Predicción del nivel de ingresos utilizando técnicas de aprendizaje automático sobre el conjunto de datos Adult (UCI)

Santiago Arenas Gómez, Wilmar Andrés Osorio Úsuga y Dayana Ramírez Areiza

***Resumen*—Este trabajo presenta el análisis y formulación del problema de predicción de ingresos personales a partir del conjunto de datos *Adult* del *UCI Machine Learning Repository*. El objetivo consiste en predecir si una persona gana más de $50 000 al año en función de variables demográficas y laborales. Se propone una aproximación basada en técnicas de aprendizaje supervisado, incluyendo modelos paramétricos, no paramétricos, ensambles y redes neuronales. Además, se realiza una revisión bibliográfica de estudios previos que emplean este mismo conjunto de datos, con el fin de identificar estrategias de modelado, métricas de evaluación y problemáticas asociadas al desbalance de clases y a la equidad algorítmica.**

***Palabras clave*—Aprendizaje automático, predicción de ingresos, clasificación supervisada, conjunto de datos Adult, UCI Machine Learning Repository.**

I. Introducción

En el contexto actual, la predicción de ingresos personales constituye una aplicación clásica del aprendizaje automático, útil para explorar relaciones entre características demográficas y socioeconómicas de la población. Este tipo de modelado se aplica en áreas como estudios de mercado, análisis de políticas públicas y evaluación de riesgo crediticio. El conjunto de datos *Adult*, proveniente del *UCI Machine Learning Repository*, ofrece una oportunidad para analizar estos factores mediante modelos de clasificación supervisada.

El dataset *Adult* fue extraído de la base de datos del Censo de Estados Unidos de 1994 y contiene información de más de 48 000 individuos. El objetivo es predecir si una persona gana más de $50 000 al año a partir de variables como edad, nivel educativo, estado civil, ocupación, raza, género y horas de trabajo por semana [1].

Este problema es relevante no solo por su valor predictivo, sino también por su potencial para evaluar la equidad y los sesgos presentes en los modelos de aprendizaje automático, ya que las variables de raza y género pueden influir indebidamente en los resultados. Por tanto, el presente proyecto busca desarrollar, comparar y analizar diferentes modelos de clasificación para predecir el nivel de ingresos de los individuos y explorar la contribución de cada característica al rendimiento del modelo.

II. Descripción del problema

El problema abordado consiste en determinar si un individuo tiene ingresos anuales superiores a $50 000 a partir de un conjunto de variables demográficas y laborales. Este es un problema de clasificación binaria que puede formularse como:

$$f(x) = \begin{cases} 1, & \text{si el ingreso} > 50K, \\ 0, & \text{en caso contrario.} \end{cases}$$

II-A. Descripción de la base de datos

El conjunto de datos *Adult* está compuesto por 48 842 registros y 14 variables predictoras. Entre ellas se incluyen variables numéricas (*edad, horas por semana, capital_gain, capital_loss*) y categóricas (*educación, estado civil, ocupación, género, raza, país de origen*, entre otras). La variable objetivo es binaria e indica si el ingreso supera o no los $50 000.

En términos de su distribución, la variable objetivo presenta un desbalance marcado: aproximadamente el 75 % de los individuos pertenecen a la clase "$\leq 50K$", mientras que únicamente alrededor del 25 % corresponden a la clase "$> 50K$". Esta desproporción implica que los modelos pueden tender a favorecer la clase mayoritaria, por lo que será necesario considerar técnicas de balanceo o métricas robustas al desbalance en las fases posteriores del proyecto.

Existen valores faltantes en atributos como *workclass*, *occupation* y *native_country*, representados con el carácter "?". Estos serán tratados mediante imputación por la moda de cada variable categórica.

Las variables categóricas se codificarán mediante *one-hot encoding*, de modo que cada categoría se represente con una variable binaria. Por su parte, las variables numéricas se normalizarán para garantizar una escala homogénea entre características, evitando que las magnitudes afecten el rendimiento de los modelos supervisados.

II-B. Aproximación desde Machine Learning

Dado que se dispone de etiquetas, se trata de un problema de aprendizaje supervisado. Se explorarán diferentes configuraciones de modelos, incluyendo técnicas paramétricas (regresión logística), no paramétricas (k-Nearest Neighbors), ensambles (Random Forest), máquinas de soporte vectorial (SVM) y redes neuronales artificiales. El rendimiento se evaluará mediante validación cruzada y métricas de precisión, sensibilidad, especificidad y F1-score.

Esta formulación permite estructurar la experimentación posterior con diversos modelos supervisados, cuya eficacia y equidad han sido ampliamente estudiadas en la literatura, como se describe en la siguiente sección.

Universidad de Antioquia – Facultad de Ingeniería, Departamento de Ingeniería de Sistemas. Proyecto desarrollado para el curso Modelos y Simulación II.

Manuscrito recibido en octubre de 2025.

III. Estado del arte

En esta sección, se presenta una revisión del *estado del arte* relacionado con la predicción de ingresos utilizando técnicas de Machine Learning sobre el dataset Adult UCI. El objetivo es identificar trabajos previos relevantes, comparar los enfoques metodológicos utilizados (preprocesamiento, modelos, métricas de evaluación) y analizar los resultados obtenidos por otros investigadores. Esta revisión permitirá contextualizar los hallazgos de nuestro proyecto, identificar las mejores prácticas y justificar las decisiones tomadas en el desarrollo de nuestros modelos.

Diversos estudios han empleado el conjunto de datos *Adult* para evaluar algoritmos de clasificación, analizar sesgos y explorar la equidad algorítmica en modelos predictivos. A continuación, se presentan algunos de los trabajos más representativos.

Kohavi [2] realizó una de las primeras evaluaciones comparativas utilizando el conjunto *Adult*, donde contrastó métodos como árboles de decisión, Naïve Bayes y redes neuronales. Su trabajo introdujo la validación cruzada como técnica estándar para la evaluación de clasificadores y reportó precisiones cercanas al 85 %.

Becker y Kohavi [1] documentaron el conjunto de datos *Adult* como parte del repositorio de aprendizaje automático de la Universidad de California, Irvine (UCI), convirtiéndose en una referencia clásica para la investigación en predicción de ingresos personales y en estudios de equidad algorítmica.

Tabla I
Estado del arte (parte I): Estudios clásicos sobre el conjunto *Adult*.

