UNIVERSIDAD NACIONAL TECNOLÓGICA DE LIMA SUR

FACULTAD DE INGENIERÍA Y GESTIÓN

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

"IMPLEMENTACIÓN DE UN MODELO PREDICTIVO BASADO EN DATAMINING PARA LA MEJORA DE LA GESTIÓN DE VENTAS EN LA DISTRIBUIDORA JIMENEZ E IRIARTE S.A"

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de

INGENIERO DE SISTEMAS

PRESENTADO POR EL BACHILLER

GUTIERREZ ALVAREZ, RENZO ANDY

VILLA EL SALVADOR

2019

DEDICATORIA

A mis padres

Por su esfuerzo, empuje, preocupación y comprensión en todo momento para seguir triunfando.

AGRADECIMIENTOS

- A mis padres por el apoyo constante para los estudios y por permitirme alcanzar sus sueños.

- A mi alma mater la Universidad Nacional Tecnológica de Lima Sur, por abrirme las puertas, y a los docentes quienes con sus conocimientos y experiencias me ayudaron a forjarme profesional y personalmente.

- A mi asesor el Dr. Frank Edmundo Escobedo Bailón que, con su conocimiento, sugerencia y empuje me apoyo a concluir el presente trabajo.

- A mis amigos Gerardo, Jehiner, Roxana, Jasón, Mirella, Alex, Jefferson, Jimmy, Max Apaza, Max Cahuana y Jhon; con los que pase gratas experiencias dentro y fuera de las aulas y por recibir su apoyo desinteresado cuando más lo necesitaba.

ÍNDICE

LISTADO DE FIGURAS

LISTADO DE TABLAS

INTRODUCCIÓN

“La evolución de las tecnologías de la información junto con la recopilación de datos y análisis, hacen posible a las empresas crear ofertas altamente personalizadas que dirigen a los consumidores a los productos o servicios correctos en el momento justo y por el precio correcto. (Davenport, Dalle Mule, & Lucker, 2012).

Según Blogadmarketing (2017) en el mercado peruano de consumo masivo lo definen los siguientes canales: moderno, digital, informal y tradicional, este último mueve el 70% de las ventas y es donde se encuentra la distribuidora Jiménez e Iriarte S.A., la cual cuenta con más de 15 años de experiencia en el rubro y con una cobertura de 5 sucursales en todo el Perú, generando diariamente en promedio 4200 pedidos que representa el 75% de cobertura del día, no obstante, la gerencia de ventas se ve en apuros debido a que no se cumple con los objetivos mensuales. Durante el 2016 se alcanzó un cumplimiento promedio del 90.56% siendo la entrada más baja el mes de setiembre con 75.34% mientras que en el 2017 y 2018 hubo un cumplimiento de 86.87%y 90.46% respectivamente; sin embargo en el análisis del ejercicio 2018 no se consideró que los objetivos disminuyeron en un 7% respecto al año 2017,por lo que realmente se puede considerar que no hubo un crecimiento respecto al ejercicio anterior, a esto se suma el alto volumen de información dificultando la generación de reportes, en consecuencia la mayoría de las decisiones se toman sobre la experiencia y/o resultados anteriores generando sobrecostes.

Teniendo en cuenta que la Gerencia Central quiere ser más competitiva dentro del rubro y según Thomas H. Davenport (2011) “Una empresa compite mediante análisis cuando hay consideración en el personal, las estrategias están basada en el análisis y se generan modelos predictivos.”, y teniendo como problemática lo descrito líneas arriba. La propuesta de la solución tecnológica planteada, implementación de un modelo predictivo basado en datamining para la gestión de ventas, pretende conocer el comportamiento del cliente, clasificándolo para posteriormente ofrecerle sugeridos y beneficios exclusivos (descuentos y

promociones) conllevado a un incremento en el volumen del pedido del cliente, reflejándose en el cumplimiento de objetivos mensuales.

La estructura que se ha seguido en el presente trabajo está conformada por tres capítulos.

- El primer capítulo hace mención al planteamiento y justificación del problema, delimitación del proyecto, formulación del problema y definición de objetivos.
- El segundo capítulo hace mención al marco teórico limitado por los antecedentes, las bases teóricas y definición de términos básicos (glosario).
- El tercer capítulo hace mención al desarrollo del trabajo que abarca el desarrollo del modelo propuesto y la presentación de los resultados.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

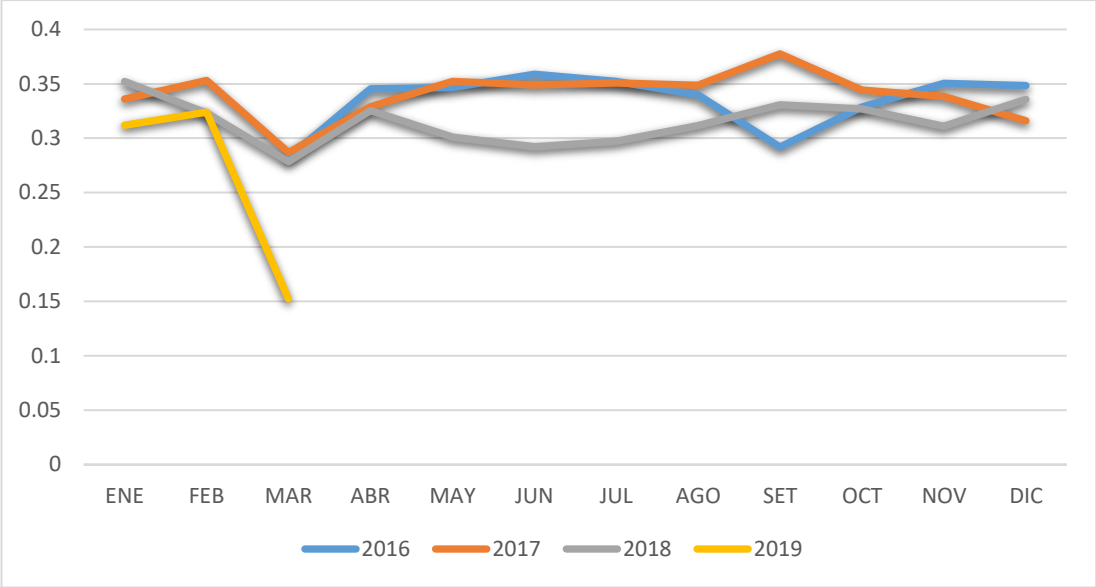
1.1. DESCRIPCIÓN DE LA REALIDAD PROBLEMÁTICA

La evolución de las tecnologías de Información de la mano con la globalización ha llevado a que las organizaciones sean entidades competitivas y productivas basadas en indicadores y/o reportes para la toma de decisiones. Sin embargo, algunas organizaciones no suelen darle relevancia a la información que disponen, sin saber que su verdadero valor es develar conocimiento. (Gómez, 2012)

La filosofía de una empresa distribuidora y/o comercializadora es la compra y venta; siendo la diferencia del costo del producto y el precio de venta al público su principal margen de ganancia (Rodriguez & Mendoza, 2011); sin embargo, el margen se concibe más beneficioso cuando está asociado al volumen de compra. Actualmente la Distribuidora Jiménez e Iriarte cuenta con el ERP Comercial Uniflex (FlexBusiness ver. 4.2.3.0) el cual cuenta con distintos módulos enfocados a cada área como: contabilidad, planeamiento, costos, distribución, créditos y cobranzas; y ventas. Según la figura 1 las ventas de la fuerza de ventas exclusivas en la sucursal de lima durante el ejercicio 2018 han crecido respecto al ejercicio anterior pero no en la proporción como se cerró, es por ello que la Gerencia Central dentro de la reunión anual con la Gerencia de Ventas indago los motivos del poco crecimiento, y dentro de las observaciones se expuso:

- La limitación de los reportes y degradación de los mismos (tiempos de respuesta 1 a 2 min.), generando incomodidad, insatisfacción debido a que no permite realizar una correcta gestión de ventas.
- No se cuenta con reportes para realizar un análisis histórico para poder clasificar al cliente y pueda acceder a nuevos beneficios.
- No se ha llegado cumplir los objetivos gerenciales y/o acuerdos comerciales con los proveedores, sin generar sobrecostos, porque no se cuenta con información para realizar gestión.

Figura 1
Comparativa de ventas (fuerza de venta exclusiva - LIMA) 2016 - 2107 – 2018 - 2019



Fuente: Propia

Como solución se generó el plan mensual de mantenimiento de base de datos (reconstrucción de índices y estadísticas) y se procedió a generar información a través de archivos Excel (conexiones ODBC) para la sede central, con lo cual hubo un mejoramiento en los tiempos de espera (10 a 90 segundos), sin embargo, siguen siendo limitados para tipo de análisis de que se requiere, por lo que es complejo formular estrategias de ventas para impulsar su desempeño sin generar sobrecostes.

1.2. JUSTIFICACIÓN DEL PROBLEMA

La importancia del presente trabajo reside en la implementación de minería de datos a través de modelos predictivos en el rubro de las ventas.

La implementación del modelo predictivo basado en datamining que se ha planteado a la gerencia de Ventas busca mejorar el desempeño y competitividad en la empresa, a través de la búsqueda de patrones repetitivos y relación en los datos, siendo favorable para todos los participantes del proceso de venta: empresa, vendedor y cliente. Los beneficios para los participantes se muestran en la tabla 1.

Tabla 1.
Beneficios del Datamining

EMPRESA	VENDEDOR	CLIENTE
Disminuir sobrecostos para el cumplimiento de acuerdos comerciales.	Incrementar el ticket promedio de ventas.	Acceder a una re categorización, que le permita acceder a descuentos y promociones personalizadas.
Tener mayor alcance de las preferencias de los clientes	Cumplir con los objetivos mensuales.	
Realizar sugeridos de acuerdo al comportamiento del cliente.	Acceder a los incentivos asociados al cumplimiento de objetivos.	
Generar nuevos reportes a través del nuevo repositorio de información.		

Fuente: Propia

Estos beneficios influyen positivamente en la experiencia del cliente impulsando la lealtad del mismo, la cual será proporcional al valor económico que genere para la distribuidora.

Adicionalmente será tomado como punto de partida para que las demás áreas a futuro implementen minería de datos y como previsión de la demanda (estimar la demanda y el volumen necesario de productos).

1.3. DELIMITACIÓN DEL PROYECTO

1.3.1. TEÓRICA

A) Gestión de ventas

Agrupa una serie de profesionales y empleados para concretar la venta de productos teniendo en cuenta la búsqueda del cliente, el conocimiento del producto, la prospección del cliente objetivo, la presentación y cierre de la venta.

B) Inteligencia de negocios

Es una aglomeración de procesos y herramientas tecnológicas para permiten generar conocimiento a partir de los datos para tomar decisiones.

C) Modelos predictivos

Son modelos matemáticos/estadísticos basados en una perspectiva de análisis, usados en la minería de datos, estos pueden ser de clasificación, regresión, asociación y agrupación.

D) Datamining

Es un conjunto de técnicas basado en modelos predictivos que permiten descubrir patrones ocultos en los datos.

1.3.2. TEMPORAL

Inicio: marzo de 2019

Fin: abril de 2019

1.3.3. ESPACIAL

El desarrollo del presente proyecto se realizará en el área de Ventas de la distribuidora Jiménez e Iriarte; ubicada en Jr. San José N° 163 – Chorrillos – Lima – Perú.

1.4. FORMULACIÓN DEL PROBLEMA

1.4.1. PROBLEMA GENERAL

¿En qué medida la implementación de un modelo predictivo basado en datamining influye en la mejora de la gestión de ventas en la Distribuidora Jiménez e Iriarte S.A.?

1.4.2. PROBLEMAS ESPECÍFICOS

- ¿De qué manera el diseño de un modelo predictivo basado en datamining permite clasificar al cliente con la finalidad de ofrecerle promociones y descuentos personalizados?
- ¿De qué manera el diseño de un modelo predictivo basado en datamining permite realizar sugeridos de venta al cliente de acuerdo a su comportamiento?
- ¿De qué manera el diseño de un modelo predictivo basado en datamining permite planificar (recomendar) la pre-venta de las fuerzas de ventas exclusivas de la sucursal de Lima?

1.5. OBJETIVOS

1.5.1. OBJETIVO GENERAL

Implementar un modelo predictivo basado en datamining para la mejora de la gestión de ventas en la Distribuidora Jiménez e Iriarte.

1.5.2. OBJETIVOS ESPECÍFICOS

- Implementar un modelo predictivo basado en datamining permite clasificar al cliente con la finalidad de ofrecerle promociones y descuentos personalizados.
- Implementar un modelo predictivo basado en datamining permite realizar sugeridos de venta al cliente conociendo su comportamiento de compra.
- Implementar un modelo predictivo basado en datamining permite planificar (recomendar) la preventa de las fuerzas de ventas exclusivas de la sucursal de Lima.

CAPÍTULO II: MARCO TEÓRICO

2.1. ANTECEDENTES

Hoy por hoy las tecnologías de información se muestran como una herramienta indispensable dentro de las distintas áreas de la organización, apoyando en la mejora de su desempeño y en la obtención de resultados correctos y más confiables. En tal sentido la minería de datos o datamining no es ajena a dicho contexto y se ve reflejado en las organizaciones que implementaron como solución tecnológica, lo cual les permitió discernir en la toma de decisiones en base a los insights y patrones extraídos de los datos.

2.1.1. ANTECEDENTES NACIONALES

A) Análisis comparativo de técnicas de minería de datos para la predicción de ventas.

Título: “Análisis comparativo de técnicas de minería de datos para la predicción de ventas”. (Roque, 2016)

Autor: Irene Leydi Roque Montalvo.

Trabajo de investigación: Universidad Señor de Sipán, Facultad de Ingeniería, Arquitectura y Urbanismo.

Roque (2016) refiere que la evaluación de los algoritmos y las técnicas de minería de datos son más relevantes que la construcción del modelo mismo, porque se adecuan más a la problemática debido a que no es lo mismo aplicar criterios de pronósticos a las series de ventas como a las series de clima. El caso de estudio se enfocó en analizar las diferentes técnicas de minería de datos con los datos históricos (08-2011 al 06-2014), para determinar la técnica más adecuada para la predicción de ventas de artículos deportivos basándose en 3 criterios de selección (confiabilidad de los pronósticos, tiempo de procesamiento para obtener la estimación y número de puntos mínimos del vector), optándose por la técnica ETS

porque representa una mayor confiabilidad (90.51%), respecto a los 84.42% de HoltWinters y 83,96% de Holt.

El trabajo llegó a las siguientes conclusiones:

- Las series de tiempo y los HoltWinters son las técnicas más adecuadas en el ámbito de las ventas teniendo en consideración los outlayers y los missing.
- La evaluación de modelos predictivos se tiene que realizar obligatoriamente debido a que brinda fiabilidad en los resultados, teniendo en cuenta el tiempo de respuesta y el grado de confianza.

Tabla 2
Comparativo de técnicas de minería de datos

	HoltWinters	**Holt**	**ETS**	**Arima**
Evaluación fundamento teórico				
Modelo parametrizado	SI	SI	SI	SI
Datos estacionales	SI	SI	SI	SI
Método estadístico	SI	SI	SI	SI
Capacidad iterativa (Aprendizaje)	NO	NO	NO	NO
Cantidad de datos de la serie	24	25	25	80
Evaluación fundamento computacional				
Procesamiento CPU	Mínimo	Mínimo	Mínimo	Mínimo
Consumo RAM	Mínimo	Mínimo	Mínimo	Mínimo
Tiempo computacional	Mínimo	Mínimo	Mínimo	Mínimo
Evaluación fundamento objetivo del modelo				
Confiabilidad de precisión pronostico	Después de pruebas	Después de pruebas	Después de pruebas	Después de pruebas

Fuente: (Roque, 2016)

B) Aplicación de minería de datos para determinar patrones de consumo en clientes de una distribuidora de suplementos nutricionales.

Título: “Aplicación de minería de datos para determinar patrones de consumo en clientes de una distribuidora de suplementos nutricionales”. (Grández, 2017)

Autor: Miguel Angel Grández Márquez.

Trabajo de investigación: Universidad San Ignacio de Loyola, Facultad de Ingeniería.

Grandéz (2017) menciona que el desempeño de una organización en la actualidad, está determinada directamente por el análisis que se realiza a los registros generados, los cuales representan un costo de oportunidad. El caso de estudio se enfocó determinar los patrones de consumo del cliente aplicando técnicas de minería de datos como reglas de asociación, clustering y redes neuronales e identificar cual se adecua a los datos almacenados (semestre 2016 y campañas nutricionales), dentro de los patrones se concluyó que:

- Los clientes entre 25 y 32 años, y que realizan la actividad física Pilates compren el producto Harbinger fitness ® correa big grip pro lifting straps con una probabilidad de 100%.
- Las clientas que hacen ejercicios con Elíptica compren el producto Nutrex ® lipo-6 cla con una probabilidad de 100%.
- La probabilidad que los clientes con más de 2 hijos y un peso entre 80,4 - 91,5 compren el producto Syntrax ® nectar medical es de 100%.

El trabajo llegó a las siguientes conclusiones:

- El tamaño de la muestra debe ser relevante, adecuado, representativo y significativo que en su mayoría de casos es de 60% para el entrenamiento y 40% para el test.
- Hay que considerar una muestra de datos significativa debido a que es influyente al grado de confiabilidad de los modelos predictivos.
- Se debe tener en cuenta un planeamiento en la recolección de datos puesto que la información debe ser analizada y refinada.

C) Minería de datos aplicada a la detección de fraude electrónico en entidades bancarias.

Título: “Minería de datos aplicada a la detección de fraude electrónico en entidades bancarias” (Ñaupas, 2016)

Autor: Carol Maribel Ñaupas Caraza.

Trabajo de investigación: Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática.

Ñaupas (2016) menciona que, debido a los avances tecnológicos, las empresas pueden gestionar grandes volúmenes de datos y descubrir conocimiento a partir de ellos, para que sean utilizados estratégicamente en la toma de decisiones además propone la utilización de la herramienta Pentaho para la solución BI debido a que es una herramienta open source centrada en procesos permitiendo crear soluciones complejas a problemas de negocio. El caso de estudio se centró en identificar patrones para catalogar como fraudulentas o integras las transacciones realizadas en los canales de banca por internet o banca móvil, para ello se usó una

población de transacciones con un rango de 3 meses y se aplicó la técnica de árboles de clasificación (algoritmo de C4.5), usando la metodología KDD (descubrimiento de Conocimiento en base de datos) la cual permitió hacer procesos iterativos con la finalidad de volver a un proceso anterior y ajustar los parámetros y/o supuestos para obtener un modelo óptimo. Para evaluar el desempeño del modelo se aplicó la matriz de confusión arrojando los siguientes resultados:

Tabla 3
Matriz de confusión - detención de fraude electrónico

accuracy: 99.02% ± 034% (mikro: 99.02%)			
	true N	**true S**	**class precisión**
pred. N	4900	44	99.11%
pred. S	5	31	86.11%
class recall	99.90%	41.33%	

Fuente: (Ñaupas, 2016)

El trabajo llegó a las siguientes conclusiones:

- Las técnicas predictivas resultan ser eficientes para descubrir conocimiento y permiten inferir como un atributo puede incidir en otros.
- La volatilidad de la variable analizada es un factor influyente en las reglas y atributos definidos para un modelo resultado de una investigación y es necesario que se vaya actualizando a la par con los expertos del negocio, con finalidad de no sucumbir dicho modelo.
- A pesar de que la técnica de árbol de decisión sea una técnica sencilla provee un alto grado exactitud para el caso de estudio porqué generó un modelo con un nivel confiabilidad de 99.02%
- Las consideraciones a tener post implementación de un modelo predictivo, debido a la volatilidad de la variable objetivo por parte del negocio.

2.1.2. ANTECEDENTES INTERNACIONALES

A) Análisis para la predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies

Título: “Análisis para la predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies” (García & Acevedo, 2010)

Autor(es): José Antonio García Bermúdez

Ángela María Acevedo Ramirez

Trabajo de investigación: Universidad Tecnológica de Pereira, Facultad de Ingenierías (Eléctrica, Electrónica, Física y Ciencias de la Computación).

García & Acevedo (2010) mencionan se ha vuelto una necesidad de las organizaciones administrar sus actividades y poseer un historial de las mismas, sin embargo, en la mayoría veces el historial crece abruptamente y realizar las mismas consultas no es eficiente. El caso de estudio se centró en encontrar afinidad entre 2 o más productos dentro de un conjunto de datos seleccionados de una gran superficie de ventas.

El trabajo llegó a las siguientes conclusiones:

- Una ventaja competitiva para las empresas es contar con un modelo que permita pronosticar el comportamiento de los clientes.
- La minería de datos se basa en la búsqueda de modelos que sé que se asemejen y puedan reflejar los movimientos que se dan con los datos.
- El tipo de dato está estrechamente ligado a los algoritmos a escoger para la minería de datos.

B) Aplicación de técnicas de minería de datos para mejorar el proceso de control de gestión en Entel.

Título: “Aplicación de técnicas de minería de datos para mejorar el proceso de control de gestión en Entel”. (Martínez, 2012)

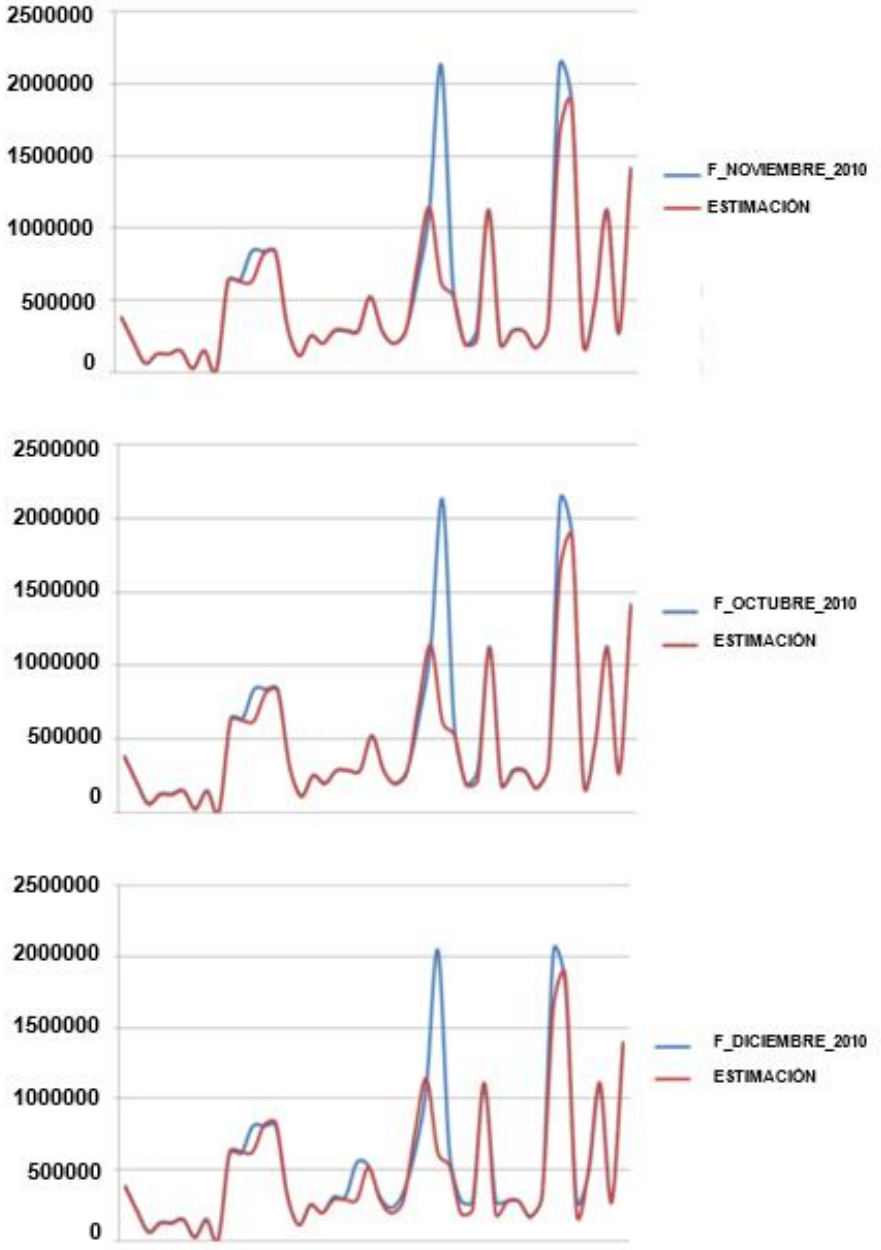
Autor: Clemente Antonio Martínez Álvarez.

Trabajo de investigación: Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas.

Martínez (2012) menciona que la aplicación de modelos de minería de datos, sirve como sustento fiable para los expertos del negocio, sin embargo, estos no reemplazan el conocimiento que ellos tienen del negocio. El caso de estudio se centró en los problemas asociados a la fuga de ingresos de una empresa de telecomunicaciones debido al no pago de los servicios privados (telefonía, internet y comunicaciones) e identificar las características del cliente no pagador, dentro de los resultados se obtuvo:

- La Implementación de un datamart para el área de aseguramientos de ingresos como repositorio único y consolidado de datos.
- El tiempo de procesamiento y almacenamiento de los datos utilizados para el cálculo de indicadores de servicios privados disminuyo en un 78%
- Poder estimar ingresos potenciales mensuales por $210 MM debido a los servicios no facturados con una confiabilidad del 80%.

Figura 2
Comparación de los precios para meses futuros



Fuente: (Martínez, 2012)

El trabajo llegó a las siguientes conclusiones:

- Es mejor tener un datamart como repositorio de información pues la información es más confiable con el fin de evitar corregir inconsistencias en la data durante la creación del modelo predictivo.

- Un datamart influye positivamente en el tiempo de procesamiento y almacenamiento de los datos utilizados para el cálculo de indicadores.
- Tras la aplicación de minería de datos se encontró relación entre las variables, lo cual permite la caracterización y proponer nuevas estrategias.

C) Minería de datos una herramienta para la toma de decisiones.

Título: “Minería de datos una herramienta para la toma de decisiones” (Calderón, 2006)

Autor: Neftalí de Jesús Calderón Méndez.

Trabajo de investigación: Universidad de San Carlos de Guatemala, Facultad de Ingeniería.

El trabajo menciona que se está perdiendo conocimiento a partir de los datos porque existe una relación inversa entre las técnicas tradicionales de análisis de información y, la velocidad de almacenamiento y análisis de los datos, así mismo el caso de estudio determinó que la minería de datos permite crear escenarios de los cuales se pueden tomar decisiones a nivel gerencial a través de técnicas de análisis de datos.

El trabajo llegó a las siguientes conclusiones:

- La minería de datos permite identificar patrones, comportamientos y reglas en los datos aplicando herramientas y técnicas.
- Los gerentes y empresarios suelen ser más eficientes en la toma de decisiones y manejo del conocimiento cuando manejan diferentes escenarios los cuales pueden

argumentarse por medio de modelos matemáticos y estadísticos.

