💼 Rúbrica Detallada de Evaluación: Decision Intelligence Capstone

Trabajo Práctico Integrador — EAyMR 2026

Cátedra: Estadística Aplicada y Modelos de Regresión (2026)
Institución: Pontificia Universidad Católica Argentina (UCA) — Facultad de Ingeniería y Ciencias Agrarias
Carreras: Licenciatura en Ciencia de Datos / Ingeniería en Inteligencia Artificial (2º Año – 2º Semestre)
Profesor Titular / Responsable: Prof. Ivan Gonzalo Vilaboa
Ponderación en Cursada: 3.0 Puntos Oficiales (30% de la nota final de cursada). Evaluado sobre una escala de 100 puntos en esta rúbrica analítica.

Esta rúbrica constituye la matriz oficial de evaluación del Trabajo Práctico Integrador (TPI). Establece los estándares analíticos, metodológicos y de ingeniería de software requeridos para la aprobación y excelencia del proyecto final.

📊 Resumen de Ponderación General (100 puntos totales → 3.0 pts en Cursada)

Dimensión Evaluada	Ponderación	Núcleo de Competencias Evaluadas
Dimensión 1: Inferencia Clásica, EDA y ANOVA	**15 pts**	Formulación de contrastes, EDA profundo, diagnóstico de normalidad/ homocedasticidad en muestras masivas, ANOVA/Welch y post-hoc.
Dimensión 2: Preparación de Datos, Gauss-Markov y PCA	**15 pts**	Diagnóstico de residuos MCO, outliers influyentes (Cook's D), VIF y PCA estricto sin data leakage (Train/ Test).
Dimensión 3: Torneo Predictivo y Machine Learning	**20 pts**	Modelos lineales regularizados, Naive Bayes, Árbol CART con poda ccp_α, Random Forest con K-Fold CV, calibración y MDI vs Permutación.
Dimensión 4: Inferencia Causal y Causal ML	**25 pts**	DAG estructural (Unidad 6), criterio Backdoor, DoWhy (ATE por IPW y regresión), Double ML y CATE en EconML, refutaciones y ensayo del choque causal vs ML.
Dimensión 5: Metodología Ágil (Scrum) y Bitácora	**10 pts**	Dinámica de equipo Scrum con Quality Gate determinístico de QA, y las 5 entradas obligatorias de `BITACORA.md` con autopsia de experimentos fallidos.
	15 pts	Web App interactiva desplegada en la nube (Streamlit/Gradio/

Dimensión Evaluada	Ponderación	Núcleo de Competencias Evaluadas
Dimensión 6: Deploy en Producción, Código y Defensa		Docker), reproducibilidad del repositorio GitHub y solvencia técnica en la defensa oral.

🎯 Matriz Analítica por Niveles de Desempeño

1. Inferencia Estadística Clásica, EDA y ANOVA (15 Puntos)

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
Sobresaliente	**14 - 15 pts**	**Diseño experimental riguroso e interpretación impecable:** Formulación matemática explícita de al menos dos hipótesis de negocio (H_0 vs H_1). Diagnóstico sofisticado de supuestos en muestras grandes ($N \geq 1.500$): reconoce la hiper-sensibilidad de Shapiro-Wilk y fundamenta la normalidad mediante asimetría, curtosis y Q-Q plots. Aplica test de Levene; ante heterocedasticidad, implementa correctamente Welch ANOVA o Kruskal-Wallis. Contrastes post-hoc (Tukey HSD) justificando el control del error tipo I por familia. Magnitudes de efecto (Cohen's d) reportadas.
Notable	**11 - 13 pts**	**Análisis correcto con detalles menores:** Formula hipótesis claras y ejecuta tests paramétricos/no paramétricos adecuados. Valida normalidad y homocedasticidad, pero recurre mecánicamente a tests formales sin discutir el impacto del tamaño muestral sobre el p-valor. Aplica

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
		ANOVA y comparaciones post-hoc de manera correcta con leves omisiones en la interpretación sustantiva de los intervalos.
Aprobado	**9 - 10 pts**	**Ejecución básica sin profundidad diagnóstica:** Plantea hipótesis estadísticas elementales. Aplica ANOVA pero omite verificar supuestos de homocedasticidad (Levene) o ignora la presencia de distribuciones altamente asimétricas. Ejecuta comparaciones post-hoc sin justificar la corrección de comparaciones múltiples.
Insuficiente	**0 - 8 pts**	**Errores conceptuales o ausencia de inferencia:** Omite contrastes de hipótesis o confunde la hipótesis nula con la alternativa. Aplica tests estadísticos inapropiados para la naturaleza de los datos. No realiza análisis de varianza o presenta resultados sin conclusiones ni validación de supuestos.

