PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

Implantación de un Sistema de Ventas que emplea una herramienta de Data Mining

Tesis para optar por el Título de Ingeniero Informático, que presenta el bachiller:

Miguel Angel Berrospi Ramirez

ASESOR: Luis Flores

Lima, Diciembre del 2012

Índice

Índice de Figuras

Índice de Tablas

Resumen

El proyecto que se presenta en este documento 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 lo antes mencionado porque la empresa a la que se aplicará 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 todo un periodo de tiempo.

Para que el objetivo final del proyecto se cumpla, se usaron herramientas de software, herramientas de planificación y de organización, estas últimas se usaron porque son herramientas estandarizadas y aceptadas internacionalmente en sus respectivos campos; además, sirvieron eficientemente para su propósito porque son una guía de pasos detalladas y específicas para cada actividad que se necesitaba en el proyecto. Con respecto a las herramientas software usadas, estas fueron seleccionadas mediante una comparación de criterios, las cuales eran necesarias por los requerimientos y necesidades planteadas en la justificación y viabilidad del proyecto.

En conclusión, el proyecto se llevó a cabo con éxito previniendo los efectos negativos o eventos inoportunos que puedan generarse durante su ejecución mediante un plan de riesgos ya incluido previamente en la planificación. Esta planificación y el planteamiento de objetivos generales y específicos con sus respectivos métodos y actividades, ayudaron a mantener una idea clara y concisa de lo que se pretendía realizar desde los inicios del proyecto.

1. Generalidades

El sector textil de venta de prendas de vestir es uno de los que más se ha impulsado en los últimos 10 años en el país, esto debido a la apertura de nuevos mercados y a la ruptura de brechas económicas, sociales y culturales. (PINTO CASTRO, 2007)

Esto ha traído consigo una fuerte competencia; a la par se han abierto nuevos mercados internacionales, nuevos nichos de mercado; es decir, nuevas oportunidades de negocio; sin embargo, no toda la industria textil ha sabido aprovecharlas, esto se debe a distintos factores tales como marketing, tecnología, mejora de procesos, etc. (PINTO CASTRO, 2007)

A causa de la expansión del sector textil de prendas de vestir, se originaron muchas organizaciones emprendedoras, muchas de ellas Pymes, las cuales se dedican a nichos de mercado específicos. (PINTO CASTRO, 2007)

Las Pymes que se enfocan a nichos específicos de mercado, solo aplican técnicas de marketing y ventas para atraer la atención de los clientes; estos últimos buscan los mejores precios y calidad; además buscan productos de acuerdo a la estación del año, entre otros. Es decir, los clientes se rigen en base a distintos aspectos para realizar una compra; aspectos que cada empresa debería tomar en cuenta para poder fidelizar a sus clientes y captar nuevos clientes.

1.1 Definición de la problemática

El uso de tecnología para el rubro empresarial textil de confección y ventas de prendas de vestir se ha vuelto una necesidad, ya que se requiere de una mejor comunicación entre los procesos críticos de la empresa (procesos de producción, procesos de Logística, procesos de ventas, procesos de compras, entre otros); además, el comportamiento del mercado textil ha cambiado, en la actualidad se necesita llegar a cada cliente de forma personalizada; también, se necesita un modelo de procesos de compra y logística sencillo para manejar las cantidades enormes de productos. Los sistemas de información brindan herramientas para agilizar los procesos descritos anteriormente; sin embargo, no todos los procesos de un área se automatizan; existen algunos que requieren de decisiones humanas.

En la actualidad, muchas de las empresas que venden productos textiles utilizan un sistema transaccional, pero el uso de un software no siempre es suficiente para alcanzar los objetivos del negocio como se mencionó en el párrafo anterior la automatización de procesos mediante sistemas de información puede ayudar a agilizar muchos procesos. (ROCKETT, 2003) Sin embargo, algunos procesos no pueden ser completamente automatizados, ya que por naturaleza dependen de decisiones tomadas (de acuerdo a cada posible escenario) por personal de la empresa.

En el proyecto actual, se presenta el escenario de una empresa textil que se dedica a la elaboración y venta de prendas de vestir. Esta empresa realiza todas las operaciones del área de ventas manualmente; es decir, el registro se realiza en guías, facturas y boletas, pero todo es manuscrito. Además, tiene como fuente de ingreso (utilidades), la venta de prendas de vestir para un público mayorista y minorista. Las utilidades por cada tipo de público varían; por un lado, a sus clientes mayoristas se les ofrece un precio menor, para que puedan llevar volúmenes grandes de productos de vestir; por otro lado, para sus clientes minoristas lo que importa es la utilidad que se pueda sacar por cada unidad de prenda de vestir.

Se han identificado dos problemas en torno a los cuales se buscarán soluciones:

- Manejo ineficiente de la información de ventas y clientes.

- Pérdida de clientes por la constante competencia y una falta de organización estratégica de la empresa para poder retenerlos.

En el primer problema, la empresa posee un registro de las ventas en documentos de ventas (Boletas de venta, Facturas, Guías de Remisión), los cuales son guardados en paquetes de documentos de ventas (conjunto de boletas, guías de remisión y facturas). Si se desea buscar cualquier venta producida en un periodo de tiempo, se tiene que revisar cada documento (búsqueda según fechas o clientes), lo que genera una pérdida de tiempo y además se corre riesgo de pérdida de información por falta de algún documento de ventas.

En el segundo problema, la empresa en los últimos 2 años ha comenzado a perder clientes, en parte por la fuerte competencia que se ha producido en el sector textil de prendas de vestir. Esta competencia obliga a la reducción de precios en sus productos para vender más y no perder más clientes. Además, la empresa tiene solo consumidores locales (compradores únicamente en el país) y estos son los que a partir de sus compras determinan el rumbo de la empresa.

En conclusión, los problemas identificados requieren soluciones que permitan sobrellevar los problemas analizados. Por esto, la empresa necesita una herramienta que le permita manejar su información de manera ordenada y otra herramienta que le permita obtener pronósticos de los comportamientos de compras de sus clientes; de esta forma quizás puedan atenuar la fuga de clientes.

1.2 Marco conceptual

En la presente sección se definirán algunos términos que se utilizarán frecuentemente en el proyecto.

1.2.1 Organización emprendedora

Es toda organización en la cual el poder se concentra en el líder, este tipo de organizaciones se enfocan en nichos de mercado riesgosos y específicos; además, normalmente en este tipo de organizaciones el líder es el dueño. (KREISER, 2006)

Todas las organizaciones lucrativas y no lucrativas siempre pasan por esta etapa y en tiempo de crisis el líder es el que siempre asume la responsabilidad. (KREISER, 2006)

1.2.2 Datos, información y Conocimiento

Los datos son individualmente características, atributos o hechos sin ninguna Información relevante. (WEISS & DAVIDSON, 2010) Estos no tienen mayor significado individualmente (dato), pero si su volumen es grande se puede interpretar como información. (WEISS & DAVIDSON, 2010)

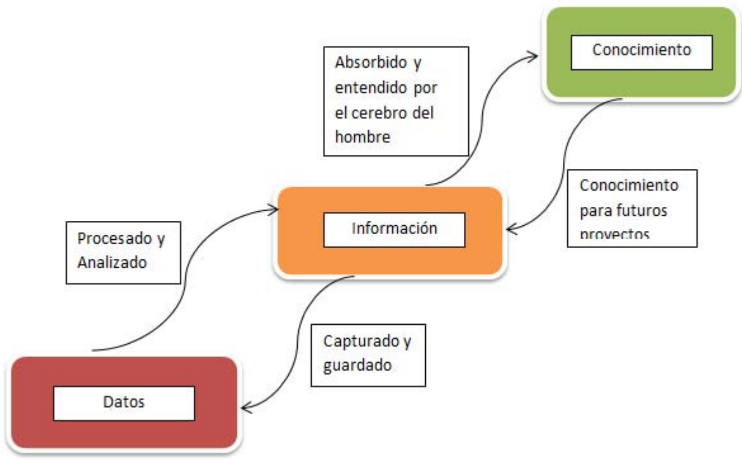


Figura 1.1: Relación entre Datos, Información y Conocimiento (LIEW, 2007)

Para el proceso de Data Mining los datos son esenciales, ya que son la fuente de hechos, de los cuales se puede llegar a obtener información analizándolos detalladamente. (WEISS & DAVIDSON, 2010)

Una de las características de la información es que el volumen de información con respecto a los datos siempre es menor; sin embargo, la información tiene mayor valor para el negocio. (LIEW, 2007)

La información siempre tiene algún significado o interpretación que se puede sacar de la enorme cantidad de datos. (LIEW, 2007)

El resultado del procesamiento de información en conjunto con la experiencia humana es el conocimiento; este tiene mayor valor para el negocio, pero su volumen decrece con respecto a la información. (LIEW, 2007) Del conocimiento obtenido del pasado y presente se puede predecir futuros comportamientos del entorno. (LIEW, 2007)

La Figura 1.1 1.1 presenta claramente la relación entre estos tres componentes; en primer lugar los datos se reúnen para formar información, luego proceden a generar conocimiento a través de la mente humana; y el proceso inverso es generar información a partir de la experiencia, el conocimiento, y finalmente se extraen datos relevantes de cualquier información.

1.2.3 Data Mining

Es un proceso no trivial que tiene como entrada datos y como salida Información, en este proceso se hace un análisis detallado a través del uso de algoritmos para descubrir patrones o comportamiento de los datos. (WEISS & DAVIDSON, 2010)

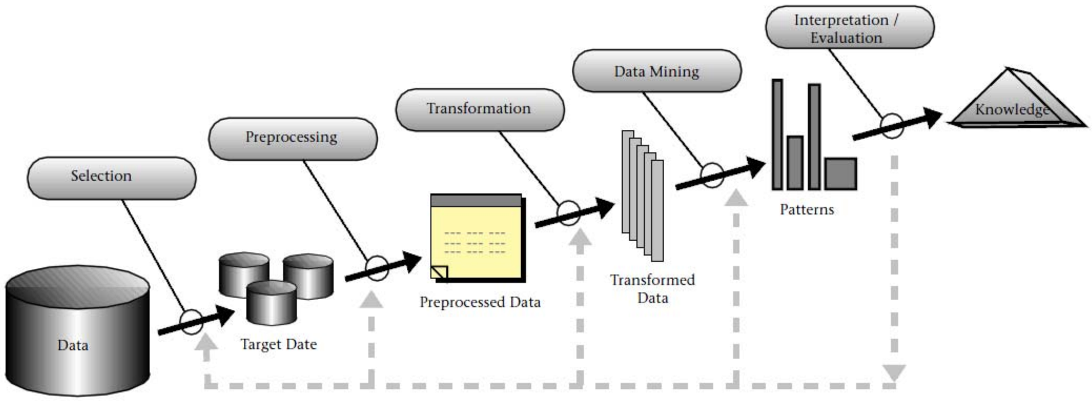


Figura 1.2: Proceso de Data Mining (FAYYAD, et al., 1996)

El Proceso de Data Mining de la Figura 1.2 es el que se definió en 1996 por Fayyad, Piatetsky- Shapiro y Smyth, como se puede apreciar consta de 4 subprocesos: Selección, Pre-procesamiento, Transformación y la parte algorítmica. (WEISS & DAVIDSON, 2010)

Los datos son ingresados al proceso de Selección, básicamente para poder seleccionar solo los datos útiles que sirven para el siguiente proceso, Pre-procesamiento de la data. (WEISS & DAVIDSON, 2010)

Luego los datos entran al proceso de transformación, donde se eleva la calidad de los datos para que queden listos para entrar en el proceso final del algoritmo; si bien no muchos toman al proceso de transformación como relevante, este es sumamente importante en todo el proceso de Data Mining, ya que sin estos no se obtienen datos de alta calidad. (WEISS & DAVIDSON, 2010)

Finalmente, tenemos el último proceso, el cual incluye el uso de algoritmos computacionales para hallar patrones en la data y finalmente generar conocimiento. (WEISS & DAVIDSON, 2010)

El proceso de Data Mining utiliza algoritmos; sin embargo, para poder usarlos correctamente se requiere de la definición de variables para hallar el comportamiento de patrones. (WEISS & DAVIDSON, 2010)

El proceso de Data Mining es comparado con KDD (Knowledge Discovery in Databases), muchas de estas comparaciones establecen que ambos términos son lo mismo y otros establecen que son diferentes; es decir, que KDD está superpuesto a Data Mining. (FAYYAD, et al., 1996)

Data Mining es usado usualmente por los estadísticos, analistas y por los administradores de sistemas de información como el proceso de descubrimiento; mientras, que el término KDD es utilizado por los especialistas en inteligencia artificial. (FAYYAD, et al., 1996)

Data Mining puede ser realizado sobre un Data Warehouse o sobre una base de datos transaccional (para el presente proyecto se realizará sobre una base de datos transaccional). (MOLINA & GARCIA, 2006) En el primer caso, los datos están pulidos por los procesos anteriores del Data Warehouse; sin embargo, en el segundo caso, no se presenta un proceso previo de limpieza de datos, por lo cual es indispensable realizar un análisis sobre cada estructura de datos. (MOLINA & GARCIA, 2006)

Según (MOLINA & GARCIA, 2006) los procesos de Data Mining se apoyan en tecnologías que le sirven de apoyo, algunas de estas son:

- Razonamiento Estadístico
  Las técnicas y métodos estadísticos son muy usados para el Data Mining, debido a que juegan un papel importante en el análisis de datos, y también en el aprendizaje automático. Muchos paquetes estadísticos usados hoy en día se han integrado a diferentes bases de datos y se comercializan como productos de Data Mining. (MOLINA & GARCIA, 2006)
- Visualización
  Las tecnologías de visualización muestran gráficamente los datos de la base de datos. Los modelos de visualización pueden ser bidimensionales,

tridimensionales e incluso multidimensionales (en bases de datos). (MOLINA & GARCIA, 2006)

- Procesamiento Paralelo
  Se ha vuelto un factor crítico en base de datos y Data Mining (Minería de Datos), ya que el rendimiento de consultas es importante para la fluidez de procesos posteriores. (MOLINA & GARCIA, 2006)

- Apoyo a la toma de decisiones
  Son básicamente herramientas que se usan para tomar decisiones eficaces y se basan en teoría de decisiones. Las herramientas de apoyo de toma de decisiones se usan para eliminar los resultados innecesarios obtenidos del proceso de Data Mining. (MOLINA & GARCIA, 2006)
- Aprendizaje automático
  Consiste en aprender las experiencias del pasado con respecto a alguna métrica de rendimiento. Por ejemplo en los videojuegos se aprende de las experiencias pasadas para aprender a jugar. En Data Mining se usan muchas de las técnicas de aprendizaje automático en los algoritmos. (MOLINA & GARCIA, 2006)

Hoy en día, Data Mining tiene muchas disciplinas tales como: Text Mining, Web Mining, Image Mining, EDM (Educational Data Mining), etc.

- EDM (Educational Data Mining)
  Es un área emergente, en la cual los métodos y técnicas para explorar los orígenes de varios sistemas de información de educación (LMS, learning management systems-Sistemas de manejo de aprendizaje, e ITS, intelligent tutoring systems- Sistemas de tutoría inteligente). (CALDERS & PECHENIZKIY, 2011) EDM contribuye al estudio de cómo los estudiantes aprenden; también permite toma de decisiones para el manejo de los datos, con el fin de mejorar la actual práctica educacional y los materiales de aprendizaje usados en la enseñanza. (CALDERS & PECHENIZKIY, 2011)
- Text Mining
  Es el proceso de extraer información relevante de datos textuales mediante algoritmos. (MOLINA & GARCIA, 2006) Esto se logra mediante un estudio de la semántica del texto y de los criterios por los cuales se identifica los datos relevantes. (MOLINA & GARCIA, 2006)

- Web Mining

  Es el uso de las técnicas de Data Mining para poder descubrir y extraer información de documentos y servicios Web. (MOLINA & GARCIA, 2006) Además, es usada para resolver los problemas de sobrecarga de información de forma directa o indirecta. (MOLINA & GARCIA, 2006)

  Los datos en la web en los últimos 10 a 12 años han crecido de una manera exponencial, es por esto que se necesita de buscadores potentes que solamente muestren la información relevante para el usuario; es decir, información indexada. (MOLINA & GARCIA, 2006) Este es uno de los objetivos que se plantea con el uso de Web Mining. (MOLINA & GARCIA, 2006)

1.2.4 Missing values (Valores Perdidos)

Se llaman así porque son un conjunto de datos guardados en una base de datos, la cual posee un registro incompleto de algunos o todos los campos necesarios para el Data Mining; es decir, se posee atributos en blanco ya que no fueron llenados. (HAN & KAMBER, 2006)

Algunas formas de combatir este problema son:

- Ignorar el registro de datos, este método no es muy efectivo; además solo puede ser aplicable cuando la cantidad de datos en el registro es muy baja, si es el caso contrario no se recomienda su uso, porque se llegaría a perder gran cantidad de datos valiosos. (HAN & KAMBER, 2006)
- Rellenar los datos faltantes manualmente, consume mucho tiempo y no es recomendable para registros con gran cantidad de datos faltantes. (HAN & KAMBER, 2006)
- Uso de una constante global para rellenar los valores faltantes, reemplazar los valores de los atributos sin datos con etiquetas, “Desconocido”, este método puede afectar el proceso de Data Mining, ya que el algoritmo puede darle otro significado a estas etiquetas. (HAN & KAMBER, 2006)
- Usar el valor más probable para rellenar el valor faltante, hacer uso de regresión, formalismo de Bayes y árbol de decisiones para predecir los valores faltantes en los registros de datos. (HAN & KAMBER, 2006)

1.2.5 Noisy Data (Datos Ruidosos)

Se define como un error aleatorio o en varianza para un valor medible; es decir que si el precio de un producto oscila entre 10 dólares y 20 dólares y hay uno que esta fuera de ese rango, este último es considerado dato ruidoso; se puede combatir suavizando los datos, para esto existen técnicas. (HAN & KAMBER, 2006) Algunas de estas son:

- Binning, este método ordena los datos con consultas a los vecinos más cercanos, luego son separados en grupos llamados “bins” o “buckets” de igual tamaño y luego los valores medios de los grupos son reemplazados por el valor más significativo del grupo, por valores de los extremos o por el valor que se encuentre en el medio del grupo; mientras más grande sea el grupo hay una mejor aproximación. (HAN & KAMBER, 2006)
- Regresión: este método consiste en encontrar la mejor línea que une a dos atributos, del resultado se puede predecir el comportamiento de los siguientes atributos, los que están muy dispersos del comportamiento de la línea de regresión lineal son aislados. (HAN & KAMBER, 2006)
- Clustering, este método consiste en agrupar datos en conjuntos o “clusters” clasificados por la similitud de los datos, los valores que caen fuera del rango de los “clusters” son aislados. (HAN & KAMBER, 2006)

1.2.6 Limpieza de datos (Data Cleaning)

Es un proceso muy grande, abarca Valores Perdidos y datos ruidosos; sin embargo, estos dos problemas no son los únicos que se presentan en este proceso; el primer paso para poder realizar una limpieza de datos es identificar las discrepancias, estas pueden ser causadas por un pobre diseño de datos, errores deliberados, decaimiento de data (datos no actualizados), etc. (HAN & KAMBER, 2006)

Para poder analizar estos problemas, se requiere tener conocimiento de metadata (estructura de datos); es decir, se requiere el dominio de los tipos de datos, valores aceptables de los atributos, si alguno de los valores de los atributos cae fuera del rango aceptable, existen dependencias entre los atributos, etc. (HAN & KAMBER, 2006) A partir de esto, se puede realizar scripts para procesar estos problemas; luego, con los datos ya procesados se deben crear reglas para poder ver los detalles de la estructura de los datos; es decir, formatos de fechas, monedas, etc. (HAN & KAMBER, 2006)

El segundo paso es la transformación de los datos, en este paso se debe transformar los datos según las reglas definidas en el paso anterior, para poder tener así formatos uniformes de datos y el proceso de algoritmia pueda procesar datos limpios. (HAN & KAMBER, 2006)

1.2.7 Knowledge Discovery in Databases (KDD, Descubrimiento de Conocimiento en Bases de Datos)

Según Fayyad (1996), KDD es un proceso no trivial para identificar patrones de datos válidos, originales, potencialmente usables y entendibles; Friedman (1997), lo consideró como una exploración automática de análisis de datos de grandes bases de datos y Hand (1998) lo definió como un proceso secundario de análisis de datos de una gran base de datos. (MAIMON & ROKACH, 2005)

Entonces se puede definir al KDD como un proceso no trivial de extracción de información a partir de datos. (MAIMON & ROKACH, 2005) Este proceso de extracción conlleva al uso de algoritmos para el análisis de la datos, también conocido como Data Mining; es decir, Data Mining es una de las fases del KDD. (MAIMON & ROKACH, 2005)

EL KDD tiene varias fases según (MAIMON & ROKACH, 2005), las cuales se mencionan a continuación:

- Desarrollo y entendimiento del dominio de la aplicación, el conocimiento relevante y los objetivos del usuario final.
- Selección del conjunto de datos a ser procesados.
- Realizar el Pre - procesamiento de datos, en esta fase se realizan las operaciones de Reducción de Dimensiones, limpieza de datos y transformación de datos.
- Escoger la apropiada tarea de Data Mining: clasificación, regresión, agrupación y resumen de los datos; según la elección hecha se podrá escoger el algoritmo en la siguiente fase.
- Escoger el algoritmo de Data Mining, esto consiste en escoger el método adecuado para buscar patrones en los datos.
- Evaluación e interpretación de los patrones a usar.
- Consolidación del conocimiento descubierto, lo cual consiste en incorporar el conocimiento en el funcionamiento del sistema para prever una acción futura y finalmente documentar el conocimiento adquirido en el proceso.

