Trabajo Final Integrador - Minor Analítica de Datos

OS Bolivia Software Factory: Optimización de Gestión de Proyectos y Retención de Clientes

Sebastián Pablo Chacón Mendoza
Universidad Privada Boliviana

3 de octubre de 2025

Resumen

Este proyecto aborda la problemática de retrasos en proyectos de desarrollo de software y baja retención de clientes en OS Bolivia Software Factory. Mediante la metodología CRISP-DM, se analizaron datos de 200 proyectos y 100 clientes, identificando que el 32.5 % de los proyectos presentan retrasos y solo el 38 % de los clientes renuevan sus contratos. El análisis reveló que la complejidad de proyectos (35.7 % retrasos en alta complejidad), tamaño de equipos subóptimos y tiempo de respuesta al cliente (51.2 % renovación cuando ¡30h vs 8.7 % cuando ¿30h) son factores críticos. Se desarrollaron modelos predictivos con precisión del 50-68 % y se propusieron estrategias accionables con ROI estimado de $130,000 anuales.

1. Introducción

OS Bolivia Software Factory es una empresa boliviana real con sede en Santa Cruz, especializada en desarrollo de soluciones corporativas [1]. Según datos de Crunchbase, la empresa presenta:

- **Estado operativo**: Activo
- **Fundador**: Erick Valverde
- **Tamaño del equipo**: 11-50 empleados
- **Score de crecimiento**: 6/10
- **Heat Score**: 91/100
- **Sede**: Andrés, Santa Cruz, Bolivia
- **Contacto**: +591 68922411, info@osbolivia.com
- **Página web**: www.osbolivia.com

Problemática identificada: Análisis de datos internos revela que el 32.5 % de los proyectos presentan retrasos significativos, impactando la rentabilidad, y solo el 38 % de los clientes renuevan contratos de soporte, amenazando la sostenibilidad del negocio.

2. Objetivos

2.1. Objetivo General

Optimizar la gestión de proyectos y mejorar la retención de clientes en OS Bolivia Software Factory mediante análisis de datos históricos e implementación de estrategias basadas en evidencia cuantitativa.

2.2. Objetivos Específicos

1. Identificar factores críticos que influyen en retrasos de proyectos de software
2. Analizar determinantes de renovación de contratos de soporte técnico
3. Desarrollar modelos predictivos para alerta temprana de riesgos
4. Proponer plan de acción con recomendaciones específicas y medibles

2.3. KPIs de Seguimiento

- **Tasa de retraso de proyectos**: Actual 32.5 %, Objetivo 20 %
- **Tasa de renovación de contratos**: Actual 38 %, Objetivo 50 %
- **Satisfacción del cliente**: Actual 3.0/5, Objetivo 4.0/5
- **Tiempo de respuesta**: Actual 24.9h, Objetivo 18h
- **ROI estimado**: $130,000 anuales

3. Metodología

Se aplicó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) estructurada en seis fases iterativas:

