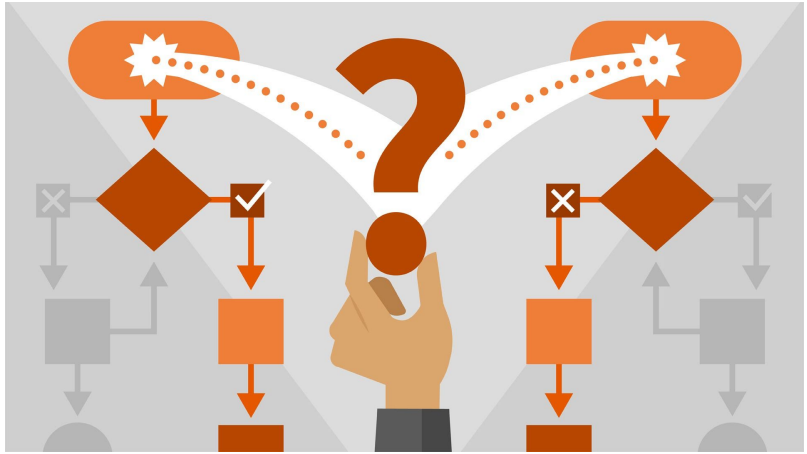
TP1: Bayes

...

Dey, Patrick
Lombardi, Matías
Vázquez, Ignacio

Tecnologías utilizadas

Ejercicio 1: Naive Bayes



Ejercicio 1

- Clasificador ingenuo de Bayes para determinar la nacionalidad de una persona
- Dataset con atributos binarios
- Dos clases: Ingles (I), Escoces (E)

Scones	Cerveza	Wiskey	Avena	Futbol	Nacionalidad
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
1	1	0	0	1	E
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

Implementación

- Parseamos el dataset y la instancia presentada
- Calculamos la probabilidad de cada clase: $P(v_j)$
- Calculamos la probabilidad de x dada la nacionalidad: $P(a_i/v_j) = \frac{x[a = a_i \wedge nacionalidad = v_j]}{\#v_j}$
- Aplicamos corrección de Laplace de ser necesario
- Calculamos la clase más probable como:

$$V_{NB} = arg\,max_{v_j \in V} \prod_{i=1}^{n} P(a_i/v_j) * P(v_j)$$

Resultados

- $x_i = (1, 0, 1, 1, 0)$
  - Probabilidad de que x_i sea inglés (I) es 23.6%
  - Algoritmo resuelve que para escoces (E) es 76.4% -> Es escoces !
- $x_i = (0, 1, 1, 0, 1)$
  - Algoritmo resuelve que para inglés (I) es 83.2% -> Es inglés !
  - Algoritmo resuelve que para escoces (E) es 16.8%

Ejercicio 2: Text classifier

Ejercicio 2

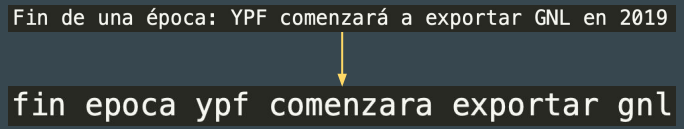
Clasificador ingenuo de Bayes para clasificación de noticias

- Analizamos cuales son las palabras más utilizadas
- Analizamos cantidad de noticias por categoría

Fecha ▲	Titular	Fuente	Categoria
1/1/19 0:00	¡Feliz 2019, Millonarios! | River Plate	La Página Millonaria	Noticias destacadas
1/1/19 0:00	¡Feliz 2019, Millonarios! | River Plate	La Página Millonaria	Noticias destacadas
1/1/19 0:00	¡Feliz 2019, Millonarios! | River Plate	La Página Millonaria	Noticias destacadas
1/1/19 0:00	¡Feliz 2019, Millonarios! | River Plate	La Página Millonaria	Noticias destacadas
1/1/19 0:00	¡Feliz 2019, Millonarios! | River Plate	La Página Millonaria	Noticias destacadas
1/1/19 0:12	Macri descongeló el salario de los funcionarios: cuánto ganará el Presidente	La Voz del Interior	Noticias destacadas
1/1/19 0:12	Macri descongeló el salario de los funcionarios: cuánto ganará el Presidente	La Voz del Interior	Noticias destacadas
1/1/19 0:12	Macri descongeló el salario de los funcionarios: cuánto ganará el Presidente	La Voz del Interior	Noticias destacadas
1/1/19 0:20	La madre del joven de Bariloche que se suicidó dijo que se retira de las redes para hacer su duelo	La Voz del Interior	Noticias destacadas
1/1/19 0:20	La madre del joven de Bariloche que se suicidó dijo que se retira de las redes para hacer su duelo	La Voz del Interior	Noticias destacadas
1/1/19 10:00	Ricky Martin y su marido anunciaron que se convirtieron en padres de una niña	Misiones Cuatro	Noticias destacadas
1/1/19 10:04	60 años de la Revolución Cubana y el legado más aciago de la historia	PanAm Post	Noticias destacadas
1/1/19 10:05	Finalmente se conoció quién era el papá del Chavo del 8, ¡muy emotivo!	La 100	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas
1/1/19 10:09	Las tres mejores funciones que WhatsApp sumó en 2018	TN	Noticias destacadas