Las fases definidas anteriormente sirven para una guía adecuada en cualquier proyecto que esté relacionado con Knowledge Discovery in Databases, el cual a veces es confundido y es relacionado al Data Mining como el mismo proceso. (MAIMON & ROKACH, 2005)

1.3 Descripción de la empresa de estudio

La empresa textil cuenta con más de 10 años en el rubro de venta de prendas de vestir. Se fundó en 1999 y se dedica a elaborar y vender polos y blusas para damas y niñas.

Para poder fabricar los polos utilizan tela punto y para las blusas usan tela plana, pero a veces suele suceder que para los polos usan una mezcla de ambos tipos de telas. Las ventas varían de acuerdo a las temporadas del año; por ejemplo, en verano se suelen vender polos manga corta de diferentes modelos, pero para el día de la madre se suelen vender polos para señoras. La empresa tiene un conjunto de clientes mayoristas fijos, las ventas de productos (prendas de vestir) a estos que son la principal fuente de utilidades, aproximadamente sesenta a setenta por ciento de las ventas; por esta, razón si el número de clientes mayoristas baja, entonces es muy probable que las ventas también bajen. Los productos que posee en stock la empresa son mayormente polos y de estos hay en existencia alrededor de 50 productos a más.

La cantidad de empleados que maneja la empresa en total es alrededor de treinta, de los cuales 8 están destinados a la parte de ventas.

A continuación se presentan algunos términos importantes del rubro textil de ventas de prendas de vestir para poder entender mejor el entorno de la empresa:

- Tela

Se llama así a la lámina que se obtiene mediante el cruce y enlace entre dos o más series de hilos textiles, unos longitudinales y los otros transversales; en general se llama tela a toda obra hecha por telar; el tejido más común está compuesto por dos series de hilos: el udimbre(longitudinal) y la trama (transversal). (ESCUELA DE DISEÑO EN EL HÁBITAT, 2008)

- Tela Plana

Está formada por una serie de hilos longitudinales entrecruzados con otra serie de hilos transversales, la cara superior del telar se llama “haz” y la inferior “envés”; en este tipo de tela se usan tramas y udimbres. (ESCUELA DE DISEÑO EN EL HÁBITAT, 2008)

Ejemplos: Denim, Corduroy, Drill, etc.

- Tela Punto

Este tipo de tela también está formada por dos series de hilos; sin embargo, solo se usan tramas o udimbres, mas no los dos. (ESCUELA DE DISEÑO EN EL HÁBITAT, 2008)

Ejemplos: Pima, Jersey, Franela, Rib, Gamuza, etc.

- Proceso de ventas

Es el proceso principal por el cual una empresa puede obtener utilidades sobre los productos que produce o sobre los servicios que brinda.

El proceso de ventas se divide en sub-procesos, los cuales se detallan a continuación:

- Contacto: La primera impresión siempre es importante para poder sellar una venta exitosa, el vendedor se debe adaptar a las costumbres y modales del cliente.
- Evaluación: Se debe de evaluar al cliente, tratando de saber cuáles son sus necesidades específicas
- Presentación del producto: El vendedor debe resumir toda la información que el cliente dio en una etapa previa, luego con la información obtenida tratar de hacer la mejor presentación del producto al cliente.
- Objeciones: Normalmente el cliente observa algunas desventajas del producto, el vendedor debe saber lidiar con estas para que la venta no se trunque.
- Cierre de venta: Es en este punto donde llega por fin la venta en sí misma, y es donde el vendedor debe poner énfasis agregando un valor agregado para que el cliente se sienta complacido con el producto y finalmente se sienta seguro de comprarlo.

1.4 Revisión del estado del arte

Para el presente proyecto se utilizan dos tipos de herramientas; por un lado, un software que facilite y automatice el proceso de ventas; por otro lado, una herramienta de software que facilite el proceso de Data Mining.

Las herramientas presentadas a continuación no están hechas a la medida exacta de la empresa textil; por esto, se realizará una adaptación de estas herramientas al modelo textil de la empresa, esto permitirá que la comunicación entre ambas herramientas pueda ser mucho más fluida y así puedan cumplir con el objetivo principal del proyecto.

1.4.1 Objetivo

El objetivo del Estado del Arte es poder seleccionar las dos herramientas (Data Mining y Software de ventas) para poder resolver los problemas de la empresa detallados en el presente proyecto. Esta selección se realizará con un conjunto de criterios definidos en base a las necesidades de la empresa.

1.4.2 Herramientas de Software para Ventas

El proyecto necesita una herramienta de software que permita realizar las operaciones de ventas; además, debe ser de fácil uso, y que permita el manejo de la cartera de clientes que la empresa posea actualmente; adicionalmente a todo esto, el software debe ser de código abierto, a continuación se detallan los productos que se evaluaron:

- **ADempiere**

  Es un ERP completo, que integra todas las funciones de una empresa: Administración de la Cadena de Suministros, Logística, producción, Ventas, Análisis de Rendimiento, Administración de relaciones con Clientes, Contabilidad y producción. (ADEMPIERE, 2012) A continuación se detallan algunas de sus principales características:

  - ADempiere está diseñado en los procesos de negocio y no en la arquitectura contable y departamental tradicional, lo que hace más fluida la interacción entre los procesos y por ende más dinámica.
  - Es de código abierto (open source).
  - Permite administrar clientes y proveedores.
  - Es un ERP y CRM, lo que permite un manejo de relaciones con los clientes para poder satisfacer sus necesidades.

- **Sviw32**

  Es un software de Punto de Venta, lo que permite el registro con código de barras (ALBARRAN, 2012), a continuación se presentan algunas de sus principales características:

  - Toma un control completo del inventario.
  - Registra las actividades de cada empleado.
  - Trabaja con sistemas operativos: Unix, Linux y Windows.

- **Openbravo**

Es uno de los ERP más conocidos en la actualidad, mejora los procesos de negocio e incrementa la productividad y la agilidad del negocio (OPENBRAVO, 2012); a continuación se detallan algunas características de este ERP:

- Posee un diseño multi-tabla (muestra diferentes tipos de datos en una sola tabla; es decir, cruza información de diferentes tipos de datos) que permite una interacción más amigable con el usuario.
- Posee una arquitectura acondicionada para trabajar con Servidor, Base de datos y clientes.

- **OpenERP**

  Es el ERP open source líder en la actualidad; además permite la integración de los procesos de las áreas de mayor demanda en empresa. (OpenERP, 2012)A continuación se detallan algunas características:

  - Permite un seguimiento de campañas de ventas.
  - Integrado con un CRM, lo que permite establecer una relación más cercana con los clientes.
  - Permite gestionar almacenes de forma fácil e intuitiva, mediante gráficos y tablas dinámicas.

- **Estrasol**

  Es un software de código abierto de pantalla táctil que permite una mejor interacción con los usuarios (vendedores); además de poseer una interfaz amigable, permite el registro rápido de las ventas. A continuación se detallan algunos de sus características:

  - Posee un control de la venta de productos.
  - Permite generar reportes de las ventas.
  - Es un software Punto de venta (POS), lo que permite el trabajo con código de barras (un control exhaustivo de los productos en tiendas).

Se realiza una comparación en la en base a los criterios de selección de la herramienta de software de ventas en la Tabla 1.1.

Luego de revisar los productos existentes, se optó por el software Open ERP como sistema transaccional para el área de ventas, ya que es una herramienta de código abierto que permite administrar de manera óptima clientes y hacer un registro de ventas actualizado, y además brinda una interacción más amigable al usuario al realizar una venta. Este último aspecto permite un fácil aprendizaje y manejo de la herramienta software por parte de los usuarios del área de ventas, y es el factor determinante para la elección del OpenERP sobre el resto de herramientas de software presentadas.

1.4.3 Herramientas de Data Mining

Hoy en día existen diversas herramientas para poder realizar el proceso de Data Mining en todo su potencial; entre estas se encuentran herramientas de código abierto y también herramientas por pago de licencia; en la Tabla 1.2, luego se realiza una comparación entre algunos productos actuales del mercado.

- **Knowledge Seeker**

  Es un producto de inteligencia de negocios con Data Mining para poder predecir y anticipar escenarios; además posee funcionalidad de diseño de estrategia (ANGOSS, 2012), entre sus principales características se tiene:
  - Es flexible, una herramienta poderosa que permite interactuar a los usuarios de forma fácil; además, posee arboles de decisión, arboles de estrategias (comparación de estrategias con múltiples variables de decisión).
  - Permite importar y exportar datos de Excel, SAS, SPSS y otros sistemas de bases de datos.
  - Permite la transformación de datos usando expresiones SQL.
- **dVelox Enterprise**

  Permite automatizar los procesos de toma de decisiones de sus procesos críticos, ofrece a los usuarios de negocio predicciones precisas de forma sencilla (APARA, 2012); a continuación se presentan algunas de sus características:
  - Analiza escenarios complejos como la prevención de fraude, retención de clientes y encuentra patrones de comportamiento y determina probabilísticamente la mejor opción para dicho escenario.
  - No requiere conocimientos avanzados de matemáticas, ni estadística para obtener el máximo rendimiento.
- **MicroStrategy Data Mining Services**

  Es una componente de la plataforma MicroStrategy BI que brinda resultados de modelos predictivos; además, brinda reportes dinámicos (MICROSTRATEGY, 2012); a continuación se detallan algunas características de esta solución:

- Predice cálculos usando funciones analíticas, las cuales incluyen: Regresión lineal, Regresión lógica, árbol de decisiones, reglas de modelos asociados y modelo de series.
- Crea reportes predictivos, flexibles, organizados para que el usuario los pueda entender fácilmente y también para que las presentaciones sean profesionales.
- Implementa un esquema de seguridad estricta para los usuarios que se encuentran dentro y fuera de la organización.

- **SQLServer Data Mining 2008**

  La familia de SQLServer es una de las más usadas en la actualidad y en la familia de productos para Inteligencia de negocios se integró una herramienta para Data Mining (MICROSOFT, 2012), a continuación se presentan algunas de sus más resaltantes características:

  - Posee una amplia gama de algoritmos tales como: Árbol de decisiones, redes neuronales, regresión lineal, regresión logística, etc.
  - Permite visualizar los modelos de Data Mining para optimizar la representación de datos.
  - Provee herramientas ETL para la limpieza y transformación de datos.
  - Permite una interacción con la interfaz gráfica agradable para el usuario, de modo que pueda entender la información procesada de manera rápida.

- **Pentaho Data Mining (Weka)**

  Esta herramienta se integra con las otras soluciones de la plataforma Pentaho, lo que la hace aún más inteligente (PENTAHO, 2012); también posee soporte para integración de datos, análisis, dashboards y reportes; a continuación se detallan algunas características:

  - Motor poderoso para manejo de grandes volúmenes de datos.
  - Posee una gran colección de algoritmos confiables y robustos.
  - Relaciones eficientes y capacidad de descubrir patrones de datos.
  - Integración simple y acelerada de datos.

Tabla 1.1: Comparación entre herramientas software para el registro de ventas (Elaboración Propia)

	Permite gestionar planes de ventas	Permite integración con otras áreas de la empresa	Propone mejoras de procesos	Permite ventas por POS (Puntos de Venta)	Permite el uso libre de los POS (Punto de Venta) sin ningún pago de por medio.	Permite una fácil interacción con el usuario (usabilidad)	Puntuación
ADempiere	5	2	2	0	0	0	9
Sviw32	0	0	0	3	0	1	4
Openbravo	5	2	2	3	0	1	13
OpenERP	5	2	2	3	3	1	16
Estrasol	0	0	0	3	0	1	4

De todos los productos comparados en la Tabla 1.2 se eligió la herramienta Pentaho Data Mining (Weka), ya que ha cumplido con todos los criterios de selección que se requieren para el presente proyecto, Además, cabe resaltar que en el presente proyecto se hace énfasis en el uso de herramientas de código abierto, ya que no se cuenta con presupuesto para adquirir licencias.

1.5 Discusión sobre los resultados del estado del arte

En el estado del arte se realizó dos comparaciones, una para el software ERP que se usará en el área de ventas de la empresa, y la otra comparación se realizó para poder elegir la herramienta software de Data Mining que se requerirá para el proyecto actual.

Se dio mayor importancia al hecho que la herramienta software sea de código abierto por los pocos recursos económicos que se cuentan para la realización del proyecto, luego era necesario que la herramienta ofrezca las funcionalidades de ventas completas y de fácil acceso; el resto de criterios tuvo menor importancia a la hora de la elección.

En el primer caso, las herramientas software de ventas que se compararon, todas eran de código abierto; sin embargo, solo dos herramientas eran reconocidas y contaban con más aceptación en el mercado, estas eran OpenBravo y OpenERP; además, ambas herramientas contaban con documentación, manuales y todas las funcionalidades que un módulo de ventas debería tener. La elección entre ambas herramientas se decidió por el costo de implantación que tenía una herramienta adicional de OpenBravo; esta herramienta adicional solo se podía conseguir mediante los partners de OpenBravo e instalarla y configurarla tenía un precio adicional.

En el segundo caso, también se realizó una comparación de los criterios para poder dar un peso o prioridad a cada uno de estos criterios; de esta comparación se priorizó los siguientes aspectos: la herramienta debe ser de código abierto, la herramienta permite transformación, integración y limpieza de datos, la cantidad de algoritmos que tiene la herramienta y trabajar con distintas base de datos; el último criterio, interfaz gráfica para el usuario, tuvo un valor menos relevante, ya que lo principal en el proyecto era usar los algoritmos. Cabe resaltar que ser una herramienta de código abierto tuvo bastante relevancia, porque el proyecto no cuenta con los recursos económicos para la adquisición de un software por pago de licencia.

La elección de la herramienta software de Data Mining no fue tan difícil, ya que el principal criterio de elección era que la herramienta debía ser de código abierto y la única que cumplía con ese requisito en el conjunto de herramientas presentado era Pentaho Weka. Sin embargo, había otros criterios de comparación adicionales que podrían equiparar la elección, pero Pentaho Weka cumplía con todos, así que se concluyó que era la herramienta que se usaría en el proyecto.

1.6 Descripción y justificación de la solución

La solución planteada para el problema mencionado en este proyecto tiene como fin brindar herramientas y ayuda para resolver todo el problema o parte de él. Esto quiere decir, que la solución del problema depende exclusivamente de la empresa y de la buena gestión que esta realice durante el proyecto.

Como ya se mencionó previamente, el problema tiene dos puntos resaltantes:

- Manejo ineficiente de la información de ventas y clientes.
- Pérdida de clientes por la constante competencia y una falta de organización estratégica de la empresa para poder retenerlos.

Para el primer caso, se plantea usar un ERP de código abierto, el cual automatice el proceso de ventas en la empresa y ayude a manejar eficientemente la información de los productos y clientes mediante reportes. Cabe resaltar que lo mejor es que adecua el proceso de ventas precario de la empresa a un nuevo modelo de ventas implantado por el ERP. Además, los problemas de búsqueda de clientes o de deudas pendientes serán más rápidos y se evitará pérdida de información, si es que se toman las medidas de seguridad adecuadas.

En la empresa seleccionada, se realizó una automatización de algunos procesos del área de ventas. Esta automatización incluye la implantación de un sistema de información de ventas de código abierto, el cual tendrá las siguientes funcionalidades:

- Realizar ventas de productos textiles (prendas de vestir).
- Gestión de clientes de la empresa.
- Gestión de productos textiles de prendas de vestir.
- Cierre de caja diaria.
- Contabilidad de caja periódicamente.
- Gestión de usuarios del software.

Tabla 1.2: Comparación entre herramientas de Data Mining (Elaboración Propia)

	Es una herramienta de código abierto.	Posee una amplia gama de algoritmos para trabajar	Trabaja con distintas bases de Datos.	Permite la limpieza, integración y transformación de datos.	Se realizan continuas actualizaciones sobre las funcionalidades del Software	Puntuación
Knowledge Seeker	0	0	3	0	2	5
dVelox Enterprise	0	0	3	3	2	8
MicroStrategy Data Mining	0	1	3	3	2	9
SQL Server Data Mining 2008	0	1	3	3	2	9
Pentaho Data Mining (Weka)	5	1	3	3	0	12

Para el segundo caso, se usó una herramienta software, la cual posee algoritmos potentes de Data Mining que procesan la información guardada en una Base de Datos y la transforman en valiosa información para que la empresa pueda tomar decisiones respecto a los resultados. La herramienta de Data Mining soló usa un algoritmo, el cuál procesa toda la información relevante que se obtiene del dispositivo de almacenamiento de datos (el cual ha sido llenado de datos previamente por el OpenERP), que luego a través de un conjunto de factores (variables de decisión) se pueda obtener un resultado que beneficie a la empresa y se pueda superar el escollo de la pérdida de clientes.

El proceso completo a seguir es el siguiente:

- Implantación del OpenERP.
- Carga de Datos a la Base de Datos transaccional implantada con el OpenERP.
- Ejecución del proceso de Data Mining.

Los pasos mostrados son consecutivos, los últimos dependen de que los primeros hayan sido realizados satisfactoriamente.

1.6.1 Justificación

La solución propuesta, facilita al usuario información del comportamiento de compra de los clientes de la empresa, tales como tipo de prendas que llevan más en un periodo de tiempo, cantidad promedio mensual de productos que cada cliente compra, conjunto de prendas de vestir que un cliente compra más, entre otros; es decir, con la herramienta Data Mining (usada en el proyecto) la empresa puede comprender como han ido evolucionando en sus compras cada uno de sus clientes. También, se debe destacar que el principal beneficiado con el proyecto es la empresa, y todos los que la conforman.

Los problemas planteados en el proyecto son: manejo ineficiente de las ventas y clientes, y pérdida de clientes por la constante competencia. Para el primer problema, el área de ventas requiere que muchos de sus procesos sean automatizados, como referencia se pueden tomar grandes, medianas y pequeñas empresas que han requerido y acudido al uso de software (ROCKETT, 2003) ; para la implantación del software del área de ventas se pudo haber seleccionado cualquier herramienta del mercado actual que permita automatizar muchos de los procesos del área de ventas; sin embargo, un software como el ERP permite una prospección a futuro, porque se puede llegar a automatizar otros procesos de la empresa que en un futuro requieran un cambio (ROCKETT, 2003), y que este se pueda dar de manera integrada; es decir, que la empresa llegará a usar un solo software, evitando conflictos por el uso de distintas herramientas de software. Para el segundo problema mencionado, este se puede solucionar, como ya se dijo en el párrafo anterior, con los resultados obtenidos del proceso algorítmico de Data Mining y una adecuada gestión ventas y clientes (esto último depende netamente de la empresa).

Con el marco teórico expuesto en el presente proyecto, se puede realizar otro proyecto para realizar el estudio del comportamiento de los clientes en la web o en una red social, en este caso se aplicaría Web Mining, que es una disciplina de Data Mining.

1.6.2 Viabilidad

Viabilidad Técnica

Para el presente proyecto se usaron métodos específicos, los cuales han sido extraídos de las metodologías seleccionadas (PMBOK, CRISP-DM, Agile OpenERP). Cada metodología usada en el proyecto tiene una versión actual y estable; sin embargo, esta puede variar en el lapso de tiempo de la ejecución del proyecto, lo que modificaría el esquema de trabajo en el proyecto. La guía de buenas prácticas para la gestión de proyectos, PMBOK, pronto se actualizó a la versión 5, esto no afectó al proyecto, ya que los métodos seleccionados del PMBOK no fueron afectados en la transición de cambios de la metodología; según el borrador de PMBOK publicado por PMI (Project Management Institute) la versión 5 tiene como principal novedad la inclusión de una nueva área de conocimiento, "área de conocimiento de los interesados (Stakeholders)". Esta modificación, no afecto de manera significativa el esquema de trabajo del proyecto. Por otro lado, también se usaron dos metodologías para el producto, CRISP-DM y Agile OpenERP. La primera metodología actualmente se encuentra en la versión 2.0 y goza de aceptación en proyectos de Data Mining a nivel mundial; esta metodología posee todos los pasos necesarios para realizar el proceso de Data Mining con éxito; además, incluye actividades específicas para cada etapa del proceso de Data Mining. (CHAPMAN, et al., 2000) Para el segundo caso, la metodología es usada específicamente para la implantación del software OpenERP (también sirve para otros software ERP); además, incluye todos los pasos necesarios para poder realizar una implantación exitosa (instalación, configuración del ERP y capacitación a los usuarios).

Para el proyecto también se usan herramientas software (OpenERP y Pentaho Weka). En el caso de la primera herramienta software, OpenERP, es una de las herramientas líderes en el mercado de software ERP de código abierto, aunque también existen versiones por pago de licencia del software. La versión OpenERP de código abierto que es usada para el presente proyecto posee metodologías de implantación, manuales de instalación y configuración, y una adaptación a distintas economías del mundo. (OpenERP, 2012) En el caso de la segunda herramienta, Pentaho Weka, pertenece al grupo Pentaho, el cual es una de las compañías que entre sus productos ofrece software de código abierto y por pago de licencia. Pentaho Weka es una herramienta poderosa que permite la factibilidad de proyectos de Data Mining; además, la herramienta cuenta con un libro, el que explica detalladamente su uso. (HAN & KAMBER, 2006)

Viabilidad Temporal

El presente proyecto tuvo una duración aproximada de 4 meses, 15 horas semanales de trabajo (en el calendario de actividades está especificado en el diagrama de Gantt del proyecto). Se planifico usar solo un único recurso humano, el tesista; el cuál dispuso de tiempo suficiente para la ejecución de todo el proyecto; además, se dispuso del tiempo de los representantes de la empresa para las entrevistas. El tiempo de estos últimos es limitado; por esto, se realizaron entrevistas planificadas para obtener la información necesaria que se requirieron en el proyecto. También, se tuvo en cuenta las postergaciones de las entrevistas; en este caso, se planifico otra reunión en la cual se realizó la entrevista.

