Apuntes de la clase del 16 de setiembre de 2025

Curso de Inteligencia Artificial

Nelson Rojas Obando
Estudiante Ingeniería en Computación
nelson.rojas@estudiantec.cr

***Resumen*—This paper summarizes the main topics discussed during the Artificial Intelligence course on September 16, 2025. It covers the quiz about concepts such as linear and logistic regression, concepts as techniques for handling outliers, and strategies to reduce high bias and high variance in machine learning models. Furthermore, it presents evaluation metrics including accuracy, precision, recall, F1-score, confusion matrix, ROC curve, and AUC, illustrated with practical case studies. Finally, the paper highlights the importance of data preprocessing tasks—such as cleaning, integration, reduction, transformation, and discretization—as essential steps to improve the quality of datasets and ensure better performance of predictive models.**

***Index Terms*—Inteligencia Artificial, Machine Learning, Métricas de evaluación, Matriz de Confusión, ROC, AUC, Data Preprocessing**

I. Introducción

La Inteligencia Artificial (IA) y el aprendizaje automático requieren no solo del diseño de modelos predictivos, sino también de procesos rigurosos para evaluar su desempeño y garantizar su aplicabilidad en escenarios reales. En este documento se abordan conceptos fundamentales que permiten comprender la relación entre los modelos, las métricas de evaluación y la calidad de los datos utilizados en su entrenamiento.

En primer lugar, se estudian métricas clásicas como la exactitud, precisión, exhaustividad y F1-score, así como métricas más avanzadas como la curva ROC y el área bajo la curva (AUC), las cuales proporcionan una visión más completa del rendimiento de un clasificador.

Además, se presenta la matriz de confusión como herramienta central para interpretar los aciertos y errores de los modelos, junto con un caso práctico aplicado a la detección de cáncer en pacientes. Asimismo, se destacan las principales tareas del preprocesamiento de datos, entre ellas la limpieza, integración, reducción, transformación y discretización, esenciales para enfrentar problemas como datos incompletos, inconsistentes o ruidosos.

II. Aspectos administrativos

Ver dos lecturas asociadas con lectura procesamiento de datos y de redes neuronales además de dos capítulos de un libro.

Se realizó el quiz #3, donde al finalizar se vieron las respuestas correspondientes. El quiz consistió en:

1. Mencione la diferencia de regresión lineal y logística.
   Respuesta:

   La **regresión lineal** y la **regresión logística** son técnicas fundamentales en el aprendizaje supervisado, pero se aplican a diferentes tipos de problemas:

   - **Regresión lineal:** se utiliza cuando la variable dependiente es *continua*. El modelo estima una relación lineal entre las variables independientes y la variable dependiente, siguiendo la forma:

     $$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$

     donde y puede tomar cualquier valor real.
   - **Regresión logística:** se emplea cuando la variable dependiente es *categórica*, típicamente binaria (0 o 1). El modelo estima la probabilidad de pertenecer a una clase utilizando la función sigmoide:

     $$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

     De esta forma, la salida está acotada en el intervalo $[0, 1]$ y se interpreta como probabilidad.

2. Describa 3 técnicas para el tratamiento de datos sobresalientes.
   Respuesta:
   Los datos sobresalientes, también conocidos como *outliers*, son observaciones que se desvían significativamente del resto de los datos. Su presencia puede afectar de manera negativa el rendimiento de los modelos de aprendizaje automático. Entre las técnicas más comunes para su tratamiento se encuentran:

   a) **Eliminación de outliers**: consiste en descartar aquellas observaciones que superan un umbral definido, por ejemplo, valores que se encuentran a más de tres desviaciones estándar de la media. Esta técnica es útil cuando los outliers provienen de errores de medición o registro.

   b) **Transformaciones de los datos**: aplicar transformaciones matemáticas, como la transformación logarítmica o la raíz cuadrada, puede reducir la influencia de valores extremos, estabilizando la varianza y mejorando la distribución de los datos.

   c) **Imputación o sustitución de valores**: reemplazar los outliers por valores más representativos, como la media, la mediana o un valor obtenido mediante interpolación. Esta técnica conserva el tamaño del conjunto de datos y es útil cuando la eliminación no es deseable.

