Universidad Tecnológica de Tijuana	FORMATO INSTRUMENTO DE EVALUACIÓN	CÓDIG	F-CA-029-A
		REVISIÓ	01/0816
		HOJA	PAGE 4 de 1

TIPO: Reactivo ORD. REM. **CALIFICACIÓN:** _____

APLICA PARA: PRIMER PARCIAL | SEGUNDO PARCIAL | TERCER PARCIAL | FINAL | EXTRAORDINARIO

MATERIA: Extracción de conocimientos en BD **MAESTRO:** M.C. Florencio López Cruz

ALUMNO: _____ **GRUPO:** TI-9B **FECHA DE APLICACIÓN:** 06/11/25

PRACTICO **VALOR:** 40 PUNTOS

Instrucciones: El examen es individual, utilice el dataset que seleccione, de las siguientes opciones

1. **Dataset de Predicción de Cancelación de Reserva de Hotel (Kaggle):**
   - **Contexto:** Determinar si una reserva realizada por un cliente de hotel será cancelada o no antes de la llegada. El conjunto de datos contiene información detallada sobre las reservas: país del cliente, duración de la estancia, número de huéspedes, tipo de habitación, fechas de llegada y salida, entre otros elementos relevantes para el negocio hotelero.
   - **Características:** Incluye una combinación de variables numéricas (como número de noches, cantidad de adultos/niños, precio promedio diario) y variables categóricas (como tipo de hotel, origen del cliente, tipo de comida, agente o compañía involucrada).
   - **Problema:** Clasificación binaria, donde el objetivo es predecir si la reserva será cancelada (1) o no cancelada (0).
   - **URL de ejemplo:** https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

2. **Predicción de Hospitalización por Diabetes (Diabetic Readmission Prediction):**
   - **Contexto:** Predecir si un paciente diabético será readmitido en el hospital dentro de 30 días.
   - **Características:** Gran cantidad de características, numéricas y categóricas, muchas con valores faltantes o "desconocidos". Requiere ingeniería de características y manejo cuidadoso de datos médicos.
   - **Problema:** Clasificación binaria.

FECHA DE ELABORACIÓN: 03/11/25
AUTORIZÓ: _____
VO.BO.: _____

SABER _% SABER HACER_40% SER_%

UNIDADES TEMÁTICAS DE REFERENCIA: II, III, IV

- **URL de ejemplo:** https://www.kaggle.com/datasets/saurabhtayal/diabetic-patients-readmission-prediction

3. **Contaminación del Aire (Air Quality Data Set)**
   - **Contexto:** Evaluar niveles de contaminantes atmosféricos en zonas urbanas.
   - **Características:** Series temporales de contaminantes (NO_2, PM2.5, O_3).
   - **Problema:** Regresión o forecasting.
   - URL de ejemplo: https://www.kaggle.com/datasets/shrutibhargava94/india-air-quality-data

4. **Predicción de Calidad del Agua (Water Potability)**
   - **Contexto:** Determinar si el agua es potable según características químicas.
   - **Características:** Variables numéricas con valores faltantes (pH, sólidos, dureza).
   - **Problema:** Clasificación binaria.
   - **URL de ejemplo:** https://www.kaggle.com/datasets/gauravduttakiit/water-potability-prediction

5. **Predicción de Cáncer de Pulmón (Lung Cancer Patient Data)**
   - **Contexto:** Identificar si un paciente es propenso a desarrollar cáncer de pulmón basándose en factores personales y clínicos como edad, historial de tabaquismo y condiciones respiratorias.
   - **Características:** Mezcla de variables categóricas (historial familiar, tipo de tos) y numéricas (edad, índice respiratorio). Puede requerir normalización y codificación.
   - **Problema: Clasificación binaria** (riesgo alto vs riesgo bajo).
   - **URL del dataset:** https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

6. **Predicción de Diabetes en Pacientes Adultos (Pima Indians Diabetes Dataset)**
   - **Contexto:** Determinar si una persona presenta diabetes tipo 2 con base en indicadores clínicos y biométricos. El dataset proviene de un estudio de mujeres adultas indígenas Pima.
   - **Características:** Exclusivamente **numéricas** como nivel de glucosa, presión arterial, índice de masa corporal, edad, número de embarazos. Contiene valores cero usados erróneamente como faltantes, requiere limpieza.
   - **Problema: Clasificación binaria** (diabética vs no diabética).
   - **URL del dataset:** https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Parte 1: Ciclo de Machine Learning
Como evaluación debe implementar un modelo de Machine Learning para el dataset seleccionado.