Implementación: preprocesamiento

- Preprocesamos el dataset
  - Todo titular pasa a minúscula
  - Removemos las tildes de los titulares
  - Removemos las palabras más utilizadas, dígitos y símbolos
  - No tenemos en cuenta los titulares cuyas categorías son “Noticias destacadas” o “Destacadas”



Internacional	3850
Nacional	3860
Destacadas	3859
Deportes	3855
Salud	3840
Ciencia y Tecnologia	3856
Entretenimiento	3850
Economia	3850
Noticias destacadas	133819
Total	160789

Implementación: ¿Por qué sacamos destacadas?

- Comparativa de cantidad de palabras en común entre categorías:

```
Palabras en común con Salud
Nacional: 707
Economia: 568
Destacadas: 845
Ciencia y Tecnologia: 526
Deportes: 620
Internacional: 782
Entretenimiento: 613
```

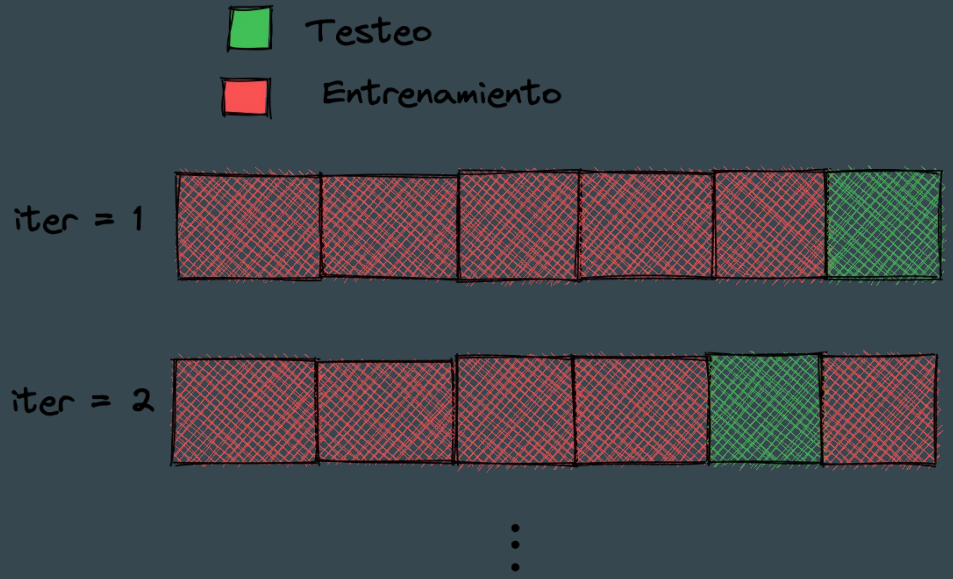
```
Palabras en común con Destacadas
Nacional: 2702
Salud: 845
Economia: 1136
Ciencia y Tecnologia: 826
Deportes: 1686
Internacional: 2293
Entretenimiento: 1473
```

Implementación: idea general para clasificar

- Tener un diccionario, para cada categoría, de todas las palabras junto con sus frecuencias
- Obtener las probabilidades utilizando únicamente las palabras presentes en un nuevo registro (el que se quiere clasificar)
- Aplicar corrección de Laplace si la palabra no está en el diccionario original
- Las probabilidades se obtienen cómo se mencionó en el primer ejercicio

