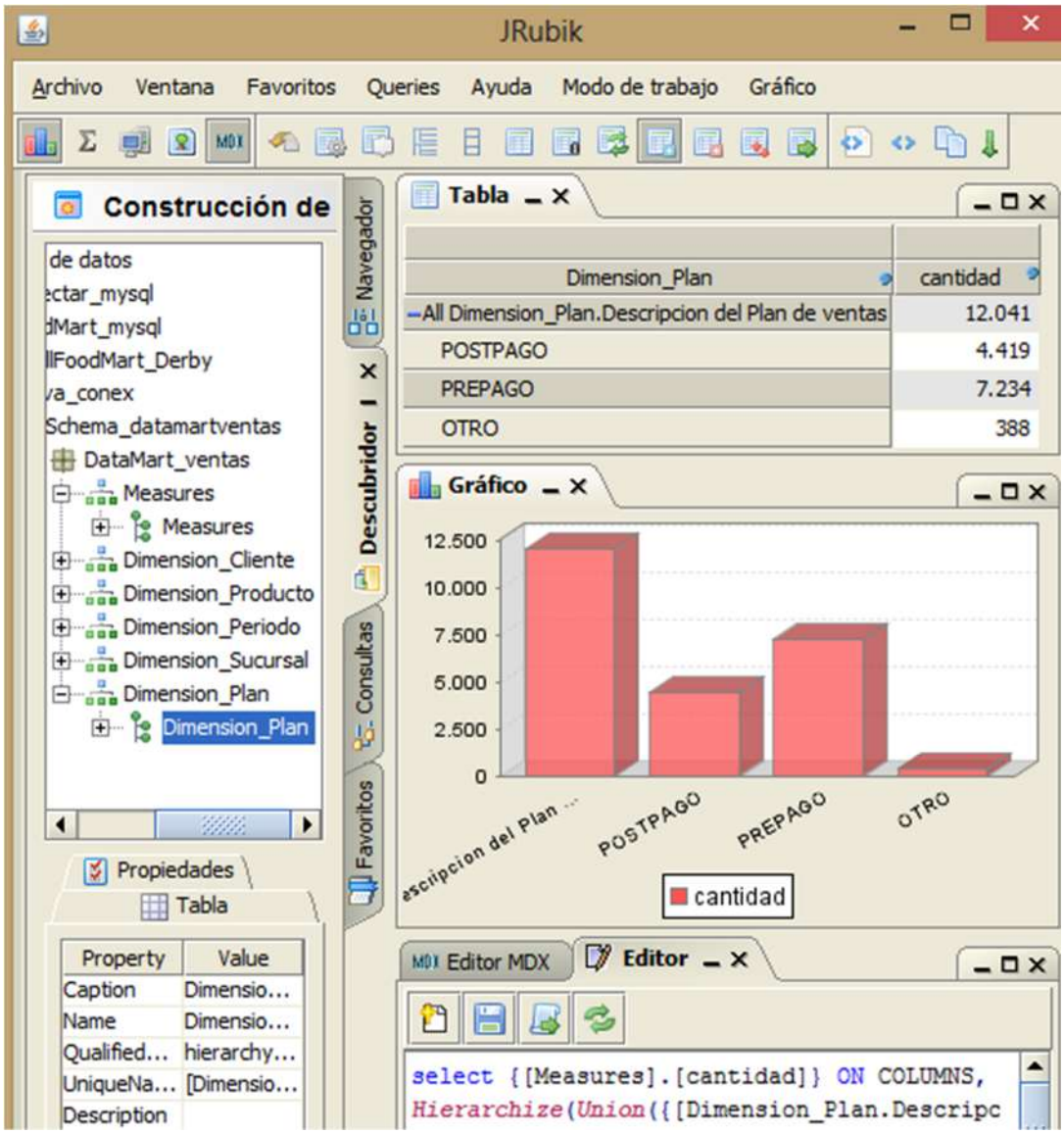


Fig. 60: Ejemplo de la exploración del cubo OLAp con JRubik.

m. Formateo de datos

En esta parte de la metodología, se ha realizado el formateo de los datos usados por el modelo de minería. Para realizar el formateo se usó funciones de cadena SQL, que han sido insertados en la sentencia SQL que extrae los datos. A continuación, se presenta la sentencia SQL que extrae los datos del data Mart de ventas.

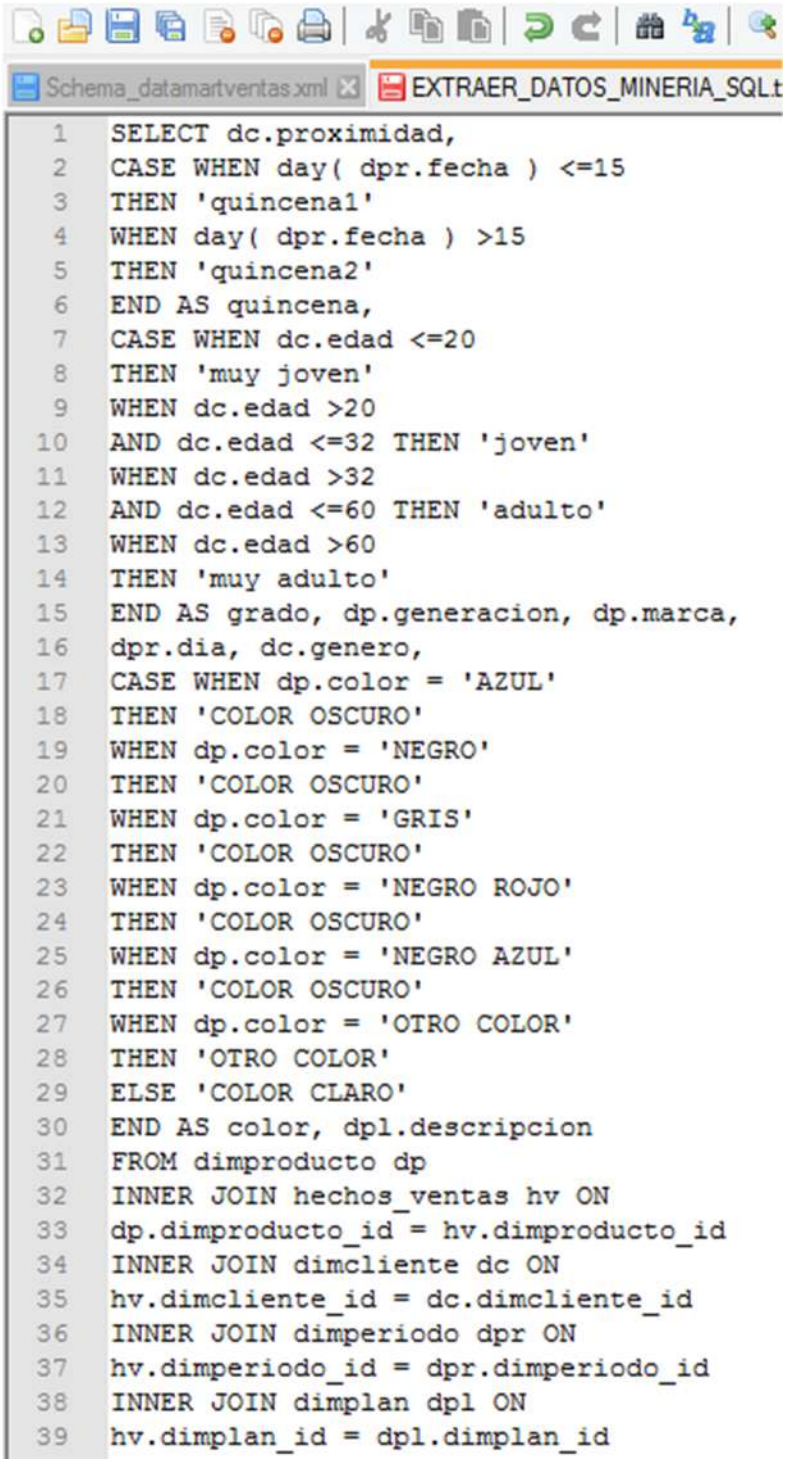


Fig. 61: Extracción de datos del dataMart para minería (usando consulta SQL)

n. Selección de la Técnica de modelado

Para seleccionar la técnica de modelado se ha tenido en cuenta el objetivo de la minería de datos. En este caso se trata de pronosticar las modalidades de venta de equipos celulares. De acuerdo a las técnicas de

minería de datos explicada en el presente documento, en el punto 2.2.5 “Técnicas de Minería” y de acuerdo al tipo de datos usados, se ha optado por las técnicas supervisadas de clasificación. Para ver nuevamente la estructura de los datos y poder describir el atributo que se ha predicho, ha sido necesario utilizar alguna herramienta de minería de datos, en este caso se ha usado la herramienta “WEKA”, versión 3.8.1, siendo además también una herramienta de código abierto “open source”.

Fig. 62: Herramienta de minería de datos WEKA

- **Exploración de los datos para minería con la herramienta WEKA**

  Para explorar los datos que han sido usados en la minería, ha sido necesario hacer una conexión con el dataMart de ventas desde Weka y ejecutar la sentencia SQL descrita en el paso “m: Formateo de datos” de la metodología.

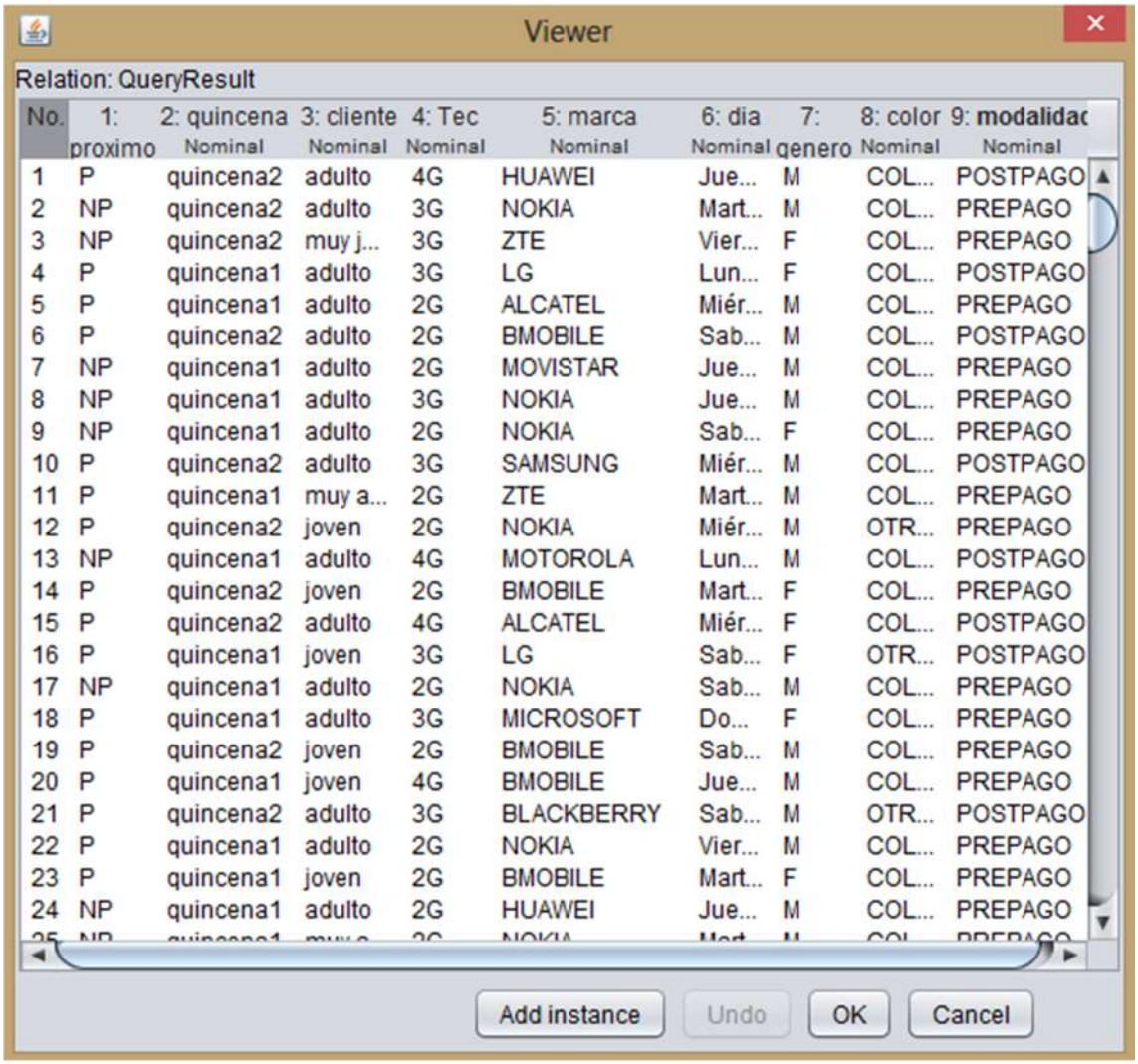


Fig. 63: Exploración de datos con WEKA

- **Descripción de los atributos de la tabla de minería**

  En la descripción del atributo “cliente”, que tiene valores vacíos “Missing” es 0%, esto se da porque el tratamiento de este tipo de datos ya se dio en la tarea “Limpieza de datos”, de la metodología. La cantidad de valores distintos que toma el atributo es 4: “Muy joven”, “joven”, “adulto” y “muy adulto”, es de tipo nominal. Ver la siguiente figura.

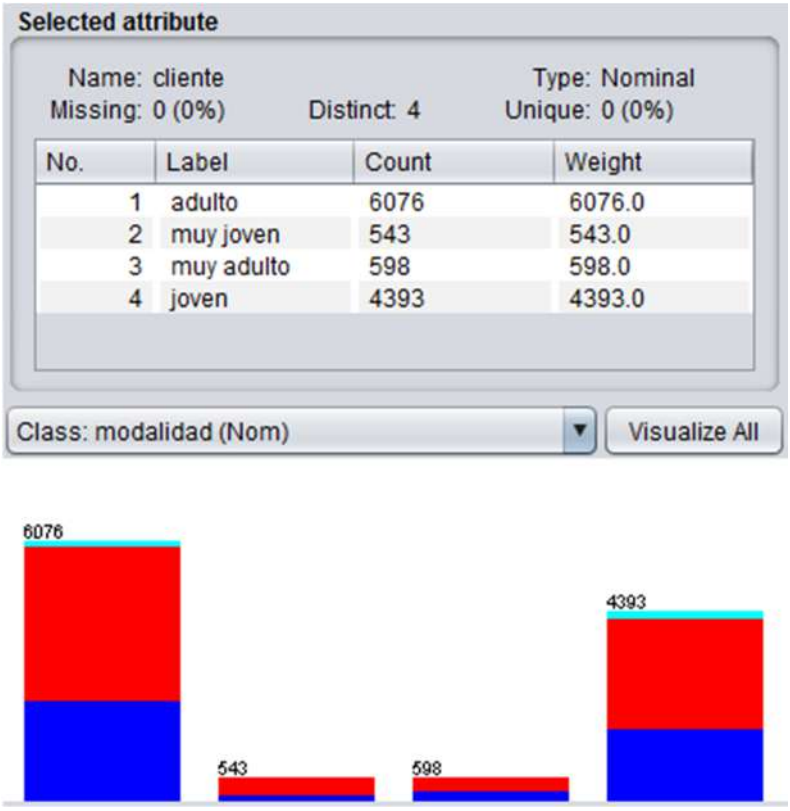


Fig. 64: Descripción del atributo "cliente" con WEKA

En caso del atributo “genero”, que corresponde a una característica del cliente, se ha tenido dos valores distintos “M: masculino”, “F: femenino” y es de tipo nominal. Ver la siguiente imagen.

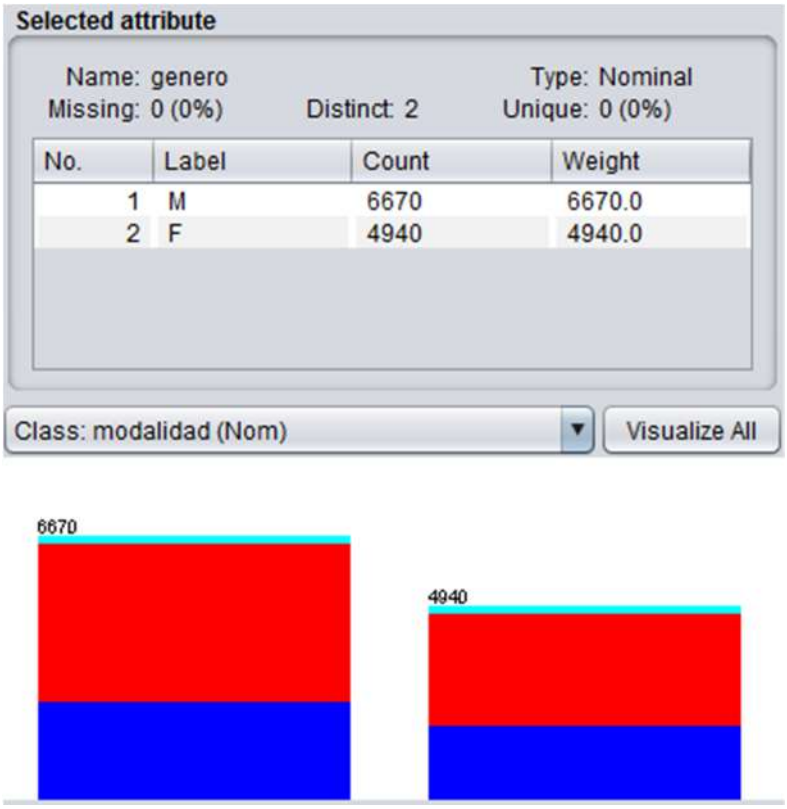


Fig. 65: Descripción del atributo “genero” con WEKA

Observando de la misma manera que en el caso anterior, el atributo “próximo”, se puede ver que tiene dos valores distintos “P”: próximo y “NP”: no próximo. Ver la siguiente imagen.

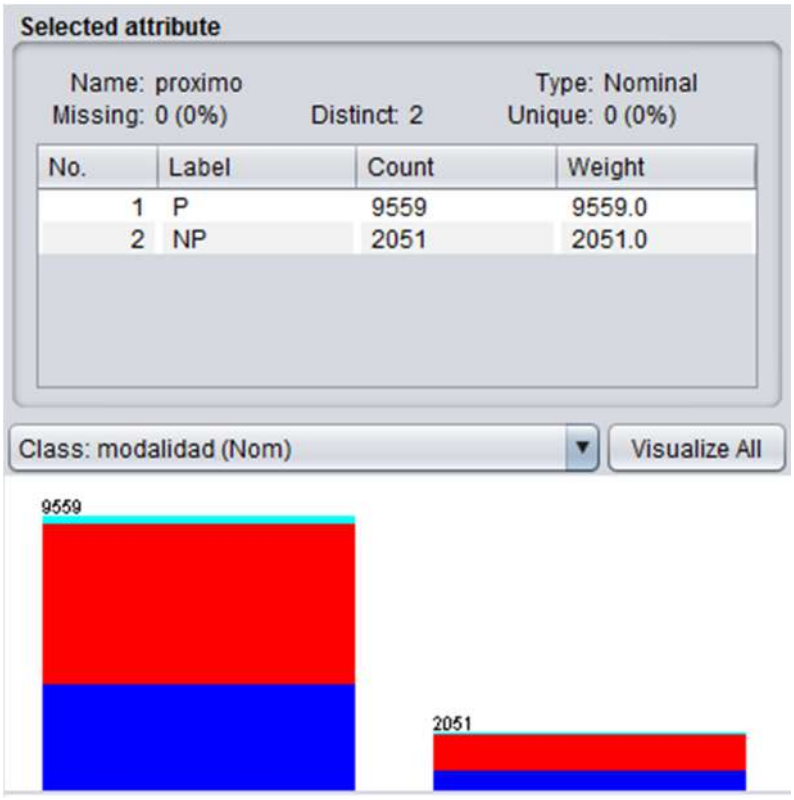


Fig. 66: Descripción del atributo “proximo” con WEKA

En el atributo “quincena” se puede ver que tiene dos valores distintos: “quincena1” y “quincena2”, es un valor de tipo nominal

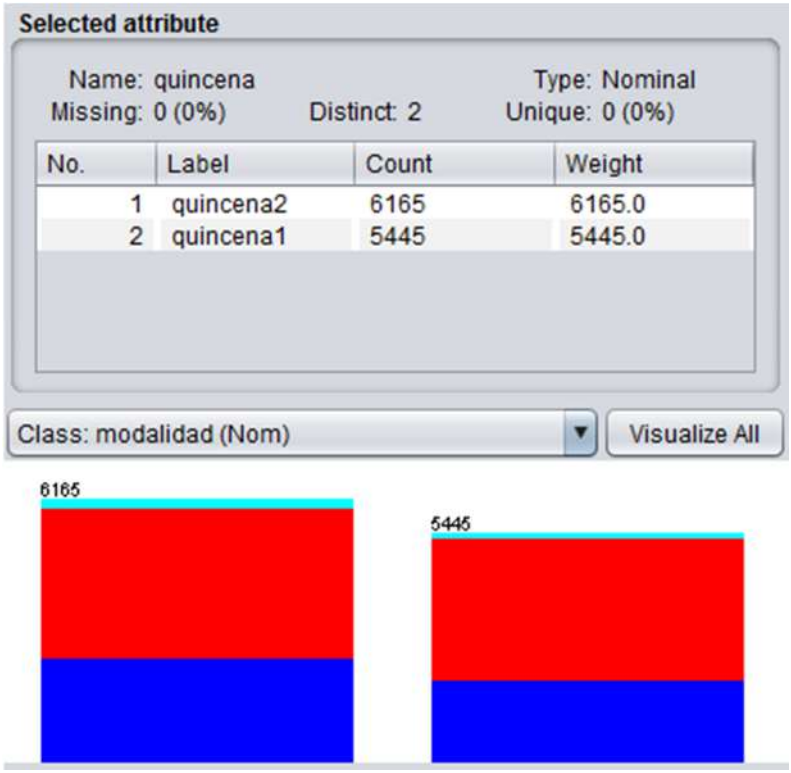


Fig. 67: Descripción del atributo “quincena” con WEKA

En el atributo “día”, se observó que tiene siete valores distintos correspondientes a los días de la semana de lunes a domingo, siendo una variable de tipo nominal.

