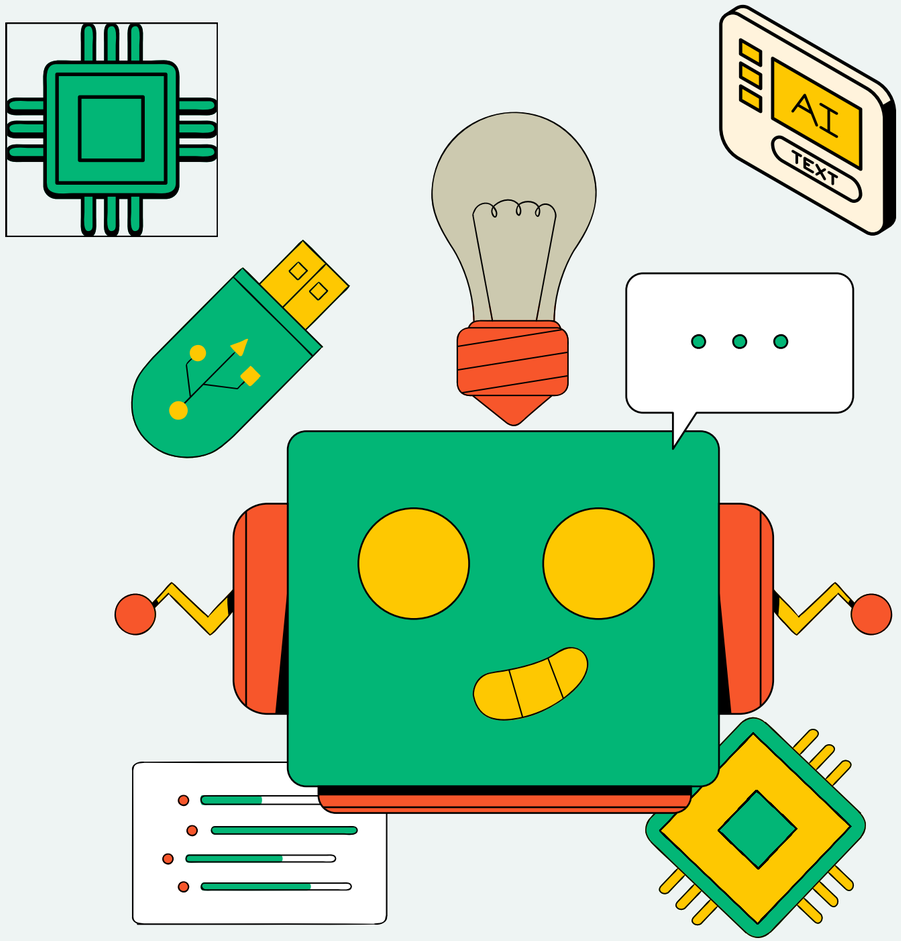
INSTITUTO TECNOLOGICO DE BUENOS AIRES



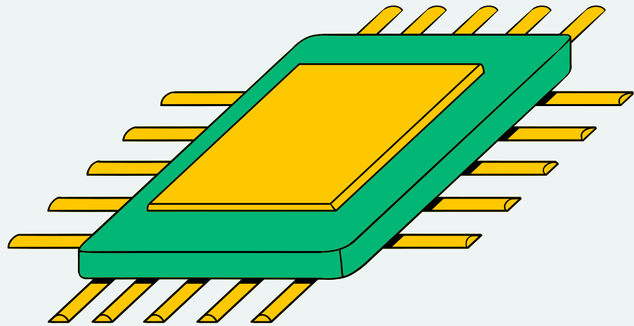
MACHINE LEARNING TEOREMA DE BAYES Y REDES BAYESIANAS

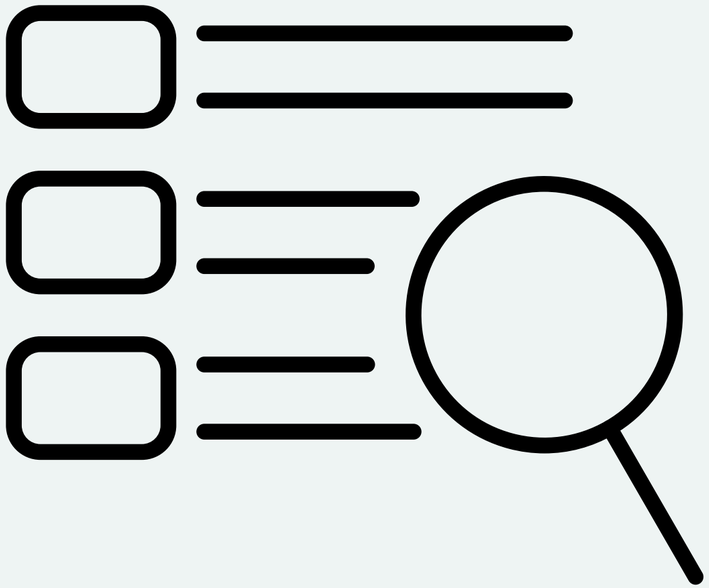
GRUPO N1

FRANCO ROSSI

TOMÁS MENDIETA

SEBASTIÁN TUESTA





TEMAS

- **Ejercicio 1: Implementar Bayes para clasificar preferencias entre ingleses y escoceses.**

- **Ejercicio 2: Categorizar noticias usando Bayes y evaluar con métricas.**

- **Ejercicio 3: Predecir admisión universitaria con Bayes basado en variables discretas.**

EJERCICIO 1

Considerando el vector de atributos binarios:
(scones, cerveza, whisky, avena, fútbol)

El vector x = (1, 0, 1, 1, 0) significa que se trata de una persona que le gustan los scones, no
toma cerveza, le gusta el whisky y la avena pero no ve futbol.
En el archivo
PreferenciasBritanicos.xls se encuentran las preferencias de 6 personas inglesas y 7
personas escocesas.

- **Implementar el clasificador ingenuo de Bayes.**

- **Determinar si corresponden a las preferencias de una persona inglesa o escocesa.**
  - **x1 = (1, 0, 1, 1, 0)**
  - **x2 = (0, 1, 1, 0, 1)**

- **Detallar paso a paso cómo calcular las probabilidades.**

scones	cerveza	wiskey	avena	futbol	Nacionalidad
0	0	1	1	1	**I**
1	0	1	1	0	**I**
1	1	0	0	1	**I**
1	1	0	0	0	**I**
0	1	0	0	1	**I**
0	0	0	1	0	**I**
1	0	0	1	1	**E**
1	1	0	0	1	**E**
1	1	1	1	0	**E**
1	1	0	1	0	**E**
1	1	0	1	1	**E**
1	0	1	1	0	**E**
1	0	1	0	0	**E**

- **Función Extraer_data(data):**

- **Esta función se encarga de leer los datos del archivo CSV (PreferenciasBritanicos.csv), que contiene las preferencias de las personas inglesas y escocesas.**

- **Se dividen los datos en dos estructuras:**

    - **`BD_1`: Contiene los atributos (scones, cerveza, whisky, avena, fútbol) de todas las personas junto con su nacionalidad (`I` o `E`).**

    - **`BD_2`: Almacena los atributos por separado para cada clase (`Ingleses` y `Escoceses`) en `BD_2[0]` y `BD_2[1]`, respectivamente. Esto permite posteriormente calcular las probabilidades condicionadas para cada clase.**

Función calcular_probabilidades(BD_1, BD_2):

