UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

DEPARTAMENTO DE INGENIERIA INDUSTRIAL

PRONÓSTICO PARA LA MOROSIDAD DE CLIENTES DE TARJETAS DE CRÉDITO DE UN RETAIL FINANCIERO, MEDIANTE EL USO DE DATOS TRANSACCIONALES E HISTORIAL DE PAGO

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

ÁLVARO IGNACIO GÁLVEZ BRAVO

PROFESOR GUÍA:

ALEJANDRA PUENTE CHANDÍA

MIEMBROS DE LA COMISIÓN:

CAROLINA SEGOVIA RIQUELME

JUAN ROMERO GODOY

SANTIAGO DE CHILE

2018

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Álvaro Ignacio Gálvez Bravo
FECHA: 04/11/2018
PROFESOR GUIA: Alejandra Puente Chandía

PRONÓSTICO PARA LA MOROSIDAD DE CLIENTES DE TARJETAS DE CRÉDITO DE UN RETAIL FINANCIERO, MEDIANTE EL USO DE DATOS TRANSACCIONALES E HISTORIAL DE PAGO

En Chile el año 2017 se comercializaron US$ 3041 MM con tarjetas de crédito, asociadas a 21 MM de cuentas activas entre la banca y otros tipos de emisores. Estos últimos, principales responsables del aumento de la cobertura hacia segmentos históricamente no bancarizados. Así, el modelo de estos actores se ha caracterizado por un mayor riesgo compensado con carteras de clientes de gran volumen, siendo la gestión de cobranza muy importante por su efecto directo en sus utilidades.

El objetivo de este trabajo es la caracterización del comportamiento de la morosidad de los clientes de la tarjeta de crédito de un *retail* financiero, de cara a potenciales mejoras en la cobranza y acciones preventivas. Así, esta memoria se aborda con metodología CRISP-DM, aproximando la morosidad de clientes mediante un modelo de pronóstico basado en cadenas de Markov y estrategias de *clustering* para la inclusión de heterogeneidad. Esta última, es realizada a nivel de grupo vía *hard clustering* y a nivel de cliente a través de lógica difusa. De esta forma la información considerada, abarca tanto el comportamiento de mora, como el historial transaccional y características sociodemográficas de clientes. Por último, dentro de los modelos de Markov, se incluye directamente el efecto del nivel de morosidad, además de los abonos a la deuda, determinando su efecto sobre el pronóstico de los próximos estados.

En relación a los resultados obtenidos, destaca la identificación de grupos de clientes que se asocian a distintos perfiles de morosidad. Así, de la aplicación de cadenas de Markov sobre los clústeres, se reconocen grupos caracterizados por un buen comportamiento pasado, cuyas probabilidades de pago ante un episodio de mora son muy altas, además de segmentos en los que cuándo cesan sus pagos, el tránsito al castigo es casi directo. Es en este último tipo de clientes que se identifica un mayor efecto del abono a la deuda, definiendo un subgrupo que en situación de mora disminuye su probabilidad de castigo. Respecto al pronóstico a nivel de cliente, se establece la necesidad de mejorar su desempeño, siendo la recomendación aplicar solo las propuestas a nivel de grupo, que radican en la priorización de las acciones más efectivas en los segmentos con mejores perspectivas, y que la estimación económica ha valorado con cota inferior en los $18 MM.

Finalmente, como trabajo futuro, se propone el estudio de la morosidad en niveles de granularidad mayor, así como el uso de los resultados obtenidos para la maximización de la esperanza del recupero, además de la inclusión directa de variables de estado relacionadas con los montos en mora y envejecimiento de la deuda. Por último, para el de comportamiento a nivel de cliente se destaca el potencial de los modelos de supervivencia, como alternativa para mejorar el desempeño de este trabajo.

A mi familia…

A mis amigos…

Y a todos los que me han acompañado

para llegar hasta aquí…

Tabla de contenido

Índice de tablas

Índice de ilustraciones

1 Introducción

No es azar que para el año 2018, las empresas de mayor valor bursátil se encuentren consistentemente ligadas a sectores tecnológicos, integrando en su modelo de negocios los datos de sus clientes para ofrecer soluciones cada vez más personalizadas. Google, Amazon y Facebook son algunos ejemplos. Estas compañías en sus orígenes harían de la información su principal negocio y activo. En este contexto, el resto de las organizaciones se han visto forzadas a imitar e incorporar estos enfoques a sus negocios.

No obstante, existe una industria en que la toma de decisiones y operación basada en información cuantitativa de los clientes; no es una tendencia reciente. En el rubro financiero, el potencial económico de la información es conocido hace varias décadas, y con la masificación de algoritmos de aprendizaje automático no ha hecho más que aumentar su prevalencia[1]. Así lo recogerían Rosenberg y Gleit (Rosenberg & Gleit, 1994) hace más de 20 años. En ese entonces, ya podrían identificar diversos casos de éxito, en los que la hoy denominada ciencia de los datos había generado importantes mejoras en la operación de este tipo de negocios. Minimización de pérdidas en la extensión de crédito, o la identificación de las personas susceptibles de aumentos o disminuciones de cupo, son algunas de las aplicaciones que hasta el día de hoy se trabajan y se intentan mejorar. Así, es evidente que el buen uso de la información disponible es crucial en la industria financiera, ya sea para acotar riesgos y o maximizar utilidades. De modo que, el uso de los datos no es tan reciente como podría pensarse.

La industria financiera chilena, es un segmento altamente competitivo del cual participan numerosos actores, ofreciendo una diversa variedad de servicios tanto a personas como empresas. En particular, dentro del segmento de consumo se encuentran las tarjetas de crédito. En el país este mercado tiene un volumen de ventas cercano a los US$ 3041 MM de (aproximadamente el 15% del PIB chileno) y registra unos 21 millones de cuentas activas distribuidas entre la banca tradicional y otros tipos de emisores, como los *retailers*. Estos últimos, son los principales responsables de la ampliación de la cobertura de estos productos hacia segmentos tradicionalmente no bancarizados. Así, el modelo de negocio de estos actores se ha caracterizado por un mayor riesgo asociado, compensando el mismo con carteras de clientes de gran volumen.

Los emisores no bancarios de tarjetas, al tomar carteras más riesgosas, enfrentan una gran presión por efectuar una correcta gestión del crédito. La cobranza de las deudas impagas es parte de este desafío, pues el hacerlo eficientemente es crítico. Siendo este el punto de partida de esta memoria.

[1] Los costos de almacenamiento en términos de US$/Mb han caído en de forma sustancial en el último tiempo. Solo en los pasados 20 años los discos rígidos experimentarían una diminución en múltiplos de hasta 100.000 su valor. Al mismo tiempo, las alternativas para estimar distintos modelos matemáticos y computacionales se han masificado de tal forma, al igual que las fuentes de datos disponibles, determinando en gran medida del auge de la ciencia de los datos.

La empresa en la que se realiza este trabajo es parte del brazo financiero de uno de los *retailers* más exitosos del país. El *holding*, tiene su cara más visible en sus negocios minoristas. No obstante, contraria a la percepción general, el brazo financiero del grupo constituye la parte central de su modelo de negocio. Siendo uno de los principales emisores y operadores de tarjetas de crédito tanto en Chile, como otros países de Latinoamérica. Así, pese a que la empresa naciera como un *retailer*, los servicios financieros se han convertido en el núcleo del negocio.

En Chile, la organización enfrenta un potencial riesgo en base al incremento de los costos asociados a la morosidad de sus clientes. Esto se explica en el aumento registrado en este apartado, al considerable tamaño de su cartera y a recientes cambios regulatorios en el cálculo de provisiones. Elementos que impactan directamente sobre la última línea, siendo esta la importancia de abordar un tema como este. Estos aspectos han determinado que el negocio haya priorizado el estudio de la cobranza y morosidad, aun cuando la proporción de compromisos no pagados es en general menor que la de sus pares, e incluso más pequeña que la de algunos actores de la banca tradicional.

Históricamente, la empresa ha sido un aliado importante de la universidad, siendo posible encontrar una diversidad de trabajos conjuntos, memorias y otros tipos de investigaciones. Problemáticas como la fuga de clientes, propensión al consumo de productos y experimentación en acciones de *marketing directo*; son temáticas que han sido abordadas. En contraste, la investigación en torno a cobranza y morosidad no tiene antecedentes recientes. Así, su novedad y relevancia justifican el desarrollo de un tema de memoria en torno a este tópico, tanto desde el punto de vista del negocio como el académico.

1.1 Descripción de la industria

Dentro de la industria financiera, el mercado de las tarjetas de crédito se encuentra conformado por: la banca tradicional, las sociedades bancarias de apoyo al giro, las cooperativas de ahorro y crédito, y los emisores de tarjetas no bancarios[2]. Todos ellos son regulados principalmente por la Superintendencia de Bancos e Instituciones Financieras (SBIF).

La competencia en el segmento de tarjetas de crédito se desarrolla de manera casi transversal entre las instituciones bancarias y las que no lo son. Esto ocurre desde que las casas comerciales, se aliaran con gestores internacionales como Visa y MasterCard, permitiendo así la compra fuera de sus negocios y no solo en aquellos comercios asociados[3]. De esta forma, para efectos de determinar la composición de la industria, a

[2] Los emisores de tarjetas de crédito no bancarios son representados principalmente por *retailers* o casas comerciales, por ejemplo, CMR Falabella, Presto, La Polar y otros.

[3] Por comercios asociados debe entenderse empresas externas al grupo emisor de la tarjeta, las cuales suscriben un acuerdo para aceptar el plástico como un medio de pago válido en un Punto de venta.

partir de la Figura 1 y Figura 2, se procede a evidenciar la competencia transversal aludida entre los diferentes tipos de actores.

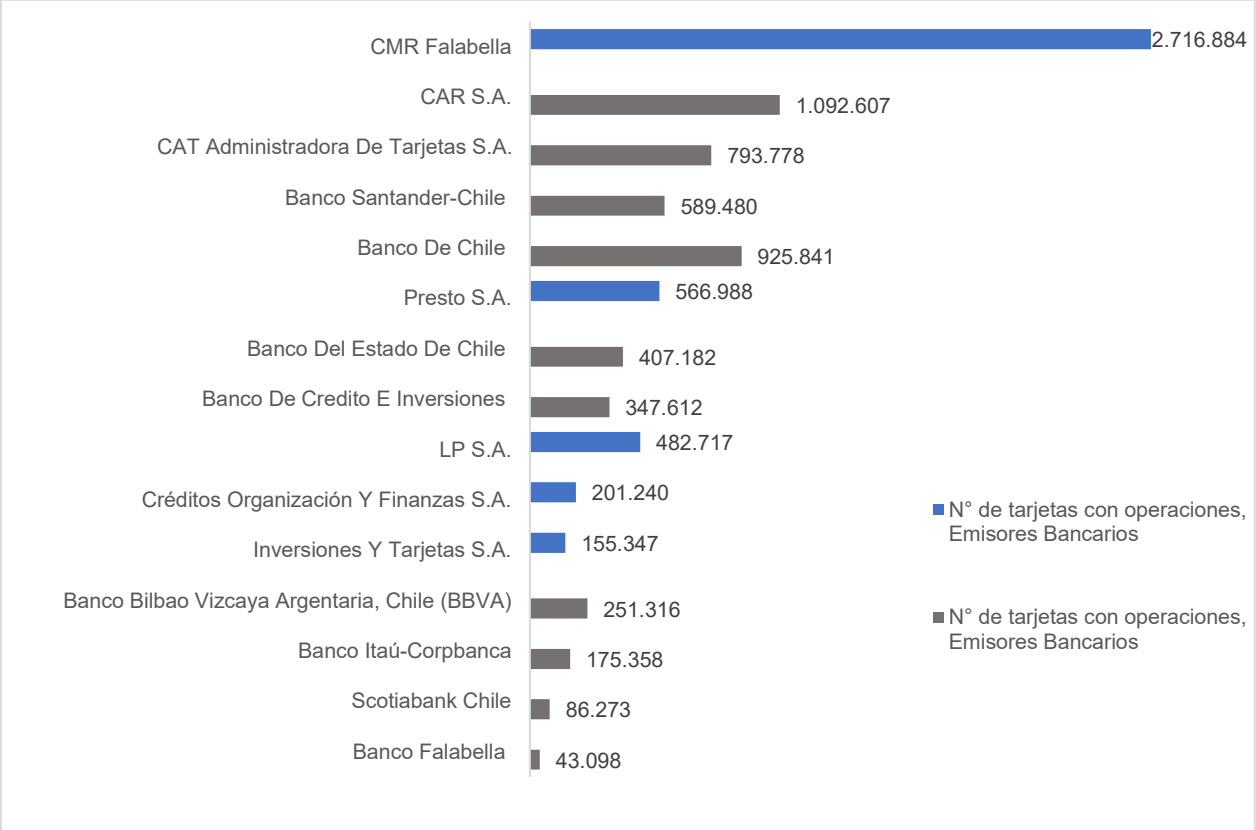


Figura 1. Número de tarjetas con operaciones individualizados por emisor y tipo de organización (marzo 2017)

Nota. Análsis basado en informe de la Superintendencia de Bancos e Instituciones Financieras Chile. (07 de agosto de 2017). Informe de Tarjetas de Crédito No Bancarias - Serie [May- 2017] y Superintendencia de Bancos e Instituciones Financieras a. (mayo de 2017). Informe de Tarjetas de Crédito - Nueva Versión Serie Mayo. Fuente: Elaboración propia.

De la Figura 1, es posible observar que el principal emisor y operador es una institución no bancaria; CMR Falabella, con un total de 2.716.884 plásticos con operaciones. Este hecho, sumado a que los primeros lugares de participación son compartidos entre los distintos tipos de emisores, da cuenta de la competitividad transversal antes mencionada. Esto sin referirse al hecho que, tanto CAR S.A. (filial de Banco Ripley) y CAT Administradora de tarjetas S.A. (Cencosud), ambas sociedades de apoyo al giro bancario, por su historia pueden ser asociadas de igual forma al *retail*.

La empresa en la que se desarrolla esta memoria se clasifica como un emisor de tarjetas de crédito no bancario, por lo tanto, regulatoriamente dicha condición tiene implicancias sobre la operación del negocio. A diferencia de los bancos que son informados por la SBIF sobre la morosidad de sus aplicantes, los emisores no bancarios, no reciben este tipo de boletines. Además, la deuda con ellos no es consolidada en el sistema financiero. Esto se traduce en que estos emisores no puedan observar tan acuciosamente el nivel de deuda de sus clientes. Por lo tanto, sería razonable que para efectos de comparar con la industria solo se considerasen aquellos emisores del mismo tipo. No obstante, dado que en términos de tarjetas activas y volúmenes de venta (ver Figura 2) la empresa

compite directamente frente a la banca, pues se encuentra en los primeros lugares. Finalmente, se opta por contrastarla con el segmento bancario, dada su posición (La organización se encuentra entre las primeras 10 instituciones con mayor monto).

Figura 2. Monto de operaciones con tarjetas individualizados por emisor y tipo de organización (marzo 2017)

Nota. Datos: Superintendencia de Bancos e Instituciones Financieras Chile. (07 de agosto de 2017). Informe de Tarjetas de Crédito No Bancarias - Serie [May- 2017] y Superintendencia de Bancos e Instituciones Financieras a. (mayo de 2017). Informe de Tarjetas de Crédito - Nueva Versión Serie Mayo. Fuente: Elaboración propia.

1.2 Descripción de la empresa

Como consecuencia de la confidencialidad de este proyecto, no es posible describir en detalle la organización. De todas formas, se profundiza en algunas características generales que permiten dar cuenta del contexto en la que esta se inserta.

La empresa es un emisor de tarjetas de crédito del tipo no bancario, siendo uno de los principales actores en esta categoría. Así, la propuesta de valor de la organización pasa por el ofrecimiento de servicios financieros de diversa índole. Estos a su vez, son potenciados por los diferentes puntos de contacto con el cliente en distintos sectores del *retail*, a través de otros negocios del *holding*.

Respecto al tipo de clientes, estos pueden caracterizarse en base a sus diferencias en comparación con los de la banca. Esto se explica en los menores requerimientos para acceder al crédito en este tipo de empresas, por lo que, junto con construir carteras más

riesgosas, también permiten que estas organizaciones lleguen masivamente a las personas. De ahí que los principales operadores se concentran en las casas comerciales más exitosas y bancos de consumo masivo.

Finalmente, en cuanto a ventajas competitivas, la empresa cuenta con una imagen de marca reconocida, que se apoya fuertemente en la filial de *retail* del *holding*. Además, la tarjeta cuenta con tres pilares fundamentales que explican la base de sus clientes: una alternativa sencilla de acceso al crédito, aún para personas no bancarizadas; un club de beneficios que fomenta el uso del plástico y privilegia las compras dentro del *holding*; y una gama de promociones y alianzas asociadas al uso de la tarjeta como medio preferente de pago.

1.3 Descripción del proyecto y justificación

Para este trabajo, se ha dispuesto abordar la morosidad de los clientes de la tarjeta de crédito del negocio. Esto se justifica en la situación que vive la empresa, consistente en el aumento de los niveles de morosidad de sus clientes, al interés en el estudio de temas de morosidad y cobranza por parte de la organización, y a los pocos antecedentes de trabajos de título que aborden una temática similar. Así, de acuerdo con cifras publicadas por la SBIF, para mayo de 2017, del total de la cartera (en MM$) aproximadamente un 7,1%, se encontraba en situación de mora, es decir entre 1 o 364 días de morosidad.

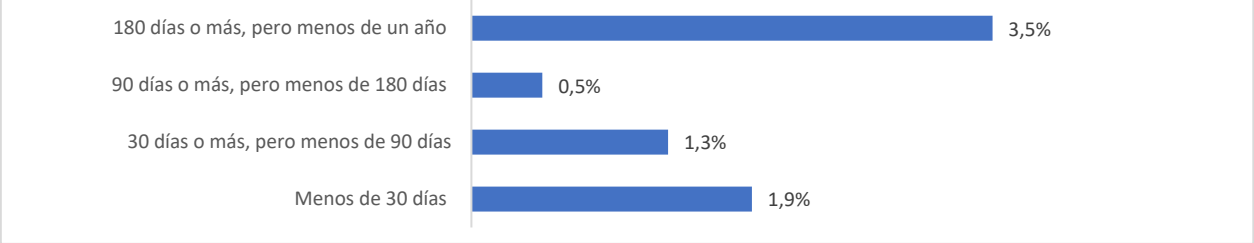


Figura 3. Distribución de tramos de mora, para la cartera de clientes de una empresa de retail financiero, mayo de 2017

Nota. Datos : Superintendencia de Bancos e Instituciones Financieras Chile. (07 de agosto de 2017). *Informe de Tarjetas de Crédito No Bancarias - Serie [May- 2017].* Fuente: Elaboración propia.

La distribución de los diferentes tramos de mora considerados por el ente regulador puede apreciarse en la Figura 3. Resulta claro que esta condición se concentra en las categorías extremas. Sin embargo, este 7,1% es poco interpretable de no ser comparado con el resto de la industria. Así, al contrastar con otros actores de *retail financiero* se evidencia que el valor presentado es menor que el de sus pares, situación que habla bien de la gestión del riesgo de crédito de la compañía. No obstante, no debe perderse de vista que dada la magnitud de la cartera de clientes de la empresa, esta cifra sigue siendo significativa. Por otra parte, aun teniendo el mejor desempeño entre los emisores no bancarios, es posible observar que el incremento en los niveles de morosidad no es transversal a la industria, lo cual es una señal de alerta (ver Figura 4).

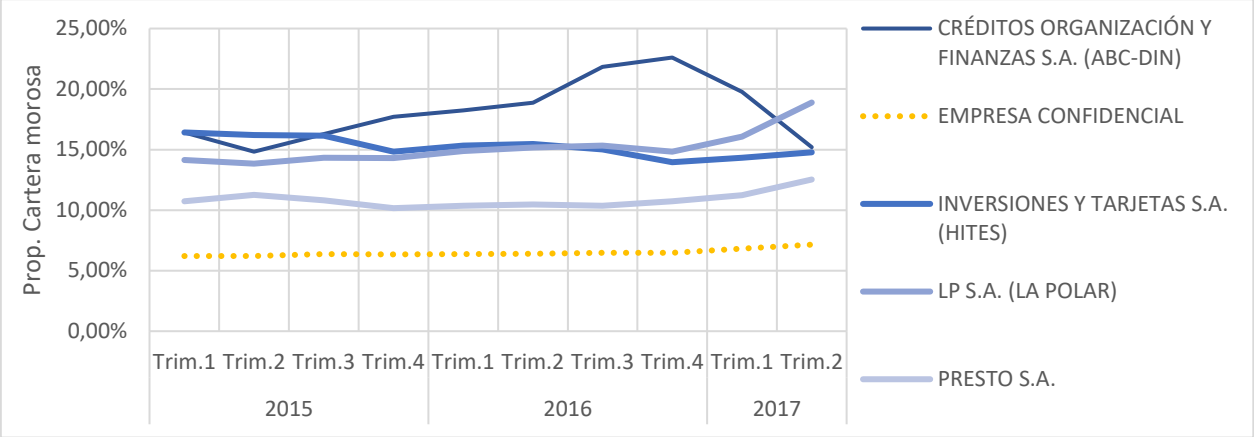


Figura 4. Evolución morosidad general Emisores no bancarios

Nota. Fuente: Elaboración propia. Datos:Superintendencia de Bancos e Instituciones Financieras Chile. (07 de agosto de 2017). *Informe de Tarjetas de Crédito No Bancarias - Serie [May- 2017]*

En contraste, si se compara la situación del negocio respecto a los emisores bancarios, se tiene un resultado diferente. A partir de la Figura 5 se puede observar que los niveles de morosidad no son tan bajos. De los principales actores del segmento bancario, la única organización con un peor desempeño corresponde a CAT Administradora de Tarjetas (Cencosud). Por el contrario, la otrora competencia directa representada por CAR S.A.[4] (Ripley), presenta un indicador mejor que la empresa de esta memoria. Asimismo, la realidad experimentada por organizaciones como Santander y Banco de Chile, cuyos niveles de morosidad en ambos casos se sitúan bajo el 3%, vislumbran el espacio de mejora que esta memoria propone capitalizar.

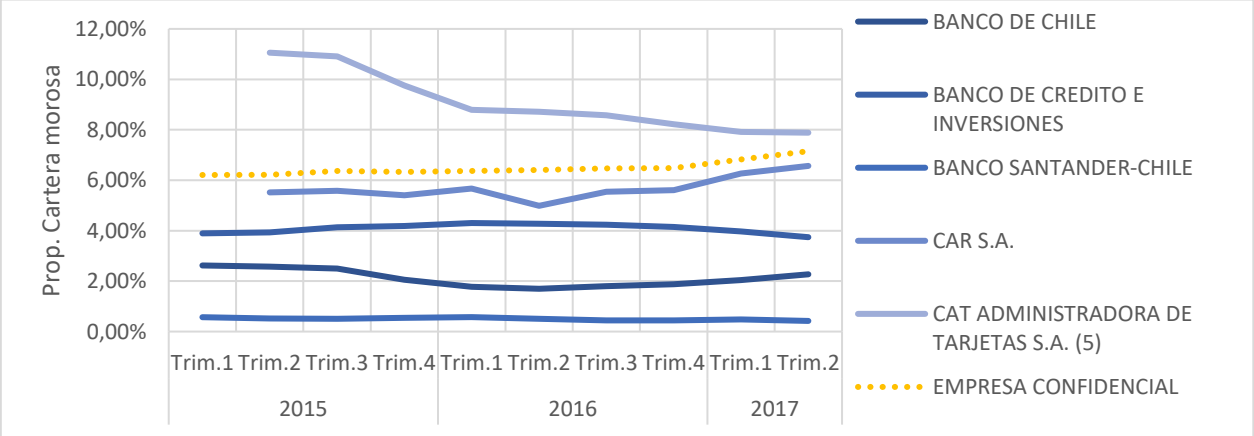


Figura 5. Evolución morosidad general Emisores bancarios

Nota. Datos: Superintendencia de Bancos e Instituciones Financieras Chile. (07 de agosto de 2017). Informe de Tarjetas de Crédito No Bancarias - Serie [May- 2017] y Superintendencia de Bancos e Instituciones Financieras a. (mayo de 2017). Informe de Tarjetas de Crédito - Nueva Versión Serie Mayo. Fuente: Elaboración propia.

Para efectos de concretizar la oportunidad de negocio de este trabajo, se identifica que los principales beneficios de una potencial reducción de la cartera morosa corresponden a: la disminución de las cuentas por cobrar que pasan a ser pérdidas, i.e. se transforman

[4] CAR.S.A. es una sociedad que administra de manera conjunta con el banco Scotiabank, la tarjeta Ripley.

en cuentas castigadas, y la baja en el costo de oportunidad asociado a la mantención de provisiones[5].

Sobre las causas del deterioro de la cartera en términos de morosidad, se tiene como hipótesis que el modelamiento existente no está capturando completamente el comportamiento de los clientes. Por este motivo, las áreas encargadas de esta gestión no pueden hacerlo de la mejor manera posible.

Hoy en día, existen dos modelos de morosidad; uno de mora temprana (1 a 89 días) y otro de mora tardía (90 a 180 días), de los cuales solo uno de ellos es empleado en gestión[6]. Estas aproximaciones, datan de hace una cantidad considerable de tiempo. Así, se teoriza que un nuevo enfoque para abordar el tema podría mejorar la situación actual de la empresa. Determinando nuevos perfiles de clientes a gestionar, y complementando el trabajo actual de la institución. Para efectos de concretizar las hipótesis planteadas, se espera que la mejor captura de la conducta sea conseguida con la adición de nuevas variables relacionadas con el comportamiento transaccional del cliente. Así, aspectos como la intensidad de uso de la tarjeta y fidelización serán considerados como parte de los análisis posteriores; incorporando una dimensión que hasta entonces no se consideraba.

En base a los elementos presentados, la alternativa de solución propuesta pretende evaluar un nuevo modelo para la identificación de clientes con riesgo de mora y sus potenciales transiciones en los diferentes niveles de gravedad. De manera de apoyar la priorización que se hace en la gestión de cobranza y prevención. Así, la propuesta de valor de este proyecto radica tanto en el mejor entendimiento de la morosidad de clientes; concretizado en un modelo de *marketing cuantitativo*, como por el uso de metodologías hasta ahora no empleadas en el área, las cuales profundizan en un tema con pocos antecedentes de trabajos de títulos anteriores. Finalmente, destaca el potencial económico del proyecto, el que para distintos escenarios de mejora en el nivel de morosidad; establece beneficios en condiciones conservadoras de alrededor de $8 MM (disminución de la probabilidad de incumplimiento en 10 puntos base respecto a la experimentada en el periodo) y que para condiciones más auspiciosas, solo por efecto de las menores chances de incumplimiento, pueden traducirse en beneficios acotados inferiormente en torno a los $185 MM (ver 5.5.1 Evaluación económica).

[5] Una provisión es un pasivo que considera el caso en que cliente no salda su deuda con la empresa. Por ello anticipándose a la eventualidad, se toman resguardos ante esta pérdida potencial. El cálculo de estos montos depende de los niveles de mora pasados. Así, reducciones en los niveles de morosidad impactan positivamente en el provisionamiento futuro (menor costo de oportunidad asociado a provisiones).

[6] Ambos modelos se encuentran basados en árboles de decisión, cuya variable a explicar es el avance a un estado de morosidad de mayor gravedad, e.g. de 30 días de mora a 60 días, siendo el más utilizado el de mora temprana.

2 Objetivos y alcances

Los objetivos de esta memoria se orientan al estudio de la morosidad de clientes de la tarjeta de crédito del negocio, con el fin de permitir la elaboración de recomendaciones que orientadas a disminuir los niveles de esta. A continuación, se presenta el objetivo general del y aquellas metas específicas que tributan al cumplimiento de este.

2.1 Objetivo general

“Caracterizar el comportamiento de morosidad de los clientes de la tarjeta de crédito de un retail financiero, mediante un modelo de pronóstico en base a datos transaccionales e historial de pago, con el fin de proponer de recomendaciones orientadas a la mejora de los niveles de morosidad experimentados por el negocio.”

2.2 Objetivos específicos

Objetivo específico 1
Identificar y definir variables que permitan dar cuenta del historial de morosidad y pago del cliente en la empresa.

Objetivo específico 2
Desarrollar un modelo de pronóstico para la evolución de los tramos de morosidad de los clientes del negocio.

Objetivo específico 3
Caracterizar grupos de clientes en base a los comportamientos de morosidad identificados, proponiendo acciones focalizadas, para efectos de reducir los niveles de morosidad del negocio.

2.3 Alcances

Se ha definido que el foco de esta memoria es el ajuste y evaluación de un modelo para el pronóstico de morosidad de clientes de la empresa. Este consta de dos segmentaciones, agrupamientos que ocurren antes del ajuste de cadenas de Markov para el modelamiento del comportamiento de mora de los clientes. Así, los límites del trabajo quedan definidos por los siguientes puntos:

- La integración del modelo a desarrollar en del proceso del negocio de la empresa, i.e. su paso a producción no es considerado parte de esta memoria.
- La información contemplada se remite a la disponible en el *Warehouse* de la empresa, datos transaccionales, comportamiento de pago, situación de morosidad e información del cliente.

- Los datos disponibles se sitúan entre el año 2012 e inicios de 2018, dependiendo la fuente que se esté observando. Por lo tanto se establece trabajar con el año 2017, en vista de la información con menor historial (datos de cobranza).
- El universo de clientes es restringido por su situación contractual. Solo se consideran aquellos que al comienzo del horizonte de evaluación no se encuentren castigados. Esto responde al hecho de que no es factible salir de este estado y su inclusión como condición inicial no presentaría cambios en el periodo en estudio.
- Los modelos a construir y ceder a la empresa serán programados en *scripts* en lenguaje Oracle SQL, Python y R. Por este motivo el diseño y creación de un aplicativo no es parte de los alcances del trabajo.
- El entregable final de este trabajo corresponde a recomendaciones de negocio basadas en los resultados obtenidos y a las posibilidades de los modelos. De esta manera, no se contempla la validación experimental de los resultados.

3 Marco teórico

En Esta sección se aborda la revisión bibliográfica efectuada, además del estado del arte sobre las temáticas aludidas. Así, en primer lugar, se presentan algunos conceptos generales relacionados con el análisis de datos y la metodología de trabajo. En forma seguida, se procede a profundizar en torno a modelos para la gestión del crédito y distintos enfoques de *data mining* y estadística para abordar problemas en este campo.

3.1 Data mining

La ciencia de los datos y específicamente el término *Data Mining* ha gozado de una creciente popularidad, como resultado de la masificación de las herramientas de análisis disponible, pero más aún por el exponencial crecimiento en la data utilizable.

Existen diferentes acepciones sobre que se debe entender por minería de datos. Por ejemplo, Fayyad, Piatetsky-Shapiro y Smith, definen que: "*Data Mining* es el proceso no trivial de identificación de patrones entendibles, válidos y potencialmente útiles para el negocio, desde grandes bases de datos" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, p. 40). De esta forma, desde la perspectiva presentada el proceso de *Data Mining* es relevante, pues refiere a cómo es posible aprovechar la información contenida en grandes volúmenes de datos para obtener conocimiento aplicable; lo que en si es el objetivo de esta memoria.

3.2 Metodología CRISP-DM

Tanto en la literatura especializada como en los trabajos de título revisados, existen ejemplos de uso de la metodología KDD[7], como base para estructurar sus investigaciones. Tal es el caso de algunas de las memorias citadas a lo largo de este trabajo (Roco Benavides, 2010; Segovia Riquelme, 2005). El proceso KDD, permite estructurar la extracción de conocimiento a partir de los datos, y se encuentra conformado por una serie de pasos propuestos por Fayyad, Piatetsky-Shapiro y Smyth (Fayyad et al., 1996). Estas etapas comienzan con la determinación de los datos referentes al problema, para luego dar paso a la limpieza y transformación de estos. En forma seguida, se contempla la aplicación de modelos de *Data Mining* de manera de contribuir a la consecución del objetivo del proyecto a través de la generación de conocimiento. Finalmente, se configura una última fase en la que los resultados obtenidos son analizados e interpretados, culminando la iniciativa. La Figura 6, muestra de manera gráfica la secuencia en que el proceso KDD es aplicado sobre un proyecto de minería de datos.

[7] KDD es acrónimo *Knowledge Discovery in Databases* es un proceso de estructuración para proyectos de minería de datos.

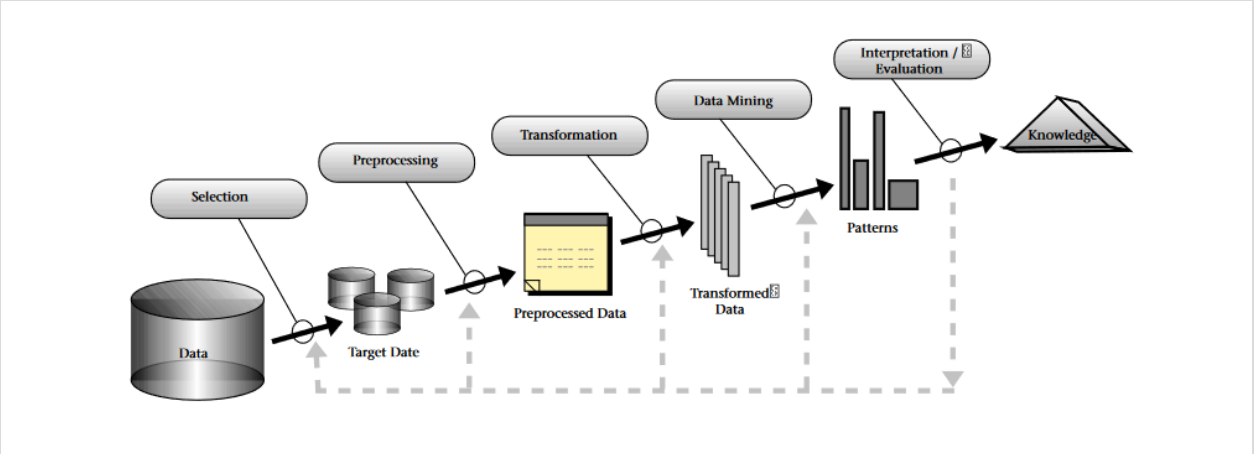


Figura 6. Etapas del proceso KDD (Knowledge Discovery in databases)

Nota. Fuente: Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37. https://doi.org/10.1609/aimag.v17i3.1230. (Fayyad et al., 1996, p. 5).

Si bien, el proceso KDD ha sido ampliamente utilizado en otros trabajos de título relacionados con datos, existen otras metodologías que permiten estructurar este tipo de proyectos. Un ejemplo es el proceso CRISP-DM, acrónimo *de Cross Industry Standard Process for Data Mining*. Las etapas de esta alternativa se encuentran consignadas en la Figura 7, y tiene como principal diferencia respecto al modelo KDD, la flexibilidad en el desarrollo de las diferentes fases. Para ello se incorpora un proceso cíclico que permite iterar los pasos en busca de mejores soluciones. Adicionalmente, explicita la necesidad de interiorizarse sobre el negocio, a fin de que nunca se pierda el contexto en el que se está trabajando (Wirth & Hipp, 2000).

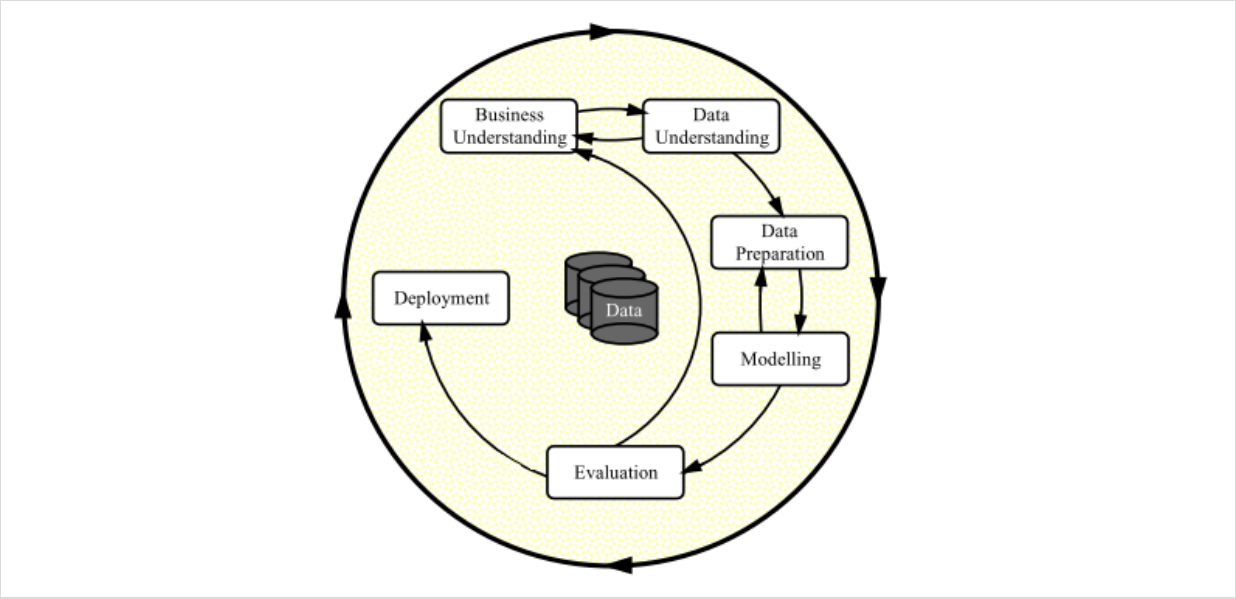


Figura 7. Fases del proceso de modelamiento CRISP-DM para Data Mining.

Nota. Fuente: Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining (PADD '00), (24959), 29–39. https://doi.org/10.1.1.198.5133. (Wirth & Hipp, 2000, p. 5)

De esta forma, basándose en la revisión efectuada por Wirth y Hipp (Wirth & Hipp, 2000), el proceso CRISP-DM puede resumirse en los pasos presentados en la Figura 7 y

explicitados a continuación (Para mayor detalle consultar ANEXO A). Nótese que, la aplicación de esta estructura al proyecto es abordada en profundidad en los capítulos destinados al marco y desarrollo metodológico (ver 4 Marco metodológico y 5 Desarrollo metodológico).

- **Comprensión del Negocio:** En esta fase se plantea la pertinencia de conocer desde una perspectiva de la organización, la lógica detrás del problema u oportunidad, al mismo tiempo que se entienden sus necesidades, de manera de focalizar los esfuerzos posteriores del proyecto.
- **Comprensión de los datos:** Al igual que el proceso KDD, el trabajar con datos requiere que exista entendimiento de sus características y se identifique la calidad de estos. Además, dentro de esta fase se incluye una primera aproximación al testeo de las hipótesis sobre el problema en estudio, de modo de detectar comportamientos en los que sea relevante profundizar en forma posterior.
- **Preparación de los datos:** En este paso se cubren los aspectos relacionados con la preparación de los datos para la aplicación de técnicas de *data mining*. Puede contemplar tanto el qué hacer con los valores faltantes, así como con la presencia de *outliers*, además de las transformaciones que puedan requerirse sobre los datos y la creación de nuevas variables.
- **Modelamiento:** Constituye la fase de construcción de modelos propiamente tal. En esta son evaluados una o más alternativas, con el propósito de buscar la mejor solución al problema planteado y calibrar los parámetros que sean susceptibles de ajustar.
- **Evaluación:** Se contempla la evaluación de los resultados obtenidos de cara a la toma de decisiones, evaluando en base a los objetivos establecidos y a métricas definidas con antelación.
- **Despliegue o implementación:** Para que los resultados sean útiles a la empresa, se considera dentro del modelo CRISP-DM un último paso en el cual se deben concretizar los *insights* de manera que sean utilizables por el negocio. De esta forma, se planifica la implementación de estos, así como su actualización , periodicidad y otras decisiones asociadas con el despliegue.

3.3 Métodos para la gestión del crédito

El problema de asistir y apoyar cuantitativamente las decisiones que se toman en el área de consumo y crédito es un tópico que ha sido recurrentemente estudiado por la literatura. La gama de decisiones cubiertas por estas metodologías, incluyen desde si se debe aprobar el financiamiento a un cliente, el cómo debe hacerse el cobro de las cuentas en mora, o que condiciones debe cumplir una persona para ser sujeto de un aumento de cupo, siendo estos algunos de los ejemplos más recurrentes.

De acuerdo con la revisión de literatura de Rosenberg y Gleit, para minimizar las pérdidas de la emisión de crédito se han desarrollado una serie de técnicas que, estadísticas o con fundamentos empíricos, intentan optimizar la exposición al riesgo que toma una compañía al proporcionar financiamiento de algún tipo. Así, estos autores clasifican las

decisiones del negocio en dos grandes grupos; una primea categoría que aborda el si se debe o no otorgar crédito a un sujeto; ejemplificado en modelos de *application scoring*. Y una segunda clase de elecciones que refiere a las acciones que se deben tomar sobre una cuenta ya activa, e.g. aumentar o disminuir cupos, siendo el modelamiento más representativo para este tipo de problemas el *behavioural scoring* (Rosenberg & Gleit, 1994, p. 589). Es en este segundo tipo de decisiones en las que se focaliza esta memoria.

3.3.1 Pronóstico de morosidad

Dado que la gestión de una cartera morosa requiere anticiparse a las posibles conductas de los clientes, el primer paso de cara a una correcta administración se basa en pronosticar el comportamiento futuro. Así, respecto a las técnicas empleadas para abordar el problema, es posible encontrar múltiples enfoques diferenciándose principalmente por la cantidad de periodos a pronosticar.

El considerar lo que ocurre en un único periodo futuro, modelamiento estático, es una estrategia que ha sido ampliamente estudiada. Para ello, se han propuesto diferentes modelos de clasificación que basados en la historia pasada de los clientes, presentan un pronóstico a partir de la extrapolación de un conjunto de ejemplos. En este contexto, el problema de identificar que clientes caerán en *default* se reduce a determinar con algún nivel de confianza, si existirá cesación de pagos en el periodo siguiente. Resultan comunes los modelos de árboles de decisión (Khandani, Kim, & Lo., 2014), así como análisis discriminante, además de regresión logística (Banasik, Crook, & Thomas, 1999). Entre las alternativas más recientes, se encuentra el uso de redes neuronales y máquinas de vectores de soporte (SVM), encontrándose casos en que estas son complementadas con en elementos de teoría de juegos y modelos no supervisados de *machine learning* (Figueroa, L'Huillier, & Weber, 2017). La Tabla 1 resume los enfoques mencionados.

Tabla 1: Modelamiento del comportamiento de morosidad, enfoque mono-periodo

Aplicación	Modelos empleados	Estrategia de Modelamiento
Base de comparación para modelo de supervivencia en morosidad (Banasik et al., 1999)	• Regresión Logística	Pronóstico de propensión al *default* o empeoramiento de situación de morosidad en un único periodo a futuro.
Pronóstico del paso a situación de mora tardía (>90 días) en un único periodo, tres meses de duración (Khandani et al., 2014) .	• Árboles de decisión (CART)	Pronóstico de la situación de morosidad en un periodo inmediatamente consecutivo, considerando potenciales interacciones no lineales en las variables explicativas.
Modelo de riesgo de *default*, con diferenciación de morosos por su "disposición a pagar" (Figueroa et al., 2017)	• *Support Vector Machine* • *k-means* con restricciones	Clasificación multiclase para el pronóstico del *default*, diferenciando entre dos tipos de *defaulters* mediante *clustering* semi-supervisado con fundamentos de teoría de juegos. Se diferencia entre clientes con problemas en su capacidad de pago respecto a aquellos sin disposición a hacerlo.

Nota. Fuente: Elaboración propia.

En la categoría de los pronósticos que abarcan más de un periodo se encuentran los denominados modelos dinámicos. Enfoques multinomiales de regresión logística (Ho Ha & Krishnan, 2012), así como modelos estadísticos de duración o supervivencia (Banasik et al., 1999; Ho Ha & Krishnan, 2012) son algunos de los ejemplos de este tipo de modelamiento. Por otra parte, se tienen otras aproximaciones con cadenas de Markov (Cyert, Davidson, & Thompson, 1962; Kuelen, Spronk, & Corcoran, 1981) las que basadas en procesos estocásticos, han llevado a aplicaciones patentadas como la de Shao et. al. (7,191,150 B1, 2013). La Tabla 2, presenta a modo general las aplicaciones y estrategias empleadas por los autores mencionados.

Tabla 2: Modelamiento del comportamiento de morosidad, enfoque multi-periodo

Aplicación	Modelos empleados	Estrategia de Modelamiento
Base de comparación para modelo de supervivencia, con lógica multi-periodo (Ho Ha & Krishnan, 2012)	• Regresión Logística multinomial	Considera que el valor que toma la variable dependiente nominal corresponde al número de meses en que es recuperada la deuda. La clasificación multiclase aborda el problema de pronosticar más de un periodo en el futuro.
Modelo de supervivencia del tiempo en mora (Banasik et al., 1999)	• Regresión de Cox	La duración es atribuida al tiempo en morosidad que experimentan los clientes. Es decir, la muerte del proceso corresponde al pago de la deuda.
Modelo de supervivencia del tiempo en mora, con heterogeneidad (Ho Ha & Krishnan, 2012)	• Regresión de Cox • *Self Organizing Maps*	La duración es atribuida al tiempo en morosidad que experimentan los clientes. Previamente se le segmenta para introducir heterogeneidad a los coeficientes de la regresión.
Gestión del proceso de cobranza basado en hitos como variables de estado (7,191,150 B1, 2013)	• Cadenas de Markov	El proceso de cobranza es modelado en estados que dan cuenta de las posibles migraciones del cliente en tramos de morosidad, renegociación, pago y término de la relación contractual.

Nota. Fuente: Elaboración propia.

Finalmente, respecto a antecedentes de otras memorias, si bien no se identifican trabajos que hagan uso de modelos como los presentados para este tema, si se considera que el problema de pronóstico tiene su símil en otros trabajos de título. Así, destacan enfoques de Markov en modelos de fuga , donde los estados corresponden a diversos niveles de la transformación RFM hasta que el cliente adquiere la definición de fugado (Roco Benavides, 2010; Segovia Riquelme, 2005), o la estimación del valor de vida del cliente (CLV) abordada con metodologías similares (Osses Godoy, 2015). En los trabajos presentados, la fuga puede compararse al alcance de la situación de castigo en el problema de morosidad, siendo este el motivo de que estas memorias sean consideradas.

3.4 Métodos de data mining y modelamiento

De las opciones discutidas como alternativas de modelamiento del comportamiento de morosidad (3.3.1 Pronóstico de morosidad), se revisaron aplicaciones de algoritmos de *machine learning* y procesos estocásticos al problema propuesto. De estas, destaca el uso de regresiones logísticas, árboles de decisión, modelos de supervivencia, enfoques

de *clustering* y cadenas de Markov. Todas estas estrategias son revisadas, con la salvedad de que las últimas dos metodologías son profundizadas en mayor detalle, dado su uso en este trabajo (para más información consultar ANEXO B).

3.4.1 Cadenas de Markov

Cadenas de Markov son un tipo de proceso estocástico, que permiten modelar mediante probabilidades la evolución de fenómenos en el tiempo. Estas son empleadas para explicar el comportamiento de un sistema a lo largo de un periodo, al identificar los estados y transiciones más probables, dado el momento en el que se sitúa el observador.

Una de las principales propiedades de este tipo de modelos corresponde a que el estado al cual evoluciona un sistema sólo depende de la situación actual en la que se encuentra y no de la historia pasada (propiedad markoviana o pérdida de memoria). Así una cadena de Markov tiene dos elementos constitutivos fundamentales:

Estados: corresponden a distintos niveles de un parámetro entre los cuales un sistema puede evolucionar. Para el caso que interesa en esta memoria, estos tienen relación con el número de periodos en mora que presenta el cliente al momento de observar el sistema. A modo de representación, estos estados suelen ser descritos como nodos en un grafo con todos los arcos que le son factibles. La Figura 8 presenta un ejemplo de la representación gráfica de una cadena de Markov que modela el comportamiento de morosidad en una patente norteamericana (7,191,150 B1, 2013)).

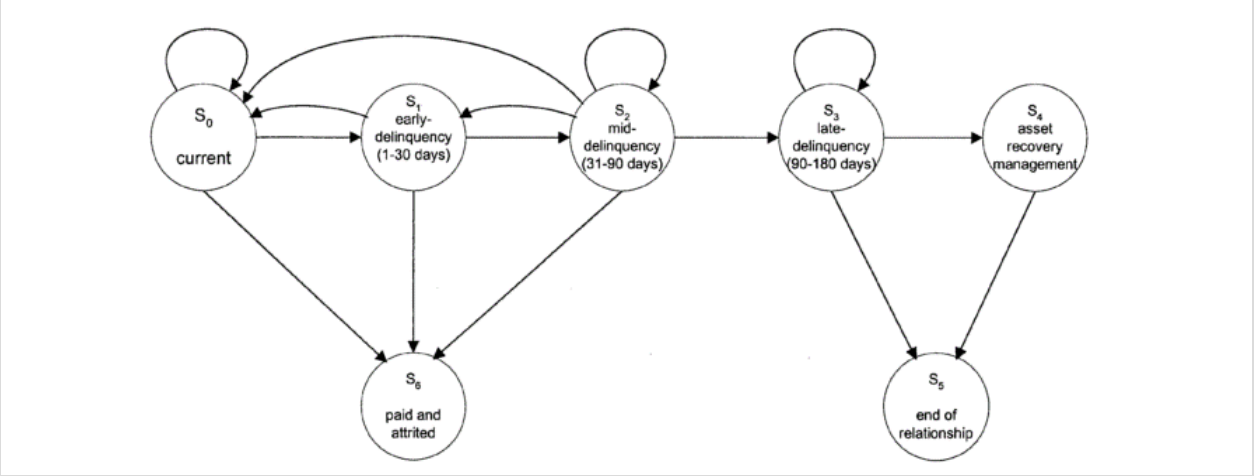


Figura 8. Modelamiento markoviano de la evolución de morosidad de clientes, para un proceso de cobranza.
Nota. Fuente: Min, S., Scott, Z., Cameron, G., Martin, R., Drossu, R., Zhang, J. (Guofeng), & Shoham, D. (2013). 7,191,150 B1. Estados Unidos. (7,191,150 B1, 2013, p. 5)

Probabilidades de transición: son el resultado directo del modelamiento mediante cadenas de Markov. Estas corresponden a la probabilidad con la que el sistema migra de un estado a otro. Por ejemplo, si se está describiendo la morosidad son las chances con las que un cliente que lleva un número de periodos sin pagar, en la etapa siguiente se

encuentre al día o aumente su permanencia en algún estado de incumplimiento. Las probabilidades descritas quedan representadas en una matriz denominada de transición, donde cada entrada (i, j) corresponde a la probabilidad de pasar del estado i al estado j.

$$\Pi = \begin{bmatrix} P_{11} & \cdots & P_{1m} \\ \vdots & P_{ij} & \vdots \\ P_{n1} & \cdots & P_{nm} \end{bmatrix} \quad (1)$$

Dentro de las propiedades básicas de este tipo de modelos, se destaca que la probabilidad de transición de un estado fijo a todas sus salidas posibles es 1 ($\sum_{j \in V} P_{ij} = 1$ propiedad de matriz estocástica). Además, para conocer las probabilidades de transición desde el momento en observación a un horizonte de n periodos, basta con multiplicar la matriz de transición Π n veces por si misma (P^n).

Finalmente, es a través de la definición de los estados correctos, la estimación de las probabilidades de transición y estacionarias que este tipo de modelos permite representar la evolución dinámica del sistema en estudio (Roco Benavides, 2010) y (Segovia, Aburto, & Goic, 2005).

3.4.1.1 Estimación de probabilidades de transición del sistema

Para efectos de desarrollar la estimación de las probabilidades de transición del sistema, al igual que lo hiciera Carlos Roco (Roco Benavides, 2010), se opta por la metodología presentada por los académicos del departamento Segovia, Aburto y Goic (Segovia et al., 2005) , que respondiera al trabajo de título de uno de ellos (Segovia Riquelme, 2005). Así, a partir de las transiciones empíricas registradas en cada periodo, se construiría una serie de matrices correspondientes a cada una de estas etapas, resumiendo el comportamiento entre los diferentes meses en estudio. Finalmente, con las matrices obtenidas, se procede a estimar las probabilidades de transición con una adaptación de la metodologías empleadas por Segovia (Segovia Riquelme, 2005, p. 43) y Roco (Roco Benavides, 2010, pp. 21–24); el método anterior ($P_{ij}(t-1, t) = T_{ij}(t-2, t-1)$ y de promedio ($P_{ij}(t-1, t)$=*prom*$\{T_{ij}(0,1), T_{ij}(1,2), \dots, T_{ij}(t-2, t-1)\}$).

Se define $P_{ij}^{(n)}(t-1, t)$ como la probabilidad de migrar desde el estado i, al estado j entre los periodos t y $t-1$, sujeto a la consideración de $(n \geq 1)$ periodos de historia pasada a través de las probabilidades empíricas de esa transición. Así se tiene que:

$$P_{ij}^{(n)}(t-1, t) = \frac{1}{n} \cdot \sum_{k=(t-1)-(n-1)}^{t-1} T_{ij}(k-1, k) \quad (2)$$

Dónde, $T_{ij}(k-1, k)$ corresponde a la probabilidad empírica de migrar desde el estado i al j, entre los periodos $k-1$ y k, i.e. la razón real entre los cambios de estado sobre el

total de transiciones ocurridas desde i en ese periodo. Nótese, que $P_{ij}^{(n)}(t-1,t)$ solo se encuentra definido cuando $t \geq n$. Además, para el caso en el que $n = 1$, las probabilidades de transición para el paso al periodo t son iguales a las chances empíricas de su par anterior $P_{ij}^{(1)}(t-1,t) = T_{ij}(t-2,t-1)$.

Finalmente, los diferentes métodos de ajuste quedan definidos por el valor que toma $P_{ij}^{(n)}(t-1,t)$, dado un valor de n determinado. Así, la Figura 9, muestra ejemplos de ajustes factibles para diferentes valores de n.

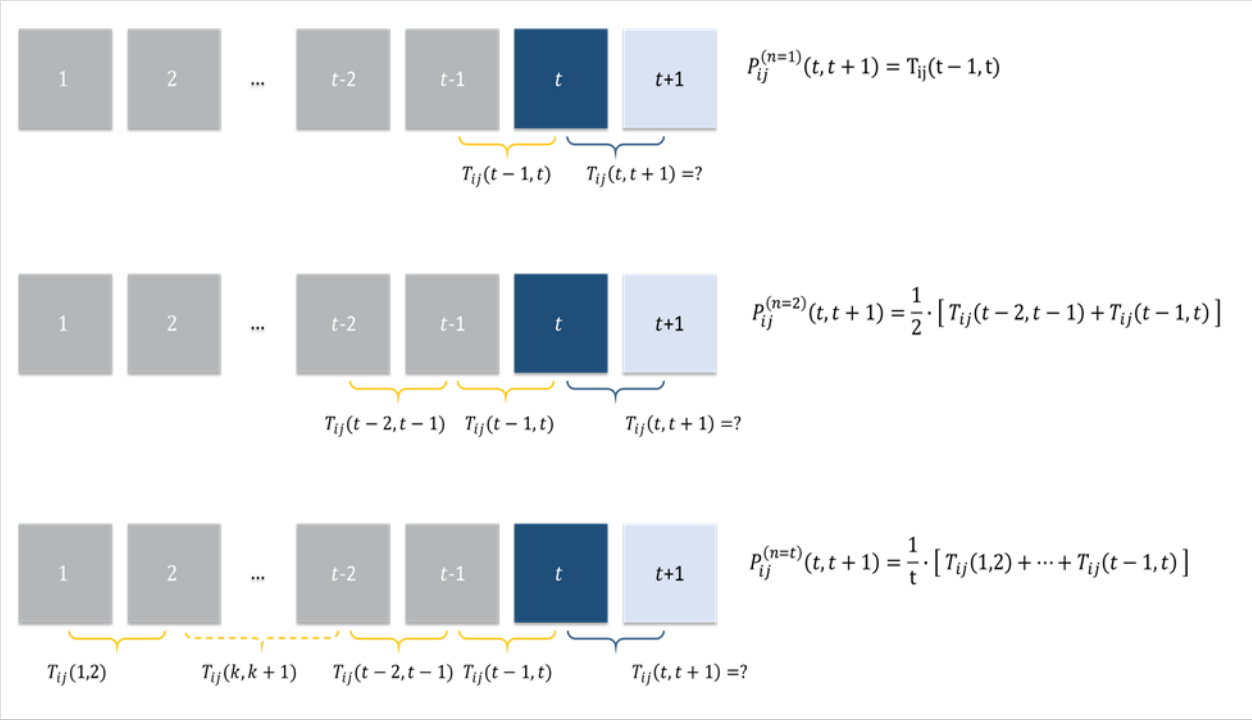


Figura 9. Diagrama de ajuste de probabilidades de transición $P_{ij}^{n}(t-1,t)$

Nota. Fuente: Elaboración propia.

3.4.2 Clustering

La tarea de clasificación de datos en forma no supervisada, i.e. en ausencia de registros etiquetados es conocida como *clustering*. En términos generales, tiene por objetivo la agrupación de un número finito de observaciones en una cantidad discreta de conjuntos similares. El fundamento radica en que los grupos generados respondan a una estructura subyacente en los datos que no es completamente conocida (Xu, 2005).

Dentro de los aspectos más importantes detrás de la aplicación de algoritmos de *clustering*, se encuentra la noción de similitud y distancia. Esta permite cuantificar el nivel de semejanza de las observaciones, permitiendo definir qué elementos deben pertenecer al mismo conjunto.

3.4.2.1 Medidas de Similitud

La distancia euclidiana es la medida de similitud más recurrente como estrategia de clustering (ver Tabla 3), no siendo este trabajo una excepción. Esta es a un caso particular de la distancia de Minkowski, también detallada en la Tabla 3. En ella se especifica sus formulaciones y algunas observaciones relacionadas a su aplicación.

Tabla 3: Medidas de similaridad y disimilitud para atributos numéricos

Medida	Formulación	Observaciones	Ejemplos
Minkowski	$D_{ij} = \left(\sum_{l=1}^{d} \left\|x_{il} - x_{jl}\right\|^{n}\right)^{\frac{1}{n}}$ (3)	Métrica invariante a traslaciones y rotaciones (si $n = 2$ es la distancia euclidiana). Atributos con valores grandes y o con mucha varianza tienden a dominar por sobre otros.	*Fuzzy c-means* basado en distancias de la familia Minkowski.
Euclidiana	$D_{ij} = \sqrt{\sum_{l=1}^{d} (x_{il} - x_{jl})^2}$ (4)	Es la métrica más utilizada, siendo a su vez un caso particular de la distancia de Minkowski con $n = 2$. Presenta una tendencia a formar clústeres esféricos en el espacio con el que se está trabajando.	*K-means*

Nota. Fuente: Adaptado de Xu, R. (2005). Survey of clustering algorithms for MANET. IEEE Transactions on Neural Networks, 16(3), 645–678. https://doi.org/10.1109/TNN.2005.845141. (2005, p. 648).

3.4.2.2 Tipos de Clusterización

Respecto a las diferentes aproximaciones al problema de clusterización, se pueden distinguir diversos enfoques. Los principales son resumidos en la Tabla 4, siendo *k-means* el más reconocible de los ellos. Además para el contexto de este trabajo, destacan los métodos particionales y difusos, por uso en la fase de modelamiento de esta memoria.

