UNIVERSIDAD PRIVADA FRANZ TAMAYO

FACULTAD DE INGENIERÍA INGENIERÍA DE SISTEMAS

Predicción de Churn y Análisis de Retención de Clientes de Internet en COTEL

CASO: COTEL R.L.

ESTUDIANTE: JUAN DAVID QUISPE INCA

MATERIA: BIG DATA

LA PAZ – BOLIVIA

I - 2025

1. Introducción

La industria de telecomunicaciones en Bolivia ha experimentado una transformación significativa en las últimas dos décadas, caracterizada por una creciente competencia entre operadoras y una demanda cada vez más exigente de servicios de conectividad de alta calidad. En este contexto, la Cooperativa de Telecomunicaciones La Paz Ltda. (COTEL), una de las principales proveedoras de servicios de Internet, telefonía y televisión por cable en el departamento de La Paz, enfrenta desafíos constantes relacionados con la retención de sus clientes.

El abandono de clientes, conocido en la industria como "churn", representa uno de los problemas más críticos para la sostenibilidad financiera de empresas de telecomunicaciones. Cada cliente que cancela su servicio no solo significa una pérdida de ingresos recurrentes, sino también el desperdicio de los recursos invertidos en su adquisición inicial y la potencial transferencia de valor hacia la competencia.

El análisis de datos masivos (Big Data) ofrece una oportunidad sin precedentes para abordar este desafío de manera proactiva. A través del procesamiento y análisis de grandes volúmenes de datos históricos sobre contratos, facturación, patrones de pago y comportamiento de clientes, es posible identificar señales tempranas de insatisfacción o riesgo de cancelación. Este proyecto busca aplicar técnicas avanzadas de ciencia de datos y aprendizaje automático para desarrollar un sistema predictivo que permita a COTEL anticiparse al churn y diseñar estrategias de retención personalizadas y efectivas.

2. Justificación

2.1 Relevancia Económica

Múltiples estudios en la industria de telecomunicaciones han demostrado que adquirir un nuevo cliente puede costar entre 5 y 7 veces más que retener uno existente. En el contexto de COTEL, donde la competencia incluye operadoras nacionales e internacionales, cada cliente perdido representa no solo la pérdida de ingresos mensuales recurrentes (aproximadamente Bs. 100-250 según el plan contratado), sino también:

- Costos de marketing y publicidad desperdiciados
- Inversión en infraestructura de instalación no recuperada
- Pérdida de valor de vida del cliente (Customer Lifetime Value - CLV)
- Riesgo de reputación negativa por recomendaciones adversas

Un modelo predictivo efectivo que identifique con 60-90 días de anticipación a clientes en riesgo puede reducir la tasa de churn en un 15-25%, lo que para una base de clientes de 50,000+ suscriptores de internet podría significar ahorros anuales de varios millones de bolivianos.

2.2 Relevancia Técnica y de Big Data

Este proyecto es particularmente adecuado para un enfoque de Big Data debido a:

- Volumen significativo de datos: Se trabajará con más de 30,000 registros de contratos históricos, incluyendo datos de facturación mensual que pueden generar cientos de miles de transacciones.
- Complejidad analítica: La predicción de churn requiere el análisis de múltiples variables interrelacionadas (comportamiento de pago, cambios de plan, suspensiones, deudas acumuladas) y la aplicación de algoritmos de machine learning supervisado.
- Procesamiento de series temporales: Los patrones de churn solo pueden identificarse mediante el análisis longitudinal de datos históricos.

- Generación de valor accionable: Los resultados del modelo pueden integrarse directamente en sistemas CRM para automatizar alertas y campañas de retención.

2.3 Relevancia Institucional y Social

Para COTEL como cooperativa, la retención de socios-clientes va más allá del aspecto económico:

- Fortalecimiento del modelo cooperativo: Retener socios contribuye a la estabilidad y crecimiento sostenible de la organización.
- Mejora de la calidad del servicio: Identificar causas de insatisfacción permite implementar mejoras operativas.
- Responsabilidad social: Mantener empleos y servicios de conectividad estables contribuye al desarrollo económico regional.

2.4 Relación con las 5V del Big Data

Este proyecto aborda de manera integral las cinco características fundamentales del Big Data:

Volumen: El proyecto trabajará con más de 30,000 registros de contratos de internet, junto con sus respectivas facturas mensuales históricas (estimadas en 200,000+ transacciones), datos de clientes y consultas de deudas, superando ampliamente el umbral de Big Data para análisis cooperativo.

Velocidad: Los datos de COTEL se actualizan mensualmente con cada ciclo de facturación. El modelo predictivo está diseñado para procesar estos datos de forma periódica (mensual o quincenal) y generar alertas en tiempo casi real sobre clientes en riesgo.

Variedad: El proyecto integra múltiples tipos de datos:

- Datos estructurados transaccionales (facturas, pagos, deudas)
- Datos maestros (información de clientes, planes comerciales)
- Datos categóricos (estados de contrato, formas de pago)
- Datos temporales (fechas de contrato, rescisión, instalación)

Veracidad: Los datos provienen directamente de los sistemas operacionales de COTEL, garantizando su confiabilidad. Sin embargo, el proyecto incluirá procesos de limpieza y validación para manejar inconsistencias, valores nulos y registros duplicados comunes en sistemas legacy.

Valor: El resultado final del proyecto genera valor tangible y medible:

- Reducción de la tasa de churn entre 15-25%
- ROI positivo en campañas de retención (menor inversión por cliente retenido)
- Mejora en la segmentación de clientes para estrategias comerciales
- Base de conocimiento para decisiones estratégicas de la gerencia

3. Planteamiento del Problema

3.1 Descripción del Problema

Actualmente, la Cooperativa de Telecomunicaciones La Paz (COTEL) no cuenta con un sistema predictivo que identifique con anticipación qué clientes de internet tienen alta probabilidad de cancelar su servicio. Las acciones de retención se llevan a cabo de manera reactiva, es decir, después de que el cliente ya ha manifestado su intención de cancelar o directamente ha procedido con la rescisión del contrato.

Esta situación genera las siguientes consecuencias negativas:

1. Pérdida inevitable de clientes: Cuando un cliente expresa su deseo de cancelar, generalmente ya ha tomado la decisión y evaluado alternativas con la competencia, reduciendo significativamente las posibilidades de retención.
2. Ineficiencia en recursos de retención: Las campañas de retención masivas (sin segmentación) resultan en:
    - Inversión desperdiciada en clientes con bajo riesgo de churn
    - Falta de recursos para atender adecuadamente a clientes de alto riesgo
    - Ofertas genéricas que no abordan las causas específicas de insatisfacción
3. Desconocimiento de patrones de churn: La empresa carece de información estructurada sobre:

    - Qué comportamientos preceden a la cancelación
    - Qué perfiles de clientes son más propensos al churn
    - Qué factores (plan, morosidad, suspensiones) tienen mayor impacto
4. Pérdida de oportunidades de mejora operativa: Sin análisis sistemático, no se identifican problemas recurrentes en la calidad del servicio, atención al cliente o pricing que podrían estar causando insatisfacción generalizada.