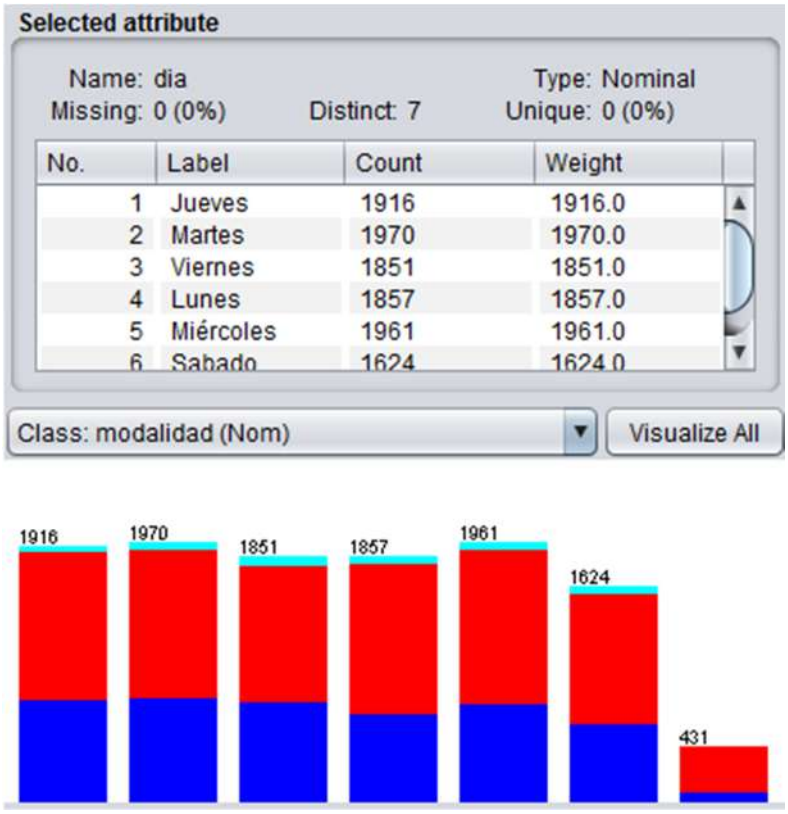


Fig. 68: Descripción del atributo “dia” con WEKA

Con respecto al atributo “marca”, se ha podido observar que tiene 13 valores distintos: “huawei”, “Nokia”, ”ZTE”, ”LG”, ”Alcatel”, ”Bmobile”, ”Movistar”, ”Samsung”, ”Motorola”, “Microsoft”, “Blackberry”, “Sony” y “otra marca”, es de tipo nominal.

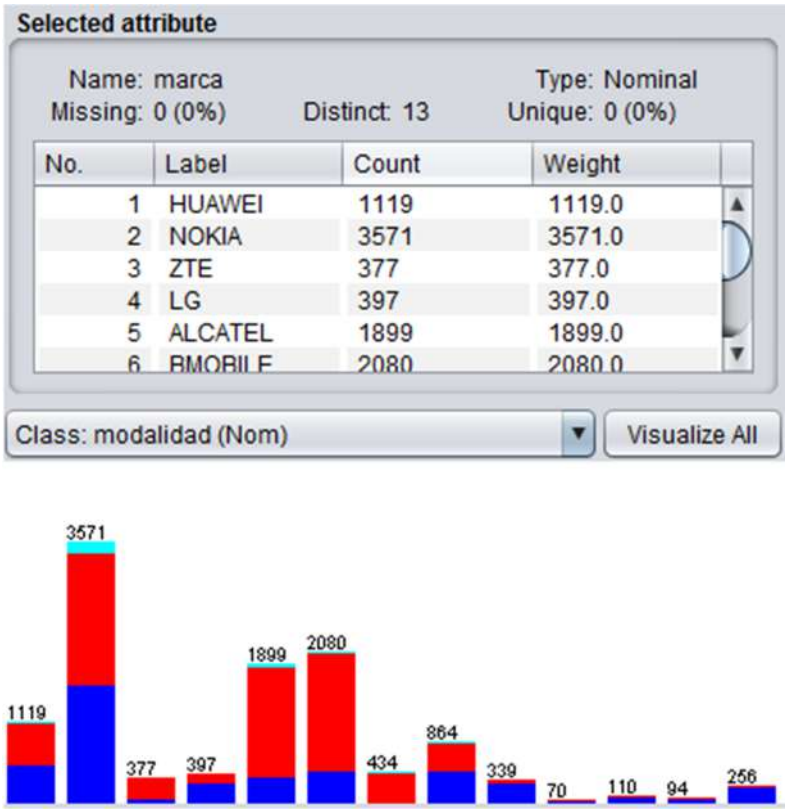


Fig. 69: Descripción del atributo “marca” con WEKA

En el atributo tecnología “Tec.” se observó que presenta tres valores distintos “2G”,”3G”, “4G” y es una variable de tipo nominal. Ver la siguiente imagen.

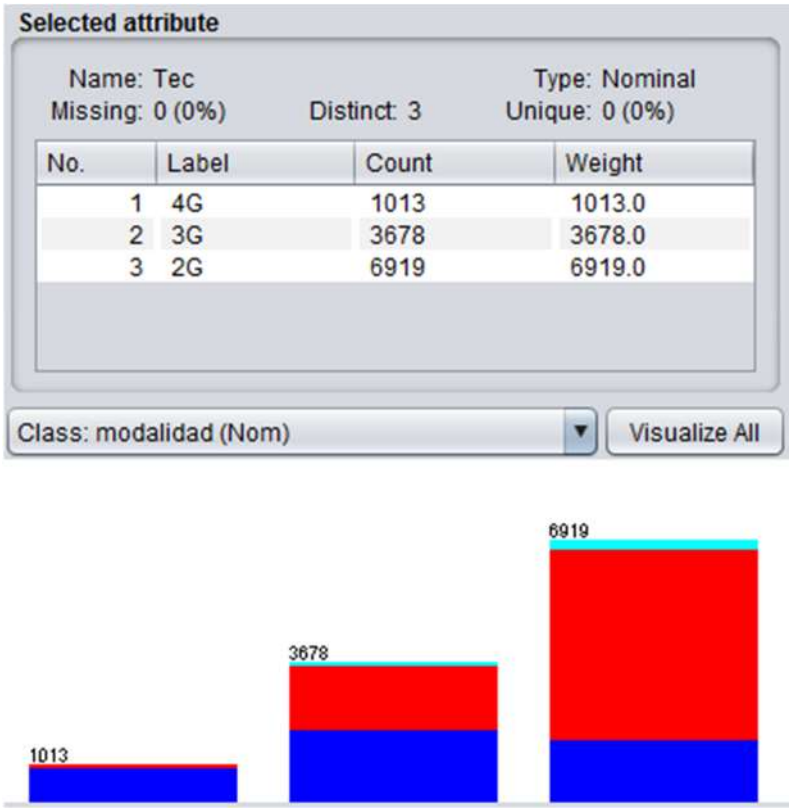


Fig. 70: Descripción del atributo “Tecnología” con WEKA

En el atributo “color” se observó que presenta tres valores distintos “Color claro”, “Color oscuro”, “Otro color” y es una variable de tipo nominal. Ver la siguiente figura.

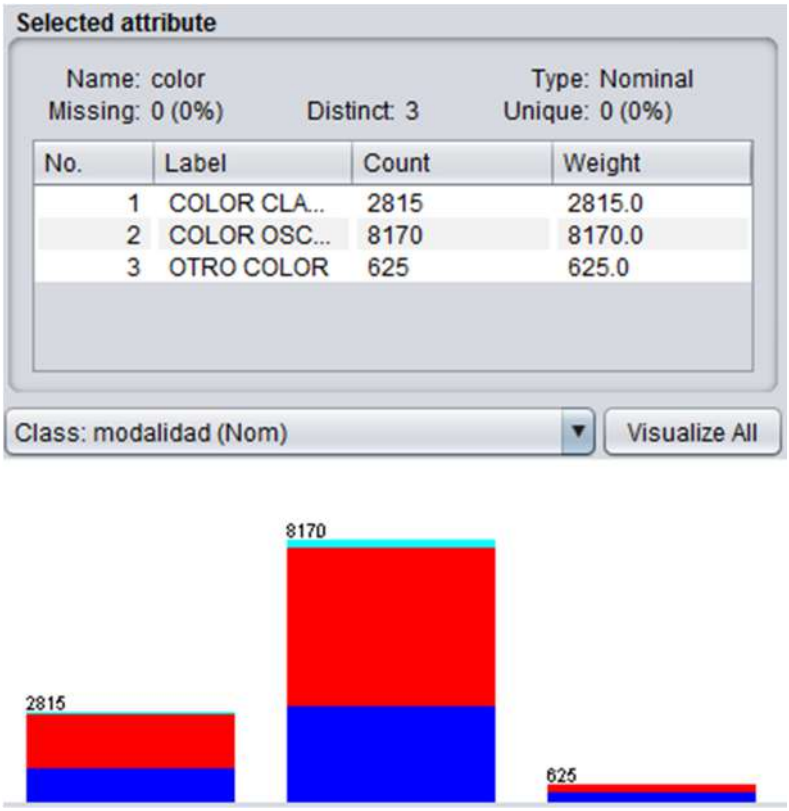


Fig. 71: Características del atributo “color” con WEKA

Finalmente, en el atributo “modalidad” se observó que presento tres valores “Prepago”, “Postpago”, “Otro” y es una variable de tipo nominal y represento al atributo clase o variable a predecir. Ver la siguiente imagen.

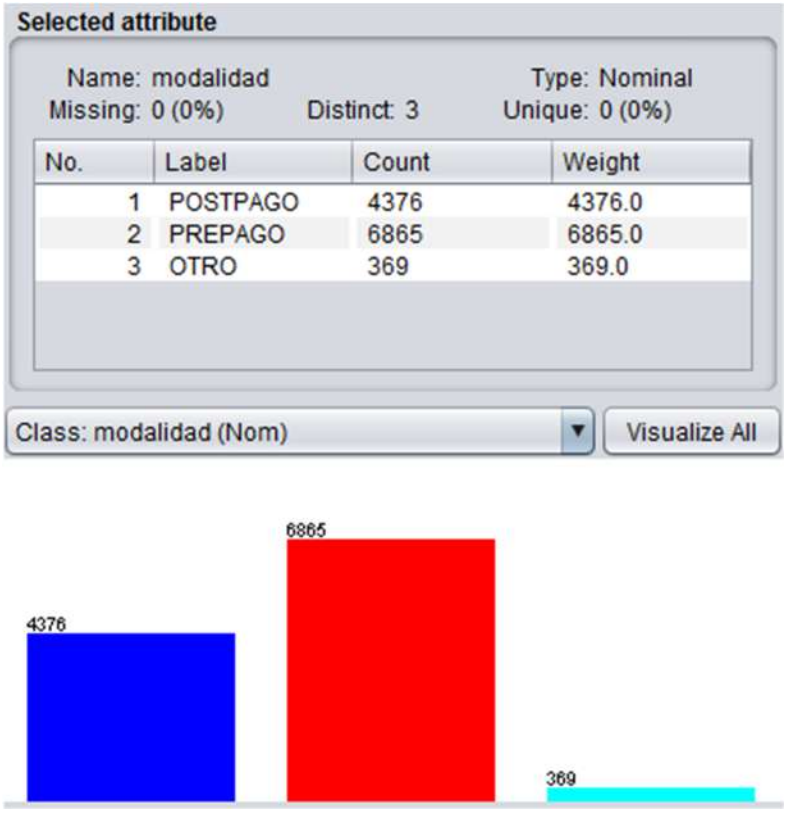


Fig. 72: Características del atributo clase “modalidad” con WEKA

La exploración de las características de los atributos de la tabla de minería ha permitido justificar la decisión de elegir a la técnica de “clasificación”, la técnica de minería usada en la presente investigación y variables o atributos son de tipo nominales.

o. Diseño de las pruebas del modelo

Del conjunto total de registros se ha seleccionado dos subconjuntos de 5805 registros cada uno, un conjunto de datos de entrenamiento y otro conjunto de pruebas. Seguidamente se pasó a probar la calidad de los resultados con los algoritmos de clasificación.

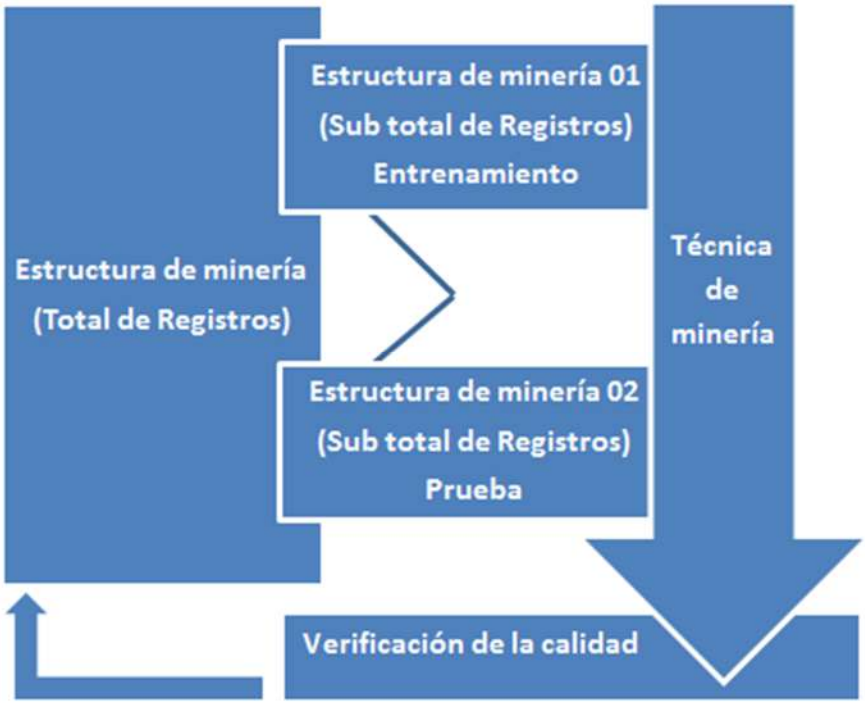


Fig. 73: diseño de pruebas del modelo de minería

Para el diseño de pruebas del modelo de minería elegido se ha hecho uso de los algoritmos de minería de datos que incluye la herramienta WEKA.

Los algoritmos de clasificación usados para comparar la calidad del modelo fueron: “REPTree”, “AttributeSelectedClassifier”, “J48”, “Jrip”. Los criterios de comparación han consistido en el tiempo de construcción del modelo en segundos y medir el porcentaje de registros clasificados correctamente. Ver la siguiente tabla Nº 16:

Tabla 19: Comparación de algoritmos de clasificación

Comparación de algoritmos de clasificación de WEKA		
Algoritmo	**Tiempo (s)**	**Porcentaje de acierto %**
REPTree	0.25	70.1886
AttributeSelectedClassifier	1.47	71.0174
J48	0.33	71.7743
Jrip	5.53	71.068

- **Elección del modelo óptimo.**

  La tabla anterior Nº16 se usó para evaluar los resultados de los algoritmos y elegir el modelo de clasificación en base al menor error posible y el menor tiempo posible para construir el modelo, en el cual se observó que el algoritmo “REPTree” puede construir el modelo en 0.25 segundos, para este caso de estudio es el más rápido y tiene un porcentaje de acierto de 70.1886 %, lo que indica que tiene un acierto menor al de todos. El algoritmo “AttributeSelectedClassifier” puede construir el modelo en 1.47 segundos y tiene una tasa de acierto de 71.0174%. El algoritmo “J48” puede construir el modelo en 0.33 segundos y tiene una tasa de acierto de 71.7743 %. El algoritmo “Jrip” puede construir el modelo en 5.53 segundos, siendo el más lento de todos y tiene una tasa de acierto 71.068 %. En conclusión, se determinó que el modelo más óptimo es el construido por el algoritmo “J48”, debido a que el porcentaje de acierto es mayor, comparado con los demás algoritmos y el tiempo de construcción del modelo es uno de los más rápidos.

p. Construcción del modelo

Después de haber elegido el modelo de clasificación y el algoritmo J48 para su construcción se ha paso a evaluar los modelos generados modificando los parámetros de ejecución del algoritmo. Se ha tomado dos parámetros diferentes que son el porcentaje de división del total de registros “porcentaje split” (porcentaje para entrenamiento y lo resto para pruebas) y la validación cruzada “cross-validation” que representa el número de tablas de registros en las cuales se evaluara el modelo. WEKA por defecto maneja un porcentaje de división de 66% y una validación cruzada de 10, estos valores se variaron para optimizar el modelo. A continuación, se presenta una tabla Nº17, donde se ve los resultados de la evaluación del modelo cuando se varía el número de pliegues de entrenamientos “folds”, con un total de 20 pruebas realizadas para optimizar el modelo.

Tabla 20: Optimización del modelo con validación cruzada

Optimización del modelo por validación cruzada	
# folds (pliegues)	**Porcentaje de acierto %**
1	71,1283
2	71,5676
3	71,5848
4	71,1283
5	71,8346
6	71,7054
7	71,8088
8	71,8949
9	71,9724
10	71,7485
11	71,8346
12	71,7829
13	71,6021
14	71,6537
15	71,8346
16	71,8691
17	71,671
18	71,8777
19	71,8519
20	71,8002

En la siguiente tabla Nº18 podemos ver los resultados cuando se varía el porcentaje de división "split" de los registros de entrenamiento, en total también 20 pruebas.

Tabla 21 : Optimización del modelo con porcentaje de división

Optimización del modelo por porcentaje de división	
Porcentaje de división % (split)	**Porcentaje de acierto %**
5	69,1178
10	71,1838
15	71,5241
20	71,2317
25	71,6665
30	71,5639

35	71,6936
40	71,5619
45	71,4957
50	71,9724
55	71,7266
60	71,6624
65	71,7696
70	71,404
75	70,7443
80	71,7485
85	71,4532
90	70,9733
95	69,6552
100	71,1838

q. Evaluación del modelo

En los resultados de las evaluaciones se pudo notar que el mejor porcentaje de acierto es 71.9724 % para una validación cruzada de 9. También, el mejor porcentaje de acierto que es de 71.9724 % para un porcentaje de división (split) de 50%. Luego de analizar los resultados de la evaluación del modelo con los dos parámetros “validación cruzada” y “porcentaje de división” se pudo determinar que el modelo optimizado tenía que ser construido con una validación cruzada de equivalente a 9 o un porcentaje de división del 50%. Para poder interpretar el modelo generado por el algoritmo J48, se presenta un ejemplo de lectura de los resultados de un árbol de decisión de WEkA Fig.74.

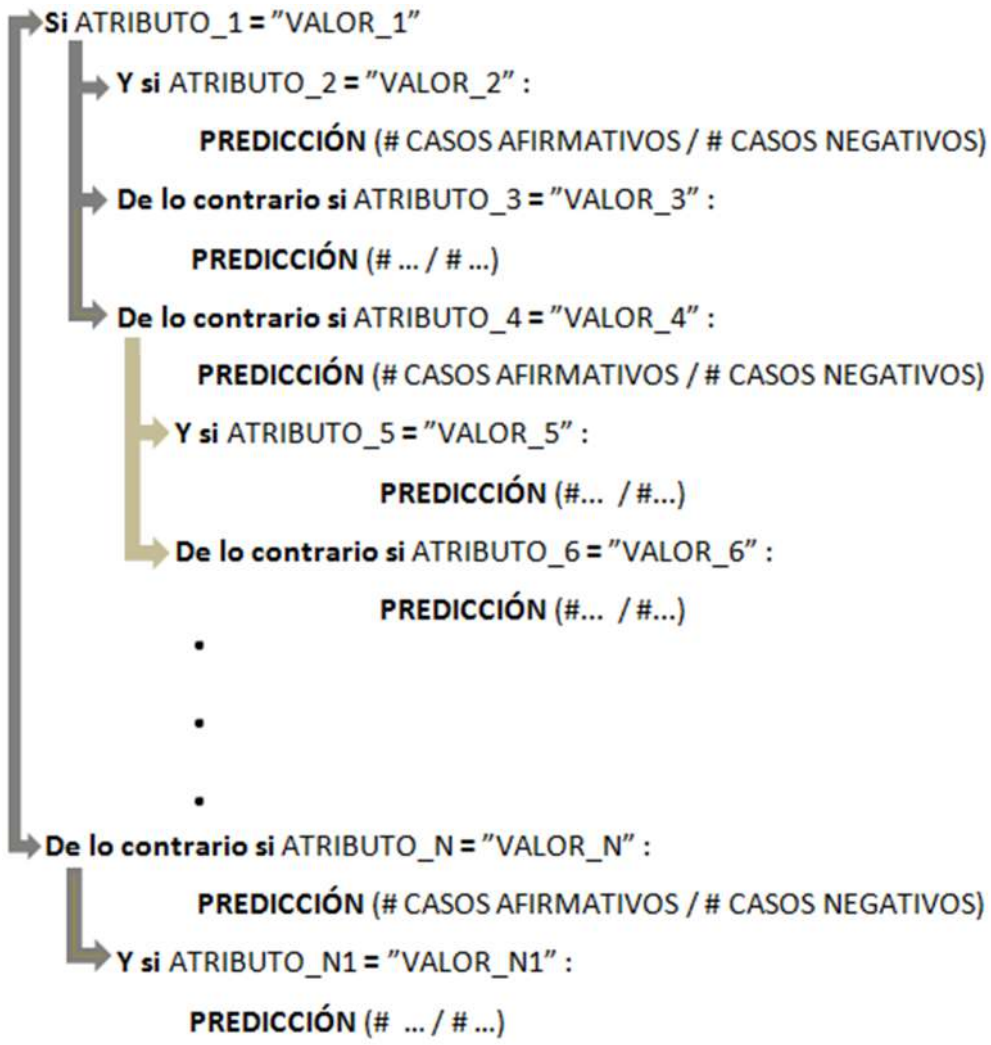


Fig. 74: Ejemplo de lectura del árbol de decisión del algoritmo J48

A continuación, se presenta la tabla Nº 21 con los resultados del modelo optimizado, con el respectivo árbol de decisión.