- **Función para implementar el clasificador ingenuo de Bayes.**

- **Se calculan tres elementos clave:**

  - **Probabilidades a priori (priori): Esto corresponde a la probabilidad de que una persona sea inglesa o escocesa, sin tener en cuenta los atributos. Se calcula a partir de los datos en BD_1.**

  - **Probabilidad de evidencia (evidencia): Esta es la probabilidad de observar cada atributo en general (sin considerar la clase). También se calcula utilizando los datos de BD_1.**

  - **Verosimilitud (verosimilitud): Aquí se calculan las probabilidades condicionadas de observar cada atributo dado que la persona pertenece a una clase específica (Inglesa o Escocesa). Esta información se obtiene de BD_2.**

Función clasificar(lista, evidencia, priori, verosimilitud):

- **Esta función implementa el proceso de clasificación de Bayes.**
- **Se toman las probabilidades calculadas en la función anterior y se aplican para clasificar nuevos vectores de atributos (en este caso, los vectores x1 = (1, 0, 1, 1, 0) y x2 = (0, 1, 1, 0, 1)).**
- **Primero se calcula la probabilidad de la evidencia utilizando la lista de atributos que se quiere clasificar.**
- **Luego, se calculan las probabilidades de verosimilitud para cada clase (Inglesa o Escocesa).**
- **Finalmente, se aplican estas probabilidades para determinar a qué clase pertenece el vector de atributos y se normalizan las probabilidades para asegurarse de que sumen 1.**

Probabilidades a priori

$$P(Ingles) = \frac{NumeroInglese}{TotalPersonas}$$

$$P(Escoces) = \frac{NumeroEscoces}{TotalPersonas}$$

Probabilidades condicionales para clase Inglesa

$$P(atributo1 = 1|Ingles) = \frac{Numerodeinglesesconatributo1 = 1}{TotalIngleses}$$

$$P(atributo2 = 1|Ingles) = \frac{Numerodeinglesesconatributo2 = 1}{TotalIngleses}$$

$$P(atributo1 = 3|Ingles) = \frac{Numerodeinglesesconatributo3 = 1}{TotalIngleses}$$

$$P(atributo1 = 4|Ingles) = \frac{Numerodeinglesesconatributo4 = 1}{TotalIngleses}$$

$$P(atributo5 = 1|Ingles) = \frac{Numerodeinglesesconatributo5 = 1}{TotalIngleses}$$

Probabilidades condicionales para clase Escocesa

$$P(atributo1 = 1|Escoces) = \frac{Numerodeescocecesconatributo1 = 1}{TotalEscocecces}$$

$$P(atributo1 = 2|Escoces) = \frac{Numerodeescocecesconatributo2 = 1}{TotalEscocecces}$$

$$P(atributo3 = 1|Escoces) = \frac{Numerodeescocecesconatributo3 = 1}{TotalEscocecces}$$

$$P(atributo4 = 1|Escoces) = \frac{Numerodeescocecesconatributo4 = 1}{TotalEscocecces}$$

$$P(atributo5 = 1|Escoces) = \frac{Numerodeescocecesconatributo5 = 1}{TotalEscocecces}$$

Probabilidades a priori

$$P(Ingles) = 47.06\%$$

$$P(Escoces) = 52.94\%$$

Probabilidades condicionales para clase Inglesa

$$P(atributo1 = 1|Ingles) = 0.50$$

$$P(atributo1 = 2|Ingles) = 0.50$$

$$P(atributo1 = 3|Ingles) = 0.375$$

$$P(atributo1 = 4|Ingles) = 0.50$$

$$P(atributo1 = 5|Ingles) = 0.50$$

Probabilidades condicionales para clase Escocesa:

$$P(atributo1 = 1|Ingles) = 0.50$$

$$P(atributo2 = 1|Escoces) = 0.56$$

$$P(atributo3 = 1|Escoces) = 0.44$$

$$P(atributo4 = 1|Escoces) = 0.67$$

$$P(atributo5 = 1|Escoces) = 0.44$$

$$P(Ingles|x1) = \frac{P(x1|Ingles) \cdot P(Ingles)}{P(x1)}$$

$P(Ingles|x1)$ es la probabilidad de que la persona sea inglesa dado el vector de atributos X1

$P(x1|Ingles)$ es la probabilidad condicional de observar el vector de atributos X1 dado que la persona es inglesa (esta se obtiene multiplicando las probabilidades de los atributos individuales).

$P(Ingles)$ es la probabilidad a priori de que una persona sea inglesa.

$P(x1)$ es la probabilidad total de observar el vector x1, independientemente de la clase.