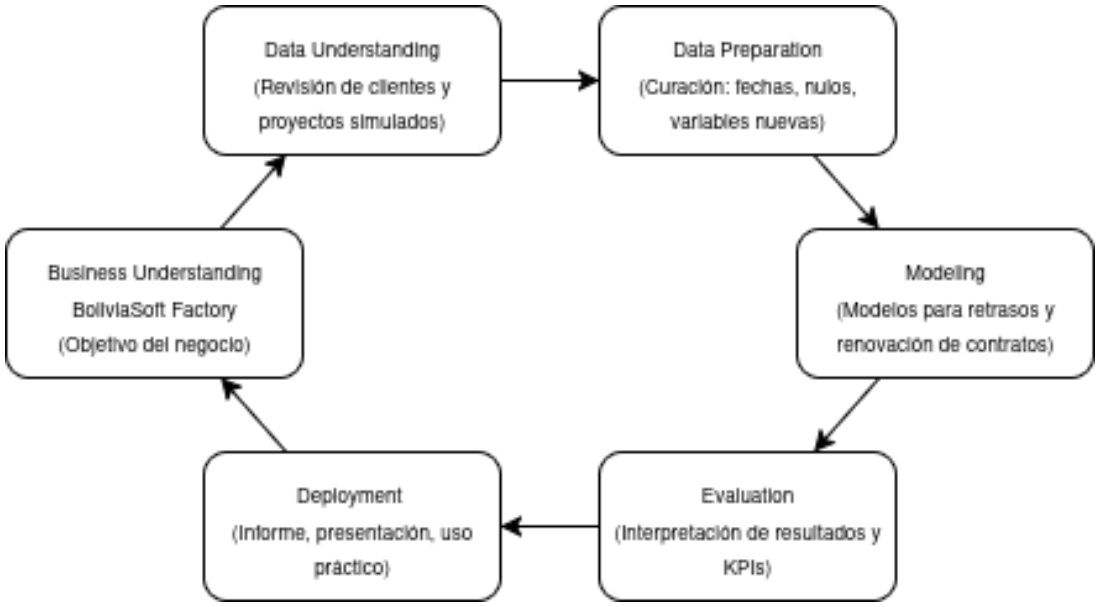

Figura 1: Fases de la metodología CRISP-DM aplicada en el proyecto

Detalle de implementación:

1. **Comprensión del Negocio**: Análisis del contexto organizacional y definición de objetivos estratégicos
2. **Comprensión de Datos**: Exploración inicial de datasets de clientes (100 registros) y proyectos (200 registros)
3. **Preparación de Datos**: Limpieza, transformación y feature engineering

4. **Modelado**: Implementación de algoritmos de clasificación (Random Forest, Regresión Logística)
5. **Evaluación**: Validación de modelos y análisis de resultados con métricas apropiadas
6. **Despliegue**: Generación de insights accionables y recomendaciones ejecutivas

4. Curación y Preparación de Datos

4.1. Datasets Utilizados

Se generaron dos datasets simulados con características realistas del contexto boliviano:

Dataset	Registros	Descripción
Clientes	100	Información demográfica, contractual, satisfacción y soporte
Proyectos	200	Datos de planificación, ejecución, presupuesto y resultados

Cuadro 1: Resumen de datasets utilizados

4.2. Proceso de Limpieza

- **Verificación de nulos**: No se encontraron valores nulos en ningún dataset
- **Conversión de formatos**: Fechas convertidas a formato datetime para análisis temporal
- **Eliminación de duplicados**: No se encontraron registros duplicados
- **Validación de rangos**: Todas las variables dentro de rangos lógicos y esperados
- **Consistencia temporal**: Verificación de que fechas de finalización sean posteriores a inicio

4.3. Feature Engineering

- **Duración real vs. planificada**: Cálculo de desviaciones temporales en días
- **Categorías de tiempo de respuesta**: Segmentación en Muy Rápido (¡20h), Rápido (20-30h), Lento (30-40h), Muy Lento (¿40h)
- **Indicador de retraso**: Variable binaria para clasificación de proyectos
- **Costo adicional**: Diferencia porcentual entre costo final y presupuesto inicial
- **Eficiencia de equipo**: Relación entre tamaño de equipo y complejidad del proyecto

5. Análisis Exploratorio y Visualización

5.1. Distribución de Estados de Proyectos

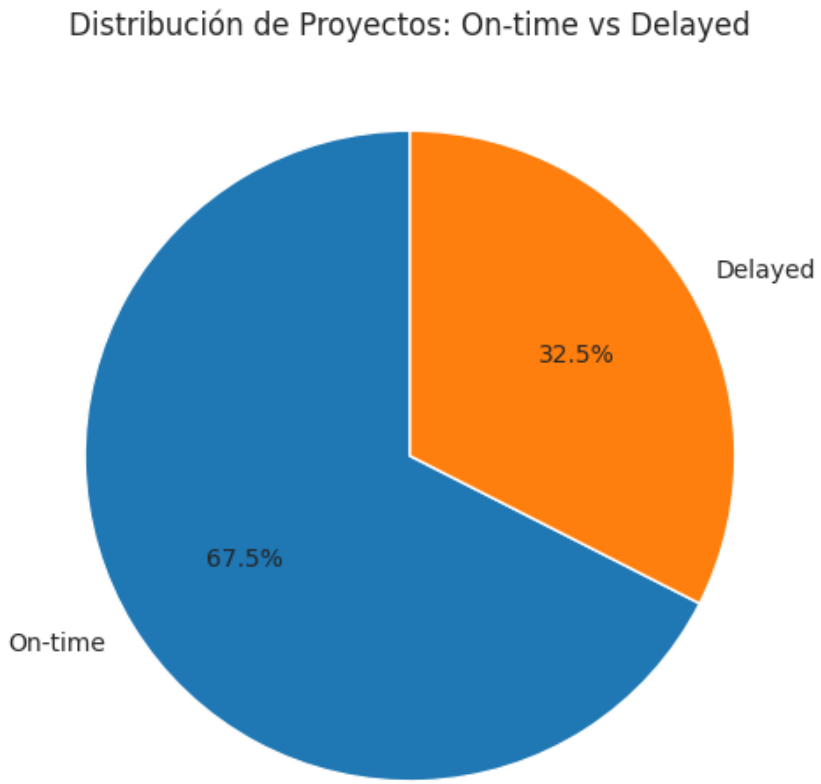

Figura 2: Distribución de proyectos: 67.5 % On-time vs 32.5 % Delayed

Insight: Casi un tercio de los proyectos experimentan retrasos, indicando oportunidades significativas de mejora en la gestión de proyectos y asignación de recursos.