3. Mencione 2 técnicas para evitar un alto sesgo y 2 técnicas para evitar alta varianza.
   Respuesta:
   En el contexto del aprendizaje automático, el **alto sesgo** (underfitting) y la **alta varianza** (overfitting) son problemas comunes. Algunas técnicas para mitigarlos son las siguientes:
   - **Para evitar un alto sesgo:**
     a) *Aumentar la complejidad del modelo*, por ejemplo, utilizando modelos polinómicos en lugar de regresión lineal simple, o redes neuronales más profundas.
     b) *Reducir la regularización excesiva*, ajustando los hiperparámetros de técnicas como L1/L2 o dropout, que en exceso limitan la capacidad de aprendizaje del modelo.
   - **Para evitar una alta varianza:**
     a) *Aumentar la cantidad de datos de entrenamiento*, mediante recolección adicional o técnicas de *data augmentation*.
     b) *Aplicar regularización*, como L1/L2, dropout o *early stopping*, con el fin de penalizar la complejidad excesiva y mejorar la generalización.
4. Anote la derivada de la función sigmoide

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

   Respuesta:

$$\sigma'(x) = \sigma(x)\left(1-\sigma(x)\right)$$

III. Métricas

Son medidas que se utilizan para indicar el rendimiento de un modelo predictvo. Constituyen la forma más objetiva de evaluar y comparar modelos de aprendizaje automático, permitiendo determinar qué tan bien se ajustan a los datos de entrenamiento y, sobre todo, qué tan bien generalizan a datos no vistos.

Asimismo, se emplean *benchmarks*, que son conjuntos de datos o pruebas estandarizadas utilizadas en la comunidad científica para comparar de manera justa el desempeño de distintos modelos.

IV. Matriz de confusión

En la matriz de confusión se colocan las clases y se realiza una clasificación según su posibilidad y veracidad, como se muestra en la figura 1. De esta forma se obtienen True Positive (verdadero positivo), False Positive (falso positivo), True negative (verdadero negativo) y False negative (falso negativo). Esta tabla puede ser N x N clases y al hacer un ploteo de esta tabla se espera que la diagonal esté dando valores verdaderos.

Un ejemplo práctico de esto es el caso de la figura 2. En el que se evaluán el resultado de embarazo en hombres y mujeres. Claramente un hombre no puede embarazarse por lo que de obtener un resultado positivo este sería un error de tipo 1. En

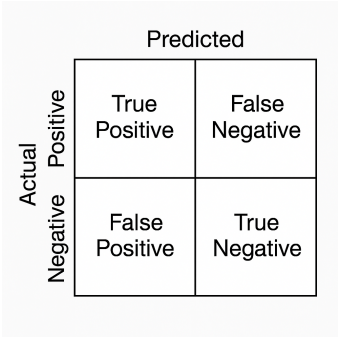

Figura 1. Enter Caption

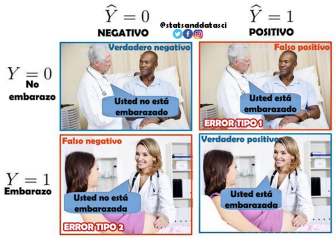

Figura 2. Enter Caption

el caso de la mujer, si existe la posibilidad por lo que se podría dar un resultado equivocado, y se conoce como error de tipo 2.

Ejemplos de métricas clásicas:

- **Accuracy (exactitud):** mide la proporción de predicciones correctas sobre el total de predicciones realizadas. Es ampliamente usada en problemas de clasificación, como en la regresión logística. Se define como:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

  donde TP (verdaderos positivos), TN (verdaderos negativos), FP (falsos positivos) y FN (falsos negativos). Este tipo de metrica otorga importancia igual a todas las clases. Es importante tomar esto en cuenta si las clases no están balanceadas. Puede no ser suficiente y da como resultado un valor porcentual (de 0 a 1). Para un modelo bien hecho se esperaría que se acerque bastante a 1.
- **Precisión (Precision):** mide la proporción de predicciones positivas correctas entre todas las predicciones positivas realizadas, como los errores de tipo 1:

$$\text{Precision} = \frac{TP}{TP+FP}$$

- **Exhaustividad (Recall):** indica la proporción de verdaderos positivos identificados correctamente sobre el total de elementos positivos, y se usa para medir los errores de tipo 2:

$$\text{Recall} = \frac{TP}{TP+FN}$$

- **F1-Score:** es la media armónica entre precisión y exhaustividad, útil cuando existe un desbalance de clases:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Caso de estudio: Dado un conjunto de 1000 pacientes se han realizado estudios para determinar la presencia de cancer. Del total de pacientes, 30 son pacientes con cáncer (clase positiva) y 970 pacientes sin cáncer (clase negativa).

