Aplicaciones de la Regresión Logística para la Predicción de Ingresos Superiores a $25,000 Pesos: Un Análisis de Factores Demográficos y Laborales

Joseph Shakalo Paz A01784107, Miguel Ángel Noriega Bedolla A01658032 y Gabriel Edid Harari A01782146, estudiantes del curso *Inteligencia artificial avanzada para la ciencia de datos I* en el ITESM.

Resumen - Se construyó y evaluó un modelo de regresión logística para estimar la probabilidad de ingresos mensuales mayores a $25,000 usando datos capturados con *Google Forms*. El flujo aplica limpieza de texto, imputación de valores faltantes, normalización de variables numéricas, codificación One-Hot en categóricas y selección de variables con importancia por permutación. El umbral de decisión se ajustó dentro de 0.05 a 0.95 para maximizar F1 mediante validación cruzada. En la prueba del 20 por ciento el modelo obtuvo AUC 0.881, F1 0.588, *recall* 0.833, *precision* 0.455 y *accuracy* 0.741. En validación cruzada *5-Fold* promedió AUC 0.953 y F1 0.776 con variación moderada. Los coeficientes muestran mayor probabilidad asociada a niveles educativos altos, más horas por semana y ciertas ocupaciones y tipos de empleo. El modelo puede servir como filtro inicial para priorizar contactos en becas y capacitación, apoyar decisiones de reclutamiento y explorar escenarios educativos simples. Sus límites son el tamaño de muestra y la clase positiva minoritaria. Para uso aplicado se sugiere ampliar datos, calibrar probabilidades y ajustar el umbral según costo de errores.

**Índice de Términos - Clasificación binaria, ROC-AUC, F1, *Cross-validation*, Ingresos, *Machine learning*, Preprocesamiento de datos, Python, Regresión logística.**

# introducción

La regresión logística es un modelo estadístico ampliamente usado en problemas de clasificación binaria. A diferencia de la regresión lineal, cuyo objetivo es predecir valores continuos, la regresión logística transforma la salida en una probabilidad entre 0 y 1 mediante la función sigmoide. Esto permite tomar decisiones al asignar una clase según un umbral definido, lo que la hace adecuada para problemas donde solo existen dos categorías de respuesta.

En este trabajo se utiliza la regresión logística para predecir si una persona percibe un sueldo mensual mayor a $25,000 pesos o si se encuentra en el grupo de ingresos iguales o menores. La información utilizada proviene de un formulario que recopila variables como edad, nivel educativo, tipo de empleo, estado civil, sector económico, horas trabajadas y otros datos demográficos. Estas variables sirven como insumo para entrenar y evaluar el modelo.

El flujo seguido incluye limpieza de datos, imputación de valores faltantes, codificación de variables categóricas mediante One-Hot Encoding, estandarización de variables numéricas y validación cruzada para garantizar un desempeño estable. También se calcula un umbral de decisión que maximiza la métrica F1, buscando un equilibrio entre precisión y exhaustividad en la clasificación.

El objetivo de este reporte es documentar de forma clara y reproducible cómo un modelo de regresión logística puede identificar patrones relacionados con el nivel de ingresos. Si bien no se busca una aplicación directa en producción, los resultados permiten visualizar su utilidad en contextos como estudios laborales, planeación educativa y análisis de equidad salarial.

# Marco teórico

En ejercicios previos de clasificación binaria, la regresión logística se había aplicado en escenarios sencillos, con pocas variables y sin procesos avanzados de limpieza o validación. Estos trabajos iniciales ayudaron a comprender la función sigmoide y la interpretación de coeficientes, pero no permitían evaluar el modelo en situaciones más complejas con múltiples predictores y datos heterogéneos.

En este proyecto se amplía ese enfoque aplicando regresión logística a un conjunto de datos con variables demográficas, educativas y laborales. El modelo estima la probabilidad de pertenecer a la clase positiva mediante la función logística:

donde 𝜃 es el vector de parámetros y 𝑋 el vector de características. La salida es una probabilidad entre 0 y 1 que se convierte en clase al fijar un umbral.