Implementación: Validación Cruzada

- Shuffle del dataset previo a particionarlo
- Particiones 70/30, 80/20 y 90/10



Implementación: Validación Cruzada

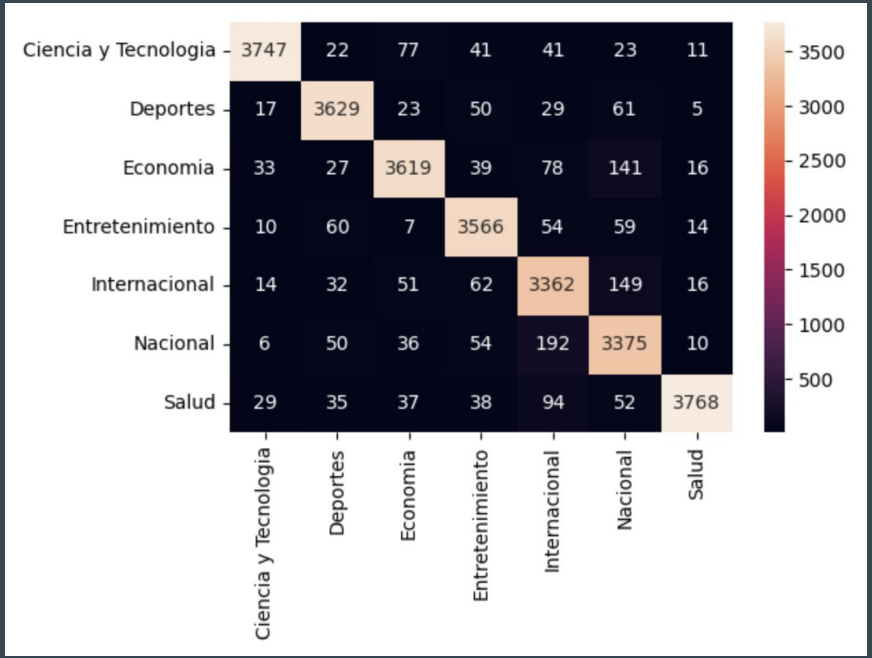
- Validación cruzada
  - Hacemos un shuffle para que no importe el orden del dataset
  - Particionamos el dataset en k partes y tomamos 1 para test y las restantes para entrenar
  - Analizamos qué tamaño deben tener estas particiones
- Por cada conjunto de particiones
  - Obtenemos la tabla de frecuencias de cada palabra por cada categoría
  - Tomamos una instancia del conjunto de testeo y le calculamos la probabilidad condicional en función de las palabras presentes en el registro a clasificar
    - Ej: “messi” tiene un 0.05 de probabilidad de aparecer en un titular de “Deportes”
  - Escribimos a un archivo las probabilidades junto con la categoría real

Implementación: Métricas

- Accuracy, Precision, TPR, FPR y F1-score:
  - Obtenemos la media de todas las las métricas por cada una de las particiones e informamos el desvío estándar
- Matriz de confusión:
  - Sumamos todas las clasificaciones hechas en todas las particiones en una sola matriz
- Curva ROC:
  - Juntamos todas las clasificaciones (vector de probabilidades) de todas las particiones
  - Calculamos las métricas requeridas variando el umbral de [0-1] con paso 0.1

