APRENDIZAJE DE MÁQUINAS (ACIF104)
AVANCE INFORME FASE 2 SUMATIVA 2

Nombre integrantes
Grupo 13:

- **Alonso Cid Riveros**
- **Scarlett Espinoza Contreras**
- **Christian Mattioni Avila**

Curso:

- **Aprendizaje Maquina - ACIF 104 – NRC 2068**

Fecha:

- **05/12/2025**

ÍNDICE

1. PROBLEMÁTICA

1.1. CONTEXTO Y RELEVANCIA:

El problema que se aborda es la predicción de niveles de ingreso, específicamente, determinar si un individuo gana más o menos de 50,000 dólares anuales basándose en un conjunto de características demográficas y laborales. Esta problemática es de alta relevancia en el contexto organizacional y de políticas públicas, ya que permite identificar los factores clave (como nivel educativo, tipo de trabajo, edad o estado civil) que influyen en la brecha salarial. Comprender estos factores es el primer paso para diseñar intervenciones efectivas que busquen promover la equidad económica

1.2. ANÁLISIS DE ANTECEDENTES:

Para justificar la relevancia de la predicción de ingresos y brecha salarial, se revisaron distintos trabajos recientes que abordan problemas socioeconómicos mediante modelos de clasificación. En primer lugar, Solís-Salazar y Madrigal-Sanabria (2022) comparan un modelo basado en XGBoost con el enfoque tradicional de Proxy Means Test para la predicción de pobreza, mostrando que los ensambles de árboles reducen significativamente los errores de clasificación en contextos con variables heterogéneas y no lineales. Este resultado refuerza la idea de que los problemas socioeconómicos, como la estimación de ingresos, se benefician de técnicas modernas de aprendizaje automático en lugar de métodos puramente estadísticos.

De forma complementaria, Curto Merino (2023) analiza la pobreza en España utilizando regresión logística, árboles de decisión y una red neuronal MLP. El autor concluye que, aunque la regresión logística mantiene una alta interpretabilidad y un desempeño competitivo, las redes neuronales logran una ligera ventaja cuando el preprocesamiento de datos es adecuado. Este hallazgo es coherente con nuestra necesidad de equilibrar explicabilidad y rendimiento al predecir si un individuo supera el umbral de 50.000 dólares anuales.

En la misma línea, Muñetón-Santa y Manrique-Ruiz (2023) aplican Random Forest, CatBoost y LightGBM para estimar un índice de pobreza multidimensional, encontrando que los ensambles de árboles capturan de mejor forma las interacciones complejas en datos tabulares. Esto es especialmente relevante para el dataset Adult, donde confluyen variables demográficas, laborales y de capital que interactúan de manera no lineal. Por otro lado, Smith Uldall y Gutiérrez Rojas (2022) desarrollan un sistema de alerta temprana para deserción escolar en Chile, comparando regresión logística, árboles y redes neuronales. Su trabajo destaca la importancia de métricas orientadas a la clase minoritaria (como recall y F1-score) cuando las decisiones tienen impacto social, lo que se alinea directamente con nuestro foco en la clase >50K en un escenario de desbalance 76/24.

Finalmente, la revisión sistemática de Pincay-Ponce et al. (2022) evidencia el uso recurrente de modelos como SVM, Random Forest y redes neuronales en el análisis de factores socioeconómicos, así como la relevancia de variables relacionadas con educación, empleo, género e ingresos familiares. Esta revisión respalda la elección del dataset Adult como un caso representativo y valida el uso de técnicas de aprendizaje automático para abordar la brecha salarial. En conjunto, estos antecedentes confirman que el problema abordado en este proyecto es pertinente, actual y coherente con la agenda de investigación en políticas públicas y analítica de datos socioeconómicos, justificando la necesidad de desarrollar un modelo interpretable y robusto para la predicción de niveles de ingreso.

1.3. REQUISITOS INICIALES:

A partir de la problemática y el análisis de antecedentes, se define el requisito principal del proyecto:

- Desarrollar un modelo de software basado en aprendizaje automático capaz de predecir la categoría de ingreso (<=50K o >50K) de un individuo, basándose en las 14 características proporcionadas en el dataset 'Adult'. (https://www.kaggle.com/datasets/wenruliu/adult-income-dataset/)
- El modelo debe priorizar el **F1-Score** de la clase minoritaria (>50K), dado el desbalance de clases (76% vs 24%) identificado en el análisis exploratorio (Celda 11 del notebook).
- El modelo debe ser interpretable, permitiendo identificar qué características son las más influyentes en la predicción (requisito de explicabilidad).
- Estos requisitos iniciales se vinculan directamente con los aprendizajes esperados del curso. En relación con el **AE2**, el proyecto exige seleccionar y aplicar técnicas de aprendizaje automático acordes a la naturaleza del problema, sus restricciones computacionales y el fuerte desbalance de clases identificado durante el análisis exploratorio. La necesidad de construir un modelo capaz de predecir ingresos en función de variables demográficas y laborales obliga a considerar distintas técnicas supervisadas, así como estrategias de preprocesamiento y balanceo, alineándose con la aplicación práctica requerida por este aprendizaje.
- Por otra parte, estos requisitos también se relacionan con el **AE3**, puesto que establecen las bases para la planificación del proyecto de software basado en aprendizaje automático. Definir de manera explícita los objetivos, restricciones, características de entrada, tipo de salida e interpretación esperada permite guiar las fases posteriores de desarrollo, tales como la preparación de los datos, la experimentación con modelos y la integración del sistema en una plataforma de uso final. De esta manera, los requisitos iniciales no solo contextualizan el problema, sino que también estructuran el plan de trabajo necesario para su implementación.

2. METODOLOGÍA

2.1. METODOLOGÍA APLICADA:

Para el desarrollo de este proyecto, se adoptó una metodología estructurada basada en las fases del modelo CRISP-DM (Cross-Industry Standard Process for Data Mining), adaptada a los requisitos específicos del aprendizaje automático:

1. **Comprensión del Problema:** Definición del objetivo (predecir ingresos >$50K) y los requisitos del proyecto.
2. **Comprensión de los Datos (EDA):** Análisis inicial para identificar variables, distribuciones, valores atípicos (outliers) y el fuerte desbalance de clases (76% vs 24%) . (Corresponde a las celdas 01-12 del notebook).
3. **Preparación de Datos:** Limpieza de datos (manejo de '?'), ingeniería de características (eliminación de fnlwgt y education), y construcción de un pipeline de preprocesamiento unificado (ColumnTransformer) para escalar datos numéricos y codificar categóricos (OneHotEncoder). (Celdas 13-17 del notebook).

4. **Modelado (Experimentación):**
    - Implementación y comparación de 3 modelos de ML (Regresión Logística, RF, SVM).
    - Implementación y comparación de 3 arquitecturas de DL (MLP Básico, MLP con Dropout, Wide & Deep).
    - Aplicación y análisis de 3 técnicas de balanceo (Baseline, SMOTE, Class Weight).
5. **Evaluación (Selección):** Selección del modelo ganador ("MLP Optimizado") basándose en las métricas clave (F1-Score y AUC-ROC) y análisis de interpretabilidad (SHAP).
6. **Refinamiento y Despliegue:** Ajuste de hiperparámetros (KerasTuner) y planificación del desarrollo (Frontend/Backend)."

2.2. PLAN DE TRABAJO

Aquí listamos las tareas, plazos y responsables.

