Universidad Internacional de La Rioja (UNIR)

ESIT

Máster en Análisis y Visualización de Datos Masivos

Metodología de análisis y segmentación de clientes usando secuencias de comportamiento.

Trabajo Fin de Máster

presentado por: Casariego Sarasquete, Nicolás-Martín

Director/a: Zanardini, Damiano

Ciudad: Madrid

Fecha: 26 de julio 2019

Resumen

Los modelos de segmentación y análisis de marketing tradicionales están limitados para agrupar y describir el comportamiento omnicanal del nuevo consumidor.

La Metodología de Segmentación y análisis de clientes basados en Secuencias de Comportamiento proporciona un novedoso enfoque que comprende:

- Su representación formal y almacenamiento,
- Su tratamiento algorítmico,
- Su tratamiento analítico con técnicas de “machine learning”,
- Su representación visual y
- La creación de segmentos basados en comportamiento.

El objetivo de la metodología es poder aplicarse en diversas industrias y sectores de negocio, y que permita segmentar y analizar clientes, ciudadanos y pacientes.

Se despliega la metodología en dos escenarios de negocio completamente diferentes para demostrar sus posibilidades prácticas de aplicación.

Los resultados muestran que la metodología complementa el abordaje tradicional con sus capacidades adicionales de segmentar y analizar el comportamiento de los clientes, permitiendo mejorar la gestión del valor, la fidelización, y la experiencia del cliente.

Palabras Clave: Segmentación por comportamiento, Análisis predictivo de clientes, Minería de secuencias de comportamiento.

Abstract

Traditional marketing analytic and segmentation models are constrained for grouping and describing the new consumer's omnichannel behavior.

The brand-new approach provided by this Customer analysis and segmentation methodology based on Behavioral sequences, covers:

- Formal representation and storage structure,
- Algorithms,
- Machine learning based analytics,
- Visualization and
- Behavioral based segment creation.

The aim of the methodology is to be applied in different industries and business sectors, and enable to segment and analyze customers, citizens and patients.

The methodology is deployed in two completely different business scenarios to demonstrate the practical capabilities of its application.

The results show that the methodology complements the traditional approach with additional capabilities to segment and analyze customer's behavior, enabling customer value, loyalty and experience improvement.

Keywords: Behavioral segmentation, Customer predictive analytics, Behavioral sequences mining.

Agradecimientos

La realización del presente Trabajo Final de Máster no hubiera sido posible sin el apoyo incondicional de mi esposa Mónica y mis hijos Agustín, Manuel y Joaquina.

Por otra parte, deseo agradecer a mi director de proyecto, Damiano Zanardini, por la confianza depositada y por su acertada orientación.

Finalmente quiero agradecer al panel de expertos, Laura Blanco, Joaquín de Aguilera y Agapito Ledezma; quienes tan exhaustiva y generosamente han contribuido a la evaluación de la metodología en sus diferentes fases.

Índice de Contenidos

Índice de Figuras

1. Síntesis

Los modelos de segmentación y análisis de marketing tradicionales son demasiado estáticos y están limitados para agrupar y describir satisfactoriamente al nuevo consumidor en su comportamiento dinámico a través de múltiples canales.

Exponemos el estado del arte en las áreas que son relevantes para dar forma a una nueva metodología: modelos de segmentación de clientes, paradigmas analíticos de clientes, algoritmos de detección de patrones en secuencias, y técnicas de visualización de secuencias.

Definimos una **Metodología de Segmentación y análisis de clientes basados en Secuencias de Comportamiento** (en adelante **MSC2**) que proporciona un novedoso enfoque para tratar el comportamiento dinámico de los clientes. La misma comprende:

- Su representación formal y sus propiedades,
- Su almacenamiento en repositorios de “big data”,
- Su tratamiento algorítmico,
- Su tratamiento analítico con técnicas de “machine learning”,
- Su representación visual y
- La creación de segmentos basados en comportamiento para su uso en campañas.

El objetivo de la Metodología MSC2 consiste en proporcionar a los decisores de marketing una nueva herramienta que complemente las actuales prácticas de segmentación de clientes, ciudadanos y pacientes.

Buscamos establecer una metodología generalizable que permita generar conocimiento de clientes a partir de la observación de su comportamiento dinámico y omnicanal.

Que permita analizar y segmentar clientes en base a su comportamiento, descubrir patrones, identificar su evolución o su ausencia, y anticiparse al mismo.

Desplegamos la metodología MSC2 de principio a fin en **dos escenarios de negocio completamente diferentes** (Tienda de moda y Servicios Sanitarios) para demostrar sus posibilidades de aplicación. Usamos dos conjuntos de datos sintéticos a partir de datos reales (que no puedo exponer por cuestiones de confidencialidad), pero que permiten mostrar sus ventajas de segmentación y análisis.

Los **beneficios para el usuario de marketing** consisten en poder comprender, analizar, segmentar y anticipar el comportamiento. Sumando este nuevo enfoque al tradicional, podemos mejorar la gestión del valor, la fidelización, y la experiencia de sus clientes.

Los **beneficios para el usuario técnico** consisten en poder reaprovechar su infraestructura tecnológica existente, y contar con nuevas variables de análisis que puede incorporar en sus sistemas de "business intelligence" corporativos.

2. Antecedentes

El presente proyecto se apoya sobre un trabajo conceptual previo (parcialmente publicado en [00] (Casariego, 2017)), realizado a partir de mi labor profesional de 30 años. Siempre enfocado en tecnología y analítica de marketing para compañías líderes tanto en Europa, como en Latinoamérica. Principalmente en los sectores de la Banca, Tarjetas de crédito, Telecomunicaciones, Automotriz, Gran consumo, Distribución, Lujo, Salud y Gobierno.

A partir de la base conceptual, en el último año hemos estado desarrollando y probando la metodología con tres empresas que han participado del pilotaje proveyendo sus datos, y su problemática en sectores tan diversos como turismo, entretenimiento y sanidad.

Hemos contado con un panel de expertos en las áreas de marketing, negocio, innovación, y minería de datos. Han participado del pilotaje aportando especificaciones de usuario, revisando la metodología y evaluando los resultados.

Aunque los pilotos están bajo el paraguas de sendos acuerdos de confidencialidad, la metodología descrita en este TFM es la misma que la utilizada en dichos proyectos, dado que al ser de mi autoría no forman parte de la confidencialidad.

El piloto correspondiente a Sanidad ha sido seleccionado finalista en la segunda convocatoria Innolabs de proyectos TIC en Healthcare correspondiente al programa de I+D+i de la Unión Europea Horizon 2020 (grant agreement No 691556 – sub-grant agreement 2018/033).

Los datos empleados en el desarrollo de los dos casos de uso han sido generados sintéticamente. Aunque su estructura y composición son de naturaleza semejante a los que se pueden obtener en los casos reales.

3. Introducción

El comportamiento del consumidor ha cambiado drásticamente en los últimos años, según [01] (Forrester, 2011) estamos inmersos en la "Era del Cliente", esto significa que el control en la relación Organización-Cliente ha cambiado de manos. Ya nada será igual que antes. Ahora nos toca servir a un Cliente (consumidor, ciudadano, paciente) empoderado, más

informado, y más influyente que nunca. Demandante de productos y servicios relevantes, personalizados, ubicuos y en tiempo real.

Ser capaz de anticipar el comportamiento de nuestros clientes es el santo grial de cada ejecutivo de negocio.

Muchos clientes exhiben un comportamiento recurrente que ahora somos capaces de identificar, organizar y almacenar por completo. Esta nueva capacidad ocurre gracias a los procesos de transformación digital, la adopción masiva de las redes sociales, las plataformas "omnicanal", la infraestructura en la nube, y la tecnología disponible de "big data" y "machine learning".

Este comportamiento observado consiste en transacciones repetitivas, compras y pagos recurrentes, navegación e interacciones en las propiedades digitales, canales, dispositivos, aplicaciones y medios sociales.

Sin ánimo de ser exhaustivo, hablamos principalmente de modelos de negocio B2B, o bien de modelos de negocio B2C en las siguientes categorías: servicios financieros, seguros, tarjetas de crédito, comercio electrónico, telecomunicaciones, medios digitales, servicios sanitarios, retail, cosmética, consumibles, etc.

Pero almacenar y recuperar dichos datos no es lo mismo que generar conocimiento, para ello necesitamos un modelo de representación del comportamiento, un modelo de segmentación de clientes basado en el comportamiento que nos facilite interactuar con ellos, descubrir patrones de comportamiento y su evolución.

Según Blasingame solamente así podremos gestionar adecuadamente la satisfacción, la experiencia y el valor de los clientes en la "Era de la relevancia" [02] (Blasingame, 2014).

4. Estado del arte

Este estudio metodológico pivota sobre cuatro dimensiones del problema planteado:

¿Cómo incorporar a los actuales modelos de segmentación de clientes los aspectos claves del nuevo consumidor: la omnicanalidad y la dinámica en tiempo "cuasi" real? Para ello hemos investigado el estado del arte en los Modelos de segmentación de clientes.

¿Cómo incorporar a los actuales modelos de datos analíticos los aspectos relacionados con la secuencia cronológica temporal y la omnicanalidad, sin perder las capacidades analíticas multidimensionales, estadísticas y predictivas? Para ello hemos investigado el estado del arte en los Paradigmas analíticos de clientes.

¿Cómo identificar patrones frecuentes en el comportamiento de grandes volúmenes de clientes, ciudadanos o pacientes, durante el “consumer journey” a lo largo de todo el ciclo de vida? Para ello hemos investigado el estado del arte en los Algoritmos de búsqueda de patrones frecuentes en secuencias de eventos.

¿Cómo representar visualmente el comportamiento observado del cliente en forma de secuencias y patrones que valga para cualquier metamodelo de negocio? ¿Cómo navegar e interactuar con dichas visualizaciones? Para ello hemos investigado el estado del arte en las Técnicas de visualización de secuencias.

4.1. Modelos de segmentación de clientes

Evolución histórica

El término “segmentación” fue acuñado en Marketing por primera vez por [03] (Smith, 1956) cuando escribió en el Marketing Journal: "La segmentación del mercado implica ver un mercado heterogéneo como un número de mercados homogéneos más pequeños en respuesta a preferencias diferentes, atribuibles al deseo de los consumidores de obtener una satisfacción más precisa de sus diferentes necesidades".

A partir de entonces la segmentación de clientes y la analítica necesaria para conocerlos, han evolucionado profundamente ligadas a la creciente disponibilidad de datos de clientes.

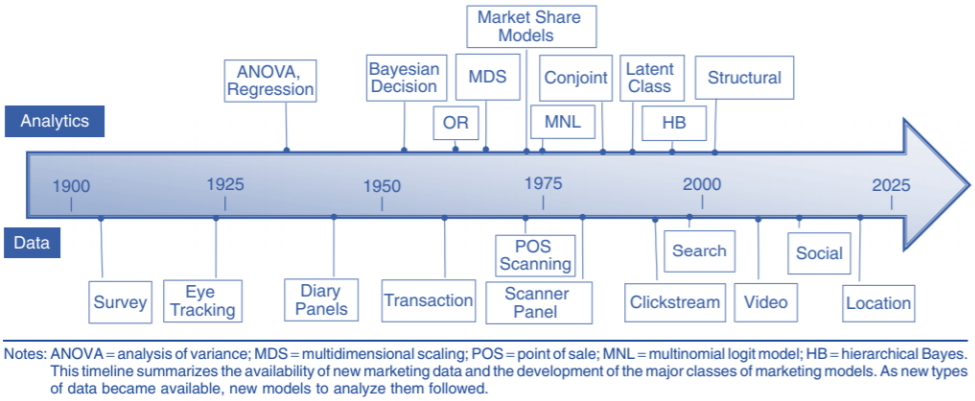


Figura 1 Línea temporal de datos y analítica de marketing [04] (Wedel & Kannan, 2016).

Según Wedel & Kannan, poniendo en perspectiva la historia de las fuentes de datos, entendemos mejor la evolución de la segmentación y las técnicas analíticas asociadas. Por

simplicidad nos referiremos a décadas completas aun a riesgo de perder precisión cronológica con el redondeo.

En los 1950s se dispone de encuestas y paneles de consumidores. En los 1960s y 1970s se incorpora la disponibilidad de censos y con ellos la segmentación sociodemográfica. En los 1980s se incorporan las transacciones bancarias y las transacciones en punto de venta; y lo que es más importante, la posibilidad de identificar al cliente en ciertas transacciones. Comienza la segmentación transaccional y por afinidad de clientes.

En los 1990s se incorpora la dimensión actitudinal y psico-gráfica dando origen a los conglomerados de clientes. Las técnicas de respuesta directa (correo directo y televenta) proporcionan la respuesta individualizada de cada campaña, y la propensión a la compra de un producto por parte de un cliente o segmento.

En los 2000s se incorpora la capacidad de trazar la navegación web y las búsquedas en entornos digitales; generando el segundo salto cuantitativo en volumen y riqueza de datos de cliente. Con ello nace la microsegmentación, o segmentación uno a uno.

En los 2010s se incorpora el contenido audiovisual, la localización en tiempo real, la omnicanalidad, y la capacidad de escuchar las redes sociales.

Simultáneamente se produce un cambio radical en el comportamiento del consumidor, cada vez menos fiel, y cada vez más exigente, impaciente, e impermeable a los mensajes publicitarios.

Ya no se trata solamente del gigantesco volumen de datos, ni de la velocidad que se debe imprimir a cada acción o reacción; sino también de la capacidad de hiper-personalizar los mensajes, productos y servicios en permanente carrera por ganar relevancia frente al cliente.

Nace entonces la segmentación basada en el comportamiento de los clientes, disparada por eventos; y juntamente con ella, el “targeting” de precisión en tiempo real.

Prácticas de la industria

Los siguientes modelos de segmentación son los más popularmente adoptados por los profesionales de marketing.

Segmentación por Valor de cliente (CLV):

Esta segmentación se basa en el concepto de valor de cliente a largo plazo (“CLV Customer lifetime value”) durante todo el ciclo de vida del cliente con la marca [05] (Fader & Toms, 2012).

Segmenta a los clientes en alto, medio y bajo valor, desde un prisma estrictamente monetario de rentabilidad e inversión por cliente a largo plazo.

Segmentación transaccional (RFM):

El modelo de segmentación RFM pivota sobre tres dimensiones: Recencia: proximidad de la última compra, Frecuencia: frecuencia de compra, y Monetización: valor monetario de las compras [06] (Chen & Sain & Guo, 2012).

Es simple de implementar y permite decidir el nivel de segmentación que se desea conseguir hasta llegar a 125 segmentos de clientes diferentes ("111" a "555"). Es un modelo muy intuitivo porque tipifica a los clientes exclusivamente en base a sus hábitos de compra.

El modelo de segmentación RF de recencia y frecuencia tiene una correspondencia con el modelo CLV [07] (Fader & Hardie & Lee 2005) a través de un mapa de iso-curvas de valor.

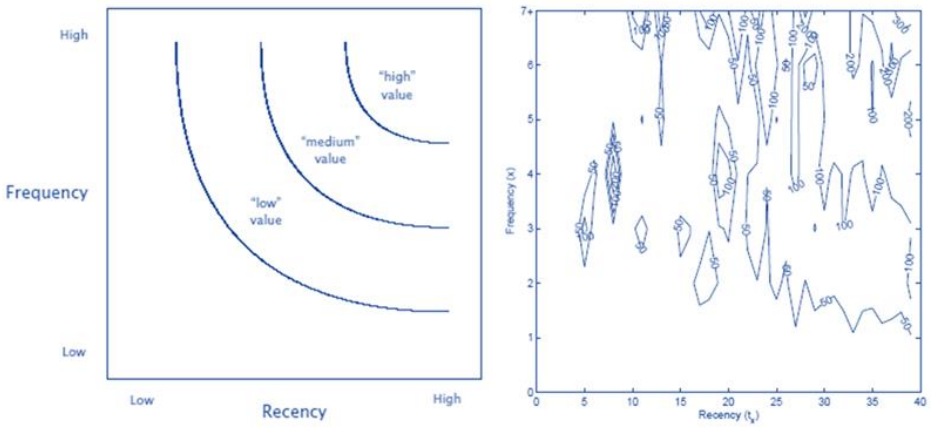


Figura 2 Modelo RF versus CLV (Fader et al. 2005).

Segmentación por satisfacción (NPS):

La medición y segmentación a través de la satisfacción del cliente ha sido objeto de varias métricas a lo largo del tiempo, Índice de satisfacción de cliente "CSAT", Índice de esfuerzo de cliente "CES", y finalmente "NPS" "Net Promoter Score" [08] (Marr B. 2012).

Figura 3 Modelo de segmentación NPS (elaboración propia).

NPS es la única métrica definida a través de un marco formal, y se ha estandarizado: la formulación de la pregunta, la escala numérica de la respuesta (de 0 a 10), la fórmula de cálculo, y la investigación cuantitativa posterior. Los clientes acaban agrupándose en

categorías de satisfacción/fidelidad: promotores (satisfechos), indiferentes y detractores (insatisfechos). Esta métrica ha sido mayoritariamente adoptada debido a su capacidad comparativa entre compañías, industrias y mercados.

Segmentación por Valor / Riesgo:

Estos modelos de segmentación son un clásico en la gestión estratégica de clientes, y no por ser simples son menos efectivos. Consideran las dos dimensiones más importantes en el tratamiento de clientes: el valor de cliente a largo plazo, y el riesgo de abandono.

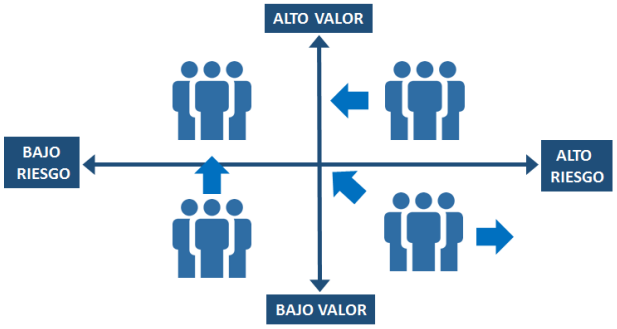


Figura 4 Modelo de segmentación Valor-Riesgo (elaboración propia).

Esta división en cuadrantes (alto valor-bajo riesgo, bajo valor-bajo riesgo, alto valor-alto riesgo, bajo valor-alto riesgo) simplifica las decisiones de desarrollo, retención, fidelización y desinversión en clientes, y permite la cuantificación máxima de la inversión de marketing en cada una de ellas.

Segmentación por Navegación o Ruta de compra:

Sin embargo, en el marketing digital y social, en lo que respecta a la navegación Web, las tácticas de segmentación y "targeting" de clientes son más dinámicas. Según Chaffey , como van estrictamente orientadas a optimizar la conversión de una determinada acción ("click-through", descarga, compra online), sólo se enfocan en la navegación inmediata anterior: origen, canal, sesiones, páginas visitadas [09] (Chaffey, 2018).

Estos modelos de clasificación de clientes nacen para optimizar el embudo de ventas en canales digitales (web, apps, comercio electrónico).

Hacia una segmentación dinámica

Si el consumidor ha cambiado su comportamiento, si los datos disponibles crecen exponencialmente y la velocidad de las interacciones es vertiginosa… ¿Por qué continuamos

utilizando modelos de segmentación basados en los requisitos y datos disponibles en los años 70s?

Si nuestros segmentos de clientes son estáticos e inmutables, no es de extrañar que la mayoría de nuestras iniciativas de marketing sean infructuosas, y que los consumidores nos perciban irrelevantes. Debemos revisitar nuestra estrategia de segmentación de clientes.

Para ello utilizamos un marco de referencia conformado por dos ejes o dimensiones. El eje horizontal representa la flexibilidad de los modelos de segmentación y su capacidad de estos para adaptarse a un entorno cambiante a través del tiempo.

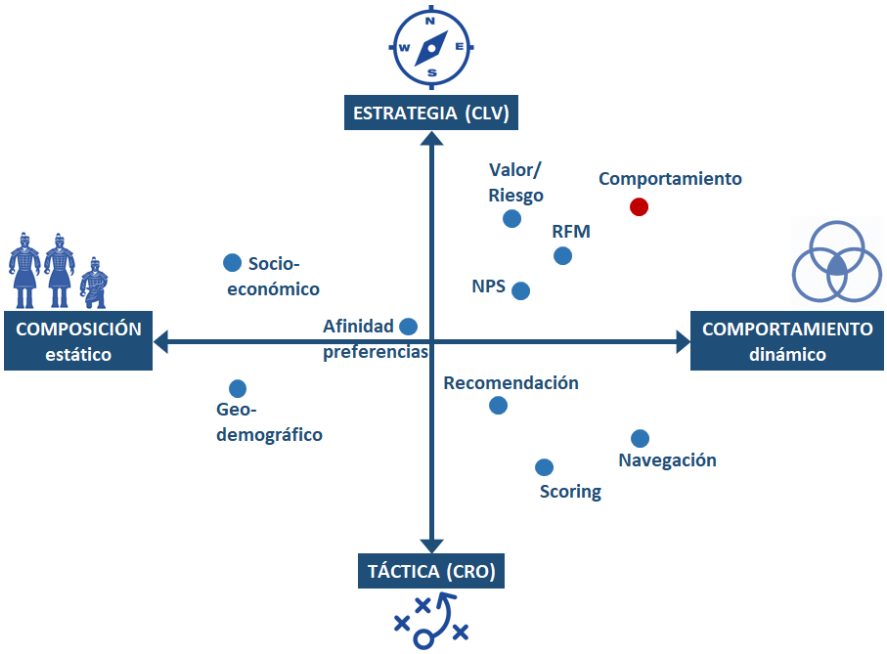


Figura 5 Mapa de tipos de segmentación [10] (Casariego, 2016)

4.2. Paradigmas analíticos de cliente

Directamente vinculados a los modelos de segmentación se encuentran los paradigmas analíticos de cliente. Nos referimos a un paradigma analítico como a una combinación de:

- Generalización: Consistente en un modelo de datos de entrada, un objetivo de análisis, una batería de algoritmos aplicables, y un modelo de los resultados de salida; tal que sean de aplicación generalizada a diferentes modelos de negocio.
- Portabilidad: que el paradigma pueda ser portado a diferentes entornos tecnológicos de almacenamiento (bases de datos) y procesamiento de datos (en la nube o en instalaciones propias), así como también pueda ser explotado con las herramientas analíticas existentes ("business intelligence", "machine learning", etc.) con moderado esfuerzo.

Los siguientes paradigmas son los más popularmente adoptados por los profesionales de la industria:

Análisis multidimensional OLAP de clientes

El análisis multidimensional definido por Kimball, también conocido como OLAP, cumple con la condición de poder ser aplicado a cualquier tipo de negocio de clientes e industrias, en tanto y en cuanto se adapte al paradigma [11] (Kimball & Ross, 2013) [12] (Kimball, 2015):

- Modelo de datos compuesto de múltiples dimensiones y medidas
- Dimensiones organizadas en jerarquías (tiempo, geografía, producto, cliente, etc.)
- Medidas susceptibles de ser agregadas (conteo, suma, promedio, mínimo, máximo, etc.)
- Visualización en tablas dinámicas “pivot” o gráficas
- Navegación “drill down” de jerarquías, y “slice & dice” de dimensiones y medidas
- Filtrado de dimensiones y medidas
- Modelo físico de datos desnormalizado: hechos, estrella, copo de nieve
- Portabilidad: Bases de datos relacionales, multidimensionales (hipercubos), o no estructuradas accesibles vía drivers SQL o MDX compatible.
- Herramientas: Power BI, Tableau software, Qlik, etc.