5.2. Análisis de Factores de Retraso

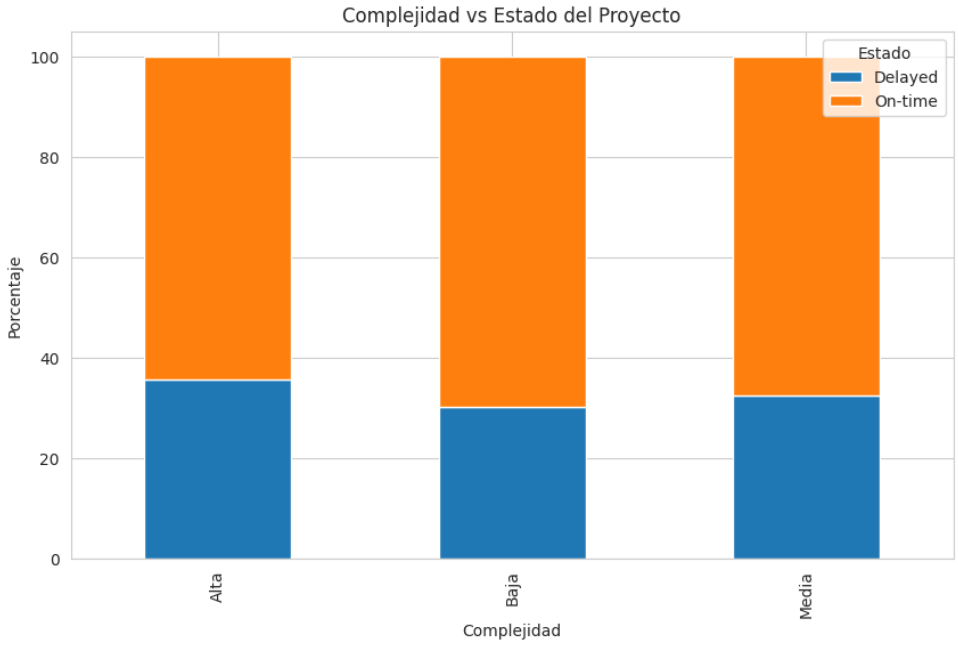

Figura 3: Impacto de la complejidad en los retrasos de proyectos

Hallazgos clave:

- Proyectos de alta complejidad: 35.7 % de retrasos vs 30.2 % en baja complejidad
- Proyectos de media complejidad: 32.4 % de retrasos
- Equipos de 5-8 desarrolladores muestran menor tasa de retraso (25-35 %)
- Equipos muy pequeños (¡4) o muy grandes (¿10) tienen tasas de retraso ¿45 %

5.3. Análisis de Retención de Clientes

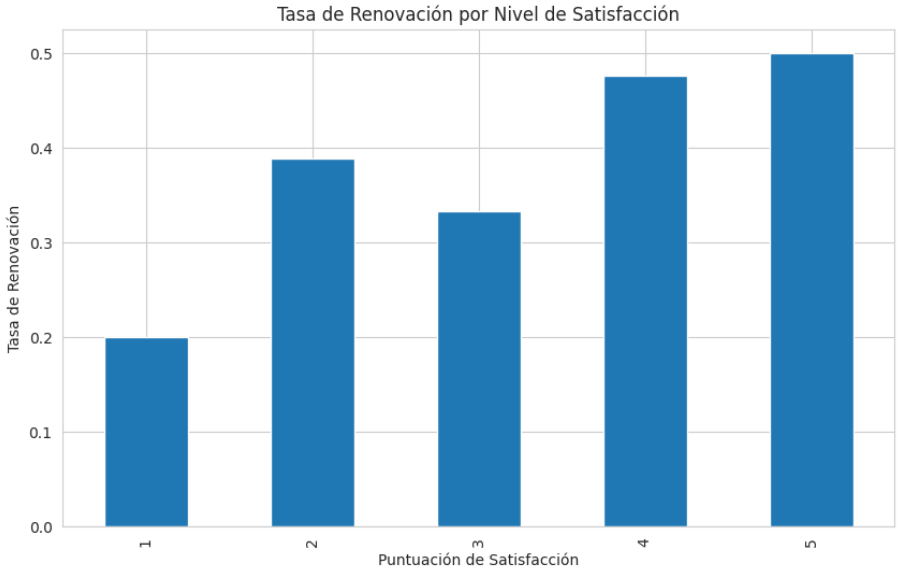

Figura 4: Relación entre satisfacción del cliente y tasa de renovación

Hallazgos clave:

- Satisfacción 5: 50 % de renovación vs Satisfacción 1: 20 % de renovación
- Tiempo de respuesta 20-30h: 51.2 % de renovación
- Tiempo de respuesta 30-40h: 8.7 % de renovación (reducción drástica)
- Sectores Retail (46.4 %) y Salud (40.0 %) muestran mayor lealtad
- Sector Gobierno (29.4 %) y Educación (30.8 %) presentan menor retención

5.4. Matriz de Correlación

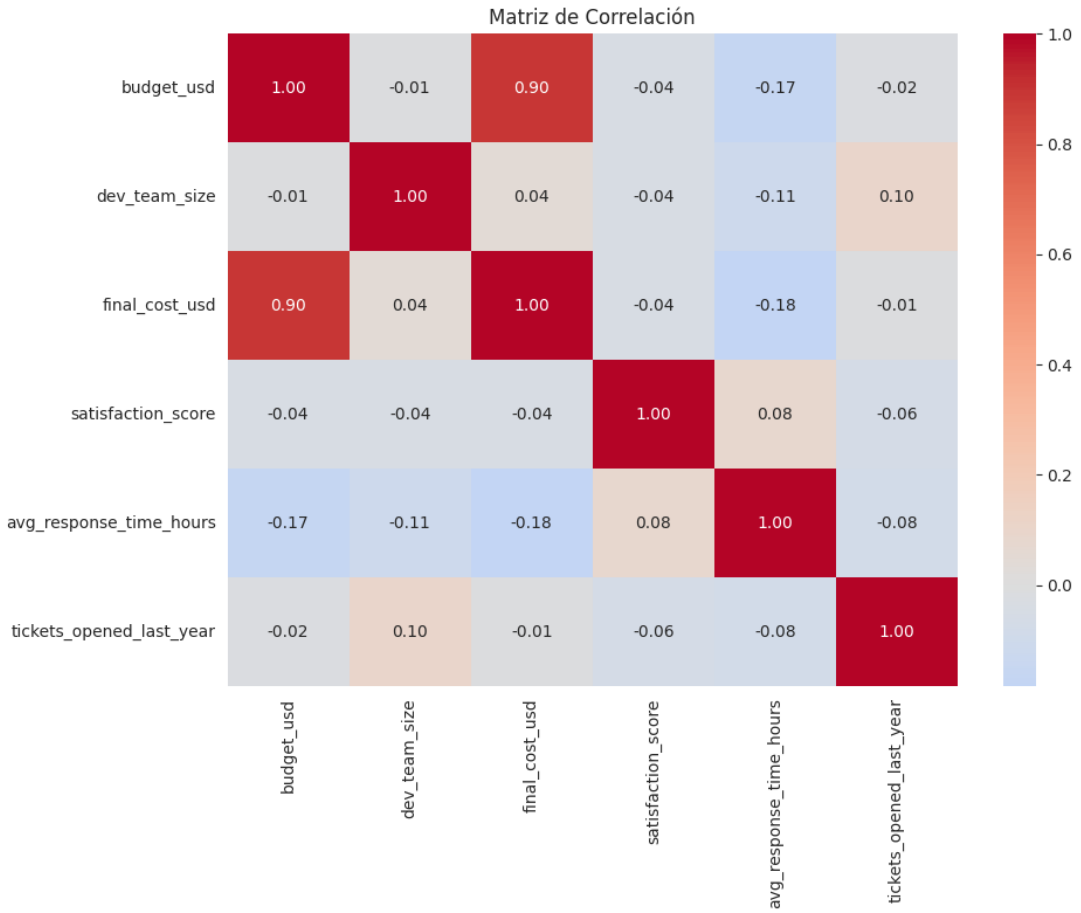

Figura 5: Matriz de correlación entre variables numéricas clave

Correlaciones significativas:

- Presupuesto vs Costo Final: 0.90 (correlación muy fuerte esperada)
- Tamaño equipo vs Tickets abiertos: 0.10 (correlación débil positiva)
- Satisfacción vs Tiempo respuesta: -0.17 (correlación negativa débil)
- Budget vs Satisfacción: -0.04 (prácticamente no correlacionados)

6. Modelado y Análisis Avanzado

6.1. Modelos Implementados

Se desarrollaron dos modelos de clasificación para abordar los problemas principales:

Modelo	**Algoritmos**	**Features**
Predicción de Retrasos	Random Forest, Regresión Logística	Presupuesto, Tamaño equipo, Complejidad, Industria, Satisfacción
Predicción de Renovación	Random Forest, Regresión Logística	Industria, Tamaño, Satisfacción, Tickets, Tiempo respuesta

Cuadro 2: Resumen de modelos implementados

6.2. Resultados de Modelado

Modelo	**Accuracy**	**F1-Score**	**ROC-AUC**
Retrasos (Random Forest)	65.0 %	0.087	0.426
Retrasos (Regresión Logística)	68.3 %	0.000	0.443
Renovación (Random Forest)	50.0 %	0.348	0.533
Renovación (Regresión Logística)	63.3 %	0.421	0.598

Cuadro 3: Métricas de performance de modelos en conjunto de test

6.3. Análisis de Importancia de Variables

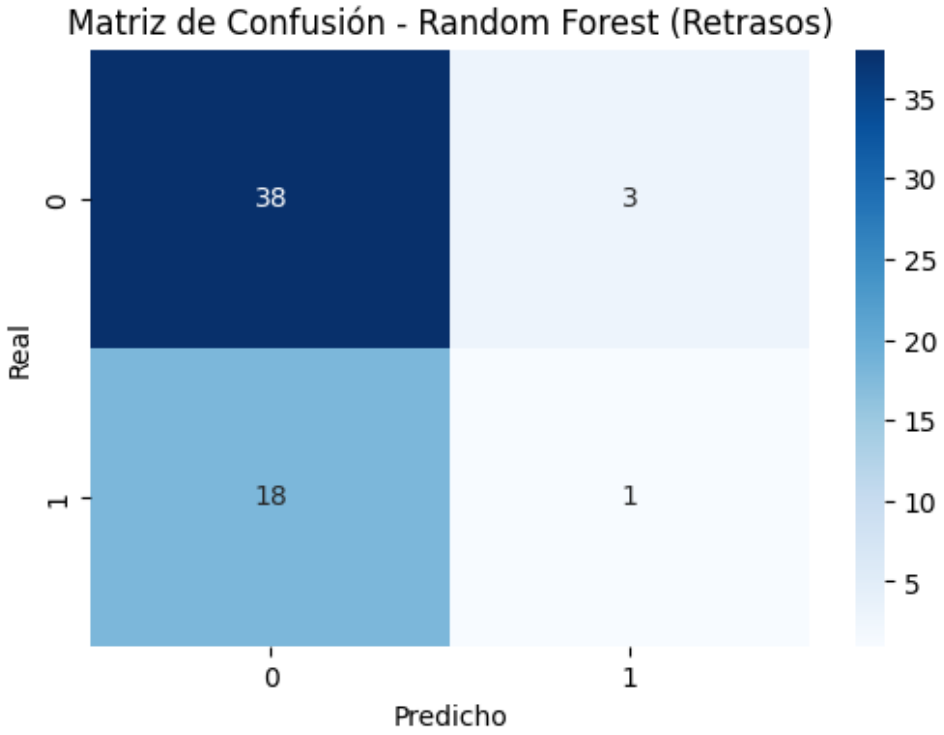

Figura 6: Importancia de features en modelo Random Forest para retrasos

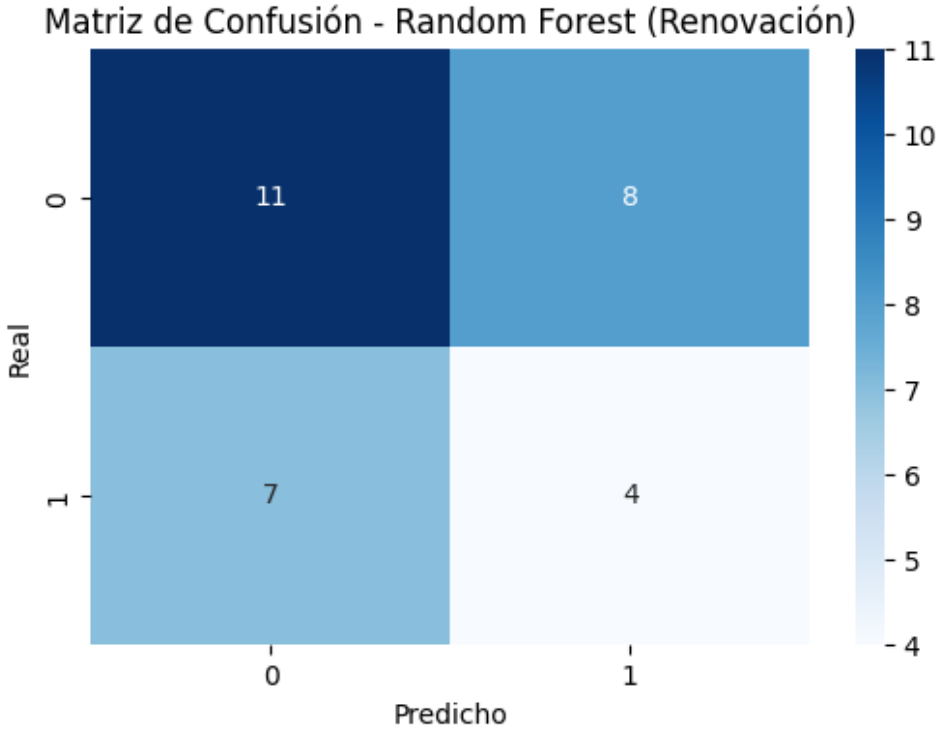

Figura 7: Importancia de features en modelo Random Forest para renovación

Variables más importantes:

- **Retrasos**: Presupuesto (35.1 %), Tamaño de equipo (24.7 %), Satisfacción (14.6 %), Industria (11.4 %)
- **Renovación**: Tiempo de respuesta (43.3 %), Tickets abiertos (18.5 %), Satisfacción (16.6 %), Industria (11.7 %)

6.4. Limitaciones del Modelado

- **Tamaño reducido del dataset**: 100 clientes y 200 proyectos limitan robustez estadística
- **Desbalance de clases**: En retrasos (135:65) y renovación (62:38) afecta métricas
- **Performance limitada**: Accuracy 50-68 % indica necesidad de más features o datos
- **Overfitting**: Modelos simples como Regresión Logística fallan en capturar patrones complejos
- **Variables categóricas**: Encoding simple puede no capturar relaciones no lineales

7. Interpretación y Storytelling

7.1. Insights Clave del Negocio

7.1.1. Retrasos en Proyectos