El flujo metodológico integra limpieza de datos, imputación de valores faltantes, One-Hot Encoding, estandarización y validación cruzada. Además, se ajusta un umbral de decisión que maximiza la métrica F1. Estos elementos hacen que el modelo sea más robusto y cercano a un escenario real.

Gracias a su simplicidad e interpretabilidad, la regresión logística sigue siendo un modelo de referencia en predicción de ingresos y en aplicaciones como riesgo crediticio, estudios médicos o análisis laboral.

# Metodología

## Captura de datos: La recolección de datos se realizó mediante un formulario en Google Forms diseñado para recopilar información sociodemográfica y laboral de los participantes. Entre las variables recolectadas se encuentran edad, nivel educativo, estado civil, ocupación, sector económico, horas trabajadas por semana, género, país de origen y relación en el hogar. Las respuestas fueron exportadas en formato CSV para su posterior procesamiento.

**Tabla 1**

5 muestras aleatorias del conjunto de datos.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Edad | Empleo | Edu. | Civil | Ocup. | Rel. | Raza | Gen. | País | Hrs | Ingresos |
| 38 | Privado | Maestría | Casado | Agro | Esposa | Blanco. | Fem. | México | 48 | >25k |
| 25 | Propia (c/emp) | Licenc. | Casado | Agro | Hijo | Mestizo | Masc. | México | 52 | ≤25k |
| 43 | Privado | Prep. | Soltero | Salud | Hijo | Afrodesc. | Fem. | México | 43 | ≤25k |
| 44 | Otros | Licenc. | Casado | Industria | Esposa | M. Oriente | Fem. | México | 43 | ≤25k |
| 18 | Propia (s/emp) | Sec/Bach | Soltero | Gobierno | Hijo | M. Oriente | Fem. | México | 53 | ≤25k |

## Preparación de los datos: El flujo metodológico inicia con la carga de los datos y la definición de la variable objetivo como binaria, donde la clase positiva corresponde a ingresos superiores al umbral establecido ($25,000). Se contemplaron diferencias de codificación y se normalizaron valores faltantes o ambiguos para asegurar consistencia en el análisis.

## Preprocesamiento: El preprocesamiento incluyó la normalización de texto (eliminación de acentos, espacios y símbolos inconsistentes) y la estandarización de categorías. En particular, la variable de ocupación se agrupó en un conjunto reducido de sectores, con la categoría “Otros” para valores fuera del dominio. Este paso redujo la variabilidad y mejoró la calidad de las entradas al modelo.

## Ingeniería de variables: Las variables numéricas se completaron con la mediana en caso de valores faltantes y después se normalizaron a una misma escala (media cero y varianza uno). Las categóricas se completaron con la moda y se transformaron mediante One-Hot Encoding. Estas transformaciones se integraron en un pipeline que asegura consistencia entre las fases de entrenamiento y producción.

## Selección de características: La selección de variables se hizo usando el método de importancia por permutación. Se midió cuánto aportaba cada columna al ROC-AUC del modelo y las que no sumaban valor o afectaban negativamente se eliminaron para simplificar y evitar ruido.

## Entrenamiento y ajuste del umbral: El clasificador no utilizó el umbral estándar de 0.5, sino que se optimizó dentro de un rango predefinido (0.05 a 0.95) para maximizar la métrica F1 en validación cruzada. Este umbral óptimo se aplicó tanto en la evaluación como en la producción.

## Validación del modelo: El entrenamiento se evaluó en dos fases. Primero, se usó una división en entrenamiento y prueba que mantiene la misma proporción de clases en ambos conjuntos (holdout estratificada, 80/20). Esta evaluación proporcionó métricas agregadas como Accuracy, Precision, Recall, F1 y ROC-AUC, junto con curvas PR (Precision-Recall) y ROC (Receiver Operating Characteristic) y matriz de confusión. Segundo, una validación cruzada externa reportó la variabilidad de F1 y ROC-AUC entre pliegues, midiendo la estabilidad del modelo ante re-muestreo.

## Generación de artefactos: El modelo final se reentrenó con todo el conjunto de datos y se guardó junto con un archivo de metadatos que documenta columnas utilizadas, el umbral óptimo y las métricas alcanzadas. Estos artefactos constituyen la entrega reproducible del sistema.