Resultados 70-30

Matriz de confusión



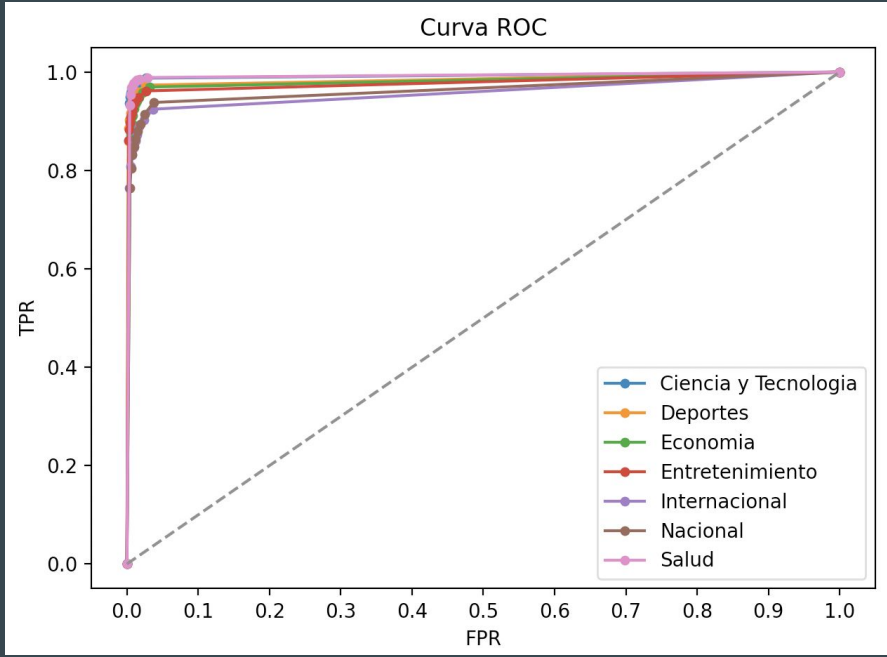
Resultados 70-30

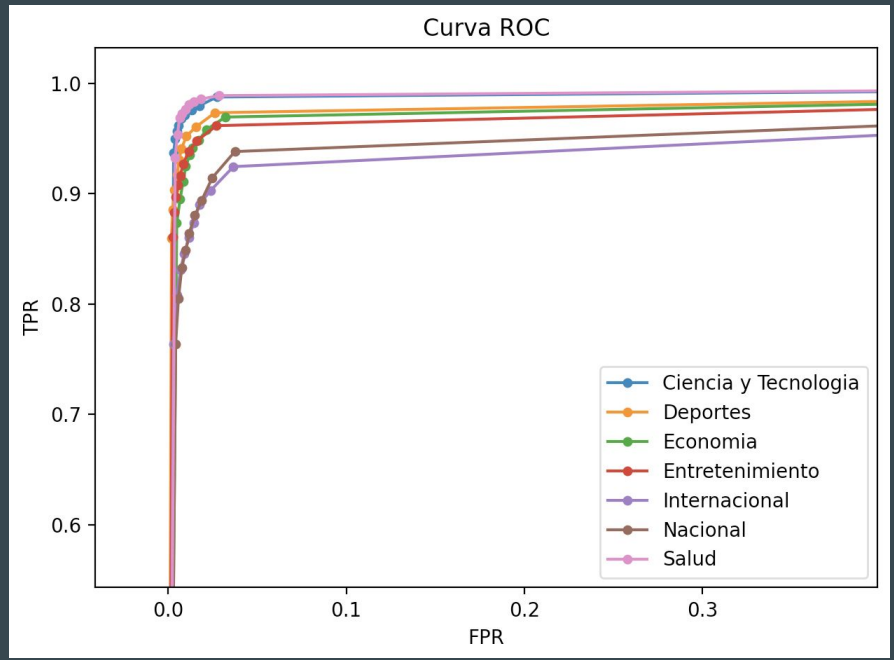
Métricas

Metric	Ciencia y Tecnologia	Deportes	Economia	Entretenimiento	Internacional	Nacional	Salud
accuracy	0.988 +- 0.008	0.985 +- 0.004	0.979 +- 0.007	0.982 +- 0.008	0.97 +- 0.007	0.969 +- 0.008	0.987 +- 0.001
precision	0.946 +- 0.002	0.952 +- 0.001	0.915 +- 0.001	0.948 +- 0.003	0.912 +- 0.008	0.907 +- 0.009	0.929 +- 0.005
fpr	0.009 +- 0.009	0.008 +- 0.007	0.015 +- 0.009	0.009 +- 0.009	0.014 +- 0.007	0.015 +- 0.009	0.012 +- 0.005
tpr	0.973 +- 0.005	0.942 +- 0.007	0.939 +- 0.004	0.927 +- 0.004	0.875 +- 0.008	0.874 +- 0.003	0.982 +- 0.002
f1	0.959 +- 0.001	0.947 +- 0.008	0.927 +- 0.003	0.937 +- 0.008	0.893 +- 0.007	0.89 +- 0.007	0.955 +- 0.009