Viabilidad Económica

El proyecto necesito de recursos humanos y materiales; en la Tabla 1.3 se tiene una lista de los recursos tangibles e intangibles usados en el proyecto.

Adicionalmente, se usaron herramientas software, las cuales no requirieron ninguna inversión de dinero, y el tesista realizo las siguientes actividades: configuración del software, capacitación al personal de la empresa y pruebas necesarias del producto en el proyecto.

Para el recurso humano se asumió un costo promedio de un practicante de ingeniería informática, 1000 soles por 120 horas; es decir, si se requirieron 15 horas semanales y el pago es 1000 soles por 120 horas, entonces por los 5 meses de trabajo del proyecto se necesitaron 300 horas hombre de trabajo, lo que equivale a 2500 soles. El costo del recurso humano resulto ser más barato que la adquisición de otro software ERP por pago de licencia, ya que estos rondan los 1500 dólares.

Tabla 1.3: Recursos tangibles e intangibles (Elaboración Propia)

<table>
<tr><td rowspan="6">Recursos tangibles e intangibles que se necesitan para el proyecto</td><td>Impresión de informes</td></tr>
<tr><td>Servicio eléctrico en el local de ventas de la empresa</td></tr>
<tr><td>Servicio de internet en el local de ventas de la empresa</td></tr>
<tr><td>Dispositivos electromagnéticos (Discos duros).</td></tr>
<tr><td>Una computadora para la implantación del Software ERP (Adquirida por la empresa)</td></tr>
<tr><td>Computadora personal portátil</td></tr>
</table>

2 Planificación

El presente proyecto tuvo una duración estimada de cinco meses, en los cuales se entregaron tanto el producto final como la documentación asociada a este. Es por esto que en este capítulo se menciona toda la planificación para los productos y el proyecto; también se explica a detalle las metodologías usadas y por último se explica por qué el proyecto es viable.

2.1 Plan de proyecto

En esta sección se presentan: objetivo general, objetivos específicos y resultados esperados, EDT (Estructura de Desglose del trabajo), Diagrama de Gantt, Gestión de riesgos, objetivo general. En general, se presenta todo lo relacionado a la planificación del proyecto.

2.2 Objetivo general

Analizar, diseñar e implantar un sistema de información de ventas en una PYME textil que emplee herramientas Data Mining para la predicción de comportamientos de compra en clientes.

2.3 Objetivos específicos

1. Seleccionar e implantar un software de código abierto para controlar la información en el área de ventas.
2. Identificar los criterios de negocio que la empresa usa para realizar sus ventas.
3. Realizar una limpieza de datos de la base de datos (Data Cleaning), con el fin de que el proceso de Data Mining pueda tener datos de entrada adecuados.
4. Adaptar un algoritmo de Árboles de Decisión que pueda analizar el comportamiento de compra de los clientes de la empresa.

2.4 Resultados esperados

1. Software de ventas implantado y configurado en la empresa.
2. Catálogo de criterios de negocio que la empresa utiliza para planificar sus ventas.
3. Proceso de limpieza (uniformización de todos los datos de la base de datos) de Data Mining de la base de datos transaccional implementado.
4. Algoritmo de Árboles de Decisión implementado para el software de Data Mining.

A continuación en la Tabla 2.1 se muestra las relaciones entre cada objetivo específico con su respectivo resultado esperado.

Tabla 2.1: Perspectiva Objetivos específicos versus resultados esperados (Elaboración Propia)

Objetivo Especifico	Resultado esperado
Seleccionar e implantar un software "código abierto" para controlar la información en el área de ventas.	Software de ventas implantado y configurado en la empresa.
Identificar los criterios de negocio que la empresa usa para realizar sus ventas.	Catálogo de criterios de negocio que la empresa utiliza para planificar sus ventas.

Objetivo Especifico	Resultado esperado
Realizar una limpieza de datos de la base de datos (Data Cleaning), con el fin de que el proceso de Data Mining pueda tener datos de entrada adecuados.	Proceso de limpieza (uniformización de todos los datos de la base de datos) de Data Mining de la base de datos transaccional implementado.
Adaptar un algoritmo de Árboles de Decisión que pueda analizar el comportamiento de compra de los clientes de la empresa.	Algoritmo de Árboles de Decisión implementado para el software de Data Mining.

2.5 Estructura de Desglose del Trabajo

En estructura de desglose del trabajo o EDT se detallan los entregables que se presentan a lo largo del proyecto a los skateholders. Se puede apreciar el EDT en la Figura 2.3.

2.6 Diagrama de Gantt

En el diagrama de Gantt de la Figura 2.4 se muestran los entregables expuestos en el EDT con programación de fecha.

2.7 Gestión de riesgos

En la Tabla 2.2 se muestra los riesgos a los que está expuesto el proyecto; asimismo se presentan planes de contingencia y mitigación por cada riesgo.

2.8 Métodos y procedimientos

En el presente proyecto se usaron métodos de algunas metodologías y herramientas actualmente usadas; a continuación se procederá a detallarlas:

Para el Proyecto: Se utilizó procesos de la guía de buenas prácticas PMBOK, la cual permite la gestión de proyectos de forma eficaz y eficiente; en la actualidad, cuenta con 42 procesos y es una de las más usadas. En el presente proyecto solo se usan los siguientes procesos:

Etapa de Iniciación

- Desarrollar el Acta de Constitución del Proyecto (Anexo N°3)
- Identificar a los Interesados (Stakeholders) (AnexoN°3)

Etapa de planificación

- Definir el Alcance del proyecto (1.6 Descripción y justificación de la solución)
- Crear EDT (Estructura Detallada del trabajo) (Figura 2.3)
- Definir las actividades del Proyecto (Figura 2.3)
- Secuenciar las actividades (Figura 2.3)
- Estimar la Duración de las Actividades (Figura 2.3)
- Desarrollar el Cronograma (Figura 2.3)

Para el producto: Se usaron dos metodología, una para Data Mining, CRISP-DM, y otra para la implementación del ERP, Agile OpenERP.

Metodología Agile OpenERP: Esta metodología tiene una parte introductoria que se divide en dos fases: la parte de planificación de las iteraciones y el análisis funcional de los procesos, luego se describe cada iteración que el producto final necesite, y finalmente se procede con la fase de cierre.

En esta metodología se trabaja con bloques funcionales, los cuales son implementados al final de cada iteración. El objetivo de esta metodología es priorizar funcionalidades, para poder atender primero aquellas de carácter urgente.

Metodología CRISP-DM: Esta metodología se categoriza en 4 niveles de abstracción: fase, tarea genérica, tarea especializada e instancia de procesos.

El primer nivel y el segundo nivel están detallados en la Figura 2.2; el tercer nivel es más especializado y describe las acciones que se realizan en las tareas genéricas del segundo nivel. El cuarto nivel es un registro de las acciones, decisiones y resultados del Data Mining.

A continuación se presentan los objetivos especificos y resultados esperados con los respectivos metodos y cada actividad tiene asociada la metodología y el método usado.

Objetivo Especifico 1: Seleccionar e implantar un software de código abierto para controlar la información en el área de ventas.

- Definir el alcance funcional de la implantación del ERP y planteamiento de calendario. (Agile OpenERP: Planificación temporal del proyecto)
- Definir objetivos de la implantación del ERP. (Agile OpenERP: Objetivos en el Despliegue por iteraciones).
- Configurar la herramienta ERP para adecuarla al modelo de la empresa. (Agile OpenERP: Implementación en Despliegue por iteraciones)
- Pruebas funcionales de cada módulo y pruebas de integración. (Agile OpenERP: Pruebas en Despliegue por iteraciones)
- Capacitación a los Usuarios para el uso de la herramienta ERP.(Agile OpenERP: Cierre en Despliegue por iteraciones)

Resultados Esperados: Software de ventas implantado y configurado en la empresa.

Objetivo Especifico 2: Identificar los criterios por los cuales la empresa decide vender sus productos textiles (prendas de vestir).

- Identificar las unidades de negocio que se ven afectadas por el proyecto actual (ventas, finanzas, entre otras). (Metodología CRISP-DM: Determinar objetivos de negocio)
- Preparar, elaborar y realizar una entrevista con los agentes directos de las ventas. (Metodología CRISP-DM: Determinar objetivos de negocio)
- Analizar y documentar los criterios por los cuales se evaluará el éxito del proyecto. (Metodología CRISP-DM: Plan de proyecto)

Resultados Esperados: Catálogo de indicadores que la empresa utiliza para planificar sus ventas.

Objetivo Especifico 3: Realizar una limpieza de datos de la base de datos (Data Cleaning), con el fin de que el proceso de Data Mining pueda tener datos de entrada adecuados.

- Seleccionar los datos de la base de datos, archivos a ser evaluados.(Metodología CRISP-DM: Seleccionar datos)
- Realizar una limpieza de datos.(Metodología CRISP-DM: Limpiar datos)
- Construir los datos de la base de datos elegida; es decir, construir atributos derivados, completar registros nuevos o transformar valores para atributos existentes.(Metodología CRISP-DM: Construir Datos)

Resultados Esperados: Proceso de limpieza (uniformización de todos los datos de la base de datos) de Data Mining de la base de datos transaccional implementado.

Objetivo Especifico 4: Adaptar un algoritmo de Árboles de Decisión que pueda analizar el comportamiento de compra de los clientes de la empresa.

- Decidir la técnica a usar en el ejercicio, teniendo en cuenta la herramienta seleccionada. (Metodología CRISP-DM: Seleccionar técnicas de modelado)
- Determinar los parámetros del modelo; es decir, las variables de entrada del algoritmo. (Metodología CRISP-DM: Seleccionar técnicas de modelado)
- Documentar las razones para elegir los parámetros en el modelo seleccionado.(Metodología CRISP-DM: Seleccionar técnicas de modelado)
- Describir el comportamiento del modelo en un documento.(Metodología CRISP-DM: Evaluar modelo)
- Realizar las conclusiones con respecto a los patrones en los datos. (Metodología CRISP-DM: Evaluar modelo)

Resultados Esperados: Algoritmo de Árboles de Decisión implementado y configurado en el software de Data Mining.

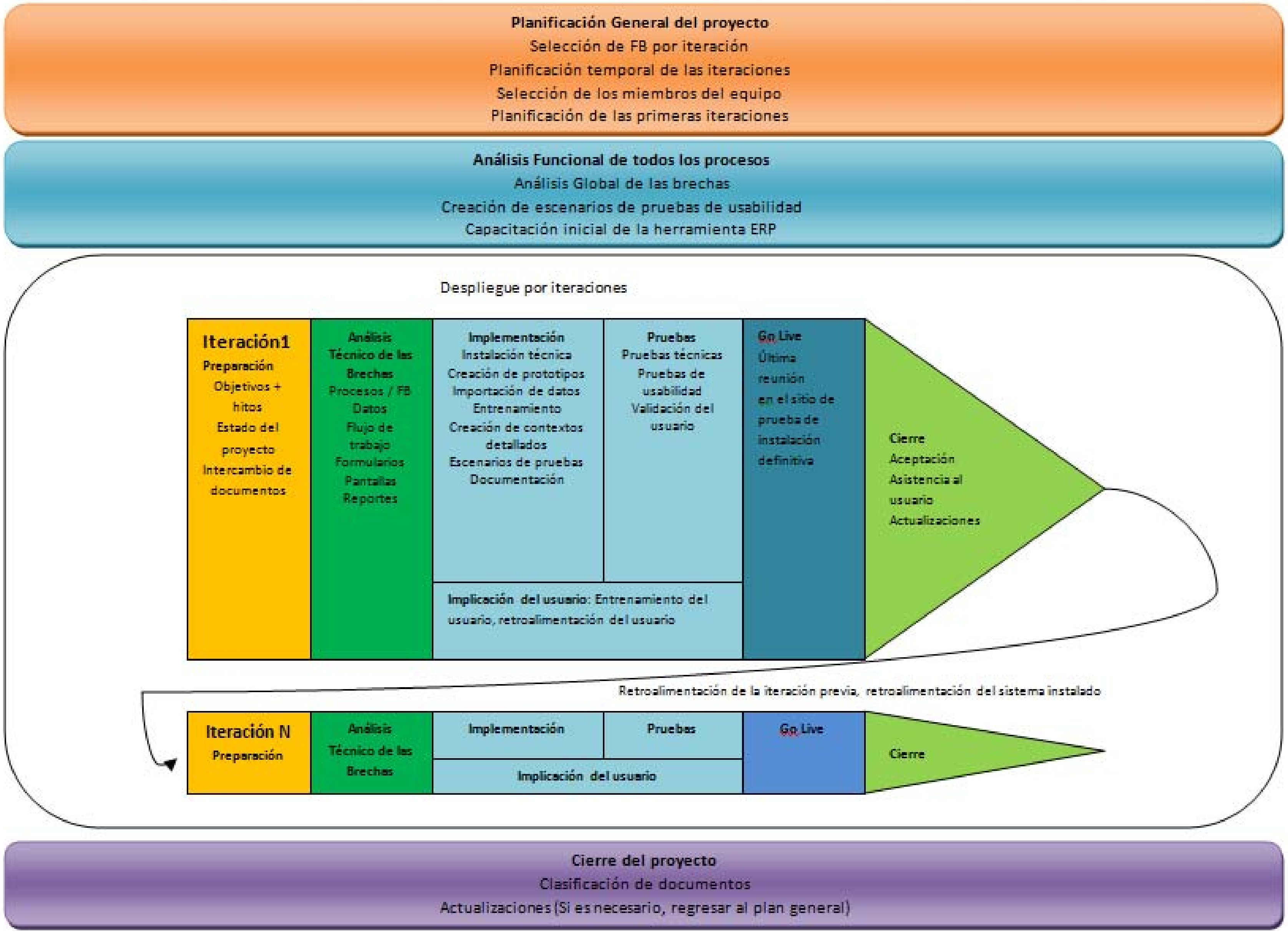


Figura 2.1: Proceso de Implantación mediante la metodología Agile OpenERP. (Elaboración Propia)

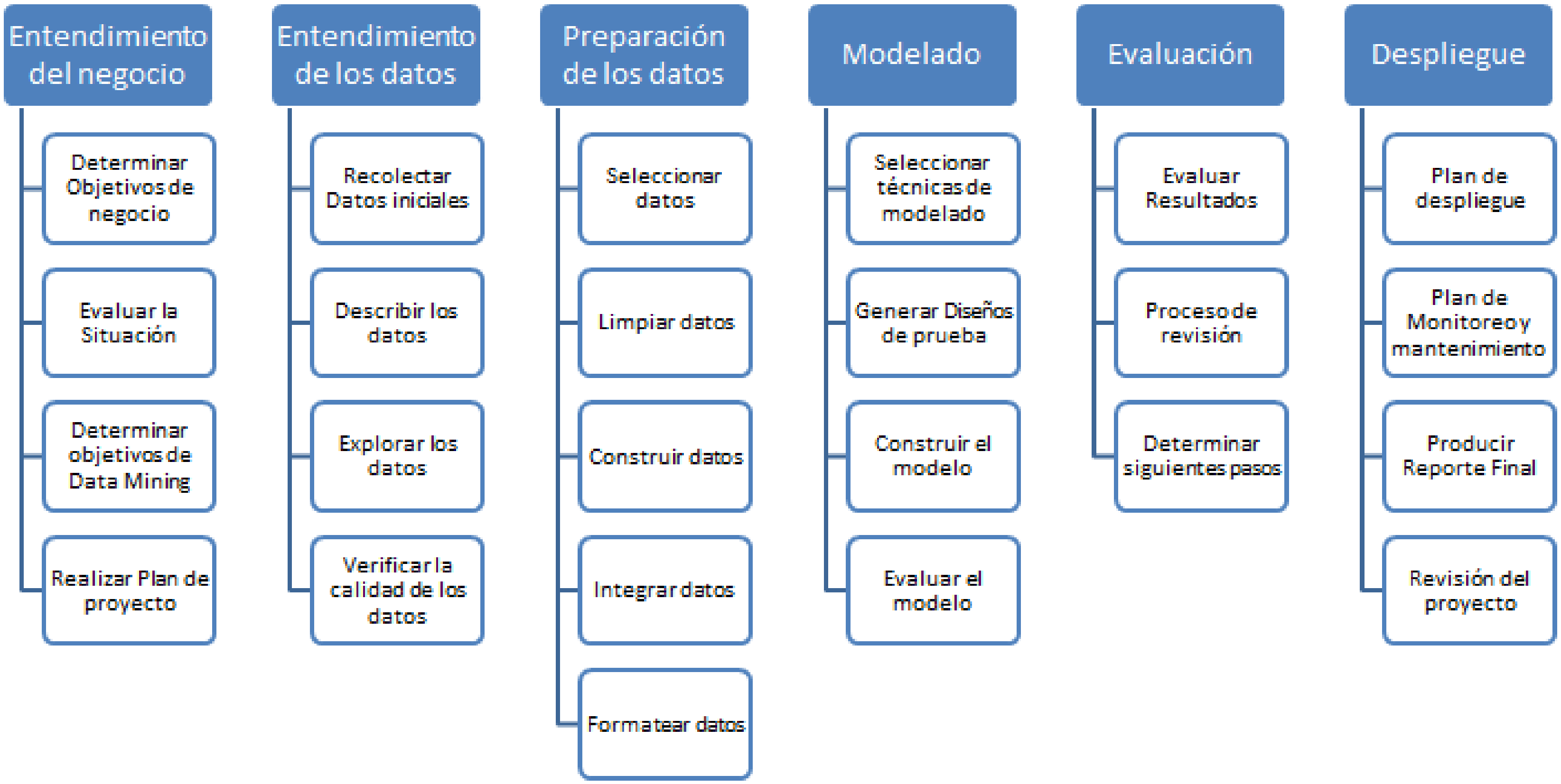


Figura 2.2: Metodología CRISP-DM (Elaboración Propia)

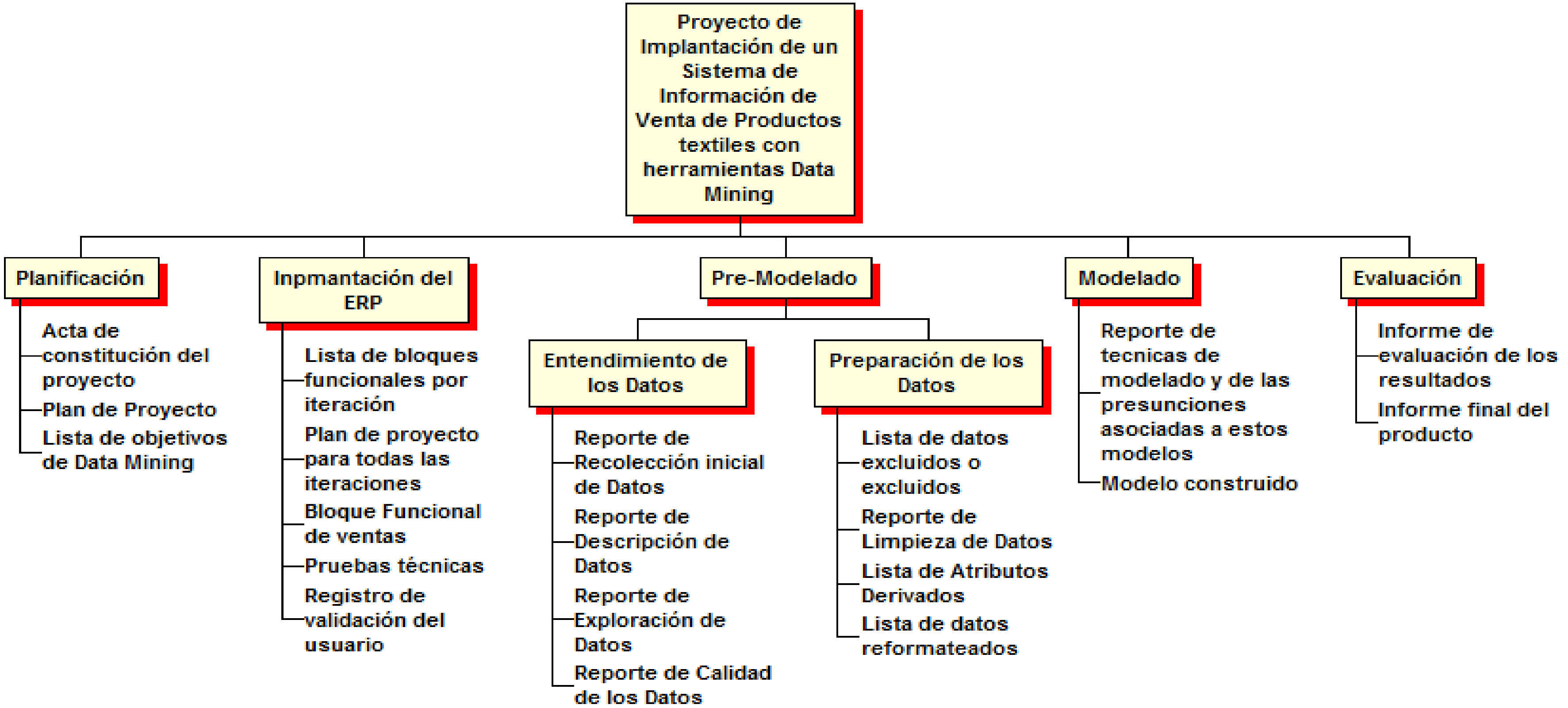


Figura 2.3: Estructura de Desglose del Trabajo (Elaboración Propia)