Análisis de conglomerados de clientes

El análisis de conglomerados de clientes, también conocido como “Clustering”, cumple con la condición de poder ser aplicado a cualquier tipo de negocio, de clientes e industrias, en tanto y en cuanto se adapte al paradigma [13] (Xu & Wunsch, 2005):

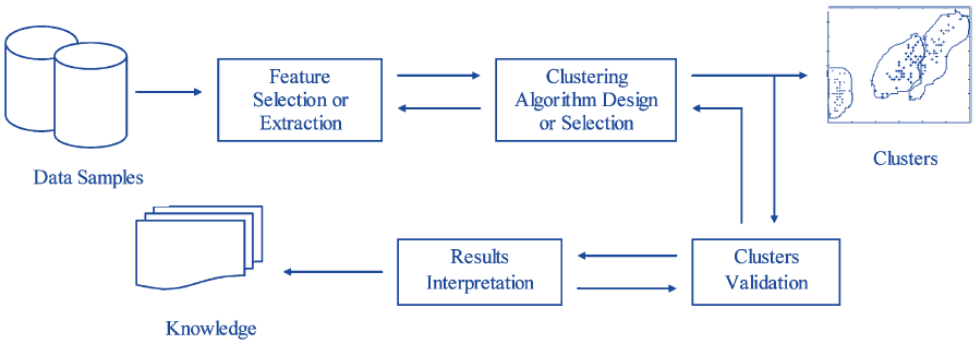


Figura 6 . Proceso de “Clustering” (Xu et al. 2005)

Es un método de aprendizaje no supervisado que permite agrupar clientes en “clusters” o conglomerados, cuyos miembros son, en cierto modo, similares entre sí. Es una colección de

clientes similares entre sí, cuyas características comunes pueden describirse, y resultan diferentes a los clientes que pertenecen a otros “clusters”.

- Modelo de datos compuesto de variables. Típicamente una tabla de clientes, con una fila por cliente, y N columnas con variables asociadas al cliente
- Los algoritmos más populares son: Algoritmo K-means, Agrupamientos jerárquicos, Agrupamientos solapados, Agrupamientos probabilísticos
- En función del algoritmo seleccionado será necesario realizar un tratamiento previo de las variables según sea su tipo
- Portabilidad: Modelo de datos: fichero CSV o JSON, tabla de Base de datos relacional
- Herramientas: Weka, R, SAS, BigML, etc.

Análisis de riesgo de fuga de clientes

Según Linoff & Berry, el análisis de riesgo de fuga de clientes normalmente se basa en algoritmos supervisados de clasificación, del tipo árboles de decisión, y cumple con la condición de poder ser aplicado a cualquier tipo de negocio de clientes e industrias, en tanto y en cuanto se adapte al paradigma [14] (Linoff & Berry, 2011):

- Modelo de datos compuesto de variables y una clase. Típicamente una tabla de clientes, con una fila por cliente, y N columnas con variables asociadas al cliente
- La clase es binaria: fuga-verdadero, fuga-falso
- Los algoritmos posibles son: ID3, CHAID, CART, C4.5
- En función del algoritmo seleccionado será necesario realizar un tratamiento previo de las variables según sea su tipo
- Portabilidad: Modelo de datos: fichero CSV o JSON, tabla de Base de datos relacional
- Herramientas: Weka, R, SAS, BigML, etc.

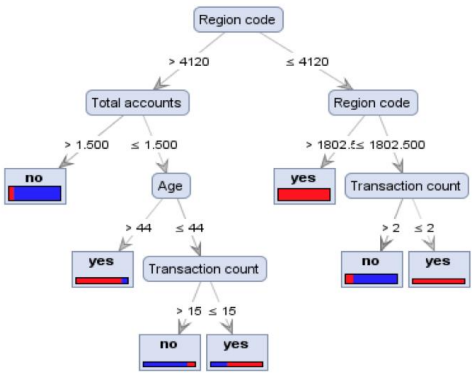


Figura 7 Árbol de decisión: Riesgo de fuga [15] (Anova analytics, 2017)

4.3. Algoritmos de búsqueda de patrones en secuencias

Bajo la hipótesis de trabajo de que el comportamiento observable de un cliente podría representarse mediante una secuencia de acciones y eventos durante el "customer journey", nos enfocamos en los algoritmos especializados en analizar secuencias y en detectar patrones dentro de las mismas.

Minería de patrones secuenciales

Los algoritmos de búsqueda de patrones en secuencias parten de las siguientes características [16] (Fournier-Viger, Chun-Wei, Uday-Kiran, Sing-Koh & Thomas, 2017):

- Disponemos de una base de datos histórica de transacciones de clientes
- Cada transacción consiste en: identificador de cliente, estampado de tiempo de la transacción, ítems comprados en la transacción
- Un cliente no tiene más de una transacción en el mismo momento
- No consideramos las cantidades compradas
- Una lista-de-ítems no puede ser vacía
- Una secuencia es una lista-de-ítems ordenada cronológicamente
- Un ítem se representa por medio de un símbolo único
- Un patrón consiste en una subsecuencia que ocurre con una frecuencia superior a un umbral determinado
- Los algoritmos propuestos para minería de patrones secuenciales son: MaxSP y VMSP
- Los algoritmos propuestos para minería de reglas de patrones secuenciales son: CMRules y ERMiner

Minería de la línea temporal de un paciente

Según Rajkomar, otro abordaje experimental equivalente, consiste en analizar la línea temporal de eventos clínicos de un paciente utilizando una representación secuencial de los eventos provenientes de múltiples fuentes de datos, sumados a un algoritmo estadístico predictivo de "Deep learning" [17] (Rajkomar A. et al., 2018).

Detección de tópicos

Los modelos probabilísticos de detección de tópicos basados en distribuciones de Dirichlet son extensamente utilizados en diferentes contextos para descubrir estructuras ocultas en largos corpus de texto. Bajo la asunción general de que los símbolos son independientes entre sí, se utiliza generalmente el concepto de bolsa de palabras.

Según Barbieri, cuando modelamos secuencias y buscamos patrones en ellas, asumimos que la generación de tópicos interpreta que un determinado símbolo que representa a un evento puede depender de un símbolo anterior en la secuencia. Estos modelos extendidos utilizan el concepto de N-gramas (secuencias de símbolos de longitud 1, 2, …, hasta N) para obtener conglomerados de tópicos más precisos [18] (Barbieri N. et al., 2013).

4.4. Técnicas de visualización de secuencias

Una parte importante de la analítica de secuencias y patrones de comportamiento observado del cliente, es la visualización y exploración de estos a través de herramientas gráficas e interactivas. Lo que sucede es que las secuencias no son una entidad de estudio muy habitual, por lo que requiere de técnicas de visualización particulares.

La siguiente es una lista de librerías de visualización aplicables a secuencias con la característica de ser de código abierto.

Diagrama Event drops

Esta librería basada en D3.js permite visualizar y navegar secuencias de eventos en diferentes planos. Permite desplazamientos laterales en la línea temporal, y "zoom" temporal hasta llegar a nivel de día y hora. Disponible bajo licenciamiento MIT, por cortesía de Marmelab y Canal Plus [19] (Marmelab), con código fuente disponible en el repositorio Github.

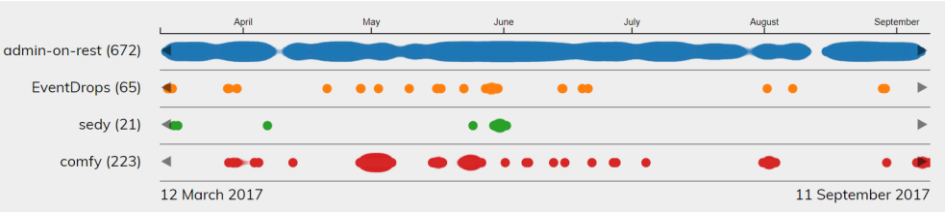


Figura 8 Visualización "EventDrops" (Marmelab)

Diagrama Collapsible tree

Esta librería basada en D3.js permite visualizar y navegar secuencias de eventos expandiendo o colapsando en profundidad las hojas del árbol. Disponible bajo licenciamiento BSD-3-Clause [20] (Bostok, 2018), con código fuente disponible en el repositorio Github.

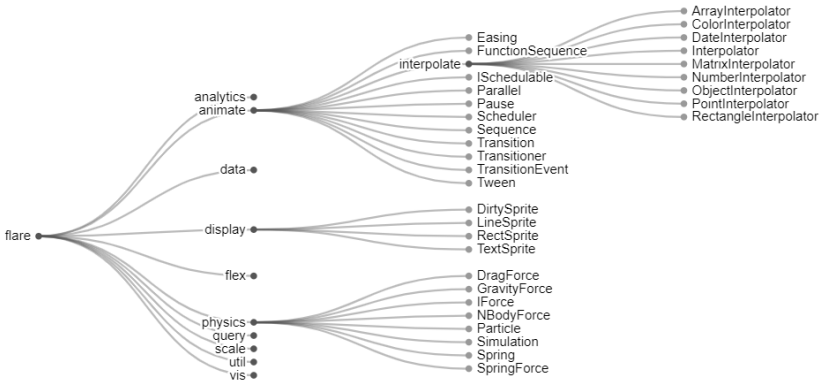


Figura 9 Visualización "Collapsible Tree" (Bostok, 2018)

Diagrama "Sequences sunburst"

Esta librería basada en D3.js permite visualizar y navegar secuencias de eventos en forma de rayos de sol, describiendo la secuencia seleccionada dinámicamente con el cálculo de porcentaje de cobertura de cada rayo. Disponible bajo licenciamiento Apache [21] (Roden, 2019), con código fuente disponible en el repositorio Github.

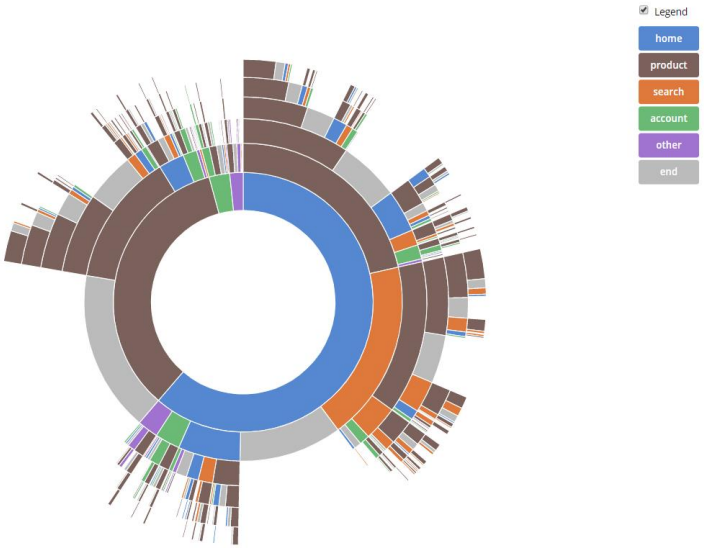


Figura 10 Visualización "Sequences sunburst" (Roden, 2019)

5. Objetivos de la metodología MSC2

Objetivo general:

El objetivo general de la nueva **Metodología MSC2 de análisis y segmentación de clientes usando secuencias de comportamiento** consiste en proporcionar a los decisores de marketing y gestión de clientes una nueva herramienta que complemente y amplíe las actuales prácticas relativas a la segmentación de clientes.

Buscamos establecer una metodología generalizable y repetible que permita generar conocimiento de clientes a partir de la observación de su comportamiento dinámico y omnicanal.

Nuestro objetivo es generar un modelo novedoso de representación, análisis y segmentación de clientes basado en comportamiento, que facilite interactuar con ellos, descubrir patrones de comportamiento, identificar su evolución, y anticiparse al mismo.

Objetivos específicos:

Para ello definimos los siguientes objetivos específicos:

Proceso generalizable

Establecer un proceso metódico generalizable para la: Captación, Representación, Almacenamiento, Análisis, Segmentación, Visualización y Anticipación del **comportamiento observado de clientes**; a través del paso del tiempo, y a través de los múltiples canales/puntos de contacto.

Entendemos el término “generalizable” como de posible aplicación a diferentes industrias y sectores: comercio electrónico, retail, banca, telecomunicaciones, turismo, ocio, medios, sanidad, programas de fidelización, etc. Así como a dinámicas de clientes, ciudadanos y pacientes.

Segmentos basados en comportamiento

Complementar la actual segmentación de clientes (CLV, RFM, NPS, Valor-Riesgo, Ruta de compra); con nuevos **segmentos de clientes basados en el comportamiento** dinámico de los clientes y sus componentes: “customer journey”, omnicanalidad, y ciclo de vida. Contrastar los segmentos actuales versus los segmentos basados en comportamiento.

Analítica y visualización de secuencias y patrones de comportamiento

Expandir el actual análisis de clientes (OLAP, Conglomerados, Riesgo de fuga, Recomendaciones); al incorporar nuevas **clasificaciones y visualizaciones basadas en secuencias de comportamiento**:

- Detección de patrones en secuencias
- Análisis de tópicos en secuencias y patrones
- Visualización de secuencias y patrones

6. Guía de trabajo

Al tratarse de una metodología eminentemente orientada a la segmentación y el análisis de clientes, utilizaremos como marco de referencia la guía de trabajo CRISP-DM, que indica como desplegar un proceso estándar de minería de datos en múltiples industrias [22] (CRISP-DM.org, 1996).

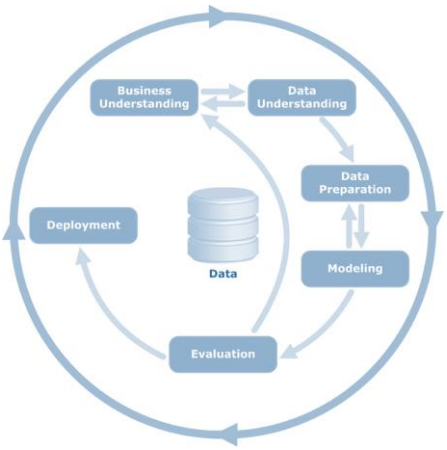


Figura 11 Metodología CRISP-DM (Crisp_DM.org, 1996)

Consistente en las siguientes fases:

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado y análisis
- Evaluación de resultados
- Despliegue de prototipo

Una vez descrita la metodología según las diferentes fases, haremos el despliegue de la metodología aplicándola a dos casos prácticos diferentes, de forma de poner a prueba su propiedad de ser generalizable. Finalmente aplicaremos la evaluación de resultados en ambos casos de aplicación.

7. Contribución de la Metodología MSC2

La contribución novedosa de la metodología MSC2 consiste en:

- Definir un **marco conceptual completo para el uso generalizado** de secuencias y patrones de comportamiento en el contexto del Marketing para el análisis y segmentación

de clientes, ciudadanos y pacientes. Dicho marco conceptual es producto de mi experiencia profesional gestionando soluciones tecnológicas y analíticas de marketing y cliente para compañías multinacionales en diferentes sectores e industrias.

- **Reutilizar, alinear y formalizar** de forma consistente una extensa **serie de mejores prácticas** ya utilizadas en diferentes campos de actuación, pero desconectadas entre sí:
    - Algoritmos de búsqueda de patrones frecuentes en secuencias de eventos
    - Paradigmas analíticos predictivos y de conglomerados
    - Técnicas de visualización de secuencias y patrones
- **Ampliar el conjunto de métricas e indicadores tradicionales de cliente**, añadiendo la componente "customer journey" con todo su contenido temporal (del ciclo de vida y el viaje del cliente a través de N periodos de tiempo), **sintetizado en una nueva familia de atributos** que contiene secuencias, patrones y tópicos de comportamiento de cliente. Estas nuevas métricas e indicadores pueden **ser incorporados** sin ningún esfuerzo a la infraestructura analítica de "business intelligence" existente en las organizaciones: informes, modelos, cuadros de mando, etc.
- **Expandir** las actuales capacidades de **análisis, segmentación, visualización y predicción del comportamiento de clientes** para la toma de decisiones por parte de los responsables de marketing y cliente. Estas capacidades adicionales se incorporan de forma natural y consistente a las prácticas habituales de **"targeting" de clientes y gestión de campañas** de marketing. Con la ventaja de que se incorporan capacidades de **visualización y análisis predictivo** que no eran posibles con anterioridad.

8. Descripción de la metodología MSC2

La nueva **Metodología MSC2** de análisis y segmentación de clientes usando secuencias de comportamiento consiste en la aplicación sistemática y ordenada de las siguientes etapas:

- Un modelo conceptual de aplicación universal (previamente desarrollado por el autor)
- Una estructura de datos genérica (entrada, almacenamiento y salida)
- Los procesos de transformación y carga
- La aplicación de algoritmos de detección de patrones en secuencias usando librerías de "SPMF Sequential Pattern Mining Framework" de código abierto
- Los metadatos de dichas secuencias y patrones que describen las nuevas entidades, así como sus métricas, permiten enriquecer el modelo de análisis y segmentación de clientes en base a su comportamiento observado
- La aplicación de técnicas de "machine learning" para agrupar y clasificar las secuencias y patrones temporales

- El análisis de los resultados utilizando herramientas de “business intelligence” estándar.
- La creación de gráficas de visualización interactivas para navegar las secuencias y patrones de comportamiento.
- La aplicación de algoritmos de detección de reglas predictivas en secuencias usando librerías de “SPMF Sequential Pattern Mining Framework” de código abierto.

8.1. Comprensión del negocio

Las organizaciones se enfrentan al nuevo consumidor (cliente, ciudadano, paciente), cuyo comportamiento es completamente diferente a lo que estábamos habituados.

Una forma gráfica de entenderlo son los mapas del viaje del consumidor [23] (Conder, 2015) que detallan el tránsito del cliente a través de los diferentes canales y puntos de contacto, realizando diferentes acciones e interacciones a lo largo del tiempo.

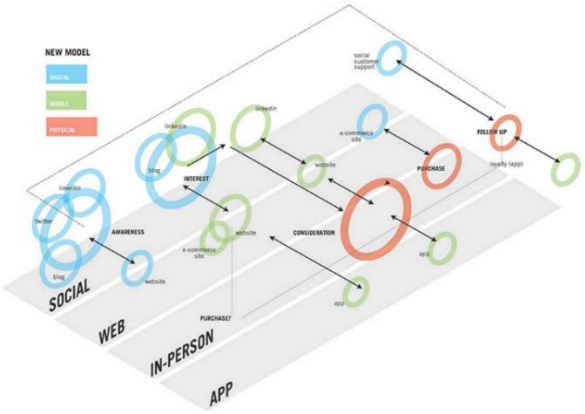


Figura 12 “Customer journey map” (Conder, 2015)

Los mecanismos de segmentación y análisis que utilizamos actualmente no son suficientes para describir, analizar y anticipar las necesidades y comportamiento del nuevo cliente.

8.2. Comprensión de los datos: Perfil de cliente 360

Desde el punto de vista de la comprensión de los datos necesarios para cumplimentar las necesidades del negocio, nos remitimos al concepto Perfil de cliente 360 grados.

Afortunadamente a partir del fenómeno “big data” las organizaciones ya cuentan o están en proceso de contar con un repositorio de datos de cliente que dé cobertura a las necesidades analíticas y de toma de decisiones.

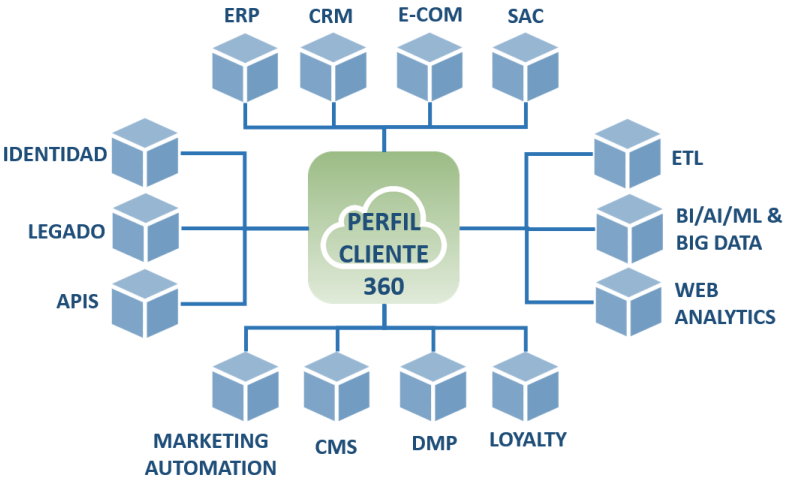


Figura 13 Perfil de cliente 360 (elaboración propia)

Independientemente de la forma de implementarlo ("Data warehouse", "Data lake", etc.), podemos asumir que las organizaciones disponen de la mayoría de los fuentes de datos de cliente:

- Perfil Cliente 360: una visión unificada 360 grados del perfil del cliente
- Identidad: el cliente identificado unívocamente en todos los sistemas y canales
- ERP: datos transaccionales de los clientes en los sistemas de gestión
- CRM: datos transaccionales de la gestión de la relación con los clientes
- E-com: datos transaccionales de las compras de comercio electrónico
- SAC: datos de las interacciones de los clientes con el SAC
- Marketing automation: datos de las comunicaciones de marketing con los clientes
- CMS: datos de los contenidos visitados por los clientes cuando están identificados
- DMP: datos de terceros asociables a nuestros clientes cuando están identificados
- Loyalty: datos transaccionales del programa de fidelización
- Legacy: datos transaccionales de los clientes en los sistemas legados
- APIS: datos transaccionales de los clientes en plataformas "Software-as-a-service"
- ETL: datos transaccionales de los clientes en sistemas disgregados
- Web analytics: datos de la navegación web de los clientes cuando están identificados
- BI, AI, ML, Big data: conjunto de métricas e indicadores analíticos de los clientes

8.3. Representación del comportamiento

Con el objetivo de avanzar sobre un nuevo modelo MSC2 de segmentación de clientes basado en su comportamiento, es necesario definir y establecer un **modelo de representación de dicho comportamiento observado**.

Definimos un modelo que cumpla con las siguientes propiedades:

- Generalización: Que sea apto para representar cualquier modelo de negocio centrado en el cliente que exhiba comportamientos repetitivos
- Cronología: Que sea apto para representar secuencias temporales cronológicas de acciones, eventos y métricas representativos del comportamiento
- Portabilidad. Que dicho modelo pueda ser portado a los entornos de almacenamiento y procesamiento de datos, así como a las herramientas analíticas más populares

Por otra parte, el modelo de representación del comportamiento debe incluir los siguientes capítulos de información estratégica y táctica de cliente que son universales en cualquier modelo de negocio o industria:

Transacciones

MSC2 incluye en este capítulo a todas aquellas transacciones donde el cliente está identificado, existe una marca temporal, y existe un intercambio de bienes, servicios, dinero, descuento, información o recompensa.

Ejemplos de estas transacciones son:

- Compra o pago de producto o servicio
- Contrato, suscripción o pago de servicio
- Redención de un descuento o promoción
- Servicio de atención al cliente
- Renovación o cancelación de servicio
- Devolución de producto, reintegro de pago de producto, servicio o suscripción
- Registro de datos personales, o aceptación o remoción de permisos RGPD
- Activación de una cuenta
- Encuesta de satisfacción
- Incidencia de reclamación
- Acumulación, canje o caducidad de puntos de recompensa
- Etc.

Interacciones

MSC2 incluye en este capítulo a todas aquellas interacciones donde el cliente está identificado, existe una marca temporal, y existe una manifestación expresa de interés por parte del cliente por interactuar con la empresa.

Ejemplos de estas interacciones son:

- Visita a una tienda física u oficina
- Visita a una tienda electrónica
- Visita a sus activos digitales: web, app, blog, chatbot, redes sociales
- Suscripción a un boletín electrónico o catálogo
- Descarga de un cupón, contenido o catálogo
- Búsqueda de un contenido, producto o servicio
- Comentario o valoración de un contenido o experiencia
- Compartir un contenido o experiencia en redes sociales
- Etc.

Comunicaciones

MSC2 incluye en este capítulo a todas aquellas comunicaciones operativas, comerciales, promocionales, o informativas, a través de cualquier canal, ya sean salientes o entrantes; donde el cliente está identificado, y existe una marca temporal.

Ejemplos de estas comunicaciones son:

- Comunicación de bienvenida
- Comunicación operativa
- Comunicación de felicitación
- Comunicación de agradecimiento
- Comunicación informativa de producto o servicio
- Comunicación del estado de cuenta
- Comunicación del estado de recompensas
- Comunicación del estado del servicio o suscripción
- Comunicación promocional con o sin incentivo
- Etc.

Métricas

Los cambios en las métricas e indicadores que se utilizan para monitoreo periódico de los clientes y el negocio incorporan conocimiento del negocio al modelo.

MSC2 incluye en este capítulo a todas aquellas métricas e indicadores de desempeño de cliente que se calculan y refrescan periódicamente, en las cuales el cliente está identificado, y existe una marca temporal relativa al momento de refresco o recálculo.

Ejemplos de estas métricas e indicadores son:

- Cambio en el indicador de valor CLV (“Customer Lifetime Value”)

- Cambio en el indicador de satisfacción NPS (“Net Promoter Score”)
- Cambio en el indicador de actividad RFM (Recencia, Frecuencia, Monetización)
- Etc.