$$x1 = (1, 0, 1, 1, 0)$$

$$P(Ingles|x1) = \frac{P(x1|Ingles) \cdot P(Ingles)}{P(x1)} = 0.2426$$

$$P(Escoces|x1) = \frac{P(x1|Escoces) \cdot P(Escoces)}{P(x1)} = 0.7574$$

$$x1 = (0, 1, 1, 0, 1)$$

$$P(Ingles|x2) = \frac{P(x2|Ingles) \cdot P(Ingles)}{P(x2)} = 0.8368$$

$$P(Escoces|x2) = \frac{P(x2|Escoces) \cdot P(Escoces)}{P(x2)} = 0.1632$$

KEY CONCEPTS

machine learning

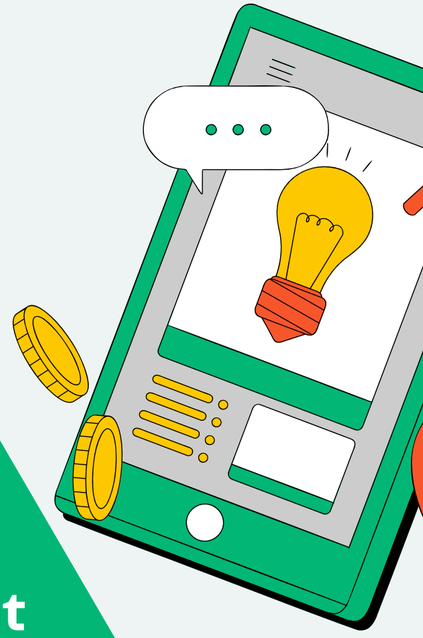
neural networks

natural language processing

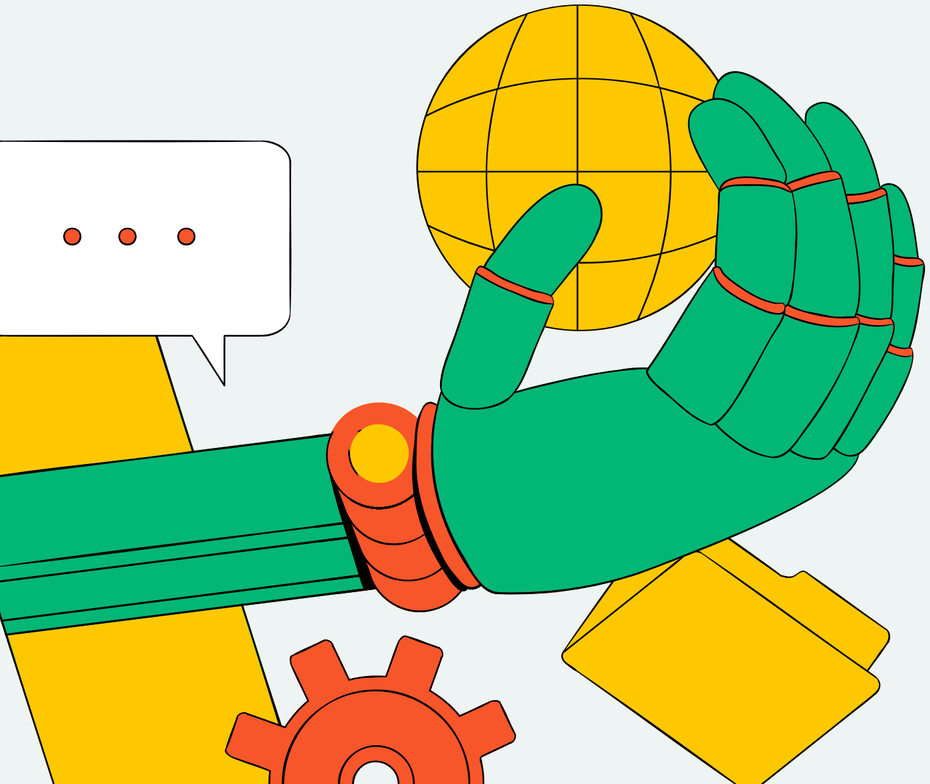
computer vision

reinforcement learning

ethical AI



EJERCICIO 2



1. Implementar un clasificador de texto utilizando el clasificador ingenuo de Bayes. Utilizar el conjunto de datos ”Noticias Argentinas” para clasificar cada noticia según su tipo.

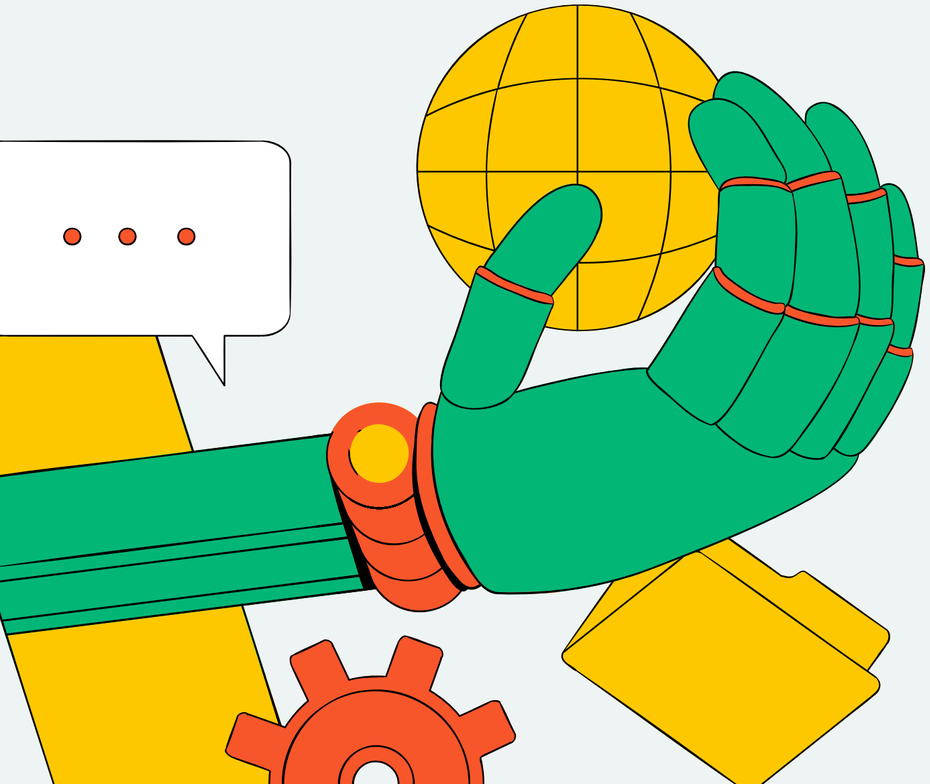
    a. **Utilizar al menos 4 categorías. Justifica la elección de las categorías utilizadas, por ejemplo en base a un análisis preliminar.**
    b. Dividir el conjunto de textos disponible para utilizar una parte de los mismos como conjunto de entrenamiento y otro como conjunto test.
    c. Construir la matriz de confusión.
    d. Calcular las medidas de evaluación Accuracy, Precisión, tasa de verdaderos positivos, tasa de falsos positivos y F1-score. Interpreta estos resultados en el contexto de las noticias clasificadas.
    e. Calcular la curva ROC y analizarla.

CATEGORÍAS

- INTERNACIONAL
- DEPORTES
- SALUD
- CIENCIA Y TECNOLOGÍA
- NACIONAL
- ECONOMÍA

Gracias a un análisis preliminar, encontramos que estas 6 categorías permiten una mejor categorización

EJERCICIO 2



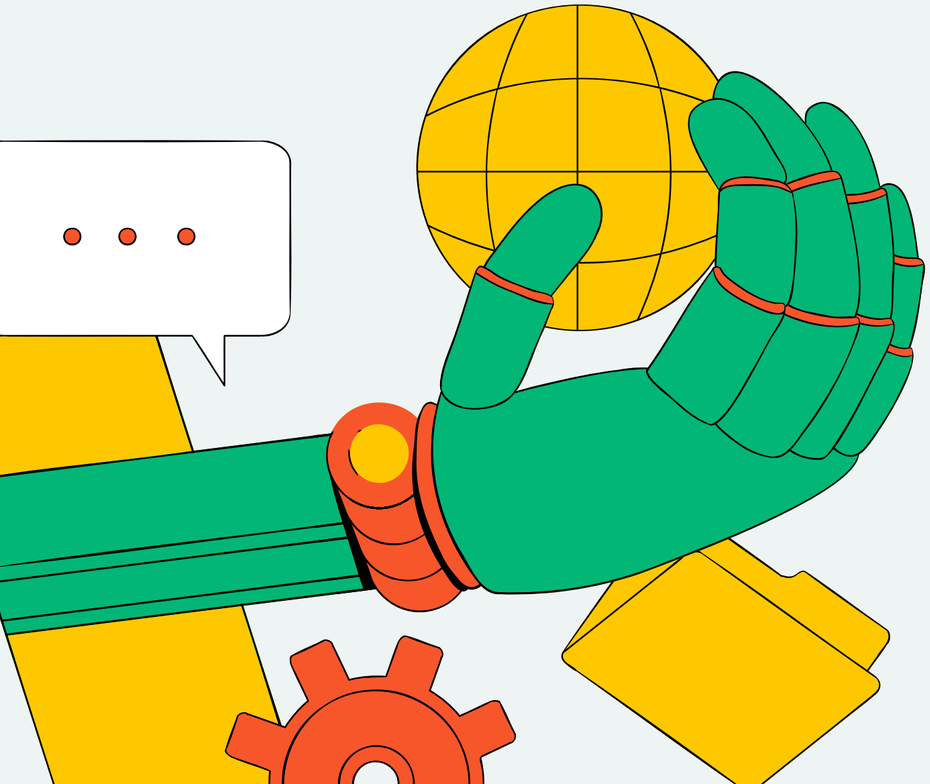
1. Implementar un clasificador de texto utilizando el clasificador ingenuo de Bayes. Utilizar el conjunto de datos ”Noticias Argentinas” para clasificar cada noticia según su tipo.

    a. Utilizar al menos 4 categorías. Justifica la elección de las categorías utilizadas, por ejemplo en base a un análisis preliminar.
    b. **Dividir el conjunto de textos disponible para utilizar una parte de los mismos como conjunto de entrenamiento y otro como conjunto test.**
    c. Construir la matriz de confusión.
    d. Calcular las medidas de evaluación Accuracy, Precisión, tasa de verdaderos positivos, tasa de falsos positivos y F1-score. Interpreta estos resultados en el contexto de las noticias clasificadas.
    e. Calcular la curva ROC y analizarla.