- Depurar los datos (missing y outliers) con la finalidad obtener modelos predictivos más confiables.

D) Minería de datos para la segmentación de clientes en la empresa tecnológica Master PC.

Título: “Minería de datos para la segmentación de clientes en la empresa tecnológica Master PC”. (Chamba, 2015)

Autor: Sairy Fernanda Chamba Jiménez.

Trabajo de investigación: Universidad Nacional de Loja, Carrera de Ingeniería de Sistemas.

Chamba (2015) menciona que el descubrimiento de patrones en el comportamiento del cliente es una de las tantas aplicaciones de minería de datos en el campo de las ventas, la cual permite elaborar estrategias de marketing personalizadas a grupos de clientes. El caso de estudio se centró en aplicar técnica de minería para determinar el nivel de lealtad del cliente en base a su historial de compra aplicando la metodología CRISP-DM, trabajando con una fuente de 5 años se obtuvo los siguientes resultados:

- La categorización a través de 4 grupos de lealtad permite replantear las estrategias de retención de clientes.
- La similitud en el comportamiento de cliente permite recomendar productos con un alto nivel de fiabilidad.

El trabajo llegó a las siguientes conclusiones:

- Resaltar cuales son los modelos predictivos más utilizados (reglas de asociación y agrupamiento) entorno a las ventas.
- La aplicación de reglas de clasificación permite a la organización generar estrategias personalizadas.
- La aplicación de reglas de asociación permite elaborar estrategias de promoción y recomendación, permitiendo tener mayor afinidad con el cliente.

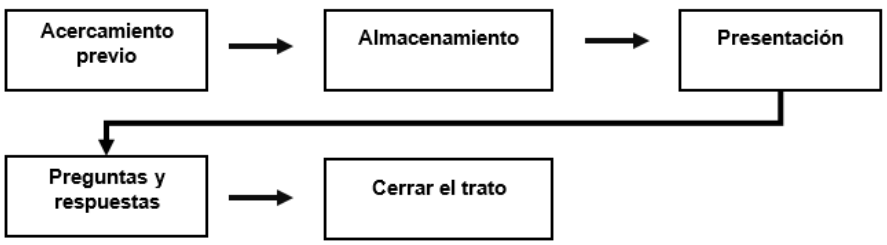
2.2. BASES TEÓRICAS

2.2.1. GESTIÓN DE VENTAS

Guillen & Sánchez (2017) menciona que la gestión de ventas es un conjunto de actividades volátiles donde interactúan diferentes elementos que contribuyen a que la venta considerando principalmente satisfacer la necesidad del cliente.

Se define el proceso de ventas, la cual esta estructura en 5 pasos, como un pilar en la gestión de ventas. Según figura 3, El vendedor atrae la atención del cliente con su acercamiento previo para posteriormente realizar un acercamiento real con la finalidad de ganar interés e estimular el deseo compra a través de una presentación, consiguiendo la confianza del cliente al contestar preguntas y resolverlas objeciones, y logrando su objetivo mediante la materialización de la venta.

Figura 3
Proceso de ventas



Fuente: (Guillén & Sánchez, 2017)

A) El papel de los sistemas de información en las ventas

Según Vega (2005) el rol de los sistemas de información en la gestión de ventas es significativo, porque permiten perfeccionar la efectividad y eficiencia del vendedor y trazar el progreso de mismo a los gerentes de venta. Así mismo los sistemas de información deben adaptarse a las particularidades de la empresa con la finalidad que sean asequibles durante la toma de decisiones

2.2.2. SISTEMAS DE INFORMACIÓN

Según Ferrer (2015) los sistemas de información están formados por un conjunto de elementos que interactúan entre ellos para conseguir un objetivo común: satisfacer las necesidades y demandas de información de la empresa. Dentro de los objetivos de los sistemas de información se tiene:

- Automatizar los procesos operativos de la organización.
- Proporcionar información que sirva de apoyo en la toma de decisiones.
- Conseguir ventajas competitivas.

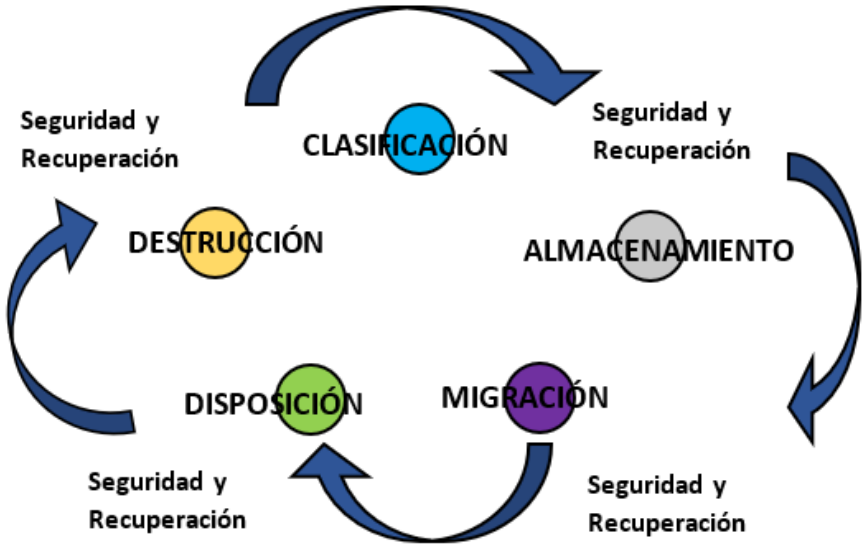
Figura 4
Niveles de toma de decisiones

Fuente: (Ferrer, 2015)

2.2.3. GESTIÓN DE DATOS

Según Ferrándiz (2018) se refiere al desarrollo y ejecución de arquitecturas, políticas, prácticas y procedimientos para administrar la información durante todo su ciclo de vida en la organización.

Figura 5
Ciclo de vida de la información



Fuente: (Ferrándiz, 2018)

La gestión de datos tiene como objetivo: entender las capacidades desde la perspectiva de personas, procesos y tecnología; y entender como cada capacidad encaja dentro de un marco global. Las capacidades de una gestión de datos son:

A) Calidad de datos

Se refiere al enfoque, políticas y procesos por el cual una organización maneja la exactitud, validez, oportunidad, completitud, unicidad y consistencia de sus datos.

B) Gobierno de datos

Proceso definido que sigue una organización para garantizar que durante de todo el ciclo de vida de los datos, estos sean de calidad, disponibilidad, usabilidad, integridad y seguridad.

C) Analytics

Gestionar el procesamiento analítico de datos y permitir el acceso a datos de soporte de decisiones para metadatos de informes y análisis: recopilación, categorización, mantenimiento, integración, control, gestión y entrega de metadatos.

D) Gestión de metadatos

Abarca la recopilación, categorización, mantenimiento, integración, control, administración y entrega de metadatos.

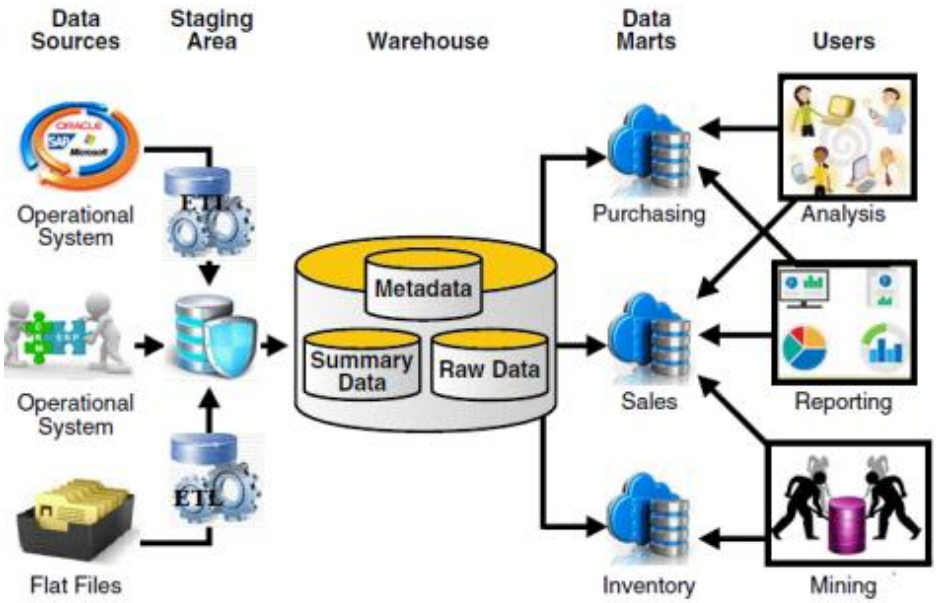
E) Integración de datos

Involucra la adquisición, extracción, transformación, movimiento, entrega, replicación, federación, virtualización y soporte operacional de los datos

F) Arquitectura de datos

Se refiere a las modelos, políticas, reglas o estándares que gobiernan qué datos son recolectados y cómo son almacenados, ordenados y puestos en uso una organización. La figura 6 muestra las capas que conforman la arquitectura de datos.

Figura 6
Arquitectura de Datos



Fuente: (Ferrándiz, 2018)

G) Privacidad de datos

- Ley de protección de datos personales.

  El pilar de la ley de protección de datos personales (LPDP – ley 29733) es garantizar el derecho fundamental a los datos personales y tratamiento del mismo sin afectar a los propietarios.

  Los datos personales se definen como datos característicos que permiten identificar a una persona (nombre, apellidos, fecha de nacimiento, dirección de correo, etc.) y se basan en los siguientes principios

  - Legalidad: Conforme a lo establecido en la LPDP.
  - Consentimiento: del titular para realizar el tratamiento de datos personales.
  - Finalidad: Promueve que los datos sean tratados solo para el propósito por el cual han sido recopilados.

- Proporcionalidad: Permite que la información que sea imprescindible y suficiente de acuerdo a la finalidad.
- Calidad: Refleja que los datos sean veraces, exactos y adecuados.
- Seguridad: La confidencialidad y seguridad de los datos personales está dada por el titular del banco de datos personales y el responsable de tratamiento.

- Seguridad de datos

  Tiene con finalidad la confidencialidad, la disponibilidad e integridad de datos promoviendo medidas preventivas y reactivas en la organización y sistemas tecnológicos.

H) Gestión de datos maestros y datos de referencia

- Gestión de datos maestros

  Método integral que se usa para definir y administrar de forma consistente los datos críticos de una organización para proporcionar un único punto de referencia. Estos datos maestros pueden incluir datos de referencia.

- Gestión de datos de referencia

  Subconjunto de los datos maestros que referencia a datos que definen conjunto permitido de valores a ser usados por otros campos de datos

2.2.4. INTELIGENCIA DE NEGOCIOS

Rodriguez & Mendoza (2011) mencionan que la inteligencia de negocios es un enfoque para la gestión empresarial que permite a la organización definir qué información es útil y relevante para la toma de decisiones.

Según Guillen (2012) la inteligencia de negocios permite tomar mejores decisiones a los usuarios de una organización a través de la administración de información aplicando un conjunto de metodologías y técnicas a los datos. Algunas de las tecnologías que forman parte de la inteligencia de negocios son:

- Integración de Datos (ETL),
- Datawarehouse / Datamart.
- Análisis OLAP (On-Line Analytical Processing).
- Reporting.
- Cuadro de mando Integral
- Dashboard
- Data mining

A) Beneficios

Según Alcalde (2018) los beneficios son:

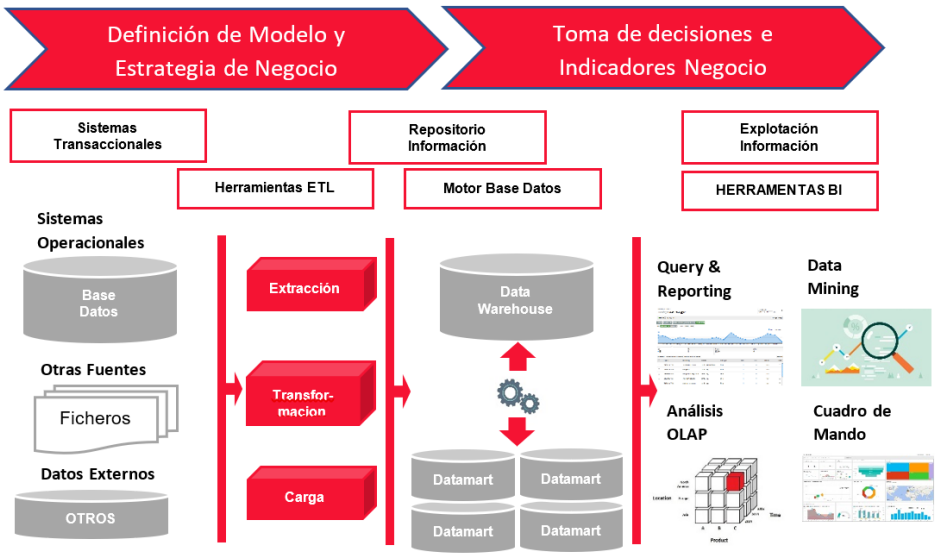
- Tomar la mejor decisión e implementar las acciones correctas.
- Complementar la intuición y la experiencia con mecanismos basados en hechos.
- Acceder y compartir fácilmente la información.
- Descubrir lo que desconocía de la empresa.
- Cuestionar en el momento correcto.
- Encontrar eficiencias y reducción de costos

B) Componentes

Según Alcalde (2018) la arquitectura de una solución BI se define con los siguientes componentes:

- Fuentes de Información: los sistemas transaccionales suministran información al Datamart o Datawarehouse.
- Herramientas ETL: involucra las actividades de extracción, transformación y carga la cual se encargar de estandarizar, filtrar y redefinir los datos.
- Motor base de datos: provee la capacidad de cálculo y consulta de grandes volúmenes de información.
- Herramientas BI: permite el análisis y navegación a través de los mismos.

Figura 7
Arquitectura de solución BI



Fuente: (Alcalde Aliaga, 2018)

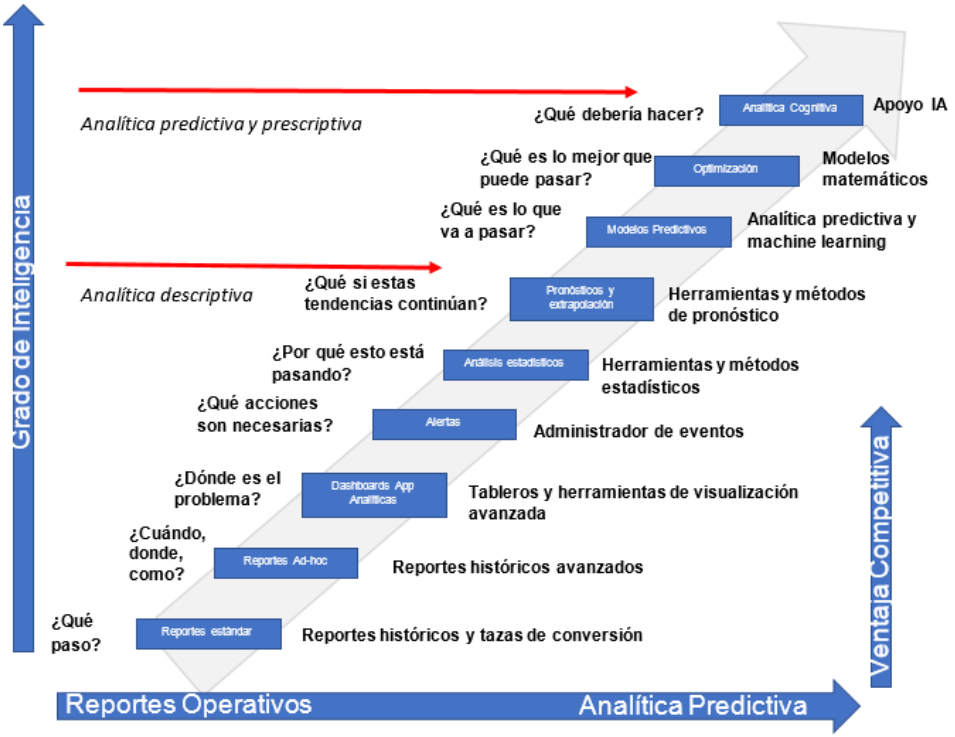
C) Nivel de madurez de BI

Según Ibarra (2014) el nivel de madurez de BI mide el alcance de implementación de BI en la empresa y ayuda a entender que no solo se trata de solamente de una implementación tecnológica, sino que depende explícitamente de cuatro pilares:

- Infraestructura: permite almacenar, distribuir y utilizar la información.
- Los procesos: determinan la generación, validación de la información y su forma de uso.
- El capital humano: relacionado con las capacidades y habilidades de las personas que hacen uso de la información.
- La cultura organizacional: definida por la empresa para fomentar el uso de la información, impulsar la optimización de procesos y capacitar al personal.

Según Alcalde (2018) el nivel de madurez de BI en una empresa se ve reflejado al nivel de acciones, los cuales permiten contestar a una serie de preguntas.

Figura 8
Nivel de madurez de BI



Fuente: (Alcalde Aliaga, 2018)

D) Metodologías de Inteligencia de Negocios

Según Esparza. Et al. (2014), existen 3 tipos de metodologías principales para el desarrollo de una solución de inteligencia de negocios, entre los cuales se tienen las siguientes:

- Ralph Kimball: Se caracteriza porque parte de un conjunto de datamarts y posteriormente se integra en un datawarehouse centralizado, orientándose más a proyectos cortos y equipos de desarrollo pequeño.

- Bill Immon: Se caracteriza porque parte de un datawarehouse y posteriormente se segmenta en distintos datamarts de acuerdo a las áreas de negocio.

- Hefesto: Se caracteriza por tener una arquitectura hibrida permitiendo adaptarse a cualquier necesidad empresarial

E) Datamart vs Datawarehouse

Según Hernandez (2008) un datawarehouse es solución tecnológica para la toma de decisiones a nivel corporativo variante en el tiempo y no volátil, mientras que un datamart es una alternativa de solución similar al datawarehouse, enfocada a un área de negocio especifica.

Tabla 4
Cuadro comparativo entre almacenes de datos

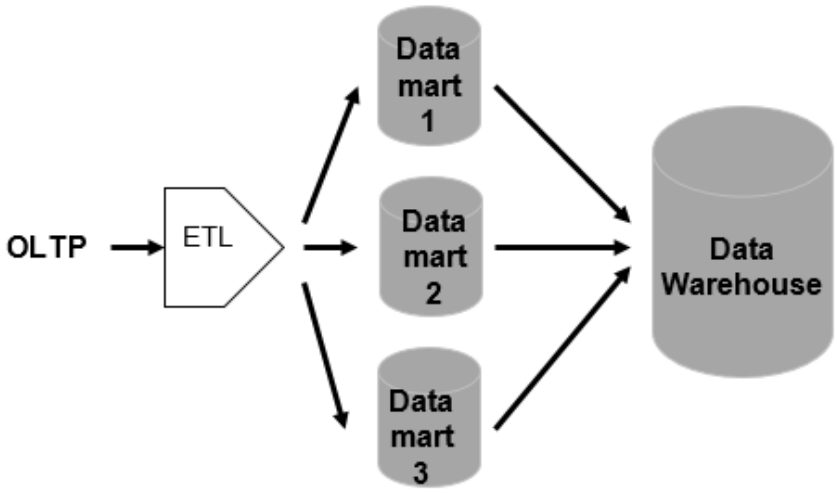
	Datawarehouse	**Datamart**
Alcance	General Estratégico Planeado	Especifico Táctico Orgánico
Datos	Históricos/detallados/agregados normalizados	Poca historia/detallados/agregados Desnormalizados
Fuentes	Muchas	Pocas
Pros y Contras	Flexible Orientada a los datos Única estructura compleja	Restrictiva Orientada a proyectos Muchas estructuras simples

Fuente: (Alcalde Aliaga, 2018)

F) Metodología Ralph Kimball

Según Esparza. Et al. (2014), la metodología Ralph Kimball se referencia como Bottom-up, debido a que el datawarehouse no es más que la unión de los diferentes datamarts, que están estructurados de una forma común a través de la estructura de bus.

Figura 9
Arquitectura botton-up

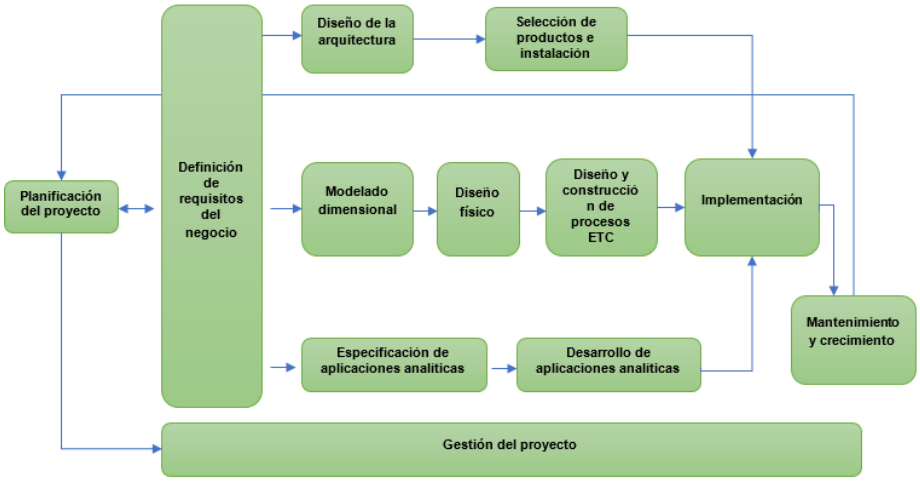


Fuente: (Esparza, Alvarez, Duque, & Quiroz, 2014)

Según Azuaje (2014) el ciclo de vida de la metodología Ralph Kimball está conformado por 4 principios básicos:

- Identificar los requerimientos del negocio.
- Edificar una infraestructura de información viable.
- Proporcionar entregas en incrementales.
- Entregar valor a los usuarios de negocio.

Figura 10
Ciclo de vida metodología Ralph Kimball



Fuente: (Alcalde Aliaga, 2018)

G) Modelado Multidimensional

Según Guillen (2012) el modelo multidimensional permite optimizar el desempeño de acceso a la información desde un punto de vista más cercano al usuario final, modelando las particularidades de los procesos que ocurren en la organización y agrupándolos en mediciones y entorno.

Elementos del modelado multidimensional:

- Tabla de hechos: Simboliza los procesos dentro de la organización, contiene los indicadores de negocio.
- Dimensión: Representan los factores por que se analiza una determinada área del negocio.

  Tipos de dimensiones:

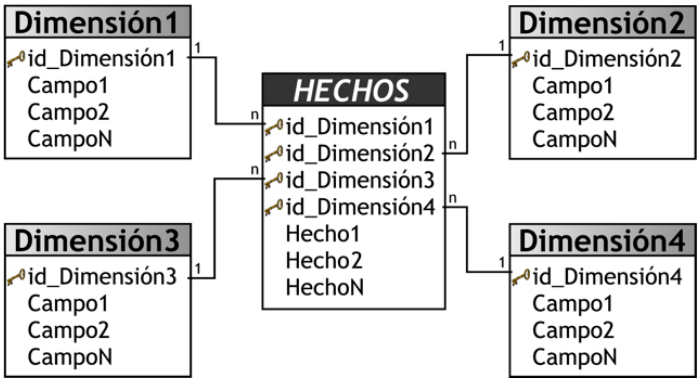
  - Tipo1: No hay historia.
  - Tipo2: Los nuevos registros son escritos en nuevas filas.
  - Tipo3: Los nuevos valores son escritos en las columnas de una fila existente.

H) Tipo de modelamiento

- Esquema estrella

  El modelamiento estrella está compuesto por una tabla de hechos y diversas tablas dimensionales, donde cada dimensión representa una característica del negocio a analizar. (Hernandez, 2008)

Figura 11
Modelo dimensional estrella

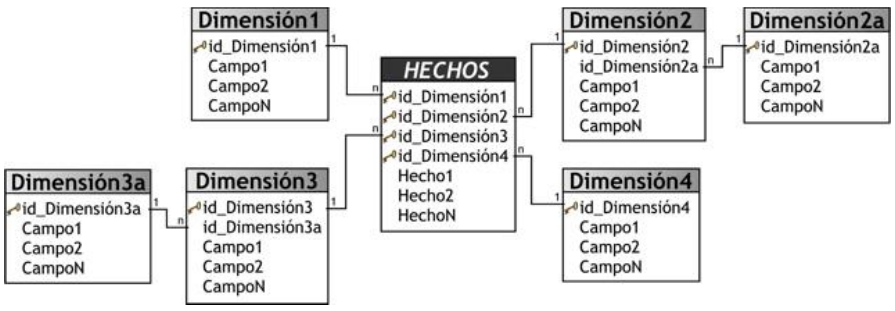


Fuente: (Bernabeu, 2009)

- Esquema Copo de nieve

  Debido a que en el modelo estrella no se considera la normalización, existe el concepto de modelamiento copo de nieve, que es un modelo estrella con las tablas de dimensiones normalizadas, es decir, una tabla dimensional hace referencia a otra tabla dimensional, así evitando la redundancia de información a nivel de base de datos. Este modelo nace para agilizar el acceso a los datos a través de su estructura. (Hernandez, 2008)

Figura 12
Modelo dimensional copo de nieve

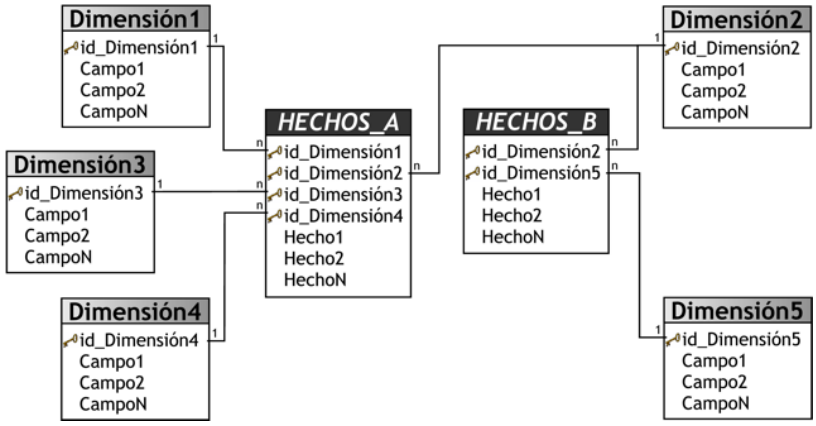


Fuente: (Bernabeu, 2009)

- Esquema constelación

  Como variante al modelo estrella se encuentra en el modelo constelación el cual se caracteriza porque hay como mínimo 2 tablas de hechos las cuales comparten algunas dimensiones. Este tipo de modelamiento nace debido a que las métricas a analizar difieren de las mismas dimensiones. (Hernandez, 2008)

Figura 13
Modelo dimensional constelación



Fuente: (Bernabeu, 2009)

I) Fuente de datos

Según De Rossi (2018) se pueden incluir: bases de datos relacionales, documentos, datos externos, etc., además menciona que se debe auditar fuentes para establecer:

- Datos contenidos en cada fuente
- Volúmenes almacenados y generados
- Tipos de datos y rango de valores
- Problemas de integridad y nulos
- Duplicidad e inconsistencias.
- Uso y actualización

J) Proceso de extracción, transformación y carga

Según Alcalde (2018) el proceso ETL consolida los datos de fuentes heterogéneas en una plataforma estándar en un formato estándar. El cual consta de 3 subprocesos: extracción, transformación y carga de datos:

- Extracción: Se encarga de extraer los datos de distintas fuentes de información, identificar los cambios en los datos, manejar de excepciones y establecer la frecuencia de extracción

- Transformación: Se encarga de transformar los datos al formato definido, considerando los subprocesos de limpieza y estandarización.