3.2 Preguntas de Investigación

Este proyecto busca responder las siguientes interrogantes:

1. ¿Qué variables tienen mayor poder predictivo sobre el churn?
   ¿Es la morosidad el factor principal? ¿Los cambios de plan indican insatisfacción? ¿La antigüedad del cliente es un factor protector?
2. ¿Existen patrones temporales identificables?
   ¿Cuántos meses antes de la cancelación comienzan a manifestarse señales de riesgo? ¿Hay estacionalidad en el churn?
3. ¿Se pueden segmentar los clientes en riesgo por causa probable de churn?
   Clientes en riesgo por: precio (buscan planes más baratos), calidad del servicio, morosidad crónica, competencia.
4. ¿Qué perfil de clientes tiene mayor tasa de churn?
   ¿Clientes nuevos vs antiguos? ¿Planes básicos vs premium? ¿Modalidad de pago prepago vs postpago?
5. ¿Es posible construir un modelo con precisión suficiente para ser operacionalmente útil?
   ¿Se puede alcanzar un recall ≥ 70% para identificar correctamente a clientes en riesgo sin generar demasiadas falsas alarmas?

3.3 Decisiones que se Beneficiarían de los Resultados

El modelo predictivo y los insights generados permitirán a COTEL tomar decisiones informadas en múltiples áreas:

Área Comercial:

- Diseñar campañas de retención proactivas dirigidas a segmentos específicos

- Crear ofertas personalizadas según la causa probable de churn
- Priorizar esfuerzos de retención en clientes de alto valor (high CLV)

Área de Atención al Cliente:

- Alertar a agentes cuando un cliente en riesgo contacta al call center
- Priorizar tickets de soporte de clientes con alto riesgo de churn
- Implementar programas de seguimiento post-instalación para clientes nuevos

Gerencia Estratégica:

- Identificar productos o zonas geográficas con mayor churn
- Evaluar el impacto de cambios de precio o políticas en la retención
- Proyectar ingresos con mayor precisión considerando churn esperado

Área Técnica:

- Detectar zonas con problemas de conectividad que generen insatisfacción
- Priorizar mantenimiento preventivo en áreas críticas

4. Objetivos de Investigación

4.1 Objetivo General

Desarrollar un modelo predictivo de churn para clientes de servicio de internet de COTEL que permita identificar con 40 a 60 días de anticipación a clientes con alta probabilidad de cancelación del servicio, facilitando la implementación de estrategias de retención proactivas y segmentadas que reduzcan la tasa de abandono y maximicen el retorno de inversión en acciones comerciales.

4.2 Objetivos Específicos

1. Construir una base de datos analítica integrada
   Integrar información de contratos de internet, facturación, historial de pagos y datos de clientes en una estructura unificada que permita análisis longitudinal y la creación de variables derivadas relevantes para la predicción de churn.
2. Identificar variables predictoras mediante análisis exploratorio de datos (EDA)
   Realizar un análisis exploratorio exhaustivo para:
   - Caracterizar la distribución de variables clave (planes, deudas, antigüedad)
   - Identificar patrones de comportamiento diferencial entre clientes que cancelaron vs. clientes activos
   - Detectar correlaciones y multicolinealidad entre variables
3. Crear dashboards interactivos para visualización y monitoreo
   Desarrollar visualizaciones de datos que incluyan:
   - Dashboard ejecutivo (tasa mensual, proyecciones)
   - Panel de monitoreo de clientes en riesgo

5. Fuentes de Datos

5.1 Descripción de Bases de Datos Disponibles

El proyecto utilizará datos operacionales reales de la Cooperativa de Telecomunicaciones La Paz (COTEL), específicamente del servicio de Internet. Las fuentes principales son:

5.1.1 Tabla de Contratos de Internet

Origen: Sistema de gestión de contratos de COTEL
Formato: CSV exportado desde base de datos relacional (PostgreSQL/Oracle)
Volumen estimado: 50,000+ contratos (históricos y actuales)
Periodo temporal: 2006 - 2025 (19 años de datos)

Variables clave:

- contrato: ID único del contrato
- cod_cliente: ID del cliente (puede tener múltiples contratos)
- plan_comercial: Plan contratado (CTL 5MB, CTL 7MB, etc.)
- forma_pago: POST PAGO / PRE PAGO
- cod_estado_contrato: Estado actual (1=ACTIVO, 2=INACTIVO, 3=SUSPENDIDO, etc.)
- anulado: Descripción del estado (EN SERVICIO, RETIRADO, INACTIVO(RESCINDIDO))
- f_contrato: Fecha de inicio del contrato
- f_rescision: Fecha de cancelación (si aplica)
- f_instalacion: Fecha de instalación del servicio
- direccion: Ubicación del servicio
- cod_servicio: Tipo de servicio (3 = Internet)

5.1.2 Tabla de Clientes

Origen: Base de datos de clientes COTEL
Formato: CSV
Volumen estimado: 30,000+ registros únicos

Variables clave:

- cod_cliente: ID único del cliente
- nombres, ape_paterno, ape_materno: Datos demográficos
- nro_documento: CI del cliente
- direccion: Dirección del cliente
- telefono_ref: Teléfono de contacto
- f_nacimiento: Fecha de nacimiento (para calcular edad)
- sexo: Género del cliente

5.1.3 Tabla de Facturación

Origen: Sistema de facturación mensual COTEL
Formato: CSV
Volumen estimado: 200,000+ facturas (histórico de varios años)
Frecuencia: Datos mensuales

Variables clave:

- contrato: ID del contrato facturado
- cod_cliente: ID del cliente
- periodo: Periodo de facturación (YYYYMM)
- fecha_emision: Fecha de emisión de factura
- monto_total: Monto total facturado
- monto_cotel: Monto correspondiente al servicio
- estado: Estado de la factura (P=Pagado, G=Generado, NP=No Pagado)
- telefono: Línea asociada (si aplica)

6. Marco conceptual

6.1 Definición y Contexto del Churn en Telecomunicaciones

El término churn (abandono o deserción de clientes) se refiere al fenómeno mediante el cual un cliente activo de una empresa de servicios decide cancelar o no renovar su contrato, resultando en la pérdida de ese cliente y sus ingresos asociados (Verbeke et al., 2012). En la industria de telecomunicaciones, el churn representa uno de los desafíos más críticos debido a la alta competitividad del mercado y los elevados costos de adquisición de nuevos clientes.

La literatura académica identifica dos tipos principales de churn:

Churn voluntario: El cliente toma la decisión activa de cancelar el servicio, generalmente motivado por insatisfacción con el precio, calidad del servicio, o atraído por ofertas de la competencia (Ahmad et al., 2019).

Churn involuntario: La cancelación ocurre por circunstancias ajenas a la voluntad del cliente, como mudanza a zonas sin cobertura, fallecimiento, o imposibilidad de pago prolongada (Hadden et al., 2007).

Este proyecto se enfoca principalmente en el churn voluntario, ya que es el que puede prevenirse mediante estrategias de retención proactivas basadas en predicción de riesgo.

