"Análisis inteligente de datos en el área de Cobranzas en una empresa de distribución eléctrica"

Autor: Chehin, Tomás Augusto
Correo: tomichehin@gmail.com

Tutor: Medina Galván, Marcelo

2025

Índice

Resumen

La morosidad en los servicios públicos constituye un desafío relevante para las organizaciones encargadas de la distribución eléctrica, ya que afecta la planificación operativa, la sostenibilidad financiera y la toma de decisiones. En este contexto, la Administración Tafí Viejo (dependiente de Lucecitas S.A., empresa encargada de la distribución de energía en la zona) gestiona un volumen considerable de información por parte de los usuarios. Sin embargo, la dispersión de estos datos y la falta de integración limitan su aprovechamiento para anticipar comportamientos de pago y orientar adecuadamente las acciones de cobranza. Este trabajo se desarrolla para analizar esta problemática y aportar herramientas analíticas que permitan mejorar la gestión basada en evidencia.

La problemática identificada surge de la necesidad de anticipar qué clientes regularizarán su situación y cuáles permanecerán en mora. La ausencia de criterios sistematizados de priorización, la dependencia de planillas manuales y la fragmentación de la información generan limitaciones operativas y decisiones reactivas, lo que afecta la eficiencia del área de cobranzas. El objetivo general fue analizar el comportamiento de los servicios morosos y desarrollar herramientas estadísticas capaces de estimar la probabilidad de continuar en mora, con el fin de optimizar la planificación operativa. Para fundamentar este propósito, el marco teórico abordó conceptos de organizaciones basadas en datos, machine learning, regresión logística binaria y técnicas de análisis multivariante como el Análisis de Correspondencias Simple (AC) y Múltiple (ACM), que permiten explorar patrones en variables categóricas.

Metodológicamente, se empleó un diseño mixto DEXPLOS. En la fase cualitativa, se realizaron entrevistas al personal del área y observación directa de los procesos, lo que permitió identificar percepciones, limitaciones y criterios internos que orientaron el análisis cuantitativo. En la fase cuantitativa se trabajó con bases reales de facturación, realizando primero un Análisis Exploratorio de Datos que evidenció diferencias marcadas entre localidades, niveles de consumo y tipos de servicio. Se observó que los servicios sin consumo activo presentan mayor probabilidad de permanecer morosos, mientras que localidades urbanas muestran mejores tasas de regularización. Posteriormente, se aplicó una regresión logística binaria que permitirá estimar la probabilidad individual de morosidad, y logró destacar variables como cantidad de facturas, consumo activo, localidad y tipo de servicio. Finalmente, el AC y el ACM permitieron identificar perfiles categóricos consistentes con los resultados anteriores.

Los resultados muestran que la morosidad presenta patrones identificables y que es posible estimar el riesgo de que un cliente moroso regularice (o no) su situación, mediante técnicas estadísticas. Esto ofrece una base sólida para mejorar la asignación de recursos, priorizar gestiones y fortalecer la toma de decisiones en la Administración Tafí Viejo.

Palabras Clave: morosidad, machine learning, gestión de cobranzas

Introducción

En el marco de la transformación digital, las organizaciones han comenzado a reconocer a los datos como uno de sus activos más valiosos. Davenport y Prusak (1998) sostienen que el conocimiento surge cuando la información se estructura y se vuelve aplicable, convirtiéndose en un recurso clave para generar ventajas competitivas. En este sentido, la literatura reciente diferencia entre organizaciones data-driven, que utilizan los datos en procesos puntuales, y aquellas data-centric, en las que la información constituye el núcleo de la estrategia y de las decisiones operativas (Asistic, 2024; Cultura Organizacional Basadas en Datos, 2024). Sin embargo, muchos entornos empresariales aún enfrentan dificultades para integrar plenamente una cultura basada en datos, lo que limita su capacidad de respuesta en contextos dinámicos.

Dentro de este escenario, el sector eléctrico ocupa un lugar de especial relevancia, ya que la distribución de energía se encuentra regulada como servicio público esencial bajo la Ley N° 24.065, lo que exige garantizar continuidad, calidad y eficiencia en su prestación. La gestión de cobranzas forma parte central de este desafío: administrar eficazmente la deuda y la morosidad de los usuarios, ya que, un alto nivel de incobrabilidad afecta directamente el flujo financiero que permite cumplir con la operación del servicio (ENRE, 2023).

En este contexto se ubica Lucecitas S.A., empresa distribuidora de energía eléctrica en la provincia de Tucumán, que cuenta con 5 Administraciones y 24 sucursales. Dentro de esta estructura, la Administración Tafí Viejo funciona como centro de operaciones responsable de cuatro sucursales: Tafí Viejo, Trancas, San Pedro y Yerba Buena. En este ámbito, el área de cobranzas enfrenta un desafío persistente: comprender las causas, patrones y características de la morosidad de los clientes, y diseñar estrategias efectivas que permitan reducir los niveles de deuda y mejorar el recupero de la deuda. Esto constituye el eje sobre el cual se centrará el presente trabajo de investigación.

La operatoria diaria de cobranzas genera grandes volúmenes de datos provenientes de diferentes fuentes, cuya integración y análisis representan una oportunidad estratégica. Sin embargo, gran parte de esta información se gestiona actualmente en planillas independientes y reportes manuales, lo que dificulta identificar tendencias, medir comportamientos de pago y evaluar el impacto de las acciones implementadas. Tal como señalan Diestra Quinto et al. (2023), el empleo de técnicas de análisis avanzado y machine learning en procesos de cobranza permite identificar patrones, predecir resultados y optimizar la asignación de recursos.

En este marco, el presente trabajo se propone analizar la deuda y el comportamiento de los clientes morosos de Lucecitas S.A. mediante el uso de herramientas de análisis de datos y modelos predictivos. El objetivo es generar información estratégica que permita comprender las variables asociadas a la morosidad,

anticipar el comportamiento de pago y proponer acciones de gestión más efectivas para fortalecer la toma de decisiones en el área de cobranzas en la Administración Tafí Viejo.

Situación Problemática

La Administración Tafí Viejo de Lucecitas S.A. concentra diariamente grandes volúmenes de datos relacionados con la deuda y los clientes morosos. Esta información proviene de diversas fuentes (facturas impagas, historiales de pago, registros de cortes, etc.) y refleja la dinámica de cobranza de las cuatro sucursales bajo su jurisdicción.

Los registros se gestionan de manera fragmentada en planillas independientes y bases no vinculadas, lo que dificulta el seguimiento global de la deuda y la identificación de tendencias entre distintos segmentos de clientes o zonas geográficas.

En la práctica, las decisiones sobre estrategias de cobranza y priorización de casos se sustentan principalmente en la experiencia de los encargados o en criterios operativos básicos, como el monto adeudado o la cantidad de facturas vencidas. Esta forma de gestión, si bien funcional en el corto plazo, limita la capacidad de análisis y la toma de decisiones basadas en evidencia, dificultando la identificación de factores determinantes del incumplimiento y la previsión de comportamientos de pago. En ausencia de un enfoque analítico, la gestión se mantiene reactiva más que preventiva, destinando recursos de manera poco eficiente.

Bajo estas condiciones, la Administración enfrenta un escenario donde la información disponible no logra transformarse en una comprensión amplia del comportamiento de los clientes morosos. La coexistencia de múltiples bases y criterios de registro dificulta obtener una lectura continua y comparativa de la evolución de la deuda, lo que limita la posibilidad de identificar con claridad qué factores influyen en el incumplimiento y cómo varía la morosidad entre localidades, tipos de servicio o niveles de consumo. Esto genera decisiones apoyadas en información parcial y restringe la capacidad de la organización para interpretar adecuadamente la dinámica del problema y atenderlo de manera integral.

Preguntas de Investigación

1. ¿De qué manera se gestiona actualmente la información vinculada a la deuda y morosidad de los clientes en Lucecitas S.A.?
2. ¿Qué variables o factores inciden en la probabilidad de que un cliente moroso regularice (o no) su situación?
3. ¿Qué herramientas de análisis inteligente de datos podrían implementarse para predecir el comportamiento de pago y diseñar estrategias más efectivas de gestión de cobranzas?

Objetivo General

El objetivo general de este trabajo es generar información estratégica sobre la deuda y el comportamiento de los clientes morosos para la toma de decisiones en el área de cobranzas, mediante herramientas de análisis inteligente de datos que permitan transformar los registros disponibles en planes de acción más efectivos.

Objetivos Específicos

1. Evaluar las fortalezas y debilidades en el manejo actual de la información relacionada con la deuda y la morosidad en Lucecitas S.A.
2. Determinar las variables y factores que inciden en la probabilidad de regularización (o no) de los clientes morosos, a partir del análisis de los datos históricos disponibles.
3. Proponer herramientas de análisis inteligente de datos que permitan estimar comportamientos de pago y optimizar las estrategias de gestión de cobranzas.

Marco Metodológico

La presente investigación se enmarcará en un enfoque mixto, dado que integrará técnicas cualitativas y cuantitativas con el propósito de obtener una comprensión más completa del fenómeno de estudio.

Para el desarrollo del trabajo se utilizará un diseño mixto exploratorio secuencial (DEXPLOS), que implicará una fase inicial de recolección y análisis de datos cualitativos, orientada a explorar el funcionamiento operativo del área de cobranzas y comprender cómo se gestionan los datos vinculados a la deuda y morosidad de los clientes. Posteriormente, en la fase cuantitativa, se trabajará con bases de datos históricas de facturas impagas, con el propósito de estimar la probabilidad de regularización de los clientes morosos y detectar variables que influyan en el comportamiento de estos.

<u>Técnicas e instrumentos de recolección de datos:</u>

- Observación directa: se llevará a cabo durante la estadía del investigador en el lugar de trabajo, con el fin de registrar los procedimientos, herramientas informáticas y criterios empleados en la gestión de cobranzas.
- Entrevistas semiestructuradas: se realizarán a un Supervisor y a un Asistente del área de cobranzas, con el objetivo de conocer el manejo actual de los datos, las principales dificultades y las necesidades estratégicas de mejora.
- Bases de datos históricas: se utilizarán los registros de facturas impagas, correspondientes al período de análisis, que servirán como fuente principal para la construcción de modelos analíticos.

Herramientas de análisis de datos:

El procesamiento de la información cualitativa se llevará a cabo mediante análisis de contenido, categorizando las respuestas de la entrevista según las dimensiones principales del problema identificado. Como apoyo visual se elaborará una nube de palabras para destacar la frecuencia de conceptos clave y un diagrama de Ishikawa (o de causa-efecto, o cola de pez) que permitirá organizar y representar gráficamente los factores asociados a las dificultades actuales en el tratamiento de la deuda.

En cuanto a la fase cuantitativa, se emplearán herramientas de análisis inteligente de datos y machine learning, tal como un modelo de regresión logística binaria, orientado a estimar la probabilidad de que un cliente moroso regularice su situación en función de las variables significativas identificadas. Todo esto seguido por un Análisis de Correspondencias (AC) y un Análisis de Correspondencias Múltiple (ACM).

El procesamiento se realizará utilizando el software RStudio, con el cual se podrá trabajar de manera local y segura sin riesgo de comprometer la privacidad de los datos, complementado con planillas de cálculo en Excel, a fin de depurar los datos y presentar los resultados mediante gráficos y tablas.

Marco Teórico

Datos – La pirámide del conocimiento:

Cuando se habla de conocimientos, se debe hacer una distinción entre datos (hechos discretos, registros), información (datos organizados, con significado) y conocimiento. Davenport y Prusak definen el conocimiento como “una mezcla fluida de experiencia estructurada, valores, información contextual e internalización experta que proporciona un marco para la evaluación e incorporación de nuevas experiencias e información”. El conocimiento “se origina y es aplicado en la mente de los conocedores”, y en la organización “queda arraigado en rutinas, procesos, prácticas y normas institucionales”.

Lo importante del conocimiento, señalan estos autores, es que está cerca de la acción y que por ello permite la toma de decisiones. Se compone de experiencia, “verdad práctica”, complejidad, criterios, reglas implícitas (intuición), así como de valores y creencias. Cada uno de estos elementos tiene sus propias características:

- **Datos:** Los datos son la mínima unidad semántica, y se corresponden con elementos primarios de información que por sí solos son irrelevantes como apoyo a la toma de decisiones.
- **Información:** La información se puede definir como un conjunto de datos procesados y que tienen un significado (relevancia, propósito y contexto), y que por lo tanto son de utilidad para quién debe tomar decisiones, al disminuir su incertidumbre.

- **Conocimiento:** El conocimiento es una mezcla de experiencia, valores, información y knowhow que sirve como marco para la incorporación de nuevas experiencias e información, y es útil para la acción.
- **Sabiduría:** En este último nivel Cleveland describió la sabiduría simplemente como "conocimiento integrado: información que se vuelve súper útil". Otros autores han caracterizado la sabiduría como "saber qué hacer correctamente". La sabiduría implica utilizar el conocimiento para el bien mayor. Por eso, la sabiduría es más profunda y más exclusivamente humana. Requiere un sentido del bien y del mal, lo correcto y lo incorrecto, lo ético y lo no ético.

Imagen N°1: La pirámide del conocimiento.

Fuente: Elaboración propia.

Análisis de Datos:

El análisis de datos es el proceso de examinar, limpiar, transformar y modelar un conjunto de datos con el objetivo de descubrir información útil, extraer conocimientos y tomar decisiones informadas. Implica la aplicación de técnicas y herramientas estadísticas, matemáticas y de visualización para identificar patrones, tendencias y relaciones en conjuntos de datos. El análisis de datos permite revelar *insights*, responder preguntas y resolver problemas, ayudando a las organizaciones y personas a comprender mejor el mundo que les rodea, optimizar procesos y tomar acciones basadas en evidencia. (Innovación Digital 360, s.f.).

Por otro lado, el análisis inteligente de datos se refiere a la aplicación de técnicas avanzadas de análisis, como el machine learning, la inteligencia artificial (IA) y el big data, para identificar patrones y extraer conocimientos significativos a partir de grandes volúmenes de datos. Este enfoque va más allá de los análisis descriptivos tradicionales, permitiendo a las organizaciones predecir comportamientos futuros y tomar decisiones proactivas.

Según Provost y Fawcett (2013), el análisis inteligente de datos es esencial para la toma de decisiones basada en datos (data-driven decision making), ya que permite descubrir relaciones ocultas en los datos y optimizar las operaciones empresariales.

Además, el uso de herramientas como dashboards interactivos facilita la visualización y comprensión de los datos, lo que mejora la capacidad de la gerencia para tomar decisiones informadas y oportunas (Few, 2012).

Organizaciones basadas en datos:

En un mundo altamente interconectado, los datos se han convertido en un recurso esencial para la transformación social. La abundancia de datos generados impone desafíos que requieren un tratamiento cuidadoso. Por lo tanto, la cultura organizacional basada en datos es crucial, ya que facilita la toma de decisiones informadas, fomenta conversaciones enriquecedoras y permite cuestionar el conocimiento establecido, creando un entorno seguro y propicio para el cambio organizacional.

"Una cultura basada en datos es aquella que basa sus decisiones en hechos derivados de datos, investigación, experiencia y aportes de fuentes creíbles, en lugar de en suposiciones y "corazonadas". En una cultura basada en datos, éstos están disponibles, son accesibles, fiables y se confían a todos los miembros de la organización para que informen sus decisiones y acciones" (Educause, 2022, p.8).

María del Pilar Villamil Giraldo (2024), Profesora asociada Departamento Ingeniería de Sistemas y coordinadora de la maestría de Ingeniería de Información Universidad de Los Andes en una revista académica menciona: "En una organización basada en datos, éstos dejan de ser recursos y se transforman en activos estratégicos de la organización que representan la manera como la organización genera valor no solo para sus propios colaboradores, sino para sus clientes. Es un ejercicio recursivo, donde los aportes y creación de experiencias diferenciales internas traducidas en agilidad, eficiencia y estandarización, se traducen en nuevas oportunidades y capacidades que la organización puede generar y transferir a sus clientes. De esta forma los datos se convierten en un pilar fundamental para movilizar y asegurar la promesa de valor de empresa y su evolución de la mano de las expectativas de sus consumidores".

