

# Guía de Estudio para ETS

## Aprendizaje de Máquina e Inteligencia Artificial

Daniel Jiménez Alcántar – ESCOM IPN

2026-01-15

### Propósito general

Comprender los fundamentos de la Inteligencia Artificial y el Aprendizaje de Máquina, así como sus enfoques, aplicaciones y métodos de validación, para analizar y evaluar modelos de aprendizaje como apoyo a la toma de decisiones.

### 1. Unidad 1: Inteligencia Artificial y Aprendizaje de Máquina

#### Contenidos

1. Inteligencia Artificial: definición, aplicaciones y perspectivas.
2. Disciplinas de la IA: representación del conocimiento, razonamiento y búsqueda.
3. Aprendizaje de Máquina: fundamentos, tipos y evaluación.
4. Métodos de validación: entrenamiento-prueba, validación cruzada y bootstrap.
5. IA como apoyo a la toma de decisiones.

#### Recomendaciones de lectura

- Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach*.
- Ammet, J. *Machine Learning and Artificial Intelligence*.
- Pajares, M., Sanz, G. & De la Cruz, J. *Aprendizaje Automático: Un Enfoque Práctico*.
- Chollet, F. *Deep Learning with Python*.

#### Conceptos clave

**Inteligencia Artificial (IA).** Rama de las ciencias computacionales que desarrolla sistemas capaces de realizar tareas asociadas a la inteligencia humana, como razonamiento, aprendizaje y toma de decisiones.

**Aprendizaje de Máquina.** Subcampo de la IA que permite a los sistemas aprender patrones a partir de datos sin programación explícita.

**Representación del conocimiento.** Modelado formal de información mediante reglas, grafos u ontologías procesables por computadora.

**Razonamiento.** Inferencia de nueva información a partir del conocimiento existente, usando enfoques lógicos o probabilísticos.

**Búsqueda.** Exploración de un espacio de estados para encontrar soluciones óptimas o aceptables.

**Aprendizaje supervisado.** Entrenamiento con datos etiquetados.

**Aprendizaje no supervisado.** Descubrimiento de patrones sin etiquetas previas.

**Sobreajuste.** Ajuste excesivo del modelo a los datos de entrenamiento que reduce la capacidad de generalización.

**Validación cruzada.** Evaluación del modelo mediante múltiples particiones de los datos.

### Cuestionario teórico

1. Defina Inteligencia Artificial y diferénciela del Aprendizaje de Máquina.
2. Mencione y explique tres aplicaciones actuales de la IA.
3. Explique la importancia de la representación del conocimiento.
4. Compare razonamiento simbólico y probabilístico.
5. ¿Qué es un problema de búsqueda? Dé un ejemplo.
6. Explique qué es el Aprendizaje de Máquina.
7. Compare aprendizaje supervisado y no supervisado.
8. ¿Qué implica evaluar una máquina de aprendizaje?
9. Diferencie conjunto de entrenamiento y prueba.
10. Justifique el uso de validación cruzada.

### Ejercicios prácticos

1. Clasifique como IA clásica o Aprendizaje de Máquina:
  - Ajedrez basado en reglas.
  - Sistema de recomendación.
  - Diagnóstico médico experto.
  - Filtro de spam.
2. Identifique el tipo de aprendizaje en:
  - Agrupamiento de clientes.
  - Predicción del precio de viviendas.
  - Detección de fraudes.
3. Un modelo obtiene 95 % en entrenamiento y 60 % en prueba:
  - Identifique el problema.
  - Proponga una solución.
4. Describa la validación cruzada k-fold con  $k = 5$ .
5. Explique el uso de ML en apoyo a la toma de decisiones en salud o transporte.

## 2. Unidad 2: Aprendizaje Supervisado

### Propósito de la unidad

Analizar y aplicar modelos de aprendizaje supervisado para problemas de regresión y clasificación, comprendiendo sus fundamentos, supuestos, ventajas, limitaciones y criterios de evaluación.

### Contenidos

1. Regresión
  - Regresión lineal
  - Regresión polinómica
  - Regresión logística
2. Redes neuronales
3. Árboles de decisión
4. K vecinos más cercanos (KNN)

## Recomendaciones de lectura

- Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach*.
- Chollet, F. *Deep Learning with Python*.
- James, G. et al. *An Introduction to Statistical Learning*.
- Pajares, M., Sanz, G. & De la Cruz, J. *Aprendizaje Automático: Un Enfoque Práctico*.

## Conceptos clave

**Aprendizaje supervisado.** Enfoque de aprendizaje donde el modelo se entrena con datos etiquetados, aprendiendo una función que relaciona entradas con salidas conocidas.

**Regresión lineal.** Modelo que asume una relación lineal entre variables independientes y una variable dependiente, generalmente expresada como:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

**Regresión polinómica.** Extensión de la regresión lineal que incorpora términos no lineales mediante potencias de las variables independientes.

**Regresión logística.** Modelo de clasificación que estima la probabilidad de pertenencia a una clase utilizando una función sigmoide.

**Redes neuronales.** Modelos computacionales inspirados en el cerebro humano, formados por capas de neuronas artificiales conectadas mediante pesos ajustables.