6.2 Customer Lifetime Value (CLV) y el Costo del Churn

El concepto de Customer Lifetime Value (CLV) o Valor de Vida del Cliente es fundamental para comprender el impacto económico del churn. El CLV representa el valor presente neto de todos los ingresos futuros esperados que un cliente generará durante su relación con la empresa (Gupta & Lehmann, 2003).

Para un proveedor de servicios de internet como COTEL, el CLV promedio de un cliente puede calcularse como:

Donde:

- ARPU (Average Revenue Per User): Ingreso promedio mensual por cliente (Bs. 100-250)
- Margen de Ganancia: Porcentaje de beneficio neto (~40-50% en telecomunicaciones)
- Vida Esperada: Años promedio que un cliente permanece activo (3-7 años típicamente)
- Costo de Adquisición: Inversión en marketing, ventas e instalación (Bs. 300-500)

Este cálculo evidencia por qué la retención de clientes es económicamente más rentable que la adquisición constante de nuevos clientes, ya que el costo de retención (campañas, descuentos, mejoras de servicio) raramente supera el 10-20% del CLV, mientras que reemplazar un cliente perdido requiere recuperar todo el costo de adquisición nuevamente.

6.3 Enfoques Metodológicos para Predicción de Churn

La investigación académica en predicción de churn ha evolucionado significativamente en las últimas dos décadas, transitando desde modelos estadísticos tradicionales hacia técnicas avanzadas de machine learning y big data analytics.

Enfoques clásicos (1990-2010):

- Regresión Logística: Modelo interpretable que estima la probabilidad de churn como función lineal de variables predictoras (Mozer et al., 2000)
- Análisis de Supervivencia: Modelos de Cox que estiman el "tiempo hasta el churn" considerando censura de datos (Lu, 2002)
- Árboles de Decisión: Algoritmos como CART o C4.5 que generan reglas interpretables para clasificación (Verbeke et al., 2012)

Enfoques modernos de Machine Learning (2010-presente):

- Random Forest: Ensamble de árboles de decisión que reduce sobreajuste y mejora precisión (Breiman, 2001)
- Gradient Boosting (XGBoost, LightGBM): Modelos secuenciales de alta precisión, líderes en competencias de ciencia de datos (Chen & Guestrin, 2016)
- Redes Neuronales Profundas: Arquitecturas DNN capaces de capturar relaciones no lineales complejas (Óskarsdóttir et al., 2019)
- Modelos de Series Temporales: LSTM y GRU para capturar patrones temporales en el comportamiento del cliente (Huang et al., 2020)

Enfoques de Big Data y Analytics Avanzados:

- Network Analytics: Análisis de redes sociales de clientes para capturar efectos de contagio del churn (Verbeke et al., 2014)
- Análisis de Texto: Procesamiento de lenguaje natural (NLP) sobre interacciones de atención al cliente para detectar insatisfacción (Coussement & Van den Poel, 2008)

- Sistemas de Recomendación: Sugerir planes o servicios complementarios para aumentar engagement y reducir churn

6.4 Variables Predictoras Clave Identificadas en la Literatura

Diversos estudios empíricos han identificado las siguientes categorías de variables como predictoras significativas del churn en telecomunicaciones:

Variables demográficas:

- Edad del cliente (clientes jóvenes tienen mayor movilidad)
- Tipo de cliente (residencial vs. empresarial)
- Ubicación geográfica (zonas urbanas vs. rurales)

Variables de comportamiento de uso:

- Promedio mensual de consumo/facturación (ARPU)
- Variabilidad en el consumo (alta variabilidad puede indicar uso irregular)
- Tendencia de uso (creciente, estable o decreciente)
- Utilización de servicios adicionales

Variables de comportamiento de pago:

- Historial de morosidad (número de facturas impagadas)
- Puntualidad en pagos (días promedio de atraso)
- Métodos de pago utilizados

Variables de servicio y satisfacción:

- Número de reclamos o tickets de soporte
- Tiempo de resolución de problemas técnicos
- Cambios de plan (upgrades vs. downgrades)
- Suspensiones temporales del servicio

Variables de relación con la empresa:

- Antigüedad del cliente (tenure)
- Número de servicios contratados (cross-selling)

- Participación en programas de lealtad
- Interacciones con atención al cliente

Variables de competencia y mercado:

- Existencia de ofertas competitivas en la zona
- Calidad de cobertura de competidores
- Eventos de marketing de la competencia

6.5 Vinculación con los Objetivos del Proyecto y Big Data

Este marco conceptual establece la base teórica para el proyecto de predicción de churn en COTEL, que se enmarca dentro del paradigma de Big Data Analytics por las siguientes razones:

1. Volumen de datos: El análisis requiere procesar 50,000+ contratos históricos con 200,000+ transacciones de facturación, superando las capacidades de análisis tradicional.
2. Variedad de fuentes: La integración de datos transaccionales (facturación), maestros (clientes, contratos), y potencialmente no estructurados (observaciones, reclamos) requiere técnicas de Big Data.
3. Velocidad de procesamiento: El modelo debe actualizar scoring de riesgo mensualmente para toda la base activa, requiriendo procesamiento eficiente de grandes volúmenes.
4. Generación de valor: La predicción precisa de churn permite estrategias de retención que pueden reducir pérdidas en millones de bolivianos anuales, justificando la inversión en infraestructura de Big Data.

El proyecto adoptará un enfoque predictivo utilizando algoritmos de machine learning supervisado, específicamente modelos de clasificación binaria que predicen la probabilidad de churn en una ventana temporal de 60-90 días. Esta elección metodológica se sustenta en:

- Disponibilidad de datos históricos etiquetados (clientes que cancelaron vs. activos)

- Necesidad de predicciones accionables (probabilidades de riesgo, no solo clasificaciones)
- Requerimiento de interpretabilidad para el equipo de negocio
- Escalabilidad para scoring de toda la base de clientes mensualmente

7. Definición de Big Data

7.1 Concepto y Evolución del Término

El término Big Data fue acuñado inicialmente en la década de 1990, pero su adopción masiva y definición formal ocurrió en la década de 2010 con la explosión de datos digitales generados por internet, redes sociales, dispositivos móviles y sensores IoT.

Según Gartner (2012), una de las definiciones más ampliamente citadas, Big Data se refiere a:

> "Activos de información de alto volumen, alta velocidad y/o alta variedad que demandan formas innovadoras y rentables de procesamiento de información que permitan una mejor comprensión, toma de decisiones y automatización de procesos."

McAfee y Brynjolfsson (2012), en su influyente artículo en Harvard Business Review "Big Data: The Management Revolution", definen Big Data como:

> "Conjuntos de datos cuyo tamaño está más allá de la capacidad de las herramientas de software de bases de datos típicas para capturar, almacenar, gestionar y analizar."