		Nombre de tarea	Duración	Comienzo	Fin
1		Planificación	5 días	mar 12/06/12	lun 18/06/12
2		Acta de Constitución del Proyecto	1 día	mar 12/06/12	mar 12/06/12
4		Plan de proyecto	2 días	mié 13/06/12	jue 14/06/12
6		Lista de Objetivos del negocio y de Data Mining	2 días	vie 15/06/12	lun 18/06/12
8		Implantación del ERP	24 días	mar 19/06/12	vie 20/07/12
9		Lista de bloques funcionales por iteración	2 días	mar 19/06/12	mié 20/06/12
12		Plan de proyecto para todas las iteraciones	2 días	jue 21/06/12	vie 22/06/12
14		Bloque funcional de ventas	14 días	lun 25/06/12	jue 12/07/12
18		Pruebas técnicas	5 días	vie 13/07/12	jue 19/07/12
21		Registro de validación del producto	1 día	vie 20/07/12	vie 20/07/12
23		Pre-Modelado	43 días	lun 20/08/12	mié 17/10/12
24		Entendimiento de los Datos	33 días	lun 20/08/12	mié 03/10/12
25		Reporte de recolección inicial de datos	5 días	lun 20/08/12	vie 24/08/12
30		Reporte de Descripción de Datos	10 días	lun 27/08/12	vie 07/09/12
41		Reporte de Exploración de datos	5 días	lun 10/09/12	vie 14/09/12
47		Reporte de calidad de los datos	13 días	lun 17/09/12	mié 03/10/12
61		Preparación de los datos	10 días	jue 04/10/12	mié 17/10/12
62		Lista de datos excluidos e incluidos	3 días	jue 04/10/12	lun 08/10/12
66		Reporte de limpieza de datos	2 días	mar 09/10/12	mié 10/10/12
69		Lista de atributos derivados	4 días	jue 11/10/12	mar 16/10/12
74		Lista de datos reformateados	1 día	mié 17/10/12	mié 17/10/12
76		Modelado	12 días	jue 18/10/12	vie 02/11/12
77		Reporte de tecnicas de modelado y de las presunciones asociadas a estos modelos	2 días	jue 18/10/12	vie 19/10/12
80		Modelo construido	10 días	lun 22/10/12	vie 02/11/12
86		Evaluación	6 días	lun 05/11/12	lun 12/11/12
87		Informe de evaluación de los resultados	3 días	lun 05/11/12	mié 07/11/12
91		Informe final del producto	3 días	jue 08/11/12	lun 12/11/12

Figura 2.4: Planificación del proyecto. (Elaboración Propia)

Tabla 2.2: Tabla de Riesgos del Proyecto (Elaboración Propia)

N°	Riesgo	Probabilidad	Impacto	Severidad	Descripción	Planes de mitigación	Planes de Contingencia
1	Poca disponibilidad del tesista.	Baja	Alto	Alta	El tesista puede recargarse de tareas adicionales a las relacionadas al proyecto de fin de carrera.	Realizar una planificación de las actividades del tesista hasta el fin del proyecto de fin de carrera.	Retirarse del curso de tesis 2.
2	La laptop del tesista es robada.	Media	Alto	Alta	El dispositivo móvil (laptop) en el cual se guarda información de las fuentes bibliográficas del proyecto es robada	Guardar toda la información relacionada al proyecto de fin de carrera en un repositorio electrónico.	Revisar la bibliografía de los entregables mandados por correo electrónico al asesor.
3	El software usado en el proyecto cambia, ya hora necesita una licencia de pago para su uso.	Baja	Alto	Alta	El software OpenERP y Pentaho Weka dejan de ser de “código abierto”, y ahora se debe de pagar una licencia para poder usar ambos software.	Adquirir la versión de prueba del software.	Cambiar de herramientas software; buscar nuevas herramientas de “código abierto”.

4	La metodología que se utiliza en el proyecto ha sido cambiada.	Baja	Medio	Media	Las metodologías que uso para mi proyecto han cambiado; es decir, se ha creado una nueva versión de estas.	Detallar cada aspecto de la metodología, y solo cambiar los métodos que uso en mi proyecto.	Leer nuevamente toda la metodología y adecuarla a mi proyecto de fin de carrera.
5	La base de datos que se usa en el proyecto deja de funcionar.	Alta	Alto	Alto	La base de datos del proyecto de fin de carrera deja de funcionar por: mal uso, dispositivo de mala calidad, sobrecalentamiento, entre otros.	Tener un respaldo de información, esto se puede hacer mediante backups.	Empezar nuevamente el registro de datos, con un nuevo disco duro.
6	La persona encargada de brindar información de la empresa esta indispuesta por un largo periodo de tiempo.	Alta	Alto	Alto	La persona encargada de dar información sobre la empresa se va de viaje, se enferma, entre otros.	Realizar entrevistas anticipadas a la persona dispuesta a dar información. Requerir información de los procesos de negocio anticipadamente.	Retirarse del curso de tesis 2.
7	Indisponibilidad del Asesor.	Baja	Medio	Media	El asesor asignado al tesista no está disponible en el curso	Preguntar sobre su disponibilidad al asesor para seguir apoyando al	Cambio de asesor.

					de Tesis2.	tesista.	
8	El proyecto de fin de carrera ya ha siendo implementado por otra entidad u otra persona.	Baja	Alto	Alta	El proyecto de fin de carrera ya ha sido implementado en una empresa textil por una consultora.	Revisar el marco teórico y estado del arte, para poder hacer una variación al proyecto.	Cambiar de tema del proyecto de fin de carrera.
9	La empresa a la que va enfocada el proyecto, quiebra económicamente y se decide su disolución.	Baja	Alto	Alto	La empresa cae en banca rota, y se decide cerrar la empresa.	Revisar el marco teórico y el estado del arte, para poder realizar variaciones al proyecto.	Cambiar de tema del proyecto de fin de carrera.

3 Implantación del ERP

La empresa mencionada en la presente tesis no cuenta con una estructura ordenada para guardar la información que día a día maneja y para poder implementar un proceso de Data Mining se necesitan datos e información de una Base de datos, ya que sobre esto trabaja el proceso algorítmico mencionado.

Entonces, la empresa necesita una solución automatizada que emplee una Base de Datos como fuente de almacenamiento, y según lo mencionado antes en el proyecto, la herramienta a usar debe ser de código abierto; es por esto, que se usará un ERP llamado OpenERP, el cual cumple todos los requisitos para la solución planteada y para el proyecto.

En lo consecuente del capítulo se procederá a describir todo lo relacionado a la implantación del OpenERP en una empresa de prendas de vestir.

3.1 Descripción del Sistema OpenERP

La herramienta OpenERP es una solución ERP que actualmente está en la versión 6.1 en un formato Web, y que además posee los siguientes paquetes:

- CRM
- EInvoicing & Payments
- Points of Sale

- Project management
- Accounting and Finance
- Employee Directory
- Gestión de ventas
- Timesheets Validation
- Gestión de almacenes
- MRP (Planificación de requerimientos de materiales)
- Purchase Management
- Todo lists
- Issues Tracker
- Recruitment Process
- Leaves management
- Expenses Management
- Assets Management
- Payroll
- Evaluación de empleados

Sin embargo, para efectos del proyecto solo se usarán funcionalidades y paquetes que estén relacionados con el módulo de ventas. Entre estos están incluidos los siguientes paquetes:

- Ventas
- Almacén
- Contabilidad
- POS Backend
- Terminal Punto de Venta
- Configuración

Finalmente, para acabar con esta breve descripción de la herramienta OpenERP, solo mencionar que es una solución software que tiene 3 presentaciones: 1 de código abierto "OpenERP Community" y 2 de licencia por pagar "OpenERP Enterprise" y "OpenERP Online".

3.2 ¿Por qué usar OpenERP y por qué el uso de la metodología Agile OpenERP?

La metodología Agile OpenERP se da por iteraciones o los llamados sprints en SCRUM; a similitud de SCRUM está metodología usa bloques funcionales, los cuales son un grupo de funcionalidades que se implantan por iteración, esto ayuda a que el proceso de implantación no utilice muchos recursos humanos a la vez.

La metodología Agile OpenERP es ideal para tipos de proyecto como el actual, ya que se implantará un solo módulo; es decir, un solo grupo de funcionalidades; además, es una metodología hecha a la medida del software OpenERP, ya que está hecha por usuarios de este mismo software, y por último solo un grupo del personal de la empresa se involucrará en el proceso.

El plan de implantación del OpenERP se encuentra en el Anexo A.

3.3 Bloques funcionales por iteración

El único módulo que se implantará en la empresa como ya se mencionó antes, es el módulo de ventas; por lo cual, solo existirá una iteración, la cual se presentará a continuación:

3.3.1 Preparación

El objetivo principal de esta iteración es implantar el módulo de ventas en la organización textil a la cual va enfocada la tesis.

Para realizar lo mencionado antes se procederá a mencionar hitos, los cuales también se encuentran en el Plan de Implantación.

- Entrevista con el personal.
- Contabilidad y clasificación de productos.
- Configuración del módulo de ventas
- Capacitación al personal de ventas de la empresa
- Pruebas del OpenERP

3.3.2 Análisis Técnico GAP

En esta sección se realiza un análisis más detallado de la iteración del módulo de ventas del OpenERP.

En primer lugar, como la implantación abarco solo el módulo de ventas, entonces se realizó el flujo de ventas de la empresa, el cual se muestra a continuación:

(1.1)Presentar productos textiles.- Este proceso se refiere básicamente a la negociación previa que existe en una venta entre el cliente y el vendedor. Básicamente algunos puntos a detallar son: el precio, la cantidad, los descuentos, etc.

(1.2)Registrar venta en documentos físicos.- Se realizará el registro de lo ya negociado (Presentar productos textiles) entre los actores: vendedor y cliente.

(1.3)Registrar cliente.- Se realiza el registro físico del cliente, si este es un nuevo consumidor.

(1.4)Realizar pago.- Se efectúa el intercambio de dinero por los productos vendidos.

(1.5)Enviar productos vendidos.- Se envían los productos solicitados al cliente asignado, si este lo ha pedido así.

De los procesos mencionados en el flujograma solo se automatizaron los siguientes: (1.2) Registrar venta en documentos físicos y (1.3) Registrar Cliente y (1.4) Registrar pago.

En segundo lugar, se realizó un conteo de todos los productos que la empresa tiene en stock; con la lista ya hecha se clasificó todos los productos por sus similares características. Además, se codificaron los productos para un manejo mucho más fácil de información.

Finalmente, se muestra una lista de FB (funcionalidades de negocio llamadas: “Bloques Funcionales”) del OpenERP asociadas a los procesos a automatizar en la Tabla 3.1.

3.3.3 Implementación

La implementación consta de dos etapas: la instalación y la configuración, las cuales se detallan a continuación.

3.3.3.1 Instalación

Para la instalación se consideró la siguiente información previa:

- La computadora en la cual se instaló el OpenERP tenía un procesador AMD Buldozer; además, ya estaba instalado el sistema operativo Windows 7 de 64 bits.
- La computadora contaba con servicio de internet durante todo el día; adicionalmente, poseía el kit básico para su correcto uso; es decir, la PC tenía en Hardware los siguientes componentes: Mouse, Teclado, Monitor LCD, CPU, Parlantes.

Luego se procedió a la descarga de los componentes del OpenERP: Servidor, Cliente Web y Postgres.

Inmediatamente después se instalaron los tres componentes: primero el Servidor, luego el Cliente Web y finalmente el gestor de Base de Datos Postgres. (OpenERP, 2012)

3.3.3.2 Configuración

Para la configuración se consideró la siguiente información previa:

- Se tenía un registro de prendas de vestir, el cual se realizó previo a este proceso en un conteo de todos los productos que se tenían en la empresa en ese entonces. Además, está lista se actualizaba cada vez que la empresa sacaba un nuevo producto al mercado.
- Se realizó una clasificación de los productos por tela, modelo, talla y características que poseía cada prenda de vestir.
- Se realizó una lista de las prendas de vestir con los precios tiene cada una en la actualidad.

Para realizar la configuración se realizaron los siguientes pasos:

Creación de la Base de Datos: Con el OpenERP ya instalado se crea una nueva Base de Datos.

Selección de módulos: Se registra como usuario administrador y luego ya dentro del sistema, se seleccionan los componentes que se desea instalar.

Registro de datos de la Empresa: Se registran todos los datos de la empresa.

Registro de la moneda: Se crea la moneda (Nuevo Sol); además se crea tasas de cambio.

Carga de idioma: Se procede a cargar el idioma por defecto que la empresa y distintos clientes usarán por defecto.

Registro de Categorías de productos: Se registran las categorías de los productos que la empresa vende.

Registro de Productos: Se registran los productos que la empresa vende y sus características.

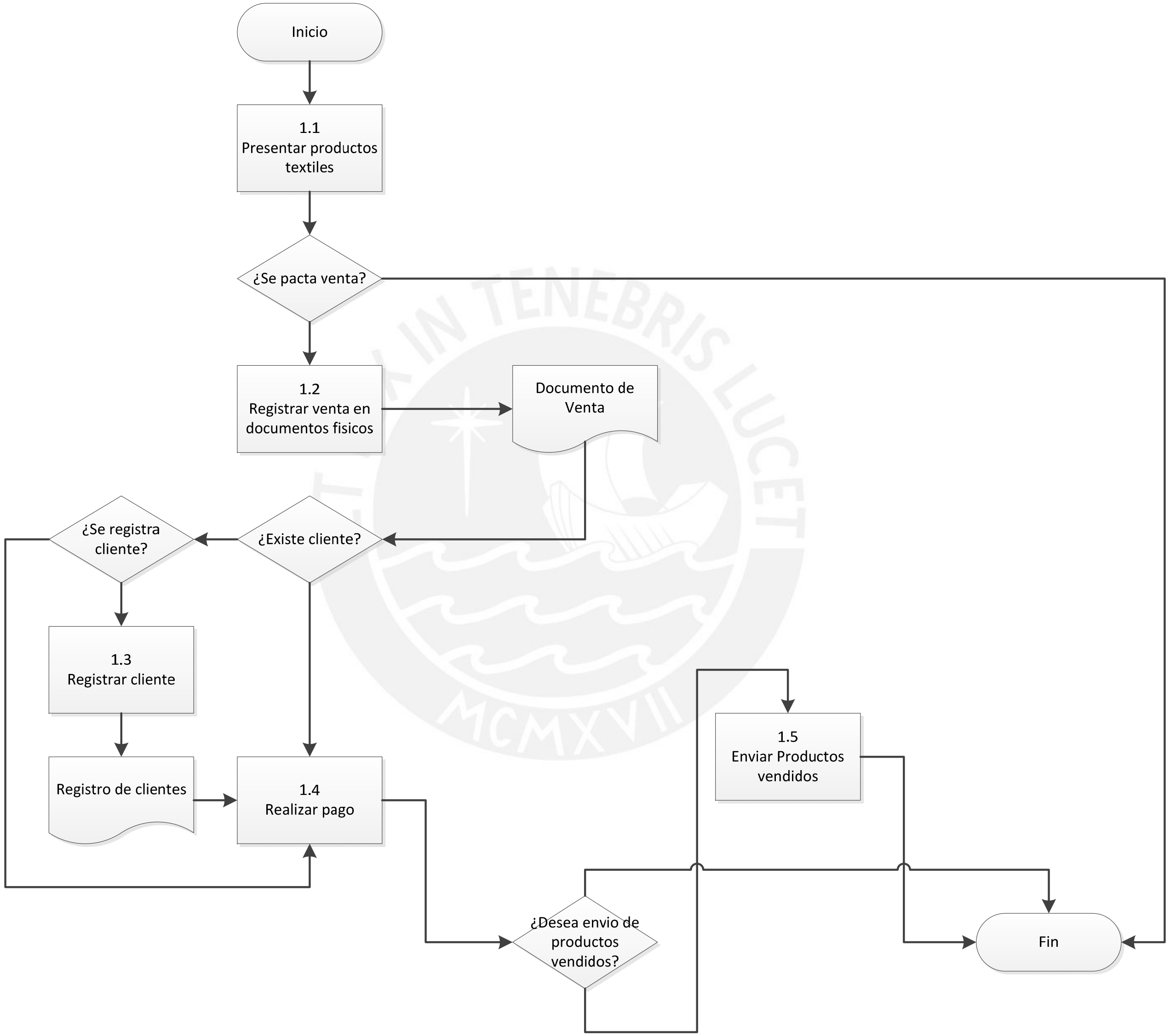


Figura 3.1: Proceso de ventas de la empresa (Elaboración Propia)

Tabla 3.1: Bloques Funcionales (Elaboración Propia)

FB (Bloques Funcionales) del OpenERP	Procesos de la Empresa
Realizar mantenimiento de Productos.	Registrar Venta
Realizar mantenimiento de Categorías de terminales de Puntos de Venta.	Registrar Venta
Realizar mantenimiento de Clientes.	Registrar Clientes
Realizar reportes y manejar indicadores de Ventas.	Registrar Venta
Realizar Ventas por Terminales de Puntos de Venta.	Registrar Venta
Realizar Pedidos de Venta.	Registrar Venta
Abrir y Cerrar registros de Caja registradora.	Registrar Venta
Realizar mantenimiento de usuarios.	Registrar Venta

Registro de Clientes: Se registran todos los clientes que la empresa posee hasta la fecha. Esto se realizó mediante el software OpenERP, ya que permite llenar los datos fácilmente y posee una interfaz amigable. Además, este procedimiento sirve para registro de nuevos clientes en el futuro; con esto se busca que los mismos usuarios puedan cargar futuros clientes por sí solos.

Registro de Usuarios: Se registran los usuarios necesarios con los permisos ya definidos.

Apertura de Cajas: Por último se realiza la apertura de las Cajas registradoras, previa creación de estas, para que se puedan a comenzar a vender y hacer pedidos.

3.4 Carga de Datos

El proyecto consta de 2 fases: la implantación del OpenERP y el proceso de Data Mining, el nexo entre estas dos fases es el uso de la Base de Datos transaccional; mientras que por el lado del OpenERP en la Base de Datos se almacenan datos;

por el otro lado, el proceso de Data Mining utiliza los datos almacenados de las tablas relacionadas a ventas para procesar con un algoritmo la información. Sin embargo, debido al poco tiempo que tiene implantado el ERP en la empresa, los datos almacenados no serán los suficientes para satisfacer el proceso algorítmico; es por esto, que se realizó una carga de datos en la Base de Datos transaccional de manera manual; es decir, venta por venta.

