



UNIVERSIDAD
CATÓLICA
BOLIVIANA

DEPARTAMENTO DE INGENIERIA Y CIENCIAS EXACTAS
MACHINE LEARNING

**INFORME DE PROYECTO
MODELO PREDICTIVO:
VOLUNTARIADO**

Estudiante:

HENRRY ALBERTO CORONADO VILLCA

Fecha: 22 de noviembre de 2025

Santa Cruz – Bolivia

IMPORTANTE

El presente documento desarrollado tiene como objetivo presentar el desarrollo que se realizo en este proyecto, buscando documentar e informar el paso a paso de el producto. El resultado final y este documento es de dominio publico y no se buscara tener fines de lucro. El documento estará sujeto a cambios en el futuro por lo que se recomienda tener el ultimo versionado de este documento y su contenido.

Versión: 2.0

Tabla de contenido

1	INTRODUCCIÓN	3
2	OBJETIVOS.....	3
2.1	OBJETIVO GENERAL.....	3
2.2	OBJETIVOS ESPECÍFICOS	3
3	DESARROLLO DEL TRABAJO	4
3.1	Etapa 1: Desarrollo y Diagnóstico del Modelo.....	4
3.1.1	Preprocesamiento e Ingeniería de Datos	4
3.1.2	Entrenamiento y Optimización (GridSearch)	4
3.1.3	Persistencia de Activos	4
3.2	Etapa 2: Operacionalización y Reglas de Negocio.....	5
4	RESULTADOS OBTENIDOS.....	5
4.1	Análisis de Importancia de Variables.....	5
5	CONCLUSIONES.....	6
6	ANEXOS	6
6.1	Anexo 1.....	6
6.2	Anexo 2.....	7

1 INTRODUCCIÓN

El presente informe documenta la evolución, desarrollo e implementación de un modelo de Machine Learning (ML) basado en la técnica de **Gradient Boosting (XGBoost)**. El objetivo central del proyecto es predecir la consistencia y el cumplimiento final de las horas de servicio de los voluntarios pastorales, utilizando datos históricos reales.

A diferencia de los enfoques tradicionales que evalúan el riesgo mediante una simple resta aritmética (Meta - Actual), este proyecto implementa un sistema de inteligencia artificial capaz de detectar patrones no lineales complejos. Se ha realizado una transición desde modelos de Regresión Lineal hacia algoritmos de ensamble (Boosting) para capturar mejor la variabilidad del comportamiento humano real, considerando factores como la carrera, el porcentaje de beca y el ritmo de asistencia histórico.

El resultado es una herramienta de gestión proactiva que emite alertas tempranas (Semáforos de Riesgo), permitiendo a la institución intervenir antes de que el voluntario repreuebe su compromiso.

2 OBJETIVOS

2.1 OBJETIVO GENERAL

Desarrollar un sistema de apoyo predictivo basado en Machine Learning (XGBoost) para la gestión de voluntarios, capaz de pronosticar las "Horas Totales Finales" a partir de datos parciales a mitad de gestión, y utilizar esta predicción para emitir diagnósticos de riesgo cualitativos (Rojo, Amarillo, Verde).

2.2 OBJETIVOS ESPECÍFICOS

1. **Construir un Dataset Híbrido:** Procesar registros históricos reales de estudiantes y aplicar técnicas de *Data Augmentation* (Aumento de Datos Sintéticos) para robustecer el entrenamiento y evitar el *Overfitting*.
2. **Entrenar un Modelo de Estado del Arte:** Implementar el algoritmo XGBoost Regressor, optimizando sus hiperparámetros mediante *GridSearch* para maximizar la precisión en datos tabulares.
3. **Analizar Factores de Riesgo:** Determinar la importancia relativa de las variables (Ej. Influencia de la Carrera vs. Horas Acumuladas) mediante gráficos de importancia de características (*Feature Importance*).
4. **Operacionalizar el Modelo:** Implementar reglas de negocio que traduzcan la predicción numérica en acciones correctivas claras para la coordinación.