FASE / TAREA	RESPONSABLE(S)	PLAZO (EJEMPLO)	ESTADO
FASE 1: PLANIFICACIÓN			
Definición de la Problemática (Req. II)	Alonso Cid	Semana En Curso	Completado
Búsqueda de Antecedentes (5 fuentes)	Scarlett Espinoza	Semana En Curso	Completado
Definición Requisitos (Req. 5d)	Christian Mattioni	Semana En Curso	Completado
Fase 2: Análisis y Modelado			
Análisis Exploratorio de Datos (EDA) (Req. 5a)	Alonso Cid	Semana En Curso	Completado
Preprocesamiento y Pipelines (Req. 5f)	Christian Mattioni	Semana En Curso	Completado
Experimentación ML (Baseline, SMOTE) (Req. 5c, 5f)	Scarlett Espinoza	Semana En Curso	Completado
Experimentación DL (MLP, Dropout, W&D) (Req. 5c)	Christian Mattioni	Semana En Curso	Completado
FASE 3: REFINAMIENTO Y EVALUACIÓN			
Ajuste Hiperparámetros (KerasTuner) (Req. 5e)	Christian Mattioni	Semana En Curso	Completado
Análisis SHAP y Curva ROC (Req. 10)	Alonso Cid	Semana En Curso	Completado
FASE 4: INFORME Y DESPLIEGUE			
Redacción Informe Sumativo (PDF)	Todo el equipo	Semana En Curso	Completado
Desarrollo API (Backend) (Req. 5g)	Alonso Cid	Semana En Curso	Completado
Desarrollo (Frontend) (Req. 5g)	Alonso Cid	Semana En Curso	Completado
Configuración GitHub (Req. 12)	Alonso Cid	Semana En Curso	Completado

3. DESARROLLO DEL PROYECTO

Para garantizar un flujo de trabajo estructurado y asegurar la trazabilidad entre el problema de negocio y la solución técnica, el desarrollo de este proyecto se rigió por las seis fases del modelo estándar **CRISP-DM**. A continuación, se detalla cómo cada etapa metodológica condicionó las decisiones y se materializa en las secciones técnicas de este informe:

- **Fase 1: Entendimiento del Negocio** El objetivo inicial fue traducir la problemática de la brecha salarial en objetivos técnicos concretos. Esta fase se consolida en el apartado **D. Requisitos del Proyecto**, donde se definieron las restricciones funcionales y, crucialmente, se determinó priorizar el **F1-Score** sobre la exactitud (*Accuracy*) como métrica de éxito, dado el contexto de desbalance del problema.
- **Fase 2: Entendimiento de los Datos** Antes de modelar, fue crítico comprender la naturaleza de la información disponible. Esta etapa corresponde a la sección **A. Análisis Exploratorio de Datos (EDA)**, donde se identificaron la estructura de las variables, la presencia de valores nulos y se confirmó el hallazgo más determinante: el severo desbalance de clases (solo 24% de ingresos altos), lo cual condicionó las estrategias posteriores.
- **Fase 3: Preparación de los Datos** La transformación de datos crudos en insumos aptos para el aprendizaje automático se aborda en dos momentos: la limpieza inicial en la **sección A** y, con mayor profundidad, en la **sección F. Elaboración de Modelos**. Aquí se aplicaron técnicas de partición (*train/test split*) y estrategias de balanceo (*SMOTE, Class Weights*) como respuesta directa para mitigar el sesgo detectado en la fase anterior.
- **Fase 4: Modelado** Esta es la etapa central del desarrollo, donde se seleccionaron y calibraron los algoritmos. Se desglosa en:
    - **B. Selección de Técnicas Candidatas:** Justificación teórica de los modelos a probar.
    - **E. Selección y Refinamiento de la Arquitectura:** Diseño específico de las redes neuronales (incluyendo la decisión de usar Dropout) y su optimización de hiperparámetros.
- **Fase 5: Evaluación** Una vez entrenados los modelos, se procedió a validar su desempeño comparativo para seleccionar al ganador. Esta etapa se refleja en la sección **C. Comparación de Técnicas**, donde se contrastaron los resultados de Machine Learning clásico frente a Deep Learning, validando cuantitativamente la superioridad de la arquitectura propuesta.
- **Fase 6: Despliegue** Finalmente, para operacionalizar el conocimiento adquirido, se construyó un prototipo funcional descrito en la sección **G. Desarrollo de Frontend y Backend**, llevando el modelo "del laboratorio a la producción" y cumpliendo con los requisitos de usabilidad.

A continuación, profundizaremos en las secciones descritas anteriormente.

A. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

El proyecto se inició con un análisis exploratorio del dataset 'Adult'. Tras la carga (Celda 1), se identificó que los valores faltantes estaban representados por el string `'?'`. Se realizó una limpieza inicial para eliminar espacios en blanco y convertir estos marcadores a `np.nan` (Celda 3), afectando a las columnas `workclass`, `occupation` y `native.country`.

Se identificaron y eliminaron características redundantes: la columna `education` se eliminó por ser una representación textual de `education.num`, y `fnlwgt` se descartó por ser un peso muestral no relevante para la predicción individual (Celda 6).

El análisis visual (Celda 8) reveló la presencia de outliers significativos en `capital.gain` y `capital.loss`, los cuales se decidieron mantener ya que son indicativos de ingresos altos.(ver Ilustración 1)

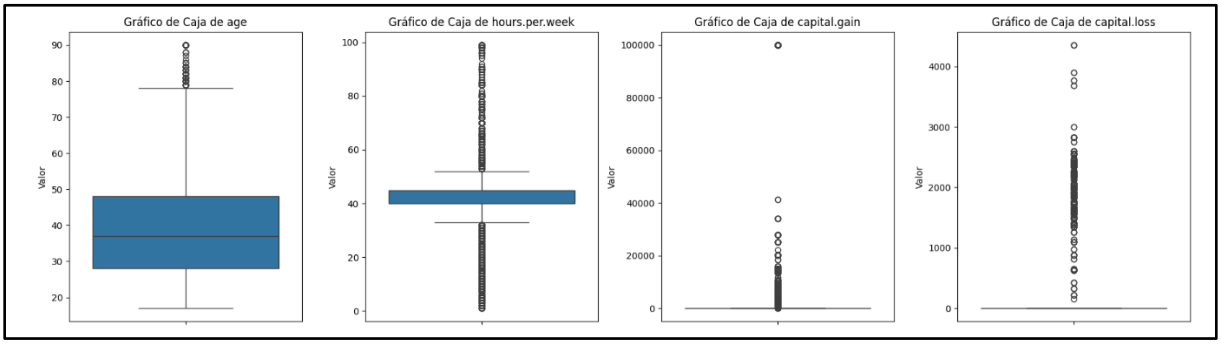

Ilustración 1 - Presencia de Outliers

Un hallazgo clave (Celda 9) fue el **fuerte desbalance de clases** en la variable objetivo `income`, con un 76% de registros en la clase `<=50K` y solo un 24% en `>50K`.(ver Ilustración 2)

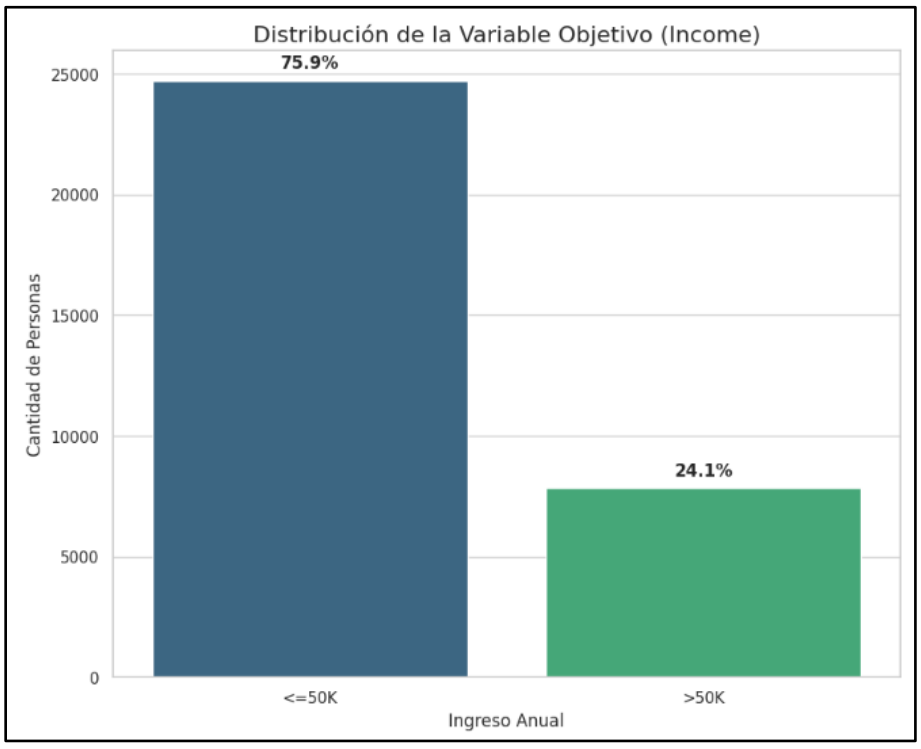

Ilustración 2- Desbalance de Clases

Finalmente, un mapa de calor (Celda 12) y gráficos bivariados (Celdas 10 y 11) mostraron correlaciones esperadas: `age`, `education.num`, `hours.per.week` y `marital.status` (específicamente 'Married-civ-spouse') mostraron una correlación positiva con ingresos más altos." (ver Ilustración 3)

