Apuntes de la Clase Semana 7 2025
Curso de Inteligencia Artificial

Rafael Vargas Solis
Apuntes del 16 de Setiembre de 2025

***Resumen*—Este documento presenta un resumen de los temas clave abordados en la clase de Inteligencia Artificial del 16 de setiembre de 2025. Comienza describiendo las diferencias fundamentales entre la regresión lineal y la regresión logística, así como las técnicas comunes para el manejo de valores atípicos y las estrategias para enfrentar el sesgo y la varianza en los modelos de aprendizaje automático. Posteriormente, se discuten las métricas de evaluación tanto para clasificación como para regresión, incluyendo la matriz de confusión, precisión, exhaustividad (recall), F1-score, curva ROC y el área bajo la curva (AUC), apoyadas en ejemplos prácticos. Finalmente, el documento resalta los problemas más frecuentes en la calidad de los datos —como la incompletitud, la inexactitud y la inconsistencia— y enfatiza la importancia de las tareas de preprocesamiento, tales como limpieza, integración, reducción, transformación y discretización, para garantizar el desarrollo de modelos predictivos robustos y confiables.**

***Index Terms*—Inteligencia Artificial, Aprendizaje Automático, Regresión, Valores Atípicos, Sesgo, Varianza, Métricas de Evaluación, Matriz de Confusión, ROC, AUC, Preprocesamiento de Datos**

I. Preguntas del quiz

1. ¿Cuál es la diferencia entre regresión lineal y regresión logística?

La *regresión lineal* se utiliza para predecir variables continuas, ajustando una recta que minimiza el error cuadrático medio. Por ejemplo, estimar el precio de una vivienda según su tamaño.

En cambio, la *regresión logística* se aplica a problemas de clasificación, donde la variable dependiente es categórica (binaria en la mayoría de casos). Utiliza la función sigmoide para mapear los valores de entrada en probabilidades entre 0 y 1. Ejemplo: predecir si un estudiante aprobará o no un curso.

2. Describa tres técnicas para el tratamiento de datos sobresalientes (outliers).

Los *outliers* o valores atípicos son observaciones que se alejan significativamente del patrón general de los datos y pueden afectar la precisión de los modelos predictivos. Existen diversas estrategias para tratarlos, entre las cuales destacan:

- **Eliminación:** Consiste en remover los outliers identificados cuando se determina que son producto de errores de medición, registros defectuosos o inconsistencias en la recolección de datos. Esta técnica debe aplicarse con cautela para no eliminar información valiosa.
- **Transformación de variables:** Aplicar funciones matemáticas como logaritmos, raíces cuadradas o escalados que reduzcan la magnitud de los valores extremos. De esta manera, se disminuye su influencia en la varianza del modelo y se mejora la distribución de los datos.
- **Winsorización (recorte):** Sustituir los valores atípicos por valores más cercanos dentro de un rango definido, usualmente basado en percentiles (por ejemplo, 1% y 99%). Esta técnica conserva la estructura general de los datos y evita que los valores extremos distorsionen los resultados.

3. Mencione dos técnicas para evitar un alto sesgo y dos para evitar alta varianza.

En el aprendizaje automático, es fundamental lograr un equilibrio entre *sesgo* y *varianza* para obtener modelos con buena capacidad de generalización. A continuación, se describen algunas técnicas para abordar ambos problemas:

Para reducir sesgo (underfitting):

- **Incrementar la complejidad del modelo:** Utilizar modelos más sofisticados, como polinomiales en lugar de lineales, redes neuronales más profundas o algoritmos no lineales, permite capturar relaciones más complejas entre las variables.
- **Incorporar nuevas variables o características:** Mediante técnicas de *feature engineering*, se pueden incluir atributos relevantes que enriquezcan la información disponible, mejorando así la capacidad predictiva del modelo.

Para reducir varianza (overfitting):

- **Aplicar regularización:** Métodos como L1 (Lasso) y L2 (Ridge) añaden penalizaciones a los coeficientes del modelo, limitando su magnitud y evitando que el modelo se ajuste excesivamente a los datos de entrenamiento.
- **Aumentar los datos o usar técnicas de ensamble:** Incrementar el tamaño del conjunto de entrenamiento o aplicar métodos como *bagging* y *random forest* mejora la estabilidad del modelo y reduce la sensibilidad al ruido de los datos.

4. ¿Cuál es la derivada de la función sigmoide $\sigma(x) = \frac{1}{1+e^{-x}}$?

La función sigmoide se define como:

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

Su derivada es:

$$\sigma'(x) = \sigma(x)\left(1 - \sigma(x)\right) \tag{2}$$

II. Métricas

Son medidas que se utilizan para indicar el rendimiento de un modelo predictivo. 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. El uso de benchmarks permite establecer un estándar de referencia que facilita la reproducibilidad y la comparación entre diferentes enfoques.