Tabla 4: Principales enfoques de clusterización

Tipo	Metodología	Ejemplos
Jerárquico	Construcción de una estructura de jerarquía entre los datos para la generación de grupos, la cual es determinada por una matriz de proximidad entre observaciones.	*Clustering aglomerativo* y divisivo
Particional	Asignación iterativa de las observaciones a un número prefijado de clústeres. La partición es obtenida al optimizar una función objetivo, siendo la más usual el error cuadrático. En vista de la imposibilidad de probar todas las combinaciones posibles, este tipo de algoritmos son implementados en base a heurísticas.	*k-means*, CLARA
Densidad	Este tipo de algoritmos se fundamenta en la idea de que un clúster corresponde a una alta densidad de observaciones, separada de otras agrupaciones por regiones de baja densidad.	DBSCAN
Fuzzy Clustering	Este tipo de métodos permite que una observación pertenezca a más de un grupo. Para ello se define un nivel de pertenencia para cada objeto en relación con cada clúster; optimizando una función objetivo que considera la partición difusa descrita.	*Fuzzy c-means* (FCM)

Nota. Fuente: Adaptado de Xu, R. (2005). Survey of clustering algorithms for MANET. IEEE Transactions on Neural Networks, 16(3), 645–678. https://doi.org/10.1109/TNN.2005.845141 , Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems, 17(2–3), 107–145. https://doi.org/10.1023/A:1012801612483 (2001, p. 113).

Adicionalmente a los tipos de *clustering* presentados en la Tabla 4, se pueden mencionar aquellos métodos basados en redes neuronales, como las redes de Kohonen y aquellos que utilizan un enfoques similares a de *Support Vector Machine* (SVM), que emplean funciones *kernel* para aprovechar espacios vectoriales de dimensiones más altas que el original para desarrollar el agrupamiento (Xu, 2005).

3.4.2.3 Clusterización de datos secuenciales

En relación con la tarea de *clusterización*, la estructura de la data juega un papel fundamental en la elección del algoritmo adecuado. De esta forma, la secuencialidad de las observaciones es un tipo de estructura que condiciona la manera en la que el *clustering* debe desarrollarse. De acuerdo con lo presentado por Xu (Xu, 2005), se pueden distinguir 3 enfoques: clusterización de la secuencia, clusterización indirecta de la sucesión y clusterización estadística del orden de los datos.

Respecto a la clusterización directa de la secuencia, esta contempla el agrupamiento de los datos en virtud del orden en el que se presentan las observaciones. Para ello, puede ser necesaria la elección de medidas de similitud diferentes y relacionadas con distancias de edición; como la distancia de Levenshtein. Este tipo de enfoques es principalmente aplicado a problemas de genética.

En segundo lugar se tiene el secuenciamiento indirecto para el desarrollo de la clusterización. Básicamente, se basa en la transformación de los datos originales, de manera que la estructura resultante de cuenta de la sucesión de los datos y al mismo tiempo permita aplicar las técnicas de *clustering* tradicionales sin mayores problemas.

Finalmente, la literatura describe un tercer enfoque, Clusterización estadística de la secuencia, que intenta a partir de la caracterización de las distribuciones de los datos y métodos estadísticos como HMMs (*Hidden Markov Models*) describir la dinámica de los grupos.

3.4.2.4 Validación de clústeres

Dado que el proceso de *clustering* corresponde a un método de clasificación no supervisado, la evaluación de los resultados obtenidos difiere de los modelos que si lo son. Esto se debe a la necesidad de determinar la validez de los mismos respecto al problema. De esta forma, se busca establecer que no se obtiene un resultado arbitrario como mera consecuencia de la convergencia del algoritmo utilizado. Así, uno de los problemas más relevantes es la evaluación en términos de ajuste a las estructuras subyacentes.

Se distinguen entre dos tipos de validación. En primer lugar, aquella basada en criterios internos, con el objetivo de evaluar los resultados en función de las cantidades, atributos inherentes al data set y al agrupamiento generado, denominándose validación interna

(Halkidi et al., 2001, p. 128). En contraste, la validación externa se basa en testear si acaso el data set se encuentra estructurado de manera aleatoria o si realmente hay un agrupamiento que subyace (Halkidi et al., 2001, p. 124).

Una de las principales funciones de la validación interna, constituye en la elección de algún hiper-parámetro inherente a la clusterización. Tal es el caso del número de grupos a ajustar cuando se trabaja con algoritmos particionales. Estos requieren de antemano el número de clústeres a construir, comúnmente denotado por k. Los resultados a obtener son altamente dependientes del mismo, por lo que existen diversos métodos y heurísticas que apuntan a obtener un valor adecuado para k. Entre las alternativas disponibles, se cuenta: la visualización de los datos proyectados en un espacio bidimensional euclidiano y la construcción de índices sujetos a distintos valores de k. En el primer caso, la técnica se encuentra limitada por la falta de explicabilidad y la pérdida de información asociada a la reducción de dimensiones (Xu, 2005, p. 664). Por este motivo, el método más empleado es el desarrollo de la segmentación con distintos valores para k, para luego elegir aquel valor que optimice un índice de validación interna previamente escogido.

3.4.2.4.1 Índices de validación interna

Entre los índices de validación interna, las nociones más elementales la constituyen el medir el nivel de semejanza entre las observaciones de un mismo clúster (cohesión), así como la diferencia entre las que pertenecen a grupos distintos (separación). A su vez, estos indicadores permiten definir métricas más complejas, de manera de ponderar de diferentes formas los niveles de cohesión y separación entre las observaciones. El presente capítulo revisa algunos de los índices más relevantes.

3.4.2.4.1.1 Within Group Sum of Squares (WGSS)

La suma de los cuadrados de las observaciones de un mismo grupo busca dar cuenta de la cohesión de los clústeres obtenidos. Se define como que tan cercanos se encuentran los registros de un segmento y matemáticamente se expresa de acuerdo con la Ecuación 5. Dónde M_i es el vector de atributos de la observación i perteneciente al clúster k-ésimo C_k, y $G^{\{k\}}$ es el baricentro del mismo grupo. Así, la Ecuación 5 se obtiene al sumar sobre todas las observaciones i pertenecientes al grupo k ($i \in I_k$).

$$WGSS^{\{k\}} = \sum_{i \in I_k} \left\| M_i - G^{\{k\}} \right\|^2 \tag{5}$$

Finalmente, se define el valor de la suma de la dispersión intragrupo (WGSS) de la partición efectuada de acuerdo a la Ecuación 6 (Desgraupes, 2013, p. 4).

$$WGSS = \sum_{k=1}^{K} WGSS^{\{k\}} \tag{6}$$

3.4.2.4.1.2 Between Group Sum of Squares (BGSS)

La suma de cuadrados entre grupos es una medida de la dispersión de los clústeres obtenidos, i.e. que tan alejados se encuentran entre si los diferentes segmentos. Para efectos de cálculo se define la medida como la dispersión de los baricentros $G^{\{k\}}$ de cada clúster, respecto al baricentro de toda la data G. Donde n_k es la cantidad de observaciones que pertenecen al clúster k-ésimo. Así la formulación queda expresada por la Ecuación 7 (Desgraupes, 2013, p. 6).

$$BGSS = \sum_{k=1}^{K} n_k \left\| G^{\{k\}} - G \right\|^2 \tag{7}$$

3.4.2.4.1.3 Coeficiente de Silhouette

Para cada observación i, se define $a(i)$ como la distancia intra-clúster media de dicha observación, respecto a las otras del mismo segmento como indica la Ecuación 8. Dónde M_i es el vector de características de dicha observación y $d(M_i, M_{i'})$ la distancia entre los vectores de los elementos i e i'. Finalmente, n_k denota la cantidad de ítems pertenecientes al clúster k-ésimo (C_k) y cuyo conjunto de observaciones se expresado por I_k.

$$a(i) = \frac{1}{n_k - 1} \sum_{\substack{i,i' \in I_k \\ i' \neq i}} d(M_i, M_{i'}) \tag{8}$$

En forma seguida, se define la distancia promedio de la observación i a los elementos pertenecientes al grupo k' de acuerdo con la Ecuación 9.

$$ð(M_i, C_{k'}) = \frac{1}{n_{k'}} \sum_{i' \in I_{k'}} d(M_i, M_{i'}) \tag{9}$$

Se define $b(i)$ como la mínima de las distancias promedios a los elementos de otros clústeres, de acuerdo con la Ecuación 10. De tal forma $b(i)$ representa la distancia promedio al clúster que constituye la mejor opción de reasignar la observación i, si esta no perteneciese al clúster C_k.

$$b(i) = \min_{k' \neq k} ð(M_i, C_{k'}) \tag{10}$$

Sigue que, para cada observación i es posible definir el cociente formulado por la Ecuación 11. De manera que, $s(i)$ es denominado coeficiente de Silhoutte. Este comprende valores ente -1 y 1. Los valores cercanos a 1 indican que el punto i se encuentra asignado al clúster correcto pues la segunda mejor opción se encuentra muy

lejana. Por su parte, valores cercanos a (-1) son indicador de que la partición podría no ser la mejor.

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \tag{11}$$

Así, s_k es definido como el promedio de los $s(i)$ de los elementos pertenecientes a un clúster C_k según la Ecuación 12, denominándose Silhouette promedio del clúster.

$$s_k = \frac{1}{n_k} \sum_{i \in I_k} s(i) \tag{12}$$

Finalmente, la medida global de los Silhouette promedio de todos los clúster corresponde a lo que se conoce como coeficiente de Silhouette, indicado por la Ecuación 13 (Desgraupes, 2013, p. 18).

$$\mathrm{C} = \frac{1}{K} \sum_{k=1}^{K} s_k \tag{13}$$

3.4.2.4.1.4 Coeficiente de Caliñski – Harabasz

Al igual que Silhouette, el índice de Caliñski-Harabasz, tiene por fundamento ponderar el nivel de cohesión entre elementos de un mismo clúster, así como la separación o dispersión entre los elementos de distintos grupos. Esta métrica es definida por el número de observaciones de cada clústeres, y los ya definidos indicadores BGSS y WGSS como indica la Ecuación 14 (Desgraupes, 2013, p. 9).

$$\mathrm{C} = \frac{N - K}{K - 1} \cdot \frac{BGSS}{WGSS} \tag{14}$$

3.4.2.5 Algoritmos de Clustering

Dentro de los enfoques de clusterización presentados en la Tabla 4, se mencionaron dos algoritmos que son relevantes para este trabajo. A continuación, ambos son revisados.

3.4.2.5.1 K-means

Método de *clustering* particional basado en una función de error que se minimiza a medida que cada observación se asigna de mejor forma, para un número predefinido de clústeres, usualmente denotado por k. El algoritmo finaliza en función de un criterio de convergencia sobre la función de error (Ecuación 15). Esta a su vez es determinada por la distancia entre el vector de características de cada ítem y el centro del clúster al cual

se encuentra asignado (centroide). Así, la partición óptima considera la minimización de las observaciones respecto al centro del segmento al que se atribuye (Xu, 2005, p. 642).

$$Min\ Error = Min \sum_{i=1}^{K} \sum_{j=1}^{N} \gamma_{ij} \left\| x_j - m_i \right\|^2 \quad (15)$$

Dónde,

$\Gamma =$ Matriz de partición tal que $\gamma_{ij} = \begin{cases} 1, & si\ x \in C_i \\ 0, & si\ no \end{cases}$, con $\sum_{i=1}^{K} \gamma_{ij} = 1\ \forall j$.

$M =$ Matriz de prototipo de los clústeres (centroides), tal que $m_i =$ es la media muestral del clúster i-ésimo definido por $m_i = \frac{1}{N} \sum_{j=1}^{N} \gamma_{ij} x_j$ siendo N_i el número de objetos en el clúster i.

Finalmente, el algoritmo en pseudo-código puede resumirse de la manera siguiente:

1. Inicializar una k-partición (k grupos) de forma aleatoria o basado en conocimiento previo. Calcular la matriz prototipo de la k-partición, $M = [m_1, \dots m_i, \dots, m_k]$.
2. Asignar cada observación al clúster C_w cuyo centroide sea el más cercano, i.e. $x_j \in C_w$ si $\left\| x_j - m_w \right\| < \left\| x_j - m_i \right\|$, para todo $j = 1 \dots N$, $i = 1, \dots, N$ e $i \neq w$.
3. Recalcular la matriz prototipo basada en la nueva partición definida en 2.
4. Repetir pasos 2.-3. hasta que no existan cambios en la partición.

3.4.2.5.2 Fuzzy c-means

Este algoritmo de *clustering*, denominado de lógica difusa, resulta de la relajación de la restricción de pertenencia exclusiva a un clúster, i.e. cada observación puede pertenecer a más de un grupo. Así, sea $x_j \in \Re^d$, $j = 1 \dots N$, donde x_j corresponden a las N observaciones a agrupar, entonces se una función de error de acuerdo con la Ecuación 16:

$$Min\ J(U, M) = \sum_{i=1}^{C} \sum_{j=1}^{N} \left(u_{i,j}\right)^m D_{i,j} \quad (16)$$

Dónde,

$U = \left[u_{i,j}\right]_{c \times N}$ Matriz de partición difusa, i.e. grados de pertenencia de las observaciones a los clústeres. Con $u_{ij} \in [0,1]$ el coeficiente de pertenecia el objeto j al clúster i.

$M = [m_1, \dots, m_c]$	Matriz prototipo (centros o medias) de la partición.
$m \in [1, \infty)$	Parámetro de fuzzificación, usualmente fijado en 2 y que determina el nivel de heterogeneidad en las pertenencias a los clústeres. Si $m \to 1$ la partición tiende a ser exclusiva, mientras que si $m \to \infty$, la matriz de partición desarrolla pertenencias del tipo $u_{ij} \to \frac{1}{c}\ \forall i, j$.
$D_{ij} = D(x_j, m_i)$	Medida de distancia entre la observación j y su vector de características x_j y el prototipo o centro del clúster i, denominado m_i.

Finalmente, el algoritmo puede ser resumido en pseudo-código de la manera siguiente:

1. Seleccionar valores apropiados para m, c, además de una constante positiva muy pequeña $0 < \epsilon \ll 1$. Con estos parámetros se inicializa aleatoriamente la matriz de prototipos M y se fija la variable $t = 0$.
2. Calcular (si $t = 0$) o actualizar (si $t > 0$) la matriz de pertenencia U, según la Ecuación 17.

$$u_{ij}^{(t+1)} = \left(\sum_{l=1}^{c} \left(\frac{D_{lj}}{D_{ij}} \right)^{\frac{1}{1-m}} \right)^{-1} \tag{17}$$

   Para $i = 1, \dots, c$ y $j = 1, \dots, N$
3. Actualizar la matriz de prototipos M, de acuerdo con la Ecuación 18.

$$m_i^{(t+1)} = \left(\sum_{j=1}^{N} \left(u_{ij}^{(t+1)} \right)^m x_j \right) \cdot \left(\sum_{j=1}^{N} \left(u_{ij}^{t+1} \right)^m \right)^{-1} \tag{18}$$

Para $i = 1, \dots, c$

4. Repetir pasos 2.-3. hasta que $\left\| M^{(t+1)} - M^{(t)} \right\| < \epsilon$.

3.5 Métodos para el preprocesamiento de datos

Esta sección aborda aquellos procedimientos y herramientas para preparar los datos, en vista de su posterior modelamiento. Para ello se tienen en cuenta diversas problemáticas que la información pude traer desde su origen.

3.5.1 Valores faltantes

Los valores faltantes (*missing values*) hacen alusión a registros en los que una o más variables no tienen un valor asignado, i.e. tienen valor nulo. Dada la naturaleza de incertidumbre acerca del dato perdido, el no considerarlos como parte del análisis no es

recomendable si no se han tomado precauciones. Esto se debe a que su omisión como parte de la fuente de datos podría responder al mismo fenómeno en estudio. Así, De acuerdo con lo expresado por el autor británico Carpenter y Goldstein (Carpenter & Goldstein, 2009), se pueden identificar tres tipos de *missing values*.

- **MCAR** (*missing completely at random*): el valor faltante se explica en un mecanismo cuyo origen es aleatorio, i.e. es independiente de las variables observadas y de parámetros omitidos. Existen test estadísticos basados en χ^2 para comprobar la naturaleza aleatoria de la ausencia de valores.
- **MAR** (*missing at random*): el valor faltante no depende de los parámetros omitidos, sino que, de las variables observables, por lo que puede ser reconstruido.
- **NMAR** (*not missing at random*): el valor faltante se explica en la existencia de parámetros no observados, correspondiendo al caso contrario al antes señalado.

Dependiendo de tipo de *missing values*, la estrategia para abordar el problema puede diferir, siendo el caso NMAR es el más crítico. Esto se debe a la imposibilidad de asegurar si la ausencia de valores tiene o no relación con el fenómeno, situación que podría introducir sesgos al análisis. Por este motivo, a continuación se presentan tres formas distintas de tratar el problema de *missing values* (Weber, 2017).

3.5.1.1 Eliminación de registros

La eliminación de registros consiste en la exclusión de aquellas filas que presenten al menos un valor nulo en alguno de sus atributos. Una de sus desventajas consiste en que para bases de alta dimensionalidad, la cantidad de casos positivos (sin *missing values*) podría ser muy escaza, reduciendo notoriamente la cantidad de datos con los que se podría trabajar. Por otra parte, si los registros que se eliminan difieren sustancialmente de aquellos que no tienen valores faltantes, es muy probable que se introduzca un sesgo.

3.5.1.2 Eliminación de atributos

La eliminación de atributos con valores nulos corresponde a no considerar aquellas variables independientes que presentan al menos, un valor faltante. Este tipo de técnicas, por lo general, es empleada solo cuando la proporción de nulos es muy grande y la variable no es capaz de aportar información.

3.5.1.3 Imputación de registros

La imputación de registros es un enfoque que a diferencia de los descritos, permite retener y no descartar por completo los datos con problemas. Así, se puede disminuir la perdida de información y potenciales sesgos de la eliminación. Dentro de las desventajas de este tipo de métodos se tiene la alteración que la imputación genera sobre la distribución de los datos. Entre las alternativas de imputación se encuentra: el reemplazo de nulos por la media o mediana de la variable, emplear una respuesta observada de una

unidad similar (*Hot deck*), la creación de una variable que indica la ausencia de valores o el estimar el valor del atributo faltante, en base a otros campos que si son conocidos.

3.5.2 *Outliers*

De acuerdo a Barnett y Lewis (Barnett & Lewis, 1994) un *outlier* se entiende como: "*An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.*".[(Una observación (o subconjunto de observaciones) que parecen ser inconsistentes con el conjunto remanente de datos]. Entre los efectos que estos valores atípicos pueden tener se cuenta: la falla en la especificación de un modelo, el sesgo de un estimador asociado a la variable y conclusiones erradas a causa de la misma situación (Ben-gal, 2005).

3.5.2.1 Técnicas para identificación de valores fuera de rango

Existen numerosas técnicas para la identificación de *outliers*, las cuales en general difieren en el nivel de compresión que se tiene sobre la data y las presunciones que se hacen acerca de ella, por ejemplo, su distribución. De acuerdo a lo propuesto por Ben-gal (Ben-gal, 2005) estas técnicas pueden clasificarse en:

- **Métodos Univariados:** los valores se asumen independientemente distribuidos (i.i.d.), y en muchos casos también se hace lo mismo con la distribución de los parámetros. Finalmente, el problema se reduce a encontrar una región de valores atípicos en que la variable no interactúa con otros atributos.
- **Métodos Multivariados:** Corresponden a aquellos que toman en cuenta la interacción de la variable en estudio con otros atributos.
- **Métodos paramétricos o estadísticos:** Asumen la distribución de la variable en estudio y o se basan en la estimación de los parámetros desconocidos de la misma.
- **Métodos no paramétricos o basados en distancia:** Tienen su fundamento en las nociones de distancia entre elementos, por lo que aplican estas medidas para determinar que valores se alejan del comportamiento común.

A continuación, se presenta una técnica del tipo multivariado (Prabhakaran, 2016) que es empleada en este trabajo. Otras alternativas pueden consultarse en ANEXO C.

3.5.2.1.1 Distancia Mahalanobis

La distancia de Mahalanobis toma en cuenta la correlación existente entre las variables al considerar la covarianza. Así, se obtiene un resultado que se basa en la distancia de cada punto a la media de la data, al mismo tiempo que toma en cuenta la correlación entre variables (Rosenmai, 2013; Ruefer, 2016).

En términos matemáticos la distancia de Mahalanobis queda formulada por la expresión: $D(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}$. Donde $\vec{x}$ e $\vec{y}$ son vectores aleatorios de una misma distribución y S la matriz de covarianza. No obstante, para su uso como método de detección de *outliers*, se asume la normalidad de las variables (Guyon & Elisseeff, 2003). Así, el cálculo queda definido como la distancia multivariada de la data, respecto a su media como indica la Ecuación 19 (Orlov, 2011):

$$D_i = \sqrt{(R_i - \mu)S^{-1}(R_i - \mu)} \tag{19}$$

Dónde R_i es el vector de la observación i, μ el vector de medias y S la matriz de covarianzas. Finalmente, una vez determinadas las distancias respecto al centro se establece un corte con el que se determinan los registros que serán considerados como *outliers* (Eidgenössische Technische Hochschule Zürich, 2012).

3.5.3 Transformación de variables

La transformación de variables permite preparar los datos para la utilización de distintas pruebas estadísticas, como test de medias, análisis de correlación y otras, permitiendo por ejemplo que se cumplan sus supuestos. Por otra parte, algunas de estas construcciones pueden llegar a incorporar más información que su valor original, aportando a la formulación de mejores modelos. Sin embargo, no debe obviarse que la transformación no siempre será imperativa y podría ser apropiado usar los valores en bruto. Así, algunas de las trasformaciones más comunes son la estandarización y normalización (ver ANEXO D).

3.6 Métodos para la selección de variables

La extracción de atributos es relevante a la hora de determinar el subconjunto final de datos a emplear en el modelamiento. A continuación, se describe brevemente algunas de las técnicas empleadas en este trabajo.

3.6.1 Análisis de correlación

El análisis de correlación está compuesto por diversas técnicas usadas para determinar la relación entre dos o más variables. Entre los métodos más clásicos se encuentra: la matriz de correlaciones y diferentes pruebas estadísticas.

3.6.1.1 Análisis de componentes principales

El Análisis de Componentes Principales (ACP) es una técnica que permite reducir la dimensión de un conjunto de datos a un espacio vectorial de menor cardinalidad. Dichos factores responden a una combinación lineal de las variables originales, por lo que se fundamenta en el cálculo de vectores y valores propios. Esta metodología responde de mejor manera cuando existen altas correlaciones entre los atributos, pues es un indicador

de la existencia de información redundante. Así, el uso de unos pocos factores permite en el mejor de los casos explicar gran parte de la variabilidad de los datos.

3.6.1.2 Tablas de contingencia y Test CHI-2

Las tablas de contingencia son una representación que permite comprobar la independencia entre dos variables categóricas. Para ello se organizan las variables en estudio de manera tabular y se verifica la existencia de diferencias entre los niveles en que se presentan los atributos. Así, se comparan los valores observados y esperados, bajo la hipótesis que si son independientes, no se observarán diferencias significativas al testar mediante una prueba χ^2 (CHI-2). Esta prueba queda resumida en la hipótesis nula H_0 que indica que no existen diferencias significativas entre las frecuencias observadas y esperadas. Y la hipótesis alternativa H_1 que refiere a la existencia de diferencias significativas entre las frecuencias.

En la tabla de contingencia, la frecuencia esperada se define por la expresión $f_e = \frac{(f_r * f_k)}{\mathrm{n}}$, donde $\mathrm{f_r}$ corresponde a la frecuencia total de la fila r de la tabla de contingencia y $\mathrm{f_k}$ a la frecuencia total de la columna k. Luego, se calcula el estadístico Chi-2, según $\chi^2 = \sum_i^n \frac{(f_o - f_e)^2}{f_e}$. Siendo f_o y f_e la frecuencia observada esperada respectivamente. Sigue que, el obtenido se compara con la magnitud teórica de la distribución χ^2 con $d.f. = (r-1) * (k-1)$ grados de libertad y un nivel de significancia arbitrario α. Finalmente, si el valor de χ^2 obtenido es igual o mayor al valor crítico, se rechaza la hipótesis nula, asumiéndose frecuencias significativamente diferentes y con ello la dependencia de las variables analizadas.

3.7 Metodologías de validación de modelos predictivos

La partición de los datos en conjuntos de entrenamiento y test permite disminuir los riesgos de *overfitting*, favorecer el ajuste de modelos generalizables (validez fuera de la muestra) y construir métricas de performance que representan de mejor manera la calidad de estos. Existen diferentes metodologías para determinar cómo separar las observaciones, siendo la más conocida la validación simple o *hold-out*. Otras alternativas como *cross-validation*, *random subsampling* y *Bootstrap* son detalladas en el ANEXO E.

3.7.1 Hold-out

Este método consiste en la partición de la data original en tres partes, por lo general, conocidas como *training*, *validation* y *test set*. El primer conjunto corresponde al grupo de datos con los cuales es entrenado el modelo, i.e. se ajustan los parámetros que son dependientes de la data. Por ejemplo, en el caso de árboles de decisión, estos son los datos que definirán las ramificaciones de este. A su vez, el set de validación cumple con dos objetivos: proporcionar un conjunto de datos sobre el cual calcular las métricas de desempeño y seleccionar un modelo, además de permitir optimizar hiper-parámetros asociados. Finalmente, una vez terminado el ajuste se procede a comprobar su

desempeño general según una métrica seleccionada, pronosticando sobre el *test set* (Souza, Matwin, & Japkowicz, 2002, p. 11).

3.8 Metodologías de evaluación de modelos predictivos

Como se revisará en el marco metodológico, esta memora considera dos tipos de pronósticos, uno a nivel de grupo y otro a nivel de cliente. Las métricas asociadas a cada uno de ellos son revisadas a continuación.

3.8.1 Evaluación a nivel de grupo

Para evaluar el ajuste a nivel grupal, se presentan tres métricas basadas en las diferencias entra las transiciones registradas y las pronosticadas (Roco Benavides, 2010, p. 49; Segovia Riquelme, 2005, p. 46). Estas medidas de evaluación son presentadas en la Tabla 5.

Tabla 5: Comparativa de métricas de evaluación a nivel de grupo

Métrica	Formulación	
Error Absoluto	$\frac{1}{N° \, Transiciones} \cdot \sum_i \sum_j \left\lvert N^*_{ij}(t, t+\Delta t) - N_{ij}(t, t+\Delta t) \right\rvert$	(20)
Error Absoluto ponderado por número de transiciones	$\frac{1}{N° \, Transiciones} \cdot \sum_i \sum_j \frac{\left\lvert N^*_{ij}(t, t+\Delta t) - N_{ij}(t, t+\Delta t) \right\rvert}{N_{ij}(t, t+\Delta t)}$	(21)
Error absoluto ponderado por número de transiciones simétrico	$\frac{1}{N° \, Transiciones} \cdot \sum_i \sum_j \frac{\left\lvert N^*_{ij}(t, t+\Delta t) - N_{ij}(t, t+\Delta t) \right\rvert}{\frac{1}{2} \cdot \left\lvert N_{ij}(t, t+\Delta t) \right\rvert + \left\lvert N^*_{ij}(t, t+\Delta t) \right\rvert}$	(22)

Nota. Fuente: Elaboracion propia.

Nótese que la primera de las métricas; el error absoluto (Ecuación 20), pese a ser la aproximación natural al evaluar el número de transiciones incorrectamente pronosticadas, tiene la problemática de no permitir comparar entre modelos de Markov con diferente número de estados. A su vez, el error absoluto ponderado (Ecuación 21) no es capaz de hacerse cargo de pronósticos en el que el número de transiciones observadas es nulo de manera que no puede asignarse un valor al error en dichos casos. Finalmente, el error absoluto ponderado por número de transiciones simétrico considera en su definición el valor del pronóstico, aun cuando el observado es nulo, de forma tal que se hace cargo de los diferentes errores que pueden ocurrir dentro de la estimación (Ecuación 22).

3.8.2 Evaluación a nivel de cliente

Para evaluar el ajuste a nivel de cliente, el problema es reducido a una clasificación supervisada, de forma tal que las métricas tradicionales pueden aplicarse directamente como metodología de evaluación, una vez introducidas las modificaciones requeridas para un problema multiclase.

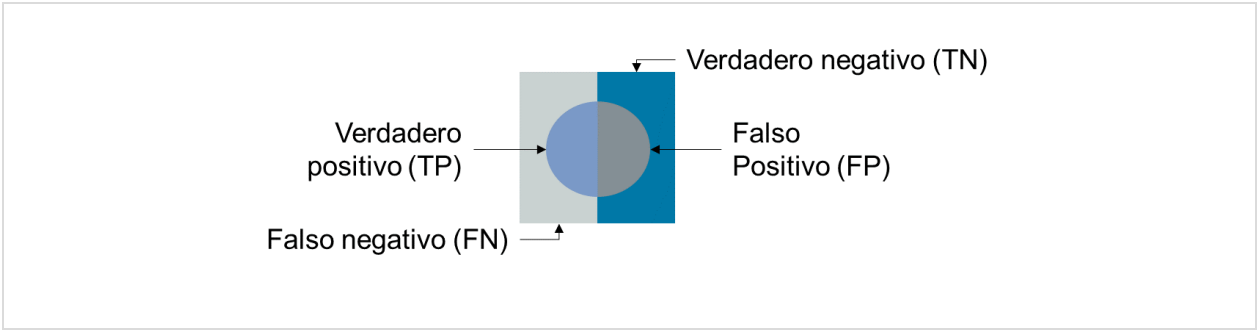


Figura 10. Posibles resultados de una clasificación binaria
Nota. Fuente: Elaboración propia.

La Tabla 6 presenta la definición de las métricas aludidas, cuyo input es clarificado a partir de los diferentes tipos de error que pueden ser cometidos en una clasificación, tal y como lo indica la Figura 10. Nótese que, dado que el problema de pronosticar el estado de morosidad de un cliente corresponde a un tipo de clasificación multiclase, el procedimiento requiere de calcular individualmente el indicador para cada una de ellas. Luego el promedio es reportado como resultado general del sistema.

Tabla 6: Comparativa de métricas de evaluación a nivel de cliente

Métrica	Formulación	
Accuracy	$= \frac{TP + TN}{TP + TN + FP + FN} =$ 	Medida que muestra el grado de exactitud de la predicción. Se calcula como la simple razón entre el número de predicciones correctas y el total de predicciones.
Precision	$= \frac{TP}{TP + FP} =$ 	Medida de rendimiento, obtenida como la relación entre las observaciones positivas predichas correctamente y el total de observaciones positivas predichas. De este modo, altos valores se relacionan con una baja tasa de falsos positivos.
Recall	$= \frac{TP}{TP + FN} =$ 	Corresponde a la ratio de observaciones positivas correctamente predichas respecto al total de observaciones positivas. De este modo, altos valores se relacionan con una baja tasa de falsos negativos.
F1-score	$= 2\frac{(Recall * Precision)}{(Recall + Precision)}$	Métrica que combina las medidas de *recall* y *precision*, tomando en cuenta tanto los falsos positivos, como negativos. A diferencia del *accuracy*, el *F1-score* presenta un mejor desempeño cuando se tienen datos con una distribución desigual de clases.

Nota. Fuente: Elaboración propia.

4 Marco metodológico

El presente capítulo, muestra como los elementos presentados en el marco teórico serán aplicados a este proyecto. Así, tomando como punto de partida la metodología CRISP-DM, se procede a estructurar el trabajo en cada una de las etapas contempladas por esta. De este modo, cada fase es evidenciada de manera concreta en los pasos a seguir a lo largo de esta memoria (ver Figura 11). Nótese que la profundización en cada una de las etapas es abordada en el capítulo 5 Desarrollo metodológico.

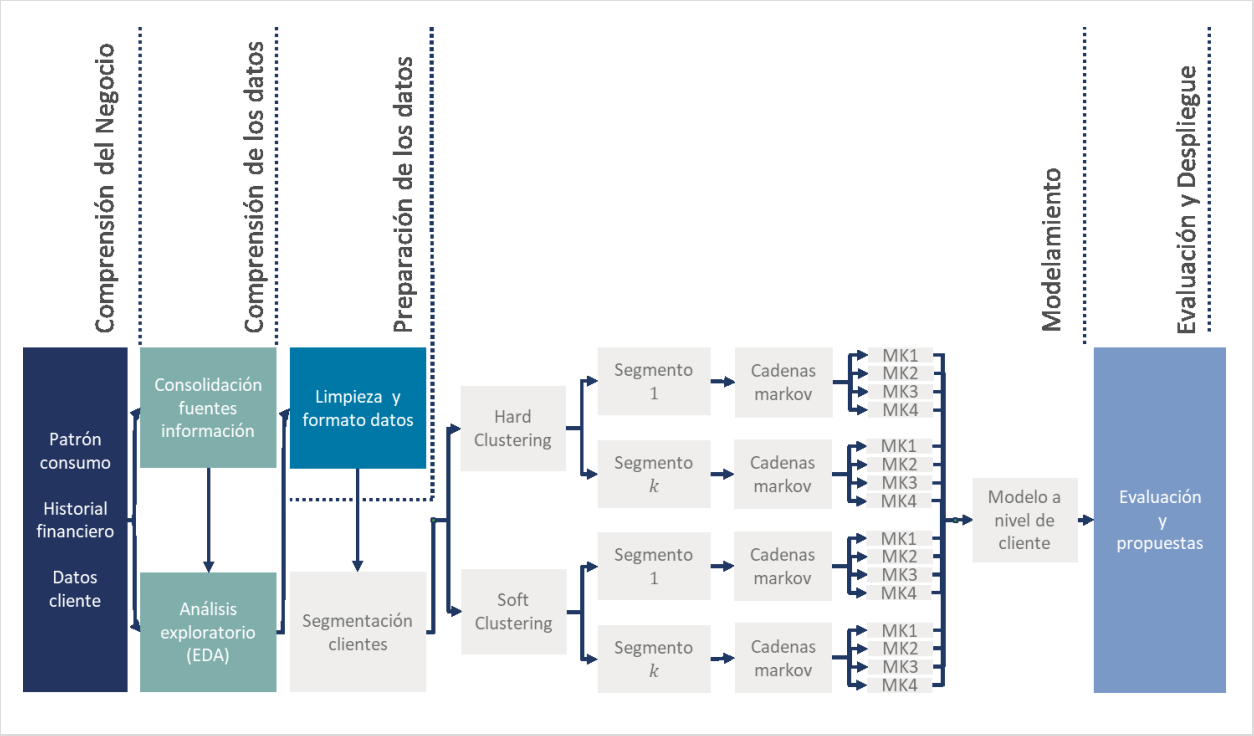


Figura 11. Diagrama de la aplicación de metodología CRISP-DM al trabajo de título.
Nota. Fuente: Elaboración propia.

4.1 Comprensión del negocio

Se comienza por profundizar en los procesos del negocio relacionados con la facturación y cobranza. Así se identifican las fuentes de datos apropiadas; aquellas que dan cuenta del proceso mencionado y del comportamiento de morosidad de los clientes en interacción con este. Adicionalmente se busca identificar reglas de negocio que puedan afectar la generación y construcción de la data que será analizada. Finalmente se cimenta una visión acabada del problema del aumento de los niveles de morosidad, permitiendo el entendimiento cabal del mismo.

4.2 Comprensión de los datos

Como pate del proceso de comprensión de los datos, una vez identificada la lógica de negocio detrás de la facturación y cobranza de la tarjeta, se comienza por la individualización de las fuentes de datos relevantes al problema, i.e. tablas de facturación,

de pagos de compras y fuentes de información sobre los clientes. En forma seguida se procede con el análisis descriptivo y su relación con los tramos de morosidad. En este punto son evaluadas de manera preliminar, algunas hipótesis que se tienen sobre las variables que impactan la morosidad de clientes. Entre estas se encuentran: medidas de fidelización, antigüedad de la cuenta, renta declarada, comportamiento transaccional y rubros. Además, se considera en este análisis; otras variables del negocio como: el tipo de tarjeta, categoría de cliente y atributos sociodemográficos.

4.3 Preparación de datos

En esta etapa, se tiene en consideración la identificación y tratamiento de valores perdidos y o atípicos. Para ello se comienza por el análisis de *missing values* y la determinación de la naturaleza de su origen, en cuanto ausencia de estos datos. Para ello se emplea la prueba MCAR descrita en el marco teórico de este trabajo de manera de tener un fundamento estadístico sobre el tipo de valores perdidos con los que se está tratando.

En forma seguida, se procede a la identificación de potenciales *outliers* haciendo uso de la distancia de Mahalanobis. Dependiendo de la cantidad de este tipo de datos se opta por la imputación y o eliminación de estos registros.

Para efectos de selección de variables, se utiliza análisis de correlación para determinar información redundante y establecer un primer corte sobre la base. En forma seguida se aplican dos pruebas estadísticas; CHI-2 y Kolmogorov-Smirnov, como primeros filtros en la selección de variables. Además, se contempla el uso de árboles de decisión como herramienta para la identificación de variables relevantes en la descripción del comportamiento de morosidad de los clientes.

Por transformaciones, a priori se explicita la necesidad de contar con un *lag* sobre las variables transaccionales que dependen de la situación contractual del cliente. Así, se capturan los impactos que cambios en el comportamiento pasado pueden tener al momento de la segmentación. De esta forma, se evita que los resultados se acoten a las reglas de negocio aplicadas cuando los clientes alcanzan un determinado tramo de morosidad.

Finalmente, las construcciones de variables más importantes consideran la definición de los tramos de morosidad de cliente, para cada combinación $(contrato, año, mes)$. Atributos como el descrito son los que permiten definir la transición entre los diferentes estados de morosidad en los pasos del modelamiento.

4.4 Modelamiento

Para efectos de modelamiento, el primer paso es la segmentación de clientes de acuerdo con las variables seleccionadas en la fase anterior. Con la conformación de estos grupos,

se procede al ajuste del modelo de cadenas de Markov para explicar las transiciones de mora en cada clúster.

Previo al ajuste del modelo de Markov sobre cada grupo, en términos de poder evaluar el desempeño y evitar sesgos metodológicos, la data es particionada de acuerdo a la técnica *hold-out*. De esta forma, cada segmento cuenta con un conjunto de entrenamiento y *testing*, con el cual efectuar el ajuste del modelo y contrastar sus resultados en el paso posterior.

Nótese que, la metodología propuesta para este trabajo contempla la evaluación de dos formas de segmentación; *hard* y *soft clustering*. Estos dos enfoques permiten aproximarse de dos maneras distintas al problema, desde una perspectiva de heterogeneidad en los comportamientos. Así, los resultados del *hard clustering* se encuentran orientados a caracterizar comportamientos a nivel de grupo, mientras que los correspondientes a *soft clustering* a generar conclusiones a nivel de cliente.

4.5 Evaluación

Para efectos de evaluación, se considerará la calidad del pronóstico en cada grupo, de forma de comprobar que nivel de precisión se tiene al pronosticar las transiciones de los clientes, siendo esta la principal métrica para evaluar el desempeño a nivel de segmento. Posteriormente, para los resultados a nivel de cliente la validación de la calidad del pronóstico se determina de acuerdo con las métricas clásicas para tareas de clasificación, considerando las modificaciones necesarias para tratar con problemas multiclase.

4.6 Despliegue

En términos de despliegue, los alcances de esta memoria no consideran la creación de un proceso de negocio a partir de los resultados obtenidos. No obstante, sí es un objetivo el efectuar recomendaciones a partir de ellos. De esta manera se proponen acciones que propenden a la disminución de la cartera morosa. Al mismo tiempo, se sensibilizan distintos escenarios en términos de beneficios para establecer el valor económico del trabajo.

5 Desarrollo metodológico

El presente capítulo, desarrolla la aplicación de la metodología de este trabajo y sus resultados. Para ello, se estructura el capítulo de acuerdo con las diferentes fases del modelo CRISP-DM (ver 3.2 Metodología CRISP-DM).

5.1 Comprensión del negocio

El problema de negocio se desarrolla en torno a las reglas que definen el proceso de facturación y cobranza de las tarjetas de crédito. Esta sucesión de operaciones considera el cómo los cargos y pagos del cliente se convierten en el monto a cancelar, pago mínimo del mes, y determinan en qué casos se produce morosidad. Además, no debe obviarse la posibilidad de que los montos cancelados pueden saldar la deuda en su totalidad, producir *revolving*[8] u ocasionar el incumplimiento de las obligaciones pactadas.

La facturación de un periodo es explicitada al cliente en el estado de cuenta mensual ($EECC_t$), en el cual se le indica cuánto debe pagar y hasta que fecha tiene para hacerlo. Así se define el monto facturado como toda la deuda que debe ser cancelada en el mes para evitar que se cobren intereses de *revolving*. Operacionalmente se conoce como $total\ mes_t$ y se encuentra compuesto por el facturado del mes anterior ($saldo\ anterior_t = total\ mes_{t-1}$), y los pagos efectuados durante dicho mes ($pagos\ mes_t$), que de ser suficientes anulan el efecto del saldo. Luego, se consideran las compras del periodo ($cargos\ mes_t$) y los intereses de *revolving* ($intereses_t$) que puedan haber ocurrido de no satisfacer el facturado del mes anterior. Así, los clientes de una tarjeta de crédito no tienen la obligación de pagar el total facturado de cada mes. La alternativa consiste en superar un pago mínimo ($Monto\ Cancelar_t$), que considera la deuda atrasada en ese instante ($Monto\ Atrasado_t$), más el pago mínimo asociado al periodo de facturación ($Pago\ Mínimo_t$). De esta forma, al superar el mínimo evita que un cliente sea considerado moroso, por lo que su monto atrasado queda definido por la Ecuación 23:

$$Monto\ Atrasado_t = \begin{cases} 0, & MC_{t-1} < PM_t \\ MC_{t-1} - PM_t, & MC_{t-1} \geq PM_t \end{cases} \quad (23)$$

MC_t = Monto a cancelar en el periodo t.
PM_t = Pagos mes del periodo t.

La Figura 12, muestra el proceso de facturación considerando como base un periodo arbitrario t. En esta se pueden observar los cargos y pagos efectuados, representados con color rojo y verde respectivamente. Estos ocurren en el intervalo

[8] *Revolving* corresponde a la situación cuando un cliente abona un monto superior al pago mínimo de su tarjeta de crédito, pero bajo el monto facturado del mes. De esta forma la institución automáticamente extiende un crédito por la diferencia entre el monto facturado y los pagos. Así el cliente no cae en cesación de pagos, pero su deuda se acrecienta por los intereses correspondientes al financiamiento de la diferencia aludida.

$(fecha\ emisión_{t-1}, fecha\ emision_t]$, consolidándose al momento de la facturación del periodo t, que en adelante se denominará $EECC_t$. Los cargos efectuados en este estado de cuenta corresponden a los movimientos que toman lugar en $EECC_t$. Por otra parte, si bien los pagos ocurren en el mismo periodo, estos responden al monto facturado y pago mínimo del periodo $t-1$, i.e. los pagos efectuados en t cubren la deuda facturada en $EECC_{t-1}$, cuyo vencimiento ocurre en pleno periodo t. Así, al momento de facturar el estado de cuenta t ($EECC_t$), recién es posible observar si existió mora o *revolving* asociado al periodo anterior.

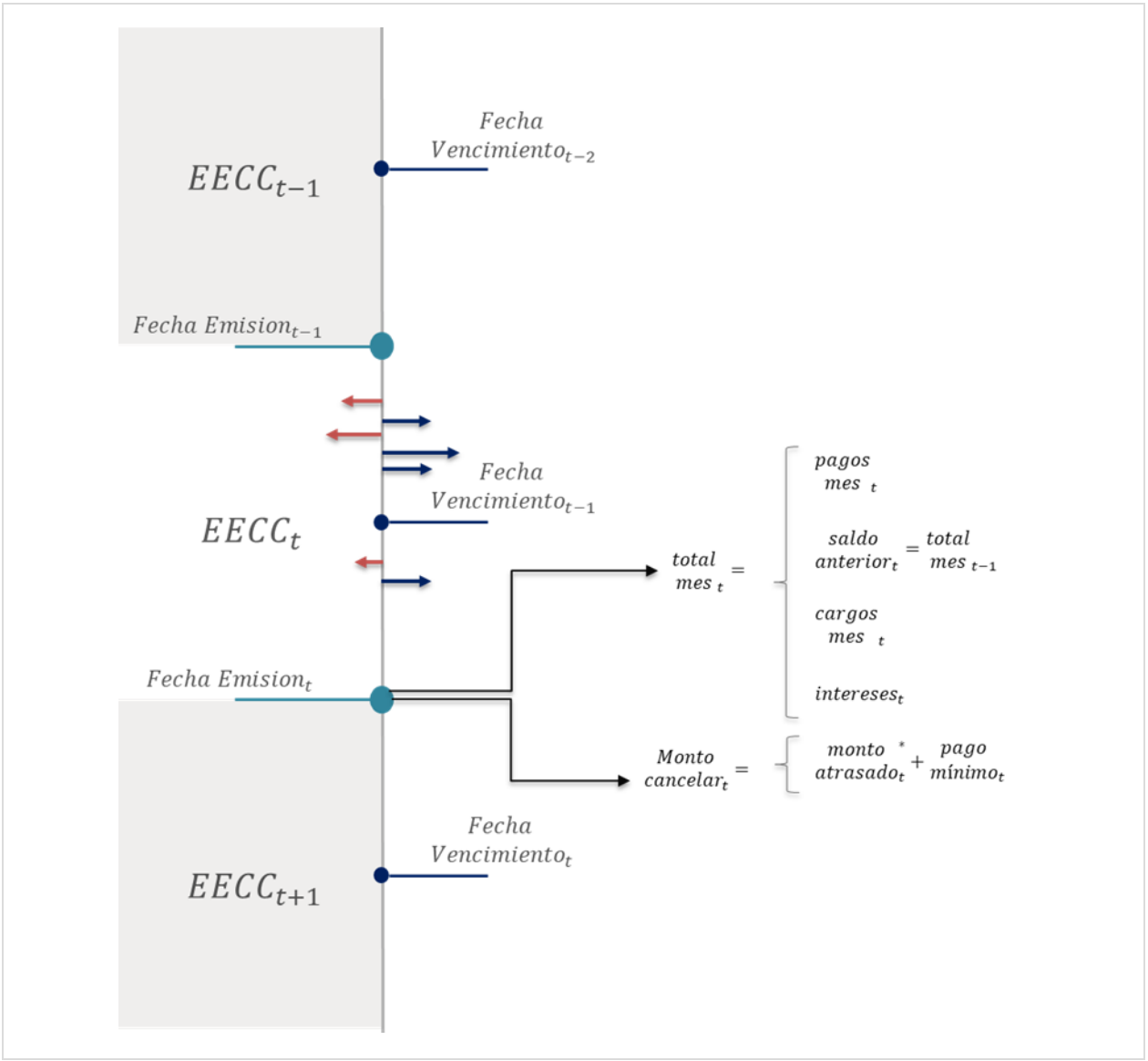


Figura 12. Proceso facturación y emisión de un estado de cuenta
Nota. Fuente: Elaboración propia.

En relación con el proceso de cobranza, la empresa ha definido para estos efectos tramos de morosidad de 30 días, a los cuales se ha adicionado la situación del cliente donde este no presenta mora y cuando la misma ha sobrepasado los 180 días, declarándose el castigo de la cuenta. Así los tramos manejados por el negocio son los presentados en la Figura 13.

Figura 13. Tramos de morosidad manejados por el negocio

Nota. Para efectos de tramos de morosidad ni la situación de castigo, ni el estado al día son considerados actualmente, pero si en el modelamiento propuesto. Fuente: Elaboración propia.

Debido a la naturaleza del negocio, existen diferentes reglas que cambian la situación del cliente a medida que este va escalando en la gravedad de su nivel de morosidad. Así, se tiene que la gestión de cobranza empieza desde el día 1. No obstante los principales hitos ocurren con el bloqueo de la tarjeta a nuevas compras en el día 60 de morosidad, así como la suspensión de la emisión del estado de cuenta al día 120. Finalmente, un cliente que sobrepasa los 180 días de morosidad le es asignada la marca castigo, ocasionando que la empresa asuma la pérdida por el no pago y el contrato no pueda ser recuperado jamás. Este cliente puede aperturar una nueva cuenta al cabo de un plazo de aproximadamente 1 año[9].

5.2 Comprensión de los datos

Los datos en los que se sustenta trabajo provienen de diversas fuentes de datos. Principalmente aquellas relacionadas con la construcción de los estados de cuenta de los clientes, su información demográfica, comportamiento transaccional e historial de pago. Del entendimiento de estas fuentes de información, se observaría una de las limitantes en el conocimiento de la situación del cliente; no es posible conocer con certeza la situación exacta del cliente en los periodos que no existe facturación. Esta situación obligó a reconstruir esos estados y determinó que en el modelamiento se optara por evaluar los efectos de la adopción de periodos de 30 y 60 días.

5.2.1 Análisis descriptivo de datos

De la descripción de la industria, se observó la posición en el mercado que ocupan los diferentes actores en la emisión de tarjetas de crédito y su situación en términos de morosidad. Sin embargo, si se toman en cuenta las fuentes de datos internas de la empresa, existen ciertas discrepancias en torno a la metodología de cálculo de estos indicadores, además del hecho de que estos corresponden a periodos diferentes. Así como primer paso, a partir de los tramos de morosidad establecidos previamente, de la Figura 14 se observa el evolutivo de la cartera. En dicho análisis solo se han considerado

[9] Si el cliente apertura una nueva cuenta, esta es asociada a un nuevo contrato, el cual tendrá su propio historial de pago y transaccional. No obstante, al momento de la evaluación para la nueva apertura, la historia pasada es considerada, además de nuevos antecedentes que puedan recogerse. Finalmente, una vez aperturada la nueva cuenta, seguirá siendo factible el cruce con la historia pasada.

a los clientes que para el año 2017 no habían sido castigados antes de empezar el periodo.

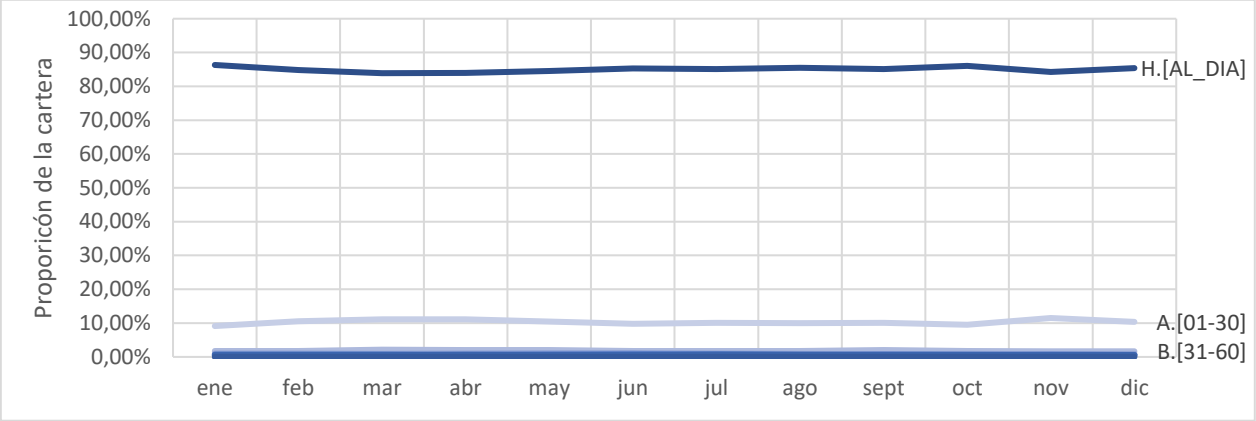


Figura 14. Evolución tramos de morosidad clientes del negocio año 2017, morosidad temprana

Nota. Fuente: Elaboración propia.

De la Figura 14 se puede observar que la cartera de clientes en situación al día se mantuvo estable a lo largo del periodo (año 2017), cifra cercana al 85% (siendo este porcentaje calculado en torno al número de contratos en cada estado). Como primera inferencia, se valida una de las claves observadas de la información provista por la SBIF. Los tramos de mora de mayor frecuencia se ubican en los niveles de menor gravedad (mora temprana). Esto se hace patente al mirar con mayor foco la misma evolución, tendencia claramente visible al observar los tramos bajo el 15% (ver Figura 15 y Figura 16**)**.

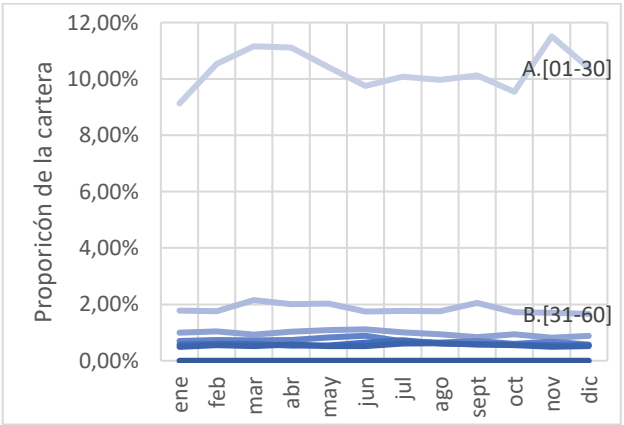


Figura 15. Evolución tramos de morosidad clientes del negocio año 2017, morosidad moderada

Nota. Fuente: Elaboración propia.

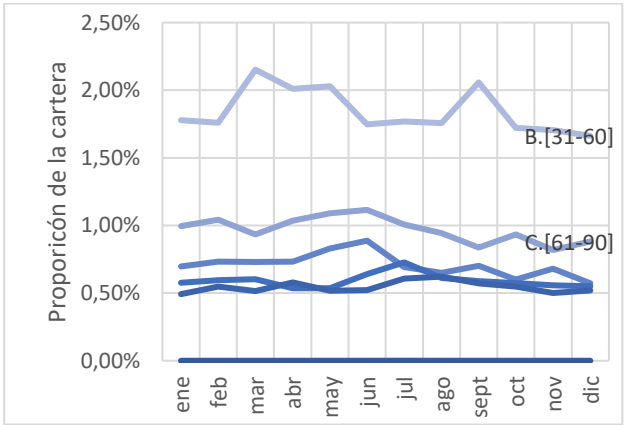


Figura 16. Evolución tramos de morosidad clientes del negocio año 2017, morosidad grave

Nota. Fuente: Elaboración propia.

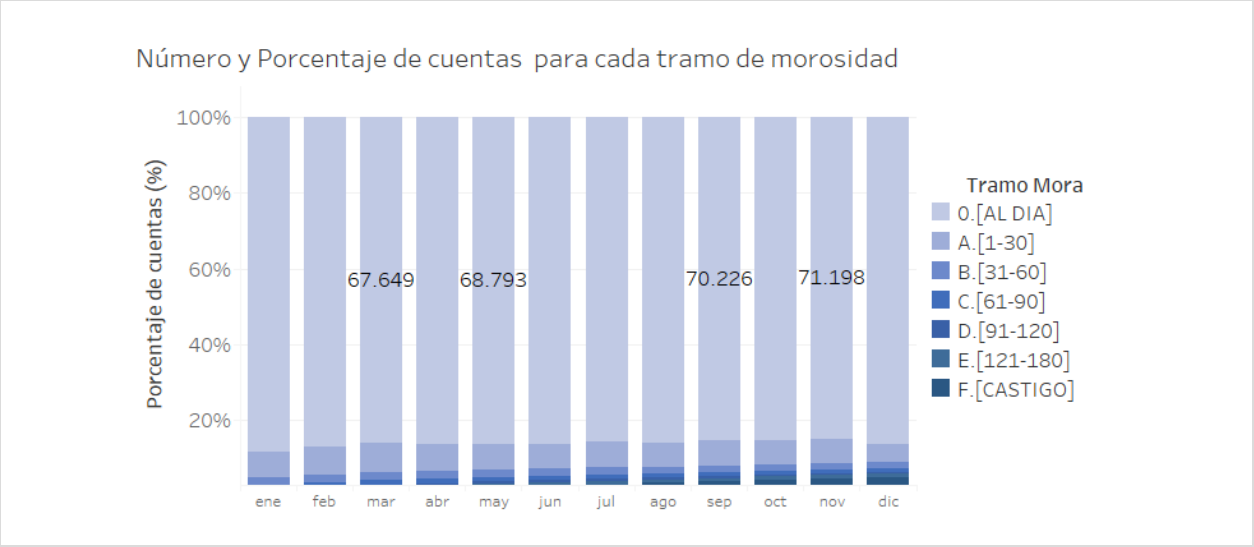


Figura 17. Tramos de morosidad para el horizonte de tiempo en estudio
Nota. Fuente: Elaboración propia.

Por otra parte, de la Figura 16 es posible apreciar que el comportamiento dominante para los tramos de morosidad de mayor gravedad no es tan claro. Así, para tramos por sobre los 91 días de morosidad no siempre se tendrá que la cantidad de personas en un tramo de menor gravedad supere a los que se encuentren en niveles más avanzados. Finalmente, la Figura 17, presenta el evolutivo descrito en términos absolutos.

La Figura 18, muestra a través de un diagrama de tipo *chord*, como se distribuyen las transiciones entre los diferentes tramos de morosidad en el mes de junio. De esta forma, el anillo representa la cantidad de clientes que se encuentran en un tramo de morosidad para junio de 2017, mientras que los arcos indican que proporción de los clientes migran hacia otro estado o se mantienen en el mismo. Nótese que los arcos son definidos como las transiciones salientes entre junio y julio del 2017.

Al observar la Figura 18, el primer elemento que predomina es la gran proporción de clientes al día, situación que ya se había caracterizado en análisis anteriores. Destaca en segunda instancia, el cómo las transiciones desde y hacia este estado, se encuentran concentradas entre los tramos más próximos en términos temporales. Dicha condición se expresa en que los estados A.[1-30] y B.[31-60] son aquellos que proporcionalmente más retornan al día. Esto aun cuando al cancelar la totalidad de una deuda es factible volver a un escenario sin deuda impaga desde cualquier nivel de morosidad, exceptuando el castigo.

Respecto a la situación de los tramos más avanzados, a medida que la mora se hace más grave, la proporción de clientes que retorna a un tramo inferior tiende a disminuir, pudiendo observarse preliminarmente que las chances de volver a un estado de deuda saludable disminuyen conforme se lleva más tiempo atrasado. Continuando con el análisis de la Figura 18, en términos del paso por un estado de morosidad, el análisis preliminar indica que quienes comienzan en un estado de baja gravedad, tienen menos probabilidad de terminar en un nivel de deuda impaga extremo.

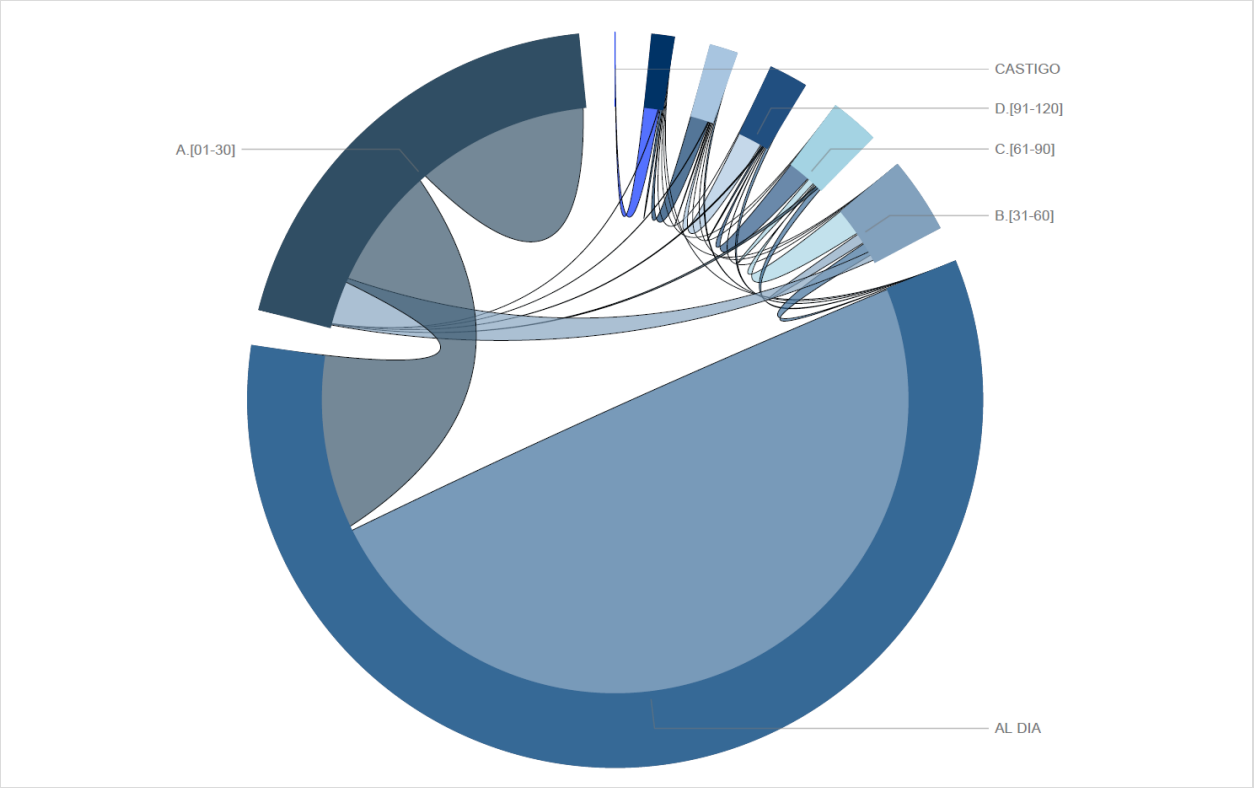


Figura 18. Transiciones morosidad de clientes cartera morosa junio 2017
Nota. Fuente: Elaboración propia.