DIVISIÓN DE DATASET

Entrenamiento	Test
70%	30%

EJERCICIO 2



1. Implementar un clasificador de texto utilizando el clasificador ingenuo de Bayes. Utilizar el conjunto de datos ”Noticias Argentinas” para clasificar cada noticia según su tipo.

    a. Utilizar al menos 4 categorías. Justifica la elección de las categorías utilizadas, por ejemplo en base a un análisis preliminar.
    b. Dividir el conjunto de textos disponible para utilizar una parte de los mismos como conjunto de entrenamiento y otro como conjunto test.
    c. **Construir la matriz de confusión.**
    d. Calcular las medidas de evaluación Accuracy, Precisión, tasa de verdaderos positivos, tasa de falsos positivos y F1-score. Interpreta estos resultados en el contexto de las noticias clasificadas.
    e. Calcular la curva ROC y analizarla.

FILTRADO DE DATA

POR CATEGORIA

Solo aquellos titulares dentro de las categorias elegidas

DROPNA()

Eliminamos filas con columnas vacias

SPLIT

División de la data en entrenamiento y testeo

FILTRADO POR STOPWORD

De los titulares quitamos palabras comunes como “día” o “que”

MINUSCULAS

Se convirtió todos las palabras de los titulares a minúsculas

PALABRAS CLAVE

PALABRAS CLAVE

FORMA 1

Uso de IA + intervención humana

FORMA 2

Uso de librerías

```python
palabra_clave={}
for i,categoria in enumerate(categorias_seleccionadas):
    df_cat_1=entrenamiento[entrenamiento["categoria"]==categoria]
    todos_los_titulares = " ".join(df_cat_1['titular'])
    texto_min=todos_los_titulares.lower()
    palabras = texto_min.split()
    palabras_filtradas = [palabra for palabra in palabras if palabra not in stop_words and len(palabra) > 1]

    # Contar las palabras más comunes
    contador_palabras = Counter(palabras_filtradas)

    # Obtener las 20 palabras más comunes
    palabras_comunes = contador_palabras.most_common(50) #300 palabras en total
    # Mostrar el resultado
    print(f"categoria:{categoria}")
    for palabra, frecuencia in palabras_comunes:
        #print(f'{palabra}: {frecuencia}')
        palabra_clave[categoria]=dict(palabras_comunes)
```

MATRIZ DE CONFUSIÓN

FORMA 1

Nuestro modelo está “sesgado” hacia la categoría Nacional

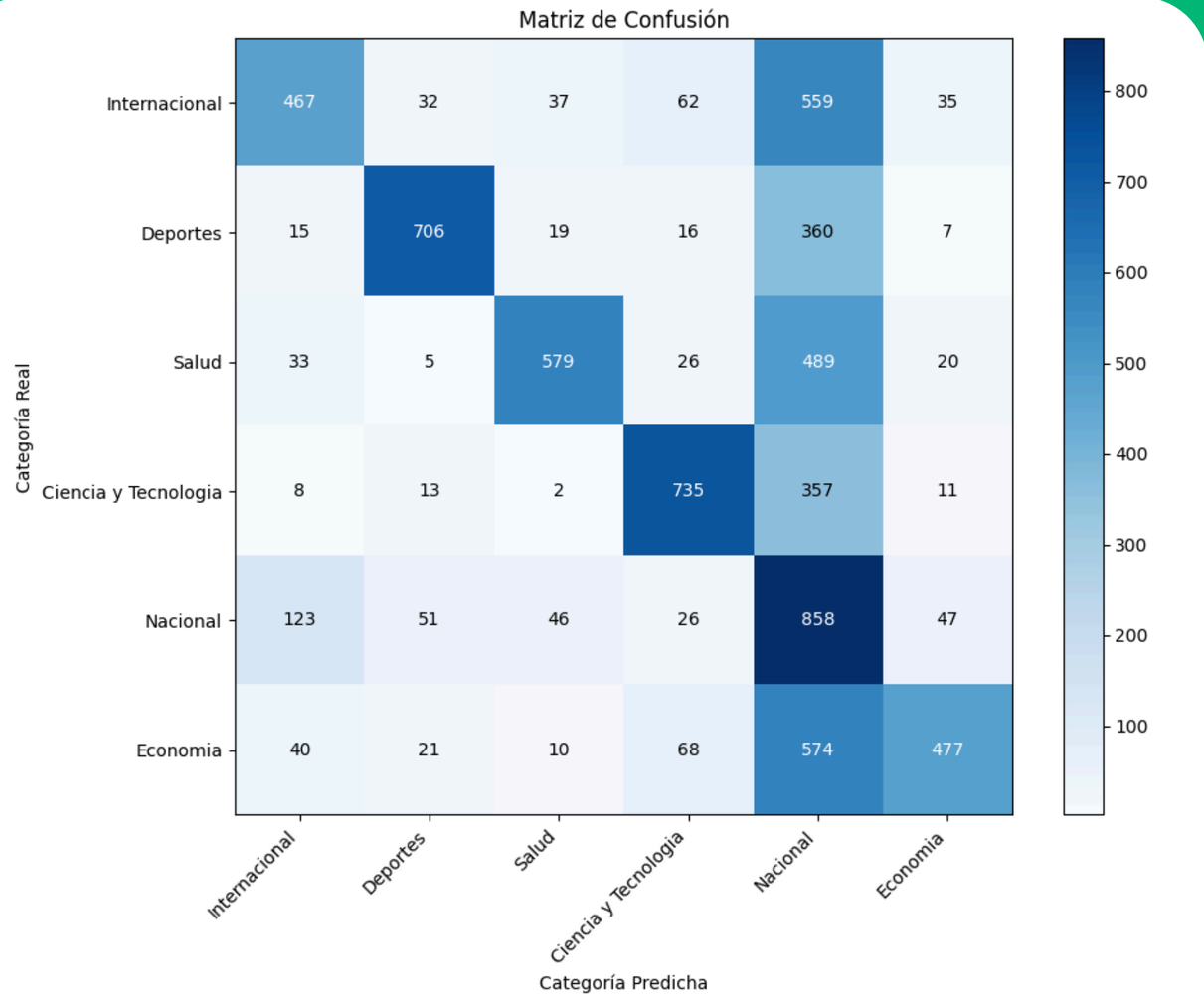
Sebastian Tuesta Pereda

La mayor coincidencia en la predicción es en la categoría “nacional”

Sebastian Tuesta Pereda

La menor coincidencia en la predicción es en la categoría “economía”