Tabla 22: Representación del modelo optimizado generado con WEKA

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     QueryResult
Instances:    11610
Attributes:   9
proximo
          quincena
          cliente
Tec
          marca
dia
          genero
          color
          modalidad
Test mode:    split 50.0% train, remainder test
=== Classifier model (full training set) ===
J48 prunedtree
------------------
Tec = 2G
|   marca = BMOBILE: PREPAGO (1603.0/330.0)
|   marca = NOKIA: PREPAGO (2549.0/1091.0)
|   marca = SAMSUNG
|   |   color = COLOR OSCURO: PREPAGO (263.0/71.0)
|   |   color = COLOR CLARO
|   |   |   quincena = quincena1: POSTPAGO (26.0/11.0)
|   |   |   quincena = quincena2
|   |   |   |   dia = Viernes: POSTPAGO (3.0)
|   |   |   |   dia = Jueves: POSTPAGO (2.0)
|   |   |   |   dia = Domingo: PREPAGO (0.0)
|   |   |   |   dia = Martes: PREPAGO (6.0/2.0)
|   |   |   |   dia = Lunes: PREPAGO (5.0/1.0)
|   |   |   |   dia = Miércoles: PREPAGO (5.0/1.0)
|   |   |   |   dia = Sabado: PREPAGO (3.0/1.0)
|   |   color = OTRO COLOR: POSTPAGO (1.0)
|   marca = BLACKBERRY: POSTPAGO (26.0/3.0)
|   marca = HUAWEI: PREPAGO (471.0/64.0)
|   marca = ALCATEL: PREPAGO (1131.0/119.0)
|   marca = MICROSOFT: PREPAGO (0.0)
|   marca = LG: PREPAGO (42.0/14.0)
|   marca = OTRA MARCA
|   |   proximo = P: OTRO (25.0/15.0)
|   |   proximo = NP: PREPAGO (6.0/2.0)
|   marca = MOVISTAR: PREPAGO (402.0/31.0)
|   marca = SONY: POSTPAGO (19.0/1.0)
|   marca = MOTOROLA
|   |   color = COLOR OSCURO: POSTPAGO (46.0/20.0)
|   |   color = COLOR CLARO: POSTPAGO (0.0)
|   |   color = OTRO COLOR: PREPAGO (6.0/2.0)
|   marca = ZTE: PREPAGO (279.0/50.0)
```

```
Tec = 3G
|   marca = BMOBILE: PREPAGO (402.0/99.0)
|   marca = NOKIA: POSTPAGO (960.0/389.0)
|   marca = SAMSUNG
|   |   color = COLOR OSCURO
|   |   |   quincena = quincena1
|   |   |   |   dia = Viernes: POSTPAGO (15.0/3.0)
|   |   |   |   dia = Jueves: PREPAGO (17.0/7.0)
|   |   |   |   dia = Domingo: PREPAGO (1.0)
|   |   |   |   dia = Martes: PREPAGO (30.0/12.0)
|   |   |   |   dia = Lunes
|   |   |   |   |   proximo = P: POSTPAGO (20.0/4.0)
|   |   |   |   |   proximo = NP: PREPAGO (2.0)
|   |   |   |   dia = Miércoles
|   |   |   |   |   cliente = muy adulto: POSTPAGO (1.0)
|   |   |   |   |   cliente = joven: PREPAGO (9.0/2.0)
|   |   |   |   |   cliente = adulto: POSTPAGO (8.0/2.0)
|   |   |   |   |   cliente = muy joven: PREPAGO (1.0)
|   |   |   |   dia = Sabado: PREPAGO (14.0/5.0)
|   |   |   quincena = quincena2: POSTPAGO (158.0/59.0)
|   |   color = COLOR CLARO
|   |   |   proximo = P
|   |   |   |   cliente = muy adulto: POSTPAGO (2.0)
|   |   |   |   cliente = joven
|   |   |   |   |   quincena = quincena1: PREPAGO (15.0/6.0)
|   |   |   |   |   quincena = quincena2: POSTPAGO (22.0/5.0)
|   |   |   |   cliente = adulto: POSTPAGO (33.0/13.0)
|   |   |   |   cliente = muy joven: PREPAGO (3.0)
|   |   |   proximo = NP: PREPAGO (17.0/7.0)
|   |   color = OTRO COLOR: POSTPAGO (56.0/7.0)
|   marca = BLACKBERRY: POSTPAGO (84.0/13.0)
|   marca = HUAWEI
|   |   dia = Viernes
|   |   |   genero = M: PREPAGO (25.0/10.0)
|   |   |   genero = F: POSTPAGO (16.0/7.0)
|   |   dia = Jueves: POSTPAGO (61.0/29.0)
|   |   dia = Domingo: PREPAGO (7.0/2.0)
|   |   dia = Martes
|   |   |   color = COLOR OSCURO: PREPAGO (31.0/9.0)
|   |   |   color = COLOR CLARO: POSTPAGO (6.0/2.0)
|   |   |   color = OTRO COLOR: POSTPAGO (2.0/1.0)
|   |   dia = Lunes
|   |   |   proximo = P
|   |   |   |   quincena = quincena1: POSTPAGO (10.0/4.0)
|   |   |   |   quincena = quincena2
|   |   |   |   |   cliente = muy adulto: PREPAGO (0.0)
|   |   |   |   |   cliente = joven: PREPAGO (7.0/1.0)
|   |   |   |   |   cliente = adulto
|   |   |   |   |   |   color = COLOR OSCURO: POSTPAGO (9.0/2.0)
|   |   |   |   |   |   color = COLOR CLARO: PREPAGO (5.0/1.0)
|   |   |   |   |   |   color = OTRO COLOR: PREPAGO (1.0)
|   |   |   |   cliente = muy joven: PREPAGO (3.0/1.0)
```

```
|   |   |   proximo = NP: PREPAGO (8.0/2.0)
|   |   dia = Miércoles
|   |   |   color = COLOR OSCURO
|   |   |   |   proximo = P: POSTPAGO (21.0/6.0)
|   |   |   |   proximo = NP: PREPAGO (7.0/2.0)
|   |   |   color = COLOR CLARO
|   |   |   |   proximo = P: PREPAGO (9.0/5.0)
|   |   |   |   proximo = NP: POSTPAGO (2.0/1.0)
|   |   |   color = OTRO COLOR: POSTPAGO (0.0)
|   |   dia = Sabado
|   |   |   cliente = muy adulto: PREPAGO (0.0)
|   |   |   cliente = joven
|   |   |   |   proximo = P: PREPAGO (8.0/2.0)
|   |   |   |   proximo = NP: POSTPAGO (3.0)
|   |   |   cliente = adulto
|   |   |   |   quincena = quincena1: POSTPAGO (5.0/1.0)
|   |   |   |   quincena = quincena2: PREPAGO (13.0/3.0)
|   |   |   cliente = muy joven: PREPAGO (3.0)
|   marca = ALCATEL: PREPAGO (712.0/233.0)
|   marca = MICROSOFT
|   |   proximo = P: PREPAGO (33.0/6.0)
|   |   proximo = NP: POSTPAGO (10.0/3.0)
|   marca = LG
|   |   color = COLOR OSCURO
|   |   |   dia = Viernes
|   |   |   |   cliente = muy adulto: POSTPAGO (0.0)
|   |   |   |   cliente = joven: POSTPAGO (5.0/2.0)
|   |   |   |   cliente = adulto: POSTPAGO (4.0/1.0)
|   |   |   |   cliente = muy joven: PREPAGO (3.0)
|   |   |   dia = Jueves: POSTPAGO (15.0/4.0)
|   |   |   dia = Domingo: PREPAGO (5.0/1.0)
|   |   |   dia = Martes
|   |   |   |   genero = M: POSTPAGO (9.0/3.0)
|   |   |   |   genero = F: PREPAGO (6.0/2.0)
|   |   |   dia = Lunes
|   |   |   |   quincena = quincena1: POSTPAGO (6.0/2.0)
|   |   |   |   quincena = quincena2: PREPAGO (12.0/3.0)
|   |   |   dia = Miércoles
|   |   |   |   quincena = quincena1: PREPAGO (14.0/4.0)
|   |   |   |   quincena = quincena2: POSTPAGO (10.0/3.0)
|   |   |   dia = Sabado
|   |   |   |   quincena = quincena1: POSTPAGO (8.0/1.0)
|   |   |   |   quincena = quincena2: PREPAGO (10.0/4.0)
|   |   color = COLOR CLARO
|   |   |   dia = Viernes
|   |   |   |   cliente = muy adulto: PREPAGO (1.0)
|   |   |   |   cliente = joven: POSTPAGO (4.0/1.0)
|   |   |   |   cliente = adulto: POSTPAGO (7.0/2.0)
|   |   |   |   cliente = muy joven: PREPAGO (2.0)
|   |   |   dia = Jueves
|   |   |   |   cliente = muy adulto: PREPAGO (0.0)
|   |   |   |   cliente = joven: PREPAGO (4.0)
```

```
|   |   |   cliente = adulto: POSTPAGO (6.0/2.0)
|   |   |   cliente = muy joven: POSTPAGO (1.0)
|   |   dia = Domingo: POSTPAGO (1.0)
|   |   dia = Martes
|   |   |   genero = M: PREPAGO (4.0/1.0)
|   |   |   genero = F: POSTPAGO (13.0/5.0)
|   |   dia = Lunes: POSTPAGO (13.0/6.0)
|   |   dia = Miércoles
|   |   |   quincena = quincena1: POSTPAGO (9.0/3.0)
|   |   |   quincena = quincena2: PREPAGO (4.0)
|   |   dia = Sabado
|   |   |   proximo = P: POSTPAGO (9.0/2.0)
|   |   |   proximo = NP: PREPAGO (2.0)
|   |   color = OTRO COLOR: POSTPAGO (41.0/11.0)
|   marca = OTRA MARCA
|   |   dia = Viernes: POSTPAGO (0.0)
|   |   dia = Jueves
|   |   |   color = COLOR OSCURO: PREPAGO (5.0/1.0)
|   |   |   color = COLOR CLARO: POSTPAGO (5.0)
|   |   |   color = OTRO COLOR: POSTPAGO (0.0)
|   |   dia = Domingo: POSTPAGO (0.0)
|   |   dia = Martes: POSTPAGO (7.0)
|   |   dia = Lunes
|   |   |   quincena = quincena1
|   |   |   |   genero = M: POSTPAGO (2.0)
|   |   |   |   genero = F: PREPAGO (3.0)
|   |   |   quincena = quincena2: POSTPAGO (5.0/1.0)
|   |   dia = Miércoles: PREPAGO (4.0/1.0)
|   |   dia = Sabado: POSTPAGO (3.0/1.0)
|   marca = MOVISTAR: PREPAGO (32.0/1.0)
|   marca = SONY: POSTPAGO (187.0/27.0)
|   marca = MOTOROLA: POSTPAGO (216.0/30.0)
|   marca = ZTE: PREPAGO (94.0/22.0)
Tec = 4G: POSTPAGO (1013.0/136.0)

Number of Leaves  :         123

Size of the tree :          170


Time taken to build model: 0.25 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.52 seconds

=== Summary ===

Correctly Classified Instances        4178               71.9724 %
Incorrectly Classified Instances      1627               28.0276 %
Kappa statistic                          0.4037
Mean absolute error                      0.2514
```

Root mean squared error 0.3612
Relative absolute error 74.3106 %
Root relative squared error 88.0871 %
Total Number of Instances 5805

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,542	0,143	0,692	0,542	0,608	0,425	0,788	0,677	POSTPAGO
	0,866	0,468	0,733	0,866	0,794	0,428	0,776	0,801	PREPAGO
	0,028	0,002	0,313	0,028	0,051	0,085	0,611	0,051	OTRO
Weighted Avg.	0,720	0,333	0,705	0,720	0,702	0,416	0,775	0,732	

=== Confusion Matrix ===