Resultados 70-30

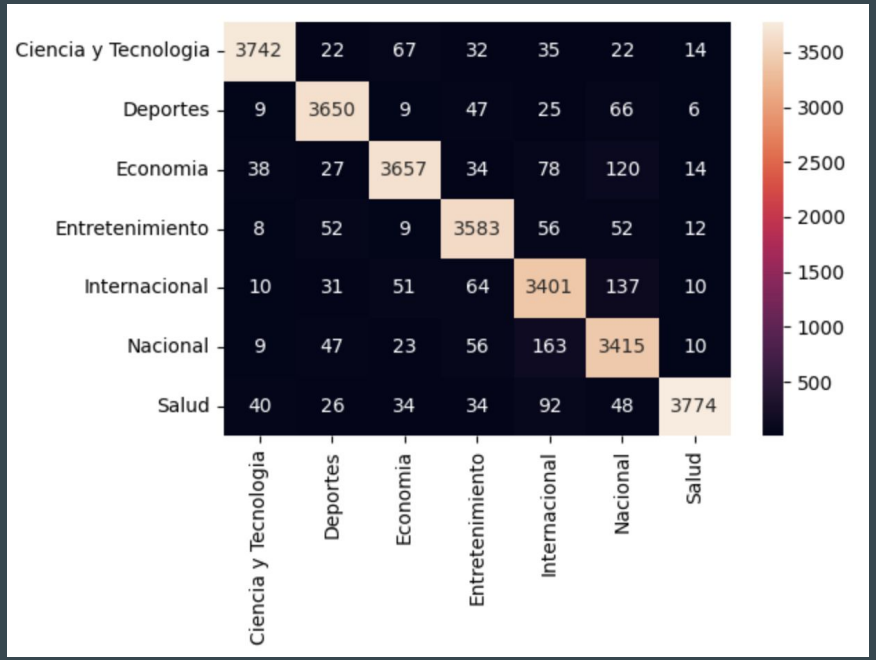
Curva Roc





Resultados 80-20

Matriz de confusión



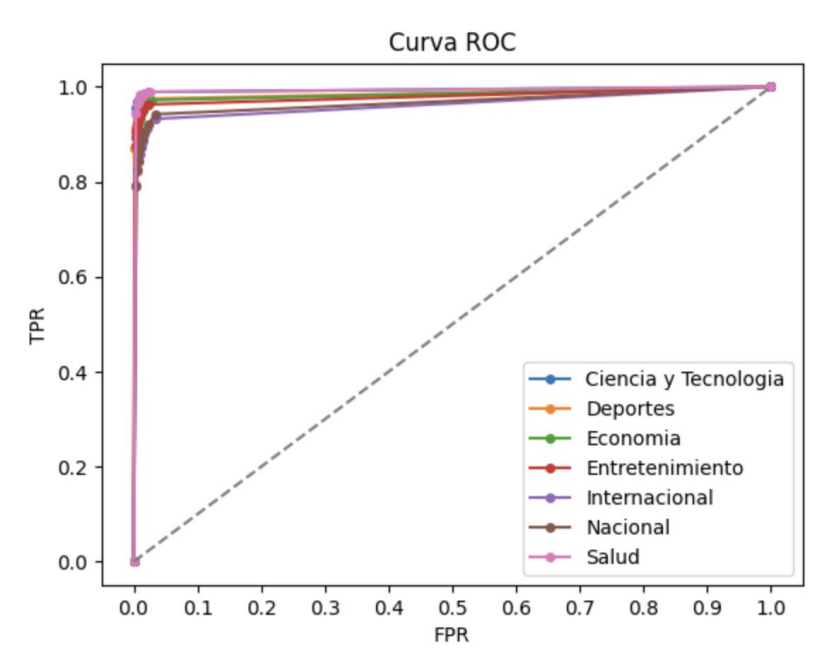
Resultados 80-20

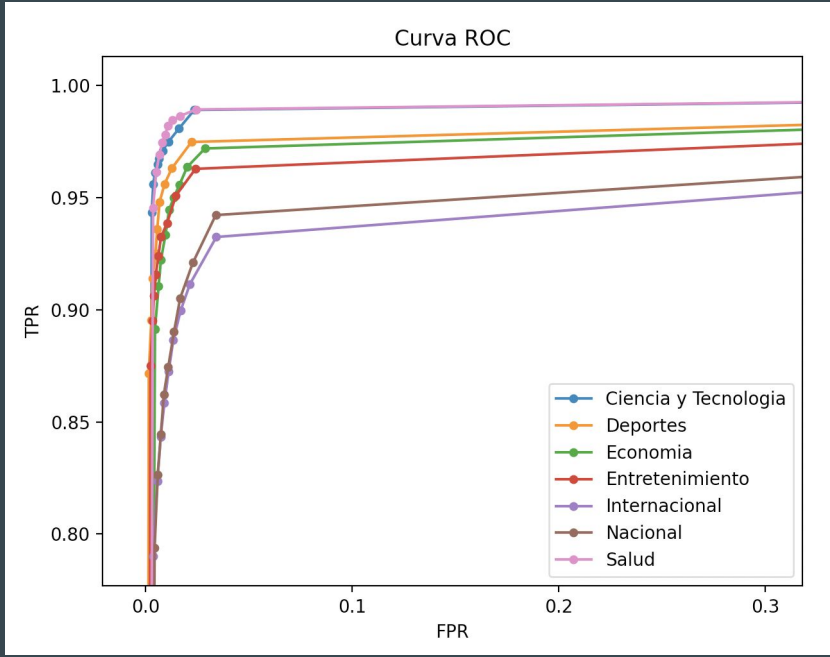
Métricas

Metric	Ciencia y Tecnologia	Deportes	Economia	Entretenimiento	Internacional	Nacional	Salud
accuracy	0.989 +- 0.003	0.986 +- 0.005	0.981 +- 0.005	0.983 +- 0.008	0.972 +- 0.005	0.972 +- 0.003	0.987 +- 0.001
precision	0.951 +- 0.009	0.958 +- 0.009	0.922 +- 0.002	0.95 +- 0.006	0.918 +- 0.006	0.917 +- 0.002	0.932 +- 0.004
fpr	0.008 +- 0.004	0.007 +- 0.004	0.013 +- 0.006	0.008 +- 0.004	0.013 +- 0.004	0.013 +- 0.007	0.012 +- 0.006
tpr	0.97 +- 0.008	0.947 +- 0.002	0.95 +- 0.007	0.931 +- 0.006	0.883 +- 0.003	0.885 +- 0.002	0.983 +- 0.003
f1	0.961 +- 0.006	0.952 +- 0.009	0.935 +- 0.005	0.94 +- 0.009	0.9 +- 0.004	0.901 +- 0.005	0.957 +- 0.002