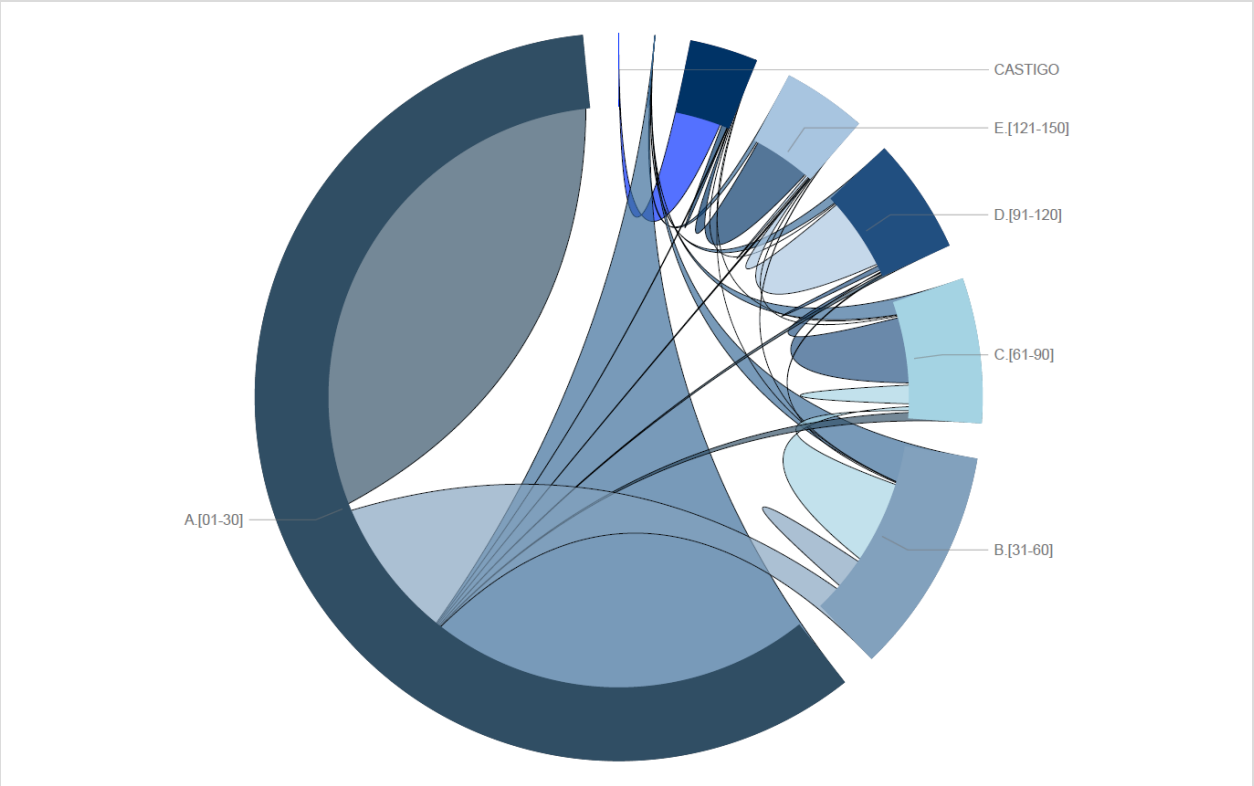


Figura 19. Transiciones morosidad de clientes cartera morosa junio 2017, sin clientes saliente del estado 0.[AL DIA]
Nota. Fuente: Elaboración propia.

Debido a que la situación al día, como era de esperar representa el general de la cartera, la Figura 19, muestra las migraciones para el mismo periodo si se excluyen aquellas que comienzan en dicho estado. Así es posible comprobar con mayor claridad algunos aspectos antes descritos. Por una parte, la menor cantidad de personas en cada tramo a medida que el nivel de morosidad se acerca al castigo. Además, se verifica que la proporción de clientes que retorna a un estado inferior adopta la misma tendencia.

En línea con los comportamientos descritos a partir de la Figura 18 y Figura 19, la interpretación de la Figura 20 es similar. Mediante una visualización de tipo *sankey*, se aprecia el estado inicial de los clientes al comienzo del periodo, el máximo tramo alcanzado durante el 2017 y su situación final al cabo de este año. De aquí es posible desprender a simple vista las menores chances de recuperar un cliente a medida que inician en una peor situación o la alcanzan. Además, se identifican que las transiciones hacia el estado al día, al castigo o al tramo de mora inmediatamente superior, predomina en todos los niveles. Por lo tanto, el abono parcial constituye una pequeña parte de las transiciones ocurridas[10].

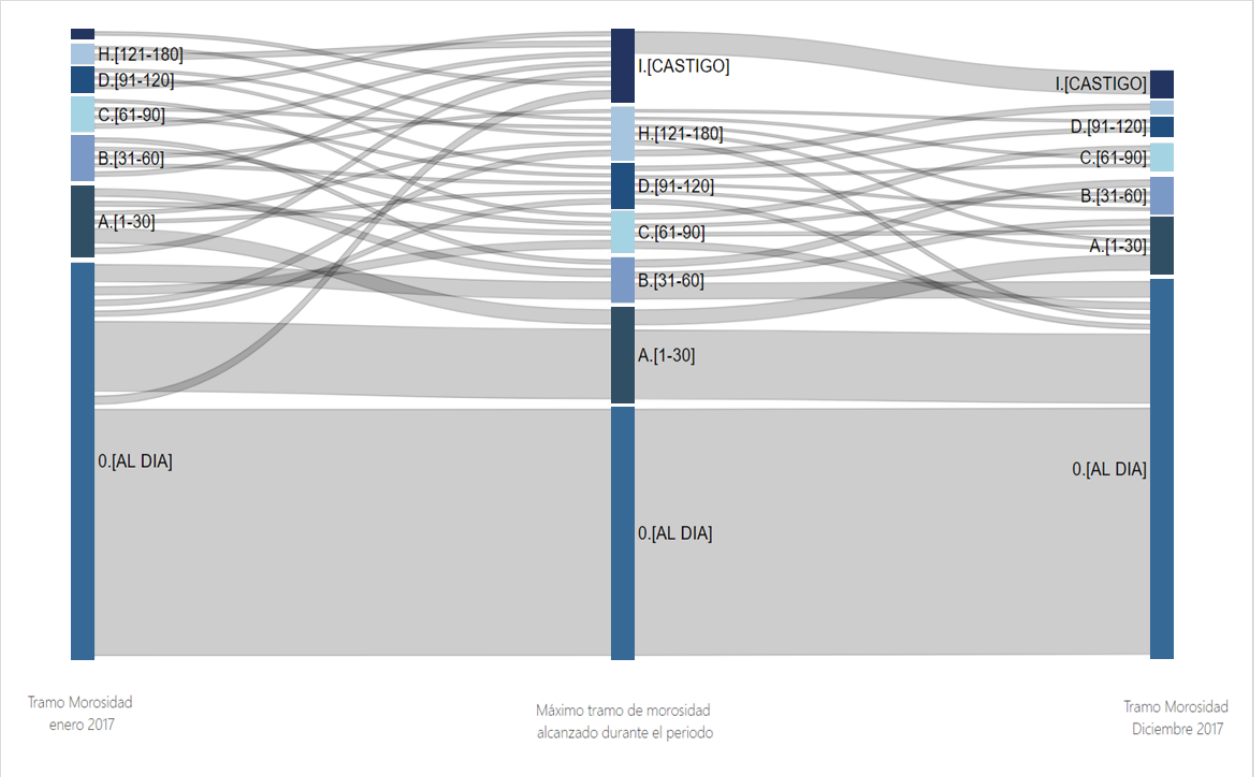


Figura 20. Máximos tramos de morosidad alcanzados para clientes en estudio año 2017

Nota. Fuente: Elaboración propia.

[10] Por abono parcial debe entenderse como el pago de una parte del monto en mora, de manera que el cliente evita avanzar a un tramo superior; al evitar un nuevo vencimiento de la deuda más antigua. Sin embargo, dado que el abono no cubre la totalidad de las deudas impagas, entonces este se mantiene en situación de morosidad.

El segundo grupo de análisis refiere a los comportamientos generales de pago. Estos son consignados en el anexo de este trabajo (ANEXO F), no obstante, sus conclusiones son discutidas a continuación, y resultan del estudio de una muestra de aproximadamente el 20%[11] de los clientes para el año 2017.

En primer lugar, se observa que los pagos tienden a concentrarse en torno a los extremos de cada mes, además de dos peaks poco perceptibles en torno a los días 10 y 20 de cada mensualidad (Figura 74). Esta situación es consistente con la distribución de las fechas de facturación y vencimiento que un cliente puede escoger para su estado de cuenta (Figura 75), y da cuenta de la tendencia a pagar con poca antelación o en el plazo límite.

Respecto a la anticipación o atraso con la que se paga el estado de cuenta, el grueso de los clientes realiza sus pagos el mismo día del vencimiento (Figura 76), lo que se condice con que la mayor parte de los clientes se encuentren al día. Adicionalmente, de los otros análisis consignados, se verifica de manera preliminar una de las hipótesis manejadas en la empresa respecto a la antigüedad de las cuentas y su relación con la prevalencia en morosidad. Existe una clara correlación entre a antigüedad de la cuenta y el promedio de días mora que presentan los clientes. Este hecho es explicable en la supervivencia de los mejores clientes y muerte de aquellos con problemas para cumplir con sus obligaciones (Figura 77), situación que con el paso del tiempo se hace más patente.

El detalle de estos análisis complementarios, son registrados en el ya referenciado ANEXO F. Otros descubrimientos preliminares consideran la poca relación entre el género y la morosidad, además de la tendencia a que esta se concentre en las tarjetas con menos prestaciones (menores descuentos, premios, etc.)[12].

5.2.2 Construcción de base analítica

La construcción de la base analítica se efectuó a partir de la consolidación de diferentes fuentes de datos. Estos orígenes contemplaron la recopilación de información acerca del cliente, registros de su historial de pago y financiero; además de su comportamiento transaccional en el negocio.

Dada la naturaleza de los datos, la construcción de la base analítica se efectuó en dos etapas. Por una parte se registró para cada cliente, identificado por su contrato, el estado de morosidad en cada mes del horizonte en estudio. Para luego determinar sus transiciones entre cada periodo. En paralelo, en una segunda base se consolidaron todos los datos a considerar en el procedimiento de *clustering*, información que originalmente

[11] Para la muestra se consideraron 726.512 clientes.

[12] Nótese que los análisis presentados son preliminares, y responden a correlaciones identificadas en las variables, por lo que no puede descartarse la presencia de factores de confusión (*confounding factors*), i.e. factor que distorsiona la medida de la asociación entre otras dos variables.

en formato panel se transformó en una única observación por cliente. En forma seguida, se aplicarían de manera estándar los algoritmos contemplados para la segmentación, en línea con lo señalado por Halkidi al presentar distintas alternativas para la clusterización de data secuencial (Halkidi et al., 2001). Este procedimiento resultaría análogo a lo desarrollado por Ho Ha y Krishnan para modelar el pago de la deuda morosa de tarjetas de crédito (Ho Ha & Krishnan, 2012). La Figura 21, muestra el cómo las fuentes de datos permitieron construir la base analítica descrita y los objetivos perseguidos en cada caso.

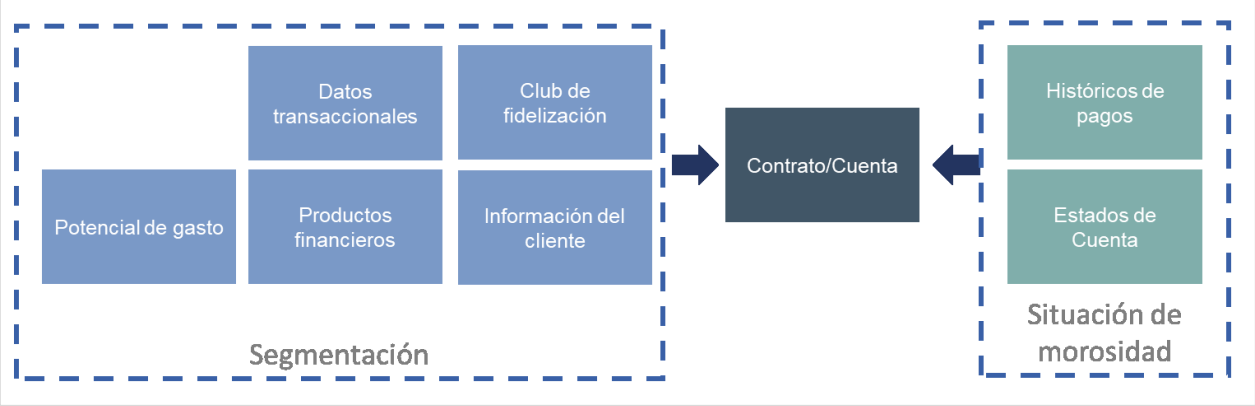


Figura 21: Orígenes contemplados en la obtención de la data
Nota.Fuente: Elaboración propia.

Para la determinación de los clientes a analizar, se comenzó por determinar una muestra de contratos para efectos de calcular estadísticos descriptivos y efectuar el seguimiento en el horizonte en estudio. Esto se construiría a partir de las siguientes definiciones: la exclusión de aquellos clientes que previo al periodo se encontrasen en situación de castigo o suspendidos sin deuda. Sobre este universo, de un total de 3.631.645 contratos, con una probabilidad del 20% se incorporaron a la muestra, sobre la cual trabaja el resto de este trabajo. Así, se culminó con un total de 726.512 contratos. Sobre estas observaciones se construiría la base analítica de manera similar al procedimiento presentado por Ho Ha y Krishnan (Ho Ha & Krishnan, 2012). De esta forma, de la consolidación de las fuentes de datos se obtendrían las 726.512 observaciones ya mencionadas, que distribuidas en 127 variables categóricas y numéricas, son detalladas en el ANEXO G.

Finalmente, como resultado directo de la construcción de la base analítica, se construyeron algunos descriptivos adicionales, los que exhibiendo datos del periodo en estudio (año 2017), permitieron determinar los pasos a seguir en términos de preparación de los datos, por ejemplo el tratamiento de *outliers* de cara a la segmentación de clientes. Los descriptivos aludidos son recogidos en las figuras siguientes (Figura 22, Figura 23, Figura 24, Figura 25 y Figura 26), considerando variables de morosidad, saturación y riesgo, siendo los estadísticos descriptivos asociados a estos mismo indicadores presentados en la Tabla 7.

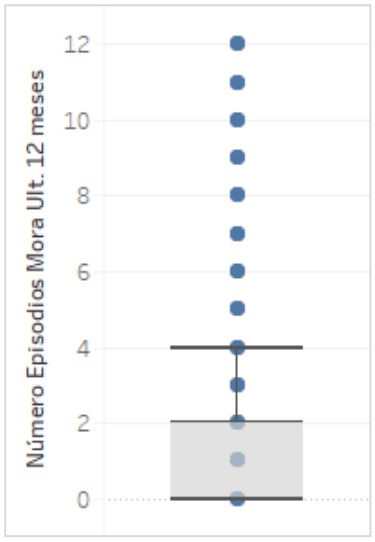


Figura 22. Distribución del número de episodios de morosidad

Nota. Fuente: Elaboración propia.

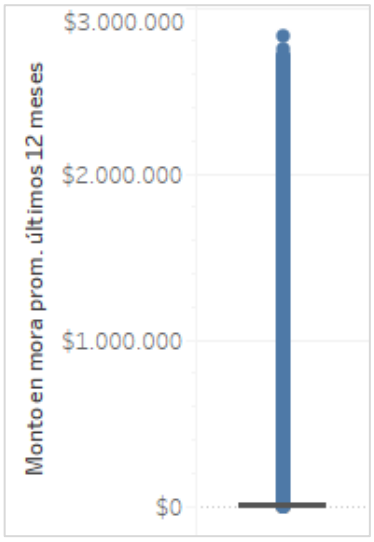


Figura 23. Distribución del monto en mora promedio

Nota. Fuente: Elaboración propia.

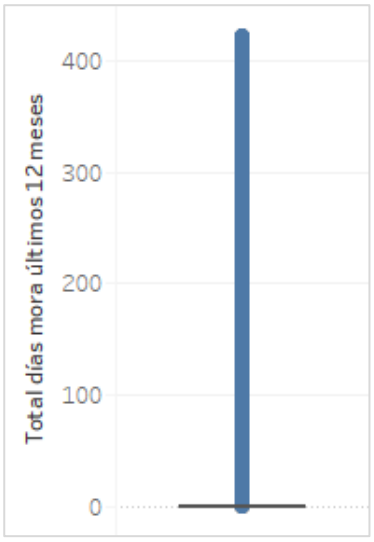


Figura 24. Distribución del total de días en situación de mora

Nota. Fuente: Elaboración propia.

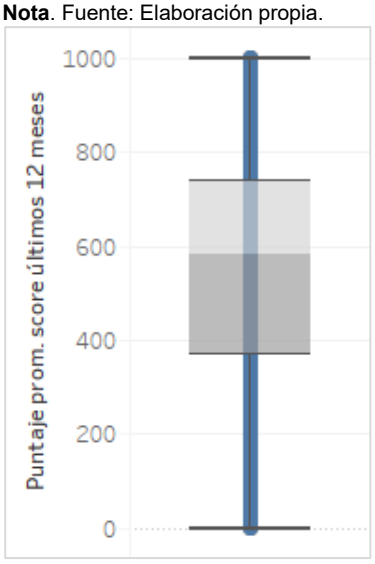


Figura 25. Distribución del score de riesgo promedio

Nota. Fuente: Elaboración propia.

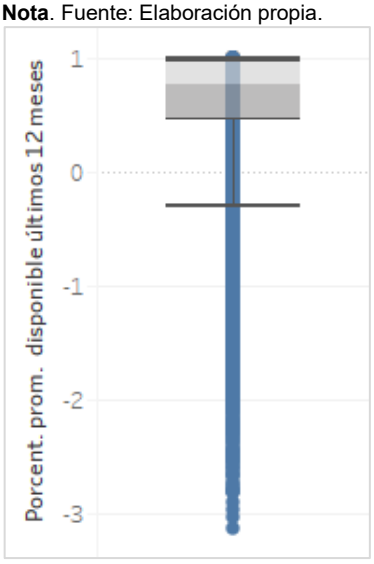


Figura 26. Distribución del promedio del porcentaje del disponible

Nota. Fuente: Elaboración propia.

Tabla 7: Estadísticos descriptivos básicos principales variables de segmentación.

	Media	Mediana	Moda	Desv. estándar	Mínimo	Máximo
Días mora	9,7	0,0	0,0	42,9	0,0	421,0
Monto en mora	23.260,0	0,0	0,0	120.361,0	0,0	2.822.581,0
Porcent. disponible	0,7	0,8	1,0	0,3	-3,1	1,0
Score riesgo	540,7	583,3	0,0	261,0	0,0	999,0
Episodios Mora	1,5	0,0	0,0	2,6	0,0	13,0

Nota. Fuente: Elaboración propia.

5.3 Preparación de los datos

Esta sección da cuenta de los diferentes procedimientos aplicados para la preparación de los datos, previo al modelamiento explicitado en la metodología. Se encuentra estructurado de acuerdo con el tipo de problema a resolver, por ejemplo, la selección de variables o el qué hacer con los valores perdidos. Sin embargo, existen tratamientos generales que no son considerados en mayor detalle dada su simplicidad. Estos comprendieron la eliminación de atributos con varianza nula, la eliminación de variables con una proporción de valores perdidos superior al 30%, y aquellos cuya calidad de información fue reportada como deficiente por el negocio.

5.3.1 Selección de datos

El proceso de selección de datos se desarrolló en paralelo a otros procedimientos para la preparación de la información. Se comenzó por un análisis de correlaciones, con el objetivo de eliminar ciertas variables redundantes en la base. De esta forma, se identifican ciertos grupos de atributos altamente correlacionados, principalmente aquellos referentes al gasto y los límites impuestos a este, el cupo y disponible. Además de otras agrupaciones de variables relacionadas con la transaccionalidad, como el número de compras en el periodo (visitas) y la acumulación de puntos. Los detalles de este análisis se encuentran en el ANEXO H (Figura 87) y responden a la evaluación del coeficiente de Pearson. Posteriormente estableciendo un corte en los valores de correlación admisibles, se eliminaron las variables consideradas redundantes (Tabla 8).

Tabla 8: Variables eliminadas por filtro de correlación

Nombre variable	Descripción
N_VISITAS_*RETAIL*_MES	Número compras en rubro *retail* en el mes.
ID_DMC_ANTIGUEDAD	Identificador de la antigüedad del cliente en la empresa.
MONTO_EMPRESA_ANIO_MOVIL_TIT	Monto gasto por titular de la cuenta últimos 12 Meses.
N_VISITAS_*RETAIL*	Número compras en rubro *retail* en el mes.
SUM_MONTO_*RETAIL*	Monto compras en rubro *retail* en el mes.
CUOTAS_AVANCES_SUM	Número de cuotas de avances pendientes.
CUOTAS_SUPERAV_SUM	Número de cuotas de super avances pendientes.
SUM_MONTO_OT	Monto gasto en POS externos al holding (On Them)
MONTO_SUPERAV_SUM	Monto total de super avances solicitados últ. 12 meses
DISPONIBLE_PROM_12M	Disponible promedio últ. 12 meses.
MONTO_PROM_DISP_SUPAVA_12M	Monto prom. disponible de super avances últ. 12 meses

Nota. Fuente: Elaboración propia.

5.3.2 Tratamiento de valores perdidos

Respecto a la presencia de valores perdidos, el enfoque empleado se basa en la adopción de un umbral de tolerancia para este tipo de datos, optando en una primera instancia por eliminar aquellas variables cuya proporción fuese superior al 30%.

Para el restante de los *missing values*, el análisis desarrollado culminaría con que estos pueden asociarse a ciertos grupos de variables (ANEXO H). En general, estos valores se

concentran en los atributos asociados al SoW[13] promedio del periodo (ver Figura 88). Así , con los patrones identificados y su visualización (Figura 89), se procede a la aplicación de un test MCAR para verificar la existencia de un potencial mecanismo subyacente en la generación de los valores perdidos.

Tabla 9: Resultados test MCAR[a]

Test MCAR	
Hipótesis Nula (H_0)	*Misising values* no son MCAR
Grados de libertad	49
Nivel de significancia (α)	0.001
p-value	0.00

Nota. [a] MCAR es la sigla del tipo de dato perdido que responden a componentes aleatorias (*Missing Completely at Random*).
Fuente: Elaboración propia.

De la Tabla 9, se observa que los resultados del test MCAR no permiten rechazar la hipótesis nula. Por lo tanto, asumir que el mecanismo de generación de los valores perdidos es aleatorio no es estadísticamente correcto. Finalmente, estas observaciones son eliminadas de la base para efectos del modelamiento, justificándose dicha acción en la alta proporción de valores perdidos y los sesgos que una imputación podría introducir en los pasos siguientes.

5.3.3 Tratamiento de valores atípicos

Para el tratamiento de valores atípicos, se aplicó un procedimiento multivariado de detección de *outliers*. Considerando así la potencial existencia de observaciones fuera del rango "normal", cuya presencia pudiese distorsionar la aplicación de los modelos propuestos. Al mismo tiempo se tomó en cuenta la interacción de las variables al momento de definir qué es un valor extraño mediante el uso de un método multivariado.

El proceso de detección de *outliers* se efectuó haciendo uso de la distancia de Mahalanobis y proyectando dichos resultados en un espacio de 2 dimensiones mediante análisis de componentes principales. En este caso, a diferencia del criterio comúnmente empleado en la literatura en el que se asume que el cuadrado de la distancia distribuye CHI-2[14] , se procede a establecer un corte diferente para determinar la cantidad de *outliers* (1%).

Debido a que al establecer un nivel de confianza $\alpha = 0{,}975$, se obtendría alrededor de un 10% de valores atípicos, para efectos de identificación de *outliers* se fijaría un corte en torno al 1%. En forma seguida, su tratamiento consistiría en su eliminación de la base. De esta manera, se procuraría la minimización del efecto de estos valores sobre las

[13] Share of Wallet (SoW) corresponde a la proporción del potencial de gasto que un cliente consume.

[14] El supuesto para la detección de *outliers* indica que la distancia de Mahalanobis (d) distribuye CHI-2 con tantos grados de libertad como la cardinalidad de las variables numéricas ($d^2 \sim \chi^2_{df=|X|}$).

estrategias de *clustering* a desarrollar en pasos siguientes, conocidas por su sensibilidad a la presencia de valores atípicos.

5.3.4 Selección de variables

Debido a una de las principales problemáticas asociadas a las metodologías de *clustering*, la denominada maldición de la dimensionalidad (*curse of dimensionality*). Resultó necesario desarrollar una selección de variables previa a la aplicación de los algoritmos mencionados. Esto, con el objeto de identificar un subconjunto de atributos relevantes, de manera que las nociones de distancia no perdieran su "sentido".

Entre los enfoques considerados para esta tarea, se comienza por el uso de metodologías de filtrado estadístico. Para ello, se construye una variable auxiliar de situación de mora, con la cual evaluar la independencia de esta respecto a los predictores. Mediante un test de Kolomogorov- Smirnov se abordan las variables numéricas y vía prueba de CHI-2 las categóricas. De la aplicación de ambas no resultó factible filtrar por ninguna de estas variables, pues todas arrojarían resultados significativos. Así, a modo de ejemplo, la Figura 27, muestra la aplicación del test KS sobre la variable categoría del cliente (variable categórica), mientras que la Figura 28 hace lo mismo para la variable EDAD (variable numérica).

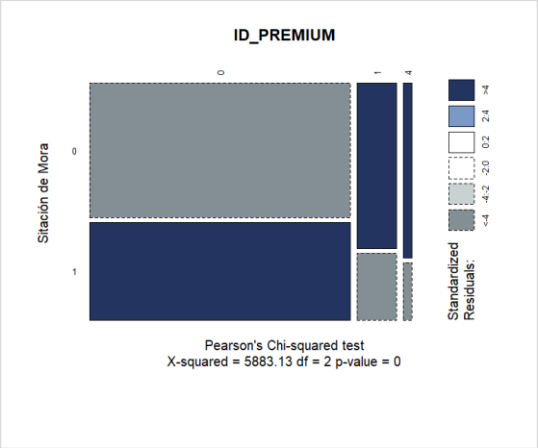


Figura 27. Test CHI-2 (χ^2) para categoría cliente

Nota. Fuente: Elaboración propia.

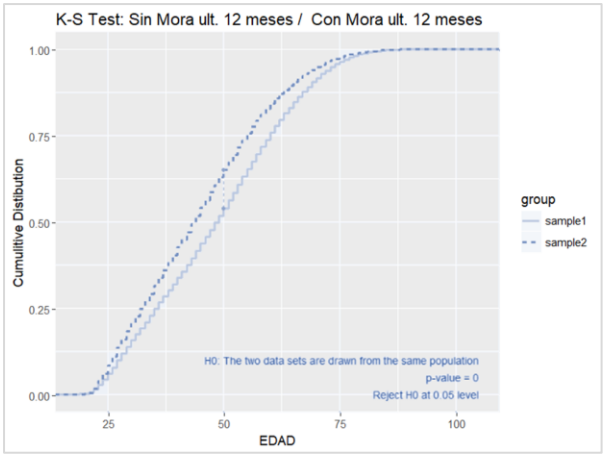


Figura 28. Test KS para la variable edad

Nota. Fuente: Elaboración propia.

Finalmente, para efectos de culminar el proceso de selección de variables, utilizando un árbol de inferencia condicional, se aprovecharía la importancia de los predictores para tomar una decisión respecto a los atributos a considerar en el *clustering*. La Figura 29, muestra los 10 primeros predictores arrojados por dicho procedimiento, de los cuales, se incorporaron 5 de ellos al algoritmo de *clustering* descrito en un capítulo próximo.

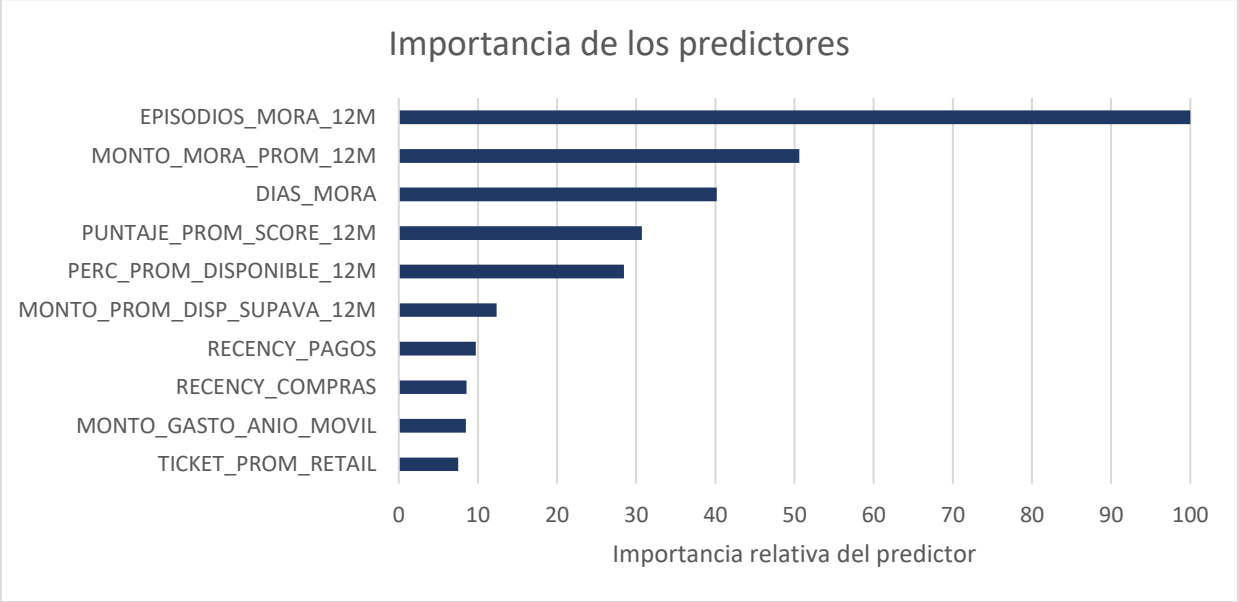


Figura 29. Top 10 de importancia de los predictores de morosidad
Nota. Fuente: Elaboración ropia.

5.4 Modelamiento

La fase de modelamiento de este trabajo se sustenta en dos etapas principales, la segmentación de los clientes según su comportamiento histórico, y el posterior modelamiento de la morosidad mediante cadenas de Markov. El presente capítulo describe el desarrollo de los pasos descritos, así como los resultados directos del proceso, cuya discusión y conclusión es abordada en forma seguida.

5.4.1 Segmentación de clientes

Como se detallara en el marco metodológico, para efectos de incorporar heterogeneidad a las cadenas de Markov, se contempló el uso de estrategias de *hard* y *soft clustering*; utilizándose *k-means* y *fuzzy c-means* respectivamente.

Las dos metodologías de *clustering* propuestas corresponden a lo que se conoce como del tipo particional y de lógica difusa. Por lo tanto, la necesidad de seleccionar de antemano el número de segmentos es imperativo, al mismo tiempo que este debe tener en consideración la cohesión y separación de los clústeres, a medida que se itera en diferentes valores. Así, la elección del valor adecuado de grupos; usualmente denotado por k, se fundamenta en el uso de métricas de validación interna, las que aplicadas sobre una muestra de la data permitieron determinar el mejor número de segmentos.

Basándose en la revisión bibliográfica, de acuerdo a lo presentado por Xu (Xu, 2005, p. 665), se utilizaría como principal criterio de definición del número de segmentos el índice de Caliñski-Harabasz. Adicionalmente, se presentan otros indicadores a modo de comparación.

Una vez establecida la cantidad de segmentos, se desarrolla la partición de la data de manera completa. Para ello se recurrió a las implementaciones de los algoritmos descritos en lenguaje R, para finalmente terminar este análisis con los resultados de la validación externa de las particiones obtenidas.

Finalmente, cada cliente perteneciente a los clústeres resultantes, se dividen manteniendo una proporción de 80%-20%, de manera de generar conjuntos de validación y entrenamientos para etapas posteriores del trabajo (modelamiento de cadenas de Markov). Este procedimiento es efectuado como lo indica la Figura 30.

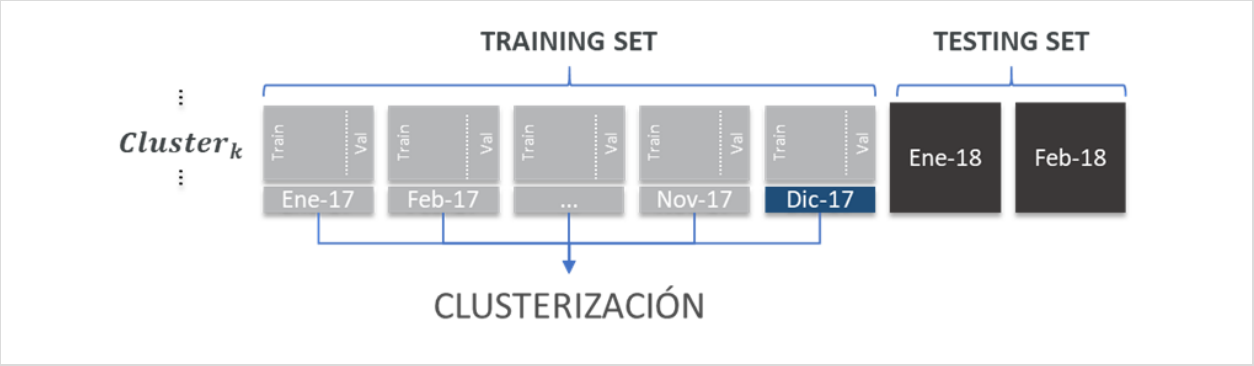


Figura 30. Partición de la data sujeta a los resultados del clustering
Nota. Fuente: Elaboración propia.

5.4.1.1 Hard Clustering (k-means)

Como se indicó, para la segmentación basada en una lógica de *hard clustering*, la elección del número de segmentos, i.e. el parámetro k, sería determinada a partir de índices de validación interna calculados en una muestra de los datos. La Figura 31, Figura 32 y Figura 33 muestran el valor de las métricas para distintos valores de k (número de clústeres).

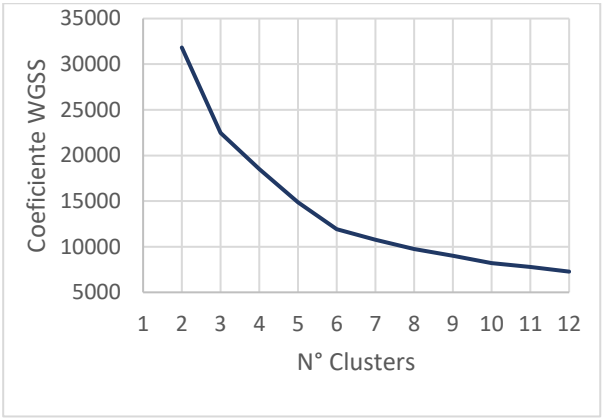


Figura 31. WGSS[a] vs número de clústeres para k-means
Nota. [a] Within group sum of squares.Fuente: Elaboración propia.

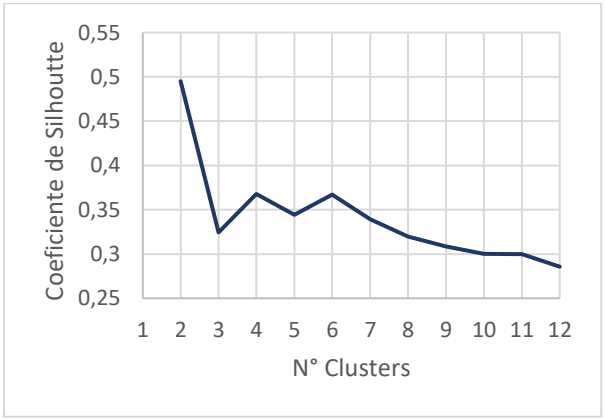


Figura 32. Coeficiente de Silhoutte vs número de clústeres para k-means
Nota. Fuente: Elaboración propia.

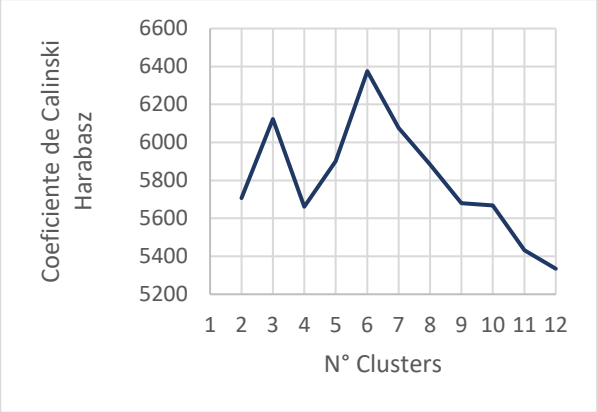


Figura 33. Coeficiente de Caliñski-Harabasz vs Número de clústeres para k-means

Nota. Fuente: Elaboración propia.

El índice de Caliñski-Harabasz determina que el número de grupos a generar se establezca en $k = 6$ (Figura 33), ya que es el k que lo maximiza. Este valor es consistente con las variaciones experimentadas por el valor del WGSS (ver 3.4.2.4.1.1 Within Group Sum of Squares (WGSS)), métrica en la que la identificación de un "codo"[15] conduce a la elección del número de segmentos óptimo, y que de acuerdo con la Figura 31 puede atribuirse de igual forma $k = 6$. Finalmente, de la aplicación del coeficiente de Silhoutte, la existencia de un máximo local para el mismo número de clústeres, valida parcialmente la decisión de asumir un total de 6 segmentos para este método (Figura 32).

Culminada la selección del número de grupos, se continúa con la aplicación del algoritmo sobre la base completa. Así, en forma seguida se aborda cualitativamente la validación externa a partir de gráficos de disimilitud, para luego continuar con la caracterización de los segmentos.

Las visualizaciones aludidas previamente, remitidas al ANEXO J, muestran la clusterización de un conjunto de datos aleatorio, en contraste con la segmentación efectuada sobre los datos. Así, la Figura 90 hace alusión al primer caso, mientras que la Figura 91 a los datos reales. De ambas visualizaciones, es posible observar la oposición entre la inexistencia de una estructura subyacente para el caso en que se tienen segmentos sobre datos aleatorios, a diferencia del caso en que se usan datos reales. De esta forma, se concluye que existe una ordenación natural en los grupos generados.

En relación con los clústeres obtenidos, estos responden a diferentes niveles en las variables incorporadas, tal y como se explicitara en el capítulo 5.3.4 Selección de variables. Por este motivo, del análisis más directo resulta la identificación de dos grupos

[15] La denominada "regla del codo" hace alusión a la identificación del punto donde la tendencia de crecimiento/decrecimiento de un indicador encuentra su inflexión, definiendo un corte para tomar una decisión sobre el mismo.

de clientes marcadamente opuestos. El clúster 5 y clúster 6 (ver Figura 34), siendo el primero de los mencionados reconocible por su conformación a partir de los clientes con peores comportamientos de morosidad. En contraste, los clientes pertenecientes al clúster 6 exhiben niveles de morosidad y riesgo más bajos, como también una mayor transaccionalidad y disponible a lo largo del periodo.

Por otra parte, destaca también la existencia de segmentos cuyos centros, difieren respecto a otros tan solo en algunas dimensiones. Es el caso de la comparativa entre el clúster 2 y clúster 5, en que las discrepancias radican en el nivel de mora y uso del disponible. Así, los usuarios de tarjetas pertenecientes al grupo 2, exhiben un nivel de endeudamiento levemente mejor que sus pares del grupo 5.

Una situación parecida es evidenciada en el comportamiento de los clústeres 1 y 3, que mantienen las mismas tendencias entre sí, aunque en diferentes escalas. Estos grupos pueden ser caracterizados como los clientes con morosidad y riesgo, pero que mantienen transaccionalidad en el negocio.

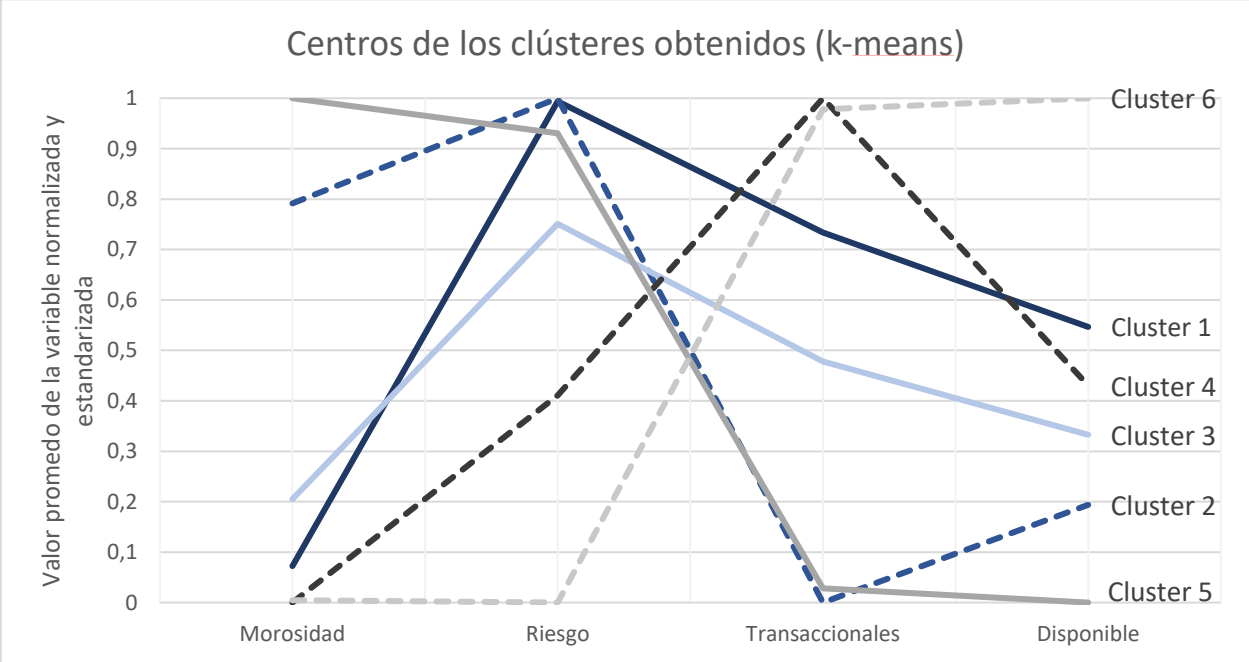


Figura 34. Centros de los clústeres caracterizados por dimensiones principales

Nota. Los centros se presentan en base a la agregación de variables más relevantes en tópicos y a su representación estandarizada y normalizada para su visualización. Se considera la el opuesto de algunas variables para facilitar la interpretación (Score de riesgo y recency de pagos). Fuente: Elaboración propia.

Tomando nuevamente como punto de partida la Figura 34, es posible caracterizar los segmentos y asignarles una denominación comercial para su gestión. Así, la Tabla 10 presenta el nombre con el que se reconocerá cada uno de los grupos obtenidos, además de sus características principales y número de clientes que lo componen. Nótese que a diferencia de la Figura 34, para mostrar sus resultados la Tabla 10 no estandariza ni centra las magnitudes de los valores, de manera de dar cuenta de las cifras reales y en bruto de las variables. Adicionalmente el detalle de otras características de cada uno de

los 6 grupos descritos es profundizado en el ANEXO I, donde se pueden observar los centros de los clústeres considerados, así como el comportamiento de otras variables que se asocian a las dimensiones presentadas en la Tabla 10.

Tabla 10: Centros hard clustering (k-means), caracterización resumen

Dimensión	Variable	Segmento	Infrecuentes con deterioro pausado	Inactivos forzados con deterioro acelerado	Activos esporádicos con deterioro moderado	Activos Saturados con deterioro pausado	Inactivos forzados de deterioro progresivo	Activos insaturados con deterioro pausado
			1	2	3	4	5	6
Tamaño	# de Clientes		73.487 (23,60%)	14.100 (2,61%)	58.429 (10,81%)	131.396 (24,31%)	2.561 (0,47%)	260.455 (48,19%)
Morosidad	# episodios Mora		0,8	9,6	6,3	0,9	10	0,4
	Monto mora		$2.624	$285.635	$62.290	$4.532	$1.418.335	$1.853
	# Días mora		1	225,8	21,3	0,9	219,1	0,3
Saturación	% del Disponible		85%	1%	33%	39%	-58%	90%
Riesgo	Score de Riesgo		166	162	307	504	203	741
			Uso tarjeta infrecuente y/o asociada a bajos montos. **Riesgo de morosidad bajo** por poca actividad.	Clientes de gasto moderado con **inactividad forzada** por morosidad. Montos de morosidad altos **(mora grave).**	Uso esporádico de la tarjeta. Niveles de saturación moderados. **Episodios de mora intermitente.**	**Uso activo de la tarjeta con niveles de saturación moderados**. Alto uso se traduce en mayor score de riesgo, aun cuando no tienen mora reciente.	Clientes de gasto alto con **inactividad forzada** por mora. Mayores montos asociados a morosidad (**mora gravísima**).	**Uso frecuente de tarjeta**, sin alcanzar altos niveles de saturación. No presentan episodios ni conductas de riesgo.

Nota. [a] **I**ndicador del comportamiento crediticio que toma valores entre 0 y 999. Dónde el valor 0 corresponde a clientes de pésimo comportamiento o nuevos en la empresa y 999 la mejor conducta crediticia. Fuente: Elaboración propia.

5.4.1.2 Soft Clustering (Fuzzy C-means)

Al igual que en la aplicación de *hard clustering*, para efectos de desarrollar la segmentación con lógica difusa, se comenzó por seleccionar el número apropiado de grupos con la misma metodología que se utilizase con *k-means*. Esta clusterización se efectuó con el algoritmo *fuzzy c-means*, que admite la pertenencia de las observaciones a más de un grupo. Así, la Figura 35 muestra la variación de la suma de cuadrados intragrupo respecto al número de clústeres, pudiendo atribuirse la mayor variación entorno a $k = 7$. Del mismo modo, de la Figura 36, se puede apreciar que el coeficiente de Silhoutte se maximiza en $k = 7$, situación que también ocurre con el índice de Caliñski-Harabasz (ver Figura 37). De esta forma, a partir de los índices de validación interna se concluye la pertinencia de fijar el número de segmentos en $k = 7$.

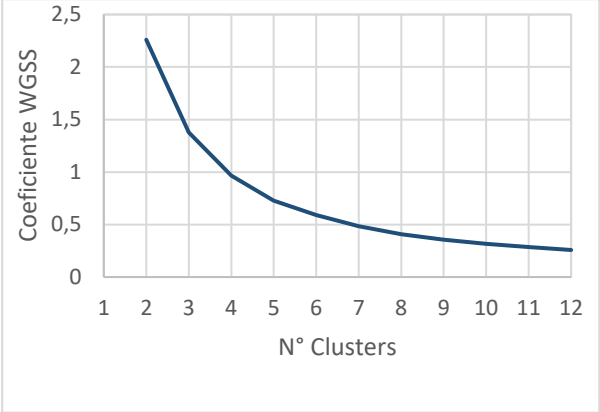


Figura 35. WGSS[a] vs número de clústeres para fuzzy c-means

Nota. [a] Within group sum of squares. Fuente: Elaboración propia.

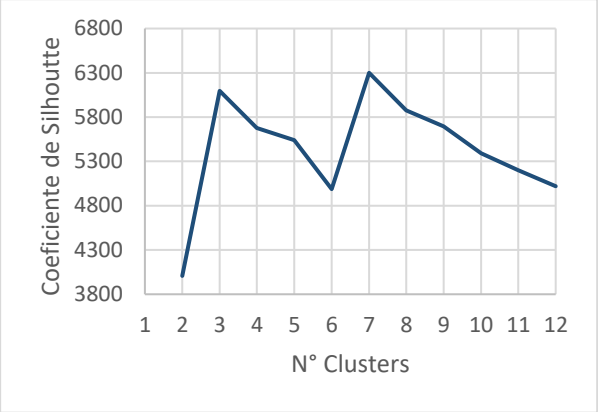


Figura 36. Coeficiente de Silhoutte vs Número de clústeres para fuzzy c-means

Nota. Fuente: Elaboración propia.

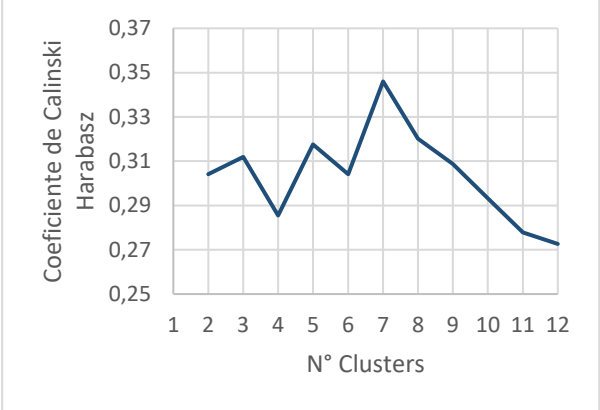


Figura 37. Coeficiente de Calinski-Harabasz vs Número de clústeres para fuzzy c-means

Nota. Fuente: Elaboración propia.

Para la validación externa, se procedería al igual que con el método anterior; documentando las visualizaciones empleadas para el análisis en el ANEXO J. De esta forma, nuevamente se comprueba de manera cualitativa, mediante gráficos de disimilitud, la existencia de una estructura subyacente en los datos reales. Este aspecto es evidenciable en la comparativa de las Figura 92 y Figura 93, del anexo ya mencionado.

Respecto a los centros resultantes de la clusterización con lógica difusa, estos son exhibidos en la Figura 38, en la cual las variables han sido agrupadas en tópicos al igual que como se hiciera antes. La conclusión directa refiere a la identificación de 5 comportamientos con marcadas diferencias, además de la existencia de dos pares que difieren en escalas (clúster 2 versus clúster 3, y clúster 4 con clúster 6).

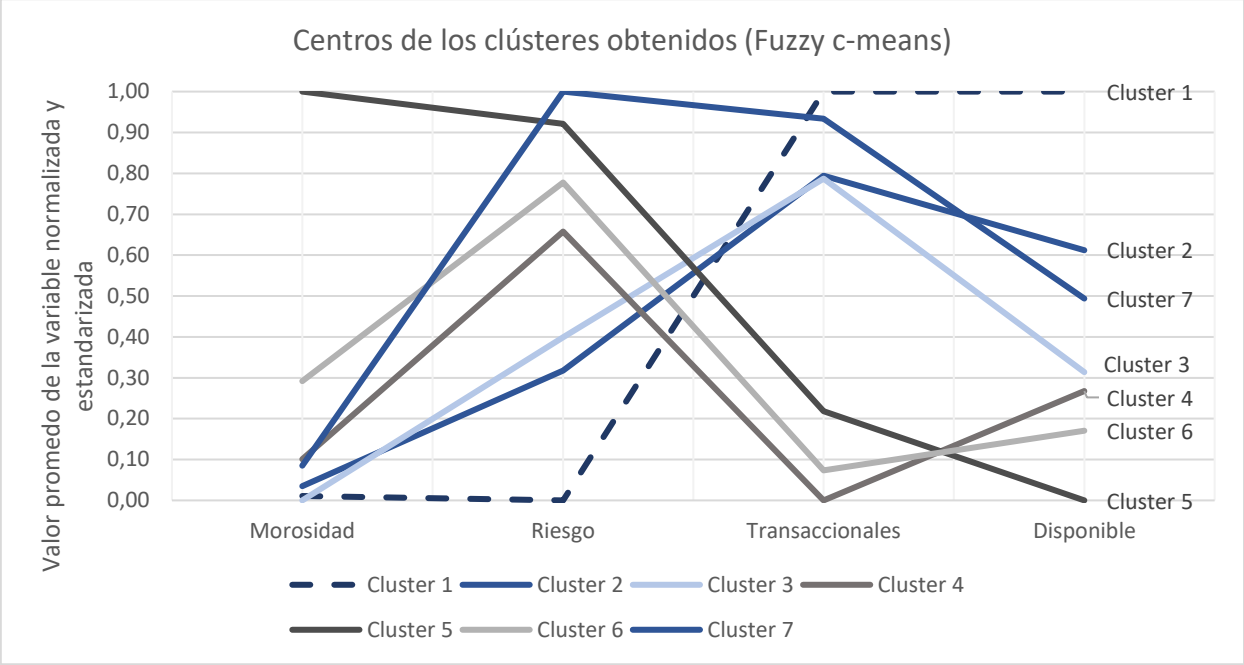


Figura 38. Centros de los clústeres caracterizados por dimensiones principales (Fuzzy c-means)

Nota. Los centros se presentan en base a la agregación de variables más relevantes en tópicos y a su representación estandarizada y normalizada para su visualización. Se considera la el opuesto de algunas variables para facilitar la interpretación (Score de riesgo y recency de pagos). Fuente: Elaboración propia.

En relación con los tamaños de los clústeres, estos resultaron más homogéneos respecto a la conformación obtenida de la aplicación de *hard clustering*. Así, la Tabla 11, presenta la composición de los grupos mencionados.

Tabla 11: Caracterización general soft clustering (c-means)

Clúster	N° de clientes	% Clientes
1	142.592	26,39%
2	141.453	26,17%
3	91.154	16,87%
4	57.071	10,56%
5	14.985	2,77%
6	37.446	6,93%
7	55.727	10,31%

Nota. Fuente: Elaboración propia.

En relación con una caracterización detallada de los segmentos y adopción de denominaciones comerciales para su gestión; esta fase de la metodología no contempla ese tipo de profundización. El motivo se explica en la finalidad de esta parte del modelamiento: el representar el comportamiento a nivel de cliente aprovechando los

niveles de pertenencia de cada observación en cada uno de los clústeres. Así, el procedimiento descrito es cubierto en la sección correspondientes a cadenas de Markov, resultados a nivel de cliente (ver sección 5.4.2.2 Nivel de cliente).

5.4.2 Cadenas de Markov

Como proceso estocástico, las cadenas de Markov permiten modelar problemas con incertidumbre. Para este trabajo se ha contemplado la evaluación de dos enfoques con los cuales abordar la morosidad de clientes. Así, en primer lugar, se refiere a la construcción de modelos que describen comportamiento a nivel de grupo, para luego evaluar pronósticos que lo hacen por cliente.

5.4.2.1 Nivel de grupo

Se ha establecido la pertinencia de evaluar distintos tipos de cadenas de Markov para representar el comportamiento de mora de los clientes. Estas aproximaciones responden a las hipótesis sobre los efectos que granularidad y variables de estado tienen sobre el desempeño de los pronósticos e *insights* que se pueden conseguir. De esta forma, la Tabla 12 presenta las denominaciones de los modelos propuestos, cuyas diferencias radican en su granularidad y variables de estado.

Tabla 12: Denominaciones para modelos de Markov propuestos

		Granularidad	
		60 días	**30 días[a]**
Variables de estado	**(Tramo morosidad)**	MK1 (Figura 39)	MK2 (Figura 40)
	(Tramo morosidad, tiempo)	MK3 (Figura 41)	MK4 (Figura 42)

Nota. [a] Debido a restricciones sobre la base de datos disponible, antes explicitada, el tramo de morosidad 121-180 días, no puede ser considerado en mayor granularidad, siendo esta una de las limitantes del trabajo. Fuente: Elaboración propia.

Las cuatro cadenas señaladas (ver Tabla 12), son representadas como grafo en las figuras siguientes. Así, la Figura 39 muestra la cadena más simple de todas, en la que la única variable de estado es el tramo de morosidad en una granularidad que define periodos de 60 días (Modelo MK1). Esta agregación es la única diferencia entre ella y la cadena representada en la Figura 40 donde los tramos de morosidad son de 30 días (Modelo MK2). En ambos casos, el avance en tramos de morosidad es representado en movimientos sobre la horizontal, los que deben leerse desde izquierda a derecha.

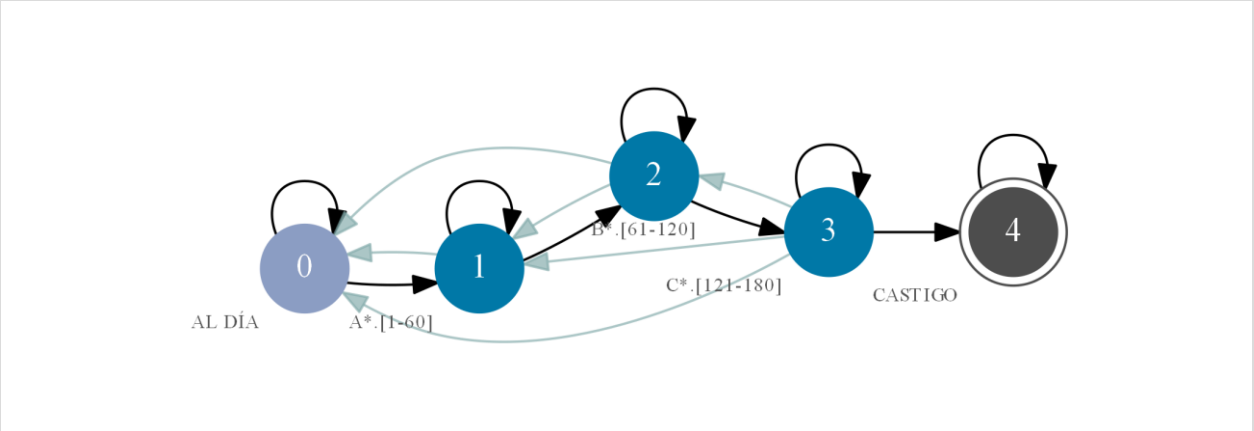


Figura 39. Representación cadena MK1
Nota. Fuente: Elaboración propia.

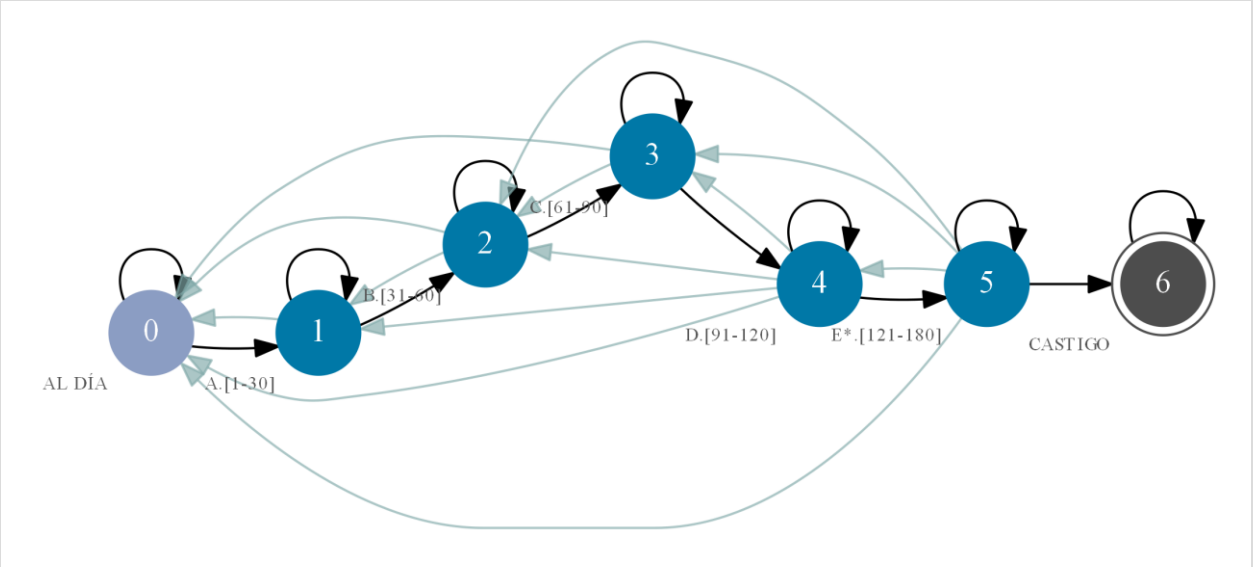


Figura 40. Representación cadena MK2
Nota. Fuente: Elaboración propia.