Lo primero es aclarar el término de organizaciones basadas en datos, que se puede llevar a organizaciones centradas en datos (data centric) o guiadas por datos (data-driven). En las primeras, la toma de decisiones se basa en los datos y toda la estructura organizacional y funcional gira en torno a ellos. En contraposición con las empresas data-driven que están centradas en aplicaciones particulares, y no se entiende la utilidad de los datos a nivel de apertura al cambio y oportunidad que se tiene de ser competitivos apalancando sus acciones y decisiones en los datos. Es por ello, que se debe empezar por resolver la pregunta ¿para qué sirven? y principalmente cómo aportan en la construcción del "negocio", tanto en la parte operativa como en la táctica y la estratégica.

Según la consultora Gartner, el concepto de organizaciones basadas en datos se refiere a aquellas que centran sus decisiones estratégicas y operativas en el análisis exhaustivo de datos. Estas organizaciones utilizan técnicas avanzadas de ciencia de

datos para transformar datos brutos en información útil, que luego es utilizada para optimizar procesos, mejorar productos y servicios, y prever tendencias futuras. (Gartner, 2023).

Toma de decisiones gerenciales:

Se puede decir que la toma de decisiones gerenciales implica identificar un problema y seguir un proceso que incluye la obtención de información, la toma de decisiones, la acción y la evaluación del desempeño de la empresa. No obstante, en la práctica, este proceso enfrenta obstáculos como la incertidumbre, que impacta en la toma de decisiones, y la falta de información sobre las alternativas y sus consecuencias, lo que dificulta la interpretación y la decisión. (Choo, 1991, como se citó en Jarrahi, 2018, citado en Diestra Quinto et al., 2023).

Machine Learning:

Machine Learning, que en español significa aprendizaje automático, es una parte de la inteligencia artificial, que busca construir programas por computadora que aprendan automáticamente en base a la experiencia adquirida. Los programas creados con Machine Learning, no necesitan ser programados con reglas explícitas que definen las tareas que debe seguir para el cumplimiento de estas, por el contrario, esta lógica mejora automáticamente en base al uso de algoritmos que analizan los datos, buscando reconocer patrones y tendencias para el entendimiento de los datos. (Bagnato, 2020, como se citó en Muñoz Cavero & Luyo Pérez, 2022, p. 28). Según explica Sandoval (2018), los algoritmos de Machine Learning se dividen en dos categorías principales: aprendizaje supervisado y no supervisado.

Aprendizaje supervisado: Es cuando entrenamos un algoritmo de Machine Learning dándole las preguntas (características) y las respuestas (etiquetas). Así en un futuro el algoritmo pueda hacer una predicción conociendo las características. Entre ellos encontramos los siguientes modelos:

- Modelos lineales: Estos tratan de encontrar una línea que se “ajuste” bien a la nube de puntos que se disponen. Aquí destacan desde modelos muy conocidos y usados como la regresión lineal (también conocida como la regresión de mínimos cuadrados), y la logística (adaptación de la lineal a problemas de clasificación cuando son variables discretas o categóricas).
- Redes neuronales: Las redes artificiales de neuronas tratan, en cierto modo, de replicar el comportamiento del cerebro, donde tenemos millones de neuronas que se interconectan en red para enviarse mensajes unas a otras. Esta réplica del funcionamiento del cerebro humano es uno de los “modelos de moda” por las habilidades cognitivas de razonamiento que adquieren.
- Modelos de árbol: Son modelos precisos, estables y más sencillos de interpretar básicamente porque construyen unas reglas de decisión que se pueden representar como un árbol. A diferencia de los modelos lineales, pueden representar relaciones no lineales para resolver problemas.
- Random Forest: Bosques Aleatorios es un tipo de clasificador que agrupa un conjunto de árboles estructurados distribuidos de manera idéntica, que arrojan

un voto unitario para la clase más popular. Su uso es tanto para tareas de clasificación como regresión, usando voto mayoritario y ponderación respectivamente. La combinación de dichos árboles claro está que bajo ciertas condiciones proporciona un mejor resultado dando como resultado un método más preciso, estable, dinámico que busca el equilibrio entre el sesgo y la varianza del bosque. (Patiño Pérez et al., 2020, p. 105).

Aprendizaje no supervisado: Aquí solo le damos las características al algoritmo, nunca las etiquetas. Se busca que agrupe los datos que le dimos según sus características. El algoritmo solo sabe que como los datos comparten ciertas características, de esa forma asume que pueda que pertenezcan al mismo grupo. Entre ellos encontramos:

- Clúster: se encarga de formar grupos diferentes dentro de los datos. Al usar algoritmos de agrupamiento se encuentra la estructura en los datos, de manera que los elementos del mismo clúster (o grupo) sean más similares entre sí que con los otros grupos generados. (Román, citado en Ramírez Mendoza, 2022).
- Análisis de Correspondencias (AC) y Análisis de Correspondencias Múltiple (ACM): son técnicas estadísticas utilizadas para estudiar las relaciones entre variables categóricas y representarlas gráficamente mediante mapas perceptuales. El AC analiza la asociación entre dos variables cualitativas, mientras que el ACM extiende esta lógica a múltiples variables categóricas. Tal como explican Aldás y Uriel (2017), ambas técnicas permiten identificar patrones, proximidades y estructuras latentes dentro de los datos, facilitando la interpretación conjunta de categorías y la comprensión de la variabilidad capturada en las dimensiones resultantes.

El desarrollo de estos modelos sigue dos fases: la fase de entrenamiento, en la que el algoritmo aprende patrones a partir de un conjunto de datos, y la fase de prueba, donde se evalúa la precisión de sus predicciones. Un alto grado de precisión indica que el algoritmo ha adquirido un buen nivel de aprendizaje.

Preparación de los datos:

Para la aplicación del modelo, es necesario seguir una serie de pasos de preparación que aseguren la calidad y relevancia de los datos utilizados. Según lo indicado en el documento Análisis Inteligente de Datos 2024 de la Universidad Austral, este proceso incluye varias fases, desde la selección y limpieza de los datos hasta su transformación, con el objetivo de obtener un conjunto de datos estructurado y listo para el modelado. A continuación, se detallan estas fases esenciales en la preparación de datos:

Fase de Selección de datos: Implica elegir las fuentes y los datos específicos que serán utilizados en el análisis. Esta fase es crucial para asegurar que se esté trabajando con la información relevante y de calidad.

- Determinar las bases de datos, archivos, APIs, u otras fuentes donde se encuentra la información.

- Recolección de datos: Extraer los datos necesarios desde las fuentes identificadas.
- Relevancia de los datos: Evaluar si los datos recolectados son pertinentes y útiles para el análisis objetivo.
- Volumen de datos: Asegurar que la cantidad de datos es suficiente para obtener resultados significativos.

Fase de Limpieza de datos: La limpieza de datos es el proceso de corregir o eliminar datos inexactos, incompletos o irrelevantes. Esta fase es esencial para asegurar la calidad y la integridad de los datos.

- Eliminación de duplicados: Identificar y remover registros duplicados que puedan sesgar el análisis.
- Manejo de valores faltantes: Decidir cómo tratar los datos ausentes, ya sea imputándolos, eliminándolos o dejándolos como están según la situación.
- Corrección de errores: Corregir datos erróneos o inexactos (errores tipográficos, valores fuera de rango, etc.).
- Normalización: Asegurar que los datos estén en un formato consistente y estandarizado (por ejemplo, fechas en el mismo formato, unidades de medida coherentes).

Fase de Exploración: Implica analizar los datos para entender sus características principales y encontrar patrones preliminares. Esta fase es clave para formular hipótesis y guiar el análisis posterior.

- Análisis descriptivo: Utilizar estadísticas descriptivas (media, mediana, moda, desviación estándar, etc.) para resumir las características principales de los datos.
- Visualización de datos: Crear gráficos y visualizaciones (histogramas, diagramas de dispersión, box plots, etc.) para identificar patrones y relaciones en los datos.
- Detección de anomalías: Identificar valores atípicos o anómalos que puedan necesitar una atención especial o una revisión adicional.
- Análisis de correlación: Evaluar las relaciones entre diferentes variables para entender cómo se influyen mutuamente.

Fase de Transformación de datos: Implica modificar los datos para que se ajusten mejor a las necesidades del análisis o del modelo que se va a utilizar. Esta fase prepara los datos para el análisis final.

- Creación de nuevas variables: Generar nuevas variables derivadas de las existentes para capturar mejor la información relevante (por ejemplo, agregar variables calculadas, como ratios o índices).
- Escalado de datos: Ajustar la escala de los datos (normalización, estandarización) para asegurar que todas las variables contribuyan de manera equitativa al análisis.
- Codificación de variables categóricas: Transformar variables categóricas en un formato adecuado para el análisis.

- Reducción de dimensionalidad: Aplicar técnicas como PCA (Análisis de Componentes Principales) para reducir el número de variables manteniendo la mayor cantidad posible de información.

Aplicación

Etapa Cualitativa:

Luego de la inmersión en la organización, de la observación directa de los procesos y de las entrevistas realizadas al Supervisor y al Asistente del área de Cobranzas, se obtuvo una comprensión profunda de la gestión actual de la deuda y de la morosidad en los clientes de Lucecitas S.A. A continuación, se desarrollan los principales hallazgos del relevamiento:

1) Fragmentación y manejo de la información:

Tanto el supervisor como el asistente coincidieron en que la empresa cuenta con un volumen importante de datos sobre facturas impagas, pagos, cortes y gestiones. Sin embargo, esa información se encuentra dispersa en distintos sistemas y archivos, lo que obliga a realizar cruces manuales en Excel para obtener una visión completa del cliente. Este proceso, aunque permite armar reportes, consume mucho tiempo y eleva la probabilidad de errores, dificultando el análisis oportuno de la deuda y la toma de decisiones.

Además, los entrevistados remarcaron que no existe una base única que consolide la información del cliente moroso, lo cual representa una de las principales debilidades del sistema actual.

2) Criterios de evaluación y priorización de clientes:

El análisis de morosidad se realiza principalmente a partir de tres variables básicas:

- Monto de deuda acumulada.
- Antigüedad de las facturas impagas.
- Zona geográfica o tipo de cliente.

Estas variables permiten una clasificación inicial, pero los entrevistados destacaron que no se aplica un modelo analítico formal que mida el riesgo o estime la probabilidad de pago. La priorización, por lo tanto, depende en gran medida del criterio y la experiencia personal del equipo de cobranzas.

Este enfoque empírico resulta funcional en la práctica, pero limita la capacidad de anticipar el comportamiento de los clientes y planificar estrategias de recupero más efectivas.

3) Comportamientos observados en los clientes morosos:

Un hallazgo relevante de las entrevistas fue la identificación de patrones de conducta entre los clientes morosos. Según los entrevistados, una parte considerable de los usuarios solo reacciona ante la amenaza o ejecución del corte, mientras que otros mantienen hábitos de pago intermitentes, alternando períodos de morosidad y regularización.

Existen también casos de reincidencia, donde el cliente vuelve a caer en mora poco tiempo después de haber saldado su deuda. Estas observaciones reflejan que la morosidad no siempre responde a la falta de recursos, sino también a factores de comportamiento, percepción del riesgo y costumbre.

Por ello, resulta necesario incorporar modelos de análisis predictivo que ayuden a anticipar estas conductas y diseñar acciones preventivas.

4) Limitaciones operativas del área de cobranzas:

Los entrevistados señalaron que la capacidad operativa del área es insuficiente frente al volumen de casos que se gestionan a diario. El equipo está compuesto por un supervisor, quien debe atender dos administraciones de manera simultánea, y un asistente, encargado del análisis de archivos, la preparación de listados y la coordinación operativa con los móviles.

Esta estructura limitada genera cuellos de botella en el análisis diario y retrasa la identificación de clientes críticos. A su vez, el trabajo manual con planillas y la falta de herramientas automatizadas incrementan la carga de trabajo y la dependencia de la experiencia individual, reduciendo la eficiencia general del proceso.

5) Uso de la información para la toma de decisiones:

En la práctica, las decisiones operativas (como a quién visitar, intimar o financiar) se toman basándose en la información disponible y en la trayectoria de cada cliente. No obstante, la dispersión de los datos y la falta de actualización en tiempo real dificultan tener una visión integral del cliente moroso. Esto provoca que algunas acciones se realicen con retraso o sin considerar toda la información relevante, afectando directamente los resultados de cobranzas y el recupero de deuda.

Ambos entrevistados coincidieron en que la creación de una herramienta integrada que unifique la información y actualice los datos automáticamente sería un avance significativo para la gestión.

6) Principales hallazgos del análisis:

De las entrevistas se desprenden los siguientes puntos críticos que caracterizan la situación actual del área:

- Dependencia de la experiencia personal: La decisión sobre las acciones a tomar se basa más en la práctica que en análisis objetivos.
- Fragmentación de datos: la información se encuentra dispersa entre varios sistemas.
- Morosidad cíclica: muchos clientes repiten ciclos de atrasos y pagos.
- Falta de automatización: el proceso manual retrasa la gestión y aumenta el margen de error.
- Recursos humanos limitados: la dotación de personal del área de Cobranzas no alcanza para el volumen de datos y gestiones.
- Ausencia de indicadores predictivos: no existen métricas que anticipen el riesgo o la probabilidad de pago.

A continuación, se presenta una nube de palabras que refleja los términos más recurrentes en las entrevistas, sintetizando las principales preocupaciones y ejes de gestión señalados por los entrevistados.

Imagen N°2: Nube de palabras

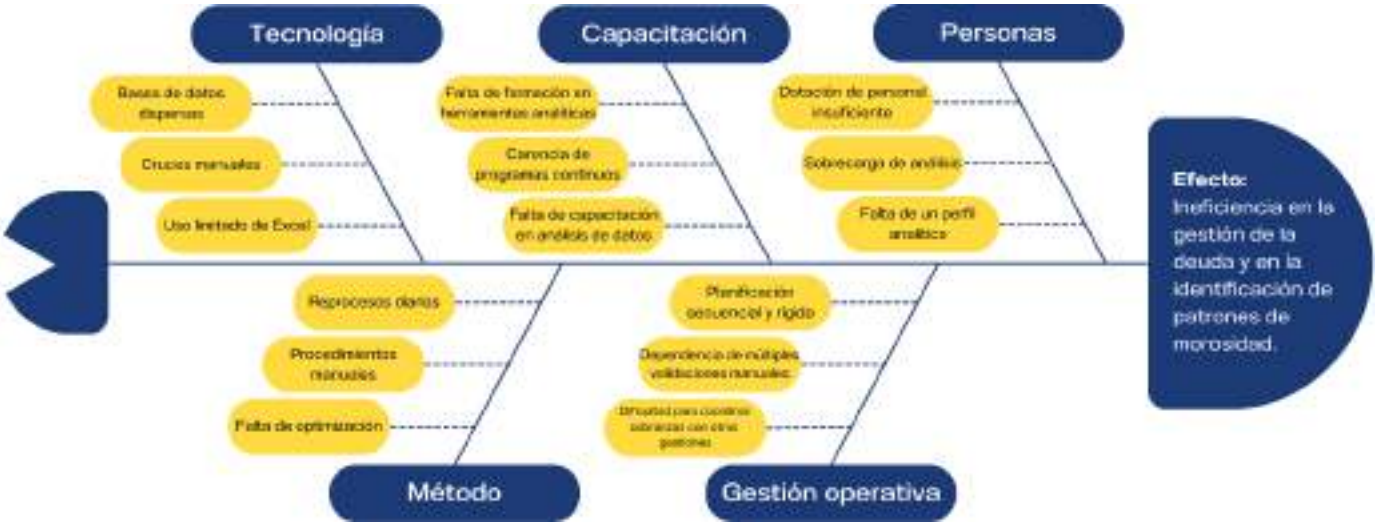
Fuente: Elaboración propia.