Sebastian Tuesta Pereda



MATRIZ DE CONFUSIÓN

FORMA 2

Nuestro modelo está “sesgado” hacia la categoría CyT

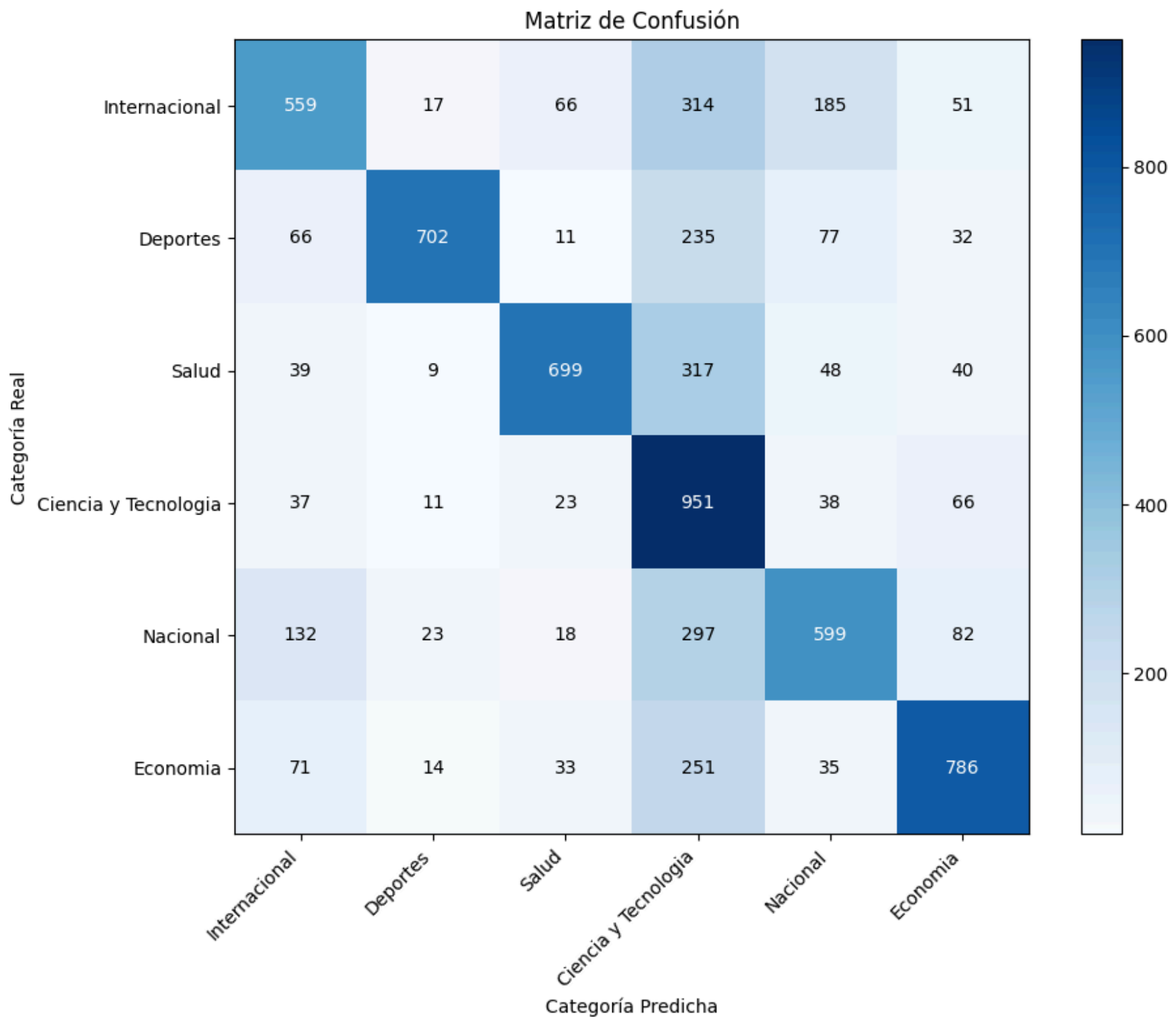
Sebastian Tuesta Pereda

La mayor coincidencia en la predicción es en la categoría “CyT”

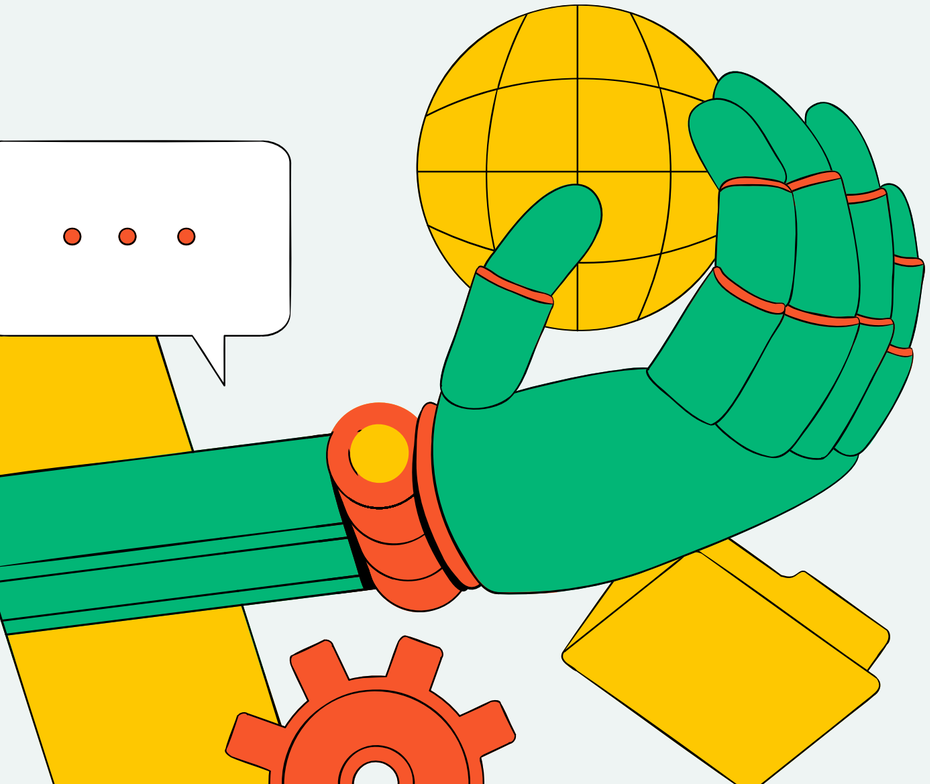
Sebastian Tuesta Pereda

La menor coincidencia en la predicción es en la categoría “Internacional”

Sebastian Tuesta Pereda



EJERCICIO 2



1. Implementar un clasificador de texto utilizando el clasificador ingenuo de Bayes. Utilizar el conjunto de datos ”Noticias Argentinas” para clasificar cada noticia según su tipo.

    a. Utilizar al menos 4 categorías. Justifica la elección de las categorías utilizadas, por ejemplo en base a un análisis preliminar.
    b. Dividir el conjunto de textos disponible para utilizar una parte de los mismos como conjunto de entrenamiento y otro como conjunto test.
    c. Construir la matriz de confusión.
    d. **Calcular las medidas de evaluación Accuracy, Precisión, tasa de verdaderos positivos, tasa de falsos positivos y F1-score. Interpreta estos resultados en el contexto de las noticias clasificadas.**
    e. Calcular la curva ROC y analizarla.

MEDIDAS DE EVALUACIÓN

FORMA 1

	PRECISIÓN	Recall	f1-score
Internacional	0.68	0.39	0.5
Deportes	0.85	0.63	0.72
Salud	0.84	0.5	0.63
CyT	0.79	0.65	0.71
Nacional	0.27	0.75	0.39
Economía	0.8	0.4	0.53

MEDIDAS DE EVALUACIÓN

FORMA 1

Conclusiones:

- **Internacional**: Esta categoría tiene un bajo recall de 0.39, lo que sugiere que el modelo no es muy efectivo para identificar titulares relacionados con Internacional. Pero, la precisión de 0.68 indica que, cuando clasifica como Internacional, la mayoría son correctas.

- **Deportes**: Muestra el mejor f1-score de 0.72, con una alta precisión de 0.85 y un recall de 0.63. Esto indica que el modelo identifica bien los titulares relacionados con deportes y es bastante preciso.