Referencia	Año	Dataset
Becker & Kohavi [1]	1996	Adult (UCI)
Kohavi [2]	1996	Adult (UCI)

Modelos	Documentación, NB, árboles, redes
Métricas	Precisión ($\approx 0{,}85$), validación cruzada
Observaciones	Publicación original del dataset; comparación clásica de clasificadores

Posteriormente, Romei y Ruggieri [3] realizaron un análisis interdisciplinario sobre la detección de discriminación en modelos de datos, destacando el uso del conjunto *Adult* como uno de los principales benchmarks para evaluar la equidad en algoritmos de clasificación. Su revisión estableció las bases conceptuales para el estudio de la equidad en el aprendizaje automático.

Ding, Hardt, Miller y Schmidt [4] revisaron críticamente las limitaciones del *Adult Dataset* en estudios de justicia algorítmica, señalando que su estructura y desbalance de clases dificultan la evaluación precisa de la equidad. Los autores propusieron nuevos conjuntos de datos alternativos para reemplazarlo, más representativos y libres de sesgos históricos.

Tabla II
Estado del arte (parte II): Fairness y análisis discriminatorio.

Referencia	Año	Dataset
Romei & Ruggieri [3]	2014	Varios (incluye Adult)
Ding et al. [4]	2021	Nuevos datasets alternativos

Modelos	Revisión de clasificadores; discusión conceptual
Métricas	Métricas de equidad, disparate impact, fairness
Observaciones	Crítica al uso de Adult; propone datasets más justos

Por su parte, Brownlee [5] exploró el desbalance de clases en el conjunto *Adult*, donde solo el 24 % de los individuos pertenecen a la categoría ">50K". Su trabajo mostró mejoras significativas en la detección de la clase minoritaria mediante técnicas de sobremuestreo como *SMOTE* y resaltó la importancia de métricas complementarias a la precisión, como el F1-score.

Tabla III
Estado del arte (parte III): Desbalance, SMOTE y mejora de modelos.

Referencia	Año	Dataset
Brownlee [5]	2020	Adult (UCI)

Modelos	RF, SVM, k-NN, LogReg + SMOTE
Métricas	F1, recall, AUC (mejoras tras SMOTE)
Observaciones	Estrategias de tratamiento del desbalance; impacto en la clase minoritaria

Estos trabajos demuestran que el conjunto de datos *Adult* sigue siendo una herramienta fundamental en la evaluación de modelos de clasificación y en el análisis de equidad algorítmica, sirviendo tanto para validar técnicas tradicionales de aprendizaje supervisado como para estudiar los sesgos inherentes en la toma de decisiones automatizada.

IV. Entrenamiento y Evaluación de Modelos

En esta sección se describe el proceso de entrenamiento y evaluación de los modelos de clasificación implementados para predecir si un individuo obtiene ingresos superiores a \$50 000 anuales (clase 1) o no (clase 0) utilizando el conjunto de datos *Adult* del UCI Machine Learning Repository. Con el fin de asegurar la generalización de los resultados y mitigar el impacto del desbalance de clases, se adoptó una metodología experimental rigurosa basada en validación cruzada y técnicas de balanceo.

IV-A. Metodología de evaluación

La estrategia empleada consta de los siguientes elementos:

- **Validación cruzada estratificada:** Se utilizó *StratifiedKFold* con $k = 5$, *shuffle*=True y *random_state*=42, garantizando que cada partición preserve la proporción original de clases. Esto es especialmente relevante debido al desbalance existente, cercano a una razón 3.15:1.
- **Balanceo de clases mediante SMOTE:** Para abordar el desbalance se aplicó *SMOTE* exclusivamente sobre los conjuntos de entrenamiento dentro de cada fold, integrado en el pipeline de cada modelo. Esta estrategia evita *data leakage*, al impedir que información sintética influya en los conjuntos de validación.
- **Normalización de variables:** Las características numéricas fueron estandarizadas con *StandardScaler*. La

normalización se realiza únicamente sobre los datos del entrenamiento de cada fold, preservando la independencia del conjunto de validación.
- **Optimización de hiperparámetros:** Se emplearon *GridSearchCV* y *RandomizedSearchCV* para la búsqueda exhaustiva y estocástica, respectivamente. En todos los casos, el criterio de selección del mejor modelo fue el F1-score, particularmente adecuado en problemas con desbalance. La estrategia de validación utilizada durante la optimización fue la misma *StratifiedKFold* de 5 particiones.
- **Métricas de desempeño:** Se reportan *Accuracy*, *Precision*, *Recall*, *F1-score* y *AUC-ROC*. Estas métricas permiten evaluar el comportamiento global y la capacidad del modelo para identificar la clase minoritaria.

A continuación se presentan los detalles de entrenamiento y los resultados obtenidos para cada modelo considerado.

IV-B. Modelo 1: Regresión Logística

La regresión logística es un modelo paramétrico que estima la probabilidad de pertenecer a la clase positiva mediante una función sigmoidal aplicada a una combinación lineal de las características. A pesar de su simplicidad, constituye una línea base sólida en tareas de clasificación binaria. En este trabajo se optimizó utilizando *GridSearchCV* con una malla centrada en la regularización y el método de optimización.

Hiperparámetros óptimos:
- $C = 0{,}1$
- `penalty` = L1
- `solver` = `saga`
- `max_iter` = 1000

Resultados de validación cruzada (5-fold): Los valores reportados corresponden a la media $\pm$ desviación estándar obtenida en las cinco particiones de validación estratificada:
- **Accuracy**: $0{,}8081 \pm 0{,}0047$
- **Precision**: $0{,}5683 \pm 0{,}0069$
- **Recall**: $0{,}8461 \pm 0{,}0086$
- **F1-score**: $0{,}6799 \pm 0{,}0070$
- **ROC-AUC**: $0{,}9049 \pm 0{,}0023$

Intervalos de confianza al 95 %: Los intervalos se calcularon como:

$$\mu \pm 1{,}96 \cdot \sigma,$$

donde μ es la media y σ la desviación estándar estimada en la validación cruzada.
- **F1-score**: $[0{,}6742,\ 0{,}6917]$
- **ROC-AUC**: $[0{,}9014,\ 0{,}9080]$

IV-C. Modelo 2: k-Nearest Neighbors (k-NN)

El algoritmo *k-Nearest Neighbors* (k-NN) es un modelo no paramétrico basado en instancias que clasifica cada muestra de prueba a partir de la mayoría de etiquetas presentes en sus k vecinos más cercanos. Aunque su implementación es sencilla y su interpretación intuitiva, su desempeño puede deteriorarse en espacios de alta dimensionalidad y su costo computacional aumenta con el tamaño del conjunto de entrenamiento. La optimización del modelo se realizó mediante *GridSearchCV*.