**Árboles de decisión.** Modelos jerárquicos que dividen el espacio de datos mediante reglas basadas en características, utilizando métricas como entropía o índice Gini.

**K vecinos más cercanos (KNN).** Algoritmo basado en instancias que clasifica o predice según la similitud con los  $k$  ejemplos más cercanos.

**Sesgo y varianza.** Conceptos que describen el equilibrio entre la capacidad de generalización y el ajuste del modelo a los datos de entrenamiento.

## Cuestionario teórico

1. Defina aprendizaje supervisado y mencione dos problemas que puede resolver.
2. Explique los supuestos de la regresión lineal.
3. ¿En qué casos la regresión polinómica es preferible a la regresión lineal?
4. Explique el funcionamiento de la regresión logística y su interpretación probabilística.
5. Describa la arquitectura básica de una red neuronal artificial.
6. ¿Qué es la función de activación y por qué es necesaria?
7. Explique cómo se construye un árbol de decisión.
8. Compare entropía e índice Gini como criterios de división.
9. Describa el algoritmo KNN y el papel del parámetro  $k$ .
10. Explique el problema del sesgo y la varianza en modelos supervisados.

## Ejercicios prácticos

1. **Regresión lineal** Genere un conjunto de datos que siga el modelo  $y = 2x + \varepsilon$ , donde  $\varepsilon \sim N(0, 1)$ . Ajuste un modelo de regresión lineal e interprete los coeficientes.
2. **Regresión polinómica** Ajuste modelos de grado 1, 2 y 3 a un conjunto de datos no lineales. Compare el error de entrenamiento y discuta el sobreajuste.
3. **Regresión logística** Utilice un conjunto de datos binario y entrene un modelo de regresión logística. Interprete la salida del modelo como probabilidad.
4. **Redes neuronales** Diseñe una red neuronal con una capa oculta para un problema de clasificación simple. Indique el número de neuronas y la función de activación utilizada.

5. **Árboles de decisión** Construya un árbol de decisión para clasificar observaciones y explique el significado de las divisiones generadas.
6. **KNN** Aplique el algoritmo KNN con distintos valores de  $k$  y analice su impacto en el rendimiento del modelo.
7. **Comparación de modelos** Compare al menos dos algoritmos supervisados para el mismo conjunto de datos e identifique cuál generaliza mejor.

### 3. Unidad 3: Aprendizaje No Supervisado

#### Propósito de la unidad

Comprender y aplicar técnicas de aprendizaje no supervisado para la identificación de patrones y estructuras ocultas en los datos, mediante métodos de agrupamiento y métricas de evaluación de la calidad de los clusters.

#### Contenidos

1. Aprendizaje no supervisado
2. Agrupamiento (clustering)
3. Algoritmo k-means
4. Expectation–Maximization (EM)
5. Agrupamiento jerárquico
6. DBSCAN

#### Recomendaciones de lectura

- Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach*.
- James, G. et al. *An Introduction to Statistical Learning*.
- Runkler, T. *Data Analytics: Models and Algorithms for Intelligent Data Analysis*.
- Aggarwal, C. *Data Clustering: Algorithms and Applications*.

#### Conceptos clave

**Aprendizaje no supervisado.** Enfoque de aprendizaje en el que los datos no cuentan con etiquetas, y el objetivo es descubrir patrones, estructuras o relaciones inherentes.

**Agrupamiento (clustering).** Técnica que consiste en agrupar observaciones de manera que los elementos dentro de un mismo grupo sean más similares entre sí que con los de otros grupos.

**k-means.** Algoritmo de agrupamiento basado en centroides que minimiza la varianza intra-grupo, asignando cada observación al centroide más cercano.

**Expectation–Maximization (EM).** Algoritmo iterativo probabilístico que estima los parámetros de modelos de mezcla, comúnmente utilizado en Gaussian Mixture Models.

**Agrupamiento jerárquico.** Método que construye una jerarquía de clusters, representada mediante un dendrograma, usando enfoques aglomerativos o divisivos.

**DBSCAN.** Algoritmo de agrupamiento basado en densidad que identifica clusters de forma arbitraria y detecta ruido, utilizando los parámetros  $\varepsilon$  y  $minPts$ .

**Distancia intra-grupo e inter-grupo.** Medidas utilizadas para evaluar la cohesión interna de los clusters y la separación entre ellos.

**Índice de silueta.** Métrica que cuantifica qué tan bien una observación se encuentra asignada a su cluster, considerando cohesión y separación.

## Cuestionario teórico

1. Defina aprendizaje no supervisado y explique en qué se diferencia del aprendizaje supervisado.
2. Explique el objetivo principal del agrupamiento de datos.
3. Describa paso a paso el funcionamiento del algoritmo k-means.
4. ¿Cuáles son las principales limitaciones de k-means?
5. Explique el principio del algoritmo Expectation–Maximization (EM).
6. Compare EM y k-means en términos de supuestos y flexibilidad.
7. Defina agrupamiento jerárquico y distinga entre enfoques aglomerativo y divisivo.
8. ¿Qué es un dendrograma y cómo se interpreta?
9. Explique el funcionamiento de DBSCAN y el significado de sus parámetros.
10. Compare DBSCAN con k-means para conjuntos de datos con ruido y formas complejas.