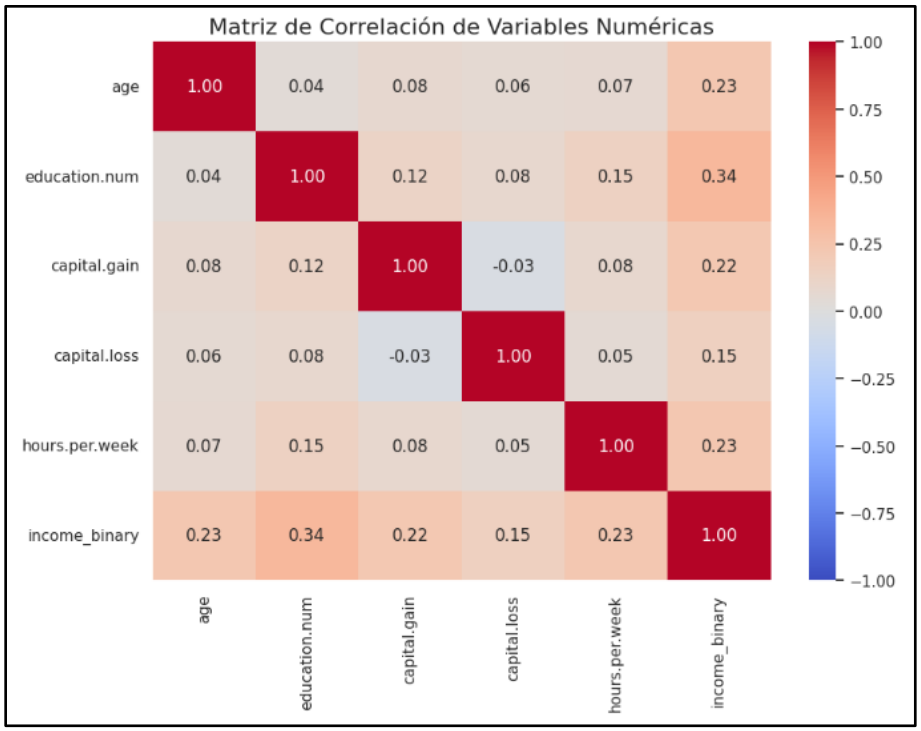

Ilustración 3 - Mapa de Calor

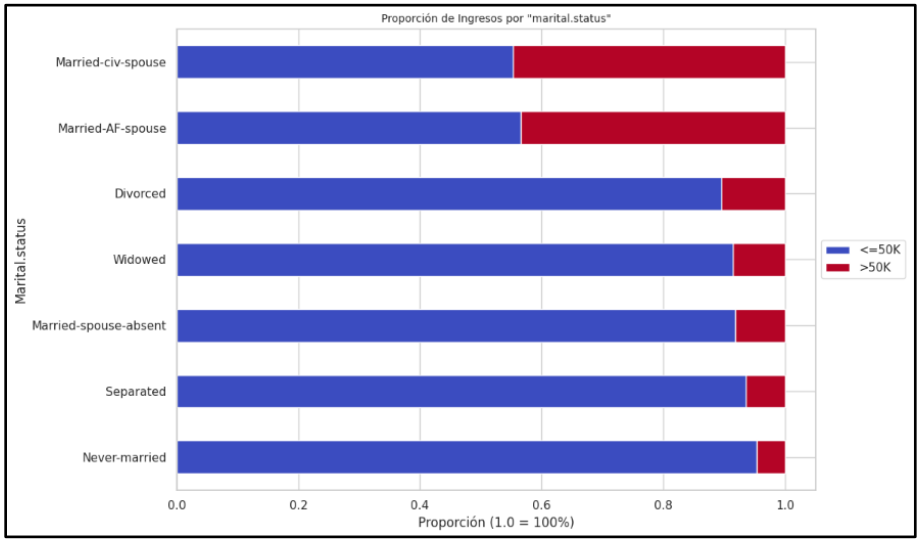

Ilustración 4 – Marital Status

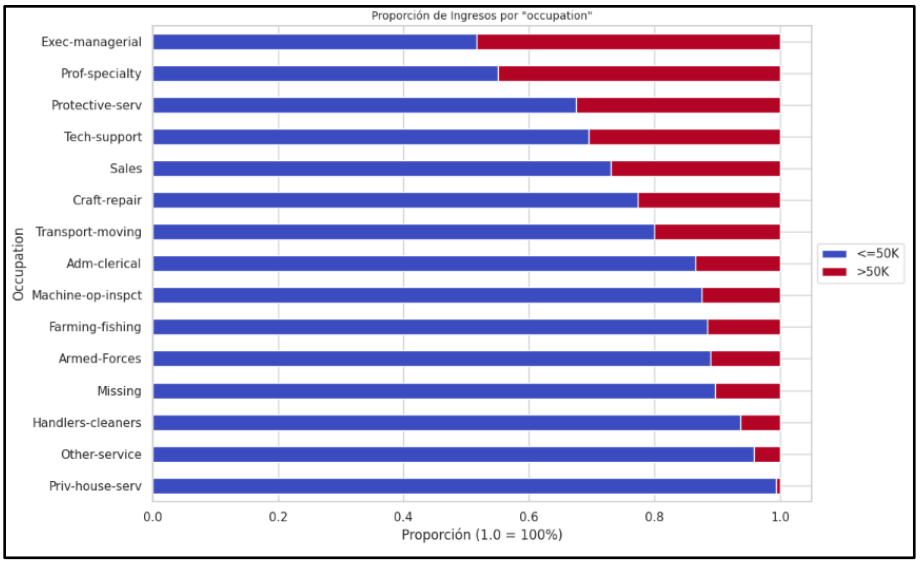

Ilustración 5 - Occupation

B. SELECCIÓN DE TÉCNICAS CANDIDATAS

Para establecer una línea base robusta (baseline), se seleccionaron tres técnicas clásicas de Machine Learning (ML) implementadas en el notebook (Celda 18) (Géron, 2019) (Murphy, 2012):

1. **Regresión Logística:** Como benchmark lineal simple y rápido.
2. **Random Forest:** Por su alta capacidad predictiva y robustez ante outliers.
3. **Support Vector Machine (SVM):** Por su efectividad en espacios de alta dimensión.

Adicionalmente, cumpliendo con los requisitos, se propusieron tres arquitecturas de Deep Learning (DL) para capturar relaciones no lineales complejas:

1. **MLP Básico:** Una red *feedforward* estándar para establecer un baseline de DL.
2. **MLP con Dropout:** Arquitectura idéntica a la básica, pero con regularización Dropout para mitigar el sobreajuste.
3. **Wide & Deep:** Una arquitectura híbrida que combina la memorización (camino *wide*) con la generalización (camino *deep*), ideal para este tipo de datos tabulares.

La selección de estos modelos se fundamenta en los siguientes 5 trabajos:

Para contextualizar el problema y fundamentar la selección de técnicas de aprendizaje automático, se realizó una revisión de literatura reciente relacionada con la predicción de pobreza, riesgo social y variables socioeconómicas mediante modelos de clasificación. Estos estudios emplean datos tabulares similares al dataset Adult, por lo que constituyen referencias válidas para este proyecto.

1. **Solís-Salazar y Madrigal-Sanabria (2022)** desarrollan un modelo de clasificación de pobreza utilizando XGBoost y lo comparan con el método tradicional Proxy Means Test. Sus resultados demuestran que los ensambles de árboles reducen significativamente los errores de clasificación, especialmente en contextos con variables heterogéneas y relaciones no lineales. Este hallazgo respalda el uso de modelos como Random Forest o Gradient Boosting en problemas de tipo

socioeconómico.