Segmentos

Los mecanismos de segmentación clásica de clientes (socioeconómico, actitudinal, afinidad, valor, riesgo, etc.) son otra manera adicional de incorporar el conocimiento del negocio dentro del modelo.

MSC2 incluye en este capítulo a todas aquellas segmentaciones de cliente que se calculan y refrescan periódicamente, en las cuales el cliente está identificado, y existe una marca temporal relativa al momento de recálculo. Ejemplo de estos eventos son:

- Cambio de valor de segmento

Contexto

Finalmente, MSC2 incorpora el contexto a través de las variables externas que no son controladas por la organización, pero que pueden influir en el comportamiento observado de los clientes. En este caso no existe un cliente identificado, pero sí contamos normalmente con una marca temporal y una región geográfica.

Ejemplos de cambios de contexto son:

- Cambio en el contexto competitivo o económico
- Cambio en el contexto climático o estacional
- Impacto de campañas de publicidad o acciones de la competencia
- Etc.

Semántica

Todas y cada una de las transacciones, interacciones, comunicaciones, métricas y segmentos poseen un significado de negocio (semántica). La metodología MSC2 enriquece el modelo indicando si se trata de un evento o acción de tipo positivo, neutro o negativo.

Ejemplos de semántica de negocio son:

- Positivo: Contrato o suscripción de servicio. Compartir un contenido en redes sociales
- Neutro: Comunicación operativa. Visita a la app
- Negativo: Encuesta de satisfacción NPS con resultado “detractor”. Reclamación

8.4. Comprensión de los datos: Modelo de datos MSC2

El modelo conceptual MSC2 de entrada consiste en un modelo analítico multidimensional de estructura de estrella, dando soporte de esta forma a tres tablas de hechos:

- Eventos y acciones: Almacenando todos los datos correspondientes a transacciones, interacciones y comunicaciones.
- Métricas y segmentos: Almacenando todos los datos correspondientes a las variaciones en métricas e indicadores, así como los cambios de segmentos.
- Contexto: Almacenando todos los datos correspondientes a situaciones de contexto: negocio, ambiente, competencia, campañas, etc.

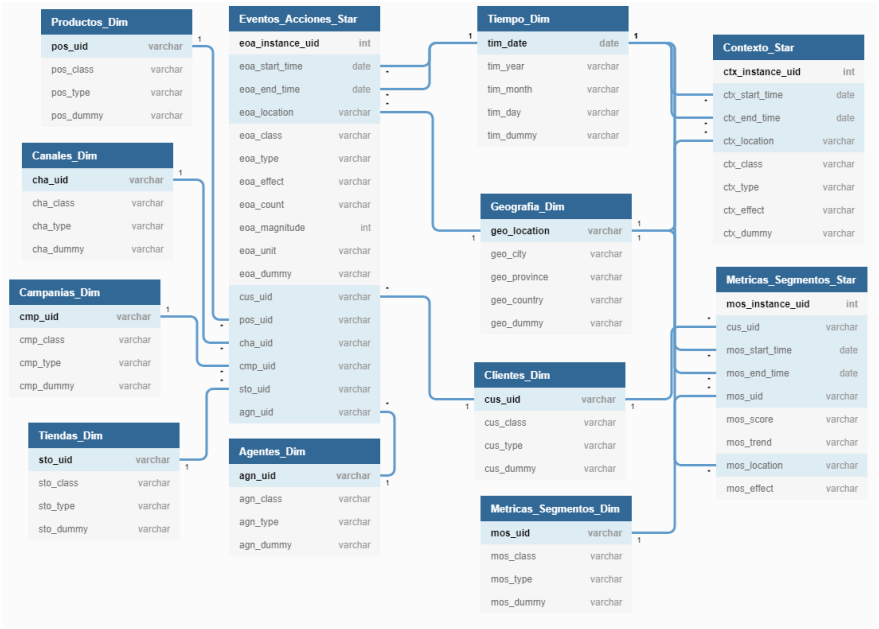


Figura 14 Modelo de datos conceptual de entrada MSC2 (diseñado con dbdiagram.io)

Completando el modelo tenemos las tablas de dimensiones, dando soporte de esta forma a las siguientes dimensiones de análisis:

- Clientes
- Productos o servicios
- Campañas de marketing
- Canales
- Tiendas
- Agentes
- Métricas o segmentos
- Geografía
- Tiempo

El modelo conceptual MSC2 de salida donde finalmente se almacena el comportamiento observado del cliente se describe a partir de las siguientes **entidades** de estudio: secuencias, tópicos de secuencias, patrones, tópicos de patrones y segmentos.

La jerarquía de agregaciones del Modelo de datos de salida es la siguiente:

- Eventos: Consiste en el nivel más básico, representa la ocurrencia de un evento, acción, cambio de métrica o segmento de un cliente en un momento dado.
- Secuencias_Día: Es el nivel inmediato superior donde son agregados todos los eventos que ocurren en el mismo momento para el mismo cliente.
- Secuencias_Cliente: Es el siguiente nivel inmediato superior donde son agregadas todas las secuencias diarias para el mismo cliente durante todo el intervalo de tiempo considerado.
- Tópicos_Secuencias: Es el último nivel de agregación donde todas las secuencias son agrupadas en conglomerados con patrones de comportamiento similares, denominados tópicos de secuencias.

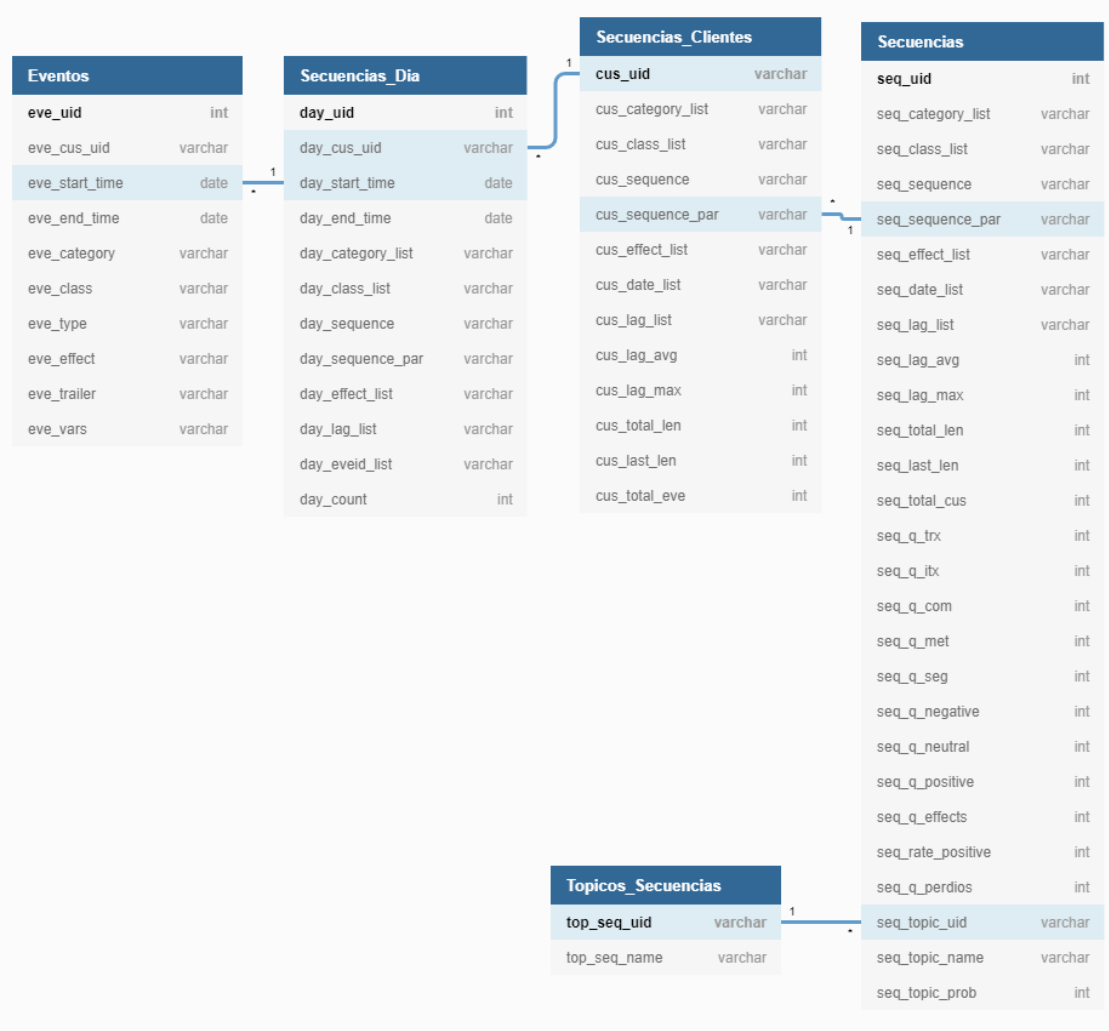


Figura 15 Modelo de datos de salida MSC2 (diseñado con dbdiagram.io)

El modelo conceptual MSC2 conforma las siguientes **entidades** de estudio:

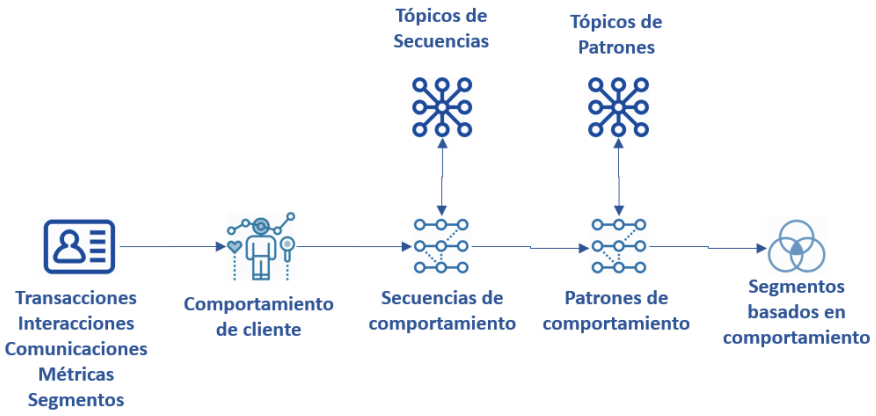


Figura 16 Entidades representadas en el modelo de datos de salida MSC2

8.5. Comprensión de los datos: Metadatos MSC2

Los metadatos son un componente vital de la metodología MSC2 puesto que es lo que permite su generalización. A través de los metadatos representamos a diferentes organizaciones reaprovechando el mismo marco de trabajo y las mismas herramientas.

Codificación simbólica MSC2 en palabras

La siguiente imagen representa el viaje de un cliente a lo largo del tiempo, donde observamos las trazas que va dejando en cada paso, en cada canal, en cada transacción, interacción, comunicación y métrica:

Figura 17 Viaje de un cliente X

La codificación en **palabras** es la traducción del modelo de datos de cada organización en la **gramática** del lenguaje que entiende la metodología de segmentación basada en comportamiento.

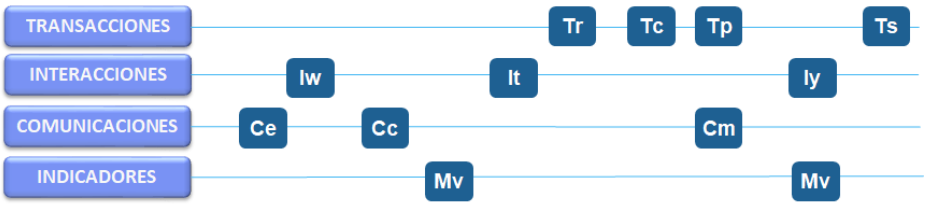


Figura 18 Viaje de un cliente X codificado en palabras

Constituyen los verdaderos **átomos** sobre los que se construyen las **secuencias y patrones** de comportamiento. Las **secuencias** de eventos traducidas a **frases** contienen las palabras que determinan el **recorrido único** de cada cliente durante su "customer journey".

La siguiente imagen representa el viaje de tres clientes a lo largo del tiempo, donde observamos que las trazas que van dejando en cada paso son parecidas, aunque tienen secuencias levemente diferentes. Si somos flexibles a pequeñas alteraciones en la distancia entre eventos, y en la omisión de una pequeña cantidad de ellos, parecerán patrones de comportamiento similares:

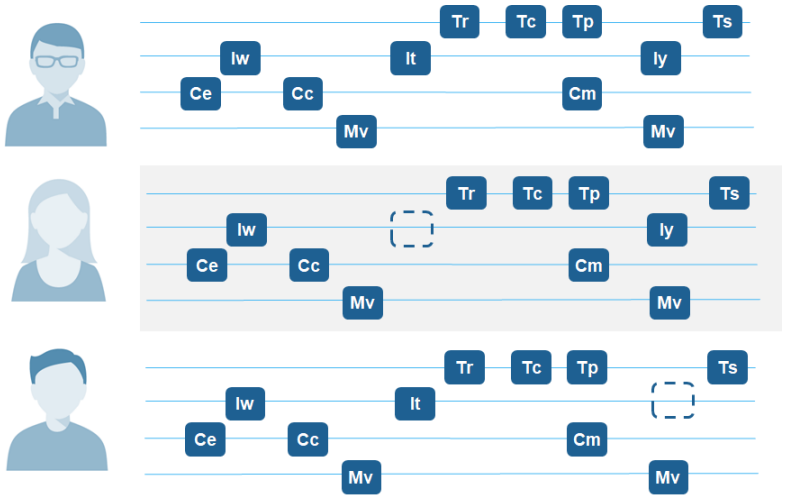


Figura 19 Viaje de tres clientes X, Y y Z codificado en palabras

Los **patrones** de secuencias de eventos traducidas en **frases** contienen las palabras que determinan el **recorrido común** de un **segmento o aglomerado** de clientes con un **comportamiento similar**.

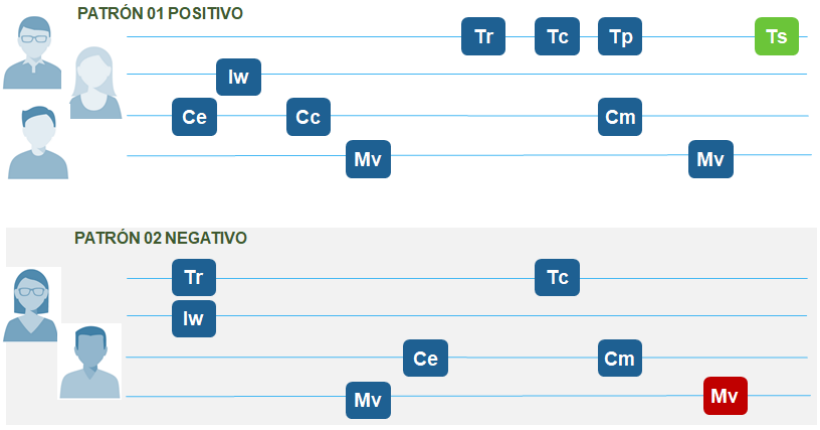


Figura 20 Dos patrones diferentes que agrupan dos segmentos de clientes

La metáfora consistente en utilizar una **gramática** para describir un comportamiento observado, nos da la **flexibilidad** que necesitamos para codificar, almacenar, analizar,

visualizar y entender los patrones de comportamiento de **modelos de negocio** completamente **distintos**.

Codificación simbólica de ejemplo

Supongamos que nuestro escenario trata de un negocio de venta de productos a través de canales físicos y online directo al consumidor.

Por ejemplo, extraemos de los diferentes sistemas de información los siguientes datos de cliente: compras, encuestas de satisfacción, visitas a nuestra App, interacciones en nuestro perfil de Facebook, campañas de email, métricas y segmentos de cliente.

Comenzamos codificando las compras (“buy”), las encuestas de satisfacción (“nps”), el uso de la App (“app”), una interacción en Facebook (“lik”). Continuamos codificando los envíos de email (“ema”), y los cambios en las métricas de cliente (“rfm”) y segmento de fidelización (“loy”). Al codificar cada evento o acción observado del cliente, podríamos obtener secuencias básicas del tipo:

```
ema/ buy/ ema/ nps

ema/ ema/ app/ buy/ ema

ema/ buy/ buy/ rfm

ema/ app/ lik/ buy/ buy/ loy
```

Donde la “/” es el delimitador de eventos o acciones dentro de una secuencia.

Adicionalmente contamos con atributos distintivos de cada uno de los eventos, acciones, métricas o segmentos que estamos codificando. Por ejemplo, la compra podría ser por el canal internet (“#ci”) o en tienda (“#ct”). En ese caso codificamos de forma extendida la palabra del evento seguida de su parámetro. Y podríamos obtener secuencias extendidas del tipo:

```
buy+#ci/ buy+#ct
```

Optimización de la codificación simbólica MSC2

La codificación simbólica MSC2 de palabras escogida **determina directamente la cardinalidad** de los **eventos, secuencias y patrones**. Dicho de otra forma, la gramática escogida nos determina la expresividad del lenguaje.

Una codificación minimalista puede acabar en una gramática de eventos muy limitada, que genera muy poca variedad de secuencias, que discrimina muy poco en cuanto al comportamiento observado de clientes, aunque tenga la propiedad de generar segmentos con gran número de clientes en él.

Una codificación extensiva, por el contrario, puede acabar en una gramática de eventos muy variada, que genera mucha cantidad de secuencias diferentes muy difíciles de agrupar, y da origen a una gran cantidad de segmentos con muy pocos clientes en cada uno.

El equilibrio consiste en obtener una cantidad de secuencias, patrones y segmentos, que sean manejables desde el punto de vista de marketing y negocio.

En cada caso nos enfrentamos a un proceso iterativo de ensayo y error hasta que logramos dar con la codificación simbólica óptima de cada modelo de negocio y cada conjunto de datos particular. Al momento podemos afirmar que las siguientes características y parámetros tienen un impacto decisivo en la configuración de la codificación simbólica óptima:

- Cantidad de secuencias que exceden el límite establecido de secuencias
- Cantidad de secuencias con menos clientes que el límite establecido por secuencia
- Cantidad de secuencias con más clientes que el límite establecido por secuencia
- Cantidad de secuencias con largo menor que el límite establecido por secuencia
- Cantidad de secuencias con largo mayor que el límite establecido por secuencia
- Cantidad de secuencias con porcentaje de positivos superior al límite establecido
- Cantidad de secuencias con palabras objetivo (semántica de negocio)
- Cantidad de secuencias acabadas en palabras objetivo (semántica de negocio)

8.6. Preparación de los datos: Flujo de procesos MSC2

Hasta el momento de esta memoria, el 50% de los procesos de datos MSC2 se automatizan y el 50% son procesos manuales de tratamiento de datos.

El flujo completo de procesos y transformaciones de principio a fin está formado por los siguientes pasos:

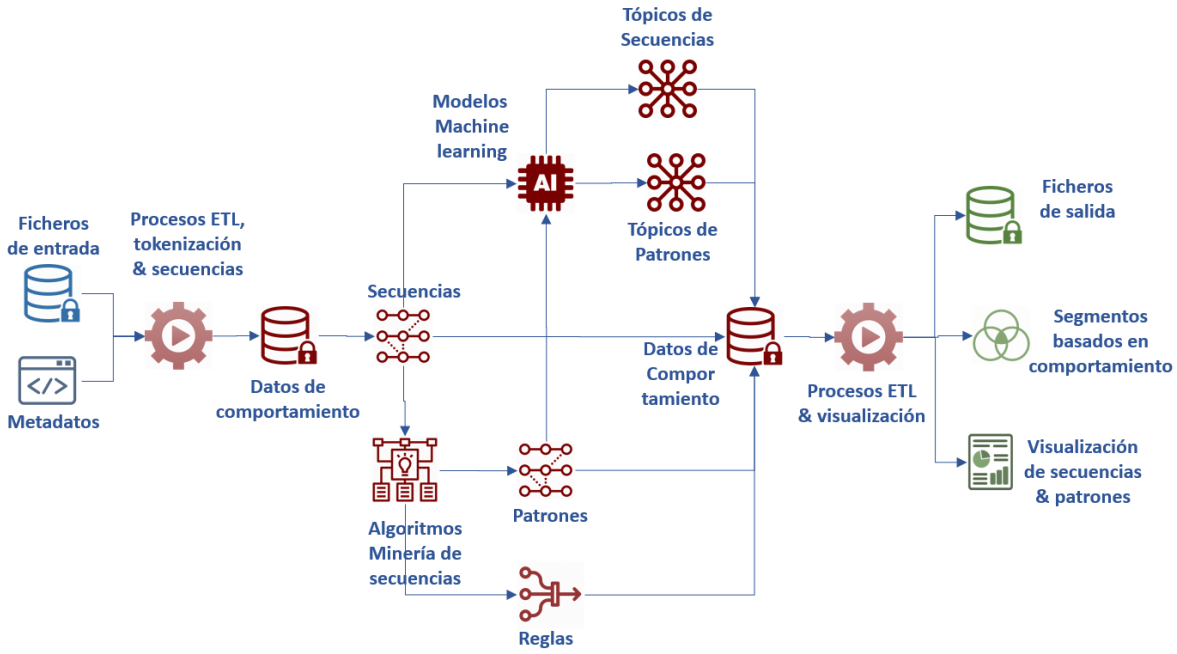


Figura 21 Flujo de procesos MSC2 completo

Procesos ETL de entrada

Este proceso ETL de extracción, transformación y carga recibe como entrada tres ficheros desnormalizados en formato CSV, y un fichero de metadatos en formato JSON, para alimentar el modelo de datos de entrada.

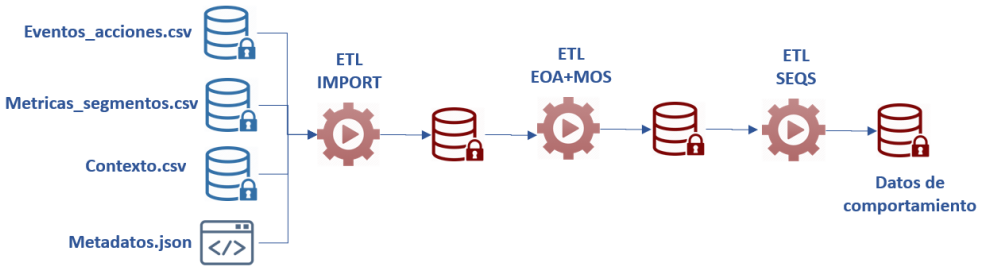


Figura 22 : Procesos ETL de entrada MSC2

Este proceso está desarrollado íntegramente en código Redshift SQL aprovechando la potencia del motor de bases de datos relacional para procesar agregaciones de grandes volúmenes de datos.

Los subprocesos son:

- **Subproceso ETL IMPORT**: Importa los ficheros CSV desde AWS S3 a las tablas de trabajo en AWS Redshift.

- **Subproceso ETL EOA+MOS**: Organiza y transforma a un formato común todas las transacciones correspondientes a Eventos, Acciones, Métricas y Segmentos, ordenadas por cliente, "timestamp", categoría, clase y tipo.
- **Subproceso ETL SEQS**: Construye las secuencias por día y por cliente a partir del total de Eventos, Acciones, Métricas y Segmentos. Calcula también las medidas relativas a la longitud, espaciado entre eventos, cantidad de eventos y cobertura de cada secuencia.

Algoritmos de minería de secuencias

Estos algoritmos reciben como entrada un fichero de secuencias de clientes en formato CSV, y un fichero de metadatos en formato JSON, para descubrir y generar subsecuencias que funcionan como patrones comunes de las secuencias de entrada; y reglas predictivas sobre secuencias aprendidas.

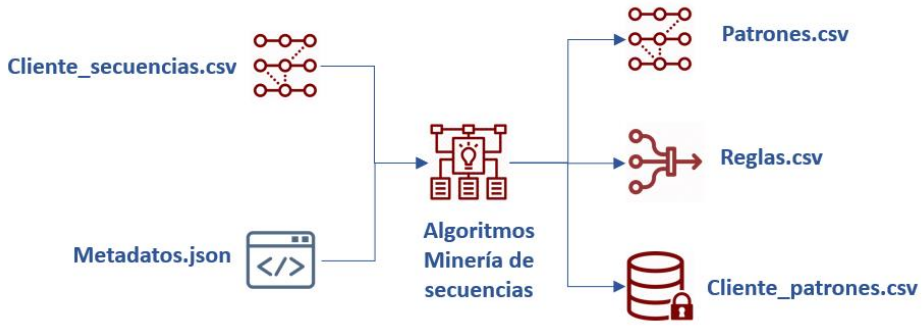


Figura 23 Proceso de minería de secuencias MSC2

Estos algoritmos forman parte de la librería que implementa modelos de "Sequential pattern mining" (Fournier-Viger, 2017) de código abierto escrito en Java.