Hiperparámetros óptimos:
- `metric` = `manhattan`
- `n_neighbors` = 11
- `weights` = `uniform`

Resultados de validación cruzada (5-fold):
- **Accuracy**: $0{,}7893 \pm 0{,}0044$
- **Precision**: $0{,}5431 \pm 0{,}0065$
- **Recall**: $0{,}7898 \pm 0{,}0105$
- **F1-score**: $0{,}6436 \pm 0{,}0067$
- **ROC-AUC**: $0{,}8638 \pm 0{,}0045$

Intervalos de confianza al 95 %:
- **F1-score**: $[0{,}6330,\ 0{,}6517]$
- **ROC-AUC**: $[0{,}8559,\ 0{,}8673]$

IV-D. Modelo 3: Random Forest

Random Forest es un método de aprendizaje conjunto (*ensemble learning*) que entrena un conjunto de árboles de decisión construidos sobre diferentes subconjuntos de los datos y de las características. La predicción final se obtiene mediante voto mayoritario en tareas de clasificación. Este enfoque reduce significativamente la varianza del modelo y mitiga el sobreajuste, convirtiéndolo en una de las técnicas más robustas para problemas tabulares. Dada la elevada cantidad de hiperparámetros y su interacción, la optimización se realizó mediante *RandomizedSearchCV*.

Hiperparámetros óptimos:
- `n_estimators` = 300
- `min_samples_split` = 10
- `min_samples_leaf` = 1
- `max_depth` = 30

Resultados de validación cruzada (5-fold):
- **Accuracy**: $0{,}8391 \pm 0{,}0044$
- **Precision**: $0{,}6334 \pm 0{,}0081$
- **Recall**: $0{,}7883 \pm 0{,}0091$
- **F1-score**: $0{,}7024 \pm 0{,}0074$
- **ROC-AUC**: $0{,}9102 \pm 0{,}0030$

Intervalos de confianza al 95 %:
- **F1-score**: $[0{,}6918,\ 0{,}7102]$
- **ROC-AUC**: $[0{,}9056,\ 0{,}9136]$

IV-E. Modelo 4: Red Neuronal Artificial (MLP)

La *Multilayer Perceptron* (MLP) es una red neuronal artificial de tipo feed-forward compuesta por capas densamente conectadas. Este tipo de modelo es capaz de aproximar funciones no lineales gracias al uso de múltiples capas ocultas y funciones de activación no lineales. En este trabajo, la arquitectura fue optimizada mediante *RandomizedSearchCV*, explorando múltiples configuraciones de profundidad y regularización.

Hiperparámetros óptimos:
- `activation` = `relu`
- `alpha` = 0.001
- `hidden_layer_sizes` = (64, 32)
- `learning_rate` = `adaptive`
- `max_iter` = 500

Resultados de validación cruzada (5-fold):

- **Accuracy**: $0{,}8142 \pm 0{,}0059$
- **Precision**: $0{,}5838 \pm 0{,}0098$
- **Recall**: $0{,}7966 \pm 0{,}0119$
- **F1-score**: $0{,}6738 \pm 0{,}0095$
- **ROC-AUC**: $0{,}8925 \pm 0{,}0047$

Intervalos de confianza al 95 %:

- **F1-score**: $[0{,}6632,\ 0{,}6881]$
- **ROC-AUC**: $[0{,}8866,\ 0{,}8998]$

IV-F. Modelo 5: SVM con Kernel RBF

Las *Support Vector Machines* (SVM) con *kernel* de base radial (RBF) son clasificadores no lineales que proyectan los datos a un espacio de mayor dimensión mediante una transformación implícita, permitiendo la construcción de un hiperplano que maximiza el margen entre clases. Estos modelos suelen ofrecer un rendimiento competitivo, pero presentan un costo computacional elevado en conjuntos de datos grandes. Por este motivo, la búsqueda de hiperparámetros mediante *RandomizedSearchCV* se realizó sobre un subconjunto estratificado equivalente al 40 % del conjunto de datos original, utilizando validación cruzada de 3 particiones.

Hiperparámetros óptimos:

- `C` = 100
- `gamma` = 0.01
- `kernel` = `rbf`
- `cache_size` = 1000
- `max_iter` = 1000

Resultados de validación cruzada (3-fold en subconjunto del 40 %):

- **Accuracy**: $0{,}7483 \pm 0{,}0196$
- **Precision**: $0{,}4815 \pm 0{,}0328$
- **Recall**: $0{,}5545 \pm 0{,}0454$
- **F1-score**: $0{,}5146 \pm 0{,}0334$
- **ROC-AUC**: $0{,}7410 \pm 0{,}0307$

Intervalos de confianza al 95 %:

- **F1-score**: $[0{,}4872,\ 0{,}5582]$
- **ROC-AUC**: $[0{,}7086,\ 0{,}7791]$

Nota: La búsqueda de hiperparámetros y la validación se realizaron sobre un subconjunto estratificado del 40 % del conjunto de datos para reducir el tiempo computacional. Si bien las métricas provienen de un subconjunto, siguen siendo representativas y estadísticamente válidas para el análisis comparativo.

IV-G. Tabla Comparativa de Resultados de Validación Cruzada

La Tabla IV resume el rendimiento de los modelos evaluados mediante validación cruzada. Se presentan la media y la desviación estándar de cada métrica. En general, el modelo **Random Forest** obtiene el mejor desempeño global en F1-score y ROC-AUC.

Tabla IV
RESULTADOS DE VALIDACIÓN CRUZADA. MEDIA ± DESVIACIÓN ESTÁNDAR.