IBM (2013) popularizó el modelo de las "3V" como características definitorias de Big Data:

- Volumen: Escala de datos (terabytes, petabytes)
- Velocidad: Análisis en tiempo real o casi real
- Variedad: Datos estructurados, semi-estructurados y no estructurados

Posteriormente, este modelo se expandió a las "5V" con la inclusión de:

- Veracidad: Calidad, precisión y confiabilidad de los datos
- Valor: Utilidad y ROI del análisis de datos

7.2 Big Data en el Contexto de este Proyecto

En el contexto específico de la predicción de churn en COTEL, Big Data no se refiere únicamente al tamaño absoluto de los datos (que, aunque significativo con 50,000+ contratos, no alcanza los petabytes de empresas tecnológicas globales), sino a la complejidad del procesamiento y análisis requerido:

1. Integración de múltiples fuentes heterogéneas: Contratos, facturación, clientes, deudas, cada una con esquemas y formatos diferentes, requieren pipelines ETL (Extract, Transform, Load) sofisticados.
2. Análisis temporal complejo: Identificar patrones de churn requiere análisis longitudinal de series temporales, cálculo de features derivadas (tendencias, variabilidades, cambios de comportamiento), y agregaciones complejas.
3. Modelado predictivo a escala: Entrenar modelos de machine learning con técnicas de validación cruzada, optimización de hiperparámetros, y scoring mensual de 40,000+ clientes activos demanda capacidades computacionales más allá de hojas de cálculo.
4. Decisiones en tiempo operativo: Aunque no es tiempo real (segundos), el modelo debe generar alertas de riesgo en ventanas de días/semanas para permitir intervenciones oportunas, lo que requiere automatización y orquestación.

Mayer-Schönberger y Cukier (2013) en su libro "Big Data: A Revolution That Will Transform How We Live, Work, and Think" argumentan que el verdadero poder de Big Data no está en el tamaño, sino en la capacidad de analizar todos los datos disponibles en lugar de muestras, y de descubrir correlaciones que permitan predicciones accionables sin necesariamente entender causalidad.

Este proyecto ejemplifica esta filosofía: en lugar de realizar encuestas a una muestra de clientes sobre su satisfacción (enfoque tradicional), se analizan todos los

datos transaccionales históricos para identificar patrones que estadísticamente predicen churn, permitiendo intervenciones proactivas.

7.3 Diferenciación de Análisis Tradicional vs. Big Data

Aspecto	Análisis Tradicional	Big Data Analytics (Este Proyecto)
Volumen	Muestras representativas (cientos a miles de registros)	Población completa (50,000+ contratos, 200,000+ transacciones)
Herramientas	Excel, Access, software estadístico básico	Python/R, bases de datos relacionales, Jupyter Notebooks, librerías de ML
Metodología	Hipótesis → Prueba estadística	Exploración → Descubrimiento de patrones → Predicción
Tiempo de análisis	Días/semanas de procesamiento manual	Minutos/horas con scripts automatizados
Actualización	Análisis puntuales, informes estáticos	Pipelines automatizados, dashboards dinámicos
Escalabilidad	Limitada (re-análisis manual)	Alta (agregar nuevos datos y re-entrenar modelos)

Este proyecto se posiciona en el paradigma de Big Data Analytics al abordar los desafíos de volumen, variedad y velocidad mediante herramientas modernas de

ciencia de datos, generando valor tangible para COTEL a través de predicciones accionables que no serían posibles con enfoques tradicionales.

8. Relación con las 5V del Big Data

8.1 Volumen

Definición: Volumen se refiere a la magnitud o escala de datos que deben ser almacenados, procesados y analizados. En el paradigma de Big Data, el volumen típicamente excede las capacidades de herramientas tradicionales de análisis.

Aplicación en este proyecto:

Escala de datos del proyecto:

- Contratos de Internet: 50,000+ registros históricos
- Facturas mensuales: 200,000+ transacciones de facturación
- Clientes únicos: 30,000+ registros con datos demográficos
- Consultas de deudas: Snapshot mensual de estado crediticio de cada cliente
- Total estimado después de integraciones: ~350,000 registros

Escalabilidad esperada: El dataset crece de forma continua con la operación de COTEL:

- Crecimiento mensual: ~2,000 nuevas facturas + ~150 nuevos contratos
- Proyección anual: +24,000 facturas adicionales + ~1,800 contratos
- En 5 años: El dataset superará los 500,000 registros, requiriendo potencialmente migración a bases de datos distribuidas (Spark, Hadoop) o data warehouses (Snowflake, BigQuery)

Justificación como Big Data: Aunque 350,000 registros pueden parecer manejables en bases de datos relacionales modernas, la complejidad surge al considerar:

1. Procesamiento de features derivadas: Cada contrato requiere cálculos de agregación sobre su historial completo de facturas (ARPU, variabilidad, tendencias), multiplicando exponencialmente las operaciones computacionales

2. Validación cruzada y entrenamiento de modelos: Algoritmos de ensemble learning (Random Forest, XGBoost) requieren entrenamiento iterativo sobre múltiples subconjuntos de datos
3. Scoring mensual: Predecir riesgo de churn para 40,000+ clientes activos cada mes requiere procesamiento batch eficiente

Métricas concretas:

- Tamaño en disco: ~500 MB de datos brutos en CSV, ~2 GB después de feature engineering
- Memoria RAM requerida: 8-16 GB para procesamiento completo del dataset en memoria (Pandas)
- Tiempo de procesamiento: ~30-45 minutos para pipeline ETL completo + entrenamiento de modelo en CPU estándar

8.2 Velocidad

Definición: Velocidad se refiere a la frecuencia con la que los datos son generados, capturados y procesados, así como la rapidez con la que deben analizarse para generar valor (Laney, 2001).

Aplicación en este proyecto:

Frecuencia de actualización de datos:

- Facturación: Generación mensual en ciclo fijo (ej: día 28 de cada mes)
- Pagos: Actualizaciones diarias conforme clientes realizan pagos
- Estados de contrato: Cambios en tiempo real (altas, bajas, suspensiones) en sistema operacional
- Deudas: Actualización continua calculada dinámicamente

Velocidad de procesamiento requerida:

Este proyecto NO requiere procesamiento en tiempo real (como sistemas de detección de fraude bancario o recomendaciones de e-commerce), sino procesamiento batch periódico:

Justificación de velocidad batch: La decisión de cancelar un servicio de internet raramente es instantánea; suele ser el resultado de insatisfacción acumulada durante semanas/meses. Por tanto, identificar clientes en riesgo con 60-90 días de anticipación mediante procesamiento mensual es completamente adecuado para permitir estrategias de retención efectivas (diseño de ofertas, contacto proactivo, mejoras técnicas).

8.3 Variedad

Definición: Variedad se refiere a la diversidad de tipos y formatos de datos que deben integrarse para el análisis. En Big Data, esto incluye datos estructurados (tablas), semi-estructurados (JSON, XML) y no estructurados (texto, imágenes, audio) (Russom, 2011).