Los subprocesos son:

- **Subproceso PATRONES**: Genera los patrones a partir de los parámetros de entrada, dejando a la salida un fichero en formato CSV: Patrones.csv
- **Subproceso CLIENTE_PATRONES**: Genera la relación de patrones asociados a cada cliente, dejando a la salida un fichero en formato CSV: cliente_patrones.csv
- **Subproceso CLIENTE_REGLAS**: Genera la relación de reglas predictivas de secuencias asociadas a cada cliente, dejando a la salida un fichero en formato CSV: cliente_reglas.csv

Modelos de "Machine learning" de Tópicos

Este algoritmo recibe como entrada un fichero de secuencias de comportamiento en formato CSV, y un fichero de patrones de comportamiento en formato CSV, para descubrir y generar conglomerados de secuencias y patrones basados en "Topic modeling" de la plataforma de software como servicio de BigML [24] (BigML, 2017).

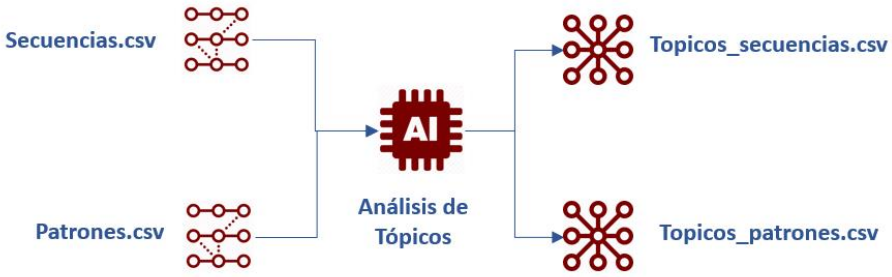


Figura 24 Modelos ML detección de tópicos en secuencias y patrones MSC2

Los subprocesos son:

- **Subproceso Tópicos de secuencias**: Genera los tópicos agrupando las secuencias de comportamiento de entrada, dejando a la salida un fichero en formato CSV: Tópicos_secuencias.csv
- **Subproceso Tópicos de patrones**: Genera los tópicos agrupando los patrones de comportamiento de entrada, dejando a la salida un fichero en formato CSV: Tópicos_patrones.csv

Procesos ETL de salida

Este proceso ETL de extracción y transformación, recibe como entrada la Base de datos de Comportamiento en AWS Redshift, y genera a la salida diferentes ficheros CSV para ser exportados y analizados en plataformas de "business intelligence" y visualización de secuencias.

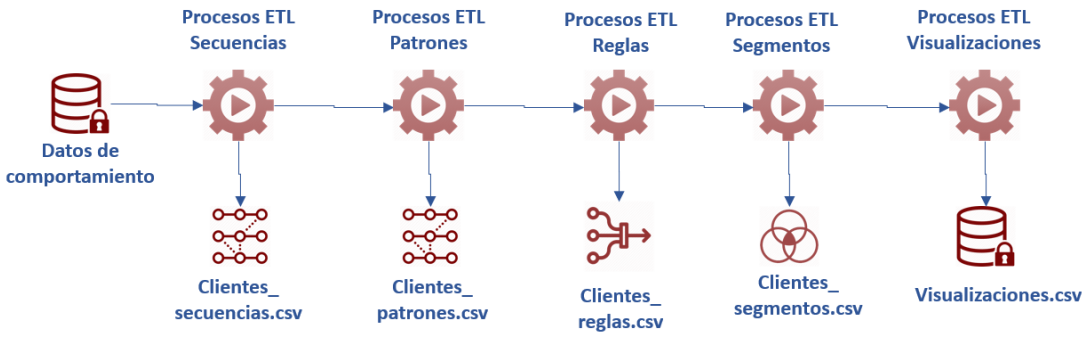


Figura 25 Procesos ETL de salida MSC2

Este proceso está desarrollado parcialmente en la herramienta de ETL de Tableau Prep [25] (Tableau Prep, 2018), y se completa con procesos manuales basados en macros.

Los subprocesos son:

- **Subproceso ETL Secuencias**: Exporta las secuencias de clientes a un fichero de salida en formato CSV: Clientes_secuencias.csv

- **Subproceso ETL Patrones**: Exporta los patrones de clientes a un fichero de salida en formato CSV: Clientes_patrones.csv
- **Subproceso ETL Reglas**: Exporta las reglas de clientes a un fichero de salida en formato CSV: Clientes_reglas.csv
- **Subproceso ETL Segmentos**: Exporta los segmentos de clientes a un fichero de salida en formato CSV: Clientes_segmentos.csv
- **Subproceso ETL Visualizaciones**: Exporta las secuencias, patrones y segmentos de clientes a un fichero de salida en formato CSV: Visualizaciones.csv

8.7. Preparación de los datos: Infraestructura y herramientas de MSC2

El siguiente esquema describe a nivel general la Infraestructura tecnológica utilizada y las herramientas de software empleadas en la metodología MSC2.

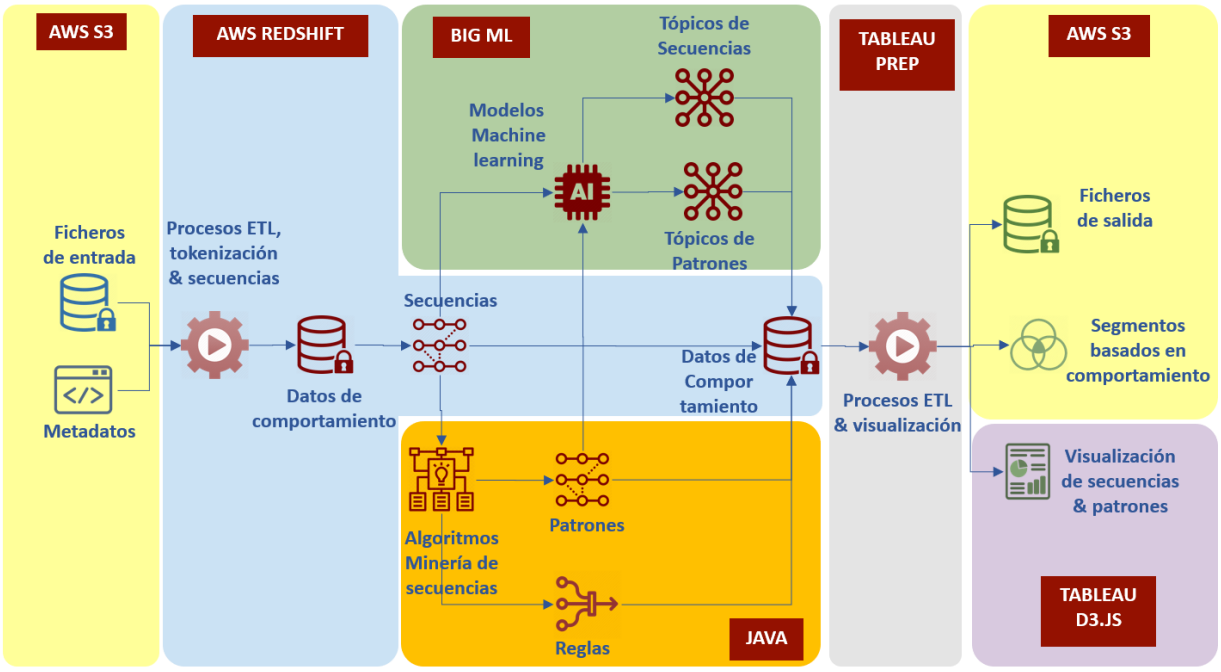


Figura 26 Infraestructura tecnológica y herramientas de software MSC2

La infraestructura y herramientas de software utilizadas son las siguientes:

- **AWS S3**: Servicio de ficheros “big data” en la nube de AWS Amazon web services
- **AWS Redshift**: Servicio de RDBMS “big data” en la nube de AWS Amazon web services
- **BigML**: Plataforma de “machine learning” en la nube ofrecido en modalidad de software como servicio

- **Entorno Java**: Entorno virtualizado con Oracle VirtualBox, con sistema operativo Linux, máquina virtual de Java, y librería SPMF de código abierto java
- **Tableau Prep**: Aplicación local de ETL de la familia Tableau software
- **Tableau**: Aplicación local de "business intelligence" de la familia Tableau software
- **D3.JS**: Librerías javascript de código abierto para visualización de secuencias con gráficas: Diagramas "Event drops", "Collapsible tree" y "Sequences sunburst"

8.8. Modelado MSC2: Secuencias y patrones de comportamiento

La **secuencia** constituye la primera entidad de estudio dentro del modelo de representación del comportamiento de clientes MSC2.

De Eventos a Secuencias de Comportamiento

La generación de secuencias es el último paso del proceso de ETL de entrada y se obtienen por medio de una **función de agregación de atributos de tipo texto**, al agrupar en orden cronológico todos los eventos, acciones, métricas y segmentos ya codificados de cada cliente.

```
680
681 insert into etl_cus_bas
682 select
683     bama_cusuid,
684     listagg(day_categories, '/') within GROUP (ORDER BY bama_start_time) as all_categories,
685     listagg(day_classes, '/') within GROUP (ORDER BY bama_start_time) as all_classes,
686     listagg(day_types, '/') within GROUP (ORDER BY bama_start_time) as all_types,
687     listagg(day_types_par, '/') within GROUP (ORDER BY bama_start_time) as all_types_par,
688     ...
689     sum(day_bass) as all_bass,
690     count(*) as all_records,
691 from
692     etl_bas_day_lag
693 group by
694     bama_cusuid
695 order by
696     bama_cusuid
697 ;
```

Figura 27 Fragmento de código SQL de agregación de secuencias MSC2 en Redshift

De esta forma disponemos la secuencia correspondiente para cada cliente en una nueva columna dentro de nuestra base de datos. A partir de allí podemos analizar los clientes agrupándolos por secuencias de comportamiento observado, y combinarlo con el resto de sus atributos.

Dichas secuencias se pueden mostrar también gráficamente. La realidad es que, dependiendo de la cantidad de registros procesados, la granularidad de la codificación simbólica elegida y la cardinalidad de combinaciones existente dentro del conjunto de datos de entrada, podemos obtener un número muy grande de secuencias diferentes que resultan muy difíciles de analizar para extraer conclusiones.

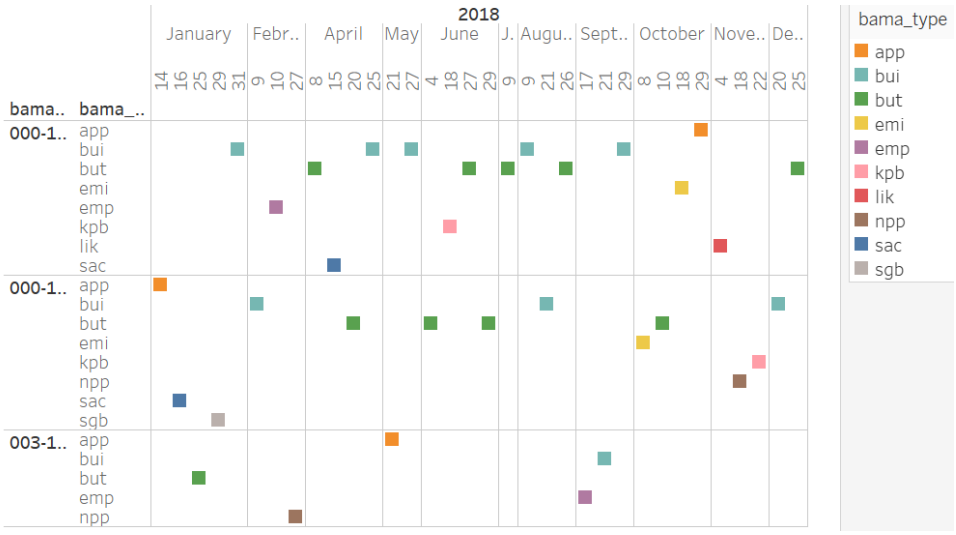


Figura 28 MSC2 Gráfica de eventos, acciones, métricas y segmentos por cliente

Además, si pensamos que las secuencias como un aglutinador de clientes en grupos o segmentos disjuntos, corremos el riesgo de acabar con una cantidad de secuencias que sean inmanejables desde el punto de vista de marketing. Típicamente las campañas de gestión de clientes consideran segmentos de clientes que no superan la decena de categorías.

Propiedades de las Secuencias de comportamiento MSC2

La gestión inteligente de clientes constituye un desafío enorme para los responsables de marketing de hoy día.

El primer reto lo constituye la **interpretación** del ingente volumen de datos que se debe gestionar. Cuando hablamos de modelar y organizar el conocimiento de centenas de miles o millones de clientes con todas sus transacciones e interacciones de los últimos 12/24 meses, el problema no está en almacenar y procesar los datos (para ello tenemos al “big data” y la nube); sino en la ausencia de una metáfora o paradigma que permita entender e interpretar ese conocimiento de forma sencilla e intuitiva.

El segundo reto consiste en la **capacidad de accionar** sobre dicho conocimiento y utilizarlo eficazmente en las diferentes campañas y acciones de marketing en forma de segmentos o “targets” de clientes.

Las secuencias de comportamiento poseen unas **propiedades** intrínsecas a su definición que las convierten en muy adecuadas para hacer frente a los retos de interpretación y capacidad de accionar:

- Representación: Poseen la capacidad de condensar varios aspectos del comportamiento tales como la temporalidad, la omnicanalidad, y la heterogeneidad de eventos, acciones, comunicaciones y métricas; en una única dimensión formal, flexible y almacenable.

- Interpretación: Mediante la codificación simbólica escogida y la determinación de la polaridad (positiva, neutra, negativa) de cada evento o secuencia desde el punto de vista del negocio, obtenemos una interpretación sin ambigüedad de cada secuencia.
- Exclusión mutua: Las secuencias funcionan como conjuntos disjuntos. Cada cliente pertenece a una única secuencia por lo tanto la intersección entre secuencias es vacía. Y la unión de todas las secuencias agrupa a todos los clientes.
- Segmentación: Dada su capacidad de aislar conjuntos funciona como una herramienta natural de segmentación de clientes.
- Conglomeración: Como la cantidad de secuencias puede ser muy extensa dependiendo de la codificación simbólica elegida y de las características del conjunto de datos en estudio, es conveniente poder agruparlas en conglomerados de similares características.
- Anticipación: Finalmente como las secuencias condensan una serie histórica de eventos o acciones puede utilizarse para entrenar algoritmos predictivos.

De Secuencias a Patrones de comportamiento MSC2

Para extraer **patrones de comportamiento** existen numerosos algoritmos que se enfocan en extraer subsecuencias frecuentes (patrones) de un conjunto de secuencias de entrada. Analizando la familia de algoritmos SPMF Sequential Pattern Mining (Fournier-Viger et al., 2017) encontramos diversos algoritmos de aplicación que varían en cuanto a su desempeño, por una parte, y el objetivo de agrupación de las secuencias frecuentes por otra.

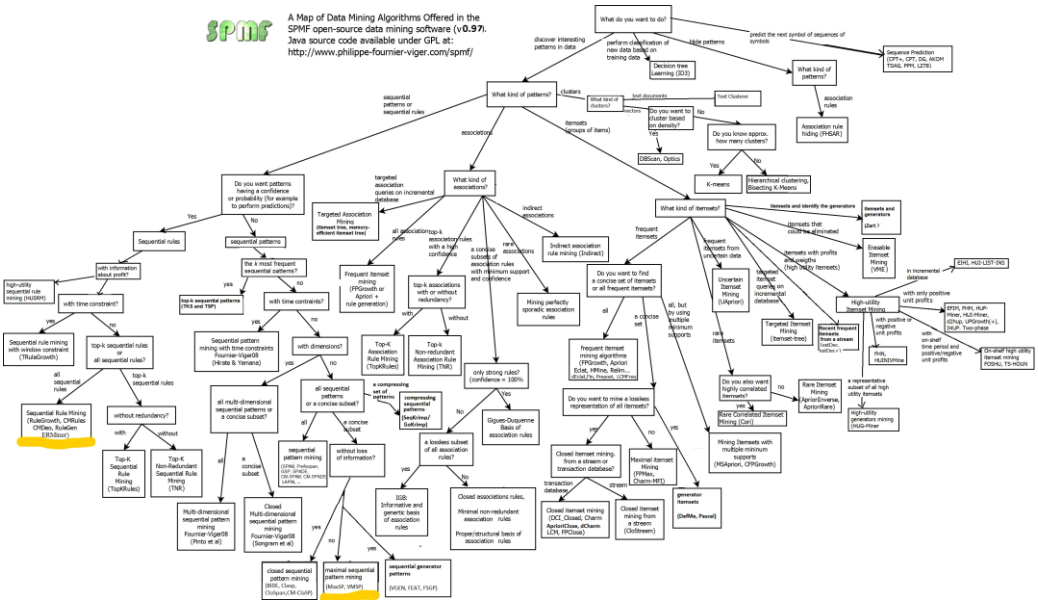


Figura 29 Mapa de algoritmos de minería SPMF (Fournier-Viger, 2015). Resaltamos en amarillo los algoritmos escogidos para MSC2

Los mismos pueden clasificarse de acuerdo a su objetivo de estudio (*Fournier-Viger, 2015*):

- Itemset Mining: Estos algoritmos descubren interesantes conjuntos de ítems (conjunto de valores) que aparecen en una base de transacciones que contenga una codificación simbólica de los registros
- Sequential Pattern Mining: Estos algoritmos descubren patrones secuenciales en un conjunto de secuencias
- Periodic Pattern Mining: Estos algoritmos descubren patrones que aparecen periódicamente en secuencias de eventos complejos pertenecientes a una base de transacciones
- Episode Mining: Estos algoritmos descubren episodios que aparecen en una secuencia simple de eventos complejos
- High-Utility Pattern Mining: Estos algoritmos descubren patrones que tienen una alta utilidad (importancia) en diferentes tipos de datos
- Stream pattern mining: Estos algoritmos descubren varios tipos de patrones en un "stream" (secuencia infinita de transacciones)
- Sequential Rule Mining: Estos algoritmos descubren reglas secuenciales probabilísticas en un conjunto de secuencias
- Sequence Prediction: Estos algoritmos predicen el próximo símbolo(s) en una secuencia basados en un conjunto de secuencias de entrenamiento

De todos ellos nos centraremos en dos algoritmos de dos subfamilias, dado que se adaptan mejor al problema de detección y anticipación de patrones de comportamiento de clientes **MSC2**: **VMSP** y **CMRules**. El algoritmo CMRules lo trataremos en la sección posterior.

Sequential pattern mining: VMSP

Esta familia de algoritmos de minería de patrones secuenciales se adapta muy bien al modelo establecido de representación del comportamiento observado de clientes. En particular nos hemos decantado por el **algoritmo VMSP**: **Efficient Vertical Mining of Maximal Sequential Patterns** (Fournier-Viger et al., 2014).

Este algoritmo al decir de sus autores es muy utilizado en aplicaciones tales como: analítica de páginas web, ejecución de programas, datos biológicos y datos sobre e-learning. En nuestro caso lo aplicaremos a la identificación de patrones de comportamiento de clientes.

Este algoritmo es muy efectivo en reconocer patrones en secuencias de comportamiento de clientes. A diferencia de los otros algoritmos que pueden presentar demasiados patrones secuenciales, el **algoritmo VMSP** extrae una representación más compacta a través de los patrones secuenciales máximos.

De esta forma se maximiza el soporte de casos (clientes) de cada patrón en la base de transacciones. Esto es lo que buscan los profesionales de marketing: segmentación de un grupo específico pero lo suficientemente nutrido para que compense realizar acciones específicas para cada grupo o segmento.

SID	Sequences
1	⟨{a, b},{c},{f, g},{g},{e}⟩
2	⟨{a, d},{c},{b},{a, b, e, f}⟩
3	⟨{a},{b},{f,g},{e}⟩
4	⟨{b},{f, g}⟩

Pattern	Sup.		Pattern	Sup.	
⟨{a}⟩	3	C	⟨{b},{g},{e}⟩	2	CM
⟨{a},{g}⟩	2		⟨{b},{f}⟩	4	C
⟨{a},{g},{e}⟩	2	CM	⟨{b},{f, g}⟩	2	CM
⟨{a},{f}⟩	3	C	⟨{b},{f},{e}⟩	2	CM
⟨{a},{f},{e}⟩	2	CM	⟨{b},{e}⟩	3	C
⟨{a},{c}⟩	2		⟨{c}⟩	2	
⟨{a},{c},{f}⟩	2	CM	⟨{c},{f}⟩	2	
⟨{a},{c},{e}⟩	2	CM	⟨{c},{e}⟩	2	
⟨{a},{b}⟩	2		⟨{e}⟩	3	
⟨{a},{b},{f}⟩	2	CM	⟨{f}⟩	4	
⟨{a},{b},{e}⟩	2	CM	⟨{f, g}⟩	2	
⟨{a},{e}⟩	3	C	⟨{f},{e}⟩	2	
⟨{a, b}⟩	2	CM	⟨{g}⟩	3	
⟨{b}⟩	4		⟨{g},{e}⟩	2	
⟨{b},{g}⟩	3	C			

C = Closed M = Maximal

Figura 30 Algoritmo VMSP (Fournier-Viger et al., 2014)

El algoritmo recibe un conjunto de datos de entrada con una base de transacciones secuenciales donde cada fila representa una secuencia con su identificador SID (cliente en nuestro problema), y tiene asociado una lista ordenada de conjuntos de ítems (eventos en nuestro problema).

A la salida genera los **patrones** descubiertos. Cada patrón tiene su soporte de casos (secuencias) correspondiente, siendo esta la métrica de desempeño de dicho patrón. También se añade la codificación de patrón cerrado (C) y/o patrón maximal (M) a título informativo.

En la metodología MSC2 se trata de adaptar el formato de las secuencias del modelo de representación del comportamiento del cliente, y traducirlo al formato de entrada al algoritmo VMSP.

```
1318 emp/bui/emp/bui/bui/nps/bui//
1319 emp/but/but/emp/but//
1320 emp/but/emp/but/but/but//
1321 emp/but/emp/but/but/emi/but//
1322 emp/but/emp/but/but/emp/but//
1323 emp/emp/bui/bui/emp/nps/bui//
1324 emp/emp/bui/bui/kpi/bui/bui/emi//
1325 emp/emp/bui/emp/bui/bui//
1326 emp/emp/bui/emp/bui/bui/sac//
1327 emp/emp/but/but/emi/but/nps//
1328 emp/emp/but/but/emi/but/nps/emp/nps//
```

```
1318 3 -1 1 -1 1 -1 3 -1 1 -1 -2
1319 3 -1 1 -1 3 -1 1 -1 1 -1 1 -1 -2
1320 3 -1 1 -1 3 -1 1 -1 1 -1 4 -1 1 -1 -2
1321 3 -1 1 -1 3 -1 1 -1 1 -1 3 -1 1 -1 -2
1322 3 -1 3 -1 2 -1 2 -1 3 -1 5 -1 2 -1 -2
1323 3 -1 3 -1 2 -1 2 -1 7 -1 2 -1 2 -1 4 -1 -2
1324 3 -1 3 -1 2 -1 3 -1 2 -1 2 -1 -2
1325 3 -1 3 -1 2 -1 3 -1 2 -1 2 -1 6 -1 -2
1326 3 -1 3 -1 1 -1 1 -1 4 -1 1 -1 5 -1 -2
1327 3 -1 3 -1 1 -1 1 -1 4 -1 1 -1 5 -1 3 -1 5 -1 -2
1328 3 -1 3 -1 1 -1 1 -1 4 -1 1 -1 5 -1 3 -1 5 -1 1 -1 2 -1 -2
```

Figura 31 Adaptación del formato de representación al formato de entrada de VMSP

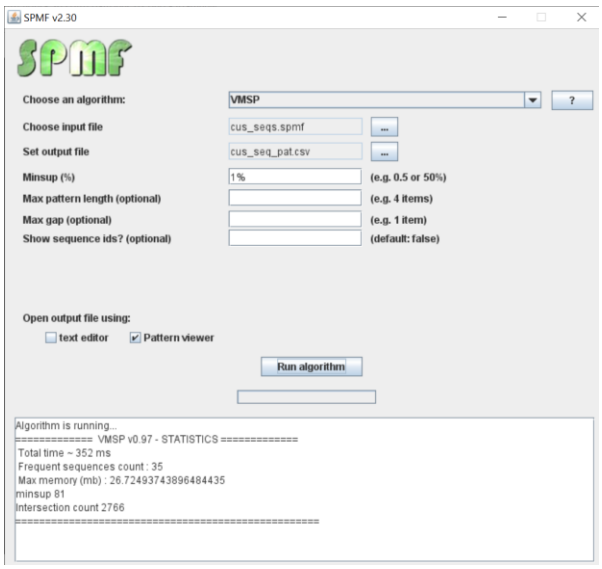
En el formato SPMF la codificación de símbolos es numérica consecutiva, y el delimitador de campo es (-1), mientras que el delimitador de registro es (-2).