En síntesis, las entrevistas revelan que la empresa dispone de datos valiosos pero subutilizados, debido a la falta de integración y de herramientas de análisis avanzadas. El conocimiento sobre el comportamiento de los clientes morosos se basa principalmente en la experiencia empírica del personal, sin apoyo en indicadores ni modelos predictivos. Por lo tanto, se vuelve necesario evolucionar hacia un enfoque analítico y proactivo, capaz de unificar las bases existentes, generar indicadores de riesgo y anticipar el comportamiento de pago de los clientes.

La incorporación de estas herramientas permitiría optimizar los recursos del área, aumentar la efectividad del recupero y reducir los niveles de morosidad, alineando la gestión con los objetivos de control y eficiencia de la organización.

Con el propósito de identificar de manera sistemática las causas que explican la menor eficiencia y la poca identificación de patrones de morosidad de Lucecitas S.A., se elaboró un diagrama de Ishikawa (o cola de pez). Esta herramienta permite visualizar la relación causa–efecto y ordenar los factores que inciden en el problema central, sin pretender jerarquizarlos de antemano.

Imagen N°3: Diagrama de Ishikawa



Fuente: Elaboración propia.

Etapa Cuantitativa:

Análisis Exploratorio de Datos (AED):

En esta segunda etapa se continuó con el relevamiento y procesamiento de las bases de datos obtenidas del sistema interno de gestión. Siguiendo las fases metodológicas expuestas en el marco teórico, se desarrolló un Análisis Exploratorio de Datos (AED) orientado a describir la composición, estructura y comportamiento de las variables relacionadas con la morosidad de los clientes.

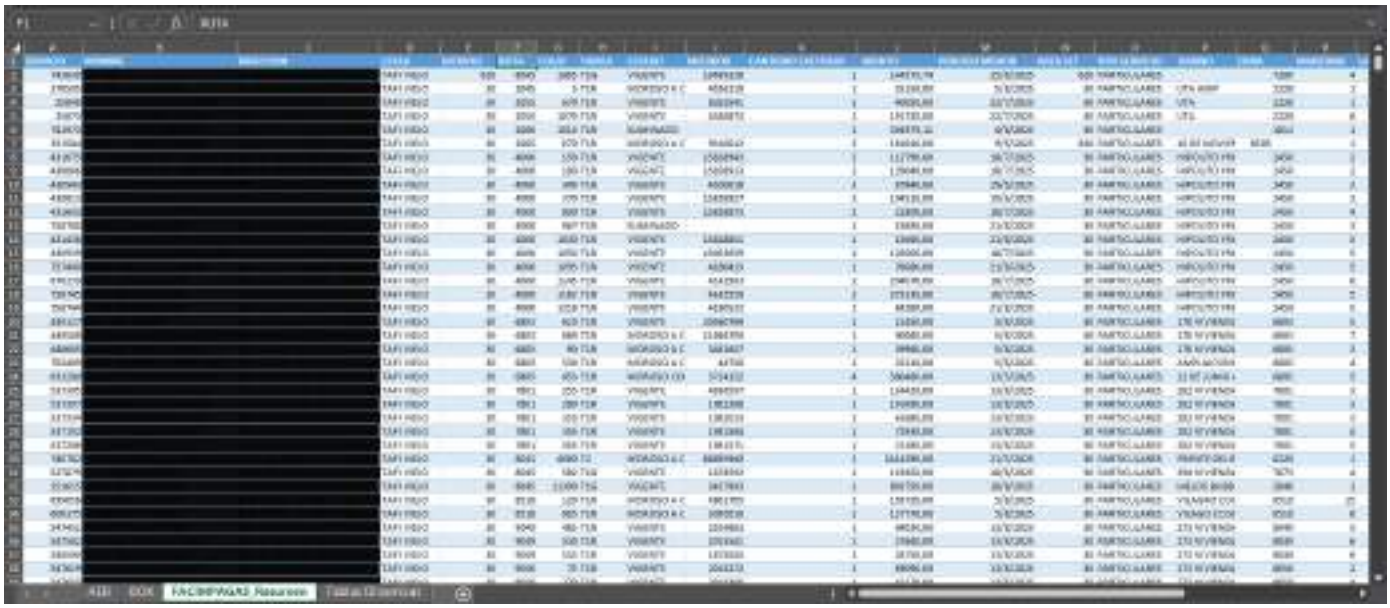
Las bases utilizadas para este análisis se denominan “FACIMPAGAS”, y registran el total de facturas impagas correspondientes a los clientes. Dichos conjuntos de datos constituyen un histórico de morosidad, reflejando la situación al 1° y al 30 de septiembre de 2025, con facturas cuyo vencimiento operativo corresponde al 31 de agosto de 2025. Este procedimiento permitió comparar ambas bases y detectar los clientes que no

regularizaron su deuda dentro del período analizado, identificando patrones de incumplimiento.

La base de datos depurada para el análisis cuenta con 27 variables, que registran información detallada sobre cada cliente, su consumo y su estado de deuda. Asimismo, incluye variables que posibilitan un análisis integral de la morosidad, como la cantidad de facturas impagas, el monto total acumulado, el estado del servicio (vigente, moroso a cortar, moroso cortado o eliminado) y el tipo de servicio.

Esta estructura de datos brinda una visión completa del comportamiento de pago de los clientes y constituye la fuente principal del análisis exploratorio y del posterior desarrollo de modelos predictivos de machine learning, orientados a optimizar la gestión de cobranzas.

Imagen N°4: Base de datos “FACIMPAGAS”



Fuente: Elaboración propia.

Clasificación de variables claves:

Tabla N°1: Definición y tipo de variables claves

Nombre de la variable	Definición	Tipo de variable
DISTRITO	Código o nombre numérico asignado a una zona geográfica específica.	Categórica nominal
BARRIO	Localización o subdivisión urbana dentro de la jurisdicción.	Categórica nominal
LOCALIDAD	Ciudad o área geográfica de referencia.	Categórica nominal

TARIFA	Categoría tarifaria según tipo y uso del servicio (ej. T1R, T1G, T2, etc.).	Categórica ordinal
TIPO SERVICIO	Clasificación técnica del tipo de conexión o servicio.	Categórica nominal
ESTADO	Situación actual del cliente: Vigente, Moroso a cortar, Moroso cortado, Eliminado.	Categórica nominal
CANTIDAD FACTURAS	Número total de facturas impagas acumuladas por el cliente.	Cuantitativa discreta
MONTO	Total de deuda acumulada por cliente (suma de todas las facturas impagas).	Cuantitativa continua
FAC CONSUMO 0	Cantidad de facturas con consumo igual a cero (posible reconexión ilegal o inactividad).	Cuantitativa discreta
¿CONSUMO >0?	Indica si el cliente tuvo consumo en el último período (“Sí” / “No”).	Categórica dicotómica
SET	Subestación transformadora asociada al servicio.	Categórica nominal
AREA / ZONA / MANZANA / LADO	Variables de ubicación física del suministro (microsegmentación geográfica).	Categóricas nominales
ADM	Administración a la que pertenece el cliente (ej. Tafí Viejo en este caso).	Categórica nominal
PERIODO MENOR	Fecha correspondiente a la factura más antigua impaga del cliente.	Temporal (fecha)
ESTADO FINAL (elaboración propia)	Identifica si el cliente regularizo su deuda (0) o permanece moroso (1) al cierre del periodo analizado	Categórica binaria

Fuente: Elaboración propia.

Resultados del Análisis Exploratorio:

El análisis exploratorio permitió obtener una primera aproximación descriptiva al comportamiento de los clientes morosos y su evolución durante el mes de septiembre de 2025.

A partir del cruce entre las bases del 1° y el 30 de septiembre, se identificaron 18.065 servicios analizados, de los cuales el 78% (14.174 servicios) regularizó su deuda dentro del período, mientras que el 22% restante (3.891 servicios) continuó en situación de morosidad.

Tabla N°2: Cantidad de clientes según su estado final

ESTADO FINAL	SERVICIOS	%
REGULARIZO	14.174	78%
SIGUE_MOROSO	3.891	22%
Total general	18.065	

Fuente: Elaboración propia.

En promedio, los clientes que regularizaron su deuda presentaban un monto adeudado de $133.098 y un promedio de 1,11 facturas impagas, mientras que los clientes que mantuvieron su morosidad registraron un monto medio de $135.456 y un promedio de 2,42 facturas impagas. Esto refleja que los clientes con mayor número de facturas acumuladas tienden a presentar un comportamiento de pago más riesgoso, incluso ante montos similares de deuda.

Tabla N°3: Promedio de montos y de facturas de clientes según su estado final.

ESTADO FINAL	Promedio de MONTO	Promedio de CANTIDAD FACTURAS
REGULARIZO	$ 133.097,96	1,11
SIGUE_MOROSO	$ 135.455,56	2,42
Total general	$ 133.605,76	1,39

Fuente: Elaboración propia.

Desde una perspectiva territorial, se observa que la mayor concentración de servicios corresponde a las localidades de Yerba Buena (6.219 servicios), Los Pocitos (2.509) y Tafí Viejo (2.384), que en conjunto representan más de la mitad del total analizado. En términos proporcionales, tanto Yerba Buena como Los Pocitos y Los Nogales exhiben los niveles más altos de regularización (85%), seguidas por El Cadillal (82%) y Cevil Redondo (79%), lo que refleja un comportamiento de pago favorable y una efectiva respuesta a las gestiones de cobranza.

Tabla N°4: Cantidad de clientes según su localidad y su estado final.

LOCALIDAD	REGULARIZO	SIGUE_MOROSO	Total general
YERBA BUENA	5.264	955	6.219
LOS POCITOS	2.128	381	2.509
TAFI VIEJO	1.708	676	2.384
CEVIL REDONDO	1.001	267	1.268
RACO	759	396	1.155
LOS NOGALES	973	175	1.148
TRANCAS	623	193	816
SAN P. DE COLAL	412	389	801
EL CADILLAL	585	129	714
CHOROMORO	401	175	576
TAPIA	140	37	177
TICUCHO	55	53	108
NUEVA ESPERAN	63	28	91
SAN JAVIER	28	28	56
VIPOS	34	9	43
Total general	14.174	3.891	18.065

Fuente: Elaboración propia.

Tabla N°5: Porcentaje de clientes según su localidad y su estado final.

LOCALIDAD	REGULARIZO	SIGUE_MOROSO
YERBA BUENA	85%	15%
LOS POCITOS	85%	15%
TAFI VIEJO	72%	28%
CEVIL REDONDO	79%	21%
RACO	66%	34%
LOS NOGALES	85%	15%
TRANCAS	76%	24%
SAN P. DE COLALAO	51%	49%
EL CADILLAL	82%	18%
CHOROMORO	70%	30%
TAPIA	79%	21%
TICUCHO	51%	49%
NUEVA ESPERANZA	69%	31%
SAN JAVIER	50%	50%
VIPOS	79%	21%
Total general	78%	22%

Fuente: Elaboración propia.

Por el contrario, las localidades de San Pedro de Colalao (49%), Ticucho (49%), San Javier (50%), Raco (34%) y Nueva Esperanza (31%) presentan los índices más elevados de morosidad relativa, lo que evidencia focos geográficos críticos donde la probabilidad de incumplimiento es considerablemente mayor. En particular, San Pedro de Colalao combina una alta proporción de morosos (49%) con un volumen significativo de casos (801 servicios), constituyéndose en una de las zonas de mayor riesgo operativo.

Estos resultados permiten identificar patrones espaciales de morosidad y segmentar la gestión de cortes y cobranzas según la respuesta histórica de cada zona. De esta manera, las áreas con altos niveles de cumplimiento pueden requerir estrategias de mantenimiento y seguimiento preventivo, mientras que las zonas con menor regularización deben priorizarse en la planificación de operativos y asignación de móviles, optimizando así la eficacia de la gestión territorial de la deuda.

Al segmentar los clientes según la variable “¿Consumo >0?”, que distingue entre servicios con consumo (morosos activos) y sin consumo (morosos pasivos), se evidencian diferencias significativas en el comportamiento de pago. Los resultados muestran que el 83% de los clientes con consumo regularizó su deuda, mientras que solo el 17% permaneció moroso. En contraste, entre los servicios sin consumo registrado, únicamente el 38% logró regularizar, y un 62% continuó en situación de morosidad.

Tabla N°6: Cantidad de clientes según su consumo y su estado final.

¿CONSUMO >0?	REGULARIZO	SIGUE_MOROSO	Total general
NO	695	1.129	1.824
SI	13.479	2.762	16.241
Total general	14.174	3.891	18.065

Fuente: Elaboración propia.

Tabla N°7: Porcentaje de clientes según su consumo y su estado final.

¿CONSUMO >0?	REGULARIZO	SIGUE_MOROSO
NO	38%	62%
SI	83%	17%
Total general	78%	22%

Fuente: Elaboración propia.

Estos valores confirman que los servicios sin consumo (en muchos casos vinculados a domicilios deshabitados o reconexiones ilegales) presentan una propensión considerablemente mayor a la permanencia en mora, representando un segmento de alto riesgo operativo para la empresa. Por el contrario, los clientes con consumo activo tienden a responder con mayor efectividad a las gestiones de cobro y corte, lo que sugiere que este grupo constituye el principal foco de recuperación potencial dentro del universo de morosos analizados.

Imagen N°5: Análisis estadístico

```
> summary(df$monto)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   36110   69190  133606  136960 24944500
```

Fuente: Elaboración propia en RStudio.

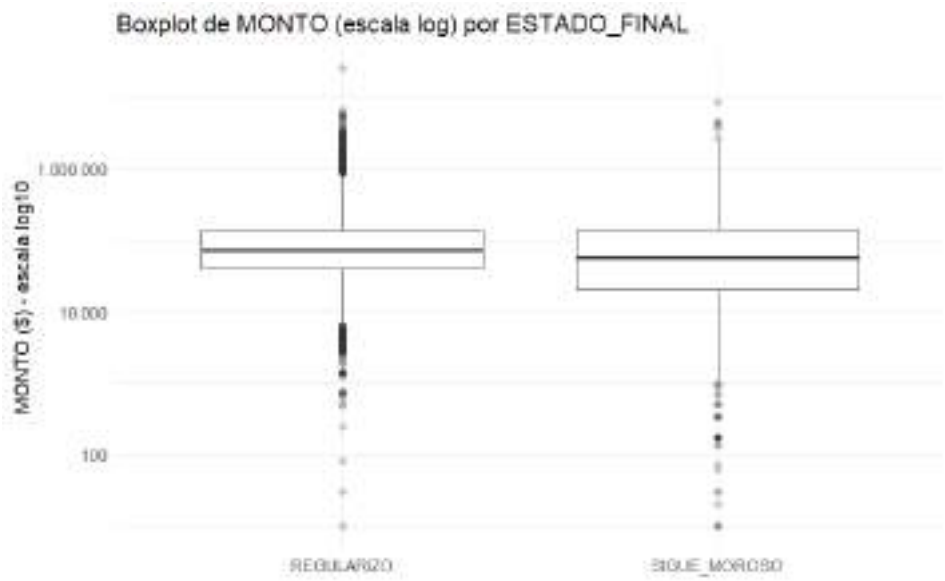
El análisis estadístico del monto adeudado por cliente permitió observar una alta dispersión y una marcada asimetría positiva en la distribución de los datos. Los resultados muestran que el 25% de los clientes registra deudas inferiores a $36.110, mientras que el 50% mantiene montos menores a $69.190. No obstante, el promedio asciende a $133.606, lo que evidencia la presencia de casos con deudas significativamente mayores que elevan la media general.

El 75% de los clientes debe menos de $136.960, pero existen valores extremos que alcanzan los $24.944.500, correspondientes a clientes institucionales o grandes consumidores, los cuales constituyen outliers dentro del conjunto de datos.