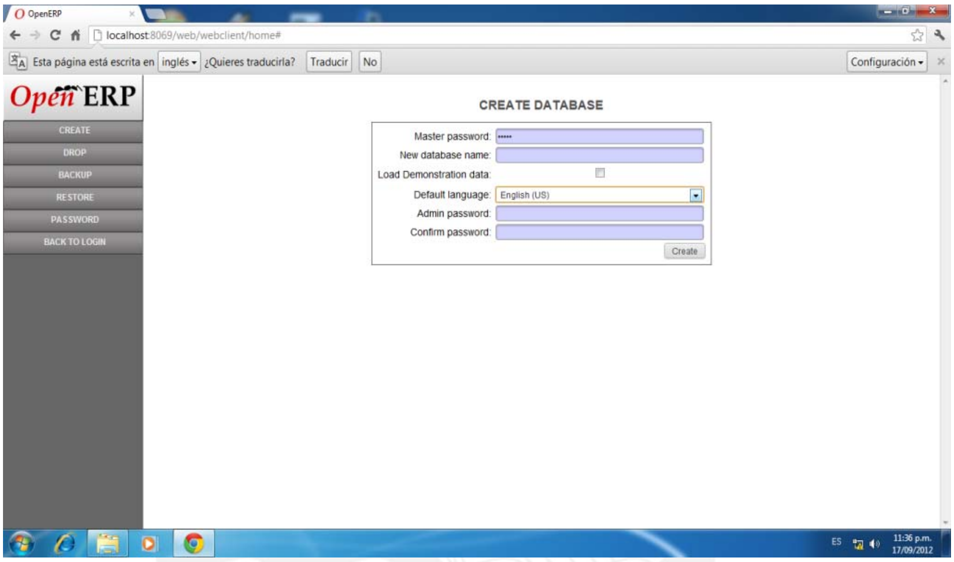


Figura 3.2: Configuración del OpenERP, creación de la Base de Datos

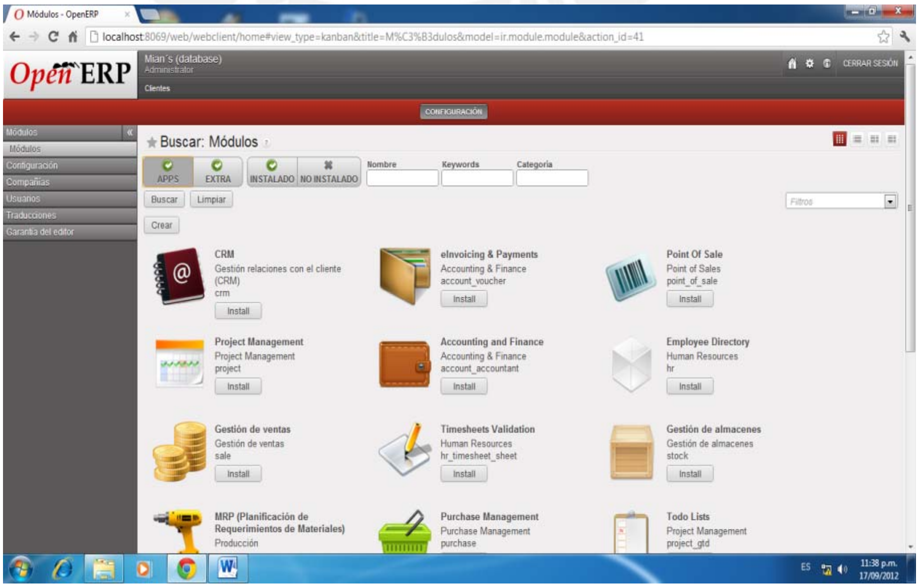


Figura 3.3: Configuración del OpenERP, Selección de módulos

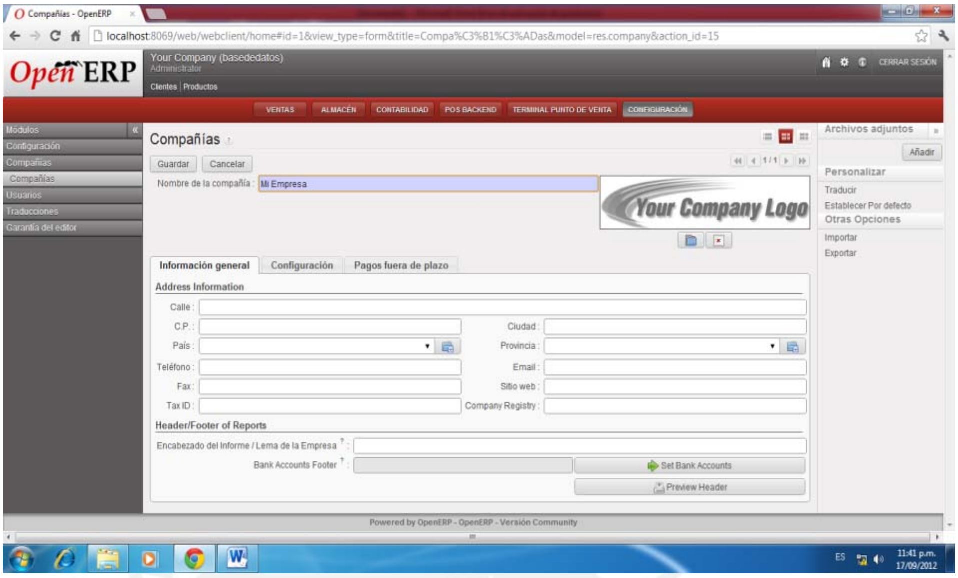


Figura 3.4: Configuración del OpenERP, Registro de los datos de la empresa

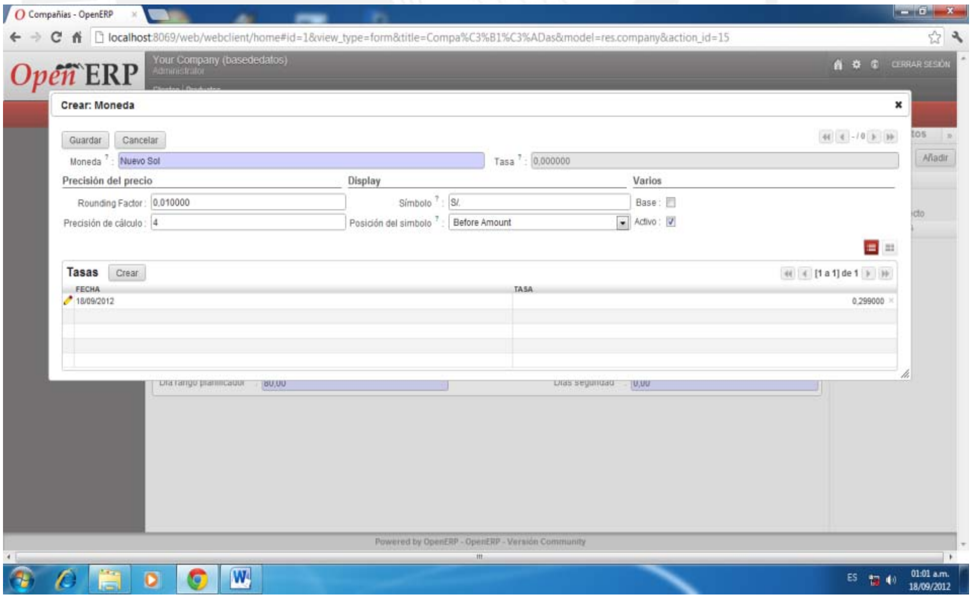


Figura 3.5: Configuración del OpenERP, moneda

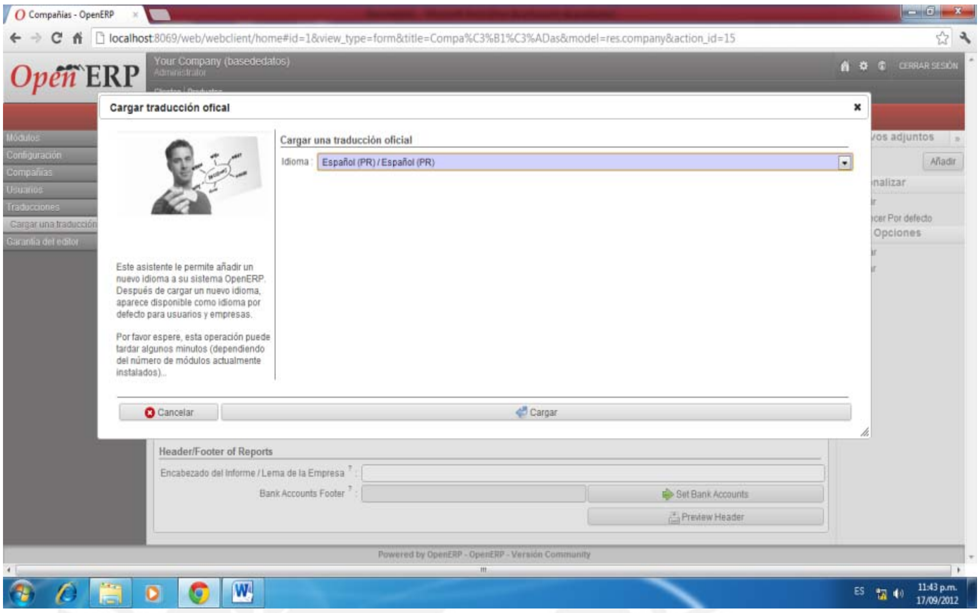


Figura 3.6: Configuración del OpenERP, idioma

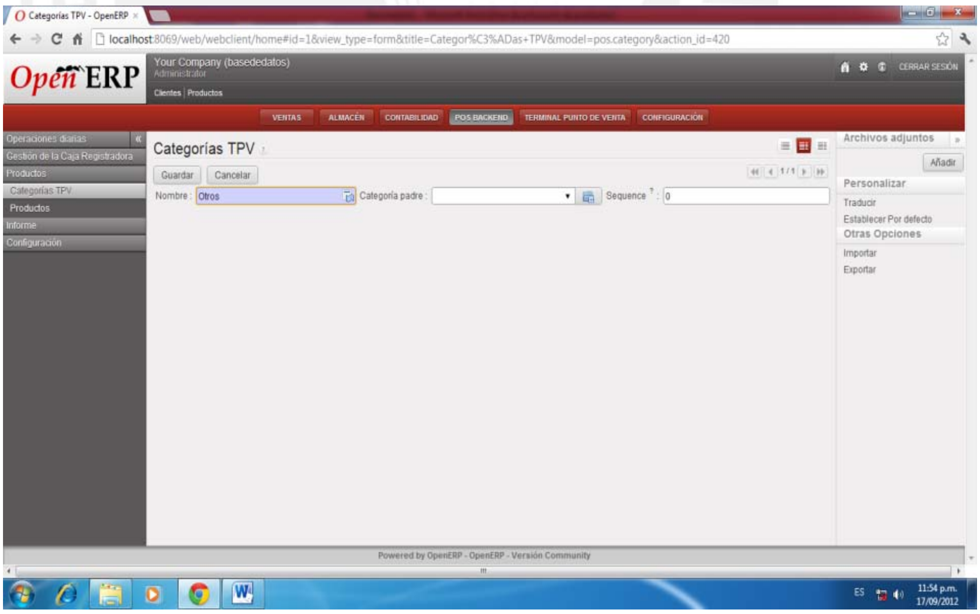


Figura 3.7: Configuración del OpenERP, Tipo de Producto

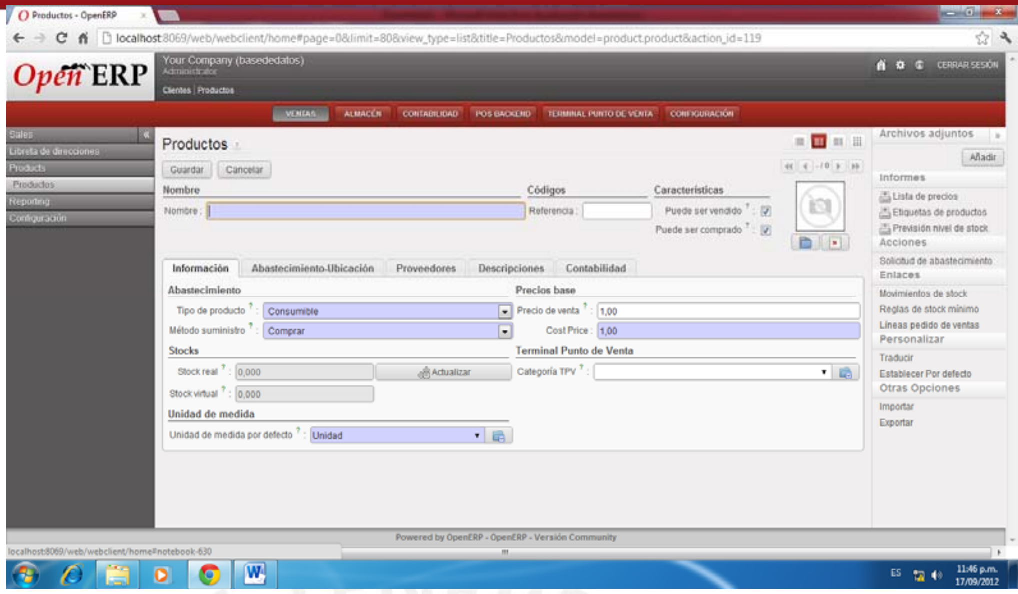


Figura 3.8: Configuración del OpenERP, Producto

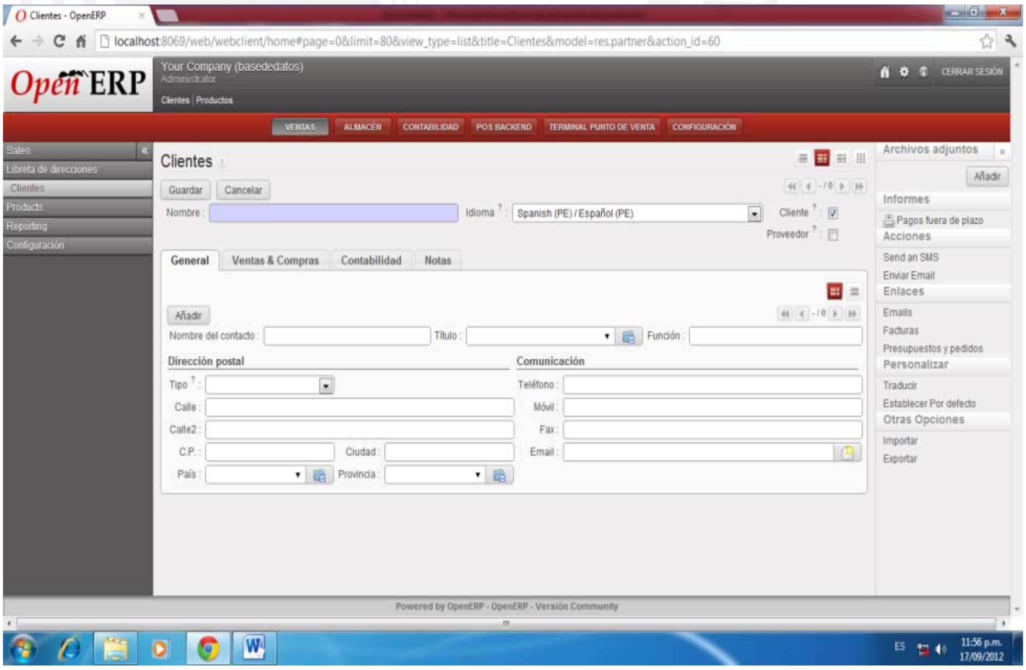


Figura 3.9: Configuración del OpenERP, clientes

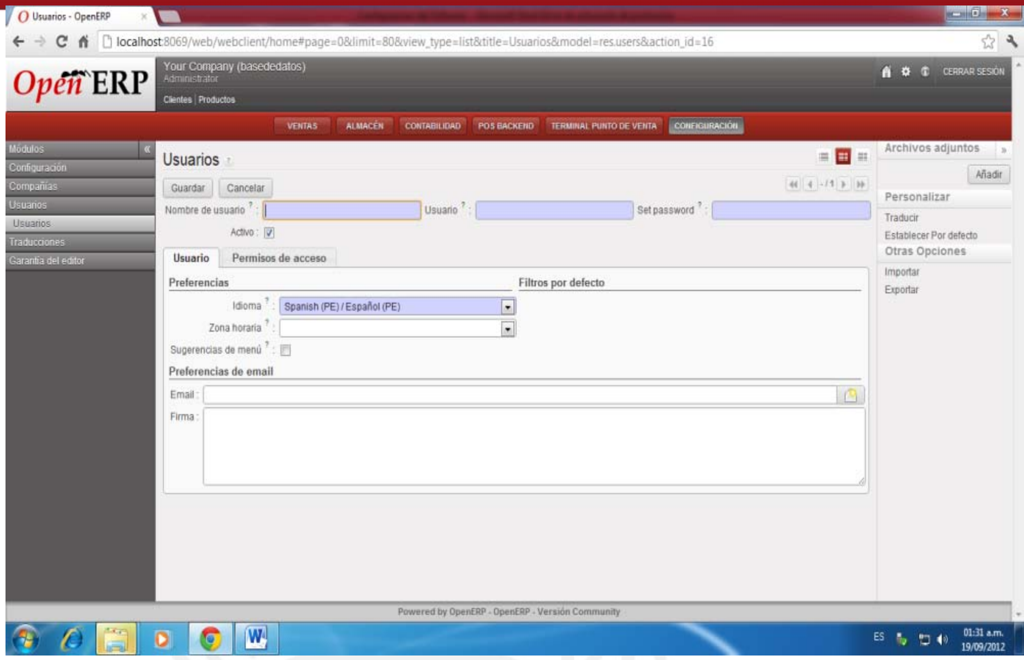


Figura 3.10: Configuración del OpenERP, usuarios (Elaboración Propia)

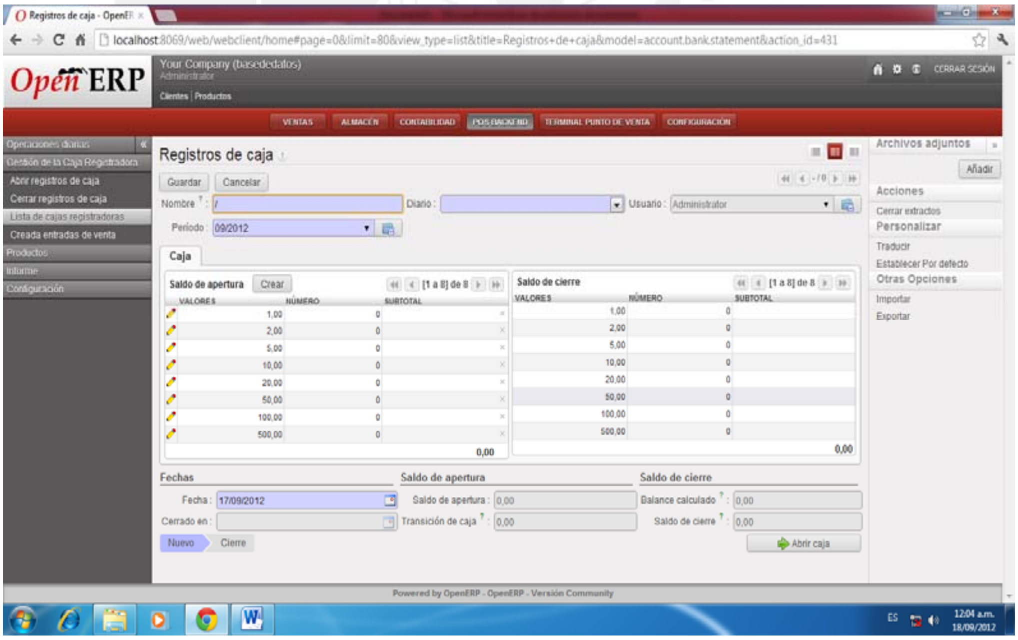


Figura 3.11: Configuración del OpenERP, Caja Registradora (Elaboración Propia)

El acumulado de la carga manual de datos será desde las ventas que se registraron el 1 de Noviembre del año 2011 en la empresa; esta carga se realizó luego de haber terminado de instalar, configurar y capacitar todo lo que respecta al OpenERP.

Para la carga manual se tomarán algunos documentos, ya hechos previamente, tales como la clasificación de los productos, boletas, guías y facturas de venta de la empresa.

El proceso de carga de ventas tiene planificado los siguientes pasos:

- Realizar Backup de la Base de Datos instalada en el servidor de la empresa y generar la réplica.
- Restaurar el archivo backup en una computadora local para su manipulación.
- Buscar las tablas referidas a ventas y analizarlas posteriormente.
- Clasificar los productos (por código) en los documentos de venta físicos.
- Carga de datos a la Base de Datos replica mediante un software implementado en el proyecto.
- Verificar los datos cargados a la Base de Datos réplica.
- Carga de información de la Base de Datos original a la Base de Datos réplica.

Para la carga de datos a la base de Datos réplica se necesitó una aplicación hecha en java mostrada en la Figura 3.12, la cual facilitó la labor de la carga de datos manual. Este software tiene como finalidad guardar las sentencias SQL que se realizan en la carga de datos en archivos de texto; luego estas sentencias fueron ejecutadas en la base de datos réplica.

3.5 Arquitectura física del OpenERP

La arquitectura del OpenERP está conformada por tres componentes:

- Base de Datos (Motor Postgres)
- Servidor
- Cliente (Gtk o Web)

La base de datos guarda toda la información que el usuario ingresa por el software; cabe resaltar que en la base de datos no se usan funciones ni procedimientos, ya que esto se maneja en el mismo servidor/procesador del OpenERP.

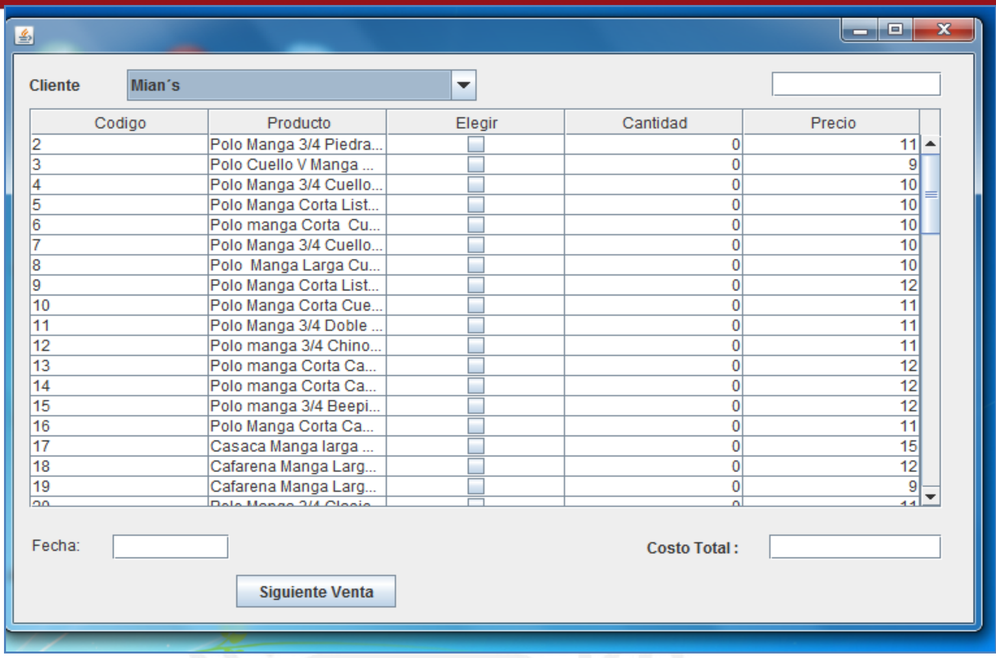


Figura 3.12: Herramienta software para la carga de datos (Elaboración Propia)

Adicionalmente, el OpenERP puede trabajar con diferentes tipos de clientes, ya que todas las funcionalidades están internamente en el servidor. (OpenERP, 2012) La arquitectura se muestra en la Figura 3.13.

3.6 Pruebas

Para las pruebas del OpenERP se pudo usar una herramienta llamada “OERPScenario”; sin embargo, el sistema operativo usado para la instalación del OpenERP es Windows, y la herramienta para las pruebas trabaja exclusivamente con RVM, el cual solo soporta Linux.

Imagen 4: Arquitectura Física del OpenERP

El plan de pruebas asignado a esta sección se encuentra en el Anexo B.

3.7 Capacitación

Los usuarios tienen un rol muy importante en este proceso, ya que son las personas que usarán el software al final de la implantación; es por esto, que se diseñó unas sesiones de capacitación teniendo en cuenta los siguientes factores:

- Las vendedoras poseían un conocimiento muy básico del uso de una computadora.

- Las vendedoras nunca antes habían usado un software de las características al que se implantó en la empresa.
- Se hizo una clasificación previa de los productos para que se puedan encontrar fácilmente estos al realizar una venta.