Luego ejecutamos el algoritmo desde la librería compilada **SPMF.jar** en un entorno virtualizado (VM, Linux, JVM) completando el conjunto de parámetros de este:

- Soporte mínimo: El porcentaje de soporte mínimo solicitado para cada patrón. En nuestro problema significa la cantidad mínima de clientes que deseamos agrupe cada patrón de comportamiento. Este parámetro es obligatorio.
- Longitud del patrón máximo: Este parámetro opcional establece un tope para el largo de los patrones resultantes. Por omisión no hay un límite.
- Gap máximo: Este parámetro opcional establece la distancia máxima tolerada entre dos símbolos de una secuencia para formar parte de un patrón. Estos salteos permiten generar patrones menos estrictos en cuanto a secuencias consecutivas, y por lo tanto más flexible para identificar patrones similares. Por omisión es un gap de uno.

El resultado de la ejecución del algoritmo VMSP nos entrega todos los patrones (**Pattern**) que cumplen con las condiciones establecidas a partir de los parámetros de entrada, juntamente con el soporte de casos correspondientes (**#SUP**).

Luego volvemos a adaptar el resultado obtenido en formato SPMF al formato de nuestra representación de patrones de comportamiento.



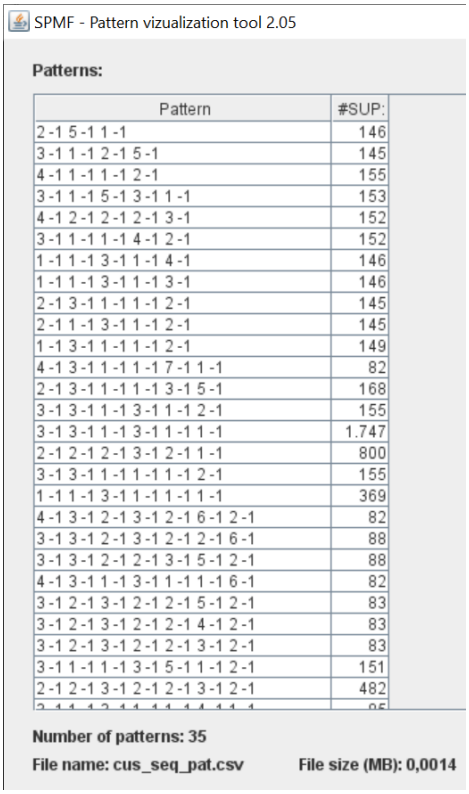


Figura 32 Ejecución del algoritmo VMSP aplicado al comportamiento de clientes

8.9. Modelado MSC2: Técnicas de “machine learning”

Una vez generadas las secuencias necesitamos un modelo de aprendizaje que permita agrupar las diferentes secuencias en un número acotado de “clusters” manejables por los responsables de marketing.

La particularidad de las secuencias en cuanto a la ausencia de atributos que no sean la propia secuencia hace descartar los algoritmos de “machine learning” no supervisados de generación de clusters o conglomerados tales como K-means, EM, etc.

Por otro lado, una secuencia puede interpretarse como una **frase perteneciente a un lenguaje generado por una gramática regular a partir del alfabeto formado por la codificación simbólica** de los eventos y acciones del cliente durante su comportamiento.

De esta forma podemos aplicar la familia de algoritmos de “**Topic modeling”** utilizados para la minería de temas y tópicos ocultos en una colección de textos o documentos utilizando modelos probabilísticos.

Los algoritmos de modelado de tópicos son utilizados principalmente en problemas de detección de estructuras en genética, bioinformática, recuperación de información, filtrado colaborativo, sistemas de recomendación de contenidos, y análisis de similitud entre documentos.

En nuestro caso lo utilizaré para **analizar similitud entre secuencias**, dado que el proceso de su codificación simbólica ya ha creado frases de un lenguaje de vocabulario controlado.

“Topic modeling”: Agrupamiento de Secuencias de comportamiento

“El objetivo principal de esta técnica de minería de textos es encontrar temas relevantes para organizar, buscar o comprender grandes cantidades de datos de texto no estructurados. Los modelos de temas se basan en el supuesto de que cualquier documento puede explicarse como una mezcla única de temas, donde cada tema es un grupo de términos concurrentes con diferentes probabilidades.” [27] (BigML, 2017).

El algoritmo de “Topic modeling” de BigML es una implementación optimizada del algoritmo LDA Latent Dirichlet Allocation, y puede encontrar los tópicos en pequeños fragmentos de texto como descripciones cortas, tweets o correos electrónicos. En nuestro caso lo aplicaremos a las secuencias de comportamiento como si se trataran de frases breves de un lenguaje.

Ejecutamos el algoritmo desde el tablero de control de **bigml.com** (modalidad de software como servicio) completando el conjunto de parámetros que requiere:

- Número de tópicos: El número de tópicos a obtener puede definirse de forma automática o manual.
- Número de términos top: Un tópico agrupa una cantidad de términos con diferentes probabilidades que deben sumar 100%. Establecemos el número máximo de términos a mostrar por tópico.
- Número de términos máximo por tópico: El número de términos diferentes por tópico puede crecer demasiado, de esta forma establecemos un máximo.
- Número de términos por nombre de tópico: Cada nombre de tópico se compone de los primeros N términos más representativos. Establecemos un límite inferior para el mismo.
- Lenguaje: Indicamos que no se trata de un lenguaje estándar para evitar la auto detección.
- Tokenización: Indicamos que el proceso de "tokenización" incluya todos los términos y no solamente las palabras.
- Remoción de palabras: En los lenguajes habituales se remueven muchas palabras que no aportan significado. Indicamos no remover ninguna palabra.
- Máximo "n-gramas": Un "n-grama" es una secuencia contigua de "n" términos. Indicamos capturar secuencias de hasta 5 gramas.

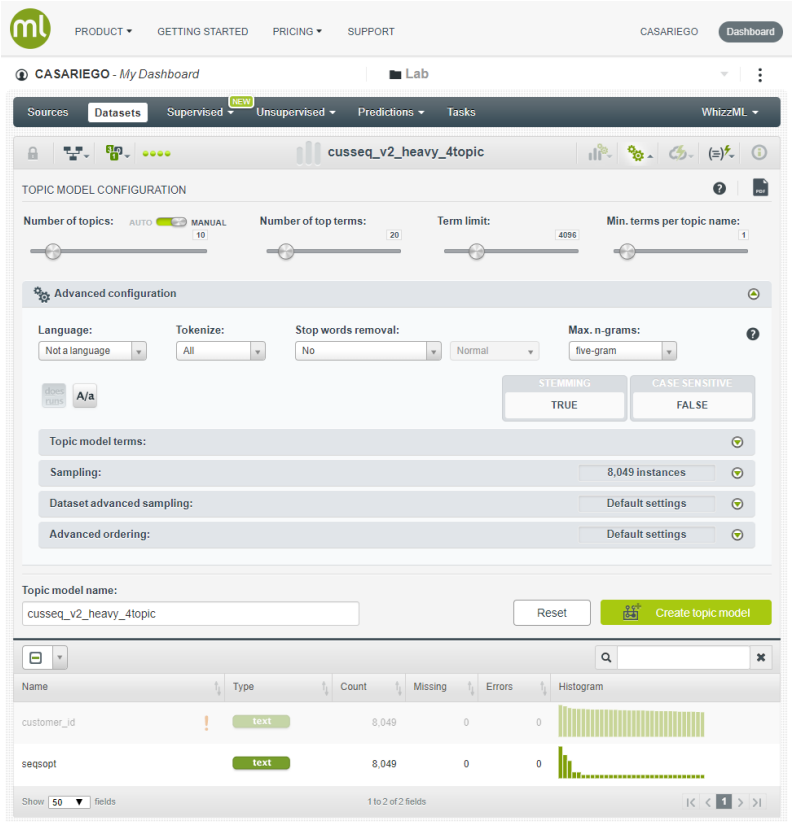


Figura 33 Algoritmo de tópicos de secuencias [27] (BigML, 2017)

Los resultados de "Topic modeling" de BigML son de fácil interpretación. Dispone de dos visualizaciones originales que nos facilitan la inspección de los modelos de tópicos de salida [27] (BigML, 2017):

- Mapa de tópicos: La visualización del Mapa de tópicos asigna los conglomerados de tópicos como círculos por cercanía temática, y brinda una visión general de la importancia de cada tema o tópico dentro del conjunto de datos de entrenamiento.
- Tabla de términos: Dado que un tópico se define por un grupo de términos con diferentes probabilidades, la Tabla de términos es una forma ideal de inspeccionar los términos destacados por tópico según su probabilidad.

8.10. Modelado MSC2: Minería de reglas de secuencias

Una de las propiedades más interesantes de las secuencias de comportamiento de cliente, es que pueden utilizarse para anticipar el siguiente evento o acción de un cliente con una cierta confianza.

Dada su propiedad de representación y almacenamiento de una secuencia histórica de eventos o acciones ordenada cronológicamente, es posible utilizar dicha información como conjunto de entrenamiento de los algoritmos que saben detectar reglas dentro de secuencias.

Por esta razón hemos seleccionado la familia de algoritmos de Sequential Rule mining para el desarrollo, y en particular el **algoritmo CMRules**.

"Sequential rule mining": CMRules

Esta familia de algoritmos de minería de reglas secuenciales se adapta muy bien al modelo establecido de representación del comportamiento observado de clientes. En particular nos decantamos por el **algoritmo CMRules**: Mining Sequential Rules Common to Several Sequences (Fournier-Viger et al., 2012).

Este algoritmo al decir de sus autores es muy utilizado en aplicaciones tales como: análisis del mercado de valores, observación del clima, gestión de sequías y datos sobre e-learning. En nuestro caso lo aplicaremos al uso generalizado de anticipación de patrones de comportamiento de clientes.

Este algoritmo es muy efectivo en reconocer reglas de secuencias de comportamiento de clientes mientras que los otros algoritmos no gestionan bien la identificación de patrones en secuencias con soportes mínimos. El **algoritmo CMRules** utiliza una combinación de

detección de patrones frecuentes con soporte mínimo, y detección de reglas de asociación temporal de símbolos.

De esta forma se maximiza el soporte de casos de cada regla en la base de transacciones, permitiendo obtener reglas verdaderamente accionables desde el punto de vista de marketing.

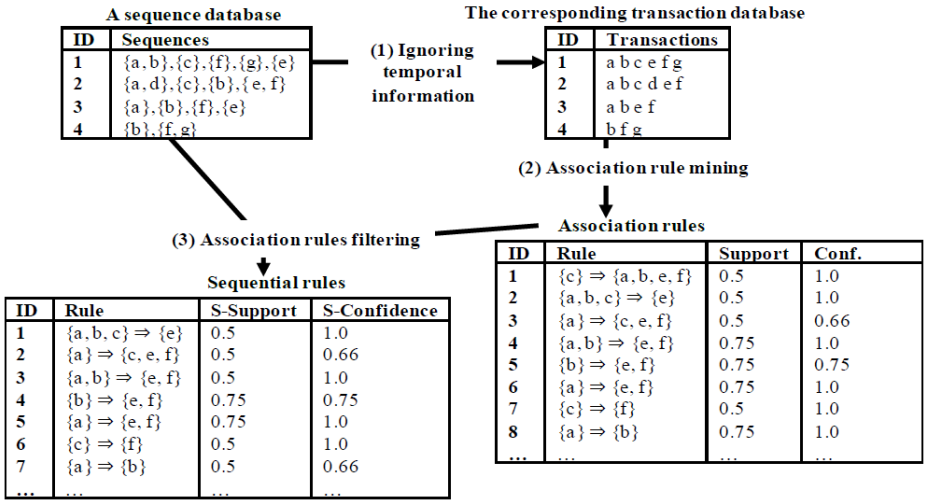


Figura 34 Algoritmo CMRules (Fournier-Viger et al., 2012)

El algoritmo recibe un conjunto de datos de entrada con una base de transacciones secuenciales donde cada fila representa una secuencia con su identificador SID (cliente en nuestro problema), y tiene asociado una lista ordenada de conjuntos de ítems (eventos en nuestro problema).

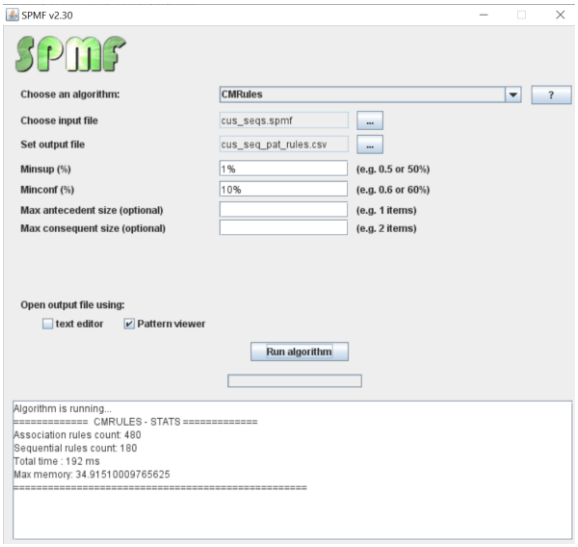
A la salida genera las **reglas** descubiertas, con sus correspondientes métricas de desempeño: soporte de casos y nivel de confianza.

En nuestro problema se trata de adaptar el formato de las secuencias del modelo de representación del comportamiento del cliente, y traducirlo al formato de entrada al **algoritmo CMRules**.

Luego ejecutamos el algoritmo desde la librería compilada **SPMF.jar** en un entorno virtualizado (VM, Linux, JVM) completando el conjunto de parámetros requeridos:

- Soporte mínimo: El porcentaje de soporte mínimo solicitado para cada regla. En nuestro problema significa la cantidad mínima de clientes que deseamos agrupe cada regla de anticipación del comportamiento. Este parámetro es obligatorio.
- Confianza mínima: El porcentaje de confianza mínimo solicitado para cada regla. Este parámetro es obligatorio.
- Tamaño de antecedente máximo: Este parámetro opcional establece un tope para la cantidad de símbolos en el antecedente de cada regla. Por omisión no hay un límite.

- Tamaño de consecuente máximo: Este parámetro opcional establece un tope para la cantidad de símbolos en el consecuente de cada regla. Por omisión no hay un límite.



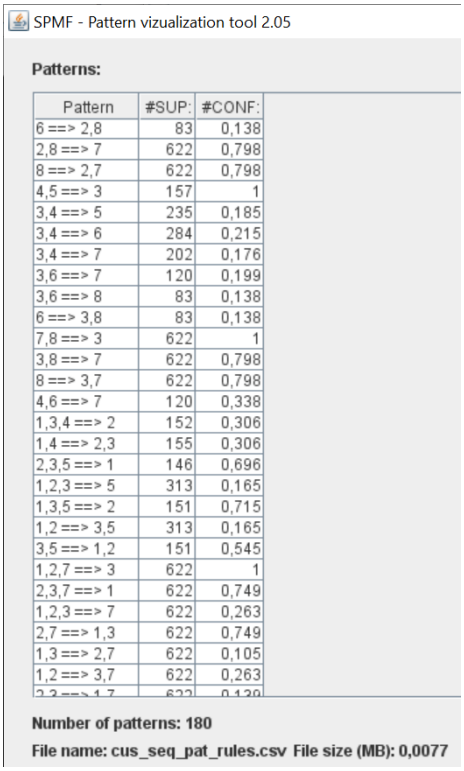


Figura 35 Ejecución del algoritmo CMRules aplicado al comportamiento de clientes

El resultado de la ejecución del algoritmo **CMRules** nos entrega todas las reglas (**Pattern**) que cumplen con las condiciones establecidas a partir de los parámetros de entrada, juntamente con el soporte de casos (**#SUP**), y el nivel de confianza (**#CONF**) correspondientes.

Luego volvemos a adaptar el resultado obtenido en formato SPMF al formato de nuestra representación de reglas de comportamiento.

```
100 3,4 ==> 6 #SUP: 284 #CONF: 0.21515151515151515
101 3,4 ==> 7 #SUP: 202 #CONF: 0.17575757575757575
102 3,6 ==> 7 #SUP: 120 #CONF: 0.19933554817275748
103 3,6 ==> 8 #SUP: 83 #CONF: 0.1378737541528239
104 6 ==> 3,8 #SUP: 83 #CONF: 0.1378737541528239
105 3,8 ==> 7 #SUP: 622 #CONF: 0.7984595635430038
106 7,8 ==> 3 #SUP: 622 #CONF: 1.0
107 8 ==> 3,7 #SUP: 622 #CONF: 0.7984595635430038
108 4,6 ==> 7 #SUP: 120 #CONF: 0.3380281690140845
109 1,3,4 ==> 2 #SUP: 152 #CONF: 0.30635245901639346
110 1,4 ==> 2,3 #SUP: 155 #CONF: 0.30635245901639346
111 1,2,3 ==> 5 #SUP: 313 #CONF: 0.16547066272688898
112 2,3,5 ==> 1 #SUP: 146 #CONF: 0.6962699822380106
113 1,3,5 ==> 2 #SUP: 151 #CONF: 0.7153284671532847
114 3,5 ==> 1,2 #SUP: 151 #CONF: 0.545201668984701
115 1,2 ==> 3,5 #SUP: 313 #CONF: 0.16547066272688898
```

	regla	soporte	confianza
2	pui/emp/kpi/seg ==> put	622	1,00
3	put/pui/kpi/seg ==> emp	622	1,00
4	put/pui/emp/seg ==> kpi	622	0,89
5	pui/emp/emi/sac ==> kpi	120	0,44
6	put/pui/kpi ==> emp	622	1,00
7	put/emp/seg ==> pui	696	1,00
8	put/pui/seg ==> emp	696	1,00
9	pui/kpi/seg ==> put	622	1,00
10	put/emi/nps ==> emp	157	1,00
11	emp/kpi/seg ==> put	622	1,00
12	put/kpi/seg ==> emp	622	1,00
13	emp/sac/seg ==> pui	83	1,00
14	pui/kpi/seg ==> emp	622	1,00
15	pui/kpi/seg ==> put/emp	622	1,00

Figura 36 Adaptación del formato de salida del algoritmo CMRules

8.11. Evaluación de resultados MSC2: Contraste de segmentos

Al final de este proceso de transformación de datos y de construcción de modelos analíticos estamos en condiciones de generar la nueva segmentación basada en comportamiento MSC2.

Mediante la aplicación selectiva de las entidades de comportamiento generadas previamente, y analizando su capacidad de clasificación y agrupamiento, procedemos a generar la nueva familia de segmentos de cliente.

Los nuevos segmentos estarán compuestos por uno o más de los siguientes componentes:

- Secuencias de comportamiento
- Tópicos de secuencias
- Reglas predictivas de secuencias
- Patrones de comportamiento
- Tópicos de patrones
- Reglas predictivas de patrones

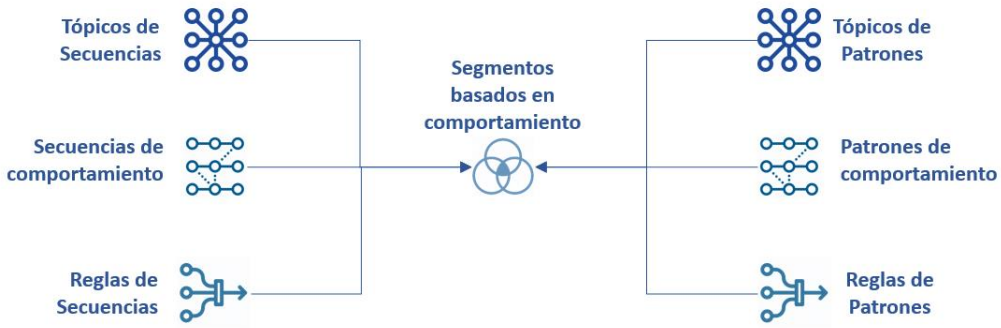


Figura 37 Componentes de los segmentos MSC2 basados en comportamiento

El principal criterio a la hora de escoger cuales de los componentes serán incluidos en la nueva segmentación de clientes, es el análisis descriptivo de los potenciales segmentos, y el contraste de estos versus los segmentos originales de marketing empleando técnicas de "profiling".

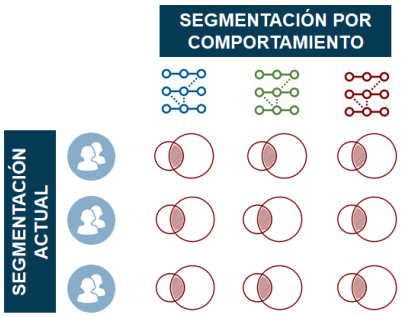


Figura 38 Segmentación actual vs. segmentación por comportamiento (elaboración propia)

Idealmente buscamos que la nueva segmentación sea transversal y complementaria de la anterior, de forma de asegurarnos que posee una capacidad clasificatoria y predictiva adicional.

Una segmentación clásica de clientes la constituyen los segmentos por valor (por ejemplo: alto, medio y bajo valor). La situación ideal es que un segmento basado en comportamiento, (por ejemplo: email/visita/compra/email/compra) atraviese lateralmente al menos dos de las categorías del segmento de valor.

De esta forma se entiende que la nueva segmentación está clasificando comportamientos de diferentes franjas de clientes, permitiendo así tomar decisiones y realizar acciones de marketing en consecuencia.

8.12. Evaluación de resultados MSC2: Técnicas de visualización

Esta novedosa segmentación de clientes MSC2 tiene dificultades a la hora de su visualización por medio de las gráficas convencionales debido a la naturaleza propia de las secuencias.

Hemos explorado diferentes diagramas existentes que se adecuan mejor a la visualización, navegación e interpretación de la segmentación basada en comportamiento.

Se debe utilizar un diagrama diferente para cada fase del proceso de generación y análisis de segmentos basados en comportamiento.

- Fase de codificación simbólica: En esta fase inicial de codificación simbólica de los eventos, acciones, métricas y segmentos de cliente, la visualización que mejor se aplica es el diagrama **Event drops**. Este permite visualizar y navegar a lo largo de la línea temporal, explorando la estructura visual de la codificación óptima escogida.
- Fase de generación de secuencias y patrones: En esta fase intermedia de generación de secuencias y patrones de comportamiento, las visualizaciones que mejor se aplican son los diagramas "**Sequences sunburst"** y "**Collapsible tree"**. Estos permiten visualizar y navegar en profundidad las secuencias de origen a fin, proporcionando una noción de la capacidad de agrupación de cada una.

9. Aplicación de la Metodología MSC2 a dos casos de uso

Con el objetivo de poner a prueba la metodología MSC2, voy a aplicarla de forma integral en dos casos de uso de diferentes sectores de negocio.

9.1. Caso A: Tienda de moda (ficticia)

El primer caso de aplicación será sobre una Tienda de moda ficticia que representa una combinación manipulada sintéticamente de la fuente de datos del E-Commerce Olist cuyos datos se encuentran disponibles en formato abierto en la plataforma Kaggle (https://www.kaggle.com/olistbr/brazilian-ecommerce), que ha sido complementada con otras fuentes de datos sintéticas simuladas a partir de los patrones detectados en el piloto de Entretenimiento.

Caso A - Comprensión del negocio:

Se trata de una Tienda de moda que cuenta con un amplio Catálogo de productos que ofrece a sus clientes a través de tiendas físicas y de su sitio de comercio electrónico.

Cuenta con un SAC (Servicio de Atención al Cliente) que recibe solicitudes y gestiona incidencias a través de diversos canales de comunicación.

El equipo de marketing realiza periódicamente Campañas de email marketing a sus clientes, enviando mensajes informativos y promocionales.

