



**UNIVERSIDAD NACIONAL  
“PEDRO RUIZ GALLO”**

**FACULTAD DE CIENCIAS FISICAS  
Y MATEMATICAS**



**Escuela Profesional de Computación e  
Informática**

**Tesis para optar por el Título Profesional de Ingeniero en  
Computación e Informática**

**IMPLEMENTACIÓN DE MINERIA DE DATOS PARA  
DETECTAR PATRONES DE COMPORTAMIENTO  
DE CLIENTES MOROSOS EN EMPRESA DE  
CREDITO CREDISERV EIRL - CHICLAYO**

**PRESENTADO POR:**

**EDWIN RONALD TORRES CHERO**

**JOSE MIGUEL FARROÑAY JULCA**

**ASESOR: ING.**

**PEDRO FIESTAS**

**LAMBAYEQUE – PERÚ 2015**



# **UNIVERSIDAD NACIONAL “PEDRO RUIZ GALLO”**

**FACULTAD DE CIENCIAS FISICAS  
Y MATEMATICAS**



## **Escuela Profesional de Computación e Informática**

**Tesis para optar por el Título Profesional de Ingeniero en  
Computación e Informática**

### **IMPLEMENTACIÓN DE MINERIA DE DATOS PARA DETECTAR PATRONES DE COMPORTAMIENTO DE CLIENTES MOROSOS EN EMPRESA DE CREDITO CREDISERV EIRL - CHICLAYO**

---

**ING. LUIS ALBERTO REYES LESCANO  
PRESIDENTE DE JURADO**

---

**ING. DENNY JOHN FUENTES ADRIANZEN  
SECRETARIO**

---

**ING. ALEJANDRO CHAYAN COLOMA  
VOCAL**



# **UNIVERSIDAD NACIONAL “PEDRO RUIZ GALLO”**



**FACULTAD DE CIENCIAS FISICAS  
Y MATEMATICAS**

**Escuela Profesional de Computación e Informática**

**Tesis para optar por el Título Profesional de Ingeniero en  
Computación e Informática**

**“IMPLEMENTACIÓN DE MINERIA DE DATOS  
PARA DETECTAR PATRONES DE  
COMPORTAMIENTO DE CLIENTES MOROSOS EN  
EMPRESA DE CREDITO CREDISERV EIRL –  
CHICLAYO”**



---

**ING. PEDRO FIESTAS RODRIGUEZ  
ASESOR DE TESIS**



---

**EDWIN RONALD TORRES CHERO**



---

**JOSE MIGUEL FARROÑAY JULCA**

## **DEDICATORIA**

Dedico este trabajo a mi Madre, ya que desde siempre ha sido el apoyo y el soporte al esfuerzo realizado y teniendo siempre la fé en mí de lograr ser cada vez mejor.

## **DEDICATORIA**

**A mi padre Miguel.**

Por los ejemplos de perseverancia, constancia, humildad que me ha infundado siempre, por el valor mostrado para salir adelante y por su amor que desde el cielo él me da.

## **AGRADECIMIENTO**

El agradecimiento a todos los que nos apoyaron en concretar este proyecto, familiares, amigos, profesores, la empresa; ya que han sido de vital importancia para su ejecución y culminación.

## **AGRADECIMIENTO**

### **A mi madre Gladys.**

Por haberme apoyado en todo momento, por sus consejos, sus valores, por la motivación constante que me ha permitido ser una persona de bien, pero más que nada, por su amor

## **RESUMEN**

El presente proyecto plantea resolver el problema de incremento de la tasa de morosidad en la empresa CREDISERV EIRL durante el año 2014 en un 10% respecto del año pasado; esto tomando como base los datos del sistema actual de créditos y control de cuotas de la empresa.

Se plantea el objetivo de aplicar técnicas de minería de datos para detectar patrones de comportamiento de clientes morosos en empresa de crédito Crediserv EIRL y de esta forma reducir los índices de morosidad de la empresa.

Se desarrolla una solución de inteligencia de negocios para detectar patrones de comportamiento en clientes morosos en la empresa de créditos Crediserv EIRL de la ciudad de Chiclayo

### **Palabras claves:**

Inteligencia de negocios, minería de datos, clientes morosos, toma de decisiones, tecnología de información

## **ABSTRACT**

This project proposes to solve the problem of increase in the delinquency rate in the company CREDISERV EIRL during 2014 by 10% over last year; this data based on the current subscription and quota control of the company.

The project proposes the aim of applying data mining techniques to detect patterns of behavior of customers defaulting on credit company Crediserv EIRL and thus reduce NPLs of the company.

A solution of business intelligence is developed to detect patterns of behavior in delinquent customers in the business of credit Crediserv EIRL Chiclayo

### **Keywords:**

Business intelligence, data mining, delinquent customers, decision making, information technology

## INDICE GENERAL

INTRODUCCIÓN .....	10
CAPÍTULO I: DATOS GENERALES DE LA ORGANIZACIÓN.....	11
1.1. Descripción de la organización .....	11
1.2. Misión, visión y objetivos de la organización .....	11
<b>1.1.1. Misión</b> .....	11
<b>1.1.2. Visión</b> .....	11
1.3. Valores corporativos.....	11
CAPÍTULO II: PROBLEMÁTICA DE LA INVESTIGACIÓN .....	12
2.1. Realidad problemática .....	12
2.2. Formulación del problema.....	13
2.3. Justificación e importancia de la investigación .....	13
2.4. Objetivos de la investigación.....	14
2.4.1. Objetivo general.....	14
2.4.2. Objetivos específicos .....	14
2.5. Limitaciones de la investigación .....	14
CAPÍTULO III: MARCO METODOLÓGICO .....	15
3.1. Tipo de investigación .....	15
3.2. Hipótesis .....	15
3.3. Variables.....	15
3.3.1. Variable independiente .....	15
3.3.2. Variable dependiente .....	15
3.4. Diseño y Contrastación de Hipótesis.....	15
CAPÍTULO IV: MARCO TEÓRICO.....	16
4.1. Antecedentes de investigación .....	16
4.1.1. Antecedentes en el contexto internacional.....	16
4.1.2. Antecedentes en el contexto nacional.....	18
4.1.3. Antecedentes en el contexto regional .....	19
4.2. Desarrollo de la temática .....	20
4.2.1. Descubrimiento del conocimiento (KDD).....	20
4.2.2. Minería de datos.....	21
4.2.3. Algoritmos de minería de datos .....	25
4.2.4. Modelo de procesos CRISP-DM .....	27
4.2.4.1. SEMMA	
4.2.4.2. KDD	
4.2.4.3. DMAMC	
4.2.4.4. CRISP-D	
4.2.4.5. Comparación entre KDD, SEMMA y CRISP-DM	
CAPÍTULO V: DESARROLLO DE LA PROPUESTA.....	46
<b>DETERMINACION DE REQUERIMIENTOS</b> .....	46
<b>DISEÑO DE DATAMART</b> .....	46
<b>PROCESO ETL</b> .....	49
<b>PROCESO DE DISEÑO DE LA BASE DE DATOS OLAP</b> .....	54
<b>DISEÑO DEL CUBO</b> .....	57
<b>IMPLEMENTACIÓN DEL CUBO</b> .....	58
<b>EXPLOTACIÓN DEL CUBO CON POWER VIEW</b> .....	59
CAPÍTULO VI: COSTOS Y BENEFICIOS .....	62



6.1. Análisis de costos .....	62
6.2. Financiamiento .....	62
CAPÍTULO VII: CONCLUSIONES .....	63
CAPÍTULO VIII: RECOMENDACIONES .....	64
CAPÍTULO IX: REFERENCIAS BIBLIOGRÁFICAS .....	65

## INDICE DE FIGURAS

Ilustración 1 - Proceso de descubrimiento del conocimiento.....	20
Ilustración 2- Técnicas de análisis de datos .....	22
Ilustración 3 - Clasificación de algoritmos de minería de datos .....	25
Ilustración 4 - Ciclo de vida de CRISP-DM .....	33
Ilustración 5 - Determinación de requerimientos .....	46
Ilustración 6 - Data Mart .....	48
Ilustración 7 - Proceso ETL .....	49
Ilustración 8 - Dimensión Cliente .....	50
Ilustración 9 - Dimensión Crédito.....	50
Ilustración 10 - Dimensión Tiempo .....	51
Ilustración 11 - Dimensión Forma de Pago.....	51
Ilustración 12 - Dimensión tipo moneda.....	52
Ilustración 13 - Dimensión Estado Crédito.....	52
Ilustración 14 - Dimensión Tipo Pago .....	53
Ilustración 15 - Dimensión tiempo.....	54
Ilustración 16 - Dimensión moneda .....	54
Ilustración 17 - Dimensión estado del crédito.....	55
Ilustración 18 - Dimensión cliente .....	55
Ilustración 19 - Dimensión tipo de pago .....	56
Ilustración 20 - Dimensión forma de pago.....	56
Ilustración 21 - Dimensión crédito.....	57
Ilustración 22 - Diseño del cubo de resumen de crédito .....	58
Ilustración 23 - Arquitectura de Power View para modelos multidimensionales.....	59
Ilustración 24 - Conectar a la solución OLAP .....	60
Ilustración 25 - Consultando datos con Power Pivot .....	60

## INTRODUCCIÓN

La creciente capacidad tecnológica de transmitir información en tiempo real, conduce a la necesidad de aprender y mejorar el nivel de vida, permite el desarrollo de la educación y de la investigación, potencia los servicios médicos, ensancha los mercados, derriba las fronteras, cuestiona las nociones habituales de tiempo y espacio.

La necesidad de datos previo a la toma de decisiones es un factor considerado desde decisiones ancestrales, la diferencia radica en que en la actualidad existe gran cantidad de información, almacenada en diferentes formatos y en diversas fuentes.

Crediserv EIRL es una empresa de la ciudad de Chiclayo, dedicada al rubro de créditos y servicios financieros, con cerca de 15 años de experiencia en el mercado.

El presente proyecto plantea resolver el problema de incremento de la tasa de morosidad en la empresa CREDISERV EIRL durante el año 2014 en un 10% respecto del año pasado; esto tomando como base los datos del sistema actual de créditos y control de cuotas de la empresa.

Se plantea el objetivo de aplicar técnicas de minería de datos para detectar patrones de comportamiento de clientes morosos en empresa de crédito Crediserv EIRL y de esta forma reducir los índices de morosidad de la empresa.

La presente investigación se justifica por el alto grado de impacto en la solución de problemas de uso de data y convertirla en información útil que genere valor en el ámbito empresarial usando técnicas avanzadas de minería de datos, las cuales serán evaluadas para medir su grado de efectividad.

## **CAPÍTULO I: DATOS GENERALES DE LA ORGANIZACIÓN**

### **1.1. Descripción de la organización**

La historia de la empresa de crédito Crediserv EIRL inicia el 01 de febrero del año 2000, como empresa individual de responsabilidad limitada en el rubro de créditos y servicios financieros, con RUC 20437345474, ubicada en calle Manuel Maria Izaga Nro. 206 de la ciudad de Chiclayo.

### **1.2. Misión, visión y objetivos de la organización**

#### **1.1.1. Misión**

Impulsar el crecimiento sostenible de nuestros clientes, colaboradores, accionistas y de la ciudad de Chiclayo

#### **1.1.2. Visión**

Ser la empresa de créditos líder en ofrecer soluciones financieras a nuestro mercado objetivo, brindando calidad de servicio, eficiencia y oportunidad

### **1.3. Valores corporativos**

- Orientación al cliente:
  - Conocer y satisfacer sus necesidades
  - Simplicidad y transparencia
  - Disponibilidad y cercanía
  - Amabilidad
- Orientación a las Personas
  - Confianza
  - Equidad
  - Reconocimiento y desarrollo
  - Trabajo en equipo
- Orientación al Logro
  - Visión global
  - Integridad
  - Proactividad
  - Responsabilidad y compromiso

## CAPÍTULO II: PROBLEMÁTICA DE LA INVESTIGACIÓN

### 2.1. Realidad problemática

En los últimos años, ha existido un gran crecimiento en la capacidad de generar y coleccionar datos, debido al gran poder de procesamiento de las computadoras y a su bajo costo de almacenamiento; sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de información "oculta", de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información (Valcárcel Asencios, 2004)

El descubrimiento de esta información "oculta" es posible gracias a la minería de datos, que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el Descubrimiento de Conocimiento (KDD) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados (Valcárcel Asencios, 2004)

Nuestro país también está inmerso en esta situación ya que el avance tecnológico y la globalización ha permitido el fácil acceso a la tecnología por parte de las medianas y pequeñas empresas de distintos rubros tanto estatales como privadas, que actualmente se enfrentan a la necesidad de manejar gran cantidad de datos y que están siendo desaprovechados sin dar un valor a la empresa. Experiencias exitosas como las de Bambos y Kola Real son ejemplos de imitación, pero detrás de las que hay mucho esfuerzo por identificar procesos seguidos por las empresas originales para luego adaptarlos a la idiosincrasia y al mercado nacional y, finalmente, generar productos y servicios nuevos o por lo menos diferenciados, (Kuramoto de Grade, 2013).

En el Perú hay cada vez más entidades crediticias, por lo que el sector es cada vez más competitivo. La cartera de crédito al consumo implica el manejo de un gran número de clientes. Las entidades financieras requieren procesar un gran número de solicitudes de crédito, por tanto es importante que la administración de riesgo de

los bancos ejerza un control efectivo sobre el proceso de evaluación de un cliente a fin de que se le otorgue o niegue un crédito (Salinas Flores, 2005)

La evaluación de riesgos financieros de un banco o de una entidad de créditos es llevada a cabo por la Gerencia de Riesgo usando por lo general técnicas subjetivas o técnicas descriptivas univariadas o bivariadas.

El índice de morosidad según datos de la Asociación de Bancos (Asbanc) en enero del 2014 en las tarjetas de crédito emitidas por bancos y financieras ascendió a 6.19%, alcanzando un pico máximo. La morosidad de la banca privada peruana subió ligeramente en julio y llegó a 2.11%, porcentaje mayor en 0.05 puntos porcentuales al registrado el mes anterior.

