

# Guía para realizar el proyecto ABP

Sitio: [Instituto Superior Politécnico Córdoba](#)  
Curso: Práctica profesionalizante II - TSCDIA - 2023  
Libro: Guía para realizar el proyecto ABP

Imprimido por: María Laura PERALTA FERRER  
Día: miércoles, 14 mayo 2025, 9:40 PM

# Tabla de contenidos

## 1. Proyecto Final: Construcción de un Modelo

- 1.1. Consignas Generales
- 1.2. Semana 1 y 2: Definición del Problema
- 1.3. Semana 3 y 4: Análisis Exploratorio (EDA) y Fairlearn
- 1.4. Semana 4 y 5: Preparando los Datos
- 1.5. Semana 6 y 7: Modelado Inicial y Experimentación con MLflow
- 1.6. Semana 8 y 9 - Despliegue del Modelo en Producción

# 1. Proyecto Final: Construcción de un Modelo

**Objetivo:** Desarrollar un modelo que resuelva un problema real, aplicando metodologías de gestión (TDSP), tracking experimental (MLflow) y equidad (Fairlearn).

## 1.1. Consignas Generales

### 1. Dataset:

- Usar un dataset real (ej.: De [Kaggle](#) o uno local relevante para su comunidad).
- Justificar la elección (impacto social, disponibilidad de características sensibles).

### 2. Entregables Obligatorios:

- Repositorio GitHub con código, documentación TDSP y datos (versionados).

### 3. Herramientas:

- **Gestión:** Jira + TDSP.
- **Técnicas:** MLflow, Fairlearn, Scikit-learn.

## 1.2. Semana 1 y 2: Definición del Problema

- **Actividades:**

1. Investigar el dataset elegido y plantear una pregunta de negocio (ej.: "¿El modelo de aprobación de créditos es justo entre géneros?").
2. Definir historias de usuario en Jira/Trello (ej.: "Como analista, quiero explorar los datos para identificar sesgos").
3. Asignar roles (Project Manager, Data Engineer, Data Scientist, Ethical Reviewer).
4. Crear el **Project Charter** ([Link a la plantilla](#))
5. Configurar repositorio GitHub con estructura TDSP ([Link al repo de ejemplo](#)).

- **Entregable:** Documento PDF con objetivo, stakeholders y métricas de éxito (técnicas y de equidad).

## 1.3. Semana 3 y 4: Análisis Exploratorio (EDA) y Fairlearn

### Objetivos:

1. Entender la estructura y calidad del dataset.
2. Identificar sesgos en datos y variables sensibles.
3. Documentar hallazgos técnicos y éticos para guiar el preprocesamiento.

### Actividades Detalladas

#### Fase 1: Comprensión Inicial del Dataset:

- Carga y Descripción General
- Identificación de Variables Clave:
  - **Variables Sensibles:** Género, etnia, edad, ubicación geográfica.
  - **Variable Objetivo:** Lo que se quiere predecir (ej.: aprobación de crédito).
  - **Features Predictoras:** Ingresos, historial crediticio, educación.

#### Fase 2: Análisis Técnico (EDA Clásico):

- **Estadísticas Descriptivas**
- **Visualizaciones Clave:**
  - Distribuciones: Histogramas, boxplots.
  - Correlaciones: Heatmap de correlaciones.
- **Herramientas Automatizadas para EDA:** Pandas Profiling o SweetViz

#### Fase 3: Análisis Ético (Detección de Sesgos)

- **Métricas de Equidad en Datos:** Disparidad Demográfica, Prueba de Chi-Cuadrado.
- **Fairlearn Dashboard** (Análisis Visual):
  - Distribución de la variable objetivo por grupo.
  - Disparidades en tasas de aprobación/rechazo.
  - Comparación de métricas básicas (ej.: proporciones).
- **Identificación de Variables Proxy**