Tabla 5
Subprocesos de transformación (proceso ETL)

Limpieza (Cleaning)	**Estandarización (Conforming)**
eliminar datos	unificar codificación
Corregir datos	Diferentes formatos
Ausencia de datos	Unidades de medida
Eliminar duplicados	Dimensiones
Datos contradictorios	Tabla Hechos / Agregadas

Fuente: (Alcalde Aliaga, 2018)

- Carga: Se encarga de migrar la información de las fuentes de datos hacia la base de datos dimensional, diferenciando la carga inicial y la carga periódica

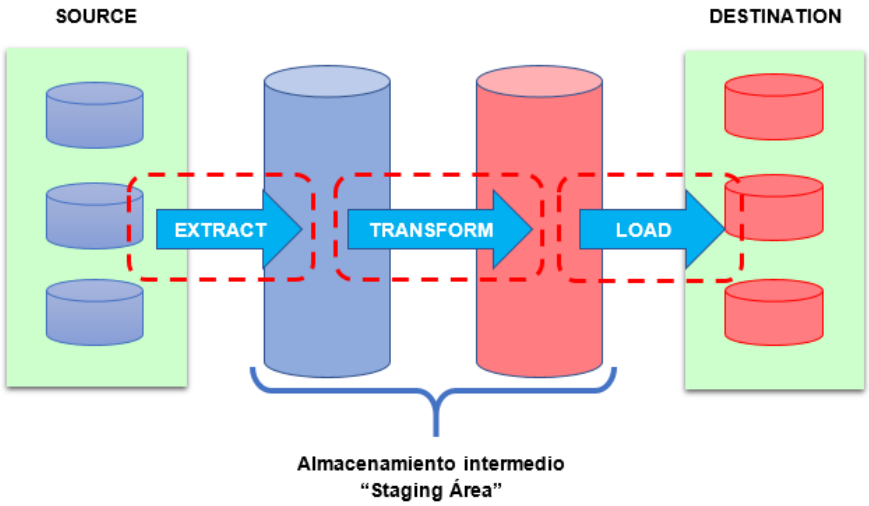
Tabla 6
Tipos de carga (proceso ETL)

Carga Inicial	**Carga Periódica**
Grandes volúmenes	Menos datos
Total	Incremental
Varios procesos	Procesos menos complejos
Histórica	Dia a día

Fuente: (Alcalde Aliaga, 2018)

La figura 14 muestra el flujo de los procesos ETL

Figura 14
Proceso ETL



Fuente: (Alcalde Aliaga, 2018)

K) Sistemas OLTP vs Sistemas OLAP

Según Guillen (2012) los sistemas OLTP (On-Line Transaction Processing) son aquellos sistemas en donde se realizan operaciones diariamente debido a que se realizan operaciones de escritura y actualización, mientras que los sistemas OLAP (On-Line Analytical Processing) se realizan operaciones de lectura y tiene como fin analizar el negocio mediante la a través de indicadores que sirven de apoyo en la toma de decisiones de una organización.

La tabla 7 muestra una comparativa entre estos 2 tipos de sistemas.

Tabla 7
Cuadro comparativo de sistemas OLTP y OLAP

	SISTEMAS OLTP	**SISTEMAS OLAP**
Datos	Valores actuales	Datos históricos y/o calculados
Organización	Por aplicación	Por áreas de la empresa
Acceso	Muy frecuente	Baja frecuencia
Actualizaciones	Actualizaciones de campos	No se actualiza, se manipula
Tiempo de respuesta	Medido por el tiempo de la transacción (segundos)	Medido por el tiempo de la consulta (minutos)
Tamaño de la BD	100MB - GB	100GB -TB
Usuarios	Miles (operativo)	Cientos (decisión)
Unidad de trabajo	Transacciones	Consultas complejas

Fuente: (Guillén F. , 2012)

2.2.5. INTELIGENCIA DE NEGOCIOS VS ANALÍTICA AVANZADA

Según Caldwell (2018) los términos de Inteligencia de negocios y Analytics en la actualidad se usan indistintamente, porque ambos describen la práctica de usar los datos para tomar mejores decisiones de negocio, sin embargo se hace la distinción del uso del mismo, en el caso del vocablo de Inteligencia de negocios cuando se hacen los procesos para producir un informe, panel o tabla dinámica para un ejecutivo, gerente intermedio o analista; y en el caso del vocablo Analytics cuando se pase por las capacidades básicas de BI y use información y datos para ayudar a sus clientes a ser altamente efectivos.

Según Alcalde (2018) los términos de Inteligencia de negocios y Analytics responden a una serie de preguntas claves, las cuales se muestran en la tabla 8.

Tabla 8
Comparativa BI-Analytics

	Inteligencia de negocios	**Analítica avanzada**
Responde a las preguntas	¿Qué paso? ¿Cuándo? ¿Quién? ¿Cuántos?	¿Por qué sucedió? ¿Pasará de nuevo? ¿Qué pasara si cambiamos X? ¿Que más nos dicen los datos que nunca pensamos medir?
incluye	Reportes (KPI's, métricas). Monitoreo y alertas automáticas. Tableros de control. Cuadros de mando. OLAP. Consultas ad hoc. Operaciones en tiempo real BI	Análisis estadístico y cuantitativo. Minería de datos. Modelos predictivos. Pruebas multivariadas. Analítica de Big Data Analítica de texto

Fuente: (Alcalde Aliaga, 2018)

2.2.6. MINERÍA DE DATOS

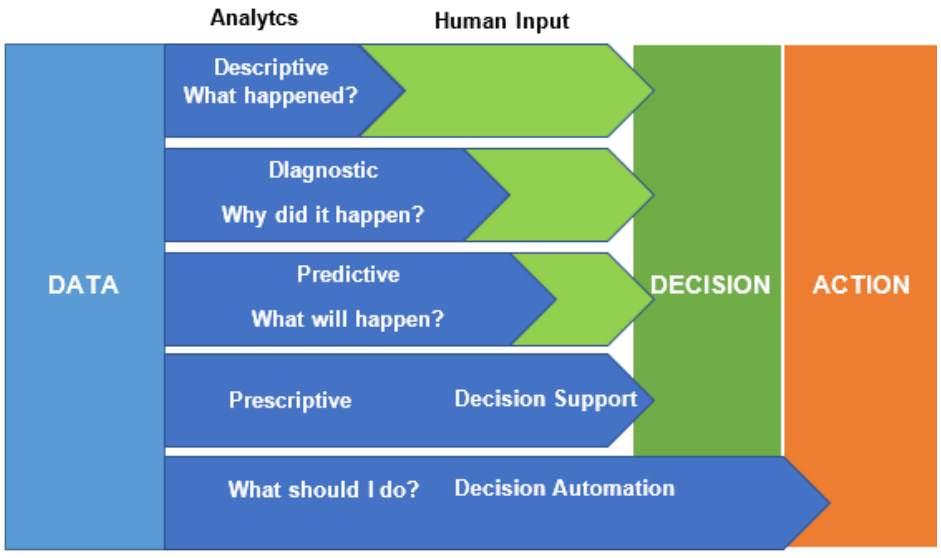
Según Benalcázar (2017), el datamining o minería de datos permite navegar dentro grandes cantidades de datos en tiempos ínfimos para identificar tendencias y patrones de comportamiento en los datos mediante aplicaciones de técnicas y modelos. Para Alcalde (2018) existen 4 objetivos de análisis que posee toda organización:

- Analítica descriptiva/diagnóstica
    - Comprender el estado pasado y actual del negocio.
    - Obtener conclusiones de acciones anteriores que sustenten una decisión estratégica.
    - Entender el "por qué" de los resultados
    - Permite visualizaciones efectivas
    - Reducir costos
    - Segmentación de clientes por ingreso
- Analítica predictiva
    - Prever comportamientos futuros.

- Estimar resultados no conocidos o costosos de obtener.
- Mejor planificación del negocio.
- Fortalece las decisiones estratégicas.
- Retención de clientes.
- Cross selling (venta cruzada).
- Predicción de desempeño.

- Analítica prescriptiva
  - Recomienda acciones para conseguir resultados convenientes, como aumentar beneficios o reducir los riesgos.
  - Integración completa con el negocio, ya que considera no sólo los datos sino su impacto en cuentas y posibles restricciones.
  - Puede redirigir la estrategia del negocio.
  - Optimización de precios.
  - Mix de productos personalizados.

Figura 15
Implicancia de la analítica en la organización



Fuente: (Alcalde Aliaga, 2018)

A) Metodologías

Según Rodríguez (2019) existen 3 metodologías dominantes para la minería de datos

- SEMMA

  Es una metodología propuesta por SAS la cual se basa en 5 fases:

  - Muestreo

    El proceso de muestreo se encarga de seleccionar una muestra representativa de la población de estudio y establecer su nivel de confianza, para aplicar el análisis, el muestreo aplicado puede ser simple y/o aleatorio con reposición.

  - Exploración

    El proceso de exploración comprende el uso de herramientas de visualización y técnicas de estadística descriptiva con la finalidad de conocer los datos para optimizar la eficiencia del modelo y determinar las variables explicativas (entradas del modelo).

  - Modificación

    El proceso de modificación establece la estandarización de los datos que van a ser usados por el modelo.

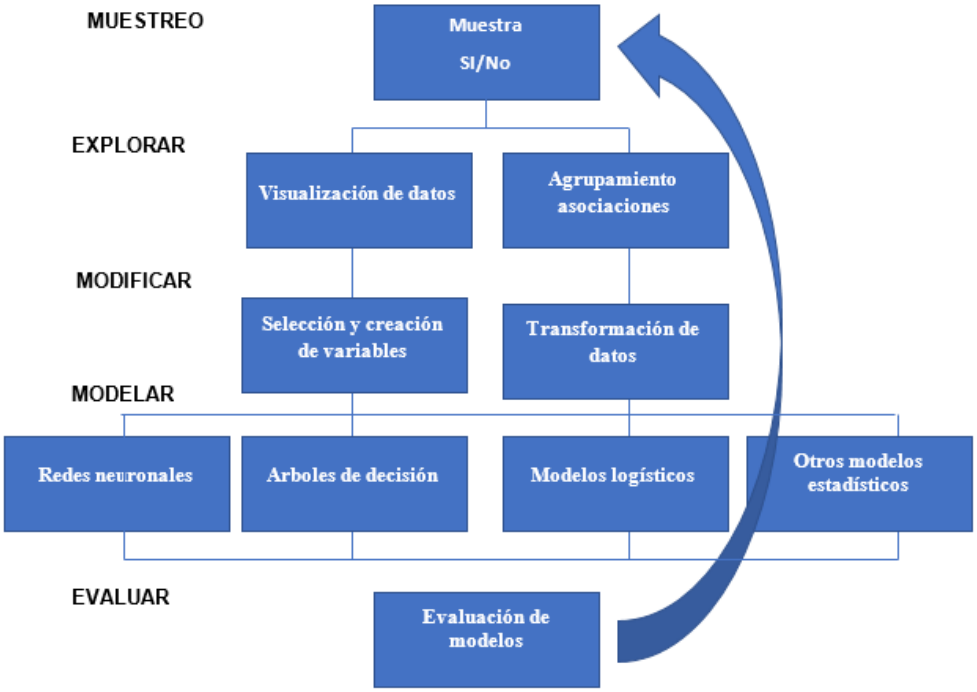
  - Modelamiento

    El proceso de modelamiento permite establecer una relación entre las variables explicativas y objetivo, fijando un nivel confianza apropiado y aplicando distintas técnicas y métodos de minería de datos.

- Evaluación

  El proceso de evaluación permite juzgar la bondad del modelo diseñado a través de otros métodos estadísticos y/o nuevas muestras.

Figura 16
Composición de fases y tareas de la metodología SEMMA



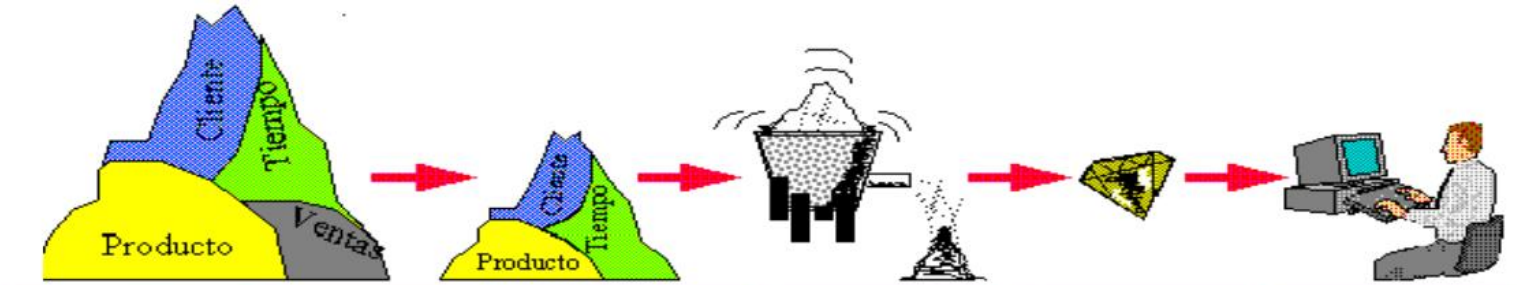
Fuente: (Mérida, 2016-2017)

- CRISP-DM

  Es una metodología sin propietario enfocada al negocio y al análisis técnico, compuesta por 5 fases:

  - Conocimiento del negocio
  - Conocimiento de los datos
  - Preparación de los datos
  - Modelamiento
  - Evaluación y análisis de resultados

Figura 17
Actividades de la metodología CRISP-DM



Conocimiento del Negocio

Determinación de Objetivos de Negocio
Background
Objetivos de negocio
Criterios de éxito de negocio

Evaluación de la Situación
Inventario de recursos
Requerimientos, presunciones y restricciones
Riesgos y contingencias
Costos y beneficios

Determinación de los Objetivos Data Mining
Objetivos Data Mining
Criterios de éxito Data Mining

Producir el Plan del Proyecto
Plan del proyecto
Evaluación inicial de herramientas y técnicas

Conocimiento de los Datos

Recolección de datos iniciales
Lista de datos a solicitar

Descripción de los datos
Reporte de descripción de los datos

Exploración de datos
Reporte de exploración de los datos

Verificar la calidad de los datos
Informe de calidad de datos

Preparación de los Datos

Selección de datos
Razonamiento para la inclusión/exclusión

Limpieza de datos
Informe de la limpieza de los datos

Construcción de datos
Atributos derivados
Registros derivados

Integración de datos
Combinación de datos

Formateo de datos
Datos reformateados

Modelamiento

Selección de la técnica de modelado
Técnicas de modelado
Presunciones del modelado

Generación de la prueba de diseño
Prueba de diseño

Construcción del modelo
Ajuste de parámetros
Modelos
Descripciones del modelo

Evaluación del modelo
Evaluación de modelos
Revisión del ajuste de parámetros

Evaluación y Análisis de Resultados

Evaluación de los resultados
Evaluación de los resultados
Data Mining en lo que concierne a criterios de éxito de negocio
Modelos aprobados

Proceso de revisión
Revisión del proceso

Determinación de los próximos pasos
Lista de posibles acciones
Decisión

Despliegue o Implementación

Desarrollo del plan
Desarrollo del plan

Plan de supervisión y mantenimiento
Supervisión y plan de mantenimiento

Informe final del producto
Informe final
Presentación final

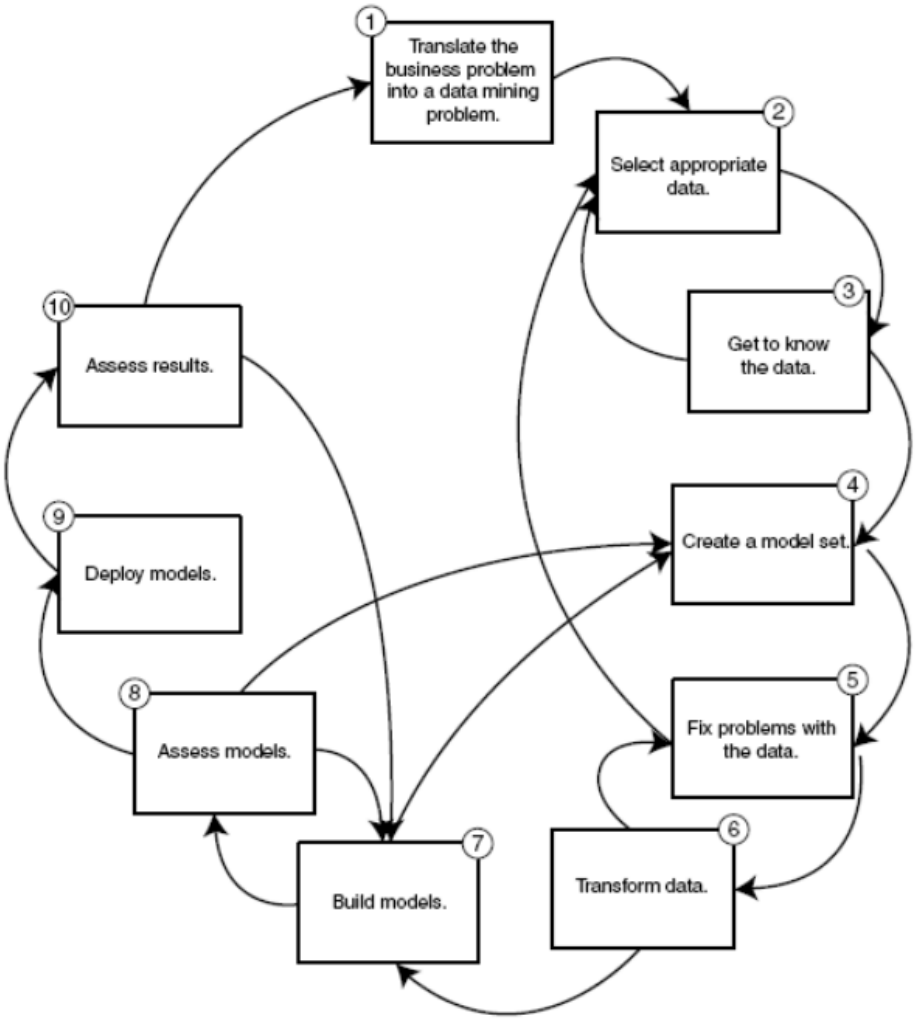
Revisión del proyecto
Documentación de la experiencia

Fuente: (Rodriguez J. , 2019)

- Berry y Linoff

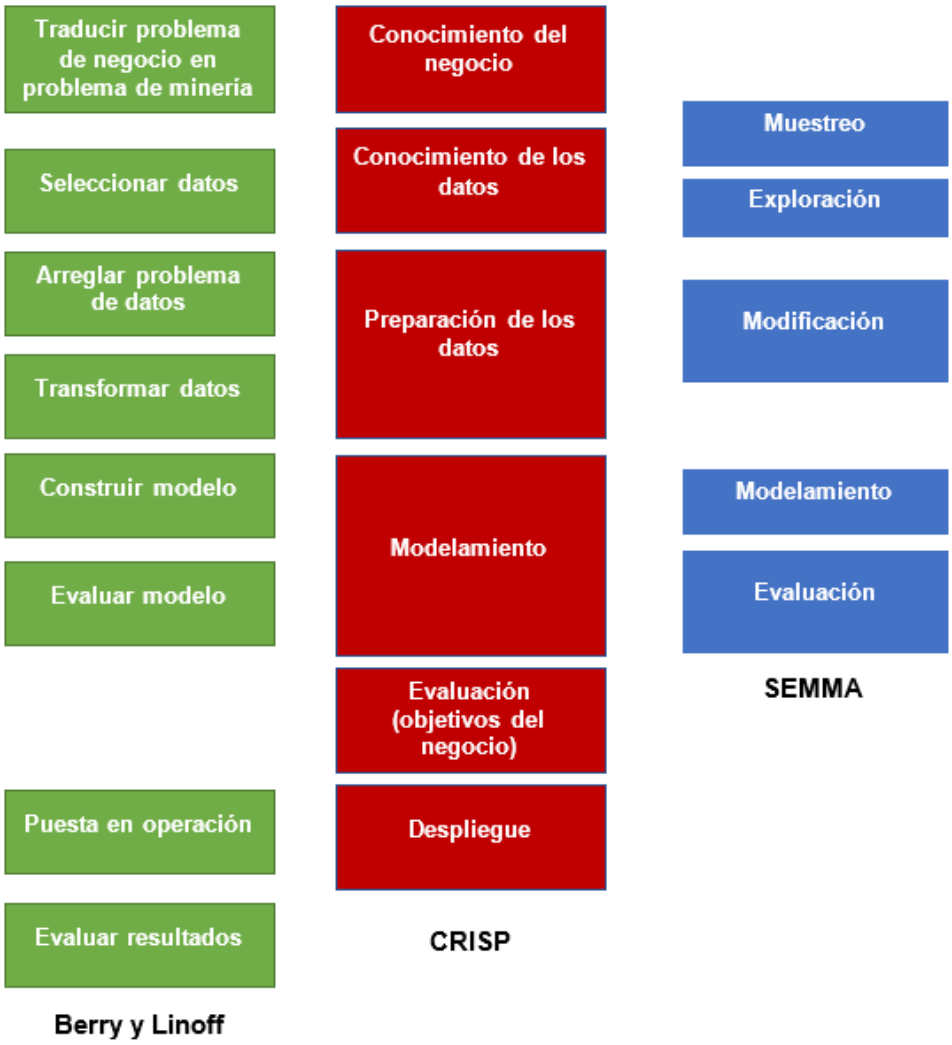
  Es una metodología propuesta por Berry y Linoff compuesta por 8 fases y 10 actividades, de la cuales algunas son bidireccionales, permitiendo volver a formular y obtener un resultado más exacto y favorable.

Figura 18
Actividades de la metodología Berry y Linoff



Fuente: (Rodriguez J. , 2019)

Figura 19
Comparación de metodologías de minería de datos



Fuente: (Rodriguez J. , 2019)

2.2.7. MODELOS PREDICTIVOS

Según Rodríguez (2019) Un modelo predictivo puede definirse como una representación matemática que permite predecir comportamientos futuros en función del conocimiento presente, en el cual se calcula la probabilidad de que un evento ocurra, por lo que un caso particular no es mandatorio, los modelos predictivos pueden ser:

- Modelos Supervisados

  Los modelos supervisados se caracterizan por tener una variable objetivo.

- Modelos No Supervisados

  Los modelos no supervisados se caracterizan por no tener una variable objetivo.

A) Aspectos éticos del análisis predictivo

Según Timón (2017) existe hay una encrucijada ética sobre el análisis predictivo, debido a que el análisis predictivo genera conocimiento sobre las personas debido a que trabaja con un conjunto de datos para determinar patrones comportamiento, sin embargo, se vuelve susceptible al aplicarse a individuos concretos, por ejemplo:

- El conocimiento inducido de análisis predictivo para una organización concluyo que no es recomendable contratar a mujeres de cierta edad y perfil socioeconómico debido a que son más propensas a quedar embarazas, perjudicando a la organización.
- El conocimiento inducido de análisis predictivo para una compañía de seguros determino que no es recomendable brindar el servicio a personas potenciales de sufrir in infarto debido a su estilo de vida.

En ambos ejemplos expuestos no se produce una invasión de la intimidad de las personas afectadas, ni prácticas ilegales, sin embargo, generan una situación discriminatoria.

B) Validación de los modelos

Según Timón (2017) la validación de los modelos, es la única forma de saber si el modelo funciona correctamente, entre las prácticas más asequibles para evaluar un modelo se tiene que dividir el modelo en 2 sub modelos (dos terceras partes para la data de entrenamiento (dataset) y la tercera parte sobrante para la data de prueba (data test)).

C) Técnicas aplicables al análisis predictivo

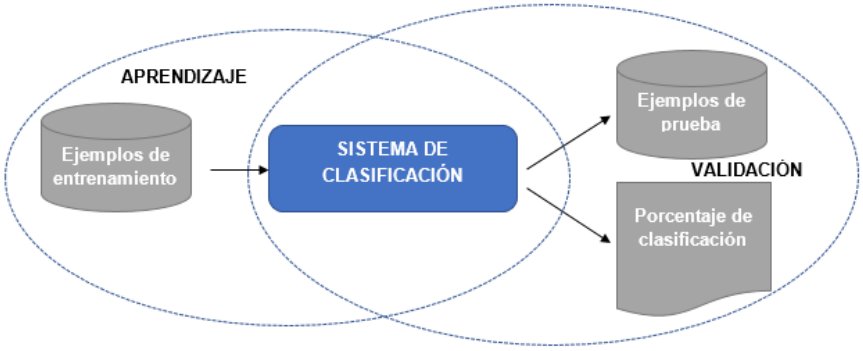
Según Rodríguez (2019) en los modelos supervisados pueden aplicarse las siguientes técnicas:

- Clasificación

  Es una técnica de minería de datos en donde en donde a partir de los datos se construye un clasificador (conjunto de reglas, un árbol de decisión, una red neuronal, etc.).

  El proceso de clasificación está basado en las siguientes etapas según la siguiente figura.

Figura 20
Etapas de proceso de clasificación



Fuente: (Rodriguez J. , 2019)

Dentro de los criterios para evaluar un clasificador se tiene la precisión, velocidad, robustez del modelo y la matriz de confusión los cuales permitirán ver el costo de una clasificación incorrecta, así como la complejidad del modelo.

- Regresión

  Es una técnica de minería de datos que tiene como objetivo modelar la relación entre las variables de estado para para obtener el valor de la variable a predecir (que en su mayoría es numérica).

  Dentro de los métodos tenemos:

  - Basado en ejemplos/instancias
  - Basado en redes neuronales
  - Análisis de regresión
  - Basado en arboles

Según Rodríguez (2019) en los modelos no supervisados pueden aplicarse las siguientes técnicas:

- Agrupación / Clustering / Segmentación

  Es una técnica de minería de datos que identifica y agrupa características similares de teniendo en cuenta maximizar la similaridad intra-cluster y minimizar la similaridad inter-cluster

  Se caracterizan por ser escalables y mostrar capacidad para tratar con datos con ruido y outliers.