## Servicio de producción: El sistema de producción carga el modelo entrenado y expone una interfaz ligera donde el usuario puede introducir variables de entrada, recibir la probabilidad de pertenecer a la clase positiva y obtener la clasificación binaria resultante. La interfaz se adapta dinámicamente al esquema de datos para evitar inconsistencias cuando las características cambian.

## Reproducibilidad: La reproducibilidad se aseguró al incluir todo el preprocesamiento en un pipeline, registrar configuraciones y métricas en archivos de metadatos, controlar dependencias y usar semillas fijas con validación estratificada. Esto hace posible repetir los resultados en distintos entornos con poca variación.

# Resultados

## Selección de características: La importancia por permutación (métrica AUC) **descartó:** Estado civil, Relación en el hogar, País de origen, Raza. **Se usaron:** Edad, Tipo de empleo, Nivel educativo más alto, Ocupación, Género, Horas trabajadas por semana.

## Desempeño en conjunto de prueba (20%): Evaluación en la partición 80/20 estratificada. **Métricas:** Accuracy: 0.7407, Precision: 0.4545, Recall: 0.8333, F1: 0.5882, ROC-AUC: 0.8810. Matriz de confusión: [[15, 6], [1, 5]].

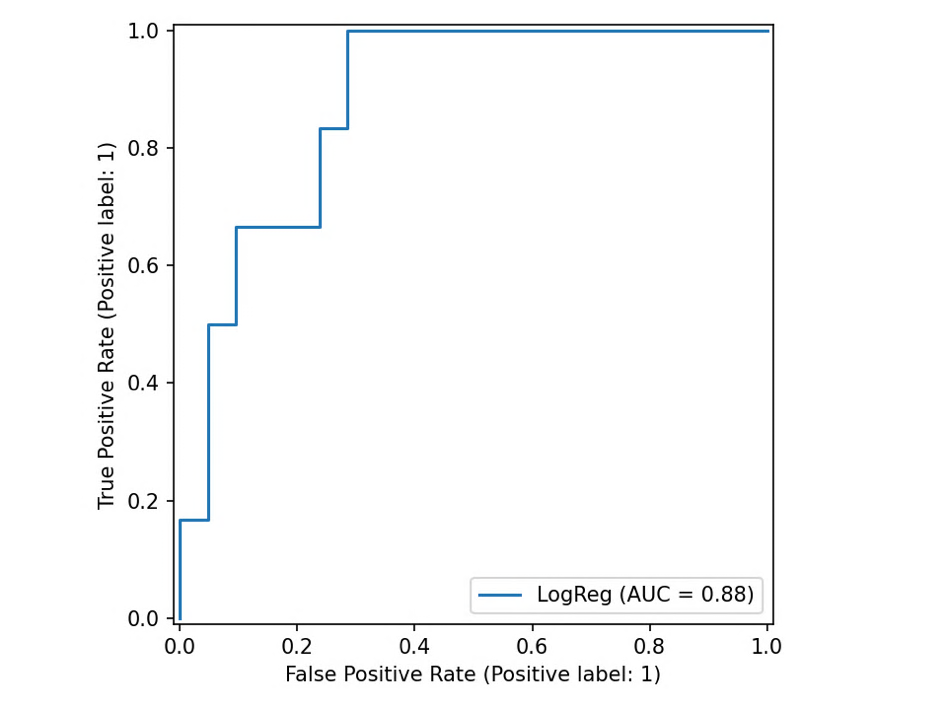


Fig. 1. Curva ROC en prueba (*holdout*). Eje X: FPR. Eje Y: TPR. AUC = 0.881.

A graph with blue lines

AI-generated content may be incorrect.

Fig. 2. Curva Precision–Recall en prueba (*holdout*). Eje X: Recall. Eje Y: Precision. AP = 0.68.

A chart of blue yellow and purple squares

AI-generated content may be incorrect.

Fig. 3. Matriz de confusión en prueba (*holdout*). TN = 15, FP = 6, FN = 1, TP = 5.

## Validación cruzada (5-Fold, umbral fijo): Evaluación con la partición cruzada: **Métricas promedio:** Accuracy: **0.8781**, Precision: **0.7005**, Recall: **0.9000**, F1: **0.7758**, ROC-AUC: **0.9533**.