Por otra parte, la Figura 41 (Modelo MK3) y Figura 42 (Modelo MK4) dan cuenta de modelos con más de una variable de estado; el tramo de morosidad y el tiempo consecutivo en mora. Para estos casos la representación en grafo asume que el tiempo consecutivo en mora se traduce en movimientos sobre la horizontal, con un máximo de 6 periodos, mientras que el tramo de morosidad se lee sobre el eje vertical hasta llegar al castigo.

Nótese que, las transiciones modeladas sobre todas las cadenas corresponden a aquellas factibles de ocurrir en el comportamiento de mora de un cliente. Por lo tanto, se asume que un deudor que abona parte de sus montos impagos tiene la posibilidad tanto de saldar su deuda por completo, así como de cancelar una fracción de sus obligaciones. Volviendo así a un tramo de mora inferior o evitando su avance en la gravedad de su situación.

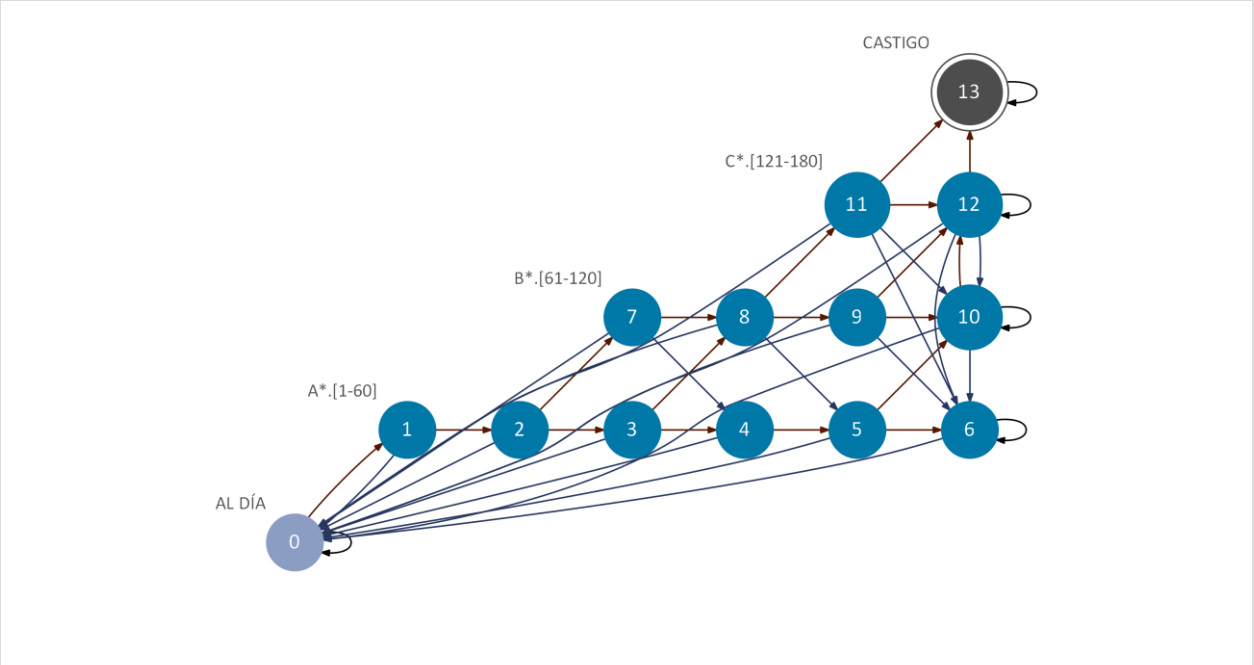


Figura 41. Representación cadena MK3
Nota. Fuente: Elaboración propia.

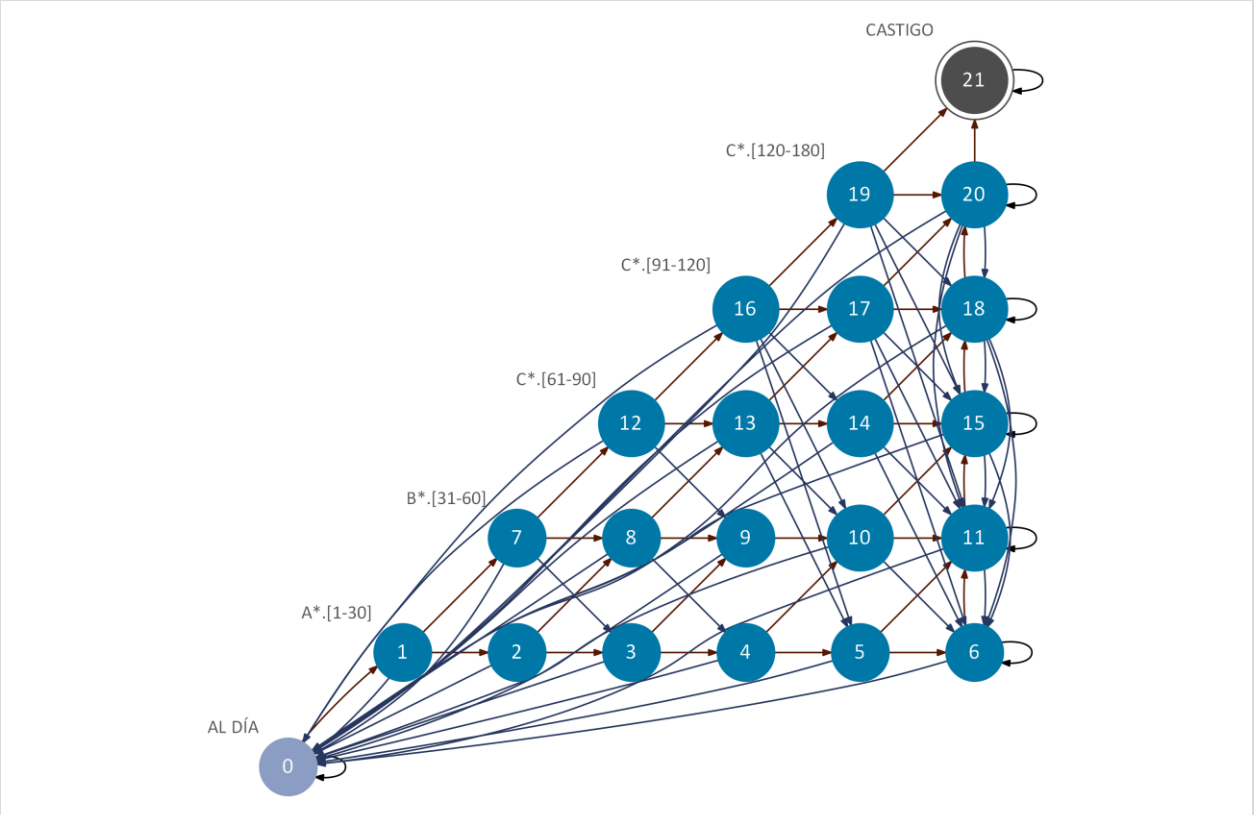


Figura 42. Representación cadena MK4
Nota. Fuente: Elaboración propia.

Cabe destacar que, para los modelos más simples MK1 y MK2, su formulación se basa directamente en las migraciones que pueden experimentar los clientes. Sin embargo, tanto el modelo MK3 y MK4 y la consideración del tiempo consecutivo en mora, tienen su

inspiración en la metodología propuesta por Osses en su método de valoración de clientes (Osses Godoy, 2015).

5.4.2.1.1 Ajuste de los modelos

La estimación de las probabilidades de transición para los diferentes modelos se basa en la metodología propuesta por Carolina Segovia (Segovia Riquelme, 2005) y más tarde replicada por Carlos Roco (Roco Benavides, 2010). El detalle de este procedimiento se encuentra registrado en el marco teórico (ver 3.4.1.1 Estimación de probabilidades de transición del sistema), cuya aplicación a esta memoria se abordó mediante un desarrollo propio, a falta de librerías con la flexibilidad de adaptarse al modelamiento propuesto. De esta forma, se implementaron dos módulos en lenguaje Python encargados de desarrollar el ajuste de los métodos y la visualización de sus resultados.

De la aplicación del método para el ajuste de las probabilidades, la Tabla 13 resume los resultados de cada uno de los modelos, cuyo detalle se presenta en el ANEXO K. De esta tabla es posible desprender las primeras conclusiones respecto a la naturaleza del ajuste en los comportamientos modelados.

En primer lugar, se verifica a partir de la Tabla 13 el comportamiento esperable en los errores absolutos, las cadenas con un menor número de estados presentan errores mayores que aquellas que en su estructura definen un mayor número de estados posibles. Esta situación tiene su excepción en el modelo MK4, donde esta tendencia se rompe al comparar con su par MK3. Siendo esta situación atribuible a la mayor dificultad de pronosticar sobre una mayor granularidad y número de estados, rompiendo la tendencia mencionada.

En relación al error promedio ponderado, al ser esta la métrica con la que trabajos anteriores comparasen entre modelos con diferentes números de estados, este indicador será relevante para explicar que cadena se comporta de mejor forma en términos de su ajuste y capacidad predictiva. Así, este error mantiene la tendencia evidenciada por su versión absoluta, pero con un relevante cambio de escala que acorta las diferencias entre los cuatro enfoques. Además, recordando las limitaciones de este indicador, se debe tener en cuenta la sub-ponderación de los fallos en la predicción de las transiciones, pues no considera el error asociado al pronóstico de transiciones cuando no las hubo en la realidad; indeterminación de la fórmula.

Respecto a los resultados obtenidos del error promedio ponderado simétrico, en primer lugar, destaca el quiebre de la tendencia a que mayor granularidad exhiba mejor desempeño cuando el modelamiento considera una única variable. Por el contrario, al considerar dos variables de estado, la tendencia mantiene la relación de orden en el error. Finalmente, existe un efecto de escala sobre los valores, resultante de la incorporación de los casos en que el pronóstico difiere de una cantidad de transiciones observadas nulas. Este hecho también es el causante de que las relaciones entre MK1 y MK2 se inviertan.

Tabla 13: Resultados métricas de ajuste modelos de Markov a nivel de cliente

		Características Modelo			Errores		
		Variable(s) de estado	Granularidad	Ventana de tiempo	Error absoluto promedio	Error prom. ponderado	Error prom. ponderado simétrico
Modelo	MK1	• Tramos de mora	60 días	2	10,88%	0,17%	**0,05%**
	MK2	• Tramos de mora	30 días*	1	**5,31%**	**0,11%**	0,07%
	MK3	• Tramos de mora • Tiempo mora[a]	60 días	1	5,74%	0,20%	0,11%
	MK4	• Tramos de mora • Tiempo mora[a]	30 días*	1	6,11%	0,38%	0,27%

Nota. [a]Tiempo mora refiere al tiempo consecutivo en mora que se considera como variable de estado. Fuente: Elaboración propia.

Para efectos de determinar un mejor modelo de pronóstico, el error absoluto cumple un papel importante en la elección de la mejor ventana de ajuste, pero no permite comparar directamente entre ellos debido al número de estados diferente en cada uno de los modelamientos. De todas formas, para tres de los cuatro casos presentados, esta ventana se sitúa en $n = 1$, siendo la única excepción la cadena MK1, que se desempeña de mejor manera en ventanas de tiempo que consideran dos periodos ($n = 2$).

Si para elegir el mejor modelo en términos de pronóstico, solo se toma en cuenta el error ponderado promedio como lo hiciesen trabajos anteriores (Roco Benavides, 2010; Segovia Riquelme, 2005), la cadena con mejor poder predictivo la constituye el modelo MK2, seguida por el MK1, MK3 y MK4. No obstante, como se mencionó esta métrica tiene la desventaja de no poder estimarse cuando el error se produce en pronosticar transiciones que no han ocurrido. Por estos motivos, se considera un tercer tipo de error, en el cual la ponderación de los errores se efectúa tanto sobre el valor pronosticado y observado de cada una de las transiciones, por tanto, se toman en consideración en la estimación casos donde se pronosticarían transiciones donde no las había. Así, el mejor de los resultados es obtenido con la cadena MK1, mientras que el modelo más consistente en todos los errores resulta el MK2. De esta forma la propuesta es que debiese tomarse como mejor modelo el MK2 dada su buena performance y consistencia.

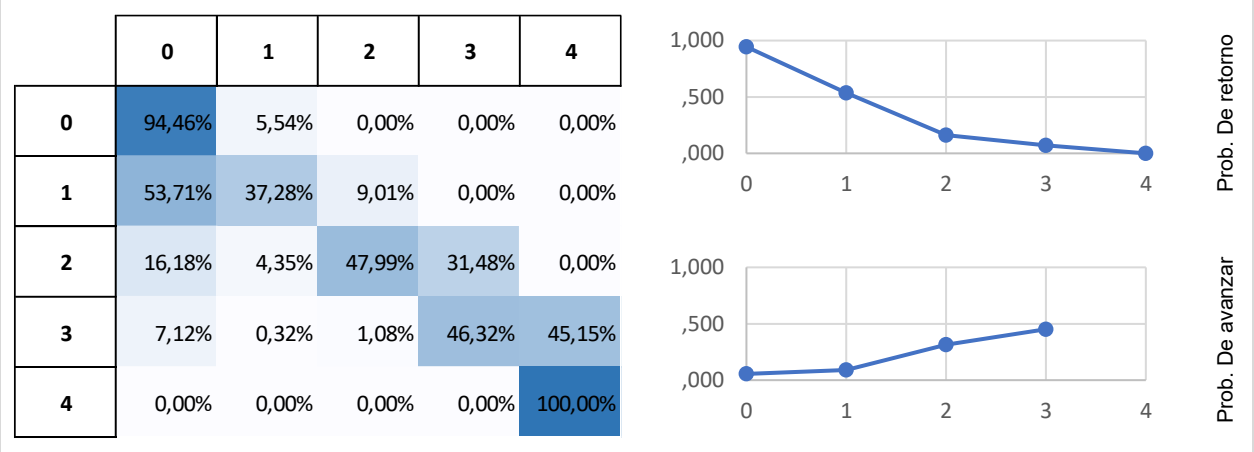
5.4.2.1.2 Transiciones representantes

Para cada modelo en su mejor ajuste, i.e. en su ventana de tiempo óptima (ver Tabla 13), se analiza su matriz representante o grafo asociado sobre los distintos clústeres. Así, se describe el comportamiento de cada grupo para complementar la segmentación anterior con las transiciones de mora y el impacto de las variables de estado sobre la misma. Estas migraciones son abordadas desde la perspectiva de tres tipos fundamentales: las transiciones de retorno, aquellas que implican una mejora en la situación de morosidad

del cliente, i.e. todas las migraciones que desde un tramo de mora superior tiene por destino uno de menor gravedad. Las migraciones de mantención que conservan el tramo de morosidad desde el cual se origina la migración, implicando que el cliente debe abonar lo suficiente para solventar los montos con vencimiento más antiguo y finalmente las transiciones de avance, i.e. todas aquellas que involucran que el cliente pase a un tramo de morosidad superior (mayor gravedad), por lo que empeora su situación como deudor.

5.4.2.1.2.1MK1 Matrices de transición representantes

Como resultado del ajuste de la cadena MK1, cuya variable de estado solo considera los tramos de morosidad en una granularidad de 60 días, se comienza por exhibir la matriz representante de la cartera completa como si no hubiese existido segmentación alguna.



	0	1	2	3	4
0	94,46%	5,54%	0,00%	0,00%	0,00%
1	53,71%	37,28%	9,01%	0,00%	0,00%
2	16,18%	4,35%	47,99%	31,48%	0,00%
3	7,12%	0,32%	1,08%	46,32%	45,15%
4	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 43. Matriz representante cartera, modelo MK1
Nota. Fuente: Elaboración propia.

De la Figura 43 destaca las semejanzas en cuanto a conclusiones entre el análisis descriptivo y el estudio de la matriz representante de toda la cartera. Así, se observa que a medida que se avanza en tramo de morosidad, la probabilidad de retornar a una situación sin deudas disminuye. Esta situación se expresa en el vector formado por la primera columna de la matriz de probabilidades de transición, $\Pi_{\bullet,0} = (\pi_{00}, \pi_{10}, \dots, \pi_{40})$, el cual resulta decreciente en cada una de sus componentes. Por el contrario, las posibilidades de avanzar hacia tramos de mayor gravedad sufren un crecimiento sostenido a medida que se acercan al castigo, traduciéndose en que el vector formado por la supra diagonal, i.e. el conjunto de elementos directamente encima de los que comprenden la diagonal de la de la matriz de probabilidades de transición, es creciente en sus componentes $(\Pi_{i,i+1} = (\pi_{01}, \pi_{12}, \dots, \pi_{34}),$ con $i \in \{1, \dots, n°_{nodos} - 1\})$. Con los elementos presentados, a continuación se muestra como difieren las matrices de transición entre los grupos conformados, observándose como los comportamientos varían en gran medida entre un segmento y otro.

	0	1	2	3	4
0	94,34%	5,66%	0,00%	0,00%	0,00%
1	66,02%	16,95%	17,03%	0,00%	0,00%
2	60,77%	2,28%	25,11%	11,83%	0,00%
3	70,48%	0,24%	0,49%	22,20%	6,59%
4	0,00%	0,00%	0,00%	0,00%	100,00%

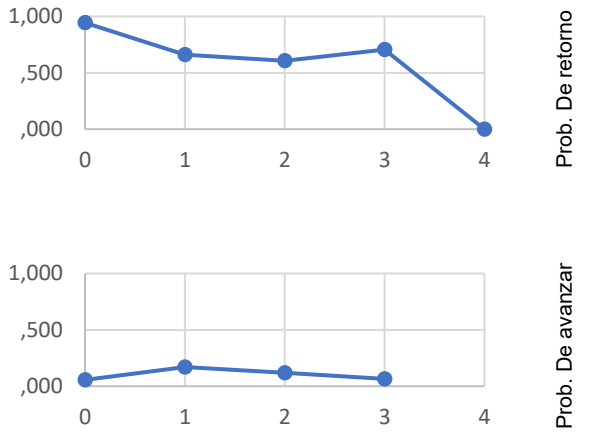


Figura 44. Matriz representante clúster 1, modelo MK1

Nota. Fuente: Elaboración propia.

	0	1	2	3	4
0	38,15%	61,85%	0,00%	0,00%	0,00%
1	11,30%	11,57%	77,13%	0,00%	0,00%
2	10,31%	2,22%	10,27%	77,19%	0,00%
3	10,45%	0,52%	0,64%	58,41%	29,98%
4	0,00%	0,00%	0,00%	0,00%	100,00%

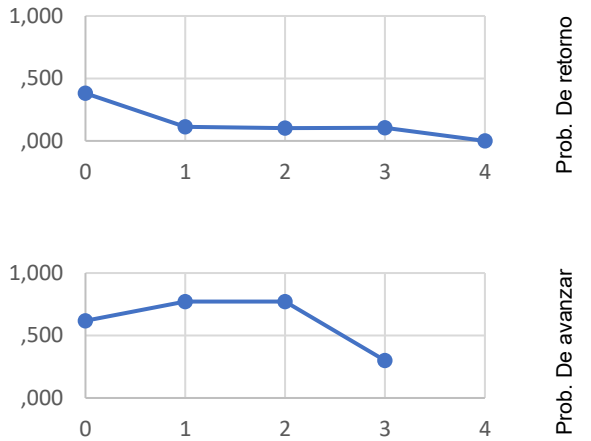


Figura 45. Matriz representante clúster 2, modelo MK1

Nota. Fuente: Elaboración propia.

	0	1	2	3	4
0	57,76%	42,24%	0,00%	0,00%	0,00%
1	45,25%	29,05%	25,69%	0,00%	0,00%
2	28,85%	4,26%	35,45%	31,45%	0,00%
3	43,51%	1,77%	3,45%	42,33%	8,94%
4	0,00%	0,00%	0,00%	0,00%	100,00%

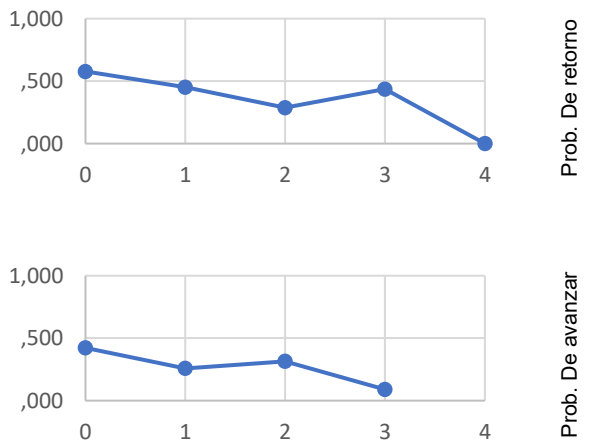


Figura 46. Matriz representante clúster 3, modelo MK1

Fuente: Elaboración propia.

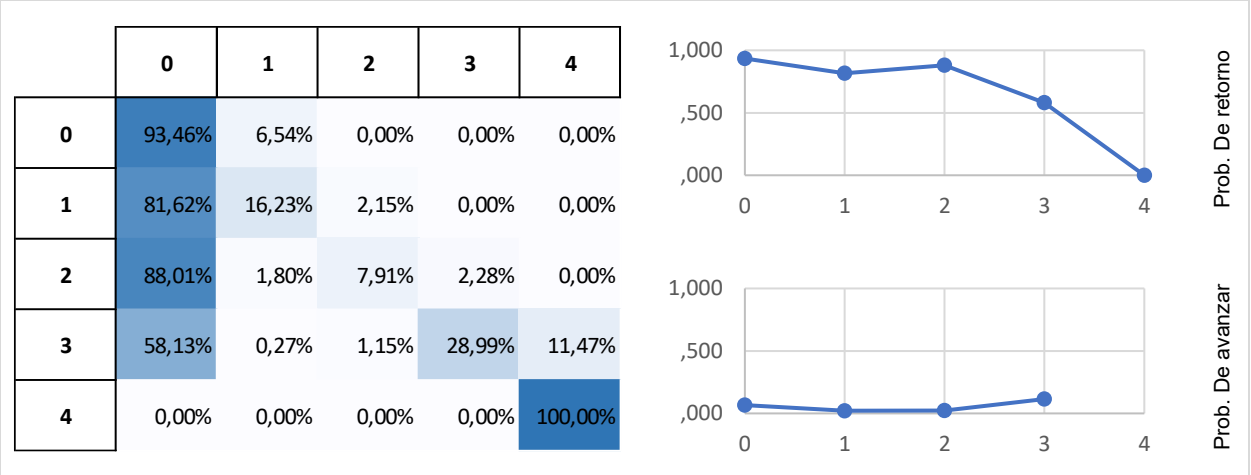


Figura 47. Matriz representante clúster 4, modelo MK1

Nota. Fuente: Elaboración propia.

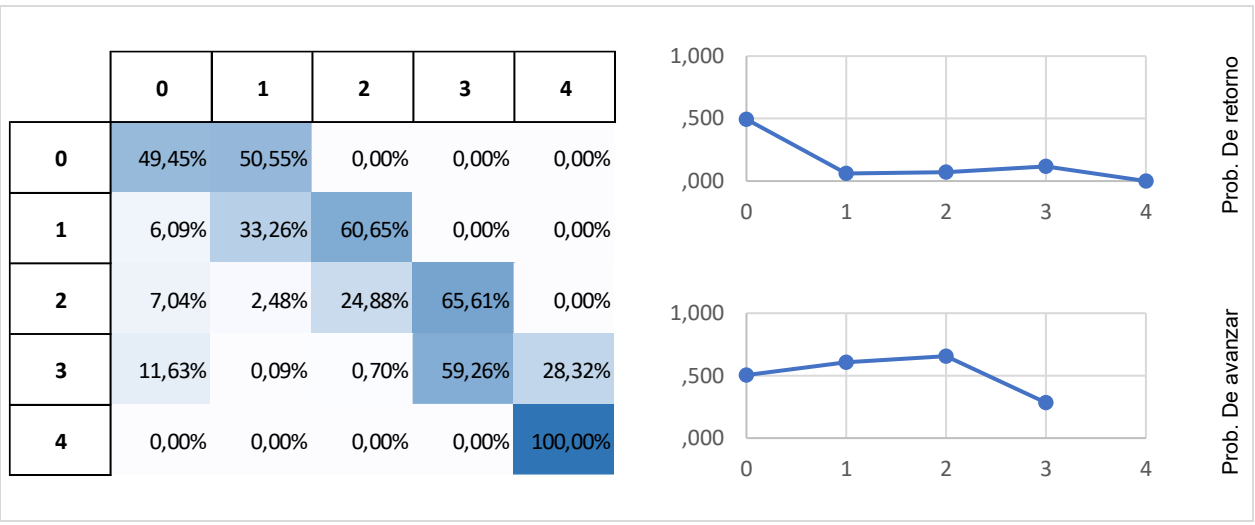


Figura 48. Matriz representante clúster 5, modelo MK1

Nota. Fuente: Elaboración propia.

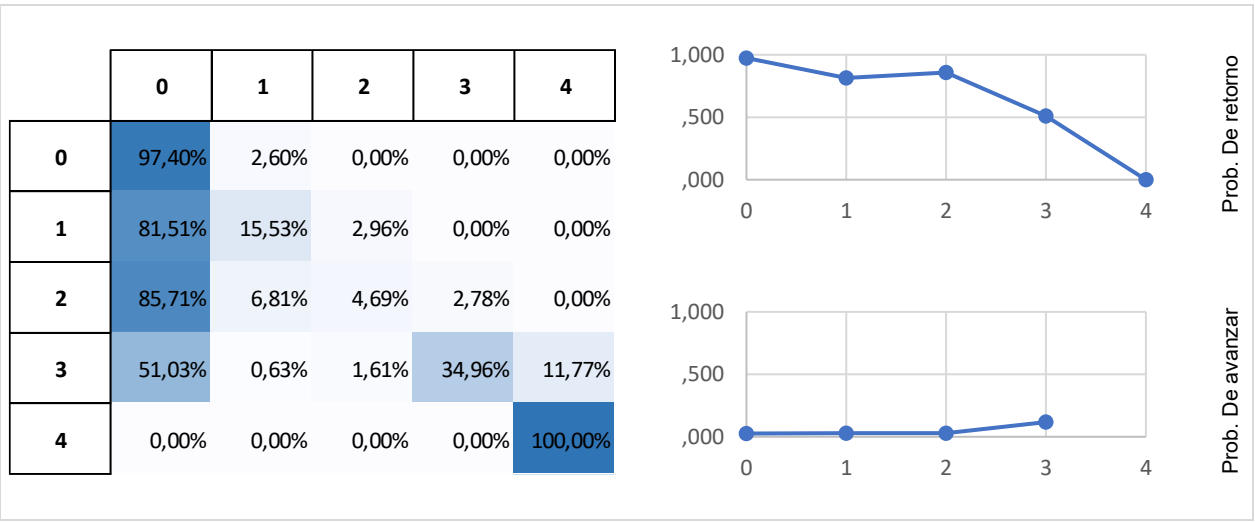


Figura 49. Matriz representante clúster 6, modelo MK1

Nota. Fuente: Elaboración propia.

Las figuras presentadas dan cuenta del comportamiento de cada uno de los segmentos, al ser representados por el modelo con variable de estado tramo de morosidad y granularidad 60 días. Los resultados generales, ya discutidos dan cuenta de un comportamiento similar al observado en el análisis descriptivo. No obstante, al diferenciar por cada uno de los segmentos, se pueden apreciar los diferentes perfiles de comportamiento de morosidad. De esta forma el análisis de estos resultados es desarrollado de acuerdo con los principales tipos de transiciones entre estados.

a) Para los segmentos de mejor comportamiento, comenzando por las probabilidades de retorno tanto para el clúster 4 (Figura 47) como el clúster 6 (Figura 49) se puede apreciar que estas decrecen conforme se avanza a un estado de mayor gravedad en morosidad. No obstante, al alcanzar el periodo comprendido entre los 61-120 días, experimentan un máximo local que rompe esa tendencia.

   Respecto a las probabilidades de avanzar de los clientes de alta transaccionalidad y buen comportamiento, según la caracterización resultante del proceso de clusterización; los grupos 4 y 6 exhiben un leve crecimiento a medida que estos segmentos se acercan a la situación de castigo. Aun así, las chances de que empeoren su situación son comparativamente bajas en relación a otros clústeres considerados en el análisis (ver Figura 47 y Figura 49 respectivamente).

b) Para los grupos de clientes riesgosos, i.e. el segmento 2 (Figura 45) y el segmento 5 (Figura 48), las probabilidades de retorno experimentan un decrecimiento sostenido a medida que aumenta el deterioro de la situación del cliente. No obstante, las chances correspondientes al grupo 2 resultan más agresivas en cuanto a desaceleración y nivel base desde el que inician, pues este segmento corresponden a los clientes asociados a ingresos más acotados, por lo que presentan los niveles de riesgo más altos. Así, los hechos exhibidos son indicador de que para los grupos de mayores niveles de riesgo, mientras más avanzada la morosidad menos chances existen de recuperarlos.

   Por otro parte, las probabilidades de avance resultan consistentemente altas, con este tipo de clientes; los de mayor riesgo. Así, la Figura 45 y Figura 48 muestran también las probabilidades para el segmento 2 y 5 respectivamente. El crecimiento sostenido de las mismas refleja la ya mencionada incremental dificultad de recuperar este tipo de clientes. No obstante, como resultado de la agregación de los tramos de morosidad, existe una sub-ponderación de estas probabilidades que se hace visible con el rompimiento de la tendencia de crecimiento hacia el final del camino.

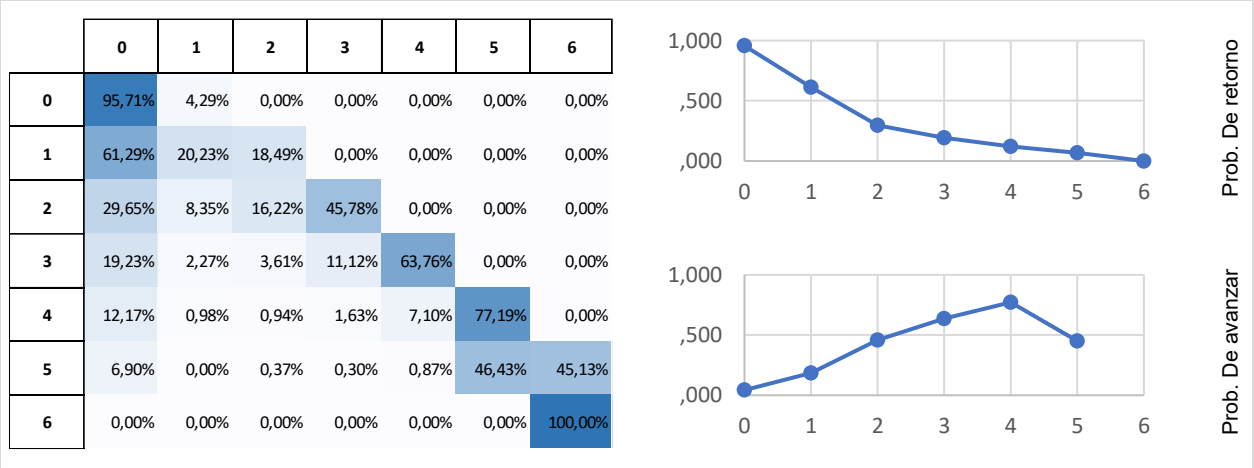
c) Para los clientes con una relación menos estrecha con el negocio, es decir, aquellos con compras esporádicas y que pueden haber presentado episodios de mora moderados (segmento 1, Figura 44 y segmento 3, Figura 46), también se verifica el comportamiento decreciente en sus probabilidades de retorno, que al igual que en los segmentos altamente transaccionales experimentan un repunte

antes de caer, diferenciándose en que este ocurre de manera más tardía; justo antes del castigo.

Finalmente, para este tipo de clientes, los segmentos con una relación menos fuerte con la empresa, las probabilidades de avance junto a presentarse en escalas pequeñas siguen comportamientos menos intuitivos, pues no presentan un crecimiento sostenido al aumentar el deterioro de la deuda. Dado que estos clientes (clúster 1, Figura 44 y clúster 3, Figura 46) tienen relaciones menos estrechas con la empresa, i.e. tienen menos interacciones con el negocio, la hipótesis es que no son conscientes de su situación hasta que esta alcanza algún hito que los hace tomar razón de esta.

5.4.2.1.2.2MK2 Matrices de transición representantes

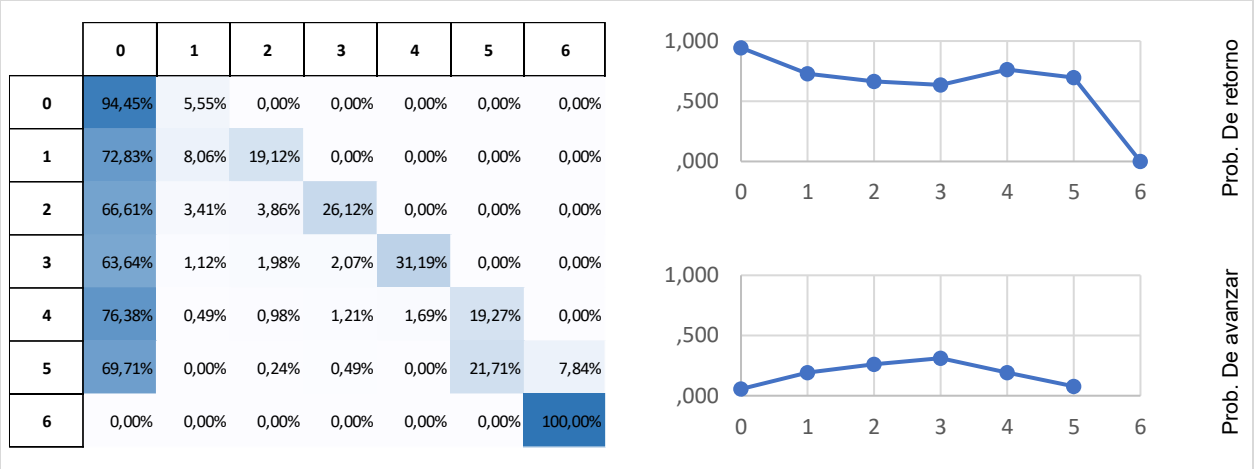
Para el modelo MK2, cuya variable de estado corresponde a los tramos de morosidad en una granularidad de 30 días, sus resultados son abordados en forma seguida. Para ello se comienza por el análisis de la cartera completa, tal y como se muestra en la Figura 50. De esta forma, en primera instancia destaca la consistencia entre el modelo MK2 y el de menor granularidad antes descrito (modelo MK1). No obstante, la gran diferencia radica en la imposibilidad de construir dos tramos de 30 días para el nivel comprendido entre 120-180 días de morosidad. Así, la probabilidad de mantenerse en dicho estado es sobre ponderada y la de avanzar subestimada, rompiendo la tendencia de crecimiento que muestra el gráfico de probabilidades de avance en la Figura 50.



	0	1	2	3	4	5	6
0	95,71%	4,29%	0,00%	0,00%	0,00%	0,00%	0,00%
1	61,29%	20,23%	18,49%	0,00%	0,00%	0,00%	0,00%
2	29,65%	8,35%	16,22%	45,78%	0,00%	0,00%	0,00%
3	19,23%	2,27%	3,61%	11,12%	63,76%	0,00%	0,00%
4	12,17%	0,98%	0,94%	1,63%	7,10%	77,19%	0,00%
5	6,90%	0,00%	0,37%	0,30%	0,87%	46,43%	45,13%
6	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 50. Matriz representante cartera, modelo MK2
Nota. Fuente: Elaboración propia.

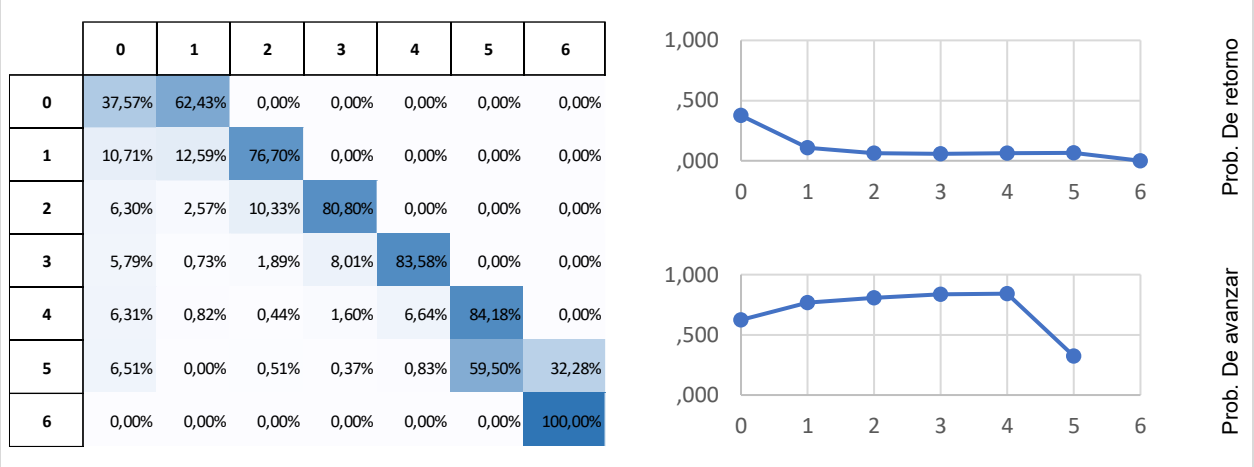
Al igual como se hiciese con el modelo MK1 y como se efectúa para las otras alternativas de modelamiento, la metodología es aplicada a cada uno de los clústeres, de manera de observar y analizar como varían las probabilidades de transición entre un grupo y otro. Así, las figuras siguientes dan cuenta de dicha heterogeneidad.



	0	1	2	3	4	5	6
0	94,45%	5,55%	0,00%	0,00%	0,00%	0,00%	0,00%
1	72,83%	8,06%	19,12%	0,00%	0,00%	0,00%	0,00%
2	66,61%	3,41%	3,86%	26,12%	0,00%	0,00%	0,00%
3	63,64%	1,12%	1,98%	2,07%	31,19%	0,00%	0,00%
4	76,38%	0,49%	0,98%	1,21%	1,69%	19,27%	0,00%
5	69,71%	0,00%	0,24%	0,49%	0,00%	21,71%	7,84%
6	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 51. Matriz representante clúster 1, modelo MK2

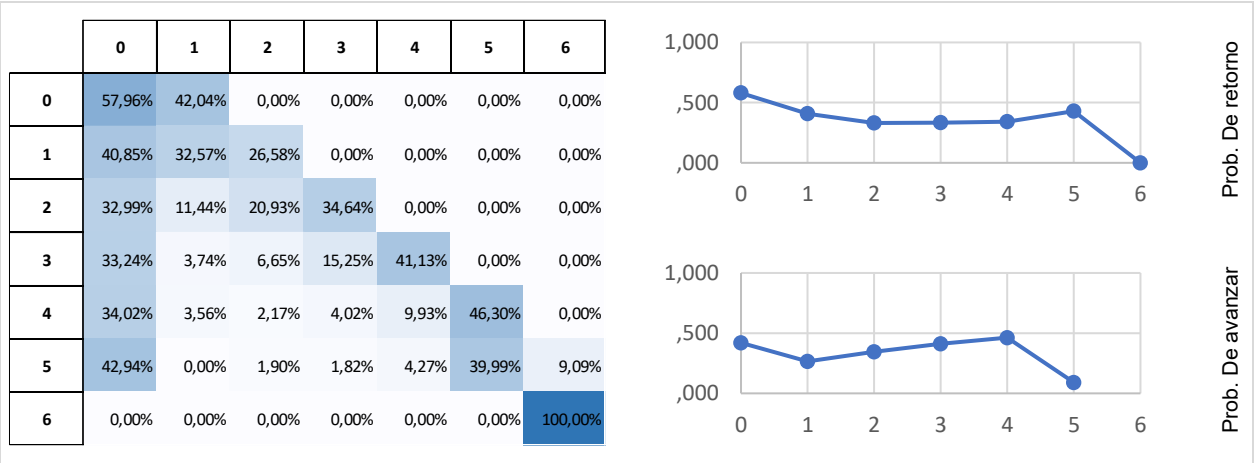
Nota. Fuente: Elaboración propia.



	0	1	2	3	4	5	6
0	37,57%	62,43%	0,00%	0,00%	0,00%	0,00%	0,00%
1	10,71%	12,59%	76,70%	0,00%	0,00%	0,00%	0,00%
2	6,30%	2,57%	10,33%	80,80%	0,00%	0,00%	0,00%
3	5,79%	0,73%	1,89%	8,01%	83,58%	0,00%	0,00%
4	6,31%	0,82%	0,44%	1,60%	6,64%	84,18%	0,00%
5	6,51%	0,00%	0,51%	0,37%	0,83%	59,50%	32,28%
6	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 52. Matriz representante clúster 2, modelo MK2

Nota. Fuente: Elaboración propia.



	0	1	2	3	4	5	6
0	57,96%	42,04%	0,00%	0,00%	0,00%	0,00%	0,00%
1	40,85%	32,57%	26,58%	0,00%	0,00%	0,00%	0,00%
2	32,99%	11,44%	20,93%	34,64%	0,00%	0,00%	0,00%
3	33,24%	3,74%	6,65%	15,25%	41,13%	0,00%	0,00%
4	34,02%	3,56%	2,17%	4,02%	9,93%	46,30%	0,00%
5	42,94%	0,00%	1,90%	1,82%	4,27%	39,99%	9,09%
6	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 53. Matriz representante clúster 3, modelo MK2

Nota. Fuente: Elaboración propia.

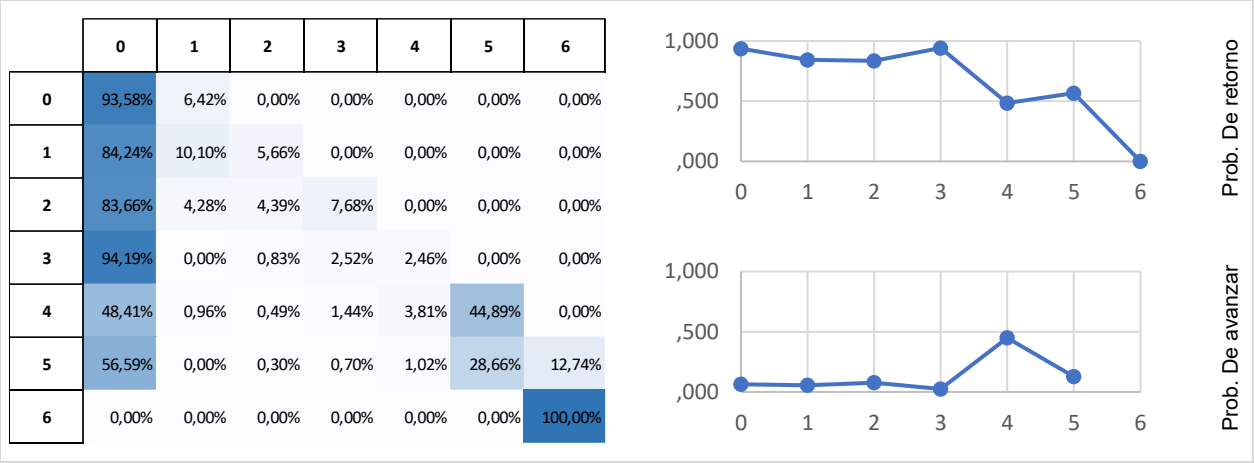


Figura 54. Matriz representante clúster 4, modelo MK2

Nota. Fuente: Elaboración propia.

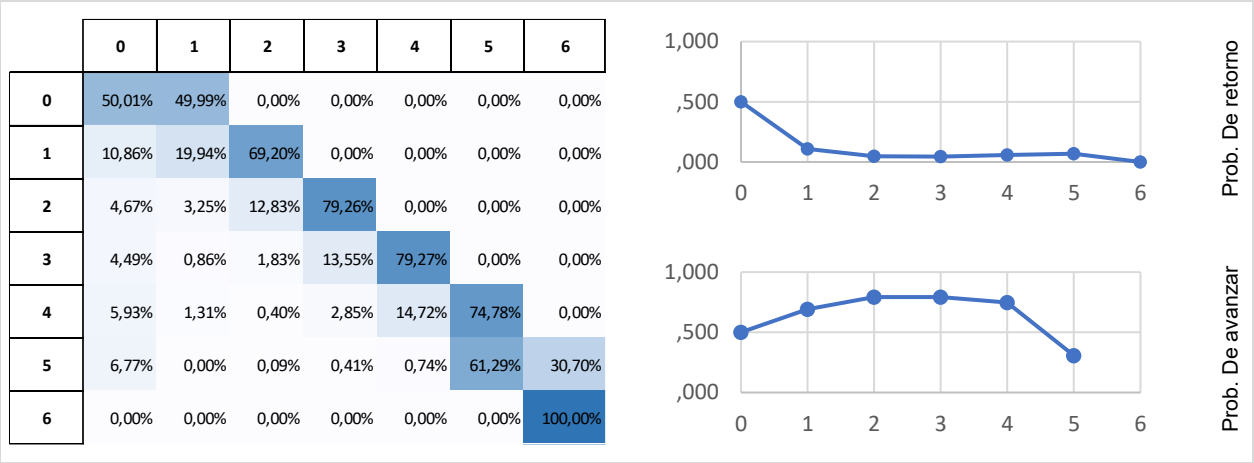


Figura 55. Matriz representante clúster 5, modelo MK2

Nota. Fuente: Elaboración propia.

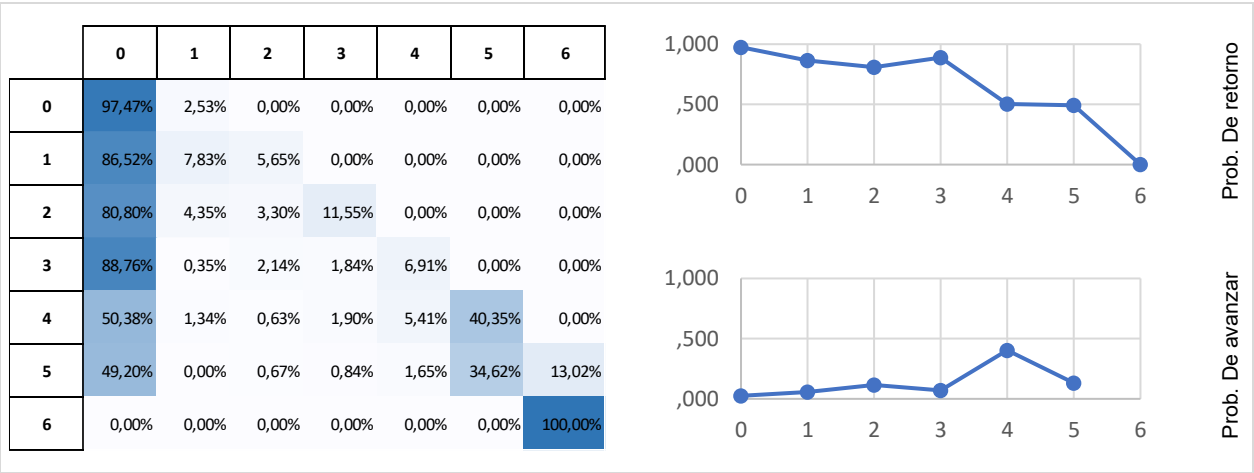


Figura 56. Matriz representante clúster 6, modelo MK2

Nota. Fuente: Elaboración propia.

A partir de los resultados del modelo MK2 sobre cada clúster, se analizan las matrices en base a las transiciones más representativas del comportamiento de morosidad.

a) Para grupos que responden a segmentos de mejor comportamiento, las probabilidades de estas presentan conductas particulares. Tal es el caso del clúster 4 (Figura 54) y clúster 6 (Figura 56), cuyas posibilidades de retorno son notoriamente superiores a las exhibidas por otros grupos. Así, para ambos casos destaca la existencia de un máximo local en la probabilidad de retorno que ocurre en el estado número 3 (60-90 días), rompiendo con la monotonía decreciente a medida que se alcanzan tramos de morosidad de mayor gravedad. Este hecho se atribuye a la mayor valoración de la tarjeta por parte de clientes de mayor transaccionalidad, quienes al pasar el umbral de los 90 días verán su cuenta suspendida.

   En relación con las probabilidades de avanzar, i.e. alcanzar tramos de mora de mayor gravedad. Estas se configuran de manera creciente a mayor gravedad en el tramo de morosidad. Por lo tanto, una vez más se tiene consistencia con el análisis descriptivo desarrollado. No obstante, no debe perderse de vista, las particularidades del último tramo, cuya agregación sesga parte del resultado en ese nivel. Así, para los segmentos 4 y 6 (ver Figura 54 y Figura 56 respectivamente) las probabilidad de avanzar de tramo son consistentemente bajas, aun cuando alcanzan situaciones de morosidad, posibilidades que solo presentan un incremento de consideración en las cercanías del castigo, cuya magnitud continúa siendo mucho menor que en segmentos de mayor riesgo.

b) Para los clientes riesgosos, representados por el clúster 2 (Figura 52) y clúster 5 (Figura 55), ambos grupos exhiben comportamientos decrecientes en la probabilidad de retorno, a medida que se avanza de tramo de morosidad. Sin embargo, las diferencias se expresan en las posibilidades de disminuir su morosidad en estados iniciales. Así, para el segmento asociado a la tarjeta abierta Visa (clúster 5) estas chances son más altas (ver Figura 55 y ANEXO I, Tabla 29).

   Por último, para los segmentos asociados a los peores comportamientos, los clústeres 2 y 5, las probabilidades de avance crecen conforme lo esperado. Así, el único gran diferenciador ente ambos grupos, es que para clientes de menor gasto y presumiblemente menor ingreso, la morosidad se dispara de manera más rápida (ver Figura 52, clúster 2).

c) En relación con los clústeres 1 y 3, Figura 51 y Figura 53 respectivamente. Para los clientes de transaccionalidad esporádica, sus probabilidades de retorno decrecen esperablemente para presentar una leve recuperación hacia el final. Este recupero se atribuye a la cercanía del castigo y las repercusiones de un potencial paso a cobranza judicial.

   Tanto el clúster 1 (Figura 51) como el clúster 3 (Figura 53), corresponden a grupos de clientes de transaccionalidad moderada, que tienen cierta mixtura en sus comportamientos de mora. Así, su probabilidad de avance crecen conforme se alcanzan tramos de mayor gravedad. Sin embargo, para los clientes asociados a la tarjeta cerrada presenta una leve mejora hacia el final del camino al castigo

5.4.2.1.2.3MK3 Representación en grafo matriz representante

Los modelos MK3 y MK4 incorporan una variable de estado adicional para representar el comportamiento de mora de los clientes. Por ello, el análisis de la matriz de transición no es directo como en los casos anteriores. Así, el detalle de las matrices de cada clúster han sido remitidos al ANEXO L. En su defecto, para el análisis principal se ha optado por proceder a partir de la representación en grafo de los modelos constituidos por más de una variable de estado.

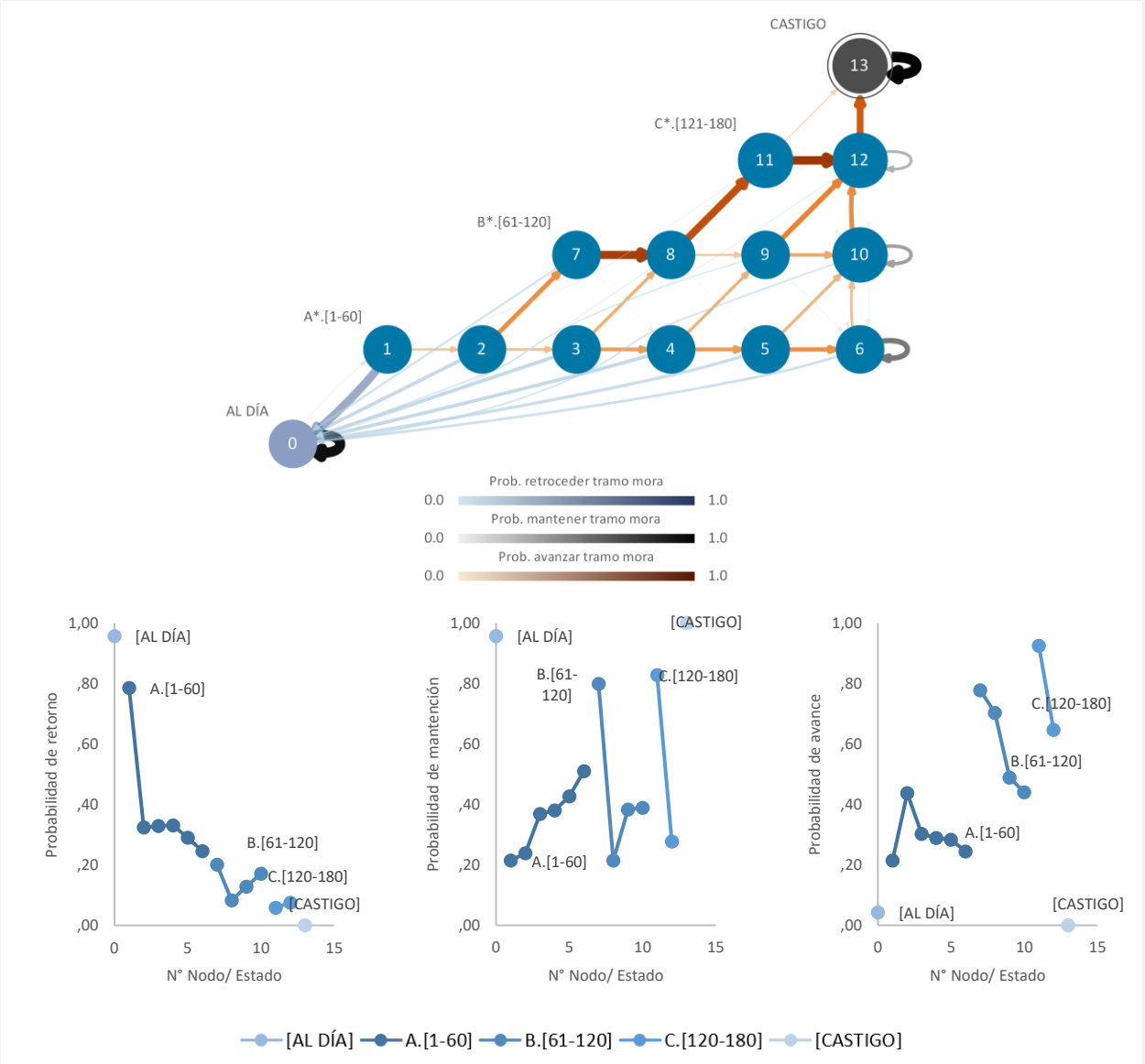


Figura 57. Representación en grafo cartera, modelo MK3
Nota. Detalalles de la matriz de transición en ANEXO L, Figura 94 . Fuente: Elaboración propia.

De la Figura 57, se puede observar la representación en grafo de la matriz de transición para el modelo en cuestión, en el que la variable tramo de morosidad es simbolizada por las filas en el diagrama, mientras que el tiempo consecutivo en mora por las columnas.

Asimismo, se muestran los tres tipos de migraciones más representativas: transiciones de retorno, mantención y avance.

Al igual que en casos anteriores, se comienza por revisar la cartera completa, en la que se destaca la prevalencia de las transiciones de retorno en los tramos inferiores, así como por el incremento en las probabilidades de avance para los niveles de mayor gravedad. Esto es evidenciable en las variaciones en la amplitud de los arcos, y los cambios del gradiente de color. Ambos atributos relacionados de manera proporcional a sus diferencias (mayores amplitudes de arco e intensidad del color representan mayores probabilidades). Esta visualización es empleada en el modelo MK4 de manera análoga.

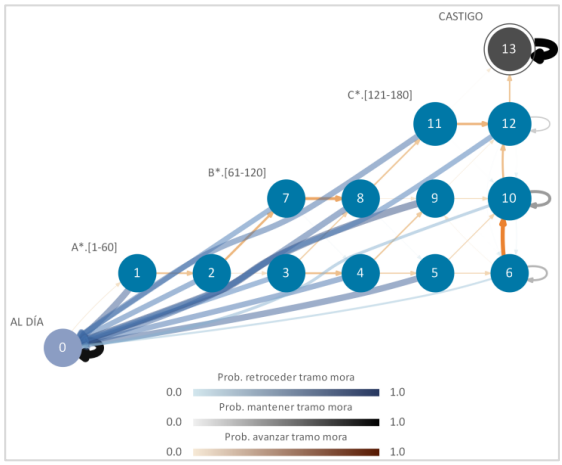


Figura 58. Representación en grafo clúster 1, modelo MK3

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO L, Figura 95 y ANEXO M,Figura 101 respectivamente. Fuente: Elaboración propia.

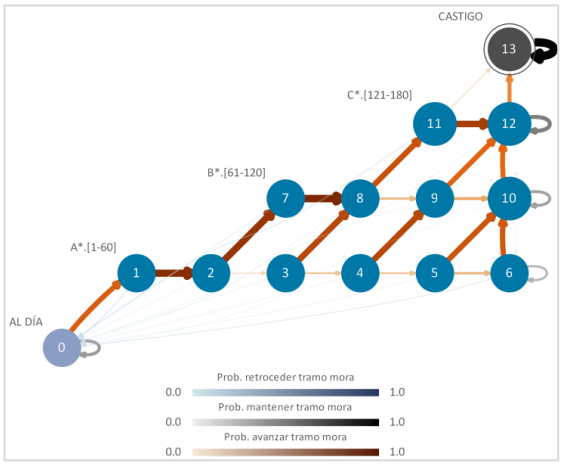


Figura 59. Representación en grafo clúster 2, modelo MK3

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO L, Figura 96 y ANEXO M,Figura 102 respectivamente. Fuente: Elaboración propia.

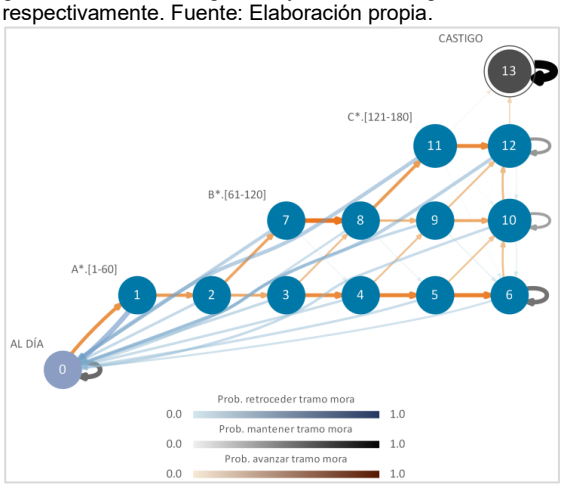


Figura 60. Representación en grafo clúster 3, modelo MK3

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO L, Figura 97 y ANEXO M,Figura 103 respectivamente. Fuente: Elaboración propia.