Para analizar la distribución del monto adeudado según el estado final del cliente, se elaboraron dos boxplots comparativos: uno con los valores originales y otro en escala logarítmica (log 10), con el fin de reducir la distorsión causada por las deudas

extremadamente altas. El gráfico en escala logarítmica permite observar mejor las diferencias entre ambos grupos y la concentración de los datos.

Imagen N°6: Boxplot de monto por estado final




```
# A tibble: 2 × 6
  estado_final     n monto_mediana    q1     q3    iqr
  <fct>        <int>         <dbl> <dbl>  <dbl>  <dbl>
1 REGULARIZO   14174         71765 39940 136050  96110
2 SIGUE_MOROSO  3891         57080 20020 142340 122320
```

Fuente: Elaboración propia en RStudio.

En este boxplot se aprecia que las líneas de las medianas son relativamente cercanas: los clientes que regularizaron su deuda presentan una mediana en torno a $71.765, mientras que los que siguen morosos tienen una mediana levemente menor, cercana a $57.080. Sin embargo, la caja del grupo “Sigue moroso” es más amplia, lo que refleja una mayor dispersión de los montos en ese segmento. En ambos casos, se observan valores atípicos hacia la parte superior, correspondientes a clientes con deudas muy elevadas.

Esta forma de la distribución representa una asimetría positiva, es decir, la mayoría de los clientes concentra deudas bajas o moderadas, mientras que un número reducido de casos presenta montos muy altos que “estiran” la distribución hacia la derecha. En otras palabras, la mayor parte de los servicios adeuda sumas menores a $150.000, pero existen algunos con deudas millonarias que elevan considerablemente el promedio general.

En síntesis, el análisis sugiere que los montos de deuda no difieren de manera sustancial entre quienes regularizan y quienes no, aunque los morosos persistentes presentan mayor variabilidad. Esto confirma la necesidad de considerar otras variables

explicativas (cantidad de facturas, consumo y zona) para modelar la probabilidad de pago en la siguiente etapa de machine learning.

Machine Learning: Modelo de predicción de regularización de morosos:

Población de estudio:

La población analizada corresponde a los servicios de la Administración Tafí Viejo y sucursales asociadas de Lucecitas S.A., tal como se exportan de la base "FACIMPAGAS".

Preparación de los datos:

De las columnas originales de la base "FACIMPAGAS", se seleccionaron las variables más representativas para la gestión de cobranzas. Durante esta etapa se aplicaron diversas transformaciones y controles:

- Transformación del monto adeudado: (monto_log), con el fin de reducir la asimetría positiva y atenuar el efecto de valores extremos (outliers) detectados en el análisis exploratorio.
- Creación de variable binaria de consumo (consumo pos): se derivó de la columna "¿CONSUMO > 0?", tomando el valor 1 cuando el servicio registró consumo activo y 0 cuando no lo tuvo.
- Cantidad de facturas (cantidad facturas): se mantuvo como variable numérica al reflejar la antigüedad o persistencia de la deuda.
- Variables categóricas: tarifa, tipo_servicio, localidad y zona fueron tratadas como factores. La variable zona fue agrupada y reducida para evitar un número excesivo de categorías: se conservaron las zonas más frecuentes y el resto se unificó bajo la categoría "OTRAS", conformando así la variable zona_red. En las variables categóricas se estableció una categoría de referencia (por ejemplo, "OTRAS" o "RESIDENCIAL"), frente a la cual se compararon los demás niveles.
- Eliminación de variables redundantes o vacías: se excluyeron columnas como ADM, MOTIVO y BARRIO, debido a su falta de cobertura o a su redundancia con otras variables como localidad o zona.
- Tratamiento de valores faltantes: se eliminaron registros con valores nulos en las variables seleccionadas mediante el comando drop_na() de R, garantizando que el modelo se ajustara solo con datos completos y consistentes.
- División de la muestra y validación: Una vez depurada la base, se creó un nuevo conjunto de datos denominado dfm, compuesto por las variables seleccionadas y sin valores faltantes. Para evaluar el desempeño del modelo de forma independiente, se dividió la muestra en dos subconjuntos: 70% de los casos para entrenamiento (train), con los cuales se ajustó el modelo, y 30% para prueba (test), utilizados posteriormente para validar su capacidad predictiva. Esta división se realizó de manera estratificada, preservando la proporción entre clientes que regularizaron y aquellos que siguieron morosos, de modo que ambas particiones mantuvieran una distribución similar de la variable dependiente.

Las imágenes de esta etapa se encuentran adjuntadas en el apéndice.

Variable objetivo o dependiente (qué queremos predecir):

Se construyó una variable binaria llamada "y" que toma:

- **1 =** "SIGUE_MOROSO" al cierre del mes,
- **0 =** "REGULARIZO" (pagó o dejó de figurar como moroso).

En otras palabras: el modelo intenta predecir la probabilidad de que un servicio siga moroso a fin de mes.

Variables explicativas:

Se usó información disponible y operativamente útil para cobranza:

- consumo_pos (binaria): indica consumo activo en el período.
- cantidad_facturas (numérica): cantidad de facturas adeudadas.
- monto_log (numérica): logaritmo del monto total adeudado.
- tarifa (categórica): categoría tarifaria (residencial, T2, T4, etc.).
- tipo_servicio (categórica): por ejemplo, PARTICULARES, GOB. MUNICIPAL, SERVICIO DE PEAJE, etc.
- localidad (categórica): localidades relevantes (Tafí Viejo, San Pedro de Colalao, etc.).
- zona (categórica, reducida): se agruparon muchas zonas y se conservaron las más frecuentes/relevantes; el resto quedó como categoría de referencia.

Modelo utilizado:

Se empleó un modelo de regresión logística binaria. El modelo de regresión logística binaria permite estimar la probabilidad de ocurrencia de un evento dicotómico; en este caso, que un servicio siga moroso **(y = 1)** o regularice su situación **(y = 0)** en función de un conjunto de variables explicativas $X_1, X_2, \dots, X_k$.

La forma funcional del modelo es:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

Donde:

- $\boldsymbol{P(Y = 1)}$ es la probabilidad estimada de que el cliente siga moroso.
- $\boldsymbol{\beta_0}$ es el intercepto del modelo.
- $\boldsymbol{\beta_i}$ son los coeficientes logísticos asociados a cada variable explicativa.
- $\boldsymbol{X_i}$ representan las variables incluidas (por ejemplo, cantidad de facturas, monto, consumo activo, etc.).
- $\boldsymbol{e}$ es la base de los logaritmos naturales.

Si se transforma esta expresión logarítmicamente, se obtiene el modelo lineal en términos del **logit**, es decir, el logaritmo del cociente entre la probabilidad de seguir moroso y la de regularizar:

$$logit\ (p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

El **logit** expresa los **odds ratios** (razones de posibilidades). Cuando $\boldsymbol{\beta_i > 0}$, la variable $\boldsymbol{X_i}$ incrementa las chances de que el cliente siga moroso; si $\boldsymbol{\beta_i < 0}$, las reduce.

Al exponenciar los coeficientes, se obtienen los odds ratios, que facilitan la interpretación práctica:

$$OR = e^{\beta_i}$$

Donde:

- $\boldsymbol{OR}$ indica cuanto se multiplican las chances de que ocurra el evento (seguir moroso) al aumentar una unidad la variable X_i.
- Si $\boldsymbol{OR > 1}$ → la variable aumenta la probabilidad de seguir moroso.
- Si $\boldsymbol{OR < 1}$ → la variable disminuye la probabilidad de seguir moroso.

Por ejemplo, un $\boldsymbol{OR = 2,3}$ para la variable *cantidad_facturas* significa que, por cada factura impaga adicional, las chances de que el cliente siga moroso se multiplican por $\boldsymbol{2,3}$ veces, manteniendo constantes las demás variables.

Verificación de los supuestos del modelo:

Para que el modelo logístico tenga validez y sus resultados sean interpretables, deben verificarse ciertos supuestos estadísticos básicos:

1. Independencia de las observaciones: Cada registro de la base corresponde a un servicio distinto, identificado mediante su número de servicio. No existen duplicados ni dependencias temporales entre registros, ya que para cada servicio se observa únicamente su estado de morosidad al cierre del mes. Por lo tanto, las observaciones pueden considerarse independientes entre sí y este supuesto se cumple adecuadamente.

Imagen N°7: Supuesto N°1

```
> # Verificar duplicados
> sum(duplicated(df))
[1] 0
```

Fuente: Elaboración propia en RStudio.

2. Ausencia de multicolinealidad: Para evaluar la correlación entre las variables numéricas incluidas en el modelo (monto_log y cantidad_facturas), se calculó la matriz de correlación, obteniéndose un valor de **r = –0.079**, prácticamente nulo. Esto indica que ambas variables no están relacionadas linealmente y no generan colinealidad entre sí. Adicionalmente, las variables categóricas fueron depuradas y reagrupadas (por ejemplo, zonas reducidas), evitando redundancias entre categorías. En consecuencia, el modelo no presenta problemas de multicolinealidad y el supuesto se considera cumplido.

Imagen N°8: Supuesto N°2

```
> cor(df[, c("monto_log", "cantidad_facturas")], use="complete.obs")
                monto_log cantidad_facturas
monto_log       1.00000000       -0.07932531
cantidad_facturas -0.07932531       1.00000000
```

Fuente: Elaboración propia en RStudio.

3. Tamaño muestral adecuado: La regresión logística requiere un número suficiente de casos en la categoría menos frecuente para asegurar estimaciones estables. En este estudio, la clase minoritaria (“sigue moroso”) cuenta con 3.891 observaciones, muy por encima de los requisitos teóricos del criterio EPV (Events Per Variable). Con más de 17.000 observaciones útiles tras el preprocesamiento, la muestra excede ampliamente el mínimo necesario para obtener un modelo robusto.

Imagen N°9: Supuesto N°3

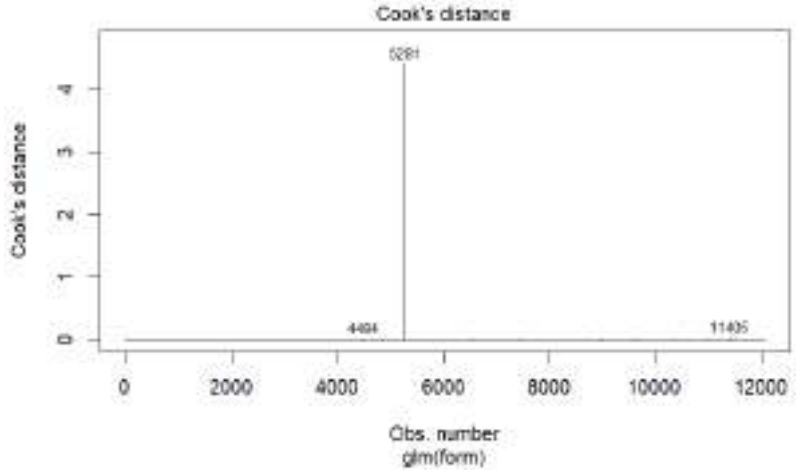
```
> table(df$y)

    0     1
14174  3891
```

Fuente: Elaboración propia en RStudio.

4. Ausencia de valores influyentes o atípicos extremos: La influencia de cada observación sobre los coeficientes del modelo se evaluó mediante el estadístico Cook’s Distance. Tal como se muestra en la Figura correspondiente, ninguna observación supera los umbrales comúnmente aceptados (1 o 4/n), y la gran mayoría de los valores se encuentran prácticamente en cero. Esto indica que no existen registros con influencia desproporcionada sobre el modelo, por lo que este supuesto también se considera satisfecho.

Imagen N°10: Supuesto N°4



Fuente: Elaboración propia en RStudio.

Resultados principales:

Imagen N°11: Variables claves con sus respectivos OR (columna *estimate*)

```
> library(broom)
> coefs_or <- broom::tidy(m1, exponentiate = TRUE)   # estimate = OR
> coefs_or %>% arrange(p.value) %>% print(n = Inf)
# A tibble: 38 x 5
   term                        estimate std.error statistic   p.value
   <chr>                          <dbl>     <dbl>     <dbl>     <dbl>
 1 cantidad_facturas           2.29e+0   0.0380     21.8     4.58e-105
 2 consumo_pos                 4.71e-1   0.0947     -7.96    1.73e- 15
 3 localidadSAN P. DE COLALAO  2.55e+0   0.139       6.69    2.28e- 11
 4 localidadTAFI VIEJO         1.84e+0   0.111       5.30    3.75e-  8
 5 monto_log                   8.94e-1   0.0279     -4.01    6.18e-  5
 6 tipo_servicioPARTICULARES   2.21e-1   0.380      -3.98    7.03e-  5
 7 zona_red5496                5.18e-1   0.217      -3.03    2.48e-  3
 8 zona_red828                 5.69e-1   0.231      -2.44    1.48e-  2
 9 localidadLOS NOGALES        7.22e-1   0.143      -2.28    2.29e-  2
10 localidadLOS POCITOS        7.77e-1   0.116      -2.17    3.02e-  2
11 zona_red1095C               4.66e-1   0.374      -2.04    4.12e-  2
12 (Intercept)                 2.64e+0   0.503       1.93    5.37e-  2
13 zona_red2034                5.36e-1   0.346      -1.79    7.38e-  2
14 zona_red4809                6.40e-1   0.263      -1.69    9.08e-  2
15 tarifaT2                    5.06e-1   0.440      -1.55    1.22e-  1
16 localidadVIPOS              1.94e-1   1.07       -1.53    1.25e-  1
17 localidadRACO               8.06e-1   0.149      -1.45    1.47e-  1
18 zona_red4813                7.17e-1   0.239      -1.39    1.64e-  1
19 zona_red477                 7.33e-1   0.240      -1.29    1.95e-  1
20 localidadTRANCAS            8.18e-1   0.163      -1.23    2.18e-  1
21 localidadEL CADILLAL        8.37e-1   0.161      -1.11    2.69e-  1
22 zona_red949                 7.47e-1   0.275      -1.06    2.90e-  1
23 tarifaTIR                   9.10e-1   0.0954     -0.991   3.22e-  1
24 zona_red87                  1.17e+0   0.267       0.591   5.55e-  1
25 localidadNUEVA ESPERANZA    1.21e+0   0.339       0.555   5.79e-  1
26 zona_red2422                8.68e-1   0.270      -0.512   6.08e-  1
27 localidadYERBA BUENA        1.05e+0   0.108       0.434   6.64e-  1
28 zona_red2160                8.93e-1   0.292      -0.388   6.98e-  1
29 localidadCHORROMORO         9.41e-1   0.179      -0.337   7.36e-  1
30 localidadTAPIA              9.41e-1   0.287      -0.212   8.32e-  1
31 localidadSAN JAVIER         1.09e+0   0.427       0.194   8.46e-  1
32 localidadTICUCHO            9.46e-1   0.482      -0.115   9.08e-  1
33 zona_red39                  9.84e-1   0.200      -0.0777  9.38e-  1
34 tipo_servicioGRANDES CLIENTES 1.95e-5 177.       -0.0613  9.51e-  1
35 tarifaT4B                   8.95e-1   209.       -0.0459  9.63e-  1
36 tarifaT3                    4.81e+5   325.        0.0403  9.68e-  1
37 tarifaT6M                   4.99e-6   325.       -0.0376  9.70e-  1
```

Fuente: Elaboración propia en RStudio.