A graph of blue bars

AI-generated content may be incorrect.

Fig. 4. F1 por fold en validación cruzada *5-Fold*. F1 = [0.7500, 0.9231, 0.8333, 0.7059, 0.6667]. Promedio = 0.7758.

A graph of blue bars

AI-generated content may be incorrect.

Fig. 5. ROC-AUC por fold en validación cruzada *5-Fold*. AUC = [1.0000, 0.9917, 0.9417, 0.9417, 0.8917]. Promedio = 0.9533.

## Coeficientes del modelo: Los coeficientes indican cómo cambia la probabilidad estimada: signo + aumenta la probabilidad de >25k, signo − la reduce. Mayor |coef| = efecto más fuerte:

## **Top 15 por |coef|:**

## Educación: **Maestría** → **+1.9892**

## Educación: **Preparatoria** → **−1.6203**

## Ocupación: **Tecnología y serv. profesionales** → **+1.3212**

## Empleo: **Propia (con empresa)** → **+1.2668**

## **Horas/semana** → **+1.2477**

## Educación: **Secundaria/Bachillerato** → **−1.2206**

## Educación: **Doctorado** → **+1.1599**

## Educación: **Primaria** → **−1.0604**

## Ocupación: **Comercio y logística** → **−0.9628**

## Empleo: **Propia (sin empresa)** → **−0.9527**

## Educación: **Universidad/Licenciatura** → **+0.9274**

## Ocupación: **Agro y alimentos** → **−0.8158**

## Ocupación: **Salud y serv. sociales** → **+0.7723**

## Género: **Prefiero no decir** → **+0.7398**

## Empleo: **Trabajo no remunerado** → **−0.7302**

# Análisis de resultados

En la prueba del 20% el modelo muestra una discriminación sólida (ROC-AUC = 0.881). El umbral se ajustó para maximizar F1 y eso se refleja en un *recall* alto (0.833) y una *precision* menor (0.455). La matriz de confusión muestra 6 falsos positivos y 1 falso negativo. En general, el sistema privilegia detectar la clase positiva, aunque incremente las alertas falsas.

Para estimar estabilidad se aplicó validación cruzada *5-Fold* con el umbral fijo. Los promedios son superiores a los del *holdout* (ROC-AUC = 0.953, F1 = 0.776) y la variación entre *folds* es moderada. La brecha frente al 20% se interpreta como efecto del tamaño reducido del conjunto de prueba más que como sobreajuste.

La selección de variables por importancia de permutación elimino variables con aporte no positivo al AUC y dejó un conjunto más pequeño orientado a lo laboral y educativo: edad, tipo de empleo, nivel educativo, ocupación, género y horas por semana. Esto redujo complejidad y ruido sin perder la capacidad para discriminar.

Las variables en el set de datos transformado siguen patrones esperados. Niveles educativos altos y más horas trabajadas se asocian con mayor probabilidad de superar $25,000. Sectores de tecnología y el trabajo propio con empresa suman, mientras comercio-logística, agro y lo propio sin empresa restan. Estos son números que el modelo usa para sumar o restar ‘puntos’. Un número alto no significa que esa variable cause el resultado.

El estudio tiene pocos datos (131 casos) y muy pocos positivos en la prueba. Eso hace que la precisión cambie mucho y que haya más incertidumbre en los resultados. Si equivocarse con un falso positivo cuesta caro, conviene ajustar el umbral o calibrar las probabilidades antes de decidir. También es útil revisar el desempeño en diferentes grupos y conseguir más datos para hacer las estimaciones más firmes.

# Conclusiones

El modelo de regresión logística cumple el objetivo de estimar la probabilidad de ganar más de $25,000 con datos sencillos. En la prueba del 20 por ciento separa bien las clases y privilegia detectar los casos positivos. Esto genera más alertas falsas, pero reduce el riesgo de dejar pasar personas que podrían estar por arriba del umbral.

Las variables con mayor aporte fueron nivel educativo, horas trabajadas por semana, tipo de empleo y ocupación. Algunas variables demográficas no aportaron y se retiraron durante la selección. El resultado es un modelo compacto y fácil de operar.