Modelo	Accuracy	Precision	Recall
Logistic Regression	$0{,}8081 \pm 0{,}0047$	$0{,}5683 \pm 0{,}0069$	$0{,}8461 \pm 0{,}0086$
k-NN	$0{,}7893 \pm 0{,}0044$	$0{,}5431 \pm 0{,}0065$	$0{,}7898 \pm 0{,}0105$
Random Forest	$\mathbf{0{,}8391 \pm 0{,}0044}$	$\mathbf{0{,}6334 \pm 0{,}0081}$	$\mathbf{0{,}7883 \pm 0{,}0091}$

Modelo	F1-score	ROC-AUC
Neural Network (MLP)	$0{,}6738 \pm 0{,}0095$	$0{,}8925 \pm 0{,}0047$
SVM (RBF)†	$0{,}5146 \pm 0{,}0334$	$0{,}7410 \pm 0{,}0307$

IV-H. Intervalos de Confianza (95 %) para Métricas Clave

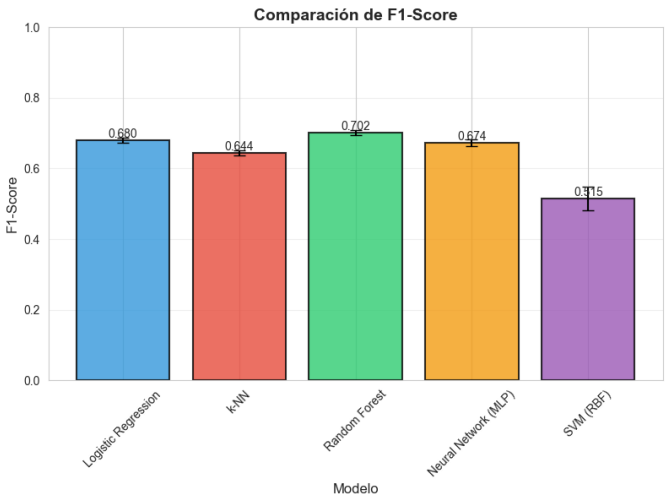
Los intervalos de confianza al 95 % permiten estimar el rango dentro del cual se encuentra el verdadero valor poblacional de cada métrica, si el experimento se repitiera múltiples veces. Estos intervalos reflejan la estabilidad del modelo y la variabilidad introducida por la validación cruzada.

Tabla V
INTERVALOS DE CONFIANZA AL 95 % PARA F1-SCORE Y ROC-AUC (MEDIA ± 1.96 DESVIACIONES ESTÁNDAR).

Modelo	F1-score (IC 95 %)	ROC-AUC (IC 95 %)
Logistic Regression	$[0{,}6742,\ 0{,}6917]$	$[0{,}9014,\ 0{,}9080]$
k-NN	$[0{,}6330,\ 0{,}6517]$	$[0{,}8559,\ 0{,}8673]$
Random Forest	$\mathbf{[0{,}6918,\ 0{,}7102]}$	$\mathbf{[0{,}9056,\ 0{,}9136]}$
Neural Network (MLP)	$[0{,}6632,\ 0{,}6881]$	$[0{,}8866,\ 0{,}8998]$
SVM (RBF)†	$[0{,}4872,\ 0{,}5582]$	$[0{,}7086,\ 0{,}7791]$

IV-I. Visualización: Gráfico de Barras — Comparación de Métricas

La Figura 1 presenta un resumen visual del rendimiento promedio de los modelos en las métricas clave. El uso de barras de error permite observar la variabilidad obtenida durante la validación cruzada.



F1-score

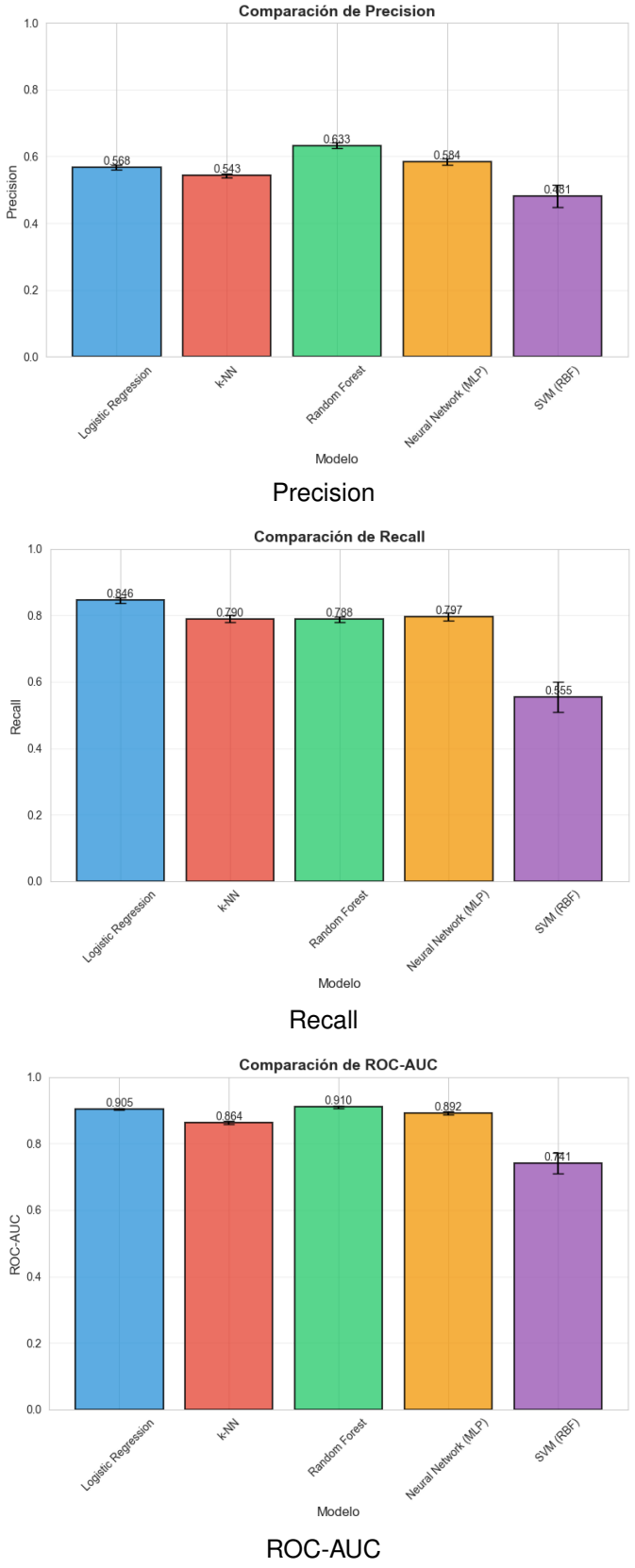


Figura 1. Comparación visual de las métricas de los modelos evaluados: F1-score, Precision, Recall y ROC-AUC.