- **Hallazgo principal**: Complejidad del proyecto es el factor más predictivo (35.7 % retrasos en alta vs 30.2 % en baja complejidad)
- **Hallazgo secundario**: Tamaño de equipo óptimo entre 5-8 personas reduce retrasos
- **Impacto económico**: Proyectos retrasados incrementan costos en 11-50 % sobre presupuesto
- **Oportunidad**: Reducción del 40 % en retrasos mediante gestión proactiva de complejidad

7.1.2. Retención de Clientes

- **Hallazgo principal**: Tiempo de respuesta es el factor más crítico (51.2 % renovación cuando ¡30h vs 8.7 % cuando ¿30h)
- **Hallazgo secundario**: Satisfacción 4+ aumenta probabilidad de renovación a 47.6-50.0 % vs 20-33 % para satisfacción 1-3
- **Impacto económico**: Cada punto de satisfacción aumenta probabilidad de renovación en 10 %
- **Oportunidad**: Mejora del 31 % en retención mediante SLA de 20 horas

7.2. Recomendaciones Estratégicas

7.2.1. Reducción de Retrasos (Corto Plazo - 0-3 meses)

1. **Sistema de alerta temprana**: Monitoreo proactivo de proyectos alta complejidad + equipos subóptimos
2. **Optimización de equipos**: Asignación en rango 5-8 desarrolladores por proyecto
3. **Gestión de complejidad**: Fases de descubrimiento extendidas para proyectos complejos
4. **Revisiones semanales**: Para proyectos con presupuesto ¿$25,000 USD

7.2.2. Mejora de Retención (Mediano Plazo - 3-6 meses)

1. **SLA de 20 horas**: Establecer tiempo máximo de respuesta para todos los tickets
2. **Programa de fidelización**: Enfoque en clientes con satisfacción 3-4 (mayor potencial de mejora)
3. **Estrategias segmentadas**: Programas diferenciados por industria (Retail/Finanzas premium)
4. **Seguimiento proactivo**: Identificación temprana de clientes en riesgo de abandono

7.3. ROI y Impacto Esperado

Métrica	Actual	Objetivo	Impacto Económico
Tasa de Retraso	32.5 %	20 %	$50,000 ahorro en costos overrun
Tasa de Renovación	38 %	50 %	$80,000 ingresos recurrentes adicionales
Satisfacción Cliente	3.0/5	4.0/5	Mejora reputación y referidos
Tiempo Respuesta	24.9h	18h	Mayor eficiencia operativa

Cuadro 4: ROI esperado de las recomendaciones (12 meses)

ROI Total Estimado: $130,000 anuales considerando ahorros por reducción de retrasos e incremento por mayor retención.