Este modelo puede usarse como filtro inicial para priorizar a quién contactar primero en campañas de becas, capacitación o seguimiento. También sirve para apoyar decisiones de reclutamiento y rangos salariales al comparar perfiles de manera consistente. En educación ayuda a explorar escenarios simples, por ejemplo, cómo cambia la probabilidad estimada cuando aumenta el nivel de estudios o las horas trabajadas.

Existen límites claros. El conjunto de datos es pequeño y la clase positiva es minoritaria, por lo que la precisión puede variar entre muestras. Las probabilidades pueden requerir calibración y el costo de los errores depende del contexto. Antes de cualquier uso operativo se deben revisar aspectos de equidad, privacidad y cumplimiento normativo.

Hay mejoras directas. Recolectar más datos y simplificar categorías poco frecuentes. Ajustar el umbral de decisión según el costo de los errores del caso de uso. Calibrar las probabilidades para que reflejen mejor la realidad. Documentar el modelo con una ficha simple y monitorear su desempeño con datos nuevos. En síntesis, el modelo es un buen punto de partida para ordenar y priorizar casos con reglas claras y con controles adecuados.

Apéndice A. Material suplementario y código fuente

1. Formulario de captura de datos: <https://forms.gle/txaRy4QaqzztcssN9>
2. Repositorio con todo el código utilizado:

<https://github.com/GabrielEdid/MiniReto_BloqueIA>

Referencias

1. GeeksforGeeks, “Logistic Regression in Machine Learning,” GeeksforGeeks, 2024. [En línea]. Disponible en: https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/

**Reflexiones Finales**

**Joseph:** En este proyecto, construimos y pusimos en producción un sistema de aprendizaje automático de punta a punta: desde la preparación de datos y la validación hasta un servicio web que entrega predicciones de forma estable. Como equipo, confirmamos que la ingeniería de features y el ajuste del umbral marcan la diferencia práctica; que métricas como F1/PR son más útiles que la accuracy en escenarios desbalanceados; y que la disciplina operativa (contratos de columnas, reproducibilidad y monitoreo básico) convierte el trabajo en una solución confiable para usuarios reales. Reconocemos el reto de sesgos y drift, y nuestro siguiente paso es fortalecer interpretabilidad y monitoreo, explorar modelos complementarios y seguir iterando para entregar valor consistente a nuestros stakeholders.

**Miguel:** Durante este proyecto entendí de verdad qué significa construir un modelo “de extremo a extremo”. Pasé de tener respuestas de un Google Forms a un pipeline reproducible: limpiar, imputar, estandarizar y one-hot antes de siquiera pensar en entrenar. Elegir Regresión Logística dejó de ser solo teoría: vi el valor de obtener probabilidades interpretables, de controlar la complejidad con L2 y de validar con K-Fold estratificado mientras afinaba hiperparámetros con ROC-AUC (mucho más útil que mirar solo accuracy). También aprendí a separar predicción de decisión: el modelo estima la probabilidad y el umbral traduce esa probabilidad en acción según la prioridad del problema. Montar el modelo en FastAPI me enseñó a pensar en uso real (inputs limpios, categorías desconocidas, mensajes de error) y a valorar la trazabilidad que dan los metadatos. Finalmente, me llevo una conciencia más clara sobre buenas prácticas: evitar fugas de información, documentar cada paso y preferir soluciones simples pero robustas que se puedan mantener y explicar.

**Gabriel:** En este proyecto aprendí a aplicar la regresión logística más allá de lo teórico. El trabajo comenzó con la recolección de datos y su limpieza, lo que me hizo ver la importancia de organizar bien la información antes de entrenar un modelo. También entendí que no todas las variables aportan valor, y que seleccionar las más útiles mejora la calidad de los resultados. Algo que me quedó claro fue que el umbral de decisión no es fijo, y que ajustarlo cambia la forma en que el modelo equilibra entre detectar más casos positivos y reducir errores. Además, interpretar los coeficientes me permitió entender qué factores están más relacionados con los ingresos, como el nivel educativo o las horas trabajadas. En general, este proyecto me enseñó que un modelo de machine learning no solo depende de la matemática, sino del proceso completo: desde cómo se recogen y procesan los datos hasta cómo se interpretan los resultados para que puedan tener un uso real y confiable.