2. **De forma complementaria, Curto Merino (2023)** compara regresión logística, árboles de decisión y una red neuronal MLP para predecir pobreza en España. El autor concluye que la regresión logística sigue siendo un modelo competitivo y con alta interpretabilidad, mientras que la red neuronal ofrece un rendimiento levemente superior cuando se cuenta con un preprocesamiento adecuado. Este trabajo justifica el uso de la regresión logística como línea base y de las redes neuronales como modelos avanzados.
3. **En el estudio de Muñetón-Santa y Manrique-Ruiz (2023)**, se aplican Random Forest, CatBoost y LightGBM para estimar el índice de pobreza multidimensional en Medellín. Los resultados muestran que los ensambles de árboles logran los mejores niveles de rendimiento, reforzando la evidencia de que estas técnicas son particularmente eficaces para capturar interacciones complejas en datos tabulares similares al Adult Income Dataset.
4. **Desde el ámbito educativo, Smith Uldall y Gutiérrez Rojas (2022)** comparan regresión logística, árboles de decisión y redes neuronales en un sistema de alerta temprana para predecir deserción escolar. Su investigación destaca que los modelos de ML superan a las técnicas tradicionales y que las métricas de sensibilidad y recall son fundamentales en problemas donde la clase minoritaria tiene alto impacto social. Esto se relaciona directamente con el desbalance 76/24 del dataset Adult, donde la clase >50K requiere especial atención.
5. **Finalmente, la revisión sistemática realizada por Pincay-Ponce et al. (2022)** confirma que en estudios con datos socioeconómicos se utilizan de manera recurrente modelos como SVM, árboles de decisión, Random Forest y redes neuronales, lo cual coincide con las técnicas implementadas en el presente proyecto. La revisión destaca también la importancia de considerar factores como educación, ingreso familiar, género y participación laboral, variables que son equivalentes a las presentes en el dataset Adult.

En conjunto, estos antecedentes demuestran que los modelos más utilizados en problemas socioeconómicos de clasificación binaria son la regresión logística (por su interpretabilidad), los ensambles de árboles (por su capacidad predictiva) y las redes neuronales (por su manejo de patrones complejos). Con base en esta evidencia, este proyecto selecciona como técnicas candidatas la regresión logística, Random Forest y la arquitectura MLP / Wide & Deep, buscando equilibrar interpretabilidad, rendimiento y capacidad de generalización sobre la clase minoritaria.

Aunque los estudios muestran una clara tendencia a favor de ensambles y redes neuronales, también evidencian que la interpretabilidad sigue siendo clave en contextos socioeconómicos. Por eso la combinación RL + RF + MLP es coherente con las necesidades del proyecto.

C. COMPARACIÓN DE TÉCNICAS (EXPERIMENTACIÓN BASE)

Se realizó una comparación exhaustiva de las 3 técnicas de ML y las 3 arquitecturas de DL. Dadas las restricciones (calidad de datos con nulos y desbalance de clases), la evaluación se centró en el F1-Score (para la clase >50K) y el AUC-ROC.

- **Comparativa ML:** (Celda 18-20) La Regresión Logística y SVM (con class_weight='balanced') superaron a Random Forest en F1-Score, demostrando que la ponderación de clases era una estrategia efectiva.
- **Comparativa DL:** (Celda 63) La arquitectura Wide & Deep (F1: 0.6822) y el MLP con Dropout (F1:

0.6828) superaron marginalmente al MLP Básico (F1: 0.6775).

Este análisis demostró que las redes neuronales, aunque computacionalmente más costosas (como se vio en el entrenamiento del SVM), ofrecían un rendimiento superior en F1 y AUC. (ver Ilustración 6)

```
======================================================
TABLA COMPARATIVA FINAL - MODELOS DEEP LEARNING
======================================================
```

	Modelo	F1-Score (>50K)	AUC-ROC	Comentarios
0	1. MLP Básico (con class_weight)	0.677502	0.903341	Baseline de Deep Learning
1	2. MLP con Dropout (30%)	0.682752	0.906770	MLP Básico + Regularización
2	3. Wide & Deep (con class_weight)	0.682229	0.906386	Arquitectura Híbrida
3	4. MLP Optimizado (KerasTuner)	0.682746	0.907666	Modelo Refinado

Ilustración 6 – Tabla Comparativa

D. REQUISITOS DEL PROYECTO (DETALLADO)

Se identificaron los siguientes requisitos funcionales y no funcionales:

- **Requisitos Funcionales:**
  - **RF-01:** El sistema debe recibir las 12 características de entrada (age, workclass, etc.) de un individuo.
  - **RF-02:** El sistema debe procesar los datos de entrada (escalar, codificar) de la misma forma que el modelo fue entrenado.
  - **RF-03:** El sistema debe retornar una predicción binaria (0 para <=50K o 1 para >50K).
- **Requisitos No Funcionales:**
  - **RNF-01 (Rendimiento):** El tiempo de predicción para una sola instancia debe ser inferior a 1 segundo.
  - **RNF-02 (Escalabilidad):** La arquitectura (ej. API) debe ser capaz de manejar N peticiones concurrentes.
  - **RNF-03 (Explicabilidad):** El sistema debe proveer una justificación de por qué el modelo tomó una decisión (cubierto con SHAP).
  - **RNF-04 (Monitoreabilidad):** Se deben registrar las predicciones y su *drift* (cambio en la distribución de datos) a lo largo del tiempo.
  - **RNF-05 (Seguridad):** El sistema debe garantizar que los datos ingresados por los usuarios sean procesados y almacenados de forma segura, evitando accesos no autorizados. Esto incluye la validación de entradas para prevenir inyecciones, el uso de canales de comunicación seguros durante el consumo del modelo y la protección del pipeline de inferencia ante manipulaciones externas.
  - **RNF-06 (Confiabilidad):** El sistema debe mantener un comportamiento estable y consistente durante su operación, asegurando que las predicciones se generen bajo las mismas condiciones que en el entrenamiento. Esto implica mecanismos que verifiquen la integridad del modelo y del preprocesador, además de un funcionamiento estable del backend y frontend bajo carga moderada.

E. SELECCIÓN Y REFINAMIENTO DE LA ARQUITECTURA (DL)

Basado en la experimentación inicial (Celda 63), se pre-seleccionó la arquitectura **MLP con Dropout** como la más prometedora debido a su alto F1-Score y estabilidad.

Justificación de la Arquitectura:

- **Entrada:** 87 neuronas (adaptada a la dimensionalidad tras el One-Hot Encoding).
- **Capas Ocultas y Regularización:** Se diseñó una red profunda (64 y 32 neuronas) con capas de **Dropout (0.3)**. La elección del Dropout es crítica para este proyecto: dado el desbalance de clases, el modelo tiende a memorizar la clase mayoritaria; el Dropout fuerza a la red a aprender características más robustas y generalizables (Goodfellow, Bengio, & Courville, 2016).
- **Salida:** Neurona Sigmoid para clasificación binaria.

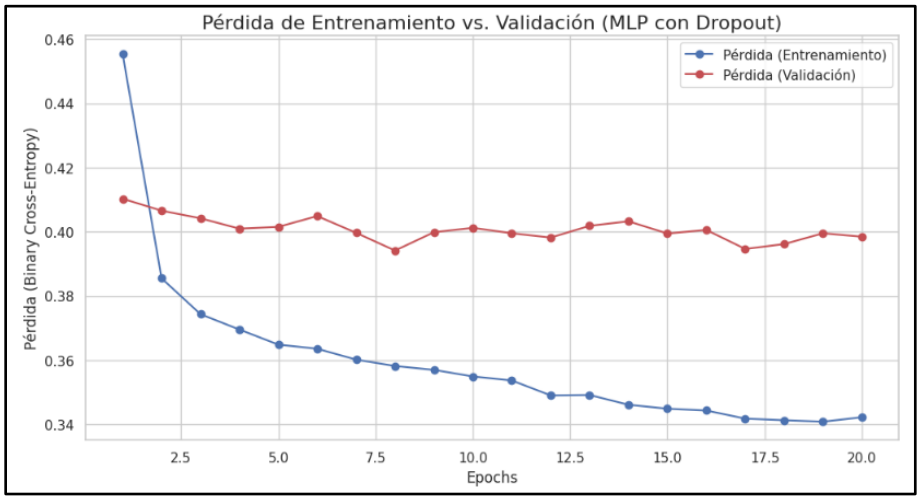

Ilustración 7 - Analisis de Convergencia

Proceso de Refinamiento (KerasTuner): Para validar si esta arquitectura manual era la óptima, se ejecutó una búsqueda de hiperparámetros (RandomSearch) explorando diferentes rangos de neuronas (32-128), tasas de dropout (0.2-0.4) y tasas de aprendizaje (ver Ilustración 7). Este proceso generó un modelo candidato 'Optimizado' para ser comparado en la fase final contra la arquitectura diseñada manualmente (ver Ilustración 8).