Resultados 80-20

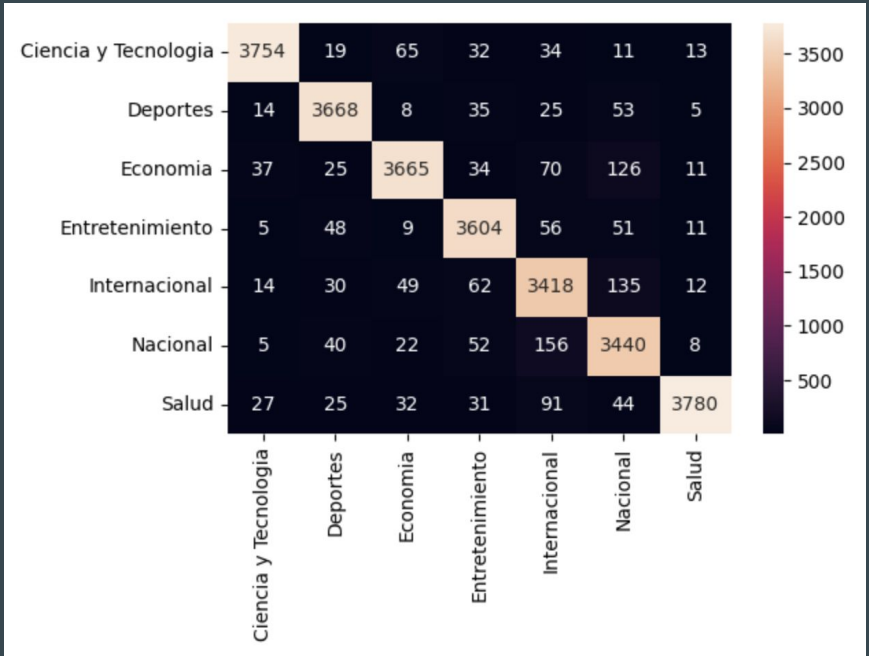
Curva Roc





Resultados 90-10

Matriz de confusión



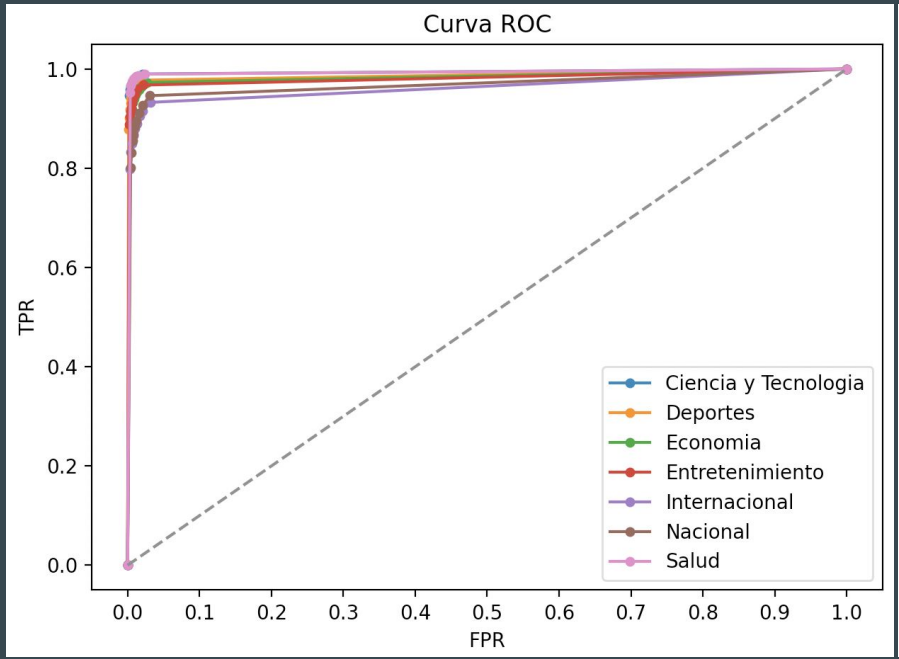
Resultados 90-10

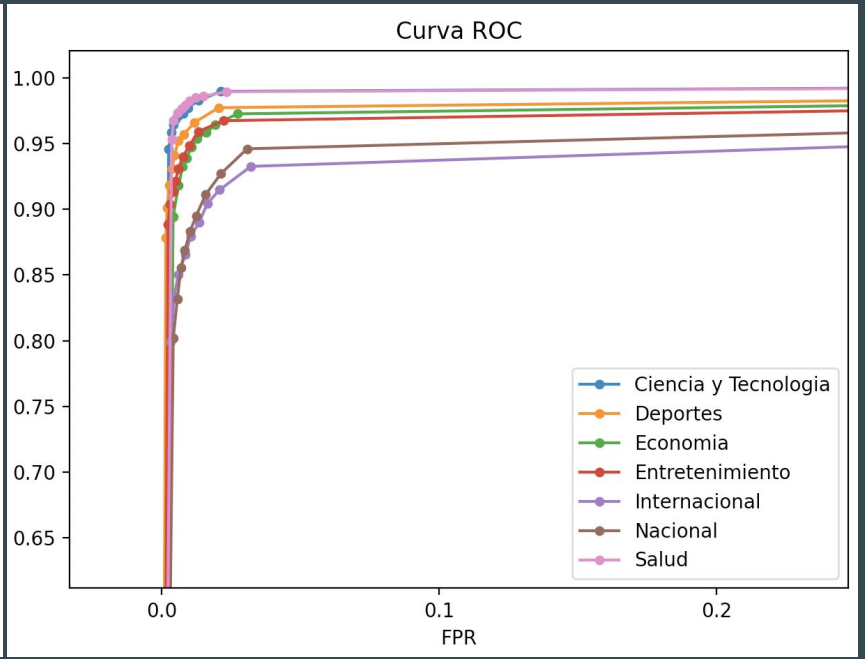
Métricas

Metric	Ciencia y Tecnologia	Deportes	Economia	Entretenimiento	Internacional	Nacional	Salud
accuracy	0.99 +- 0.004	0.988 +- 0.005	0.982 +- 0.005	0.984 +- 0.008	0.973 +- 0.008	0.974 +- 0.002	0.989 +- 0.006
precision	0.956 +- 0.008	0.963 +- 0.006	0.924 +- 0.002	0.952 +- 0.003	0.919 +- 0.003	0.924 +- 0.008	0.938 +- 0.003
fpr	0.008 +- 0.004	0.006 +- 0.009	0.013 +- 0.007	0.008 +- 0.009	0.013 +- 0.008	0.012 +- 0.003	0.011 +- 0.001
tpr	0.974 +- 0.001	0.951 +- 0.002	0.952 +- 0.005	0.936 +- 0.003	0.888 +- 0.002	0.891 +- 0.004	0.984 +- 0.002
f1	0.965 +- 0.002	0.957 +- 0.001	0.938 +- 0.003	0.944 +- 0.008	0.903 +- 0.004	0.907 +- 0.007	0.961 +- 0.009