1) Importancia y sentido de las variables:

A partir de la tabla de OR (ya convertidos desde los coeficientes logísticos), los hallazgos más relevantes para gestión (aquellos con p value < 0.05) son:

Tabla N°8: Interpretación de OR de las variables claves

Variable	OR (estimate)	Interpretación
cantidad_facturas	2.29	Por cada factura impaga adicional, las chances de que el cliente **siga moroso se multiplican por 2.29**. Es el predictor más fuerte del modelo.
consumo_pos	0.47	Los clientes con consumo activo tienen **un 53% menos** de probabilidades de seguir morosos que los que no tienen consumo.
***localidad* SAN P. DE COLALAO**	2.53	Los clientes de esta localidad tienen **2,5 veces más chances de seguir morosos** que los de otra localidad.
***localidad* TAFI VIEJO**	1.84	Los de **Tafí Viejo tienen 84% más probabilidades de continuar morosos** que los de otras localidades.
monto_log	1.09	A medida que aumenta el monto de deuda, **crece un 9%** la probabilidad de seguir moroso.
***tipo_servicio* PARTICULARES**	0.22	Los servicios particulares tienen una probabilidad **78% menor** de seguir morosos frente a los servicios del Gobierno (categoría base).
zona_red3496	0.51	En esta zona, las chances de seguir moroso son **49% menores** a otras zonas.
zona_red828	0.57	En esta zona, se encuentra una probabilidad **43% menor** de continuar moroso.
***localidad* LOS NOGALES**	0.72	**Reducción del 28%** en las probabilidades de seguir moroso.
***localidad* LOS POCITOS**	0.77	Probabilidad **23% menor** de permanecer moroso.
zona_red1095C	0.46	Disminuye un **54% la probabilidad de morosidad** respecto a las otras zonas.

Fuente: Elaboración propia.

2) Capacidad global del modelo (AUC):

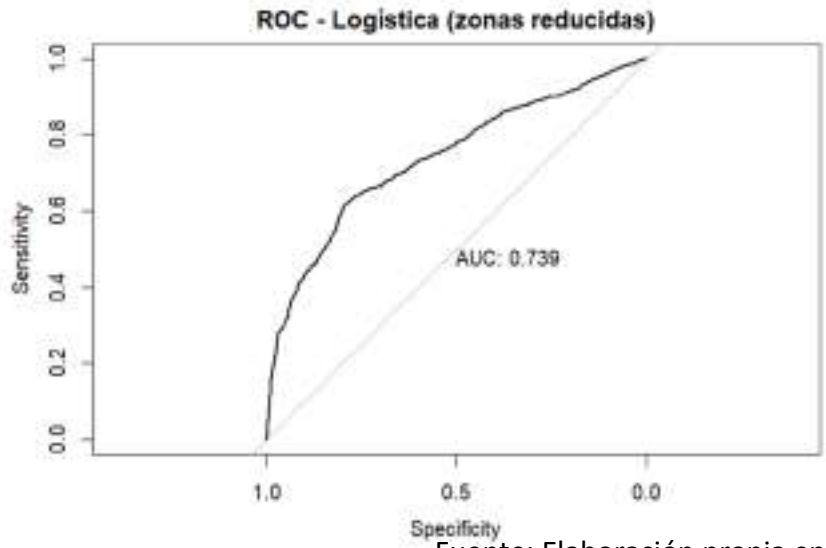
Para evaluar el poder de discriminación del modelo, es decir, qué tan bien separa "seguirá moroso" vs. "regularizó", usamos la curva ROC y su área (AUC).

¿Qué es AUC? Es un número entre 0,5 y 1:

- 0,5 = adivinar al azar.
- 0,7–0,8 = "aceptable a bueno".

- 0,8–0,9 = muy bueno.
- >0,9 = excelente.

Imagen N°12: Curva ROC



Fuente: Elaboración propia en RStudio.

Resultado**: AUC ≈ 0,739** → el modelo tiene capacidad moderada/buena para distinguir morosos persistentes de regularizados. Por lo tanto, es un modelo útil para priorizar servicios en la gestión diaria. No es perfecto (ningún modelo lo es), pero mejora significativamente a seleccionar al azar o por una sola regla fija.

3) Umbral de decisión:

El modelo de regresión logística desarrollado devuelve una probabilidad individual para cada cliente, que representa el riesgo de continuar moroso al cierre del mes. Este valor, que varía entre 0 y 1, expresa de manera intuitiva qué tan probable es que un cliente no regularice su deuda: cuanto más cercano esté a 1, mayor es el riesgo de persistir en la morosidad; cuanto más próximo a 0, mayor es la probabilidad de que pague o se normalice.

De esta forma, el modelo transforma los datos históricos en un indicador de riesgo concreto que puede utilizarse para priorizar acciones de cobranza y planificar la gestión de manera más eficiente. Sin embargo, para poder convertir esas probabilidades en decisiones prácticas (por ejemplo, decidir a quién contactar), es necesario definir un umbral de decisión. El umbral es el punto de corte que separa a los clientes de alto riesgo de los de bajo riesgo, y su valor determina la sensibilidad del modelo.

Por ello, se seleccionó el umbral que maximiza el F1-score (≈ 0,37), un criterio que permite lograr un compromiso entre precisión y capacidad de detección, priorizando el rendimiento global del modelo frente al desbalance de clases. Este criterio proviene del campo del machine learning y se utiliza cuando el objetivo es identificar

con mayor certeza a la clase minoritaria, (en este caso, 19% de los servicios eran morosos), sin aumentar en exceso los falsos positivos.

Con este umbral, el modelo alcanza:

- Exactitud = 0,84 (proporción total de predicciones correctas)
- Sensibilidad = 0,26 (proporción de morosos realmente identificados)
- Especificidad = 0,97 (proporción de regularizados correctamente identificados)

En términos simples, este umbral prioriza la correcta identificación de los clientes que efectivamente regularizan, mientras mantiene un nivel acotado de falsos positivos. Sin embargo, la capacidad de detección de morosos es moderada debido a la baja sensibilidad obtenida.

De este modo, todo cliente con una probabilidad igual o superior a 0,37 será considerado de alto riesgo de morosidad, mientras que aquellos con una probabilidad inferior se clasificarán como probables regularizadores.

El uso del umbral F1 resulta útil en contextos con clases desbalanceadas, ya que prioriza el rendimiento general del modelo y la eficiencia operativa. No obstante, este criterio tiende a favorecer la correcta clasificación de regularizadores por encima de la detección de morosos, debido a la sensibilidad obtenida.

Es importante remarcar que la elección del umbral depende de las prioridades operativas de la Administración. No obstante, si la Administración Tafí Viejo prioriza detectar la mayor cantidad posible de morosos, el umbral podría ajustarse a valores más bajos (por ejemplo, 0,25), aumentando la sensibilidad del modelo. De esta forma, se ampliaría la cobertura de detección, aun a costa de incluir algunos falsos positivos, lo cual puede ser ventajoso en contextos donde la prevención del impago resulta más valiosa que la eficiencia operativa.

Conclusiones del Modelo:

En conclusión, el modelo queda entrenado y listo para su aplicación en los próximos períodos, permitiendo asignar a cada servicio una probabilidad de morosidad personalizada y clasificar automáticamente los casos según su nivel de riesgo.

Esto convierte la herramienta en un sistema predictivo capaz de anticipar el comportamiento de los clientes, optimizar la toma de decisiones y orientar los recursos hacia los casos con mayor impacto potencial, cumpliendo así con el objetivo general de generar información estratégica sobre la deuda y mejorar la gestión de cobranzas en la Administración Tafí Viejo y sus sucursales.

Análisis de Correspondencia Simple:

El Análisis de Correspondencias Simple (AC), o bivariado, es una técnica estadística que permite estudiar la relación entre dos variables categóricas. A partir de una tabla de contingencia, el AC construye un mapa geométrico donde cada categoría se representa como un punto.

La distancia entre los puntos refleja qué tan similares o diferentes son las categorías:

- **Puntos cercanos** → categorías que aparecen juntas con frecuencia
- **Puntos alejados** → categorías que rara vez coinciden.

Población analizada y variables incluidas:

Se utilizó la misma base depurada aplicada previamente en el modelo de regresión logística, correspondiente a la población total de clientes incluidos en el dataset FACIMPAGAS_RESUMEN, compuesta por 18.065 registros, donde cada fila representa un cliente con su estado de pago y variables descriptivas.

Para este análisis preliminar se seleccionaron dos variables cualitativas centrales:

1. Estado final del cliente (variable dependiente del modelo logístico):
   - REGULARIZÓ
   - SIGUE_MOROSO
2. Consumo posterior al vencimiento (variable explicativa clave):
   - Con consumo posterior
   - Sin consumo posterior

Ambas variables ya habían demostrado relevancia en la regresión logística; el AC permite evaluarlas desde un enfoque exclusivamente descriptivo y geométrico.

Resultados:

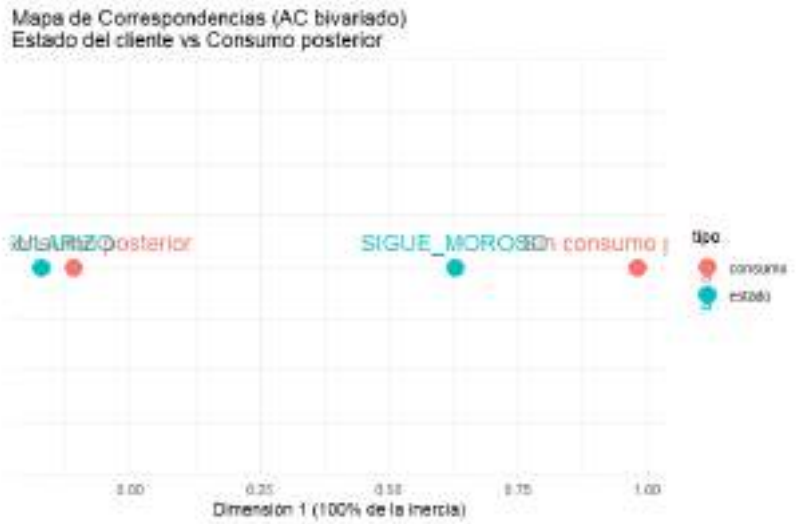
Tabla N°9: Tabla de contingencia

	Sin consumo posterior	Con consumo posterior
REGULARIZO	695	13479
SIGUE_MOROSO	1129	2762

Fuente: Elaboración propia en RStudio.

La prueba de Chi-cuadrado resultó significativa: $\boldsymbol{X^2 = 1952,7,\ p < 0.001}$ confirmando asociación entre ambas variables.

Imagen N°13: Mapa de Correspondencias



Fuente: Elaboración propia en RStudio.

Como la tabla 2×2 solo posee una dimensión significativa, todas las categorías se proyectan sobre un eje único, donde la proximidad refleja similitud en patrones de comportamiento.

El gráfico muestra:

1. <u>Regularizó está muy cerca de "Con consumo posterior":</u>
   Esto indica que los clientes que tienen consumo activo se comportan de manera similar, y tienen alta probabilidad de regularizar.
2. <u>Sigue Moroso está muy cerca de "Sin consumo posterior":</u>
   Esto nos dice que, aquellos clientes sin consumo presentan perfiles similares a los morosos persistentes, y tienen bajan probabilidad de regularizar.

Conclusión:

El AC simple confirma de forma descriptiva y geométrica un patrón ya sugerido por la regresión logística: el consumo posterior al vencimiento es un indicador conductual fundamental que separa dos perfiles claros de clientes:

- **Clientes que continúan con consumo** → mayor probabilidad de regularización.
- **Clientes sin consumo activo** → mayor probabilidad de mantenerse morosos.

La única dimensión del AC bivariado captura el 100% de la asociación, evidenciando que esta variable es altamente discriminante.

Este resultado simple motivó la ampliación del análisis hacia un Análisis de Correspondencias Múltiple (ACM), incorporando simultáneamente otras variables cualitativas (tarifa, zona, tipo de servicio, etc.) para identificar perfiles más complejos y patrones multivariados de comportamiento.

Análisis de Correspondencia Múltiple:

El Análisis de Correspondencias Múltiple (ACM) es una técnica estadística multivariante de carácter exploratorio y descriptivo, utilizada para identificar asociaciones entre categorías de variables cualitativas y representar dichas relaciones en un espacio de baja dimensión, generalmente en un plano bidimensional.

A diferencia de la regresión logística (que busca predecir una variable dependiente), el ACM tiene como objetivo visualizar patrones de comportamiento entre grupos de categorías, permitiendo construir perfiles de individuos o segmentos con características similares.

En este caso, el ACM se aplicó para profundizar la interpretación del comportamiento de los clientes morosos, complementando al AC simple y a la regresión logística binaria. Mientras la regresión permitió estimar probabilidades de regularización, el ACM ofrece una visión estructural y gráfica de cómo se agrupan las distintas categorías de variables (tarifa, tipo de servicio, consumo, etc.) en torno a los dos resultados posibles: REGULARIZÓ o SIGUE MOROSO.

El método transforma la información categórica en un conjunto de ejes factoriales (Dimensiones), donde cada punto representa una categoría, y las distancias entre puntos reflejan la similitud o asociación entre ellas.

Población y variables utilizadas:

La población es la misma que se utilizó en la regresión logística, garantizando coherencia entre ambos modelos. Las variables seleccionadas para el ACM son de tipo categórico, dado que este método requiere variables cualitativas o discretizadas. Se definieron de la siguiente manera:

Tabla N°10: Descripción de variables y su rol en el ACM

Tipo	Variable	Descripción	Rol en el ACM
Variable suplementaria	estado_final	Resultado del cliente: REGULARIZÓ / SIGUE MOROSO	No participa en el cálculo, pero se proyecta en el mapa
Activa	tarifa	Tipo de tarifa aplicada al cliente (T1R, T1G, T6M, TDI, etc.)	Permite distinguir perfiles tarifarios
Activa	tipo_servicio	Categoría de servicio (PARTICULARES, GOB.	Identifica el tipo de cliente

		MUNICIPAL, GRANDES CLIENTES, PEAJE, etc.)	
Activa	consumo_pos	Indica si el cliente tuvo consumo posterior al vencimiento (1 = Sí / 0 = No)	Refleja la actividad del servicio
Activa	zona / localidad	Ubicación geográfica del cliente	Analiza diferencias regionales (con menor peso por alta dispersión)

Fuente: Elaboración propia.

Estas variables son las mismas utilizadas en la regresión logística, pero tratadas aquí como modalidades categóricas sin suponer relación lineal entre ellas.

El motivo de incluir las mismas variables es lograr una comparabilidad directa entre los resultados del modelo predictivo (logístico) y los resultados descriptivos (ACM), fortaleciendo la validez del análisis.

Resultados del análisis:

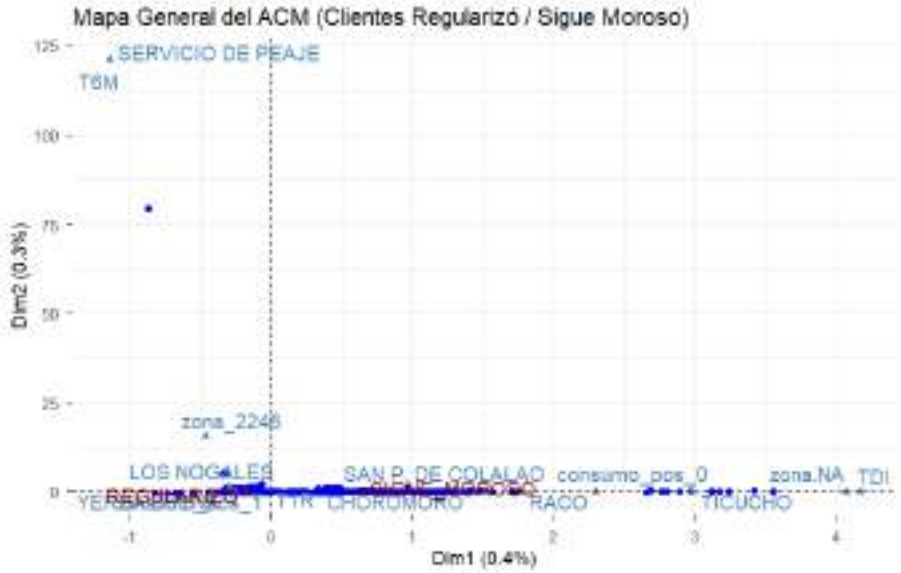
1) Inercia y estructura factorial:

El ACM mostró que las dos primeras dimensiones (Dim1 y Dim2) explican conjuntamente un 0,7% de la inercia total. Aunque el porcentaje parece bajo, esto es habitual cuando existen muchas categorías dispersas (como localidades y zonas). Por ello, la interpretación se centró en las categorías con mayor contribución y mejor calidad de representación (cos^2).