En CREDISERV EIRL la tasa de morosidad del año 2014 se ha incrementado en 10% respecto del año pasado; esto tomando como base los datos del sistema actual de créditos y control de cuotas de la empresa.

Sin embargo existen, técnicas que permiten encontrar patrones de comportamiento basados en un conjunto de variables independientes, que son factibles de aplicar en el mercado peruano como son los denominados árboles de clasificación, la red neuronal artificial (RNA), entre otras (Salinas Flores, 2005).

## **2.2. Formulación del problema**

¿Es posible detectar patrones de comportamiento de usuarios potencialmente morosos mediante la aplicación de minería de datos?

## **2.3. Justificación e importancia de la investigación**

La investigación se justifica porque la minería de datos se hace necesaria en importantes áreas, tales como la economía, el cuidado de la salud, la investigación científica, etcétera. En estas áreas existe una gran cantidad de datos que sólo han sido analizados parcialmente, y que contienen una gran cantidad de información que aún no ha sido explorada (Tumero, 2011).

La presente investigación se justifica por el alto grado de impacto en la solución de problemas de uso de data y convertirla en información útil que genere valor en el ámbito empresarial usando técnicas avanzadas de minería de datos, las cuales serán evaluadas para medir su grado de efectividad.

Con la utilización de la Minería de Datos, podrá evaluar el aspecto comercial en cuanto a demanda y oferta, permitiéndole tomar decisiones más acertadas en cuanto a ofertas a los usuarios, evitando la pérdida de dinero que se generaría si se tomaran decisiones equivocadas.

## **2.4. Objetivos de la investigación**

### **2.4.1. Objetivo general**

Aplicar minería de datos para detectar patrones de comportamiento de clientes morosos en empresa de crédito Crediserv EIRL

### **2.4.2. Objetivos específicos**

- Analizar el proceso de gestión de préstamos en pequeñas empresas de créditos.
- Aplicar técnicas de limpieza y recolección de datos desde las diversas fuentes de información de la empresa.
- Seleccionar y aplicar el modelo de minería de datos para determinar el que mejor identifica patrones de comportamiento.
- Desarrollar una aplicación que muestre los resultados de la técnica de minería de datos seleccionada

## **2.5. Limitaciones de la investigación**

Para el desarrollo del proyecto tuvimos todas las condiciones, por lo que no se consideran limitaciones

## **CAPÍTULO III: MARCO METODOLÓGICO**

### **3.1. Tipo de investigación**

Descriptiva

### **3.2. Hipótesis**

La aplicación de minería de datos permite detectar patrones de comportamiento de usuarios potencialmente morosos en la empresa de crédito Crediserv EIRL

### **3.3. Variables**

#### **3.3.1. Variable independiente**

Aplicación de minería de datos

#### **3.3.2. Variable dependiente**

Patrones de comportamiento de usuarios potencialmente morosos en la empresa de crédito Crediserv EIRL

### **3.4. Diseño y Contrastación de Hipótesis**

<b>X: Aplicación de minería de datos</b>	<b>Y: Patrones de comportamiento de usuarios potencialmente morosos en la empresa de crédito Crediserv EIRL</b>
1- Clasificación	Disminuir tiempos para determinar cliente morosos potenciales
2- Regresión	Minimizar riesgos en la pérdida de créditos

## **CAPÍTULO IV: MARCO TEÓRICO**

### **4.1. Antecedentes de investigación**

#### **4.1.1. Antecedentes en el contexto internacional**

**TITULO :** Creación de perfiles de deudores de crédito universitario, para mejoramiento de campañas de cobranza, usando minería de datos (Lagos Vera, 2011)

**AUTOR :** Lagos Vera, Carolina

**AÑO :** 2011

#### **RESUMEN**

El trabajo de tesis determinó un conjunto de características comunes que presentaban los estudiantes de la Universidad de la Frontera Chile, que presentaban morosidad en su historial de pagos a la universidad. El proyecto utiliza la técnica de minería de datos clustering o agrupamiento, con la metodología CRISP-DM.

Se concluyó que con la implementación de un proyecto como este, se logrará tener perfiles de alumnos de acuerdo a sus características académicas, socio-económicas, demográficas, entre otras, los cuales ayudarán a establecer qué campaña es más adecuada para un perfil en particular, optimizando el proceso de cobranza.



**TITULO :**        **Aplicación de Minería de Datos para Predecir Fuga  
de Clientes en la Industria de las Telecomunicaciones**  
(Barrientos, 2013)

**AUTOR :**        Barrientos, Francisco

**AÑO     :**        2013

#### RESUMEN

El trabajo de tesis presenta una metodología para predecir la fuga de clientes en un ambiente multiplataforma en la industria de telecomunicaciones. Se utilizan diversos algoritmos de minería de datos como redes neuronales, support vector machines y árboles de decisión para evaluar la calidad de su predicción, como el porcentaje de aciertos en la variable predicha.

La investigación concluye generando un modelo permite que la empresa en lugar de generar acciones al azar de la base de 9000 clientes del producto NGN, se focalice en un número reducido de 114 clientes de la empresa; que el modelo predijo como posibles fugas y por ende, generar estrategias más personalizadas para aumentar su retención.

#### 4.1.2. Antecedentes en el contexto nacional

**TITULO :**        **Implementación de un modelo de minería de datos  
para mejorar la toma de decisiones comerciales en la  
empresa Star Perú S.A.C. (Lopez Lopez & Velez  
Rojas, 2009)**

**AUTOR :**        Lopez Lopez & Velez Rojas

**AÑO     :**        2009

#### RESUMEN

El objetivo de la investigación fue implementar un Modelo de Minería de Datos para mejorar la toma de decisiones en la empresa STAR PERU S.A.C.

La muestra estuvo conformada por 10 directivos y trabajadores de la empresa STAR PERU S.A.C. del área comercial.

Analizaron la base de datos para obtener los campos que contenían datos útiles; Emplearon, modelos de clustering y árboles de decisión.

Se logró recolectar y clasificar los datos revisando y analizando las estructuras de éstas en las áreas involucradas, identificando los elementos básicos.

Esta investigación es de utilidad para el presente trabajo de tesis ya que se investigaron tendencias comerciales utilizando algoritmos de minería de datos de tipo clustering y árboles de decisión, evaluando su nivel de adaptación al problema y seleccionando el más adecuado.

#### 4.1.3. Antecedentes en el contexto regional

**TITULO :**        **Aplicación de redes neuronales artificiales para el pronóstico de la demanda de agua potable en la empresa EPSEL S.A de la ciudad de Lambayeque**  
(Vidaurre Siadén, 2012)

**AUTOR :**        Vidaurre Siadén, Yasmin

**AÑO:**        2011, Chiclayo – Lambayeque

##### RESUMEN

El trabajo de tesis demostró que el empleo de técnicas computacionales basadas en inteligencia artificial, como las Redes Neuronales Artificiales, con Arquitectura perceptrón de Multicapa; reducen el nivel de error de las predicciones de la demanda de agua potable. Se utilizaron redes neuronales artificiales como método eficaz y potencial de predicción aplicable a funciones de planificación en la alta dirección para empresas del sector saneamiento.

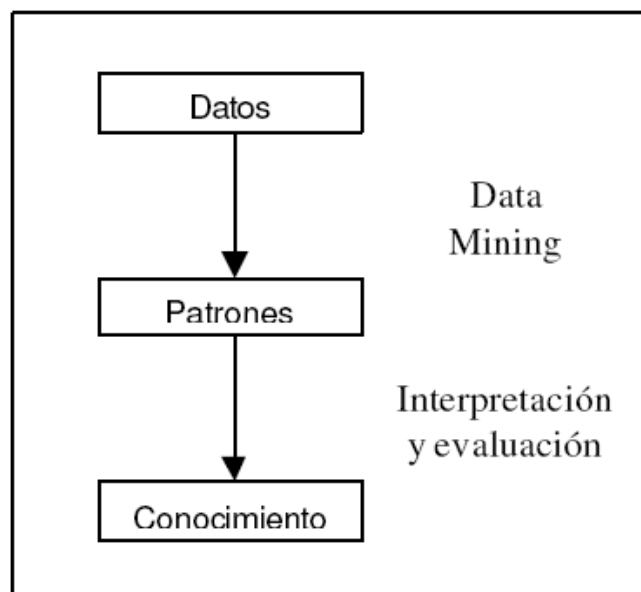
Se concluyó que el error del modelo de predicción de la cantidad demandada de agua potable a mediano plazo aplicando Redes Neuronales Artificiales con Arquitectura perceptrón de Multicapa, es significativamente menor que el obtenido por la empresa

## 4.2. Desarrollo de la temática

### 4.2.1. Descubrimiento del conocimiento (KDD)

(Valcárcel Asencios, 2004) En su investigación, cita que el autor (Molina, Data Mining "Torturando los datos hasta que confiesen", 2001), lo define como: La extracción no trivial de información potencialmente útil a partir de un gran volumen de datos, en el cual la información está implícita, donde se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones, para conseguirlo harán falta técnicas de aprendizaje, estadística y bases de datos”.

Las tareas comunes en KDD son la inducción de reglas, los problemas de clasificación y clustering, el reconocimiento de patrones, el modelado predictivo, la detección de dependencias, etc. Los datos recogen un conjunto de hechos (una base de datos) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto). El KDD involucra un proceso iterativo e interactivo de búsqueda de modelos, patrones o parámetros, los cuales descubiertos han de ser válidos, novedosos para el sistema y potencialmente útiles.



Fuente: (Valcárcel Asencios, 2004)

Ilustración 1 - Proceso de descubrimiento del conocimiento

#### **4.2.2. Minería de datos**

(Tumero, 2011) Nos dice que la minería de datos es el proceso que tiene como propósito descubrir, extraer y almacenar información relevante de amplias bases de datos, a través de programas de búsqueda e identificación de patrones y relaciones globales, tendencias, desviaciones y otros indicadores aparentemente caóticos que tienen una explicación que pueden descubrirse mediante diversas técnicas de esta herramienta.

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

##### **Objetivo**

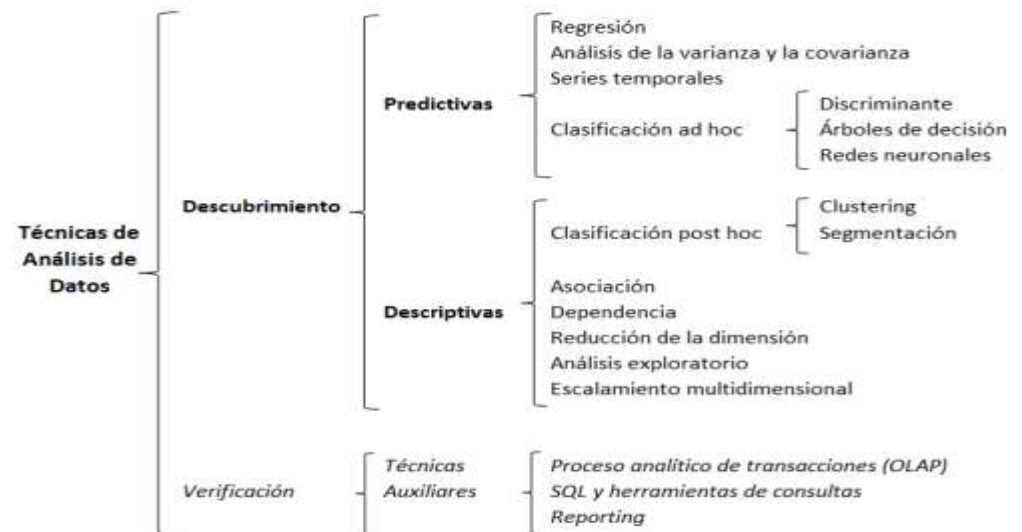
El objetivo principal de la minería de datos es resolver problemas analizando los datos que se encuentran en la base de datos determinando un patrón de comportamiento que define las características buscadas.

##### **Técnicas Descriptivas en minería de datos**

(Dandretta, 2002) Nos dice que hay tres formas de ver este punto, la primera se denomina caracterización de los datos (Data Caracterización), el cual realiza un resumen de las características generales de una clase particular de datos; los resultados suelen representarse en términos de reglas de caracterización.

La segunda es la discriminación de datos (Data Discrimination), que es una comparación entre las características generales de los objetos de una clase respecto a las de otro conjunto contrastante. Finalmente, también se puede aplicar una combinación de ambas. Análisis de asociación. Es el descubrimiento de reglas de asociación que muestran condiciones del tipo atributo-valor que ocurre con frecuencia dentro de un conjunto de datos.

La minería mediante reglas de asociación es el proceso de búsqueda interesante de correlaciones entre un conjunto grande de datos.



Fuente: (Valcárcel Asencios, 2004)

Ilustración 2- Técnicas de análisis de datos

El descubrimiento de reglas de asociación en grandes volúmenes de transacciones de negocios, puede facilitar el proceso de toma de decisiones. Por ejemplo, una regla de asociación descubierta en un conjunto de transacciones de libros de computación puede ser como sigue:

Sistema Operativo—→Linux [soporte = 3%, confianza =45%]

Esta regla refleja un modelo de compra para libros de computación, donde el consumidor que compra libros de sistemas operativos, tiende a comprar libros de Linux al mismo tiempo.

El soporte y la confianza son dos medidas que reflejan la utilidad y la certeza de la regla descubierta. En el ejemplo estos índices indican que el 45%de las transacciones que contienen libros de sistemas operativos también contienen libros de Linux y que el 3% de todas las transacciones contiene a ambos ítems.

**Análisis de clústers:** aquí se analizan objetos sin consultar clases conocidas. En general, las clases no se presentan en los datos de entrenamiento simplemente porque no se conocen. El proceso trabaja agrupando objetos según el principio de “maximizarla similitud dentro de una clase y minimizar la similitud entre clases”. Un clúster es una colección de objetos de datos mutuamente similares. Clustering es el proceso de agrupamiento de objetos. El análisis de clustering, tiene una gran variedad de aplicaciones, incluyendo procesos de imágenes, análisis de transacciones comerciales y reconocimiento de patrones.