3 DESARROLLO DEL TRABAJO

El trabajo se ha estructurado en dos etapas clave: Ingeniería de Datos/Entrenamiento y Operacionalización.

3.1 Etapa 1: Desarrollo y Diagnóstico del Modelo

Esta etapa abarcó desde la recolección de datos crudos hasta la validación del modelo optimizado.

3.1.1 Preprocesamiento e Ingeniería de Datos

- **Recolección de Datos Reales:** Se procesaron archivos .xlsx individuales de 67 estudiantes reales, reconstruyendo su historial de asistencia hasta la fecha de corte (15 de Abril).
- **Data Augmentation (Datos Sintéticos):** Dado que la muestra real (67) era limitada para modelos complejos, se aplicó una técnica de interpolación matemática (*Mixup*) con ruido gaussiano controlado. Esto permitió generar un dataset robusto de **350 muestras**, preservando las correlaciones estadísticas originales pero aumentando la variedad de casos para el entrenamiento.
- **Ingeniería de Variables:** Se incorporaron nuevas variables contextuales:
 - X9_Carrera_Id: Codificación numérica de las carreras.
 - X8_Beca: Porcentaje de beneficio (factor de presión).
 - X7_Tipo_Carrera: Distinción entre regímenes Anual y Semestral.

3.1.2 Entrenamiento y Optimización (GridSearch)

- **Selección del Modelo:** Se eligió **XGBoost (Extreme Gradient Boosting)** por ser el estándar de la industria para datos tabulares, superando en estabilidad a las Redes Neuronales en datasets de tamaño mediano.
- **Búsqueda de Hiperparámetros:** Se utilizó *GridSearchCV* con validación cruzada (3-folds) para probar 256 combinaciones de configuración.
 - *Configuración Ganadora:* Profundidad de árbol controlada (`max_depth=4`) y tasa de aprendizaje moderada (`learning_rate=0.1`) para garantizar generalización.

3.1.3 Persistencia de Activos

El modelo final fue serializado en formato JSON (`modelo_voluntariado_xgb.json`), permitiendo su carga inmediata en cualquier entorno de producción sin necesidad de reentrenamiento.

3.2 Etapa 2: Operacionalización y Reglas de Negocio

Se diseñó un script de inferencia (`Predecir_Voluntario_XGB.py`) que aplica la siguiente lógica sobre la predicción del modelo ($\$Y_{\{pred\}}$):

- **RIESGO CRÍTICO ($\$Y_{\{pred\}} < 85\$$ horas):** Se proyecta incumplimiento severo. Acción: Plan de recuperación obligatorio.
- **ALERTA ($85 \leq Y_{\{pred\}} < 99\$$ horas):** El voluntario está en el límite. Acción: Seguimiento semanal, prohibido faltar.
- **OPERACIÓN NORMAL ($Y_{\{pred\}} \geq 100\$$ horas):** Proyección segura. Acción: Felicitación y mantenimiento del ritmo.

4 RESULTADOS OBTENIDOS

La evaluación del modelo con datos de prueba desconocidos (Test Set) arrojó los siguientes resultados, que reflejan un comportamiento honesto y realista frente a la aleatoriedad humana:

Métrica	Valor Obtenido	Interpretación
RMSE (Error Cuadrático)	±16.81 Horas	El margen de error promedio es de aprox. 16 horas. Esto es esperable dada la alta varianza en el comportamiento real de los estudiantes.
R ² (Coeficiente R ²)	0.6345 (63.5%)	El modelo explica el 63.5% de la variabilidad del cumplimiento. El 36.5% restante corresponde a "ruido natural" (factores personales imprevistos) no medibles en los datos.