Aplicación en este proyecto:

Tipos de datos integrados:

1. Datos estructurados (mayoría del proyecto):

- Tablas relacionales: Formato tabular clásico (filas y columnas) con esquemas definidos
    - Contratos: 15+ columnas con tipos de datos específicos (int, varchar, date)
    - Facturas: 10+ columnas con transacciones mensuales
    - Clientes: 20+ columnas con información demográfica

2. Datos categóricos:

- Variables nominales: plan_comercial, forma_pago, cod_estado_contrato
- Variables ordinales: nivel de riesgo crediticio (bajo, medio, alto)
- Requerimiento: One-Hot Encoding o Label Encoding para modelado

3. Datos temporales:

- Fechas y timestamps: f_contrato, f_rescision, f_instalacion, fecha_emision
- Series temporales: Secuencias de facturas mensuales por cliente

- Requerimiento: Conversión a features numéricos (antigüedad en meses, días desde último evento, tendencias)

4. Datos semi-estructurados (campo observaciones):

- Formato: Texto libre sin esquema predefinido
- Ejemplos:
    - "LINEA RETIRADA P/MORA OT. 03141950"
    - "PUNTO COTEL"
    - "PERSONA TRAMITE ANTONIO VELASQUEZ 248-1685"
- Desafío: Requiere parsing, extracción de entidades (nombres, números de orden, motivos), y potencialmente NLP para categorización
- Aplicación actual: Este proyecto inicialmente NO procesará este campo por limitaciones de tiempo, pero se identifica como oportunidad de mejora futura mediante técnicas de text mining

5. Datos con formatos inconsistentes (direcciones):

- Variabilidad:
    - "C RIO GUAPORE Nro.2155"
    - "AV ALFONSO UGARTE Nro.126 PISO:1 Dpto.1"
    - "MZNO N Nro.1634 MZNO N"
- Desafío: Falta de estandarización (abreviaciones, errores tipográficos, formatos variables)
- Solución: Normalización mediante regex y extracción de componentes clave (tipo de vía, zona, número)

Formatos de almacenamiento:

- Fuente original: Bases de datos relacionales (PostgreSQL/Oracle) de COTEL
- Formato de trabajo: Archivos CSV para portabilidad
- Formato optimizado: Parquet para procesamiento eficiente (compresión columnar)
- Formato de salida: JSON para integración con dashboards web

8.4 Veracidad

Definición: Veracidad se refiere a la calidad, precisión, confiabilidad y consistencia de los datos. En Big Data, donde el volumen es inmenso, garantizar la veracidad es crítico para evitar el principio "garbage in, garbage out" (Fisher et al., 2012).

Aplicación en este proyecto:

Evaluación de calidad de datos:

Fortalezas (alta veracidad):

1. Datos transaccionales críticos:
    - Facturación auditada por requerimientos fiscales (SIN, impuestos)
    - Pagos con trazabilidad bancaria completa
    - Estados de contrato registrados automáticamente por workflows del sistema
    - Nivel de confianza: 95%+
2. Datos de identificación:
    - cod_cliente y contrato son claves primarias únicas
    - nro_documento (CI) verificado al momento de contratación
    - Nivel de confianza: 98%+

Debilidades (veracidad cuestionable):

1. Datos demográficos opcionales:
    - f_nacimiento: ~30% valores nulos, ~5% valores anómalos (edad < 18 o > 100 años)
    - telefono_ref: ~40% faltante
    - email: ~70% faltante (no obligatorio en contratos antiguos)
    - Nivel de confianza: 60-70%
2. Datos descriptivos:
    - direccion: Errores tipográficos, abreviaciones no estandarizadas
    - observaciones: Campo libre sin validación
    - Nivel de confianza: 50-60%
3. Datos históricos migrados:

    - Contratos pre-2010 con campos incompletos por migraciones de sistemas legacy
    - Posibles duplicados de clientes con IDs diferentes
    - Nivel de confianza: 70-80%

Impacto en el modelo:

- Variables con baja veracidad se excluirán o usarán con cautela
- Se priorizarán features derivados de datos transaccionales (alta veracidad)
- Se realizará análisis de sensibilidad: ¿cómo cambia el modelo si excluimos variables dudosas?

8.5 Valor

Definición: Valor se refiere a la utilidad, relevancia y ROI (Return on Investment) que se obtiene del análisis de Big Data. Es la V más importante, ya que justifica toda la inversión en infraestructura y recursos (Marr, 2015).

Aplicación en este proyecto:

Resultados esperados:

1. Valor económico directo cuantificable:

Escenario base (sin modelo):

- Base de clientes de internet: 40,000 activos
- Tasa de churn anual: 15% (típica en ISPs)
- Clientes perdidos/año: 6,000
- ARPU promedio: Bs. 150/mes
- Pérdida anual de ingresos: 6,000 × Bs. 150 × 12 = Bs. 10,800,000

Escenario con modelo predictivo (reducción de churn del 20%):

- Tasa de churn reducida a: 12%
- Clientes perdidos/año: 4,800
- Clientes retenidos adicionales: 1,200
- Ingresos protegidos: 1,200 × Bs. 150 × 12 = Bs. 2,160,000/año

Inversión requerida:

- Desarrollo del proyecto (tiempo de data scientist): Bs. 80,000
- Infraestructura y herramientas (servidores, licencias): Bs. 20,000
- Campañas de retención (descuentos, mejoras técnicas): Bs. 300,000/año
- Total inversión primer año: Bs. 400,000

Recuperación de inversión (payback period): ~2.2 meses

2. Valor económico indirecto:

Optimización de recursos comerciales:

- Sin modelo: Contacto aleatorio a 8,000 clientes/año → Tasa de éxito 12% → 960 retenidos
- Con modelo: Contacto dirigido a 3,000 clientes de alto riesgo → Tasa de éxito 35% → 1,050 retenidos
- Ahorro en costos operativos: Bs. 100,000/año (menos llamadas, mejor uso del tiempo del equipo)

Reducción de costos de adquisición:

- Costo de adquirir nuevo cliente: Bs. 400 (marketing, instalación)
- Clientes retenidos que no deben reemplazarse: 1,200
- Ahorro indirecto: Bs. 480,000/año

3. Valor estratégico (decisiones mejor informadas):

A nivel operacional:

- Priorización de recursos técnicos en zonas de alto riesgo
- Alertas automáticas al call center cuando cliente en riesgo contacta
- Asignación dinámica de presupuesto de retención

A nivel táctico:

- Identificación de planes problemáticos → rediseño de productos
- Detección de zonas geográficas con churn elevado → mejoras de infraestructura

- Segmentación de clientes para campañas personalizadas

A nivel estratégico:

- Proyecciones financieras más precisas (incorporando churn esperado)
- Evaluación de impacto de decisiones de pricing antes de implementarlas
- Benchmarking: comparar churn de COTEL vs. competencia

4. Valor de conocimiento (insights accionables):

El modelo no solo predice *quién* se irá, sino *por qué*:

Feature importance revelará:

- ¿Es el precio el problema principal? → Ajustar tarifas o crear planes más económicos
- ¿Es la calidad del servicio? → Priorizar inversión en infraestructura técnica
- ¿Es falta de uso? → Campañas educativas sobre beneficios del servicio
- ¿Es la competencia? → Análisis competitivo y mejora de propuesta de valor

Segmentación de clientes:

- Identificar perfiles de alto valor (high CLV) para proteger prioritariamente
- Detectar clientes "recuperables" vs. "perdidos definitivamente"
- Crear arquetipos de clientes en riesgo para personalización

5. Valor intangible (transformación organizacional):

Cultura data-driven:

- Capacitación del equipo de COTEL en análisis de datos
- Establecimiento de KPIs basados en evidencia (no intuición)
- Democratización de datos mediante dashboards accesibles

Replicabilidad:

- La metodología puede extenderse a telefonía y TV Cable
- El pipeline de datos puede reutilizarse para otros proyectos (detección de fraude, segmentación de marketing, optimización de precios)

9. Justificación del Enfoque Aplicado

9.1 Elección del Tipo de Análisis: Predictivo

Tipos de análisis de datos:

La taxonomía estándar de análisis de datos identifica cuatro niveles (Gartner, 2012):

1. Descriptivo: ¿Qué pasó? (Reportes históricos, dashboards de KPIs)
2. Diagnóstico: ¿Por qué pasó? (Análisis de causas raíz, correlaciones)
3. Predictivo: ¿Qué pasará? (Modelos de forecasting, machine learning)
4. Prescriptivo: ¿Qué deberíamos hacer? (Optimización, simulación de escenarios)

Justificación de la elección predictiva:

Este proyecto adopta un enfoque predictivo porque:

1. Naturaleza proactiva del problema:

- El objetivo NO es entender por qué los clientes cancelaron en el pasado (diagnóstico)
- El objetivo ES identificar quién tiene alta probabilidad de cancelar en el futuro (predicción)
- Esta anticipación permite intervenciones preventivas antes de que el churn ocurra

2. Disponibilidad de datos históricos etiquetados:

- COTEL tiene 19 años de datos con labels naturales (churn = 1 si canceló, 0 si sigue activo)
- Esta estructura de datos es ideal para machine learning supervisado
- No hay necesidad de generar labels manualmente (como en problemas de detección de anomalías)

3. Requerimiento de accionabilidad:

- El modelo debe generar scoring continuo de riesgo (0-100%) para priorización
- No basta con clasificar binariamente (churn/no churn); se necesitan probabilidades para asignar recursos proporcionalmente
- Las predicciones deben actualizarse mensualmente para reflejar cambios en comportamiento

4. Horizonte temporal específico:

- Se busca predecir churn en ventana de 60-90 días
- Este horizonte es lo suficientemente largo para diseñar e implementar estrategias de retención
- Es lo suficientemente corto para que las señales de riesgo sean detectables en los datos

9.2 Justificación de Herramientas Tecnológicas

Stack tecnológico seleccionado:

Python como lenguaje principal

Justificación:

1. Ecosistema de ciencia de datos más maduro:
    - Pandas: Manipulación de datos tabulares (equivalente a SQL + Excel con esteroides)
    - NumPy: Operaciones numéricas vectorizadas de alto rendimiento
    - Scikit-learn: Suite completa de algoritmos de ML con API consistente
2. Visualización de datos:
    - Matplotlib/Seaborn: Gráficos estáticos de alta calidad para EDA
    - Plotly: Gráficos interactivos para dashboards web
    - Sweetviz/Pandas Profiling: Reportes automáticos de EDA

9.3 Garantías de Reproducibilidad y Claridad

Reproducibilidad

Definición: La capacidad de otro investigador (o el mismo en el futuro) de ejecutar el código y obtener exactamente los mismos resultados.

Estrategias implementadas:

- Control de versiones con Git:
- Gestión de dependencias con requirements.txt:
- Instalación de entorno reproducible:
- Documentación exhaustiva:

Claridad del Flujo de Trabajo

Definición: La capacidad de cualquier persona (técnica o no técnica) de entender qué hace el proyecto y cómo ejecutarlo.

10. Objetivo del Análisis

10.1 Propósito Central del Análisis

El objetivo principal de este análisis es desarrollar un modelo predictivo que identifique con 60-90 días de anticipación qué clientes de servicio de internet de COTEL tienen alta probabilidad de cancelar su contrato, permitiendo a la cooperativa implementar estrategias de retención proactivas, personalizadas y costo-efectivas antes de que el churn ocurra.

Este objetivo se vincula directamente con los objetivos específicos establecidos en la primera entrega:

1. Construir una base de datos analítica integrada que consolide información de múltiples fuentes operacionales de COTEL
2. Identificar variables predictoras mediante análisis exploratorio que revelen patrones de comportamiento diferencial entre clientes que cancelaron vs. activos

3. Desarrollar modelos de machine learning capaces de clasificar clientes en categorías de riesgo (alto, medio, bajo)
4. Crear visualizaciones interactivas que faciliten la toma de decisiones del equipo comercial y gerencial

10.2 Tipo de Análisis: Enfoque Híbrido

Este proyecto implementa un enfoque analítico híbrido que combina tres niveles:

A. Análisis Exploratorio de Datos (EDA)

Objetivo: Comprender la estructura, distribución y calidad de los datos antes de modelar.

Actividades:

- Análisis univariado: Distribuciones de variables numéricas (histogramas, boxplots) y categóricas (gráficos de barras)
- Análisis bivariado: Relaciones entre variables predictoras y la variable objetivo (churn)
- Detección de outliers y valores anómalos
- Análisis de correlaciones y multicolinealidad
- Identificación de valores faltantes y estrategias de imputación

Herramientas: Pandas Profiling, Seaborn, Matplotlib, Sweetviz

B. Análisis Descriptivo

Objetivo: Caracterizar el fenómeno del churn en COTEL mediante estadísticas agregadas.

Actividades:

- Cálculo de tasa de churn histórica (mensual, trimestral, anual)
- Segmentación de clientes por características (edad, antigüedad, plan, zona)
- Análisis de cohortes: comparar comportamiento de clientes que ingresaron en diferentes periodos

- Cuantificación del impacto económico del churn (ingresos perdidos, CLV afectado)

Entregables: Dashboards descriptivos con KPIs clave para gerencia

C. Análisis Predictivo (Enfoque Principal)

Objetivo: Construir modelos de machine learning que predigan la probabilidad de churn de cada cliente.

Actividades:

- Preparación de datos: Limpieza, feature engineering, encoding de categóricas
- División temporal de datos: Train (70%) / Validation (15%) / Test (15%)
- Entrenamiento de múltiples algoritmos:
    - Regresión Logística: Modelo baseline interpretable
    - Random Forest: Modelo ensamblado robusto
    - XGBoost: Modelo de gradient boosting de alta precisión
- Optimización de hiperparámetros mediante GridSearchCV
- Evaluación con métricas apropiadas: Recall, Precision, F1-Score, ROC-AUC
- Análisis de feature importance para interpretabilidad

Entregable: Modelo productivo capaz de generar scoring mensual de riesgo de churn

Estos datasets son extraídos de bases de datos relacionales (PostgreSQL/Oracle) mediante consultas SQL y exportados a formato CSV para análisis.