### **Técnicas Predictivas en minería de datos**

(Dandretta, 2002) En sus investigaciones nos dice:

**Clasificación y predicción:** son dos tipos de análisis de datos, aquellos que pueden ser usados para clasificar datos y los que se usan para predecir tendencias. La clasificación de datos predice clases de etiquetas mientras la predicción de datos predice funciones de valores continuos.

Aplicaciones típicas incluyen análisis de riesgo para préstamos y predicciones de crecimiento. Algunas técnicas para clasificación de datos incluyen: clasificación bayesianas. K-Nearest Neighbor, algoritmos genéticos, entre otros.

**Árboles de decisión.** Definen un conjunto de clases, asignando a cada dato de entrada una clase y determina la probabilidad de que ese registro pertenezca a la clase.

Podemos distinguir dos tipos de árboles, el primero es el árbol de decisión de clasificación, donde cada registro a clasificar fluye por una rama del árbol. La rama a seguir es determinada por una serie de preguntas definidas por los nodos de la rama. Cuando el registro llega a un nodo hoja, se le asigna a la clase del nodo hoja.

El segundo es el árbol de decisión de regresión, cuando el registro llega a un nodo hoja, a la variable de salida de ese nodo, se le asigna el promedio de los valores de la variable de salida de los registros que cayeron en ese nodo hoja durante el proceso de entrenamiento.

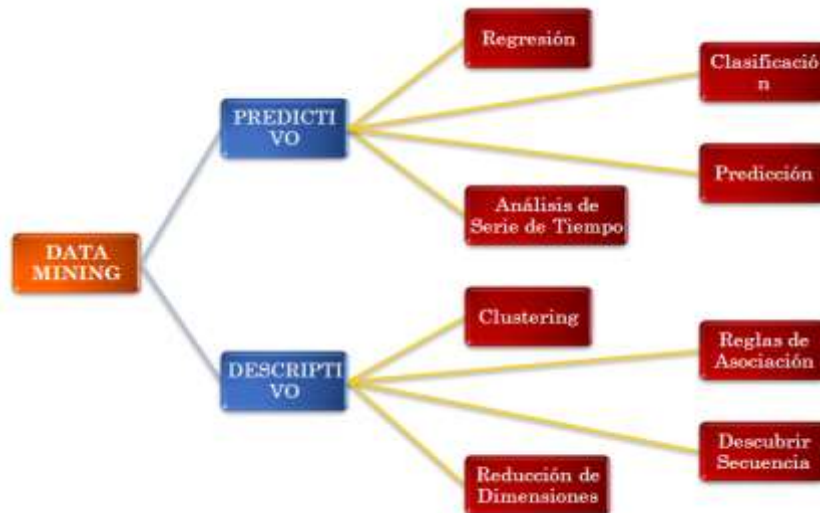
**Redes Neuronales.** Son modelos predictivos no lineales que aprenden a través del entrenamiento.

Existen diferentes tipos de redes neuronales, las más conocidas son las simples y multicapas. Las tareas básicas de las redes neuronales son reconocer, clasificar, agrupar, asociar, almacenar patrones, aproximación de funciones, sistemas (predicción, control, entre otros) y optimización.



#### 4.2.3. Algoritmos de minería de datos

Según (Valcárcel Asencios, 2004) la aplicación de técnicas de minería de datos en grandes bases de datos persiguen los siguientes resultados:



(Kimball, 1998)

Ilustración 3 - Clasificación de algoritmos de minería de datos

**Clasificación:** Se trata de obtener un modelo que permita asignar un caso de clase desconocida a una clase concreta (seleccionada de un conjunto redefinido de clases), como son los árboles de clasificación (CART), cuyos resultados pueden expresarse mediante reglas ejecutables directamente del SQL o el método de Bayesiano.

**Regresión:** Se persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable (modelos de regresión logística).

**Agrupamiento (clustering):** Hace corresponder cada caso a una clase, con la peculiaridad de que las clases se obtienen directamente de los datos de entrada utilizando medidas de similitud. Es decir, agrupan a los datos bajo diferentes métodos y criterios. Las técnicas más usadas son las clásicas (distancia mínima) y las redes neuronales (método de Kohonen o método de Neural-Gas).

**Resumen:** Se obtienen representaciones compactas para subconjuntos de los datos de entrada (análisis interactivo de datos, generación automática de informes, visualización de datos).

**Modelado de Dependencias:** Se obtienen descripciones de dependencias existentes entre variables. El análisis de relaciones (por ejemplo las reglas de asociación), en el que se determinan relaciones existentes entre elementos de una base de datos, podría considerarse un caso particular de modelado de dependencias.

**Análisis de Secuencias:** Se intenta modelar la evolución temporal de alguna variable, con fines descriptivos o predictivos (redes neuronales multicapas).

#### 4.2.4. Modelo de procesos

Son diversos los modelos de proceso que han sido propuestos para el desarrollo de proyectos de Data Mining tales como SEMMA (Sample, Explore, Modify, Model, Assess) [SAS, 2003], KDD (Knowledge Discovery in Databases) , DMAMC (Definir, Medir, Analizar, Mejorar, Controlar) [Sixsigma, 2005], o CRISP-DM (Cross Industry Standard Process for Data Mining) [CRISP-DM, 2000], sin embargo uno de los modelos principalmente utilizados en los ambientes académico e industrial es el modelo CRISPDM.

Mencionaremos estas metodologías dominantes para el proceso de la minería de datos ahondando en la de CRISP-DM, que será la utilizada finalmente en este proyecto.

##### 4.2.4.1. SEMMA

SEMMA es el acrónimo a las cinco fases: (Sample, Explore, Modify, Model, Assess) La metodología es propuesta por SAS Institute Inc, la define como: “... proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones de negocios desconocidos...”

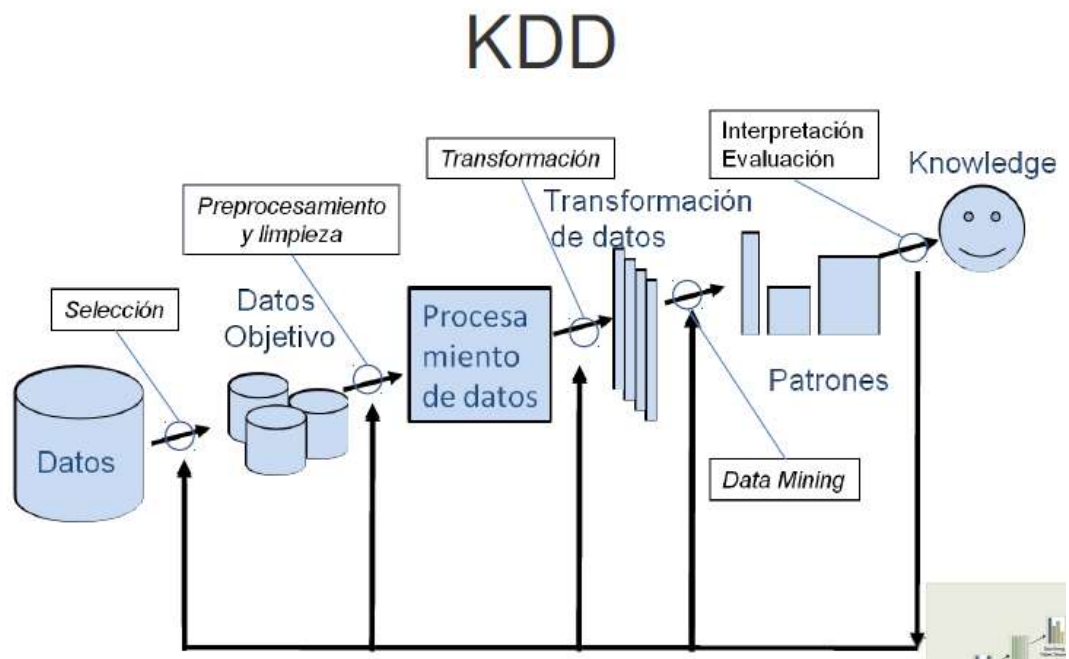
##### Fases y actividades SEMMA



#### 4.2.4.2. KDD (Knowledge Discovery in Databases)

Es una metodología propuesta por Fayyad [3] en 1996, propone 5 fases: Selección, preprocesamiento, transformación, minería de datos y evaluación e implantación. Es un proceso iterativo e interactivo.

La Extracción de conocimiento está principalmente relacionado con el proceso de descubrimiento conocido como *Knowledge Discovery in Databases* (KDD), que se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información [1]. No es un proceso automático, es un proceso iterativo que exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones. Es un proceso que extrae información de calidad que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos. La siguiente figura ilustra las etapas del proceso KDD:



Como muestra la figura anterior, las etapas del proceso KDD se dividen en 5 fases y son:

1. **SELECCIÓN DE DATOS.** En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos.
2. **PREPROCESAMIENTO.** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.
3. **TRANSFORMACIÓN.** Consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.
4. **DATA MINING.** Es la fase de modelamiento propiamente tal, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u “ocultos” en los datos.
5. **INTERPRETACIÓN Y EVALUACIÓN.** Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

Además de las fases descritas, frecuentemente se incluye una fase previa de análisis de las necesidades de la organización y definición del problema, en la que se establecen los objetivos de la minería de datos. También es usual incluir una etapa final, donde los resultados obtenidos se integran al negocio para la realización de acciones comerciales.

#### **4.2.4.3. DMAMC (Definir, Medir, Analizar, Mejorar, Controlar)**

Six Sigma propone cinco fases en su secuencia metodológica:

- Definir (DEFINE)
- Medir (MEASURE)
- Analizar (ANALIZE)
- Mejorar (IMPROVE)
- Controlar (CONTROL)

##### **Fase 0: DEFINIR.**

- Se establecen las necesidades de la empresa.
- Se identifican los procesos que deben mejorarse
- Pasos clave:
  - Crear un enunciado del problema
  - Identificar parámetros críticos para la satisfacción (CTQ's) \*\*
  - Definir estándares de desempeño
  - Negociar estándares de desempeño (cartera)

##### **Fase 1: MEDIR.**

- Determinar las características del producto que son críticas para la satisfacción del cliente (CTQ's; Customer Total Qualities)
- “Enfocar” o buscar las variables claves del proceso que explican la variación indeseable de las características CTQ's
- Completar un análisis del sistema de medición.
- Establecer una línea base: estimar la capacidad del proceso a corto y largo plazo
- Algunas de las herramientas son:
  - Mapa del proceso
  - AMFE (Análisis de modo de falla y efectos)
  - Planillas de control para datos medidos
  - Gage R&R (Repetibilidad y reproducibilidad de la medición)
  - Capacidad de proceso

## **Fase 2: ANALIZAR.**

- Afinar la búsqueda para las variables claves del proceso
  - Ensayo de hipótesis
  - Correlación/regresión
- Confirmar las métricas necesarias para medir los CTQ's
- En algunos casos, recomendar el rediseño del proceso del producto
- Algunas de las herramientas son:
  - Intervalos de confianza
  - Potencia y tamaño de la muestra
  - Análisis multivariantes
  - Ensayo de hipótesis
  - Correlación/regresión
  - Análisis de la varianza (ANOVA)

## **Fase 3: MEJORAR.**

- Terminar la búsqueda de variables claves del proceso
- Determinar el efecto de las variables claves del proceso en la variación indeseable en las características CTQ
- Establecer los niveles de desempeño para las variables de proceso que reducen la variación indeseable en cada CTQ
  - Caracterización
  - Optimización
- Algunas de las herramientas son:
  - Bloques totalmente aleatorios
  - Experimentos factoriales completos
  - Experimentos factoriales fraccionados
  - Optimización de la respuesta
  - Metodología de superficie de respuesta

#### **Fase 4: CONTROLAR.**

- Asegurar que las nuevas condiciones del proceso estén documentadas y monitoreadas con métodos de control estadísticos de procesos
- Después de un período de “asentamiento”, volver a estimar la capacidad del proceso.
- Dependiendo de los resultados de los análisis de seguimiento, repetir una o más de las fases precedentes
- Algunas de las herramientas son:
  - Planes de control
  - EVOP
  - Control estadístico de procesos
  - Análisis de capacidad de proceso



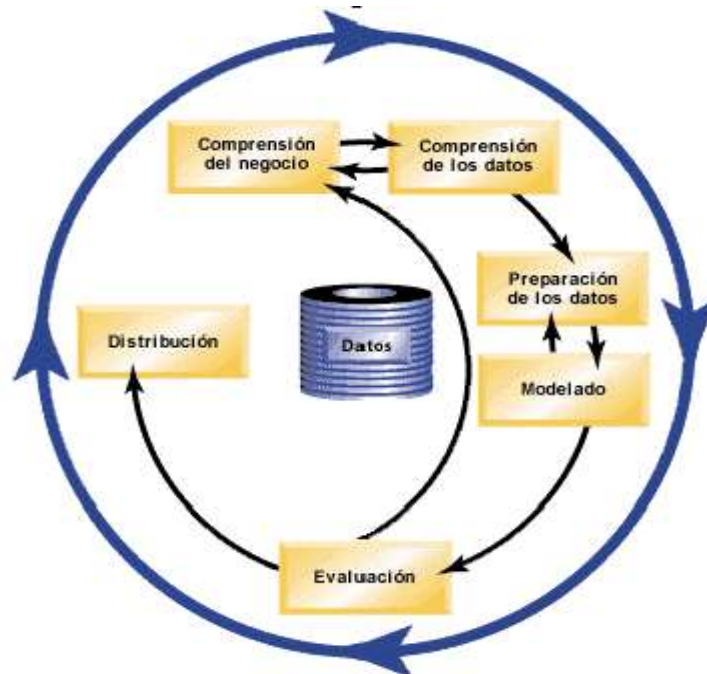
#### 4.2.4.4. CRISP-DM

Según (IBM, 2012)

CRISP-DM, que son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar trabajos de minería de datos.

- Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.
- Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

El ciclo vital del modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario.



