Actividad: Predicción de Diabetes con Machine Learning

Objetivo:

Desarrollar un modelo de Machine Learning para predecir si un paciente tiene diabetes basándose en características clínicas. La actividad incluirá desde la exploración de datos hasta la evaluación del modelo, promoviendo el razonamiento crítico sobre la calidad de los datos y el rendimiento del modelo.

Duración estimada:

2 horas

Parte 1: Introducción a los Datos y Conceptos Clave

Explicación breve

¿Qué es Machine Learning?
Machine Learning (ML) es una rama de la inteligencia artificial (IA) que permite a las computadoras aprender y mejorar automáticamente a partir de la experiencia sin ser programadas explícitamente para cada tarea. En lugar de seguir instrucciones rígidas, los algoritmos de ML identifican patrones en los datos y toman decisiones o realizan predicciones basadas en esos patrones.

El proceso típico de ML incluye:

1. **Recolección de datos:** Obtener un conjunto de datos relevante.
2. **Preprocesamiento:** Limpiar, transformar y organizar los datos para su análisis.
3. **Elección del modelo:** Seleccionar el algoritmo de ML adecuado.
4. **Entrenamiento:** Alimentar el modelo con datos etiquetados o no etiquetados para que aprenda.
5. **Evaluación:** Medir el rendimiento del modelo usando métricas específicas.
6. **Predicción:** Usar el modelo entrenado para hacer predicciones sobre nuevos datos.

Diferencias entre aprendizaje supervisado y no supervisado:

El aprendizaje supervisado se basa en datos etiquetados, es decir, cada entrada tiene una salida conocida, y el objetivo es que el modelo aprenda a hacer predicciones precisas basadas en esas etiquetas. Este tipo de aprendizaje se usa para tareas como la clasificación (asignar una categoría a un dato) y la regresión (predecir valores numéricos). Algunos algoritmos comunes en este enfoque son la regresión lineal, la regresión logística, los árboles de decisión, las máquinas de soporte vectorial (SVM) y las redes neuronales. Ejemplos de aplicación incluyen la clasificación de imágenes, la detección de fraudes, el diagnóstico médico y la predicción de ventas.

Por otro lado, el aprendizaje no supervisado trabaja con datos no etiquetados. Su objetivo es encontrar patrones ocultos o estructuras dentro de los datos sin una salida predefinida. Este tipo de aprendizaje se utiliza para tareas como el clustering (agrupación de datos similares) y la reducción de dimensionalidad (simplificación de conjuntos de datos complejos). Algunos algoritmos típicos incluyen K-Means, DBSCAN y PCA (Análisis de Componentes Principales). Las aplicaciones comunes de este enfoque incluyen la segmentación de clientes, el análisis de sentimientos, la compresión de datos y la detección de anomalías.

Clasificación: concepto y aplicaciones

La clasificación es una técnica de Machine Learning supervisado cuyo objetivo es asignar una categoría o etiqueta a una observación basándose en sus características. El modelo de clasificación aprende a partir de un conjunto de datos etiquetados y luego usa ese conocimiento para predecir la categoría de nuevos datos no vistos.

Las aplicaciones de la clasificación son muy amplias, especialmente en el ámbito biosanitario:

- **Diagnóstico médico:** Clasificación de imágenes médicas para detectar tumores, fracturas o anomalías.
- **Predicción de enfermedades:** Determinar la probabilidad de que un paciente desarrolle una enfermedad basándose en sus antecedentes médicos.
- **Análisis de muestras:** Clasificar células como benignas o malignas en estudios de histopatología.
- **Detección de anomalías:** Identificar parámetros fuera de lo normal en análisis de laboratorio.

¿Cómo funcionan los modelos de clasificación en medicina?

En el ámbito médico, los modelos de clasificación se entrenan con conjuntos de datos clínicos que contienen información sobre pacientes, como síntomas, resultados de pruebas, antecedentes médicos y diagnósticos previos. Estos modelos pueden utilizar algoritmos como árboles de decisión, regresión logística, redes neuronales o máquinas de soporte vectorial.