```
   a    b    c   <-- classified as
 1170  983    5 |    a = POSTPAGO
 458 3003    6 |    b = PREPAGO
  63  112    5 |    c = OTRO
```

r. Evaluación de los resultados

En esta parte de la metodología se evaluó los resultados que nos muestra el algoritmo de clasificación y la forma en como decide el algoritmo con los datos que se le proporciono. Se evaluó lo siguiente, según la lectura del árbol generado por el algoritmo:

→ Si se vende al cliente un equipo celular con tecnología “2G” y si la marca es “BMOBILE”, de un total de 1933 ventas, 1603 son vendidos como PREPAGO; de lo contrario si la marca es “NOKIA” de un total de 3640 ventas, 2549 son vendidos como prepago; de lo contrario si la marca es “SAMSUNG”, y si el color es “color oscuro” de un total de 334 ventas 263 son vendidos como “PREPAGO”; de lo contrario si el color es “color claro” y si la quincena es la primera quincena del mes “quincena1” de un total de 37 ventas, 26 ventas se realizan como “POSTPAGO” ; de lo contrario si la quincena es la segunda quincena del mes “quincena2” se tiene lo más probable que se

venda “PREPAGO” son los días “Lunes”, “Martes” y “Miércoles”; de lo contrario si la marca es “BLACKBERRY” de un total de 29 ventas 26 ventas son “POSTPAGO”; lo contrario si la marca es “HUAWEI” de un total de 535 ventas 471 ventas son en “PREPAGO”; de lo contrario si la marca es “ALCTEL” de un total de 1250 ventas 1131 ventas son “PREPAGO”; lo contrario si la marca es ”MICROSOFT” no ocurre ventas; de lo contrario si la marca es “LG” de un total de 56 ventas 42 ventas son en “PREPAGO”; de lo contrario si la marca es “MOVISTAR” de un total de 433 ventas 402 ventas son en “PREPAGO”; de lo contrario si la marca es “SONY” de 20 ventas 19 ventas son en “POSTPAGO”; de lo contrario si la marca es “MOTOROLA ”lo más vendidos el “Color oscuro” de 66 ventas 46 ventas son en “POSTPAGO”; de lo contrario si la marca es “ZTE” de 329 ventas 279 ventas en “PREPAGO”

→ Si las ventas al cliente son equipos celular con tecnología “3G” y si la marca es “BMOBILE” de 501 ventas, 402 ventas son en “PREPAGO”; de lo contrario si la marca es “NOKIA” de 1349 ventas 960 ventas se venden en “POSTPAGO”; de lo contrario si la marca es “SAMSUNG” y si el color es “Color oscuro” y si la quincena es la primera quincena del mes “quincena1” los mejores días para vender “PREPAGO” son los días “martes”, “jueves” y “sábado”, mientras “lunes” y “viernes” se venden más “POSTPAGO”; de lo contrario si la quincena es la segunda quincena de la semana “quincena2” de 217 ventas 158 ventas son “POSTPAGO”; de lo contrario, siguiendo con la marca “SAMSUNG” si el color es “color claro” y si el clientes es próximo al punto de venta “P” y si es ”joven” y si es la primera quincena del mes, de 21 ventas 15 ventas son en “PREPAGO”; lo contrario si es la segunda quincena “quincena2” de un total de 27 ventas, 22 ventas son en “POSTPAGO”, pero si el cliente es “adulto” de un total de 46 ventas, 33 ventas son “POSTPAGO”; por el contrario si el cliente no es próximo al punto de venta “NP” de 24 ventas 17 ventas son “PREPAGO”; siguiendo con la marca “SAMSUNG”, si por el contrario el color no es claro ni tampoco

oscuro, de un total de 63 ventas 56 ventas son “POSTPAGO”. Continuando con la marcas de la venta de celulares “3G”, si por el contrario la marca es “BLACKBERRY” de un total de 97 ventas 84 ventas son “POSTPAGO”; de lo contrario si la marca es “HUAWEI” los días en que más se vendió fue “jueves” de 90 ventas 69 ventas fueron “POSTPAGO” ; pero si el día es “martes” se venden más en color oscuro de un total de 40 ventas 31 ventas son equipos “PREPAGO”, si son de otro color se venden en “POSTPAGO” pero en menor cantidad; si el día es “Lunes” y el cliente es próximo “P” al punto de venta y si es la primera quincena del mes “quincena1” de 14 ventas 10 ventas son “POSTPAGO”, además si cliente es “joven” se vendieron de un total de 8 ventas 7 ventas fueron “PREPAGOS”; de lo contrario si el cliente es adulto y los equipos son de color oscuro, se vendieron de un total de 11 ventas 9 fueron en “POSTPAGO”; pero si el cliente es no próximo al punto de venta “NP” de un total de 10 ventas 8 ventas son en “PREPAGO”. Continuando con la marca “HUAWEI” si el día es “miércoles” la venta de equipos en colores oscuros a clientes próximos “P” fueron de un total de 27 ventas 21 fueron “POSTPAGO” mientras que a clientes no próximos de un total de 9 ventas 7 ventas son “PREPAGO”; de lo contrario si el color es claro de un total de 14 ventas 9 ventas fueron a clientes próximos “P” en “PREPAGO”. siguiendo con la marca “HUAWEI”, si el día es “sábado” y si el cliente es próximo “P” y es “joven” de un total de 10 ventas 8 ventas son “PREPAGO”; de lo contrario si es “joven” y es no próximo “NP” ocurren 3 ventas en “POSTPAGO”; de lo contrario si es día “sábado” y si el cliente es “adulto” y es la primera quincena del mes “quincena1” de un total de 6 ventas 5 ventas son en “POSTPAGO”; pero si es la segunda quincena del mes “quincena2” de un total de 16 ventas 13 son ventas “PREPAGO”. Volviendo a nodo del árbol “marca” con respecto a la tecnología “3G”, si la marca es “ALCATEL” de un total de 945 ventas 712 ventas son “PREPAGO”; de lo contrario si la marca es “MICROSOFT” y el cliente es próximo “P” de 39 ventas 33 ventas son “PREPAGO” y si no es próximo “NP” de 13

ventas 10 ventas son “POSTPAGO”; de contrario si la marca es “LG” y el color es oscuro y el día es viernes y el cliente es “joven”, de un total de 7 ventas 5 ventas son “POSTPAGO”, pero si el día fuera “jueves” de un total de 19 ventas 15 ventas son “POSTPAGO”, pero si el día es “Martes” y el género del cliente es masculino “M” de un total de 12 ventas 9 ventas son “POSTPAGO”, respecto al mismo día y si el género fuera femenino “F” de un total de 8 ventas 6 ventas son “PREPAGO”, pero si el día es “lunes” y además es la segunda quincena del mes “quincena2” de un total de 15 ventas 12 ventas son “PREPAGO”, pero si fuera día “miércoles” y es la primera quincena del mes “quincena1” de un total de 18 ventas 14 ventas son “PREPAGO”, también se puede ver que si el día es “sábado” y es la primera quincena del mes “quincena1” de un total de 9 ventas 8 ventas son en “POSTPAGO” , respecto al mismo día y si fuera “quincena2” de un total de 14 ventas 10 ventas son en “PREPAGO”. Continuando con la descripción de la rama del árbol, cuando la marca es “LG” y además si el nodo color: “COLOR CLARO” y si el día es “viernes” y además el cliente es adulto, de un total de 9 ventas 7 ventas son “POSTPAGO”, también se da un caso similar cuando el día es “jueves”, también se puede ver que si el día es “martes” y el género es “femenino” de un total de 18 ventas 13 ventas son en “POSTPAGO” y si el día es “lunes” de un total de 19 ventas 13 ventas son “POSTPAGO”, también se puede ver si el día es “miércoles” y es la primera quincena del mes “quincena1” de un total de 12 ventas 9 ventas son “POSTPAGO”, pero si el día es “sábado” y el cliente es próximo “P” al punto de venta de un total de 11 ventas 9 ventas son “POSTPAGO”; pero volviendo al nodo color, y si el color no es color “color claro” ni color “color oscuro” se tiene que de un total de 52 ventas 41 ventas son “POSTPAGO”. Volviendo al nodo: marca, dentro del nodo: tecnología del árbol, se puede ver si es que la marca es “MOVISTAR”, de un total de 33 ventas 32 ventas son en “PREPAGO”; lo contrario si la marca es “SONY”, de un total de 214 ventas 187 ventas son “POSTPAGO”; por el contrario, si la marca es “MOTOROLA” de un total de 246 ventas

216 ventas son en “POSTPAGO”; de lo contrario si la marca es “ZTE” de un total de 116 ventas 94 ventas son en “PREPAGO”.

→ Finalmente, volviendo al nodo principal del árbol: tecnología “tec”, se puede ver que, si la tecnología es “4G”, de un total de 1149 ventas 1013 ventas son “POSTPAGO”.

Finalmente, en tabla anterior Nº19 también se puede observar los indicadores que genera el algoritmo J48:

- ✓ Correctamente clasificados: 4178 instancias que representa el 71.9724 %.
- ✓ Incorrectamente clasificados: 1627 instancias que representan 28.0276 %.

Por lo tanto, se puede concluir que el algoritmo tiene un nivel de acierto de **71.9724 %,** que en aproximado seria **72**%. También se observó la matriz de confusión generada, en donde se puede observar que: 1170 ventas son clasificados correctamente como “POSTPAGO”,3003 ventas son clasificados correctamente como “PREPAGO” y 5 ventas son clasificados otra modalidad de venta “OTRO”. También se puede ver en la matriz de confusión que: 983 instancias fueran clasificadas como “PREPAGO “cuando en realidad eran “POSTPAGO”, 5 instancias fueran clasificadas como “OTRO” cuando eran “PREPAGO”, 6 instancias fueron clasificados como “OTRO” cuando eran “PREPAGO”. También se puede ver que: 458 instancias fueron clasificadas como “POSTPAGO” cuando eran “PREPAGO”, 63 instancias fueron clasificados como “POSTPAGO” cuando eran “PREPAGO”, 112 instancias fueron clasificados como “PREPAGO” cuando eran de otra modalidad “OTRO”

s. Revisión del proceso

Después de revisar nuevamente, la metodología usada, se puede indicar que se ha usado cada uno de sus fases y tareas de la metodología. Esto también ha sido posible gracias al uso de herramientas de software y la base teórica de la investigación en cada una de las fases y tareas de la metodología.

t. Determinación de las próximas etapas

Después de revisar los resultados por los involucrados en toma de decisiones en el área de ventas de la empresa, en este caso, representado por el personal que representa la población en estudio, se decidió pasar a la siguiente etapa que es la planificación de la implementación.

u. Planificación de la implementación

Después de presentar los resultados del modelo, quedo a cargo de los interesados en el área de ventas, en aplicar los resultados a sus planes de marketing y evaluar la automatización del modelo de minería.

v. Planificar el monitoreo y el mantenimiento

Para el mantenimiento y monitoreo del proceso de minería se designó un personal de informática por parte de la empresa, que se encargara de dar mantenimiento a las herramientas instaladas y verificar las tareas del proceso de minería.

w. Creación de un reporte final

A continuación, se presenta un resumen de los resultados más importantes del modelo de minería, después de analizar cuidadosamente los resultados presentados en el punto anterior: “r. Evaluación de Resultados” de la metodología CRISP-DM.

“Resumen de los resultados más importantes del algoritmo J48”

Es importante aclarar, que, para hacer el resumen de las ocurrencias más importantes, se ha tenido en cuenta la mayor cantidad de instancias que ocurren respecto a la variable de predicción o clase “modalidad”. Esto permite pronosticar en base al nivel de ocurrencias de la clase, condicionado por los diferentes valores que toman los atributos. Estos resultados son muy importantes debido a que le ha permitido a la empresa pronosticar en base a la experiencia de los datos que posee, basándose en los resultados que muestra el modelo. El resumen de los resultados se muestra a continuación:

- Se puede vender celulares con tecnología “2G”, de la marca “BMOBILE” con una probabilidad de 82.92% en modalidad de “PREPAGO”.
- Se puede vender celulares con tecnología “2G” de la marca “NOKIA” con una probabilidad de 70% en modalidad “PREPAGO”.
- Los clientes tienen una tendencia a comprar celulares con tecnología “2G” de la marca de la marca “SAMSUNG” que sean de colores oscuros con una probabilidad con una probabilidad de 78.7% en modalidad “PREPAGO”.
- La primera quincena “quincena1” o la segunda quincena “quincena2”, así como también los días de la semana, tienen una influencia mínima en la venta de celulares con tecnología “2G”, debido a la baja ocurrencia de casos, tanto en modalidad “PREPAGO” y “POSTPAGO”.