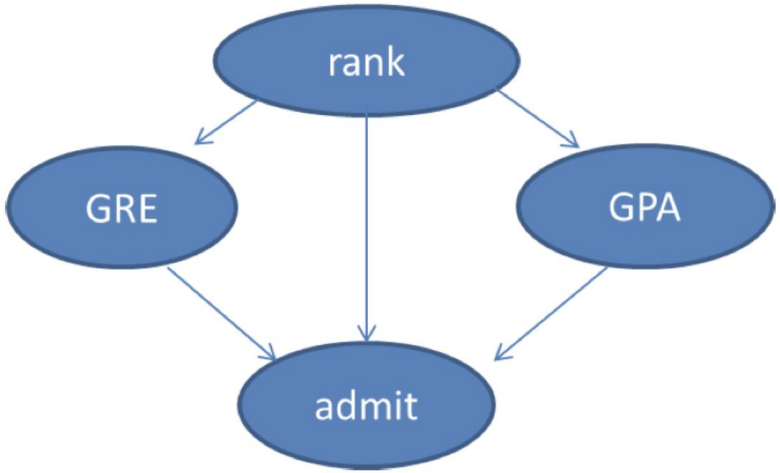
Resultados 90-10

Curva Roc



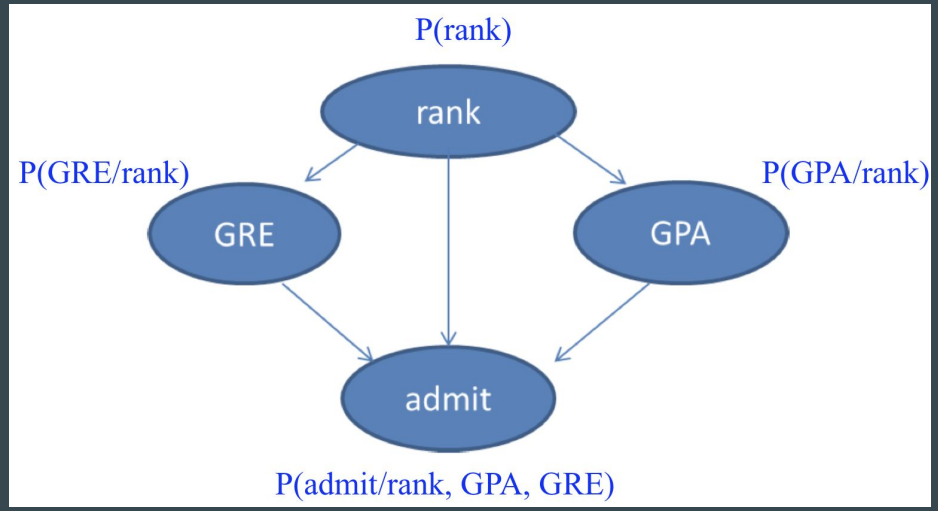


Ejercicio 3: Bayesian Network



Ejercicio 3

- Implementación de una red bayesiana para determinar la admisión de alumnos
- Dataset
  - admit ∈ {0,1}
  - GRE ∈ {GRE < 500, GRE ≥ 500}
  - GPA ∈ {GPA < 3, GPA ≥ 3}
  - rank ∈ {1,2,3,4}



Implementación

- Parseo archivo csv, categorizando columnas GPA y GRE en las categorías de la diapositiva anterior
- Si el valor del atributo es -1, quiere decir que puede tomar cualquier valor
- Cálculo de probabilidades a priori y condicionales en base a grafo dado
- Cálculo de probabilidades pedidas.

Implementación: cálculo probabilidad pedida

- Probabilidad de que una persona que proviene de una escuela con rango 1 no haya sido admitida en la universidad.

$$P(a = 0/r = 1) = \frac{\sum_{gre \in \{0,1\}} \sum_{gpa \in \{0,1\}} P(a = 0, r = 1, gre, gpa)}{P(r = 1)}$$

- Como la variable admit depende de GRE y GPA, debemos considerar todos los casos posibles. Uno de los términos del numerador sería:

$$P(a = 0/r = 1, gre = 0, gpa = 0) * P(gre = 0/r = 1) * P(gpa = 0/r = 1) * P(r = 1)$$

Resultados

a) Probabilidad de que una persona que proviene de una escuela con rango 1 no haya sido admitida en la universidad.

Resultado = 44%

b) Calcular la 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.

Resultado = 21%

Aprendizaje paramétrico

- Dados:
  - Estructura (grafo de la red)
  - Datos de entrenamiento
- Se estiman las probabilidades a priori y condicionales de cada una de las variables
- Estas probabilidades permiten clasificar nuevos registros

¡Muchas gracias!

Dey, Patrick
Lombardi, Matías
Vázquez, Ignacio