Fuente: (IBM, 2012)

Ilustración 4 - Ciclo de vida de CRISP-DM

## **Fase 1 - Comprensión del negocio**

Debe dedicar tiempo a explorar las expectativas de su organización con respecto a la minería de datos. Intente implicar a la mayor cantidad de personas que sea posible en estas discusiones y documente los resultados. El paso final de la fase de CRISP-DM trata de cómo producir un plan de proyecto utilizando la información que se contiene en esta documentación.

Aunque este estudio pueda parecer prescindible, no lo es. Conozca las razones comerciales para que sus esfuerzos en minería de datos aseguren que todos los usuarios están de acuerdo antes de asignar recursos.

### **Determinación de los objetivos comerciales**

Es el punto en el que las soluciones se especifican. Como resultado de sus investigaciones y reuniones, debe crear un objetivo principal concreto acordado por los patrocinadores del proyecto y otras unidades comerciales que se vean afectadas por los resultados.

Este objetivo se puede traducir de forma eventual de algo tan nebuloso como “reducir la tasa de abandono de clientes” a objetivos específicos de minería de datos que dirigirán el análisis.

### **Valoración de la situación**

Ahora que ha definido un objetivo comercial claro, es hora de realizar una valoración de su situación actual. Este paso implica cuestiones como:

- ¿Qué tipos de datos están disponibles para el análisis?
- ¿Dispone del personal necesario para completar el proyecto?
- ¿Cuáles son los principales factores de riesgo?
- ¿Dispone de planes de contingencia para cada factor de riesgo?

### **Determinación de los objetivos de minería de datos**

Ahora que el objetivo comercial ha quedado claro, es hora de traducirlo en una realidad de minería de datos. Por ejemplo, el objetivo comercial para “reducir el abandono” se puede traducir en un objetivo de minería de datos que incluye:

- Identificación de clientes de mayor valor en función de los datos de compra recientes
- Creación de un modelo utilizando datos disponibles de clientes para pronosticar las posibilidades de abandono de cada uno
- Asignación de un rango a cada cliente basado en las posibilidades de abandono y valor del cliente.

Estos objetivos de minería de datos, si se cumplen, se pueden utilizar por la empresa para reducir el abandono entre los clientes de mayor valor.

### **Producción de un plan de proyecto**

En este punto, ya está listo para producir un plan para el proyecto de minería de datos. Las cuestiones que haya planteado hasta el momento y los objetivos comerciales y de minería de datos que haya formulado formarán la base de este plan.

## **Fase 2 - Comprensión de los datos**

La fase de comprensión de datos de CRISP-DM implica estudiar más de cerca los datos disponibles de minería. Este paso es esencial para evitar problemas inesperados durante la siguiente fase (preparación de datos) que suele ser la fase más larga de un proyecto.

La comprensión de datos implica acceder a los datos y explorarlos con la ayuda de tablas y gráficos, de esta forma podrá determinar la calidad de los datos y describir los resultados de estos pasos en la documentación del proyecto.

## Recopilación de datos iniciales

En este punto en CRISP-DM, puede acceder a los datos de diferentes fuentes como:

- Datos existentes. Incluye una amplia variedad de datos, como datos transaccionales, datos de encuesta, registros Web, etc. Tenga en cuenta si los datos existentes son suficientes para adaptarse a sus necesidades.
- Datos adquiridos. ¿Su organización utiliza datos adicionales, como datos demográficos? Si no los utiliza, considere si son necesarios.
- Datos adicionales. Si las fuentes anteriores no satisfacen sus necesidades, es posible que necesite realizar encuestas o realizar seguimientos adicionales para servir de complemento a los almacenes de datos actuales.

## Descripción de los datos

Existen muchas formas de describir datos, pero la mayoría de datos se centra en la cantidad y calidad de los datos; la cantidad de datos disponible y el estado de los datos. A continuación se incluyen algunas características clave para describir datos.

- **Cantidad de datos.** En la mayoría de técnicas de modelado, los tamaños de datos tienen un equilibrio relacionado. Los grandes conjuntos de datos pueden producir modelos más precisos, pero también pueden aumentar el tiempo de procesamiento. Considere utilizar un subconjunto de datos. Cuando tome notas para el informe final, asegúrese de incluir estadísticos de tamaños para todos los conjuntos de datos y recuerde tener en cuenta tanto el número de registros como los campos (atributos) cuando describa los datos.

- **Tipos de valores.** Los datos pueden incluir una variedad de formatos, como numérico, categórico (cadena) o Booleano (verdadero/falso). Si presta atención al tipo de valor puede evitar posteriores problemas durante la fase de modelado.
- **Esquemas de codificación.** Con frecuencia, los valores de la base de datos son representaciones de características como género o tipo de producto. Por ejemplo, un conjunto de datos puede utilizar M y F para representar masculino y femenino, mientras que otro puede utilizar los valores numéricos 1 y 2. Registre los esquemas incoherentes en el informe de datos.

### **Exploración de los datos**

Esta fase de CRISP-DM permite explorar los datos con las tablas, gráficos y otras herramientas de visualización disponibles. Estos análisis pueden ayudarle a describir los objetivos de minería de datos generados durante la fase de comprensión comercial.

También pueden ayudarle a formular hipótesis y dar forma a las tareas de transformación de datos que tienen lugar durante la preparación de los datos.

### **Verificación de calidad de los datos**

Los datos no suelen ser perfectos. De hecho, la mayoría de los datos contienen errores de codificación, valores perdidos u otro tipo de incoherencias que hacen que los análisis resulten difíciles en algunas ocasiones. Una forma de evitar posibles problemas es realizar un análisis de calidad de los datos disponibles antes de proceder al modelado.

### **Fase 3 – Preparación de los datos**

La preparación de datos es uno de los aspectos más importantes y con frecuencia que más tiempo exigen en la minería de datos. De hecho, se estima que la preparación de datos suele llevar el 50-70 % del tiempo y esfuerzo de un proyecto. Dedicar los esfuerzos adecuados a las primeras fases de comprensión comercial y comprensión de datos puede reducir al mínimo los gastos indirectos relacionados, pero aún deberá dedicar una buena cantidad de esfuerzo para preparar y empaquetar los datos para la minería. Dependiendo de su organización y sus objetivos, la preparación de datos suele implicar las tareas siguientes:

- Fusión de conjuntos y/o registros de datos.
- Selección de una muestra de un subconjunto de datos.
- Agregación de registros.
- Derivación de nuevos atributos.
- Clasificación de los datos para el modelado.
- Eliminación o sustitución de valores en blanco o ausentes.
- División en conjuntos de datos de prueba y entrenamiento.

#### **Selección de datos**

En función de la recopilación de datos inicial realizada en la fase CRISP-DM anterior, ahora puede comenzar a seleccionar los datos relevantes a sus objetivos de minería de datos. De forma general, existen dos formas de seleccionar datos:

- Selección de elementos (filas) implica la toma de decisiones como las cuentas, productos o clientes que se van a incluir.
- Selección de atributos o características (columnas) implica la toma de decisiones sobre el uso de características como la cantidad de las transacciones o los ingresos por hogar.

### **Limpieza de datos**

La limpieza de datos implica observar más de cerca los problemas en los datos que ha seleccionado incluir en el análisis. El informe de calidad de datos preparado durante la fase de comprensión de datos contiene detalles sobre los tipos de problemas concretos de sus datos.

### **Construcción de nuevos datos**

Con frecuencia, necesitará construir nuevos datos. Por ejemplo, puede ser de gran utilidad crear una nueva columna con la adquisición de una garantía ampliada en cada transacción.

Existen dos formas de construir nuevos datos:

- Derivación de atributos (columnas o características)
- Generación de registros (filas)

### **Integración de datos**

No es raro disponer de varios orígenes de datos para el mismo conjunto de cuestiones comerciales. Por ejemplo, puede tener acceso a los datos de un crédito hipotecario, así como a los datos demográficos para el mismo conjunto de clientes. Si estos conjuntos de datos contienen el mismo identificador único (como el número de seguridad social), puede fusionarlos.

Existen dos métodos básicos para integrar los datos:

- La fusión de datos implica unir dos conjuntos de datos con registros similares, pero con atributos diferentes. Los datos se fusionan utilizando el mismo identificador clave en cada registro (como el ID de usuario). Los datos resultantes aumentan las columnas o las características.
- La adición de datos implica integrar dos o más conjuntos de datos con atributos similares, pero con registros diferentes. Los datos se integran en función de los campos similares (como el nombre de producto o la longitud del contrato).

## **Formato de datos**

Como paso final antes de la construcción del modelo, es muy útil comprobar si algunas técnicas requieren aplicar un formato concreto o la clasificación de los datos. Por ejemplo, no es extraño que un algoritmo de secuencia requiera que los datos estén clasificados de forma previa antes de ejecutar el modelo. Incluso si el modelo puede ejecutar la clasificación de forma automática, puede ahorrar tiempo si utiliza un nodo “Ordenar” antes del modelado.

## **Fase 4 – Modelado**

Este es el punto donde todo el duro trabajo anterior comienza a tener sentido. Los datos que ha preparado se incorporan a las herramientas analíticas y los resultados comenzarán a arrojar algo de luz al problema planteado en Comprensión del negocio.

El modelado se suele ejecutar en múltiples iteraciones. Normalmente, los analistas de datos ejecutan varios modelos utilizando los parámetros por defecto y ajustan los parámetros o vuelven a la fase de preparación de datos para las manipulaciones necesarias por su modelo. Es extraño que las cuestiones relativas a la minería de datos de una empresa se solucionen satisfactoriamente con un modelo y ejecución únicos. Esto es lo que hace la minería de datos tan interesante; existen muchas formas para resolver un problema concreto.

## **Selección de la técnica de modelado**

Aunque pueda tener algunos conocimientos acerca de los tipos de modelado que sean los más adecuados para las necesidades de su organización, es el momento de tomar la decisión de los tipos de modelado que se van a utilizar. La determinación del modelado más adecuado se basará en las siguientes consideraciones:



- Los tipos de datos disponibles para la minería. Por ejemplo, ¿los campos de interés son categóricos (simbólicos)?
- Sus objetivos de minería de datos. ¿Sólo quiere tener un mejor conocimiento de los almacenes de datos transaccionales y descubrir patrones de compras interesantes? ¿Necesita producir una puntuación indicando, por ejemplo, las posibilidades de impago de un préstamo a un estudiante?
- Requisitos específicos de modelado. ¿Necesita el modelo un tipo o un tamaño de datos concreto? ¿Necesita un modelo con unos resultados fácilmente presentables?

### **Generación de un diseño de comprobación**

Como paso final antes de generar el modelo, debe volver a tener en cuenta cómo se comprobarán los resultados del modelo. Existen dos partes para generar un diseño de comprobación global:

- Descripción de los criterios de “bondad” de un modelo
- Definición de los datos en los que se comprobarán estos criterios

La bondad de un modelo se puede medir de varias formas. Para modelos supervisados, como C5.0 y C&R Tree, las mediciones de bondad suelen calcular la tasa de error de un modelo concreto. Para modelos sin supervisión, como redes de conglomerados de Kohonen, las mediciones pueden incluir criterios como facilidad de interpretación, distribución o el tiempo de procesamiento necesario.

### **Generación de los modelos**

En este punto, debe tener la preparación suficiente para generar los modelos que haya considerado. Tómese el tiempo necesario para experimentar con diferentes modelos antes de llegar a conclusiones definitivas. La mayoría de analistas de datos suelen generar varios modelos y comparar los resultados antes de aplicarlos o integrarlos.

Para poder registrar su progreso con una amplia variedad de modelos, asegúrese de registrar los ajustes y datos utilizados para cada modelo. De esta forma podrá analizar los resultados con otras personas y comprobar sus pasos si fuera necesario. Al final del proceso de generación de modelos dispondrá de tres tipos de información que puede utilizar en la toma de decisiones de minería de datos:

- Configuración de parámetros incluye las notas que ha tomado sobre los parámetros que producen los mejores resultados.
- Los modelos reales producidos.
- Descripciones de resultados de modelos, incluyendo problemas de datos y rendimiento que hayan ocurrido durante la ejecución del modelo y exploración de los resultados

### **Evaluación del modelo**

Ahora que ha definido un conjunto de modelos iniciales, obsérvelos detenidamente para determinar cuáles son los más precisos o eficaces para considerarse finales. Finales puede significar varias cosas, como “listo para aplicar” o “ilustra patrones interesantes”. Si consulta el plan de pruebas que ha creado previamente, puede ayudarle a crear esta valoración desde el punto de vista de su organización.

### **Fase 5 – Evaluación**

En este punto, habrá completado la mayor parte de su proyecto de minería de datos. También habrá determinado, en la fase de modelado, que los modelos son técnicamente correctos y efectivos en función de los criterios de rendimiento de minería de datos que ha definido previamente.

Sin embargo, antes de continuar, debe evaluar los resultados de sus esfuerzos utilizando los criterios de rendimiento comercial establecidos en el inicio del proyecto. Es la clave para asegurar que su organización pueda utilizar los resultados que ha obtenido. La minería de datos produce dos tipos de resultados:

- Los modelos finales seleccionados en la fase anterior de CRISP-DM.
- Las conclusiones o interferencias obtenidas de los modelos y del proceso de minería de datos.

### **Evaluación de los resultados**

En esta etapa, formalizará su evaluación en función de si los resultados del proyecto cumplen los criterios del rendimiento comercial. Este paso requiere una clara comprensión de los objetivos comerciales, por lo que debe estar seguro de incluir factores de toma de decisiones en la evaluación del proyecto.

### **Proceso de revisión**

Las metodologías eficaces suelen incluir tiempo para reflexionar sobre los aciertos y errores del proceso que se acaba de completar. La minería de datos no es muy diferente. Una parte fundamental de CRISP-DM es aprender de su propia experiencia para que sus proyectos de minería de datos sean más efectivos.