  Existen principalmente 2 distintas aproximaciones de clustering:

  - Jerárquico

    Crea una disgregación jerárquica de los datos basándose en algún criterio, en donde obtendremos un clustering distinto de acuerdo a nivel de corte.

- Método aglomerativo

  Comienza con tantos clústeres como individuos y va formando grupos según la similitud teniendo en cuenta la distancia del vecino más próximo y más lejano y la distancia entre los centroides.

- Método de división

  Comienza con único cluster (toda la BD) y consiste en ir dividiendo el cluster según la similitud de sus componentes.

- Partición

  Construye distintas particiones y las evalúa de acuerdo a algún criterio, en los clustering de partición se suele definir el número de agrupaciones.

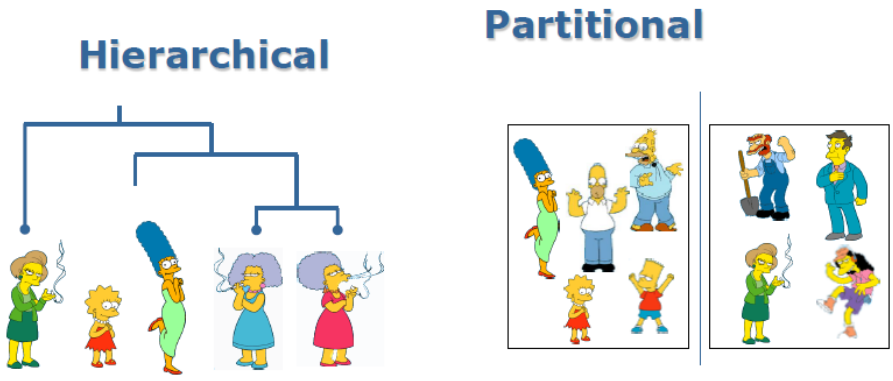
  Métodos Heurísticos

  - K-means: cada cluster se representa por el centro del cluster.

    - Algoritmo k-means: es un algoritmo iterativo en el que las instancias se van moviendo entre clusters hasta que se alcanza el conjunto de clusters deseados.

  - K-medoids: cada cluster se representa por uno de los objetos incluidos en el cluster.

Figura 21
Distinción entre las aproximaciones de clustering



Fuente: (Rodriguez J. , 2019)

- Asociación

  Es una técnica de minería de datos que utiliza para descubrir patrones frecuentes, secuencias, asociaciones, correlaciones o estructuras casuales entre el conjunto de datos.

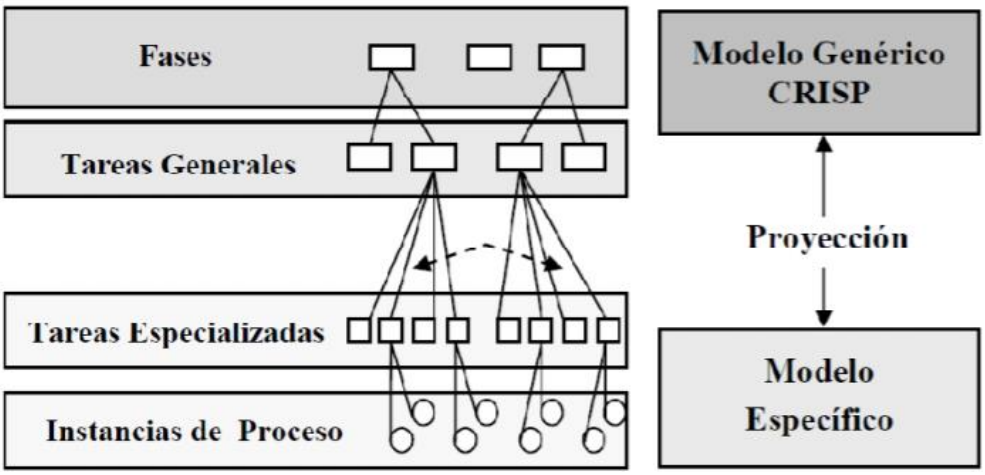
  - Algoritmo A priori

    Es un algoritmo simple pero robusto que encuentra las asociaciones más frecuentes, iterando sobre la base de datos hasta que las asociaciones obtenidas no tienen soporte mínimo.

2.2.8. METODOLOGÍA CRISP-DM

Según Gallardo (2009) la metodología CRISP-DM cuenta con 4 niveles de abstracción (ver figura 22) conformado por tareas que van desde el punto de vista general hasta el especifico y organiza el desarrollo de un proyecto de datamining, en una serie de seis fases (figura 23):

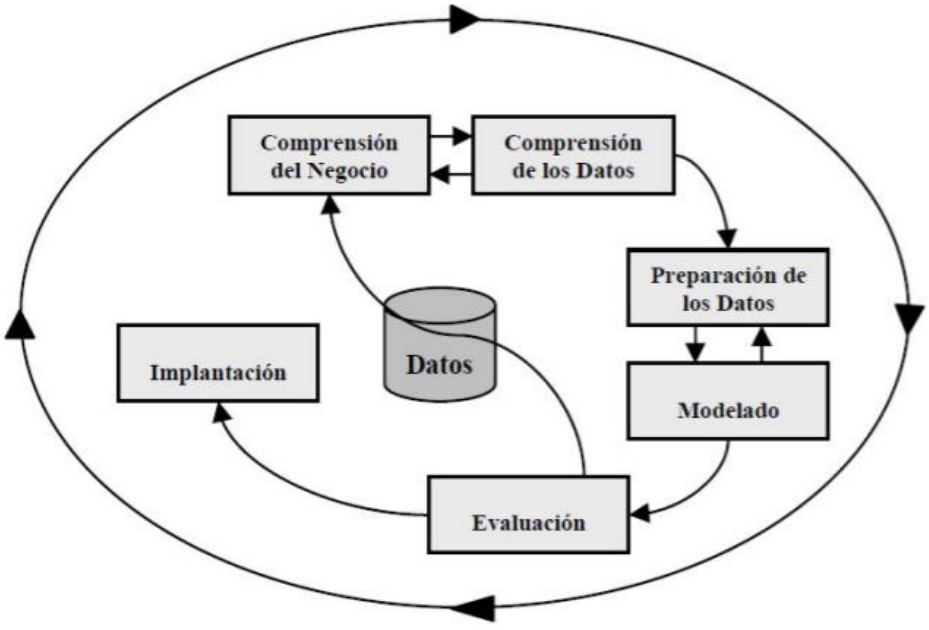
Figura 22
Esquema de los 4 niveles de CRISP-DM



Fuente: (Gallardo, 2009)

Cada fase maquila tareas generales que se proyectan en tareas específicas la cuales describen las acciones a realizar, sin embargo, nunca se propone como ejecutarlas.

Figura 23
Modelo del proceso CRISP-DM



Fuente: (Gallardo, 2009)

A continuación, se describen cada una de las fases en que se divide CRISP-DM.

A) Comprensión del negocio

Esta fase se caracteriza porque a partir del conocimiento adquirido del negocio se homologa en un problema de datamining y en un plan preliminar que tiene como fin alcanzar los objetivos del negocio. Las principales tareas que componen esta fase son las siguientes:

- Determinar los objetivos del negocio.

  Se define la problemática a resolver, el por qué la necesidad de utilizar datamining y los criterios de éxito.

- Evaluación de la situación.

  Se Analiza la situación del negocio, definiendo los requisitos del problema, tanto en términos de negocio como en términos de datamining.

- Determinar de los objetivos de DM.

  Se homologa los objetivos del negocio en criterios de éxito para el proyecto de minería de datos.

- Producir un plan del proyecto

  Se diseña un plan de trabajo que describa el paso a paso para desarrollar el proyecto de minería de datos.

B) Compresión de los datos

Esta fase comprende la recolección de datos con la finalidad de familiarizarse, identificar su calidad y tener una idea de la problemática. Las principales tareas a desarrollar en esta fase del proceso son:

- Recolección de datos iniciales.

  Se lleva un registro de los datos adquiridos, técnicas usadas para su recolección y los problemas y/o soluciones generados en el proceso.

- Descripción de los datos.

  Se describe cada dato desde el significado de cada campo hasta la descripción del formato inicial.

- Exploración de datos.

  Se aplican estadísticos básicos a los datos con la finalidad de encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas.

- Verificación de la calidad de los datos.

  Se realizan validaciones sobre los datos con la finalidad de determinar inconsistencias, valores nulos y valores fuera de rango los cuales pueden generar ruido el proceso.

C) Preparación de los datos

Esta fase se encarga de adaptar los datos para aplicar técnicas de datamining. Las principales tareas a desarrollar en esta fase del proceso son:

- Selección de datos.

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

- Limpieza de los datos.

  Se aplican diversas técnicas como normalización, discretización y tratamiento de valores ausentes con la finalidad de obtener datos de calidad y sean adecuados para la fase de modelación.

- Estructuración de los datos.

  Se puede generar nuevos atributos a partir de atributos y/o transformar los valores de atributos existentes.

- Integración de los datos.

  Se encarga de crear nuevas estructuras, a partir de los datos seleccionados.

- Formateo de los datos.

  Se encarga de adecuar los datos sin transgredir su significado, con la finalidad de facilitar el empleo de alguna técnica de datamining en particular.

D) Modelado

En esta fase se discierne las técnicas de modelado más apropiadas para el proyecto de minería de datos, teniendo en cuenta la disposición de los datos adecuados, el cumplimiento de los requisitos del problema y conocimiento de la técnica.

Anticipadamente al modelado de los datos, se debe fijar un método de evaluación el cual permitirá definir el grado de bondad del modelo.

Las principales tareas de esta fase son las siguientes:

- Selección de la técnica de modelado

  Consiste en seleccionar la técnica de minería de datos más apropiada al tipo de problema a resolver teniendo en cuenta la relación entre los objetivos de minería de datos y las herramientas existentes.

- Generación del plan de prueba.

  Consiste en generar una serie de actividades que permitan probar la calidad y validez del modelo minería de datos.

- Construcción del Modelo.

  Se encarga de ejecutar la técnica seleccionada sobre los datos preparados para generar uno o más modelos a través de un proceso iterativo con la finalidad de encontrar el modelo con mejor rendimiento.

- Evaluación del modelo.

  Se encarga de interpretar los modelos teniendo en cuenta 3 factores: adecuación a los datos y los criterios de éxito de minería de datos

E) Evaluación

En esta fase se evalúa el modelo, considerando los criterios de éxito del problema y criterios de los expertos del negocio. Las tareas principales tareas en esta fase del proceso son las siguientes:

- Evaluación de los resultados.

  Se encarga de evaluar el modelo en relación a los objetivos del negocio y ver que tanto afecta al negocio.

- Proceso de revisión

  Se encarga de valorar todo el proceso de minería de datos e identificar que actividades pueden ser retocados.

- Determinación de futuras fases.

  Se encarga de establecer criterios para discernir si se pasa al proceso de implementación o es necesario retocar y/o iniciar desde cero cualquiera de los procesos anteriores.

F) Implementación

En esta fase se encargar de homologar el conocimiento obtenido de minería de datos una vez validado en acciones dentro del proceso de negocio.

Las principales tareas que se ejecutan en esta fase son las siguientes:

- Plan de implementación.

  Se encarga transformar los resultados de minería de datos a estrategias de negocio.

- Informe Final.

  Se puede realizar de 2 formas como resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación que explaye los resultados logrados en el proyecto.

- Revisión del proyecto:

  Se encarga de evaluar lo correcto y lo incorrecto, así como discernir en lo que se puede mejorar.

2.2.9. PRINCIPALES HERRAMIENTAS DE ANÁLISIS PREDICTIVO

Según Digital Guide (2018) las mejores herramientas con GUI en el rubro de datamining son las siguientes:

A) Weka

WEKA (Waikato Environment for Knowledge Analysis) es un software de código abierto desarrollado por la Universidad de Waikato en la década de los noventa implementada en Java y compatible con Windows, macOS y Linux, es capaz de procesar en ella los datos solicitados. Asimismo, cuenta con un sinnúmero de funciones de aprendizaje automático y secunda tareas tan relevantes del datamining como el análisis de clústeres, correlación o regresión, así como la clasificación de datos, punto fuerte este último al usar redes de neuronas artificiales, árboles de decisión y algoritmos ID3 o C4.5.

B) Orange Data Mining

Orange nació de un proyecto como proyecto de la Universidad de Liubliana hace más de 20 años, desarrollado inicialmente en C++, para después ampliarse con Python.

Orange ofrece aplicaciones de gran utilidad para el análisis de datos y de texto, así como características de aprendizaje automático, trabajando con operadores para la clasificación, regresión y clustering a través de una programación visual la cual facilita a los usuarios en el proceso de datamining para tomar decisiones rápidamente en el ámbito profesional.

C) Knime

El software KNIME (Konstanz Information Miner), desarrollado por la universidad de Constanza, está basado en Java y preparado con

Eclipse, y se destaca por contar con una amplia gama de módulos y paquetes los cuales permiten descubrir estructuras ocultas de datos.

Dentro de las bondades de KNIME se tiene: la integración de numerosos procedimientos de aprendizaje automático y de datamining, y la eficiencia en el tratamiento previo de los datos.

D) SAS Studio

SAS (Statistical Analysis System) es un producto de SAS Institute, se caracteriza por su gran escalabilidad debido a que permite aumentar progresivamente su eficiencia aumentando los recursos de hardware o de cualquier otro tipo y porque dispone de una interfaz de usuario gráfica.

E) RapidMiner

RapidMiner, antes conocida como YALE, siglas de “Yet Another Learning Environment”, es un software de minería de datos basada en Java que contiene más de 500 operadores con diferentes enfoques para mostrar las conexiones en los datos: hay opciones para datamining, textmining o webmining, análisis de sentimiento o minería de opinión; que se caracteriza ca por permitir el acceso gratuito y por su fácil manejo dado que no requiere un conocimiento elaborado en programación.

La herramienta está formada por tres grandes módulos: RapidMiner Studio, RapidMiner Server y RapidMiner Radoop, cada uno encargado de una técnica diferente de minería de datos. Asimismo, RapidMiner prepara los datos antes del análisis y los optimiza para su rápido procesamiento.

A continuación, se muestra un cuadro comparativo entre la lista de software mencionado anteriormente:

Tabla 9
Comparativa de software de datamining

	Características	**Lenguaje de programación**	**Sistema operativo**	**Precio / Licencia**
RapidMiner	Apto para todos los procesos. Destaca en el análisis predictivo	Java	Windows, macOS, Linux	Freeware, diferentes versiones de pago
WEKA	Muchos métodos de clasificación	Java	Windows, macOS, Linux	Software libre (GPL)
Orange	Crea una visualización de datos atractiva sin que se requieran muchos conocimientos previos para ello	Núcleo del software: C++, ampliación y lenguaje de entrada: Python	Windows, macOS, Linux	Software libre (GPL)
KNIME	Software de datamining de código abierto que ha democratizado el acceso a los análisis predictivos	Java	Windows, macOS, Linux	Software libre (GPL) (a partir de la versión 2.1)
SAS	Caro, pero potente para grandes empresas	Lenguaje SAS	Windows, macOS, Linux	Freeware limitado a instituciones públicas, el precio se establece tras solicitud, diferentes modelos disponibles

Fuente: (Digital Guide, 2018)

Figura 24
Magic Quadrant for Data Science and Machine Learning Platforms

Fuente: (Idoine, Krensky, Brethenoux, & Linden, 2019)

F) R Studio

Según Timón (2017) RStudio es un IDE de software libre para desarrollo que hace uso de complementos llamados paquetes que proporcionan funcionalidades de modelado en R.

A través de RStudio se pueden adjuntar conjuntos de datos al entorno desde múltiples formatos (como .csv, .xlsl, .txt, .json, .dbf o .xml) y hacer el tratamiento de los datos como normalización, estandarización y agregación antes de generar los modelos predictivos.

2.3. DEFINICIÓN DE TÉRMINOS BÁSICOS

- ANÁLISIS ESTADÍSTICO

  Es la ciencia de recopilar, explorar y presentar grandes cantidades de datos con la finalidad de encontrar patrones y tendencias implícitas. (SAS, s.f.)

- ATRIBUTOS

  Constituyen criterios de análisis de las métricas en un cubo dimensional, siendo en su mayoría en los campos de las tablas de dimensiones. (Esparza, Alvarez, Duque, & Quiroz, 2014)

- BUSINESS ANALYTICS

  Se enfoca en el análisis a futuro con base en la información de la empresa y modelos predictivos para apoyar la toma de decisiones y mejorar la competitividad del negocio. (ESAN, 2017)

- CONOCIMIENTO

  Es una composición de experiencia, valores, información y know-how que sirve como marco para la incorporación de nuevas experiencias e información. (Sinnexus, s.f.)

- DATAMART

  Es un repositorio de datos que contiene información de un área en específica. (Córdova, 2013)

- DATAMINING

  Es el conjunto de técnicas y tecnologías aplicadas a los datos con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los mismos en un determinado contexto. (Sinnexus, s.f.)

- DATAWAREHOUSE

  Es una colección de datamart ligada a la información de una empresa la cual es usada como soporte para el proceso de tomas de decisiones gerenciales. (Córdova, 2013)

- DATO

  Es la mínima unidad semántica, que corresponde como elemento primario de información que por sí solos son irrelevantes como apoyo a la toma de decisiones. (Sinnexus, s.f.)

- DATOS MAESTROS

  Representan entidades claves del negocio, cuyos datos se han acordado compartir a través de toda la empresa, por ejemplo: cliente, producto, empleado, etc. (Ferrándiz, 2018)

- DATOS DE REFERENCIA

  Son un subconjunto del maestro de datos que referencia a datos que definen un conjunto permitido de valores a ser usados por otros campos de datos, por ejemplo: código postal, clasificación de industria, género, estado civil, etc. (Ferrándiz, 2018)

- DIMENSIÓN

  Es una colección de miembros, unidades o individuos del mismo tipo (Bustos & Mosquera, 2013)

- DISTRIBUIDORA

  Es una compañía que compra bienes o servicios y los comercializa a otras compañías para obtener ganancias. (Invitado, 2012)

- ELEMENTO

  Es cada una de las personas, animales, cosas o entidades de las cuales se reúnen los datos. (Mori Paiva, 2018)

- ERP

  Es un sistema estructurado que brinda soporte a la parte operativa de la organización, permitiendo monitorear a nivel de registro el estado de los procesos de la empresa, los avances de cada área, entre otros. (Córdova, 2013)

- ESTIMADOR

  Un estimador es un estadístico que se usa en la muestra, para tasar un parámetro desconocido de la población. (Mori Paiva, 2018)

- ETL

  Es el responsable de extraer la información, personalizarla y cargarla en el almacén de datos, los cuales manejan un gran volumen de datos y permiten gestionar la carga de trabajo. (Sánchez, 2014)

- FORECASTING

  Es una técnica que utiliza datos históricos como entradas para realizar estimaciones predictivas en dirección a las tendencias futuras. (Chen, 2018)

- FUERZA DE VENTAS

  Está conformada por un conjunto de empleados que se dedican a realizar las ventas de una compañía. (Vega, 2005)

- GRANULARIDAD

  El concepto de granularidad se basa en el nivel de detalle sobre la información que se desea almacenar, siendo un factor importante debido a que impacta en la toma de decisiones. (Quillén, 2017)

- INFORMACIÓN

  La información se puede definir como un conjunto de datos procesados y que tienen un significado. (Sinnexus, s.f.)

- INSIGHTS

  Es el valor obtenido a través del uso de análisis. (Marrs, 2016)

- METADATO

  El concepto de metadatos se refiere a aquellos datos que describen el contenido de los archivos o la información de los mismos. (PowerData, 2016)

- MODELADO DIMENSIONAL

  Presenta los datos de un marco de trabajo e intuitivo, para permitir su acceso con un alto rendimiento, cada modelo dimensional está compuesto por una tabla con una llave combinada, llamada tabla de hechos y un conjunto de tablas más pequeñas llamadas tabla de dimensiones. (Bustos & Mosquera, 2013)

- MUESTRA

  Es un subconjunto representativo de elementos provenientes de una población. (Mori Paiva, 2018)

- OUTLIERS

  Los outliers se definen como valores atípicos, que son una observación que se aparta de un patrón general de la muestra, son valores extremos que se desvían de otras observaciones en los datos. (Santoyo, 2017)

- PARÁMETRO

  Es el conjunto de todos los elementos que se desean analizar y que presentan una o varias características en común. (Mori Paiva, 2018)

- POBLACIÓN

  Es el conjunto de todos los elementos que se desean analizar y que presentan una o varias características en común. (Mori Paiva, 2018)

- PROCESO DE VENTAS

  Según Sánchez (2014) indica que el proceso de ventas es una secuencia lógica, que emprende el vendedor con un comprador potencial y que tiene por objeto producir alguna reacción deseada en el cliente (usualmente la compra).

- SESGO

  Es la diferencia entre la esperanza de un estimador θ y el verdadero valor del parámetro θ a estimar. Un estimador θ es insesgado, si: $E\ \theta = \theta$. (Mori Paiva, 2018)

- SISTEMA OLAP

  Se denomina al proceso en el que se emplean herramientas sofisticadas que permiten agilizar el proceso de análisis de grandes cantidades de datos desde diferentes fuentes, organizada en perspectivas (dimensiones) y métricas, permitiendo ejecutar análisis complejos de datos, en base a los cuales se tomarán las decisiones del negocio. OLAP, permite a los usuarios una fácil y amigable navegación por la información, obteniendo el nivel de granularidad. (Córdova, 2013)

- SISTEMA OLTP

  Son sistemas para la captura de transacciones cotidianas (ventas, compras, control de almacén, cuenta corriente, generación de notas de crédito, control de la producción, contabilidad, etc.) y fuente principal de las soluciones analíticas. Se encarga de dar soporte a los procesos diarios de ingreso y mantenimiento de datos en tiempo real. Proporciona soporte muy limitado y poco eficiente para la elaboración de reportes y, en consecuencia, para la toma de decisiones de la empresa. (Córdova, 2013)

- TABLA DE HECHOS

  Es una colección de piezas de datos y datos de contexto. Cada hecho representa una parte del negocio, una transacción o un evento. (Bustos & Mosquera, 2013)

- TOMA DE DECISIONES

  La Toma de decisiones consiste básicamente en elegir una opción entre las disponibles, a los efectos de resolver un problema actual o potencial. (Guillén F. , 2012)

- VARIABLE

  Es el dato registrado producto de la apreciación de una característica en un individuo o unidad elemental. (Mori Paiva, 2018)

CAPÍTULO III: DESARROLLO DEL TRABAJO DE SUFICIENCIA PROFESIONAL

Se implementará la metodología CRISP-DM en el desarrollo del presente trabajo debido a que es una metodología neutral, estructurada y fácil de implementar, y según Moine, Gordillo, & Haedo (2011) es una de las principales metodologías a implementar porque:

- Contempla el análisis y compresión del problema previo al proceso de minería.
- Puntualiza las actividades específicas en cada fase del proceso.
- Propone actividades de planificación para las distintas áreas de la gestión del proyecto.

3.1. MODELO DE SOLUCIÓN PROPUESTO

Está basado en los procesos de la metodología CRISP-DM.

3.1.1. CONOCIMIENTO DEL NEGOCIO

Permite comprender la problemática que se desea resolver, a través de la recolección de los datos correctos e interpretación de los resultados.

A) Determinar los objetivos del negocio

La Gerencia Central y la Gerencia de Ventas de la distribuidora Jiménez e Iriarte S.A. es consciente que se vive en un mundo globalizado, por lo cual se debe responder a los cambios con rapidez y precisión, en consecuencia, ve la necesidad de contar con una solución tecnológica que permita generar nuevas oportunidades de negocio y ayude a comprender mejor el comportamiento del cliente para atenderlo de manera oportuna y correcta, sin generar sobrecostes.

Dentro de los objetivos propuestos por la empresa se tiene:

- Agrupar a los clientes nuevos en base a su comportamiento de compra.
- Sugerir la compra de productos de acuerdo al comportamiento del cliente.

La Gerencia de Ventas establece como criterios de éxito:

- Conocer el comportamiento de compra del cliente.
- Cumplir con los objetivos mensuales asignados a la fuerza de venta exclusiva.

B) Evaluación de la situación

Actualmente se cuenta con un Data Center con discos de estado sólido e interconectados a la red LAN vía fibra canal a 10Gb/s, en el cual se contiene el servidor virtual que tiene instalado SQL Server 2014 y almacena el detalle de las ventas desde el año 2016 (marzo) hasta la fecha, por lo que supone que se cuenta con la cantidad de registros para poder resolver el problema.

- **Inventario de recursos**
  - Software
    SQL Server 2014
    Pentaho PDI (Open Source)
    R (Open Source)

  - Hardware
    1 servidor Virtual - datamart (Intel® Xenon ® CPU E5-2697Av4 @2.6 GHz (8 procesadores), 64 GB RAM, 1 TB de almacenamiento, Windows Server 2012 R2 x64).
    1 servidor Virtual - minería de datos (Intel® Xenon ® CPU E5-2697Av4 @2.6 GHz (16 procesadores), 128 GB RAM, 1 TB de almacenamiento, Windows Server 2012 R2 x64).

- Fuente de datos

  No se cuenta con un repositorio exclusivo para las consultas.

- Recursos humanos

  Se cuenta con personal capacitado para generar un repositorio de datos y modelos predictivos, así como expertos dentro del rubro de ventas.

Tabla 10
Stakeholders del proyecto

Stakeholders	**Cargo**	**Función que desempeña en el proyecto**
Rivelino Aguirre	Gerente del Área de Ventas	Brinda información sobre la gestión de ventas y los alcances de los proveedores.
Juan Isla	Jefe de Venta	Análisis de la situación actual de las ventas realizadas.
José Fernández	Analista de datos	Análisis de requerimiento y carga de datos al datamart
Renzo Gutierrez	Analista de datos/Analista BA	Generación y validación de modelos predictivos.

Fuente: Propia

- **Requisitos, supuestos y restricciones**
  - Requisitos funcionales
    - Contar con una herramienta que optimice la gestión de ventas (preventa), debido a que actualmente se hace uso de Microsoft Excel, a través de OBDC, para la obtención de reportes y tablas dinámicas; la

incorporación de un datamart es clave debido a que será la fuente verídica de información.

- El repositorio debe contemplar lo siguiente
    - Las cantidades de productos vendidas por sucursal / grupos de venta / fuerzas de venta / supervisor / vendedor / Cliente en un tiempo determinado.
    - El monto total de las ventas de los productos en un determinado tiempo.
    - La cantidad de productos bonificados en un determinado tiempo.

  Todas estas variables permitirán comprender mejor el comportamiento del cliente para poder categorizarlo para acceder a promociones y descuentos, y sugeridos.