- **Salud**: Aunque la precisión es alta (0.84), el recall es bajo (0.5), lo que sugiere que el modelo tiene dificultades para detectar todos los titulares de salud

- **CyT (Ciencia y Tecnología):** Esta categoría también tiene un buen desempeño con un f1-score de 0.71, un recall de 0.65, y una precisión de 0.79, lo que significa que identifica una buena cantidad de titulares de esta categoría con una precisión bastante alta.

- **Nacional**: Tiene el desempeño más bajo, con una precisión muy baja de 0.27, lo que indica que el modelo clasifica erróneamente la mayoría de los titulares como Nacional. Sin embargo, el recall es alto (0.75), lo que significa que la mayoría de los titulares de Nacional se identifican

- **Economía**: Con una precisión de 0.8 y un recall de 0.4, el modelo clasifica correctamente la mayoría de los titulares etiquetados como Economía, pero no “recupera” una gran cantidad de ellos.

MEDIDAS DE EVALUACIÓN

FORMA 2

	PRECISIÓN	Recall	f1-score
Internacional	0.62	0.47	0.53
Deportes	0.9	0.63	0.74
Salud	0.82	0.61	0.7
CyT	0.4	0.84	0.54
Nacional	0.61	0.52	0.56
Economía	0.74	0.66	0.7

MEDIDAS DE EVALUACIÓN

FORMA 2

Conclusiones:

- **Internacional**: El modelo identifica correctamente el 62% de los titulares que clasifica como Internacional, pero solo detecta el 47% de los titulares que realmente pertenecen a esta categoría.

- **Deportes**: Este es el mejor resultado en cuanto a precisión. El 90% de las predicciones positivas para Deportes son correctas, y el modelo identifica el 63% de los titulares de esta categoría.

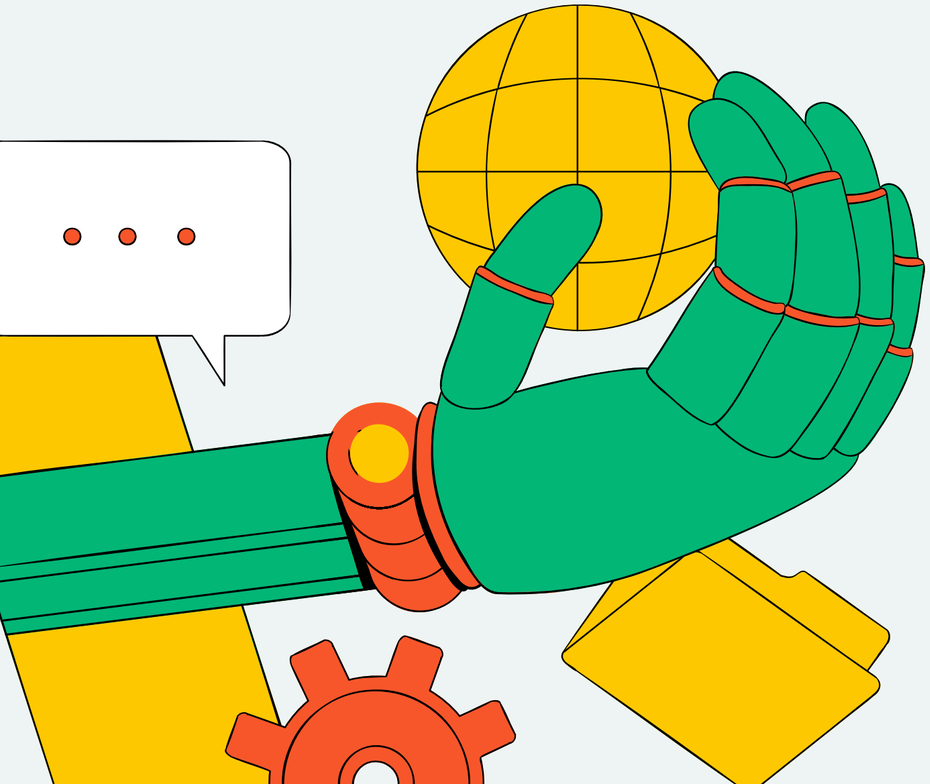
- **Salud**: El modelo es bastante preciso (82%) en la categoría de Salud y captura un 61% de los titulares correctos.

- **CyT (Ciencia y Tecnología):** Aunque el modelo captura una gran parte de los titulares de esta categoría (84% de recall), no es muy preciso, ya que solo el 40% de sus predicciones como positivas son correctas.

- **Nacional**: La precisión y el recall son moderados. El 61% de las predicciones de titulares de Nacional son correctas, y el modelo identifica el 52% de los titulares reales de esta categoría

- **Economía**: El modelo identifica correctamente el 74% de los titulares que clasifica como Economía y detecta el 66% de los titulares reales de esta categoría

EJERCICIO 2



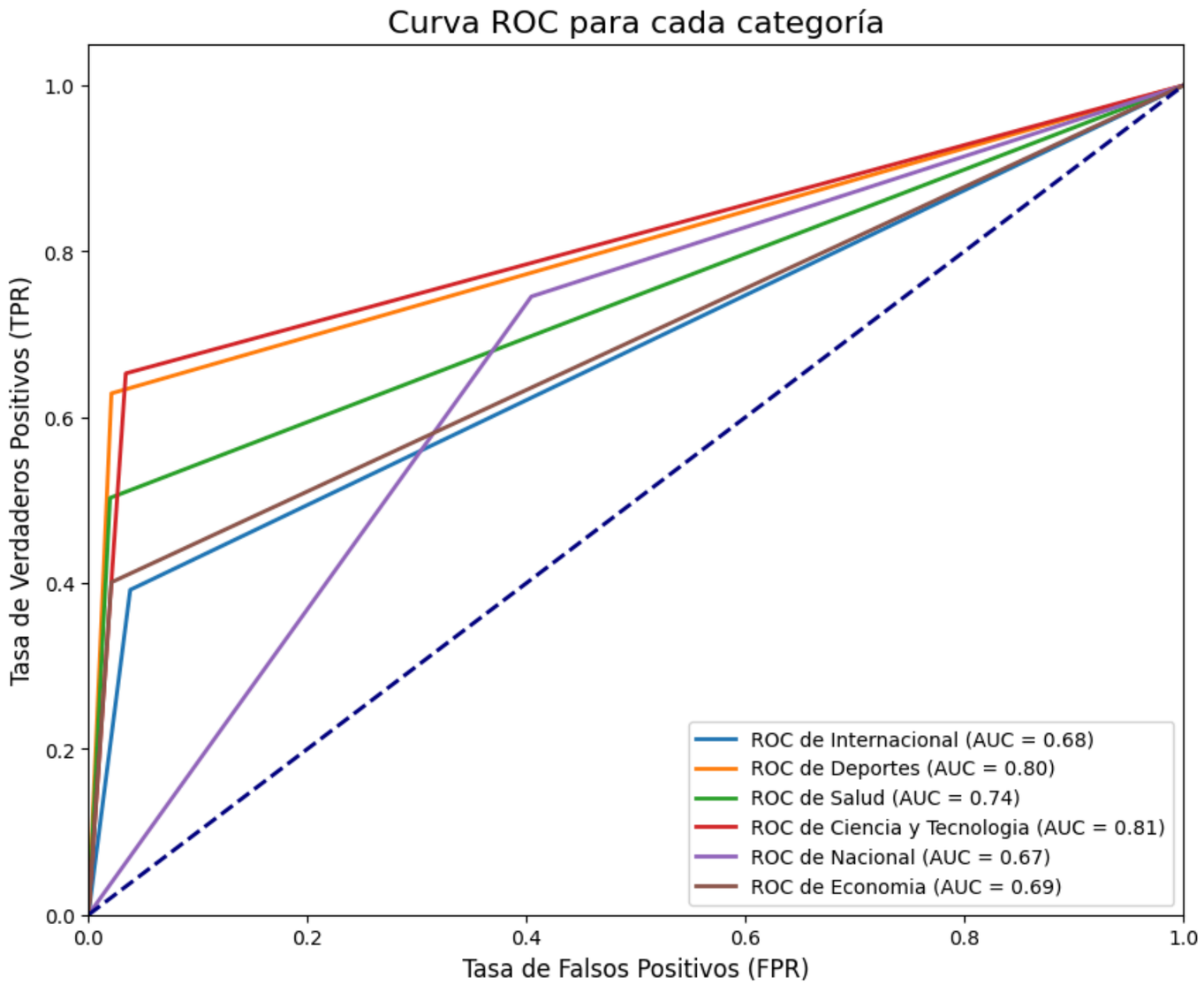
1. Implementar un clasificador de texto utilizando el clasificador ingenuo de Bayes. Utilizar el conjunto de datos ”Noticias Argentinas” para clasificar cada noticia según su tipo.