Las capacitaciones fueron individuales (una vez por persona) y constantes, para que se puedan ir adaptando rápidamente al sistema software. En conclusión, la capacitación duró aproximadamente una semana con un software de prueba y finalmente, para verificar que la capacitación se hizo con éxito, cada personal capacitado firmo un acta de compromiso por la capacitación instruida.

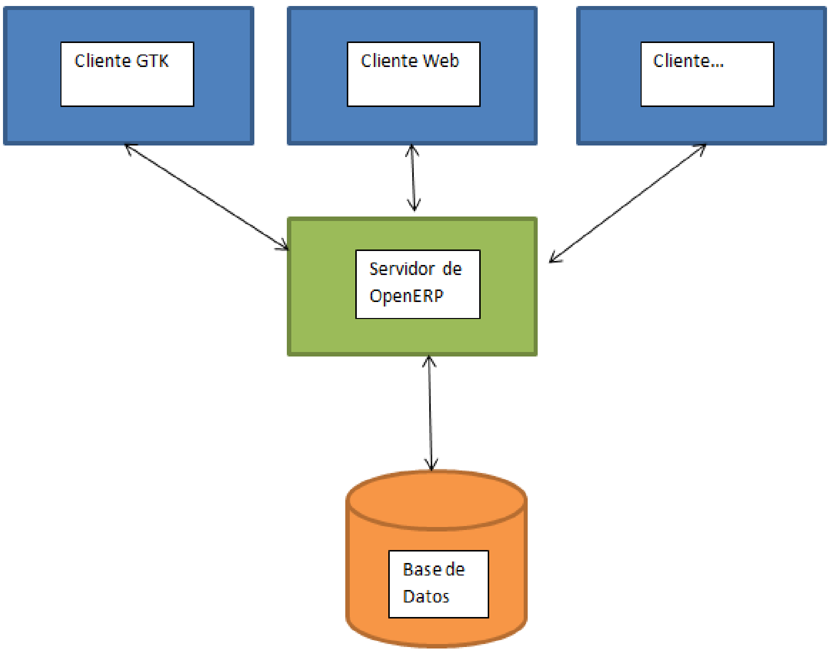


Figura 3.13: Arquitectura física del OpenERP (Elaboración Propia)

4 Data Mining

Hoy en día en la industria textil se usan muchas herramientas tecnológicas, como software para procesos o áreas, también software que abarca toda una empresa; es decir, ERPs, CRMs, MRPs, etc. (ICON-INSTITUT GmbH Private Sector, 2009) Debido a esto, muchos datos se concentran en las bases de datos transaccionales, pero estos no son explotados; es decir, no llegan a ser información que le sirva para hacer frente a la competencia. (ICON-INSTITUT GmbH Private Sector, 2009)

El uso de Data Mining ayudará a que los datos sean más explotados y que la empresa pueda obtener información valiosa de sí misma. Es por esto, que en el presente capítulo se desarrolla todo el proceso de Data Mining para la empresa textil; posteriormente, al final del capítulo se realiza las conclusiones.

4.1 Objetivos del Negocio

Para la metodología CRISP-DM es necesario especificar los objetivos que se tendrán presente en esta parte del proyecto. A continuación se mencionan los objetivos involucrados en el proyecto:

- Mantener a los clientes mayoristas cuando ellos son propensos a moverse a un competidor.

El criterio por el cual se mide el éxito del objetivo mencionado es:

- Intervalo de cantidad de clientes con respecto a meses anteriores.

4.2 Objetivos del Data Mining

Los objetivos del Data Mining suelen relacionarse mucho con los objetivos del negocio; sin embargo, estos son más técnicos.

- Predecir la compra de productos textiles de los clientes, obteniendo datos de sus compras de un año.

4.3 Recolección de datos iniciales

Los datos que se obtienen para el proceso de Data Mining son los que se encuentran registrados en algunas tablas de la base de datos del OpenERP implantado en la empresa; además existe otro conjunto de datos que se obtiene por la carga de datos que se realiza en algunas tablas de una base de datos idéntica en estructura (tablas y relaciones entre estas, más no en registros o filas) a la que usa el OpenERP; este proceso de carga está más detallado en el capítulo anterior en la sección de carga de datos. A continuación se detalla cómo se obtienen los datos paso a paso:

En primer lugar, se realiza una copia de respaldo de la Base de Datos transaccional asociada al ERP usado en la empresa, a esta copia de respaldo se le asigna el nombre “BDBACKUP”. En esta copia “BDBACKUP” se insertan un conjunto de sentencias SQL (carga de datos); estas contienen los registros de las ventas, líneas de ventas y movimientos realizados en el almacén a partir de las ventas; el modo como se realizó la carga de datos se encuentra detallada en el capítulo 3, en la sección de carga de datos.

En segundo lugar, una vez finalizado el periodo dentro del cual estaba previsto guardar los datos que se realizaran para el proceso del Data Mining (1 de Abril al 31 de octubre), se realiza una nueva copia de respaldo de la Base de Datos usada por la empresa; a esta copia se le asignará un nuevo nombre “BDBACKUP2”.

En tercer lugar, se ingresan en la Base de Datos “BDBACKUP2” todos los registros de datos guardados por carga de datos en la Base de Datos “BDBACKUP”.

Finalmente, una vez que en la Base de Datos “BDBACKUP2” se encuentren todos los registros necesarios para el posterior proceso de Data Mining, entonces se procede a seleccionar los datos que son utilizados en el pre-modelado; es así que se crea una nueva tabla “VENTAS”, la cual contiene solo los posibles datos que serán usados en el proceso algorítmico de Data Mining.

Un pequeño conjunto de datos de la tabla “VENTAS” se muestran en la Tabla 4.6.

Los datos de la tabla “VENTAS” guardan un registro de exactamente 12 meses; desde el 1 de noviembre del 2011 al 31 de octubre del 2012.

Los datos incluidos en la tabla “VENTAS” fueron elegidos sobre otros datos de las tablas del ERP (pos_order, pos_order_line, res_partner) por el hecho de estar relacionados directamente con los productos que la empresa vende, con las ventas y clientes de la empresa.

Las tablas que utiliza el OpenERP son en total 343; sin embargo se usan 3 tablas específicamente en el área de ventas que sirven para el posterior proceso de Data Mining: POS_ORDER, POS_ORDER_LINE y RES_PARTNER.

Tabla 4.1: Tabla de la base de datos del OpenERP, POS_ORDER (Elaboración Propia)

Tabla	POS_ORDER
id	Identificador de una venta
create_uid	Identificador del usuario del OpenERP
create_date	Día del registro de la venta
write_date	Día del registro de la venta
write_uid	Identificador del usuario del OpenERP
sale_journal	Identificador de la caja de ventas
account_move	cuenta de movimiento (si la operación se hace por banco)
date_order	Día del registro de la orden de la venta

Tabla	POS_ORDER
partner_id	Identificador del cliente
nb_print	Identificador del Voucher de impresión
user_id	identificador del usuario del OpenERP
name	Nombre de la orden de venta
invoice_id	Identificador de la factura
company_id	Identificador de la compañía
note	Nota de texto
state	Estado de la venta (Pagado, Cancelado, Reservado)
shop_id	Identificador de compras
pricelist_id	Identificador de precio de lista
picking_id	Identificador del movimiento en inventario

Tabla 4.2: Tabla de la base de datos del OpenERP, POS_ORDER_LINE (Elaboración Propia)

Tabla	POS_ORDER_LINE
id	Identificador de la línea de venta
create_uid	Identificador del usuario que realizo la línea de venta
create_date	Día de la venta
write_date	Día de la venta
write_uid	Identificador del usuario que realizo la línea de venta

Tabla	POS_ORDER_LINE
notice	Notas de la línea de venta
product_id	Identificador del producto
product_name	Nombre del producto
order_id	Identificador de la venta
price_unit	Precio unitario
price_subtotal	Subtotal de la línea de venta (cantidad por el precio)
company_id	Identificador de la compañía
price_subtotal_incl	Subtotal modificada de la línea de venta (cantidad por el precio menos el descuento)
qty	Cantidad del producto
discount	Descuento del producto en la línea de venta
name	Nombre de la línea de orden de venta

Tabla 4.3: Tabla de la base de datos del OpenERP, RES_PARTNER (Elaboración Propia)

Tabla	RES_PARTNER
id	Identificador del cliente
create_uid	Identificador del usuario que creo el registro del cliente
create_date	Día de creación del registro
write_date	Día de modificación del registro del

Tabla	RES_PARTNER
	cliente
write_uid	Identificador del usuario que creo el modifico el registro del cliente
comment	Comentarios
color	Color
date	Día creación del registro del cliente
active	Estado del cliente (Activo/Desactivo)
customer	Si es cliente o no
credit_limit	Límite de crédito
user_id	Identificador del usuario que creo el registro del cliente
name	Nombre del cliente
title	Título del cliente (Sr. Sra. O Srta.)
company_id	Identificador de la compañía
website	Página web de la compañía
employee	Si es empleado o no
suplier	Si es proveedor o no
debit_limit	Límite de línea de débito
state	Provincia de residencia del cliente

De los datos expuestos en los cuadros: Tabla 4.1, Tabla 4.2, Tabla 4.3 solo se escogen los datos mostrados en las columnas del cuadro Tabla 4.4

Tabla 4.4: Tabla de las columnas usadas de las tablas 4.1, 4.2 y 4.3 para la selección de datos (Elaboración Propia)

Tabla	Columna	Descripción
RES_PARTNER	Id	Identificador del cliente
RES_PARTNER	Name	Nombre del cliente
RES_PARTNER	State	Provincia de residencia del cliente
POS_ORDER_LINE	product_name	Nombre del producto
POS_ORDER	Id	Identificador de la venta al cliente
POS_ORDER_LINE	create_date	Día de la venta
POS_ORDER_LINE	Qty	Cantidad de productos en una línea de venta
POS_ORDER_LINE	price_subtotal_incl	Subtotal de línea de venta (cantidad por precio)

También se usan sentencias SQL para poder obtener el porcentaje de clientes total de la base de datos transaccional del ERP como se muestra en el cuadro.

A partir del Tabla 4.4 se obtiene la lista de clientes por cada departamento**¡Error! La autoreferencia al marcador no es válida.**.

Tabla 4.5: Lista de clientes por departamento y su respectivo porcentaje (Elaboración Propia)

Departamento	Cantidad	Porcentaje
Lambayeque	19	0.2

Amazonas	2	0.02
Puno	1	0.01
Ancash	1	0.01
Piura	10	0.1
Arequipa	10	0.1
San Martin	5	0.05
Tacna	7	0.07
Ica	1	0.01
Lima	1	0.01
Madre de Dios	2	0.02
La Libertad	11	0.11
Cusco	9	0.09
Loreto	3	0.03
Cajamarca	9	0.09
Huánuco	2	0.02
Tumbes	2	0.02
Ayacucho	1	0.01
Apurímac	1	0.01

4.4 Descripción de los datos

Los campos: Tipo Tela, Manga, Tela, Talla, Modelo y Adicional de la **¡Error! La autoreferencia al marcador no es válida.** son extraídos del campo product_name del

La única fuente de datos que se usa a partir de aquí es la tabla "VENTAS", la cual contiene 3513 filas de registros de ventas, también creada a partir de los datos de las tablas: Tabla 4.1, Tabla 4.2 y Tabla 4.3. La tabla “VENTAS” se muestra en la Tabla 4.6.

En la Tabla 4.6 se muestran los siguientes campos:

Código.- Es el identificador de cada cliente, este se encuentra en un formato numérico (integer: numero entero).

Cliente.- Es el nombre de cada persona natural o jurídica que realiza compras en la empresa, en este campo se detallan los nombres de las personas y sus apellidos, si son personas naturales; caso contrario, se tiene el nombre jurídico de la persona, o empresa. Este campo está en formato tipo String; es decir, cadena de caracteres.

Departamento.- Es el nombre del estado de la república del Perú, donde se encuentra la residencia actual del cliente. El campo está en formato tipo String; es decir; cadena de caracteres.

Venta.- Es el identificador de un registro de venta efectuado en la empresa. Este campo está en formato Integer (números enteros).

Tipo prenda.- Es el tipo de ropa que la empresa vende; es decir, casacas, polos, cafarenas y capuchas. Este campo está en formato String, y básicamente solo nos brinda información de la cadena de texto.

Manga.- Se refiere a la parte de las prendas que cubren las extremidades superiores del ser humano; es decir, los brazos y antebrazos. Existen 4 tipos de mangas: Larga, tres cuartos, corta y Cero.

Tela.- La empresa en la actualidad usa distintos tipos de tela, entre estas tenemos: Pima, Rip, Viscosa, Hidrosedal, Fresh Terry, Full Licra, Turbo y Cuadrille. En este campo solo se describe el nombre de la tela y este está en formato String: cadena de texto.

Talla.- Solo se menciona la talla de la prenda de vestir; las tallas son: L (Grande), XL (extra-grande), M (mediano), S (Estándar) y Niñas (Tallas para niñas). Este campo esta como String: cadena de texto; solo se mencionan las tallas.

Modelo.- Describe a un conjunto de prendas de vestir que tiene el mismo diseño, pero con algunas ligeras variaciones. Entre los modelos de prendas de vestir de la empresa se encuentran: Modelo Clásico, Modelo Camisero, Modelo Capa, etc. Este campo esta como String: cadena de texto; es decir solo se mencionan los modelos de cada prenda de vestir.

Adicional.- Característica de la prenda de vestir que resalta y la hace diferente de otras prendas en el mismo modelo o diferente de otras prendas de vestir de todas las que se encuentran en la empresa. Este campo se encuentra como String: Cadena de texto; es decir, solo se mencionan las características de las prendas de vestir; cabe resaltar que no todas las prendas de vestir poseen este campo lleno.

Fecha.- Es la fecha en la que se produjo la venta del producto al cliente. Está en formato date: formato de fecha e incluye el día mes y año.

Cantidad.- Es el número prendas de vestir que se adquirió de un determinado producto. Este campo es un entero y nunca es nulo; además, su rango suele estar entre 1 y 100 por cada fila.

Subtotal.- Es la sumatoria total del precio por la cantidad de prendas de vestir adquiridas por cada producto. Este campo es un float: soporta números enteros y no enteros; su rango suele estar entre 0 y 1000.00.

4.5 Exploración de los datos

En esta sección se realizan gráficos para ver las estadísticas de los datos que se tienen para el proceso algorítmico posterior de Data Mining.

Los gráficos que se presentan en las figuras: Figura 4.2, Figura 4.3 muestran la cantidad de comprada por cada prenda de vestir en los meses de Abril y Mayo.

4.6 Verificar la calidad de los datos

Para validar la calidad de los datos se realizan las siguientes preguntas:

¿Los datos de la Base de Datos están completos?

Los datos usados en la tabla "VENTAS" contienen todo lo necesario para lograr el objetivo final del Data Mining, ya que poseen tanto los detalles de venta, las ventas realizadas, los montos gastados, la cantidad vendida, los productos y por último las fechas en las que se realiza cada operación.

¿Son correctos?, ¿o estos contienen errores? y, ¿si hay errores, que tan

Comunes son estos?

Solo se han encontrado errores en los valores de la columna talla, ya que se repiten algunos valores tales como las tallas S y S”. Este error no es tan común, ya que solo está presente en un producto y en las ventas asociadas a este.

¿Hay valores omitidos en los datos?

Si existen valores omitidos en la tabla “VENTAS”, los cuales se presentan en la columna ADICIONAL.

¿Cómo se representan estos,

Donde ocurre esto, y que tan comunes son estos?

Estos valores se presentan como valores vacíos de cadenas de texto; estos valores son comunes ya que se presentan por cada venta en la que se encuentra incluida una prenda de vestir que no tiene el campo ADICIONAL diferente de la cadena vacía.

Los valores que se encuentran en la Base de Datos relacionados a la columna ADICIONAL no contribuyen directamente a los resultados del Data Mining, ya que no se considera a este una variable de decisión del problema planteado: “Pérdida de clientes”.

4.7 Seleccionar los datos

A partir de esta sección solo se consideraron los datos que se usan en el modelo de Data Mining; es decir, las variables de decisión que se usan para la predicción del comportamiento de clientes.

Como el objetivo principal del Data Mining es predecir el comportamiento de los clientes, entonces los clientes serán una de las variables de decisión y todo lo que este asociado a este directamente; es decir, las variables que fueron tomadas en cuenta son la provincia, el nombre del cliente y el identificador del cliente; además de estas se usaron las variables referidas a las ventas: cantidad y subtotal. Y por último, como las variables están ligadas a los productos, entonces se consideró solo Tela, ya que este atributo define si un cliente adquiere los productos porque busca precios bajos, telas de baja calidad, o si busca mejor calidad en la tela.

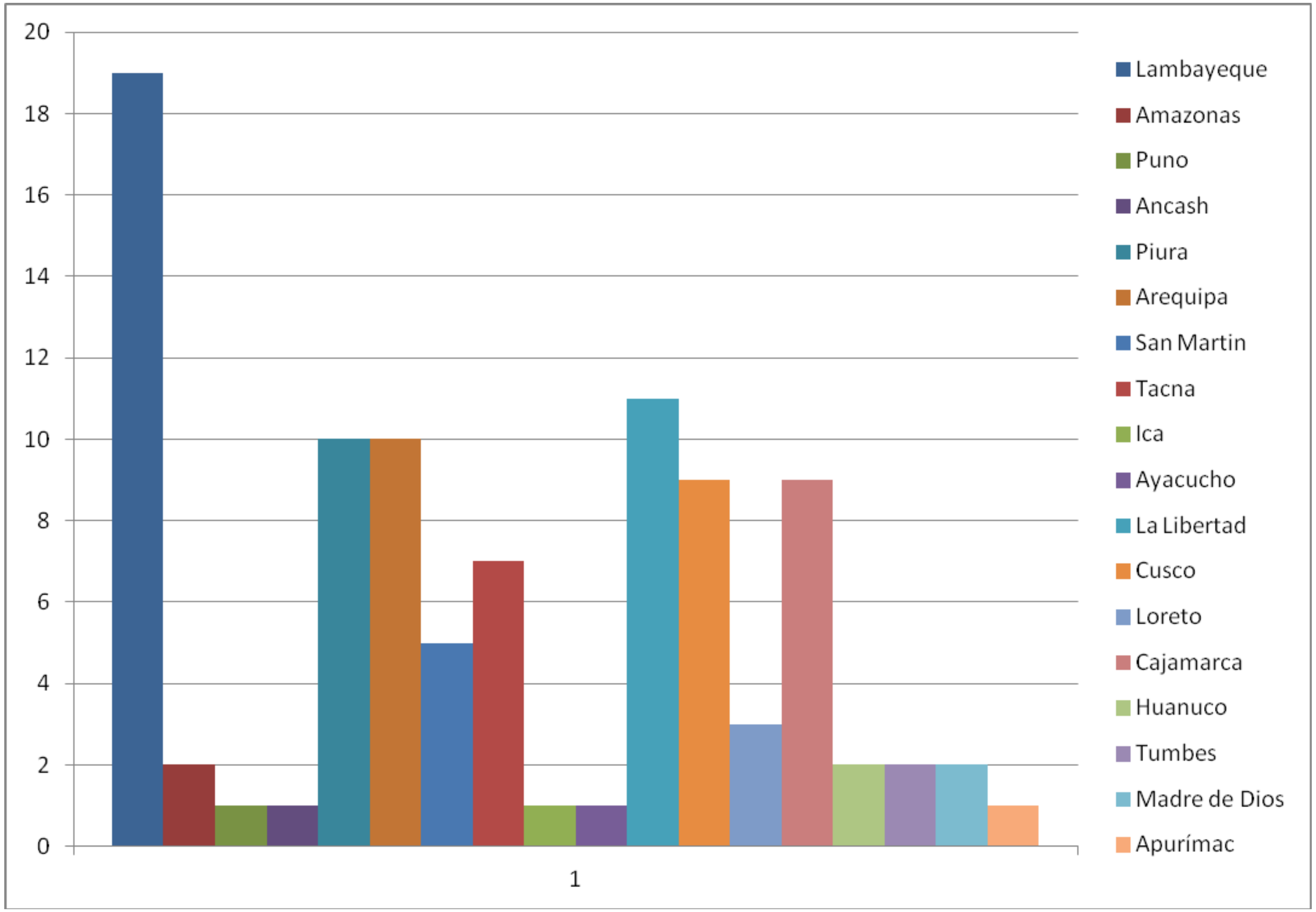


Figura 4.1: Clientes por departamento o provincia (Elaboración Propia)

Tabla 4.6: Lista de ventas por producto de clientes mayoristas de la empresa (Elaboración Propia)

Código	Cliente	Departamento	Venta	Tipo Prenda	Manga	Tela	Talla	Modelo	Adicional	Fecha	Cantidad	Sub total
54	Jessica Cinthia Gutierrez Cahuana	Huánuco	24012	Polo	Corta	VISCOSA	L	Listado		02/04/2012	15	150
52	Liliana Requejo Segovia	Lambayeque	24013	Polo	tres cuartos	PIMA	M	Camisero		02/04/2012	20	220
52	Liliana Requejo Segovia	Lambayeque	24013	Polo		PIMA	L	Encaje		02/04/2012	20	200
52	Liliana Requejo Segovia	Lambayeque	24013	Polo	Corta	PIMA	L	Encaje		02/04/2012	10	100

Código	Cliente	Departamento	Venta	Tipo Prenda	Manga	Tela	Talla	Modelo	Adicional	Fecha	Cantidad	Sub total
52	Liliana Requejo Segovia	Lambayeque	24013	Polo	tres cuartos	VISCOSA	L		Cuello:V: Bolsillo	02/04/2012	30	300
55	Centro Comercial Torres Barboza Hnos SAC	Piura	24014	Polo	Corta	PIMA	XL	Clásico		02/04/2012	20	260
55	Centro Comercial Torres Barboza Hnos SAC	Piura	24014	Cafarena	Larga	RIP	S			02/04/2012	25	250

Código	Cliente	Departamento	Venta	Tipo Prenda	Manga	Tela	Talla	Modelo	Adicional	Fecha	Cantidad	Sub total
55	Centro Comercial Torres Barboza Hnos SAC	Piura	24014	Polo	Corta	FULL LICRA	S		Cuello:Cuadrado	02/04/2012	15	165
55	Centro Comercial Torres Barboza Hnos SAC	Piura	24014	Polo	Corta	FULL LICRA	S	Clásico		02/04/2012	20	200

Código	Cliente	Departamento	Venta	Tipo Prenda	Manga	Tela	Talla	Modelo	Adicional	Fecha	Cantidad	Sub total
55	Centro Comercial Torres Barboza Hnos SAC	Piura	24014	Polo	Cero	FULL LICRA	S	Olimpico		02/04/2012	25	250
55	Centro Comercial Torres Barboza Hnos SAC	Piura	24014	Polo	Corta	PIMA	S	Camisero		02/04/2012	10	100

4.8 Limpieza de los datos

En esta sección se eleva la calidad de datos, para esto se seleccionó los datos que puedan tener fallas o causar fallas en el proceso algorítmico de Data Mining; luego se procedió a solucionar estos problemas con técnicas sencillas o con técnicas más complejas propias del Data Mining. En este caso en particular no se detectaron grandes fallas; sin embargo, si hubo una y se detalla a continuación.