2. Preparación de Datos, Supuestos de Gauss-Markov y PCA (15 Puntos)

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
Sobresaliente	**14 - 15 pts**	**Diagnóstico econométrico exhaustivo y PCA libre de fuga de datos:** Ajuste del modelo MCO de referencia con reporte completo (R^2, estadístico F, p-valores). Diagnóstico formal de Gauss-Markov: Breusch-Pagan/White para heterocedasticidad, Jarque-Bera y Durbin-Watson. Distingue con precisión analítica entre residuos estudentizados, apalancamiento (h_{ii}) y observaciones influyentes mediante Distancia de Cook ($D_i > 4/n$). Tratamiento fundado de multicolinealidad con VIF. **PCA metodológicamente perfecto:** estandarización y proyección ajustadas exclusivamente sobre el subconjunto de entrenamiento (Train) y aplicadas a Test sin data leakage. Scree plot, regla del codo/Kaiser e interpretación conceptual clara de las cargas (loadings) de los componentes retenidos.
Notable	**11 - 13 pts**	**Diagnóstico sólido y reducción adecuada:** Ejecuta análisis de

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
		residuos, calcula Cook's D y VIF identificando variables colineales. Implementa PCA correctamente explicando la varianza acumulada, aunque la interpretación semántica de los autovectores es superficial o la separación Train/Test en el pipeline de PCA presenta pequeñas inconsistencias documentales.
Aprobado	**9 - 10 pts**	**Tratamiento estándar pero superficial:** Calcula residuos de regresión lineal pero no aplica tests formales de heterocedasticidad. Elimina outliers de forma indiscriminada sin calcular Distancia de Cook ni leverage. Aplica PCA a todo el dataset antes de la partición (incurriendo en data leakage leve) o no interpreta los componentes resultantes.
Insuficiente	**0 - 8 pts**	**Omisión de diagnósticos o errores graves:** No valida supuestos de regresión. Ignora la presencia de multicolinealidad severa ($\text{VIF} > 10$). Aplica transformaciones arbitrarias sin justificación o no

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
		implementa reducción de dimensionalidad por PCA.

3. Torneo de Modelos Predictivos, Machine Learning y Calibración (20 Puntos)

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
Sobresaliente	**18 - 20 pts**	**Arquitectura de Machine Learning de vanguardia y rigor comparativo:** Partición estricta Train (80%) y Test (20%) con validación cruzada K-Fold (o Stratified K-Fold) en Train. Ajusta y compara rigurosamente las 4 familias: 1) Lineal regularizado (Ridge/ Lasso con senda de penalización o Logística con Odds Ratios y CIs); 2) Naive Bayes evaluando el impacto del supuesto de independencia; 3) Árbol CART individual implementando poda por complejidad de costo (Cost-Complexity Pruning con ccp_α óptimo obtenido por CV); 4) Random Forest optimizado vía GridSearchCV/ RandomizedSearchCV con reporte del error Out-of-Bag (OOB). Matriz comparativa exhaustiva en Test (R^2, RMSE, MAE o AUC-ROC, PR-AUC, F1, Brier Score). Curvas de calibración probabilística (reliability diagrams). **Análisis crítico de importancia:** expone matemáticamente el sesgo de MDI (Gini) hacia variables

