

# Actividad 3

## Métricas de Evaluación

**Profesor: Pablo Estévez**

Auxiliar: Pablo Cornejo

Ayudantes: Diego Castillo, Andrés González, Sebastián Guzmán, Francisco Soto

## Instrucciones Generales

- **Importante:** Todos los gráficos realizados **deben** tener una etiqueta en el eje X y en el eje Y, además de un título. Las etiquetas y los títulos son libres a ser determinados por cada uno, sin embargo, deben estar relacionados con el contexto del gráfico.
- Por favor comentar los códigos. Si hay subsecciones, se recomienda comentar los lugares donde comienza cada subsección.
- Utilicen nombres significativos para variables, funciones y clases. Esto les ayudará principalmente a ustedes cuando lean sus códigos en un futuro y también al resto de la gente con la que trabajen luego.

## Instrucciones Específicas

- Las preguntas 1 y 2 se responden en el mismo Jupyter notebook entregado junto a la actividad. Esto vale tanto para el código como para las respuestas escritas.
- No se pueden utilizar clases ni funciones de **sklearn**.

## P1: Implementación de métricas

En esta sección implementaremos métricas de clasificación. Para su implementación considere que tiene un vector de etiquetas reales llamado `true_labels` y un vector de etiquetas predichas llamado `predicted_labels`.

1. Implemente una función que reciba como entrada las etiquetas reales y las etiquetas predichas (en ese orden). La función debe retornar el número de verdaderos negativos, falsos positivos, falsos negativos y verdaderos positivos en este orden.
2. Implemente una función que reciba como entrada las etiquetas reales y las etiquetas predichas (en ese orden) y grafique una matriz de confusión. Esta debe contener las etiquetas reales en el eje  $y$  y las etiquetas predichas en el eje  $x$ . Se recomienda utilizar la función `heatmap` de `seaborn`.
3. Implemente una función para cada una de las siguientes métricas. Las entradas de estas funciones deben ser el número de verdaderos negativos, falsos positivos, falsos negativos y verdaderos positivos en este orden:
  - a) Recall (o Sensibilidad)
  - b) Especificidad
  - c) Tasa de falsos positivos
  - d) Tasa de falsos negativos
  - e) Precision
  - f) F1-Score
4. Utilizando dos funciones de la parte anterior, implemente una función que tome como argumento un vector con las etiquetas reales de los datos y un vector con la probabilidad de que cada dato sea de la clase positiva y grafique una curva ROC.

## P2: Análisis de métricas

Para esta pregunta sólo se pueden utilizar las funciones creadas en la pregunta 1. Considere un problema para el cual usted tiene un clasificador que permite detectar si una persona tiene una enfermedad levemente peligrosa o no. Si la persona tiene esta enfermedad, entonces es sometida a un tratamiento muy invasivo pero que permite curar la enfermedad. Por el contrario, si la persona tiene la enfermedad y no es tratada, entonces las secuelas de la enfermedad se manifiestan pero son menores. En este ejemplo clasificar a una persona sana como enferma es caro ya que se sometería a la persona a un tratamiento invasivo sin necesitarlo, sin embargo, clasificar a una persona enferma como sana de todas formas tiene un costo pero menor que la otra opción. Este es un claro ejemplo en el cual los dos tipos de clasificaciones erróneas (falsos positivos y falsos negativos) son asimétricos en términos de costos.

Considere que la clase positiva corresponde a predecir que la persona tiene la enfermedad y por lo tanto se le realiza el tratamiento (el caso negativo es lo contrario). Responda las siguientes preguntas utilizando los datos generados en el notebook de la actividad. Los datos corresponden a lo siguiente:

- `y_true` es el vector de etiquetas reales (la condición real de cada paciente).
- `predicted_proba` es la predicción de nuestro clasificador para cada paciente. Esta predicción corresponde a la probabilidad de que la persona sea de la clase positiva.

Responda:

1. Considerando un umbral de probabilidad igual a 0.5 para la clasificación (probabilidades mayores a 0.5 se consideran clase positiva y menores a 0.5 negativas), obtenga las distintas métricas calculadas en la pregunta P1.3 y grafique la matriz de confusión (P1.2).
2. Utilizando los vectores entregados, grafique una curva ROC utilizando la función generada en la P1.4.
3. Responda. Considerando la asimetría de la clasificación, ¿es mejor fijar un umbral mayor o menor a 0.5 para clasificar los datos?, ¿por qué?
4. Fije un umbral arbitrario que siga su decisión de la pregunta P2.3 (mayor o menor a 0.5) y calcule nuevamente la matriz de confusión. Responda, ¿se obtuvo el resultado que esperaba?, ¿qué sucede si continúa incrementando/disminuyendo el umbral?