IV-J. Visualización: Curvas ROC

Las curvas ROC (*Receiver Operating Characteristic*) permiten evaluar el rendimiento de los clasificadores binarios al representar la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) para distintos umbrales de decisión. El *Área bajo la curva* (AUC) resume en un único valor el comportamiento global del modelo; valores más cercanos a 1 indican un mejor desempeño. La Figura 2 muestra las curvas ROC obtenidas para los modelos evaluados.

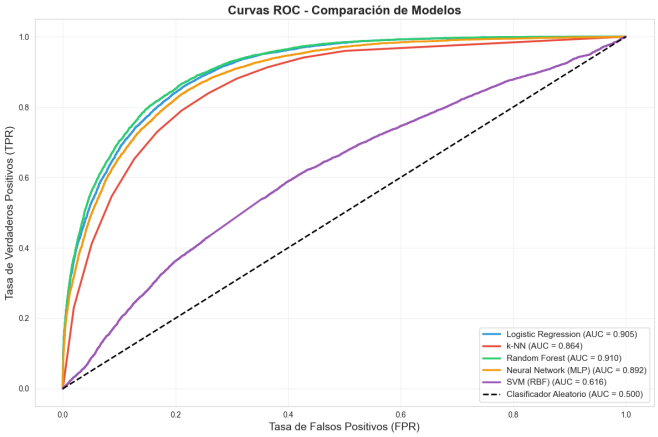
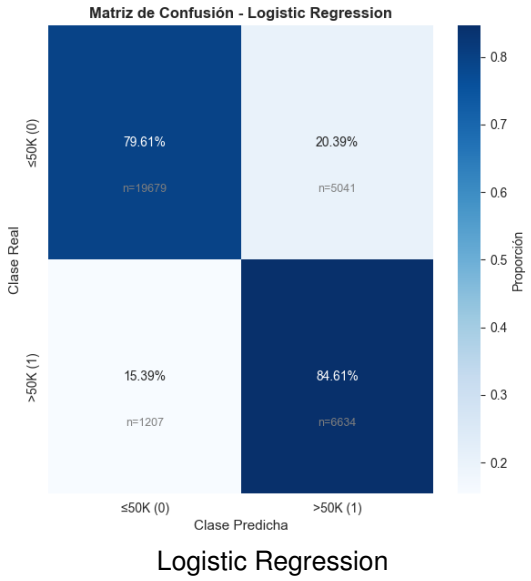


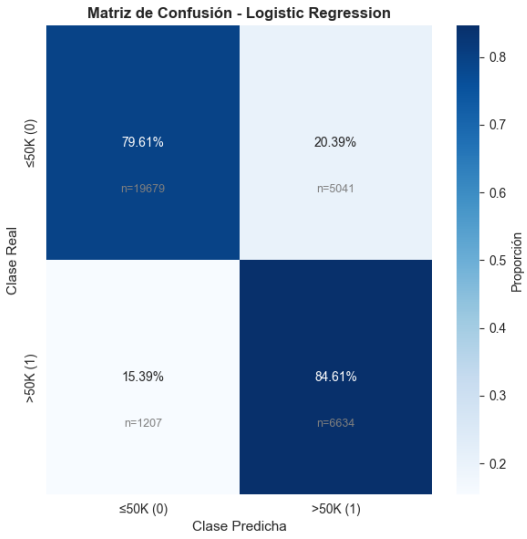
Figura 2. Curvas ROC de los modelos evaluados. Se incluye el valor del área bajo la curva (AUC) para cada clasificador.

IV-K. Visualización: Matrices de Confusión

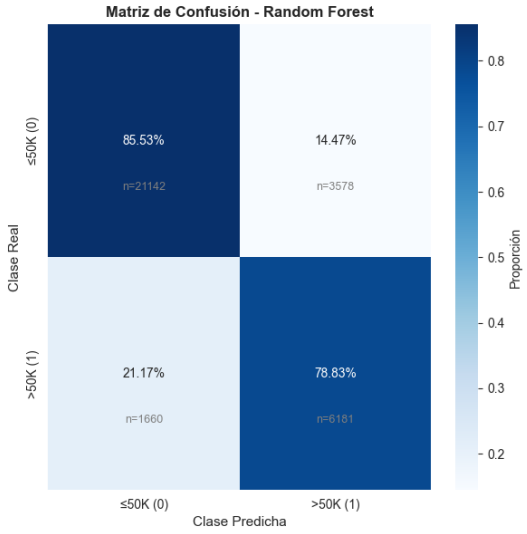
Las matrices de confusión permiten analizar detalladamente el desempeño de un clasificador al mostrar los conteos de verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN). Debido al desbalance del dataset, se presentan matrices de confusión normalizadas, complementadas con los valores absolutos dentro de cada celda, lo que facilita evaluar tanto el rendimiento relativo como el impacto real de los errores de clasificación en cada modelo.



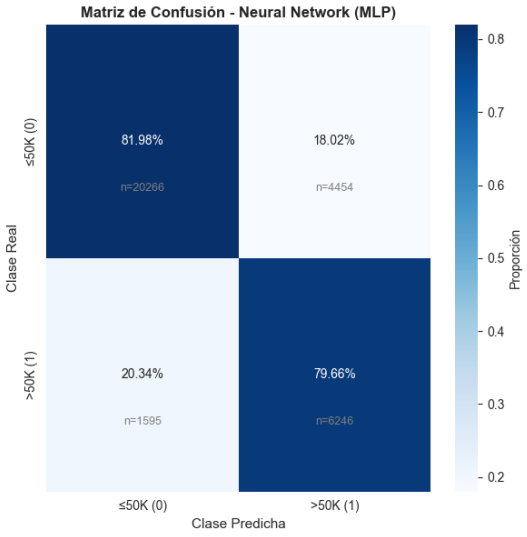
Logistic Regression



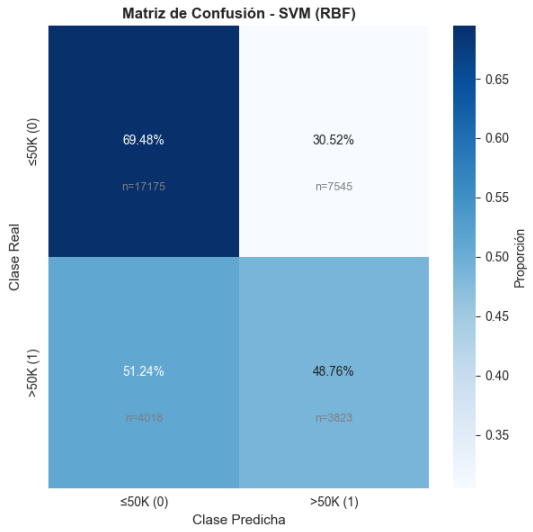
k-NN



Random Forest



Neural Network (MLP)



SVM (RBF)

Figura 3. Matrices de confusión normalizadas para cada modelo evaluado. Los valores indican proporciones por clase y conteos absolutos.