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
		continuas y demuestra cómo Permutation Feature Importance en Test proporciona una jerarquía insesgada.
Notable	**15 - 17 pts**	**Comparación multimodelo competente:** Entrena las 4 familias de algoritmos con K-Fold CV y evalúa métricas en Test de forma apropiada. Poda el árbol de decisión y optimiza hiperparámetros en Random Forest. Extrae Feature Importance, aunque la comparación entre MDI y Permutation Importance carece de profundidad analítica sobre los sesgos de cardinalidad o se omiten las curvas de calibración probabilística.
Aprobado	**12 - 14 pts**	**Modelado básico con validación limitada:** Compara al menos 2 familias de modelos pero utiliza división simple Train/Test sin validación cruzada sistemática. No realiza poda por complejidad del árbol o deja que sobreajuste. Extrae solo la importancia de variables por defecto (MDI) sin cuestionar sus limitaciones.
Insuficiente	**0 - 11 pts**	

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
		Sobreajuste evidente o diseño experimental defectuoso: Entrena un único modelo sin partición adecuada o evalúa en los mismos datos de entrenamiento. Mezcla datos de Test en la optimización de hiperparámetros. No reporta métricas de generalización comparativas.

4. Inferencia Causal, Causal ML y Choque Metodológico (25 Puntos)

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
Sobresaliente	**23 - 25 pts**	**Maestría en causalidad estructural y modelado Causal ML:** Formalización teórica impecable del DAG en DOT/NetworkX según lo aprendido en la Unidad 6, con identificación explícita de tratamiento T, resultado Y, confusores Z, mediadores M y colisionadores C. Demuestra analíticamente el cumplimiento del Criterio Backdoor y explica la trampa de condicionar en mediadores o colisionadores. Flujo en 4 pasos con `DoWhy`: estimación de ATE mediante regresión ajustada y Propensity Score Weighting (IPW), contrastando la magnitud del sesgo frente a la regresión ingenua. Implementación de **Double Machine Learning (DML)** con `EconML` utilizando Random Forest como modelos auxiliares para estimar el **CATE** y graficar la heterogeneidad de efectos por subgrupos. Batería completa de refutaciones superadas (Random Common

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
		Cause, Placebo Treatment, Data Subset). **Ensayo del Choque Causal vs Predictivo:** discusión lúcida y madura sobre por qué la variable más predictiva de ML puede ser inútil o peligrosa como palanca causal de intervención.
Notable	**19 - 22 pts**	**Modelado causal sólido y refutaciones completas:** Construye un DAG plausible y aplica el criterio Backdoor en DoWhy. Estima ATE por al menos un método y aplica EconML para estimar CATE con leves errores en la interpretación de los estimadores ortogonales. Supera las pruebas de refutación estocástica y elabora una reflexión crítica adecuada sobre correlación vs causalidad.
Aprobado	**15 - 18 pts**	**Comprensión causal elemental pero procedimental:** Construye un DAG simplificado sin justificar suficientemente los arcos. Ejecuta DoWhy de forma mecánica sin comparar con la regresión ingenua. No implementa Double ML o no logra graficar la heterogeneidad de

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
		efectos (CATE). Ejecuta solo una prueba de refutación.
Insuficiente	**0 - 14 pts**	**Confusión severa entre asociación y causalidad:** No formula un DAG causal o comete el error grave de condicionar sobre mediadores/ colisionadores sin advertirlo. Interpreta coeficientes de regresión predictiva como si fueran efectos causales directos. No utiliza librerías de inferencia causal o falla en la ejecución de DoWhy.

5. Metodología de Equipo Ágil (Scrum) y Bitácora de Ingeniería (10 Puntos)

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
Sobresaliente	**9 - 10 pts**	**Cultura de ingeniería ágil y trazabilidad viva:** Organización explícita del equipo bajo los 4 roles Scrum (Project Lead/Product Owner, Backend & ML Engineer, Frontend & UX Engineer, QA & Validation Engineer). Evidencia de un protocolo de calidad estricto (**Quality Gate determinístico** con tests unitarios `pytest` y linters antes de integrar a main). Archivo `BITACORA.md` ejemplar con las **5 entradas completas**: marco estratégico, matriz de supuestos y decisiones estadísticas justificadas, **autopsia de al menos dos experimentos fallidos**, formalización del debate causal, y **trazabilidad ética del uso de asistentes de IA** (prompts, alucinaciones detectadas y tests para neutralizarlas).
Notable	**8 pts**	**Estructura ágil clara y bitácora detallada:** Asignación visible de roles y evidencia de revisiones de código cruzadas. `BITACORA.md` completa con la mayoría de las secciones desarrolladas, documentando decisiones metodológicas y al