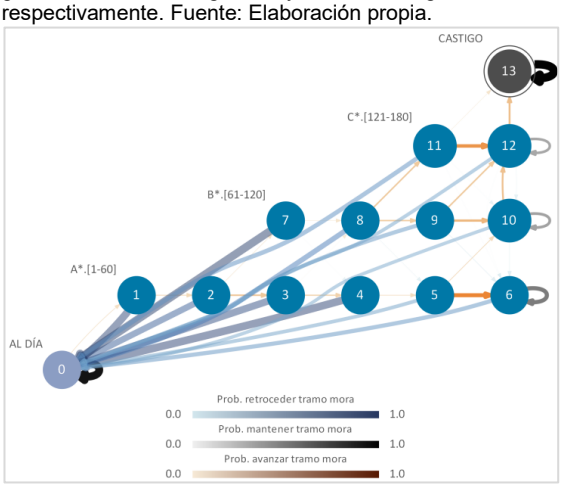


Figura 61. Representación en grafo clúster 4, modelo MK3

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO L, Figura 98 y ANEXO M,Figura 104 respectivamente. Fuente: Elaboración propia.

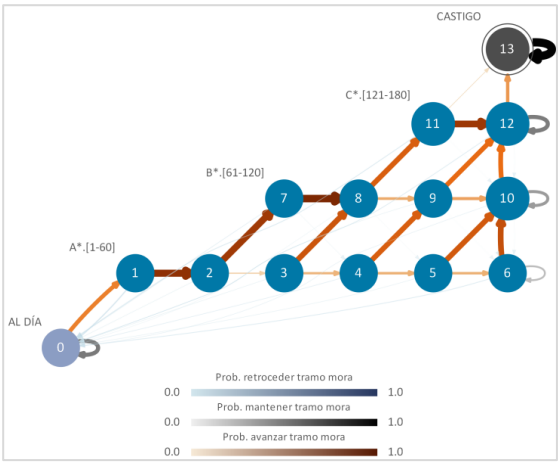


Figura 62. Representación en grafo clúster 5, modelo MK3

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO L, Figura 99 y ANEXO M,Figura 105 respectivamente. Fuente: Elaboración propia.

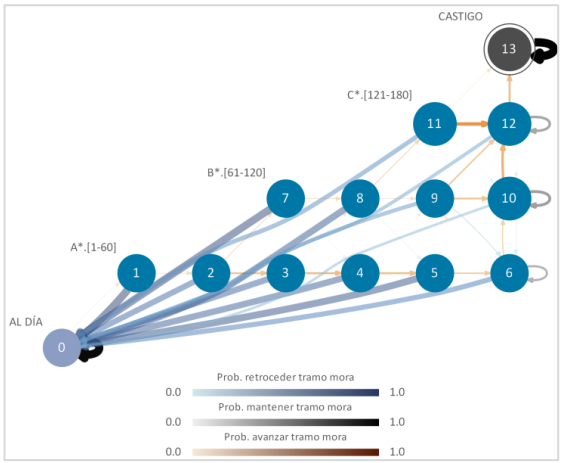


Figura 63. Representación en grafo clúster 6, modelo MK3

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO L, Figura 100 y ANEXO M,Figura 106 respectivamente. Fuente: Elaboración propia.

Tomando como punto de partida los resultados exhibidos, complementados en el ya señalado ANEXO M, se procede con el análisis de acuerdo con los comportamientos de morosidad identificables y con la cadena empleada (modelo MK3).

a) En relación con las probabilidades de retorno, comenzando por los clientes de mejor comportamiento, i.e. los pertenecientes al clúster 4 y 6 (Figura 61 y Figura 63), estos presentan un nivel comparativamente superior a los otros grupos y una tendencia decreciente en el tramo de mora al aproximarse al castigo, que se traduce en mayores posibilidades de recuperarse, así como menos chances de *default*. No obstante, la incorporación del abono al modelo tiene como consecuencia la identificación de un efecto positivo de esta para la morosidad temprana, al incrementar las probabilidades de volver a encontrarse al día, mientras que en mora tardía dicho fenómeno no es tal y es superado por el mayor tiempo en situación de morosidad.

   Respecto a las probabilidades de mantención, estas presentan una relación creciente respecto al número de abonos, que no siendo un indicador de la regularización de la deuda, si dan cuenta de una disposición o esfuerzo de pago. Así, para los grupo 4 y 6, Figura 61 y Figura 63 respectivamente, son las que presentan comportamientos más alejados de la tendencia descrita, pues en un comienzo estas disminuyen, ya que los abonos mejoran las chances de retornar al día, mientras que a más periodos aumenta las probabilidades de mantención.

   Por último, las probabilidades de avance de los clústeres 4 (Figura 61) y 6 (Figura 63), como en todos los modelamientos presentados, resultan crecientes en tramo, pero con niveles acotados respecto a los grupos con peores comportamientos. Por su parte el abono resulta favorable tan solo en morosidad temprana, pues siendo estos los mejores clientes, un tiempo prolongado en mora da cuenta de una situación subyacente más allá de una falta disposición u olvido.

b) Para los grupos de alto riesgo, clúster 2 (Figura 59) y clúster 5 (Figura 62), las conclusiones comienzan por el bajo nivel de las posibilidades de retorno cuando se declara la mora. Sin embargo, el efecto del abono para estos casos resulta especialmente beneficioso, pues permite diferenciar dentro del universo de clientes con problemas, a aquellos con mayor disposición a regularizar su deuda.

   Respecto a las probabilidades de mantención, los grupos riesgosos se ven favorecidos por el abono, tanto en probabilidades de retorno como mantención. Así, para el clúster 2 (Figura 59), estas chances crecen consistentemente. Por su parte, para el grupo 5 (Figura 62) el crecimiento se produce en mora tardía, ya que en tramos tempranos el abono reparte su efecto sobre retorno y mantención.

   El nivel base de probabilidades de avance más alto lo presentan los clientes más riesgosos, el clúster 2 (Figura 59) y clúster 5 (Figura 62), correspondiendo a aquellos grupos con mayores chances de deteriorar su situación. Por otra parte, su nivel de deterioro resulta progresivo al empeoramiento de su situación, por lo que en tramos superiores es más probable que se exacerbe su estado. Sin embargo, donde existen marcadas diferencias con otros grupos, ocurre en el efecto del abono, el cual tiene un mayor peso para este tipo de clientes, quienres disminuyen sus probabilidades de avance conforme efectúan un mayor número de pagos. Así, para dichos segmentos este es un hecho diferenciador dentro de ellos, aspecto que cómo se desarrolla posteriormente se propone accionar.

c) Para las probabilidades de retorno de los clientes de poca transaccionalidad y riesgo moderado, se identifica que para el caso del clúster 1 (Figura 58) existe una disminución a medida que se avanza de nivel de morosidad, que en tramos tempranos es atenuada por el abono, mejorando las posibilidades de estos clientes. Por su parte, para los pertenecientes al grupo 3 (Figura 60), estas chances siguen un decrecimiento estricto sin que el abono modifique la tendencia.

   En relación con sus probabilidades de mantención, para el caso del segmento 1 (Figura 58) y segmento 3 (Figura 60), su evolución resulta similar a la que experimenta el resto de los clientes, siendo la principal diferencia el nivel base en torno al cual se mueven estas probabilidades, que se sitúa entre los dos extremos presentados.

   Finalmente respecto a probabilidades de avance, para el caso de las personas con menor relación con la empresa; compradores esporádicos que no hacen un uso elevado de su disponible, el clúster 1 (Figura 58), mantiene el comportamiento evidenciado por los clientes más responsables, diferenciándose de ellos solo en escala. Así, sus probabilidades de avanzar crecen conforme se deteriora su situación y el abono solo es relevante mientras no haya transcurrido demasiado tiempo. Por otra parte, para los clientes del grupo 3 (Figura 60), al igual que sus pares del grupo 1, presentan probabilidades crecientes en tramo de morosidad, diferenciándose en mayor medida por el efecto del abono, el cual se sostiene en el tiempo, aún en tramos de morosidad avanzados.

5.4.2.1.2.4 MK4 Representación en grafo matriz representante

Al igual que con el modelo MK3, el detalle de las matrices de transición representantes de la cadena MK4 se remite a los apéndices de este trabajo (ver ANEXO N). Sin embargo, el análisis de las migraciones asociadas es profundizado mediante su representación en grafo. Así, se comienza con el análisis de la cartera completa, describiendo sus comportamiento general para luego profundizar de manera específica en cada clúster, grupos que a su vez son abordados en mayor detalle en el ANEXO O.

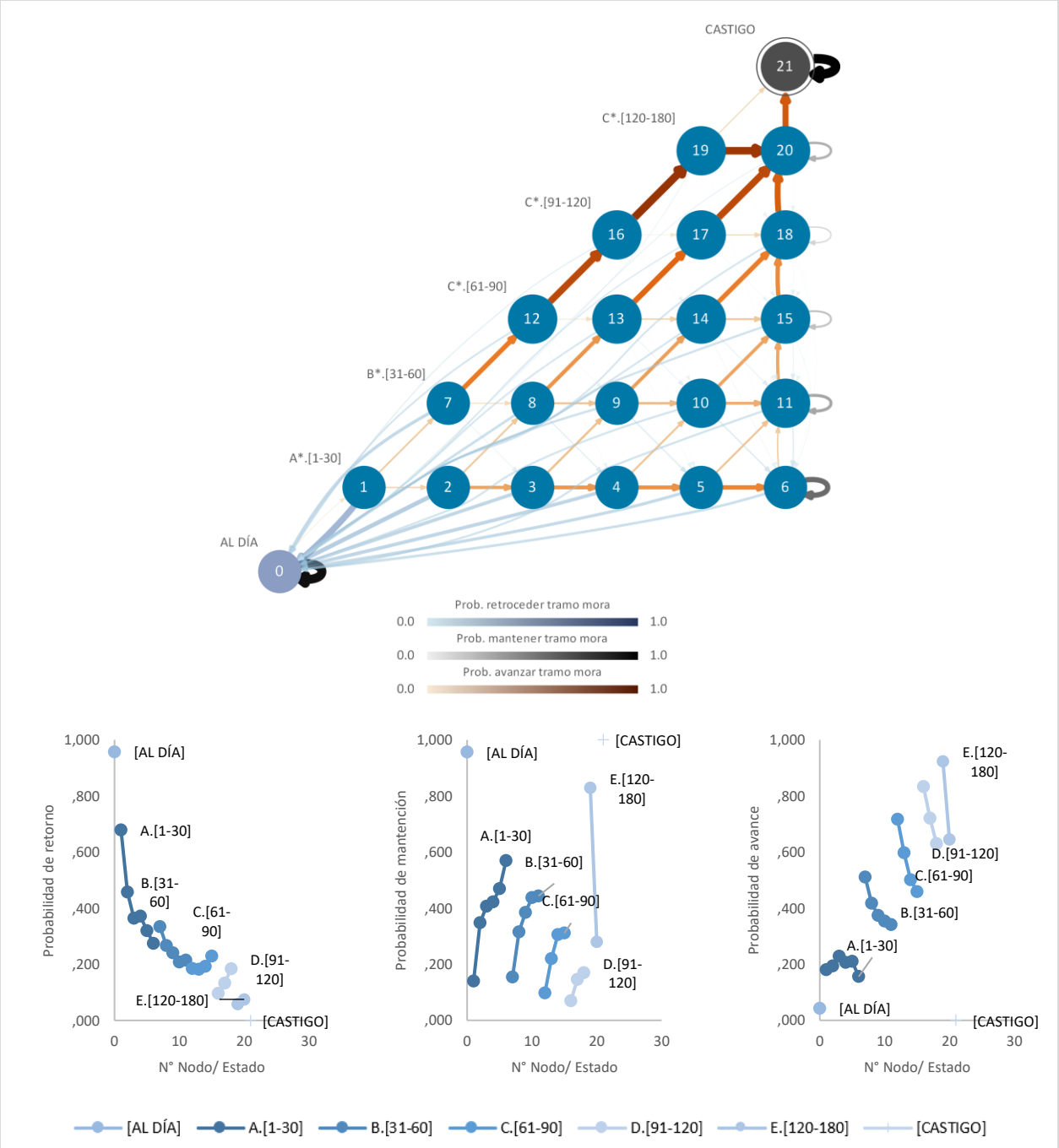


Figura 64. Representación en grafo cartera, modelo MK4

Nota. Detalalles de la matriz de transición en ANEXO N, Figura 107. Fuente: Elaboración propia.

Para efectos de la representación en grafo, las visualizaciones contruidas muestran los tres tipos de transición fundamentales dentro de los diferentes comportamientos de morosidad: las migraciones de retroceso, mantención y avance (Para mayor detalle consultar 5.4.2.1.2 Transiciones representantes). Como se eplicitó en el capitulo anterior, las transiciones descritas son simbolizadas por los colores empleados y la amplitud del arco, de manera que una mayor prevalencia de tonos azulados indica una tendencia hacia transiciones de retroceso, mientras que los anaranjados muestran la prevalencia del avance en tramo de morosidad. Así el análisis siguiente caractiza el comportamiento de los diferentes clústers en base a los señalado.

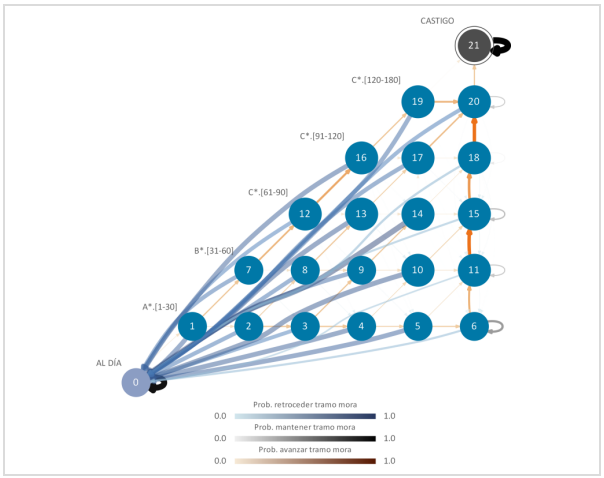


Figura 65. Representación en grafo clúster 1, modelo MK4

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO N, Figura 108 y ANEXO O, Figura 114 respectivamente. Fuente: Elaboración propia.

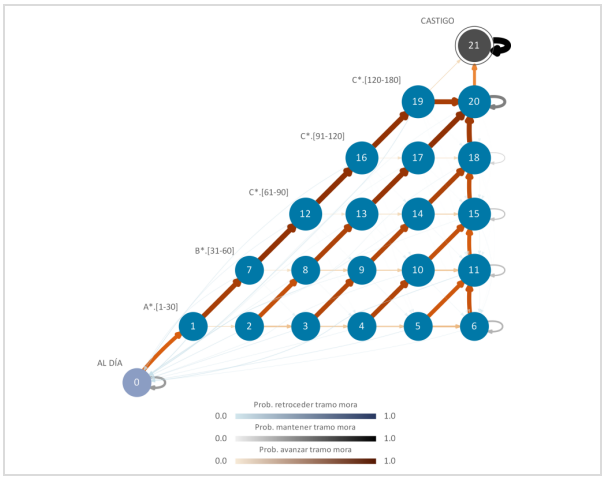


Figura 66. Representación en grafo clúster 2, modelo MK4

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO N, Figura 109 y ANEXO O, Figura 115 respectivamente. Fuente: Elaboración propia.

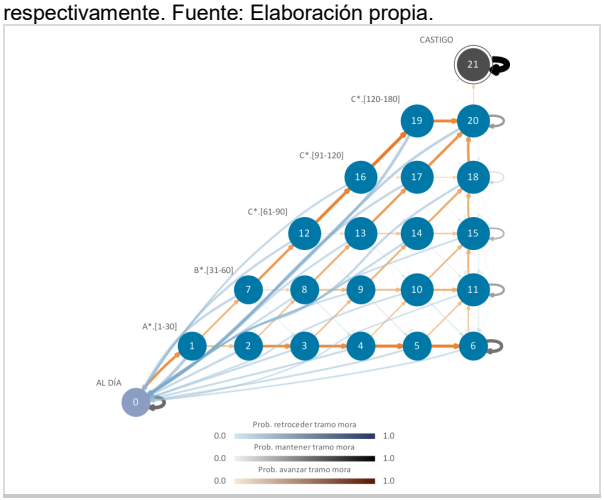


Figura 67. Representación en grafo clúster 3, modelo MK4

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO N, Figura 110 y ANEXO O, Figura 116 respectivamente. Fuente: Elaboración propia.

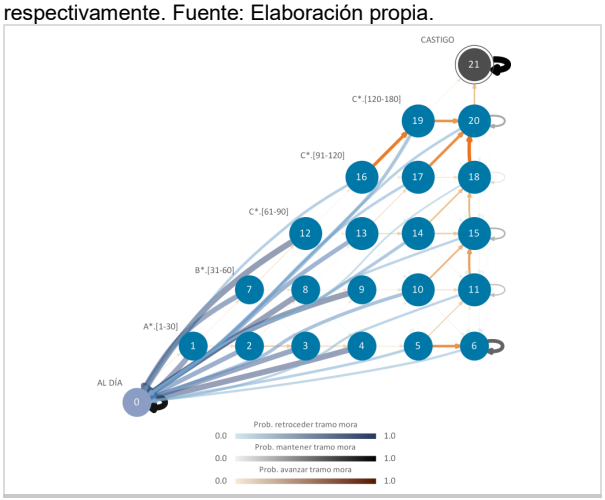


Figura 68. Representación en grafo clúster 4, modelo MK4

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO N, Figura 111 y ANEXO O, Figura 117 respectivamente. Fuente: Elaboración propia.

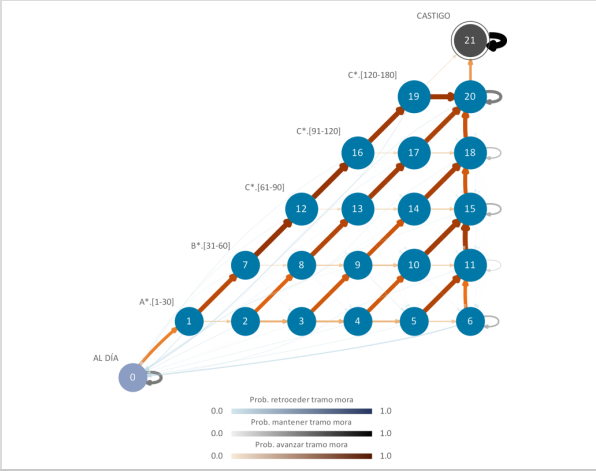


Figura 69. Representación en grafo clúster 5, modelo MK4

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO N, Figura 112 y ANEXO O, Figura 118 respectivamente. Fuente: Elaboración propia.

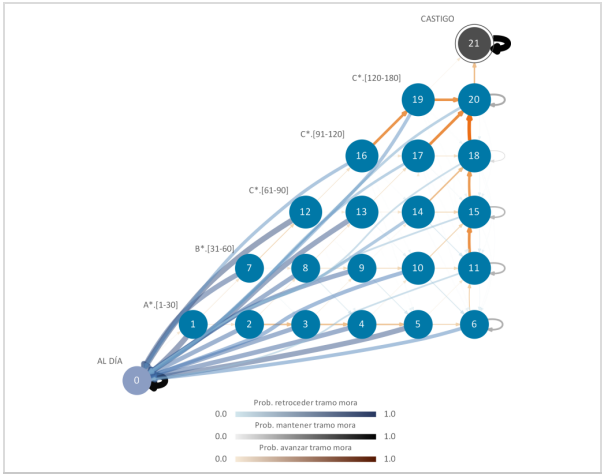


Figura 70. Representación en grafo clúster 6, modelo MK4

Nota. Detalalles de la matriz de transición y representación en grafo en ANEXO N, Figura 113 y ANEXO O, Figura 119 respectivamente. Fuente: Elaboración propia.

a) En relación con las probabilidades de retorno, comenzando por los clientes de mejor comportamiento, i.e. los pertenecientes a los clústeres 4 (Figura 68) y 6 (Figura 70), estos exhiben un comportamiento decreciente en un inicio, para luego sufrir una leve recuperación hacia la mitad de su transición al castigo. Siendo esta condición prevalente en los tramos de morosidad tempranos. De esta forma, controlando por el efecto del tramo de morosidad del cliente, previo a los 90 días, esta no tiene impactos de relevancia en el decrecimiento de la probabilidad de retornar, contrastando con el comportamiento una vez superada la barrera descrita, ya que el descenso esperable es evidenciado. Por su parte, el efecto del abono, tienen incidencia positiva en los primeros tramos de morosidad, en los cuales puede asociarse a una mayor disposición al pago. Situación contraria a la observada en tramos superiores en la que no influye en el mejoramiento de las probabilidades Presumiblemente se ha atribuido a una mayor contribución de la imposibilidad de cancelar por sobre una falta de disposición.

   En relación con las probabilidades de mantención, estas son relativamente homogéneas entre los distintos segmentos de morosidad, siendo el elemento diferenciador la escala en la que se presentan estas chances. Así, para el clúster 4 (Figura 68) y clúster 6 (Figura 70), el efecto del abono es creciente a medida que se efectúan un mayor número de pagos, situación que ocurre hasta alcanzar un máximo, para luego decrecer. El nivel de ambos grupos resulta el más alto de los diferentes perfiles, pudiendo observarse en última instancia la ya mencionada sobre ponderación resultante de la existencia del tramo [121-180] días mora.

   Sobre las probabilidades de avanzar, de manera global se aprecia un crecimiento a medida que se tienen tramos de morosidad de mayor gravedad y una disminución conforme se abona lo suficiente para mantener el nivel. Así, para los segmentos de buen comportamiento o alta transaccionalidad, los clústeres 4 y 6 (Figura 68 y Figura 70 respectivamente), todos ellos exhiben una baja en las

probabilidades de avance hasta alcanzar un mínimo hacia la mitad del periodo, para luego experimentar un crecimiento sostenido hasta el castigo.

b) Para los segmentos de alto riesgo e historial de morosidad; el clúster 2 (Figura 66) y clúster 5 (Figura 69), cabe mencionar el menor nivel base en el que se sitúan las probabilidades de retorno una vez declarada la situación de mora. Por su parte, el efecto de abono tiene una mayor prevalencia en los tramos tempranos, en cuanto a impacto en la disminución de las chances de disminuir el nivel de incumplimiento. De esta forma, el pago parcial de la deuda se sitúa como indicador de disposición a pagar y diferenciador de un subgrupo dentro de aquellos clientes con comportamiento más riesgosos (score de riesgo y número de episodios mora).

   Como se mencionó, las probabilidades de mantención son similares entre los distintos segmentos de morosidad, diferenciándose principalmente en escala. Así, para el clúster 2 (Figura 66) y clúster 5 (Figura 69), al igual que con los grupos anteriores el efecto del abono es creciente a medida que se tienen más pagos, destacándose de los otros perfiles de clientes por el *nivel* base más bajo en el que se sitúan estas probabilidades.

   Finalmente, para los segmentos de peor comportamiento, los clústeres 2 y 5 (Figura 66 y Figura 69), son los más favorecidos en la medida que se abona a la deuda, pues sus probabilidades de avanzar de tramo disminuyen conforme abonan más a su deuda, siendo el efecto más fuerte en el grupo asociado a montos de menor cuantía (clúster 2). No obstante, a pesar de esta situación, siguen siendo los grupos que por nivel base tienen las mayores probabilidades de avance.

c) En el caso de los segmentos con comportamientos moderados en riesgo, morosidad y transaccionalidad, para el caso de los clientes pertenecientes al clúster 1 (Figura 65), sus probabilidades de retorno resultan crecientes conforme se efectúan un mayor número de pagos. Sin embargo, existe un cambio de tendencia importante al alcanzar el nivel de 6 o más periodos consecutivos. Esta situación es atribuible a la agregación de todas las variables de estado que son resumidas en dicho nodo. Por el contrario para el caso del clúster 3 (Figura 67), es decir, aquellos clientes con clientes de historia mora moderada, pero con mayor transaccionalidad, el abono no tiene mayores impactos sobre la probabilidad de retorno, por lo que esta chance decrece aun cuando se ha pagado parcialmente la deuda, siendo la única excepción el tramo final que por limitaciones de datos no pudo ser aperturado en la misma granularidad que el resto.

   Respecto a probabilidades de mantención, la similitud entre los distintos segmentos tiene como principal diferencia la escala. Así, para el clúster 1 (Figura 65) y clúster 3 (Figura 67) se presenta un aumento en las probabilidades de mantención, creciente en el número de abonos, cuyo nivel base se sitúa entre los grupos de mejor comportamiento y aquellos asociados a más riesgo.

Finalmente, en relación con las probabilidades de avance, el comportamiento entre los dos grupos representantes de las conductas de riesgo moderado y transaccionalidad esporádica no es uno solo. Así, para los clientes del clúster 1 (Figura 65) , estos exhiben una disminución en las probabilidades de avance hasta alcanzar un mínimo en la mitad del periodo, para luego experimentar un crecimiento sostenido hasta el castigo. Por el contrario, para el caso de los clientes del segmento 3 (Figura 67), el efecto de los tramos de morosidad es el esperado, incrementando las probabilidades de avanzar, a medida que se tiene un tramo superior. Al mismo tiempo estas chances mantienen una relación inversa con los abonos, que disminuyen las posibilidades de empeorar la situación del cliente.

5.4.2.2 Nivel de cliente

Como se explicitó, el pronóstico a nivel de cliente es abordado con la combinación de la lógica difusa con los modelos de Markov. Para ello se aprovecha los niveles de pertenencia de cada observación a los clústeres obtenidos. Así, la ponderación de estos valores sobre las matrices de transición de cada grupo, permiten definir una matriz única para cada cliente. Este procedimiento es ejemplificado en la Figura 71. En ella es posible apreciar la construcción de la matriz de un cliente j, considerando las probabilidades de transición que se corresponden con los clústeres en sus matrices Π_k y el nivel de pertenencia del cliente a cada uno de estos ($u_{i,j}$).

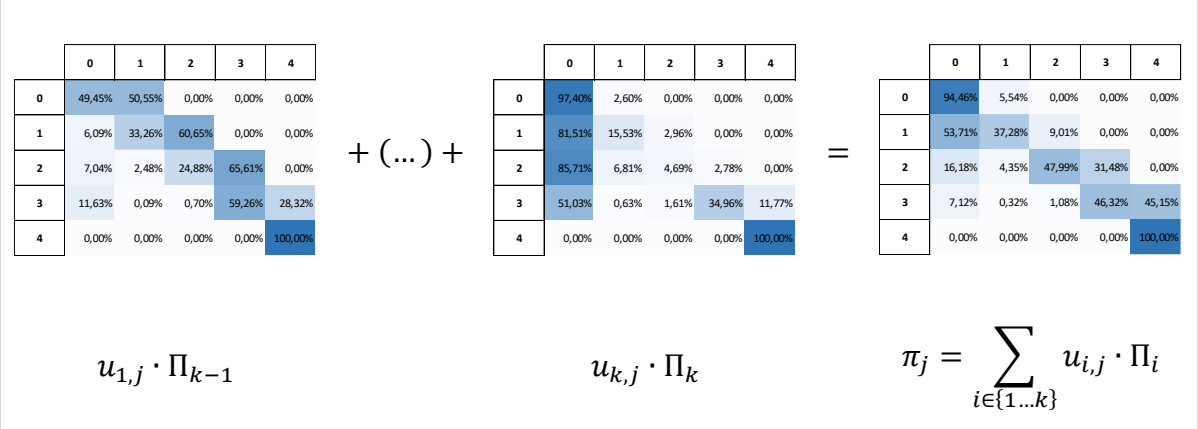


Figura 71. Ejemplo de construcción matriz de transiciones a nivel de cliente
Nota. Fuente: Elaboración propia.

El procedimiento observado en la Figura 71 es repetido para cada uno de los clientes considerados en el estudio, con cada uno de los modelos propuestos. De esta forma se hace posible comparar el impacto de las variables de estados sobre los pronósticos obtenidos.

En relación con la metodología de evaluación, al representarse el avance entre los diferentes estados como un problema de clasificación, se emplean las métricas tradicionales para este tipo de tareas: *precision*, *recall* y *F1-score*. Para ello se consideran las modificaciones necesarias para manejar las múltiples clases (para mayor detalle ver 3.8.2 Evaluación a nivel de cliente). Así, la Tabla 14 presenta los resultados

correspondientes a dos estrategias diferentes, pues junto con testear el método propuesto en la metodología, se compararía lo que habría sucedido en el caso de emplear los resultados de grupo para pronosticar a nivel de cliente. Nótese que los indicadores presentados son calculados sobre una muestra de testeo, la cual considerando el entrenamiento con los comportamientos experimentados por los clientes en 2017, se sitúa a comienzos de 2018 para evaluar los dos primeros meses como set de testeo (ver Figura 30).

Tabla 14: Resultados métricas de ajuste modelos de Markov a nivel de cliente

		Métricas equiponderadas			**Métricas ponderadas por N° de transiciones**		
Método de Pertenencia	**Modelo Markov**	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1-score**
Hard Clustering	MK2	0,59	0,61	0,56	0,84	0,90	0,87
Soft Clustering	MK2	0,53	0,45	0,48	0,83	0,90	0,86

Nota. Fuente: Elaboración propia.

De la Tabla 14, se observa que contra intuitivamente el mejor pronóstico resulta de usar los resultados de grupo para dicha tarea, aunque la diferencia es menor. Se debe notar que al equiponderar la performance esta se ubica en torno al 50% en todas las métricas. Dado que el problema a resolver es multiclase, este valor es suficiente para desempeñarse mejor que el pronóstico aleatorio. Así, el mejor pronóstico lo obtiene el *hard clustering*.

Si se analizan las métricas de performance ponderadas, el rendimiento se ve notoriamente incrementado para los mismos parámetros. No obstante, esto se debe que los mejores resultados del modelo se concentran en la capacidad de anticipar transiciones extremas, como el paso al castigo o el retorno y salida de la situación al día; las migraciones más frecuentes. Dado que el modelo es menos impreciso en las transiciones menos habituales, su desempeño se ve favorecido al evaluarse de esta forma.

Por los motivos presentados, como se indica en los pasos futuros y recomendaciones, se propone que para efectos de llevar este modelo a una situación productiva es ineludible el mejorar su performance.

5.5 Evaluación y despliegue

La revisión de los modelos propuestos con métricas acordes a cada metodología es una tarea en la que ya se ha profundizado en este trabajo. No obstante, resulta relevante desarrollar una perspectiva de negocio sobre estos resultados, pues como indican los objetivos de esta memoria, el mejorar la situación de la empresa es la finalidad principal.

Desde una a perspectiva general, la segmentación de clientes permitió identificar grupos con comportamientos de morosidad heterogéneos, diferencias que son evidenciadas al ajustar las cadenas sobre estos clústeres. Así, para los resultados a nivel de grupo, se consiguen buenos niveles de pronóstico, al mismo tiempo que se capturan relaciones existentes entre las variables de estado, la morosidad y los atributos que definen cada segmento. Por el mismo motivo, las recomendaciones de negocio consideradas en los objetivos de la memoria son abordadas a partir de ellos.

De los resultados de pronóstico a nivel de cliente, a pesar de que ponderados por el número de transiciones alcanzan niveles de performance cercanos al 80%, no es claro que estos sean aceptables. El hecho de que estos valores respondan a un mejor nivel de pronóstico en migraciones extremas y a una poca capacidad predictiva en las intermedias, termina por condicionar que la recomendación sea el no llevar a producción este modelo, sin antes de completar alguno de los trabajos futuros propuestos.

Respecto a las propuestas comerciales, como se mencionara, estas responden a los resultados del pronóstico a nivel de grupo, y al aprovechamiento que estos pueden tener para el cumplimiento del objetivo general, la generación de acciones que propendan a la disminución de los niveles de morosidad. Estas propuestas son recogidas en el capítulo final de este trabajo, el cual se encuentra destinado a las conclusiones y recomendaciones.

Finalmente, antes de proceder con las conclusiones del trabajo, se detalla la evaluación de este desde una perspectiva económica y de los beneficios potenciales, que la aplicación de esta memoria puede tener para el negocio.

5.5.1 Evaluación económica

Para efectos de estimar el potencial económico de esta memoria, se retoma una de las ideas fuerza presentadas en la justificación del trabajo: los beneficios de la disminución de los niveles de morosidad y sus efectos sobre los costos asociados al problema.

En el caso de una hipotética disminución de los niveles de mora, este hecho se traduciría de manera directa en un recupero mayor y una menor cantidad de clientes castigados. De esta forma, las pérdidas asociadas a deudas incobrables se reduciría, al mismo tiempo que menos consumidores serían dados de baja como resultado de su incumplimiento. Por otra parte, en vista de que los montos a provisionar dependen de los niveles de morosidad (probabilidad de incumplimiento y proporción de la deuda perdida), el disminuir las cantidades que se deben provisionar es también un beneficio, pues liberaría recursos para otros usos como consecuencia del costo de oportunidad asociado a la reserva de estos montos. De este modo, descontando otros beneficios menos directos, se procede a estimar el potencial económico de este trabajo a través de una aproximación sobre los efectos de la disminución de los niveles de morosidad en los castigos y provisiones efectuados en el periodo en estudio.

La metodología propuesta para la estimación de los beneficios económicos se basa en considerar diferentes escenarios de mejora en los niveles de morosidad, casos en los que se asumen diferentes grados de disminución en la probabilidad de incumplimiento y de perdida esperada cuando se produce esa situación. De esa forma, se estima el impacto sobre los costos incurridos en el periodo, pudiendo traducir las variaciones en las probabilidades en diferencias de costos incurridos.

Del compendio de normas contables emitido por la SBIF (Superintendencia de Bancos e Instituciones Financieras Chile, 2007), se construye la aproximación del cálculo de provisiones de la empresa. La Ecuación 23 muestra dicha formulación.

$$Provision = EG \cdot \left(\frac{PI_{grupo}}{100}\right) \cdot \left(\frac{PDI_{grupo}}{100}\right) \quad (24)$$

Con:

EG	$=$	Monto de la exposición grupal.
PI_{grupo}	$=$	Probabilidad de incumplimiento del grupo.
PDI_{grupo}	$=$	Porcentaje de pérdida dado el incumplimiento esperada asociada al segmento grupal.

Por otra parte, respecto al castigo, este corresponde al monto total adeudado con el que se superan los 180 días y que contablemente se asume que no fue ni será cancelado. Simplificando el problema, este puede ser entendido como la exposición (EG_{grupo}) con la que los clientes son castigados.

Con estos elementos en consideración y tomando como punto de partida las provisiones y castigos aplicadas para la base de clientes del año 2017 (periodo en estudio), se procede a sensibilizar el impacto que potenciales disminuciones en la probabilidad de incumplimiento (PI) y porcentaje de pérdida dado el incumplimiento (PDI) tendría sobre los valores anuales de estos apartados. Nótese que dado que la estimación es efectuada sobre el año 2017, y en una muestra de clientes , el beneficio calculado resulta de carácter anual y solo aplicable al grupo y periodo en estudio. Así, la Tabla 15 exhibe el análisis de sensibilidad que variaciones de puntos base (bp o ‰) sobre PI y PDI implican sobre las estimaciones de provisiones y castigos para el periodo. Así, se teoriza como los impactos monetarios de esta memoria pueden variar dependiendo del escenario considerado.

De esta forma considerando un escenario conservador donde solo la probabilidad de incumplimiento tiene una disminución de 10‰ se obtiene un valor para el proyecto en torno a los $18 millones. Nótese que esto considera la disminución en los montos castigados y en provisiones de acuerdo con la menor PD_{grupo}.

Por su parte, si se evalúa con un criterio de mayor acidez, y solo se considera el efecto sobre el castigo. Entonces, el único beneficio del proyecto lo constituye el dinero que no se asume como pérdida, ascendiendo dicho monto a las $ 8 millones, para disminuciones de 10 ‰ de la morosidad de la cartera.

Tabla 15: Beneficios esperados sensibilizados por disminución de PI[a] y PDI[b] en MM$[c]

Puntos base variación PI^a \ Puntos base variación PDI^b	0	-10	-20	-30	-40	-50	-100	-150	-200	-250
0	0	11	22	33	44	55	110	164	219	274
-10	18	29	40	51	62	73	128	183	237	292
-20	37	48	59	70	81	92	146	201	256	310
-30	55	66	77	88	99	110	165	219	274	328
-40	74	85	96	107	118	128	183	238	292	347
-50	92	103	114	125	136	147	201	256	310	365
-100	185	196	206	217	228	239	293	347	402	456
-150	277	288	299	310	320	331	385	439	493	547
-200	370	380	391	402	413	423	477	531	584	638
-250	462	473	483	494	505	515	569	622	676	729

Nota. [a] Probabilidad de Incumplimiento [b] Porcentaje de pérdida dado el incumplimineto. [c]Millones de Pesos. Fuente : Elaboración propia.

Adicionalmente, se reconoce que además de las utilidades antes señaladas, existen otras fuentes de beneficio económico no capturadas por esta metodología de estimación. Estas corresponden principalmente al costo de oportunidad que se tiene sobre los clientes, que de estar al día podrían utilizar su tarjeta en otras interacciones con la empresa.

Finalmente, en vista de que el ejercicio académico desarrollado por este trabajo se configura sobre una muestra del total de clientes de la empresa y a que la metodología presentada para la estimación del valor es una estimación del potencial real. Se procede a establecer que los valores presentados son referenciales y constituyen una cota inferior de los beneficios potenciales de una implementación exitosa de esta memoria, cuya validación empírica ha sido propuesta como un trabajo futuro (ver 6.5 Trabajos futuros).

6 Conclusión

Como capítulo final de este trabajo, las conclusiones obtenidas son sistematizadas en diversas dimensiones que recogen las reflexiones para cada una de ellas. Estas áreas de análisis responden las temáticas relacionadas con: el cumplimiento de objetivos, metodología y modelos empleados, resultados y métricas de desempeño, propuestas comerciales y trabajos futuros.

6.1 Conclusiones sobre la metodología y modelos

6.1.1 Datos e información

En vista que el trabajo desarrollado aborda elementos propios de varias áreas de la compañía, el proceso de preparación de la data ocuparía un lugar importante, debido a la necesidad de comprender un proceso en el que actúan varias divisiones. Esta situación se manifestaría; en la dificultad para contar con datos en un máximo nivel de granularidad (situación del cliente cada 30 días según su periodo de facturación). Dado que la gestión de cobranza se efectúa bajo estos criterios, dicha situación hubiese sido el escenario ideal. En su defecto, se modeló bajo la máxima granularidad posible y dimensionando los periodos en tramos más agregados para homogeneizar su duración. Los efectos de estas acciones fueron la sobre y sub ponderación de algunos comportamientos; probabilidades de avance sub-ponderadas por la agregación de tramos y de mantención sobre-ponderadas por lo mismo. Por lo tanto, se concluye que mayores niveles de granularidad permitirían análisis futuros más acabados sobre los tramos de morosidad, ya que haría los modelos comparables con la gestión actual en todos los tramos y no se introduciría la problemática de mezclar granularidades en un mimo modelo, aspecto que finalmente se optó por reportar y considerar en todos los análisis de este trabajo.

6.1.2 Segmentación y clasificación

Se verifica la utilidad del uso de estrategias no supervisadas para el agrupamiento de observaciones, en este caso clientes bajo un criterio de similitud. Al mismo tiempo, se evidencia la necesidad de desarrollar estos procedimientos de manera iterativa, conforme su mayor dificultad es la validación de sus resultados, en vista de que no pueden ser ratificados sin algún grado de conocimiento experto. Así, se destaca la importancia de métricas que permiten abordar teóricamente el proceso de comprobación, a modo de obtener una primera aproximación al resultado final. Este resultado aun así debe ser visado por personas con conocimiento del negocio, al mismo tiempo que se consideran criterios de accionabilidad, por ejemplo para evitar la generación de grupos muy pequeños, o la posibilidad de que los centros de los clústeres no sean representativos del conjunto (presencia de outliers ensucie análisis posteriores). Además, la consideración de variables relevantes al problema se sitúa como otro de los puntos fundamentales, ya que como se recoge en la literatura y se evidenció en este trabajo, puede ocasionar una gran variabilidad de los resultados de la segmentación, como distintos números de clústeres para diferentes dimensionalidades y la caracterización de comportamientos no relacionados con el problema.

6.1.3 Cadenas de Markov

En relación con las cadenas de Markov, se destaca su flexibilidad en la construcción de modelos que permiten un análisis multi-periodo, considerando la evolución de todas las migraciones posibles en la determinación de un pronóstico. No obstante, esta virtud, puede ser también una desventaja dado que, para modelos con muchos estados, la interpretación directa de la matriz de transición es dificultosa, abordándose este problema en este trabajo con herramientas de visualización, además de la identificación de subconjuntos de migraciones específicas.

Para el caso aplicado, se destaca la posibilidad del análisis multi-periodo que se concretiza en la caracterización de grupos que transicionan progresivamente al castigo y de otros segmentos que no. Esto se expresa en las probabilidades exhibidas por sus matrices de transición, que darían cuenta del pronóstico en más de un periodo en el futuro. Al mismo tiempo, la inclusión de una variable de estado adicional al tramo de morosidad permitió identificar conductas especiales dentro de un segmento (efectos particulares del abono).

6.2 Conclusiones sobre los resultados obtenidos

Respecto a los resultados obtenidos, las primeras conclusiones refieren a aquellos que son consecuencia directa de la segmentación desarrollada. De esta, se obtendría 6 grupos caracterizados por diferentes comportamientos basado en variables relacionadas con morosidad, transaccionalidad y riesgo. Estos clústeres son retomados a continuación (Tabla 16).

Tabla 16: Segmentos de morosidad construidos

Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5	Clúster 6
Infrecuentes con deterioro pausado	Inactivos forzados con deterioro acelerado	Activos esporádicos con deterioro moderado	Activos Saturados con deterioro pausado	Inactivos forzados de deterioro progresivo	Activos insaturados con deterioro pausado

Nota. Fuente: Elaboración propia.

Estos grupos al ser combinados con los modelos de Markov introducirían la heterogeneidad deseada dentro del modelamiento; estableciendo claras diferencias entre las probabilidades de transición de un grupo y otro. Por otra parte, los segmentos permiten identificar a través de su análisis, la correlación entre altos usos del disponible y la morosidad, así como la importancia de episodios de morosidad pasados como indicadores de potenciales situaciones futuras, y como el impacto del nivel de gasto varía según el segmento al que pertenezca el cliente.

De las características más relevantes de los clústeres obtenidos, se destaca la importancia de los segmentos 2 y 5, que junto con agrupar a los clientes con algunos de los riesgos más altos, (mayores scores promedios de riesgo), son los que presentan mayor incidencia de mora, con un mayor número de episodios y montos; que finalmente

en el modelamiento se asocian a los comportamientos de deterioro más progresivo. Por otra parte, se remarca la volumetría de los grupos 4 y 6 que concentran el grueso de los clientes (más del 60%), quienes transaccionalmente más activos (mayor ticket promedio, monto gastado y menor recencia de compras), dan cuenta de la una parte importante de la cartera que históricamente ha presentado acotados episodios de morosidad. Nótese que en vista del objetivo de contención de la cobranza, i.e. evitar el avance en morosidad, dado el deterioro progresivo de algunos segmentos, los grupos más importantes lo constituyen los segmentos 2 y 5, a pesar de no ser los más numerosos.

De las consecuencias de la granularidad sobre las probabilidades de transición y calidad de los pronósticos, es importante retomar, los sesgos que esta introduce sobre las chances de migración, por cuanto las probabilidades de avanzar de tramo son sub-ponderadas en tramos agregados y las de mantención sobre-ponderadas para dichos casos. Considerando que la gestión de cobranza de la empresa es organizada de manera mensual, no se evaluó la posibilidad de asumir pronósticos y tramos cada 60 días. Por otra parte, en relación con los efectos de la granularidad sobre el pronóstico, basándose en los efectos sobe las métricas de desempeño, tramos de más amplios, i.e. menores niveles impactan de manera positiva sobre la cantidad de transiciones acertadas, atribuyéndose dicho comportamiento al resultado de la agregación sobre la performance del pronóstico. No obstante, la magnitud del efecto puede ser fuertemente atenuado por el tipo de modelamiento empleado en la cadena (Definición de variables de estado para construcción de la cadena).

Sobre las variables de estado aplicadas en cada uno de los modelamientos, estas determinan el tipo de conclusiones que se obtienen de cada ajuste. Además, tienen un fuerte impacto en la magnitud del error cometido en cada uno de los pronósticos. Así, se establece que, para métricas comparables, los modelamientos más simples consiguen mejores desempeños, mientras que, en términos absolutos, cadenas con un mayor número de estados fallan menos transiciones. De esta forma, el modelo con mejor tasa de pronóstico en términos absolutos y transversalidad (error absoluto) es el modelo MK2, mientras que el de mejor desempeño general (error ponderado por número de transiciones simétrico) es MK1. Por otra parte, la inclusión de una variable de estado adicional al tramo de morosidad, el tiempo consecutivo en mora, tiene como consecuencia la captura de diferencias en las migraciones entre tramos de morosidad iguales, discrepancias que responden a las disposiciones a regularizar su situación en cada uno de los grupos, y que quedan determinadas por el número de abonos que se realizan para evitar agravar la situación de morosidad.

En relación con los resultados del modelamiento a nivel de cliente, las métricas de evaluación indican una performance por debajo de lo esperado, pues en promedio no sobrepasa el 50% de *precision* y *recall*, para los mejores casos, pues el modelamiento es capaz de capturar las transiciones extremas, pero no las que ocurren entre tramos intermedios. Por este motivo, no se recomienda el paso a producción de este modelo, sin antes desarrollar mejoras en la performance (ver 6.5 Trabajos futuros).

Sobre las relaciones entre las variables estudiadas y el fenómeno de morosidad, estas interacciones son abordadas en base a las transiciones más importantes para proceso; las de retorno, mantención y avance. De esta forma los principales *insights* son recogidos de acuerdo con estos tres tipos de migraciones y tres clasificaciones macro de los clústeres desarrollados.

La macro clasificación de los clústeres se ajusta a los resultados de la segmentación con enfoque grupal, teniéndose que, por clientes de buen comportamiento se consideran aquellos que en promedio presentaran menos de 1 episodio de morosidad en los últimos 12 meses y su score de riesgo se encontrase entre los mejores promedios (>500). Que por clientes de mal comportamiento se toman en cuenta aquellos con las mayores incidencias de morosidad (más de 7 episodios de mora en los últimos meses y score de riesgo <210). Y que por clientes de poco riesgo y frecuencia se consideraran quienes tuviesen baja incidencia de mora y/o riesgo y un comportamiento transaccional acotado (recencia de compras en torno a los 3 y 4 meses e incidencias de morosidad bajo los 6 episodios en los últimos 12 meses). Así, finalmente se procede a concluir sobre la relación de las variables estudiadas y el fenómeno de morosidad en base a sus probabilidades.

Probabilidades de retorno

- Clientes de **buen comportamiento** disminuyen sus probabilidades de retornar a medida que avanzan de tramo de morosidad. No obstante, la excepción la constituye el momento en que se produce la suspensión de la cuenta, estado en el cual se experimenta un repunte en esta probabilidad.
- Clientes de **mal comportamiento** experimentan probabilidades de retorno estrictamente decrecientes al avanzar en tramos de morosidad, siendo aquellos con las tarjetas más restrictivas quienes ven disminuidas estas chances en mayor cuantía.
- Clientes de **poco riesgo y frecuencia** experimentan probabilidades de retorno decrecientes con excepción de tramos de mayor gravedad, en los que existe un leve repunte atribuible a la amenaza de castigo.

Probabilidades de avanzar

- Clientes de **buen comportamiento** no tienen grandes probabilidades de avance exceptuando tramos finales de morosidad, en los que las posibilidades de empeorar su situación se hacen patentes.
- Clientes de **mal comportamiento** presentan probabilidades de avance crecientes conforme se progresa en tramos de morosidad de mayor mora. No obstante, dentro de este grupo existen diferencias al considerar otras variables como el abono.
- Clientes de **poco riesgo y frecuencia** experimentan mayores probabilidades de morosidad hacia mitad de su recorrido de mora, presumiblemente por su menor relación con la empresa.

Efecto abono

- Para clientes de **poco riesg**o y o **buen comportamiento** el abono se asocia a menores probabilidades de avance hasta un cierto número de periodos consecutivos, para luego perder su efecto, ya que para este grupo la morosidad no es un problema de disposición si no que de incidencias.
- Para clientes de **mal comportamiento**, el abono tiene mayor impacto en reducción de probabilidades de avance. El efecto se produce aún con las mayores tasas y es más relevante en segmentos de menor gasto y tarjetas con menos prestaciones.
- El efecto de abono es más relevante en clientes esporádicos de riesgo moderado; **poco riesgo y frecuencia**, observándose monotonía en la disminución de sus probabilidades de avance.

6.3 Propuestas comerciales

A partir de los resultados del modelamiento a nivel de grupo y cliente, se proponen algunas recomendaciones de negocio orientadas a propiciar la disminución de los niveles de mora, y así reducir sus efectos sobre la última línea.

En términos de propuestas comerciales, estas se basan en los descubrimientos en el comportamiento de grupo, mientras que para el caso del pronóstico a nivel de cliente adquieren un carácter hipotético. Principalmente, debido a que este modelo tendría un desempeño que admite mejoras, pero que de perfeccionarse podría aplicarse como se indica en este apartado. Ese último elemento es considerado dentro de los trabajos futuros a presentar de manera posterior (ver 6.5 Trabajos futuros). Así la presente sección recoge las recomendaciones de negocio propuestas a partir de este trabajo. Cabe destacar que en vista de los mejores resultados del modelo MK2 en términos transversales, las propuestas presentadas consideran la aplicación del mencionado para el pronóstico y validación en el tiempo de los descubrimientos. Además, se contempla el combinarlo con la identificación de hitos, como los abonos para la ejecución de las recomendaciones.

6.3.1 Jerarquización a nivel de grupo

Con el objetivo de mejorar la gestión de cobranza para aquellos clientes que han caído en situación de morosidad, el primer grupo de recomendaciones considera el enriquecimiento de la jerarquización con la que se efectúa la asignación de la gestión.

La situación actual considera la asignación de la cartera morosa en base a las capacidades de cada uno de los canales de gestión, condicionado al tramo de morosidad alcanzado. Así, en los niveles de menor gravedad, solo se aplican gestiones masivas y de menores costos. Por el contrario, en tramos de mora superior, estas gestiones pueden considerar acciones de mayor impacto, como llamadas desde un centro de atención telefónica (*call center*) e incluso cobradores en terreno.

La propuesta contempla que en la asignación actual se incorpore un nuevo criterio de jerarquía. De esta forma se mantiene el fundamento de la gestión de cobranza; contener el avance de los clientes a tramos de morosidad de mayor gravedad y al mismo tiempo se incorpora los nuevos resultados, los cuales son recopilados en las siguientes prioridades.

- **Primera prioridad:** se propone aplicar las acciones de cobranza más efectivas a los deudores que cumpliendo criterio de monto mínimo, se han asignado a los grupos de máximo riesgo; Inactivos forzados con deterioro acelerado (clúster 2) e Inactivos forzados de deterioro progresivo (clúster 5). Este criterio se justifica en la rápida transición hacia el castigo de estos grupos, por lo que, bajo la máxima de contener el avance de los clientes, se alinea con la estrategia actual de cobranza.
- **Segunda prioridad:** se establece para aquellos clientes que, a través del abono, han suavizado sus probabilidades de avance se debe capitalizar la mayor disposición a regularizar su situación, antes de que el efecto abono sea atenuado por el deterioro de su deuda Por este motivo este grupo es presentado en segundo lugar en la jerarquía.
- **Tercera prioridad:** para los clientes de buen comportamiento histórico (clústeres 4 y 6), en base a sus resultados se propone una gestión pasiva, i.e. que no implique quitar recursos de los segmentos prioritarios, pues hasta los 90 días sus probabilidades de retorno aún no han experimentado una baja significativa. Asimismo, atendiendo al incremento de sus probabilidades en los tramos cercanos al castigo, dado que comparativamente a los otros grupos este deterioro no es tan exacerbado se recomienda el mantener este grupo al final de la jerarquía.

6.3.2 Jerarquización a nivel de cliente

Como segunda parte de las recomendaciones, se encuentra la jerarquización a nivel de cliente. Esta es una propuesta complementaría a las prioridades establecidas por la pertenencia a un grupo, que busca ofrecer resultados más personalizados. La idea es que además de los presentado, se consideren las probabilidades de cada deudor, siendo el criterio de prioridad que los clientes tengan en el análisis de dos periodos mayores probabilidades de avanzar de tramo, que de retornar al día o tramos inferiores. Esta aplicación es ejemplificad de manera ficticia por la Tabla 17. En ella se observa cómo clientes del mismo segmento, respondiendo a su heterogeneidad presentan pronósticos diferentes, al mismo tiempo que su prioridad pasa es determinada por su proyección de más de un periodo. De esta manera, se centralizan los esfuerzos en quienes no retornarán después de su avance. En este caso, el ejemplo lo constituye el cliente AAA, que respecto al cliente BBB se espera que empeoré su situación en un horizonte de tiempo mayor.

Tabla 17: Caso de uso resultados a nivel de cliente y jerarquización

ID	Grupo	Grupo 1	Grupo 2	Grupo 3	Estado Actual (t)	Estado Sig. $(t+1)$	Estado Subsig. $(t+2)$	Prioridad
AAA	1	80%	10%	10%	0.[AL DIA]	A.[1-60]	B.[61-120]	1°
BBB	1	50%	30%	20%	0.[AL DIA]	A.[1-60]	0.[AL DIA]	2°
CCC	2	30%	50%	20%	0.[AL DIA]	A.[1-60]	0.[AL DIA]	3°

Nota. Fuente: Elaboración propia.

En vistas de que, al término de este trabajo, parte de los próximos pasos contemplan el mejorar la capacidad predictiva de los pronósticos en este nivel, esta recomendación queda sujeta a la realización de dichos perfeccionamientos.

6.3.3 Triggers

En vista del efecto que tiene la suspensión de la cuenta para clientes de alta y moderada transaccionalidad, el aumento de sus probabilidades de retorno y disminución de las de avance al producirse ese hito, se propone un set de recordatorios orientados a informar y o rememorar las consecuencias del avance en tramo de morosidad, y consecuente suspensión de la cuenta. Así, clientes con una alta valoración del uso de la tarjeta y que ante la suspensión saldan sus deudas, serán informados de antemano de las implicancias de no hacerlo[16], de manera de disminuir el avance al anticipar el hito (Suspensión). De manera que esta recomendación se constituye en un plano educacional y de alerta al cliente.

6.4 Potencial económico

El potencial económico de este trabajo se explica en los efectos sobre dos costos relacionados con la morosidad. Costes que de producirse una disminución de los niveles de no pago se verían afectados de manera favorable.

El primero de estos costos corresponde a los castigos que no se concretarían, disminuyendo las pérdidas directas del no pago. Mientras que el otro beneficio, sería el aumento del capital disponible, al evitar parte del costo de oportunidad ocasionado por el provisionamiento sobre una cartera con mejores comportamientos de pago. Así, como se presentara en el escenario conservador de la evaluación económica (5.5.1 Evaluación económica), mejoras de los niveles de morosidad traducidos en disminuciones de 10‰ (10 puntos bases) en las probabilidades de incumplimiento, estiman beneficios cercanos

[16] Clientes en el estado suspendido no se encuentran habilitados para comprar y o aprovechar los beneficios de sus cuentas, al mismo tiempo que presentan mayores dificultades para regularizar su situación.

a los $8 millones si solo se consideran el efecto directo sobre el castigo y que si se suman las consecuencias de provisionamiento sobrepasa el doble del beneficio anterior, ascendiendo a unos $18 millones. Nótese que los resultados presentados corresponden a el valor del trabajo estimado sobre la muestra de clientes del estudio y solo considerando el año 2017 (beneficio anual), por lo que constituye una cota inferior del potencial del trabajo.

6.5 Trabajos futuros

En vista de los resultados obtenidos y propuestas desarrolladas, se establecen cuatro líneas de acción principales, cuya ejecución podrían complementar y mejorar esta memoria.

- **Reconstrucción de la data:** En vista de los efectos de la granularidad y su relación en las probabilidades de transición y calidad del pronóstico, un potencial próximo paso lo constituye la reconstrucción en tramos de 30 días para mejorar la aplicabilidad del modelo dentro del flujo de cobranza.

- **Incorporación explicita del efecto de acciones:** Dada la naturaleza del análisis, en próximas iteraciones se debe considerar el tratamiento de los clientes dentro del flujo de cobranza para que las acciones se guíen considerando el contrafactual de haber sido gestionado respecto a no serlo y sus implicancias en el recupero. Siendo este otro punto fundamental, al incluir en la jerarquización de los clientes en la asignación de cobranza el esperado de recuperación.

- **Mejoras en la performance de pronósticos**: Como resultado de los deficientes resultados a nivel de cliente, en términos de pronosticar su comportamiento individual, se propone la evaluación de otras metodologías recopiladas en la revisión de la literatura. En particular, como lo hiciera Banasik, Crook y Thomas (Banasik et al., 1999) y Ho Ha y Krishnan (Ho Ha & Krishnan, 2012), se sugiere que para este enfoque se consideren modelos de duración en sus diferentes variantes.

- **Integración con optimización del recupero:** Dado que la gestión de cobranza busca maximizar la recuperación de las deudas vencidas, evitando que los clientes avancen en los tramos de morosidad respectivos, el paso siguiente corresponde en integrar los modelos de pronósticos con modelos de optimización que permitan incorporar directamente el monto que la empresa no pierde con su gestión de cobranza.

6.6 Conclusiones sobre objetivos planteados

Para terminar con el capítulo destinado a las conclusiones de esta memoria, se finaliza con la validación y justificación del cumplimiento de los objetivos planteados.

En primer lugar, en relación con la meta general de este trabajo, esta se da por alcanzada. La creación de oportunidades para propiciar la disminución de la morosidad, y reducir los costos asociados a ella, es a grandes rasgos conseguida gracias a la caracterización del fenómeno de morosidad, mediante la construcción de un modelo de pronóstico eficaz para describir comportamientos de grupo y elaborar propuestas comerciales.

En segundo lugar, seguida a la validación del cumplimiento del objetivo general, los específicos se evalúan como conseguidos, justificándose como sigue:

- Se identificaron las principales variables explicativas del comportamiento de morosidad de los clientes, tales como la mora pasada, nivel de endeudamiento, nivel de uso de la tarjeta y disponible, además de la propia situación de mora del cliente y abono incorporadas de manera directa en el modelamiento.
- Se construyó un modelo de pronóstico del estado de la morosidad de los clientes con un enfoque multi-periodo.
- Del perfilamiento de segmentos de comportamiento de morosidad, se derivarían recomendaciones comerciales orientadas a la disminución de la mora.

Finalmente, con los elementos antes señalados se considera que los objetivos de este trabajo han sido desarrollados a cabalidad, siendo las oportunidades de mejora y trabajos futuros un punto de partida para nuevos ejercicios académicos, y al mismo tiempo el cierre de esta memoria.

7 Bibliografía

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but When will Borrowers Default. *The Journal of the Operational Research Society*, *50*(12), 1185. https://doi.org/10.2307/3010627

Barnett, V., & Lewis, T. (1994). Outliers in statistical data. *John Wiley & Sons*, *3*(1).

Ben-gal, I. (2005). Outlier Detection. *Data Mining and Knowledge Discovery Handbook*, 131–146. https://doi.org/10.1007/0-387-25465-x_7

Carpenter, J. R., & Goldstein, H. (2009). *Handling missing data*.