IV-L. Conclusiones y Modelos Guardados

El análisis comparativo de los cinco modelos evaluados permitió identificar diferencias claras en su capacidad para manejar el desbalance de clases y capturar las relaciones presentes en el conjunto de datos. En términos de desempeño global, considerando especialmente el F1-Score y el área bajo la curva ROC (ROC-AUC), el **Random Forest** se consolidó como el modelo de mejor rendimiento general. Este clasificador obtuvo el mayor F1-Score ($0{,}7024 \pm 0{,}0074$) y el valor más alto de ROC-AUC ($0{,}9102 \pm 0{,}0030$), lo que indica un buen equilibrio entre sensibilidad y precisión, así como una capacidad robusta para discriminar entre ambas clases.

La **Regresión Logística** también mostró un comportamiento competitivo, particularmente en Recall, evidenciando su utilidad como modelo base y su capacidad para capturar relaciones lineales en los datos. Por su parte, el MLP y el k-NN obtuvieron resultados intermedios, mientras que el SVM con kernel RBF, a pesar de su potencial teórico, presentó un desempeño inferior debido a las restricciones computacionales y a la necesidad de entrenarlo en un subconjunto del dataset.

Para facilitar su reutilización y despliegue, todos los modelos optimizados fueron serializados y almacenados en el directorio `modelos_entrenados/`. Estos archivos contienen tanto la arquitectura final como los hiperparámetros óptimos seleccionados durante el proceso de búsqueda.

Mejor modelo según F1-Score: Random Forest F1-Score: $0{,}7024 \pm 0{,}0074$, ROC-AUC: $0{,}9102 \pm 0{,}0030$

Modelos serializados:

- `modelos_entrenados/logistic_regression.pkl`
- `modelos_entrenados/knn.pkl`
- `modelos_entrenados/random_forest.pkl`
- `modelos_entrenados/neural_network.pkl`
- `modelos_entrenados/svm.pkl`

IV-L1. Resumen de Hiperparámetros Óptimos: A continuación se presenta un resumen de los mejores hiperparámetros identificados para cada modelo durante el proceso de optimización:

1. **Regresión Logística**
   - `classifier__C`: 0.1
   - `classifier__max_iter`: 1000
   - `classifier__penalty`: l1
   - `classifier__solver`: saga
2. **k-Nearest Neighbors**
   - `classifier__metric`: manhattan
   - `classifier__n_neighbors`: 11
   - `classifier__weights`: uniform
3. **Random Forest**
   - `classifier__n_estimators`: 300
   - `classifier__min_samples_split`: 10
   - `classifier__min_samples_leaf`: 1
   - `classifier__max_depth`: 30
4. **Red Neuronal Artificial (MLP)**
   - `classifier__max_iter`: 500
   - `classifier__learning_rate`: adaptive
   - `classifier__hidden_layer_sizes`: (64, 32)
   - `classifier__alpha`: 0.001
   - `classifier__activation`: relu
5. **SVM (Kernel RBF)**
   - `classifier__kernel`: rbf
   - `classifier__gamma`: 0.01
   - `classifier__C`: 100

V. Reducción de Dimensión

En esta sección se analiza el impacto de la reducción de dimensionalidad sobre el rendimiento de los modelos de clasificación. Se aplican dos técnicas ampliamente utilizadas: el Análisis de Componentes Principales (PCA), como método lineal, y el Uniform Manifold Approximation and Projection (UMAP), como método no lineal. El objetivo es determinar si una representación de menor dimensionalidad mejora la eficiencia computacional o el desempeño predictivo, particularmente en los dos modelos con mejor rendimiento identificados previamente (Random Forest y Regresión Logística).

V-A. Análisis de Relevancia de Características

Se realizó un análisis de relevancia de características con el fin de identificar su contribución individual a la predicción del ingreso. Para ello se utilizaron tres métodos estadísticos: *Mutual Information*, Chi-cuadrado y ANOVA F-statistic. Los puntajes obtenidos se normalizaron y promediaron para construir una medida de importancia más robusta.

Los principales hallazgos fueron los siguientes:

- Las características con mayor relevancia corresponden a `marital-status_Married-civ-spouse`, `capital-gain`, `education-num` y `age`.
- Se identificaron 24 características con importancia baja (por debajo del percentil 25), en su mayoría relacionadas con países de origen poco representados y categorías específicas de ocupación o educación.

Las Figuras 4 y 5 presentan la visualización de estos resultados.

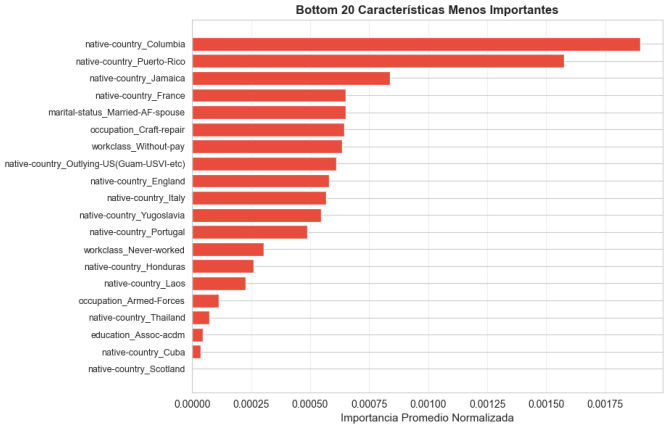


Figura 4. Características con menor relevancia según el promedio normalizado de importancia.

