

Instrucciones Generales

Este taller evalúa su capacidad de aplicar técnicas de Machine Learning supervisado y no supervisado en contextos reales de consultoría económica. El taller consta de tres puntos independientes que simulan proyectos de consultoría para distintos clientes. En cada punto se espera rigor técnico, pero también **capacidad de comunicar resultados a audiencias no técnicas** y una reflexión sobre las implicaciones éticas y de responsabilidad de las decisiones algorítmicas.

Importante

Estructura de evaluación:

- **Punto 1 — Clasificación:** 1.75 puntos (1.50 base + 0.25 competencia).
- **Punto 2 — Regresión:** 1.75 puntos (1.50 base + 0.25 competencia).
- **Punto 3 — Segmentación:** 1.50 puntos.
- **Nota mínima garantizada:** Todos los grupos que entreguen un trabajo completo y de calidad obtienen hasta **4.50/5.00**.
- **Bonificación competitiva (0.50):** Los **dos mejores grupos** en clasificación (0.25) y los **dos mejores grupos** en regresión (0.25) reciben la bonificación adicional, evaluada como se describe a continuación.

¿Cómo funciona la competencia?

Los profesores entregaremos la **base de datos completa** de cada punto. Ustedes trabajan con toda la base para explorar, limpiar, hacer feature engineering y entrenar modelos. Sin embargo, para la evaluación competitiva, nosotros tomaremos la base original, **la remuestreamos y haremos una nueva partición train/test** que ustedes nunca verán. Luego **re-ejecutaremos su notebook completo** sobre la nueva partición y evaluaremos las predicciones sobre el nuevo test set.

Esto significa que:

1. Su notebook debe ser **completamente reproducible**: desde la carga de datos hasta la exportación del modelo final.
2. Su pipeline de preprocessamiento y feature engineering debe funcionar sobre *cualquier* partición de los mismos datos, no solo la que ustedes usaron.
3. No pueden hacer ajustes manuales ni “hardcodear” decisiones basadas en observaciones específicas.

¿Qué entregar?

Entregable

Cada grupo entrega **un solo archivo comprimido** (.zip) con:

1. **Un notebook** (.ipynb) **por punto**: Contiene **todo** el pipeline — carga de datos, limpieza, feature engineering, entrenamiento, evaluación e interpretación. Debe estar ordenado, comentado y ser ejecutable de principio a fin.
2. **Modelo exportado** (.pk1) para los Puntos 1 y 2: Exporten su mejor modelo final usando joblib o pickle.
3. **Informes ejecutivos** (PDF): Un informe corto por punto, dirigido al cliente del contexto. Sin código, sin fórmulas. Solo resultados, visualizaciones e interpretación.

Atención

Sobre la reproducibilidad: Nosotros tomaremos su .ipynb, cambiaremos la semilla de partición y/o remuestreamos los datos, y lo ejecutaremos de principio a fin. Si su notebook no corre limpiamente, **no podrán participar en la competencia** (pero conservan su nota base). Asegúrense de:

- No usar rutas absolutas de su computador (usen rutas relativas).
- Incluir un requirements.txt si usan librerías no estándar.
- Verificar que el notebook corre en un kernel limpio (*Restart and Run All*).

Punto 1: Clasificación para Focalización de Programas Sociales (1.75 pts)

Contexto del cliente

El Banco Interamericano de Desarrollo (BID) los ha contratado como consultores para mejorar el sistema de focalización de programas de asistencia social en Costa Rica. Actualmente, el gobierno utiliza un modelo de *Proxy Means Test* (PMT) para clasificar hogares según su nivel de vulnerabilidad económica y asignar subsidios. Sin embargo, el modelo vigente presenta problemas de precisión: algunos hogares vulnerables son excluidos del programa (*falsos negativos*) y otros que no lo necesitan reciben el subsidio (*falsos positivos*).

El BID necesita un modelo de clasificación **binaria** que, a partir de características observables del hogar, prediga si un hogar es **pobre** o **no pobre**. Ustedes deben construir la variable objetivo de la siguiente manera:

$$Y_i = \begin{cases} 1 & \text{si el hogar es pobre (Target original } \in \{1, 2\}\} \\ 0 & \text{si el hogar no es pobre (Target original } \in \{3, 4\}\} \end{cases}$$

Nota: Tengan en cuenta que la base está a nivel individuo, por ende, la variable Target puede presentar valores diferentes al interior del mismo hogar. Para asignar la categoría de la variable objetivo al hogar, considere la categoría binaria **más frecuente** al interior del hogar. También, revisen las definiciones de las variables y piensen en cómo agregarlas para construir la información a nivel hogar.