### Entregables de la Semana 3 y 4

- **Crear el documento de definición de datos:** ["data\\_definition"](#)
- **Jupyter Notebook con EDA:**
  - Código ejecutable + comentarios explicativos.
  - Gráficos interactivos (Plotly) o estáticos (Matplotlib/Seaborn).
- **Informe PDF de Hallazgos:**
  - **Sección Técnica:**
    - Distribuciones, correlaciones, valores faltantes.
    - Ejemplo: "El 30% de los datos de ingresos están incompletos en el grupo 'género no binario'".
  - **Sección Ética:**
    - Disparidades identificadas (ej.: "La tasa de aprobación para mujeres es un 18% menor que para hombres").
    - Variables proxy detectadas (ej.: "El código postal explica el 40% de la varianza en la etnia").

## 1.4. Semana 4 y 5: Preparando los Datos

**Objetivo:** Aprender a limpiar y transformar los datos para que sean *justos* y *útiles* antes de entrenar un modelo

### Paso 1: Limpieza de Datos

**Problemas Comunes:**

- **Valores faltantes:** ¡Datos incompletos! (ej.: una fila sin "ingreso").
- **Outliers:** Valores extremos (ej.: alguien con 200 años de edad).
- **Errores tipográficos:** "Femenno" en lugar de "Femenino".

### Paso 2: Codificar Variables

**¿Qué es Codificar?**

Transformar datos *no numéricos* en números

**Ejemplo:**

- **Género:** "Femenino" → 0, "Masculino" → 1, "No binario" → 2.

Evita codificar variables sensibles si no es necesario (podrían causar sesgos).

### Paso 3: Mitigar Sesgos en los Datos

Imaginá que el 90% de los datos de aprobación de créditos son de hombres. El modelo podría aprender que es mejor aprobar créditos a hombres.

Balanceo de datos:

1. **Sobremuestreo:** Agregar más datos de grupos minoritarios.
2. **Submuestreo:** Reducir datos de grupos mayoritarios.

### Paso 4: Feature Engineering General

**Transformaciones Matemáticas:**

- **Normalización:** Escalar variables a un rango común (ej.: 0 a 1).
- **Logaritmos:** Para manejar valores extremos.

**Agregar Interacciones entre Variables:**

- **Ejemplo:** Combinar *edad* e *ingreso* para crear una variable de "poder adquisitivo por edad".

**Codificación de Texto.**

### Entregables de la Semana 4 y 5

- **Crear el documento de reporte de datos** ([data\\_summary](#))
- **Script de Preprocesamiento:**
  - Código que limpie, codifique y balancee los datos.
- **Dataset Procesado** (*datos\_procesados.csv*):
  - Datos listos para entrenar el modelo.

- **Informe Ético** (PDF):

- Explicación de cómo se mitigaron los sesgos



## 1.5. Semana 6 y 7: Modelado Inicial y Experimentación con MLflow

**Objetivo:** Entrenar modelos básicos, compararlos y registrar todo en MLflow para garantizar trazabilidad.

### 1. Selección de Modelos

### 2. Entrenamiento Básico

### 3. MLflow Tracking: Registrar Experimentos

### 4. Comparar Modelos en MLflow UI

## Entregables Semana 6 y 7

1. **Jupyter Notebook** con código de entrenamiento y registro en MLflow.
2. **Capturas de MLflow UI** mostrando los experimentos.
3. **Documento PDF** explicando la elección del mejor modelo.
4. **Crear el documento de reporte de modelos** ( [baseline\\_models](#) y [model\\_report](#) )

## 1.6. Semana 8 y 9 - Despliegue del Modelo en Producción

**Objetivo:** Llevar el modelo a un entorno real (ej.: API, app web) usando MLflow.

**1. Empaquetar el Modelo con MLflow**

**2. Desplegar como API Local**

### Entregables Semana 8 y 9

1. **Script de Despliegue** ([deploy.py](#)).
2. **Captura de la API funcionando** (ej.: Postman o respuesta de curl).
3. **Informe de despliegue** con pasos seguidos y problemas enfrentados.
4. **Documento de despliegue de modelos e informe de salida** ([deploymentdoc](#) y [exitreport](#))