Como ya se mencionó antes en la tabla “VENTAS” se encuentra algunas incongruencias como las tallas: " L"," M","ÑA","XL","SA"," S","S"". En esta clasificación se puede apreciar que existen dos tallas repetidas “S” y “S”” solo que en una va aumentada un carácter adicional, las comillas dobles. Para resolver este problema se actualiza en toda la tabla “VENTAS” los valores que tengan el atributo “S”” y se los convierte en “S”.

La razón por la que no hubo demasiadas fallas en los datos fue por las pocas tablas que se usaron para formar la tabla “VENTAS”; es decir, si en algunas de esas tablas hubiese habido problemas por valores nulos, cadenas de texto vacías, números fuera de rango entonces hubiese causado ruido en el proceso algorítmico.

4.9 Construcción de los datos

Esta parte es una de las más importantes en el Data Mining, ya que se define las variables que se usan para el proceso algorítmico.

De acuerdo a los objetivos del Data Mining se busca la predicción del comportamiento de los clientes; por otro lado, según el proyecto se busca evitar la pérdida de clientes; entonces lo que en el modelo se busca es identificar la pérdida de clientes con las variables que se poseen; para que se logre esto se plantea la siguiente estrategia.

De los 12 meses que se tienen de periodo de prueba se seleccionan 7 muestras; en cada una de estos periodos de pruebas se posee un cuatrimestre histórico y un bimestre en el que se pronostica el comportamiento de acuerdo al histórico. Se plantean dos meses de pronósticos, ya que se define a un cliente en fuga de la empresa si es que este no ha realizado compras en la empresa por más de dos meses. Entonces se analizan las siete muestras para poder predecir el comportamiento de los clientes en los siguientes dos meses de la última muestra; es decir: noviembre y diciembre del 2012.

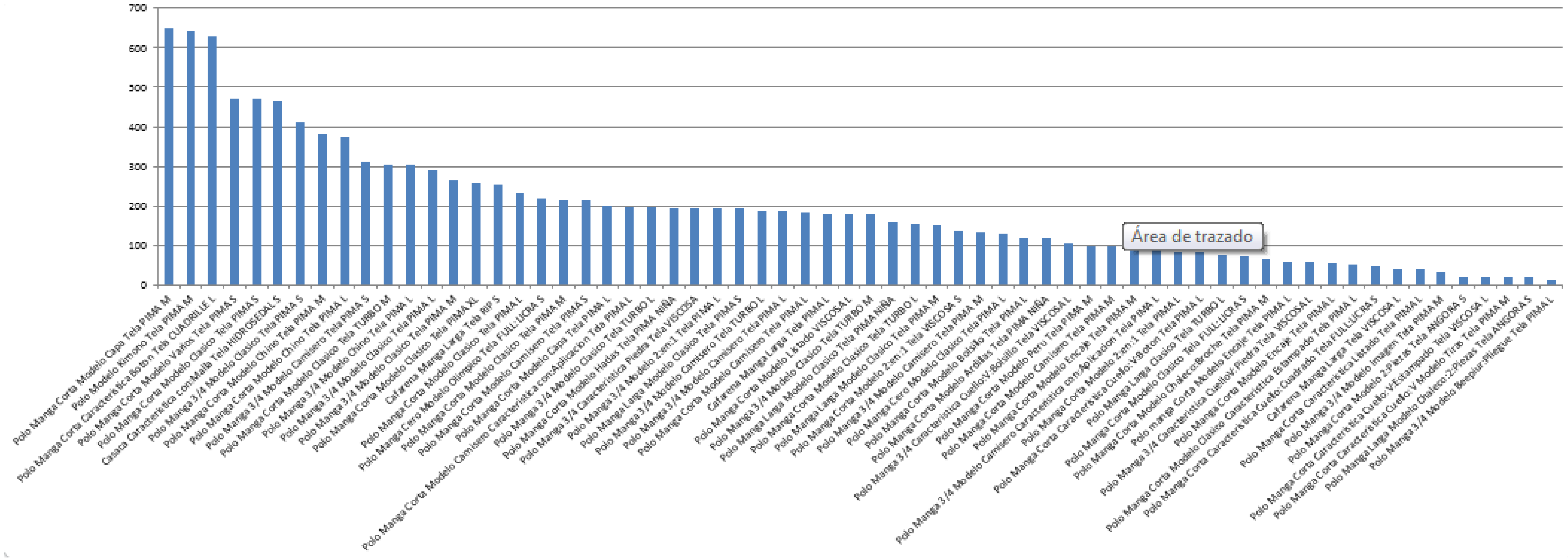


Figura 4.2: Gráfico de ventas realizadas en el mes de Abril por producto (Elaboración Propia)

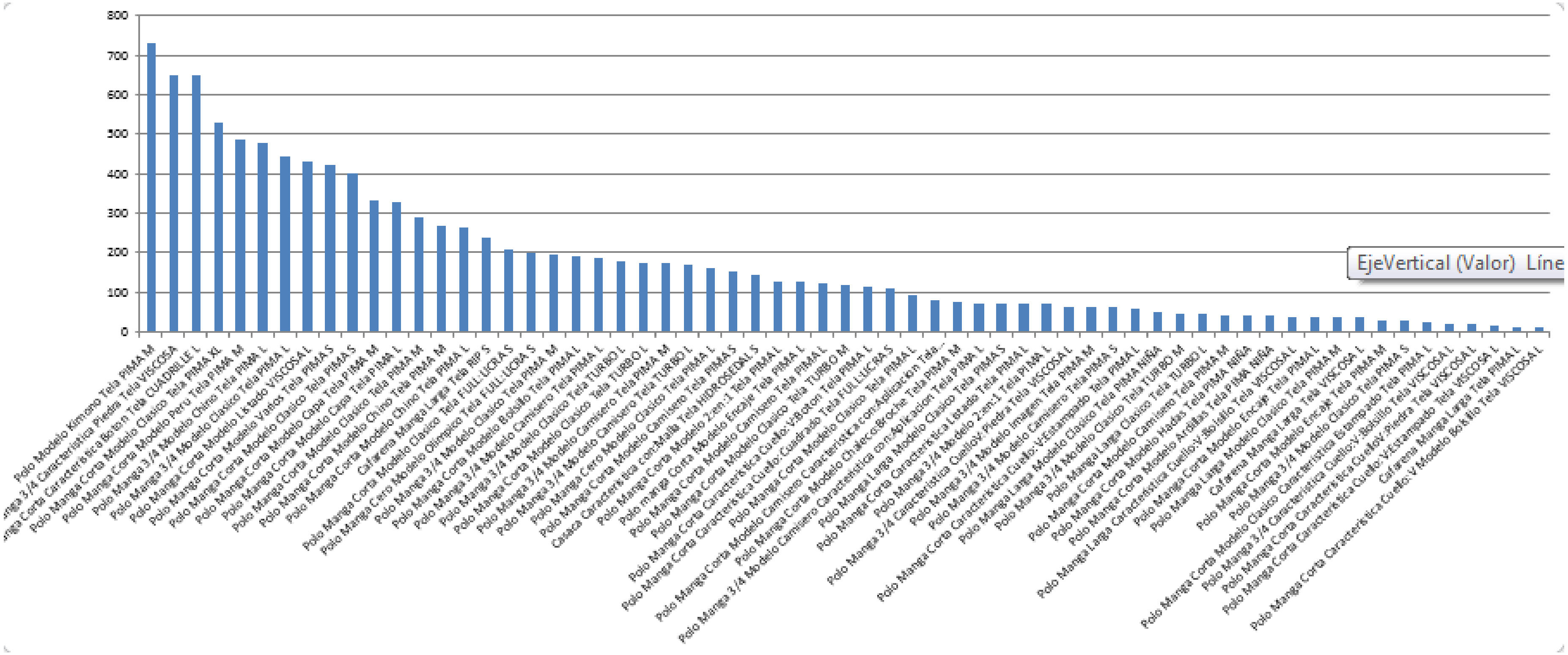


Figura 4.3: Gráfico de ventas realizadas en el mes de Mayo por producto (Elaboración Propia)

Entonces en estas siete muestras se designan las muestras tal como se detalla en la Tabla 4.7, entonces se entiende por este gráfico que en la Muestra 1, los meses Abril, Mayo y Junio son los meses históricos y los meses a pronosticar son el mes de Julio y Agosto; tal como se indicó inicialmente si un cliente no realiza compras en dos meses, para este caso los meses P0 y P1; entonces la empresa ha perdido el cliente. Para la muestra 2 y 3 funciona de la misma forma mencionada anteriormente.

Tabla 4.7: Meses de prueba (Elaboración Propia)

Meses Históricos				Meses Pronostico		
H0	H1	H2	H3	P0	P1	Muestras
Noviembre	Diciembre	Enero	Febrero	Marzo	Abril	Muestra 1
Diciembre	Enero	Febrero	Marzo	Abril	Mayo	Muestra 2
Enero	Febrero	Julio	Abril	Mayo	Junio	Muestra 3
Febrero	Marzo	Abril	Mayo	Junio	Julio	Muestra 4
Marzo	Abril	Mayo	Junio	Julio	Agosto	Muestra 5
Abril	Mayo	Junio	Julio	Agosto	Septiembre	Muestra 6
Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Muestra 7

Ahora se seleccionan las variables derivadas y no derivadas que serán usadas en el proceso de Data Mining; estas variables se obtienen de la tabla “VENTAS” y mediante una sentencia SQL se seleccionan las variables que se utilizaran; estas variables se muestran en la Tabla 4.8. De estas variables solo las que ocasionan ruido son ID_PROVEEDOR y PROVEEDOR, entonces las demás variables son de utilidad para el proceso algorítmico de Data Mining.

Tabla 4.8: Variables construidas para utilizar en el proceso algorítmico de Data Mining (Elaboración Propia)

NOMBRE	DESCRIPCION	FUENTES	FORMATO
ID_PROVEEDOR	CODIGO UNICO CLIENTE	ESQUEMA.VENTAS	INTEGER
PROVEEDOR	NOMBRE DEL CLIENTE	ESQUEMA.VENTAS	VARCHAR(200)
PROVINCIA	RESIDENCIA DEL CLIENTE	ESQUEMA.VENTAS	FLOAT
SUBTOTAL_H3	MONTO COMPRADO EN EL MES HISTORICO 3	ESQUEMA.VENTAS	FLOAT
SUBTOTAL_H2	MONTO COMPRADO EN EL MES HISTORICO 2	ESQUEMA.VENTAS	FLOAT
SUBTOTAL_H1	MONTO COMPRADO EN EL MES HISTORICO 1	ESQUEMA.VENTAS	FLOAT
SUBTOTAL_P0	MONTO COMPRADO EN EL MES PRONOSTICADO 0	ESQUEMA.VENTAS	FLOAT
SUBTOTAL_P1	MONTO COMPRADO EN EL MES PRONOSTICADO 1	ESQUEMA.VENTAS	FLOAT

NOMBRE	DESCRIPCION	FUENTES	FORMATO
SUBTOTAL_TRIM_PROM	PROMEDIO DEL MONTO COMPRADO EN LOS 3 MESES HISTORICOS	ESQUEMA.VENTAS	FLOAT
FREQ_MES_H3	NUMERO DE COMPRAS EN EL MES HISTORICO H3	ESQUEMA.VENTAS	INTEGER
FREQ_MES_H2	NUMERO DE COMPRAS EN EL MES HISTORICO H2	ESQUEMA.VENTAS	INTEGER
FREQ_MES_H1	NUMERO DE COMPRAS EN EL MES HISTORICO H1	ESQUEMA.VENTAS	INTEGER
FREQ_MES_P0	NUMERO DE COMPRAS EN EL MES HISTORICO P0	ESQUEMA.VENTAS	INTEGER
FREQ_MES_P1	NUMERO DE COMPRAS EN EL MES HISTORICO P1	ESQUEMA.VENTAS	INTEGER
TRIM_FREQ_PROM	NUMERO DE COMPRAS PROMEDIO ENTRE LOS 3 MESES HISTORICOS	ESQUEMA.VENTAS	FLOAT

NOMBRE	DESCRIPCION	FUENTES	FORMATO
Y	DECISION DE CLIENTE EN FUGA O CLIENTE AUN FIDELIZADO	ESQUEMA.VENTAS	VARCHAR(50)
PROM_MES_H3	PROMEDIO DEL MONTO DE LAS VENTAS EN EL MES H3	ESQUEMA.VENTAS	FLOAT
PROM_MES_H2	PROMEDIO DEL MONTO DE LAS VENTAS EN EL MES H2	ESQUEMA.VENTAS	FLOAT
PROM_MES_H1	PROMEDIO DEL MONTO DE LAS VENTAS EN EL MES H1	ESQUEMA.VENTAS	FLOAT
PROM_MES_P0	PROMEDIO DEL MONTO DE LAS VENTAS EN EL MES P0	ESQUEMA.VENTAS	FLOAT
PROM_MES_P1	PROMEDIO DEL MONTO DE LAS VENTAS EN EL MES P1	ESQUEMA.VENTAS	FLOAT
MAX_SUBTOTAL_MES_H3	MAXIMO MONTO DE VENTAS EN EL MES H3	ESQUEMA.VENTAS	FLOAT

NOMBRE	DESCRIPCION	FUENTES	FORMATO
MAX_SUBTOTAL_MES_H2	MAXIMO MONTO DE VENTAS EN EL MES H2	ESQUEMA.VENTAS	FLOAT
MAX_SUBTOTAL_MES_H1	MAXIMO MONTO DE VENTAS EN EL MES H1	ESQUEMA.VENTAS	FLOAT
MAX_SUBTOTAL_MES_P0	MAXIMO MONTO DE VENTAS EN EL MES P0	ESQUEMA.VENTAS	FLOAT
MAX_SUBTOTAL_MES_P1	MAXIMO MONTO DE VENTAS EN EL MES P1	ESQUEMA.VENTAS	FLOAT
MIN_SUBTOTAL_MES_H3	MINIMO MONTO DE VENTAS EN EL MES H3	ESQUEMA.VENTAS	FLOAT
MIN_SUBTOTAL_MES_H2	MINIMO MONTO DE VENTAS EN EL MES H2	ESQUEMA.VENTAS	FLOAT
MIN_SUBTOTAL_MES_H1	MINIMO MONTO DE VENTAS EN EL MES H1	ESQUEMA.VENTAS	FLOAT
MIN_SUBTOTAL_MES_P0	MINIMO MONTO DE VENTAS EN EL MES P0	ESQUEMA.VENTAS	FLOAT
MIN_SUBTOTAL_MES_P1	MINIMO MONTO DE VENTAS EN EL MES P1	ESQUEMA.VENTAS	FLOAT

4.10 Formateo de los datos

Se realizaron pruebas antes de la construcción del modelo final, y en estas pruebas la herramienta Pentaho Weka requería que la variable que me define a un cliente como fuga o aún fidelizado sea de tipo nominal (valores de tipo discreto); sin embargo, ya que en un principio este valor estaba solo como 0 o1 no podía ser de tipo nominal; entonces se optó por convertir esa variable a una cadena de texto, la cual mostraría el mensaje FUGA si el cliente sería una potencial fuga para la empresa; de lo contrario muestra el mensaje FIEL.

4.11 Selección de la técnica de modelado

Ahora se procede a seleccionar la técnica de modelado y el algoritmo que se planeó usar en el proceso algorítmico de Data Mining.

Las técnicas de modelado que se usan para Data Mining están mostradas en la Figura 4.4.

Las técnicas mencionadas en la Figura 4.4 son usadas para predicción y descripción; para este trabajo de fin de carrera se necesita predecir el comportamiento de los clientes; por lo tanto se usan las técnicas de predicción. Cabe resaltar que las técnicas de modelado engloban un conjunto de algoritmos que las ejecutan lo mismo pero de diferente forma; es decir, los algoritmos al final tienen como meta un mismo fin pero de distinta forma. (MOLINA & GARCIA, 2006)

En este caso se usa la técnica de modelado “Arboles de Decisiones” y se usa el algoritmo “J48” (este algoritmo es el mismo algoritmo C4.5, solo que en Pentaho Weka tiene el nombre J48). Se escogió el modelado de “Árboles de decisión” porque su aprendizaje es más robusto al ruido, y a la vez es fácil de usar. (MOLINA & GARCIA, 2006) Además, se escogió el algoritmo J48 porque soporta tanto valores continuos como valores discretos, y los datos que se usan como variables contienen valores nominales (discretos) y numéricos (no nominales). (MOLINA & GARCIA, 2006)

4.12 Construcción del modelo

En esta etapa se realiza la prueba principal con el algoritmo, para esto se seleccionan los datos que se necesitan en el software PENTAHO WEKA tal como se aprecia en la Figura 4.5.

Luego se selecciona a la variable que determina si un cliente es o no un cliente que fuga de la empresa; para esto se selecciona a la variable “Y”, y se la convierte en clase, entonces la herramienta la reconocerá como la variable a predecir, tal como se muestra en la Figura 4.6.

Luego se selecciona el algoritmo que se usa en la prueba final, como ya se dijo antes el algoritmo a usar es el J48, tal como se muestra en la Figura 4.7.

Finalmente, se ajustan algunos valores estadísticos de la herramienta software PENTAHO WEKA para el algoritmo J48, tal como se muestra en la Figura 4.8.

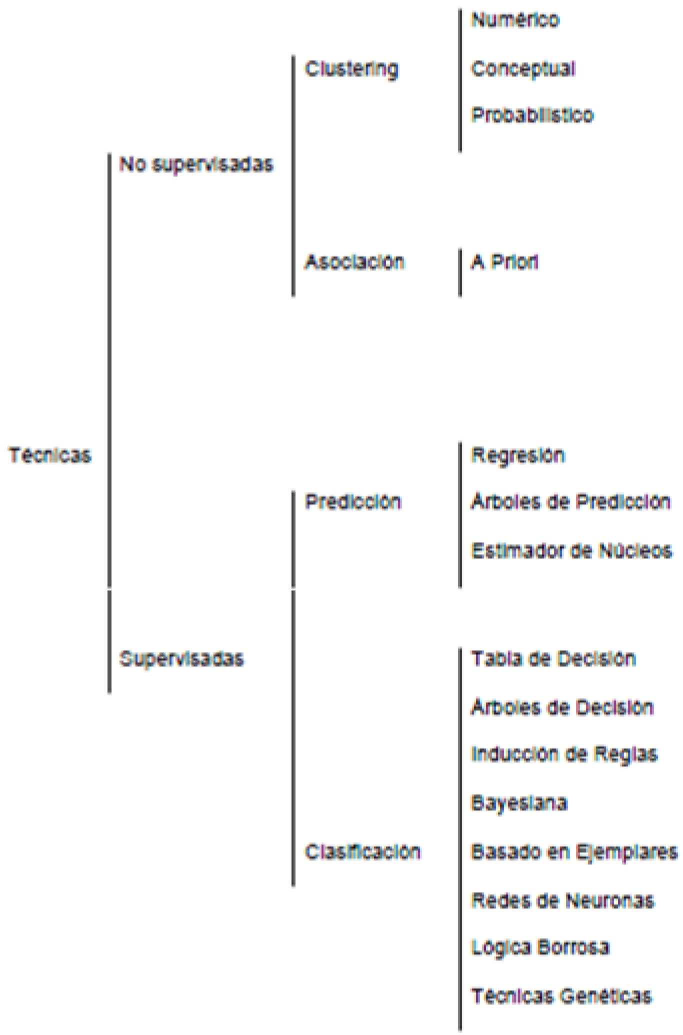


Figura 4.4: Técnicas de modelado de Data Mining (MOLINA & GARCIA, 2006)

Y el resultado obtenido del software PENTAHO WEKA mostrado en es un conjunto de reglas aplicables para los siguientes dos meses Noviembre y Diciembre del año 2012.

Los resultados de la Tabla 4.9 se representan de la siguiente forma:

De la provincia de Tacna 11 de 14 clientes llegan a ser fieles, y los 3 restantes suelen ser clientes que ya no compran en la empresa.