### **Determinación de los pasos siguientes**

Por ahora ha obtenido unos resultados, ha evaluado su experiencia de minería de datos y se debe estar preguntando, ¿qué viene a continuación? Esta fase le ayuda a responder esa pregunta en términos de objetivos comerciales de minería de datos. Básicamente, llegados a este punto dispone de dos opciones:

- Continuar con la fase de desarrollo. La siguiente fase le ayudará a incorporar los resultados del modelo a su proceso comercial y producir un informe final. Incluso si sus esfuerzos invertidos en la minería de datos no han sido satisfactorios, debe utilizar la fase de

desarrollo de CRISP-DM para crear un informe final para su distribución al patrocinador del proyecto.

- Volver y refinar o sustituir los modelos. Si encuentra que los resultados son casi óptimos, pero no lo suficiente, considere otro tipo de modelado. Puede utilizar sus conocimientos adquiridos en esta fase para refinar los modelos y producir mejores resultados.

En este punto, su decisión incluye la precisión y relevancia de los resultados de modelado. Si los resultados se adaptan a sus objetivos comerciales de minería de datos, puede pasar a la fase de aplicación. Con independencia de la decisión que tome, asegúrese de registrar el proceso de evaluación.

## **Fase 6 - Distribución**

La distribución es el proceso que consiste en utilizar sus nuevos conocimientos para implementar las mejoras en su organización. Además, la distribución puede significar que utilice los conocimientos adquiridos en minería de datos para aplicar modificaciones en su organización. Por ejemplo, es posible que descubra patrones de alarma en sus datos que indican un cambio en el comportamiento de los clientes de más de 30 años. Es posible que estos resultados no se integren formalmente en sus sistemas de información, pero serán de gran utilidad para la planificación y toma de decisiones de marketing.

#### 4.2.4.5. Comparación entre KDD, SEMMA y CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	xxxxx	Conocimiento del negocio
Selección	muestra	Conocimiento de los datos
Preprocesamiento	exploración	
Transformación	Modificación	
Minería de datos	Modelo	
interpretación / evaluación	evaluación	
Post KDD	xxxxx	

## CAPÍTULO V: DESARROLLO DE LA PROPUESTA

### DETERMINACION DE REQUERIMIENTOS

Para

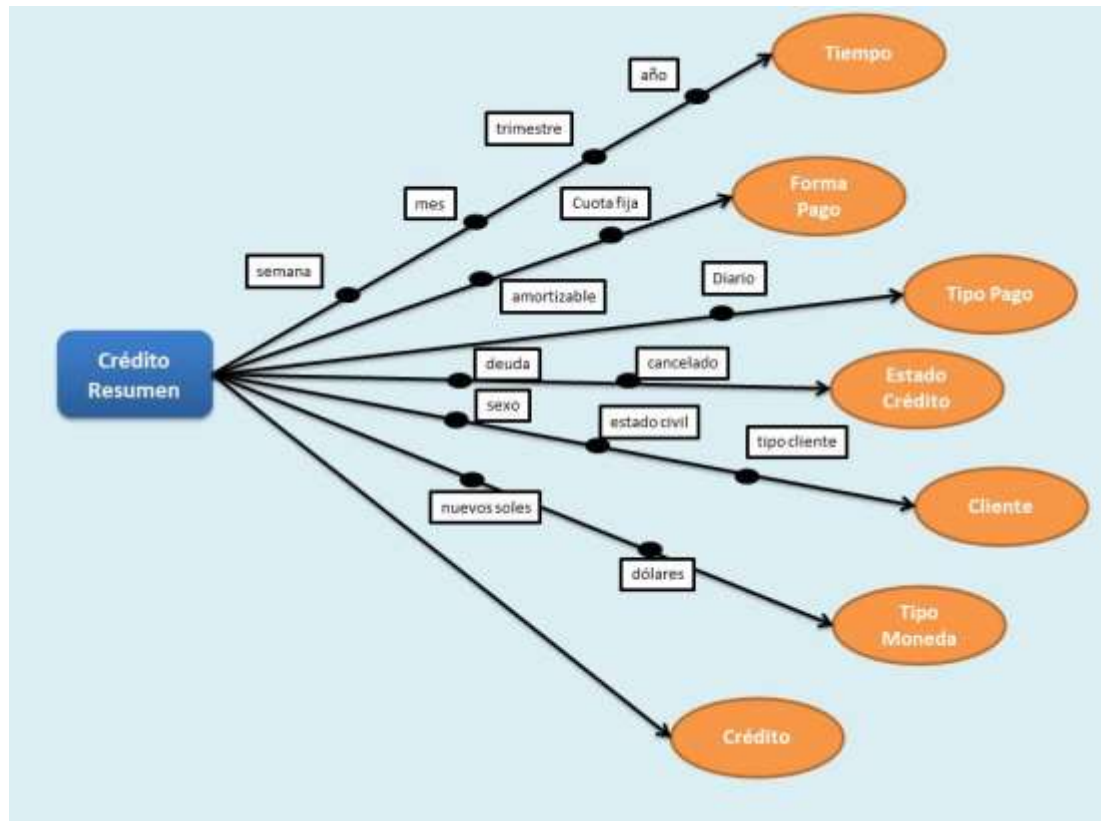


Ilustración 5 - Determinación de requerimientos

### DISEÑO DE DATAMART

Se determinó la dimensionalidad de cada indicador y luego se especificaron los diferentes grados de detalle (atributos) dentro de cada concepto del negocio (dimensión), así también como la granularidad de cada indicador (variable o métrica) y las diferentes jerarquías que dan forma al Modelo Dimensional del Negocio (BDM).

Las construcciones primarias son tablas de hechos (hecho creditoResumen). La tabla de Hecho contiene métricas derivadas del hecho de créditos: importe de créditos (en nuevos soles), número de cuotas y cantidad de días de deuda.

El modelo dimensional debe ser estructurado alrededor de un proceso del negocio (hecho resumen de crédito). La granularidad de la tabla de hechos, debe ser lo más atómico posible, esto permite mayor flexibilidad y extensibilidad.

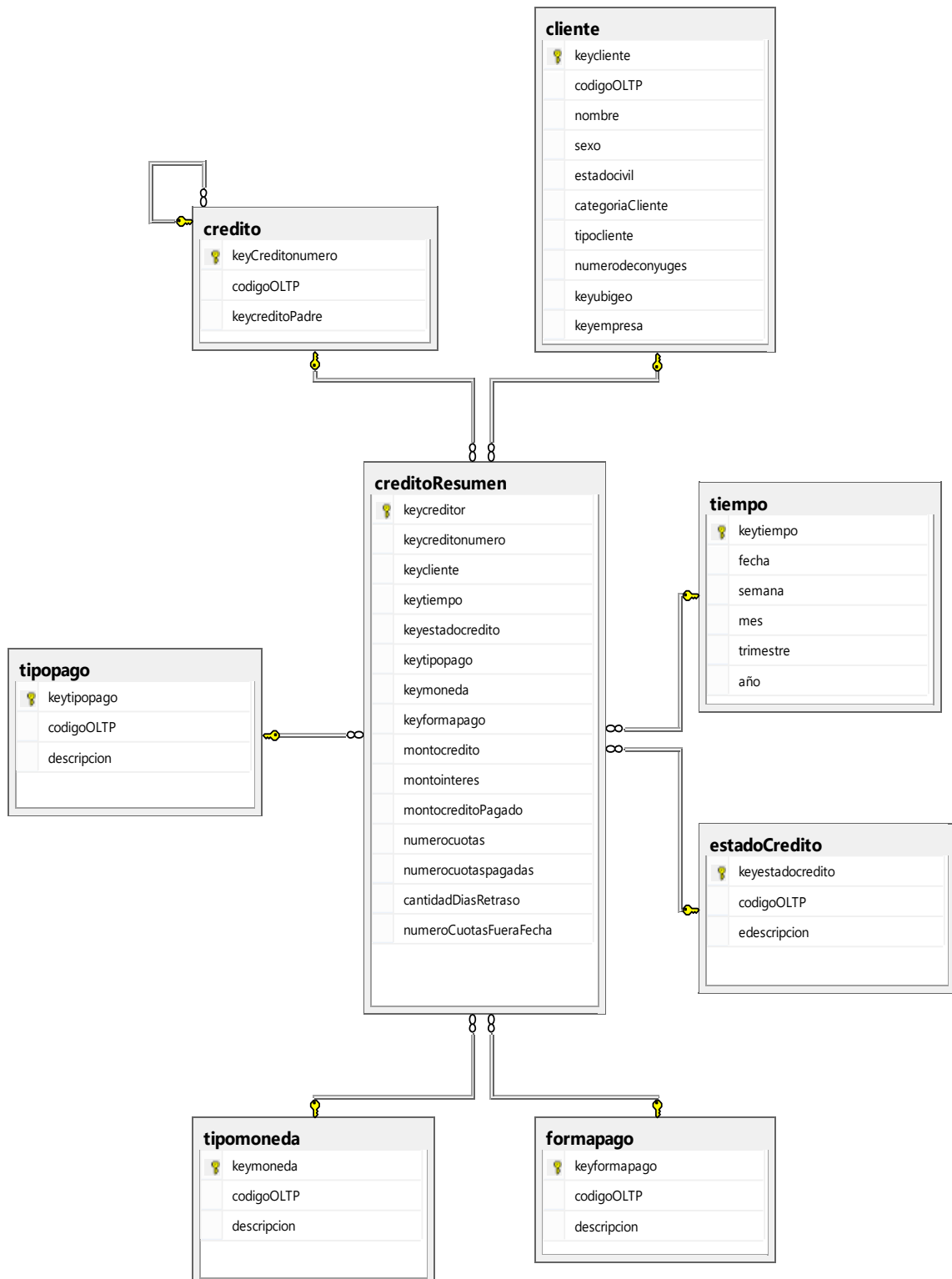
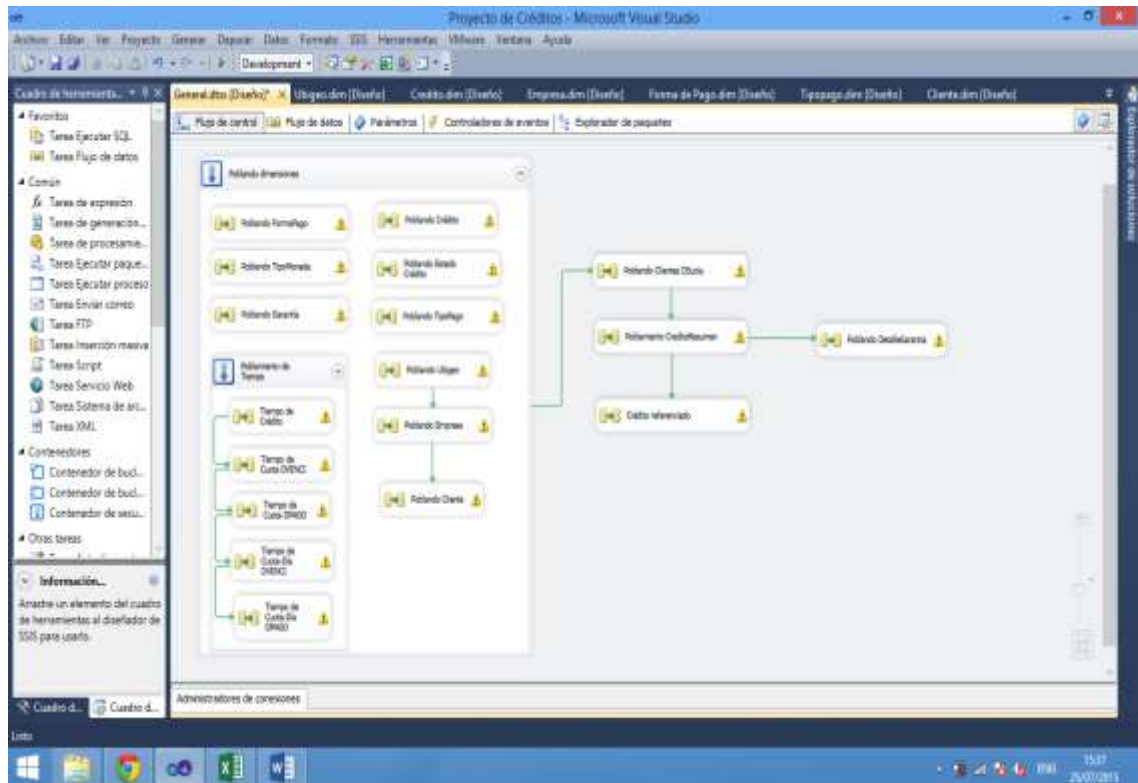


Ilustración 6 - Data Mart



## PROCESO ETL

Poblamiento del DataMart para análisis de las colocaciones de crédito de la financiera Crediserv EIRL, se define el poblamiento de las dimensiones y luego el poblamiento de la FACT TABLE mediante el proceso de extracción, transformación y carga de datos



### Ilustración 7 - Proceso ETL

Se definen las siguientes dimensiones:

- Forma de pago
- Tipo de moneda
- Garantía
- Crédito
- Estado de crédito
- Cliente
- Tiempo