Estructura de costos del programa

El equipo de operaciones del BID les ha compartido la siguiente información sobre los costos asociados al programa de subsidios:

Concepto	Costo (USD)
Subsidio anual por hogar beneficiario	\$1,200
Costo administrativo de incluir un hogar en el programa (verificación, registro, seguimiento)	\$150
Costo estimado de visita de verificación cuando un hogar apela su exclusión	\$80
<i>Cuando un hogar pobre es excluido del programa :</i>	
Pérdida de bienestar estimada por año sin asistencia	\$2,500
Costo de atención de emergencia social posterior (salud, alimentación de crisis)	\$900
Costo político y reputacional para el programa (estimado por caso público)	\$400
<i>Cuando un hogar no pobre recibe el subsidio :</i>	
Subsidio desperdiciado	\$1,200
Costo administrativo irrecuperable	\$150

Adicionalmente, el BID les indica que el presupuesto total del programa es limitado: cada dólar mal asignado es un dólar que no llega a quien lo necesita. Sin embargo, el mandato institucional del BID establece que **la protección de los hogares más vulnerables es la prioridad principal** del programa.

Datos

Utilizarán la base de datos **Costa Rican Household Poverty Level Prediction** del BID, disponible en Kaggle: <https://www.kaggle.com/c/costa-rican-household-poverty-prediction/data>

La base contiene información a nivel individual y de hogar con 143 variables. La variable objetivo original (Target) tiene cuatro categorías que ustedes deben binarizar como se indica arriba. La unidad de análisis es el *hogar*, no el individuo: deberán agregar la información individual a nivel de hogar.

Requerimientos

- a) **Exploración y preprocessamiento (0.15 pts):** Análisis exploratorio. Identifiquen y traten valores faltantes, variables irrelevantes e inconsistencias. Documenten las decisiones de limpieza. Agreguen la información individual a nivel de hogar de forma justificada.
- b) **Feature engineering (0.15 pts):** Construyan al menos 5 variables nuevas a partir de las existentes que consideren relevantes para predecir la pobreza del hogar. Justifiquen cada variable desde el conocimiento económico del problema.
- c) **Elección de la métrica de optimización (0.25 pts):** A partir de la estructura de costos presentada:
- Calculen el **costo total aproximado** de un falso negativo y de un falso positivo. Expliquen su razonamiento.
 - Con base en esa asimetría de costos, **argumenten cuál métrica de clasificación** (accuracy, precision, recall, F1, F_β , u otra) es la más adecuada para optimizar el modelo en este contexto.
 - Discutan: ¿hay costos que la tabla **no incluye** pero que ustedes consideran relevantes? (e.g., costos intangibles, costos intergeneracionales, costos de legitimidad del programa). ¿Cómo cambiaría su elección de métrica si los incluyeran?
- d) **Modelamiento (0.30 pts):** Entrenen al menos tres modelos de clasificación distintos (e.g., Regresión Logística, KNN, SVM, Random Forest, Gradient Boosting). Para cada modelo:
- Realicen *tuning* de hiperparámetros usando cross-validation, **optimizando la métrica que eligieron en el literal anterior**.
 - Reporten: accuracy, precision, recall, F1-score y matriz de confusión.
 - Comparen los modelos usando la métrica elegida y justifiquen la selección del modelo final.
- e) **Interpretabilidad y explicabilidad (0.35 pts):** Para su mejor modelo:
- Identifiquen las variables más importantes para la predicción global.
 - Seleccionen **3 hogares específicos** (uno correctamente clasificado, uno con falso negativo y uno con falso positivo) y expliquen, **a nivel de ese hogar individual**, por qué el modelo tomó esa decisión. ¿Qué variables “empujaron” la predicción hacia una clase u otra?
 - Discutan: si un funcionario del gobierno pregunta “*¿por qué este hogar fue excluido del programa?*”, ¿su modelo puede dar una respuesta satisfactoria?

Importante

No les estamos pidiendo explícitamente una herramienta o librería particular para la explicabilidad. Pero sí esperamos que **puedan explicar decisiones a nivel de observación individual**, no solo importancias globales. Investíguenlo.