- Los celulares con tecnología “2G” de la marca “BLACKBERRY”, se venden más en “POSTPAGO” con una probabilidad de 89%
- Los celulares con la tecnología “2G” de la marca “HUAWEI”, se venden en “PREPAGO” con una probabilidad del 88%.
- Los celulares con la tecnología “2G” de la marca “ALCATEL”, se venden en “PREPAGO” con una probabilidad del 90.5%.

- ☞ Los celulares con la tecnología “2G” de la marca “LG”, se venden en “PREPAGO” con una probabilidad del 75% (42/14).
- ☞ Los celulares con la tecnología “2G” de la marca “MOVISTAR”, se venden en “PREPAGO” con una probabilidad del 92.8% (402/31).
- ☞ Los celulares con tecnología “2G” de las marcas “SONY” y “MOTOROLA” se venden más en “POSTPAGO”.
- ☞ Los celulares con la tecnología “2G” de la marca “ZTE”, se venden en “PREPAGO” con una probabilidad del 84.8% (279/50).
- ☞ Los celulares con tecnología “3G” de la marca “BMOBILE” se venden más en “PREPAGO” con una probabilidad de 80.2%(402/99).
- ☞ Los celulares con tecnología “3G” de la marca “NOKIA” se venden más en “POSTPAGO” con una probabilidad de 71.2%(960/389).
- ☞ Los celulares marca “SAMSUNG” con tecnología “3G” y de colores oscuros se venden más en “POSTPAGO” en la segunda quincena de cada mes con una probabilidad de 72.8%(158/59).
- ☞ Los celulares marca “SAMSUNG” con tecnología “3G” que no sean de colores claros u oscuros se venden más en “POSTPAGO” con una probabilidad de 72.8%(158/59).
- ☞ Los celulares marca “BLACKBERRY” con tecnología “3G” se venden más en “POSTPAGO” con una probabilidad de 86.6% (84/13).
- ☞ Los celulares de marca “HUAWEI” con tecnología “3G” se venden más en “PREPAGO” los días “viernes” con una probabilidad de 71.4%(25/10), mientras que en “POSTPAGO” se venden más los días “jueves” con una probabilidad de 67.7% (61/29).
- ☞ Los celulares de la marca “ALCATEL” con tecnología “3G” se venden más en “PREPAGO” con una probabilidad de 75% (712/233).
- ☞ Los celulares de marca “MICROSOFT” se venden más en “PREPAGO” a clientes próximos “P” con una probabilidad de 84.6% (33/6).
- ☞ Los celulares de la marca “LG” con tecnología “3G” se venden más en colores oscuros, los días “jueves” en modalidad “POSTPAGO”. También los celulares que no son de colores claros u oscuros se

vendieron en modalidad “POSTPAGO” con una probabilidad 79% (41/11).

- ☞ Los celulares de marca “MOVISTAR” con tecnología “3G” se venden más en “PREPAGO” con una probabilidad de 97% (32/1).
- ☞ Los celulares de la marca “SONY” se venden más en modalidad “POSTPAGO” con una probabilidad de 87% (187/27).
- ☞ Los celulares de la marca “MOTOROLA” se venden más en modalidad “POSTPAGO” con una probabilidad de 88% (216/30).
- ☞ Los celulares de la marca “ZTE” se venden más en “PREPAGO” con una probabilidad de 81% (94/22).
- ☞ Los celulares con tecnología “4G” se venden más en modalidad “POSTPAGO” con una probabilidad de 88% (1013/136).

x. Revisión del proyecto

En esta última tarea de la metodología, consiste en identificar y analizar los puntos que fueron bien realizados, los que fueron mal realizados, y los que podrían mejorarse. Para no redundar, todos estos puntos fueron mencionados en el CAPITULO V: Conclusiones y Recomendaciones, del presente documento.

3.2. Tratamiento y análisis de datos y presentación de resultados

En esta parte, se describe la manera en cómo se han obtenido, tratado y analizado los resultados, de la comparación del pronóstico del modelo con las ventas futuras, descritos en la tabla Nº 26 y, también al aplicar el instrumento de la encuesta antes y después de aplicar el modelo. Esto ha permitido medir el impacto del modelo de minería de datos y demostrar la hipótesis de la presente investigación.

Tabla 23: Comparación del pronóstico del modelo y las ventas futuras.

Predicción	Ventas registradas en periodos futuros (2017-2018)
Se puede vender celulares con tecnología “2G”, de la marca “BMOBILE” con una probabilidad de 82.92% en modalidad de “PREPAGO”.	La empresa ha vendido solo 3 marcas de celulares en la tecnología 2G. Debido a que la empresa pasó de ser operadora de “MOVISTAR” a ser operadora de “CLARO” a partir del año 2017, algunas marcas de celulares fueron descontinuados. Se vendieron celulares de la marca “AZUMI”,”HUAWEI”.”MOTOROLA” con tecnología “2G”, todos en modalidad “PREPAGO” (100%)
Se puede vender celulares con tecnología “2G” de la marca “NOKIA” con una probabilidad de 70% en modalidad “PREPAGO”.	
Los celulares con la tecnología “2G” de la marca “HUAWEI”, se venden en “PREPAGO” con una probabilidad del 88%.	
Los celulares con la tecnología “2G” de la marca “ALCATEL”, se venden en “PREPAGO” con una probabilidad del 90.5%.	
Los celulares con la tecnología “2G” de la marca “MOVISTAR”, se venden en “PREPAGO” con una probabilidad del 92.8%.	
Los celulares con la tecnología “2G” de la marca “LG”, se venden en “PREPAGO” con una probabilidad del 75%.	
Los celulares con la tecnología “2G” de la marca “ZTE”, se venden en “PREPAGO” con una probabilidad del 84.8%.	
Los clientes tienen una tendencia a	

comprar celulares con tecnología “2G” de la marca de la marca “SAMSUNG” que sean de colores oscuros con una probabilidad con una probabilidad de 78.7% en modalidad “PREPAGO”.	
Los celulares con tecnología “2G” de las marcas “SONY” y “MOTOROLA” se venden más en “POSTPAGO”.	No se registran venta de celulares de gama baja (2G), en plan “POSTPAGO”
Los celulares con tecnología “2G” de la marca “BLACKBERRY”, se venden más en “POSTPAGO” con una probabilidad de 89%	
Los celulares con tecnología “3G” de la marca “BMOBILE” se venden más en “PREPAGO” con una probabilidad de 80.2%.	No se registran ventas de la marca “BMOBILE” en tecnología “3G”. El producto esta descontinuado en este periodo
Los celulares de la marca “ALCATEL” con tecnología “3G” se venden más en “PREPAGO” con una probabilidad de 75%.	Todos los celulares de la marca “ALCATEL”, con tecnología “3G” fueron vendidos en “PREPAGO”
Los celulares de marca “HUAWEI” con tecnología “3G” se venden más en “PREPAGO” los días “viernes” con una probabilidad de 71.4% mientras que en “POSTPAGO” se venden más los días “jueves” con una probabilidad de 67.7%.	No se registran ventas de la marca “HUAWEI” en tecnología “3G”. El producto esta descontinuado en este periodo.
Los celulares de marca “MICROSOFT” se venden más en “PREPAGO” a clientes próximos “P” con una probabilidad de 84.6%.	No se registran ventas de la marca “MICROSOFT” en tecnología “3G”. El producto esta descontinuado en este periodo.

Los celulares de marca “MOVISTAR” con tecnología “3G” se venden más en “PREPAGO” con una probabilidad de 97%.	No se registran ventas de la marca “MOVISTAR” en tecnología “3G”. El producto esta descontinuado en este periodo.
Los celulares de la marca “ZTE” se venden más en “PREPAGO” con una probabilidad de 81%.	Todos los celulares de la marca “ZTE”, con tecnología “3G” fueron vendidos en “PREPAGO”.
Los celulares con tecnología “3G” de la marca “NOKIA” se venden más en “POSTPAGO” con una probabilidad de 71.2%.	No se registran ventas de la marca “SAMSUNG” en tecnología “3G”. El producto esta descontinuado en este periodo; sin embargo, se vendieron otras marcas “AZUMI” “LANIX” en colores oscuros en modo, todos en “POSTPAGO”
Los celulares marca “SAMSUNG” con tecnología “3G” y de colores oscuros se venden más en “POSTPAGO” en la segunda quincena de cada mes con una probabilidad de 72.8%.	
Los celulares marca “SAMSUNG” con tecnología “3G” que no sean de colores claros u oscuros se venden más en “POSTPAGO” con una probabilidad de 72.8%.	
Los celulares marca “BLACKBERRY” con tecnología “3G” se venden más en “POSTPAGO” con una probabilidad de 86.6%.	
Los celulares de la marca “SONY” se venden más en modalidad “POSTPAGO” con una probabilidad de 87%.	
Los celulares de la marca “MOTOROLA” se venden más en	

modalidad “POSTPAGO” con una probabilidad de 88%.	
Los celulares con tecnología “4G” se venden más en modalidad “POSTPAGO” con una probabilidad de 88%.	Se vendieron celulares de la marca “ZTE” con tecnología “4G”, de las cuales el 84% son en modalidad “POSTPAGO”. Se vendieron celulares de la marca “SAMSUNG” con tecnología “4G”, de las cuales el 77% son en modalidad “POSTPAGO”. Se vendieron celulares de la marca “LANIX” con tecnología “4G”, de las cuales el 83% son en modalidad “POSTPAGO”. Se vendieron celulares de la marca “AZUMI” con tecnología “4G”, de las cuales el 84% son en modalidad “POSTPAGO”. Se vendieron celulares de la marca “HUAWEI” con tecnología “4G”, de las cuales el 91% son en modalidad “POSTPAGO”. Se vendieron celulares de la marca “LG” con tecnología “4G”, de las cuales el 88% son en modalidad “POSTPAGO”. Se vendieron celulares de la marca “APPLE” con tecnología “4G”, de las cuales el 96% son en modalidad “POSTPAGO”. Se vendieron celulares de la marca “MOTOROLA” con tecnología “4G”, de las cuales el 98% son en modalidad “POSTPAGO”. Se vendieron celulares de la marca

	"IPHONE" con tecnología "4G", de las cuales el 100% son en modalidad "POSTPAGO".

Para ver el beneficio de la aplicación del modelo al momento de realizar pronósticos de ventas, se elaboró una tabla para comparar los recursos requeridos, antes y después de aplicar el modelo. Es importante indicar que, en la comparación se considera que la información esta adecuadamente almacenada, para que sea procesada en la construcción del modelo, que incluye la extracción, transformación y carga de datos (ETL) para el dataMart, exploración con WEKA y elaboración del reporte final del pronóstico. La comparación se realiza haciendo uso de los resultados de la tabla Nº 10, en el capítulo III: Materiales y métodos, donde se describe los recursos usados para hacer el pronóstico, inicialmente.

Tabla 24: Comparación de recursos usados para el pronóstico, antes y después de aplicar el modelo

Preparar la información	Medios de uso	Tiempo (antes)	Tiempo (después)
Reporte de ventas de campo	Se usaba hojas de cálculo MS Excel. Ahora se carga y se extrae datos del dataMart directamente.	2 semanas	1.5 minutos
Reporte de ventas de tienda	Se usaba hojas de cálculo y el sistema transaccional. Ahora se carga y se extrae datos del dataMart directamente.	1 semana	1.5 minutos
Disgregar la información, según los criterios de análisis (Local de venta, marca, generación, plan,	Se usaba hojas de cálculo MS Excel. Ahora se realiza filtros con consultas SQL o MDX, pre configuradas	1 semana	1 min

periodo, etc.)	en el dataMart, para los criterios de análisis		
Juntar y cruzar la información con el periodo anterior	Se usó hojas de cálculo MS Excel. Ahora, el algoritmo del modelo de minería, procesa los datos y presenta los resultados del pronostico	5 días aprox.	0.33 segundos aprox.

La medición de la variable modelo de minería, se realizó mediante la aplicación de una encuesta de conformidad de acuerdo al ANEXO Nº 02, a las 4 personas implicadas en el análisis de datos de ventas. El cuestionario está basado en los criterios de: Usabilidad y funcionalidad del modelo. Se consideró que una repuesta afirmativa es equivalente a 1 y una repuesta negativa equivalente a un peso de 0.

Tabla 25: Resultado de conformidad con el modelo de minería

(*)	Pregunta	Resp (0)	Resp (1)	total
1	¿El modelo describe de manera clara las características de las ventas y productos de equipos celulares, que interesan evaluar para realizar pronósticos?	0	4	4
2	El modelo de minera hace uso de toda los datos almacenados en el data Mart, y los procesa de manera rápida	0	4	4
3	¿Los resultados mostrados por el modelo es fácil de interpretar, para tomar decisiones en la venta de equipos celulares?	0	4	3
4	Las configuraciones de las herramientas de software usadas para genera el modelo son fáciles de realizar	2	2	2
5	La herramienta WEKA se conecta de manera fácil con la data Mart de ventas, para generar el modelo.	1	3	4