## Poblamiento de la Dimensión Tiempo

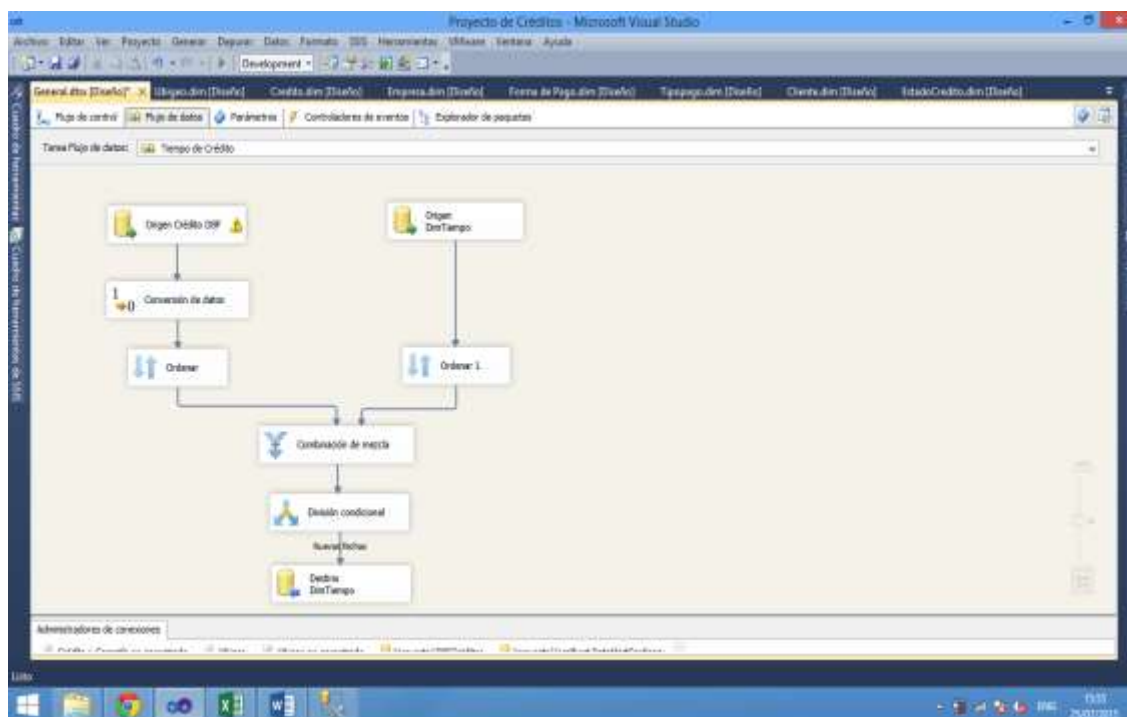


Ilustración 10 - Dimensión Tiempo

## Poblamiento de la Dimensión Forma de Pago

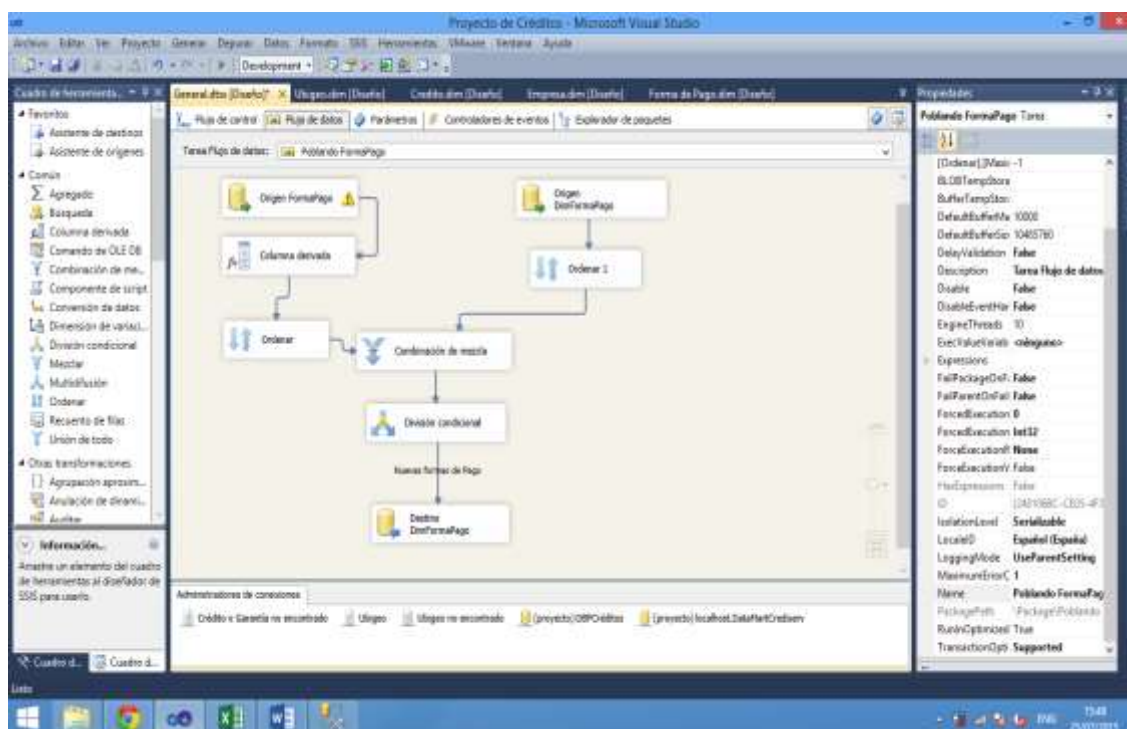


Ilustración 11 - Dimensión Forma de Pago

## Poblamiento de la Dimensión Tipo moneda

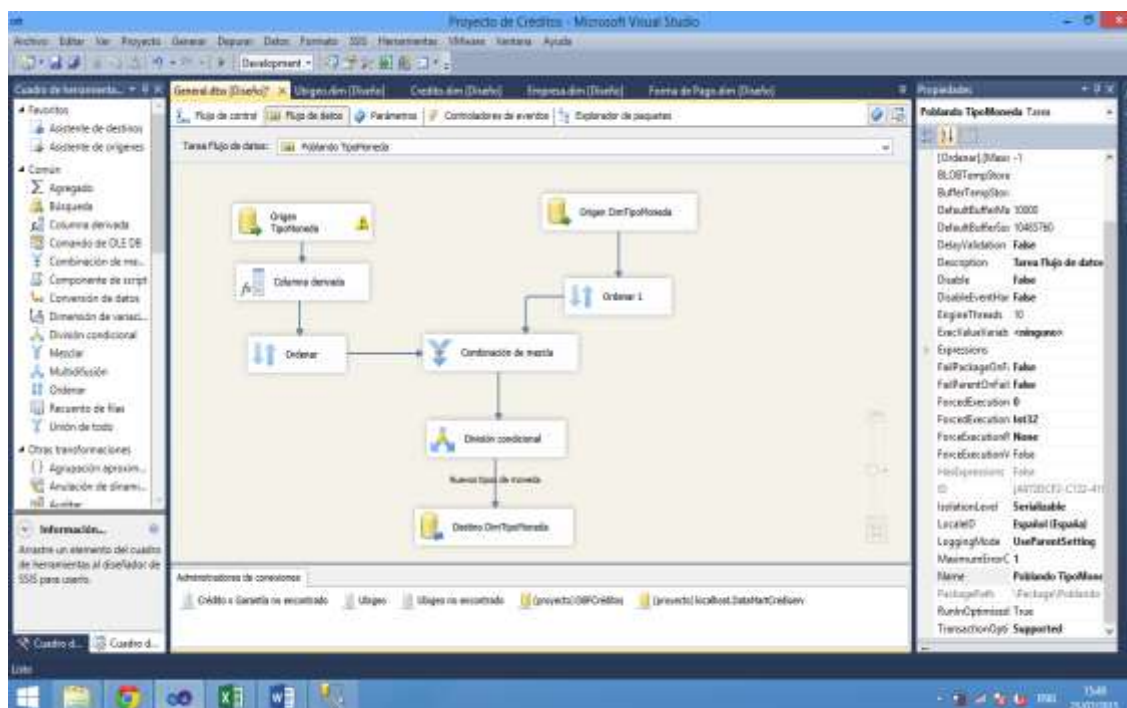


Ilustración 12 - Dimensión tipo moneda

### Poblamiento de la Dimensión Estado Crédito

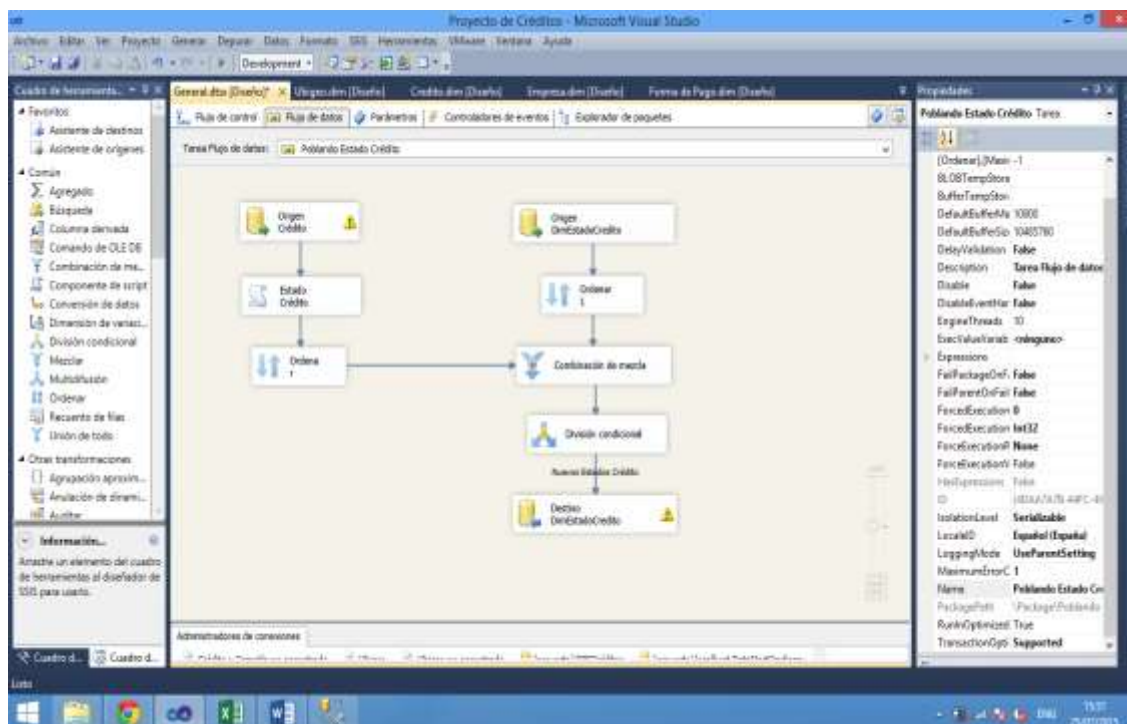


Ilustración 13 - Dimensión Estado Crédito

### Poblamiento de la Dimensión Tipo Pago





## PROCESO DE DISEÑO DE LA BASE DE DATOS OLAP

Las dimensiones del cubo se diseñaran a partir de las tablas dimensión del DataMart poblado en el proyecto de ETL

