



**TECNOLÓGICO
NACIONAL DE MÉXICO**



Tecnológico Nacional de México

Campus Culiacán

Carrera:

Ingeniería en Sistemas Computacionales

Nombre de la materia:

Inteligencia Artificial

Tarea 2: Aprendizaje automático

Alumnos:

Aguilar Recio Jesús Octavio

Flores Fernández Emily Karely

Nombre del maestro:

Zuriel Dathan Mora Félix

Grupo:

9:00 – 10:00

Proceso de Aprendizaje Automático

El **aprendizaje automático (Machine Learning, ML)** es una rama de la inteligencia artificial que permite a los sistemas mejorar su rendimiento en tareas específicas mediante la experiencia adquirida a partir de los datos. Un modelo de ML es tan bueno como los datos con los que es entrenado, por lo que la calidad y preparación de los datos juegan un papel crucial en su desempeño.

1. Adquisición de datos.

La **adquisición de datos** en aprendizaje automático consiste en recopilar información de diversas fuentes, como bases de datos, archivos CSV, API, imágenes o grabaciones de audio. Es fundamental asegurar que los datos sean representativos, de calidad y cumplan con normativas como el GDPR. Además, deben estar libres de sesgos que puedan afectar el rendimiento del modelo.

Fuentes Comunes de Datos en ML

- **Bases de datos estructuradas** (SQL, NoSQL, BigQuery).
- **APIs y Web Scraping** (Twitter API, Wikipedia Scraping).
- **Archivos CSV, JSON, Excel** (Datos de sensores, registros históricos).
- **Imágenes y vídeos** (Dataset de visión computacional como ImageNet).
- **Audio y lenguaje natural** (Conjuntos de datos de voz como Librispeech).

Consideraciones en la Adquisición de Datos

- **Calidad de los datos** : Deben ser relevantes y representativos.
- **Cantidad de datos** : Más datos pueden mejorar la precisión, pero también aumentar los costos computacionales.
- **Accesibilidad y legalidad** : Respetar regulaciones como **GDPR** (protección de datos en la UE).
- **Ética y sesgo** : Evitar conjuntos de datos que reflejan discriminación o sesgo.

2. Preprocesamiento de Datos.

El **preprocesamiento de datos** es el paso en el que se limpian y transforman los datos para hacerlos utilizables. Se manejan valores faltantes, se eliminan duplicados y se corrigen errores. También se convierten datos categóricos en numéricos y se normalizan las variables para garantizar que estén en escalas comparables. En algunos casos, se crean nuevas características relevantes para mejorar la precisión del modelo. Sin este proceso, un modelo de aprendizaje automático puede producir resultados inexactos.

Pasos Clave en el Preprocesamiento

1. Limpieza de datos

- **Manejo de datos faltantes :**
 - Eliminar datos incompletos.
 - Rellenar con valores estadísticos (media, mediana).
- **Eliminación de duplicados :** Evita sesgos en los resultados.
- **Corrección de errores :** Unificación de formatos ("M" y "Masculino" deben ser lo mismo).

2. Conversión de tipos de datos

- **Datos categóricos** → Números (One-Hot Encoding, Label Encoding).
- **Fechas y tiempos** → Formato estándar ISO 8601.

3. Normalización y Estandarización

- **Normalización (Escalado Min-Max) :** Convierte valores entre 0 y 1.
- **Estandarización (Z-score) :** Media = 0, desviación estándar = 1.

4. Ingeniería de características

- Creación de nuevas variables a partir de los datos originales (Ej: extraer día de la semana desde una fecha).

3. Entrenamiento del modelo.

El **entrenamiento del modelo** es la etapa del ML en la que se utiliza un algoritmo de aprendizaje automático para aprender patrones a partir de los datos preprocesados. En este proceso también el modelo ajusta sus parámetros internos para minimizar el error en las predicciones.

Pasos del entrenamiento del modelo

1. Selección del algoritmo:

Se debe de seleccionar un algoritmo dependiendo del problema (clasificación, regresión, clustering, etc.) algunos ejemplos de algoritmos podrían ser regresión lineal, árboles de decisión, redes neuronales, entre otros.

2. División de datos:

Los datos se dividen en conjuntos de entrenamiento y validación. El conjunto de entrenamiento es utilizado para ajustar el modelo, mientras que el de validación se usa para evaluar su rendimiento durante el entrenamiento.

3. Ajuste de parámetros:

En este paso el modelo ajusta sus parámetros para minimizar una función de pérdida.

4. Iteración y optimización:

Es el proceso de entrenamiento suele ser iterativo. En cada iteración, el modelo mejora sus predicciones basándose en los errores cometidos en la iteración anterior.

Algunas consideraciones para este proceso son:

Sobreajuste: El modelo puede aprender demasiado los detalles del conjunto de entrenamiento, perdiendo generalización. Para evitarlo, se pueden usar técnicas como la regularización o la validación cruzada.

Subajuste: El modelo no aprende lo suficiente de los datos, resultando en un rendimiento pobre. Esto puede deberse a un modelo demasiado simple o a un entrenamiento insuficiente.