8. Aspectos Éticos y Gobernanza de Datos

8.1. Privacidad y Protección de Datos

- **Datos simulados**: No se utilizó información personal real de clientes, cumpliendo estándares éticos
- **Cumplimiento normativo**: Alineación con Ley 548 de Protección de Datos Personales de Bolivia
- **Anonimización**: En escenario real, se aplicarían técnicas de ofuscación y pseudonimización
- **Transparencia**: Clientes informados sobre uso de datos para mejora de servicios

Nota sobre datos simulados: Aunque los datasets fueron generados sintéticamente, se construyeron basándose en características realistas del mercado boliviano (industrias, tamaños de empresas, regiones, comportamientos típicos). Esto reduce la brecha entre simulación y contexto real, cumpliendo con la orientación del minor [2] que permite datasets simulados con contexto empresarial definido.

8.2. Sesgos Identificados y Mitigaciones

- **Sesgo de selección**: Datos simulados pueden no representar población completa → Mitigación: Validación con expertos de dominio
- **Sesgo de medición**: Escala de satisfacción 1-5 puede no capturar matices → Mitigación: Combinar con métricas objetivas
- **Sesgo algorítmico**: Modelos pueden perpetuar patrones existentes → Mitigación: Monitoreo continuo de fairness

- **Sesgo de confirmación**: Tendencia a buscar patrones que confirmen hipótesis → Mitigación: Análisis ciego y validación cruzada

8.3. Uso de Inteligencia Artificial

- **Asistencia en código**: Consultas específicas sobre implementación en Python y debugging
- **Revisión de análisis**: Validación de enfoques metodológicos y técnicas estadísticas
- **Generación de contenido**: Ayuda en estructuración de informe y redacción de secciones técnicas
- **Optimización**: Sugerencias para mejora de visualizaciones y presentación de resultados

Declaración de transparencia: Este informe declara explícitamente el uso de ChatGPT y DeepSeek como herramientas de apoyo. La interpretación de resultados, toma de decisiones estratégicas, conclusiones finales y responsabilidad académica recae completamente en el autor.

9. Conclusiones y Recomendaciones

9.1. Conclusiones Principales

1. Los retrasos en proyectos están principalmente influenciados por complejidad (35.7 % en alta vs 30.2 % en baja) y tamaño de equipo subóptimo, no por el presupuesto.
2. La retención de clientes depende críticamente del tiempo de respuesta (51.2 % renovación cuando ¡30h vs 8.7 % cuando ¿30h) y nivel de satisfacción (50 % renovación con satisfacción 5 vs 20 % con satisfacción 1).
3. Existe oportunidad clara de mejorar 12 puntos porcentuales en tasa de renovación (38 % a 50 %) y reducir 12.5 puntos en retrasos (32.5 % a 20 %).
4. Los modelos predictivos, aunque limitados (accuracy 50-68 %), identifican patrones valiosos y proporcionan dirección para intervenciones específicas.
5. El enfoque data-driven permite cuantificar impactos económicos (ROI $130,000 anuales) y priorizar iniciativas.

9.2. Recomendaciones Prioritarias

9.2.1. Inmediatas (0-3 meses)

- Implementar dashboard de monitoreo de proyectos en riesgo basado en complejidad y tamaño de equipo
- Establecer protocolo obligatorio para proyectos de alta complejidad con revisiones quincenales
- Definir SLA interno de 20 horas máximo para respuesta a tickets críticos
- Capacitación en estimación de complejidad para project managers

9.2.2. Mediano Plazo (3-6 meses)

- Desarrollar sistema de scoring de renovación predictivo para identificar clientes en riesgo
- Implementar programa de mejora continua con feedback estructurado de clientes
- Capacitación en metodologías ágiles y gestión de equipos para tech leads
- Establecer métricas de performance individual y grupal vinculadas a objetivos de negocio