	Total	3	17	

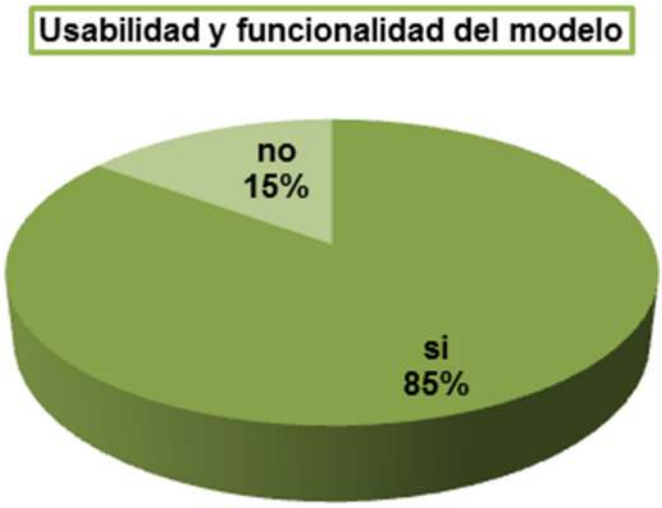


Fig. 75: Gráfica de nivel de conformidad con el modelo de minería

Para medir el comportamiento de la variable “Impacto en el pronóstico de ventas en la empresa CELLSERVICE”, se aplicó también una encuesta de conformidad descrito en el ANEXO Nº 01, a las 4 personas encargadas de tomar decisiones en el área de ventas, se ha puesto a cada pregunta un peso a las respuestas, se tomó como referencia la escala de Liker.

Tabla 26: Criterios de evaluación de la encuesta del Pre y Post -Test

CRITERIO	ABREVIATURA	PUNTAJE
Totalmente de acuerdo	**TDA**	**5**
De acuerdo	**DA**	**4**
Ni de acuerdo ni en desacuerdo	**NAD**	**3**
En desacuerdo	**ED**	**2**
Totalmente en desacuerdo	**TDD**	**1**

En las siguientes tablas Nº22 y Nº23 de los resultados del Pre – test y Post – test, “F.” representa la frecuencia para un determinado criterio, “Ẋ1” el promedio ponderado de la pregunta en el pre – test y “Ẋ2” el promedio ponderado de la pregunta en el post – test, respectivamente.

Tabla 27: Resultado del Pre - Test

		TDA	DA	NAD	ED	TDD	
(*)	Pregunta	F.	F.	F.	F.	F.	$\dot{X}1$
1	¿Usted conoce toda información tiene almacenada, de sus ventas, clientes, productos?	0	0	0	1	3	1,25
2	¿Analiza frecuentemente la información que posee?	0	0	1	2	1	2
3	¿En el proceso de análisis de la información hace uso de archivos en diferentes formatos?	1	3	0	0	0	4,25
4	La manera como se registra, almacena y consulta los datos de ventas, clientes y productos ¿permite encontrar la información que usted necesita?	0	0	2	2	0	2,5
5	¿Es complicado integrar datos históricos de ventas, clientes, productos para sus análisis?	1	2	1	0	0	4
6	¿La preparación de datos para el análisis requiere de poco recurso humano?	0	0	0	3	1	1,75
7	¿El análisis de datos de ventas, clientes, productos se realiza de manera rápida?	0	0	0	3	1	1,75
8	¿Tiene dificultades para ver comportamiento de datos de sus ventas?	1	1	2	0	0	3,75
9	¿Los resultados obtenidos de análisis son de mucha ayuda en la toma de decisiones?	0	0	3	1	0	2,75
10	¿Usted utiliza su información almacenada para realizar pronósticos?	0	0	0	3	1	1,75
11	¿Los pronósticos usados son fáciles de realizar?	0	0	0	1	3	1,25
12	¿Confiable realizar el pronóstico de ventas en base a la información que posee?	0	0	1	3	0	2,25
13	¿Los pronósticos de venta basada en la información que tiene le permiten minimizar errores en las estrategias de ventas?	0	0	1	3	0	2,25
14	El equipo de la fuerza de ventas establece estrategias basadas en pronósticos	0	0	0	4	0	2

Tabla 28: Resultado del Post - Test

		TDA	DA	NAD	ED	TDD	
(*)	Pregunta	F.	F.	F.	F.	F.	$\dot{X}2$
1	¿Usted conoce toda información tiene almacenada, de sus ventas, clientes, productos?	1	3	0	0	0	4,25
2	¿Analiza frecuentemente la información que posee?	2	2	0	0	0	4,5
3	¿En el proceso de análisis de la información hace uso de archivos en diferentes formatos?	0	0	0	4	0	2
4	La manera como se registra, almacena y consulta los datos de ventas, clientes y productos ¿permite encontrar la información que usted necesita?	2	2	0	0	0	4,5
5	¿Es complicado integrar datos históricos de ventas, clientes, productos para sus análisis?	0	0	1	3	0	2,25
6	¿La preparación de datos para el análisis requiere de poco recurso humano?	3	1	0	0	0	4,75
7	¿El análisis de datos de ventas, clientes, productos se realiza de manera rápida?	1	3	0	0	0	4,25
8	¿Tiene dificultades para ver comportamiento de datos de sus ventas?	0	0	1	3	0	2,25
9	¿Los resultados obtenidos de análisis son de mucha ayuda en la toma de decisiones?	3	1	0	0	0	4,75
10	¿Usted utiliza su información almacenada para realizar pronósticos?	1	3	0	0	0	4,25
11	¿Los pronósticos usados son fáciles de realizar?	0	3	1	0	0	3,75
12	¿Confiable realizar el pronóstico de ventas en base a la información que posee?	2	2	0	0	0	4,5
13	¿Los pronósticos de venta basada en la información que tiene le permiten minimizar errores en las estrategias de ventas?	2	2	0	0	0	4,5
14	El equipo de la fuerza de ventas establece estrategias basadas en pronósticos	2	1	1	0	0	4,25

- Diferencia de medias Pre – Test y Post – Test

Tabla 29: Resultado de diferencia de medias del Pre y Post - Test

Nº. Pregunta	Ẋ1: Pre - test	Ẋ2: Post - test	Ẋ2 - Ẋ1
1	1,25	4,25	3
2	2	4,5	2,5
3	4,25	2	-2,25
4	2,5	4,5	2
5	4	2,25	-1,75
6	1,75	4,75	3
7	1,75	4,25	2,5
8	3,75	2,25	-1,5
9	2,75	4,75	2
10	1,75	4,25	2,5
11	1,25	3,75	2,5
12	2,25	4,5	2,25
13	2,25	4,5	2,25
14	2	4,25	2,25
Total			**21.25**

Como se puede ver en la tabla Nº 24, el resultado de la diferencia de medias es de un valor positivo de **21.5** puntos.

CAPÍTULO IV. ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Después de ver los resultados de las pruebas del Pre – test y post test aplicados a la población, se ha pasado a demostrar la hipótesis, en la que se observa si modelo de minería de datos produce un impacto o no.

4.1. Análisis de resultados

Se ha considerado el tipo de la investigación, el material de estudio, la comprobación de la hipótesis por medio de la medición de las variables de estudio y la técnica de recolección de datos aplicados a los que toman decisiones en el área de ventas de la empresa. La población de estudio está constituida por la empresa “CELLSERVICE EIRL”, en la cual se ha desarrollado el modelo de minería de datos. Se realizó un muestreo no-probabilístico donde la elección de la muestra depende del criterio del investigador. La muestra está representada por las personas encargadas de tomar decisiones, para los procesos de ventas. La muestra está representada por:

- ✓ Gerente general: 1 persona.
- ✓ Gerente de Comercial: 1 persona.
- ✓ Control económico: 1persona.
- ✓ Control Interno: 1 persona.

4.1.1 Comprobación de la hipótesis

En el presente estudio se aplicó la Prueba de Hipótesis para dos muestras dependientes. En este caso se trata de dos muestras que contienen los mismos individuos en dos condiciones que se trata de diferenciar. Para ello se usará la siguiente formula:

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}}$$

Dónde:

$\bar{d}$: Es la media de las diferencias entre los valores de las muestras del pre – test y post – test.

S_d:Es la desviación estándar de las diferencias de medias.

n:Número de elementos o preguntas de las encuestas. $n = 14$

El valor de t, para comparar, se obtiene la tabla t – Student. En este caso se tiene que los grados de libertad es de $n - 1 = 13$. Para este estudio se consideró un nivel de significancia de $\alpha = 0.05$ o 5%. Por tanto $\alpha/2 = 0.025$

a. Formulación de la Hipótesis

- Hipótesis nula (H_0): El Modelo de Minería de Datos no tiene un impacto en el Pronóstico de Ventas de la Empresa CELL SERVICE E.I.R.L, en el periodo 2012 – 2016.

  H_0: $\mu_1 = \mu_2 \therefore \mu_1 - \mu_2 = 0$
- Hipótesis alternativa (H_1): El Modelo de Minería de Datos tiene un impacto en el Pronóstico de Ventas de la Empresa CELL SERVICE E.I.R.L, en el periodo 2012 – 2016.

  H_1: $\mu_1 \neq \mu_2 \therefore \mu_1 > \mu_2$ ó $\mu_1 < \mu_2$

b. Ubicación de la región critica

Para la región de la hipótesis nula o de rechazo se tiene que:

$$t: < -t_{\frac{\alpha}{2}} ; t_{\frac{\alpha}{2}} > =< -0.025; 0.025 >$$

$$\therefore t: < -2.16; 2.16 >$$

Para la región de aceptación de la hipótesis se tiene que:

$$t: < -\infty; -t_{\frac{\alpha}{2}}], [t_{\frac{\alpha}{2}}; \infty + > =< -\infty; -0.025], [0.025; \infty + >$$

$$\therefore t: < -\infty; -2.16], [2.16; \infty + >$$

c. Determinación y ubicación de valor esperado en la región crítica

Luego de aplicar la fórmula para el cálculo de$t$ sé que:

$$t = \frac{1{,}52}{1{,}85/\sqrt{13}} = \mp 3.08$$

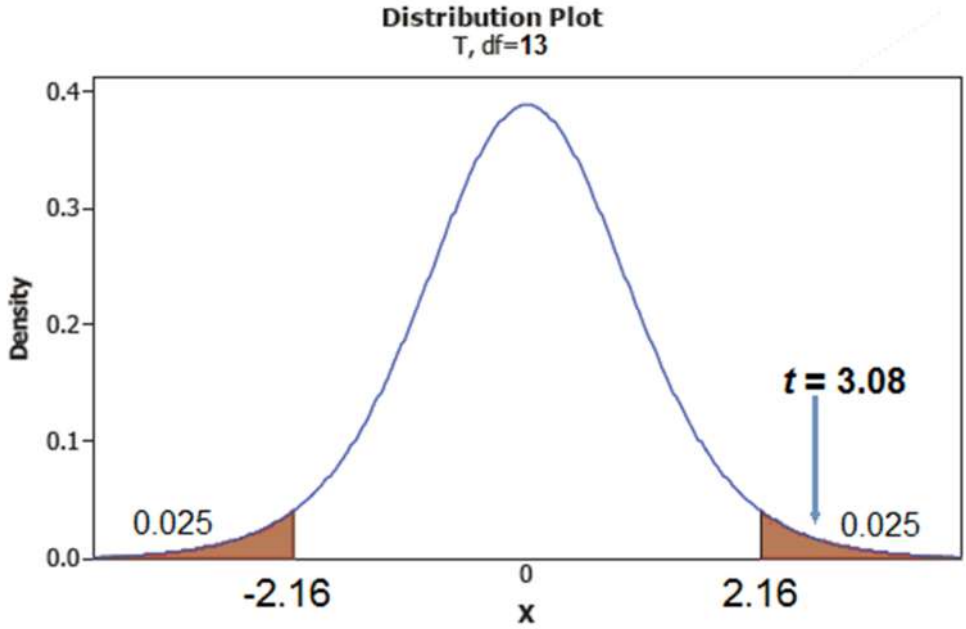


Fig. 76: Ubicación del valor esperado en la región crítica1

Luego para H_1: $\mu_1 \neq \mu_2 \ \therefore \ \mu_1 > \mu_2$ ó $\mu_1 < \mu_2$

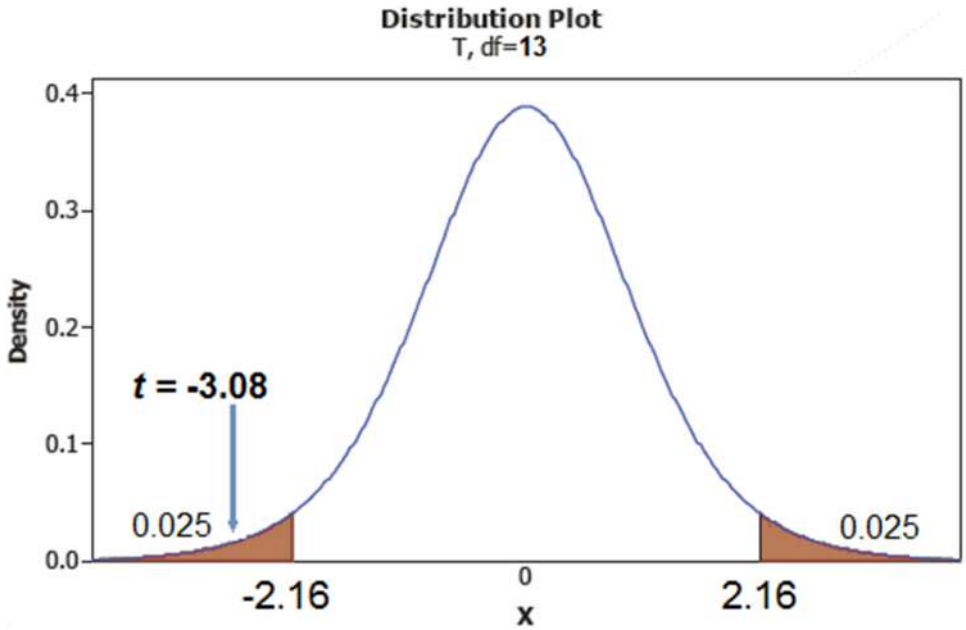


Fig. 77: Ubicación del valor esperado en la región crítica 2

d. Aceptación o rechazo de la hipótesis

Luego de ubicar el valor de $t = \pm 3.08$ en la curva se tiene que el valor es mayor a 2.16 y es menor que -2.16, por lo tanto se encuentra en la zona de aceptación. De esta manera se rechaza la hipótesis nula $H_{0:}$" El Modelo de Minería de Datos no tiene un impacto en el Pronóstico de Ventas de la Empresa CELL SERVICE E.I.R.L, en el periodo 2012 – 2016" y se acepta la hipótesis $H_{1:}$ "El Modelo de Minería de Datos tiene un impacto en el Pronóstico de Ventas de la Empresa CELL SERVICE E.I.R.L, en el periodo 2012 – 2016".

4.2. Discusión de resultados

Después de realizar las operaciones de cálculo con los datos recogidos de las encuestas del Pre – Test y Post - test, se logró demostrar que el modelo de minería de datos si produce un impacto, en este caso un impacto positivo, demostrado en los resultados de conformidad del Post – test. Esto se da porque, de acuerdo a los resultados descritos en la tabla Nº 22, el modelo pronostica con datos muy cercanos a los datos de ventas del periodo 2017 y lo que va del 2018, a pesar de que la empresa a descontinuado algunas marcas de celulares, de manera que le permite al personal que toma decisiones en el área de ventas, basarse en el modelo y pronosticar el comportamiento de sus ventas y mejorar las estrategias de marketing.

Evaluar la problemática ha permitido definir una forma para integrar los datos de ventas y aplicar la metodología de minería de datos CRISP – DM para construir el modelo.

Siguiendo la metodología del modelo, se ha tenido que entender los formatos de los datos almacenados en la base de datos del sistema de ventas transaccionales y las hojas de cálculo. La preparación de los datos ha consistido filtrar datos con errores de ingreso, datos redundantes, datos

vacíos que posteriormente se usaron para poblar la base de datos del data mart usando procesos ETL.

La construcción del data mart ha permitido realizar consultas y generar reportes de manera ágil en el análisis de ventas y generar el modelo de minería.

La evaluación del modelo de minería se realizó en función de los resultados mostrados por el algoritmo utilizado para generar modelo, descrito en el capítulo III. Materiales y métodos: q. “Evaluación del modelo”, finalmente se realizó un reporte de los resultados del modelo que fue revisado por los usuarios implicados en la toma decisiones respecto a las ventas y que fue aceptado como un modelo que les ayuda a mejorar el análisis y realizar pronósticos de los comportamientos de sus ventas. El reporte final de los resultados del modelo es presentado en el capítulo III. Materiales y métodos: W. “Creación de un reporte final”.

El presente trabajo a rescatado las experiencia de otros trabajos de minería de datos realizados como ocurre en “Análisis para la Predicción de Ventas utilizando Minería de Datos en almacenes de Ventas de grandes superficies” [*2*], que al igual que el presente trabajo, hace uso de herramientas de minería de datos de la categoría open source y que el modelo generado ha servido para realizar predicciones de ventas. En el trabajo “Análisis de patrones de compra de tiendas retail utilizando business intelligence” [*3*]”, de la misma manera se puede decir que la el análisis de datos en base de datos reales, permite encontrar información relevante al negocio de acuerdo al objetivo. En el trabajo “Algoritmos de minería de datos en la recolección de inteligencia” [*4*], también se utiliza el algoritmo de árbol de decisión, que permiten visualizar los atributos que determinan el comportamiento de los datos. En el trabajo “Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la modalidad abierta y a distancia de la UTPL”, de la misma manera, se hace uso de la metodología de minería de datos CRISP – DM y el modelo generado permitió determinar estrategias en la toma de