Cyert, R. M., Davidson, H. J., & Thompson, G. L. (1962). Estimation of the Allowance for Doubtful Accounts by Markov Chains. *Management Science*, *8*(3), 287–303. https://doi.org/10.1287/mnsc.8.3.287

Desgraupes, B. (2013). Clustering indices. *University of Paris Ouest-Lab Modal'X*, *1*(April), 34. Recuperado a partir de ftp://apache.cs.uu.nl/mirror/CRAN/web/packages/clusterCrit/vignettes/clusterCrit.pdf

Eidgenössische Technische Hochschule Zürich. (2012). Finding Multivariate Outlier. Recuperado 15 de septiembre de 2017, a partir de https://stat.ethz.ch/education/semesters/ss2012/ams/slides/v2.2.pdfhttps://stat.ethz.ch/education/semesters/ss2012/ams/slides/v2.2.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Figueroa, N., L'Huillier, G., & Weber, R. (2017). *Adversarial classification using signaling games with an application to phishing detection*. *Data Mining and Knowledge Discovery* (Vol. 31). Springer US. https://doi.org/10.1007/s10618-016-0459-9

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, *3*(3), 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, *17*(2–3), 107–145. https://doi.org/10.1023/A:1012801612483

Ho Ha, S., & Krishnan, R. (2012). Predicting repayment of the credit card debt. *Computers and Operations Research*, *39*(4), 765–773. https://doi.org/10.1016/j.cor.2010.10.032

Khandani, A. E., Kim, A. J., & Lo., A. W. (2014). Consumer credit-risk models via machine-learning algorithms. https://doi.org/https://doi.org/10.1016/j.jbankfin.2010.06.001

Kuelen, J. A. M. van, Spronk, J., & Corcoran, A. W. (1981). On the Cyert-Davidson-Thompson Doubtful Accounts Model. *Management Science*, *27*(1), 108–112. https://doi.org/10.1287/mnsc.27.1.108

Min, S., Scott, Z., Cameron, G., Martin, R., Drossu, R., Zhang, J. (Guofeng), & Shoham, D. (2013). *7,191,150 B1*. Estados Unidos.

Orlov, A. I. (2011). Mahalanobis distance. En *Encyclopedia of Mathematics*. Recuperado a partir de http://www.encyclopediaofmath.org/index.php?title=Mahalanobis_distance&oldid=17720

Osses Godoy, A. A. (2015). *Desarrollo de un método de valoración de clientes en una empresa del sector automotriz*. Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas. Recuperado a partir de http://repositorio.uchile.cl/handle/2250/133525

Patel, M. (2016). k-means clustering for Outlier detection. Recuperado 18 de septiembre de 2017, a partir de https://rpubs.com/maulikpatel/228345

Poblete, B. (2016). *CC5206-1 Introducción a la Mineria de Datos 2016, Primavera. Clase 6: Clasificación (2/3)*. Recuperado a partir de https://www.u-cursos.cl/ingenieria/2016/2/CC5206/1/material_docente/bajar?id_material=1534929

Prabhakaran, S. (2016). Outlier Treatment. Recuperado 18 de mayo de 2018, a partir de http://r-statistics.co/Outlier-Treatment-With-R.html

Roco Benavides, C. E. (2010). *Modelamiento Predictivo para el Aumento de Consumo de Tarjeta de Crédito Sobre el Análisis de Comportamiento Transaccional de Clientes de una Institución Financiera*. Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas. Recuperado a partir de http://www.repositorio.uchile.cl/handle/2250/103714

Rosenberg, E., & Gleit, A. (1994). Quantitative Methods in Credit Management: A Survey. *Operations Research*, *42*(4), 589–613. https://doi.org/10.1287/opre.42.4.589

Rosenmai, P. (2013). Using Mahalanobis Distance to Find Outliers. Recuperado 15 de septiembre de 2017, a partir de http://eurekastatistics.com/using-mahalanobis-distance-to-find-outliers/

Ruefer, S. (2016). Outlier Detection with Mahalanobis Distance.

Segovia, C., Aburto, L., & Goic, M. (2005). Caracterización del proceso de fuga de clientes utilizando información transaccional.

Segovia Riquelme, C. A. (2005). *Caracterización del proceso de fuga de clientes de un retail banking utilizando información transaccional*. Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas.

Souza, J., Matwin, S., & Japkowicz, N. (2002). Evaluating data mining models: A pattern language. *Proceedings of the 9th Conference on Pattern Language of Programs, Illinois, USA*, 1–23. Recuperado a partir de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.12.7271&rep=rep1&type=pdf

Superintendencia de Bancos e Instituciones Financieras Chile. (2007). Compendio de Normas Contables, Anexo 1, Cap. B-1. Recuperado a partir de https://www.sbif.cl/sbifweb/internet/archivos/norma_6545_1.pdf

Weber, R. (2017). KDD Process, Feature Selection, Pre-processing, Transformation. Support Material IN4521 Introduction to Data Mining. Santiago.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining (PADD '00)*, (24959), 29–39. https://doi.org/10.1.1.198.5133

Xu, R. (2005). Survey of clustering algorithms for MANET. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141

8 Anexos y Apéndices

ANEXO A Tareas asociadas a las fases del proceso CRISP-DM

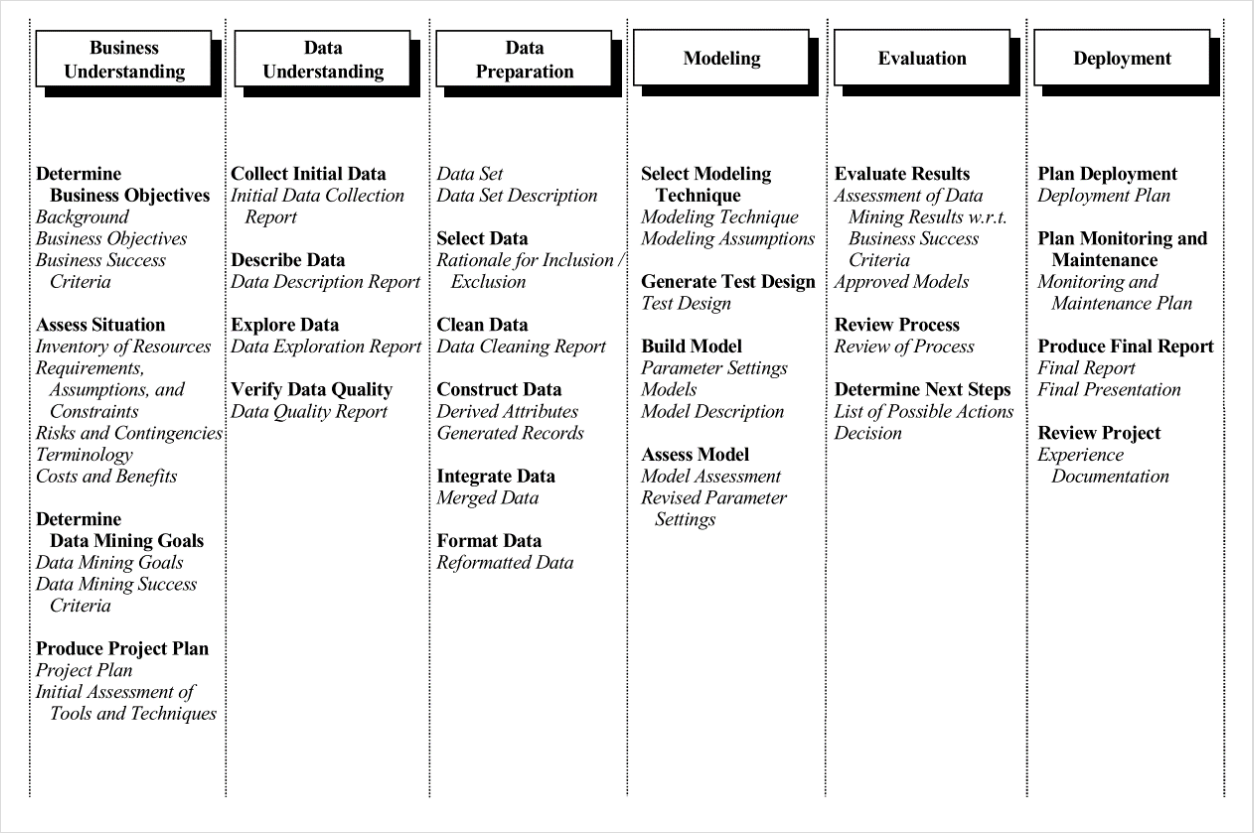


Figura 72. Visión general de los pasos del proceso CRISP-DM y las diferentes tareas asociadas a cada uno de ellos.

Nota. Fuente: Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining (PADD '00), (24959), 29–39. https://doi.org/10.1.1.198.5133 (Wirth & Hipp, 2000, fig. 3)

ANEXO B Alterativas de modelamiento administración de crédito

Regresiones Logísticas

Es un tipo de modelo en que la variable dependiente es categórica y representa la pertenencia de un registro a una clase. Este método se asocia a modelos de elección discreta y corresponde a un clasificador estadístico. Dependiendo del número de categorías presentadas se considerará un modelo logit binario o multinomial. La función logística se encuentra acotada entre 0 y 1 para cada una de las clases, por lo que, los resultados se interpretan como una probabilidad. Así, para el caso del logit binario, la probabilidad de que ocurra uno de los sucesos $\{S\}$ para el individuo $\{i\}$,queda descrita por la Ecuación 25.

$$P_i(S) = \frac{e^{\beta^T X}}{e^{\beta^T X} + 1} \tag{25}$$

Donde, $X = \{x_1, x_2 \dots x_j \dots x_n\}$ e un vector de características del individuo de dimensión n y β un vector a estimar mediante maximización de verosimilitud, para determinar el peso de los parámetros.

Árboles de decisión

Los árboles de decisión son una metodología de minería de datos que permite explicar una variable dependiente, tanto si esta es de tipo numérica como categórica, denominándose para cada uno de los casos como arboles de regresión o clasificación respectivamente. Estos modelos están constituidos por una estructura jerárquica que particiona la data en base a los valores que pueden tomar las variables independientes, ramificándose hasta que alcanza un criterio de detención.

Uno de los modelos más populares para la construcción de árboles de decisión es el método CART (Khandani, Kim, & Lo, 2010, pág. 19). Este se basa en la partición binaria recursiva de la data en base a un vector de características $(x_1 \dots x_n)$, cuyas relaciones son categorizados hacia alguno de los nodos hijos en base a reglas, por ejemplo $x_j < L_j$ con $j = 1 \dots n$. El parámetro $\{L_j\}$ es escogido en base a la minimización de alguna distancia entre la variable dependiente y el valor predicho por el árbol, a partir de una métrica (usualmente el error cuadrático medio). Para evitar la partición excesiva y sobre ajuste a los datos es necesario tener un criterio de detención, siendo un ejemplo el índice de Gini.

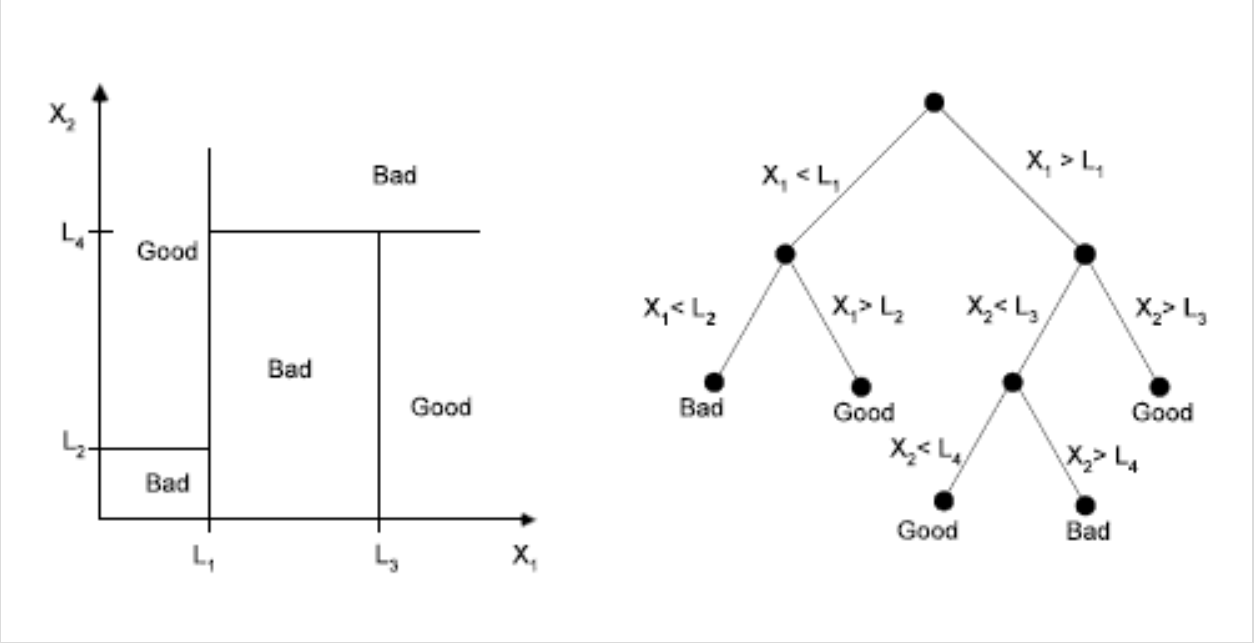


Figura 73. Ejemplo de modelo CART para una variable dependiente discreta con dos outcome, $good$ y bad y dos variables independientes $\{x_1, x_2\}$

Nota. Fuente: Khandani, A. E., Kim, A. J., & Lo., A. W. (2014). Consumer credit-risk models via machine-learning algorithms. https://doi.org/https://doi.org/10.1016/j.jbankfin.2010.06.001.(Khandani et al., 2014, p. 20).

Modelos de supervivencia

El análisis de supervivencia consiste en atribuir una distribución de probabilidad para el tiempo que le toma a una observación alcanzar un hito, usualmente denominado muerte. Para ello se emplea una función de verosimilitud que permite construir el modelo, asumiendo una probabilidad de muerte $f(t)$ y otra de supervivencia $S(t)$. No obstante, para efectos de desarrollar el análisis, la estimación ocurre a través de la función de Hazard, definida en base a la distribución acumulada $\mathrm{F(t)} = \int_0^t f(u)du$ y la supervivencia $S(t)$, tal y como lo indica la Ecuación 26 (Ha & Krishnan, 2012).

$$h(t) = \frac{1}{P(T > t)} \lim_{\Delta t \to 0} \frac{\mathrm{P(t \leq T \leq t + \Delta t)}}{\Delta t} = \frac{f(t)}{S(t)} \tag{26}$$

A la función $h(t)$, estimada mediante máxima verosimilitud, se le incorpora heterogeneidad al introducir un nivel base denominado h_0. Este valor puede ser acelerado en virtud del tiempo y variables explicativas a través de una función del tipo $\psi(z)$, usualmente explicitada como $\psi(z) = \exp(b_1 z_1 + b_2 z_2 + \cdots + b_n z_n)$. A su vez, se establecen dos tipos de modelos: acelerados y aquellos que asumen monotonía en el tiempo. Así, para el primero (ver Ecuación 27) h_0 actúa como función del tiempo, para acelerar o ralentizar la vida del sistema (Banasik, Crook, & Thomas, 1999, pág. 1186) . Mientras que el segundo caso define una relación con el tiempo como indica la Ecuación 28.

$$h(t) = \psi(z) h_0\big(t\psi(z)\big) \tag{27}$$

$$h(t) = \psi(z) h_0(t) = e^{b \cdot z}\, h_0(t) \tag{28}$$

ANEXO C Métodos para la identificación de outliers

Tabla 18: Comparativa de métodos para la identificación de outliers

Método	Descripción
Valor extremo	Método de tipo univariado, basado en la identificación de los elementos más alejados de la media, permitiendo marcar como *outliers* aquellos que se encuentran más lejanos al estimador mencionado. Nótese que procedimientos de este tipo pueden ser complementados con el uso de *boxplots* para determinar cuántos elementos extremos deben ser extraídos.
Score basado en percentil	En la categoría de los métodos univariados, este contempla el asumir la distribución subyacente de la variable (normal, *t-student*, CHI-2, IQR, etc.), para luego computar el score normalizado basado en el estadístico de la distribución. Finalmente, se identifican aquellas observaciones que yacen más allá de un percentil determinado, en función del valor de su estadístico.
K-means para detección de *outliers*	Usando algoritmos de *clustering* como *k-means*, esta técnica, considera el particionar en k grupos las data y asignar cada punto al centro más próximo. Una vez que todos los elementos han sido imputados a un grupo, computando la distancia de cada objeto al centro del clústeres al que fue asignado, se procede a elegir aquellos registros cuya distancia es mayor (Patel, 2016), definiendo un umbral para su identificación.

Nota. Fuente: Elaboración propia.

ANEXO D Métodos para la transformación de variables

Tabla 19: Comparativa de métodos para la transformación de variables

Método	Descripción	Formulación
Z-Score	Representa el valor de la variable, en términos del número de desviaciones estándar sobre las que se encuentra y su diferencia respecto a la media. De esta forma, cuando el valor se encuentra sobre la media, Z adquiere un valor positivo, mientras que cuando dicha cifra está por debajo del promedio, Z es negativo.	$$Z = \frac{X - E(X)}{\sigma(X)} \quad (29)$$
Max – Min	La técnica Max-Min es una estrategia de que transforma linealmente una variable X en una variable Y de acuerdo con la siguiente de la Ecuación 30. De esta forma, se modifica el rango de valores de la variable X, modificando sus límites inferior y superior a 0 y 1. En ocasiones puede convenir normalizar los datos en un rango de valores predefinido, en tal caso se recurre a la Ecuación 31.	$$\mathrm{Y} = \frac{(\mathrm{X} - \mathrm{X}^-)}{(\mathrm{X}^+ - \mathrm{X}^-)} \quad (30)^a$$ $$\mathrm{Y} = \frac{(\mathrm{X} - \mathrm{X}^-)(\mathrm{X}^{+\prime} - \mathrm{X}^{-\prime})}{(\mathrm{X}^+ - \mathrm{X}^-)} + \mathrm{X}^{-\prime} \quad (31)^b$$

Nota. [a] X^+corresponde al máximo de la variable X, mientras que X^-el mínmo. [b] X^+ y X^- mantienen su deficnión, al mismo tiempo que se establecen dos valores arbitrarior para acotar el rango de salida [$X^{+\prime}, X^{-\prime}$]. Fuente: Elaboración propia.

ANEXO E Metodologías para el entrenamiento de modelos

Tabla 20: Comparativa de metodologías de partición de datos para entrenamiento

Método	Descripción
Random subsampling	Basada en la ejecución repetida de *hold out* (ver 3.7.1 Hold-out) , intenta reducir la variabilidad introducida por la partición con la que se entrenan los modelos (Poblete, 2016). Luego, las métricas de desempeño son promediadas a partir de las repeticiones, pues en cada iteración el conjunto de *training* y *validation* varia por el muestreo. El método culmina con la evaluación del modelo de mejor desempeño, al aplicarse sobre el conjunto de prueba o evaluación, ya que sus hiper parámetros y mejor metodología fueron elegidas en el paso anterior.
K-folds Cross Validation	Este método particiona los datos en k subconjuntos disjuntos denominados *folds*. Cada uno de ellos es asumido como conjunto de validación, mientras que los $k-1$ restantes son empleados como *training*. Así, se evalúa la calidad de los modelos sobre el conjunto k-ésimo no considerado en entrenamiento. Esta operación es repetida de manera que cada *fold* sea considerado como validación. Los resultados de cada prueba son promediados y con este indicador se efectúa la selección de modelos y ajuste de hiper-parámetros. Este proceso sistemático y asegura que todos los datos que no pertenecen al conjunto de prueba sean considerados como entrenamiento y validación. Finalmente, para el modelo óptimo, se calcula el desempeño general del mejor modelo sobre el conjunto de test (Souza et al., 2002, p. 13)
Leave One Out	Esta metodología, corresponde a un caso particular de *k-folds cross validation*, pues se basa en aplicar este último con tantas particiones como observaciones existen en el conjunto entrenamiento-validación, i.e. con una cardinalidad igual $k = n$. Así, todos los datos son empleados como set de validación de manera sistemática, siendo estos los únicos que pertenecen a este conjunto en su iteración respectiva.
Bootstrap	Considerando una muestra de tamaño d, esta se efectúa con reemplazo y se usa para entrenar un modelo M_i , luego se procede a calcular las medidas de desempeño tanto sobre este mismo, como sobre el conjunto de test. Finalmente se estima el error final para la iteración i de acuerdo con la expresión, $e_i = 0.632 \cdot e_{training} + 0.368 \cdot e_{testing}$. Posteriormente, esta operación es repetida para cada iteración i$\in \{1..m\}$, determinando así el desempeño del modelo.

Nota. Fuente Elaboración propia.

ANEXO F Análisis descriptivo

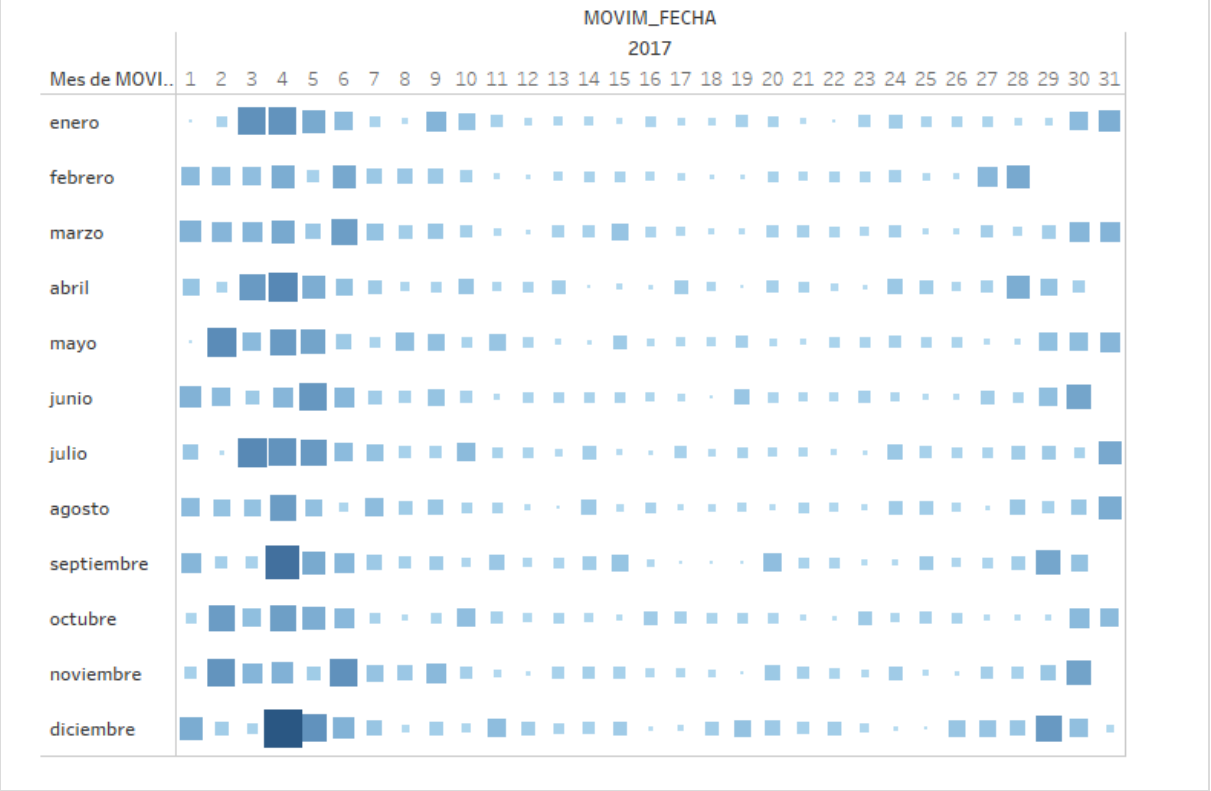


Figura 74. Concentración de pagos efectuados para muestra de la cartera de clientes, año 2017.

Nota. Fuente: Elaboración propia.

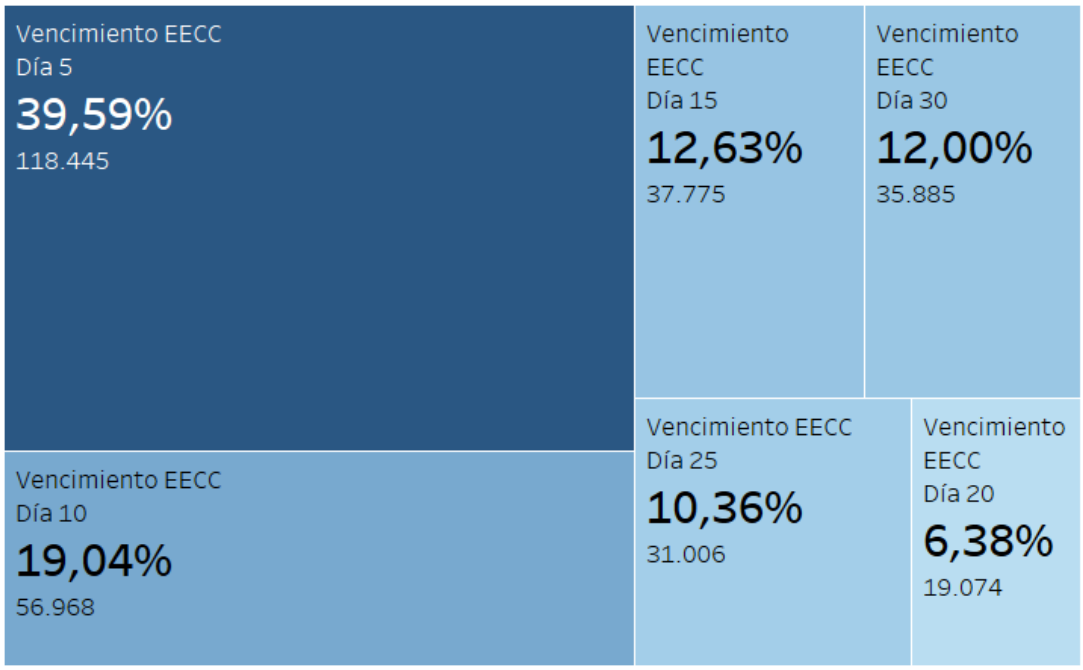


Figura 75. Distribución fecha de vencimiento de los clientes

Nota. Fuente: Elaboración propia.

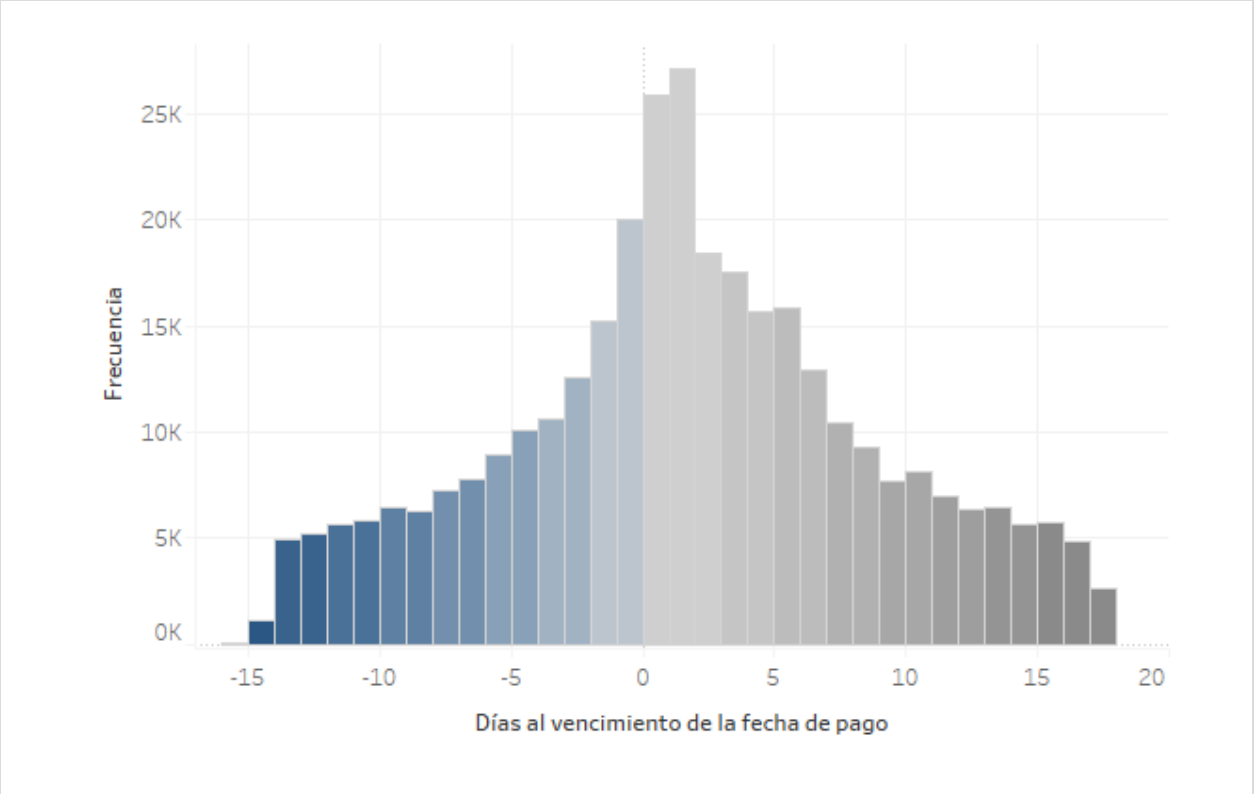


Figura 76. Distribución de día al vencimiento con la que cancela la facturación mensual

Nota. Fuente: Elaboración propia.

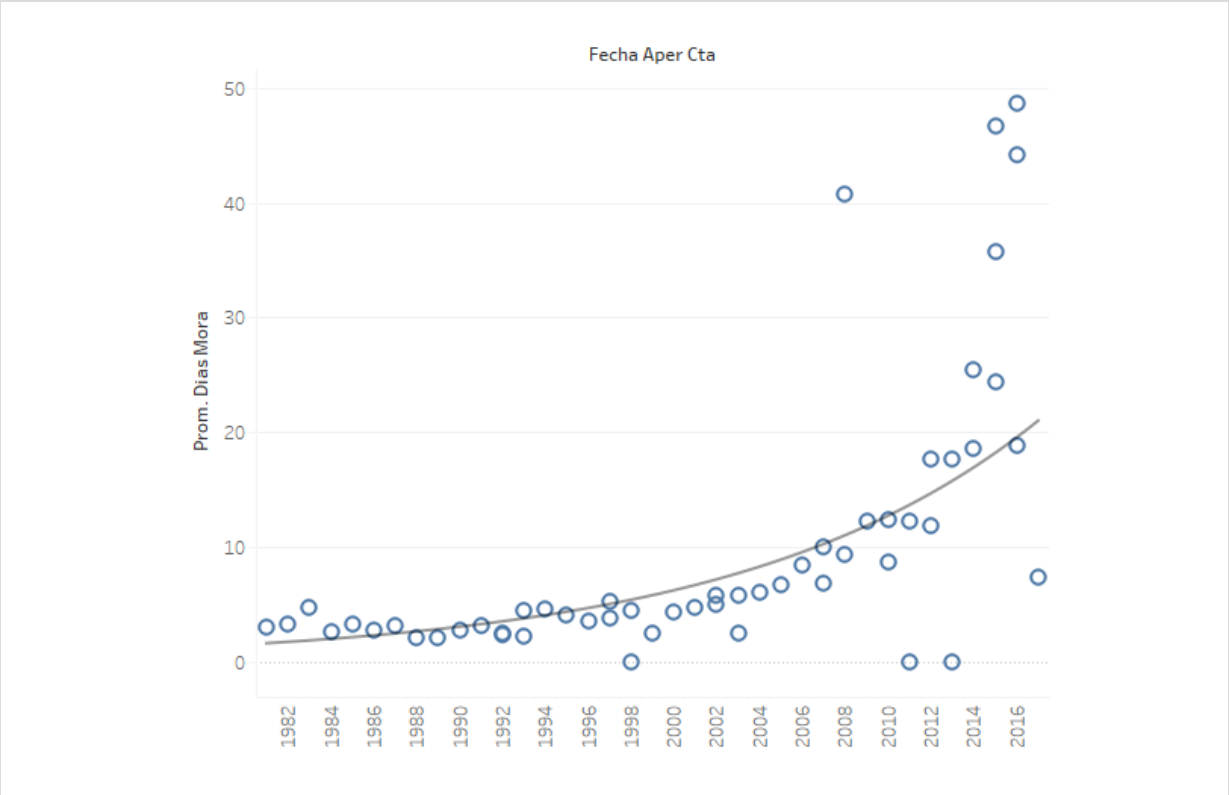


Figura 77. Promedio días de morosidad respecto al año de apertura de la cuenta

Nota. Fuente: Elaboración propia.

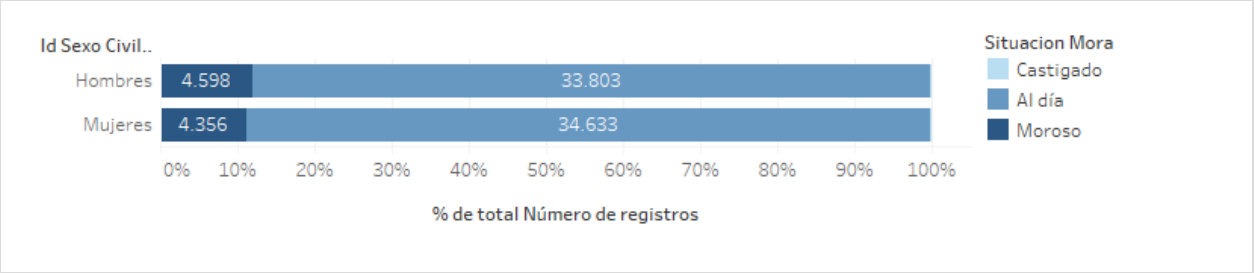


Figura 78. Situación de morosidad respecto al género del cliente

Nota. Fuente: Elaboración propia.

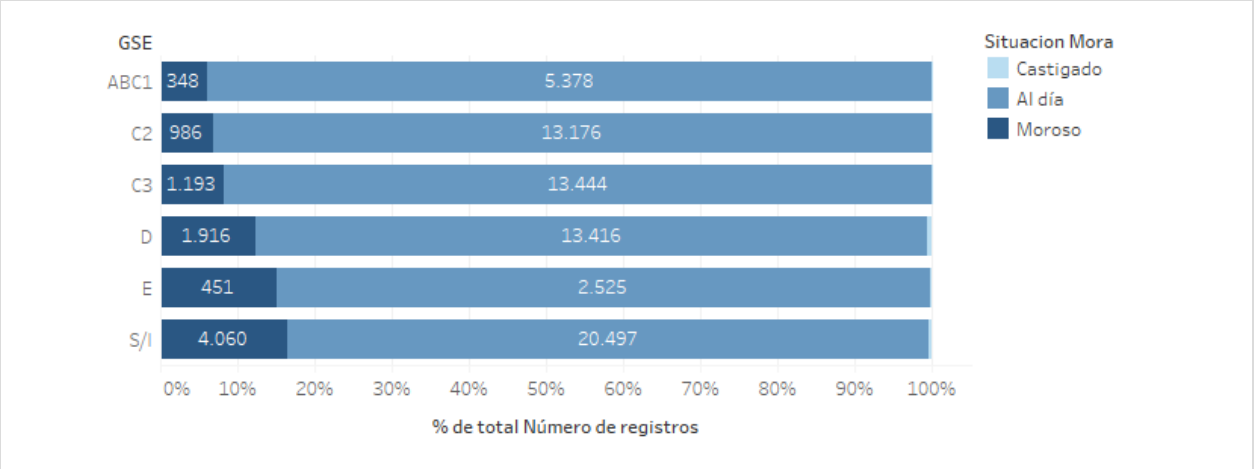


Figura 79. Situación de morosidad respecto al grupo socioeconómico (GSE) del cliente

Nota. Fuente: Elaboración propia.

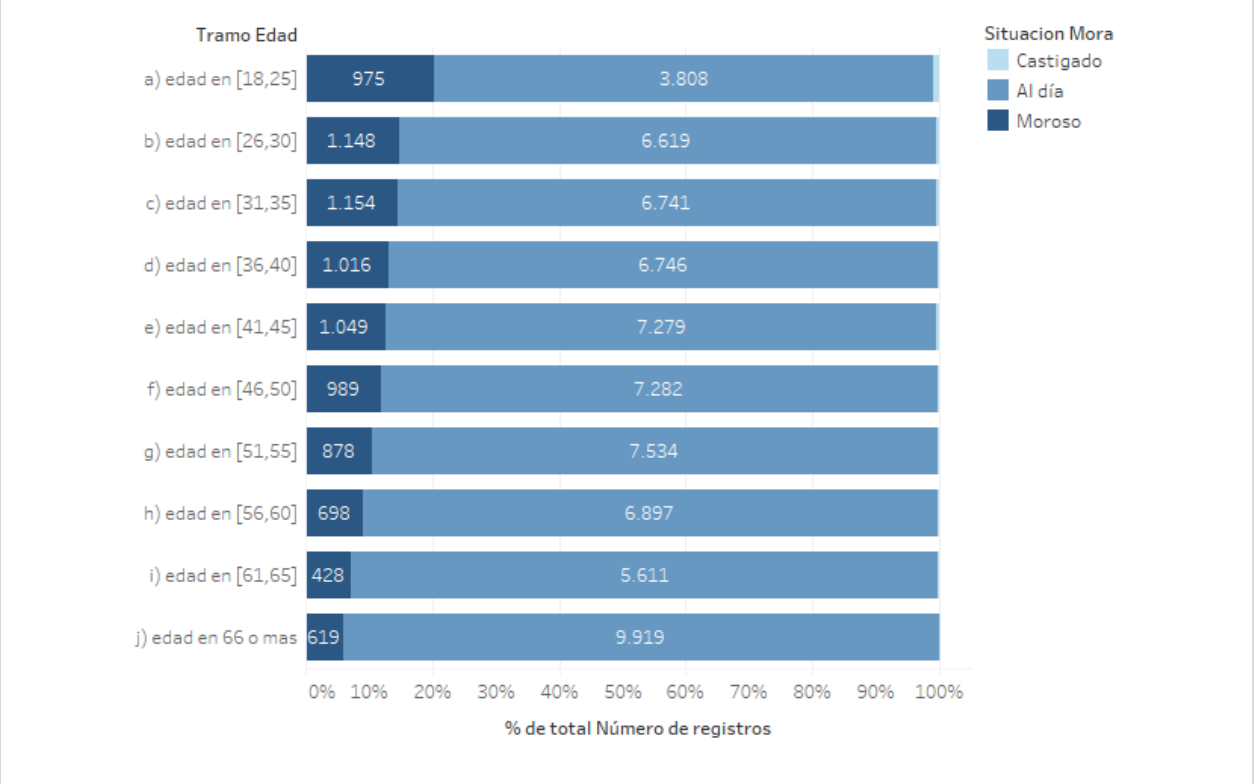


Figura 80. Situación de morosidad respecto al tramo de edad del cliente

Nota. Fuente: Elaboración propia.

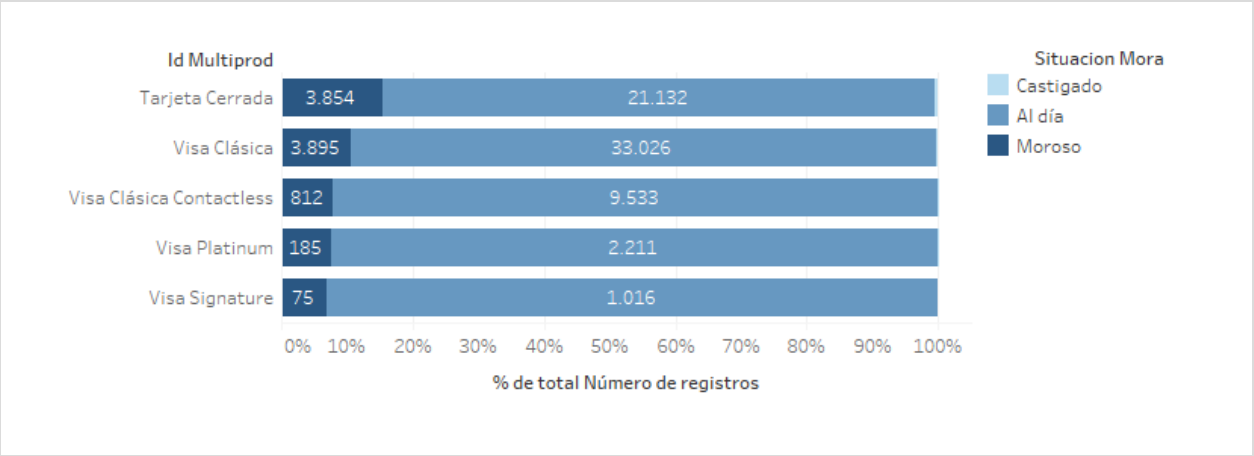


Figura 81. Situación de morosidad respecto al tipo de tarjeta del cliente

Nota. Fuente: Elaboración propia.

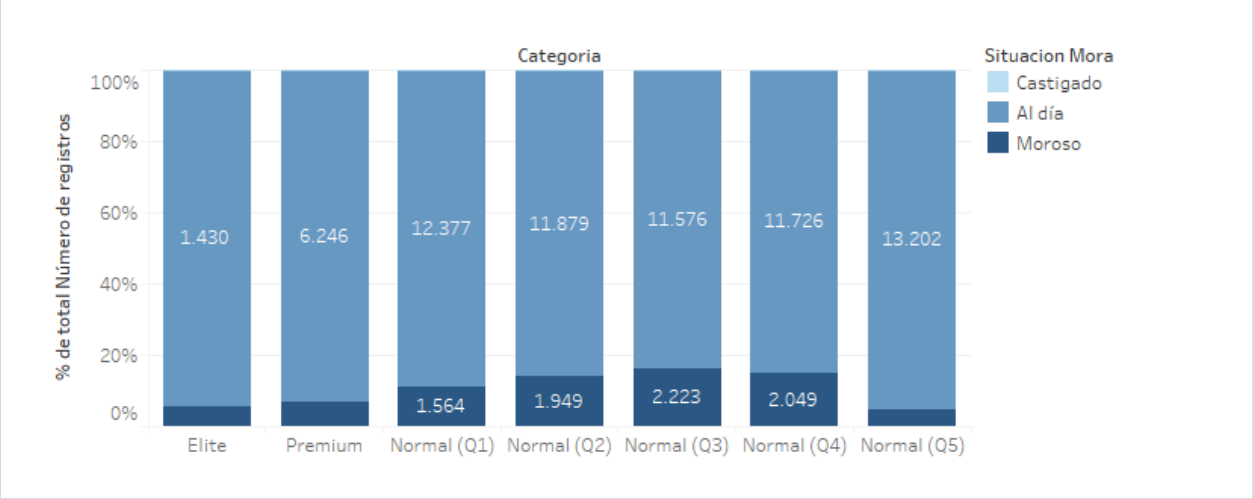


Figura 82. Situación de morosidad respecto a la categoría de cliente

Nota. Fuente: Elaboración propia.

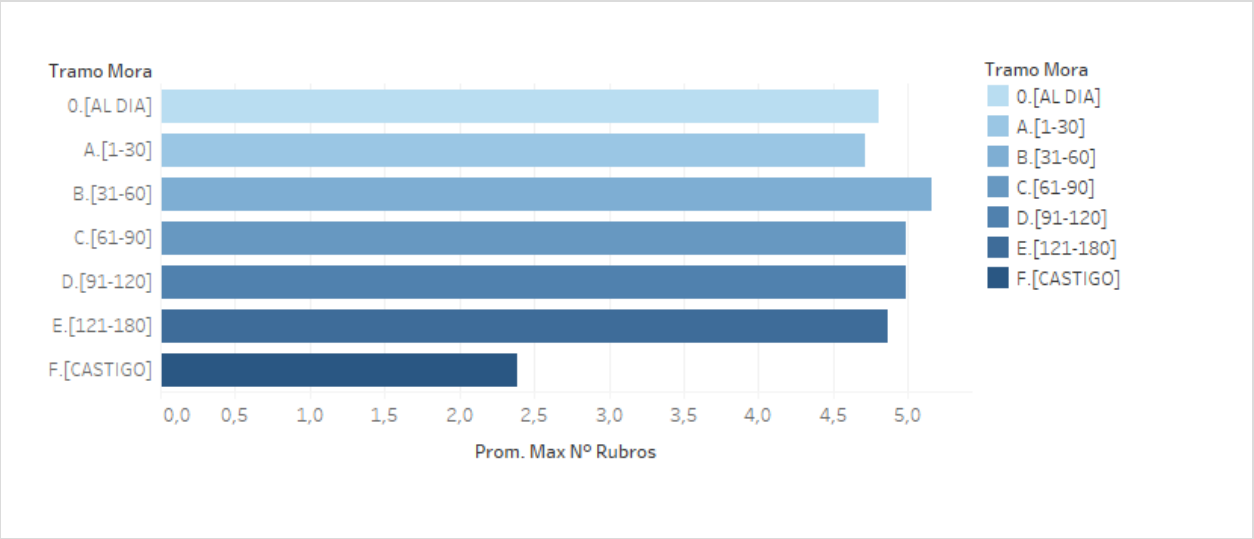


Figura 83. Tramo de morosidad versus Número de rubros promedio en el que transaccionan los clientes

Nota. Fuente: Elaboración propia.

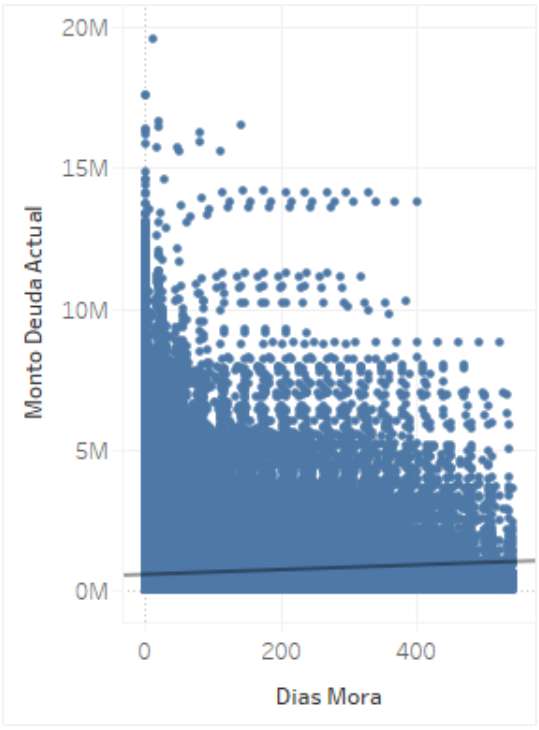


Figura 84. Monto deuda compras respecto a días mora registrados al cierre de mes

Nota. Fuente: Elaboración propia.

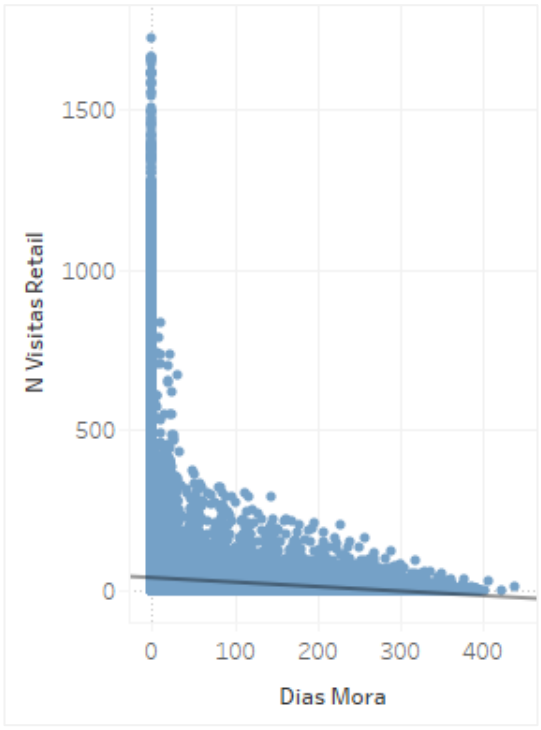


Figura 85. Frecuencia de compras respecto a días mora registrados al cierre de mes

Nota. Fuente: Elaboración propia.

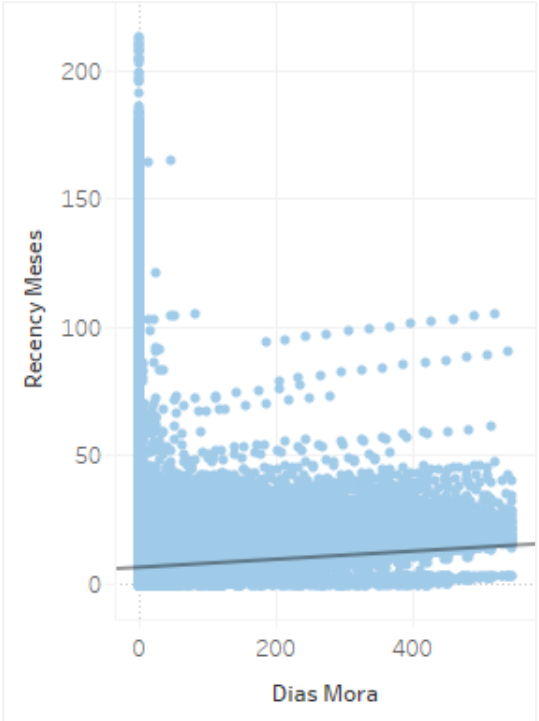


Figura 86. Recency compras respecto a días mora registrados al cierre de mes

Nota. Fuente: Elaboración propia.

ANEXO G Lista de variables disponibles base analítica

Tabla 21: Lista de variables de caracterización del cliente disponibles

Nombre de la variable	Descripción
ACTIVO_SUCURSAL	Variable binaria de actividad en sucursales.
ACTIVO_WEB	Variable binaria de actividad web.
CATEGORIA	Tipo de cliente según nivel de gasto (3 niveles).
DESC_DMC_REGION	Región de origen.
DIR_GEORREF_X	Dirección en coordenadas geográficas (Latitud).
DIR_GEORREF_Y	Dirección en coordenadas geográficas (Longitud).
EDAD	Edad del cliente.
ESTADO_CIVIL	Estado civil del cliente (Soltero, Casado, etc.).
ID_COMUTRA	Comuna de trabajo.
ID_DMC_ANTIGUEDAD	Identificador de la antigüedad del cliente.
ID_EDUCACION	Identificador del nivel de educación del cliente.
ID_HABIPAG	Identificador del hábito de pago.
ID_NUMERO_HIJAS	Identificador del número de hijas.
ID_NUMERO_HIJOS	Identificador del número de hijos.
ID_PREMIUM	Identificador de la categoría del cliente.
ID_PROFESION	Identificador de la profesión.
ID_SEXO	Identificador del sexo del cliente.
MONTO_PROM_RENTA_12M	Monto promedio de rentra declarada últ. 12 meses.

Nota. Las variables TRAMO_EDAD y FECHA_NACIMIENTO es omitida por ser transformación de las presentadas Fuente: Elaboración propia.

Tabla 22: Lista de variables de información de la cuenta disponibles

Nombre de la variable	Descripción
ANTIGUEDAD_CTA_AGNO	Antigüedad de la cuenta en años.
CANT_ADICIONALES	N°de tarjetas adicionales al titular.
CANT_PROM_CARGAS_12M	Cantidad promedio de cargas últ. 12 meses.
CUOTA_PROM_PONDE	Cuota promedio ponderada de las compras.
DISPONIBLE_PROM_12M	Disponible promedio últ. 12 meses.
FECHA_APER_CTA	Fecha de apertura de la cuenta.
FECHA_RENEGOCIA	Fecha de últ. Renegociación de la deuda.
FECHA_ULT_CUPO	Fecha de últ. Cupo registrado.
FECHA_ULT_PAGO	Fecha de últ. Pago del estado de cuenta.
ID_DMC_RANGO_CUPO	Identificador del rango del cupo de la cuenta.
ID_MULTIPROD	Identificador del tipo de tarjeta de la cuenta titular.
MONTO_PROM_CUPO_12M	Monto promedio del cupo últ. 12 meses.
MONTO_PROM_DEUDA_12M	Monto promedio deuda últ. 12 meses.
MONTO_PROM_DEUDA_FACT_12M	Monto promedio deuda facturad últ. 12 meses.
N_ADICIONALES	Número de adicionales.
PERC_PROM_DISPONIBLE_12M	Porcentaje promedio del disponible últ. 12 meses.
RECENCY_CUPO	Recencia del cupo.
RECENCY_PAGOS	Recencia de pagos.
RECENCY_RENEGOCIA	Recencia de renegociación.
TIENE_DESGRAVAMEN_12M	Variable binaria para desgravamen de la tarjeta.
TIENE_SEGURO_NO_DESG_CESTAR	Variable binaria para cesantía de la tarjeta.
TRAMO_MONTO_PROM_CUPO_12M	Tramo del monto promedio del cupo últ. 12 meses.
TRAMO_PERC_PROM_DISPONIBLE_12M	Tramo % promedio del disponible últ. 12 meses.

Nota. Fuente: Elaboración propia.

Tabla 23: Lista de variables de morosidad disponibles

Nombre de la variable	Descripción
DELINQUENCY_AMOUNT_PROM_12M	Monto promedio en mora últ. 12 meses.
DELINQUENCY_NUMBER_12M	N° meses con mora últ. 12 meses.
DELINQUENCY_RECENCY	Recencia meses últ. episodio mora .
DIAS_AL_VENCIM_PROM_12M	N°promedio de día de respecto a vencimiento.
DIAS_MORA	N° días mora últ. 12 meses.
MONTO_MORA	Monto total mora últ. 12 meses.
MONTO_MORA_PROM_12M	Monto mora promedio últ. 12 meses.
NUM_PAGOS_12M	N° pagos o abonos últ. 12 meses.

Nota. Fuente: Elaboración propia.

Tabla 24: Lista de variables de potencial de gasto[a]

Nombre de la variable	Descripción
SOW_AUTOMOTRIZ_12M	SoW rubro automotriz últ. 12 meses.
SOW_COMBUSTIBLE_12M	SoW rubro combustible últ. 12 meses.
SOW_COMUNICACIONES_12M	SoW rubro comunicaciones últ. 12 meses.
SOW_EDUCACION_12M	SoW rubro educación últ. 12 meses.
SOW_ENTRETENCION_12M	SoW rubro entretención últ. 12 meses.
SOW_FARMACIAS_12M	SoW rubro farmacias últ. 12 meses.
SOW_MEJ_HOGAR_12M	SoW rubro mejoramiento del hogar últ. 12 meses.
SOW_PROM_12M	SoW General últ. 12 meses.
SOW_RECAUDACION_12M	SoW rubro recaudación últ. 12 meses.
SOW_RESTAURANT_12M	SoW rubro restaurant últ. 12 meses.
SOW_SALUD_12M	SoW rubro salud últ. 12 meses.
SOW_SEGUROS_12M	SoW rubro seguros últ. 12 meses.
SOW_SUPERMERCADOS_12M	SoW rubro supermercados últ. 12 meses.
SOW_TIENDAS_DPTO_12M	SoW rubro tiendas departamentales últ. 12 meses.
SOW_TRANSPORTE_12M	SoW rubro transporte últ. 12 meses.
SOW_VIAJES_12M	SoW rubro viajes últ. 12 meses.
SOW_VIVIENDA_12M	SoW rubro vivienda últ. 12 meses.

Nota. [a] Share of Wallet o es el la fracción del potencial de gasto de un cliente que es consumido con la tajeta, suele ser abreviado SoW. Fuente: Elaboración propia.

Tabla 25: Lista de variables de productos financieros disponibles

Nombre de la variable	Descripción
CUOTAS_AVANCES_PROM	N° cuotas de avances promedio últ. 12 meses.
CUOTAS_AVANCES_SUM	N° cuotas de avances total últ. 12 meses.
CUOTAS_SUPERAV_PROM	N° cuotas de super avances prom. últ. 12 meses.
CUOTAS_SUPERAV_SUM	N° cuotas de super avances prom. últ. 12 meses.
DISPONIBLE_AVANCE_PROM_12M	Promedio disponible para avances últ. 12 meses.
MONTO_AVANCES_PROM	Monto promedio de avances últ. 12 meses.
MONTO_AVANCES_SUM	Monto total avances últ. 12 meses.
MONTO_PROM_CUPO_SUPAVA_12M	Monto promedio cupo sup. avances últ. 12 meses.
MONTO_PROM_DISP_SUPAVA_12M	Monto prom. Disp. super avances últ. 12 meses
MONTO_SUPERAV_PROM	Monto promedio de super avances últ. 12 meses.
MONTO_SUPERAV_SUM	Monto total super avances últ. 12 meses.
N_AVANCES	Número de Avances últ. 12 meses.
N_SUPERAV	Número de super avances últ. 12 meses.
RECENCY_AVANCES	Recencia de avances

Nota. Variable FECHA_ULT_AVANCE omitada en pos de su Recency Fuente: Elaboración propia.

Tabla 26: Lista de variables transaccionales disponibles

Nombre de la variable	Descripción
ACTIVO_NEGOCIO	Variable binaria clte. activos en el *holding*.
ACTIVO_KIOSKO	Variable binaria clte. activos en *kioskos*.
ACTIVO_ONTHEM	Variable binaria clte. activos fuera el *holding*.
ACTIVO_MEJ_HOGAR_ONUS	Variable binaria clte. activos mej. hogar *holding*.
ACTIVO_SUPERMERC_ONUS	Variable binaria clte. activos supermerc. *holding*.
ACTIVO_VIAJES	Variable binaria clientes activos en rubro viajes.
FECHA_ULT_COMPRA	Fecha última compra.
FECHA_ULT_COMPRA_12M	Fecha última compra.
MONTO_GASTO_ANIO_MOVIL	Monto gasto últ. 12 meses (año móvil).
MONTO_GASTO _ANIO_MOVIL_ADIC	Monto gasto titular últ. 12 meses (año móvil).
MONTO_GASTO _ANIO_MOVIL_TIT	Monto gasto adiciona últ. 12 meses (año móvil).
MONTO_VIAJES	Monto gasto en viajes últ. 12 meses.
N_NEGOCIOS	Número de negocios holding con transacciones.
N_VIAJES	N° compras en viajes últ. 12 meses.
N_VISITAS_TIENDA_DEPT_ONUS	N° compras tiendas Dept. *holding* últ. 12 meses.
N_VISITAS_OT	N° compras fuera del *holding* últ. 12 meses.
N_VISITAS_*RETAIL*	N° compras *retail* últ. 12 meses.
N_VISITAS_*RETAIL*_MES	N° compras promedio *retail* por mes últ. 12 meses.
N_VISITAS_MEJ_HOGAR_ONUS	N° compras mej. hogar *holding* últ. 12 meses.
N_VISITAS_SUPERMERC_ONUS	N° compras supermercados holding últ. 12 meses.
NUM_RUBROS	N° rubros distintos con compras últ. 12 meses.
PERC_GASTO_ADIC	% del gasto asociado a adicionales últ. 12 meses.
RECENCY_COMPRAS	Recencia de compras final periodo en estudio.
RECENCY_PROM_12M	Recencia promedio de compras últ. 12 meses.
SUM_MONTO_TIENDA_DEPT_ONUS	Monto gasto Tienda Dptos. *holding* últ. 12 meses.
SUM_MONTO_OT	Monto gasto fuera *holding* últ. 12 meses.
SUM_MONTO_*RETAIL*	Monto gasto rubro *retail* últ. 12 meses.
SUM_MONTO_MEJ_HOGAR_ONUS	Monto gasto en mej. hogar. *holding* últ. 12 meses.
SUM_MONTO_SUPERMERC_ONUS	Monto gasto en supermerc. *holding* últ. 12 meses.
TICKET_PROM_RETAIL	Ticket promedio rubro *retail* últ. 12 meses
TRAMO_PROM_RECENCY_12M	Tramo de recencia de compras últ. 12 meses

Nota. Fuente: Elaboración propia.

Tabla 27: Lista de variables de fidelización disponibles

Nombre de la variable	Descripción
CON_CANJES	Variable binaria para clientes con canjes.
N_CANJES	N° canjes en el último año.
STOCK_PUNTOS	N° puntos acumulados en stock.
SUM_PTOS_CANJE	Total puntos canjeados en últ. año.

Nota. Fuente: Elaboración propia.

Tabla 28: Lista de variables identificadoras

Nombre de la variable	Descripción
ID_DMC_CLIENTE	Identificador anonimizado del cliente.
CONTRATO	N° de contrato de la cuenta.
ID_AGNO_MES	Identificador del año y mes de la observación.

Nota. Fuente: Elaboración propia.