- Requisitos no funcionales
    - El datamart (repositorio de información) debe ser construida sobre el Motor de base de datos SQL Server
    - La herramienta de inteligencia de negocios debe ser desarrollada bajo una plataforma Open Source.
    - La herramienta de minería de datos debe ser desarrollada bajo una plataforma Open Source.

        - Se debe contar ciertas medidas de seguridad, debido a la a la data sensible con la cual se va a trabajar.
    - Supuestos
        - No debe haber volatilidad en los honorarios y/o costes presupuestados en el proyecto.
        - Los datos no deben contar con missing y outliers.
        - No hacer uso de toda la fuente de información para el proyecto de minería de datos.
    - Restricciones
        - Se debe disponer de todas las contraseñas para acceder a los datos.
        - Se debe contar con la autorización del cliente para el tratamiento de los datos sensibles.
        - Solo se trabajará con información de la fuerza exclusiva – Lima para la aplicación de minería de datos.

- **Riesgos y contingencias**
    - El riesgo que presenta el proyecto es el tiempo de desarrollo del mismo por las pruebas de los distintos modelos y la calidad de los datos desde las fuentes de información.
    - El plan de contingencia consiste en redefinir los tiempos con la finalidad de cumplir con los plazos establecidos y realizar un tratamiento de los datos con la finalidad de estandarizarlos.

- **Terminología**

  Ver Capitulo II: Marco teórico - definición de términos básicos.

- **Costes y beneficios**

  El proyecto no supone ningún coste adicional frente a lo presupuestado (ver tabla 11). En cuanto a los beneficios, se espera que un crecimiento a nivel de las ventas a través del redescubrimiento del cliente y que el cliente perciba lo importante que es para la distribuidora.

Tabla 11
Presupuesto

RUBROS	Tipo de recurso		TOTAL
	Efectivo	***En especie***	
RECURSOS HUMANOS	**10000**	**0**	**10000**
Analista de datos	4000	0	4000
Analista BA	6000	0	6000
RECURSOS MATERIALES	**0**	**0**	**0**
Laptop	0	0	0
Servidor	0	0	0
SERVICIOS	**900**	**0**	**900**
Impresión	300	0	300
Transporte	150	0	150
Internet	150	0	150
Luz	300	0	300
OTROS	**2200**	**0**	**2200**
Compra de equipos	0	0	0
Viáticos	200	0	0
Licencias	0	0	0
Sobrecostos	2000	0	0
Bibliografía - únicamente libros	0	0	0
TOTAL	**13100**	**0**	**13100**

Fuente: Propia

C) Determinación de los objetivos de datamining

Dentro de los objetivos propuestos del datamining se tiene:

- Clasificar al cliente para que pueda acceder ciertos beneficios (bonificaciones y/o descuentos).
- Poder ofrecer sugeridos de compra de acuerdo al comportamiento del cliente.

Los criterios de éxito del datamining son:

- Predecir la clasificación del cliente con una fiabilidad del 85% para poder realizar sugeridos y que pueda acceder a promociones y/o recomendaciones personalizadas.

D) Producir un plan de proyecto

El proyecto se desmenuzará en 7 etapas para facilitar su organización:

- Comprensión la situación del negocio. Tiempo estimado 5 días. (etapa 1)
- Comprensión de los datos a través de análisis de la estructura y calidad de los datos. Tiempo estimado: 5 días. (etapa 2)
- Preparación de los datos mediante la selección, limpieza, conversión y formateo de los datos. Tiempo estimado: 10 días. (etapa 3)
- Elección y ejecución de las técnicas de modelado sobre los datos. Tiempo estimado: 6 días. (etapa 4)
- Análisis de los resultados obtenidos, si fuera necesario repetir la etapa 4. Tiempo estimado: 5 días. (etapa 5)
- Generación de informes con los resultados obtenidos en función de los objetivos y los criterios de éxito de negocio establecidos. Tiempo estimado: 7 días. (etapa 6)
- Promoción de los resultados. Tiempo estimado: 7 días. (etapa 7)

La herramienta a utilizar para ejecutar el proyecto de minería de datos es R-Studio, entre sus principales características reside contar con un IDE de programación y también por su facilidad de trabajar con librerías externas (packages) que añaden más funcionalidades a R, dentro de los paquetes importantes tenemos. rodbc, roughtsets, sqldf, etc.

En cuanto a las técnicas que se van a emplear para la extracción de conocimiento, R ofrece los siguientes tipos de tareas de minería de datos a través de sus paquetes:

- Predictivas: Clasificación o Regresión
- Descriptivas. Agrupamiento (clustering) o Reglas de asociación

3.1.2. CONOCIMIENTO DE LOS DATOS

Implica estudiar más de cerca los datos disponibles de minería, con la finalidad de evitar problemas inesperados durante la siguiente fase.

A) Recolección de datos

Hace referencia a la recopilación inicial de los datos de la empresa, en una base de datos alterna y su adecuación para su futuro procesamiento. Como base de datos alterna se prosiguió con el diseño de un datamart para el área de ventas, para posteriormente generar una consulta respecto a la información de la fuerza de venta exclusiva de la sucursal de Lima con registros de los años 2016, 2017, 2018 y 2019 (enero-febrero), ignorando las bonificaciones, la cual será la fuente de información del proyecto.

- Criterios de selección

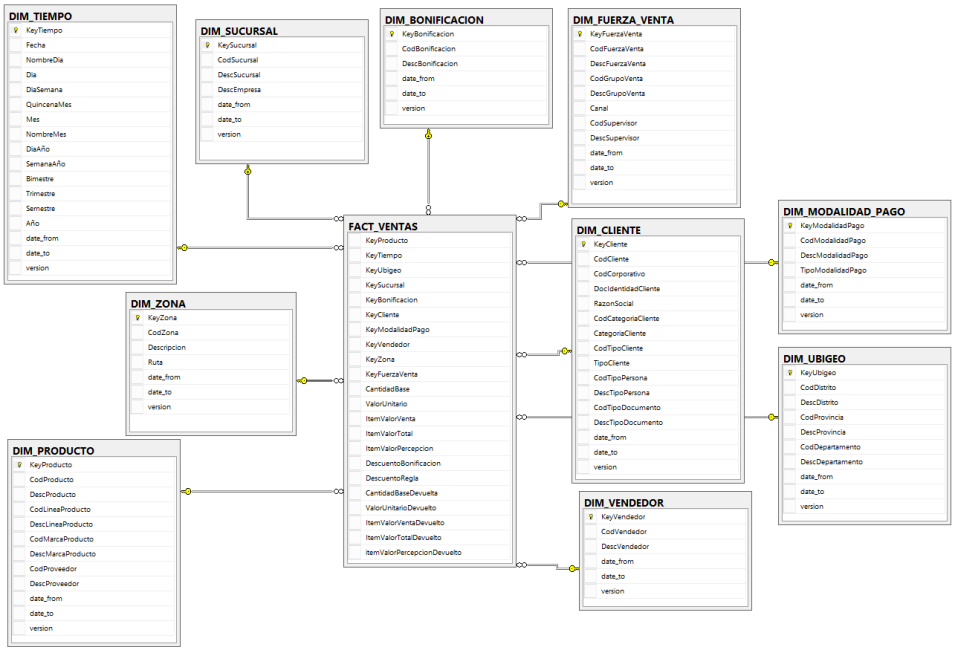
  Considerando que el objetivo de la minería de datos es predecir la clasificación del cliente de acuerdo a su comportamiento de compra, se considerará la información de las ventas que son almacenadas en las tablas del datamart (cliente, fuerza de venta, modalidad de pago, producto, sucursal, tiempo, ubigeo,

vendedor, zona y tiempo), el cual ha sido diseñado como parte de un requerimiento funcional por parte de la gerencia.

- Creación del datamart

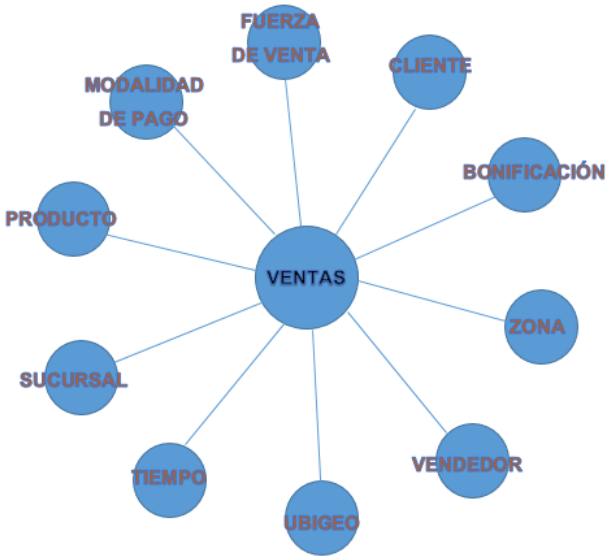
  Considerando las ventas como fuente de información se procede a seleccionar las dimensiones, tabla de hechos y métricas, y se muestra en la siguiente figura.

Figura 25
Modelo dimensional estrella - Datamart Ventas



Fuente: Propia

Figura 26
Modelado grafico de alto nivel



Fuente: Propia

- Selección del motor de base de datos y la herramienta de inteligencia de negocios

  En esta etapa se selecciona el Gestor de base de datos y la herramienta de inteligencia de negocios necesaria para la implementación del Datamart de ventas para su posterior instalación.

  - Motor de base de datos

    En cumplimiento de los requerimientos funcionales y no funcionales se optó con el motor de base de datos SQL Server 2014, debido a que se cuenta con una cartera de licencias disponibles y el tratamiento de los datos debe de ser más sencillo debido a que el sistema transaccional también trabaja sobre un motor de base de datos SQL Server.

  - Herramienta de inteligencia de negocios

    En cumplimiento de los requerimientos funcionales y no funcionales se optó por la Suite Pentaho debido a las ventajas sobre otras herramientas Business Intelligence (figura 27).

Figura 27
Cuadro comparativo de herramientas de extracción

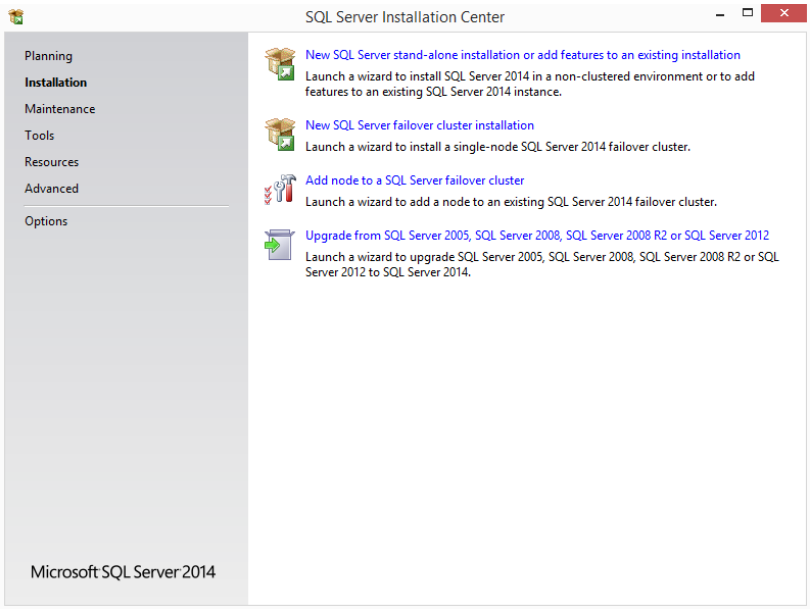
Características de la Herramienta	Business Objects	Microsoft SQL Server	IBM	pentaho	Java Clove	Java Octopus
¿Forma parte de una plataforma integrada de inteligencia de negocios?	Sí	No	Sí	Sí	No	No
¿La herramienta de extracción posee una interface grafica de uso?	Sí	No	Sí	Sí	No	Sí
¿Soporta diversos tipos de bases de datos?	No	Sí	No	Sí	Sí	Sí
¿Permite cargas desde ficheros excel, xml y planos?	Sí	No	Sí	Sí	Sí	Sí
¿Requiere una fácil instalación de la herramienta?	Sí	Sí	Sí	No	No	No
¿La plataforma posee una herramienta de explotación, herramientas de *reporting*, herramientas de consultas y análisis?	Sí	No	Sí	Sí	No	No
¿Se encuentra fácilmente consultoras de sistemas para la herramienta?	Sí	Sí	Sí	No	No	No

Fuente: (Rodriguez & Mendoza, 2011)

- Instalación del motor de base de datos y de la herramienta de inteligencia de negocios
    - Motor de base de datos

      Una vez descargado el instalador del SQL Server 2014, iniciar sesión en el servidor virtual con un usuario con privilegios de administrador, cerrar todas las aplicaciones y ejecutar, nos mostrara el asistente de instalación (figura 28).

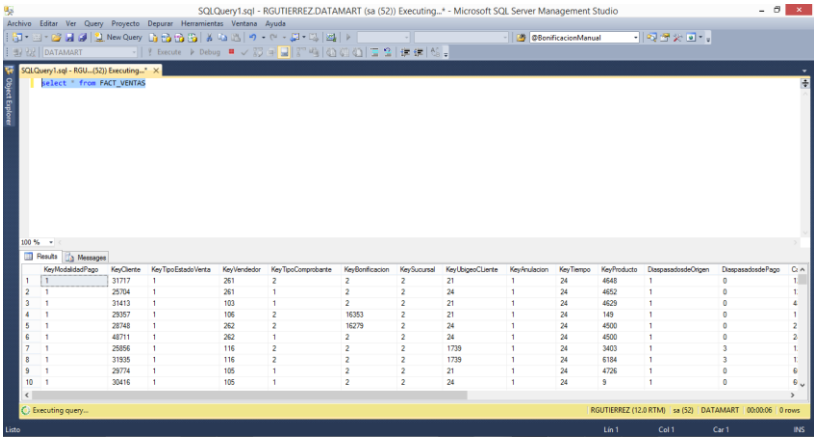
Figura 28
Asistente de instalación SQL Server 2014



Fuente: Propia

Seleccionar los servicios de motor de base de datos, configurar la instancia, configurar el motor de base de datos, y hacer clic en siguiente para continuar con la instalación. Cuando la instalación, haga clic en cerrar para salir del asistente de instalación y ejecutar SQL Server 2014, se mostrará el entorno de trabajo del SQL (figura 29)

Figura 29
Entorno de trabajo del SQL Server 2014

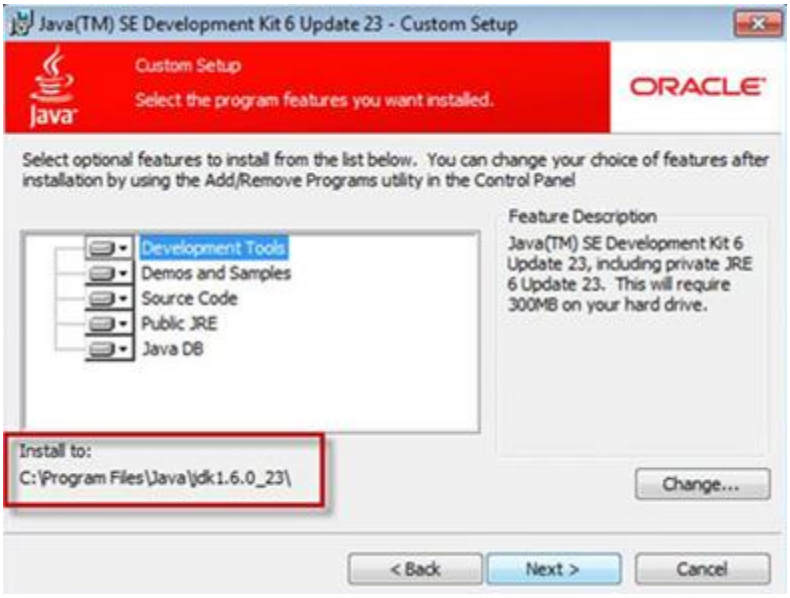


Fuente: Propia

- Herramienta de inteligencia de negocios

  Para configurar Pentaho Data Integration (PDI) primero hay que configurar el entorno Java, para ello descargamos Java ™ Developer Kit 6 Update 23 (JDK) o una versión posterior y procedemos con la instalación (figura 30)

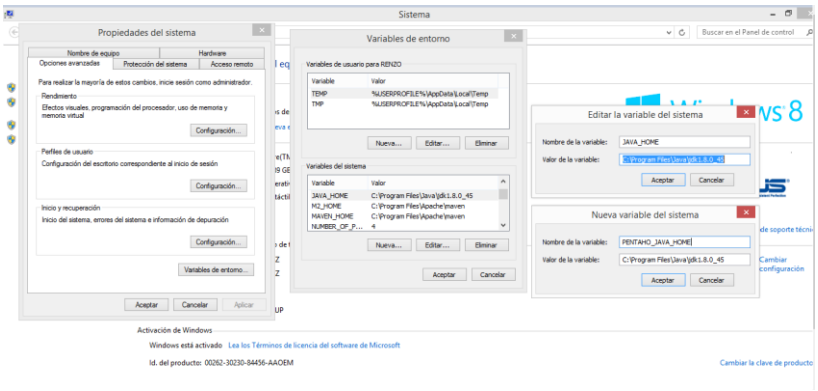
Figura 30
Entorno de instalación del JDK Java



Fuente: Propia

Posteriormente se procede a configurar las variables de entorno en el servidor virtual (figura 31) con la finalidad de poder ejecutar correctamente el PDI).

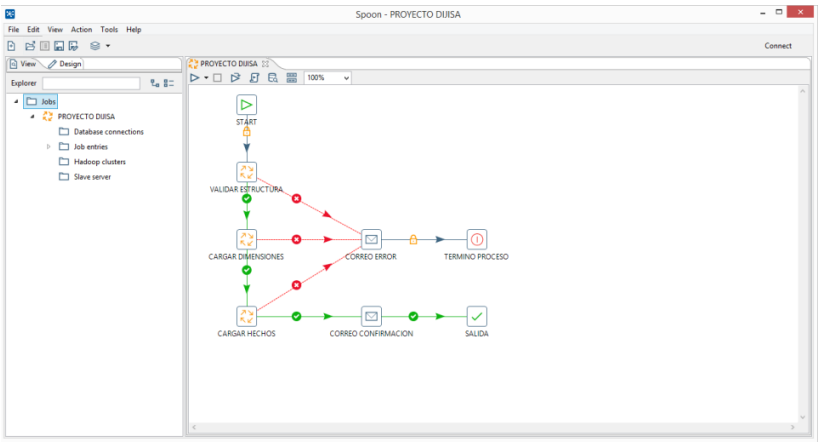
Figura 31
Configuración de variable de entorno (Java - Pentaho)



Fuente: Propia

Posteriormente se procede a descargar el PDI, para este proyecto se trabajó con la versión 7.025, una vez descargado descomprimir y ejecutar el archivo spoon.bat para ejecutar el PDI, se mostrará el entorno de trabajo (figura 32).

Figura 32
Entorno de trabajo de Pentaho Data Integration (PDI)

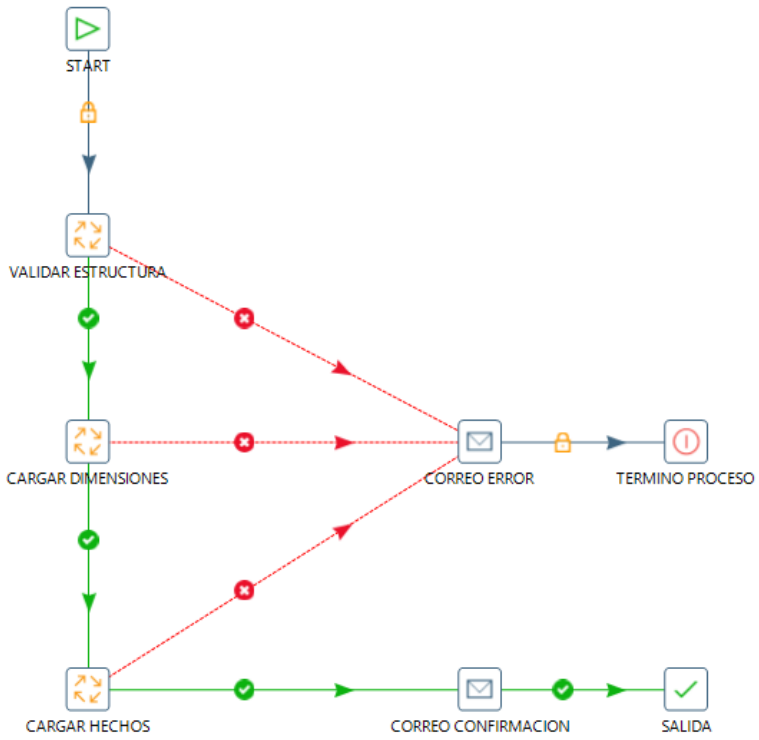


Fuente: Propia

- Diseño e implementación del subsistema ETL
    - Esquema de integración de datos

      Está diseñado en Pentaho Data Integration (PDI) el cual puede programarse para su ejecución, así mismo permite enviar correo de confirmación tras la finalización correcta del proceso ETL o correo de error cuando el proceso de ETL se ve interrumpido.

Figura 33
Esquema de integración de datos

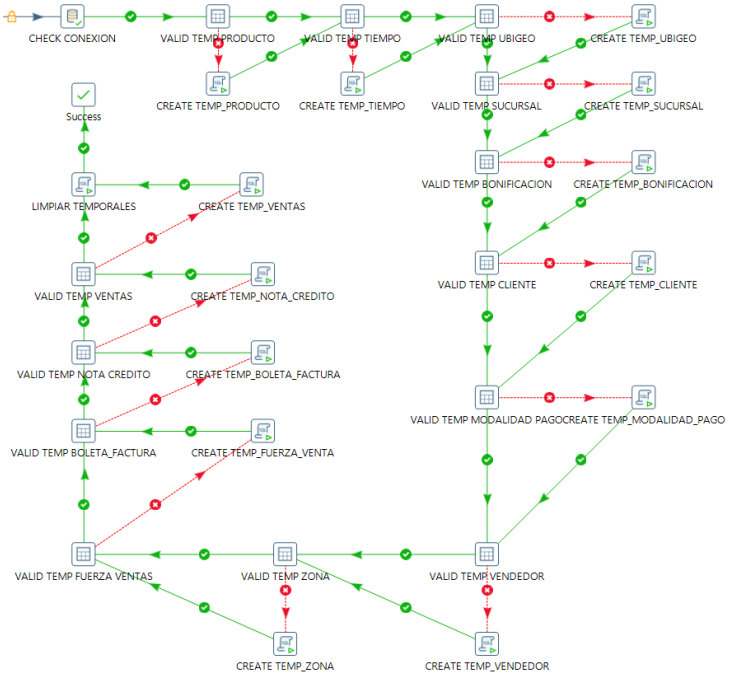


Fuente: Propia

    - Esquema de validación de estructura

      Hace una revisión de la base de datos y las tablas y si no hay respuesta o no existe procede a crearlas, para posteriormente relacionar y limpiar las tablas.

Figura 34
Esquema de validación de estructura

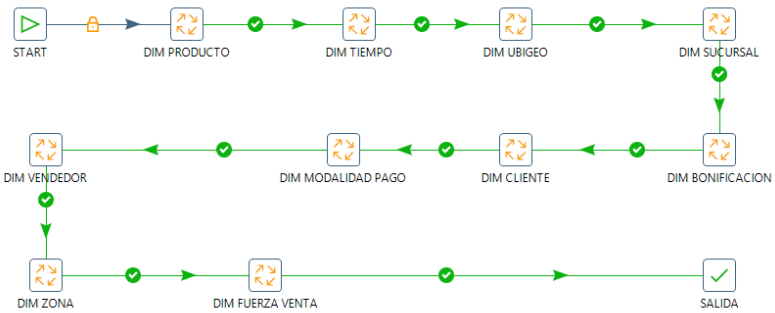


Fuente: Propia

- Esquema de carga de dimensiones

  Inicia con la configuración de conexión a la base de datos transaccional y base de datos de repositorio para la extracción y tratamiento de los datos (validación de campos null, operaciones con campos texto, recortes de texto, concatenar texto) y posterior registro en el repositorio.

Figura 35
Esquema de carga de dimensiones

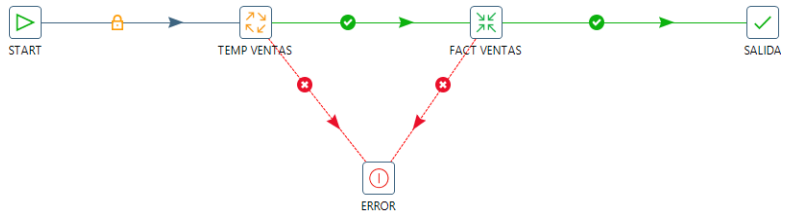


Fuente: Propia

- Esquema de carga de hechos

  Inicia con la configuración de conexión a la base de datos transaccional y base de datos de repositorio para la extracción y tratamiento de los datos (validación de campos null, operaciones con campos texto, recortes de texto y concatenar texto) para posteriormente realizar uniones con las dimensiones y terminar haciendo filtros de los campos a registrar en el repositorio.

Figura 36
Esquema general de carga de hechos



Fuente: Propia

B) Descripción de los datos

Los datos se encuentran almacenados en un almacén de datos con un esquema dimensional estrella (ver figura 25), y será fuente de la vista con la que se trabajará en el proyecto de minería de datos.

A continuación, se describirá las tablas que conforman la vista y se detalla con los atributos que se van usar:

- Tabla modalidad de pago

  Detalla información sobre las modalidades de pago con la que se puede emitir un comprobante. La tabla cuenta con 2 registros (contado-crédito), a continuación, se describe los atributos que posee la tabla.

Tabla 12
Estructura tabla Modalidad de Pago

TABLA	**MODALIDAD_PAGO**				
columna	*tipo de dato*	*tam.*	*key*	*null*	*descripción*
KeyModalidad Pago	int		PK	N	Código principal
TipoModalidad Pago	nvarchar	50		N	Nombre del tipo de modalidad de pago

Fuente: Propia

- Tabla cliente

Detalla información del cliente a quien se le vende, desde que tipo de cliente (natural o jurídica) hasta la categoría de cliente y giro de negocio. La tabla cuenta con 26049 registros, a continuación, se describe los atributos que posee la tabla.

Tabla 13
Estructura dimensión Cliente

TABLA	**CLIENTE**				
columna	*tipo de dato*	*tam.*	*key*	*null*	*descripción*
KeyCliente	int		PK	N	Código principal
CodCliente	nvarchar	20		N	Código de cliente
RazonSocial	nvarchar	100		N	Nombre del cliente
CodTipo Cliente	nvarchar	10		N	Código de tipo de cliente
TipoCliente	nvarchar	50		N	Nombre del tipo de cliente (bodega, mercado, etc.)