### Dimensión Tiempo

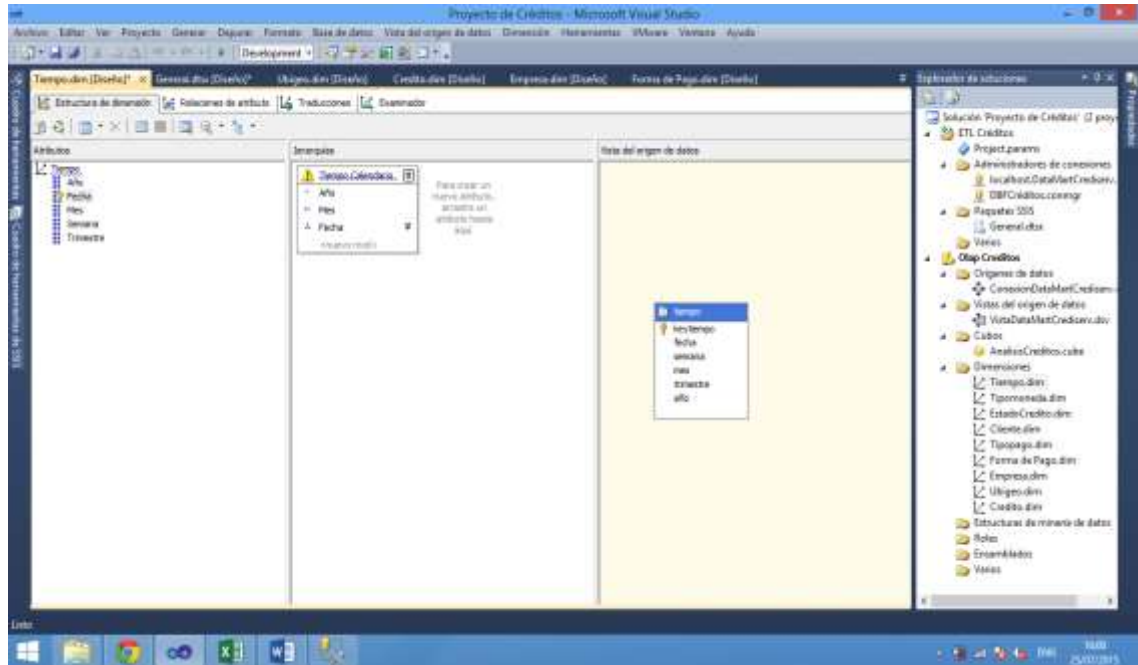


Ilustración 15 - Dimensión tiempo

### Dimensión Tipo Moneda

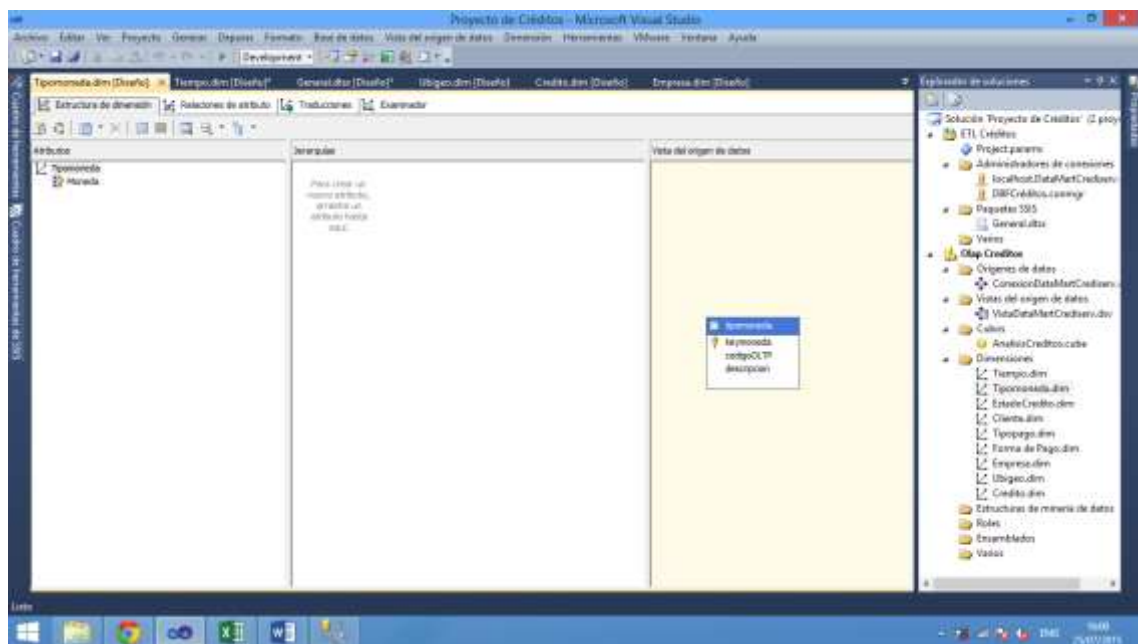
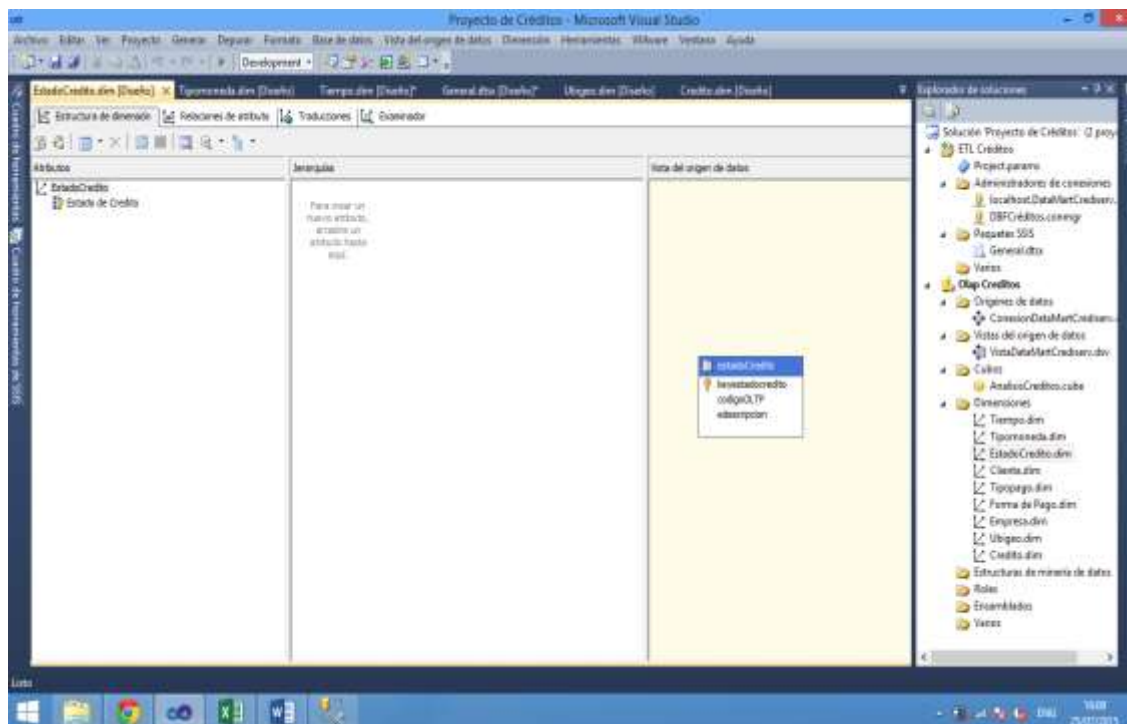


Ilustración 16 - Dimensión moneda



## Dimensión Cliente

## Dimensión Tipo Pago

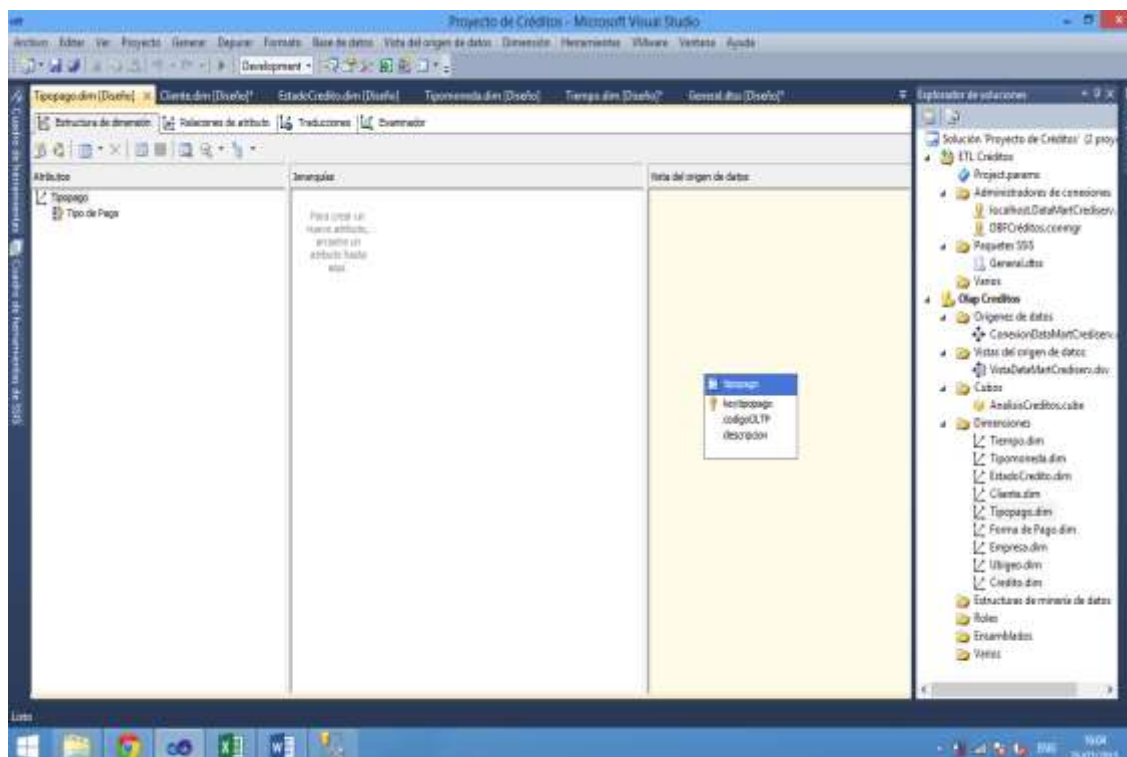


Ilustración 19 - Dimensión tipo de pago

## Dimensión Forma Pago

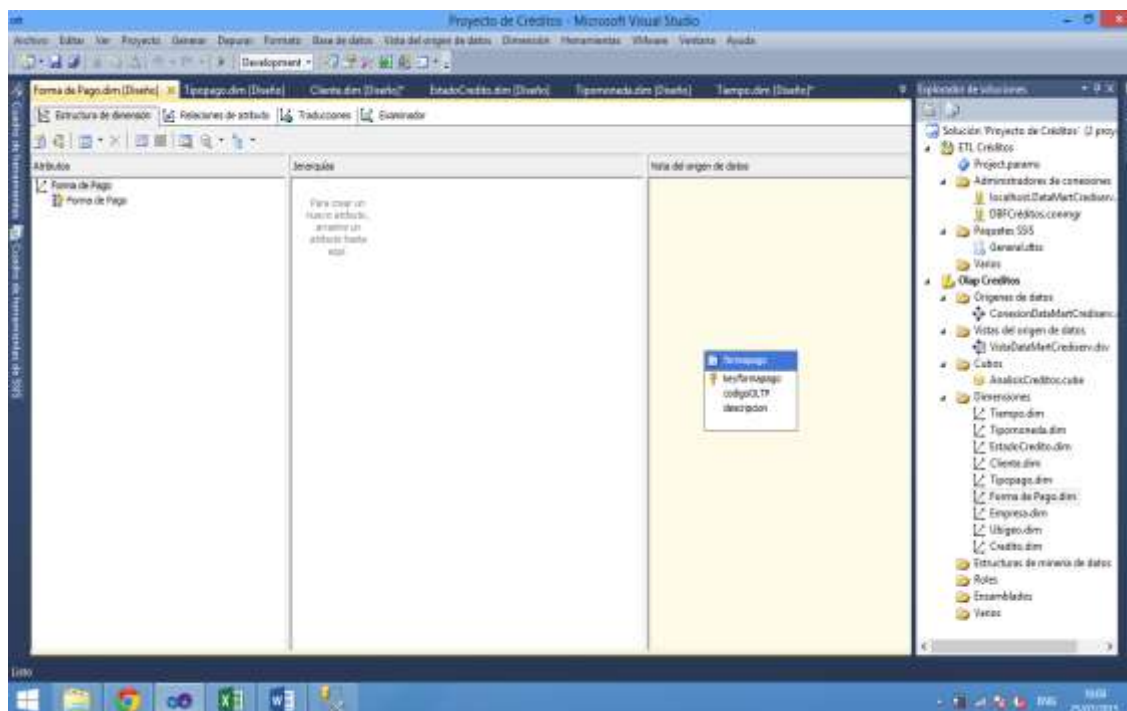


Ilustración 20 - Dimensión forma de pago



## Dimensión Crédito

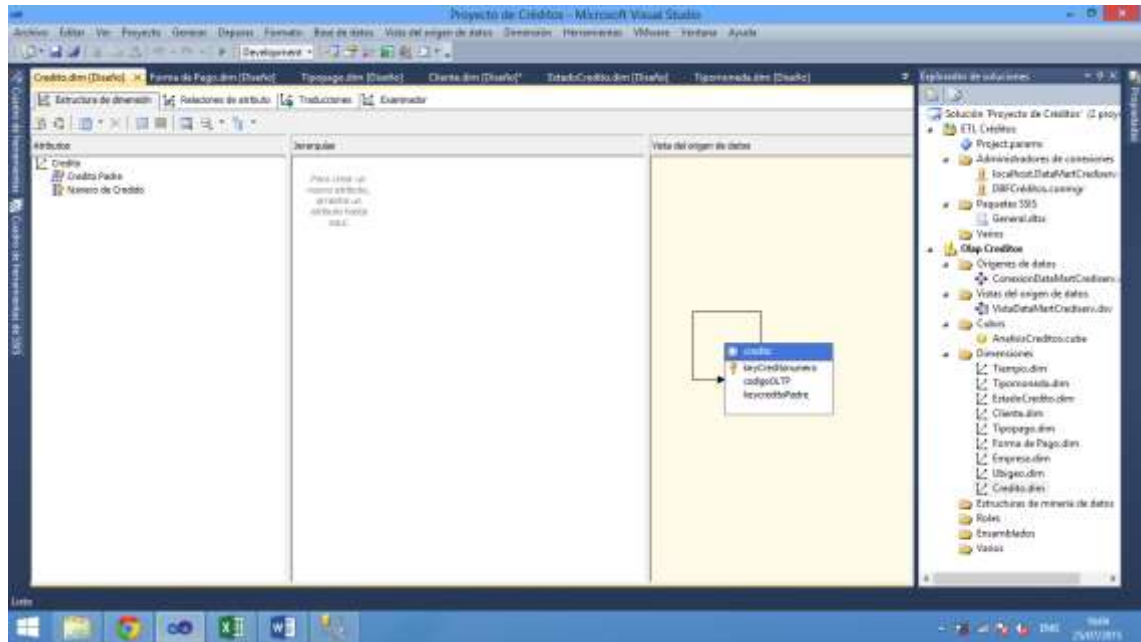


Ilustración 21 - Dimensión crédito

## DISEÑO DEL CUBO

Una vez definidas cada una de las dimensiones del modelo multi dimensional, se procede a diseñar el cubo a partir de cada una de las dimensiones diseñadas anteriormente.

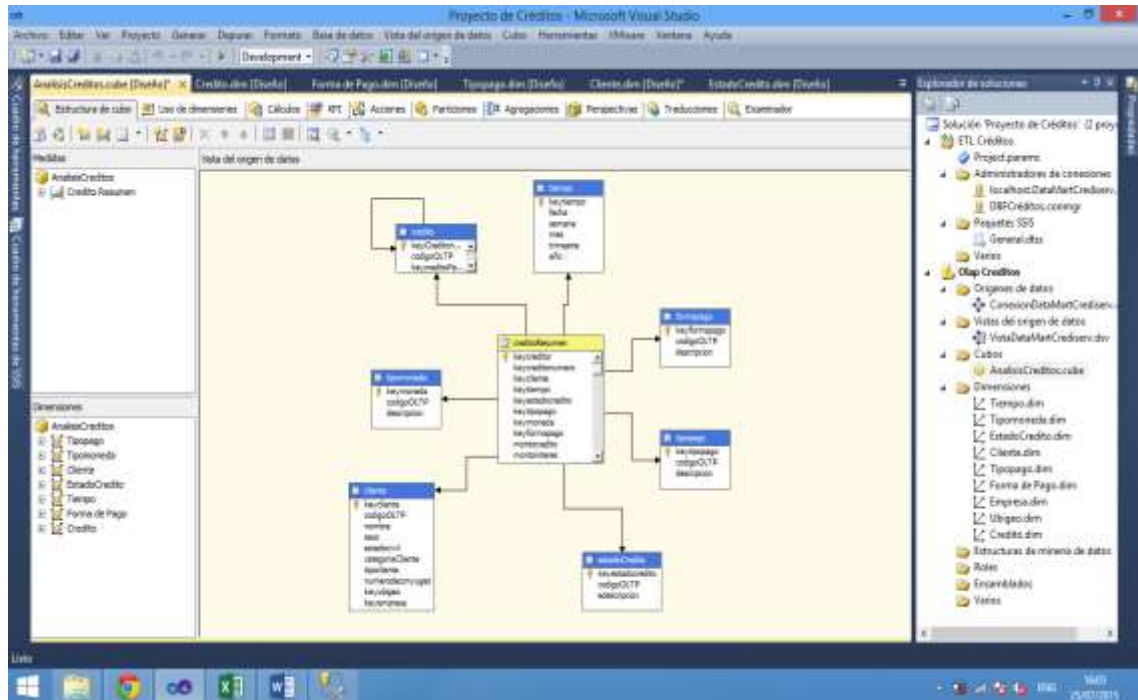


Ilustración 22 - Diseño del cubo de resumen de crédito

## IMPLEMENTACIÓN DEL CUBO

Lo más importante a tener en cuenta para implementar el cubo es considerar que la tabla contiene todas las n-tuplas, con los valores de las dimensiones, o índice del cubo, y los valores de las métricas previamente calculados para el cruce de valores del índice en cuestión

Dimensión	Jerarquía	Operador	Expresión de ROL	Parámetro
Estado de Crédito	Montos			
PASADO	236393.36	236393.36	236393.36	121
PRESENTE	236393.36	236393.36	236393.36	121
PROYECTADO	236393.36	236393.36	236393.36	121
Forma de Pago	Montos			
PASADO	236393.36	236393.36	236393.36	121
PRESENTE	236393.36	236393.36	236393.36	121
PROYECTADO	236393.36	236393.36	236393.36	121
Tiempo	Montos			
PASADO	236393.36	236393.36	236393.36	121
PRESENTE	236393.36	236393.36	236393.36	121
PROYECTADO	236393.36	236393.36	236393.36	121
Ubicacion	Montos			
PASADO	236393.36	236393.36	236393.36	121
PRESENTE	236393.36	236393.36	236393.36	121
PROYECTADO	236393.36	236393.36	236393.36	121

## EXPLOTACIÓN DEL CUBO CON POWER VIEW

Power View es un cliente Web ligero que se inicia en el explorador desde un archivo de origen de datos de informe (.rsds) en una biblioteca de SharePoint. El origen de datos de informe actúa como un puente entre el cliente y el origen de datos back-end.