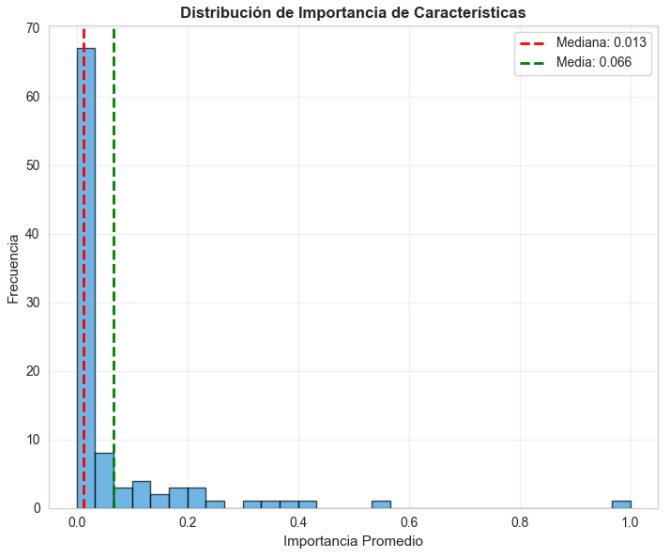


Figura 5. Distribución de la importancia media de las características y comparación entre métodos.

V-B. Extracción de Características con PCA (Reducción Lineal)

El Análisis de Componentes Principales (PCA) se aplicó para reducir la dimensionalidad de los datos de manera lineal. Antes de su aplicación, se utilizó `StandardScaler` para normalizar las características y asegurar que todas contribuyeran equitativamente al análisis. Se determinó que **82 componentes principales** son suficientes para retener el **95.74 %** de la varianza explicada. Esto corresponde a una reducción del **15.5 %** respecto a las 97 características originales.

Los dos modelos con mejor rendimiento (Random Forest y Regresión Logística) fueron reentrenados utilizando estas características reducidas. La Figura 6 muestra la varianza explicada por componente.

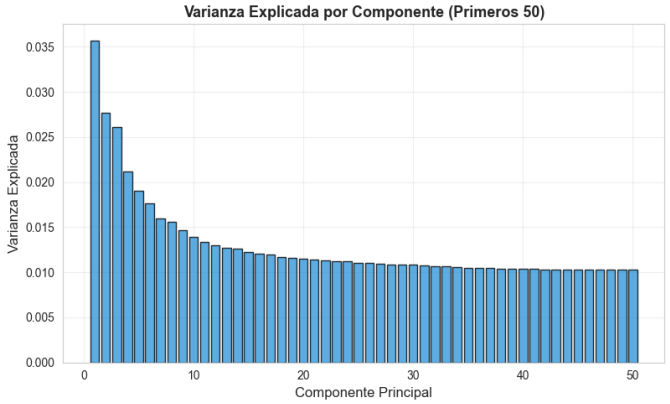


Figura 6. Varianza explicada obtenida mediante PCA.

V-C. Extracción de Características con UMAP (Reducción No Lineal)

Para explorar reducciones no lineales, se utilizó *Uniform Manifold Approximation and Projection* (UMAP). Se configuró para reducir el conjunto a **20 componentes**, lo cual representa una reducción del **79.4 %** respecto a las 97 características originales. A diferencia de PCA, UMAP busca preservar la estructura topológica de los datos, lo cual puede resultar beneficioso en problemas con relaciones altamente no lineales.

Los dos modelos principales también fueron reentrenados con estas características reducidas. La Figura 7 ilustra la proyección bidimensional obtenida.

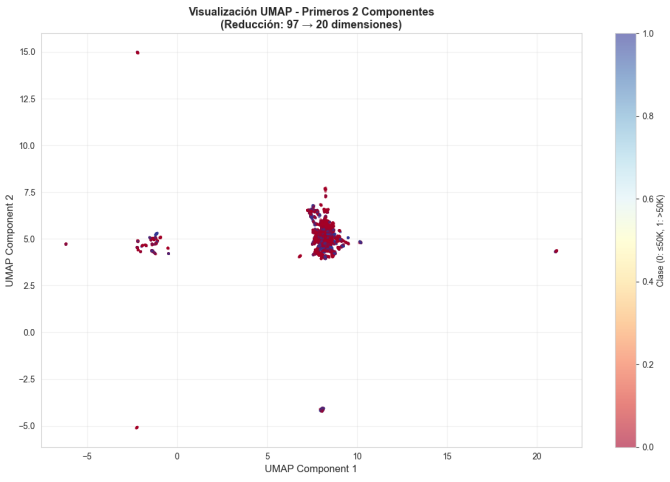


Figura 7. Proyección bidimensional de UMAP coloreada por clase de ingreso.

V-D. Comparación Global: Original vs PCA vs UMAP

La Tabla VI resume el impacto de PCA y UMAP en el rendimiento de Random Forest y Regresión Logística. Se muestran las dimensiones resultantes, el porcentaje de reducción y las métricas F1-Score y ROC-AUC obtenidas mediante validación cruzada.

Los resultados muestran una disminución en el rendimiento al aplicar reducción de dimensionalidad, especialmente en el caso de UMAP. PCA conserva mejor la estructura necesaria para modelos lineales como la Regresión Logística, aunque aún representa una ligera degradación comparada con el conjunto original.

Tabla VI
RENDIMIENTO CON CARACTERÍSTICAS ORIGINALES, PCA Y UMAP

Modelo	Método	Dim	Red.	F1	AUC
RF	Original	97	0 %	0.7024	0.9102
RF	PCA	82	15.5 %	0.6588	0.8807
RF	UMAP	20	79.4 %	0.6169	0.8458
LR	Original	97	0 %	0.6799	0.9049
LR	PCA	82	15.5 %	0.6740	0.9022
LR	UMAP	20	79.4 %	0.4674	0.6996

La Figura 8 muestra una comparación visual del deterioro en las métricas.