Fuente: Propia

- Tabla ubigeo cliente

  Detalla información del ubigeo del cliente a quien se le vende. La tabla cuenta con 57 registros, a continuación, se describe los atributos que posee la tabla.

Tabla 14
Estructura tabla Ubigeo

TABLA	**UBIGEO**				
columna	*tipo de dato*	*tam.*	*key*	*null*	*descripción*
KeyUbigeo Cliente	int		PK	N	Código principal
Distrito	nvarchar	50		N	Nombre del distrito
Provincia	nvarchar	50		N	Nombre de la provincia relacionada al distrito

Fuente: Propia

- Tabla tiempo

  Detalla información de la fecha que se realizó la venta, en su forma desagregada. La tabla cuenta con 916 registros, a continuación, se describe los atributos que posee la tabla.

Tabla 15
Estructura dimensión tiempo

TABLA	**TIEMPO**				
columna	*tipo de dato*	*tam.*	*key*	*null*	*descripción*
KeyTiempo	int		PK	N	Código principal
Fecha	date			N	Fecha de generación de documento
Año	nvarchar	10		N	Año de la fecha

Fuente: Propia

- Tabla producto

  Detalla información de los productos que se realizó la venta, en su forma desagregada, desde la marca, línea hasta el proveedor del producto. La tabla cuenta con 583 registros, a continuación, se describe los atributos que posee la tabla.

Tabla 16
Estructura dimensión producto

TABLA	**PRODUCTO**				
columna	*tipo de dato*	*tam.*	*key*	*null*	*descripción*
KeyProducto	int		PK	N	Código principal
CodProducto	nvarchar	20		N	Código de producto
Producto	nvarchar	100		N	Nombre del producto
LineaProducto	nvarchar	100		N	nombre de la línea del producto
MarcaProducto	nvarchar	100		N	Nombre de la marca del producto

Fuente: Propia

- Tabla ventas

  Detalla información de las ventas y métricas. La tabla cuenta con 5686333 registros, a continuación, se describe los atributos que posee la tabla.

Tabla 17
Estructura tabla de hechos ventas

TABLA	**VENTAS**				
columna	*tipo de dato*	*tam.*	*key*	*null*	*descripción*
KeyModalidad Pago	int		FK	N	Código principal modalidad de pago
KeyCliente	int		FK	N	Código principal cliente
KeyUbigeo	int		FK	N	Código principal

				ubigeo cliente
KeyTiempo	int	FK	N	Código principal tiempo
KeyProducto	Int	FK	N	Código principal producto
Cantidad	int		N	Cantidad de ítems comprados.
Soles	decimal (18,4)		N	Valor Venta total neto de compra sin IGV

Fuente: Propia

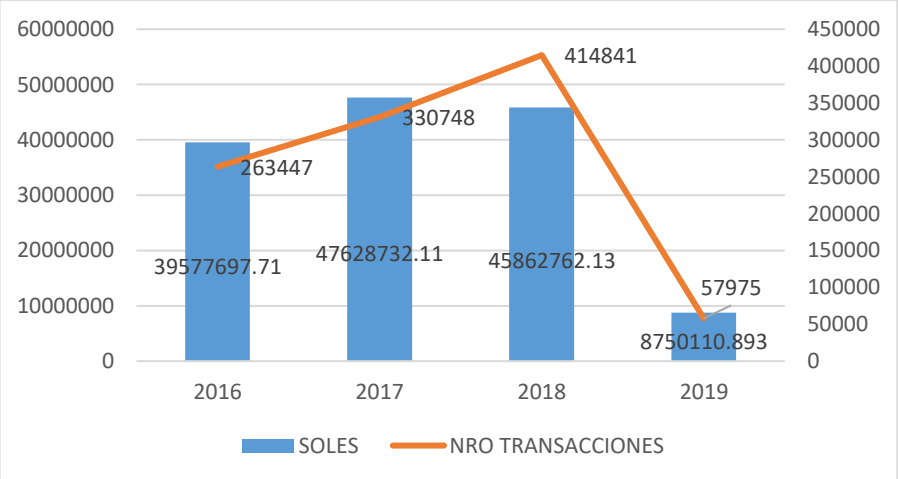
C) Exploración de los datos

Permite determinar la consistencia y completitud de los datos.

- Ventas vs número de transacciones por año

  La figura 37 refleja que los ingresos no tienen relación directa con el número de transacciones, tal como se refleja en el 2018 cuando los ingresos disminuyeron respecto al 2017 en un 4%, sin embargo, el número de transacciones aumento en un 25%.

Figura 37
Ventas vs transacciones - agrupadas por año

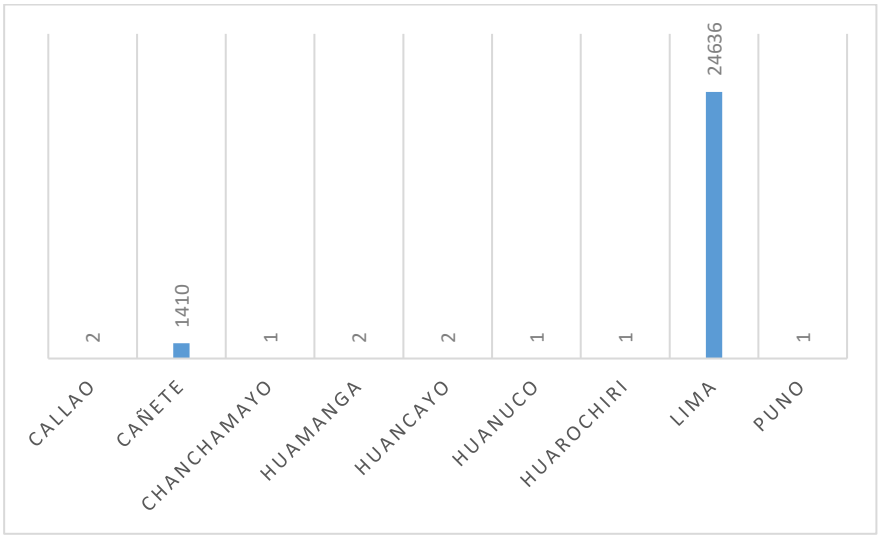


Fuente: Propia

- Clientes por provincia

El 94.53% de los clientes finales pertenecen a la provincia de Lima, el resto a otras provincias del país.

Figura 38
Número de clientes por provincia

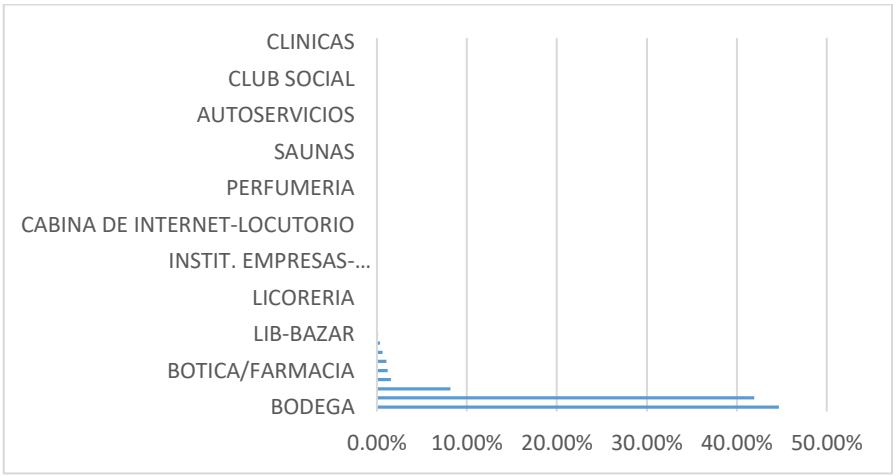


Fuente: Propia

- Ventas por tipo de negocio

  El 44.66% de las ventas es realizada por las bodegas seguido de los puestos de mercado y mayoristas, adicionalmente se visualiza que hay tipos de negocio que generan ingresos menores a 1%.

Figura 39
Porcentaje de ventas por tipo de negocio



Fuente: Propia

D) Verificación de la calidad de los datos

Después de explorar los datos se puede afirmar que los datos están completos, debido a que la información parte de un datamart el cual ha sido diseñado como parte del proyecto, en donde los valores nulos y/o errores se enmendaron y/o estandarizaron en el proceso ETL durante el desarrollo del mismo.

Al realizar un análisis de las ventas por provincia se detectó registros de ventas en las provincias como Cañete y Puno desde la sucursal de Lima, pero al ser ínfimos en comparación con los registros de la provincia Lima serán registros ignorados dentro del proceso de minería de datos.

Al realizar un análisis de ingreso de ventas por tipo de negocio se descubrió que los tipos de negocio como: bodega, botica/farmacia, minimarket, otros y puesto de mercado son los más significativos por lo que se realizará una discretización de los datos en el apartado siguiente para ser considerados.

3.1.3. PREPARACIÓN DE LOS DATOS

Implica adaptar los datos recolectados para prepararlos para aplicar técnicas de minería de datos (limpieza de datos, transformación de datos, relación de variables, número de registros, etc.).

A) Selección de datos

En términos de volumen de datos, se van a utilizar casi todos los registros contemplados en la vista, debido a que se filtrara el atributo Provincia, con la finalidad de no contar con outliers, así se alinearán los registros a los objetivos de minería de datos contemplados en el paso 1: Comprensión del negocio.

Los campos seleccionados para el análisis son los siguientes:

- Ventas

  KeyModalidadPago, KeyCliente, KeyProducto, KeyUbigeo, KeyTiempo, Cantidad, Soles

- Modalidad pago

  KeyModalidadPago, TipoModalidadPago

- Cliente

  KeyCliente, CodCliente, TipoCliente

- Ubigeo

  KeyUbigeoCliente, Distrito

- Tiempo

  KeyTiempo, Fecha

- Producto

  KeyTiempo, CodProducto, LineaProducto, MarcaProducto

Aplicando los filtros (verificación de datos) tenemos como resultado el tamaño inicial del conjunto de datos seleccionados (ver tabla 18)

Tabla 18
Resumen de datos iniciales-fuente de Dataming

Fuente de datos	2016-2019
Número de clientes	24636
Número de transacciones	1020337

Fuente: Propia

B) Limpieza de los datos

La fuente de información (dataset) con la que se cuenta para el proyecto de minería de datos está alineada a los objetivos de minería de datos y es alimentada por el datamart de ventas, en el cual durante su desarrollo se realizó la transformación de datos (limpieza) con la finalidad de estandarizar los registros, por lo que no hay necesidad de hacer una limpieza más profunda sobre los registros.

C) Estructuración de los datos

Con la finalidad conocer el comportamiento del cliente y la segmentación del mismo se implantará el método RFM, para ello se crearon nuevas variables a partir de las variables fecha, número de transacciones (Venta única por día y cliente) y soles:

- Recencia: Días transcurridos desde la última compra, comparando con la fecha que se empezó el proyecto (28/02/2019)
- Frecuencia: Número de transacciones en promedio generadas en el tiempo de estudio.
- Monto: Valor de las compras totales realizadas por el cliente en el tiempo de estudio.

El cálculo de las nuevas variables se realizó en R para ello se procedió a cargar los datos en la variable RFMCliente y se ejecutó la siguiente consulta:

```
#cargar los registros la variable dataset para procesarlos

library(RODBC)

sql<-"select * from resumenVentasFiltrado"

cn<-odbcConnect("local",uid="sa",pwd="d1j1s@..")

dataset<-sqlQuery(cn,sql)
```

```
#asegurase que la variable fecha es tipo date

dataset$Fecha<-as.Date(dataset$Fecha,format="%d/%m/%y" )

#obtener los usuarios unicos

RFMCliente <- with( dataset, data.frame( CodCliente = sort(unique(CodCliente))))

#añadimos la columna recencia

RFMCliente <- cbind(RFMCliente,recencia = aggregate( round( as.numeric(
difftime(as.Date("2019-02-28"), dataset$Fecha, units="days")) ),
list(dataset$CodCliente), min )$x)

#añadimos la columna recencia

RFMCliente <- cbind(RFMCliente,frecuencia = with( dataset, as.numeric( by(
Fecha, CodCliente, function(x) length(unique(x)) ) ) ))

#añadimos la columna monto

RFMCliente <- cbind(RFMCliente,monto = with( dataset,
as.numeric(by(Soles,CodCliente,sum) )))
```

La tabla 19 muestra los 10 primeros registros generados.

Tabla 19
Clientes con variables RFM

CodigoCliente	recencia	frecuencia	monto
8910	1	2	274.782
35394	3	8	972.6198
37003	236	1	36.8496
59561	223	3	157.3063
61783	1049	1	34.1836
64433	0	39	3076.8171
92391	1	96	14904.5038
112572	10	9	463.5817
115026	910	15	713.3131
119605	8	7	218.8858

Fuente: Propia

Por un lado, las variables como Distrito, TipoModalidadPago y TipoCliente servirán para caracterizar los grupos generados, mientras que las variables como CodProducto, LineaProducto y

MarcaProducto servirán para generar asociaciones dentro de los grupos para realizar sugeridos de venta.

Así mismo se realizó la discretización las siguientes variables:

- Distrito: Se realizó la agrupación de los distritos según su ubicación:

Tabla 20
Discretización del campo distrito

UBICACIÓN	DESCRIPCIÓN
BALNEAREO	PUCUSANA, PUNTA HERMOSA, PUNTA NEGRA, SAN BARTOLO, SANTA MARIA DEL MAR
CENTRO	BARRANCO, BREÑA, JESUS MARIA, LA VICTORIA, LIMA, LINCE, MAGDALENA DEL MAR, MIRAFLORES, PUEBLO LIBRE, RIMAC, SAN BORJA, SAN ISIDRO, SAN LUIS, SAN MIGUEL, SANTIAGO DE SURCO, SURQUILLO
ESTE	ATE, CIENEGUILLA, EL AGUSTINO, LA MOLINA, SAN JUAN DE LURIGANCHO, SANTA ANITA
NORTE	COMAS, LOS OLIVOS, PUENTE PIEDRA, SAN MARTIN DE PORRES
SUR	CHORRILLOS, LURIN, PACHACAMAC, SAN JUAN DE MIRAFLORES, VILLA EL SALVADOR, VILLA MARIA DEL TRIUNFO

Fuente: Propia

- TipoCliente: Se realizó la agrupación de los tipos de negocio: agrupándolas según los ingresos generados

Tabla 21
Discretización del campo tipo de cliente

TIPOCLIENTE	DESCRIPCIÓN
BODEGA	BODEGA
MERCADO	PTO-MERCADO
MAYORISTA	MAYORISTA
BOTICA	BOTICA/FARMACIA
PANADERIA	PANADERIA-PASTELERÍA
OTROS	OTROS, LOCUTORIO, UNVERSIDADES, COLEGIOS, ETC.

Fuente: Propia

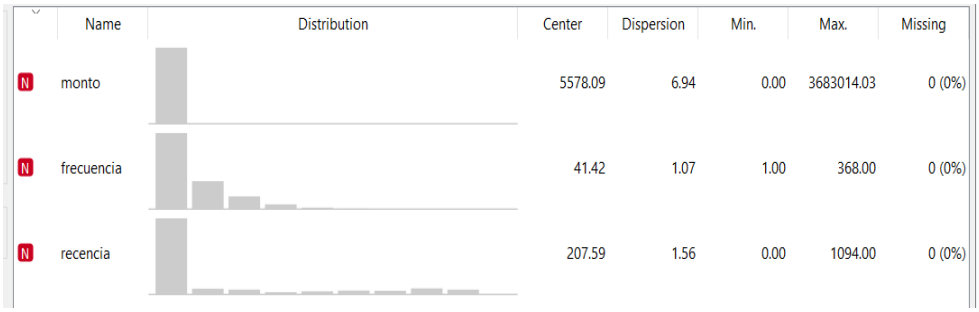
D) Integración de los datos

Se realizó la fusión del conjunto de datos del cliente: CodCliente con las nuevas variables creadas: Recencia, Frecuencia y Monto, a través de R, los primeros resultados se muestran en la tabla 19.

E) Formateo de los datos

Revisando los valores de recencia, frecuencia y monto tenemos que todas las variables tienen una distribución con un sesgo a la izquierda (ver figura 40):

Figura 40
Estadísticos RFM



Fuente: Propia

Es por ello que los registros se normalizaron (ver tabla 22) en base al conocimiento y experiencia de los expertos del negocio.

Tabla 22
Normalización de variables RFM

Escala	Nombre de escala	Recencia (días)	Frecuencia	Monto (soles)
5 ptos.	Muy alto	[0,3]	[72, a más]	[5556.89, a más]
4 ptos.	Alto	[4,13]	[40,71]	[2486.04,5556.88]
3 ptos.	Medio	[14,64]	[16,39]	[937.09,2486.03]
2 ptos.	Bajo	[65-501]	[4,15]	[244.93,937.08]
1 pto.	Muy Bajo	[502, a más]	[0,3]	[0,244.92]

Fuente: Propia

- Puntuación Recencia: Las puntuaciones más altas indican las transacciones más recientes
- Puntuación Frecuencia: Las puntuaciones más altas indican mayor número de transacciones.
- Puntuación Monto: Las puntuaciones más altas indican mayor valor monetario.

3.1.4. MODELAMIENTO

Implica seleccionar las técnicas de modelado más apropiadas para el proyecto de minería de datos.

A) Selección de la técnica de modelado

Debido a que el objetivo del negocio es entender el comportamiento del cliente, la técnica que más se adecua es el de clustering utilizando K-means++, así mismo se usa el algoritmo LEM2 para ver el nivel de relación entre las variables nominales y los clústeres encontrados, y por último el algoritmo a priori para determinar las relaciones entre los productos para realizar sugeridos entre los grupos encontrados.

B) Generación del plan de prueba.

El procedimiento que se empleará para probar la calidad y validez del modelo una vez generado será a través de la validación interna, en el caso del algoritmo K-means, debido a que no se cuenta con información externa para su validación.

La validación interna evalúa que tan buena es la estructura del clustering sin necesidad de información ajena al propio algoritmo y u resultado, basándonos en la suma de cuadrados intragrupos y suma de error al cuadrado.

En el caso del Algoritmo LEM2 el dataset se dividirá en 2 bloques (data de entrenamiento (60%) y data de prueba (40%).

Y, por último, el caso del algoritmo a priori considera el nivel de confianza y soporte y se evalúa a través del criterio Lift y Loevinger.

C) Construcción y evaluación del modelo

- Instalación de R

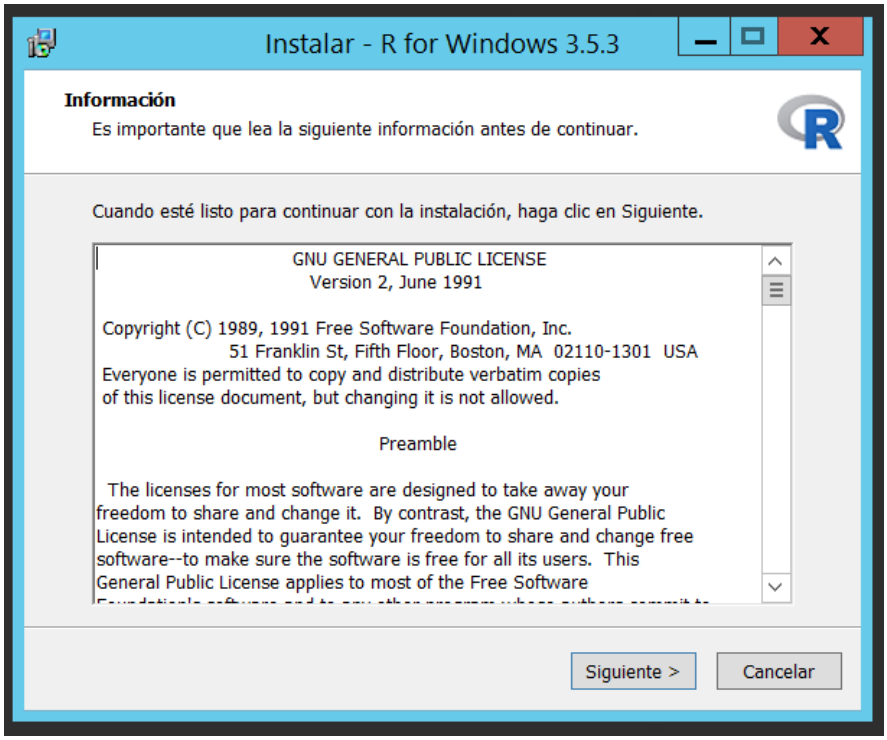
  Ingresar a https://cran.r-project.org/bin/windows/base/ y descargar la versión para Windows (ver figura 41).

Figura 41
Página Web - descargar R

Fuente: Propia

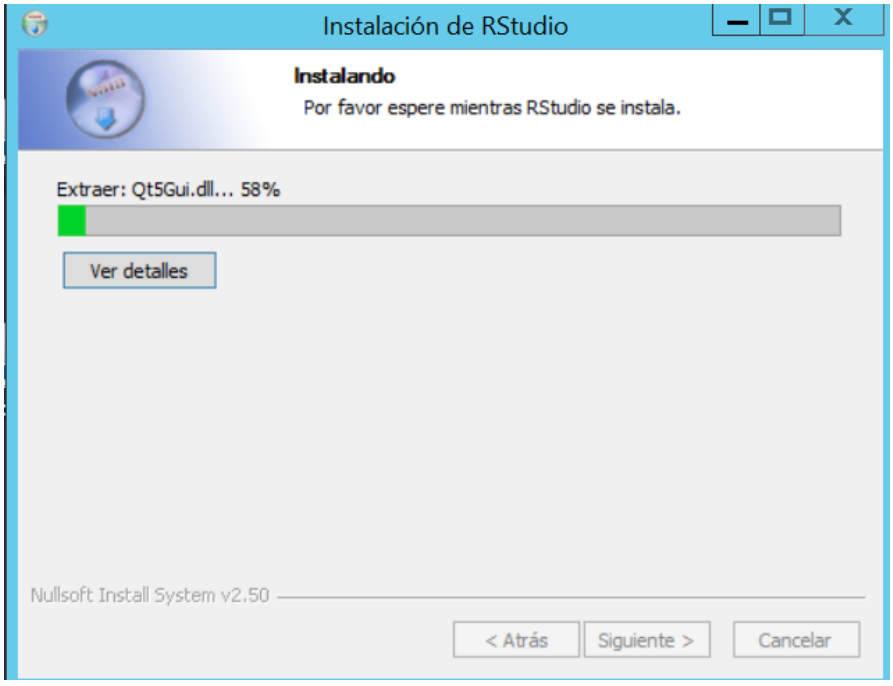
Descargar y ejecutar el instalador de R, (ver figura 42 Y 43)

Figura 42
Instalación R-Core



Fuente: Propia

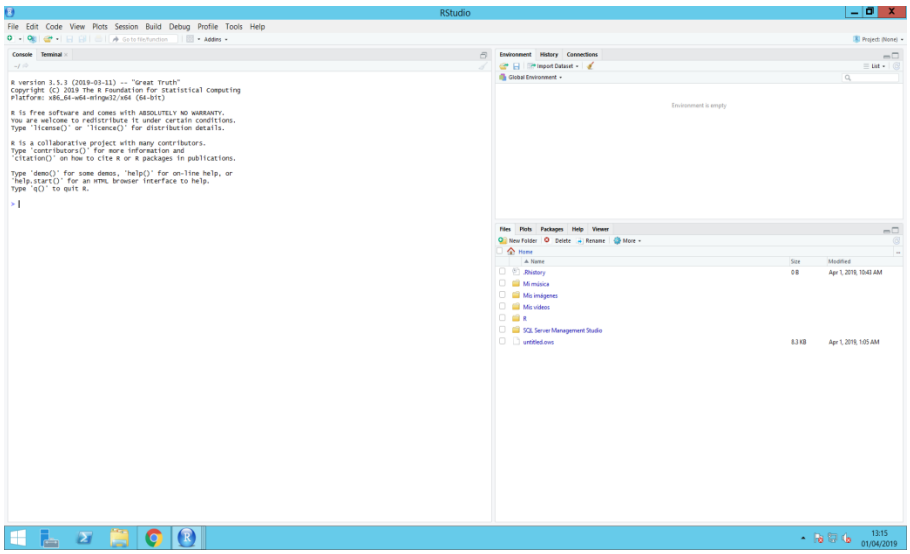
Figura 43
Instalación RStudio



Fuente: Propia

Una vez instalado R y RStudio se mostrará la interfaz de trabajo (ver figura 44)

Figura 44
Interfaz de trabajo RStudio



Fuente: Propia

- Configuraciones iniciales

Dentro de las configuraciones iniciales esta la instalación de los paquetes con los cuales se va a trabajar, a través de las siguientes consultas.

```
install.packages("RODBC")

install.packages("RoughSets")

install.packages("arules")

install.packages("arulesViz")

install.packages("Rppc")
```

Una vez configurada las librerías nos conectamos a la base de datos para recuperar información, el cual será el dataset del proyecto, con la siguiente consulta:

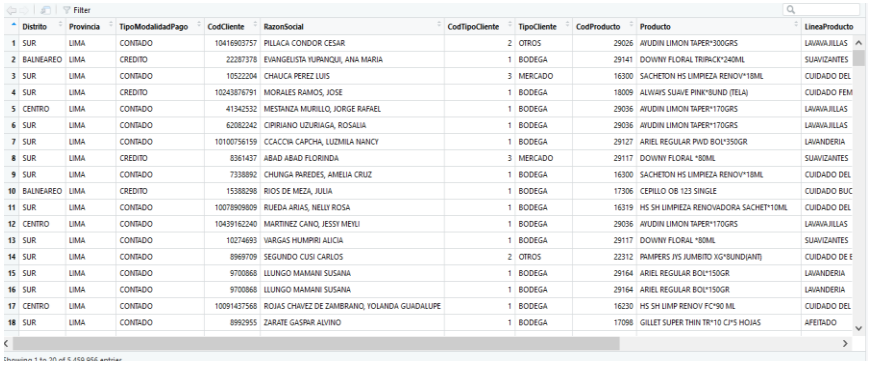
```
#cargar los registros la variable dataset para procesarlos

library(RODBC)
```

```
sql<-"select * from resumenVentasFiltrado "

cn<-odbcConnect("local",uid="sa",pwd="d1j1s@..")

dataset<-sqlQuery(cn,sql)
```

Para poder visualizar los registros (ver figura 45) se ejecutó la siguiente consulta:

```
View(dataset)
```

Figura 45
Visualización de registros en RStudio



	Distrito	Provincia	TipoModalidadPago	CodCliente	RazonSocial	CodTipoCliente	TipoCliente	CodProducto	Producto	LineaProducto
1	SUR	LIMA	CONTADO	10416903757	PILLACA CONDOR CESAR	2	OTROS	29026	AYUDIN LIMON TAPER*300GRS	LAVAVAJILLAS
2	BALNEAREO	LIMA	CREDITO	22287378	EVANGELISTA YUPANQUI, ANA MARIA	1	BODEGA	29141	DOWNY FLORAL TRIPACK*240ML	SUAVIZANTES
3	SUR	LIMA	CONTADO	10522204	CHAUCA PEREZ LUIS	3	MERCADO	16300	SACHETON HS LIMPIEZA RENOV*18ML	CUIDADO DEL
4	SUR	LIMA	CREDITO	10243876791	MORALES RAMOS, JOSE	1	BODEGA	18009	ALWAYS SUAVE PINK*BUND (TELA)	CUIDADO FEM
5	CENTRO	LIMA	CONTADO	41342532	MESTANZA MURILLO, JORGE RAFAEL	1	BODEGA	29036	AYUDIN LIMON TAPER*170GRS	LAVAVAJILLAS
6	SUR	LIMA	CONTADO	62082242	CIPIRIANO UZURIAGA, ROSALIA	1	BODEGA	29036	AYUDIN LIMON TAPER*170GRS	LAVAVAJILLAS
7	SUR	LIMA	CONTADO	10100756159	CCACCIA CAPCHA, LUZMILA NANCY	1	BODEGA	29127	ARIEL REGULAR PWD BOL*350GR	LAVANDERIA
8	SUR	LIMA	CREDITO	8361437	ABAD ABAD FLORINDA	3	MERCADO	29117	DOWNY FLORAL *80ML	SUAVIZANTES
9	SUR	LIMA	CONTADO	7338892	CHUINGA PAREDES, AMELIA CRUZ	1	BODEGA	16300	SACHETON HS LIMPIEZA RENOV*18ML	CUIDADO DEL
10	BALNEAREO	LIMA	CREDITO	15388298	RIOS DE MEZA, JUUA	1	BODEGA	17306	CEPILLO OB 123 SINGLE	CUIDADO BUC
11	SUR	LIMA	CONTADO	10078909809	RUEDA ARIAS, NELLY ROSA	1	BODEGA	16319	HS SH LIMPIEZA RENOVADORA SACHET*10ML	CUIDADO DEL
12	CENTRO	LIMA	CONTADO	10439162240	MARTINEZ CANO, JESSY MEYLI	1	BODEGA	29036	AYUDIN LIMON TAPER*170GRS	LAVAVAJILLAS
13	SUR	LIMA	CONTADO	10274693	VARGAS HUMPIRI ALICIA	1	BODEGA	29117	DOWNY FLORAL *80ML	SUAVIZANTES
14	SUR	LIMA	CONTADO	8969709	SEGUNDO CUSI CARLOS	2	OTROS	22312	PAMPERS JYS JUMBITO XG*BUND(ANT)	CUIDADO DE E
15	SUR	LIMA	CONTADO	9700868	LLUNGO MAMANI SUSANA	1	BODEGA	29164	ARIEL REGULAR BOL*150GR	LAVANDERIA
16	SUR	LIMA	CONTADO	9700868	LLUNGO MAMANI SUSANA	1	BODEGA	29164	ARIEL REGULAR BOL*150GR	LAVANDERIA
17	CENTRO	LIMA	CONTADO	10091437568	ROJAS CHAVEZ DE ZAMBRANO, YOLANDA GUADALUPE	1	BODEGA	16230	HS SH LIMP RENOV FC*90 ML	CUIDADO DEL
18	SUR	LIMA	CONTADO	8992955	ZARATE GASPAR ALVINO	1	BODEGA	17098	GILLET SUPER THIN TR*10 CJ*5 HOJAS	AFEITADO

Fuente: Propia

- Algoritmo K-MEANS

Para trabajar con el algoritmo k.means, debemos seleccionar los parámetros RFM los cuales serán analizados para formar los clúster.