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
		menos un experimento fallido, aunque la trazabilidad de prompts o los tests de QA son breves.
Aprobado	**6 - 7 pts**	**Cumplimiento parcial de la metodología:** Roles declarados en el repositorio pero con poca evidencia de dinámicas de revisión o control de calidad de QA. `BITACORA.md` presente pero tratada como un resumen tardío del trabajo en lugar de un diario vivo de experimentación; omite la autopsia de experimentos fallidos.
Insuficiente	**0 - 5 pts**	**Ausencia de bitácora o dinámica desorganizada:** `BITACORA.md` vacía, inexistente o copiada del enunciado. No se definen roles de equipo ni se evidencia ningún criterio de aseguramiento de calidad o reproducibilidad en el desarrollo.

6. Deploy en Producción, Código y Defensa Oral (15 Puntos)

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
Sobresaliente	**14 - 15 pts**	**Producto de software en producción y oratoria ejecutiva de alto impacto: Web App desplegada en vivo y accesible públicamente** (Streamlit Cloud, Render, Hugging Face o Docker funcional) con los tres módulos operativos: diagnóstico exploratorio/PCA interactivo, simulador predictivo con intervalos y simulador causal de políticas ($do(T = t)$) con visualización de CATE y semáforo de refutaciones. Repositorio GitHub profesional con estructura limpia (`/src` , `/app` , `/tests`), `README.md` exhaustivo y gestión con `uv` / `requirements.txt` . **Defensa oral brillante:** exposición ajustada a 12-15 minutos, diapositivas ejecutivas impecables (10-12 slides PDF), demostración fluida en vivo y dominio técnico individual sobresaliente de cada integrante al responder las preguntas del tribunal.
Notable	**11 - 13 pts**	**Despliegue operativo y sólida defensa oral:** Aplicación web funcionando en la nube con simulador predictivo y causal básico. Repositorio bien organizado y reproducible. Exposición oral clara con buen uso del tiempo y soporte visual, respondiendo satisfactoriamente a las preguntas del tribunal docente con buen dominio conceptual.
Aprobado	**9 - 10 pts**	**Despliegue local con deficiencias o defensa titubeante:** La aplicación web solo corre localmente (no está desplegada en la nube) o presenta errores de interfaz/reactividad. Repositorio ejecutable pero desordenado o con dependencias no fijadas. Exposición oral aceptable pero

Nivel	Rango	Criterios de Evaluación y Evidencias Observables
		con desbalances evidentes de participación entre los integrantes.
Insuficiente	**0 - 8 pts**	**Proyecto no ejecutable y defensa deficiente:** No entrega aplicación web ni deploy. Repositorio no reproducible (código roto, dependencias faltantes). Exposición oral desorganizada, lectura literal de diapositivas o incapacidad de responder a preguntas conceptuales elementales de la materia.

⚖️ Conversión de Puntajes y Calificación Final de Cursada

El puntaje obtenido en esta rúbrica (sobre 100 puntos) se convierte directamente a la escala oficial de la asignatura:

$$\text{Puntos Obtenidos en la Cursada (máx. 3.0)} = \frac{\text{Puntaje Total Obtenido (máx. 100)}}{100} \times 3.0$$

Puntaje Rúbrica TPI	Puntos en Cursada (máx. 3.0)	Nivel Académico Institucional
90 a 100 pts	**2.7 a 3.0 pts**	**Sobresaliente / Distinción:** Proyecto de excelencia profesional apto para portfolio técnico.
75 a 89 pts	**2.25 a 2.65 pts**	**Notable / Muy Bueno:** Dominio sólido de todos los ejes teóricos y computacionales.
60 a 74 pts	**1.8 a 2.2 pts**	**Aprobado / Suficiente:** Cumple satisfactoriamente con los requisitos mínimos de la materia.
< 60 pts	**< 1.8 pts**	**Insuficiente:** No alcanza los estándares mínimos; debe reelaborar para instancia de Recuperatorio.