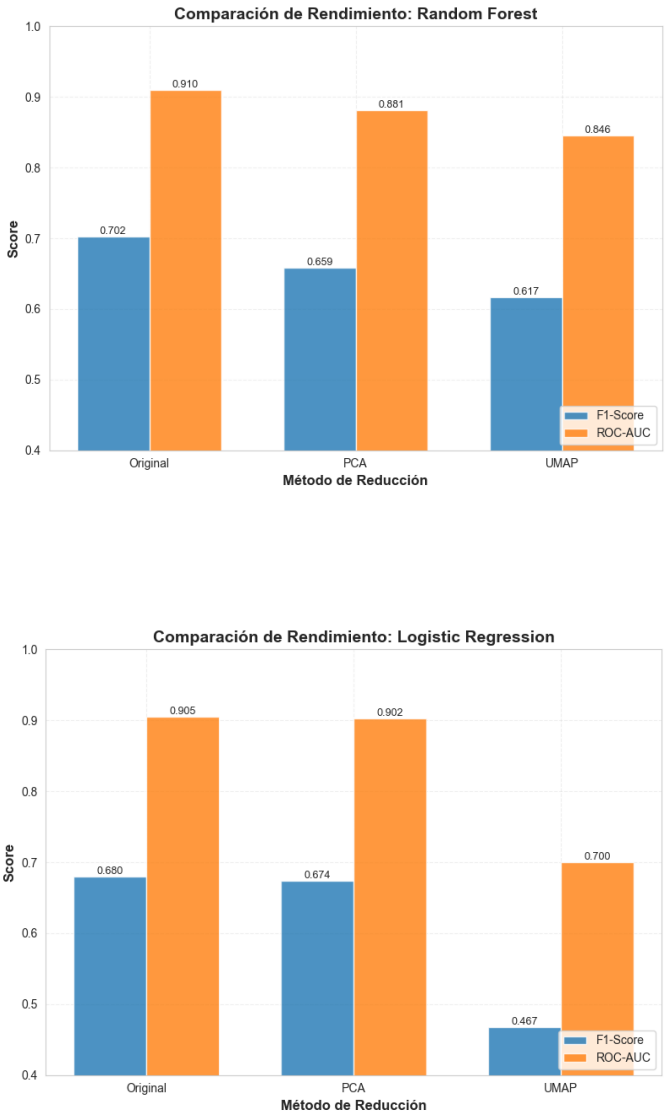


Figura 8. Comparación de F1-Score y ROC-AUC para Random Forest y Regresión Logística con distintos métodos de reducción.

V-E. Conclusiones de la Sección de Reducción de Dimensión

El análisis de reducción de dimensionalidad permite extraer varias conclusiones relevantes para este proyecto:

- **Relevancia de características:** Las variables asociadas al estado civil, la ganancia de capital, el nivel educativo y la edad fueron identificadas consistentemente como las más influyentes en la predicción del ingreso. Asimismo, se encontraron 24 características con relevancia muy baja, lo que indica que una *selección de características* podría resultar más adecuada que una extracción lineal o no lineal.
- **Impacto de PCA:** Aunque PCA logró reducir la dimensionalidad en un 15.5 % manteniendo más del 95 % de la varianza explicada, ambos modelos presentaron una ligera disminución en rendimiento. Esto sugiere que las relaciones lineales capturadas por PCA no conservan completamente la información discriminativa necesaria para el problema.
- **Impacto de UMAP:** La reducción más agresiva realizada con UMAP (79.4 %) produjo una caída considerable en el desempeño tanto de Random Forest como de la Regresión Logística. Esto indica que, pese a su capacidad para preservar estructuras no lineales, la proyección utilizada no retuvo suficiente información relevante para la clasificación.
- **Recomendación final:** Para este conjunto de datos y los modelos evaluados, mantener las características originales proporciona el mejor rendimiento en F1-Score y ROC-AUC. La reducción de dimensionalidad no mejoró el desempeño y, en el caso de UMAP, lo deterioró notablemente. Una estrategia más prometedora sería aplicar *selección de características* basada en su importancia individual, o ajustar más exhaustivamente los hiperparámetros de las técnicas de reducción si se desea explorar esta vía en futuros trabajos.

VI. Conclusiones

El presente trabajo abordó el problema de predicción del nivel de ingresos en el conjunto de datos *Adult* del UCI Machine Learning Repository mediante la aplicación de múltiples técnicas de aprendizaje supervisado. A partir de un proceso riguroso de preprocesamiento, validación cruzada, optimización de hiperparámetros y análisis comparativo, se obtuvieron diversos hallazgos relevantes.

En primer lugar, se determinó que el conjunto de datos presenta desafíos inherentes, tales como el desbalance de clases, la presencia de variables categóricas de alta cardinalidad y relaciones no lineales entre características. La aplicación de *SMOTE* dentro de los pipelines de aprendizaje permitió mitigar parcialmente el desbalance, contribuyendo a una evaluación más justa de los modelos.

En cuanto al rendimiento predictivo, el modelo **Random Forest** se posicionó como la mejor alternativa, obteniendo los valores más altos de F1-Score y ROC-AUC entre todos los modelos evaluados. La Regresión Logística se desempeñó de manera competitiva y demostró ser un modelo robusto y eficiente, especialmente considerando su simplicidad y menor costo computacional. Los modelos k-NN, MLP y SVM mostraron comportamientos adecuados, aunque inferiores en comparación con las dos mejores alternativas.

El análisis de reducción de dimensionalidad reveló que, para este conjunto de datos, tanto PCA como UMAP no mejoraron el desempeño de los modelos. Aunque PCA conservó más del 95 % de la varianza y UMAP produjo proyecciones no lineales más compactas, ambas técnicas introdujeron pérdidas de información relevantes para la tarea de clasificación. En consecuencia, se concluyó que trabajar con el conjunto original de características (tras un preprocesamiento adecuado) proporciona mejores resultados que aplicar transformaciones de reducción dimensional.

Finalmente, el análisis de importancia de características destacó que solo un subconjunto reducido de variables aporta información significativa a la predicción. Esto sugiere que enfoques de *selección de características* podrían constituir una vía prometedora para futuros trabajos, así como la evaluación de técnicas más avanzadas de balanceo, regularización o interpretabilidad. También sería pertinente explorar modelos basados en *gradient boosting* modernos (como XGBoost, LightGBM o CatBoost), que podrían mejorar aún más el rendimiento observado.

En conjunto, los resultados obtenidos evidencian la utilidad del aprendizaje automático para la predicción del nivel de ingresos y ofrecen una línea base sólida para investigaciones posteriores orientadas a mejorar la precisión, la eficiencia y la equidad de los modelos aplicados.

Referencias

[1] B. Becker and R. Kohavi, "Adult Data Set," *UCI Machine Learning Repository*, 1996. DOI: 10.24432/C5XW20.

[2] R. Kohavi, "Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid," in *Proc. Second Int. Conf. Knowledge Discovery and Data Mining (KDD)*, Portland, OR, USA, 1996, pp. 202–207.

[3] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, vol. 29, no. 5, pp. 582–638, 2014. DOI: 10.1017/S0269888913000039.

[4] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring Adult: New Datasets for Fair Machine Learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[5] J. Brownlee, "Imbalanced Classification with the Adult Income Dataset," *Machine Learning Mastery*, 2020. [Online]. Available: https://machinelearningmastery.com/imbalanced-classification-with-the-adult-income-dataset/