Para ello revisamos el punto de preparación de datos: estructuración de datos, donde se realizó el cálculo de las variables recencia, frecuencia y monto.

Teniendo en consideración el punto preparación de datos: formateo de datos, se procedió a normalizar las variables según la tabla 22, a través de la siguiente consulta:

```
#normalizando valores RFM

#variable recencia

rango5r<-which(RFMClienteNormalizado$recencia>=502)
```

```
rango4r<-which(RFMClienteNormalizado$recencia>=65 &
RFMClienteNormalizado$recencia<502)

rango3r<-which(RFMClienteNormalizado$recencia>=14 &
RFMClienteNormalizado$recencia<65)

rango2r<-which(RFMClienteNormalizado$recencia>=4 &
RFMClienteNormalizado$recencia<14)

rango1r<-which(RFMClienteNormalizado$recencia>=0 &
RFMClienteNormalizado$recencia<4)

#asignación de valores recencia

RFMClienteNormalizado[rango5r,"recencia"]<-1

RFMClienteNormalizado[rango4r,"recencia"]<-2

RFMClienteNormalizado[rango3r,"recencia"]<-3

RFMClienteNormalizado[rango2r,"recencia"]<-4

RFMClienteNormalizado[rango1r,"recencia"]<-5

#variable frecuencia

rango5f<-which(RFMClienteNormalizado$frecuencia>=72)

rango4f<-which(RFMClienteNormalizado$frecuencia>=40 &
RFMClienteNormalizado$frecuencia<72)

rango3f<-which(RFMClienteNormalizado$frecuencia>=16 &
RFMClienteNormalizado$frecuencia<40)

rango2f<-which(RFMClienteNormalizado$frecuencia>=4 &
RFMClienteNormalizado$frecuencia<16)

rango1f<-which(RFMClienteNormalizado$frecuencia>=0 &
RFMClienteNormalizado$frecuencia<4)

#asignación de valores frecuencia

RFMClienteNormalizado[rango5f,"frecuencia"]<-5

RFMClienteNormalizado[rango4f,"frecuencia"]<-4

RFMClienteNormalizado[rango3f,"frecuencia"]<-3

RFMClienteNormalizado[rango2f,"frecuencia"]<-2

RFMClienteNormalizado[rango1f,"frecuencia"]<-1
```

```
#variable monto

rango5m<-which(RFMClienteNormalizado$monto>=5556.89)

rango4m<-which(RFMClienteNormalizado$monto>=2486.04 &
RFMClienteNormalizado$monto<5556.89)

rango3m<-which(RFMClienteNormalizado$monto>=937.09 &
RFMClienteNormalizado$monto<2486.04)

rango2m<-which(RFMClienteNormalizado$monto>=244.93 &
RFMClienteNormalizado$monto<937.09)

rango1m<-which(RFMClienteNormalizado$monto>=0 &
RFMClienteNormalizado$monto<244.93)

#asignación de valores monto

RFMClienteNormalizado[rango5m,"monto"]<-5

RFMClienteNormalizado[rango4m,"monto"]<-4

RFMClienteNormalizado[rango3m,"monto"]<-3

RFMClienteNormalizado[rango2m,"monto"]<-2

RFMClienteNormalizado[rango1m,"monto"]<-1
```

La tabla 23 muestra los 10 primeros resultados.

Tabla 23
Registros RFM normalizados

CodigoCliente	recencia	frecuencia	monto
8910	5	1	2
35394	5	2	3
37003	2	1	1
59561	2	1	1
61783	1	1	1
64433	5	3	4
92391	5	5	5
112572	4	2	2
115026	1	2	2
119605	4	2	1

Fuente: Propia

Para hallar el número de clúster óptimo se usaron 2 validadores distintos:

- Suma de cuadrados intragrupos

  Se ejecuto la siguiente consulta:

  set.seed(80)

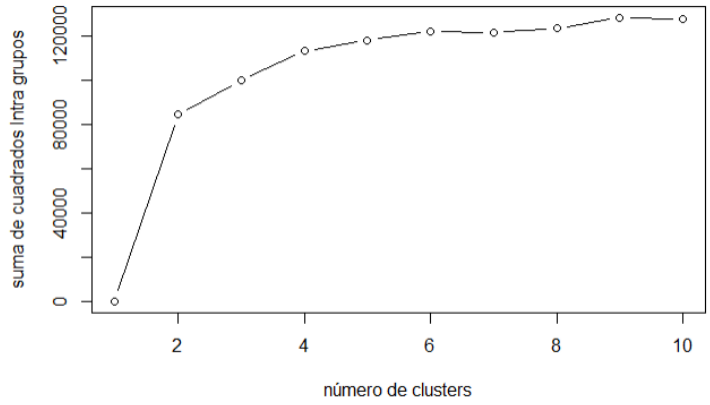
  sumbt<-kmeans(RFM, centers = 1)$betweenss

  for(i in 2:10) sumbt[i] <- kmeans(RFM, centers = i)$betweenss

  plot(1:10, sumbt, type = "b", xlab = "número de clúster", ylab = "suma de cuadrados Intra grupos")

  View(sumbt)

Figura 46
Grafica Suma de cuadrados intragrupos



Fuente: Propia

Tabla 24
Resultados de la Suma de Error Intragrupos

Clúster	Suma de error Intragrupos
1	-2.03E-08
2	8.44E+04
3	9.99E+04
4	1.13E+05
5	1.18E+05
6	1.22E+05
7	1.22E+05
8	1.24E+05
9	1.29E+05
10	1.28E+05

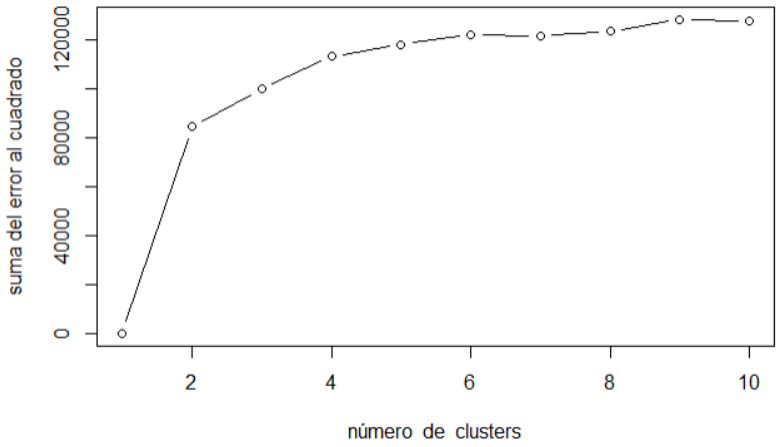
Fuente: Propia

- Curva de distorsión o Suma de error al cuadrado

  Se ejecuto la siguiente consulta:

  ```
  set.seed(80)

  wss<-(nrow(RFM)-1)*sum(apply(RFM ,2,var))

  for(i in 2:10) {wss[i] <- sum(kmeans(RFM, centers = i,nstart =25)$withinss)}

  plot(1:10, sumbt, type = "b", xlab = "número  de clúster", ylab = "suma del error al cuadrado")

  View(wss)
  ```

Figura 47
Grafica Curva de distorsión



Fuente: Propia

Tabla 25
Resultados de la Suma de Error al Cuadrado

Clúster	Suma del Error al Cuadrado
1	146482.08
2	62041.07
3	45834.74
4	32408.86
5	27445.73
6	22739.48
7	20420.32
8	18170.69
9	16673.25
10	15670.42

Fuente: Propia

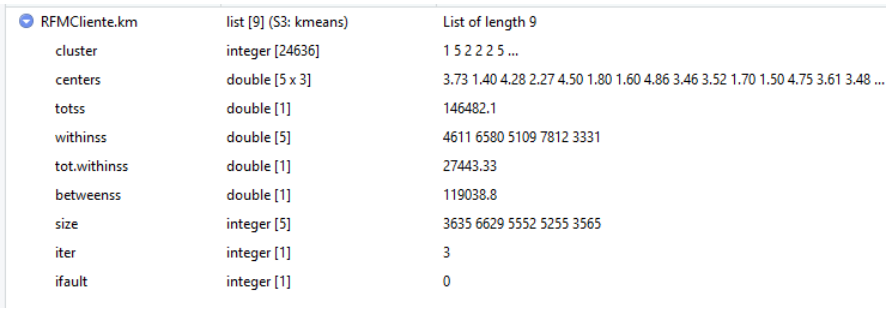
Luego de evaluar las gráficas se observa que el K (número de clúster) optimo es 5 debido a que hay una variación suave.

Ahora, vamos a construir el modelo para K =5 y nstart=25.

Se ejecuto la siguiente consulta:

```
set.seed(80)

RFMCliente.km <- kmeans(RFM, centers = 5,nstart = 25)
```

Figura 48
Resultados Kmeans



Fuente: Propia

A continuación, en la tabla 26 se muestran los resultados de la segmentación a través de clúster

Tabla 26
Resultados de clústeres

Clúster	Recencia	Frecuencia	Monto
1	3.727923	1.800825	1.703989
2	1.400060	1.602353	1.498265
3	4.281340	4.855548	4.753242
4	2.272502	3.455947	3.614462
5	4.502384	3.515568	3.478261

Fuente: Propia

Para ver que clúster es más representativo para la empresa se realizó el cálculo de la distancia al punto cero, obteniendo los resultados de la tabla 27, considerando que a mayor distancia del punto cero el clúster se vuelve más representativo, la fórmula para el cálculo es la siguiente:

```
centros<-aggregate(RFM ,by = list(RFMCliente.km$cluster), mean)

for(i in c(1:NROW(centros))){

centros$distanciaCero[i]<-sqrt(sum((centros$recencia[i]-
0)^2,(centros$frecuencia[i]-0)^2,(centros$monto[i]-0)^2))}
```

Tabla 27
Resultados de clústeres con Nivel de Representatividad (clientes leales)

Clúster	Recencia	Frecuencia	Monto	Distancia Cero	Nivel de fidelidad	# instancias
1	3.727923	1.800825	1.703989	4.477048	Baja	3635
2	1.400060	1.602353	1.498265	2.602403	Muy Baja	6629
3	4.281340	4.855548	4.753242	8.031160	Muy Alta	5552
4	2.272502	3.455947	3.614462	5.492920	Media	5255
5	4.502384	3.515568	3.478261	6.687973	Alta	3565

Fuente: Propia

A Continuación, asignamos los clústeres identificados al dataset RFMClienteNormalizado, a través de la siguiente consulta:

```
cl<-cbind(RFMCliente.km$cluster)

for(i in 1:length(cl)){

RFMClienteNormalizado$Agrupacion[i]<-cl[i]}
```

La figura 49 muestra los resultados de la consulta generada.

Figura 49
Dataset con asignación de Clúster

CodCliente	recencia	frecuencia	monto	Agrupacion
8910	5	1	2	1
35394	5	2	3	5
37003	2	1	1	2
59561	2	1	1	2
61783	1	1	1	2
64433	5	3	4	5
92391	5	5	5	3
112572	4	2	2	1
115026	1	2	2	2
119605	4	2	1	1
121089	2	3	2	2

Fuente: Propia

Una vez identificados los grupos se procedió a crear reglas de asociación con el algoritmo LEM2, considerando los grupos generados (clúster) como variable de decisión.

- Algoritmo LEM2

Como primer punto se procedió a categorizar las variables continuas/numéricas (RFM), según la tabla 22, y la variable de decisión, según la tabla 27, a través de la siguiente consulta:

```
#categorizar variables RFM y Agrupación

#recencia

dataset_LEM2_groupNormalizado$recencia<-
factor(dataset_LEM2_groupNormalizado$recencia)
```

```
levels(dataset_LEM2_groupNormalizado$recencia)<-c("Muy bajo","Bajo","Medio","Alto","Muy Alto")

#frecuencia

dataset_LEM2_groupNormalizado$frecuencia<-factor(dataset_LEM2_groupNormalizado$frecuencia)

levels(dataset_LEM2_groupNormalizado$frecuencia)<-c("Muy bajo","Bajo","Medio","Alto","Muy Alto")

#monto

dataset_LEM2_groupNormalizado$monto<-factor(dataset_LEM2_groupNormalizado$monto)

levels(dataset_LEM2_groupNormalizado$monto)<-c("Muy bajo","Bajo","Medio","Alto","Muy Alto")

#agrupación

dataset_LEM2_groupNormalizado$Agrupacion<-factor(dataset_LEM2_groupNormalizado$Agrupacion)

levels(dataset_LEM2_groupNormalizado$Agrupacion)<-c("Baja","Muy Baja","Muy Alta","Media","Alta")
```

Figura 50
Categorización de variables RFM y decisión

CodCliente	Distrito	TipoModalidadPago	TipoCliente	recencia	frecuencia	monto	Agrupacion
8910	BALNEAREO	CONTADO	BODEGA	Muy Alto	Muy bajo	Bajo	Baja
35394	SUR	CONTADO	BODEGA	Muy Alto	Bajo	Medio	Alta
37003	SUR	CONTADO	BODEGA	Bajo	Muy bajo	Muy bajo	Muy Baja
59561	SUR	CONTADO	BODEGA	Bajo	Muy bajo	Muy bajo	Muy Baja
61783	SUR	CONTADO	BODEGA	Muy bajo	Muy bajo	Muy bajo	Muy Baja
64433	CENTRO	CONTADO	BODEGA	Muy Alto	Medio	Alto	Alta
92391	SUR	CONTADO	MERCADO	Muy Alto	Muy Alto	Muy Alto	Muy Alta
112572	SUR	CONTADO	MERCADO	Alto	Bajo	Bajo	Baja
115026	SUR	CONTADO	BODEGA	Muy bajo	Bajo	Bajo	Muy Baja
119605	SUR	CONTADO	BODEGA	Alto	Bajo	Muy bajo	Baja

Fuente: Propia

Como segundo punto se procedió a dividir los datos en 60% por cada clúster para datos de entrenamiento y 40% por cada clúster para datos de prueba.

- Para ello se guardó los índices de cada grupo en una variable, a través de la siguiente consulta:

```
#crear las agrupaciones

agrupacion_1<-which(dataset_LEM2_groupNormalizado$Agrupacion
=="Baja")

agrupacion_2<-which(dataset_LEM2_groupNormalizado$Agrupacion
=="Muy Baja")

agrupacion_3<-which(dataset_LEM2_groupNormalizado$Agrupacion
=="Muy Alta")

agrupacion_4<-which(dataset_LEM2_groupNormalizado$Agrupacion
=="Media")

agrupacion_5<-which(dataset_LEM2_groupNormalizado$Agrupacion
=="Alta")
```

- Posteriormente se eligió aleatoriamente (60% de los datos) a través de los índices de las agrupaciones por medio de la siguiente consulta:

```
entrenamiento_1<-sample(1:length(agrupacion_1),size =
0.6*length(agrupacion_1))

entrenamiento_2<-sample(1:length(agrupacion_2),size =
0.6*length(agrupacion_2))

entrenamiento_3<-sample(1:length(agrupacion_3),size =
0.6*length(agrupacion_3))

entrenamiento_4<-sample(1:length(agrupacion_4),size =
0.6*length(agrupacion_4))

entrenamiento_5<-sample(1:length(agrupacion_5),size =
0.6*length(agrupacion_5))
```

- A continuación, se guardó los índices aleatorios escogidos en el paso anterior para agruparlos en la variable entrenamiento según la siguiente consulta:

```
#agrupar las agrupaciones de entrenamiento

entrenamiento<-
c(agrupacion_1[entrenamiento_1],agrupacion_2[entrenamiento_2],ag
rupacion_3[entrenamiento_3],agrupacion_4[entrenamiento_4],agrup
acion_5[entrenamiento_5])
```

- Para finalizar con la separación de los datos en 2 grupos (entrenamiento y prueba), a través de la siguiente consulta:

```
dataset_entrenamiento<-
dataset_LEM2_groupNormalizado[entrenamiento,]

dataset_prueba<-dataset_LEM2_groupNormalizado[-entrenamiento,]
```

Como tercer punto se procede a crear la tabla de decisión con los datos de la tabla de entrenamiento, a través de la siguiente consulta:

```
library("RoughSets")

library("Rcpp")

decision.table<-
SF.asDecisionTable(dataset=sub_dataset_entrenamiento,decision.attr =
7,indx.nominal=c(1,7))
```

Donde:

Dataset: es la fuente de entrenamiento para generar el modelo.

decisión.attr: Es la columna de variable de decisión definida al inicio, mayormente esta va en la última columna.

indx.nominal: involucra todos los atributos nominales que se van a considerar dentro de la tabla de decisión.

La figura 51 muestra la tabla de decisión generada con la data de entrenamiento.

Figura 51
Tabla de decisión

Distrito	TipoModalidadPago	TipoCliente	recencia	frecuencia	monto	Agrupacion
SUR	CONTADO	BODEGA	Alto	Bajo	Bajo	Baja
SUR	CONTADO	BODEGA	Alto	Medio	Bajo	Baja
SUR	CONTADO	BODEGA	Alto	Muy bajo	Muy bajo	Baja
SUR	CONTADO	MERCADO	Medio	Medio	Bajo	Baja
SUR	CONTADO	BODEGA	Medio	Muy bajo	Muy bajo	Baja
SUR	CONTADO	BODEGA	Alto	Bajo	Muy bajo	Baja
CENTRO	CONTADO	BODEGA	Muy Alto	Bajo	Bajo	Baja
SUR	CONTADO	BODEGA	Medio	Medio	Bajo	Baja
SUR	CONTADO	BODEGA	Alto	Bajo	Bajo	Baja
SUR	CONTADO	BODEGA	Alto	Muy bajo	Muy bajo	Baja
SUR	CONTADO	BODEGA	Medio	Medio	Bajo	Baja
SUR	CONTADO	BODEGA	Alto	Muy bajo	Muy bajo	Baja
CENTRO	CONTADO	OTROS	Alto	Bajo	Bajo	Baja
SUR	CONTADO	BODEGA	Medio	Bajo	Bajo	Baja
SUR	CONTADO	BODEGA	Alto	Muy bajo	Muy bajo	Baja
ESTE	CREDITO	BODEGA	Alto	Medio	Bajo	Baja

Fuente: Propia

Como cuarto punto se genera las reglas de inducción con la tabla de decisión usando el algoritmo LEM2, a través de la siguiente consulta:

```
rulesLEM2<-RI.LEM2Rules.RST(decision.table)
```

La figura 52 muestra la estructura de las reglas generadas por el algoritmo LEM2

Figura 52
Reglas de induccion-LEM2

```
A set consisting of  76  rules:
1. IF monto is Bajo and frecuencia is Bajo and recencia is Medio THEN  is Baja;
                (supportSize=486; laplace=0.991853360488798)
2. IF monto is Bajo and frecuencia is Bajo and recencia is Alto THEN  is Baja;
                (supportSize=320; laplace=0.987692307692308)
3. IF frecuencia is Muy bajo and recencia is Medio THEN  is Baja;
                (supportSize=331; laplace=0.988095238095238)
4. IF frecuencia is Bajo and recencia is Muy Alto and monto is Bajo THEN  is Baja;
                (supportSize=260; laplace=0.984905660377358)
5. IF frecuencia is Muy bajo and recencia is Alto THEN  is Baja;
                (supportSize=178; laplace=0.978142076502732)
6. IF monto is Muy bajo and recencia is Medio THEN  is Baja;
                (supportSize=373; laplace=0.989417989417989)
7. IF monto is Bajo and frecuencia is Medio and recencia is Medio THEN  is Baja;
                (supportSize=151; laplace=0.974358974358974)
8. IF monto is Muy bajo and recencia is Muy Alto THEN  is Baja;
                (supportSize=145; laplace=0.973333333333333)
9. IF recencia is Alto and monto is Muy bajo THEN  is Baja;
                (supportSize=226; laplace=0.982683982683983)
10. IF TipoCliente is BODEGA and recencia is Alto and monto is Bajo THEN  is Baja;
                (supportSize=353; laplace=0.988826815642458)
... and 66 other rules.
```

Fuente: Propia

Como quinto paso se sometió los datos de prueba a las reglas generadas y se evaluó la precisión del mismo.

```
#convertir a tabla de decisión la data prueba

decision.table_rueba<-SF.asDecisionTable(dataset
=sub_dataset_prueba)

#someter la tabla de decisión prueba a las reglas

pred.vals<-predict(rulesLEM2,decision.table_rueba)

#obtener el valor asignado real en la data_prueba

GrupoReglaPruebaReal.vals<-data.frame(dataset_prueba$Agrupacion)

#medir la precisión real vs simulado

mean(pred.vals==GrupoReglaPruebaReal.vals)
```

Como último paso se repitió 10 veces el proceso con la finalidad de evaluar la calidad de modelo, teniendo en cuenta que en cada interacción los datos de entrenamiento y prueba son seleccionados aleatoriamente en la razón de 60%y 40 % respectivamente.

Los primeros 5 resultados arrojaron un nivel de precisión superior a lo esperado, según la tabla 28

Tabla 28
Resultados Algoritmo LEM2 (60%-40%)

CASO	NRO DE REGLAS	PRECISIÓN
1	76	0.9998502
2	73	0.9998502
3	77	0.9994007
4	77	0.9992508
5	78	0.9992508

Fuente: Propia

A continuación, se muestra las 10 primeras reglas del caso 3

- IF monto is Bajo and frecuencia is Bajo and recencia is Medio THEN is Baja;

  (supportSize=465; laplace=0.991489361702128)

- IF monto is Bajo and recencia is Alto THEN is Baja;

  (supportSize=441; laplace=0.991031390134529)

- IF frecuencia is Muy bajo and recencia is Medio THEN is Baja;

  (supportSize=318; laplace=0.987616099071207)

- IF frecuencia is Bajo and recencia is Muy Alto and monto is Bajo THEN is Baja;

  (supportSize=262; laplace=0.98501872659176)

- IF frecuencia is Muy bajo and recencia is Alto THEN is Baja;

  (supportSize=176; laplace=0.977900552486188)

- IF monto is Muy bajo and recencia is Muy Alto THEN is Baja;

  (supportSize=167; laplace=0.976744186046512)

- IF Distrito is SUR and recencia is Medio and monto is Bajo THEN is Baja;

  (supportSize=589; laplace=0.993265993265993)

- IF monto is Muy bajo and recencia is Medio THEN is Baja;

  (supportSize=356; laplace=0.988919667590028)

- IF recencia is Alto and monto is Muy bajo THEN is Baja;

  (supportSize=216; laplace=0.981900452488688)

- IF monto is Medio and frecuencia is Bajo and recencia is Alto THEN is Baja;

  (supportSize=50; laplace=0.927272727272727)

Una vez identificados la agrupación de clientes y las reglas de asociación de clientes en base a ciertos atributos (distrito,

modalidad de pago, etc.) es posible recomendar que productos puede adquirir un cliente.

- Algoritmo APRIORI

  En primer lugar, se subdividió la data a nivel de los grupos (clústeres), a través de la siguiente consulta:

  ```
  dataset_APRIORI_groupNormalizadoFiltro_“n”<-subset(
  dataset_APRIORI_group,Agrupacion==“n”)
  ```

  Donde “n” es el número de clúster.

  Posteriormente se busca las asociaciones entre los atributos, por temas de estudio solo nos centraremos en el nivel de fidelidad medio debido a que es más fácil reposicionar un cliente regular que un cliente potencialmente dado de baja, sin embargo, su ejecución abarca todos los grupos definidos.

  Nos centraremos en 2 tipos de asociaciones.