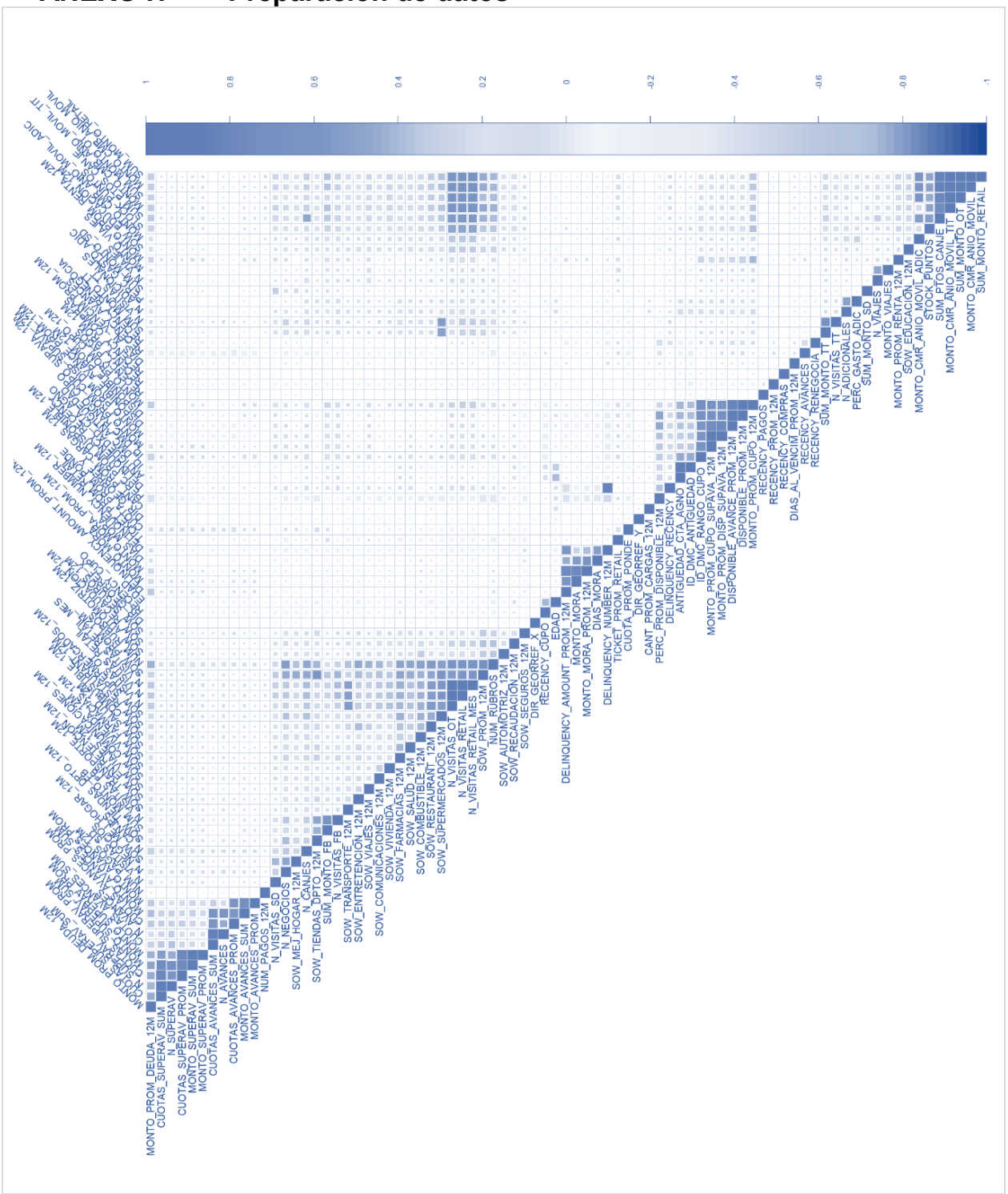


Figura 87. Matrices de correlaciones para base analítica

Nota. Fuente: Elaboración propia.

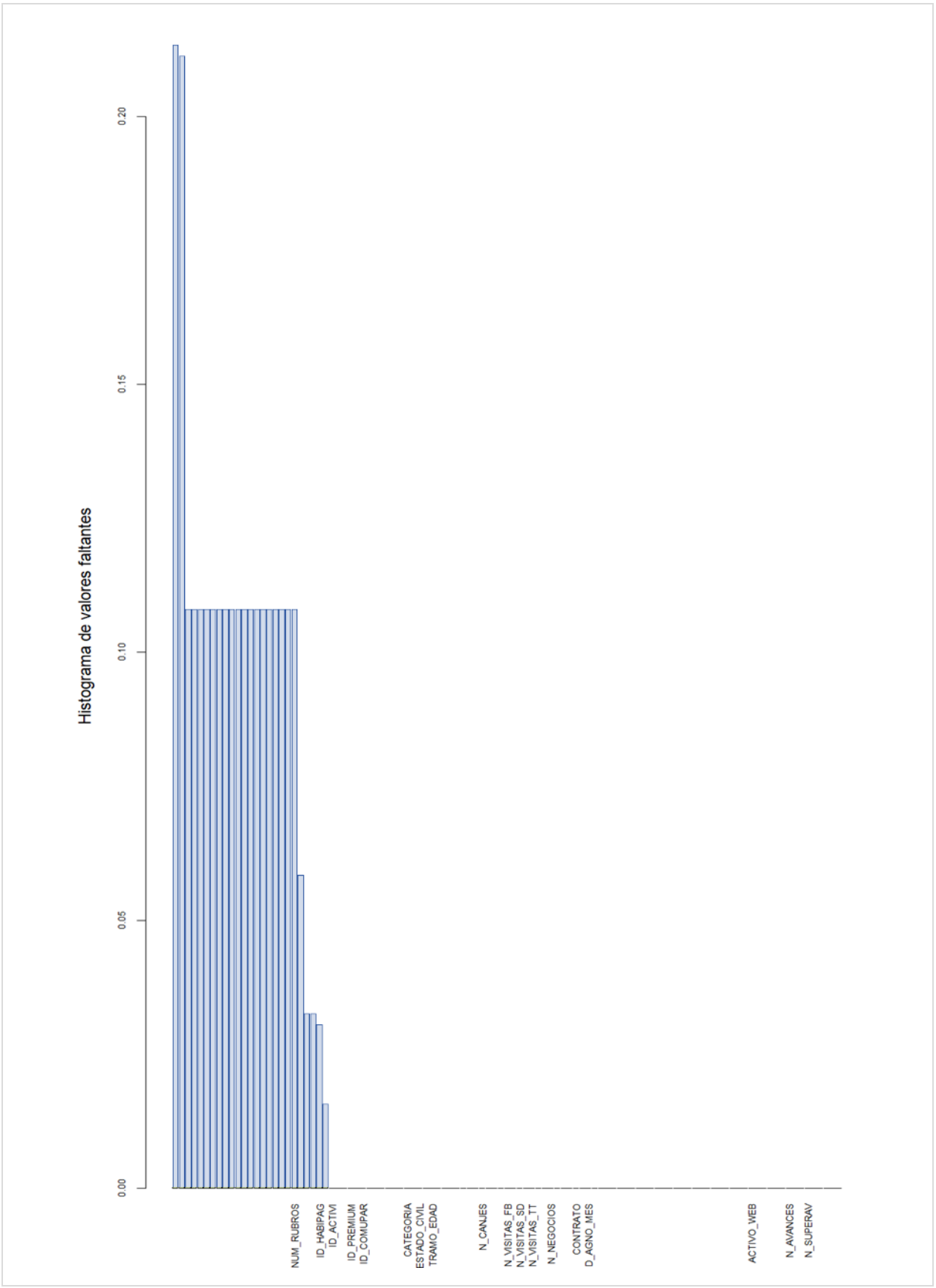


Figura 88. Histograma de valores perdidos por variable

Nota. Fuente: Elaboración propia.

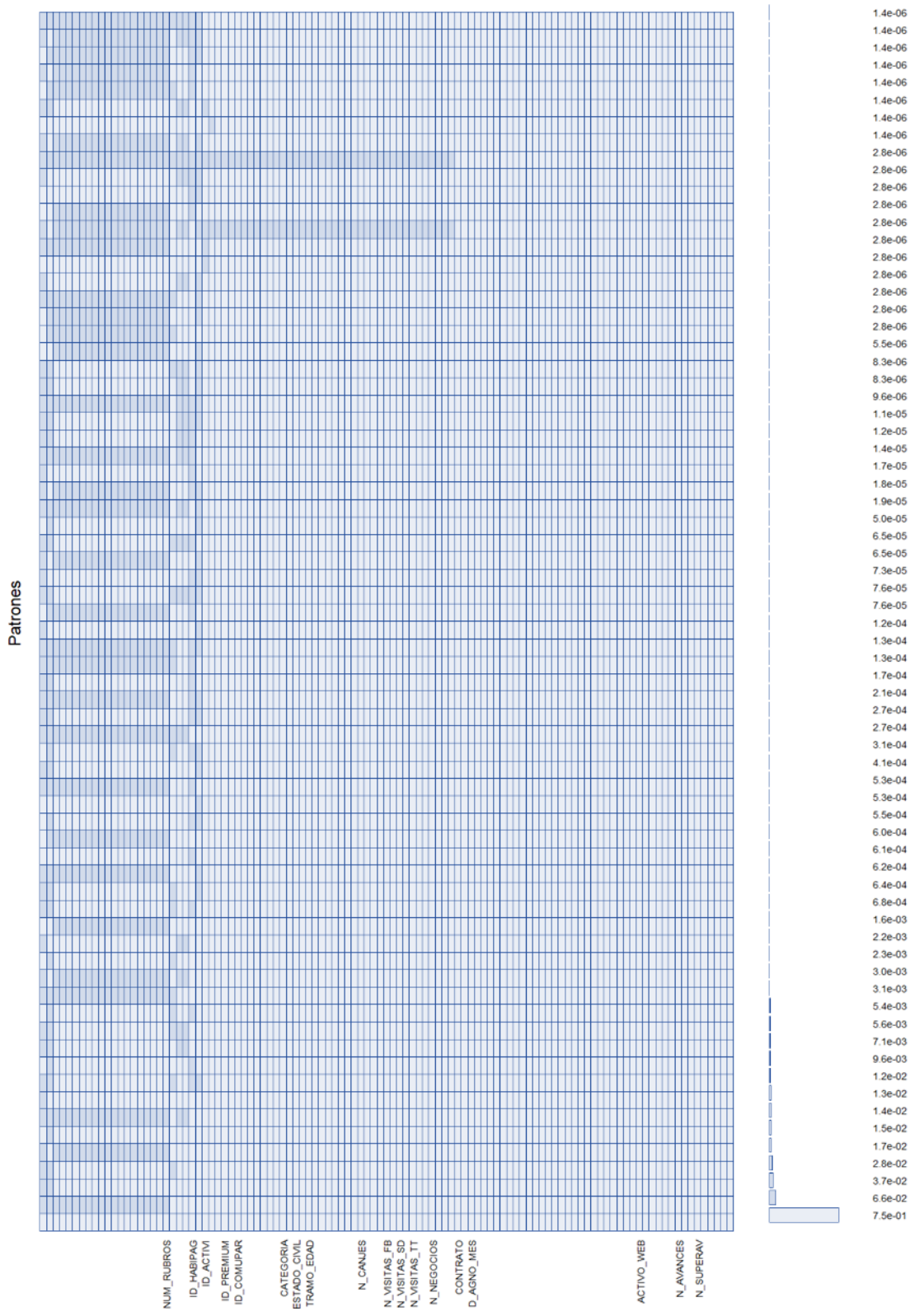


Figura 89. Patrones de valores perdidos según frecuencia relativa
Nota. Fuente: Elaboración propia.

ANEXO I Hard clustering K-means

Tabla 29: Centros hard clustering (k-means), caracterización extendida

Dimensión	Variable	Segmento: Infrecuentes con deterioro pausado 1	Inactivos forzados con deterioro acelerado 2	Activos esporádicos con deterioro moderado 3	Activos Saturados con deterioro pausado 4	Inactivos forzados de deterioro progresivo 5	Activos insaturados con deterioro pausado 6
Tamaño	# de Clientes	73.487	14.100	58.429	131.396	2.561	260.455
Morosidad	# episodios Mora	0,8	9,6	6,3	0,9	10,0	0,4
	Monto mora	$2.624	$285.635	$62.290	$4.532	$1.418.335	$1.853
	# Días mora	1,0	225,8	21,3	0,9	219,1	0,3
	Recencia de pagos (meses)	2,0	6,8	1,0	0,2	5,9	0,6
Saturación	% del Disponible	85%	1%	33%	39%	-58%	90%
	Monto Cupo	$743.340	$428.502	$631.105	$734.772	$1.312.167	$1.645.391
Potencial	Share of Wallet	14%	25%	22%	35%	23%	20%
Demográfico	Edad	44,4	38,0	42,9	43,9	42,8	50,9
	% Mujeres	52%	45%	52%	56%	37%	50%
Contrato	% Tarjeta Cerrada[a]	49%	47%	44%	22%	4%	18%
	% Tarjeta Visa[b]	51%	53%	56%	78%	96%	82%
	% con Adicionales	9%	6%	11%	14%	15%	22%
Riesgo-Rentabilidad	Score	166	162	307	504	203	741
	Cuota Promedio	2,4	2,6	3,0	3,7	2,6	2,8
Productos Financieros	% con Super Avance (SAv)	2%	4%	7%	20%	19%	13%
	% con Avance (Av)	11%	18%	23%	36%	24%	24%
	Monto Super Avance (SAv)	$15.601	$29.586	$73.506	$183.264	$184.871	$147.379
	Monto Avance (Av)	$16.368	$22.705	$33.841	$54.120	$75.296	$66.188
Fidelización	% Clientes con canje	5%	19%	25%	50%	56%	46%
Transaccional	Recencia compras (meses)	3,4	9,3	4,1	0,6	10,1	1,7
	Monto gasto año móvil	$368.132	$252.803	$601.798	$1.910.329	$424.187	$1.518.418
	Ticket promedio *Retail*	$64.886	$27.595	$41.160	$44.455	$29.254	$55.025

Nota. [a] Tarjeta que solo se encuentra habilitada para compras en negocios del holding y en alianzas con la empresa. [b] Tarjeta habilitada para compras en comercios que aceptan pagos con tarjetas. Fuente: Elaboracion propia.

ANEXO J Validación externa de las clusterización obtenidas

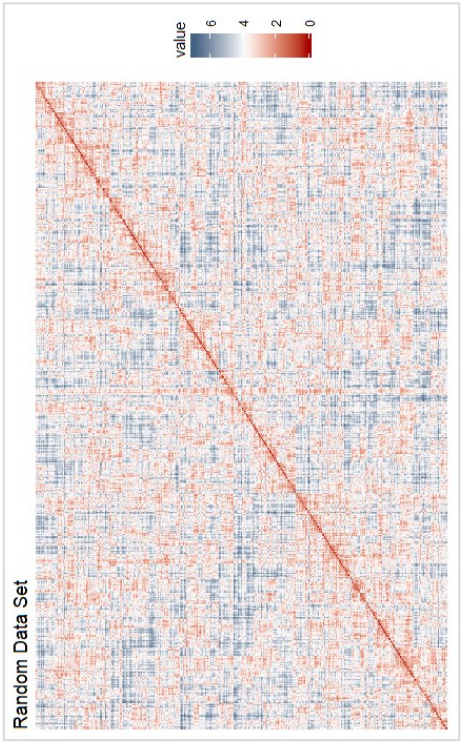


Figura 90. Matriz de disimilitud sobre data set aleatorio (k-means)

Nota. Fuente:Elaboración propia.

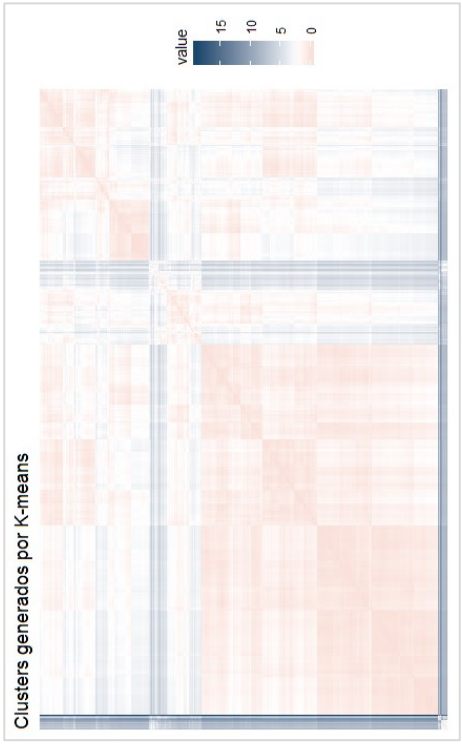


Figura 91. Matriz de disimilitud sobre resultados k-means

Nota. N° de segmentos $k = 6$. Fuente:Elaboración propia.

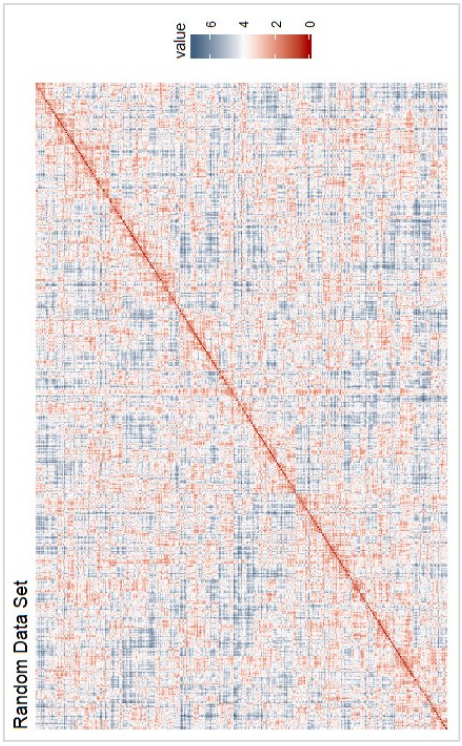


Figura 92. Matriz de disimilitud sobre data set aleatorio (c-means)

Nota. Fuente:Elaboración propia.

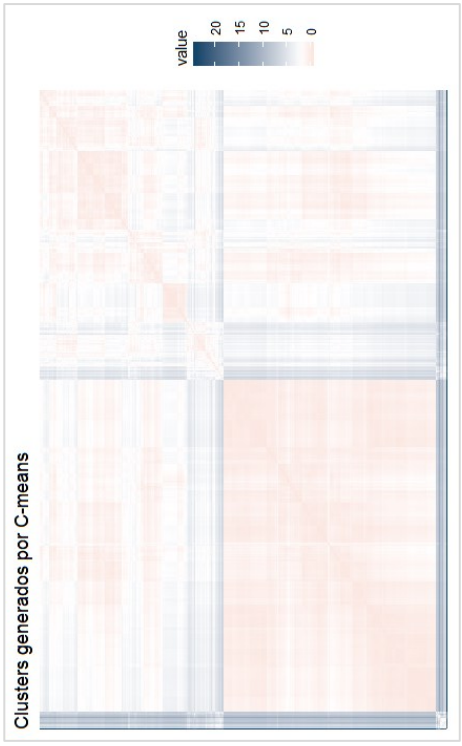


Figura 93. Matriz de disimilitud sobre resultados c-means

Nota. N° de segmentos $k = 7$. Fuente:Elaboración propia.

ANEXO K Detalle ajuste de modelos sobre hard clustering

Tabla 30: Resultados ajuste Modelo MK1 por segmentos vía hard clustering

Clúster	Error absoluto promedio	Error prom. ponderado	Error prom. ponderado simétrico
clust_1	3,09%	0,04%	0,01%
clust_2	20,50%	0,32%	0,06%
clust_3	18,38%	0,04%	0,01%
clust_4	3,36%	0,03%	0,00%
clust_5	18,65%	0,56%	0,22%
clust_6	1,32%	0,02%	0,00%
Promedio Modelo	**10,88%**	**0,17%**	**0,05%**

Nota. Modelo MK1 considera por variables de estado los tramos de morosidad en una granularidad de 60 días y una ventana de tiempo de ajsute de $n = 2$. Fuente: Elaboración propia.

Tabla 31: Resultados ajuste Modelo MK2 por segmentos vía hard clustering

Clúster	Error absoluto promedio	Error prom. ponderado	Error prom. ponderado simétrico
clust_1	1,87%	0,02%	0,01%
clust_2	9,50%	0,15%	0,06%
clust_3	6,71%	0,02%	0,01%
clust_4	2,41%	0,00%	0,00%
clust_5	10,58%	0,48%	0,31%
clust_6	0,79%	0,00%	0,00%
Promedio Modelo	**5,31%**	**0,11%**	**0,07%**

Nota. Modelo MK1 considera por variables de estado los tramos de morosidad en una granularidad de 30 días y una ventana de tiempo de ajsute de $n = 1$. Fuente: Elaboración propia.

Tabla 32: Resultados ajuste Modelo MK3 por segmentos vía hard clustering

Clúster	Error absoluto promedio	Error prom. ponderado	Error prom. ponderado simétrico
clust_1	1,97%	0,03%	0,02%
clust_2	10,19%	0,22%	0,09%
clust_3	7,20%	0,03%	0,01%
clust_4	2,43%	0,01%	0,01%
clust_5	11,83%	0,92%	0,54%
clust_6	0,80%	0,00%	0,00%
Promedio Modelo	**5,74%**	**0,20%**	**0,11%**

Nota. Modelo MK1 considera por variables de estado los tramos de morosidad y tiempo consecutivo en mora en una granularidad de 60 días y una ventana de tiempo de ajsute de $n = 1$. Fuente: Elaboración propia.

Tabla 33: Resultados ajuste Modelo MK4 por segmentos vía hard clustering

Clúster	Error absoluto promedio	Error prom. ponderado	Error prom. ponderado simétrico
clust_1	2,11%	0,05%	0,04%
clust_2	10,26%	0,45%	0,21%
clust_3	7,60%	0,05%	0,02%
clust_4	2,58%	0,01%	0,01%
clust_5	13,28%	1,70%	1,32%
clust_6	0,84%	0,01%	0,01%
Promedio Modelo	**6,11%**	**0,38%**	**0,27%**

Nota. Modelo MK1 considera por variables de estado los tramos de morosidad y tiempo consecutivo en mora en una granularidad de 30 días y una ventana de tiempo de ajsute de $n = 1$. Fuente: Elaboración propia.

ANEXO L Matrices de transición para modelos de dos variables (MK3)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	95,71%	4,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	78,58%	0,00%	21,42%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	32,44%	0,00%	0,00%	23,82%	0,00%	0,00%	0,00%	43,74%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	32,92%	0,00%	0,00%	0,00%	36,82%	0,00%	0,00%	0,00%	30,26%	0,00%	0,00%	0,00%	0,00%	0,00%
4	33,13%	0,00%	0,00%	0,00%	0,00%	37,99%	0,00%	0,00%	0,00%	28,88%	0,00%	0,00%	0,00%	0,00%
5	29,04%	0,00%	0,00%	0,00%	0,00%	0,00%	42,65%	0,00%	0,00%	0,00%	28,31%	0,00%	0,00%	0,00%
6	24,58%	0,00%	0,00%	0,00%	0,00%	0,00%	51,03%	0,00%	0,00%	0,00%	24,39%	0,00%	0,00%	0,00%
7	20,08%	0,00%	0,00%	0,00%	2,11%	0,00%	0,00%	0,00%	77,81%	0,00%	0,00%	0,00%	0,00%	0,00%
8	8,28%	0,00%	0,00%	0,00%	0,00%	1,89%	0,00%	0,00%	0,00%	19,53%	0,00%	70,30%	0,00%	0,00%
9	12,88%	0,00%	0,00%	0,00%	0,00%	0,00%	4,85%	0,00%	0,00%	0,00%	33,44%	0,00%	48,83%	0,00%
10	17,08%	0,00%	0,00%	0,00%	0,00%	0,00%	4,75%	0,00%	0,00%	0,00%	34,16%	0,00%	44,00%	0,00%
11	5,84%	0,00%	0,00%	0,00%	0,00%	0,00%	0,82%	0,00%	0,00%	0,00%	0,82%	0,00%	81,17%	11,34%
12	7,53%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%	0,00%	0,00%	0,91%	0,00%	26,71%	64,73%
13	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 94. Matriz representante cartera completa, modelo MK3

Nota. Fuente: Elaboración propia.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	94,45%	5,55%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	79,13%	0,00%	20,87%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	65,48%	0,00%	0,00%	7,59%	0,00%	0,00%	0,00%	26,93%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	62,06%	0,00%	0,00%	0,00%	22,76%	0,00%	0,00%	0,00%	15,18%	0,00%	0,00%	0,00%	0,00%	0,00%
4	67,75%	0,00%	0,00%	0,00%	0,00%	14,59%	0,00%	0,00%	0,00%	17,65%	0,00%	0,00%	0,00%	0,00%
5	76,54%	0,00%	0,00%	0,00%	0,00%	0,00%	11,61%	0,00%	0,00%	0,00%	11,85%	0,00%	0,00%	0,00%
6	23,69%	0,00%	0,00%	0,00%	0,00%	0,00%	24,97%	0,00%	0,00%	2,50%	48,84%	0,00%	0,00%	0,00%
7	63,47%	0,00%	0,00%	0,00%	2,15%	0,00%	0,00%	0,00%	34,37%	0,00%	0,00%	0,00%	0,00%	0,00%
8	74,34%	0,00%	0,00%	0,00%	0,00%	1,78%	0,00%	0,00%	0,00%	4,80%	0,00%	19,08%	0,00%	0,00%
9	86,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,21%	0,00%	0,00%	0,00%	9,70%	0,00%	4,09%	0,00%
10	35,71%	0,00%	0,00%	0,00%	0,00%	0,00%	4,04%	0,00%	0,00%	0,00%	36,29%	0,00%	23,96%	0,00%
11	70,48%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,34%	0,00%	28,34%	0,83%
12	63,49%	0,00%	0,00%	0,00%	0,00%	0,00%	1,11%	0,00%	0,00%	0,00%	0,83%	0,00%	15,40%	19,17%
13	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 95. Matriz representante clúster 1, modelo MK3

Nota. Fuente: Elaboración propia.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	37,57%	62,43%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	13,21%	0,00%	86,79%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	6,46%	0,00%	0,00%	10,30%	0,00%	0,00%	0,00%	83,24%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	5,79%	0,00%	0,00%	0,00%	21,19%	0,00%	0,00%	0,00%	73,02%	0,00%	0,00%	0,00%	0,00%	0,00%
4	5,87%	0,00%	0,00%	0,00%	0,00%	21,10%	0,00%	0,00%	0,00%	73,04%	0,00%	0,00%	0,00%	0,00%
5	5,59%	0,00%	0,00%	0,00%	0,00%	0,00%	27,86%	0,00%	0,00%	0,00%	66,56%	0,00%	0,00%	0,00%
6	9,38%	0,00%	0,00%	0,00%	0,00%	0,00%	25,95%	0,00%	0,00%	0,00%	64,67%	0,00%	0,00%	0,00%
7	7,94%	0,00%	0,00%	0,00%	1,69%	0,00%	0,00%	0,00%	90,37%	0,00%	0,00%	0,00%	0,00%	0,00%
8	4,63%	0,00%	0,00%	0,00%	0,00%	0,98%	0,00%	0,00%	0,00%	26,94%	0,00%	67,46%	0,00%	0,00%
9	3,37%	0,00%	0,00%	0,00%	0,00%	0,00%	1,69%	0,00%	0,00%	0,00%	34,26%	0,00%	60,68%	0,00%
10	3,59%	0,00%	0,00%	0,00%	0,00%	0,00%	2,94%	0,00%	0,00%	0,00%	34,42%	0,00%	59,06%	0,00%
11	9,89%	0,00%	0,00%	0,00%	0,00%	0,00%	0,94%	0,00%	0,00%	0,00%	1,27%	0,00%	78,55%	9,34%
12	3,79%	0,00%	0,00%	0,00%	0,00%	0,00%	0,18%	0,00%	0,00%	0,00%	0,71%	0,00%	48,15%	47,16%
13	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 96. Matriz representante clúster 2, modelo MK3

Nota. Fuente: Elaboración propia.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	57,96%	42,04%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	61,00%	0,00%	39,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	32,15%	0,00%	0,00%	33,17%	0,00%	0,00%	0,00%	34,68%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	32,05%	0,00%	0,00%	0,00%	45,27%	0,00%	0,00%	0,00%	22,68%	0,00%	0,00%	0,00%	0,00%	0,00%
4	30,21%	0,00%	0,00%	0,00%	0,00%	47,40%	0,00%	0,00%	0,00%	22,39%	0,00%	0,00%	0,00%	0,00%
5	29,51%	0,00%	0,00%	0,00%	0,00%	0,00%	49,36%	0,00%	0,00%	0,00%	21,13%	0,00%	0,00%	0,00%
6	23,84%	0,00%	0,00%	0,00%	0,00%	0,00%	53,59%	0,00%	0,00%	0,00%	22,57%	0,00%	0,00%	0,00%
7	40,35%	0,00%	0,00%	0,00%	5,58%	0,00%	0,00%	0,00%	54,08%	0,00%	0,00%	0,00%	0,00%	0,00%
8	27,55%	0,00%	0,00%	0,00%	0,00%	6,21%	0,00%	0,00%	0,00%	27,99%	0,00%	38,25%	0,00%	0,00%
9	32,92%	0,00%	0,00%	0,00%	0,00%	0,00%	9,33%	0,00%	0,00%	0,00%	34,53%	0,00%	23,22%	0,00%
10	30,77%	0,00%	0,00%	0,00%	0,00%	0,00%	10,14%	0,00%	0,00%	0,00%	34,10%	0,00%	24,98%	0,00%
11	45,80%	0,00%	0,00%	0,00%	0,00%	0,00%	3,03%	0,00%	0,00%	0,00%	2,15%	0,00%	46,68%	2,34%
12	41,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,88%	0,00%	0,00%	0,00%	6,74%	0,00%	38,18%	13,20%
13	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 97. Matriz representante clúster 3, modelo MK3

Nota. Fuente: Elaboración propia.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	93,58%	6,42%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	90,16%	0,00%	9,84%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	75,88%	0,00%	0,00%	19,17%	0,00%	0,00%	0,00%	4,95%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	85,50%	0,00%	0,00%	0,00%	13,72%	0,00%	0,00%	0,00%	0,78%	0,00%	0,00%	0,00%	0,00%	0,00%
4	92,39%	0,00%	0,00%	0,00%	0,00%	7,61%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	43,67%	0,00%	0,00%	0,00%	0,00%	0,00%	47,96%	0,00%	0,00%	0,00%	8,38%	0,00%	0,00%	0,00%
6	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	95,56%	0,00%	0,00%	0,00%	0,77%	0,00%	0,00%	0,00%	3,67%	0,00%	0,00%	0,00%	0,00%	0,00%
8	62,19%	0,00%	0,00%	0,00%	0,00%	3,15%	0,00%	0,00%	0,00%	16,78%	0,00%	17,89%	0,00%	0,00%
9	48,92%	0,00%	0,00%	0,00%	0,00%	0,00%	5,48%	0,00%	0,00%	0,00%	27,37%	0,00%	18,23%	0,00%
10	40,52%	0,00%	0,00%	0,00%	0,00%	0,00%	4,71%	0,00%	0,00%	0,00%	33,19%	0,00%	21,58%	0,00%
11	53,42%	0,00%	0,00%	0,00%	0,00%	0,00%	0,87%	0,00%	0,00%	0,00%	1,08%	0,00%	41,54%	3,09%
12	42,18%	0,00%	0,00%	0,00%	0,00%	0,00%	0,81%	0,00%	0,00%	0,00%	3,02%	0,00%	31,03%	22,96%
13	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 98. Matriz representante clúster 4, modelo MK3

Nota. Fuente: Elaboración propia.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	50,01%	49,99%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	14,57%	0,00%	85,43%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	5,03%	0,00%	0,00%	14,64%	0,00%	0,00%	0,00%	80,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	4,95%	0,00%	0,00%	0,00%	27,86%	0,00%	0,00%	0,00%	67,20%	0,00%	0,00%	0,00%	0,00%	0,00%
4	6,59%	0,00%	0,00%	0,00%	0,00%	30,32%	0,00%	0,00%	0,00%	63,09%	0,00%	0,00%	0,00%	0,00%
5	5,04%	0,00%	0,00%	0,00%	0,00%	0,00%	29,82%	0,00%	0,00%	0,00%	65,13%	0,00%	0,00%	0,00%
6	9,10%	0,00%	0,00%	0,00%	0,00%	0,00%	22,41%	0,00%	0,00%	0,00%	68,48%	0,00%	0,00%	0,00%
7	6,52%	0,00%	0,00%	0,00%	3,01%	0,00%	0,00%	0,00%	90,47%	0,00%	0,00%	0,00%	0,00%	0,00%
8	3,41%	0,00%	0,00%	0,00%	0,00%	1,41%	0,00%	0,00%	0,00%	32,14%	0,00%	63,04%	0,00%	0,00%
9	4,74%	0,00%	0,00%	0,00%	0,00%	0,00%	0,95%	0,00%	0,00%	0,00%	36,49%	0,00%	57,82%	0,00%
10	5,18%	0,00%	0,00%	0,00%	0,00%	0,00%	1,76%	0,00%	0,00%	0,00%	36,22%	0,00%	56,84%	0,00%
11	9,79%	0,00%	0,00%	0,00%	0,00%	0,00%	0,11%	0,00%	0,00%	0,00%	1,25%	0,00%	81,11%	7,73%
12	8,57%	0,00%	0,00%	0,00%	0,00%	0,00%	0,08%	0,00%	0,00%	0,00%	1,82%	0,00%	49,26%	40,27%
13	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 99. Matriz representante clúster 5, modelo MK3

Nota. Fuente: Elaboración propia.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	97,47%	2,53%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	93,07%	0,00%	6,93%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	69,62%	0,00%	0,00%	21,31%	0,00%	0,00%	0,00%	9,07%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	74,36%	0,00%	0,00%	0,00%	23,58%	0,00%	0,00%	0,00%	2,06%	0,00%	0,00%	0,00%	0,00%	0,00%
4	76,37%	0,00%	0,00%	0,00%	0,00%	22,72%	0,00%	0,00%	0,00%	0,91%	0,00%	0,00%	0,00%	0,00%
5	83,06%	0,00%	0,00%	0,00%	0,00%	0,00%	16,94%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	61,01%	0,00%	0,00%	0,00%	0,00%	0,00%	25,74%	0,00%	0,00%	0,72%	12,54%	0,00%	0,00%	0,00%
7	90,21%	0,00%	0,00%	0,00%	1,62%	0,00%	0,00%	0,00%	8,17%	0,00%	0,00%	0,00%	0,00%	0,00%
8	80,10%	0,00%	0,00%	0,00%	0,00%	3,54%	0,00%	0,00%	0,00%	5,92%	0,00%	10,44%	0,00%	0,00%
9	48,54%	0,00%	0,00%	0,00%	0,00%	0,00%	12,37%	0,00%	0,00%	0,00%	21,00%	0,00%	18,09%	0,00%
10	30,91%	0,00%	0,00%	0,00%	0,00%	0,00%	5,25%	0,00%	0,00%	0,00%	34,58%	0,00%	29,26%	0,00%
11	53,90%	0,00%	0,00%	0,00%	0,00%	0,00%	0,92%	0,00%	0,00%	0,00%	1,05%	0,00%	41,24%	2,88%
12	41,88%	0,00%	0,00%	0,00%	0,00%	0,00%	0,80%	0,00%	0,00%	0,00%	2,79%	0,00%	30,83%	23,71%
13	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 100. Matriz representante clúster 6, modelo MK3

Nota. Fuente: Elaboración propia.

ANEXO M Detalles representación en grafo modelo MK3

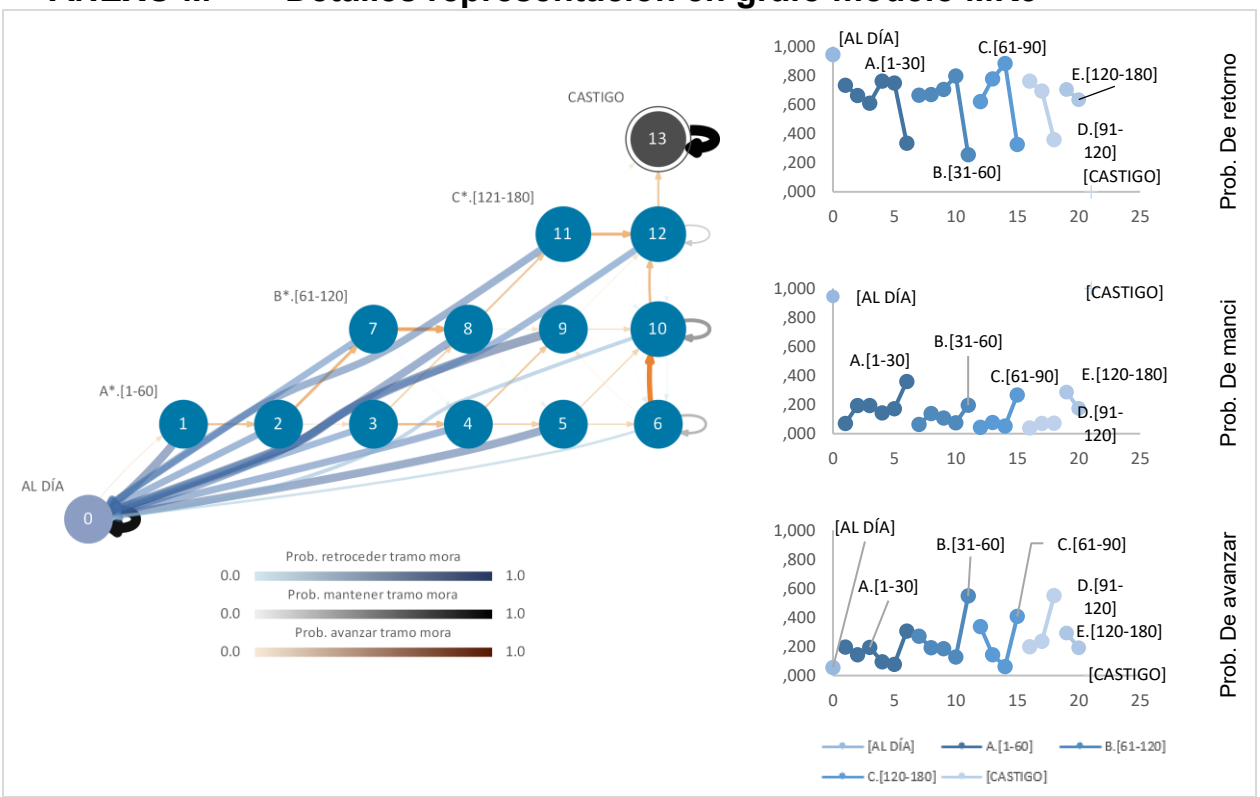


Figura 101. Detalle representación en grafo clúster 1, modelo MK3

Nota. Fuente: Elaboración propia.

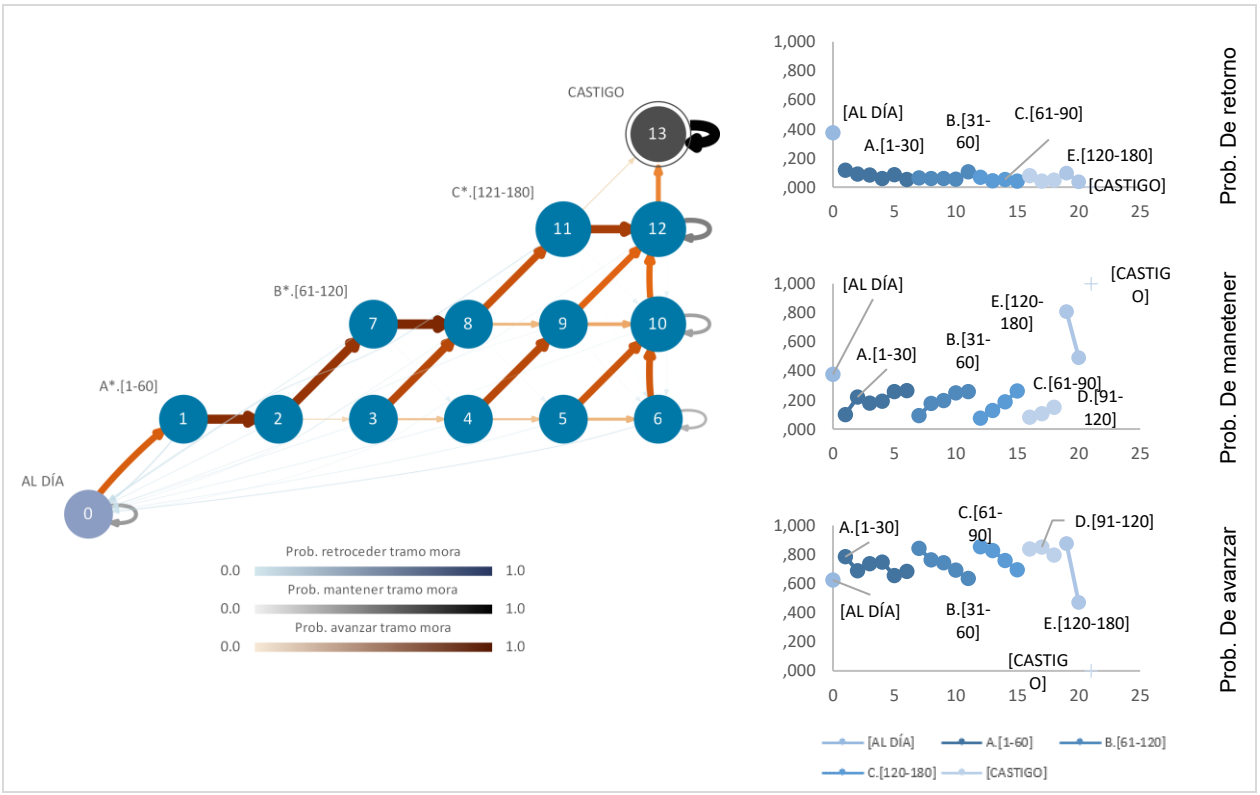


Figura 102. Detalle representación en grafo clúster 2, modelo MK3

Nota. Fuente: Elaboración propia.

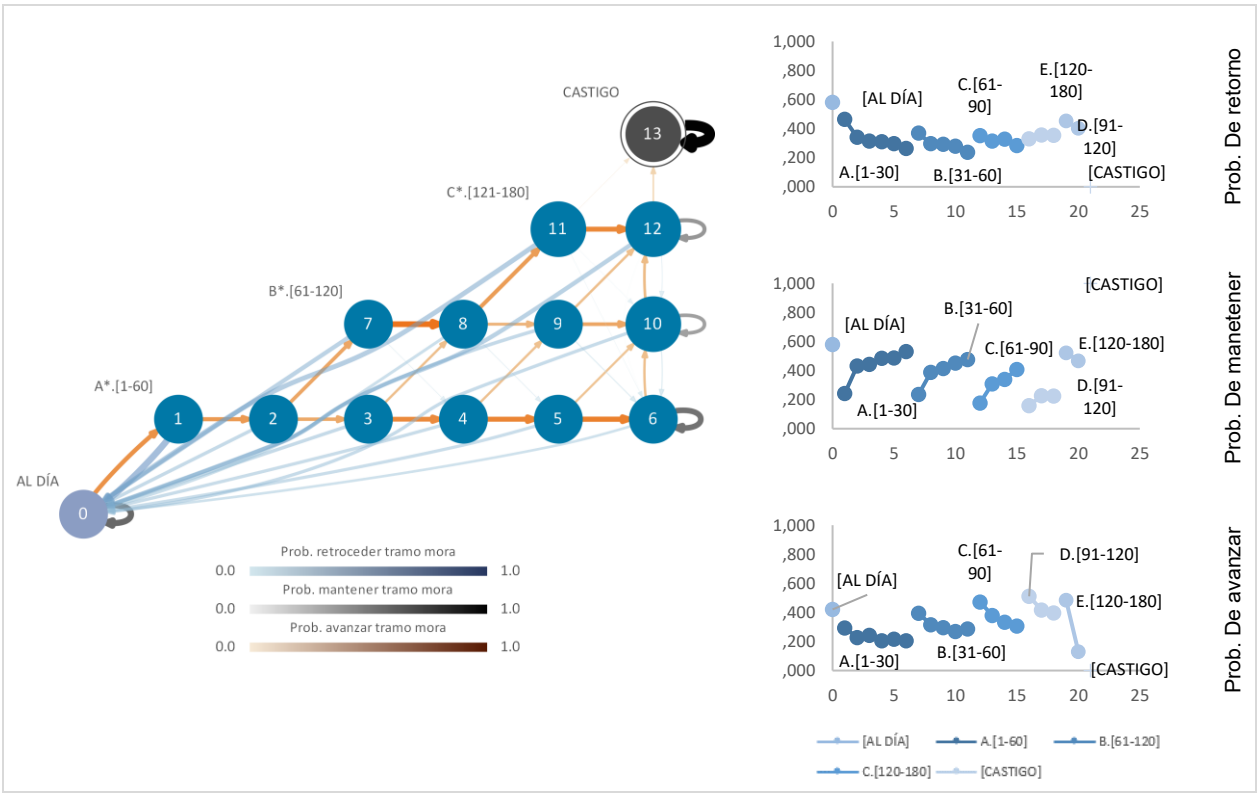


Figura 103. Detalle representación en grafo clúster 3, modelo MK3

Nota. Fuente: Elaboración propia.

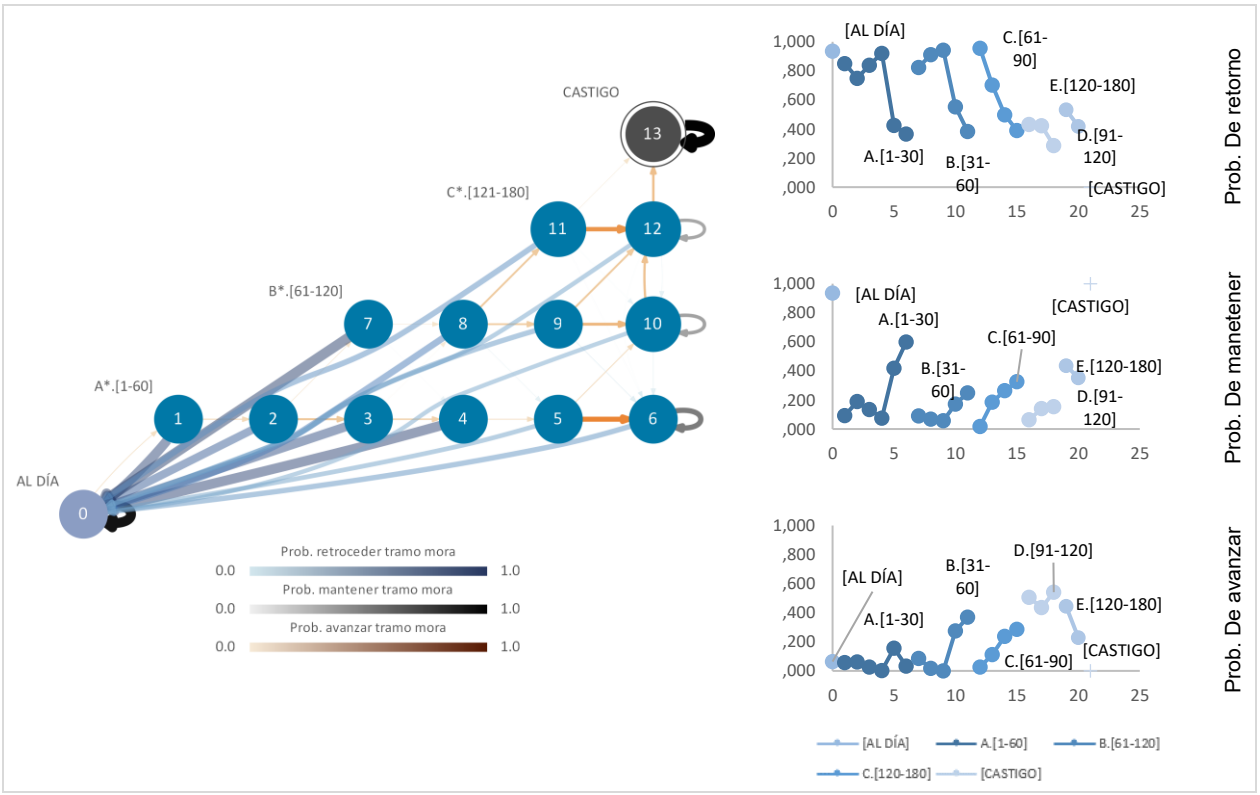


Figura 104. Detalle representación en grafo clúster 4, modelo MK3

Nota. Fuente: Elaboración propia.

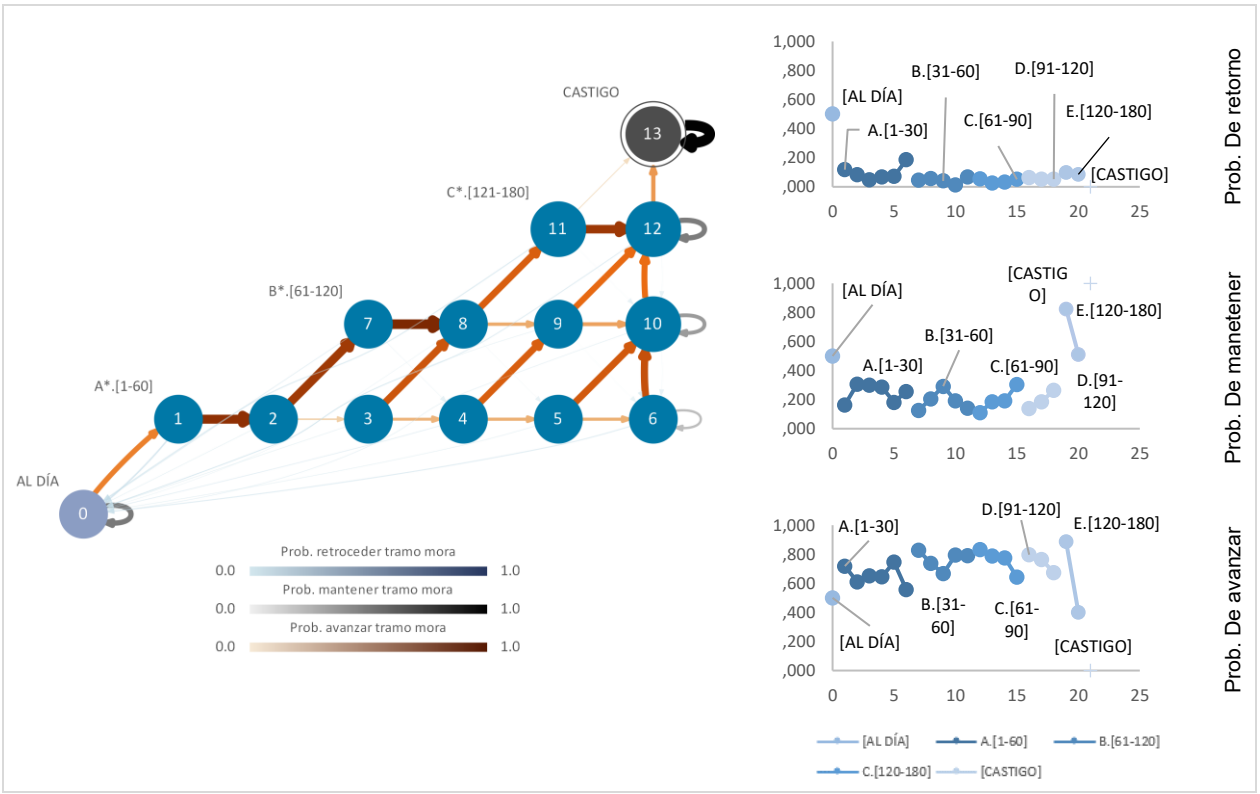


Figura 105. Detalle representación en grafo clúster 5, modelo MK3

Nota. Fuente: Elaboración propia.

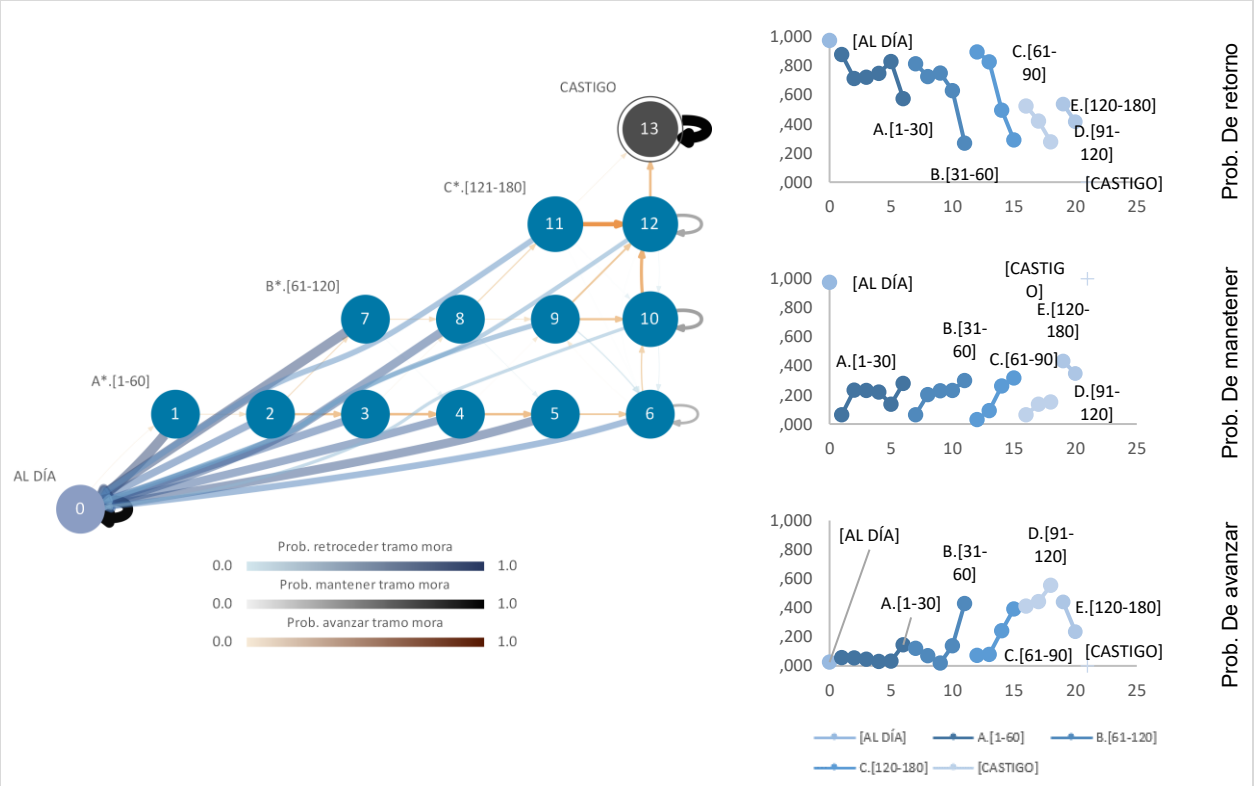


Figura 106. Detalle representación en grafo clúster 6, modelo MK3

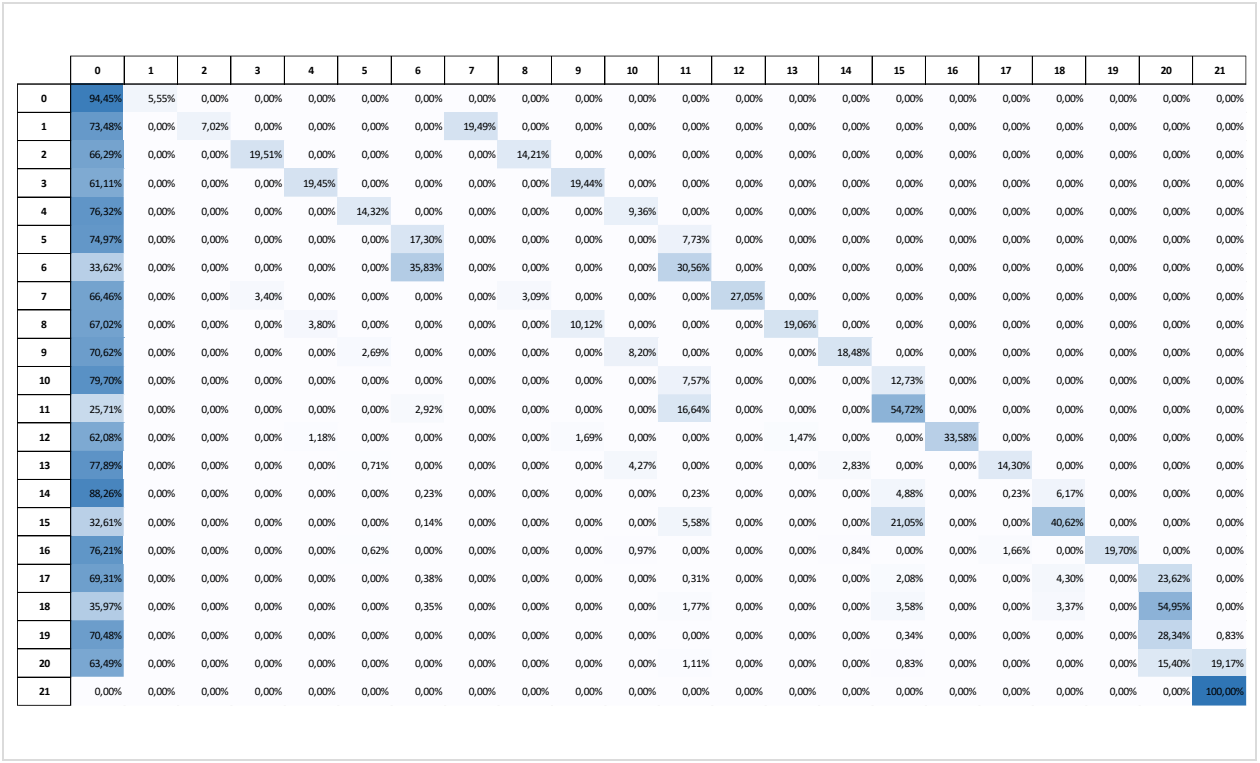
Nota. Fuente: Elaboración propia.