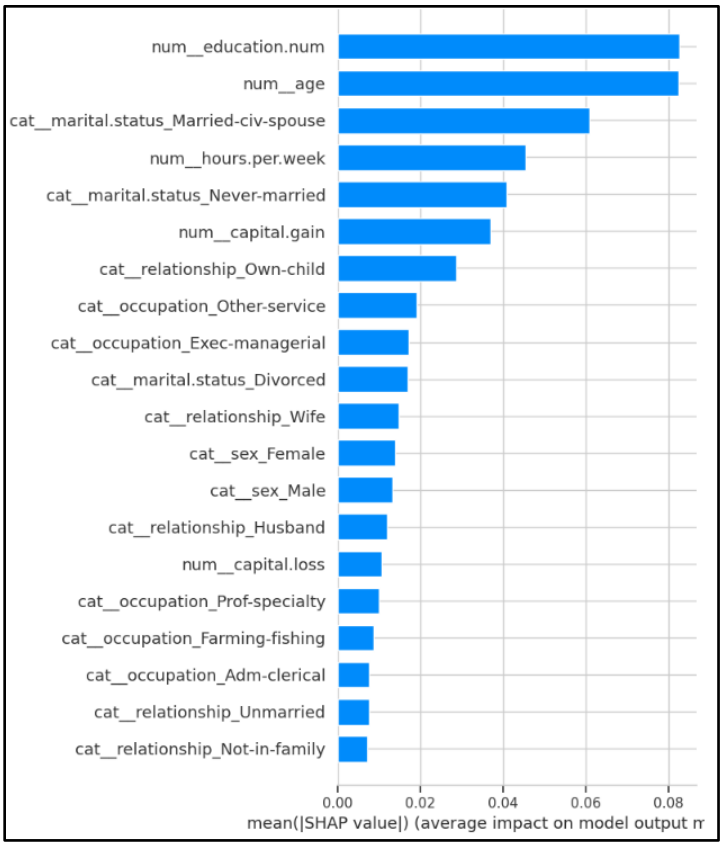

Ilustración 8 – Caracteristicas Importantes.

F. ELABORACIÓN DE MODELOS (BALANCEO)

El modelo se construyó sobre datos divididos en 70% para entrenamiento (22,792 muestras) y 30% para prueba (9,769 muestras), usando stratify=y para preservar la distribución de clases. Un 20% de los datos de entrenamiento se usó como conjunto de validación durante el *fit*.

Dado el desbalance de 76/24 en la clase income, se implementaron y analizaron tres técnicas de balanceo en los modelos de ML y DL:

1. **Baseline (Sin Balanceo):** (Celda 23) Sirvió como punto de partida.
2. **SMOTE (Sobremuestreo):** (Celda 25) Mejoró significativamente el Recall de la clase minoritaria, pero a costa de la Precisión.
3. **Class Weight (Ponderación):** (Celdas 27, 37) Ofreció el mejor equilibrio entre F1-Score y AUC-ROC, penalizando al modelo por errores en la clase >50K. Esta fue la técnica seleccionada para toda la experimentación de Deep Learning."

```
--- Entrenando Modelos Baseline (Sin Balanceo) ---
Entrenando Regresión Logística...

Resultados para Regresión Logística (Baseline):
              precision    recall  f1-score   support

       <=50K       0.88      0.93      0.90      7417
        >50K       0.73      0.60      0.66      2352

    accuracy                           0.85      9769
   macro avg       0.80      0.77      0.78      9769
weighted avg       0.84      0.85      0.85      9769

Entrenando Random Forest...

Resultados para Random Forest (Baseline):
              precision    recall  f1-score   support

       <=50K       0.88      0.92      0.90      7417
        >50K       0.70      0.61      0.65      2352

    accuracy                           0.84      9769
   macro avg       0.79      0.76      0.78      9769
weighted avg       0.84      0.84      0.84      9769

Entrenando SVM (Kernel RBF)...

Resultados para SVM (Kernel RBF) (Baseline):
              precision    recall  f1-score   support

       <=50K       0.88      0.94      0.91      7417
        >50K       0.75      0.59      0.66      2352

    accuracy                           0.85      9769
   macro avg       0.81      0.76      0.78      9769
weighted avg       0.85      0.85      0.85      9769


--- Tabla Comparativa Baseline ---
| Modelo              |   F1-Score (>50K) |   AUC-ROC |   Tiempo (seg) |
|:--------------------|------------------:|----------:|---------------:|
| Regresión Logística |            0.6600 |    0.9006 |         0.4212 |
| Random Forest       |            0.6536 |    0.8894 |        19.3491 |
| SVM (Kernel RBF)    |            0.6587 |    0.8974 |       133.2206 |
```

Ilustración 9 – Entrenando de Modelos sin Balanceo.

```
--- Entrenando Modelos Balanceados (con SMOTE) ---
Entrenando Regresión Logística con SMOTE...

Resultados para Regresión Logística (con SMOTE):
              precision    recall  f1-score   support

       <=50K       0.94      0.79      0.86      7417
        >50K       0.56      0.84      0.67      2352

    accuracy                           0.80      9769
   macro avg       0.75      0.82      0.77      9769
weighted avg       0.85      0.80      0.81      9769

Entrenando Random Forest con SMOTE...

Resultados para Random Forest (con SMOTE):
              precision    recall  f1-score   support

       <=50K       0.89      0.89      0.89      7417
        >50K       0.66      0.66      0.66      2352

    accuracy                           0.84      9769
   macro avg       0.78      0.78      0.78      9769
weighted avg       0.84      0.84      0.84      9769

Entrenando SVM (Kernel RBF) con SMOTE...

Resultados para SVM (Kernel RBF) (con SMOTE):
              precision    recall  f1-score   support

       <=50K       0.94      0.80      0.86      7417
        >50K       0.57      0.85      0.68      2352

    accuracy                           0.81      9769
   macro avg       0.76      0.82      0.77      9769
weighted avg       0.85      0.81      0.82      9769


--- Tabla Comparativa con SMOTE ---
| Modelo              |   F1-Score (>50K) |   AUC-ROC |   Tiempo (seg) |
|:--------------------|------------------:|----------:|---------------:|
| Regresión Logística |            0.6729 |    0.8996 |         2.2688 |
| Random Forest       |            0.6597 |    0.8839 |        40.0380 |
| SVM (Kernel RBF)    |            0.6800 |    0.8976 |       420.4616 |
```

Ilustración 10 – Entrenando Modelos Balanceados (con SMOTE)

```
--- Entrenando Modelos Balanceados (con class_weight) ---
Entrenando Regresión Logística (CW)...

Resultados para Regresión Logística (CW) (con class_weight):
              precision    recall  f1-score   support

       <=50K       0.94      0.79      0.86      7417
        >50K       0.56      0.84      0.67      2352

    accuracy                           0.80      9769
   macro avg       0.75      0.81      0.77      9769
weighted avg       0.85      0.80      0.81      9769

Entrenando Random Forest (CW)...

Resultados para Random Forest (CW) (con class_weight):
              precision    recall  f1-score   support

       <=50K       0.89      0.91      0.90      7417
        >50K       0.69      0.63      0.66      2352

    accuracy                           0.84      9769
   macro avg       0.79      0.77      0.78      9769
weighted avg       0.84      0.84      0.84      9769

Entrenando SVM (Kernel RBF) (CW)...

Resultados para SVM (Kernel RBF) (CW) (con class_weight):
              precision    recall  f1-score   support

       <=50K       0.94      0.79      0.86      7417
        >50K       0.56      0.85      0.68      2352

    accuracy                           0.80      9769
   macro avg       0.75      0.82      0.77      9769
weighted avg       0.85      0.80      0.82      9769


--- Tabla Comparativa con class_weight ---
| Modelo                    |   F1-Score (>50K) |   AUC-ROC |   Tiempo (seg) |
|:--------------------------|------------------:|----------:|---------------:|
| Regresión Logística (CW)  |            0.6719 |    0.9009 |         0.4076 |
| Random Forest (CW)        |            0.6591 |    0.8879 |        21.0437 |
| SVM (Kernel RBF) (CW)     |            0.6766 |    0.8997 |       186.0037 |
```

Ilustración 11 – Entrenando Modelos Balanceados (con Class Weight)

G. DESARROLLO DE FRONTEND Y BACKEND

Para cumplir con los requisitos de usabilidad y escalabilidad, se implementó un prototipo funcional basado en micro-servicios web utilizando el framework **Streamlit**. La arquitectura del sistema integra los siguientes componentes:

1. **Backend y Motor de IA:** Se desarrolló un pipeline de inferencia en Python que gestiona la comunicación entre la interfaz y el modelo. El sistema utiliza carga en caché para optimizar el acceso al modelo neuronal (.keras) y al preprocesador (.joblib). La función make_prediction() procesa los datos de entrada aplicando las mismas transformaciones del entrenamiento antes de consultar al modelo.
2. **Frontend (Interfaz de Usuario):** La interfaz gráfica permite la interacción intuitiva mediante un panel de control con selectores validados para las 12 variables sociodemográficas. Los resultados se presentan visualmente, indicando la clasificación de ingreso (>50K o <=50K) junto con su probabilidad de confianza.
3. **Explicabilidad (XAI):** Se integró la librería **SHAP** para proporcionar transparencia en las decisiones. Cada predicción se acompaña de un **Gráfico de Fuerza (Force Plot)** dinámico que visualiza qué

características específicas (como edad o estado civil) influyeron positiva o negativamente en el resultado final.

