Apuntes Semana 7 - 18/09/2025

1st Darío Espinoza Aguilar
2020109109
Computer Engenieering
darioespinoza477@estudiantec.cr

***Abstract*—Este documento corresponde a los apuntes de la clase del 18 de septiembre de 2025, donde se repasan los conceptos de métricas y fórmulas para evaluar los modelos. Además, se hace un repaso del proceso de preparación y procesamiento de datos antes de poder entrenar el modelo.**

***Index Terms*—Métricas, data cleaning, procesamiento de datos, datasets**

I. Introducción

Se mencionó que se iba a dejar la Tarea 2 la otra semana y el proyecto 1 la semana que le sigue. Además, nos dio la invitación al evento de Ingeniería para lo que quiera asistir.

En la parte de noticias, se mencionó que Xbox Gaming esta desarrollado un IA para desarrollar juegos viejos, más que un desarrollador es un porteador de juego viejos para que puedan ser jugables.

El profe menciona que se lanzó un protocolo nuevo para pago a través de agentes. El protocolo es AP2, fue lanzado por Google y lo que se busca es que los pagos por medios electrónicos se puedan realizar sin necesidad de intervención humana. Menciono la posibilidad de hacer pasantía en la empresa donde él trabaja.

Se hablo de las próximas evaluaciones que vamos a tener, el profesor menciono que la mayoría de las tareas programadas ya sean proyectos o tareas van a ser de modelos clasificatorios ya que tiene afinidad por ellos y que se van a tener tareas de investigación.

II. Repaso de Métricas

Son las métricas asociadas a un modelo que nos indica el rendimiento de un modelo predictivo. La forma más objetiva de evaluar y comparar un modelo. En todos los modelos que se sacan siempre hay métricas o benchmarks que nos indican que tan bueno es el modelo.

Se repaso lo que es la matriz de confusión. Que de los algoritmos de clasificación tenemos 2 etiquetas se pueden ver como positivo y negativo, y de esas 2 etiquetas se pueden tener 4 4 posibles valores que la matriz de confusión nos ayuda a visualizar esos valores. En esta matriz se tienen los Target Class que es la etiqueta tenemos en el dataset y el predicted class que es la predicción de nuestro modelo. Los cuatros combinaciones

- **TP:** True positive
- **TN:** True negative
- **FP:** False positive
- **FN:** False negative

A partir de estos se pueden hacer varias combinaciones para calcular ciertas métricas

1) Accuracy: Clasificación correcta entre todos los intentos. La fórmula es la siguiente:

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

Es útil cuando los errores por clase son igual de importantes. Otorga importancia igual a todas las claases.

2) Precision: Mide los errores tipo 1 (FP). Tasa de predicciones positivas correctas entre todas las predicciones positivas.

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

3) Recall: Mide los errores de tipo 2 (FN). Tasa de predicciones correctas entre todos los ejemplos positivos del conjunto de datos.

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

4) F1-Score: Esta es un métrica que contempla ambos errores. Comúnmente utilizada en problemas de clasificación, especialemnte cuando tenemos desequelibrio de clases.

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

A. ROC (Receiver Operating Characteristic)

El área bajo la curva siempre tiene que ser > 0.5. Si es $= 5$ lo que vamos a tener es un random clasifier, es un clasificador aleatorio. Si es cada vez mayor a 0.5 el modelo va siendo mejor.

III. Respaso de procesamiento de datos

En la práctica podemos tener ciertos problemas con nuestros datos:

- **Incompletitud:** valores faltantes en atributos importantes
- **Inexactitud o ruido:** errores y valores atípico en las transacciones
- **Inconsistencia:** discrepancia en lo código de departamentos o categoría

A. ¿Por qué los datos 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.
  - Conocido como **datos faltantes disfrazados**.
- Inconsistencia en convenciones de nombres, códigos o formatos.
- Tuplas duplicadas que requieren procesos de **data cleaning**.

B. ¿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 del equipo
- Datos inconsistentes con otros registros pueden ser eliminados
- Historial o modificaciones de datos pasados pueden no haberse registrado
- Valores faltantes en atributos claves pueden necesitar ser **inferidos**

C. ¿Por qué los datos pueden ser inconsistentes?

- Diferencias en convenciones de nombres o códigos usados para clasificar elementos
- Formatos de entrada distintos para un mismo atributo
- Conflictos entre bases de datos o sistemas que manejan el
- Errores al integrar datos de múltiples fuentes heterogéneas
- Actualizaciones parciales o incorrectas que dejan registros contradictorios

D. Principales tareas en el preprocesamiento de datos

- **Data Cleaning** Eliminación de ruido, corrección de inconsistencias, tratamiento de valores faltantes
- **Data Integration** Combinación de datos de múltiples fuentes heterogéneas en un repositorio coherente
- **Data Reduction** Reducción de volumen mediante selección de atributos, reducción de dimensionalidad o discretización
- **Data Transformation** Normalización, estandarización, agregación, construcción de nuevas variables
- **Data Discretization** Transformación de atributos continuos en atributos categóricos

1) Data cleaning (missing values):

- Ignorar tuplas con valores faltantes (riesgo si la perdida de datos es significtiva)
- Completar manualmente los valores (costoso y poco práctico en grandes datasets)
- Usar un valor global constante.
- Rellenar con la media (normal), mediana o moda
- Inferir valores mediante modelos estadistico o de ML (regresión, k-NN, árboles de decisión)

2) Data Cleaning (Noisy Data):

- Agrupar valores en intervalos(bins).
- Se puede utilizar con datos muy ruidosos, se reemplaza el valor por: la media del bin, la mediana del bin, el borde más cercano del bin.

3) Data Integration (Redundancy Handling):

- Se ajusta una función matemática (lineal o no lineal) para suavizar el ruido de los datos.
- Aplicar técnicas de filtrado para suavizar fluctuaciones, se puede utilizar la media móvil (utilizar los últimos 7 elementos para ir calculando la media).