## Ejercicios prácticos

1. **k-means** Genere un conjunto de datos bidimensional con tres grupos bien definidos. Aplique k-means con  $k = 3$  y analice los centroides obtenidos.
2. **Selección de  $k$**  Aplique k-means con diferentes valores de  $k$  y utilice el método del codo o el índice de silueta para seleccionar el valor óptimo.
3. **Expectation–Maximization** Ajuste un modelo de mezcla gaussiana (GMM) a un conjunto de datos y compare los clusters obtenidos con los de k-means.
4. **Agrupamiento jerárquico** Aplique un método aglomerativo y visualice el dendrograma. Corte el dendrograma para obtener distintos números de clusters y compare los resultados.
5. **DBSCAN** Aplique DBSCAN a un conjunto de datos con ruido. Identifique los clusters detectados y los puntos considerados como ruido.
6. **Comparación de métodos** Compare k-means, EM y DBSCAN para el mismo conjunto de datos y discuta cuál es más adecuado según la estructura de los datos.
7. **Aplicación práctica** Proponga un problema real donde el aprendizaje no supervisado sea más adecuado que el aprendizaje supervisado y justifique su elección.

## 4. Unidad 4: Modelos de Predicción

### Propósito de la unidad

Analizar modelos de predicción basados en enfoques probabilísticos y de ensamble, comprendiendo sus fundamentos teóricos, supuestos, ventajas y aplicaciones en problemas reales de toma de decisiones.

### Contenidos

1. Modelos de predicción
2. Redes Bayesianas
3. Bosques Aleatorios (Random Forest)

### Recomendaciones de lectura

- Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach*.
- Koller, D. & Friedman, N. *Probabilistic Graphical Models*.
- Breiman, L. *Random Forests*.
- James, G. et al. *An Introduction to Statistical Learning*.

## Conceptos clave

**Modelos de predicción.** Modelos cuyo objetivo es estimar el valor de una variable de interés a partir de un conjunto de variables explicativas, considerando incertidumbre y generalización.

**Redes Bayesianas.** Modelos probabilísticos gráficos que representan un conjunto de variables aleatorias y sus dependencias condicionales mediante un grafo dirigido acíclico (DAG).

**Probabilidad condicional.** Medida de la probabilidad de que ocurra un evento dado que otro evento ya ha ocurrido.

**Inferencia bayesiana.** Proceso de actualización de creencias sobre un evento utilizando el teorema de Bayes y evidencia observada.

**Bosques Aleatorios (Random Forest).** Método de ensamble basado en múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de datos y características.

**Bagging.** Técnica de ensamble que reduce la varianza del modelo entrenando múltiples modelos sobre muestras con reemplazo del conjunto de entrenamiento.

**Importancia de variables.** Medida que indica la contribución relativa de cada característica en la predicción del modelo.

**Generalización.** Capacidad de un modelo para mantener un buen desempeño sobre datos no vistos durante el entrenamiento.

## Cuestionario teórico

1. Explique el objetivo de los modelos de predicción en el Aprendizaje de Máquina.
2. Defina qué es una red bayesiana y describa sus componentes.
3. ¿Qué es un grafo dirigido acíclico (DAG) y por qué es fundamental en redes bayesianas?
4. Explique el concepto de probabilidad condicional.
5. Describa el proceso de inferencia bayesiana.
6. Compare redes bayesianas y modelos discriminativos.
7. Explique el principio de funcionamiento de los bosques aleatorios.
8. ¿Por qué los Random Forest reducen el sobreajuste frente a un solo árbol?
9. Explique el concepto de bagging y su relación con Random Forest.
10. ¿Qué información proporciona la importancia de variables?

## Ejercicios prácticos

1. **Redes Bayesianas** Considere un sistema de diagnóstico médico con variables relacionadas con síntomas y enfermedades. Dibuje un posible grafo bayesiano e identifique dependencias condicionales.
2. **Inferencia bayesiana** Suponga una prueba médica con una tasa de falsos positivos conocida. Calcule la probabilidad de que un paciente tenga la enfermedad dado un resultado positivo.
3. **Random Forest** Entrene un modelo de bosque aleatorio para un problema de clasificación o regresión. Analice el impacto del número de árboles en el desempeño.
4. **Comparación de modelos** Compare el desempeño de un árbol de decisión y un Random Forest sobre el mismo conjunto de datos.
5. **Importancia de variables** Analice las variables más relevantes identificadas por un Random Forest y explique su significado en el contexto del problema.
6. **Aplicación práctica** Proponga un caso real donde el uso de redes bayesianas o bosques aleatorios sea más adecuado que otros modelos predictivos.

## Observaciones para el ETS

- Se evalúa comprensión conceptual y razonamiento probabilístico.
- Es común justificar la elección del modelo de predicción.

- Random Forest suele compararse con árboles de decisión.
- Se evalúa análisis y razonamiento, no memorización.
- Se espera justificación conceptual clara.