4. **Monitoreo:** El sistema incluye un módulo de registro (*logging*) que almacena cada consulta en un archivo persistente. Esto permite visualizar estadísticas de uso y la distribución de las predicciones en tiempo real, facilitando la detección temprana de desviaciones en los datos o en el comportamiento del modelo.

Ilustración 12 – App Funcionando en entorno Local

5. **Seguridad:** Para asegurar que la aplicación sea adecuada para entornos reales, se incorporan medidas básicas de protección en la comunicación entre frontend y backend. El backend valida todas las entradas del usuario antes de enviarlas al pipeline de inferencia, evitando que valores inesperados o malformados afecten el modelo o produzcan errores de ejecución. Asimismo, se contempla el uso de canales de comunicación cifrados (HTTPS) durante un despliegue productivo, con el fin de proteger la transmisión de datos sensibles como edad, estado civil u ocupación. Estas medidas permiten que el sistema opere de forma segura incluso en escenarios donde interactúan múltiples usuarios simultáneamente.

6. **Escalabilidad:** El prototipo está diseñado para ejecutarse inicialmente de forma local, pero su arquitectura permite escalar a un despliegue en la nube mediante contenedores o servicios de hosting de aplicaciones (por ejemplo, AWS, Render o Google Cloud). El backend puede manejar múltiples peticiones concurrentes al cargar el modelo y el preprocesador en memoria de forma persistente, evitando tiempos de carga repetitivos. Además, la separación entre frontend (Streamlit) y backend permite distribuir la carga, ejecutar varias réplicas del servicio y balancear tráfico en escenarios de alto uso. Esto asegura que el sistema pueda crecer sin comprometer tiempos de respuesta ni estabilidad.

4. CONSIDERACIONES DE DESPLIEGUE Y ÉTICA

4.1. ANÁLISIS DE FAIRNESS Y CONSIDERACIONES ÉTICAS

El despliegue de un sistema de inteligencia artificial para la predicción de ingresos conlleva responsabilidades éticas significativas. Dado que el modelo trabaja con datos socioeconómicos sensibles,

es imperativo analizar las implicaciones de su transición de un entorno experimental a uno productivo para garantizar un uso justo y responsable.

4.1.1. IDENTIFICACIÓN DE VARIABLES SENSIBLES Y SESGOS HISTÓRICOS

El dataset 'Adult' incluye explícitamente atributos protegidos como `sex` (sexo), `race` (raza) y `native.country` (origen nacional). Si bien el análisis exploratorio demostró que estas variables poseen poder predictivo estadístico, su uso directo conlleva el riesgo de perpetuar sesgos históricos presentes en los datos originales de 1994, una época marcada por brechas salariales estructurales significativas. Entrenar el modelo sin cuestionar estas correlaciones implica aceptar esas desigualdades pasadas como reglas válidas para el futuro.

4.1.2. RIESGO DE SESGO ALGORÍTMICO EN PRODUCCIÓN

Existe un riesgo latente de que el modelo aprenda a penalizar injustamente a ciertos subgrupos (como mujeres o minorías raciales) no por sus capacidades individuales, sino por correlaciones engañosas o prejuicios sistémicos en los datos de entrenamiento. Si este modelo se utilizara en escenarios reales, como el otorgamiento de créditos bancarios o el filtrado automático de currículums, un sesgo no detectado podría automatizar la discriminación a gran escala, excluyendo sistemáticamente a grupos subrepresentados y generando consecuencias legales y éticas graves para la organización.

4.1.3. ESTRATEGIAS DE MITIGACIÓN PROPUESTAS

Para asegurar un despliegue ético y robusto, se define la siguiente estrategia de mitigación en tres niveles:

- **Pre-Procesamiento (Datos):** Se recomienda aplicar técnicas como **'Reweighting'** (reasignar pesos a las muestras de entrenamiento) para equilibrar la importancia de los grupos subrepresentados antes de que el modelo aprenda de ellos. Alternativamente, se puede optar por **'Fairness through Unawareness'**, eliminando deliberadamente las variables sensibles del conjunto de entrada, aunque esto debe hacerse con cuidado para evitar variables *proxy* (variables que correlacionan con la protegida).
- **Validación (Métricas):** Antes de la puesta en marcha, se auditará el modelo utilizando métricas de paridad, como la *Tasa de Falsos Positivos* diferenciada por grupo, para asegurar que el error se distribuya equitativamente.
- **Transparencia (Operación):** El sistema en producción utilizará el módulo de explicabilidad (SHAP) integrado en la aplicación. Esto permite que cada predicción sea auditada por un humano, visualizando si una variable sensible jugó un rol determinante en la decisión, evitando así el efecto de "caja negra".

5. ESTRATEGIA DE MONITOREO Y MANTENIMIENTO

Un modelo de Machine Learning no es un artefacto estático; su rendimiento tiende a degradarse a medida que los datos del mundo real cambian respecto a los datos de entrenamiento. Para garantizar la robustez del sistema a largo plazo, se define la siguiente estrategia de monitoreo continuo:

5.1. DETECCIÓN DE DATA DRIFT (DESVIACIÓN DE DATOS)

- **Concepto:** Se vigilará si cambia la distribución estadística de las variables de entrada (ej. si la edad promedio de los usuarios aumenta significativamente o si cambia la distribución geográfica).
- **Métrica:** Se implementará el cálculo mensual del **PSI (Population Stability Index)** para las 5 variables más influyentes según SHAP. Si el PSI de alguna variable supera el umbral de 0.2, se activará una alerta automática para revisión técnica.

5.2. MONITOREO DE PREDICTION DRIFT

- **Concepto:** Se analizará la distribución de las salidas del modelo en el tiempo.
- **Acción:** Se comparará semanalmente el porcentaje de predicciones positivas (>50K) versus el histórico. Un cambio brusco en esta proporción sin una justificación económica externa podría indicar un fallo en el modelo o un cambio en la población objetivo.

5.3. MÉTRICAS DE DESEMPEÑO (GROUND TRUTH)

- **Concepto:** En la medida que se obtengan etiquetas reales (ej. verificando los ingresos reales meses después de la predicción), se recalculará el rendimiento del modelo.
- **Umbral Crítico:** Se monitoreará específicamente el **F1-Score** de la clase minoritaria. Si este indicador cae por debajo de 0.65 (actualmente 0.68), se desencadenará un proceso de re-entrenamiento del modelo con datos más recientes.

6. RESULTADOS

La evaluación final del desempeño de los modelos se realizó utilizando datos no vistos durante el entrenamiento, correspondientes al conjunto de prueba (30%). Esto garantiza que las métricas obtenidas reflejan la capacidad real de generalización del modelo.

La etapa final de evaluación se llevó a cabo utilizando el conjunto de prueba (*Test Set*, 30%). Tras comparar cuatro estrategias de Deep Learning, los resultados arrojaron un hallazgo interesante sobre la efectividad de la regularización:

6.1. DESEMPEÑO CUANTITATIVO

Contrario a la hipótesis inicial de que la arquitectura híbrida o el modelo optimizado automáticamente serían superiores, el modelo **MLP con Dropout (30%)** emergió como el de mejor desempeño global.