2) Mapa General del ACM:

El mapa general muestra la proyección conjunta de las variables activas y la variable suplementaria estado_final. Los puntos en azul representan las categorías activas, mientras que los en rojo corresponden a los resultados REGULARIZÓ y SIGUE MOROSO.

Imagen N°14: Mapa general del ACM



Fuente: Elaboración propia en RStudio.

Eje 1 (Dim1): distingue entre clientes activos y regulares (izquierda) y morosos persistentes o sin consumo (derecha). Se observa que:

- A la izquierda, se ubican las categorías vinculadas con clientes residenciales, tarifas T1R y T1G, localidades como Yerba Buena y Los Nogales, todos cercanos a REGULARIZÓ.
- A la derecha, aparecen categorías asociadas a SIGUE MOROSO, como la localidad San P. de Colalao, Raco, Ticucho, y principalmente consumo_pos_0 (sin consumo).

Eje 2 (Dim2): refleja diferencias secundarias entre tipos de servicio y tarifas. Se observa que:

- SERVICIO DE PEAJE y T6M se ubican en los extremos superiores del gráfico. Estos puntos se encuentran alejados del centro, lo que sugiere perfiles muy específicos que poseen comportamientos de morosidad particulares.

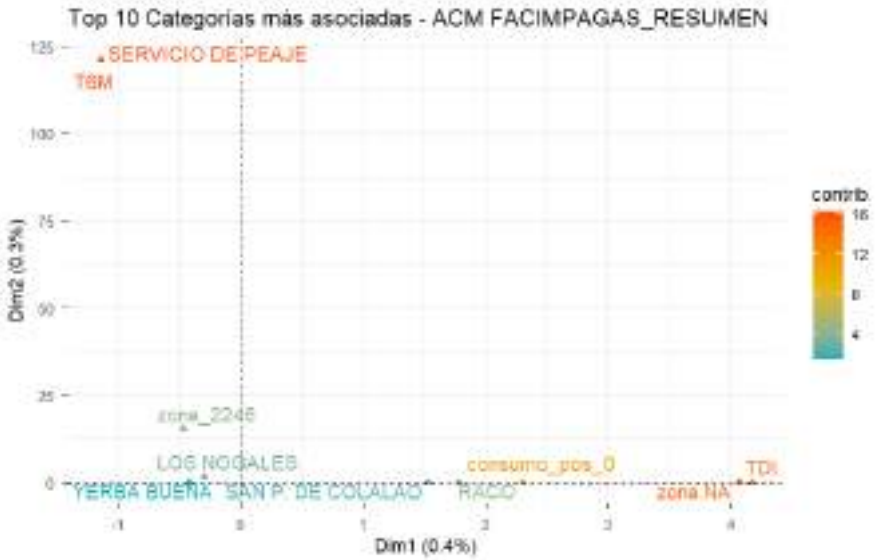
El mapa general permite visualizar con claridad los dos polos conductuales de los clientes: Un grupo residencial y activo, asociado a REGULARIZÓ. Un grupo con consumo nulo y localización periférica, asociado a SIGUE MOROSO.

Estos resultados confirman gráficamente las relaciones que el modelo logístico había cuantificado mediante probabilidades y odds ratios.

3) Categorías con mayor contribución al ACM:

El gráfico presenta las categorías que más contribuyen a la construcción de las dos primeras dimensiones (Dim1 y Dim2) del ACM. Las categorías más relevantes se visualizan mediante un gradiente de color, donde los tonos más cálidos (rojos/naranjas) representan mayor contribución, es decir, más peso estructural en la diferenciación de los perfiles de clientes.

Imagen N°15: Categorías con mayor contribución al ACM



Fuente: Elaboración propia en RStudio.

En conjunto, el gráfico muestra que la estructura del ACM está definida principalmente por cuatro factores:

a) Tipo de tarifa (TDI, T6M).
b) Tipo de cliente/servicio (SERVICIO DE PEAJE).
c) Nivel de consumo (consumo_pos_0).
d) Localización geográfica o administrativa (zonas sin identificar y rurales).

Estos factores explican la oposición central del modelo:

- Clientes con tarifa TDI, o con consumo nulo o bajo, se asocian a la persistencia en la morosidad.
- Clientes residenciales de zonas urbanas, como Los Nogales o Yerba Buena, se agrupan en el extremo opuesto, representando comportamientos más regulares.

Conclusiones del Análisis:

El Análisis de Correspondencias Múltiple permitió confirmar y visualizar los mismos patrones identificados por la regresión logística binaria, pero desde una perspectiva estructural y relacional.

Las asociaciones visualizadas entre consumo_pos, tarifa y tipo de servicio coinciden con las variables significativas del modelo logístico. Esto refuerza la robustez del diagnóstico y la consistencia de los hallazgos.

El ACM permitió distinguir dos grupos nítidos: clientes activos y residenciales, vinculados a REGULARIZÓ. Y, además, clientes sin consumo y con tarifas industriales o especiales, vinculados a SIGUE MOROSO. Este contraste es fundamental para orientar las políticas de cobranza y segmentar estrategias.

Recomendaciones

A partir de los hallazgos obtenidos en las etapas cualitativa y cuantitativa, se proponen las siguientes acciones destinadas a fortalecer la gestión de cobranzas en la Administración Tafí Viejo de Lucecitas S.A:

1. Integración y consolidación de la información de morosidad:
    - Implementar una base única e integrada que consolide facturas impagas, consumos, cortes, reconexiones e historial del cliente.
    - Estandarizar criterios de registro entre áreas y sucursales.
    - Automatizar la actualización periódica para evitar errores y reducir trabajo manual.

2. Continuidad y perfeccionamiento del modelo de regresión logística:
    - Actualizar el modelo periódicamente (mensual o bimestral) incorporando nuevos datos.
    - Evaluar distintos umbrales de decisión según las prioridades operativas del período.
    - Incluir nuevas variables predictivas: historial de cortes, días promedio de atraso, comportamiento post–corte, tipo de deuda, estacionalidad.
    - Considerar técnicas de balanceo de clases (pesos ajustados o sobremuestreo) cuando se priorice la detección de morosos.

3. Incorporación de un Análisis Discriminante (AD) como complemento:
    - Comparar su capacidad de clasificación con la regresión logística.
    - Identificar combinaciones lineales de variables con fuerte poder discriminante.
    - Analizar posibles patrones alternativos entre morosos persistentes y regularizadores.

4. Repetición y ampliación del AC y del ACM:
    - Repetir los análisis periódicamente para monitorear cambios en perfiles.
    - Ampliar el ACM incorporando nuevas variables cualitativas (modalidad de pago, edad del servicio, historial de corte).

- Utilizar los resultados para futuros clusters de clientes, facilitando estrategias segmentadas.

5. Desarrollo de tableros de monitoreo y control:
- Diseñar dashboards (Looker Studio o Power BI) para visualizar:
  - ✓ Probabilidad de morosidad por cliente.
  - ✓ Comportamiento por localidad y zona.
  - ✓ Efectividad de cortes y reconexiones.
  - ✓ Evolución del AUC, sensibilidad y exactitud.
  - ✓ Deuda por segmento y alertas automáticas para clientes de alto riesgo.

6. Fortalecimiento del proceso interno de cobranzas:
- Automatizar listados operativos diarios o semanales.
- Reducir la dependencia del conocimiento tácito del personal mediante procedimientos estandarizados.
- Evaluar la necesidad de refuerzo de dotación operativa en el área de Cobranzas.
- Mejorar la articulación Supervisor–Asistente mediante reportes consolidados.

7. Evaluación y retroalimentación continua del sistema predictivo:
- Recalibrar el modelo mensual o bimestralmente y comparar predicciones con resultados reales.
- Incorporar indicadores de costo-beneficio: costo por visita, costo de falsos positivos, monto recuperado.
- Utilizar la retroalimentación para perfeccionar la herramienta predictiva y optimizar la toma de decisiones.

Estas recomendaciones buscan promover una gestión de cobranzas más eficiente, preventiva y basada en evidencia, fortaleciendo la capacidad analítica y operativa de la Administración Tafí Viejo y contribuyendo a la evolución hacia una cultura organizacional centrada en datos.

Conclusiones

El presente trabajo permitió obtener una comprensión integral, profunda y fundamentada del fenómeno de la morosidad en la Administración Tafí Viejo de Lucecitas S.A., integrando una visión cualitativa del funcionamiento operativo con un análisis cuantitativo exhaustivo basado en datos reales de facturación, consumo, y comportamiento de pago. Esta aproximación mixta posibilitó construir un diagnóstico sólido y generar conocimiento aplicado con valor directo para la toma de decisiones de la organización.

La etapa cualitativa reveló, en primer lugar, que la empresa dispone de una gran cantidad de información relevante para la gestión de cobranzas, pero la fragmentación de los registros y la ausencia de un sistema integrado dificultan su aprovechamiento efectivo. La dependencia de planillas manuales, la falta de estandarización y la elevada carga administrativa generan cuellos de botella que reducen la capacidad del área para anticipar situaciones de riesgo. Asimismo, las entrevistas mostraron que la toma de decisiones se apoya principalmente en la experiencia del personal y en criterios operativos tradicionales (monto de deuda, antigüedad, ubicación), lo que limita el alcance estratégico de la gestión y dificulta la detección temprana de patrones de morosidad. Esta información contextual fue esencial para orientar la fase cuantitativa y comprender las necesidades reales del área.

La etapa cuantitativa permitió transformar estas percepciones iniciales en un análisis estructurado del comportamiento de los clientes. El Análisis Exploratorio de Datos reveló que la morosidad no se distribuye de manera uniforme: existen diferencias claras entre localidades, segmentos de consumo y tipos de servicio. Localidades como San Pedro de Colalao, Ticucho y Raco presentan niveles de morosidad persistentemente elevados, mientras que zonas urbanas como Yerba Buena o Los Nogales exhiben tasas significativamente menores. Asimismo, la variable “consumo” resultó uno de los factores más determinantes: los servicios sin consumo presentan una probabilidad sustancialmente mayor de permanecer en mora, un hallazgo consistente tanto con el criterio operativo de los entrevistados como con la evidencia numérica.

Sobre esta base, el modelo de regresión logística permitió cuantificar las relaciones entre variables y estimar probabilidades individuales de morosidad. El análisis confirmó que la cantidad de facturas impagas, la ausencia de consumo, el tipo de servicio y ciertas ubicaciones geográficas inciden significativamente en el riesgo de incumplimiento. El desempeño del modelo, con un AUC de 0,739, mostró una capacidad de discriminación moderada pero útil para la clasificación operativa. Si bien la sensibilidad del modelo fue relativamente baja con el umbral que maximiza el F1-score, este comportamiento es coherente con el desbalance natural de las clases y con la necesidad de priorizar eficiencia en determinadas instancias de gestión. El modelo constituye así una herramienta aplicable que permite priorizar casos, ordenar operativos y orientar recursos hacia los segmentos con mayor impacto potencial.

Complementariamente, el Análisis de Correspondencias Simple y Múltiple permitió visualizar las relaciones estructurales entre categorías cualitativas, revelando perfiles consistentes de comportamiento. El ACM mostró dos polos claramente diferenciados: por un lado, clientes residenciales de zonas urbanas, con consumo activo y alta probabilidad de regularización; por el otro, clientes sin consumo, pertenecientes a zonas rurales o con tarifas especiales, vinculados al riesgo de persistencia en la morosidad. Esta convergencia entre los métodos fortalece la validez interna de los resultados y demuestra que los patrones detectados no son fruto de un único análisis, sino de una estructura subyacente presente en los datos.

En conjunto, los resultados obtenidos permiten concluir que la morosidad en la Administración Tafí Viejo no es un fenómeno aleatorio, sino predecible y segmentable. Existen factores claramente asociados al riesgo de incumplimiento, y su identificación permite diseñar estrategias diferenciadas según el perfil del cliente. La combinación de análisis cualitativos, estadísticos y predictivos brinda una base sólida para la toma de decisiones y constituye un avance significativo hacia un modelo de gestión proactivo, basado en evidencia y orientado a la eficiencia operativa.

Finalmente, el estudio demuestra la utilidad concreta de incorporar herramientas de análisis de datos y modelos predictivos en la gestión de un servicio público esencial como la energía eléctrica. La regresión logística, el análisis discriminante potencial, el AC, el ACM y los futuros tableros de monitoreo (recomendados) representan oportunidades tangibles para optimizar recursos, reducir la morosidad y mejorar la respuesta operativa de la Administración. La continuidad de estas herramientas, sumada a una consolidación de la información y al refuerzo de procesos internos, permitirá avanzar hacia una cultura organizacional basada en datos que fortalezca la sostenibilidad financiera y operativa del servicio. Estas conclusiones no solo sintetizan los hallazgos del trabajo, sino que constituyen una invitación a continuar profundizando el análisis y perfeccionando los mecanismos de gestión de la deuda en periodos futuros.

Referencias

- Aldás, J., & Uriel, E. (2017). Análisis multivariante aplicado con R. Paraninfo.
- Asistic. (2024). *Organizaciones data-driven vs. data-centric*. Recuperado de https://www.asistic.com/
- Arroyo Morales, A. (s.f.). *Metodología de la investigación en las ciencias empresariales*. Universidad Nacional de San Antonio Abad del Cusco.
- Davenport, T., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Harvard Business School Press.
- Diestra Quinto, M., et al. (2023). Aplicación de *machine learning* en la gestión de cobranzas. *Revista de Investigación en Ciencias Empresariales, 12*(2), 55-70.
- Educause. (2022). *Creating a data-informed culture*. EDUCAUSE Review. Recuperado de https://www.educause.edu/
- ENRE. (2023). *Informe anual del Ente Nacional Regulador de la Electricidad*. Buenos Aires: ENRE.
- Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten*. Analytics Press.
- Gartner. (2023). *Data-driven organizations: Strategy and practices*. Gartner Research.
- Hernández Sampieri, R., Fernández Collado, C., & Baptista, P. (2018). *Metodología de la investigación* (6ª ed.). McGraw-Hill.

- Hernández-Sampieri, R., & Mendoza Torres, C. P. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta*. McGraw-Hill Interamericana.
- Innovación Digital 360. (s.f.). *Análisis de datos: Concepto y aplicaciones*. Recuperado de https://www.innovaciondigital360.com/
- Muñoz Cavero, J., & Luyo Pérez, F. (2022). *Machine learning aplicado a la gestión de datos empresariales*. Editorial Académica Española.
- Patiño Pérez, J., Rodríguez, L., & Morales, F. (2020). Modelos de clasificación supervisada: Árboles de decisión y *random forest*. *Revista Colombiana de Computación, 21*(2), 101-115.
- Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.
- Ramírez Mendoza, J. (2022). Algoritmos de *clustering* y segmentación en *machine learning*. *Revista Latinoamericana de Tecnología, 15*(1), 77-89.
- Sandoval, J. (2018). *Introducción al machine learning*. Editorial Alfaomega.
- Villamil Giraldo, M. del P. (2024). Organizaciones basadas en datos: Retos y oportunidades. *Revista de Ingeniería de Información, 18*(3), 12-19.

Apéndice

Guía de entrevista Entrevistado: Supervisor de Cobranzas	**Objetivo:** Obtener una visión integral del funcionamiento del área de cobranzas, identificando cómo se gestionan las deudas de los clientes, qué dificultades enfrenta el área y qué criterios se aplican en la toma de decisiones para mejorar el recupero y reducir la morosidad.

1) Sobre el funcionamiento actual del área de cobranzas:

- ¿Cuándo un cliente pasa a ser considerado moroso o ingresa en proceso de gestión de deuda?
- ¿Podrías describir cómo funciona el área de cobranzas en el día a día, especialmente en la gestión de clientes morosos?
- ¿Cómo se distribuyen las tareas y responsabilidades dentro del equipo de trabajo?
- ¿Qué procesos siguen para gestionar la deuda y hacer el seguimiento de los pagos?
- ¿Cómo se priorizan los clientes o zonas a gestionar según el nivel de morosidad?