El origen de datos back-end puede ser un libro PowerPivot en SharePoint, un modelo tabular de un servidor de Analysis Services en modo tabular o un modelo multidimensional en un servidor de Analysis Services que se ejecuta en modo multidimensional.

Los informes de Power View pueden guardarse en una galería o una biblioteca de SharePoint y se comparten con otros miembros de la organización.

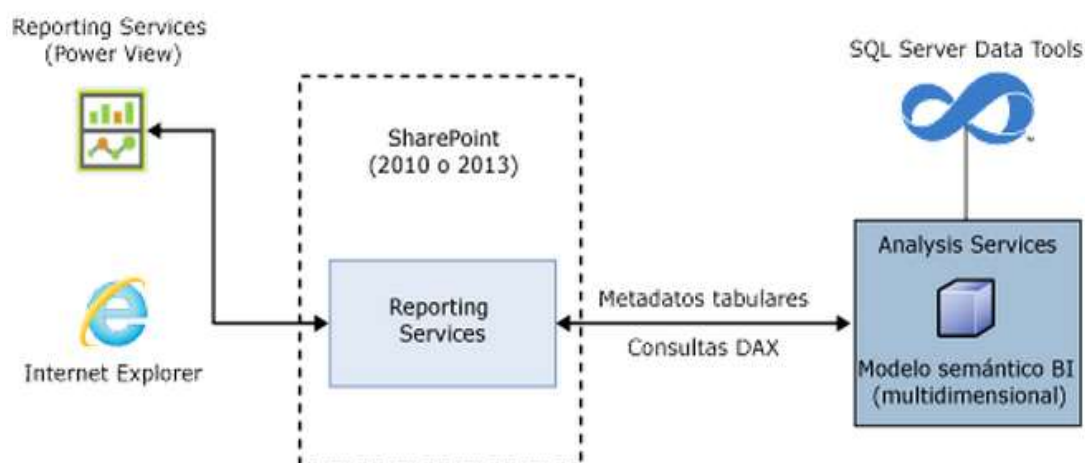


Ilustración 23 - Arquitectura de Power View para modelos multidimensionales

### Conexión a la solución OLAP

Es necesario conectar la herramienta PowerPivot para Excel a la fuente de datos multidimensional.

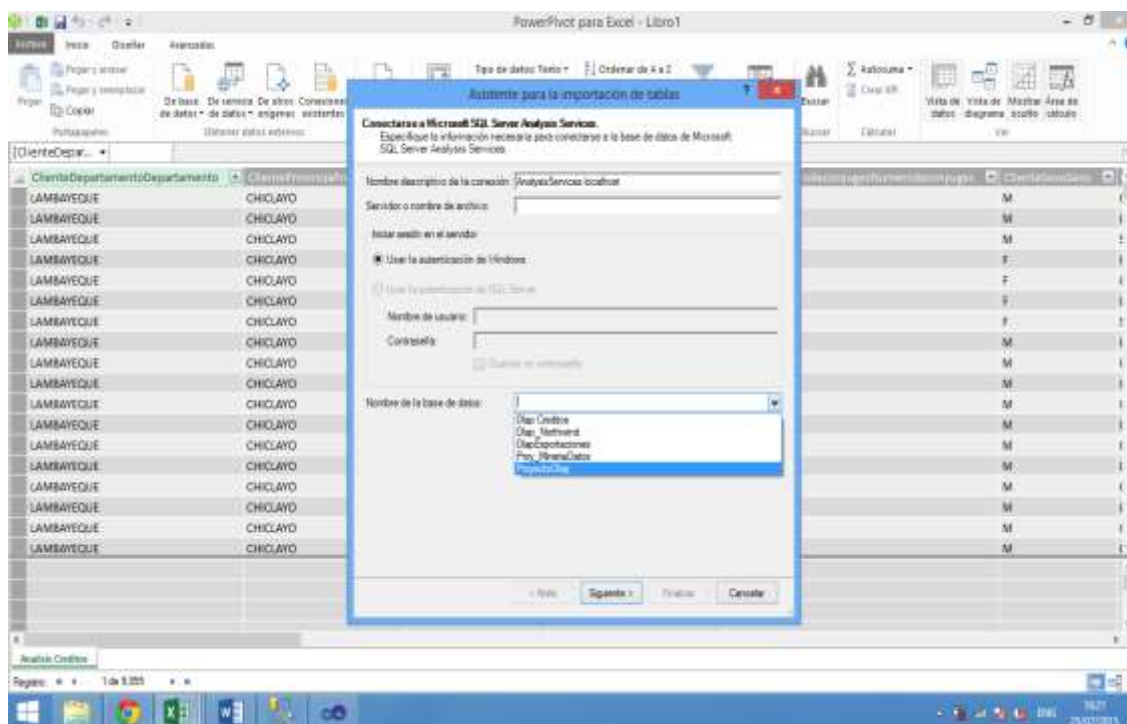


Ilustración 24 - Conectar a la solución OLAP

## Consultando datos con Power Pivot

Los datos de morosidad pueden visualizarse con la la herramienta PowerPivot para Excel

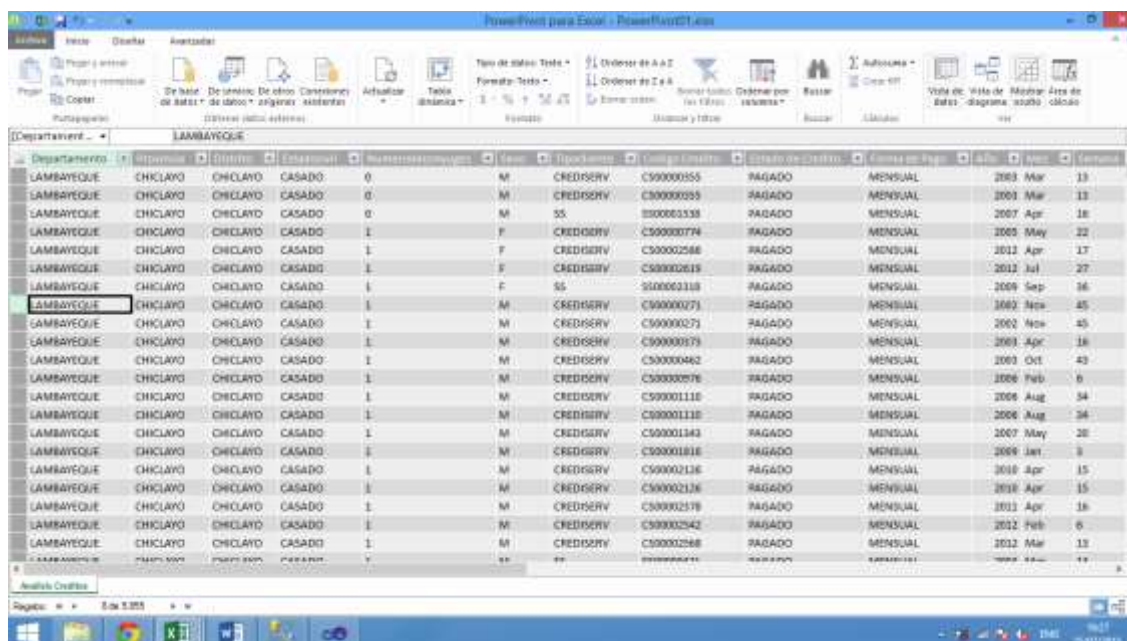


Ilustración 25 - Consultando datos con Power Pivot



## CAPÍTULO VI: COSTOS Y BENEFICIOS

### 2.1. Análisis de costos

Nombre de tarea	Costo
<b>Proyecto de Tesis</b>	<b>S/. 8.480,00</b>
Revisión bibliográfica	S/. 200,00
Formulación del problema	S/. 80,00
Justificación	S/. 40,00
Antecedentes	S/. 160,00
Estado del arte	S/. 320,00
Objetivos	S/. 80,00
Marco teórico	S/. 600,00
Definición de términos	S/. 120,00
Marco metodológico	S/. 320,00
Marco Administrativo	S/. 120,00
Presentación de proyecto	S/. 40,00
<b>Desarrollo de propuesta</b>	<b>S/. 6.440,00</b>
Analizar el proceso de otorgamiento de prestamos	S/. 2.400,00
Seleccionar los Algoritmos de Minería de Datos	S/. 1.200,00
Analizar el rendimiento de modelos de minería	S/. 1.200,00
Desarrollar una aplicación	S/. 800,00
Evaluar el resultado del algoritmo	S/. 800,00

### 2.2. Financiamiento

El proyecto será financiado en su totalidad por el equipo desarrollador del proyecto.

## **CAPÍTULO VII: CONCLUSIONES**

- Se realizó el análisis del proceso de gestión de préstamos en la empresa de créditos Crediser EIRL de la ciudad de Chiclayo, determinando que existe un 10% de morosidad durante el año 2014
- Utilizando herramientas de Microsoft Excel y de Microsoft SQL Server se aplicaron técnicas para limpieza de datos. Estos datos fueron recolectados desde las diversas fuentes de información de la empresa servidor de base de datos, hojas de cálculo en Microsoft Excel y documentos impresos
- Se seleccionó la técnica de Descripción de Clases como modelo de minería de datos a aplicar, por ser el que mejor identifica patrones de comportamiento. La Discriminación de Datos es una comparación entre las características generales de los objetos de una clase respecto a las de otro conjunto contrastante.
- Se desarrolló una aplicación de minería de datos utilizando el software Microsoft SQL Server donde se muestran patrones de comportamiento de clientes morosos en la empresa de créditos Crediser EIRL de la ciudad de Chiclayo como resultado de la técnica de minería de datos seleccionada.

## **CAPÍTULO VIII: RECOMENDACIONES**

- Capacitar al personal analista de créditos de la empresa Crediserv EIRL de la ciudad de Chiclayo para utilizar la aplicación desarrollada como herramienta para el análisis y otorgamiento de créditos
- Desarrollar programas de extracción automática de datos desde las fuentes de datos transaccionales hacia la base de datos multi-dimensional
- Evaluar el rendimiento de otras técnicas de minería de datos para análisis del comportamiento de clientes morosos tales como regresión, resumen, análisis de secuencias



## CAPÍTULO IX: REFERENCIAS BIBLIOGRÁFICAS

- Barrientos, F. (Setiembre de 2013). Aplicación de Minería de Datos para Predecir Fuga de Clientes en la Industria de las Telecomunicaciones. *Revista Ingeniería de Sistemas*.
- Dandretta, G. H. (2002). *Web mining: implementando técnicas de data minning en un servidor web*. Universidad de Belgrano.
- IBM. (2012). *Manual CRISP-DM de IBM SPSS Modeler*. EEUU: IBM.
- Kimball, R. (1998). *The Data Warehouse Lifecycle Toolkit*. Wiley India.
- Kuramoto de Grade, J. (29 de Agosto de 2013). El Perú recién le da importancia a la ciencia, tecnología e innovación. *El Comercio*.
- Lagos Vera, C. (2011). Creación de perfiles de deudores de crédito universitario, para mejoramiento de campañas de cobranza, usando minería de datos. Santiago, Santiago, Chile.
- Lopez Lopez, G., & Velez Rojas, E. (Octubre de 2009). Implementación de un modelo de minería de datos para mejorar la toma de decisiones comerciales en la empresa Star Perú S.A.C. Chimbote, Ancash, Peru.
- Salinas Flores, J. (2005). Reconocimiento de patrones de morosidad para un producto crediticio usando la técnica de árbol de clasificación CART. Lima, Lima, Perú.
- Tumero, I. (2011). *Mineria de Datos, el arte de sacar conocimiento de base de datos*. Puerto Ordaz.
- Valcárcel Asencios, V. (2004). Data Mining y el descubrimiento del conocimiento. *Industrial Data*, 83-86.
- Vidaurre Siadén, Y. (Abril de 2012). Aplicación de redes neuronales artificiales para el pronóstico de la demanda de agua potable en la empresa EPSEL S.A de la ciudad de Lambayeque. Chiclayo, Chiclayo, Lambayeque.
- Usama Fayyad, From Data Mining to Knowledge Discovery in Databases.
- Gregory Piatetsky-Shapiro, From Data Mining to Knowledge
- Padhraic Smyth, American Association for Artificial Intelligence, 1996 Discovery in Databases and