11. Descripción del Dataset

11.1 Nombre y Origen de los Datasets

El proyecto utiliza cuatro datasets operacionales de la Cooperativa de Telecomunicaciones La Paz (COTEL), todos provenientes de sus sistemas de gestión interna:

1. Dataset de Contratos de Internet (contratos_internet.csv)

2. Dataset de Clientes (clientes.csv)
3. Dataset de Facturación (facturacion_internet.csv)
4. Dataset de Deudas (deudas_clientes.csv)

Estos datasets son extraídos de bases de datos relacionales (PostgreSQL/Oracle) mediante consultas SQL y exportados a formato CSV para análisis.

11.2 Dimensiones de los Datasets

Dataset	Registros	Variables	Periodo Temporal
Contratos Internet	50,247	18	2006-01-27 a 2025-10-23
Clientes	30,156	22	Acumulado histórico
Facturación	214,389	14	2019-01 a 2025-10
Deudas	12,458	6	Snapshot actual

TOTAL INTEGRADO	~307,250	60+ (post-join)	19 años de historia

11.3 Tipos de Datos Presentes

El conjunto de datos integrado incluye los siguientes tipos de variables:

Datos Numéricos:

- Continuos: monto_total, monto_cotel, deuda_telefonia, arpu, antiguedad_meses, edad_cliente
- Discretos: facturas_mora, facturas_totales, total_servicios, cuotas_modem

Datos Categóricos:

- Nominales: plan_comercial, forma_pago, cod_estado_contrato, anulado, sexo, tipo_personeria
- Ordinales: riesgo_crediticio (bajo, medio, alto), segmento_edad (joven, adulto, adulto mayor)

Datos Temporales:

- Fechas: f_contrato, f_rescision, f_instalacion, fecha_emision, f_nacimiento
- Periodos: periodo (formato YYYYMM, ej: 202310)

Datos de Texto:

- Texto libre: observaciones, observaciones2, motivo_anulacion
- Direcciones: direccion, direccion_esp (formato semi-estructurado)

Datos Identificadores:

- Claves primarias: contrato, cod_cliente, factura_interna
- Claves foráneas: cod_acci_contrato, cod_dosificacion

Datos Booleanos/Binarios:

- Indicadores: tiene_internet, tiene_tvcable, es_empresa, mora_cronica

- Variable objetivo: churn (0 = activo, 1 = cancelado)

11.4 Variables Objetivo

Variable Objetivo Principal:

- churn (binaria): Indica si el cliente canceló el servicio (1) o permanece activo (0)
    - Derivada de: cod_estado_contrato y anulado
    - Criterio: churn = 1 si anulado en ['RETIRADO', 'INACTIVO(RESCINDIDO)']
    - Distribución estimada: 30% churn, 70% activos (desbalanceada)

Variables Objetivo Secundarias (para análisis exploratorio):

- fecha_churn (fecha): Fecha exacta de cancelación (solo para clientes churned)
- meses_hasta_churn (numérico): Tiempo desde observación hasta cancelación (para análisis de ventana temporal)

11.5 Descripción Detallada de Variables por Dataset

Dataset 1: Contratos de Internet

Nombre Variable	Descripción	Tipo de Dato	Ejemplo
contrato	Identificador único del contrato	INTEGER	89002406
username	Usuario de acceso (mayormente vacío)	VARCHAR (50)	-

clave	Contraseña (no usado)	VARCHAR (50)	-
cod_tipo_plan	Código del tipo de plan	INTEGER	103
cod_categoria	Categoría del cliente	INTEGER	2
cod_tipo_pago	Tipo de modalidad de pago	INTEGER	NULL
cod_cliente	Identificador único del cliente	VARCHAR (12)	000000017963 1
f_contrato	Fecha de firma del contrato	DATE	2006-01-27
observaciones	Notas sobre el contrato	TEXT	"PUNTO COTEL"
cod_promotor	Código del promotor/vendedor	INTEGER	NULL
cod_estado_contrato	Estado actual del contrato	INTEGER	2
cod_estado_servicio	Estado del servicio técnico	INTEGER	0

cod_acci_contrato	Código de acción del contrato	INTEGER	35001596
f_rescision	Fecha de cancelación	DATE	NULL
f_instalacion	Fecha de instalación técnica	DATE	2006-02-03
tipo_conexion	Tipo de tecnología (1=cobre)	INTEGER	1
cod_forma_pago	Forma de pago (1=efectivo, 2=banco)	INTEGER	2
telefono	Número de teléfono asociado	VARCHAR (10)	2328132
f_reinstalacion	Fecha de reinstalación	DATE	NULL

Dataset 2: Clientes

Nombre Variable	Descripción	Tipo de Dato	Ejemplo
cod_cliente	Identificador único del cliente	VARCHAR (12)	0000000138006

ape_paterno	Apellido paterno	VARCHAR (50)	TORREZ
ape_materno	Apellido materno	VARCHAR (50)	MURIEL
nombres	Nombres completos	VARCHAR (100)	MARIA MARCELA
nombre_pila	Nombre para facturación	VARCHAR (150)	TORREZ MURIEL MARIA
direccion	Dirección principal	TEXT	PLAN 88 MANSANO 221
cod_documento	Tipo de documento (CI, NIT)	VARCHAR (10)	CI
nro_documento	Número de documento	VARCHAR (20)	4785503
tipo_personeria	Persona natural (N) o jurídica (J)	CHAR(1)	N
telefono_ref	Teléfono de referencia	VARCHAR (15)	NULL

abonado	Indicador de cliente activo	CHAR(1)	A
direccion_esp	Dirección específica adicional	TEXT	NULL
nro_ruc	Número de RUC (empresas)	VARCHAR (15)	NULL
sexo	Género del cliente	CHAR(1)	F
estado_civil	Estado civil	VARCHAR (20)	C
fax	Número de fax	VARCHAR (15)	NULL
casilla	Casilla postal	VARCHAR (20)	NULL
email	Correo electrónico	VARCHAR (100)	NULL
nombre_factura	Nombre para facturación	VARCHAR (150)	TORREZ MURIEL MARIA

ruc_factura	RUC para facturación	VARCHAR (15)	NULL
f_nacimiento	Fecha de nacimiento	DATE	1965-04-26
complemento	Complemento de CI	VARCHAR (5)	NULL

Dataset 3: Facturación

Nombre Variable	Descripción	Tipo de Dato	Ejemplo
cod_concesion	Código de concesión/zona	INTEGER	1
factura_interna	Número interno de factura	BIGINT	201310000108
cod_dosificacion	Código de dosificación SIN	INTEGER	62
contrato	Código del contrato facturado	INTEGER	1000161
periodo_desde	Inicio del periodo facturado	INTEGER	20131001

periodo_hasta	Fin del periodo facturado	INTEGER	20131031
telefono	Teléfono asociado	VARCHAR(10)	2486064
fecha_envio	Fecha de envío de factura	DATE	2013-10-31
fecha_emision	Fecha de emisión	DATE	2013-10-31
periodo	Periodo en formato YYYYMM	INTEGER	201310
monto_total	Monto total de la factura	DECIMAL(10,2)	110.27
cod_mensaje	Código de mensaje al cliente	INTEGER	1
monto_cf	Monto consumo fijo	DECIMAL(10,2)	110.27
estado	Estado de pago (P/G/NP)	CHAR(2)	P