De la provincia de Cajamarca (Si el promedio cuatrimestral de compras supera los S/. 1773 entonces todos en ese subgrupo suelen ser fieles a la empresa; caso contrario, el 90% de los clientes son fieles y solo el 10% deja de comprar en la empresa).

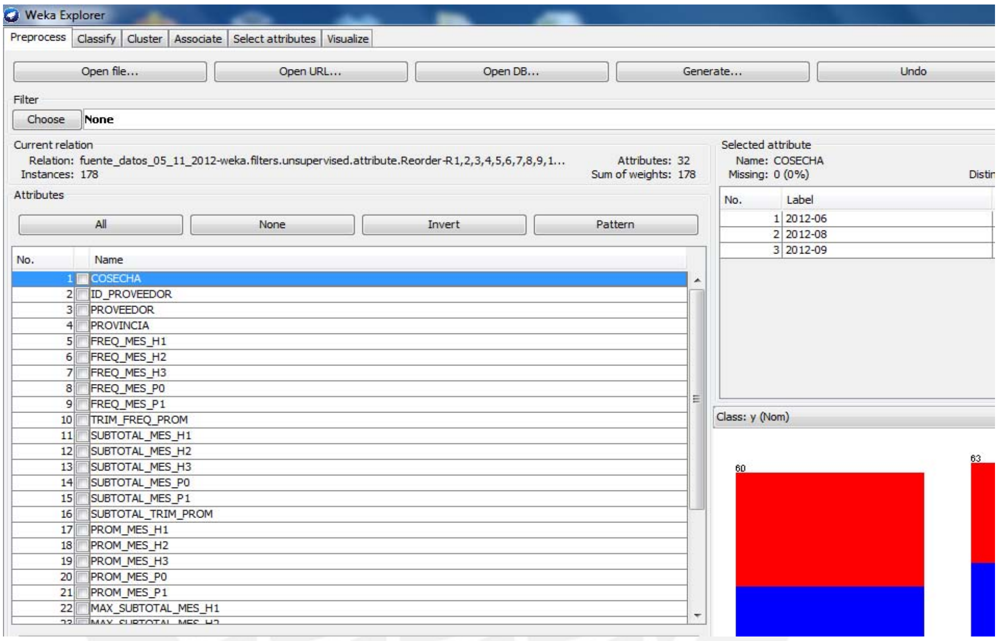


Figura 4.5: **Selección de datos en la herramienta PENTAHO WEKA** (Elaboración Propia)

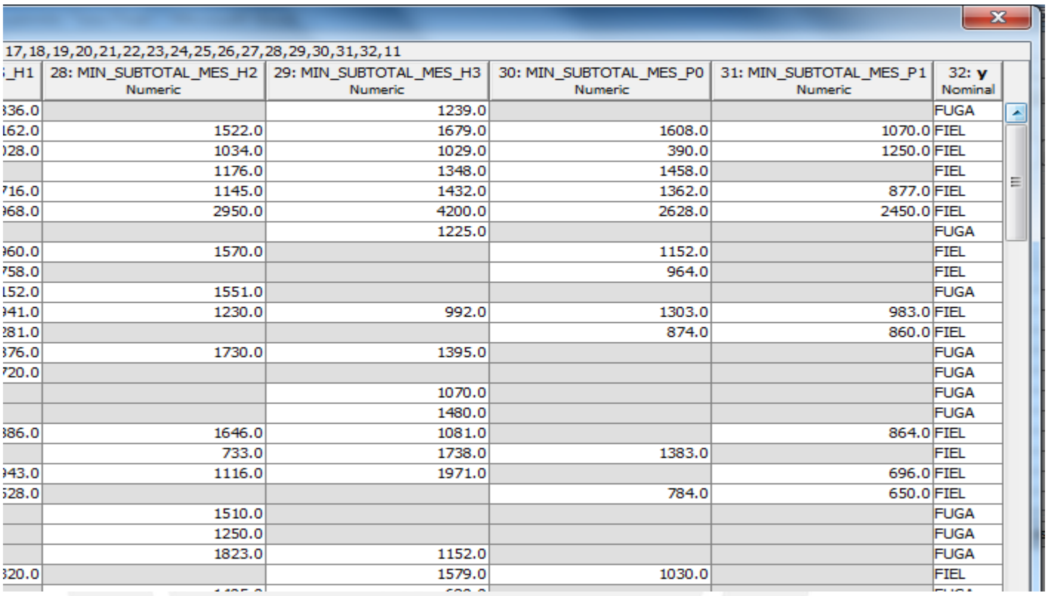
De la provincia de Lambayeque (Si la veces que el cliente fue de compras en el mes H1 es menor o igual a 2; entonces, 3 de 4 personas son fieles a la empresa; caso contrario este único cliente deja de comprar en la empresa; por otro lado, si la cantidad de veces que el cliente fue de compras en el mes H1 es mayor o igual a 3 entonces 10 de 13 clientes son fieles a la empresa).

De la provincia de Huánuco todos los clientes desertan de la empresa y ya no realizan compras.

De la provincia de Piura (Si la frecuencia trimestral de compras es 1 o ninguna, entonces la fuga del cliente es 100% probable; caso contrario, los clientes fieles son 7 de 8).

De la provincia de Cusco (Si la frecuencia de compras en el mes H3 es 1 o 0 entonces 8 de 11 clientes son clientes que ya no compran más; caso contrario, también 8 de 11 clientes son fieles aún a la empresa).

De la provincia de Arequipa 18 de 24 clientes son fieles a la empresa.De la provincia de Ancash, Loreto, Madre de Dios, Ayacucho, Amazonas y Apurímac todos los clientes llegan a no comprar más en la empresa.


17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,11

_H1	28: MIN_SUBTOTAL_MES_H2 Numeric	29: MIN_SUBTOTAL_MES_H3 Numeric	30: MIN_SUBTOTAL_MES_P0 Numeric	31: MIN_SUBTOTAL_MES_P1 Numeric	32: y Nominal
36.0		1239.0			FUGA
62.0	1522.0	1679.0	1608.0	1070.0	FIEL
28.0	1034.0	1029.0	390.0	1250.0	FIEL
	1176.0	1348.0	1458.0		FIEL
16.0	1145.0	1432.0	1362.0	877.0	FIEL
68.0	2950.0	4200.0	2628.0	2450.0	FIEL
		1225.0			FUGA
60.0	1570.0		1152.0		FIEL
58.0			964.0		FIEL
52.0	1551.0				FUGA
41.0	1230.0	992.0	1303.0	983.0	FIEL
81.0			874.0	860.0	FIEL
76.0	1730.0	1395.0			FUGA
20.0					FUGA
		1070.0			FUGA
		1480.0			FUGA
86.0	1646.0	1081.0		864.0	FIEL
	733.0	1738.0	1383.0		FIEL
43.0	1116.0	1971.0		696.0	FIEL
28.0			784.0	650.0	FIEL
	1510.0				FUGA
	1250.0				FUGA
	1823.0	1152.0			FUGA
20.0		1579.0	1030.0		FIEL

Figura 4.6: Definir la variable a predecir (Elaboración Propia)

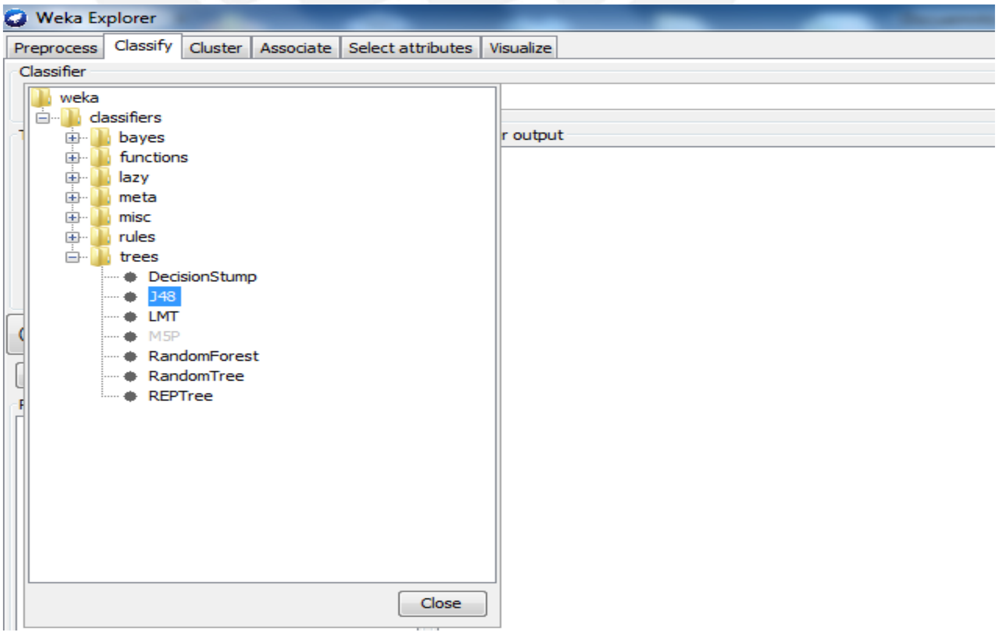


Figura 4.7: Selección del algoritmo a usar (Elaboración Propia)

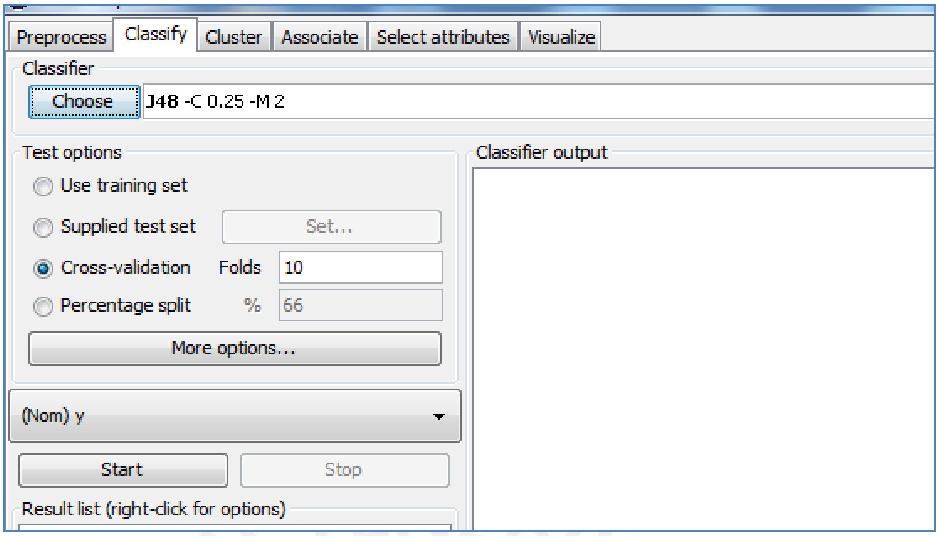


Figura 4.8: Ventana para la modificación de variables en PENTAHO WEKA (Elaboración Propia)

De la provincia de la Libertad (Si la frecuencia de compras cuatrimestral es menor o igual a 2 y el promedio de compras cuatrimestral es menor a S/. 784 entonces son fieles a la empresa, pero si el promedio de compras cuatrimestral es mayor a S/784 entonces 12 de 13 clientes son propensos a no comprar más; por otro lado, si la frecuencia de compras cuatrimestral es mayor a 2 entonces esos clientes son fieles a la empresa).

De la provincia de San Martín (Si los clientes compran más de S/. 2661 en el promedio de compras del cuatrimestre entonces son propensos a irse; caso contrario se quedan aún como clientes).

4.13 Evaluación de resultados

Los resultados presentados en la Tabla 4.9, muestran claramente que existe una fuga de clientes muy frecuente, y los diversos nodos o condicionales mostrados en la misma tabla indican las posibles razones o los indicios a que un cliente ya no compre más en la empresa. Por lo tanto el modelo planteado ha logrado su objetivo; sin embargo, estos resultados no son los finales, ya que esta muestra solo incluye datos históricos del 1 de Abril del 2012 al 31 de octubre del mismo año. Se podría considerar un conjunto de datos con más credibilidad si hubiese por lo menos un año de datos históricos; y el producto final contemplará esta cantidad

Tabla 4.9: Resultado del modelo de Data Mining (Elaboración Propia)

```
PROVINCIA = Tacna: FIEL (11.0/3.0)
PROVINCIA = Cajamarca
|  PROM_SUBTOTAL_CUATRIMESTRAL <= 1773: FIEL (3.0)
|  PROM_SUBTOTAL_CUATRIMESTRAL > 1773: FUGA (9.0/1.0)
PROVINCIA = Lambayeque
|  FREQ_MES_H1 <= 2
|  |  FREQ_MES_H1 <= 1: FIEL (3.54/1.08)
|  |  FREQ_MES_H1 > 1: FUGA (8.85/1.15)
|  FREQ_MES_H1 > 2: FIEL (10.62/3.23)
PROVINCIA = Huánuco: FUGA (4.0)
PROVINCIA = Piura
|  PROM_FREQ_CUATRIMESTRAL <= 1.75: FUGA (4.0)
|  PROM_FREQ_CUATRIMESTRAL > 1.75: FIEL (7.0/1.0)
PROVINCIA = Cusco
|  FREQ_MES_H3 <= 1: FUGA (8.0/2.0)
|  FREQ_MES_H3 > 1: FIEL (8.0/2.0)
PROVINCIA = Arequipa: FIEL (18.0/6.0)
PROVINCIA = La Libertad
|  PROM_FREQ_CUATRIMESTRAL <= 2.75
|  |  PROM_SUBTOTAL_CUATRIMESTRAL <= 784: FIEL (2.0)
|  |  PROM_SUBTOTAL_CUATRIMESTRAL > 784: FUGA (12.0/1.0)
|  PROM_FREQ_CUATRIMESTRAL > 2.75: FIEL (2.0)
PROVINCIA = Ancash: FUGA (2.0)
PROVINCIA = Loreto: FUGA (6.0)
PROVINCIA = Madre de Dios: FUGA (2.0)
PROVINCIA = San Martin
|  PROM_SUBTOTAL_CUATRIMESTRAL <= 2661: FIEL (2.0)
|  PROM_SUBTOTAL_CUATRIMESTRAL > 2661: FUGA (2.0)
PROVINCIA = Ayacucho: FUGA (2.0)
```

de datos, pero en el presente documento solo se considera 7 meses de datos. Además, con un año de datos históricos se pueden considerar variables adicionales como las temporadas, etc.

5 Observaciones, conclusiones y recomendaciones

En este capítulo final se presentan las observaciones realizadas durante el desarrollo del proyecto, las conclusiones obtenidas y las recomendaciones que se han considerado pertinentes.

5.1 Observaciones

- El ERP usado (OpenERP) es una herramienta que posee una variedad de módulos que pueden ser explotados por cualquier empresa que lo requiera; y además, es de código abierto.
- La herramienta Pentaho Weka usada en el proyecto tenía limitaciones en cuanto a los algoritmos, ya que herramientas de software por pago de licencia poseen algoritmos más potentes; sin embargo, para el proyecto era suficiente el uso del algoritmo usado.
- El producto final posee un bajo grado de escalabilidad, ya que para otras situaciones u otros tipos de empresa se buscan distintas necesidades; por esto, es importante definir los objetivos del Data Mining primero antes de realizar cualquier paso siguiente.
- El proceso de Data Mining requiere de varias pruebas para obtener un valor final aceptable; esto conlleva a realizar ajustes a valores adicionales a las variables de decisión (variables del algoritmo).

5.2 Conclusiones

- La herramienta OpenERP de código abierto es de fácil uso; sin embargo esto no quiere decir que una empresa con mayor número de transacciones diarias la pueda usar normalmente, ya que la versión gratuita posee algunas desventajas que pueden ocasionar problemas a futuro en la empresa.
- La implantación de un ERP no incluye solamente la instalación del software en la empresa, sino que también conlleva posibles problemas por resistencia de los usuarios al cambio, capacitación de usuarios, comprensión de las tablas de la base de datos, etc.
- 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 obtenidos 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.

5.3 Recomendaciones

- El producto final puede ser analizado desde otra perspectiva si es que se posee más datos de los que se tenía para este proyecto; es decir, los resultados pueden variar si se tiene datos mayores a un año de historia y/o si se tiene otro objetivo de Data Mining.
- Si la cantidad de datos hubiese sido mayor, entonces la cantidad de variables de decisión pudiese haber sido mayor, lo cual conlleva a poder analizar otro algoritmo o conjunto de algoritmos.
- El tema de este trabajo o el área de estudio (Data Mining) puede resultar interesante para cualquier estudiante de Ingeniería Informática, por lo que sería muy conveniente que se cree un área de estudio del tema.
- El producto final puede tener una amplia aceptación en el mercado, ya que actualmente las empresas necesitan manejar conocimiento que les permita sobresalir en la industria a la cual está enfocada.

Bibliografía

BERRY, M. J. A. & LINOFF, G. S.,

2004. *Data Mining Techniques.* Segunda Edición ed. Indianapolis: Wiley.

HAN, J. & KAMBER, M.,

2006. *Data Mining Concepts and Techniques.* Segunda Edición ed. Burlington: Elsevier.

MAIMON, O. & ROKACH, L.,

2005. *DECOMPOSITION METHODOLOGY FOR KNOWLEDGE DISCOVERY AND DATA MINING Theory and Applications.* Primera Edición ed. Singapore: World Scientific.

WEISS, G. M. & DAVIDSON, B. D.,

2010. Data Mining. *The Handbook of Technology Management,* Volumen 3, pp. 1-17.

WIDENER, T.,

1996. Mining Business Databases. *Communications of the ACM,* Noviembre, 39(11), pp. 42-48.

LIEW, A.,

2007. Understanding Data, Information, Knowledge And Their Inter-Relationships. *Journal of Knowledge Management Practice.* 8(2), pp. 11-14.

CALDERS, T. & PECHENIZKIY, M.,

2011. Introduction to the Special Section on Educational Data Mining. *SIGKDD Explorations,* 13(2), pp. 1-6.

FAYYAD, U., PIATETSKY-SHAPIRO, G. & SMYTH, P.,

1996. From Data Mining to Knowledge Discovery in databases. *AI Magazine,* pp. 37-54.

KREISER, P. M.,

2006. *The Evolution of Competitive Strategies in Global Forestry Industries Comparative Perspectives.* Dordrecht, Springer, pp. 191-194.

ROWLEY, J.,

2007. The Wisdom hierachy: representations of the DIKW hierachy. *Journal of information Science,* 33(163), pp. 163-180.

ADEMPIERE,

2012. *aDempiere.*

<http://www.adempiere.com/images/a/ad/Porque_usar_ADempiere_v2_Spanish.pdf>
[Último acceso: 27 Abril 2012].

ALBARRAN, R.,

2012. *The POS System SVIEW32.* <http://sviw32.sourceforge.net/about.html>
[Último acceso: 27 Abril 2012].

ANGOSS,

2012. *Angoss.*

<http://www.angoss.com/predictive-analytics-software/products/data-analysis-software>
[Último acceso: 14 Abril 2012].

APARA,

2012. *Apara.* <http://www.aparasw.com/index.php/es/dvelox-enterprise>
[Último acceso: 12 Abril 2012].

CHAPMAN, P. y otros,

2000. *Guía paso a paso de Minería de Datos.* s.l.:s.n.

<http://www.dataprix.com/book/export/html/107>

[Último acceso: 15 Abril 2012].

ESCUELA DE DISEÑO EN EL HÁBITAT,

2008. *IndumentariayModa.* <http://diseniodeindumentaria2.files.wordpress.com/2008/04/tejidos-de-caladapdf.pdf>
[Último acceso: 27 Abril 2012].

ICON-INSTITUT GmbH Private Sector,

2009. *ESTUDIOS DE MERCADO E IDENTIFICACIÓN DE OPORTUNIDADES.* <http://www.mincetur.gob.pe/Comercio/ueperu/licitacion/pdfs/Informes/115.pdf>
[Último acceso: 06 Octubre 2012].

MICROSOFT,

2012. *SQL Server.* <http://www.microsoft.com/sqlserver/en/us/solutions-technologies/business-intelligence/data-mining.aspx>
[Último acceso: 15 Abril 2012].

MICROSTRATEGY,

2012. *Microstrategy.* <http://www.microstrategy.com/Software/Products/Service_Modules/DataMining_Services/>
[Último acceso: 10 Abril 2012].

MOLINA, J. & GARCIA, J.,

2006. *Técnicas de análisis de datos* . <http://www.giaa.inf.uc3m.es/docencia/II/ADatos/apuntesAD.pdf>
[Último acceso: 23 Abril 2012].

OPENBRAVO,

2012. *Estrasol.*<http://www.estrasol.com.mx/brochure-pos.php>
[Último acceso: 27 Abril 2012].

OPENBRAVO,

2012. *OPENBRAVO ERP.*

<http://www.openbravo.com/product/erp/>
[Último acceso: 27 Abril 2012].

OpenERP,

2012. *OpenERP.*

<http://www.openerpspain.com/gestion-de-ventas>
[Último acceso: Abril 27 2012].

PENTAHO,

2012. *Pentaho Weka Proyect.*

<http://www.cs.waikato.ac.nz/ml/weka/>
[Último acceso: 10 Abril 2012].

PINTO CASTRO, J.,

2007. *Evaluación de la oportunidad de desarrollo de las mipymes del sub-sector confecciones: a propósito de la firma del TLC.* <http://economia.unmsm.edu.pe/organizacion/iiec/archivos/revistasiie/PC_10/PC10_CAP09.pdf>
[Último acceso: 20 Abril 2012].

ROCKETT, L.,

2003. *Las TI importan, si uno quiere.* <http://www.iese.edu/es/files/Art_EN_Rocket_ITMatters_Oct03_ESP_tcm5-7433.pdf>
[Último acceso: 12 Septiembre 2012].