4. Evaluación del modelo.

La **evaluación del modelo** es la fase en la que se mide su rendimiento utilizando un conjunto de datos que no fue utilizado durante el entrenamiento (conjunto de prueba). Esto permite determinar si el modelo es capaz de generalizar bien a nuevos datos.

Métricas de Evaluación Comunes

1. Clasificación:

- Precisión (Accuracy): Proporción de predicciones correctas sobre el total.
- Precisión (Precision): Proporción de verdaderos positivos sobre todos los positivos predichos.
- Recall (Sensibilidad): Proporción de verdaderos positivos sobre todos los positivos reales.
- F1-Score: Media armónica de precisión y recall.

2. Regresión:

- Error Cuadrático Medio (MSE): Promedio de los errores al cuadrado.
- Error Absoluto Medio (MAE): Promedio de los errores absolutos.
- Coeficiente de Determinación (R^2): Mide cuánto de la variabilidad de los datos es explicada por el modelo.

3. Clustering:

- Silueta (Silhouette Score): Mide cuán bien separados están los clusters.
- Inercia (Inertia): Suma de las distancias al cuadrado de cada punto a su centroide.

Técnicas de Evaluación

- Validación Cruzada (Cross-Validation): Divide los datos en múltiples subconjuntos y entrena/evalúa el modelo en cada uno para obtener una estimación más robusta del rendimiento.
- Matriz de Confusión: Muestra los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos en problemas de clasificación.

Consideraciones en la Evaluación

- Sesgo y Varianza: Un modelo con alto sesgo puede no capturar la complejidad de los datos, mientras que uno con alta varianza puede sobreajustarse.
- Generalización: El modelo debe funcionar bien en datos no vistos, no solo en los datos de entrenamiento.

5. Implementación del modelo.

La **implementación del modelo** es la fase en la que el modelo entrenado y evaluado se despliega en un entorno real para hacer predicciones o tomar decisiones basadas en nuevos datos. Esta fase es crucial para llevar el modelo del laboratorio al mundo real.

Pasos Clave en la Implementación

1. Serialización del Modelo:

El modelo se guarda en un formato que puede ser cargado y utilizado posteriormente, como un archivo .pkl (Pickle) en Python o .h5 para modelos de redes neuronales.

2. Despliegue en Producción:

El modelo se integra en sistemas o aplicaciones, como APIs, aplicaciones móviles, o sistemas en la nube (por ejemplo, AWS, Google Cloud, Azure).

3. Monitoreo y Mantenimiento:

Una vez implementado, el modelo debe ser monitoreado para asegurar que sigue funcionando correctamente. Esto incluye:

- Monitoreo del rendimiento: Detectar si el rendimiento del modelo disminuye con el tiempo (por ejemplo, debido a cambios en los datos).
- Actualizaciones: Reentrenar el modelo con nuevos datos si es necesario.

Consideraciones en la Implementación

- Escalabilidad: El modelo debe ser capaz de manejar grandes volúmenes de datos y solicitudes en tiempo real.
- Latencia: El tiempo que tarda el modelo en hacer una predicción debe ser aceptable para la aplicación.
- Seguridad: Proteger el modelo y los datos de posibles ataques o accesos no autorizados.
- Ética y Cumplimiento: Asegurar que el modelo cumple con regulaciones y no causa daños o discriminación.

Tabla de similitudes y diferencias entre el modelo cognitivo y aprendizaje automático.

Componente del Modelo Cognitivo	Etapas del Aprendizaje Automático	Similitudes	Diferencias
Percepción (recepción de estímulos sensoriales)	Adquisición de datos (recolección de información de diversas fuentes)	Ambos procesos implican la recopilación de información del entorno.	En el modelo cognitivo, la percepción es biológica y subjetiva, mientras que en ML los datos son recopilados de manera estructurada y objetiva.
Atención (filtrado de información relevante)	Procesamiento de datos (limpieza y transformación de datos)	Se eliminan datos irrelevantes y se prioriza la información útil.	En los humanos, la atención depende del contexto y la experiencia; en ML, se usan algoritmos para limpiar y transformar datos.
Memoria (almacenamiento y recuperación de información)	Entrenamiento del modelo (ajuste de parámetros a partir de datos históricos)	Ambos almacenan información y la usan para mejorar el desempeño en futuras tareas.	La memoria humana es flexible y asociativa, mientras que en ML el almacenamiento se realiza en estructuras de datos y modelos matemáticos.
Razonamiento (toma de decisiones basada en información)	Evaluación del modelo (medición del desempeño y ajuste de hiperparámetros)	Ambos procesos analizan información para determinar la mejor respuesta posible.	El razonamiento humano es adaptable y basado en experiencias; en ML, la evaluación es cuantitativa y depende de métricas específicas.
Ejecutar acciones (actuar con base en decisiones)	Implementación del modelo (uso del modelo en entornos reales)	Tanto el ser humano como un modelo de ML se aplica lo aprendido para tomar decisiones en el mundo real.	La acción humana implica juicio y emociones; En ML, la implementación sigue reglas predefinidas sin subjetividad.

Referencias.

<https://hastie.su.domains/ElemStatLearn/>