decisiones. En el trabajo “Propuesta de modelo de detección de fraudes de energía eléctrica en clientes residenciales de lima metropolitana aplicando minería de datos”, también realiza el proceso de minería de datos en información histórica y además se concluye que la preparación de los datos para construir el modelo es lo que más demora en aprox. 60% del tiempo total. Finalmente, en el trabajo “Desarrollo de un dataMart para mejorar la toma de decisiones en el área de ventas de la corporación furukawa” al igual que en el presente trabajo permitió mejorar el análisis de ventas, la generación de reportes y que este caso la base de datos multidimensional ha servido como fuente de datos para el modelo de minería.

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

- ✓ De acuerdo a los resultados descritos en la tabla Nº 22, el modelo pronostica con datos muy cercanos a los datos de ventas del periodo 2017 y lo que va del 2018, a pesar de que la empresa a descontinuado algunas marcas de celulares. Esto permite al personal que toma decisiones en el área de ventas, basarse en el modelo y pronosticar el comportamiento de sus ventas y mejorar las estrategias de marketing. Además, también, según los resultados en la tabla Nº 24, los recursos usados para hacer pronóstico, por el modelo, en cuanto a tiempo es muy optimo, se pasó de una escala de semanas o días a tan solo minutos o segundos.

- ✓ El modelo de minería generado, haciendo uso de datos almacenados del periodo del año 2012 al 2016 produce un impacto positivo en el pronóstico de ventas de la empresa “CellService”, de esta manera pudiendo mejorar las estrategias de marketing en la venta de equipos celulares. Tal como se pudo observar en el nivel de conformidad con el modelo y los pronósticos de ventas en la tabla de diferencia de medias del Pre y Post test en la “Tabla Nº 24”, teniendo un valor de 21.5 puntos de mayor conformidad, el cual fue ratificado mediante la demostración de la hipótesis con un valor de “t” igual 3.08.

- ✓ De la tabla Nº 22, también se puede decir, que, con el avance de la tecnología de los equipos celulares, el gusto de los clientes es también cambiante y las empresas buscaran alinearse a estas necesidades, pudiendo descontinuar algunas líneas de productos, lo que afectaría de alguna manera los resultados de los modelos de pronósticos de ventas basados en minería, que hacen uso de datos históricos.

- ✓ Se estudió la problemática del análisis de ventas de equipos celulares y la forma de aplicar pronósticos, que sirvió para seleccionar la metodología adecuada, en este caso, CRISP-DM y las herramientas de software libre,

para el proceso de minería, tales como: Mysql, Pentaho data Integration, Schema Workbench, JRubik, Weka.

- ✓ Comprender los datos de ventas almacenados, permitió ver las fuentes de datos que es el servidor de base de datos Mysql y archivos en MS Excel, describirlos, ver su calidad, que fueron tomados en cuenta para generar el modelo, siguiendo la metodología.

- ✓ La preparación de datos para la minería se alcanzó gracias al uso de filtros con sentencias SQL y herramientas de software para ETL como Pentaho Data integration. Los datos preparados sirvieron para poblar el dataMart de ventas.

- ✓ La elaboración de un Data Mart de ventas sirvió para explorar los 11610 registros de datos de ventas de equipos celulares, procesarlos y analizarlos, en grandes cantidades, de manera rápida, lo que permitió construir la estructura de minería de datos usada por el modelo.

- ✓ Se logró generar y evaluar del modelo de minería generado por la herramienta WEKA, eligiendo el modelo que tenía el mayor porcentaje de acierto, en este caso un total de 71.9724 %, lo que representa el grado de acierto de los pronósticos de las ventas de equipos celulares.

5.2 Recomendaciones

- ✓ El modelo generado es un caso particular, en la que se hace uso de datos propios de ventas, de una determinada empresa, para mejorar el análisis de ventas aplicando pronósticos; sin embargo, se pueden construir otros modelos, para casos similares o en otras problemáticas de análisis datos como en el sector financiero, ciencias de la salud, seguridad ciudadana, sector climático, entre otros.

- ✓ Un pronóstico será más concreto y confiable si se realiza haciendo uso de datos históricos reales apropiadamente almacenados.

- ✓ El procedimiento de la minería de datos, no necesariamente es igual para todos los casos, depende mucho de la problemática. Por ello es importante estudiar bien la situación y los datos que se quiere utilizar.

- ✓ Las herramientas de software usados en el presente trabajo fueron de la categoría open source; pero existen herramientas de la categoría comercial que pueden ayudar a realizar los procesos básicos de la minería de datos.

- ✓ Existen técnicas de minería de datos, aparte de los árboles de decisión, que son muy potentes y pueden ayudar a solucionar problemas de análisis y pronósticos más complicados.

- ✓ Las herramientas de software libre, son una buena alternativa para las empresas pymes, cuando quieren automatizar procesos, no necesariamente de minería de datos, ya que son de licencia gratuita; la inversión radicaría en la capacitación que tendrían que recibir para su desarrollo.

REFERENCIAS BIBLIOGRÁFICAS

[1] Usama Fayyad, Gregory Piatetsky Shapiro, and Padhraic Smyth. (1996, Noviembre) From Data Mining to Knowledge Discovery in Databases | Fayyad | AI Magazine. [Online]. http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230/1131

[2] José Antonio García Bermudes and Angela Maria Acevedo Ramirez. (2010) Análisis para predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies. [Online]. http://repositorio.utp.edu.co/dspace/bitstream/handle/11059/1339/006312G216.pdf?sequence=1&isAllowed=y

[3] Daniel Adasme Alarcón and Eduardo Salomón Díaz. (2013, Diciembre) Análisis de patrones de compra de tiendas retail utilizando business intelligence. [Online]. http://repositorio.uchile.cl/bitstream/handle/2250/115078/Adasme%20A.%2c%20Daniel.pdf?sequence=1&isAllowed=y

[4] Maria del Socorro Buendia Campos. (2014) ALGORITMOS DE MINERÍA DE DATOS EN LA RECOLECCIÓN DE INTELIGENCIA. [Online]. http://docplayer.es/4701046-Algoritmos-de-mineria-de-datos-en-la-recoleccion-de-inteligencia.html

[5] Briceño Ordoñez and Karla Fernanda. (2013) Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL. [Online]. http://dspace.utpl.edu.ec/bitstream/123456789/7897/1/Ordonez%20Brice%c3%b1o%20Karla-%20Informatica.pdf

[6] Coaguila Flores and Johanna Denise. (2014) Propuesta de modelo de detección de fraudes de energía eléctrica en clientes residenciales de Lima Metropolitana aplicando minería de datos. [Online]. http://www.repositorioacademico.usmp.edu.pe/bitstream/usmp/1266/1/flores_cjd.pdf

[7] Miguel Angel Berrospi Ramírez. (2013, Diciembre) Implantación de un sistema de ventas que emplea una herramienta de data mining. [Online].

http://tesis.pucp.edu.pe/repositorio/handle/123456789/5002

[8] Alex Jesús Durand Mendoza. (2014) Desarrollo de un datamart para mejorar la toma de decisiones en el área de ventas de la corporación Furukawa. [Online]. http://repositorio.untecs.edu.pe/handle/UNTELS/100

[9] Juan Miguel Moine, Ana Silvia Haedo, and Silvia Gordillo. (2011, Mayo) Estudio comparativo de metodologías para minería de datos. [Online]. http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1

[10] Juan Miguel Moine. (2013, Agosto) Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo. [Online]. http://sedici.unlp.edu.ar/bitstream/handle/10915/29582/Documento_completo.pdf?sequence=1

[11] kumar Gorakala suresh. (2015, octubre) Cross Industry Standard for Data Mining. [Online]. http://www.analyticbridge.com/profiles/blogs/cross-industry-standard-for-data-mining

[12] María N Moreno García, Luis A Miguel Quintales, Francisco J García Peñalvo, and M José Polo Martín. (2001) APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN LA CONSTRUCCIÓN Y VALIDACIÓN DE MODELOS PREDICTIVOS Y ASOCIATIVOS A PARTIR DE ESPECIFICACIONES DE REQUISITOS DE SOFTWARE. [Online]. http://www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf

[13] Jesús García Herrero and José Manuel Molina López. (2012) TÉCNICAS DE ANÁLISIS DE DATOS APLICACIONES PRÁCTICAS UTILIZANDO MICROSOFT EXCEL Y WEKA. [Online]. http://www.academia.edu/6078707/T%C3%89CNICAS_DE_AN%C3%81LISIS_DE_DATOS_APLICACIONES_PR%C3%81CTICAS_UTILIZANDO_MICROSOFT_EXCEL_Y_WEKA

[14] Microsoft Developer Network. (2017) Estructuras de minería de datos (Analysis Services). [Online]. https://msdn.microsoft.com/es-es/library/ms174757(SQL.90).aspx?tduid=(4a3847123dacd4bb274dbc74a89aa54e)(256380)(2459594)(TnL5HPStwNw-FmmO9OC.m3xOD7fPRHVcNQ)()#DataMiningStructure

[15] Willan J Staton, Michael J Etzel, and Bruce J Walker, *Fundamentos de Marketing*, Decimocuarta edición ed., Marcela I. Rocha Martínez, Ed. México: McGRAW-HILL/INTERAMERICANA EDITORES, S.A., 2007.

[16] BRUCE WALKER, WILLIAM J. STANTON, and MICHAEL J. ETZEL, "FUNDAMENTOS DE MARKETING 14ª ed.," in *FUNDAMENTOS DE MARKETING*.: Mc Graw Hill-Interamericana, 2007, p. 188.

[17] Rafael Augusto Florez Quintero. (2011) Pronóstico de Ventas Exitoso: ¿Cuantitativos o Cualitativos? [Online]. http://biblioteca.unitecnologica.edu.co/notas/tesis/0056227.pdf

[18] Jorge Humberto Frausto Enríques. (2009) Instituto Tecnológico y de Estudios Superiores de Monterrey. [Online]. https://repositorio.itesm.mx/bitstream/handle/11285/569469/DocsTec_10231.pdf?sequence=1&isAllowed=y

[19] Margaret Rouse. (2015) searchdatacenter.techtarget.com. [Online]. http://searchdatacenter.techtarget.com/es/definicion/MySQL

[20] ecured. Software libre. [Online]. https://www.ecured.cu/Pentaho_Data_Integration

[21] Sherman Wood. (2007) mondrian - pentaho. [Online]. https://mondrian.pentaho.com/documentation/workbench.php

[22] sourceforge.net. (2005) Rubik. [Online]. http://rubik.sourceforge.net/spanish/jrubik_es.html

[23] ecured. Weka. [Online]. https://www.ecured.cu/Weka

[24] Security Training Share. (2018) Best Open Source Data Mining Tools. [Online]. https://securityonline.info/8-best-open-source-data-mining-tools-weka-rapid-miner-orange-knime-jhepwork-apache-mahout-elki-rattle/

[25] Microsoft. (2018) Data Mining (SSAS). [Online]. https://msdn.microsoft.com/en-us/library/bb510516(v=sql.120).aspx

[26] SAS Institute. SAS® Enterprise Miner. [Online]. https://www.sas.com/en_us/software/enterprise-miner.html

[27] weka.sourceforge.net. weka.classifiers.rules. [Online]. http://weka.sourceforge.net/doc.dev/weka/classifiers/rules

[28] WordReference.com. (2018) Online Language Dictionaries. [Online]. http://www.wordreference.com/definicion/operatividad

[29] foromarketing. (2017) Análisis de venta. [Online]. http://www.foromarketing.com/diccionario/analisis-de-venta/

[30] crecenegocios. El proceso de marketing. [Online]. https://www.crecenegocios.com/el-proceso-de-marketing/

[31] Microsoft. (2017, Mar.) Modelos de mineria de datos. [Online]. https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/mining-models-analysis-services-data-mining

[32] DefiniciónABC. (2017-2018) Definición de Impacto. [Online]. https://www.definicionabc.com/general/impacto.php

[33] Gustavo R. Rivadera. (2010, Mayo) La metodología de Kimball para el diseño de almacenes de datos (Data warehouses). [Online]. www.ucasal.edu.ar/htm/ingenieria/cuadernos/./5-p56-rivadera-formateado.pdf

ANEXOS

ANEXO 1: Encuesta realizada a las 4 personas implicadas en la toma de decisiones el área de ventas:

ENCUESTA DE CONFORMIDAD CON EL PROCESO DE ANALISIS DE VENTAS

Objetivo de la encuesta.

El objetivo es medir el grado de conformidad del personal con el proceso de análisis de información de ventas.

Instrucciones.

Lea con cuidado cada una de las preguntas que se le hace a continuación y marque con un aspa (x) el casillero con la repuesta que a usted le parece apropiada.