- f) **Informe ejecutivo para el BID (0.30 pts):** Máximo 2 páginas dirigido al equipo directivo del BID:
- Resumen del modelo seleccionado y su desempeño en lenguaje accesible.
 - Justificación de por qué la métrica elegida es la correcta para este programa.
 - Visualizaciones claras (no más de 3).
 - Discusión de implicaciones éticas: ¿qué pasa con los hogares mal clasificados? ¿El modelo podría discriminar por alguna variable sensible?
 - Recomendación concreta.
- g) **Competencia (+0.25 pts):** La bonificación competitiva se asigna considerando **dos dimensiones**:
- I. **Elección de la métrica:** Los profesores evaluaremos si la métrica elegida por el grupo es coherente con la estructura de costos del problema. Grupos que elijan una métrica inadecuada (e.g., optimizar accuracy cuando los costos son claramente asimétricos) no podrán acceder a la bonificación, independientemente del desempeño de su modelo.
 - II. **Desempeño del modelo:** Entre los grupos que hayan elegido correctamente la métrica, aquellos 2 que tengan mejor desempeño en dicha métrica (evaluada sobre el test set generado por los profesores) reciben los +0.25 puntos. Ayuda: la métrica debería ser un F_β con $\beta \in \{0.5, 1, 2\}$

Los profesores remuestrearemos la base, haremos una nueva partición y re-ejecutaremos su pipeline completo.

Punto 2: Regresión para Consultoría en Desarrollo Global (1.75 pts)

Contexto del cliente

La Organización Mundial de la Salud (OMS) y el Banco Mundial los contratan como consultores para desarrollar un modelo predictivo de **esperanza de vida al nacer** a nivel de país-año. El objetivo es identificar los factores socioeconómicos, sanitarios e institucionales que mejor predicen la esperanza de vida, y generar un modelo que permita proyectar esta variable para países con datos incompletos o para escenarios contrafactuales de política pública.

El equipo técnico del Banco Mundial necesita un modelo que:

- Prediga con precisión la esperanza de vida a partir de indicadores disponibles.
- Permita entender cuáles indicadores tienen mayor peso predictivo.
- Sea lo suficientemente robusto para generalizar a observaciones país-año no incluidas en el entrenamiento.

Datos

Utilizarán la base de datos **Life Expectancy (WHO)**, disponible en Kaggle:

<https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>

La base contiene datos de 193 países entre 2000 y 2015, con 22 variables que incluyen indicadores de salud (mortalidad infantil, cobertura de inmunización, VIH/SIDA), económicos (PIB per cápita, gasto en salud), educativos (años de escolaridad) y demográficos. La variable objetivo es **Life expectancy (años)**.

Requerimientos

- a) **Exploración y preprocesamiento (0.20 pts):** Análisis exploratorio. Identifiquen patrones, valores faltantes y relaciones entre variables. Documenten su estrategia de imputación y justifiquen por qué es apropiada en el contexto de datos de panel de países.
- b) **Feature engineering (0.15 pts):** Construyan al menos 3 variables nuevas relevantes. Pueden incluir interacciones, transformaciones no lineales, indicadores regionales, o ratios con sentido económico.
- c) **Modelamiento (0.40 pts):** Entrenen al menos tres modelos de regresión distintos (e.g., OLS, Ridge, Lasso, Elastic Net, Random Forest, Gradient Boosting, KNN Regressor). Para cada modelo:
 - Realicen *tuning* de hiperparámetros usando cross-validation.
 - Reporten: RMSE, MAE, R^2 en train y en validación.
 - Analicen los residuales del mejor modelo: ¿hay patrones? ¿Heterocedasticidad? ¿Países con errores sistemáticos?
- d) **Interpretación económica (0.40 pts):** Para su mejor modelo:
 - Identifiquen las 5 variables más importantes y discutan su relevancia desde la teoría del desarrollo económico.
 - Realicen un ejercicio de *análisis contrafactual sencillo*: elijan un país específico y simulen cómo cambiaría la predicción de esperanza de vida si se duplicara el gasto en salud per cápita, manteniendo todo lo demás constante. Interpreten.
 - Expliquen, para al menos 2 países con las peores predicciones (mayor error absoluto), por qué el modelo falló.
- e) **Informe ejecutivo para el Banco Mundial (0.35 pts):** Máximo 2 páginas dirigido al equipo directivo. Debe incluir resumen del modelo, al menos una visualización de predicciones vs. valores reales, discusión de limitaciones y recomendaciones de política.
- f) **Competencia (+0.25 pts):** Los profesores remuestrearemos la base, haremos una nueva partición y re-ejecutaremos su pipeline. La métrica será el **RMSE**. Los dos mejores grupos (menor RMSE) reciben +0.25 puntos.