- **Mejor F1-Score:** El MLP con Dropout alcanzó un **0.6836**, superando al modelo *Wide & Deep* (0.6812) y al modelo ajustado con *KerasTuner* (0.6804). Esto indica que la regularización simple fue la estrategia más efectiva para generalizar sobre la clase minoritaria.
- **Mejor AUC-ROC:** Este modelo también lideró marginalmente en capacidad de discriminación con un AUC de **0.9071**, empatando técnicamente con el modelo optimizado.
- **Selección Final:** Se seleccionó el **MLP con Dropout** como el modelo definitivo. Su arquitectura,

aunque más simple que la *Wide & Deep*, demostró ser la más robusta, sugiriendo que para este dataset en particular, evitar el sobreajuste mediante Dropout es más crítico que capturar interacciones de características complejas (ver Ilustración 13).

```
======================================================
TABLA COMPARATIVA FINAL - MODELOS DEEP LEARNING
======================================================
```

	Modelo	F1-Score (>50K)	AUC-ROC	Comentarios
0	1. MLP Básico (con class_weight)	0.677392	0.902375	Baseline de Deep Learning
1	2. MLP con Dropout (30%)	0.683609	0.907146	MLP Básico + Regularización
2	3. Wide & Deep (con class_weight)	0.681199	0.905831	Arquitectura Híbrida
3	4. MLP Optimizado (KerasTuner)	0.680447	0.907056	Modelo Refinado

Ilustración 13 – Tabla Comparativa Final – Modelos Deep Learning

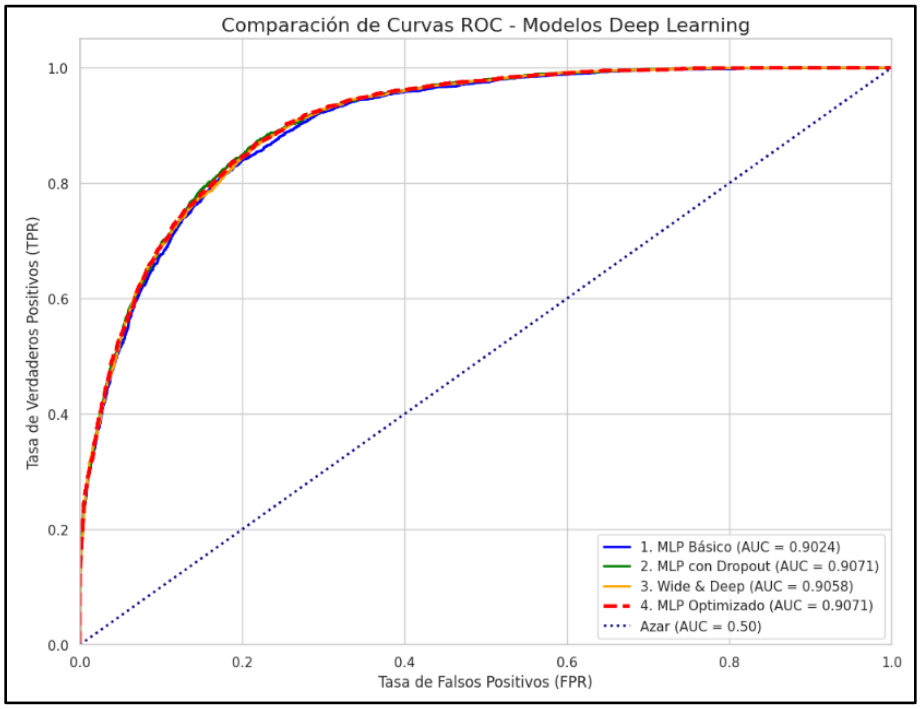

Ilustración 14 – Grafico Curva ROC

6.2. ANÁLISIS INTERPRETATIVO (EXPLICABILIDAD CON SHAP)

Para cumplir con el requisito de explicabilidad, se utilizó SHAP (SHapley Additive exPlanations) (Géron, 2019) . El análisis reveló qué factores influyen más en la predicción de altos ingresos (ver Ilustración 15 e Ilustración 16) :

- **Estado Civil (marital.status / relationship):** El gráfico de resumen (Beeswarm) muestra que la característica cat_relationship_Husband es el predictor más fuerte. Estar casado (True) impulsa fuertemente la probabilidad hacia >50K.

- **Edad (age) y Educación (education.num):** Ambas variables muestran una relación lineal positiva: a mayor edad y mayor nivel educativo, mayor es la probabilidad de ingresos altos.
- **Ganancias de Capital (capital.gain):** Aunque muchos individuos tienen valor 0, aquellos con ganancias de capital altas (puntos rojos en el gráfico) tienen un impacto positivo extremadamente alto en la predicción."

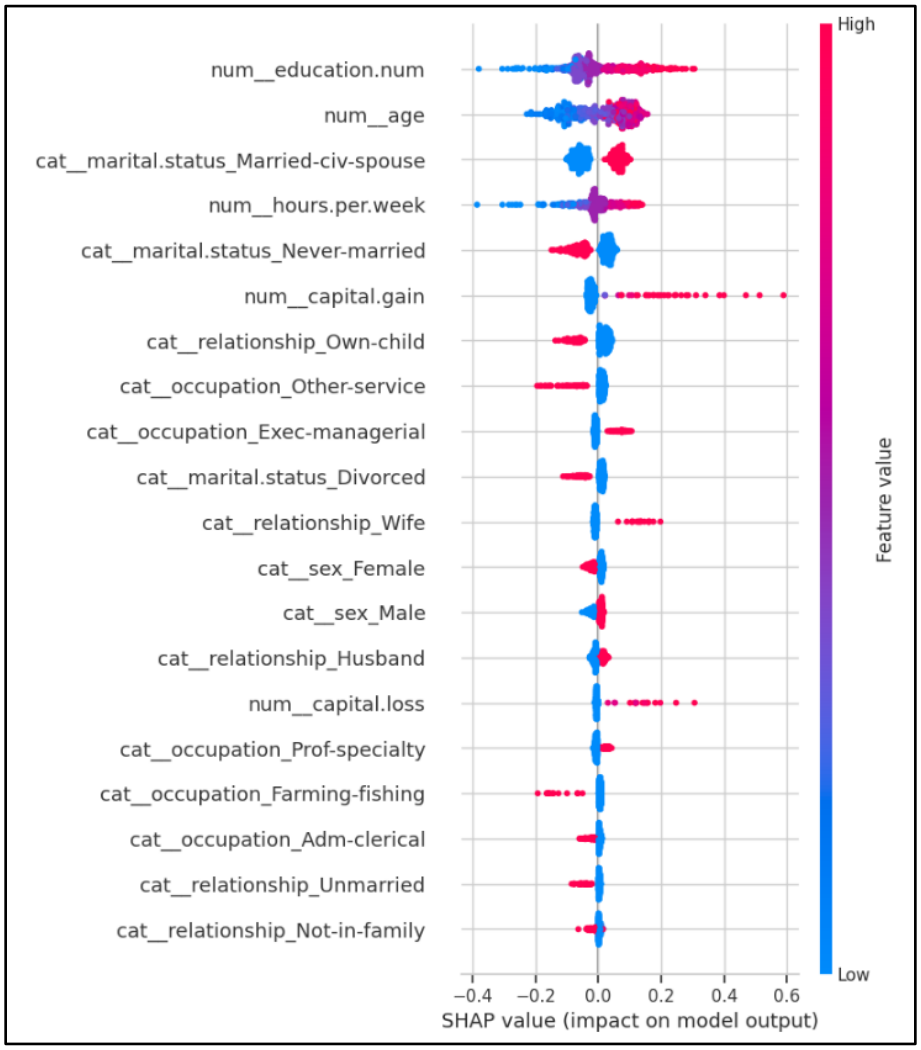

Ilustración 15 - Grafico Beeswarm

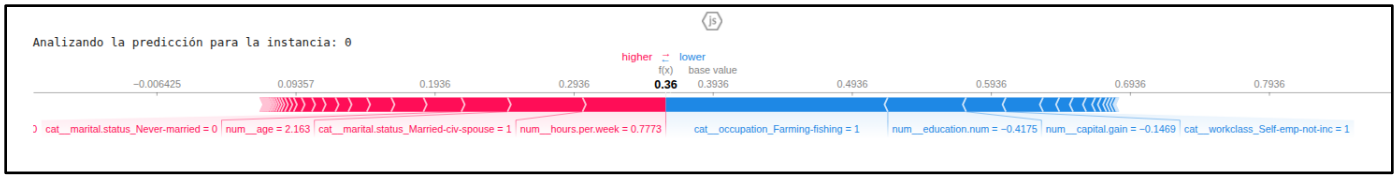

Ilustración 16 – Grafico Force Plot

6.3. DISCUSIÓN E INTERPRETACIÓN ANALÍTICA

Más allá de las métricas de rendimiento, el análisis de los resultados ofrece *insights* profundos sobre la naturaleza del problema y la arquitectura de la solución:

6.3.1. ¿POR QUÉ EL MODELO CON DROPOUT SUPERÓ A ARQUITECTURAS MÁS COMPLEJAS?