Regularmente se realizan encuestas de satisfacción dirigidas a los clientes utilizando la mecánica NPS “Net Promoter Score”.

La Tienda cuenta con una aplicación móvil para interactuar con sus clientes.

La Tienda es muy activa en su perfil de Facebook donde interactúa con sus clientes y fans.

Finalmente disponen de un sistema de inteligencia de cliente que monitorea diferentes métricas y segmentos de cliente.

Caso A - Perfil de cliente 360:

En esta Tienda ficticia simulamos las siguientes fuentes de datos que completan el Perfil de cliente 360 grados:

- CRM: Identidad y datos de cliente.
- VENTAS TIENDA: Transacciones de venta realizadas en tiendas físicas a través del ERP.
- COMERCIO ELECTRÓNICO: Transacciones de venta realizadas en el sitio de comercio electrónico.
- CLIMA: Complemento para las transacciones de venta proveniente de una fuente de datos abiertos de AEMET con información meteorológica y climatológica.
- SAC: Transacciones de incidencias realizadas a través del Servicio de atención al cliente.

- EMAIL MARKETING: Comunicaciones enviadas a los clientes a través de la plataforma de Email marketing.
- MOBILE APP: Interacciones de los clientes a través de la aplicación móvil.
- FACEBOOK: Interacciones de los clientes a través de la red social Facebook en las propiedades digitales de la tienda.
- ENCUESTAS DE SATISFACCIÓN: Transacciones de cliente a través de las encuestas de satisfacción.
- INTELIGENCIA DE CLIENTE: Métricas y segmentos de cliente proporcionados a través del sistema analítico de Inteligencia de cliente.

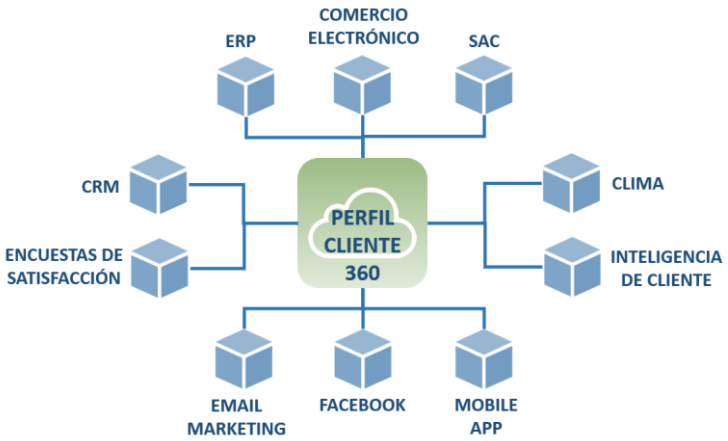


Figura 39 Caso A: Perfil de cliente 360 grados de la Tienda

Caso A - Representación del comportamiento:

En el caso de la Tienda, MSC2 incluye las transacciones observadas del cliente:

- Compra de producto en tienda física
- Compra de producto en tienda electrónica
- Servicio de atención al cliente
- Encuesta de satisfacción
- Incidencia de reclamación

En este caso el contexto es incorporado a las transacciones como el estado de la climatología en el día de la compra.

MSC2 también incluye las interacciones observadas del cliente. En este caso son:

- Interacciones con la Aplicación móvil
- Interacciones con los contenidos de la red social Facebook

MSC2 también incluye las comunicaciones desde y hacia el cliente. En este caso son:

- Comunicaciones informativas vía email
- Comunicaciones promocionales vía email

MSC2 también incorpora el conocimiento del negocio a través de las métricas y segmentos. En este caso las métricas y segmentos son:

- Cambio en el indicador principal de cliente
- Cambio de segmento principal de cliente

Caso A - Modelo de datos:

En el caso de la Tienda, el modelo conceptual MSC2 de entrada consiste en:

- Entidad principal Cliente
- Asociación Cliente-Transacciones
- Asociación Cliente-SAC
- Asociación Cliente-NPS
- Asociación Cliente-Comunicaciones
- Asociación Cliente-Interacciones
- Asociación Cliente-KPIs
- Asociación Cliente-Segmentos

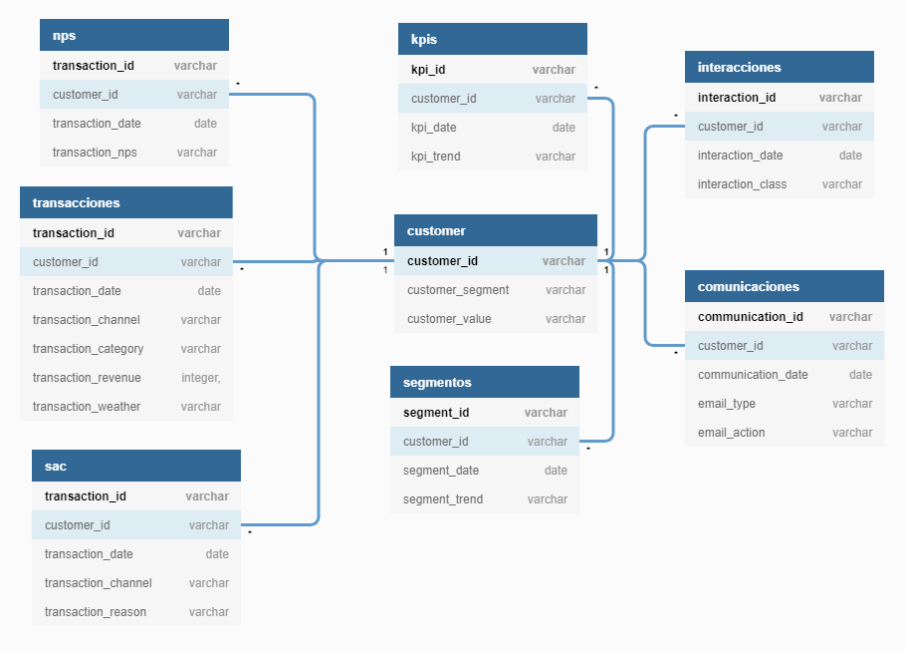


Figura 40 Caso A: Modelo de datos conceptual de entrada de la Tienda

Caso A - Metadatos: Codificación

En el caso de la Tienda, comenzamos codificando las transacciones.

Tenemos la transacción de “compra” con el tipo de canal asociado, para ello definimos una codificación básica utilizando la palabra “**buy**”. Y codificamos un parámetro para diferenciar las compras de tipo offline “**buy+#co**”, de las compras de tipo online “**buy+#ci**”, y así disponer de una codificación extendida.

Categoría: Evento_acción & Clase: Transacción							
category	class	type	token	param	param code	token + param code	effect
eoa	**trx**	**evento compra ?**	**buy**				
eoa	trx	compra.tipo = offline ?	buy	#c	#co	buy+#co	positivo
eoa	trx	compra.tipo = online ?	buy	#c	#ci	buy+#ci	positivo
eoa	**trx**	**evento nps ?**	**nps**				
eoa	trx	nps.respuesta = promotor ?	nps	#r	#rp	nps+#rp	positivo
eoa	trx	nps.respuesta = detractor ?	nps	#r	#rd	nps+#rd	neutro

Figura 41 Caso A: codificación de transacciones

También codificamos la transacción de "encuesta de satisfacción" utilizando la palabra "**nps**" en la codificación básica. Y cidficamos un parámetro para diferenciar las encuestas con respuesta promotor "**nps+#rp**", de las encuestas con respuesta detractor "**nps+#rd**".

Continuamos codificando las interacciones:

Categoría: Evento_acción & Clase: Interacción							
category	class	type	token	param	param code	token + param code	effect
eoa	**itx**	**evento usaApp ?**	**app**				neutro
eoa	**itx**	**evento like ?**	**lik**				positivo

Figura 42 Caso A: codificación de interacciones

Así es que tenemos la interacción de "uso del App" con una codificación básica utilizando la palabra "**app**". Y las interacciones en Facebook usan una codificación básica utilizando la palabra "**lik**".

Continuamos codificando las comunicaciones:

Categoría: Evento_acción & Clase: Comunicación							
category	class	type	token	param	param code	token + param code	effect
eoa	**com**	**evento email ?**	**ema**				
eoa	com	email.action = none ?	ema	#a	#an	ema+#an	negativo
eoa	com	email.action = open ?	ema	#a	#ao	ema+#ao	positivo
eoa	com	email.action = click ?	ema	#a	#ac	ema+#ac	positivo

Figura 43 Caso A: codificación de comunicaciones

Así es que tenemos la comunicación de "email" con el tipo de acción asociado, para ello definimos una codificación básica utilizando la palabra "**ema**". Y codificamos un parámetro para diferenciar las acciones de tipo nulo "**ema+#an**", de las acciones de apertura "**ema+#ao**", de las acciones de click "**ema+#ac**", y así disponer de una codificación extendida.

Categoría: Metrica_segmento & Clase: Métrica							
category	class	type	token	param	param code	token + param code	effect
mos	**met**	**metrica rfm ?**	**rfm**				
mos	met	rfm.cambio = mejor ?	rfm	#m	#mm	rfm+#mm	positivo
mos	met	rfm.cambio = peor ?	rfm	#m	#mp	rfm+#mp	negativo

Figura 44 Caso A: codificación de métricas

Luego codificamos los cambios de métricas. Así es que tenemos la métrica “RFM” con el tipo de cambio asociado, para ello definimos una codificación básica utilizando la palabra “**rfm**”. Y codificamos un parámetro para diferenciar los cambios a mejor “**rfm+#mm**”, de los cambios a peor “**rfm+#mp**”.

Continuamos codificando los cambios de segmento:

Categoría: Metrica_segmento & Clase: Segmento							
category	class	type	token	param	param code	token + param code	effect
mos	**seg**	**segmento loyalty ?**	**loy**				
mos	seg	loyalty.cambio = azul ?	loy	#s	#sa	loy+#sa	negativo
mos	seg	loyalty.cambio = oro ?	loy	#s	#so	loy+#so	positivo

Figura 45 Caso A: codificación de segmentos

Así es que tenemos el segmento “Loyalty” con el tipo de cambio asociado, para ello definimos una codificación básica utilizando la palabra “**loy**”. Y codificamos un parámetro para diferenciar los cambios a segmento azul “**loy+#sa**”, de los cambios a segmento oro “**loy+#so”.**

Bama C..	Bama T..	Bama Trailer	
trx	bui	+#ci+#pa+#ec	1,750
		+#ci+#pa+#ef	6,820
		+#ci+#pa+#en	12,806
		+#ci+#pb+#ec	721
		+#ci+#pb+#ef	2,890
		+#ci+#pb+#en	5,492
	but	+#co+#pa+#ec	2,546
		+#co+#pa+#ef	10,049
		+#co+#pa+#en	18,939
		+#co+#pb+#ec	1,073
		+#co+#pb+#ef	4,352
		+#co+#pb+#en	8,113
	npd	+#rd	251
	npp	+#rn	1,056
		+#rp	187
	sac	+#ic+#mi	452
		+#ic+#mq	81
		+#ie+#mi	566
		+#ie+#mq	109
		+#it+#mi	57
		+#it+#mq	39
itx	app		4,676
	lik		10,052
com	emi	+#ei+#ac	3,193
		+#ei+#ao	12,422
	emp	+#ep+#ac	4,662
		+#ep+#ao	18,772
met	kpb		3,008
	kpm		738
seg	sgb		3,950
	sgm		993
Grand Total			140,815

Figura 46 Caso A: Codificación extendida inicial

En todos los casos incorporamos la columna “**efecto**” para establecer el significado que le asigna el usuario de negocio a sus eventos, acciones, métricas y segmentos. Estos pueden ser “**positivos**”, “**neutros**” o “**negativos**”. Y así se incorporan en la codificación simbólica.

Aplicamos una primera pasada del proceso de codificación extendida a todos los datos para obtener una primera cuantificación.

Caso A - Metadatos: Codificación optimizada

Al analizar la distribución de frecuencias de la codificación extendida, vemos que genera una agrupación demasiado atomizada de los eventos, acciones, métricas y segmentos. Por lo tanto, ensayamos una nueva codificación básica que "recodifica" las palabras incorporando los atributos o propiedades más significativas.

			Bama Effect			
Bama Category	Bama Class	Bama Type	n	o	p	Grand ..
eoa	trx	bui			30,479	30,479
		but			45,072	45,072
		npd	251			251
		npp		1,056	187	1,243
		sac	229	1,075		1,304
	itx	app		4,676		4,676
		lik			10,052	10,052
	com	emi			15,615	15,615
		emp			23,434	23,434
mos	met	kpb		2,400	608	3,008
		kpm	738			738
	seg	sgb		3,187	763	3,950
		sgm	993			993
Grand Total			2,211	12,394	126,210	140,815

Figura 47 Caso A: Recodificación básica

Caso A – Decisiones previas

Varias de las lecciones aprendidas durante la validación de la metodología MSC2 aplicada a los conjuntos de datos reales, consisten en tomar una serie de decisiones previas que faciliten la máxima utilidad de los resultados.

- Marco temporal: Es necesario definir a priori el marco temporal del estudio. Normalmente se encuentra entre 12, 18 o 24 meses dependiendo de las características del negocio en concreto y de su estacionalidad. En este caso conservamos las transacciones correspondientes a **12 meses**.
- Foco en grupo de clientes "heavy": La metodología MSC2 pretende descubrir patrones de comportamiento en largas secuencias de eventos de cliente, por lo que la misma no tiene ningún interés para clientes cuyas secuencias no superan un determinado umbral. En este caso definimos dicho umbral entre 4 a 10 compras. Comportamientos superiores a 10 compras constituyen una anomalía que vamos a desechar del conjunto de datos.
- Remoción de emails enviados no leídos: Las organizaciones tienden a saturar a sus clientes con envíos masivos de email que ni siquiera son abiertos. La experiencia nos indica que es prudente conservar solamente las comunicaciones que son abiertas o clicadas, para evitar multitud de eventos que constituyen "ruido" en las secuencias.

- Codificación optimizada: Realizar varios intentos de codificación básica y extendida hasta encontrar la más adecuada para cada conjunto de datos.
- Mínimo soporte de cantidad de clientes por secuencia: La metodología MSC2 pretende identificar patrones de comportamiento comunes a grandes grupos de clientes. En este caso filtramos las secuencias que no agrupan al menos a 70 clientes.

Caso A – Flujo de procesos

En este caso hemos preparado un proceso previo para compatibilizar los conjuntos de datos sintéticos con la entrada esperada por la metodología MSC2 en sus procesos ETL de entrada.

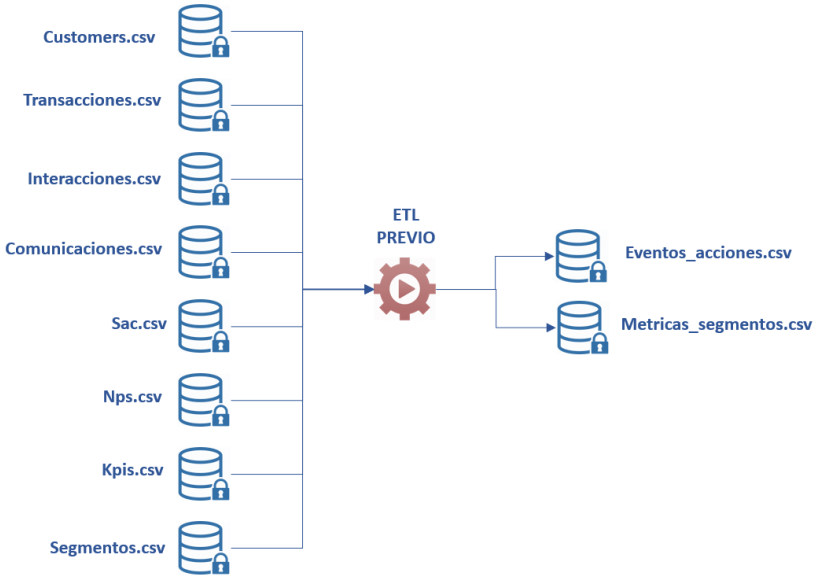


Figura 48 Caso A: Flujo de procesos de entrada

Caso A – Infraestructura y herramientas

En este caso hemos utilizado exactamente el mismo entorno de infraestructura tecnológica y herramientas que el presentado en la metodología MSC2.

Caso A – Modelado: Secuencias y patrones de comportamiento

La generación de secuencias es el último paso del proceso de ETL de entrada. De esta forma creamos la secuencia correspondiente para cada cliente, obteniendo un total de 39 secuencias.

A continuación, procesamos las secuencias resultantes utilizando el algoritmo el algoritmo **VMSP.** Los parámetros de ejecución del algoritmo utilizados son los siguientes:

- Soporte mínimo: 1% del conjunto de datos de entrada.
- Longitud del patrón máximo: 9 tokens máximo.

- Gap máximo: 2, esto presupone que tolero salteos de hasta 1 token en las secuencias.

A diferencia de las secuencias, los patrones no son clasificadores exclusivos. De esta forma obtenemos un total de 30 patrones de comportamiento.

Seqsopt	
emp/emp/but/emp/but/but	1,671
bui/bui/bui/emp/bui/but	800
seg/emp/but/bui/kpi/emp/but/but	622
bui/bui/emp/bui/bui/emp/bui	482
emp/but/but/emp/but	449
bui/but/bui/but/but/bui/emp	330
but/but/but/emp/but	313
bui/bui/emp/bui/bui/bui	303
emi/emp/bui/emp/bui/bui	225
emp/emp/bui/emp/bui/bui	222
but/but/emp/but/but/but	219
emi/emp/but/emp/but/but	197
bui/bui/emp/bui/bui/emi/bui/sac/kpi	120
bui/emp/but/but/emp/nps	96
emp/emp/bui/emp/bui/bui/sac	88
emp/emp/bui/bui/kpi/bui/bui/emi	88
emp/emp/bui/bui/emp/nps/bui	88
emp/but/emp/but/but/emp/but	86
emp/but/emp/but/but/emi/but	85
emp/but/emp/but/but/but	84
emp/bui/emp/bui/bui/nps/bui	83
emp/bui/emp/bui/bui/emp/bui	83
emp/bui/emp/bui/bui/emi/bui	83
emp/bui/bui/sac/emp/bui/seg/bui	83
emi/emp/but/emp/but/but/sac	82

Patrón	Longitud	Soporte
emp/emp/but/emp/but/but	6	1747
bui/bui/bui/emp/bui/but	6	800
seg/emp/but/bui/kpi/emp/but/but	8	622
bui/bui/emp/bui/bui/emp/bui	7	482
bui/but/bui/but/but/bui/emp	7	406
but/but/emp/but/but/but	6	369
emi/emp/bui/emp/bui/bui	6	295
emi/emp/but/emp/but/but	6	267
emp/but/emp/but/but/but	6	234
emp/bui/emp/bui/bui/bui	6	228
bui/emp/but/but/emp/nps	6	168
emp/emp/but/but/emp/but	6	158
emp/but/but/but/emp/but	6	158
emp/emp/but/but/emi/but/nps/emp/nps	9	157
but/emp/but/but/emp/but	6	154
but/nps/emp/but	4	153
emp/bui/bui/sac/bui	5	153
but/emp/but/but/emi/but	6	153
emp/bui/bui/bui/emp	5	152
emp/but/emp/but/but/sac	6	152
emp/but/but/emp/nps/but/bui	7	151
but/emp/but/but/bui	5	149
emp/bui/sac/emp	4	148
emp/but/emp/but/bui	5	148
emp/but/nps/but	4	146
bui/but/emp/but/bui	5	145
bui/bui/emp/bui/bui/emi/bui/sac/kpi	9	120
emp/emp/bui/emp/bui/bui/sac	7	82
emp/emp/bui/bui/kpi/bui/bui/emi	8	82
emp/emp/bui/bui/emp/nps/bui	7	81

Figura 49 Caso A: Secuencias y Patrones de comportamiento MSC2

Caso A – Modelado: Técnicas de "machine learning"

"Topic modeling" de Secuencias

A partir de las secuencias de comportamiento, ejecutamos el algoritmo no supervisado de "Topic modelling" de BigML con el siguiente conjunto de parámetros:

- Número de tópicos: Manualmente establecemos un número de 10 tópicos.
- Número de términos top: Establecemos el número máximo de 20 términos.
- Lenguaje: Indicamos que no se trata de un lenguaje.
- Tokenización: Indicamos que incluya todos los términos.
- Remoción de palabras: Indicamos no remover ninguna palabra.
- Máximo "n-gramas": Indicamos capturar secuencias de hasta 5 gramas.

Los resultados del proceso se aprecian a través de dos visualizaciones.

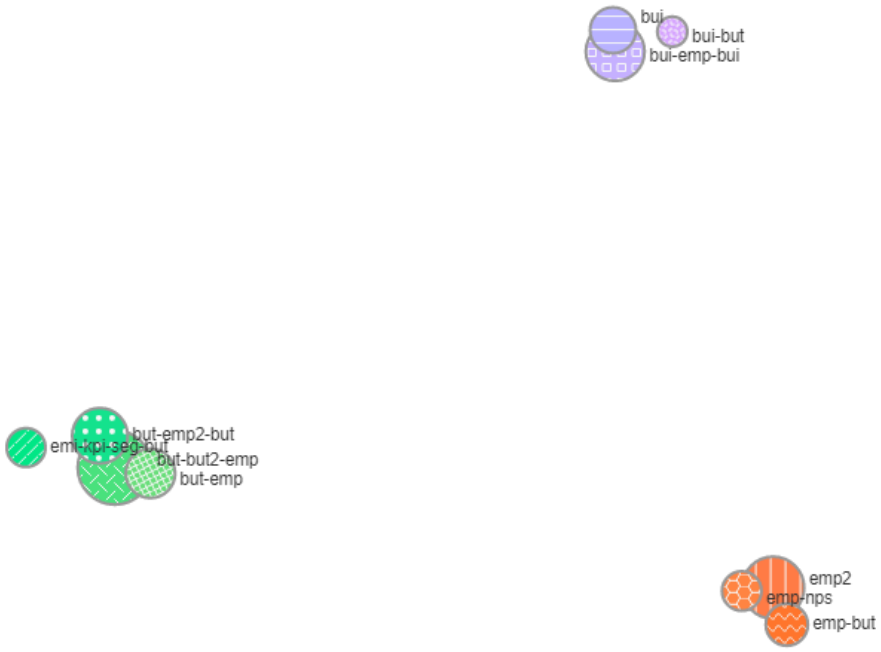


*Figura 50 Caso A: Mapa de **tópicos de secuencias***

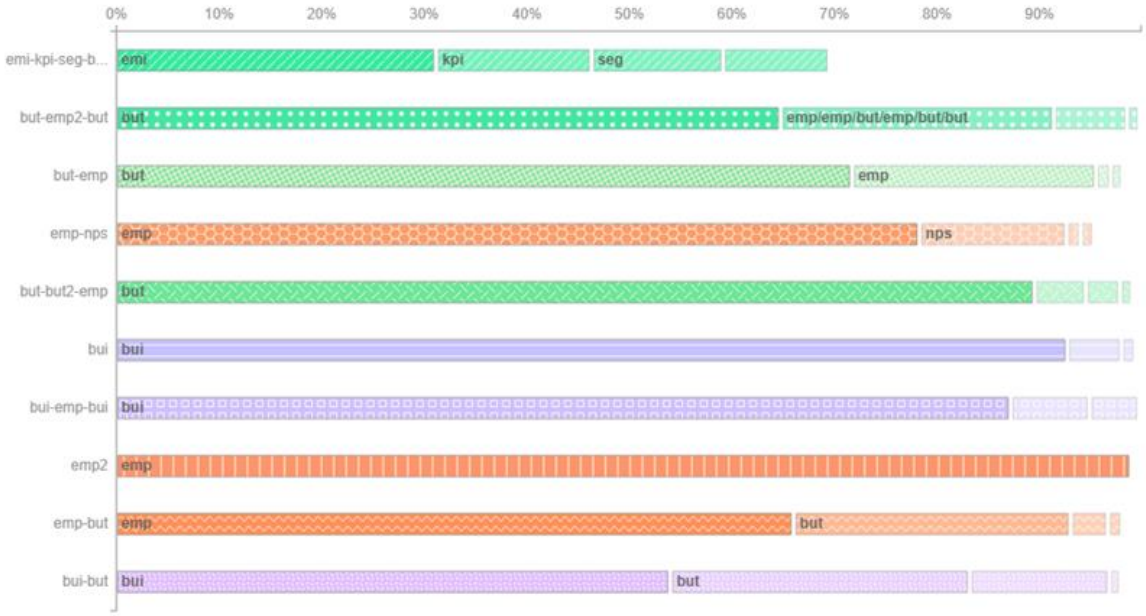


*Figura 51 Caso A: Tabla de **términos de secuencias***

A continuación, asignamos los tópicos a las secuencias de comportamiento y vemos la potencia de clasificación que poseen.