El proceso funciona así:

1. **Recopilación de datos:** Se obtiene información de historias clínicas electrónicas, laboratorios o estudios de imagen.
2. **Preprocesamiento:** Limpieza, normalización y transformación de datos para prepararlos para el modelo.
3. **Entrenamiento:** El modelo aprende a identificar patrones entre los síntomas y el diagnóstico basado en ejemplos etiquetados.
4. **Evaluación:** Se mide el rendimiento del modelo con métricas como precisión, sensibilidad, especificidad y área bajo la curva ROC.
5. **Predicción:** Se usa el modelo para predecir diagnósticos en pacientes nuevos basándose en sus datos clínicos.

Por ejemplo, un modelo puede predecir si un paciente tiene riesgo de diabetes en función de su edad, peso, nivel de glucosa y presión arterial.

Introducción a los datos tabulares en el ámbito biosanitario

Los datos tabulares en el ámbito biosanitario son aquellos que se organizan en forma de tablas, con filas que representan observaciones (pacientes, estudios, muestras) y columnas que contienen variables (edad, peso, resultados de pruebas, diagnósticos). Estos datos suelen extraerse de sistemas de información hospitalaria (HIS), registros médicos electrónicos (EHR) o laboratorios clínicos (LIS).

Algunas características comunes de los datos tabulares biosanitarios son:

- **Datos demográficos:** Edad, género, peso, altura.
- **Resultados de laboratorio:** Niveles de glucosa, hemoglobina, colesterol.
- **Historial médico:** Diagnósticos previos, tratamientos, cirugías.
- **Síntomas reportados:** Fiebre, dolor, presión arterial.

El análisis de estos datos permite construir modelos predictivos, realizar estudios epidemiológicos, optimizar la toma de decisiones clínicas y personalizar tratamientos médicos.

Dataset:
Usaremos el dataset **PIMA Indians Diabetes** (disponible en Kaggle o en sklearn). Contiene datos como:

- Niveles de glucosa
- Presión arterial
- Índice de masa corporal (IMC)
- Edad
- Historial de embarazos (en mujeres)
- Entre otros

Tarea para los alumnos:

- Reflexionar sobre qué variables podrían influir en la diabetes.
- Discutir posibles desafíos (datos faltantes, correlaciones, sesgo en los datos).

Parte 2: Implementación en Python

Paso 1: Carga y Exploración de los Datos (20 min)

Los alumnos deben:

- Cargar los datos en un DataFrame de pandas.
- Inspeccionar las primeras filas.
- Obtener estadísticas descriptivas (media, desviación estándar, valores faltantes).

Código base:

Preguntas para reflexionar:

1. ¿Hay valores extremos o posibles errores en los datos?
2. ¿Cómo podríamos mejorar la calidad del dataset?

Paso 2: Preparación de Datos

- Separar características (`X`) y variable objetivo (`y`).
- Normalizar los datos con `StandardScaler`.
- Dividir en entrenamiento y prueba (80%-20%).

Código:

Paso 3: Entrenamiento del Modelo (30 min)

Los alumnos probarán diferentes modelos y compararán su desempeño.

Código base:

Preguntas para los alumnos:

1. ¿El modelo es suficientemente preciso para ser usado en un entorno médico?
2. ¿Qué pasaría si hubiera más datos o datos de mejor calidad?
3. ¿Cómo podríamos mejorar el modelo? (Ejemplo: probar otro algoritmo, ajustar hiperparámetros)

Parte 3: Reflexión y Discusión

Los alumnos debatirán sobre:

- Importancia del balance entre precisión y sensibilidad en aplicaciones médicas.
- Riesgos de sesgo en los datos de salud.
- Posibles usos del modelo en la práctica clínica.

Extensión:
Podemos probar con **otro modelo (SVM, KNN, regresión logística)** o realizar un **análisis de características importantes** con `feature_importances_` de RandomForest.