    a. Utilizar al menos 4 categorías. Justifica la elección de las categorías utilizadas, por ejemplo en base a un análisis preliminar.
    b. Dividir el conjunto de textos disponible para utilizar una parte de los mismos como conjunto de entrenamiento y otro como conjunto test.
    c. Construir la matriz de confusión.
    d. Calcular las medidas de evaluación Accuracy, Precisión, tasa de verdaderos positivos, tasa de falsos positivos y F1-score. Interpreta estos resultados en el contexto de las noticias clasificadas.
    e. **Calcular la curva ROC y analizarla.**

CURVA ROC

FORMA 1

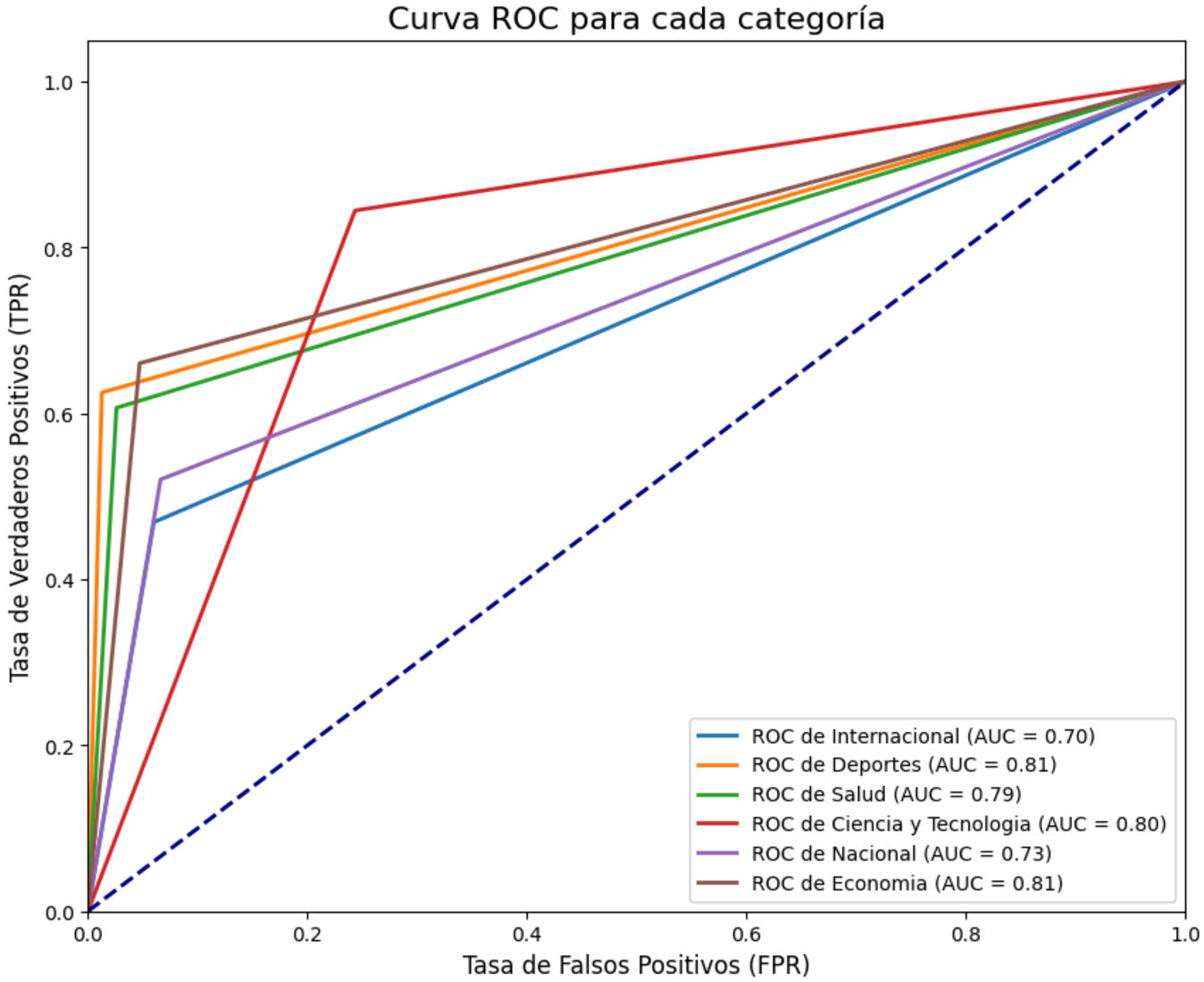
Tasa de Verdaderos Positivos (TPR): Indica cuántos ejemplos positivos reales fueron correctamente identificados como positivos.
Tasa de Falsos Positivos (FPR): Indica cuántos ejemplos negativos fueron incorrectamente clasificados como positivos.



CURVA ROC

FORMA 2

Tasa de Verdaderos Positivos (TPR): Indica cuántos ejemplos positivos reales fueron correctamente identificados como positivos.
Tasa de Falsos Positivos (FPR): Indica cuántos ejemplos negativos fueron incorrectamente clasificados como positivos.