Seq Topic	Seqsopt	
bui	bui/bui/emp/bui/bui/emi/bui/sac/kpi	120
	bui/but/bui/but/but/bui/emp	330
	emi/emp/bui/emp/bui/sac/bui	82
	emp/bui/bui/sac/emp/bui/seg/bui	83
	emp/bui/emp/bui/bui/emi/bui	83
	emp/bui/emp/bui/bui/nps/bui	83
	emp/emp/bui/bui/kpi/bui/bui/emi	88
bui-but	bui/bui/bui/emp/bui/but	800
	emi/bui/but/bui/but/but/bui/emp	76
	emp/emp/but/emp/but/but/sac/bui	76
bui-emp-bui	bui/bui/emp/bui/bui/bui	303
	bui/bui/emp/bui/bui/emp/bui	482
	emi/emi/bui/emp/bui/emi/bui/bui/emi/emp	76
	emi/emp/bui/emp/bui/bui	225
but-but2-emp	bui/but/emp/but/but/emi/bui	73
	but/but/but/emp/but	313
	but/but/emp/but/but/but	219
	but/but/emp/but/but/emp/but	75
	emi/emp/but/but/kpi/but	82
	emi/emp/but/emp/but/but	197
	emi/emp/but/emp/but/but/sac	82
	emp/but/but/emp/but	449
	seg/emp/but/bui/kpi/emp/but/but	622

*Figura 52 Caso A: **Tópicos y secuencias***

“Topic modeling” de Patrones

A partir de los patrones de comportamiento, ejecutamos el algoritmo no supervisado de “Topic modelling” de BigML con el siguiente conjunto de parámetros:

- Número de tópicos: Manualmente establecemos un número de 8 tópicos.
- Número de términos top: Establecemos un número máximo de 20 términos.
- Lenguaje: Indicamos que no se trata de un lenguaje.
- Tokenización: Indicamos que incluya todos los términos.
- Remoción de palabras: Indicamos no remover ninguna palabra.
- Máximo “n-gramas”: Indicamos capturar secuencias de hasta 5 gramas.

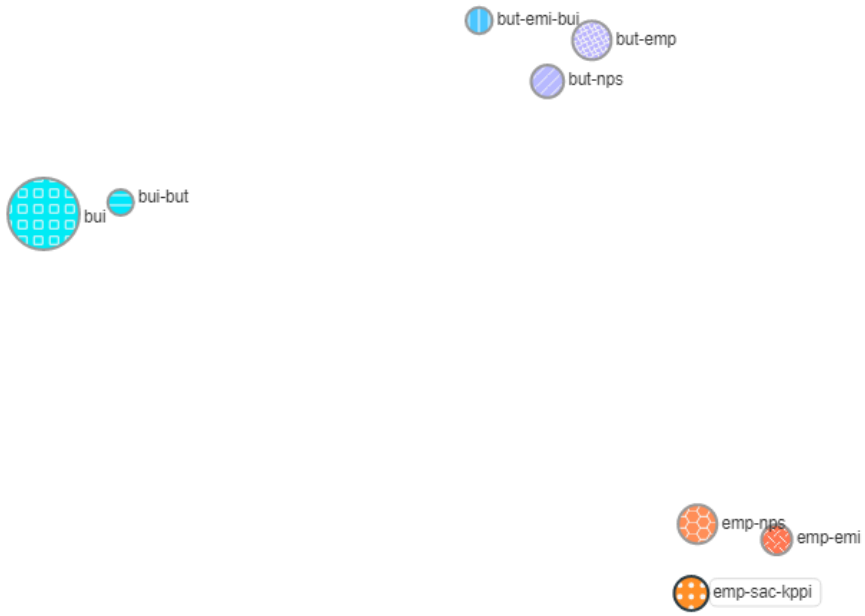


*Figura 53 Caso A: Mapa de **tópicos de patrones***

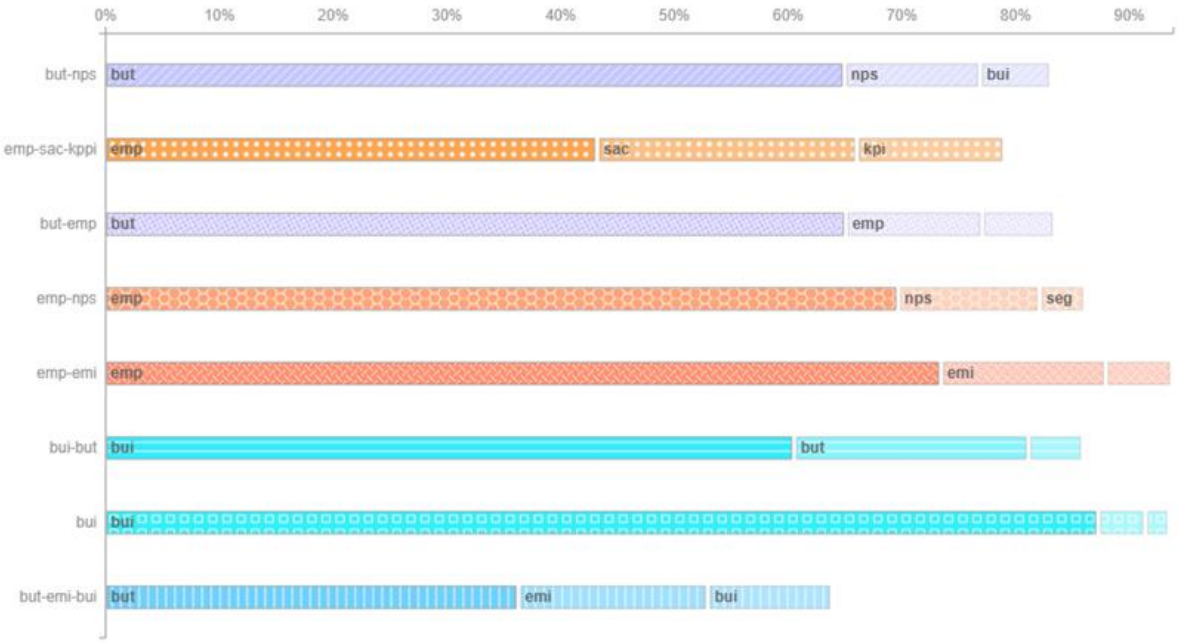


*Figura 54 Caso A: Tabla de **términos de patrones***

Caso A – Modelado: Minería de reglas de secuencias

A continuación, procesamos las secuencias resultantes utilizando el algoritmo **CMRules** utilizando los siguientes parámetros:

- Soporte mínimo: 1% del conjunto de datos de entrada.

- Confianza mínima: 10%.

El algoritmo genera 180 reglas predictivas que se ajustan a los parámetros solicitados.

regla	soporte	confianza	pre	q_pre	pos	q_pos
emp/kpi/seg ==> put	622	1,00	emp/kpi/seg	3	put	1
put/kpi/seg ==> emp	622	1,00	put/kpi/seg	3	emp	1
emp/sac/seg ==> pui	83	1,00	emp/sac/seg	3	pui	1
pui/kpi/seg ==> emp	622	1,00	pui/kpi/seg	3	emp	1
pui/kpi/seg ==> put/emp	622	1,00	pui/kpi/seg	3	put/emp	2
put/pui/seg ==> kpi	622	0,89	put/pui/seg	3	kpi	1
put/emp/seg ==> kpi	622	0,89	put/emp/seg	3	kpi	1
put/pui/seg ==> emp/kpi	622	0,89	put/pui/seg	3	emp/kpi	2
put/emp/seg ==> pui/kpi	622	0,89	put/emp/seg	3	pui/kpi	2
pui/emp/seg ==> put	696	0,89	pui/emp/seg	3	put	1
pui/emp/seg ==> kpi	622	0,80	pui/emp/seg	3	kpi	1
pui/emp/seg ==> put/kpi	622	0,80	pui/emp/seg	3	put/kpi	2

Figura 55 Caso A: Ejemplos de reglas predictivas de secuencias

Caso A – Evaluación de resultados: Contraste de segmentos

Con el objetivo de contrastar la capacidad adicional de segmentación de la metodología MSC2, procedemos primero a analizar la segmentación tradicional de la Tienda. Para ello disponemos de las correspondientes distribuciones de **Valor (alto, bajo), y Segmento (clásico, oro).**

valor

Customer Value	
alto	2,029
bajo	6,020
Grand Total	8,049

segmento

Customer Segment	
clasico	5,230
oro	2,819
Grand Total	8,049

segmento:valor

	Customer Value	
Customer Segment	alto	bajo
clasico	195	5,035
oro	1,834	985
Grand Total	2,029	6,020

Figura 56 Caso A: Segmentación tradicional de clientes

Ahora aplicamos el análisis comparativo de la segmentación basada en el comportamiento de los clientes de la Tienda utilizando la metodología MSC2.

Contrastamos los segmentos tradicionales de clientes (Valor y Segmento) comparado con la nueva segmentación basada en los tópicos de secuencias identificados previamente.

La nueva segmentación atraviesa las anteriores, mostrando una capacidad clasificatoria adicional. También las distribuciones tradicionales se ven alteradas dentro de determinados tópicos de comportamiento.

Seq Topic	Customer Value alto	bajo
bui	227	642
bui-but	255	697
bui-emp-bui	274	812
but-but2-emp	506	1,606
but-emp	80	260
emp2	439	1,320
emp-but	95	296
emp-nps	153	387
Grand Total	2,029	6,020

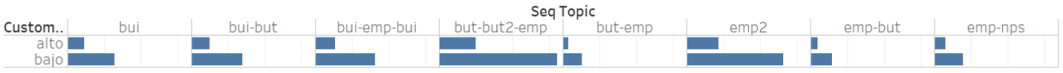


Figura 57 Caso A: Contraste Valor versus tópicos de secuencias

Seq Topic	Customer Segment clasico	oro
bui	563	306
bui-but	580	372
bui-emp-bui	720	366
but-but2-emp	1,372	740
but-emp	240	100
emp2	1,133	626
emp-but	268	123
emp-nps	354	186
Grand Total	5,230	2,819

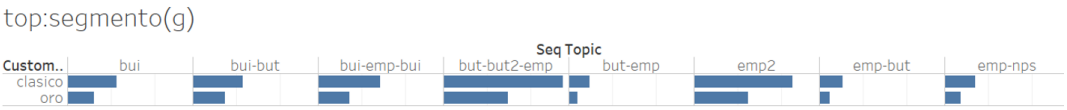


Figura 58 Caso A: Contraste Segmento versus tópicos de secuencias

Caso A – Evaluación de resultados: Técnicas de visualización

Para poder visualizar, analizar y navegar las secuencias de comportamiento, las procesamos usando la librería “Sequences sunburst”.

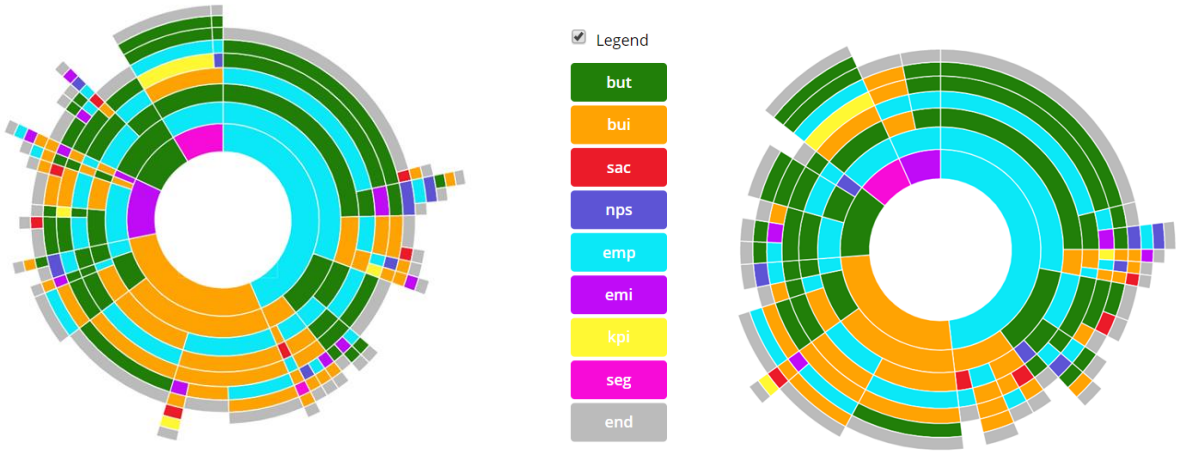


*Figura 59 Caso A: Visualización de **secuencias** de comportamiento (izquierda) y de **patrones** de comportamiento (derecha)*

Observamos que ambas tienen siluetas similares, pero los patrones son más relajados que las secuencias.

9.2. Caso B: Servicio Sanitario (ficticio)

El segundo caso de aplicación es sobre un Servicio sanitario ficticio representado por un conjunto de datos sintéticos simuladas a partir de los patrones detectados en el piloto de Sanidad y reflejando también la dificultad de disponer de datos de comportamiento en dicha industria.

Caso B - Comprensión del negocio:

Se trata de un proveedor de Servicios Sanitarios que atiende a una comunidad de pacientes a través de diferentes servicios profesionales especializados en atención sanitaria.

Cada centro cuenta con su propio sistema de información, los cuales alimentan el registro de actividad de los pacientes a través de los diferentes servicios.

Caso B - Perfil de paciente 360:

En este caso simulamos las fuentes de datos que completan el Perfil de paciente 360 grados:

- Ficha de paciente: Identidad y datos de paciente.
- Servicios especializado H-C-U-D: Registro de la actividad del paciente en cada servicio especializado H, C, U y D.

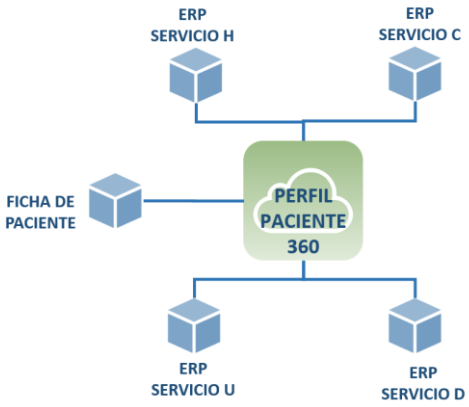


Figura 60 Caso B: Perfil de Paciente 360 grados

Caso B - Representación del comportamiento:

MSC2 registra todas las transacciones del paciente:

- Alta de ficha de paciente.

- Registro de atención profesional especializada en Servicios H, C, U y/o D.

MSC2 también registra el conocimiento del negocio a través de los segmentos de paciente:

- Cambio en la edad del paciente.

Caso B - Modelo de datos:

En el caso del Servicio Sanitario, el modelo conceptual MSC2 de entrada consiste en:

• Entidad principal Cliente • Asociación Cliente-Transacciones servicio H • Asociación Cliente-Transacciones servicio C	• Asociación Cliente-Transacciones servicio U • Asociación Cliente-Transacciones servicio D

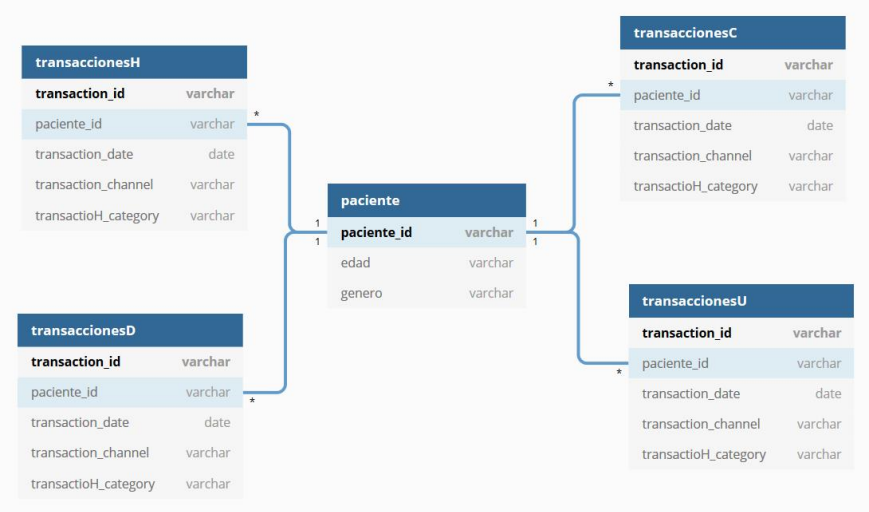


Figura 61 Caso B: Modelo de datos conceptual de entrada Servicio sanitario

Caso B - Metadatos: Codificación

En el caso del Servicio Sanitario, comenzamos codificando las transacciones.

Tenemos la transacción de “registro_actividad” con el diagnóstico asociado, para ello definimos una codificación básica utilizando las palabras “**sanH**”, “**sanC**”, “**sanU**” y “**sanD**” según sea la fuente. Y codificamos un parámetro para diferenciar los diagnósticos “**sanH+#d1**”, y así disponer de una codificación extendida.

Categoría: Evento_acción & Clase: Transacción							
category	class	type	token	param	param code	token + param code	effect
eoa	**trx**	**evento actividad en servicio H ?**	**sanH**				
eoa	trx	actividad.diagnóstico = d1		#d	#d1	sanH+#d1	positivo
eoa	trx	actividad.diagnóstico = d2		#d	#d2	sanH+#d2	negativo
eoa	**trx**	**evento actividad en servicio C ?**	**sanC**				
eoa	trx	actividad.diagnóstico = d1		#d	#d1	sanC+#d1	positivo
eoa	trx	actividad.diagnóstico = d2		#d	#d2	sanC+#d2	negativo
eoa	**trx**	**evento actividad en servicio U ?**	**sanU**				
eoa	trx	actividad.diagnóstico = d1		#d	#d1	sanU+#d1	positivo
eoa	trx	actividad.diagnóstico = d2		#d	#d2	sanU+#d2	negativo
eoa	**trx**	**evento actividad en servicio D ?**	**sanD**				
eoa	trx	actividad.diagnóstico = d1		#d	#d1	sanD+#d1	positivo
eoa	trx	actividad.diagnóstico = d2		#d	#d2	sanD+#d2	negativo

Figura 62 Caso B: Codificación de transacciones

Continuamos codificando los cambios de segmento etario:

Categoría: Metrica_segmento & Clase: Segmento							
category	class	type	token	param	param code	token + param code	effect
mos	**seg**	**segmento etario ?**	**segE**				
mos	seg	edad.cambio = 1		#e	#e0	segE+#e0	neutro
mos	seg	edad.cambio = i					
mos	seg	edad.cambio = N		#e	#e9	segE+#e9	neutro

Figura 63 Caso B: Codificación de segmentos

Tenemos el segmento “Edad” con el tipo de cambio asociado, para ello definimos una codificación básica utilizando la palabra “**segE**”. Y codificamos un parámetro para diferenciar los cambios de segmento por rango de edad “**segE+#e3**”**.**

En todos los casos incorporamos en la codificación simbólica la columna “**efecto**” con valores “**positivos**”, “**neutros**” o “**negativos**”.

Aplicamos una primera pasada del proceso de codificación básica a todos los datos y obtenemos una primera cuantificación:

			Bama Effect	
Bama Category	Bama Class	Bama Typo	n	o
eoa	trx	sanC		245.867
		sanD		33.110
		sanH		1.854
		sanU	11.214	
mos	seg	segE		72.794
Grand Total			**11.214**	**353.625**

Figura 64 Caso B: Codificación básica inicial

Caso B - Metadatos: Codificación optimizada

Al analizar la distribución de frecuencias de la codificación básica, identificamos que una de las fuentes de datos tiene un volumen de eventos desproporcionado en relación con las

demás. Por lo tanto, decidimos enfocarnos solamente en aquellos pacientes que han mostrado actividad en al menos un par de servicios específicos. De esta forma la codificación optimizada es la misma, pero el universo de trabajo es más limitado.

			Bama Effect	
Bama Category	Bama Class	Bama Typo	n	o
eoa	trx	sanC		53.464
		sanD		33.092
		sanH		1.650
		sanU	8.475	
mos	seg	segE		11.861
Grand Total			8.475	100.067

Figura 65 Caso B: Codificación básica limitada a pacientes multiservicio

Caso B – Decisiones previas

En el caso del Servicio Sanitario, debido a sus particularidades, las decisiones previas que facilitan la máxima utilidad de los resultados son los siguientes:

- Marco temporal: En este caso conservamos las transacciones correspondientes a **48 meses**.
- Foco en grupo de clientes "multiservicio": En este caso definimos el universo de análisis para los pacientes que al menos tienen actividad en dos servicios especializados diferentes.
- Codificación optimizada: Se conserva la codificación básica.
- Mínimo soporte de cantidad de pacientes por secuencia: En este caso filtramos las secuencias que agrupan al menos a 10 pacientes.

Caso B – Flujo de procesos

En este caso hemos preparado un proceso previo para compatibilizar los conjuntos de datos sintéticos con la entrada esperada por la metodología MSC2 en sus procesos ETL de entrada.

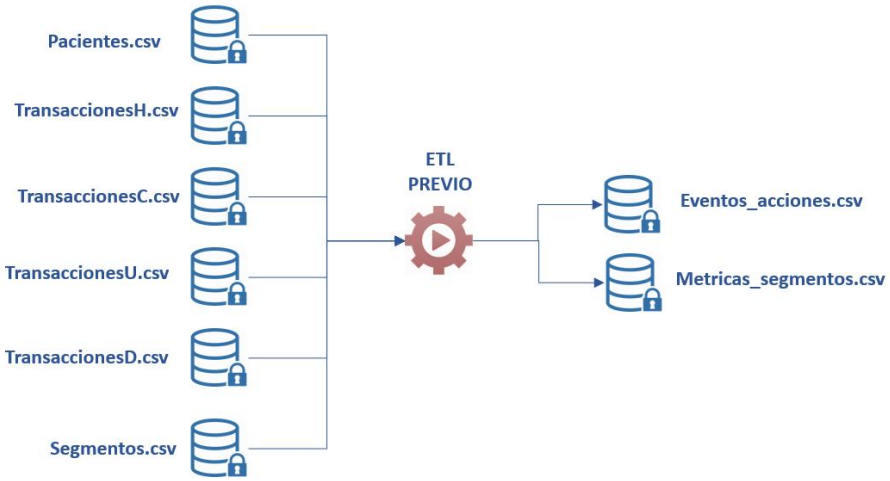


Figura 66 Caso B: Flujo de procesos de entrada

Caso B – Infraestructura y herramientas

En este caso hemos utilizado exactamente el mismo entorno de infraestructura tecnológica y herramientas que el presentado en la metodología MSC2.

Caso B – Modelado: Secuencias y patrones de comportamiento

La generación de secuencias es el último paso del proceso de ETL de entrada. Creamos la secuencia correspondiente para cada paciente, obteniendo un total de 22 secuencias diferentes con suficiente soporte de pacientes.

Secuencia	
sanU/sanU	279
sanU/sanC	83
sanU2/sanU	63
sanU/sanC/sanC	63
sanU/sanU/sanU	45
sanU/sanC/sanC/sanC	42
sanU/sanC3/sanC	41
sanU/sanH	31
sanU/sanH/sanU	28
sanH/sanU/sanU	25
sanU2/sanU/sanU	23
sanU2/sanC	20
sanU/sanC/sanC3/sanC	20
sanC/sanC/sanU	17
sanC/sanU/sanC	16
sanC3/sanU	14
sanU/sanU/sanU/sanU	13
sanC/sanU/sanC/sanC	12
sanU2/sanU2/sanU	11
sanU2/sanC/sanC	11
sanU/sanC/sanC3/sanC3/sanC	11
sanH/sanU	11
Grand Total	879

Figura 67 Caso B: Tabla de secuencias MSC2 con sumatorio de pacientes

A continuación, procesamos las secuencias resultantes utilizando el algoritmo **VMSP** con los siguientes parámetros de ejecución:

- Soporte mínimo: 1% del conjunto de datos de entrada.
- Longitud del patrón máximo: 8 “tokens” máximo.
- Gap máximo: 1, esto presupone que no tolero salteos en las secuencias.