En general, las métricas pueden dividirse en:

- **Métricas de clasificación**: accuracy, precision, recall, F1-score, ROC y AUC.
- **Métricas de regresión**: error cuadrático medio (MSE), error absoluto medio (MAE) o coeficiente de determinación (R^2).

III. Matriz de Confusión

La *matriz de confusión* organiza los resultados de un modelo de clasificación en función de las predicciones realizadas y las clases reales. Se definen cuatro componentes:

- **True Positive (TP)**: positivos correctamente clasificados.
- **False Positive (FP)**: negativos clasificados incorrectamente como positivos (error tipo I).
- **True Negative (TN)**: negativos correctamente clasificados.
- **False Negative (FN)**: positivos clasificados incorrectamente como negativos (error tipo II).

En problemas multiclase, la matriz puede extenderse a $N \times N$. Un clasificador ideal concentra todos los valores en la diagonal principal.

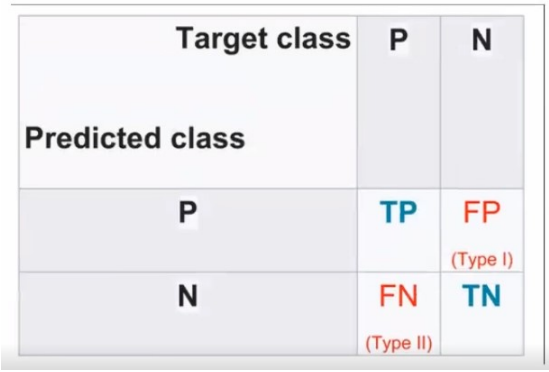

Fig. 1. Ejemplo de matriz de confusión en clasificación binaria.

IV. Métricas Clásicas

A partir de la matriz de confusión se derivan las métricas más utilizadas:

A. Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Mide la proporción de predicciones correctas. Es útil en datos balanceados, pero engañosa en clases desbalanceadas.

B. Precisión

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Indica qué proporción de predicciones positivas fueron correctas. Relevante cuando los falsos positivos son costosos.

C. Recall (Sensibilidad)

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Mide la capacidad del modelo para identificar correctamente los positivos. Importante en contextos donde los falsos negativos son críticos.

D. F1-Score

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{6}$$

Es la media armónica entre precisión y recall, usada en casos de desbalance de clases.

V. Caso de Estudio

Se evaluó un modelo de detección de cáncer con 1000 pacientes.

- Clase positiva: 30 pacientes con cáncer.
- Clase negativa: 970 pacientes sin cáncer.

Matriz de confusión:

	Cáncer	No cáncer
Cáncer	25 (TP)	20 (FP)
No cáncer	5 (FN)	950 (TN)

Resultados:

- **Accuracy:** $\frac{25+950}{1000} = 97.5\%$
- **Recall:** $\frac{25}{25+5} = 83.3\%$
- **Precisión:** $\frac{25}{25+20} = 55\%$
- **F1-Score:** $\frac{2 \cdot 0.55 \cdot 0.833}{0.55+0.833} \approx 66.2\%$

A pesar del alto valor de accuracy, las métricas muestran limitaciones en la detección de la clase positiva.

VI. Métricas Avanzadas

A. Curva ROC

La curva ROC (*Receiver Operating Characteristic*) muestra el desempeño de un clasificador binario para distintos umbrales. Representa la Tasa de Verdaderos Positivos (TPR) frente a la Tasa de Falsos Positivos (FPR).

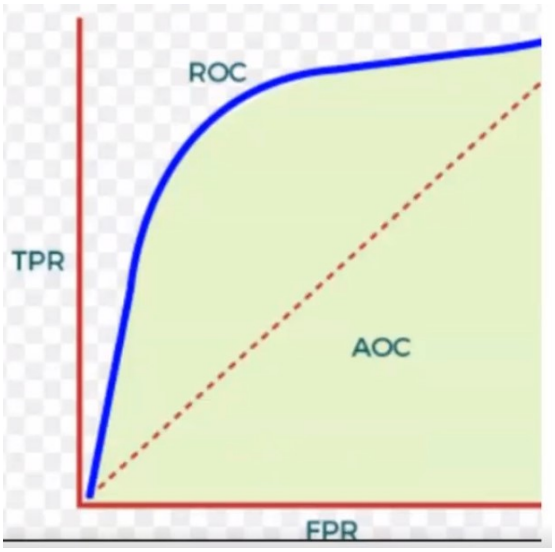

Fig. 2. Ejemplo de curva ROC y cálculo de AUC.