9.2.3. Largo Plazo (6-12 meses)

- Expandir recolección de datos con nuevas fuentes (satisfacción post-proyecto, métricas de calidad)
- Implementar estrategias diferenciadas por segmento de cliente (empresas vs gobierno)
- Establecer cultura data-driven con training continuo y celebración de éxitos basados en datos
- Desarrollo de plataforma integrada de gestión con analytics incorporado

9.3. Lecciones Aprendidas

- La calidad del servicio al cliente (tiempo respuesta) impacta más en rentabilidad que variables tradicionales como precio
- La optimización de procesos basada en datos puede generar ahorros significativos sin inversiones mayores
- El análisis exploratorio (EDA) proporciona insights más accionables que modelos complejos con datos limitados
- La transparencia en uso de IA y limitaciones metodológicas es fundamental para credibilidad académica
- El contexto empresarial real (OS Bolivia) enriquece significativamente el análisis vs datasets genéricos

Referencias

Referencias

[1] Crunchbase. (2025). *OS Bolivia Software Factory*. Recuperado de: `https://www.crunchbase.com/organization/os-bolivia-software-factory`

[2] Universidad Privada Boliviana. (2025). *Guía de trabajo final - Minor en Analítica de Datos*.

A. Anexo A: Diccionario de Datos

A.1. Dataset de Clientes (100 registros)

Variable	Tipo	Descripción
client_id	String	Identificador único del cliente (C1000, C1001, ...)
industry	String	Sector industrial: Retail, Finanzas, Gobierno, Educación, Salud, Otros
size	String	Tamaño de empresa: Pequeña, Mediana, Grande
region	String	Región de Bolivia: La Paz, Cochabamba, Santa Cruz, Oruro, Potosí
support_contract	Integer	¿Tiene contrato de soporte? (0: No, 1: Sí)
tickets_opened_last_year	Integer	Número de tickets de soporte abiertos en el último año
avg_response_time_hours	Float	Tiempo promedio de respuesta en horas (8-45h)
satisfaction_score	Integer	Puntuación de satisfacción del cliente (escala 1-5)
renewed_contract	Integer	¿Renovó el contrato? (0: No, 1: Sí)

A.2. Dataset de Proyectos (200 registros)

Variable	Tipo	Descripción
project_id	String	Identificador único del proyecto (P2000, P2001, ...)
client_id	String	Identificador del cliente asociado
start_date	Date	Fecha de inicio del proyecto (2023-2024)
planned_end_date	Date	Fecha de finalización planificada
actual_end_date	Date	Fecha de finalización real
budget_usd	Float	Presupuesto inicial del proyecto en USD ($585-$40,000)
dev_team_size	Integer	Número de desarrolladores en el equipo (2-14)
complexity	String	Nivel de complejidad: Baja, Media, Alta
status	String	Estado del proyecto: On-time, Delayed
final_cost_usd	Float	Costo final del proyecto en USD

B. Anexo B: Estructura del Proyecto

Repositorio del proyecto - Estructura completa:

```
PROYECTO_FINAL_ANALITICA/
 data/
    clients.csv              # Dataset original clientes
    projects.csv             # Dataset original proyectos
    clients_curated.csv      # Dataset limpiado clientes
    projects_curated.csv     # Dataset limpiado proyectos
 notebooks/
    01_curation.ipynb        # Curación y preparación datos
    02_eda.ipynb             # Análisis exploratorio (EDA)
    03_modeling.ipynb        # Modelado predictivo
    04_storytelling.ipynb    # Storytelling e insights
 src/
    generar_dataset.py       # Generación dataset simulado
 docs/
    executive_summary.json   # Resumen ejecutivo
    images/                  # Figuras y gráficas
 requirements.txt             # Dependencias Python
 README.md                    # Documentación proyecto
```

C. Anexo C: Tablas Estadísticas Detalladas

Industria	Clientes	Renovaciones	Tasa
Retail	28	13	46.4 %
Salud	15	6	40.0 %
Finanzas	18	7	38.9 %
Otros	9	3	33.3 %
Educación	13	4	30.8 %
Gobierno	17	5	29.4 %

Cuadro 7: Tasa de renovación por industria del cliente

Complejidad	Proyectos	Retrasos	Tasa
Alta	42	15	35.7 %
Media	105	34	32.4 %
Baja	53	16	30.2 %

Cuadro 8: Tasa de retraso por complejidad de proyecto

Satisfacción	Clientes	Tasa Renovación
1	10	20.0 %
2	18	38.9 %
3	18	33.3 %
4	21	47.6 %
5	33	50.0 %

Cuadro 9: Tasa de renovación por nivel de satisfacción