ANEXO N Matrices de transición para modelos de dos variables (MK4)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	95,71%	4,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	67,93%	0,00%	14,03%	0,00%	0,00%	0,00%	0,00%	18,04%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	45,82%	0,00%	0,00%	34,82%	0,00%	0,00%	0,00%	0,00%	19,36%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	36,46%	0,00%	0,00%	0,00%	40,66%	0,00%	0,00%	0,00%	0,00%	22,88%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
4	37,18%	0,00%	0,00%	0,00%	0,00%	42,19%	0,00%	0,00%	0,00%	0,00%	20,63%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	32,04%	0,00%	0,00%	0,00%	0,00%	0,00%	46,94%	0,00%	0,00%	0,00%	0,00%	21,02%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	27,47%	0,00%	0,00%	0,00%	0,00%	0,00%	56,95%	0,00%	0,00%	0,00%	0,00%	15,59%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	33,47%	0,00%	0,00%	6,08%	0,00%	0,00%	0,00%	0,00%	9,35%	0,00%	0,00%	0,00%	51,10%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
8	26,75%	0,00%	0,00%	0,00%	10,16%	0,00%	0,00%	0,00%	0,00%	21,38%	0,00%	0,00%	0,00%	41,71%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9	24,13%	0,00%	0,00%	0,00%	0,00%	12,60%	0,00%	0,00%	0,00%	0,00%	25,87%	0,00%	0,00%	0,00%	37,40%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10	20,86%	0,00%	0,00%	0,00%	0,00%	0,00%	13,15%	0,00%	0,00%	0,00%	0,00%	30,61%	0,00%	0,00%	0,00%	35,37%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
11	21,52%	0,00%	0,00%	0,00%	0,00%	0,00%	12,34%	0,00%	0,00%	0,00%	0,00%	32,02%	0,00%	0,00%	0,00%	34,12%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
12	18,50%	0,00%	0,00%	0,00%	1,70%	0,00%	0,00%	0,00%	0,00%	1,78%	0,00%	0,00%	0,00%	6,32%	0,00%	0,00%	71,70%	0,00%	0,00%	0,00%	0,00%	0,00%
13	18,26%	0,00%	0,00%	0,00%	0,00%	1,80%	0,00%	0,00%	0,00%	0,00%	5,39%	0,00%	0,00%	0,00%	14,82%	0,00%	0,00%	59,73%	0,00%	0,00%	0,00%	0,00%
14	19,30%	0,00%	0,00%	0,00%	0,00%	0,00%	3,76%	0,00%	0,00%	0,00%	0,00%	7,27%	0,00%	0,00%	0,00%	19,55%	0,00%	0,00%	50,13%	0,00%	0,00%	0,00%
15	23,02%	0,00%	0,00%	0,00%	0,00%	0,00%	4,04%	0,00%	0,00%	0,00%	0,00%	6,53%	0,00%	0,00%	0,00%	20,53%	0,00%	0,00%	45,88%	0,00%	0,00%	0,00%
16	9,66%	0,00%	0,00%	0,00%	0,00%	0,92%	0,00%	0,00%	0,00%	0,00%	0,80%	0,00%	0,00%	0,00%	1,26%	0,00%	0,00%	3,97%	0,00%	83,39%	0,00%	0,00%
17	13,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,74%	0,00%	0,00%	0,00%	0,00%	0,74%	0,00%	0,00%	0,00%	1,48%	0,00%	0,00%	11,60%	0,00%	72,10%	0,00%
18	18,48%	0,00%	0,00%	0,00%	0,00%	0,00%	1,30%	0,00%	0,00%	0,00%	0,00%	1,46%	0,00%	0,00%	0,00%	2,76%	0,00%	0,00%	12,97%	0,00%	63,05%	0,00%
19	5,83%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,82%	0,00%	0,00%	0,00%	0,21%	0,00%	0,00%	0,82%	0,00%	81,00%	11,32%
20	7,51%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,12%	0,00%	0,00%	0,00%	0,35%	0,00%	0,00%	0,90%	0,00%	26,61%	64,50%
21	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 107. Matriz representante cartera completa, modelo MK4

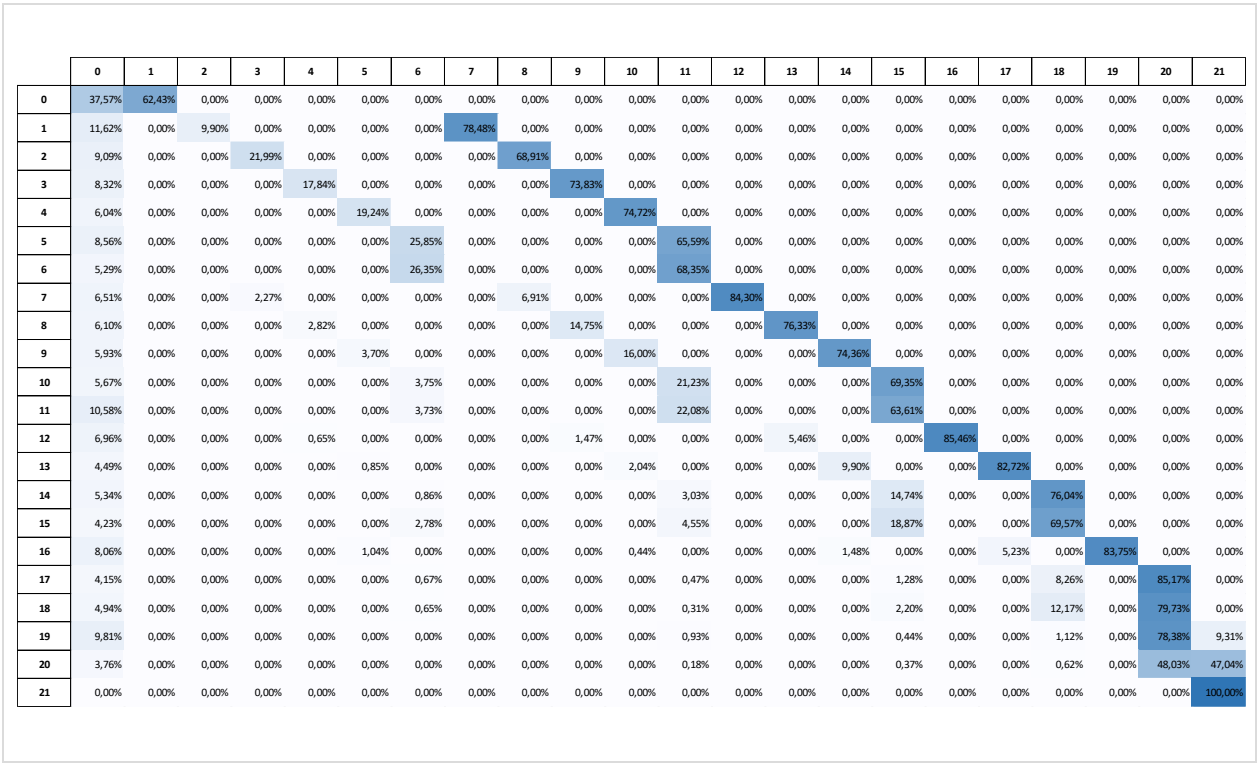
Nota. Fuente: Elaboración propia.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	94,45%	5,55%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	73,48%	0,00%	7,02%	0,00%	0,00%	0,00%	0,00%	19,49%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	66,29%	0,00%	0,00%	19,51%	0,00%	0,00%	0,00%	0,00%	14,21%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	61,11%	0,00%	0,00%	0,00%	19,45%	0,00%	0,00%	0,00%	0,00%	19,44%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
4	76,32%	0,00%	0,00%	0,00%	0,00%	14,32%	0,00%	0,00%	0,00%	0,00%	9,36%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	74,97%	0,00%	0,00%	0,00%	0,00%	0,00%	17,30%	0,00%	0,00%	0,00%	0,00%	7,73%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	33,62%	0,00%	0,00%	0,00%	0,00%	0,00%	35,83%	0,00%	0,00%	0,00%	0,00%	30,56%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	66,46%	0,00%	0,00%	3,40%	0,00%	0,00%	0,00%	0,00%	3,09%	0,00%	0,00%	0,00%	27,05%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
8	67,02%	0,00%	0,00%	0,00%	3,80%	0,00%	0,00%	0,00%	0,00%	10,12%	0,00%	0,00%	0,00%	19,06%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9	70,62%	0,00%	0,00%	0,00%	0,00%	2,69%	0,00%	0,00%	0,00%	0,00%	8,20%	0,00%	0,00%	0,00%	18,48%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10	79,70%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	7,57%	0,00%	0,00%	0,00%	12,73%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
11	25,71%	0,00%	0,00%	0,00%	0,00%	0,00%	2,92%	0,00%	0,00%	0,00%	0,00%	16,64%	0,00%	0,00%	0,00%	54,72%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
12	62,08%	0,00%	0,00%	0,00%	1,18%	0,00%	0,00%	0,00%	0,00%	1,69%	0,00%	0,00%	0,00%	1,47%	0,00%	0,00%	33,58%	0,00%	0,00%	0,00%	0,00%	0,00%
13	77,89%	0,00%	0,00%	0,00%	0,00%	0,71%	0,00%	0,00%	0,00%	0,00%	4,27%	0,00%	0,00%	0,00%	2,83%	0,00%	0,00%	14,30%	0,00%	0,00%	0,00%	0,00%
14	88,26%	0,00%	0,00%	0,00%	0,00%	0,00%	0,23%	0,00%	0,00%	0,00%	0,00%	0,23%	0,00%	0,00%	0,00%	4,88%	0,00%	0,23%	6,17%	0,00%	0,00%	0,00%
15	32,61%	0,00%	0,00%	0,00%	0,00%	0,00%	0,14%	0,00%	0,00%	0,00%	0,00%	5,58%	0,00%	0,00%	0,00%	21,05%	0,00%	0,00%	40,62%	0,00%	0,00%	0,00%
16	76,21%	0,00%	0,00%	0,00%	0,00%	0,62%	0,00%	0,00%	0,00%	0,00%	0,97%	0,00%	0,00%	0,00%	0,84%	0,00%	0,00%	1,66%	0,00%	19,70%	0,00%	0,00%
17	69,31%	0,00%	0,00%	0,00%	0,00%	0,00%	0,38%	0,00%	0,00%	0,00%	0,00%	0,31%	0,00%	0,00%	0,00%	2,08%	0,00%	0,00%	4,30%	0,00%	23,62%	0,00%
18	35,97%	0,00%	0,00%	0,00%	0,00%	0,00%	0,35%	0,00%	0,00%	0,00%	0,00%	1,77%	0,00%	0,00%	0,00%	3,58%	0,00%	0,00%	3,37%	0,00%	54,95%	0,00%
19	70,48%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,34%	0,00%	0,00%	0,00%	0,00%	28,34%	0,83%
20	63,49%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,11%	0,00%	0,00%	0,00%	0,83%	0,00%	0,00%	0,00%	0,00%	15,40%	19,17%
21	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 108. Matriz representante clúster 1, modelo MK4

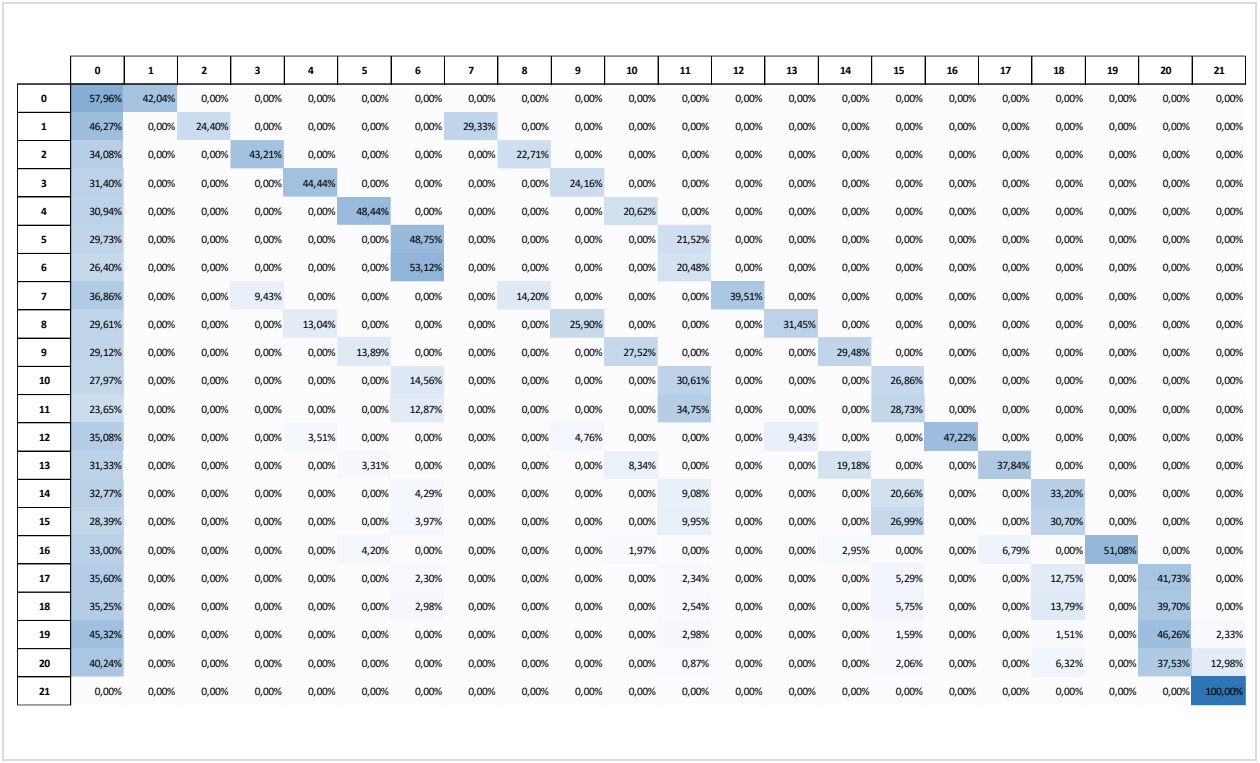
Nota. Fuente: Elaboración propia.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	37,57%	62,43%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	11,62%	0,00%	9,90%	0,00%	0,00%	0,00%	0,00%	78,48%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	9,09%	0,00%	0,00%	21,99%	0,00%	0,00%	0,00%	0,00%	68,91%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	8,32%	0,00%	0,00%	0,00%	17,84%	0,00%	0,00%	0,00%	0,00%	73,83%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
4	6,04%	0,00%	0,00%	0,00%	0,00%	19,24%	0,00%	0,00%	0,00%	0,00%	74,72%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	8,56%	0,00%	0,00%	0,00%	0,00%	0,00%	25,85%	0,00%	0,00%	0,00%	0,00%	65,59%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	5,29%	0,00%	0,00%	0,00%	0,00%	0,00%	26,35%	0,00%	0,00%	0,00%	0,00%	68,35%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	6,51%	0,00%	0,00%	2,27%	0,00%	0,00%	0,00%	0,00%	6,91%	0,00%	0,00%	0,00%	84,30%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
8	6,10%	0,00%	0,00%	0,00%	2,82%	0,00%	0,00%	0,00%	0,00%	14,75%	0,00%	0,00%	0,00%	76,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9	5,93%	0,00%	0,00%	0,00%	0,00%	3,70%	0,00%	0,00%	0,00%	0,00%	16,00%	0,00%	0,00%	0,00%	74,36%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10	5,67%	0,00%	0,00%	0,00%	0,00%	0,00%	3,75%	0,00%	0,00%	0,00%	0,00%	21,23%	0,00%	0,00%	0,00%	69,35%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
11	10,58%	0,00%	0,00%	0,00%	0,00%	0,00%	3,73%	0,00%	0,00%	0,00%	0,00%	22,08%	0,00%	0,00%	0,00%	63,61%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
12	6,96%	0,00%	0,00%	0,00%	0,65%	0,00%	0,00%	0,00%	0,00%	1,47%	0,00%	0,00%	0,00%	5,46%	0,00%	0,00%	85,46%	0,00%	0,00%	0,00%	0,00%	0,00%
13	4,49%	0,00%	0,00%	0,00%	0,00%	0,85%	0,00%	0,00%	0,00%	0,00%	2,04%	0,00%	0,00%	0,00%	9,90%	0,00%	0,00%	82,72%	0,00%	0,00%	0,00%	0,00%
14	5,34%	0,00%	0,00%	0,00%	0,00%	0,00%	0,86%	0,00%	0,00%	0,00%	0,00%	3,03%	0,00%	0,00%	0,00%	14,74%	0,00%	0,00%	76,04%	0,00%	0,00%	0,00%
15	4,23%	0,00%	0,00%	0,00%	0,00%	0,00%	2,78%	0,00%	0,00%	0,00%	0,00%	4,55%	0,00%	0,00%	0,00%	18,87%	0,00%	0,00%	69,57%	0,00%	0,00%	0,00%
16	8,06%	0,00%	0,00%	0,00%	0,00%	1,04%	0,00%	0,00%	0,00%	0,00%	0,44%	0,00%	0,00%	0,00%	1,48%	0,00%	0,00%	5,23%	0,00%	83,75%	0,00%	0,00%
17	4,15%	0,00%	0,00%	0,00%	0,00%	0,00%	0,67%	0,00%	0,00%	0,00%	0,00%	0,47%	0,00%	0,00%	0,00%	1,28%	0,00%	0,00%	8,26%	0,00%	85,17%	0,00%
18	4,94%	0,00%	0,00%	0,00%	0,00%	0,00%	0,65%	0,00%	0,00%	0,00%	0,00%	0,31%	0,00%	0,00%	0,00%	2,20%	0,00%	0,00%	12,17%	0,00%	79,73%	0,00%
19	9,81%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,93%	0,00%	0,00%	0,00%	0,44%	0,00%	0,00%	1,12%	0,00%	78,38%	9,31%
20	3,76%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,18%	0,00%	0,00%	0,00%	0,37%	0,00%	0,00%	0,62%	0,00%	48,03%	47,04%
21	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 109. Matriz representante clúster 2, modelo MK4

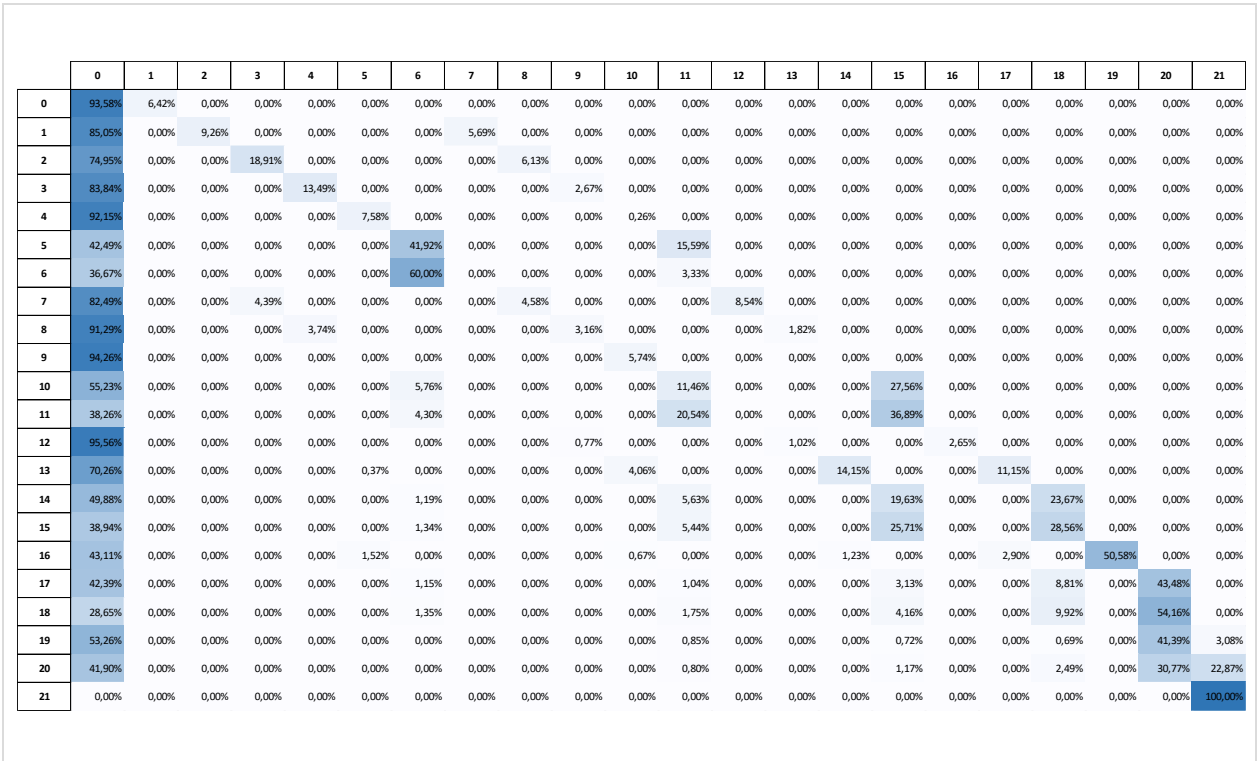
Nota. Fuente: Elaboración propia.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	57,96%	42,04%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	46,27%	0,00%	24,40%	0,00%	0,00%	0,00%	0,00%	29,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	34,08%	0,00%	0,00%	43,21%	0,00%	0,00%	0,00%	0,00%	22,71%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	31,40%	0,00%	0,00%	0,00%	44,44%	0,00%	0,00%	0,00%	0,00%	24,16%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
4	30,94%	0,00%	0,00%	0,00%	0,00%	48,44%	0,00%	0,00%	0,00%	0,00%	20,62%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	29,73%	0,00%	0,00%	0,00%	0,00%	0,00%	48,75%	0,00%	0,00%	0,00%	0,00%	21,52%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	26,40%	0,00%	0,00%	0,00%	0,00%	0,00%	53,12%	0,00%	0,00%	0,00%	0,00%	20,48%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	36,86%	0,00%	0,00%	9,43%	0,00%	0,00%	0,00%	0,00%	14,20%	0,00%	0,00%	0,00%	39,51%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
8	29,61%	0,00%	0,00%	0,00%	13,04%	0,00%	0,00%	0,00%	0,00%	25,90%	0,00%	0,00%	0,00%	31,45%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9	29,12%	0,00%	0,00%	0,00%	0,00%	13,89%	0,00%	0,00%	0,00%	0,00%	27,52%	0,00%	0,00%	0,00%	29,48%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10	27,97%	0,00%	0,00%	0,00%	0,00%	0,00%	14,56%	0,00%	0,00%	0,00%	0,00%	30,61%	0,00%	0,00%	0,00%	26,86%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
11	23,65%	0,00%	0,00%	0,00%	0,00%	0,00%	12,87%	0,00%	0,00%	0,00%	0,00%	34,75%	0,00%	0,00%	0,00%	28,73%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
12	35,08%	0,00%	0,00%	0,00%	3,51%	0,00%	0,00%	0,00%	0,00%	4,76%	0,00%	0,00%	0,00%	9,43%	0,00%	0,00%	47,22%	0,00%	0,00%	0,00%	0,00%	0,00%
13	31,33%	0,00%	0,00%	0,00%	0,00%	3,31%	0,00%	0,00%	0,00%	0,00%	8,34%	0,00%	0,00%	0,00%	19,18%	0,00%	0,00%	37,84%	0,00%	0,00%	0,00%	0,00%
14	32,77%	0,00%	0,00%	0,00%	0,00%	0,00%	4,29%	0,00%	0,00%	0,00%	0,00%	9,08%	0,00%	0,00%	0,00%	20,66%	0,00%	0,00%	33,20%	0,00%	0,00%	0,00%
15	28,39%	0,00%	0,00%	0,00%	0,00%	0,00%	3,97%	0,00%	0,00%	0,00%	0,00%	9,95%	0,00%	0,00%	0,00%	26,99%	0,00%	0,00%	30,70%	0,00%	0,00%	0,00%
16	33,00%	0,00%	0,00%	0,00%	0,00%	4,20%	0,00%	0,00%	0,00%	0,00%	1,97%	0,00%	0,00%	0,00%	2,95%	0,00%	0,00%	6,79%	0,00%	51,08%	0,00%	0,00%
17	35,60%	0,00%	0,00%	0,00%	0,00%	0,00%	2,30%	0,00%	0,00%	0,00%	0,00%	2,34%	0,00%	0,00%	0,00%	5,29%	0,00%	0,00%	12,75%	0,00%	41,73%	0,00%
18	35,25%	0,00%	0,00%	0,00%	0,00%	0,00%	2,98%	0,00%	0,00%	0,00%	0,00%	2,54%	0,00%	0,00%	0,00%	5,75%	0,00%	0,00%	13,79%	0,00%	39,70%	0,00%
19	45,32%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,98%	0,00%	0,00%	0,00%	1,59%	0,00%	0,00%	1,51%	0,00%	46,26%	2,33%
20	40,24%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,87%	0,00%	0,00%	0,00%	2,06%	0,00%	0,00%	6,32%	0,00%	37,53%	12,98%
21	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 110. Matriz representante clúster 3, modelo MK4

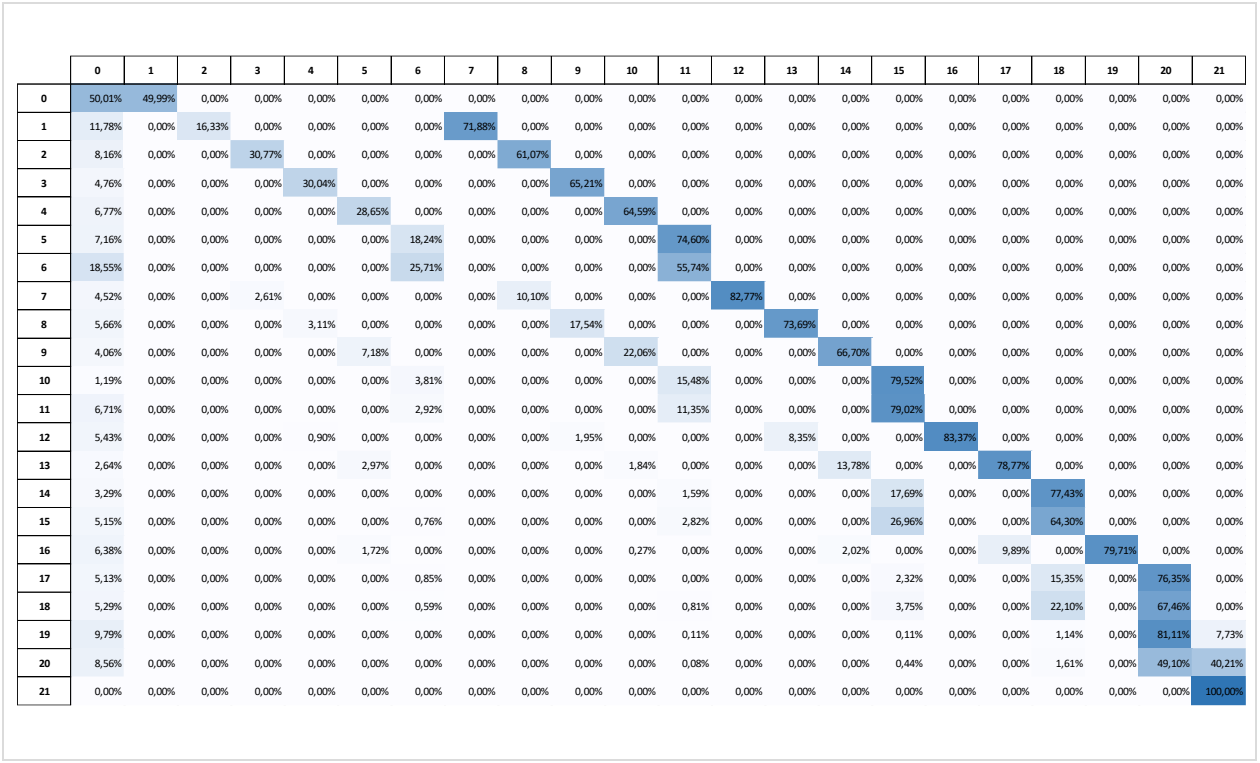
Nota. Fuente: Elaboración propia.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	93,58%	6,42%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	85,05%	0,00%	9,26%	0,00%	0,00%	0,00%	0,00%	5,69%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	74,95%	0,00%	0,00%	18,91%	0,00%	0,00%	0,00%	0,00%	6,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	83,84%	0,00%	0,00%	0,00%	13,49%	0,00%	0,00%	0,00%	0,00%	2,67%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
4	92,15%	0,00%	0,00%	0,00%	0,00%	7,58%	0,00%	0,00%	0,00%	0,00%	0,26%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	42,49%	0,00%	0,00%	0,00%	0,00%	0,00%	41,92%	0,00%	0,00%	0,00%	0,00%	15,59%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	36,67%	0,00%	0,00%	0,00%	0,00%	0,00%	60,00%	0,00%	0,00%	0,00%	0,00%	3,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	82,49%	0,00%	0,00%	4,39%	0,00%	0,00%	0,00%	0,00%	4,58%	0,00%	0,00%	0,00%	8,54%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
8	91,29%	0,00%	0,00%	0,00%	3,74%	0,00%	0,00%	0,00%	0,00%	3,16%	0,00%	0,00%	0,00%	1,82%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9	94,26%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	5,74%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10	55,23%	0,00%	0,00%	0,00%	0,00%	0,00%	5,76%	0,00%	0,00%	0,00%	0,00%	11,46%	0,00%	0,00%	0,00%	27,56%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
11	38,26%	0,00%	0,00%	0,00%	0,00%	0,00%	4,30%	0,00%	0,00%	0,00%	0,00%	20,54%	0,00%	0,00%	0,00%	36,89%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
12	95,56%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,77%	0,00%	0,00%	0,00%	1,02%	0,00%	0,00%	2,65%	0,00%	0,00%	0,00%	0,00%	0,00%
13	70,26%	0,00%	0,00%	0,00%	0,00%	0,37%	0,00%	0,00%	0,00%	0,00%	4,06%	0,00%	0,00%	0,00%	14,15%	0,00%	0,00%	11,15%	0,00%	0,00%	0,00%	0,00%
14	49,88%	0,00%	0,00%	0,00%	0,00%	0,00%	1,19%	0,00%	0,00%	0,00%	0,00%	5,63%	0,00%	0,00%	0,00%	19,63%	0,00%	0,00%	23,67%	0,00%	0,00%	0,00%
15	38,94%	0,00%	0,00%	0,00%	0,00%	0,00%	1,34%	0,00%	0,00%	0,00%	0,00%	5,44%	0,00%	0,00%	0,00%	25,71%	0,00%	0,00%	28,56%	0,00%	0,00%	0,00%
16	43,11%	0,00%	0,00%	0,00%	0,00%	1,52%	0,00%	0,00%	0,00%	0,00%	0,67%	0,00%	0,00%	0,00%	1,23%	0,00%	0,00%	2,90%	0,00%	50,58%	0,00%	0,00%
17	42,39%	0,00%	0,00%	0,00%	0,00%	0,00%	1,15%	0,00%	0,00%	0,00%	0,00%	1,04%	0,00%	0,00%	0,00%	3,13%	0,00%	0,00%	8,81%	0,00%	43,48%	0,00%
18	28,65%	0,00%	0,00%	0,00%	0,00%	0,00%	1,35%	0,00%	0,00%	0,00%	0,00%	1,75%	0,00%	0,00%	0,00%	4,16%	0,00%	0,00%	9,92%	0,00%	54,16%	0,00%
19	53,26%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,85%	0,00%	0,00%	0,00%	0,72%	0,00%	0,00%	0,69%	0,00%	41,39%	3,08%
20	41,90%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,80%	0,00%	0,00%	0,00%	1,17%	0,00%	0,00%	2,49%	0,00%	30,77%	22,87%
21	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 111. Matriz representante clúster 4, modelo MK4

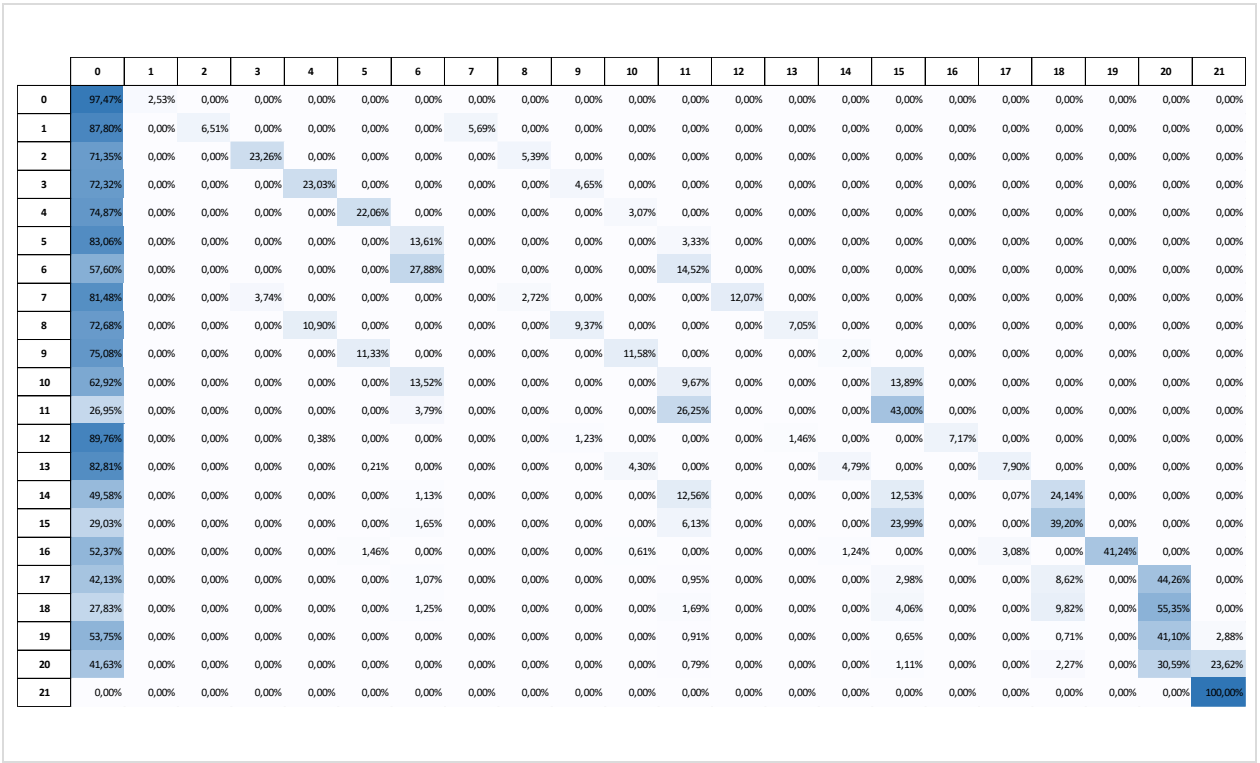
Nota. Fuente: Elaboración propia.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	50,01%	49,99%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	11,78%	0,00%	16,33%	0,00%	0,00%	0,00%	0,00%	71,88%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	8,16%	0,00%	0,00%	30,77%	0,00%	0,00%	0,00%	0,00%	61,07%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	4,76%	0,00%	0,00%	0,00%	30,04%	0,00%	0,00%	0,00%	0,00%	65,21%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
4	6,77%	0,00%	0,00%	0,00%	0,00%	28,65%	0,00%	0,00%	0,00%	0,00%	64,59%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	7,16%	0,00%	0,00%	0,00%	0,00%	0,00%	18,24%	0,00%	0,00%	0,00%	0,00%	74,60%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	18,55%	0,00%	0,00%	0,00%	0,00%	0,00%	25,71%	0,00%	0,00%	0,00%	0,00%	55,74%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	4,52%	0,00%	0,00%	2,61%	0,00%	0,00%	0,00%	0,00%	10,10%	0,00%	0,00%	0,00%	82,77%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
8	5,66%	0,00%	0,00%	0,00%	3,11%	0,00%	0,00%	0,00%	0,00%	17,54%	0,00%	0,00%	0,00%	73,69%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9	4,06%	0,00%	0,00%	0,00%	0,00%	7,18%	0,00%	0,00%	0,00%	0,00%	22,06%	0,00%	0,00%	0,00%	66,70%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10	1,19%	0,00%	0,00%	0,00%	0,00%	0,00%	3,81%	0,00%	0,00%	0,00%	0,00%	15,48%	0,00%	0,00%	0,00%	79,52%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
11	6,71%	0,00%	0,00%	0,00%	0,00%	0,00%	2,92%	0,00%	0,00%	0,00%	0,00%	11,35%	0,00%	0,00%	0,00%	79,02%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
12	5,43%	0,00%	0,00%	0,00%	0,90%	0,00%	0,00%	0,00%	0,00%	1,95%	0,00%	0,00%	0,00%	8,35%	0,00%	0,00%	83,37%	0,00%	0,00%	0,00%	0,00%	0,00%
13	2,64%	0,00%	0,00%	0,00%	0,00%	2,97%	0,00%	0,00%	0,00%	0,00%	1,84%	0,00%	0,00%	0,00%	13,78%	0,00%	0,00%	78,77%	0,00%	0,00%	0,00%	0,00%
14	3,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,59%	0,00%	0,00%	0,00%	17,69%	0,00%	0,00%	77,43%	0,00%	0,00%	0,00%
15	5,15%	0,00%	0,00%	0,00%	0,00%	0,00%	0,76%	0,00%	0,00%	0,00%	0,00%	2,82%	0,00%	0,00%	0,00%	26,96%	0,00%	0,00%	64,30%	0,00%	0,00%	0,00%
16	6,38%	0,00%	0,00%	0,00%	0,00%	1,72%	0,00%	0,00%	0,00%	0,00%	0,27%	0,00%	0,00%	0,00%	2,02%	0,00%	0,00%	9,89%	0,00%	79,71%	0,00%	0,00%
17	5,13%	0,00%	0,00%	0,00%	0,00%	0,00%	0,85%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,32%	0,00%	0,00%	15,35%	0,00%	76,35%	0,00%
18	5,29%	0,00%	0,00%	0,00%	0,00%	0,00%	0,59%	0,00%	0,00%	0,00%	0,00%	0,81%	0,00%	0,00%	0,00%	3,75%	0,00%	0,00%	22,10%	0,00%	67,46%	0,00%
19	9,79%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,11%	0,00%	0,00%	0,00%	0,11%	0,00%	0,00%	1,14%	0,00%	81,11%	7,73%
20	8,56%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,08%	0,00%	0,00%	0,00%	0,44%	0,00%	0,00%	1,61%	0,00%	49,10%	40,21%
21	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 112. Matriz representante clúster 5, modelo MK4

Nota. Fuente: Elaboración propia.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	97,47%	2,53%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
1	87,80%	0,00%	6,51%	0,00%	0,00%	0,00%	0,00%	5,69%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
2	71,35%	0,00%	0,00%	23,26%	0,00%	0,00%	0,00%	0,00%	5,39%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	72,32%	0,00%	0,00%	0,00%	23,03%	0,00%	0,00%	0,00%	0,00%	4,65%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
4	74,87%	0,00%	0,00%	0,00%	0,00%	22,06%	0,00%	0,00%	0,00%	0,00%	3,07%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
5	83,06%	0,00%	0,00%	0,00%	0,00%	0,00%	13,61%	0,00%	0,00%	0,00%	0,00%	3,33%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
6	57,60%	0,00%	0,00%	0,00%	0,00%	0,00%	27,88%	0,00%	0,00%	0,00%	0,00%	14,52%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
7	81,48%	0,00%	0,00%	3,74%	0,00%	0,00%	0,00%	0,00%	2,72%	0,00%	0,00%	0,00%	12,07%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
8	72,68%	0,00%	0,00%	0,00%	10,90%	0,00%	0,00%	0,00%	0,00%	9,37%	0,00%	0,00%	0,00%	7,05%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
9	75,08%	0,00%	0,00%	0,00%	0,00%	11,33%	0,00%	0,00%	0,00%	0,00%	11,58%	0,00%	0,00%	0,00%	2,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
10	62,92%	0,00%	0,00%	0,00%	0,00%	0,00%	13,52%	0,00%	0,00%	0,00%	0,00%	9,67%	0,00%	0,00%	0,00%	13,89%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
11	26,95%	0,00%	0,00%	0,00%	0,00%	0,00%	3,79%	0,00%	0,00%	0,00%	0,00%	26,25%	0,00%	0,00%	0,00%	43,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
12	89,76%	0,00%	0,00%	0,00%	0,38%	0,00%	0,00%	0,00%	0,00%	1,23%	0,00%	0,00%	0,00%	1,46%	0,00%	0,00%	7,17%	0,00%	0,00%	0,00%	0,00%	0,00%
13	82,81%	0,00%	0,00%	0,00%	0,00%	0,21%	0,00%	0,00%	0,00%	0,00%	4,30%	0,00%	0,00%	0,00%	4,79%	0,00%	0,00%	7,90%	0,00%	0,00%	0,00%	0,00%
14	49,58%	0,00%	0,00%	0,00%	0,00%	0,00%	1,13%	0,00%	0,00%	0,00%	0,00%	12,56%	0,00%	0,00%	0,00%	12,53%	0,00%	0,07%	24,14%	0,00%	0,00%	0,00%
15	29,03%	0,00%	0,00%	0,00%	0,00%	0,00%	1,65%	0,00%	0,00%	0,00%	0,00%	6,13%	0,00%	0,00%	0,00%	23,99%	0,00%	0,00%	39,20%	0,00%	0,00%	0,00%
16	52,37%	0,00%	0,00%	0,00%	0,00%	1,46%	0,00%	0,00%	0,00%	0,00%	0,61%	0,00%	0,00%	0,00%	1,24%	0,00%	0,00%	3,08%	0,00%	41,24%	0,00%	0,00%
17	42,13%	0,00%	0,00%	0,00%	0,00%	0,00%	1,07%	0,00%	0,00%	0,00%	0,00%	0,95%	0,00%	0,00%	0,00%	2,98%	0,00%	0,00%	8,62%	0,00%	44,26%	0,00%
18	27,83%	0,00%	0,00%	0,00%	0,00%	0,00%	1,25%	0,00%	0,00%	0,00%	0,00%	1,69%	0,00%	0,00%	0,00%	4,06%	0,00%	0,00%	9,82%	0,00%	55,35%	0,00%
19	53,75%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,91%	0,00%	0,00%	0,00%	0,65%	0,00%	0,00%	0,71%	0,00%	41,10%	2,88%
20	41,63%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,79%	0,00%	0,00%	0,00%	1,11%	0,00%	0,00%	2,27%	0,00%	30,59%	23,62%
21	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Figura 113. Matriz representante clúster 6, modelo MK4

Nota. Fuente: Elaboración propia.

ANEXO O Detalles representación en grafo modelo MK4

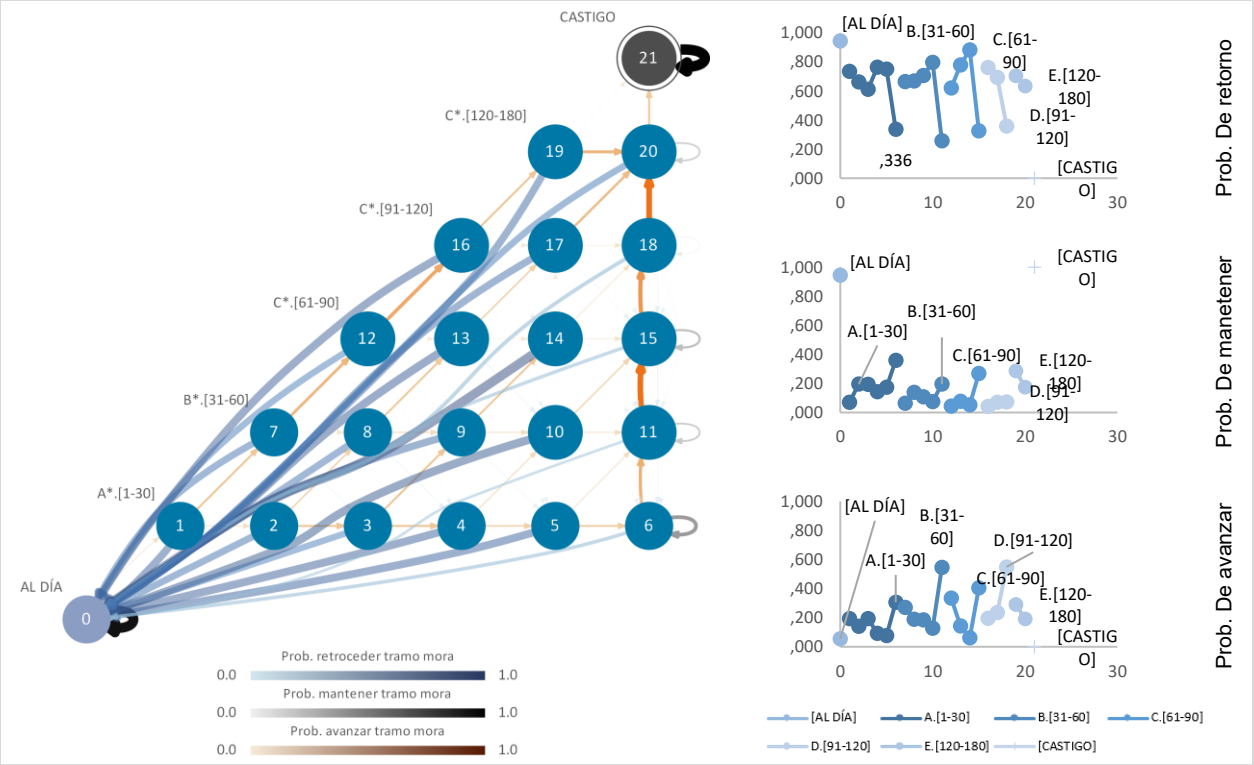


Figura 114. Detalle representación en grafo clúster 1, modelo MK4

Nota. Fuente: Elaboración propia.

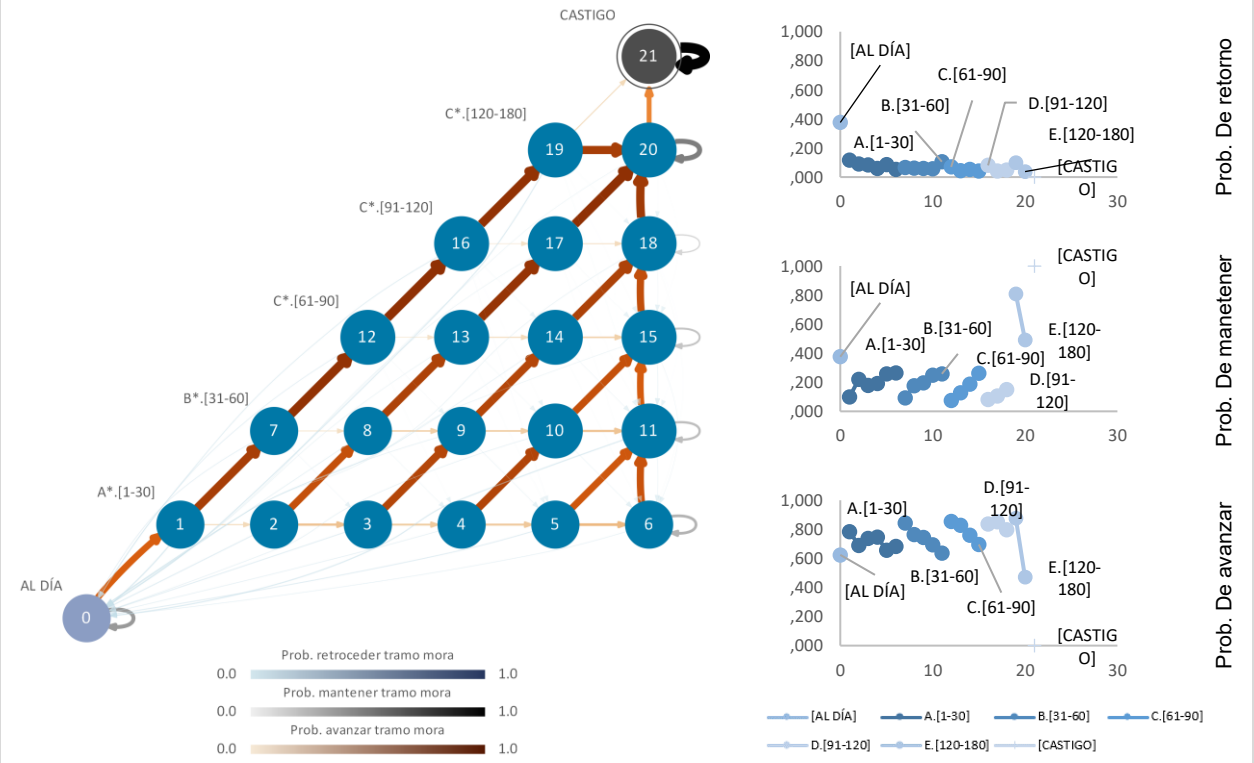


Figura 115. Detalle representación en grafo clúster 2, modelo MK4

Nota. Fuente: Elaboración propia.

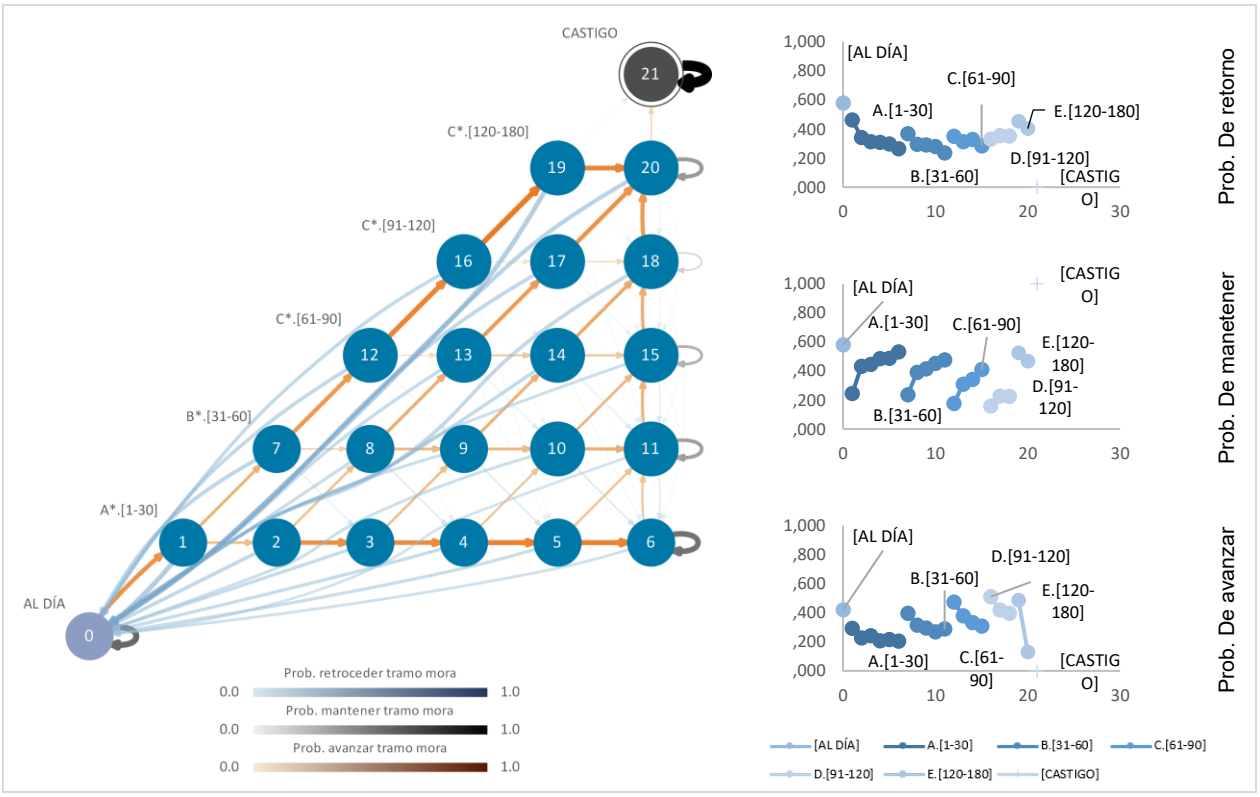


Figura 116. Detalle representación en grafo clúster 3, modelo MK4

Nota. Fuente: Elaboración propia.

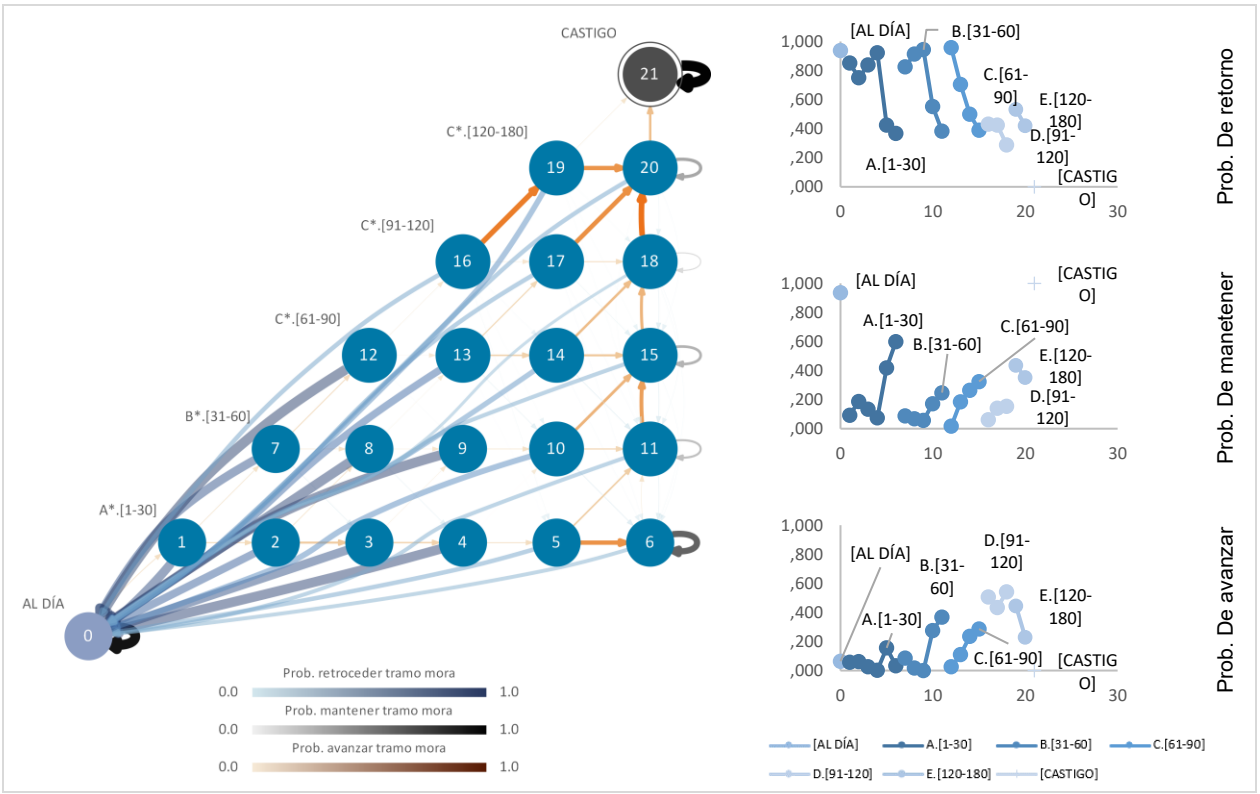


Figura 117. Detalle representación en grafo clúster 4, modelo MK4

Nota. Fuente: Elaboración propia.

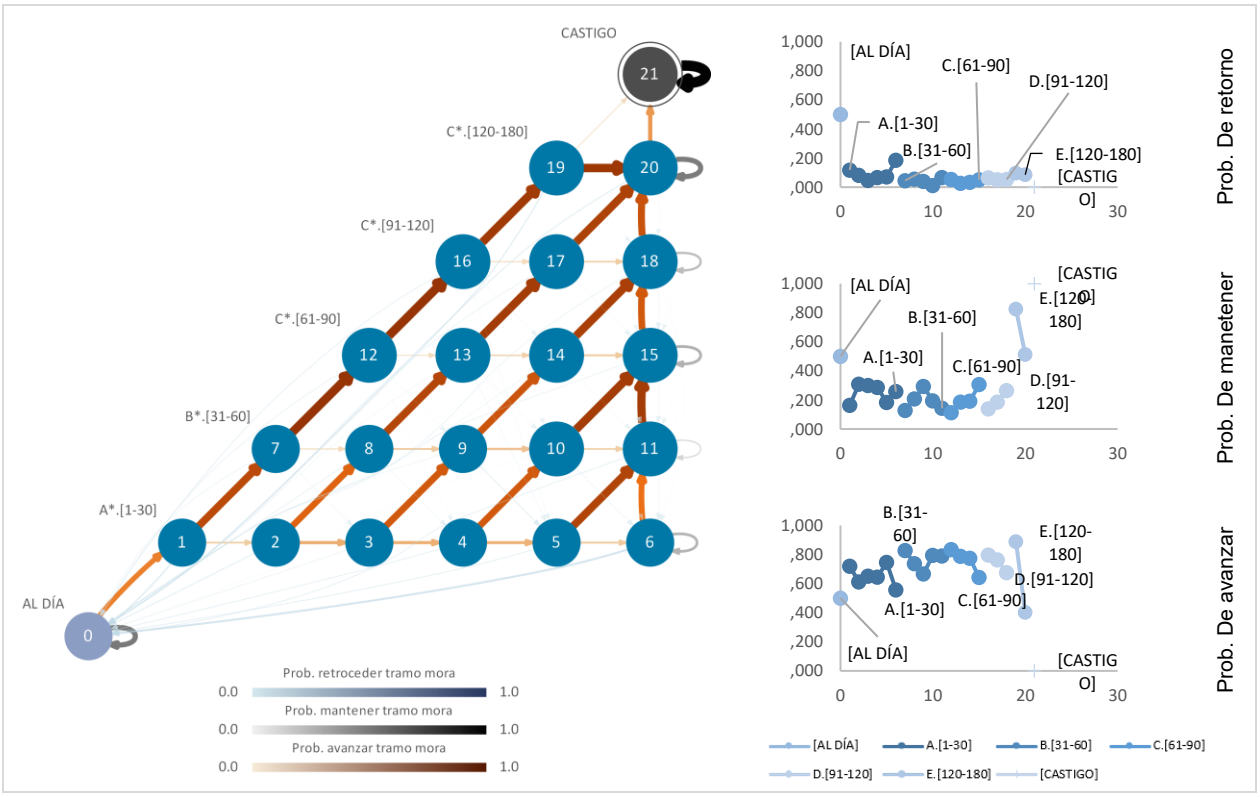


Figura 118. Detalle representación en grafo clúster 5, modelo MK4

Nota. Fuente: Elaboración propia.

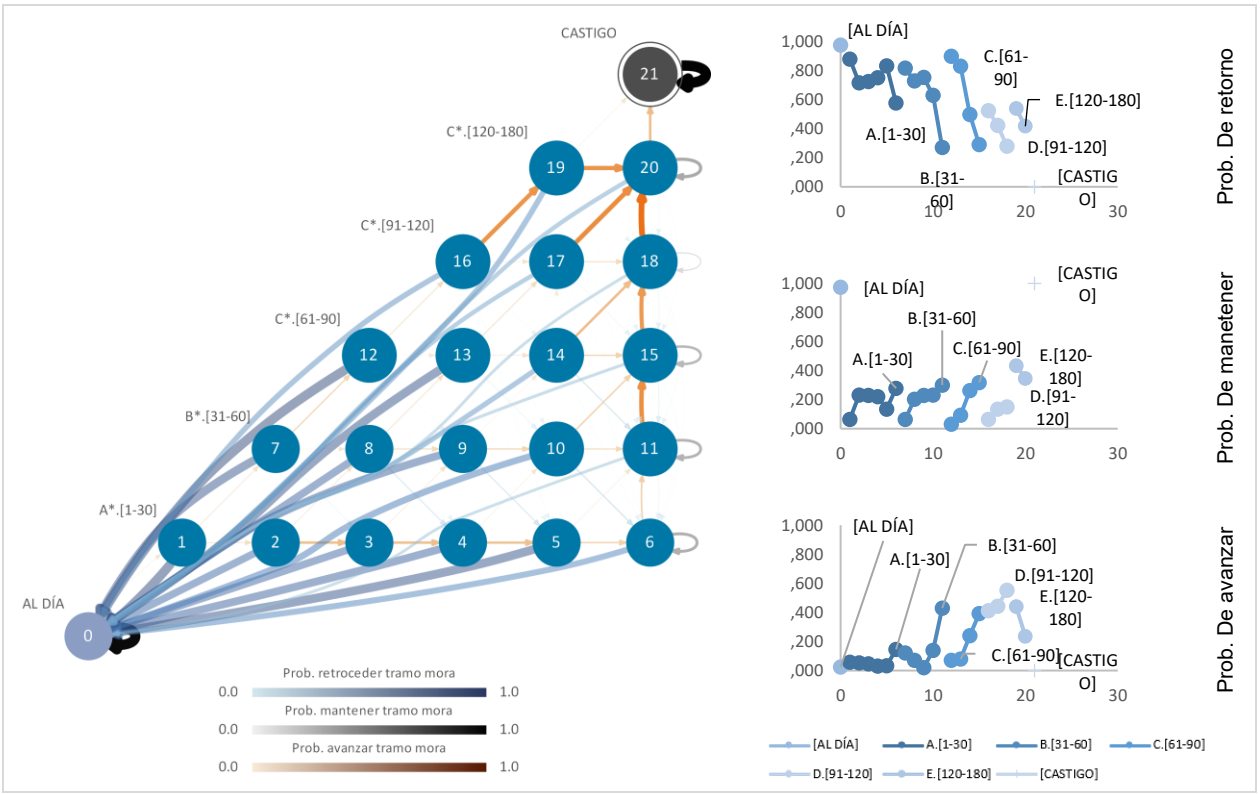


Figura 119. Detalle representación en grafo clúster 6, modelo MK4

Nota. Fuente: Elaboración propia.