2) Sobre la gestión de datos actual:

- ¿Qué tipo de datos son los más relevantes para el área en su operativa diaria? (por ejemplo: montos adeudados, antigüedad de la deuda, zona, tipo de cliente)
- ¿Cómo se recolectan, organizan y utilizan actualmente esos datos en el área?
- ¿Qué dificultades encuentran al manejar grandes volúmenes de información sobre clientes y pagos?

3) Toma de decisiones:

- Cuando debe tomar una decisión operativa o estratégica, ¿en qué aspectos se basa?
- ¿De qué manera utiliza los datos o reportes disponibles para respaldar sus decisiones?
- ¿Considera que cuenta con la información necesaria y en el momento adecuado para decidir?

4) Problemas y desafíos observados:

- ¿Cuáles son los mayores desafíos o problemas que observa actualmente en la gestión de la deuda y el seguimiento de clientes morosos?
- ¿En qué etapas del proceso identifica más dificultades o cuellos de botella?
- ¿Qué cambios cree que podrían mejorar la eficiencia y la efectividad del área de cobranzas?

5) Herramientas útiles y necesidades:

- ¿Considera que el área cuenta con las herramientas de análisis adecuadas para gestionar la deuda? ¿Por qué?
- Si pudiera implementar nuevas herramientas de gestión o control, ¿cuáles serían más útiles para optimizar el trabajo?
- ¿Qué características debería tener una herramienta de apoyo para mejorar la toma de decisiones?
- ¿Qué indicadores considera más importantes para evaluar el desempeño del área en términos de morosidad y recupero?

6) Conclusión:

- ¿Hay algún tema que no hayamos tratado y que consideres relevante para comprender mejor el funcionamiento del área de cobranzas y la gestión de la deuda?

Guía de entrevista Entrevistado: Asistente de Cobranzas.	**Objetivo:** Profundizar en la gestión y análisis de datos del área, relevando qué herramientas utiliza, cómo procesa y presenta la información, y qué indicadores considera relevantes.

1) Sobre el análisis de datos:

- ¿Qué herramientas y métodos utilizas actualmente para analizar los datos relacionados con la gestión de la deuda y los clientes morosos?
- ¿Qué tipo de análisis realizas con esos datos? (por ejemplo: segmentación de clientes, ratios de morosidad, tendencias por zona o por monto).
- ¿Cuáles han sido los hallazgos más significativos que obtuviste a partir del análisis de información sobre la deuda?
- ¿Cómo decidís qué variables o indicadores son los más relevantes para cada informe o análisis?

2) Desafíos y oportunidades:

- ¿Cuáles son los principales desafíos que enfrentás al analizar los datos o al elaborar informes sobre morosidad?
- ¿Considerás que el análisis de datos podría ser más eficiente con alguna herramienta adicional (por ejemplo: software de análisis o tablero de control)? ¿Cuál sería útil?
- ¿Qué limitaciones observás en las bases de datos o en el sistema actual para analizar la deuda y el comportamiento de los clientes?

3) Sobre los informes periódicos:

- ¿Cuáles son los principales procesos que siguen para gestionar la cobranza y hacer el seguimiento de los pagos de los clientes morosos?
- ¿Cómo se priorizan las zonas o los clientes a gestionar según el nivel de deuda o el riesgo de morosidad?
- ¿Qué tipo de información incluís habitualmente en los informes de cobranzas?
- ¿Qué indicadores clave (KPI) considerás más importantes para evaluar la gestión? (por ejemplo: monto recuperado, porcentaje de mora, antigüedad de deuda).
- ¿Sentís que los informes actuales reflejan completamente la situación del área de cobranzas? ¿Qué aspectos podrían mejorarse o incorporarse para tener una visión más completa?

4) Toma de decisiones basada en datos:

- ¿Qué tipo de decisiones importantes se toman en el área a partir de los informes y análisis que elaborás?
- ¿Considerás que los datos actuales permiten tomar decisiones informadas y en el momento oportuno?
- ¿Con qué frecuencia revisás la información con el Supervisor o con otras áreas relacionadas?
- Cuando se debe tomar una decisión operativa, ¿en qué aspectos te basás?
- ¿De qué manera utilizás los datos o reportes disponibles para respaldar esas decisiones?
- ¿Considerás que contás con toda la información necesaria para hacerlo de forma precisa y actualizada?

5) Conclusión:

- ¿Hay algún aspecto del análisis de datos, los informes o el seguimiento de clientes morosos que no hayamos mencionado y que consideres relevante destacar?

Fórmula General del Análisis de Correspondencias Simple:

El AC se basa en la variabilidad de la tabla respecto de un escenario de independencia. Esa variabilidad se conoce como inercia, y su medida fundamental proviene de la estadística Chi-cuadrado:

$$Inercia = \frac{X^2}{n}$$

Donde:

- $\boldsymbol{X^2}$ es el estadístico de Chi-cuadrado de la tabla.
- $\boldsymbol{n}$ es el total de observaciones.

El método toma esta inercia y la descompone en una o más dimensiones, en una tabla 2×2, como en el análisis bivariado aplicado aquí:

$$Dimensionalidad = \min(I - 1, J - 1) = 1$$

Por eso el resultado se representa en un único eje, que explica el 100% de la relación entre las variables.

Fórmula General del Análisis de Correspondencias Múltiple:

El ACM parte de la matriz disyuntiva completa $\boldsymbol{X}$, que representa a $\boldsymbol{n}$ individuos y $\boldsymbol{p}$ variables cualitativas con un total de J categorías.

Cada fila de $\boldsymbol{X}$ corresponde a un individuo, y cada columna a una categoría posible (con valores 1 si el individuo pertenece a esa categoría y 0 en caso contrario).

A partir de esta matriz se construye la matriz de Burt, definida como:

$$B = X'X$$

Donde:

- $\boldsymbol{B}$: matriz de Burt (contiene todas las tablas de contingencia entre las variables)
- $\boldsymbol{X'}$: transpuesta de la matriz disyuntiva X.
- $\boldsymbol{X'X}$: producto que resume la frecuencia conjunta de cada par de categorías.

Luego se realiza una descomposición en valores singulares (SVD) de la matriz B:

$$B = PDP'$$

Donde:

- $\boldsymbol{P}$ es la matriz de autovectores (coordenadas factoriales de las categorías)
- $\boldsymbol{D}$ es una matriz diagonal que contiene los autovalores (X_1, X_2, X_3,…)

Cada autovalor X_i representa la inercia explicada por la Dimensión i, es decir, la proporción de la variabilidad total que cada eje resume.

Finalmente, las coordenadas factoriales de las categorías se calculan como:

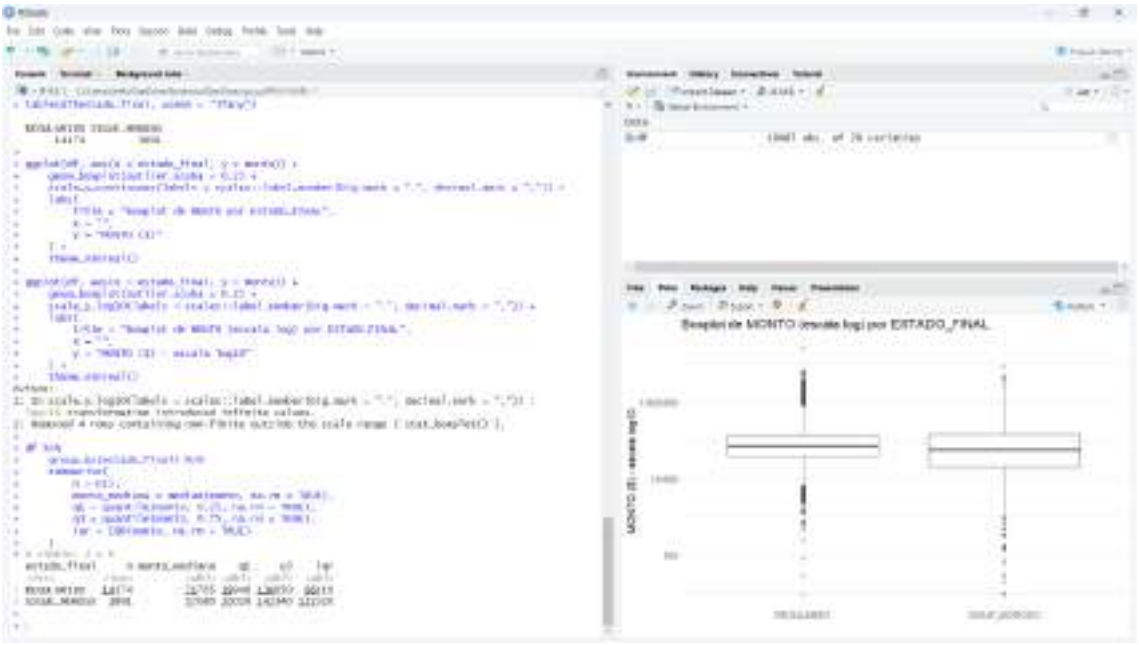
$$F = P\sqrt{D}$$

Estas coordenadas permiten representar las categorías en un espacio de baja dimensión (por ejemplo, en el plano Dim1, Dim2), donde:

- La distancia entre puntos refleja su grado de asociación o similitud.
- Los ejes (dimensiones) son combinaciones lineales de las variables originales que resumen la estructura de correspondencia entre ellas.

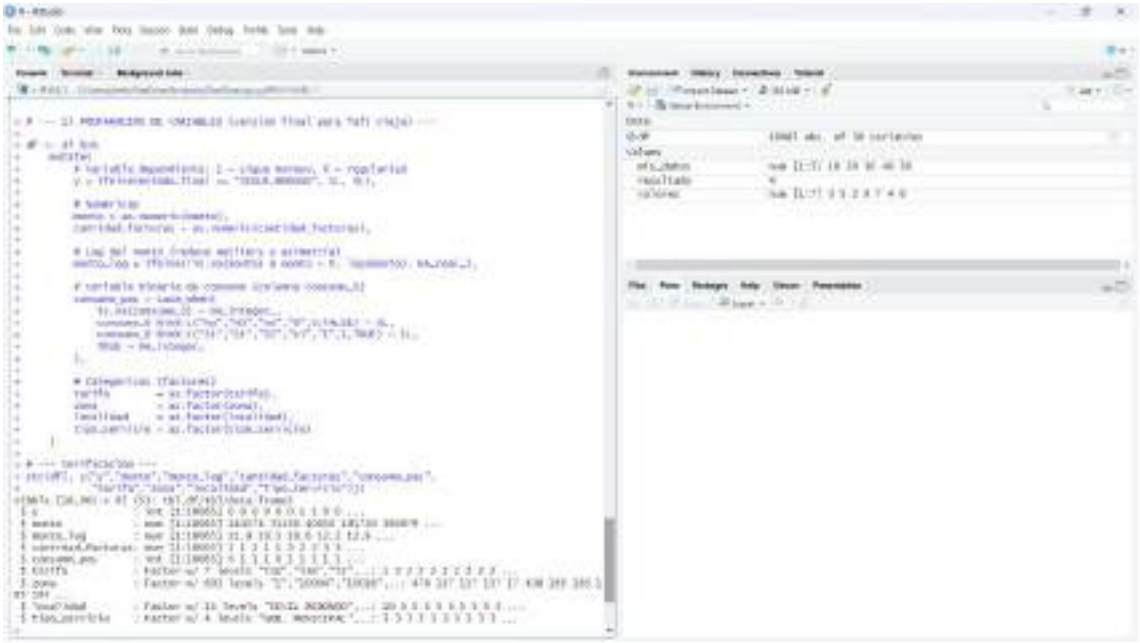
<u>Códigos de RStudio utilizados para este trabajo:</u>

Imagen N°16: Boxplot en R



Fuente: Elaboración propia en RStudio.

Imagen N°17: Preparación de datos en R

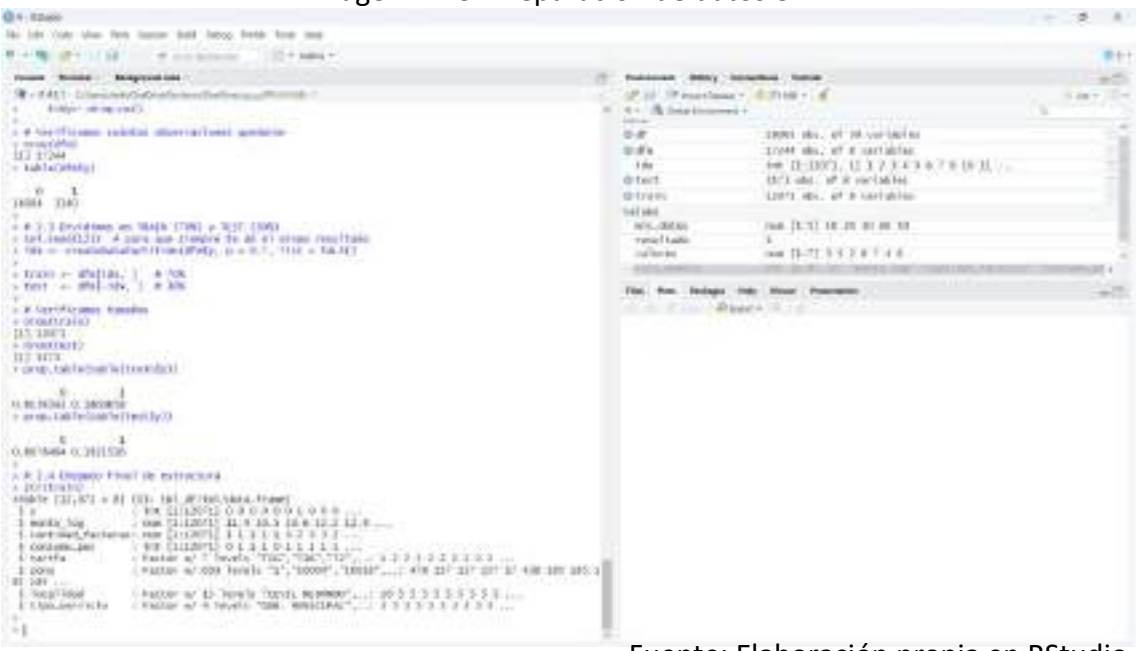


Fuente: Elaboración propia en RStudio.

Imagen N°18: Preparación de datos en R

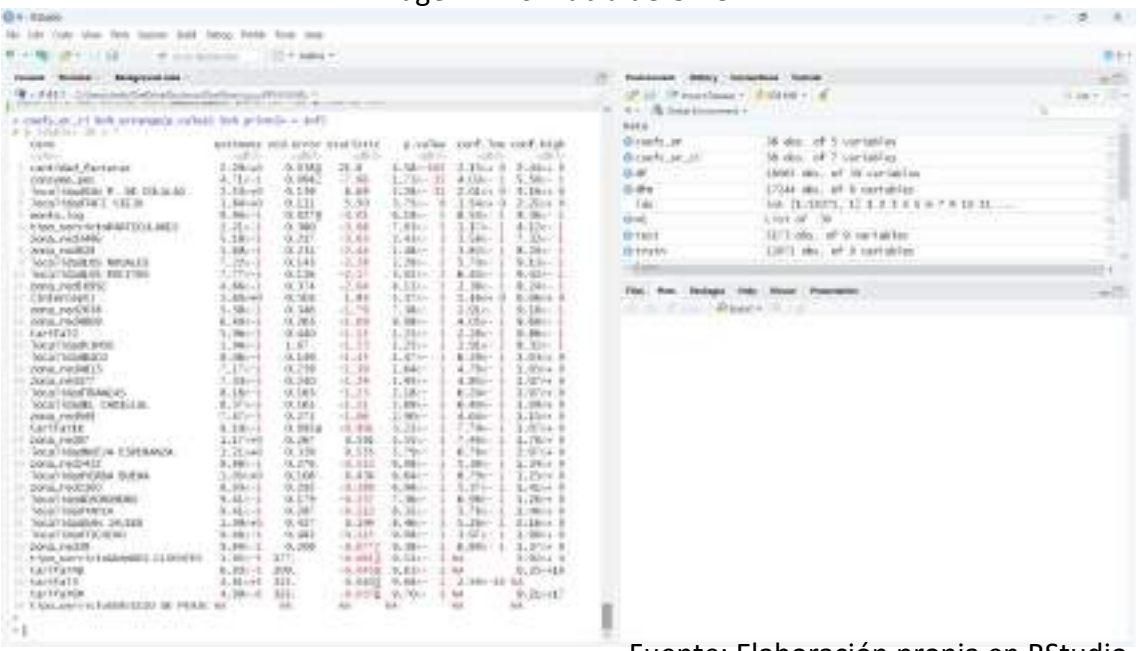
Fuente: Elaboración propia en RStudio.