Matriz de confusión:

Predicción/Objetivo	Cáncer	No cáncer
Cáncer	$25\,(TP)$	$20\,(FP)$
No cáncer	$5\,(FN)$	$950\,(TN)$

Métricas de evaluación:

- **Accuracy:**

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}} = \frac{25 + 950}{1000} = 0{,}975\,(97{,}5\,\%)$$

- **Recall (Sensibilidad):**

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{25}{25 + 5} = 0{,}833\,(83{,}3\,\%)$$

Indica la capacidad del modelo para identificar correctamente a los pacientes con cáncer. A pesar del alto *accuracy*, el *recall* muestra espacio de mejora en la detección de la clase positiva (cáncer).

- **Precisión:**

$$\text{Precisión} = \frac{TP}{TP + FP} = \frac{25}{25 + 20} = 0{,}55\,(55\,\%)$$

El modelo presenta una baja precisión, lo cual implica que muchas predicciones positivas son en realidad falsos positivos. Este valor debe considerarse con cautela, dependiendo del contexto clínico.

- **F1-Score:**

$$F1 = \frac{2 \cdot \text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}} = \frac{2 \cdot 0{,}55 \cdot 0{,}833}{0{,}55 + 0{,}833} \approx 0{,}662\,(66{,}2\,\%)$$

El *F1-score* refleja un balance bajo entre precisión y sensibilidad, indicando que la capacidad del modelo para clasificar correctamente la clase minoritaria (cáncer) aún no es adecuada.

Otras métricas no tan básicas:

Receiver Operating Characteristic (ROC)

La **curva ROC** es una representación gráfica que muestra el rendimiento de un clasificador binario a diferentes umbrales de decisión. En el eje x se representa la tasa de falsos positivos (FPR) y en el eje y la tasa de verdaderos positivos (TPR o *Recall*). Una curva más cercana a la esquina superior izquierda indica un mejor desempeño del modelo.

Area Under the Curve (AUC)

El **AUC** corresponde al área bajo la curva ROC. Su valor varía entre 0 y 1, donde un valor de 0,5 indica un modelo sin capacidad de discriminación (equivalente a un clasificador aleatorio), mientras que un valor cercano a 1 representa un modelo con alto poder de discriminación.

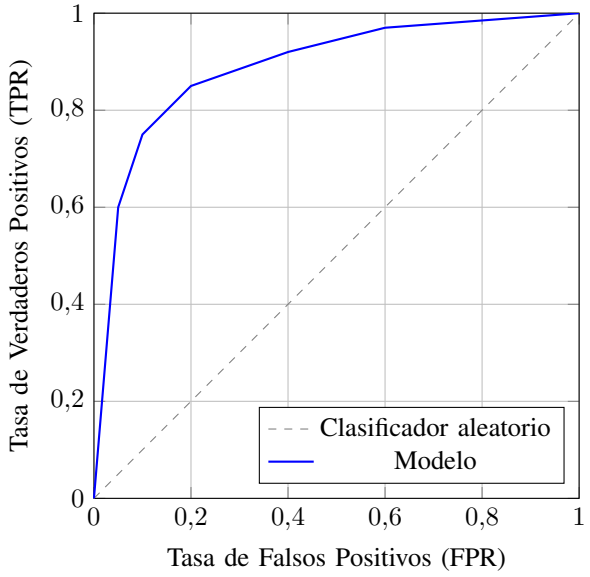

Figura 3. Ejemplo de curva ROC. Un área bajo la curva (AUC) más cercano a 1 indica mejor rendimiento.

De darse un caso en el que la curva, como por ejemplo en la figura 3, se acercara mucho a la recta perdería valor porque estaría dando los resultados incorrectos, de una forma casi que aleatoria. El área bajo la curva debe de ser de al menos 0,5.

Procesamiento de datos

Problemas encontrados

- Incompletitud: valores faltantes en atributos importantes, ej. Si el producto estaba en oferta.
- Inexactitud o ruido: errores y valores atípicos en las transacciones.
- Inconsistencia: discrepancias en los códigos de departamentos o categorías.