Carga y Exploración de Datos (EDA)
- Cargar el dataset elegido utilizando Pandas.
- Realizar un análisis exploratorio de datos (EDA) exhaustivo:
  - Descripción estadística de las características (media, desviación estándar, percentiles, etc.).
  - Identificación de tipos de datos y conversión si es necesario.
  - Detección y visualización de valores faltantes (mapas de calor, gráficos de barras).
  - Análisis de la distribución de las características numéricas (histogramas, boxplots).
  - Análisis de la distribución de las características categóricas (gráficos de barras, conteos).
  - Análisis de correlaciones entre características (mapas de calor de correlación).

Preprocesamiento y Transformación de Datos

- Manejo de valores faltantes: Implementar una estrategia adecuada (imputación por media, mediana, moda, regresión, eliminación, etc.) y justificar la elección.
- Codificación de características categóricas: Aplicar One-Hot Encoding, Label Encoding u otra técnica apropiada. Justificar la elección.
- Escalado de características numéricas: Aplicar estandarización (StandardScaler) o normalización (MinMaxScaler) si es necesario. Justificar la elección.
- División del dataset en conjuntos de entrenamiento y prueba (train-test split).

Selección y Entrenamiento de Modelos

- Entrenar al menos dos de los siguientes algoritmos de Machine Learning en el conjunto de entrenamiento:
    - Regresión Lineal (si el problema es de regresión)
    - Regresión Logística (si el problema es de clasificación)
    - K-Vecinos Más Cercanos (k-NN)
    - *Opcional para bonificación:* Un tercer modelo de Scikit-learn de su elección (ej. Árbol de Decisión, Random Forest, SVM).
- Realizar ajuste de hiperparámetros básico (ej. utilizando GridSearchCV o RandomizedSearchCV) para al menos uno de los modelos.

Evaluación de Modelos

- Evaluar el rendimiento de cada modelo entrenado en el conjunto de prueba.
- **Para clasificación:** Calcular y reportar métricas como:
    - Precisión (Accuracy)
    - Recall
    - F1-Score
    - Matriz de Confusión
    - Curva ROC y AUC.
- **Para regresión:** Calcular y reportar métricas como:

- Error Cuadrático Medio (MSE)
- Error Absoluto Medio (MAE)
- R-cuadrado.

- Comparar el rendimiento de los modelos y seleccionar el "mejor" modelo para el despliegue, justificando la elección.

Parte 2: Despliegue Web del Modelo (30% de la calificación total)
En esta parte de su evaluación deberá integrar el modelo entrenado y preprocesado en una aplicación web interactiva.

Persistencia del Modelo y el Preprocesador

- Guardar el modelo entrenado (el "mejor" de la Parte 1) y los objetos de preprocesamiento (ej. StandardScaler, OneHotEncoder, LabelEncoder o un Pipeline completo) utilizando joblib o pickle.
- **Resultados Esperados:** Archivos .pkl o .joblib que contengan el modelo y los transformadores.

Desarrollo de la Aplicación Web:

- Desarrollar una aplicación web simple utilizando **Flask** o **Streamlit** (usted elige la tecnología que desee implementar).
- La aplicación debe permitir a un usuario:
  - Ingresar los valores de las características necesarias para una nueva predicción. La interfaz de usuario debe ser intuitiva y clara.
  - Hacer clic en un botón para obtener la predicción.
  - Mostrar claramente el resultado de la predicción al usuario.

Formato de Entrega y Reporte

- Reporte Detallado en PDF
- Repositorio de Código y Dataset