  - Asociación por marca, distrito y tipo de negocio

    Subdividir los campos a usar a través de la siguiente consulta:

    ```
    aso_marca_distrito_tiponegocio<-
    dataset_APRIORI_groupNormalizadoFiltro_4[,c("Distrito","TipoCli
    ente","MarcaProducto")]
    ```

    Aplicamos el algoritmo A priori para determinar el nivel de asociaciones.

    ```
    rules_aso_marca_distrito_tiponegocio<-
    inspect(apriori(aso_marca_distrito_tiponegocio,parameter =
    list(supp=0.05,conf=0.7,target="rules")))
    ```

Figura 53
Reglas de asociación Nivel de fidelidad-Medio (marca, distrito, tipo negocio)

	lhs		rhs	support	confidence	lift	count
[1]	{}	=>	{TipoCliente=BODEGA}	0.74518292	0.7451829	1.0000000	815596
[2]	{}	=>	{Distrito=SUR}	0.81579931	0.8157993	1.0000000	892885
[3]	{MarcaProducto=GILLETTE}	=>	{TipoCliente=BODEGA}	0.05065551	0.7620893	1.0226876	55442
[4]	{MarcaProducto=GILLETTE}	=>	{Distrito=SUR}	0.05235310	0.7876289	0.9654689	57300
[5]	{MarcaProducto=PAMPERS}	=>	{TipoCliente=BODEGA}	0.05751167	0.7257947	0.9739820	62946
[6]	{MarcaProducto=PAMPERS}	=>	{Distrito=SUR}	0.07067669	0.8919368	1.0933287	77355
[7]	{MarcaProducto=ACE}	=>	{TipoCliente=BODEGA}	0.06263459	0.7536941	1.0114216	68553
[8]	{MarcaProducto=ACE}	=>	{Distrito=SUR}	0.06595669	0.7936695	0.9728734	72189
[9]	{MarcaProducto=DOWNY}	=>	{TipoCliente=BODEGA}	0.06432122	0.7617950	1.0222926	70399
[10]	{MarcaProducto=DOWNY}	=>	{Distrito=SUR}	0.07051314	0.8351296	1.0236950	77176
[11]	{MarcaProducto=ARIEL}	=>	{TipoCliente=BODEGA}	0.06551082	0.7512285	1.0081128	71701
[12]	{MarcaProducto=ARIEL}	=>	{Distrito=SUR}	0.06692974	0.7674996	0.9407946	73254
[13]	{MarcaProducto=AYUDIN}	=>	{TipoCliente=BODEGA}	0.07079729	0.7010939	0.9408346	77487
[14]	{MarcaProducto=AYUDIN}	=>	{Distrito=SUR}	0.07692617	0.7617871	0.9337923	84195
[15]	{MarcaProducto=PANTENE}	=>	{TipoCliente=BODEGA}	0.08356304	0.7619932	1.0225586	91459
[16]	{MarcaProducto=PANTENE}	=>	{Distrito=SUR}	0.08940046	0.8152234	0.9992940	97848
[17]	{Distrito=CENTRO}	=>	{TipoCliente=BODEGA}	0.11808594	0.7297097	0.9792357	129244
[18]	{TipoCliente=MERCADO}	=>	{Distrito=SUR}	0.15099713	0.8308239	1.0184170	165265
[19]	{MarcaProducto=H&S}	=>	{TipoCliente=BODEGA}	0.17799507	0.7564388	1.0151048	194814

Fuente: Propia

- Asociación por producto, tipo de negocio y modalidad de pago

  Subdividir los campos a usar a través de la siguiente consulta:

  aso_producto_tiponegocio_modalidad_pago<-dataset_APRIORI_groupNormalizadoFiltro_4[,c("TipoModalidadPago","TipoCliente","CodProducto")]

  Aplicamos el algoritmo A priori para determinar el nivel de asociaciones.

  rules_aso_producto_tiponegocio_modalidad_pago<-inspect(apriori(aso_producto_tiponegocio_modalidad_pago,parameter = list(supp=0.05,conf=0.7,target="rules")))

Figura 54
Reglas de asociación Nivel de fidelidad-Medio
(producto, tipo de negocio, modalidad de crédito)

	lhs		rhs	support	confidence	lift	count
[1]	{}	=>	{TipoCliente=BODEGA}	0.74518292	0.7451829	1.0000000	815596
[2]	{}	=>	{TipoModalidadPago=CONTADO}	0.79936336	0.7993634	1.0000000	874896
[3]	{Producto=SACHETON HS LIMPIEZA RENOV*18ML}	=>	{TipoCliente=BODEGA}	0.07171005	0.7993523	1.0726927	78486
[4]	{Producto=SACHETON HS LIMPIEZA RENOV*18ML}	=>	{TipoModalidadPago=CONTADO}	0.07653603	0.8531476	1.0672838	83768
[5]	{TipoCliente=MERCADO}	=>	{TipoModalidadPago=CONTADO}	0.14244338	0.7837591	0.9804791	155903
[6]	{TipoModalidadPago=CREDITO}	=>	{TipoCliente=BODEGA}	0.14482166	0.7218106	0.9686355	158506
[7]	{TipoCliente=BODEGA}	=>	{TipoModalidadPago=CONTADO}	0.60036126	0.8056562	1.0078724	657090
[8]	{TipoModalidadPago=CONTADO}	=>	{TipoCliente=BODEGA}	0.60036126	0.7510493	1.0078724	657090
[9]	{TipoCliente=BODEGA,Producto=SACHETON HS LIMPIEZ...	=>	{TipoModalidadPago=CONTADO}	0.06161951	0.8592870	1.0749642	67442
[10]	{TipoModalidadPago=CONTADO,Producto=SACHETON H...	=>	{TipoCliente=BODEGA}	0.06161951	0.8051046	1.0804120	67442

Fuente: Propia

3.2. RESULTADOS

Implican la evaluación del modelo, teniendo en cuenta los criterios de éxito definidos en el paso anterior (conocimiento del negocio), así como los aspectos a mejorar.

3.2.1. EVALUACIÓN Y ANÁLISIS DE RESULTADOS

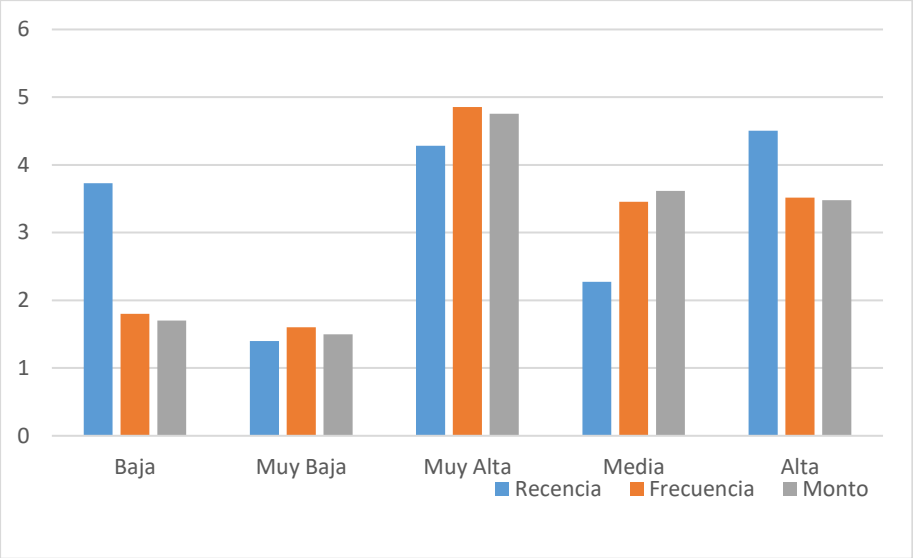
A) Evaluación de los resultados

Se establecieron 2 criterios de éxito en el negocio:

- Conocer el comportamiento de compra del cliente.

  En base a los resultados de la agrupación de clientes con el algoritmo Kmeans, se descubrieron 5 niveles de lealtad (ver figura 55) posibles debido a que se puede clasificar al cliente según su frecuencia de compra, monto comprado y última compra.

Figura 55
Agrupación Clientes



Fuente: Propia

Tabla 29
Perfil de Fidelidad

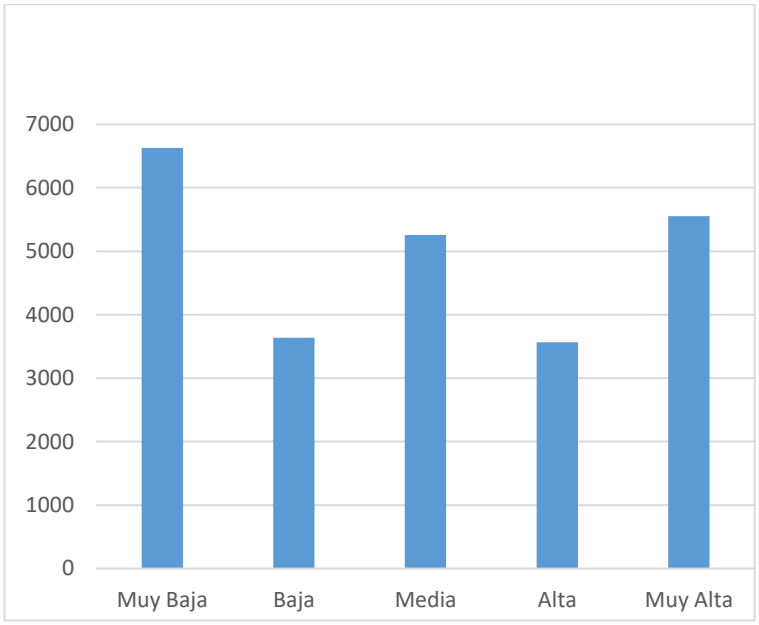
GRUPO	PUNTUACION RFM			Características
	R	F	M	
Muy Alta	4.28	4.86	4.75	Son clientes campeones porque compraron recientemente, con frecuencia y gastan más.
Alta	4.50	3.52	3.48	Son clientes leales debido a que gastan más de lo regular y son responsivos a las promociones.
Media	2.27	3.46	3.62	Son clientes nuevos y/o prometedores que no compran muy menudo, pero gastan regularmente
Baja	3.73	1.80	1.70	Son clientes que necesitan atención debido a que compran poco y con baja frecuencia.
Muy Baja	1.40	1.60	1.49	Son clientes hibernadores o

	perdidos debido a que gastan poco y no compran hace mucho tiempo.

Fuente: Propia

Cada uno de los grupos descritos posee cierta cantidad de clientes a los cuales se les aplicara un marketing dirigido. La figura 56 muestra la distribución de los clientes según su nivel de fidelidad.

Figura 56
Número de clientes según nivel de fidelidad



Fuente: Propia

El jefe de venta y los supervisores deben de poner un alto interés en los clientes de fidelidad baja y media que suman más de la tercera parte del universo de clientes con la finalidad de recuperarlos.

- Cumplir con los objetivos mensuales asignados a la fuerza de venta exclusiva.

  Con la finalidad de cumplir con los objetivos mensuales, se tuvo que interpretar mejor las características (ubicación, tipo de negocio, modalidad de pago y variables RFM) de cada grupo (clúster) creado, permitiendo al experto del negocio sugerirle la reclasificación del mismo, conllevando a un potencial crecimiento, reflejado en el cumplimiento de los objetivos.

  Por lo que será entregado un conjunto de recomendaciones de productos para los clientes con fidelidad media, así como un informe de los 20 productos que se han comprado con mayor frecuencia dentro del grupo.

B) Proceso de revisión

El proceso hasta este punto se ha ejecutado tal y como estaba previsto, sin embargo, por premuras de tiempo no se han sometido a pruebas los demás grupos de clústeres identificados para deducir las demás reglas de asociación, y también solo se ha trabajado con un único algoritmo para cada proceso [agrupación (Kmeans), inducción (LEM2), y asociación (Apriori)], pudiendo usar otros algoritmos que permitan encontrar nuevos resultados.

C) Determinación de futuras fases.

El siguiente paso a realizar en el proyecto es el de ejecutar la etapa de despliegue e implementación.

3.2.2. DESPLIEGUE E IMPLEMENTACIÓN

Implica transformar el conocimiento en acciones dentro del proceso de negocio.

A) Plan de implementación

Para poder implantar este proyecto es necesario:

- Tener acceso a la base de datos transaccional - SQL Server.
- Trabajar con Pentaho Data Integration para el diseño del datamart, sin embargo, es posible usar otras herramientas ETL, pero conllevaría tiempo en el diseño y posiblemente la revisión del acápite preparación de datos: limpieza de datos.
- Tener acceso al datamart de ventas diseñado en el motor de base de datos SQL, sin embargo, si se trabaja con otro motor de base de datos, solo es necesario configurar un nuevo ODBC, debido a que es la configuración realizada R para consultar la base de datos.

Tabla 30
Cronograma de actividades

Cronograma de actividades	**ENE 15 - 31**	**FEB 1 - 14**	**FEB 15 - 28**	**MAR 1 - 15**	**MAR 16 - 31**	**ABR 1 - 8**
Definición del tema						
Anteproyecto	X	X				
Conocimiento del negocio	X	X				
Preparación de los datos	X	X	X			
Modelado		X	X			
Evaluación			X	X	X	
Implementación				X	X	X
Presentación						X

Fuente: Propia

CONCLUSIONES

Este proyecto se enfocó en identificar el comportamiento del cliente de acuerdo a sus compras, definiendo para ello una metodología de desarrollo. En tal sentido se cumplió con los objetivos planteados, obteniéndose los resultados asociados a cada uno de ellos:

- La implementación de un datamart para el área de ventas, está basado en los requerimientos del negocio, con el afán de consolidar la información y hacer más fácil la limpieza de datos en el desarrollo del modelo predictivo.
- Se detallaron 5 niveles de fidelidad en el presente trabajo: Muy Alta, Alta, Media, Baja y Muy Baja, y un modelo con un 99.9% de fiabilidad estos resultados permitieron al jefe de ventas y supervisores gestionar con la gerencia y los proveedores promociones y/o descuentos personalizados.
- La aplicación del algoritmo de asociación Apriori sobre el conjunto de transacciones del grupo media, permitió elaborar reglas de asociación importantes con niveles de confianza superior al 85% con la finalidad de realizar sugeridos en las ventas.
- Las recomendaciones de productos generadas para el grupo media permiten al vendedor sugerir la compra de ciertos productos a los clientes con el mismo nivel de fidelidad,
- Las recomendaciones permiten re categorizar a los clientes pertenecientes al grupo de fidelidad baja y muy baja para que puedan acceder a dichos beneficios, conllevando un aumento en el ticket de venta y reflejándose en el cumplimiento de los objetivos.

RECOMENDACIONES

- Considerar el resguardo del datamart debido a que es la fuente del modelo predictivo planteado.

- Es importante tener políticas de gestión de datos con la finalidad que la información almacenada sea la más transparente posible debido a que minimiza los tiempos de desarrollo (limpieza de datos) de un modelo predictivo.

- Debido a la premura del tiempo solo se trabajó con un único algoritmo en cada proceso, si bien los resultados fueron los esperados es necesario hacer una comparativa con 2 o más algoritmos con la finalidad de encontrar el óptimo.

- Terminar de analizar los grupos faltantes con la finalidad de encontrar nuevas reglas de asociación.

- Se debe realizar el proceso cada 3 meses con la finalidad de mantener actualizada la categorización del cliente.

BIBLIOGRAFÍA

Alcalde Aliaga, M. E. (Setiembre de 2018). *Fundamentos de Business Intelligence - Diplomado de Business Intelligence [diapositiva].* Recuperado el 2 de Marzo de 2019, de https://isil.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_33134_1&content_id=_1499018_1

Azuaje, A. (2014). *Metodología de Kímball.* Caracas. Obtenido de https://docs.google.com/document/d/1qtaKD91OjqDAIHmCq-hcp2AjT922FTAgP8j9cbdadpU/edit

Benalcázar, J. (2017). *Análisis comparativo de metodologías de minería de datos y su aplicabilidad a la industria de servicios.* Universidad De las Américas, Quito, Ecuador.

Bernabeu, D. (6 de Mayo de 2009). *Datawarehouse manager*. Recuperado el 5 de Marzo de 2019, de http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/arquitectura-del-data-warehouse/34-datawarehouse-manager

Blogadmarketing. (10 de Octubre de 2017). *BLOGS UPC*. Obtenido de Blog de Administración y Marketing: https://blogs.upc.edu.pe/blog-de-administracion-y-marketing/noticias/consumo-masivo-el-mercado-peruano-y-la-gestion-de

Bustos, S., & Mosquera, V. (2013). Análisis, diseño e implementación de una solución business intelligence para la generación de indicadores y control de desempeño, en la empresa OTECEL S.A, utilizando la metodología Hefesto V2.0. *tesis de pregrado.* Escuela Politécnica del Ejército, Sangolquí, Ecuador.

Calderón, N. d. (2006). Minería de datos una herramienta para la toma de decisiones. *tesis de pregrado.* Unversidad de San Carlos de Guatemala, Guatemala, Guatemala.

Caldwell, C. (14 de Junio de 2018). *Tendencias de BI - Business Intelligence vs Analytics: ¿Cuál es la diferencia?* Recuperado el 07 de Marzo de 2019, de LogiAnalytics: https://www.logianalytics.com/bi-trends/by-the-numbers-the-latest-stats-in-embedded-analytics/

Chamba, S. (2015). Minería de datos para la segmentación de clientes en la empresa tecnológica Master PC. *tesis de pregrado.* Universidad Nacional de Loja, Loja, Ecuador.

Chen, j. (7 de Agosto de 2018). *Forecasting*. Recuperado el 8 de Marzo de 2019, de https://www.investopedia.com/terms/f/forecasting.asp

Córdova, J. (2013). Análisis,diseño e implementación de una solución de inteligencia de negocios para el área de importaciones en una empresa comercializadora/importadora. *tesis de pregrado.* Universidad Pontifica Catolica del Perú, Lima, Lima, Perú.

Davenport, T. H. (Diciembre de 2011). Competir mediante el analisis. págs. 18-28.

Davenport, T. H., Dalle Mule, L., & Lucker, J. (Febrero de 2012). Sepa qué quieren sus clientes antes que ellos mismos. *Haward Business Review*, 50-56.

De Rossi, A. (Octubre de 2018). *Taller de soluciones de BI - Diplomado de Business Intelligence [diapositiva].* Recuperado el 3 de Marzo de 2019, de https://isil.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_33143_1&content_id=_1499048_1

Definista, C. G. (s.f.). *Definición de Comercializadora*. Obtenido de https://conceptodefinicion.de/comercializadora/

Digital Guide. (30 de Enero de 2018). *Software de data mining: realiza análisis de datos más efectivos*. Recuperado el 8 de Marzo de 2019, de https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas/

ESAN. (21 de Noviembre de 2017). *Business Intelligence vs Business Analytics: ¿Hay diferencias?* Recuperado el 9 de Marzo de 2019, de https://www.esan.edu.pe/apuntes-empresariales/2017/11/business-intelligence-vs-business-analytics-hay-diferencias/

Esparza, D., Alvarez, C., Duque, L., & Quiroz, D. (2014). Análisis, diseño e implementación de un Datamart utilizando herramientas Open Source para la unidad administrativa y financiera de la ESPE. *tesis de pregrado.* Universidad de las Fuerzas Armadas ESPE, Sangolqui, Ecuador.

Ferrándiz, L. A. (Setiembre de 2018). *Gestíon de Datos- Diplomado Business Intelligence [diapositiva].* Recuperado el 2 de Marzo de 2019, de https://isil.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_33135_1&content_id=_1499021_1

Ferrer Mos, S. (12 de Febrero de 2015). *tatis*. Obtenido de http://pertutatis.cat/la-piramide-de-los-diferentes-tipos-de-sistemas-de-informacion/

Gallardo, J. (2009). Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM). *tesis de doctorado.* Universidad Politecnica de Madrid, Madrid, España.

García, J., & Acevedo, Á. (2010). Análisis para la predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies. *tesis de grado.* Universidad Tecnológica de Pereira, Pereira, Colombia.

Gartner. (Setiembre de 2018). *Gartner, Inc.* Obtenido de Gartner Data & Analytics: https://www.gartner.com/binaries/content/assets/events/keywords/business-intelligence/mda18/mda18_brochure_6.pdf

Gómez, A. (2012). Inteligencia de negocios, una ventaja competitiva para las organizaciones. *Ciencia y Tecnología - Universidad Nacional de Trujillo - Escuela de Postgrado, 8*(22). Obtenido de http://revistas.unitru.edu.pe/index.php/PGM/article/view/193

Grández, M. (2017). Aplicación de minería de datos para determinar patrones de consumo futuro en clientes de una distribuidora de suplementos nutricionales. *tesis de pregrado.* Universidad San Ignacio de Loyola, Lima, Lima, Perú.

Guillén, F. (2012). Desarrollo de un datamart para mejorar la toma de decisiones en el área de tesorería de la municipalidad provincial de cajamarca. *tesis de pregrado.* Universidad Privada del Norte, Cajamarca, Perú.

Guillén, S., & Sánchez, K. (2017). *Evaluación de la gestión del área de ventas de la empresa constructora JSM S.A.C para proponer medidas correctivas que incrementen la rentabilidad económica, periodo 2015-2016.* universidad Católica Santo Toribio de Mogrovejo, Chiclayo, Perú.

Hernandez, L. (2008). Diseño y construcción de un datamart para la manteción de indicadores de sostenibilidad de la industria del salmón. *tesis de maestria.* Universidad de Chile, Santiago de Chile, Chile.

Hsue, C. C., & Shyang, C. Y. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *partment of Information Management, National Yunlin University of Science and Technology*, págs. 4176-4184.

Ibarra, A. (Agosto de 2014). *Por qué debes conocer el nivel de madurez del Business Intelligence de tu empresa*. Recuperado el 4 de Marzo de 2019, de http://www.nebi.co/bi/por-que-debes-conocer-el-nivel-de-madurez-del-business-intelligence-de-tu-empresa/

Idoine, C., Krensky, P., Brethenoux, E., & Linden, A. (28 de Enero de 2019). *Magic Quadrant for Data Science and Machine Learning Platforms*. Recuperado el 07 de Marzo de 2019, de https://www.gartner.com/doc/reprints?id=1-65WC0O1&ct=190128&st=sb

Invitado, A. (22 de Marzo de 2012). Recuperado el 8 de Marzo de 2019, de https://www.emprendices.co/que-es-un-distribuidor/

Kendall, J. E., & Kendall, K. E. (2005). *Análisis y Diseño de Sistemas* (Sexta ed.). Mexico: Prentica Hall Inc.

Li, S. (24 de Setiembre de 2017). *A Gentle Introduction on Market Basket Analysis—Association Rules*. Recuperado el 8 de Marzo de 2019, de https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce

Marrs, M. (10 de Mayo de 2016). *The Difference Between Data, Analytics, and Insights*. Recuperado el 8 de Marzo de 2019, de http://info.localytics.com/blog/difference-between-data-analytics-insights

Martínez, C. (2012). Aplicación de técnicas de minería de datos para mejorar el proceso de control de gestión de Entel. *Tesis de maestria.* Universidad de Chile, Santiago de Chile, Chile.

Matute, G. (12 de Abril de 2013). *¿Por qué es importante aplicar la inteligencia de negocios?* Recuperado el 12 de Marzo de 2019, de https://www.esan.edu.pe/conexion/actualidad/2013/04/12/inteligencia-negocios-empresa/

Mérida, J. (2016-2017). Adaptación de estándares de dirección de proyectos particularizados para la minería de datos. *tesis de maestría.* Universidad de Oviedo, Oviedo, España.

Moine, J., Gordillo, S., & Haedo, A. (2011). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. *XVII Congreso Argentino de Ciencias de la Computación (CACIC 2011)*, (pág. 8). Buenos Aires, Argentina. Obtenido de http://hdl.handle.net/10915/18749

Mori Paiva, H. A. (Noviembre de 2018). *Estadistica par los negocios - Diplomado de Business Intelligence [diapositiva].* Recuperado el 8 de Marzo de 2019, de https://isil.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_33140_1&content_id=_1499038_1

Ñaupas, C. (2016). Mineria de datos aplicada a la detección de fraude electrónico en entidades bancarias. *Tesis de pregrado.* Universidad Nacional Mayor de San Marcos, Lima, Lima, Perú.

PowerData. (19 de Marzo de 2016). *El valor de la gestión de datos, ¿Qué son los metadatos y cuál es su utilidad?* Recuperado el 8 de Marzo de 2019, de https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/que-son-los-metadatos-y-cual-es-su-utilidad

Quillén, R. (2017). Sistema de soporte de decisiones con tecnología Datawarehouse para la gestión de la información de la empresa Mallku Import SAC - Juliaca 2016. *tesis de pregrado.* Universidad Nacional del Altiplano, Puno, Puno, Perú.

Rodriguez, J. (Enero de 2019). *Analitica de datos & data mining - Diplomado de Business Intelligence [diapositiva].* Recuperado el 5 de Marzo de 2019, de https://isil.blackboard.com/webapps/blackboard/content/listContent.jsp?course_id=_33141_1&content_id=_1499041_1

Rodriguez, K., & Mendoza, A. (2011). Análisis diseño e implementacíon de una solucíon de inteligencia de negocios para el área de compras y ventas de una empresa comercializadora de electrodomésticos. *Tesis de pregrado.* Universidad Pontificia Católica del Perú, Lima, Lima, Perú.

Roque, I. (2016). Análisis comparativo de técnicas de minería de datos para la predicción de ventas. *tesis de pregrado.* Universidad Señor de Sipán, Pimentel, Chiclayo, Perú.

Sánchez, O. (2014). Modelo de inteligencia de negocios para la toma de decisiones en la empresa San Roque S.A. *Tesis de pregrado.* Universidad Privada Antenor Orrego, Trujillo, Perú.

Santoyo, S. (11 de Setiembre de 2017). *A Brief Overview of Outlier Detection Techniques*. Recuperado el 8 de Marzo de 2019, de https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561

SAS. (s.f.). *Análisis estadístico - Mire a su alrededor. Las estadísticas están por doquier.* Recuperado el 08 de Marzo de 2019, de https://www.sas.com/es_pe/insights/analytics/statistical-analysis.html

Sinnexus. (s.f.). *Datamining (Minería de datos)*. Recuperado el 08 de Marzo de 2019, de https://www.sinnexus.com/business_intelligence/datamining.aspx

Sinnexus. (s.f.). *Datos, información, conocimiento*. Recuperado el 08 de Marzo de 2019, de https://www.sinnexus.com/business_intelligence/piramide_negocio.aspx

Timón, C. (2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source. *tesis de pregrado.* Universitat Oberta de Catalunya, Barcelona, España.

Vega, D. (2005). Gestión estrategica del dpto de ventas aplicada en una empresa comercial - farmaceutica. *tesis de pregrado.* Universidad Nacional Mayor de San Marcos, Lima, Lima, Perú.