Se dio como comparación el caso de la diabetes y la presión sanguínea. Si se registran mediciones, se espera que tenga un valor sino tiene presión sanguínea entonces no tendría sentido o el sujeto estaría muerto, sería un valor que no aporta pero es importante no elimiar el registro ya que podrían haber otras features que si aporten valor.

Por esta razón, los datasets requiren preprocesamiento antes de aplicar técnicas de minería o aprendizaje. Es un problema del mundo real.

Porque pueden ser inexactos?

- Instrumentos de recolección de datos defectuosos
- Errores humanos o computacionales en la entrada de datos
- Usuarios que ingresan valores falsos para campos obligatorios (ej. Fecha por defecto '1 de enero' para ocultar cumpleaños)
  - Conocido como datos faltantes disfrazados
- Inconsistencias en convenciones de nombres, códigos o formatos (ej. fechas con distintos formatos)

- Tuplas duplicadas que requieren procesos de data cleaning

Por qué los datos pueden estar incompletos?

- Atributos de interés no siempre están disponibles
- No se incluyen porque no se consideraron importantes en el momento de la entrada
- Datos relevantes no se registran por malentendidos o fallos de equipo
- Datos inconsistentes con otros registros pueden ser eliminados
- Historial o modificaciones de datos pasados pueden no haberse registrado
- Valores faltantes en atributos clave pueden necesitar ser inferidos.

Por que los datos pueden ser inconsistentes?

- Diferencias en convenciones de nombres o códigos usados para clasificar elementos
- Formatos de entrada distintos para un

Principales tareas en el preprocesamiento de datos:

- **Data cleaning (limpieza de datos):** eliminación de ruido, corrección de inconsistencias y tratamiento de valores faltantes.
- **Data integration (integración de datos):** combinación de información proveniente de múltiples fuentes heterogéneas en un repositorio coherente y unificado.
- **Data reduction (reducción de datos):** disminución del volumen mediante selección de atributos relevantes, reducción de dimensionalidad o discretización.
- **Data transformation (transformación de datos):** incluye normalización, estandarización, agregación y construcción de nuevas variables.
- **Data discretization (discretización de datos):** transformación de atributos continuos en atributos categóricos para facilitar el análisis y la modelización.

Data Cleaning: tratamiento de valores faltantes y ruido:

- **Tratamiento de valores faltantes:**
  - Ignorar tuplas con valores faltantes (puede llevar a la pérdida de datos).
  - Completar manualmente los valores faltantes (costoso y poco práctico en grandes *datasets*).
  - Usar un valor global constante (por ejemplo: "desconocido", ∞).
  - Rellenar con la media, mediana o moda. También puede hacerse por clase.
    - Ejemplo: en clasificación de clientes por riesgo crediticio, reemplazar con el ingreso promedio de clientes en la misma categoría de riesgo.
  - Inferir valores mediante modelos estadísticos o de aprendizaje automático (regresión, k-NN, árboles de decisión).
- **Binning (agrupación en intervalos):**
  - Reemplazar cada valor por:
    - La media del bin.
    - La mediana del bin.
    - El borde más cercano del bin.
  - Ejemplo: valores de salarios.
- **Suavizado de ruido:**
  - Ajustar una función matemática a los datos (puede ser lineal o no lineal).
    - Ejemplo: ventas mensuales con fluctuaciones; se ajusta una regresión lineal para capturar la tendencia general y se consideran ruido los valores que se desvían demasiado.
  - Aplicar técnicas de filtrado, como la media móvil.
    - Ejemplo: datos diarios de temperatura con ruido; se calcula la media móvil de 7 días.

Data Integration: manejo de redundancia:

- La misma información puede estar registrada varias veces o con valores distintos.
  - Ejemplo: un cliente registrado como "Juan Pérez" y también como "J. A. Pérez".
  - Direcciones almacenadas como "San José, Costa Rica" en una base de datos y como "SJ-CR" en otra.
- Uso de pruebas estadísticas para detectar redundancia o asociación entre variables:
  - **Prueba de correlación χ^2 para datos nominales:** mide la asociación entre variables categóricas.
  - Hipótesis de independencia:

$$H_0 : P(A_i \cap B_j) = P(A_i)\,P(B_j)$$

  - Las variables se consideran independientes si:

$$\chi^2_{\text{calculado}} \leq \chi^2_{\alpha,df}$$