Punto 3: Segmentación de Clientes para Estrategia Comercial (1.50 pts)

Contexto del cliente

Una fintech latinoamericana de pagos digitales y microcréditos (piensen en Nequi, Nubank, Rappi Pay o similar) los ha contratado como consultores estratégicos. La empresa ha crecido rápidamente y tiene una base grande de usuarios, pero trata a todos sus clientes de la misma forma: misma comunicación, mismas ofertas, mismos productos. El CMO (*Chief Marketing Officer*) quiere implementar una estrategia de **marketing diferenciado** basada en segmentos de clientes para personalizar campañas, diseñar ofertas adecuadas, reducir churn y priorizar esfuerzos de retención.

Datos

Utilizarán la base de datos “segmentación.csv”,

La base contiene información demográfica y comportamental de clientes: edad, género, ingreso anual, años como miembro, frecuencia de compra, monto de última compra, puntaje de gasto, categoría preferida, entre otros.

Requerimientos

- a) **Exploración y preprocesamiento (0.15 pts):** Análisis exploratorio. Estandarización, tratamiento de outliers y codificación de variables categóricas. Justifiquen cada decisión.
- b) **Determinación del número de segmentos (0.25 pts):** Utilicen al menos dos métodos para determinar el número óptimo de clusters (e.g., método del codo, coeficiente de silueta, Gap statistic). Presenten los resultados gráficamente y justifiquen su elección final. No se trata solo de “lo que diga el codo”: argumenten por qué ese número tiene sentido desde la estrategia de negocio.
- c) **Modelamiento (0.25 pts):** Implementen al menos dos algoritmos de clustering (e.g., K-Means, DBSCAN, Gaussian Mixture Models, Clustering Jerárquico). Comparen y seleccionen el que mejor capture la estructura. Si usan reducción de dimensionalidad (PCA, t-SNE) para visualizar, documéntenlo.
- d) **Perfilamiento de segmentos (0.35 pts):** Para cada segmento:
 - Describan su perfil en lenguaje de negocio.
 - Asígnenle un **nombre comercial** descriptivo (e.g., “Jóvenes digitales de alto valor”, “Clientes dormidos”).
 - Calculen métricas clave por segmento.
 - Visualicen los segmentos en al menos 2 gráficos informativos.
- e) **Propuesta comercial (0.50 pts):** Para cada segmento, propongan:
 - **Producto/oferta recomendada:** ¿Qué producto financiero o servicio le ofrecerían?
 - **Canal de comunicación:** ¿Cómo les hablan? (push notification, email, llamada, redes sociales).
 - **Tono y mensaje:** Un ejemplo concreto de mensaje de marketing (1-2 oraciones).
 - **Estrategia de retención:** Para los segmentos en riesgo de churn, propongan una acción específica.

Presenten esta propuesta como un “playbook” comercial que el CMO pueda llevar directamente a su equipo.

Entregable

Formato del Punto 3: El informe ejecutivo de este punto debe tener formato de *presentación de consultoría*: máximo 4 páginas o slides exportadas a PDF. Visualmente claro, con gráficos y tablas-resumen, escrito para una audiencia de negocio.

Criterios Transversales de Evaluación

Criterio	Descripción
Rigor técnico	Correcta implementación de modelos, validación cruzada, métricas apropiadas. Código limpio y reproducible.
Conocimiento del dominio	Las decisiones de modelamiento reflejan comprensión del contexto económico, social o de negocio.
Comunicación	Los informes ejecutivos son claros, concisos y dirigidos a la audiencia correcta.
IA Responsable	Se discuten implicaciones éticas, sesgos potenciales y limitaciones del modelo.
Creatividad	Feature engineering original, visualizaciones informativas, propuestas innovadoras.

Especificaciones Técnicas

- **Lenguaje:** Python (recomendado: scikit-learn, pandas, matplotlib/seaborn).
- **Formato del notebook:** .ipynb. Debe ser ejecutable de principio a fin en un kernel limpio.
- **Modelo exportado:** Archivo .pkl generado con joblib o pickle (Puntos 1 y 2).
- **Informes ejecutivos:** PDF.
- **Entrega:** Vía Bloque Neón o el medio indicado. **Un solo envío por grupo** en formato .zip.
- **Estructura del .zip:**

```
Grupo_XX/
Punto1_Clasicacion.ipynb
Punto1_modelo.pkl
Punto1_Informe_BID.pdf
Punto2_Regresion.ipynb
Punto2_modelo.pkl
Punto2_Informe_BancoMundial.pdf
Punto3_Segmentacion.ipynb
Punto3_Playbook_Comercial.pdf
requirements.txt
```