B. Área Bajo la Curva (AUC)

El AUC mide el área bajo la curva ROC:

- AUC = 0.5: clasificador aleatorio.
- AUC cercano a 1: modelo con gran poder de discriminación.

VII. Procesamiento de Datos

A. Problemas encontrados en la calidad de datos

En escenarios reales, los datos suelen presentar problemas que afectan directamente la efectividad de los algoritmos de minería y aprendizaje automático. Los principales son:

- **Incompletitud:** valores faltantes en atributos importantes. Ejemplo: si un producto estaba en oferta y no se registró la variable.
- **Inexactitud o ruido:** errores de medición, valores atípicos o entradas anómalas en transacciones.
- **Inconsistencia:** discrepancias en nombres, códigos o formatos. Ejemplo: fechas almacenadas como `DD/MM/AAAA` en una base de datos y como `MM-DD-YYYY` en otra.

Un caso ilustrativo es la recolección de datos médicos: en la medición de presión sanguínea, un valor faltante no implica que el registro deba eliminarse, ya que otras características (edad, peso, historial clínico) sí aportan información valiosa. Esto demuestra que los datasets requieren **preprocesamiento** antes de aplicar técnicas de minería o aprendizaje.

B. Causas de datos defectuosos

- Instrumentos de recolección defectuosos.
- Errores humanos o computacionales en la entrada de datos.
- Valores falsos en campos obligatorios (ejemplo: fecha por defecto "1 de enero" para ocultar cumpleaños), conocidos como **datos faltantes disfrazados**.
- Inconsistencias en convenciones de nombres, códigos o formatos.
- Registros duplicados que requieren procesos de *data cleaning*.

C. Causas de incompletitud

- Atributos de interés no siempre disponibles o considerados irrelevantes en el momento de captura.
- Fallos técnicos o malentendidos durante la recolección.
- Eliminación de registros por inconsistencias.
- Ausencia de historial o modificaciones no registradas.
- Valores faltantes en atributos clave que deben ser inferidos.

D. Causas de inconsistencias

- Diferencias en convenciones de nombres o códigos.
- Formatos de entrada distintos para un mismo atributo.
- Conflictos entre bases de datos heterogéneas.
- Errores en la integración de fuentes múltiples.
- Actualizaciones parciales que dejan registros contradictorios.

E. Principales tareas en el preprocesamiento

- **Data cleaning:** eliminación de ruido, corrección de inconsistencias y tratamiento de valores faltantes.
- **Data integration:** combinación de datos provenientes de múltiples fuentes heterogéneas en un repositorio unificado.
- **Data reduction:** reducción de volumen mediante selección de atributos, reducción de dimensionalidad o discretización.
- **Data transformation:** normalización, estandarización, agregación y construcción de nuevas variables.
- **Data discretization:** transformación de atributos continuos en categorías o intervalos.

F. Data Cleaning: Tratamiento de valores faltantes y ruido

1) Valores faltantes:

- Ignorar tuplas con valores faltantes (riesgoso si se pierde mucha información).
- Completar manualmente (costoso en grandes datasets).
- Usar un valor global constante (ej. "desconocido").
- Rellenar con la media, mediana o moda, también por clase.
- Inferir valores mediante modelos estadísticos o de ML (regresión, k-NN, árboles de decisión).

2) Binning: **Binning** agrupa valores en intervalos (*bins*) y reemplaza cada valor por:

- La media del bin.
- La mediana del bin.
- El borde más cercano del bin.

Ejemplo: salarios ruidosos [2950, 3000, 3020, 8000]. El bin (2900–3100) se reemplaza por la media (2990), mientras que 8000 queda como posible outlier.

3) Suavizado de ruido:

- Ajustar una función matemática (lineal o no lineal) para suavizar fluctuaciones. Ejemplo: regresión lineal en ventas mensuales.
- Aplicar técnicas de filtrado como la **media móvil**:

$$MA_7(t) = \frac{1}{7}\sum_{i=0}^{6} x_{t-i}$$

donde x_t es el valor en el día t. Esto genera una curva suavizada que refleja la tendencia real.

G. Data Integration: Manejo de redundancia

La misma información puede estar registrada varias veces o con diferencias. Ejemplo: un cliente como "Juan Pérez" en una base de datos y "J. A. Pérez" en otra. Se aplican pruebas estadísticas como la **chi-cuadrado** (χ^2) para detectar redundancia o asociaciones entre variables categóricas:

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

Si $\chi^2_{calculado} \leq \chi^2_{\alpha,df}$, se acepta la hipótesis de independencia.

REFERENCES

[1] A. Burkov, *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019. [Online]. Available: https://themlbook.com/