1) ¿Usted conoce toda información tiene almacenada, de sus ventas, clientes, productos?

   ☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo

   ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

2) ¿Analiza frecuentemente la información que posee?

   ☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo

   ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

3) ¿En el proceso de análisis de la información hace uso de archivos en diferentes formatos?

   ☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo

   ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

4) La manera como se registra, almacena y consulta los datos de ventas, clientes y productos ¿permite encontrar la información que usted necesita?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo

ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

5) ¿Es complicado integrar datos históricos de ventas, clientes, productos para sus análisis?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo

ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

6) ¿La preparación de datos para el análisis requiere de poco recurso humano?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo

ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

7) ¿El análisis de datos de ventas, clientes, productos se realiza de manera rápida?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo

ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

8) ¿Tiene dificultades para ver comportamiento de datos de sus ventas?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo

ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

9) ¿Los resultados obtenidos de análisis son de mucha ayuda en la toma de decisiones?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

10) ¿Usted utiliza su información almacenada para realizar pronósticos?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

11) ¿Los pronósticos usados son fáciles de realizar?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

12) ¿Confiable realizar el pronóstico de ventas en base a la información que posee?

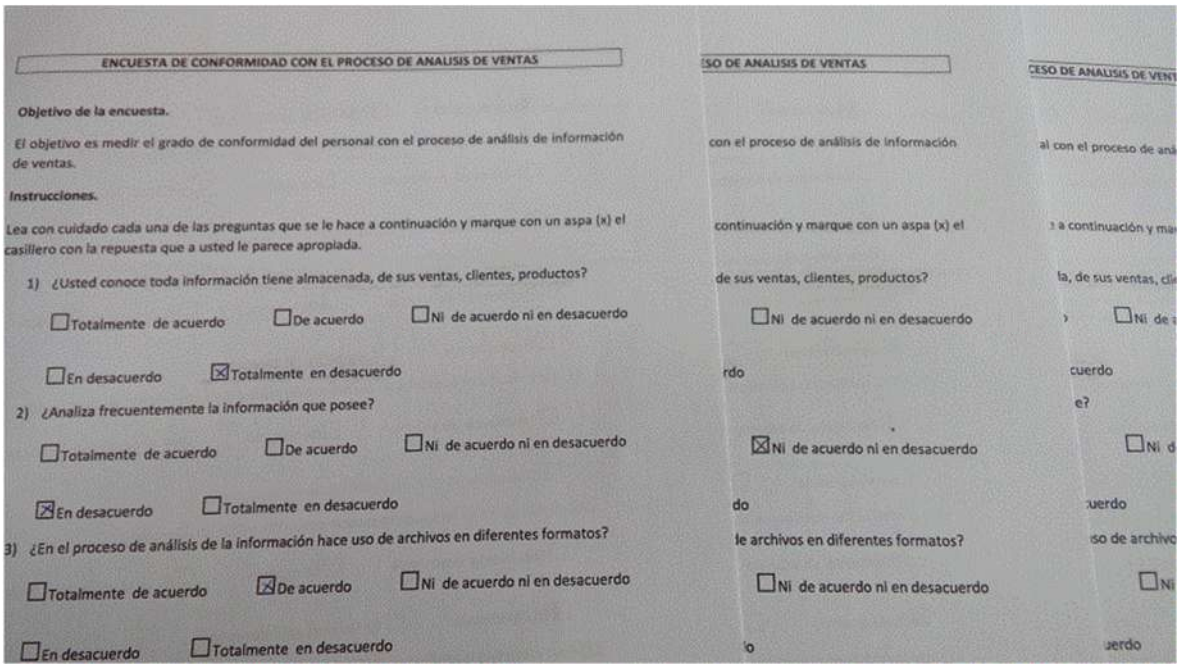
☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

13) ¿Los pronósticos de venta basada en la información que tiene le permiten minimizar errores en las estrategias de ventas?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

14) El equipo de la fuerza de ventas establece estrategias basadas en pronósticos

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo



ENCUESTA DE CONFORMIDAD CON EL PROCESO DE ANALISIS DE VENTAS

Objetivo de la encuesta.

El objetivo es medir el grado de conformidad del personal con el proceso de análisis de información de ventas.

Instrucciones.

Lea con cuidado cada una de las preguntas que se le hace a continuación y marque con un aspa (x) el casillero con la repuesta que a usted le parece apropiada.

1) ¿Usted conoce toda información tiene almacenada, de sus ventas, clientes, productos?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo ni en desacuerdo ☐En desacuerdo ☒Totalmente en desacuerdo

2) ¿Analiza frecuentemente la información que posee?

☐Totalmente de acuerdo ☐De acuerdo ☐Ni de acuerdo ni en desacuerdo ☒En desacuerdo ☐Totalmente en desacuerdo

3) ¿En el proceso de análisis de la información hace uso de archivos en diferentes formatos?

☐Totalmente de acuerdo ☒De acuerdo ☐Ni de acuerdo ni en desacuerdo ☐En desacuerdo ☐Totalmente en desacuerdo

ANEXO 2: Encuesta de conformidad de con el modelo de minería:

ENCUESTA DE CONFORMIDAD CON EL MODELO DE MINERIA DE DATOS

Objetivo de la encuesta.

El objetivo es medir la conformidad con el modelo de minería de datos.

Instrucciones.

Lea con cuidado cada una de las preguntas que se le hace a continuación y marque con un aspa (x) el casillero con la repuesta que a usted le parece apropiada.

1) ¿El modelo describe de manera clara las características de las ventas y productos de equipos celulares, que interesan evaluar para realizar pronósticos?

   ☐Si ☐No

2) ¿El modelo de minera hace uso de todos los datos almacenados en el data Mart, y los procesa de manera rápida?

   ☐Si ☐No

3) ¿Los resultados mostrados por el modelo es fácil de interpretar, para tomar decisiones en la venta de equipos celulares?

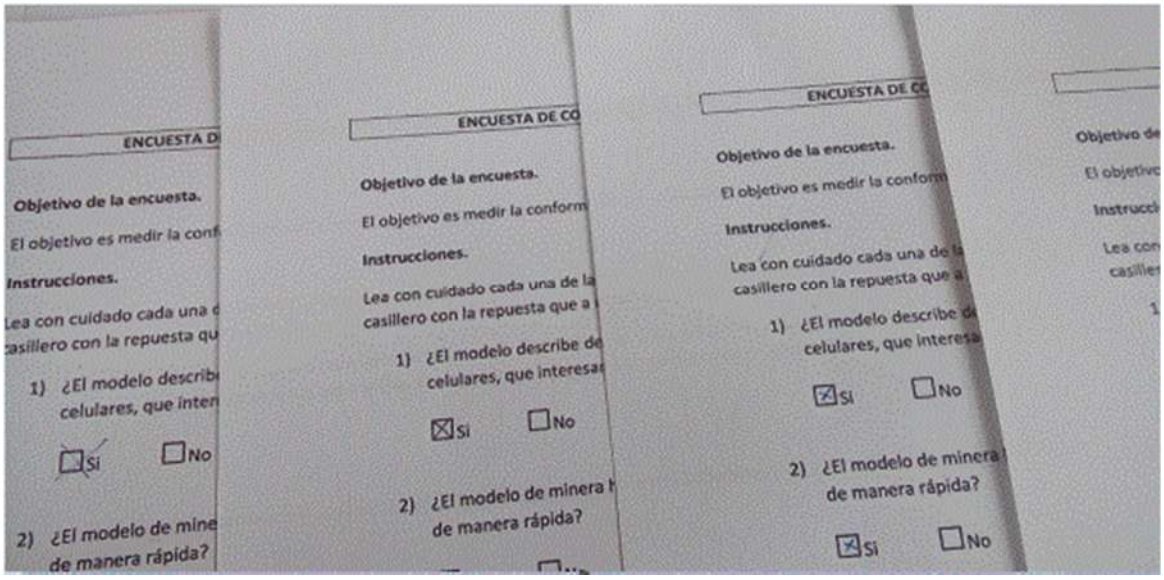
   ☐Si ☐No

4) ¿Las configuraciones de las herramientas de software usadas para genera el modelo son fáciles de realizar?

   ☐Si ☐No

5) ¿La herramienta WEKA se conecta de manera fácil con el data Mart de ventas, para generar el modelo?

   ☐Si ☐No

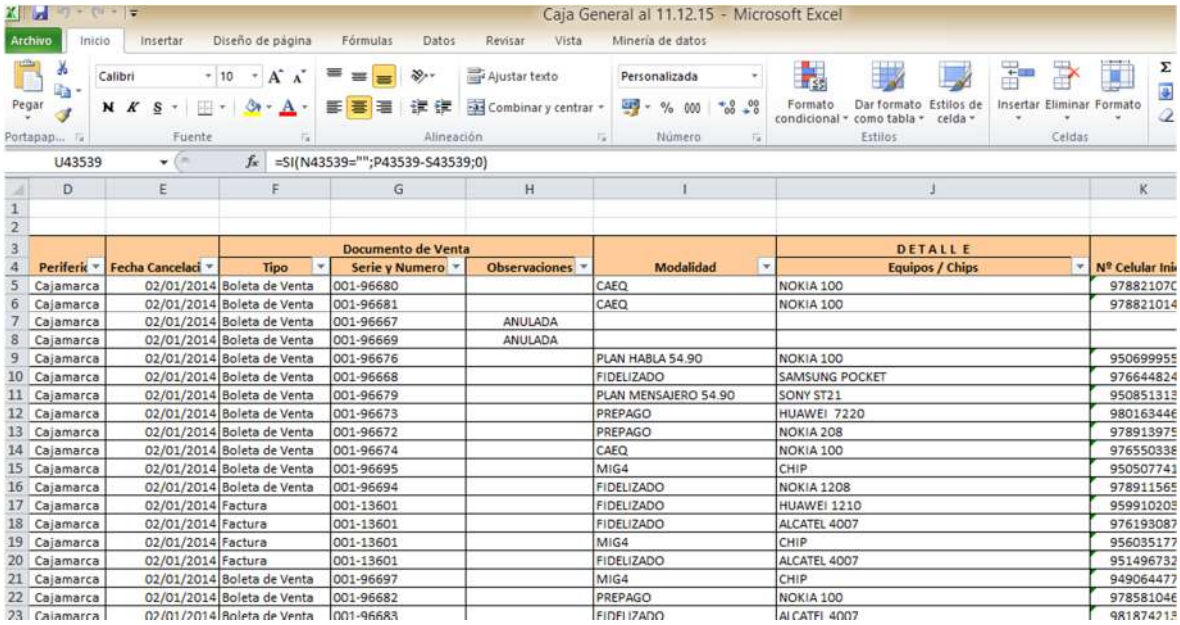


ANEXO 3: Espacios de trabajo y configuraciones del proceso de minería de datos:

i. Base de datos del sistema transaccional de ventas de equipos celulares

Tabla	Acción	Registros[1]	Tipo	Cotejamiento	Tamaño
activacion		230	MyISAM	latin1_swedish_ci	28.9 KB
area		0	MyISAM	latin1_swedish_ci	1.0 KB
cajapedido		1	MyISAM	latin1_swedish_ci	2.0 KB
cambio		3	MyISAM	latin1_swedish_ci	2.0 KB
cargo		16	MyISAM	latin1_swedish_ci	2.4 KB
categoria		0	MyISAM	latin1_swedish_ci	1.0 KB
cliente		23,211	MyISAM	latin1_swedish_ci	2.2 MB
cobranza		0	MyISAM	latin1_swedish_ci	1.0 KB
cobranzaseries		0	MyISAM	latin1_swedish_ci	1.0 KB
codigoequipo		21,404	MyISAM	latin1_swedish_ci	3.4 MB
comision		0	MyISAM	latin1_swedish_ci	1.0 KB
comprobante		340	MyISAM	latin1_swedish_ci	64.5 KB
consolidado		0	MyISAM	latin1_swedish_ci	1.0 KB
credito		1	MyISAM	latin1_swedish_ci	2.0 KB
ctipodocemisor		6	MyISAM	latin1_swedish_ci	2.2 KB
ctipodocidentificacion		6	MyISAM	latin1_swedish_ci	2.3 KB
ctiponota		14	MyISAM	latin1_swedish_ci	2.9 KB
cuenta		4	MyISAM	latin1_swedish_ci	2.2 KB
deposito		0	MyISAM	latin1_swedish_ci	1.0 KB

ii. Documentos de ventas en formato de MS Excel



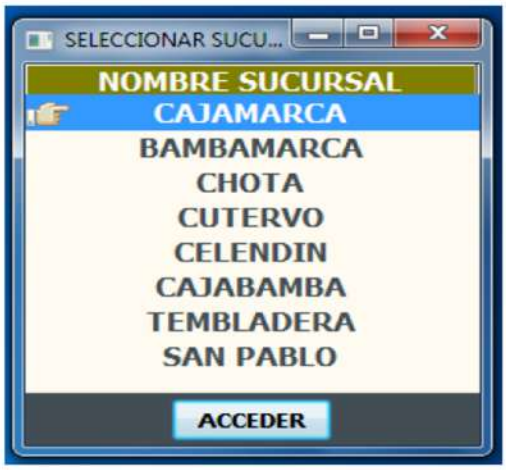
Caja General al 11.12.15 - Microsoft Excel

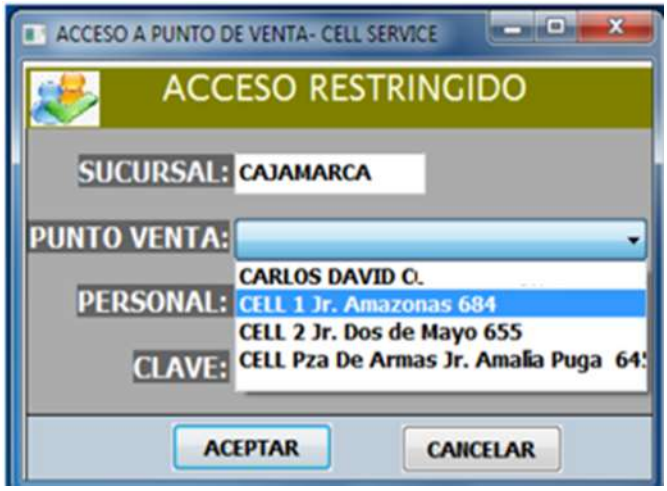
Archivo | Inicio | Insertar | Diseño de página | Fórmulas | Datos | Revisar | Vista | Minería de datos

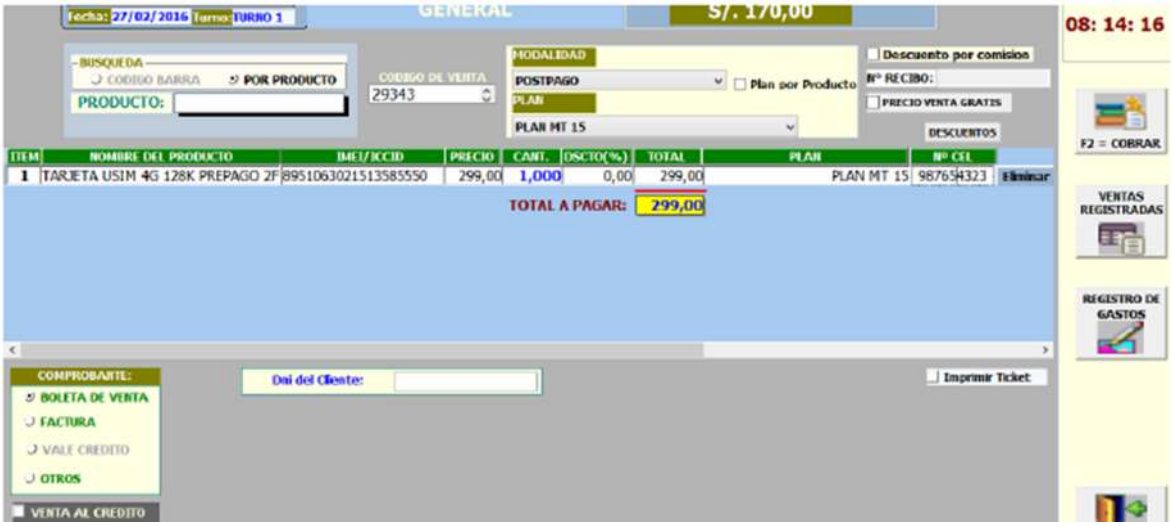
U43539 =SI(N43539="";P43539-S43539;0)

	Periferi	Fecha Cancelaci	Documento de Venta Tipo	Serie y Numero	Observaciones	Modalidad	DETALLE Equipos / Chips	Nº Celular Ini
5	Cajamarca	02/01/2014	Boleta de Venta	001-96680		CAEQ	NOKIA 100	978821070
6	Cajamarca	02/01/2014	Boleta de Venta	001-96681		CAEQ	NOKIA 100	978821014
7	Cajamarca	02/01/2014	Boleta de Venta	001-96667	ANULADA			
8	Cajamarca	02/01/2014	Boleta de Venta	001-96669	ANULADA			
9	Cajamarca	02/01/2014	Boleta de Venta	001-96676		PLAN HABLA 54.90	NOKIA 100	950699955
10	Cajamarca	02/01/2014	Boleta de Venta	001-96668		FIDELIZADO	SAMSUNG POCKET	976644824
11	Cajamarca	02/01/2014	Boleta de Venta	001-96679		PLAN MENSAJERO 54.90	SONY ST21	950851313
12	Cajamarca	02/01/2014	Boleta de Venta	001-96673		PREPAGO	HUAWEI 7220	980163446
13	Cajamarca	02/01/2014	Boleta de Venta	001-96672		PREPAGO	NOKIA 208	978913975
14	Cajamarca	02/01/2014	Boleta de Venta	001-96674		CAEQ	NOKIA 100	976550338
15	Cajamarca	02/01/2014	Boleta de Venta	001-96695		MIG4	CHIP	950507741
16	Cajamarca	02/01/2014	Boleta de Venta	001-96694		FIDELIZADO	NOKIA 1208	978911565
17	Cajamarca	02/01/2014	Factura	001-13601		FIDELIZADO	HUAWEI 1210	959910203
18	Cajamarca	02/01/2014	Factura	001-13601		FIDELIZADO	ALCATEL 4007	976193087
19	Cajamarca	02/01/2014	Factura	001-13601		MIG4	CHIP	956035177
20	Cajamarca	02/01/2014	Factura	001-13601		FIDELIZADO	ALCATEL 4007	951496732
21	Cajamarca	02/01/2014	Boleta de Venta	001-96697		MIG4	CHIP	949064477
22	Cajamarca	02/01/2014	Boleta de Venta	001-96682		PREPAGO	NOKIA 100	978581046
23	Cajamarca	02/01/2014	Boleta de Venta	001-96683		FIDELIZADO	ALCATEL 4007	981874215

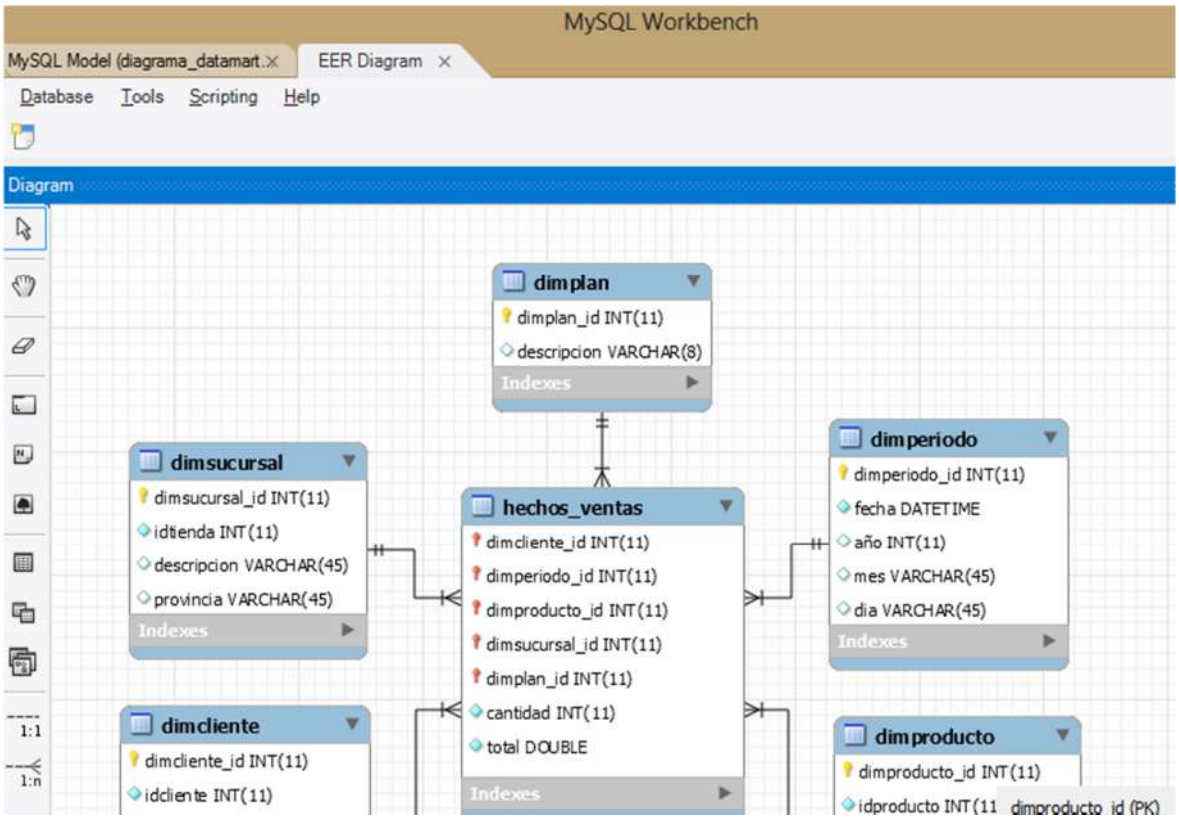
iii. Sistema de ventas de celulares





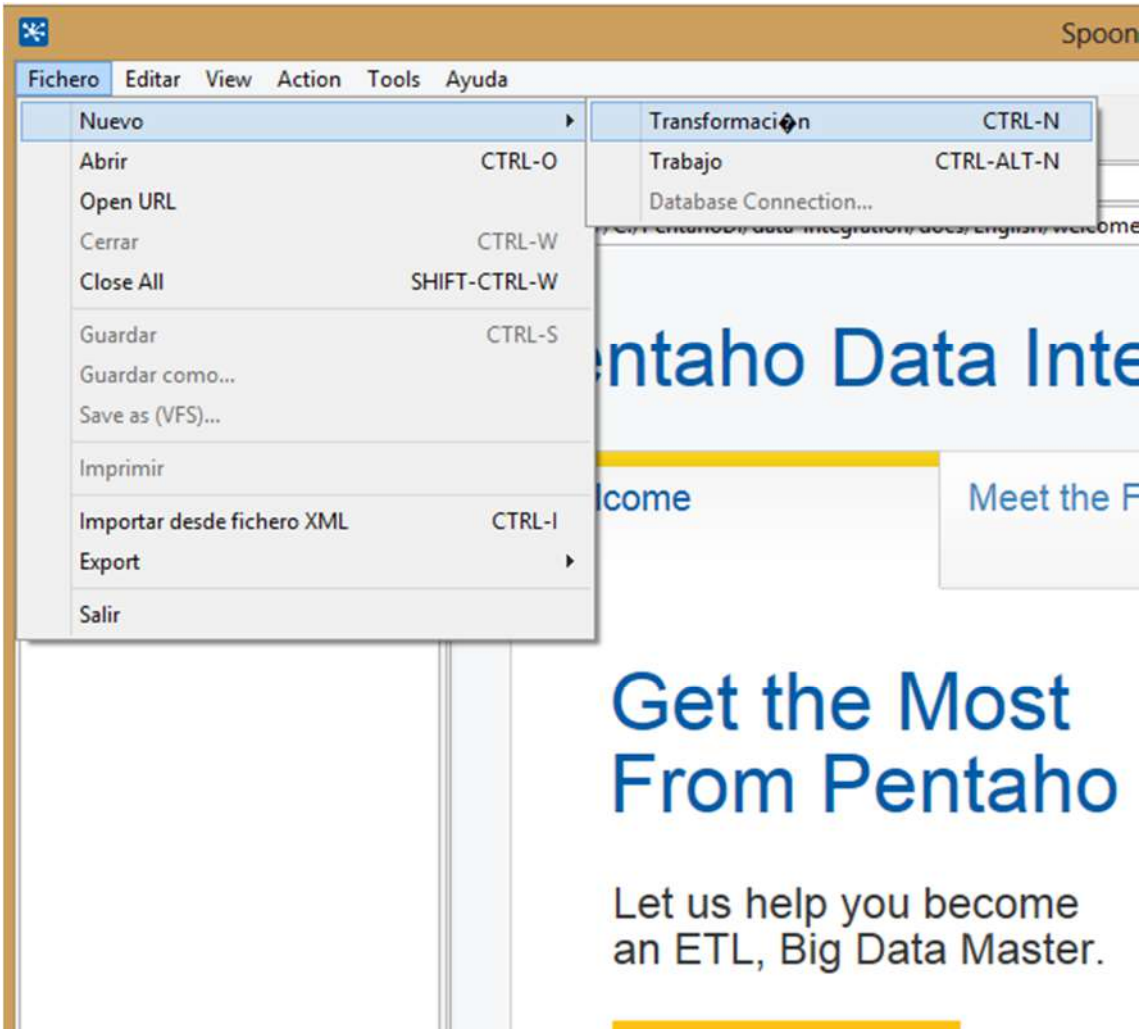


iv. Desarrollo del data mart de ventas



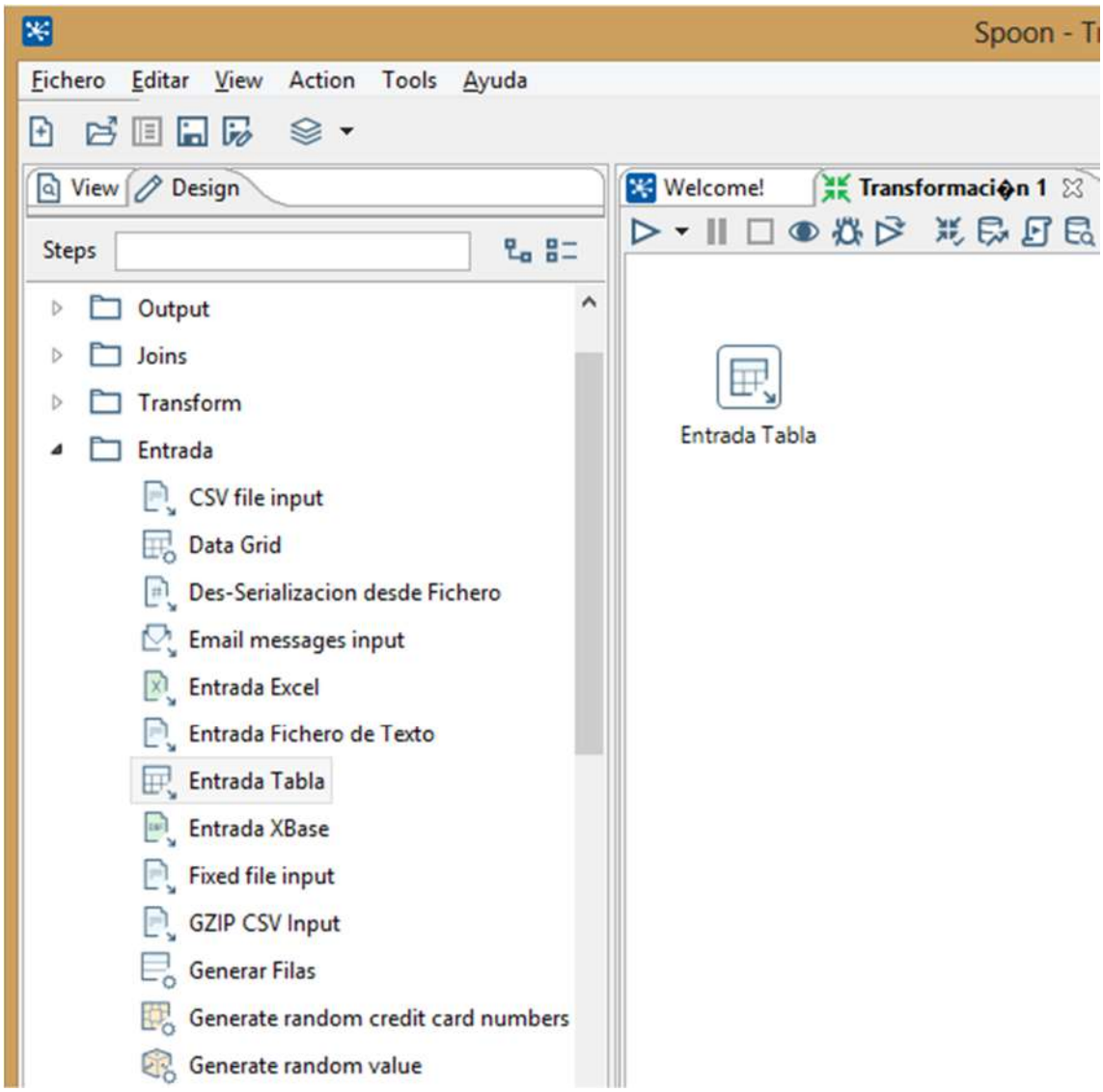
v. Configuración del proceso ETL para poblar el data mart

Generando el proceso ETL

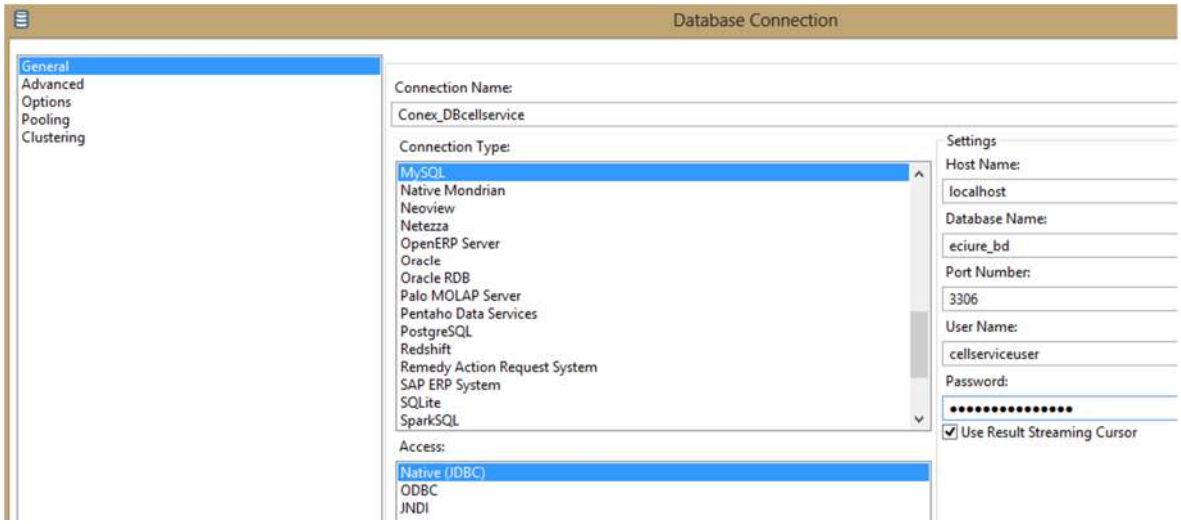


Nos ubicamos en la parte del diseño y seleccionamos los objetos que incluirán el paquete ETL.

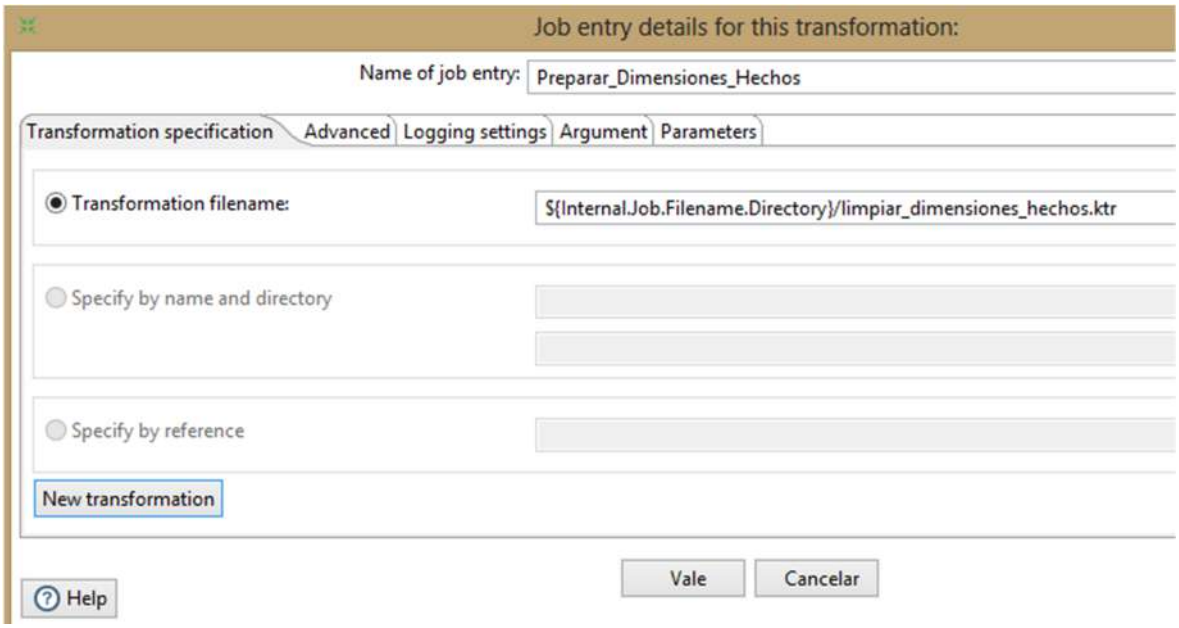
Como los datos se van a extraer directamente de la base de datos, seleccionamos la categoría “Entrada” y seleccionamos el objeto “Entrada Tabla”.

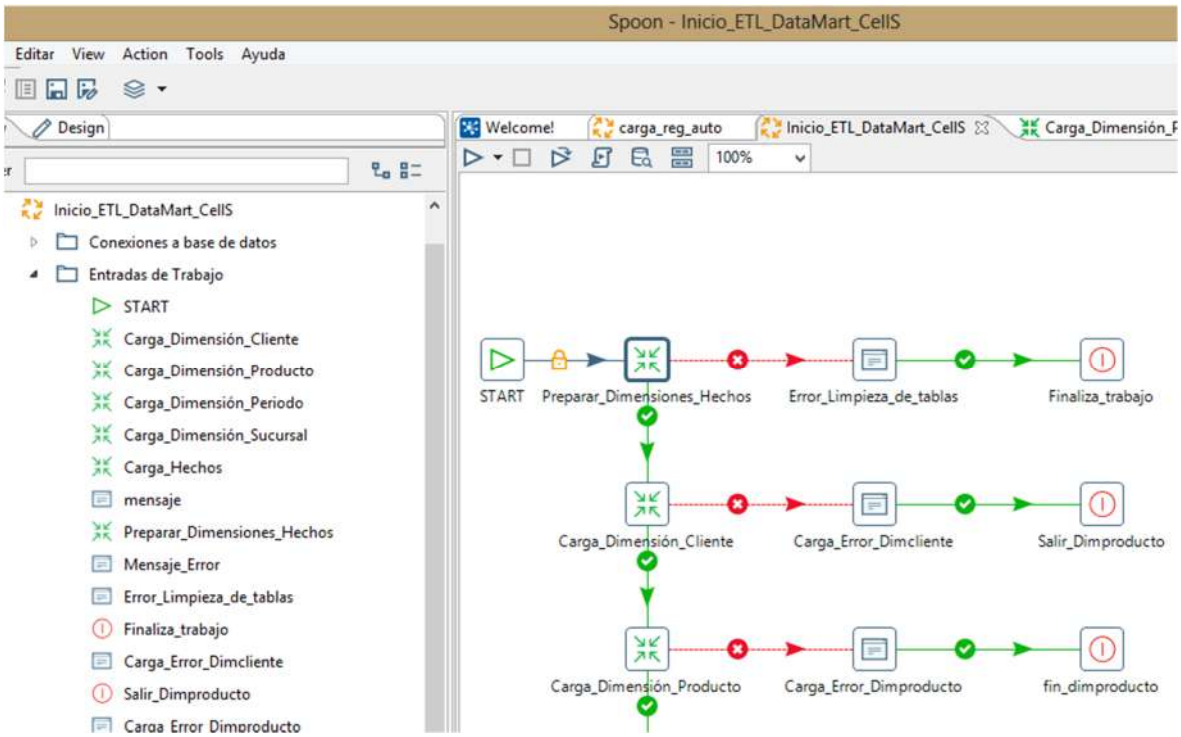


El objetivo del proceso ETL es poblar el dataMart, comenzaremos definiendo el origen de datos para cada uno de las dimensiones y de la tabla hechos.

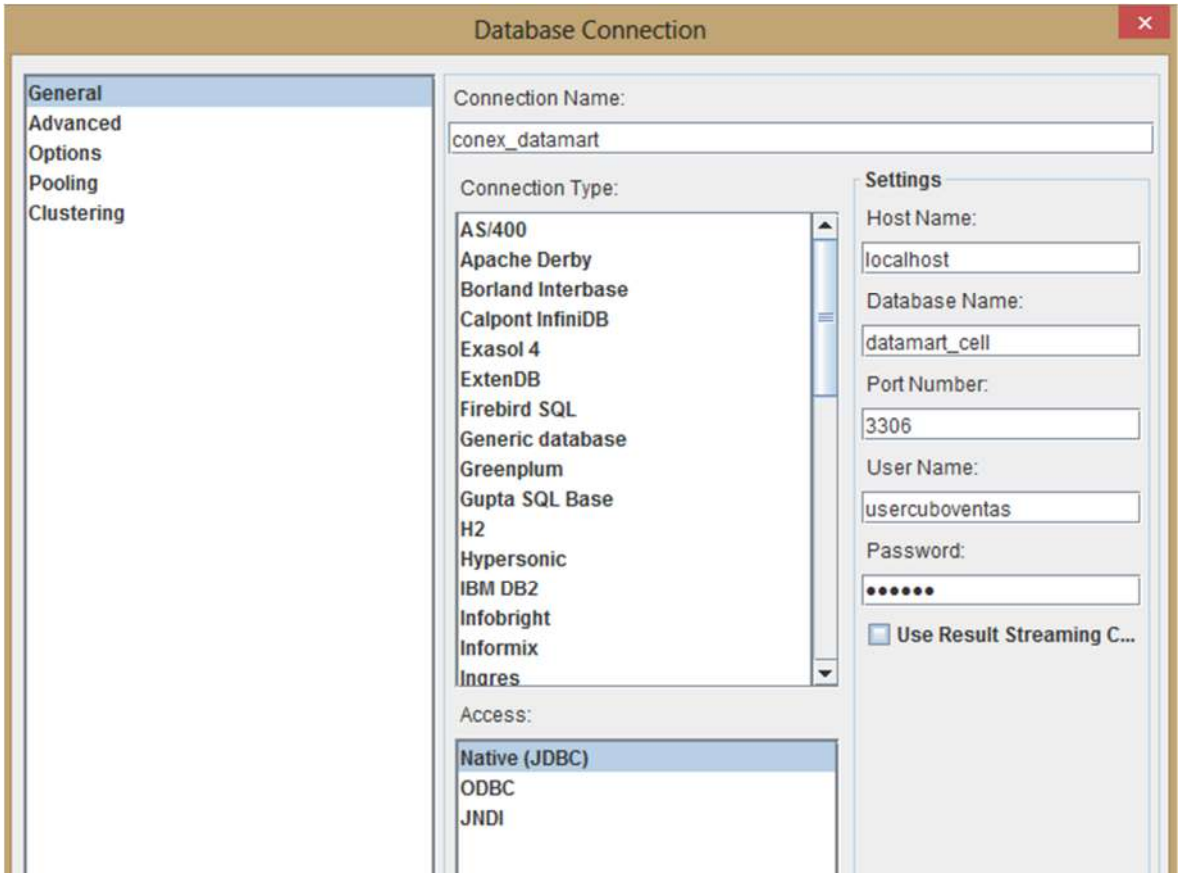


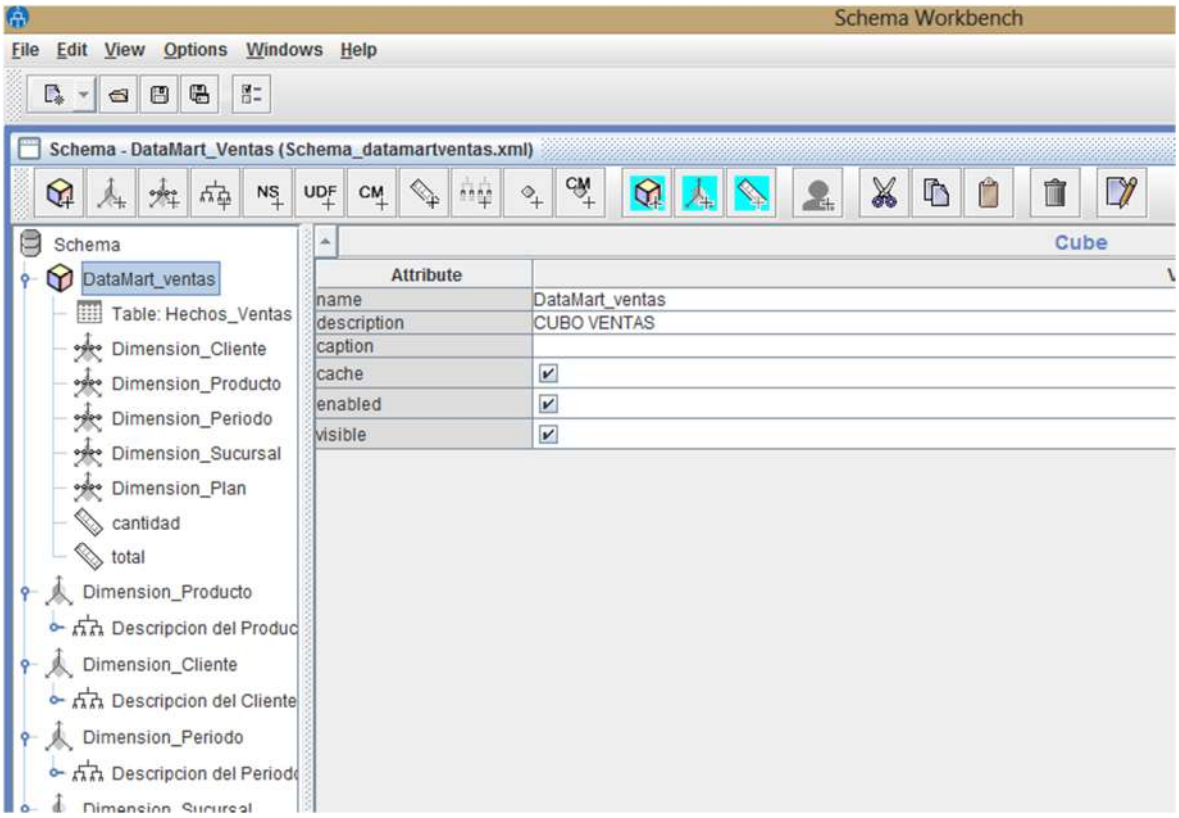
#	idproducto	nombre	generacion	tipo	marca	modelo	color
1	1	TERMINAL GSM LG A200 NEGRO GRIS	2G	CELULAR	LG	A200	OTRO COLOR
2	3	TERMINAL 3G MOVISTAR NEON AZUL	3G	CELULAR	MOVISTAR	NEON	AZUL
3	4	TERMINAL GSM HUAWEI G7206 NEGRO	2G	CELULAR	HUAWEI	G7206	NEGRO
4	5	TERMINAL GSM HUAWEI G7206 BLANCO	2G	CELULAR	HUAWEI	G7206	BLANCO
5	6	TERMINAL GSM HUAWEI G3512 NEGRO NARANJA	2G	CELULAR	HUAWEI	G3512	OTRO COLOR
6	7	TERMINAL GSM NOKIA 111 NEGRO	2G	CELULAR	NOKIA	111	NEGRO
7	8	TERMINAL GSM NOKIA 306 BLANCO	2G	CELULAR	NOKIA	306	BLANCO
8	9	TERMINAL GSM HUAWEI G2800S NEGRO	2G	CELULAR	HUAWEI	G2800S	NEGRO
9	10	TERMINAL GSM NOKIA 100 NEGRO	2G	CELULAR	NOKIA	100	NEGRO
10	11	TERMINAL GSM LG A200 NEGRO VIOLETA	2G	CELULAR	LG	A200	OTRO COLOR
11	12	TERMINAL GSM MOTOROLA EX116 GRIS	2G	CELULAR	MOTOROLA	EX116	GRIS
12	13	TERMINAL GSM NOKIA 100 AZUL	2G	CELULAR	NOKIA	100	AZUL
13	14	TERMINAL GSM BLACKBERRY 8520 NEGRO	2G	CELULAR	BLACKBERRY	8520	NEGRO
14	15	TERMINAL GSM MOVISTAR ONDA ROJO	2G	CELULAR	MOVISTAR	ONDA	ROJO
15	16	TERMINAL GSM MOVISTAR ONDA AZUL	2G	CELULAR	MOVISTAR	ONDA	AZUL
16	17	TERMINAL GSM MOVISTAR ONDA AMARILLO	2G	CELULAR	MOVISTAR	ONDA	AMARILLO
17	18	TERMINAL 3G MOTOROLA A953 NEGRO	3G	CELULAR	MOTOROLA	A953	NEGRO
18	20	TERMINAL GSM 3G HUAWEI E173 MORADO	3G	CELULAR	HUAWEI	E173	OTRO COLOR
19	21	TERMINAL MODEM 3G HUAWEI E303C BLANCO	3G	CELULAR	HUAWEI	E303C	BLANCO
20	22	TERMINAL MODEM 3G HUAWEI E303C NEGRO	3G	CELULAR	HUAWEI	E303C	NEGRO
21	23	TERMINAL GSM SAMSUNG E2220 GRIS TEXTO	2G	CELULAR	SAMSUNG	S2220	GRIS
22	24	TERMINAL GSM SAMSUNG E2220 NEGRO	2G	CELULAR	SAMSUNG	S2220	NEGRO
23	25	TERMINAL GSM SAMSUNG E2230 NEGRO	2G	CELULAR	SAMSUNG	E2230	NEGRO



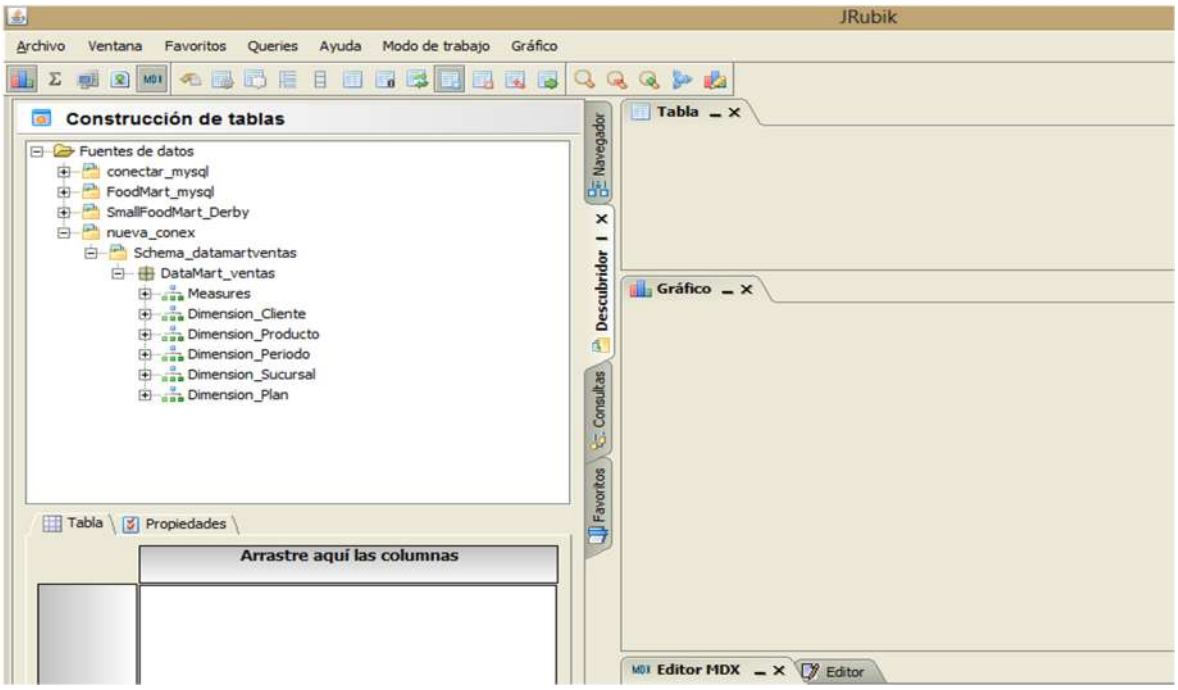


vi. Diseño del cubo del data mart de ventas con schemaworkbench

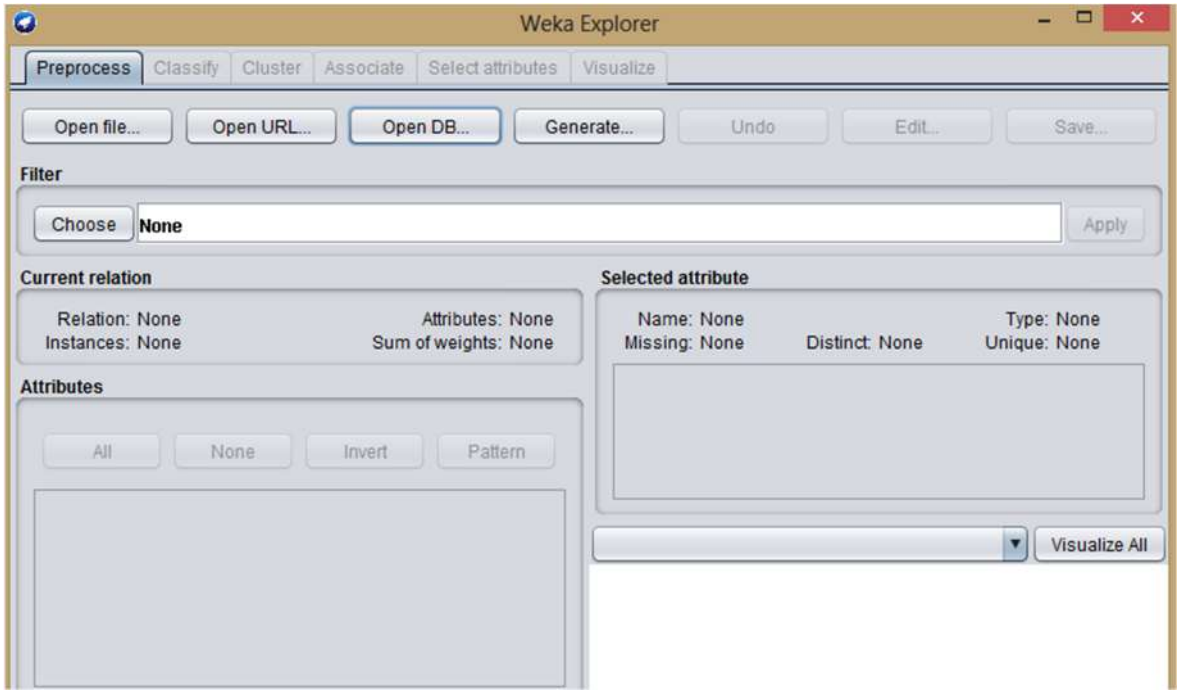




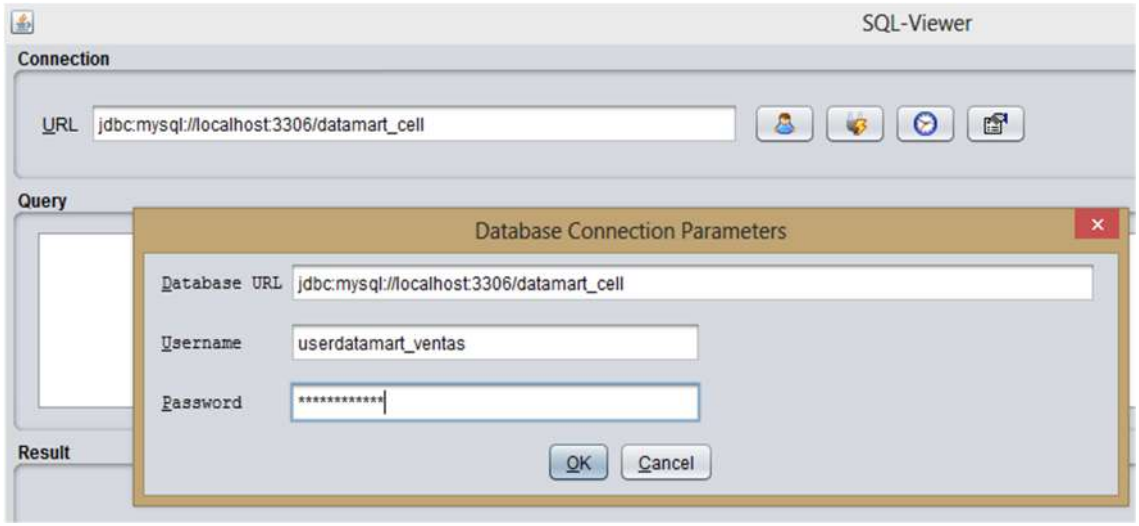
vii. Exploración de datos del data mart con Jrubik



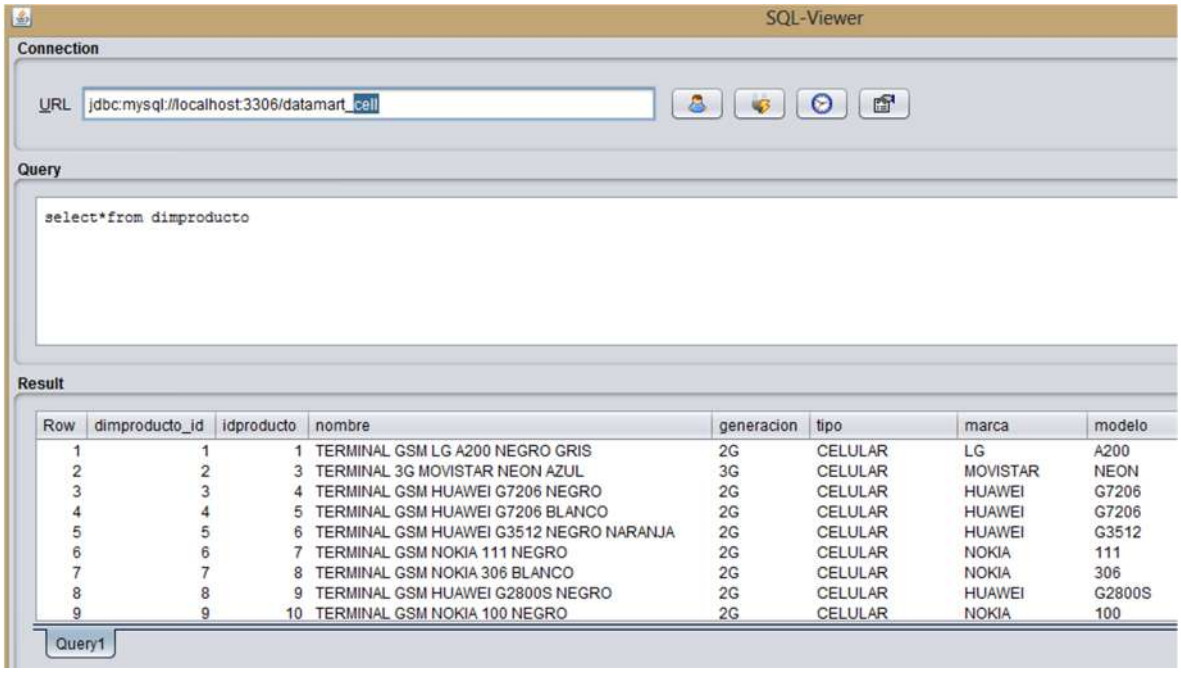
viii. Espacio de trabajo de la herramienta de minería de datos WEKA



Conectando la herramienta weka con el data mart de ventas

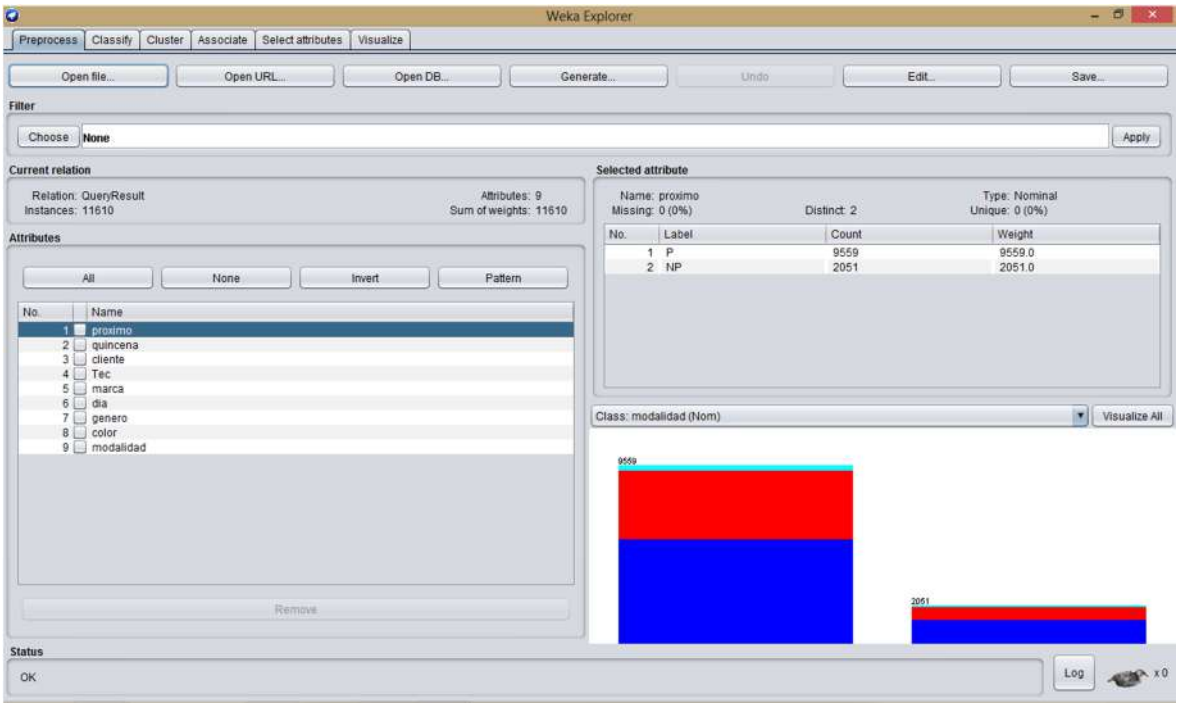


Exploración de tablas de las dimensiones del data mart

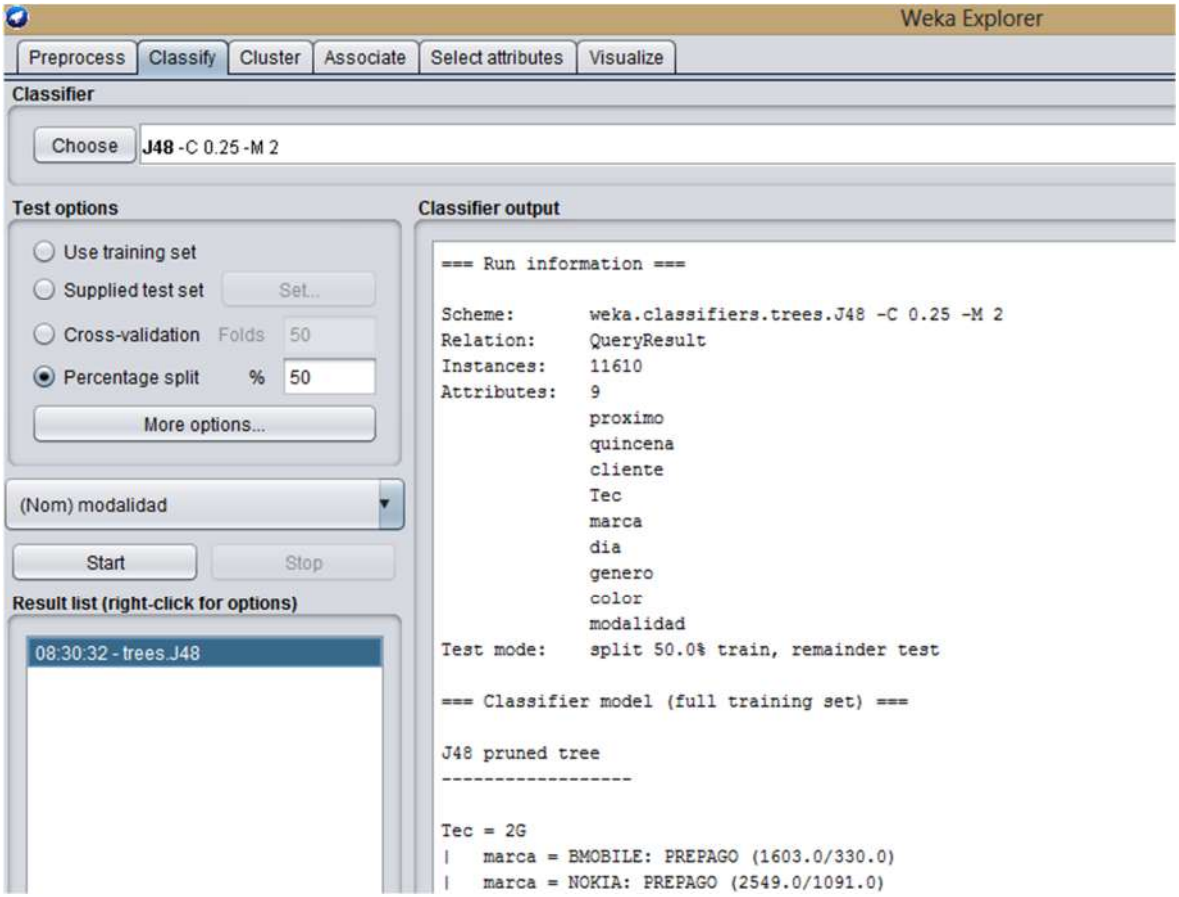


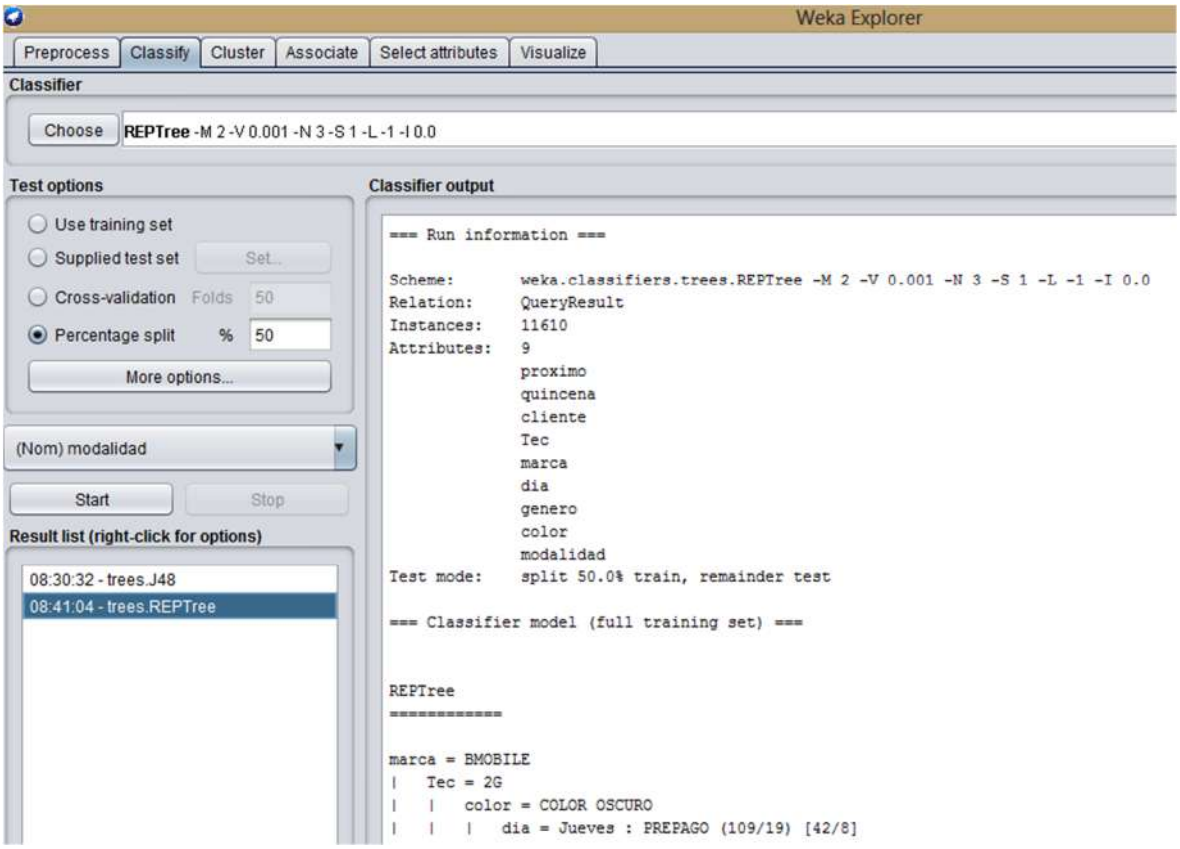
Row	dimproducto_id	idproducto	nombre	generacion	tipo	marca	modelo
1	1	1	TERMINAL GSM LG A200 NEGRO GRIS	2G	CELULAR	LG	A200
2	2	3	TERMINAL 3G MOVISTAR NEON AZUL	3G	CELULAR	MOVISTAR	NEON
3	3	4	TERMINAL GSM HUAWEI G7206 NEGRO	2G	CELULAR	HUAWEI	G7206
4	4	5	TERMINAL GSM HUAWEI G7206 BLANCO	2G	CELULAR	HUAWEI	G7206
5	5	6	TERMINAL GSM HUAWEI G3512 NEGRO NARANJA	2G	CELULAR	HUAWEI	G3512
6	6	7	TERMINAL GSM NOKIA 111 NEGRO	2G	CELULAR	NOKIA	111
7	7	8	TERMINAL GSM NOKIA 306 BLANCO	2G	CELULAR	NOKIA	306
8	8	9	TERMINAL GSM HUAWEI G2800S NEGRO	2G	CELULAR	HUAWEI	G2800S
9	9	10	TERMINAL GSM NOKIA 100 NEGRO	2G	CELULAR	NOKIA	100

Configuración de la tabla de la estructura de minería de datos para el modelo en Weka, con la variable o atributo de predicción “modalidad”:

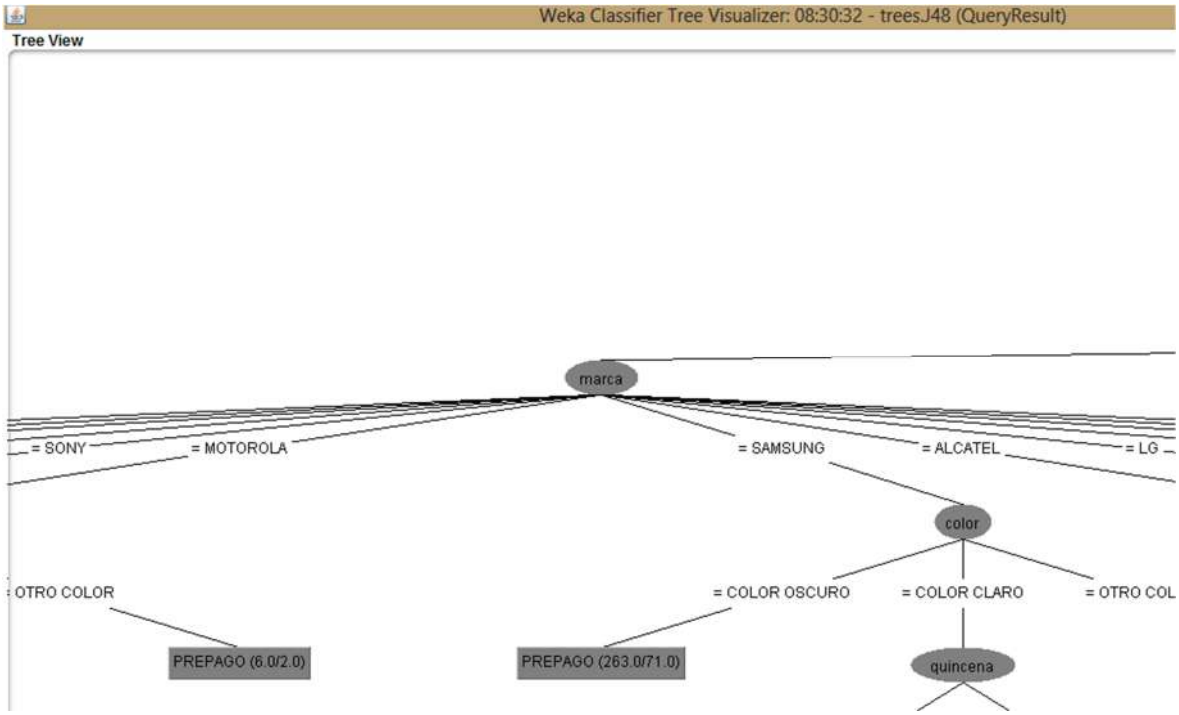


Pruebas del modelo con los algoritmos de Weka:





Ejemplo de la vista del árbol generado por el algoritmo de árbol de decisión en Weka:



Exploración de ventas del periodo 2017-2018, para comparar con los resultados del modelo.