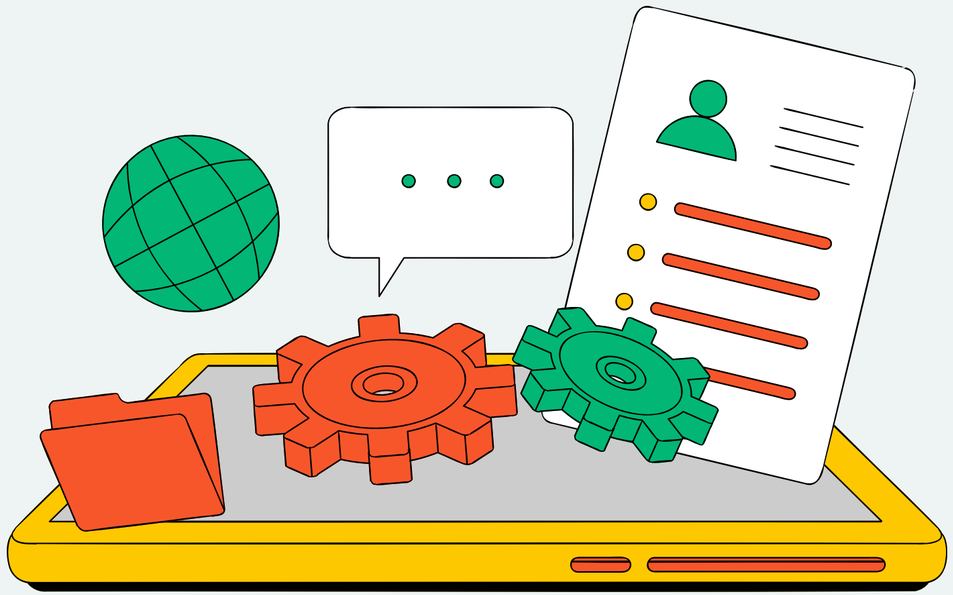


REDES BAYESIANAS

Útiles para modelar relaciones

Uso de Grafos Aciclicos Dirigidos

Uso de T. de Bayes
Uso de T. de Factorización



CAMINO RECORRIDO

1-DATOS

Se nos aporta un archivo .CSV

2-CSV

Análisis de las calificaciones de cada alumno y su admisión

3-MÉTRICAS

Tanto GRE como GPA deben discretizarse

4-PROBABILIDADES

Utilización de los datos para calcular probabilidades condicionales

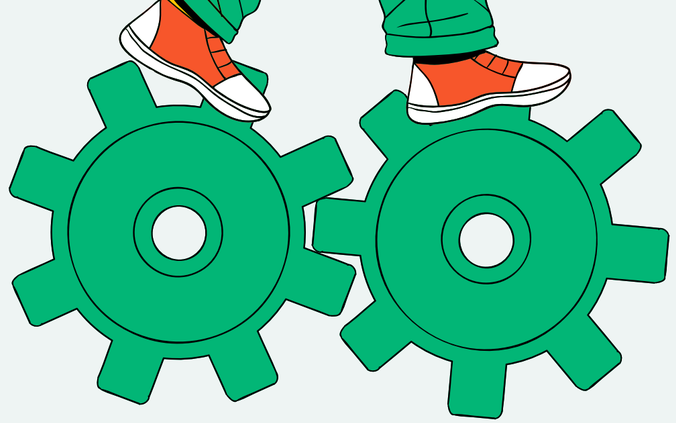
5-FACTORIZACIÓN

Hallar las probabilidades conjuntas mediante las probabilidades condicionales

6-RESULTADOS

Se obtienen las propabilidades de admisión dada la evidencia

DATOS



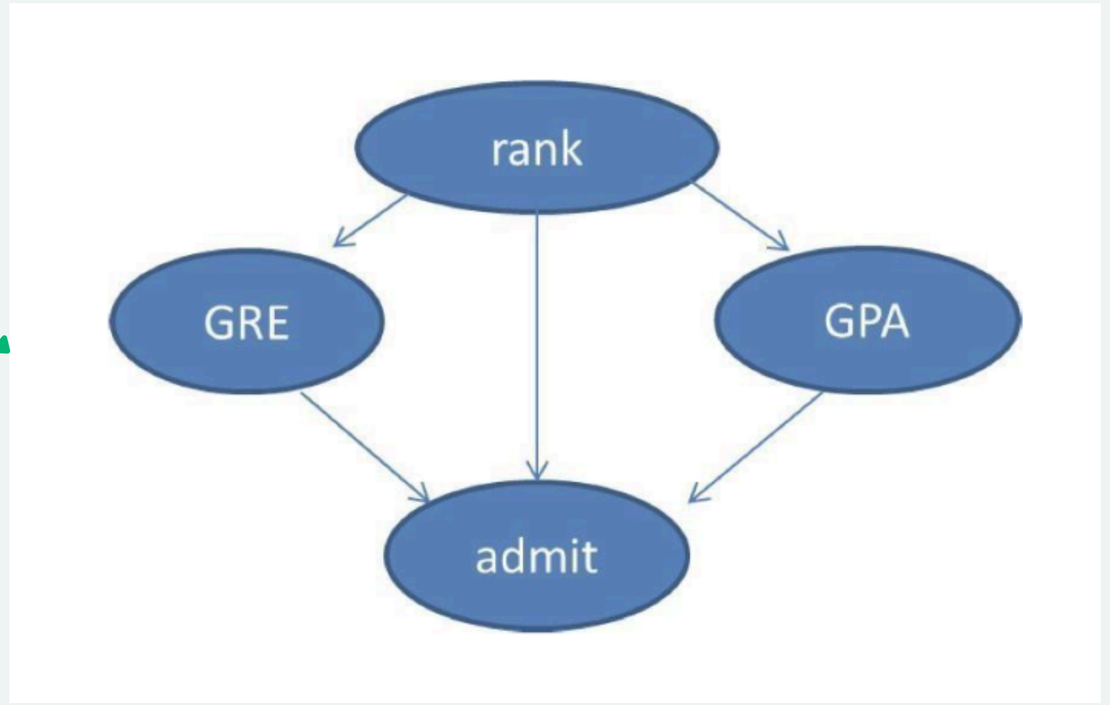
FORMATO

Se presentan los datos en una tabla con el formato:

Admit	GRE	GPA	Rank
...	...	...	...

DAG

Relaciones entre variables



DISCRETIZACIÓN

01

MOTIVO

Se discretiza para covertir a las variables GRE y GPA en dos grupos que el modelo pueda distinguir. Esto permite simplificar las cuentas.

02

FORMATO

El formato de la discretización consiste en tomar los valores posibles de las variables a discretizar y separarlas en grupos según un criterio.

03

IMPLEMENTACIÓN

Se implementa mediante el código, una condición que establece que:
High = GRE ≥ 500, Low= GRE < 500
High = GPA ≥ 3, Low= GRE < 3

RELACIÓN ENTRE VARIABLES

DAG

LA ADMISIÓN (ADMIT) DEPENDE DEL PRESTIGIO DE LA ESCUELA (RANK), EL PROMEDIO ACADÉMICO (GPA) Y EL PUNTAJE DEL EXAMEN (GRE).

TANTO EL GPA COMO EL GRE DEPENDEN DEL RANK DE LA ESCUELA.

PROBABILIDAD CONJUNTA

El calculo de la probabilidad conjunta se realiza mediante el teorema de Factorización

Se pueden calcular las probabilidades conjuntas a partir de las probabilidades condicionales. Para saber que probabilidades condicionales necesitamos, miramos los nodos del DAG y cada uno de sus padres.

CALCULO DE LAS PROBABILIDADES

Las probabilidades que debemos calcular para usar en el teorema de la factorización son:

P(rank) = es una probabilidad marginal
P(GPA | rank) = P(GPA, rank) / P(rank)
P(GRE | rank) = P(GRE, rank) / P(rank)
P(admit | rank, GPA, GRE) =P(admit, rank, GPA, GRE) / P(rank, GPA, GRE)

RESULTADOS

Caso 1: Probabilidad de que una persona que proviene de una escuela con rango 1 no haya sido admitida en la universidad.
P= 0.011946721311475 41

Caso 2: Probabilidad de que una persona que fue a una escuela de rango 2, tenga GRE = 450 y GPA = 3.5 sea admitida en la universidad.
P= 0.011037527593818984

En esencia, el proceso de aprendizaje en este caso es un **aprendizaje a partir de datos o aprendizaje estadístico.** El modelo (la red bayesiana) "aprende" las relaciones entre las variables a partir de las observaciones en los datos, representadas por las probabilidades condicionales estimadas.