El hecho de que el **MLP con Dropout (30%)** superara tanto a la arquitectura híbrida *Wide & Deep* como al modelo optimizado automáticamente revela una característica crítica de los datos: **la presencia de ruido y solapamiento entre clases**. En el dataset 'Adult', existen individuos con características demográficas casi idénticas, pero con ingresos diferentes. En este escenario, los modelos más complejos o sobre-optimizados tienden a "memorizar" estas inconsistencias (sobreajuste). La capa de Dropout actuó como un filtro de robustez, "apagando" aleatoriamente neuronas durante el entrenamiento y forzando a la red a aprender patrones generalizables en lugar de depender de correlaciones débiles o ruidosas (Goodfellow, Bengio, & Courville, 2016). Esto demuestra que, para datos tabulares con desbalance, la **regularización** suele ser más determinante que la complejidad arquitectónica.

6.3.2. INTERPRETACIÓN DE LAS VARIABLES CLAVE (SHAP) EN EL CONTEXTO DEL PROYECTO

El análisis de explicabilidad (Figura SHAP) desvela la lógica socioeconómica que el modelo ha aprendido:

- **La Hegemonía del Capital sobre el Trabajo**
  La variable `capital.gain` (ganancia de capital) muestra un impacto positivo extremo y abrupto. Esto implica que el modelo ha detectado que la acumulación de activos financieros (inversiones) es un discriminador de riqueza mucho más potente que las variables laborales tradicionales. En el contexto del proyecto, esto significa que el modelo es excepcionalmente sensible a detectar perfiles de "inversionista".

- **El Factor de Estabilidad Social**
  La fuerte influencia positiva de `marital.status: Married-civ-spouse` y `age` sugiere que el modelo asocia los altos ingresos con la estabilidad y la acumulación de experiencia. El algoritmo penaliza la "juventud" y la "soltería" a menos que estén compensadas por altos niveles educativos (`education.num`).

6.3.3. IMPLICANCIAS OPERATIVAS

Estas dependencias tienen una consecuencia directa para el despliegue: el modelo es **conservador**. Tenderá a clasificar correctamente a perfiles tradicionales de altos ingresos (mayores de 40 años, casados, con inversiones), pero podría tener una tasa de error más alta (falsos negativos) en perfiles atípicos emergentes, como jóvenes profesionales solteros con altos salarios, pero sin ganancias de capital acumuladas. Reconocer este sesgo operativo es vital para la toma de decisiones basada en este modelo.

7. PROPUESTA DE MEJORAS

7.1. IDENTIFICACIÓN DE LIMITACIONES

Analizamos por qué no llegamos a un F1-Score de 0.80 o más

- **Techo de Rendimiento por "Ruido" en los Datos:** A pesar de probar múltiples arquitecturas (MLP, Wide & Deep) y técnicas de balanceo, el F1-Score para la clase >50K se estancó alrededor de 0.68. Esto sugiere que existe un límite inherente en la capacidad predictiva de las variables actuales, es decir, hay personas con las mismas características demográficas (misma edad, educación, ocupación) que tienen ingresos diferentes, lo que hace imposible una clasificación perfecta sin más datos.
- **Ineficacia de la Búsqueda Aleatoria (Random Search):** Una limitación técnica importante fue que el proceso de *refinamiento* automático (usando KerasTuner con RandomSearch y 10 intentos) no logró superar al diseño manual con Dropout. Esto indica que el espacio de búsqueda definido fue insuficiente o que la búsqueda aleatoria no logró converger al óptimo global en el tiempo asignado.

2. Mejoras Propuestas

Proponemos soluciones técnicas concretas para la Fase 3

- **Mejora 1: Optimización Bayesiana de Hiperparámetros:** En lugar de usar una búsqueda aleatoria (RandomSearch), se propone implementar **Optimización Bayesiana**. Esta técnica utiliza los resultados de los intentos previos para decidir cuál es la mejor combinación de hiperparámetros a probar después. Esto permitiría explorar el espacio de búsqueda de manera mucho más eficiente, ajustando no solo neuronas y dropout, sino también la tasa de aprendizaje y el *batch size* con mayor precisión, lo que podría desbloquear el rendimiento extra que el modelo manual con Dropout ya ha insinuado.
- **Mejora 2: Estrategia de Ensamble (Stacking o Voting):** Dado que el modelo **MLP con Dropout** y el modelo **Wide & Deep** obtuvieron resultados extremadamente similares (F1: 0.6836 vs 0.6812) pero utilizan arquitecturas internas diferentes (una profunda vs. una híbrida), es muy probable que cometan errores en instancias distintas. Se propone implementar un **Ensamble de Modelos (Stacking)**. Esta técnica entrenaría un "meta-modelo" (como una Regresión Logística simple) que tome las predicciones de ambas redes neuronales como entrada para emitir el veredicto final, lo que teóricamente reduciría la varianza y aumentaría el F1-Score final.

8. CÓDIGO FUENTE EN GITHUB

El desarrollo técnico completo de este proyecto, incluyendo los notebooks de análisis (.ipynb), los scripts de preprocesamiento y la documentación técnica asociada, se encuentra alojado y versionado en el siguiente repositorio público de GitHub:

- **Enlace al Repositorio**

  Link: https://github.com/MaidoniaN/ACIF104-Sumativa2-Grupo13

  El repositorio está organizado siguiendo buenas prácticas de desarrollo y cuenta con un archivo README.md en la raíz que detalla la estructura del proyecto, lista las librerías necesarias (en requirements.txt) y proporciona instrucciones paso a paso para la instalación del entorno y la ejecución de los modelos."

- **Video demostrativo APP**

  Link: https://youtu.be/LIrMImaIZcQ

  Adicionalmente, como una buena práctica, hemos realizado un video demostrativo de la Aplicación,

el cual ha sido subido a Youtube en el siguiente enlace, aquí se demuestra como el modelo es implementado en una APP que mezcla Front-End y Back-End más el modelo.

9. BIBLIOGRAFÍA

- Cid R., A., Espinoza C, S., & Mattioni A., C. (2025). *ACIF 104 - Aprendizaje Maquina - Informe Formativa 3.* Santiago.
- Cid R., A., Espinoza C., S., & Mattioni A., C. (2025). *ACIF 104 - Aprendizaje Maquina - Informe Formativa 4.* Santiago.
- Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O' REILLY.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* London: The MIT Press.
- Merino, C. (2023). *Comparación de diferentes técnicas para predecir la pobreza (Trabajo de fin de grado). Universidad de Valladolid.* Obtenido de https://uvadoc.uva.es/handle/10324/63022
- Muñetón-Santa, & Manrique-Ruiz. (2023). *Predicting Multidimensional Poverty with Machine Learning Algorithms: An Open Data Source Approach Using Spatial Data. Social Sciences, 12(5), 296.* Obtenido de https://intellectum.unisabana.edu.co/handle/10818/62548
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* London: The MIT Press.
- Pincay-Ponce et al. (2022). *Analítica de datos de factores socioeconómicos que inciden en el rendimiento escolar: Revisión sistemática*. Obtenido de https://www.researchgate.net/publication/370802446
- Silva, O. (2025). *Consolidado Fase 1 y Fase 2 (Material de Clase).* Santiago: Universidad Andres Bello.
- Silva, O. (2025). *Libros Jupyter Notebook vistos en Clase.* Universidad Andres Bello.
- Smith Uldall, & Gutiérrez Rojas. (2022). *Una aplicación de aprendizaje automático en políticas públicas: Predicción de alerta temprana de deserción escolar en el sistema de educación pública de Chile. Multidisciplinary Business Review,*. Obtenido de https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-39922022000100020
- Solís-Salazar, & Madrigal-Sanabria. (2022). *Una propuesta de aprendizaje automático para predecir la pobreza. Revista Tecnología en Marcha, 35(4), 84–94.* Obtenido de https://revistas.tec.ac.cr/index.php/tec_marcha/article/view/5766