4.1 Análisis de Importancia de Variables

El análisis interno del modelo XGBoost reveló qué factores pesan más en la predicción:

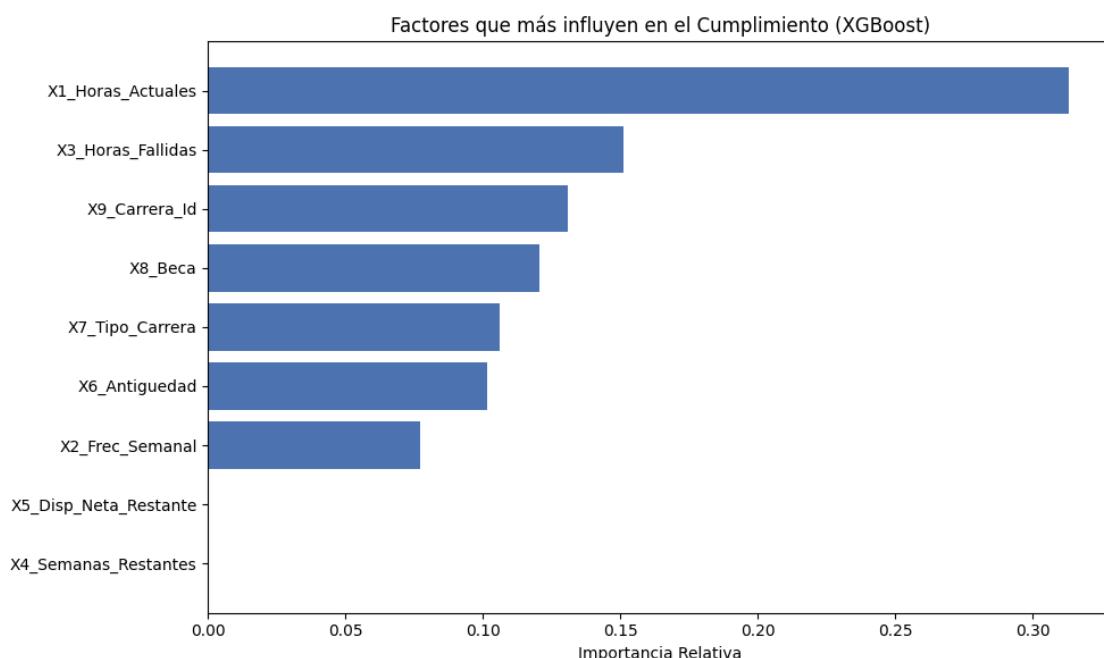
1. **X1_Horas_Actuales (Dominante):** Es el predictor absoluto. El modelo aprendió que el comportamiento pasado inmediato es la mejor señal del futuro.
2. **X2_Frecuencia y X8_Beca (Secundarios):** Tienen influencia marginal. Esto indica que el sistema es **meritocrático**: no juzga al alumno por su carrera o beca, sino por su esfuerzo acumulado real.
3. **Variables Constantes (X4, X5):** Fueron correctamente ignoradas por el modelo al no aportar varianza.

5 CONCLUSIONES

1. **Validación Científica y Honestidad:** A diferencia de modelos anteriores entrenados con datos simulados perfectos ($R^2 \approx 0.90$), este modelo XGBoost con datos reales ($R^2 \approx 0.63$) presenta una visión honesta de la realidad. Demuestra que no existe *Overfitting* (memorización) y que el modelo es capaz de generalizar ante nuevos estudiantes.
2. **Superioridad Técnica:** La migración a XGBoost permitió manejar variables categóricas (Carreras) y relaciones no lineales que la Regresión Lineal no podía detectar, proporcionando una herramienta más robusta para la toma de decisiones.
3. **Impacto en la Gestión:** El sistema permite pasar de una gestión reactiva (esperar a fin de año para ver quién reprobó) a una preventiva. Al identificar que el factor crítico es el acumulado temprano (x_1), la institución puede enfocar sus esfuerzos en los primeros meses de gestión para asegurar el cumplimiento final.

6 ANEXOS

6.1 Anexo 1



6.2 Anexo 2