Dataset 4: Deudas

Nombre Variable	Descripción	Tipo de Dato	Ejemplo
cod_cliente	Identificador del cliente	VARCHAR(12)	000000017963 1
deuda_telefonia	Deuda del servicio de telefonía	DECIMAL(10,2)	0.00
deuda_internet	Deuda del servicio de internet	DECIMAL(10,2)	250.50
deuda_tvcable	Deuda del servicio de TV Cable	DECIMAL(10,2)	0.00
deuda_total	Suma de todas las deudas	DECIMAL(10,2)	250.50
facturas_vencidas	Cantidad de facturas impagas	INTEGER	3
antiguedad_deuda_dias	Días desde la factura más antigua	INTEGER	92

11.6 Variables Derivadas

Además de las variables originales, el proyecto creará las siguientes variables derivadas:

Nombre Variable	Descripción	Tipo	Fórmula/Criterio
antiguedad_meses	Antigüedad del cliente en meses	INTEGER	(HOY - f_contrato) / 30
edad_cliente	Edad del cliente en años	INTEGER	(HOY - f_nacimiento) / 365
arpu	Ingreso promedio mensual	DECIMAL	AVG(monto_total) por cliente
std_arpu	Desviación estándar del ARPU	DECIMAL	STDDEV(monto_total)
cv_arpu	Coeficiente de variación	DECIMAL	std_arpu / arpu
tasa_pago	% facturas pagadas a tiempo	DECIMAL	COUNT(estado='P') / COUNT(*)
facturas_mora	Facturas en mora	INTEGER	COUNT(estado='NP')
meses_deuda	Meses de deuda acumulada	DECIMAL	deuda_internet / arpu

tiene_internet	Tiene servicio de internet	BOOLEAN	Derivado de contratos
tiene_tvcable	Tiene servicio de TV	BOOLEAN	Derivado de contratos
total_servicios	Cantidad de servicios	INTEGER	tiene_internet + tiene_tvcable + tiene_telefonia
tipo_bundle	Tipo de paquete	CATEGORICAL	SOLO/DUO/TRIO
es_cliente_nuevo	Cliente < 12 meses	BOOLEAN	antiguedad_meses < 12
mora_cronica	Mora en ≥3 facturas	BOOLEAN	facturas_mora >= 3
segmento_edad	Segmento por edad	CATEGORICAL	<30=JOVEN, 30-50=ADULTO, etc.
riesgo_crediticio	Nivel de riesgo	CATEGORICAL	Basado en meses_deuda
churn	Cliente canceló servicio	BOOLEAN (TARGET)	anulado IN ('RETIRADO', 'INACTIVO')

12. Proceso de Extracción de Datos

12.1 Metodología de Obtención de Datos

Los datos utilizados en este proyecto fueron obtenidos mediante acceso directo a las bases de datos operacionales de COTEL, previa autorización institucional. Este acceso fue facilitado por el área de Tecnologías de Información de la cooperativa, bajo estrictos protocolos de confidencialidad y uso ético de información de clientes.

Método de Extracción: Consultas SQL Directas

La extracción se realizó mediante consultas SQL ejecutadas directamente sobre la base de datos relacional de COTEL (PostgreSQL 12.x), utilizando credenciales de solo lectura para garantizar la integridad de los sistemas de producción.

Herramientas utilizadas:

- pgAdmin 4: Cliente de administración de PostgreSQL para exploración inicial del esquema de base de datos
- Python 3.10+: Para automatización de extracción y procesamiento
- SQLAlchemy 2.0: Biblioteca ORM para conexión segura a PostgreSQL
- Pandas 1.5+: Para carga, manipulación y exportación de datos
- python-dotenv: Para gestión segura de credenciales mediante variables de entorno

12.2 Flujo Detallado de Adquisición de Datos

Exploración del Esquema de Base de Datos

1. Acceso inicial: Conexión VPN a la red interna de COTEL
2. Mapeo de tablas: Identificación de tablas relevantes en el esquema public y cotel_operaciones
3. Documentación: Entrevistas con administradores de BD para comprender relaciones entre tablas

Tablas identificadas:

- tbl_contratos_internet (50K+ registros)

- tbl_clientes (30K+ registros)

- tbl_facturacion_mensual (200K+ registros)

- tbl_deudas_consolidadas (12K+ registros)

- tbl_planes_comerciales (catálogo)

- tbl_estados_contrato (catálogo)

13. Formatos y Fuentes de Datos

13.1 Formatos de Archivos

Todos los datasets fueron extraídos y almacenados en formato CSV (Comma-Separated Values) por las siguientes razones:

Ventajas del formato CSV para este proyecto:

- Portabilidad: Compatible con múltiples herramientas (Python, R, Excel, Tableau, Power BI)
- Legibilidad: Inspección rápida en editores de texto plano
- Interoperabilidad: Estándar universal para intercambio de datos tabulares
- Versionamiento: Fácil seguimiento de cambios en sistemas Git

14. Evidencia de Volumen y Variedad

14.1 Comprobación del Volumen Mínimo

El proyecto supera ampliamente el requisito mínimo de 30,000 registros establecido para proyectos de Big Data:

Tabla Resumen de Volumen por Dataset

Dataset	Registros	Columnas

Contratos de Internet	50,247	18
Clientes	30,156	22
Facturación	214,389	14
Deudas	12,458	6
TOTAL INTEGRADO	307,250	60+

14.2 Justificación de Variedad

El proyecto cumple con el criterio de variedad al integrar múltiples fuentes, tipos de datos y formatos:

Variedad de Fuentes

Fuente	Descripción	Naturaleza
Sistema de Contratos	Gestión de altas/bajas de servicios	Transaccional (eventos)
Base de Clientes	Datos maestros demográficos	Maestro (estático)
Sistema de Facturación	Generación mensual de facturas	Transaccional (periódico)

Motor de Deudas	Cálculo consolidado de morosidad	Analítico (derivado)

14. Referencias Bibliográficas de las Secciones Añadidas

Sobre extracción de datos y ETL:

- Kimball, R., & Caserta, J. (2004). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley.
- Redman, T. C. (2008). *Data Driven: Profiting from Your Most Important Business Asset*. Harvard Business Press.

Sobre ética y protección de datos:

- European Union. (2016). *General Data Protection Regulation (GDPR)*. Regulation (EU) 2016/679.
- Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A*, 374(2083), 20160360.
- Estado Plurinacional de Bolivia. (2011). *Ley General de Telecomunicaciones, Tecnologías de Información y Comunicación, Nº 164*.

Sobre formatos de datos y estándares:

- Shafranovich, Y. (2005). *Common Format and MIME Type for Comma-Separated Values (CSV) Files*. RFC 4180, IETF.
- Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergeau, F. (2008). *Extensible Markup Language (XML) 1.0* (Fifth Edition). W3C Recommendation.

Sobre verificación de calidad de datos:

- Batini, C., & Scannapieco, M. (2016). *Data and Information Quality: Dimensions, Principles and Techniques*. Springer.
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5-33.