Imagen N°19: Preparación de datos en R



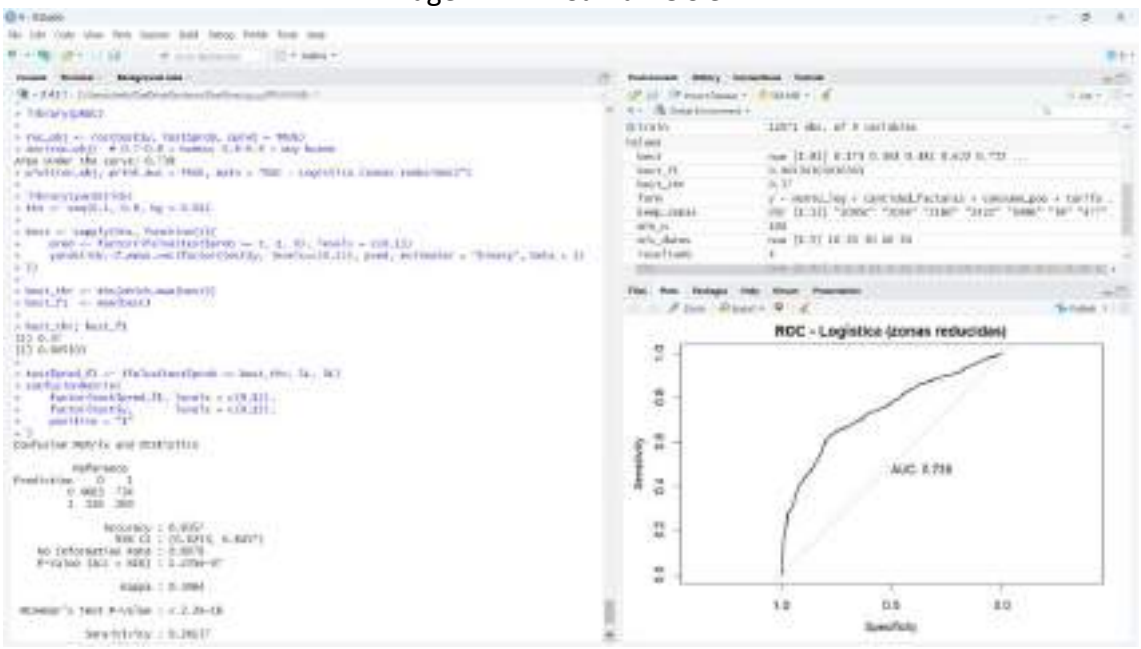
Fuente: Elaboración propia en RStudio.

Imagen N°20: Tabla de OR en R



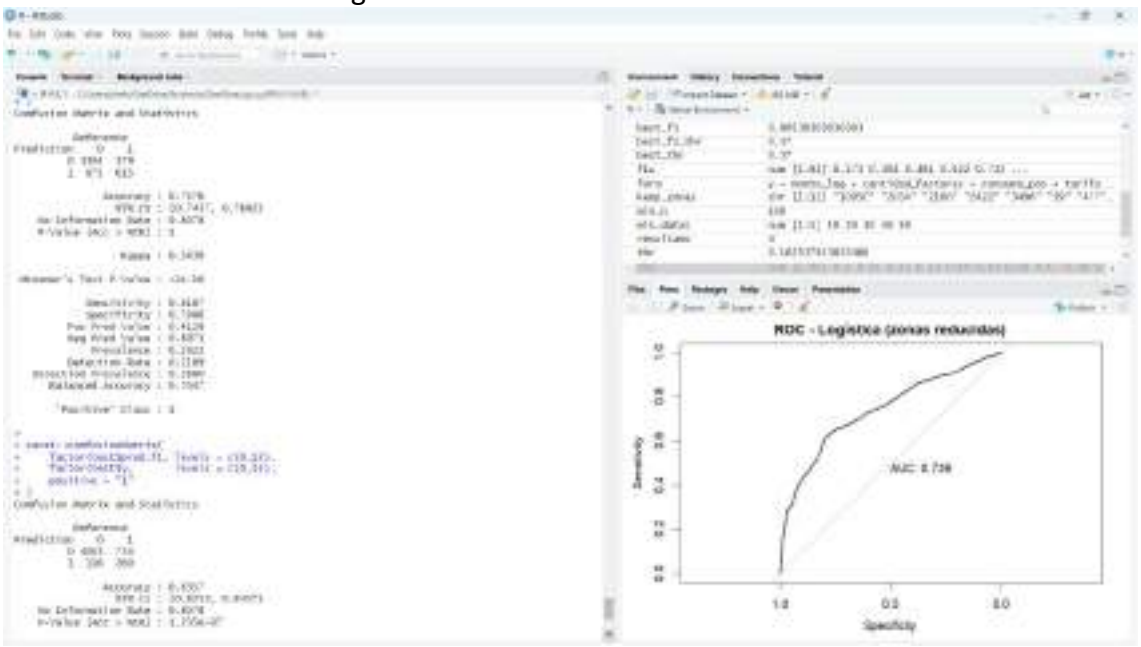
Fuente: Elaboración propia en RStudio.

Imagen N°21: Curva ROC en R



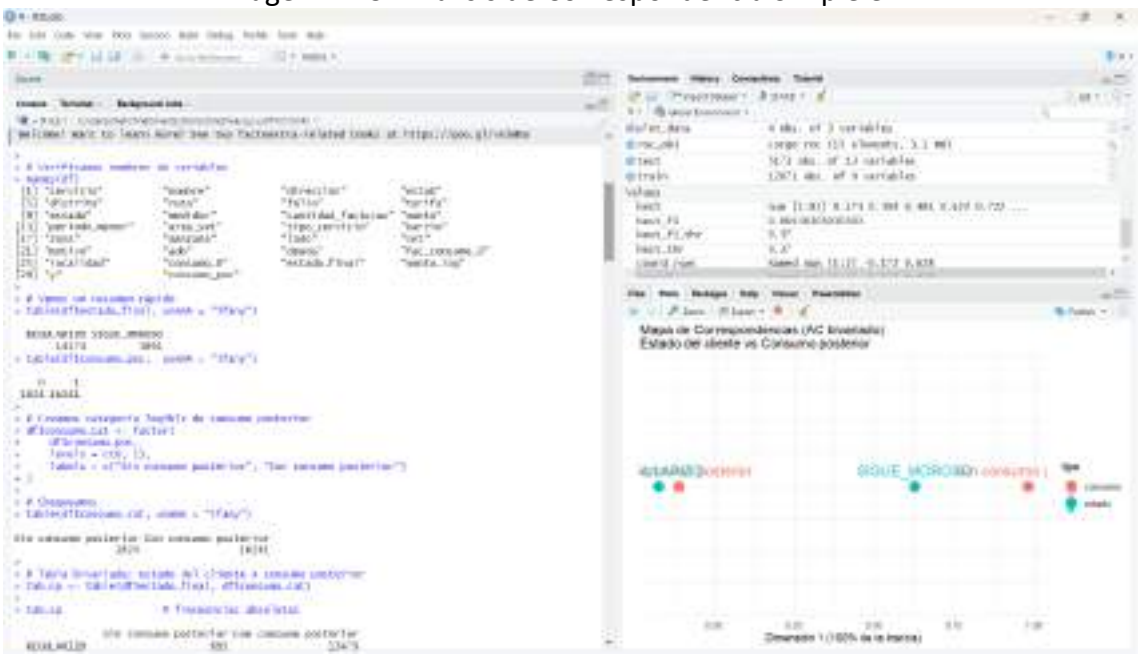
Fuente: Elaboración propia en RStudio.

Imagen N°22: Umbral de decisión F1 en R



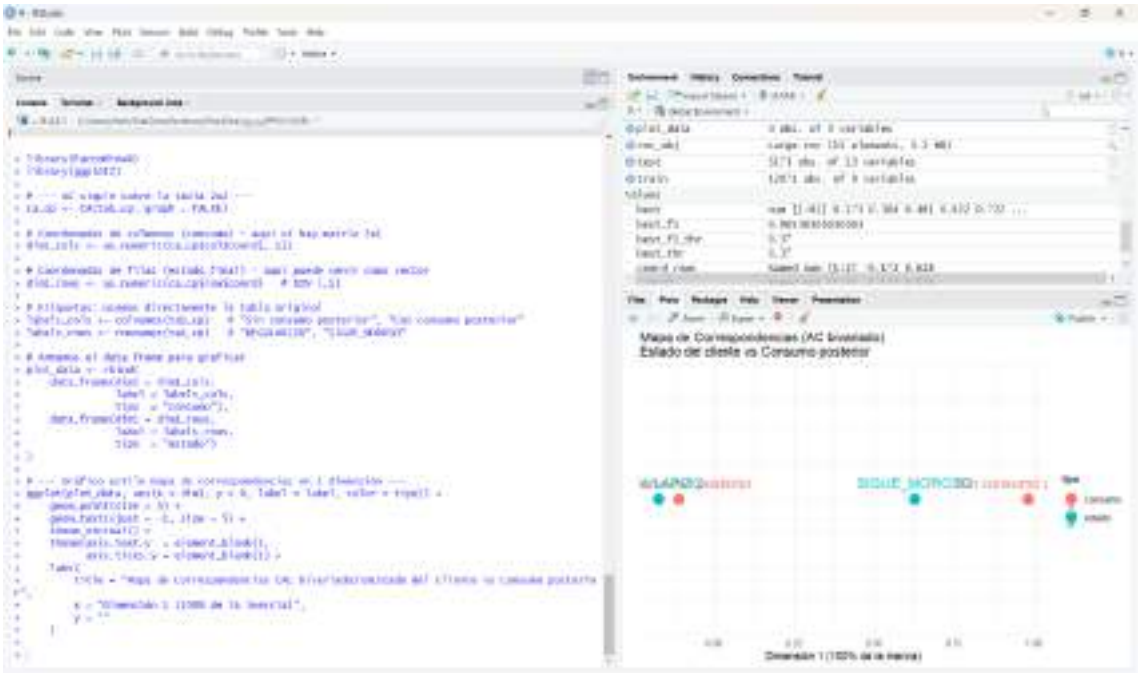
Fuente: Elaboración propia en RStudio.

Imagen N°23: Análisis de Correspondencia Simple en R



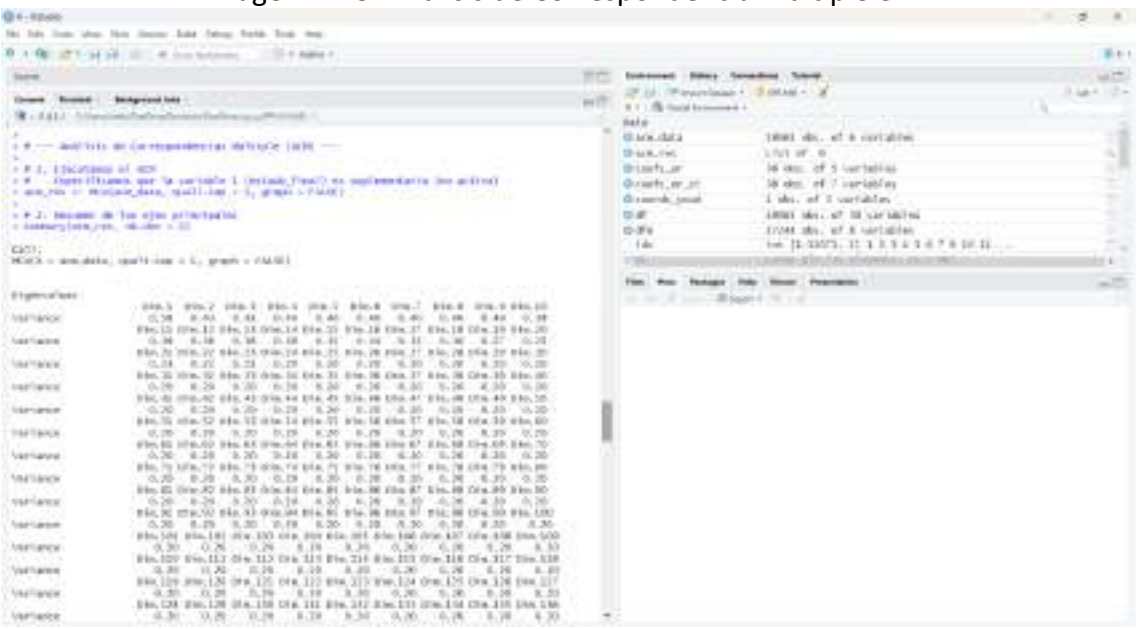
Fuente: Elaboración propia en RStudio.

Imagen N°24: Análisis de Correspondencia Simple en R



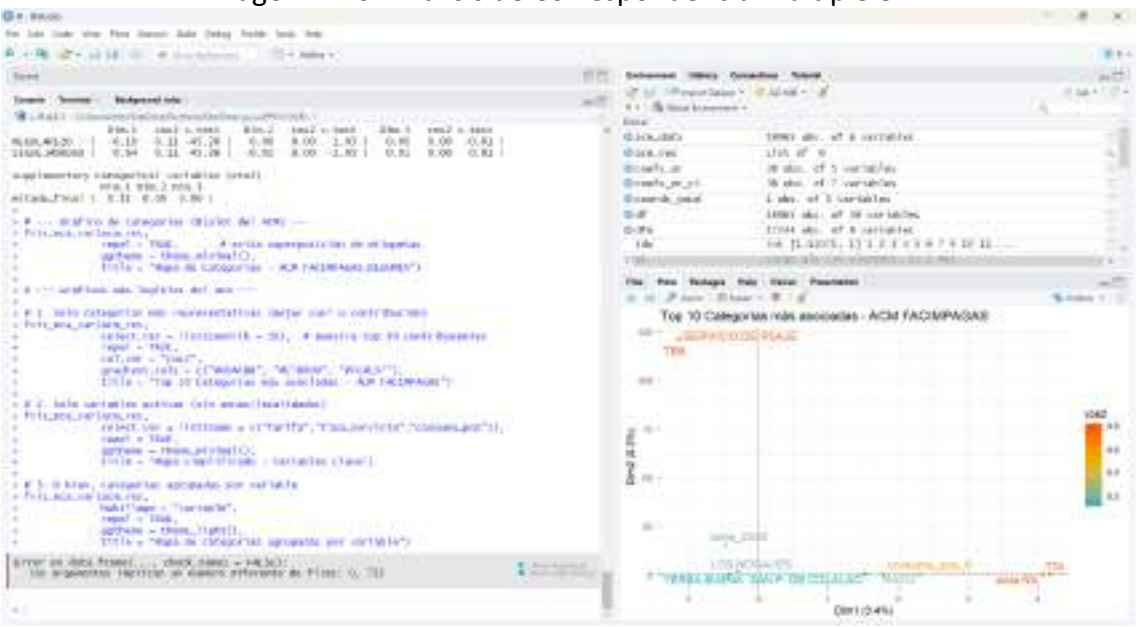
Fuente: Elaboración propia en RStudio.

Imagen N°25: Análisis de Correspondencia Múltiple en R



Fuente: Elaboración propia en RStudio.

Imagen N°26: Análisis de Correspondencia Múltiple en R



Fuente: Elaboración propia en RStudio.