De esta forma obtenemos un total de 18 patrones de comportamiento de salida.

Patron	Largo	
sanU/sanU	5	438
sanC/sanC/sanC	5	410
sanC3/sanC3/sanC3/sanC3/sanC3/sanC	6	396
sanH/sanU	5	227
sanC3/sanC3/sanC3/sanC/sanC	6	227
sanC/sanC/sanU	5	222
sanD/sanD	5	220
sanU/sanC3/sanC	5	189
sanC/sanC3/sanC	5	177
sin_patron	0	151
sanU2	0	145
sanU/sanC/sanC	5	142
sanC/sanU/sanC	5	107
sanC/sanC3/sanC3/sanC3	5	84
sanU/sanC3/sanC3/sanC3	5	81
sanC3/sanU	6	77
sanU/sanH	5	45
sanD3/sanD	6	34
sanC3/sanC/sanC3	6	6
Grand Total		3.378

Figura 68 Caso B: Patrones de comportamiento

Caso B – Modelado: Técnicas de “machine learning”

A partir de las secuencias de comportamiento, ejecutamos el algoritmo no supervisado de “Topic modelling” de BigML con el siguiente conjunto de parámetros:

- Número de tópicos: Manualmente establecemos un número de 5 tópicos.
- Número de términos top: Establecemos el número máximo de 20 términos.
- Lenguaje: Indicamos que no se trata de un lenguaje.
- Tokenización: Indicamos que incluya todos los términos.
- Remoción de palabras: Indicamos no remover ninguna palabra.
- Máximo “n-gramas”: Indicamos capturar secuencias de hasta 5 gramas.

Los resultados de “Topic modeling” los visualizamos de la siguiente forma:

Figura 69 Caso B: Mapa de tópicos de secuencias

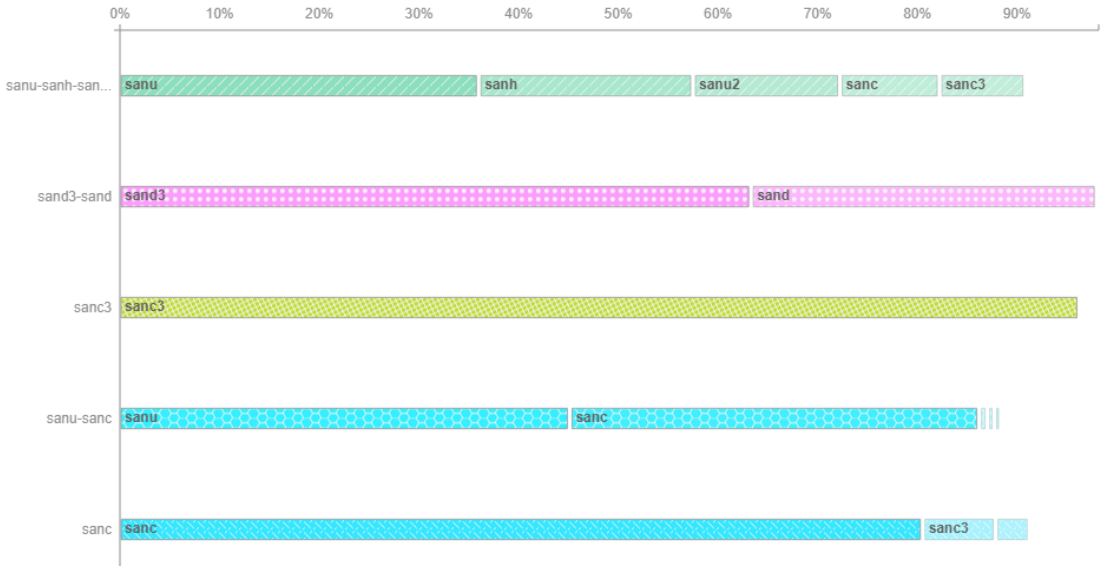


Figura 70 Caso B: Tabla de términos de secuencias

A continuación, asignamos los tópicos a las secuencias de comportamiento y comprobamos su potencia de clasificación.

Topico	
sanc	542
sanc3	831
sand3-sand	378
sanu-sanc	1.039
sanu-sanh-sanu2	588
Grand Total	3.378

Topico	Secuencia	
sanc	sanC/sanU	10
	sanC/sanU/sanC/sanC	12
	sanU2/sanC	20
	sanU2/sanC/sanC	11
sanc3	sanU/sanC3/sanC3/sanC	10
sanu-sanc	sanC3/sanU	14
	sanC/sanC/sanU	17
	sanC/sanU/sanC	16
	sanU2/sanU	63
	sanU/sanC	83
	sanU/sanC3/sanC	41
	sanU/sanC/sanC	63
	sanU/sanC/sanC3/sanC	20
	sanU/sanC/sanC/sanC	42
	sanU/sanH	31
	sanU/sanU	279
	sanU/sanU/sanU	45
	sanU/sanU/sanU/sanU	13
sanu-sanh-sanu2	sanH/sanU	11
	sanH/sanU/sanU	25
	sanU2/sanU2/sanU	11
	sanU2/sanU/sanU	23
	sanU/sanC/sanC3/sanC3/sanC	11
	sanU/sanH/sanU	28
Grand Total		899

Figura 71 Caso B: Tópicos de secuencias y Secuencias

Caso B – Modelado: Minería de reglas de secuencias

Luego procesamos las secuencias resultantes utilizando el algoritmo el algoritmo **CMRules** con los parámetros correspondientes:

- Soporte mínimo: 1% del conjunto de datos de entrada.
- Confianza mínima: 60%.

El algoritmo genera 10 reglas predictivas que se ajustan a los parámetros solicitados.

regla	soporte	confianza	pre	q_pre	pos	q_pos
sanU2,sanC3 ==> sanC	389	0,99	sanU2,sanC3	3	sanC	1
sanH,sanU,sanC3 ==> sanC	613	0,99	sanH,sanU,sanC3	3	sanC	1
sanH,sanC3 ==> sanC	665	0,99	sanH,sanC3	3	sanC	1
sanU,sanC3 ==> sanC	1.595	0,98	sanU,sanC3	3	sanC	1
sanC3 ==> sanC	1.868	0,97	sanC3	4	sanC	1
sanD3 ==> sanD	360	0,97	sanD3	3	sanD	1
sanH ==> sanU	893	0,92	sanH	4	sanU	1
sanU2 ==> sanU	525	0,80	sanU2	4	sanU	1
sanH ==> sanC	791	0,76	sanH	4	sanC	1
sanH,sanU ==> sanC	708	0,75	sanH,sanU	3	sanC	1

Figura 72 Caso B: Ejemplos de reglas predictivas de salida

Caso B – Evaluación de resultados: Contraste de segmentos

Contrastamos la capacidad adicional de segmentación de la metodología MSC2, analizando la segmentación tradicional del conjunto de datos de Servicio sanitario. Para ello disponemos de las correspondientes distribuciones de **Rango de edad,** y **Género.**

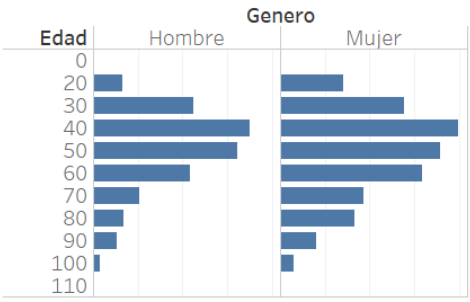


Figura 73 Caso B: Segmentación tradicional de pacientes

Aplicamos el análisis comparativo de la segmentación basada en el comportamiento de los pacientes del Servicio sanitario utilizando la metodología MSC2.

Comenzamos con el contraste del segmento tradicional de pacientes (Rango de edad, Género) comparado con la nueva segmentación basada en los tópicos de secuencias identificados previamente.

La nueva segmentación atraviesa la anterior, lo que muestra una capacidad clasificatoria diferente; y adicionalmente la distribución tradicional se ve alterada en determinados tópicos de secuencias de comportamiento.

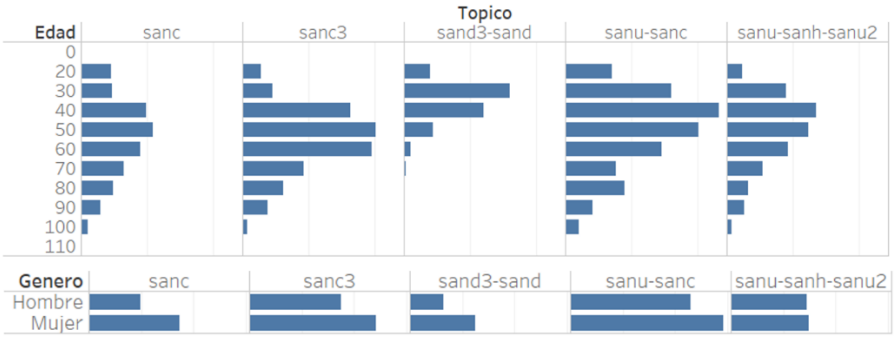


Figura 74 Caso B: Contraste Rango de edad y Género versus Tópicos de secuencias

Continuamos con el contraste ahora comparado con la nueva segmentación basada en los patrones de secuencias. Nuevamente la segmentación dinámica atraviesa la tradicional; y aquí también la distribución tradicional original se ve alterada.

Figura 75 Caso A: Contraste Segmento y Género versus Patrones de secuencias

Caso B – Evaluación de resultados: Técnicas de visualización

Utilizamos la librería de visualización interactiva “Sequences sunburst” para poder analizar y navegar las nuevas secuencias de comportamiento.

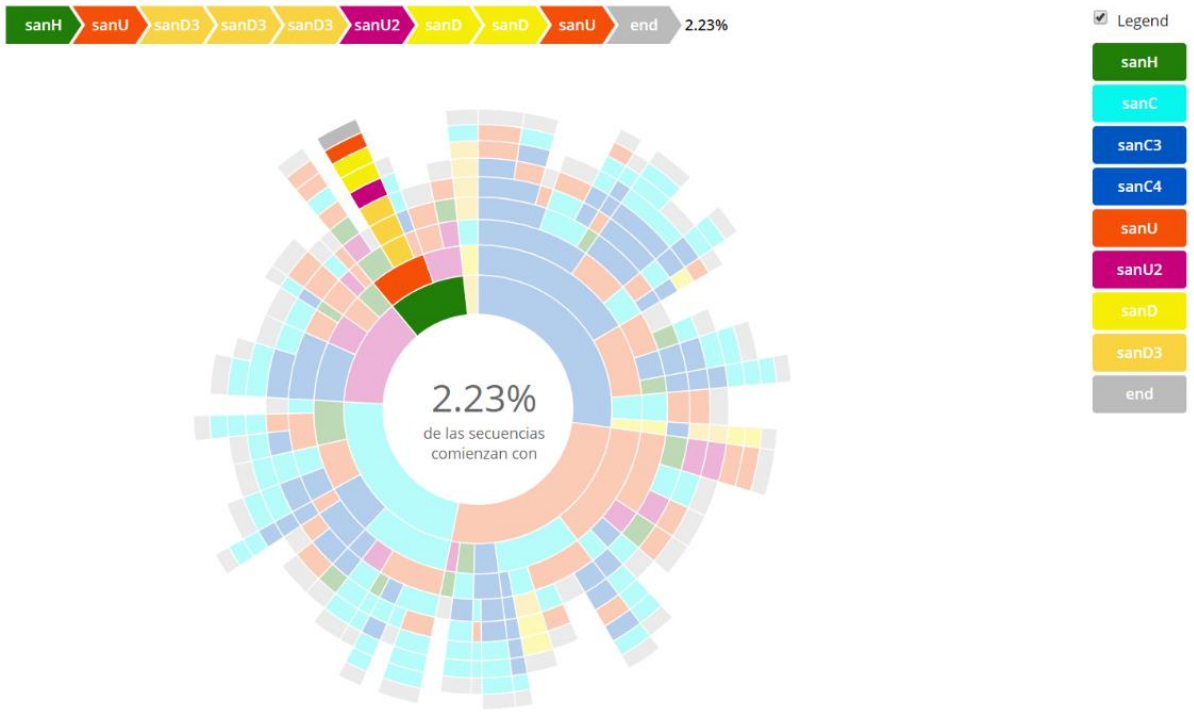


Figura 76 Caso B: Visualización “Sequences sunburst” de una secuencia de comportamiento

En este caso visualizamos la selección de una trayectoria concreta de actividades sanitarias específicas, con su correspondiente ruta de “tokens” y el conjunto de pacientes cubierto por dicha trayectoria.

10. Evaluación de los resultados MSC2

La evaluación de la capacidad de segmentación y análisis adicional de la metodología MSC2 ha sido contrastada mediante las siguientes propiedades:

Generalización:

Hemos aplicado la metodología completa en dos modelos de negocio diferentes, mostrando que se trata de una herramienta cuyo uso puede generalizarse sin mayor dificultad. Y que puede aplicarse a varios y diversos escenarios de segmentación y análisis de clientes, ciudadanos y pacientes.

Complementariedad:

Hemos contrastado la capacidad de segmentación de la nueva metodología comparada con la segmentación tradicional de clientes. Los resultados muestran que esta nueva forma de agrupar a los clientes no sustituye la anterior, sino que la complementa.

La complementa porque descubre la existencia de micro segmentos dentro de los anteriores, sobre los cuales es posible accionar desde un punto de vista de negocio.

Visualización:

La visualización de las secuencias y patrones es determinante para la comprensión y aceptación de la nueva segmentación propuesta por parte de la comunidad de usuarios, tanto de negocio como de tecnología.

Después de los sucesivos pilotos, ensayos y pruebas nos hemos decantado por los algoritmos de visualización de secuencias y patrones que mejor funcionan.

Estas herramientas han demostrado ser una interfaz analítica muy eficaz e intuitiva.

Opinión del Panel de expertos:

Agapito Ledezma Espino, PhD

Profesor Titular del Departamento de Informática de la Universidad Carlos III de Madrid. Área de Conocimiento: Ciencia de la Computación e Inteligencia Artificial. Especialista en Aprendizaje Automático e Inteligencia Artificial.

“La metodología MSC2 puede ser considerada como una guía de trabajo para el modelado, análisis y visualización de patrones de comportamiento de clientes. Dicha metodología está alineada con la metodología CRISP-DM para procesos de minería de datos. Además,

incorpora coherencia y orden en los procesos y algoritmos. Por otro lado, resulta interesante que también incluya buenas prácticas para el análisis, detección y agrupamiento de patrones en diferentes dominios".

Laura Blanco Solari

Directora de marketing senior, especialista en transformación digital e inteligencia de clientes. Extensa trayectoria como Directora de Clientes para Europa en Value Retail, Directora de Marketing en Las Rozas Village, Directora de Innovación en Grey Group, y Directora General en MRM//McCann y en Contrapunto BBDO. Pionera en desarrollo de estrategias de clientes, CRM y marketing basado en datos para instituciones financieras y negocios de venta directa.

"Creo que MSC2 es una metodología que, por primera vez, permite formalizar y sistematizar el análisis y la segmentación dinámica de clientes con una clara visión de negocio. Además, crea un espacio y un lenguaje común entre marketing/ventas y tecnología, entre equipos técnicos y de negocio, a la hora de abordar proyectos de innovación alrededor del cliente y su relación con marcas y productos a lo largo del tiempo.

Los resultados de la metodología se interpretan sin dificultad por los usuarios de negocio y la incorporación de inteligencia artificial a la generación de campañas de marketing y ventas es muy pragmática y eficaz.

La incorporación de variables innovadoras, con una gran capacidad descriptiva, permiten determinar patrones de comportamiento muy interesantes y ayudan a descubrir oportunidades, "moments of truth", sobre todo anticipar comportamientos, mejorando claramente el resultado del Plan de Marketing y el ROI".

Joaquín de Aguilera, PhD

Codirector del Customer analytics farm, Director académico del Master in digital marketing, y Profesor asociado de la Escuela de Ciencias humanas y Tecnología del Instituto de Empresa. Especialista senior en marketing y comunicación.

"He colaborado en la coordinación de la aplicación de la metodología MSC2 durante su pilotaje en tres compañías de sectores diferentes. Esto nos permite afirmar que la misma constituye un marco de trabajo ordenado, que se centra en las necesidades de segmentación para obtener resultados prácticos que eran casi imposibles de obtener hasta ahora. El resultado de la metodología permite describir y accionar sobre nuevos segmentos de clientes a través de campañas de marketing".

11. Conclusiones

Los resultados muestran que la metodología complementa el abordaje tradicional con sus capacidades adicionales de segmentar y analizar el comportamiento de los clientes, ciudadanos y paciente, permitiendo mejorar la gestión de su valor, su fidelización, y su experiencia.

La nueva segmentación generada por medio del empleo de la metodología MSC2 contribuye a la mejor y más adecuada segmentación porque incorpora las siguientes nuevas características:

- **Dinámica**:

Esta segmentación dinámica incorpora el concepto de temporalidad y cronología en el comportamiento de los clientes. Un elemento que normalmente no recoge la segmentación tradicional, que es mucho más estática. Esta dinamicidad está mucho más en línea con la velocidad del comportamiento del nuevo consumidor.

- **Descriptiva**:

Esta segmentación es auto explicativa, los usuarios de negocio pueden entender los componentes principales del comportamiento y su secuencia. También pueden descubrir el "customer journey" a través de múltiples canales y puntos de contacto a lo largo del ciclo de vida del cliente. Esta novedosa visión transversal puede incorporarse fácilmente a los modelos analíticos existentes en la actualidad.

- **Accionable**:

La propia semántica de cada segmento o tópico de comportamiento permite definir una estrategia específica de relación con el cliente, y una acción táctica concreta en cualquiera de los canales de contacto en tiempo casi real. De esta forma permite gestionar a los clientes cuyo comportamiento pone en riesgo la fidelidad o su valor, a una velocidad más adecuada.

- **Granular**:

Esta segmentación descubre micro segmentos de comportamiento homogéneo, lo cual permite gestionar comunicaciones y experiencias más relevantes y personalizadas con los clientes. Siempre manteniendo un equilibrio entre el soporte de cantidad de clientes del segmento, y la cantidad de segmentos o tópicos. Este equilibrio es muy necesario para poder gestionar campañas de marketing de forma efectiva.

- **Anticipativa**:

La naturaleza temporal de la segmentación basada en el comportamiento permite identificar reglas predictivas con cierto nivel de confianza. Dichas reglas permiten a los decisores de negocio anticiparse a potenciales movimientos futuros de los clientes, ciudadanos y pacientes. De esta forma se puede optimizar la experiencia global del cliente, al tiempo que se gestionan las oportunidades o las amenazas de negocio.

- **Complementaria**:

Finalmente, esta nueva segmentación es complementaria de las actuales en uso. Nos las reemplaza, sino que las enriquece al permitir distinguir diferentes patrones de comportamiento dentro de los actuales segmentos de cliente.

También hemos identificado ciertas limitaciones en la metodología MSC2.

- **Limitaciones**:

Durante la evaluación de la metodología MSC2 hemos identificado una serie de escenarios donde la misma tiene una limitada aportación. Ocurre en los modelos de negocio de compra esporádica, donde no es posible recoger suficientes trazas de eventos para cada cliente, ni existen patrones de actividad. También ocurre en los demás modelos de negocio donde hay grupos de clientes con escasa actividad observable, por el mismo motivo anterior.

12. Líneas de trabajo futuras

Las posibles líneas futuras del presente trabajo son:

- **Optimización de la codificación básica**:

Identificar y adaptar un algoritmo de búsqueda en un espacio de soluciones para resolver el problema de la codificación óptima de secuencias. De forma que se maximice la creación automática de secuencias con mayor capacidad de segmentación efectiva.

- **Temporalidad entre eventos**:

Identificar y adaptar un algoritmo de regresión para poder estimar de forma automática la temporalidad entre eventos de cada cliente. Esta información enriquece las reglas actuales de predicción.

- **Ausencia de eventos para alimentar modelos predictivos**:

Extender el algoritmo de generación de secuencias para incorporar un token distinguido que represente la ausencia de evento o acción del cliente. Esto permitirá analizar y predecir patrones de clientes “dormidos” dentro de las secuencias.

- **Software de automatización**:

Finalmente, otra línea futura de trabajo consiste en desarrollar una pieza de software para automatizar el flujo de datos y procesos mediante la integración de herramientas de ETL y APIs de las plataformas en la nube, y de “machine learning”.

13. Referencias y enlaces

[00] Casariego N. (2017). Un modelo de datos común para la representación del comportamiento observado del cliente. Revista ICONO14 Revista Científica De Comunicación Y Tecnologías Emergentes, 15(2), 55-91. https://doi.org/10.7195/ri14.v15i2.1078

[01] Forrester (2011). The Age of the Customer. https://go.forrester.com/blogs/category/age-of-the-customer/

[02] Blasingame J.(2014). The Age of the Customer: Prepare for the Moment of Relevance. SBN Books.

[03] Smith W. R. (1956), Product differentiation and Marketing segmentation as alternative Marketing Strategies. American Marketing Association, Marketing Management Journal 1995 reprint.

[04] Wedel M. & Kannan P. K. (2016). Marketing analytics for Data-rich environments. AMA/MSI Journal of Marketing. http://dx.doi.org/10.1509/jm.15.0413

[05] Fader P.S. & Toms S. (2012). Customer centricity: focus on the right customers for strategic advantage. Ed. Wharton Digital Press.

[06] Chen D. & Sain S.L. & Guo K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Macmillan Publishers, Journal Database Marketing & Customer Strategy Management Vol. 19(3), pp. 197–208.

[07] Fader P.S. & Hardie B.G.S. & Lee K.L. (2005). RFM and CLV: Using Iso-Value Curves for Customer Base Analysis, Journal of Marketing Research, 2005, Vol. 42, No. 4, pp. 415-430. https://doi.org/10.1509/jmkr.2005.42.4.415

[08] Marr B. (2012). Key Performance Indicators. The 75 measures every manager needs to know. FT Publishing Pearson, 2012, pp. 85-89, 97-100, 159-163.

[09] Chaffey D. (2018). How to use (Advanced) Segments in Google Analytics. https://www.smartinsights.com/google-analytics/google-analytics-segmentation/segmenting-google-analytics/

[10] Casariego N. (2016). Segmentación basada en comportamiento de clientes. IE School of Human Sciences & Technology, Customer Analytics Farm.

[11] Kimball, R., & Ross, M. (2013). The data warehouse toolkit the complete guide to dimensional modeling. New York, NY: Wiley.

[12] Kimball, R. (2015). Design Tip #176 Dimensional Models – Logical or Physical?. https://www.kimballgroup.com/2015/07/design-tip-176-dimensional-models-logical-or-physical/

[13] Xu R. & Wunsch D.C. (2005). Survey of Clustering Algorithms. IEEE Transactions on Neural Networks 16(3):645 - 678. DOI: 10.1109/TNN.2005.845141

[14] Linoff G. & Berry M.J.A. (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship management. Wiley.

[15] Anova analytics (2017). Customer churn predictive models. https://www.r-bootcamp.com/customer-churn-predictive-models/

[16] Fournier-Viger P., Chun-Wei J., Uday Kiran R., Sing-Koh Y., Thomas R. (2017). A Survey of Sequential Pattern Mining. http://www.philippe-fournier-viger.com/dspr-paper5.pdf

[17] Rajkomar A. et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine (2018) 1:18. doi:10.1038/s41746-018-0029-1

[18] Barbieri N. et al. (2013). Probabilistic topic models for sequence data. Mach Learn (2013) 93:5–29. DOI 10.1007/s10994-013-5391-2

[19] Marmelab. Recuperado de: https://marmelab.com/EventDrops/

[20] Bostok M. (2018). Recuperado de: https://observablehq.com/@d3/collapsible-tree

[21] Roden K. (2019). Recuperado de: https://bl.ocks.org/kerryrodden/766f8f6d31f645c39f488a0befa1e3c8

[22] CRISP-DM.org (1996). Recuperado de: https://data.sngular.com/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science

[23] Conder P. (2015). The Customer Journey of Past and Present. Lenatti blog, recuperado de: https://www.lenati.com/blog/2015/02/customer-journey-past-and-present/

[24] BigML (2017). Topic modeling. Recuperado de: https://bigml.com/features/topic-model

[25] Tableau Prep (2018). Combine, shape, and clean your data for analysis with Tableau Prep. Recuperado de: https://www.tableau.com/products/prep