# Trabajo Práctico 1

Aprendizaje Automático

Giuliano Scaglioni (57244)

# Índice

# Ejercicio 1

## Variables

**Variables explicativas**: programas que escucha (toman valores sí/no)
**Variable a explicar**: edad del oyente (toma valores viejo/joven)

## Probabilidades

| **P(D/h) (conjunto de entrenamiento)** | | | | |
|---|---|---|---|---|
| **Programa 1** | **Programa 2** | **Programa 3** | **Programa 4** | **Edad** |
| 95% | 5% | 2% | 20% | Joven |
| 3% | 82% | 34% | 92% | Viejo |

| **P(h) (a priori)** | |
|---|---|
| **Edad** | **Probabilidad** |
| Joven | 10% |
| Viejo | 90% |

## Predicción

Se calcula la probabilidad P(h/D) para cada una de las posibles clases (viejo y joven) y se elige la que mayor probabilidad tiene, es decir, se evalúa

$$argmax_{h \in H} P(D/h)P(h)$$

Luego se normalizan los resultados a 1 para encontrar P(h/D) pues

$$P(Joven/D) + P(Viejo/D) = 1$$

| **P(h/D) (a posteriori)** | | | |
|---|---|---|---|
| **D** | **Edad** | **P(D/h)P(h)** | **P(h/D)** |
| (Si, No, Si, No) | Joven | 0,95*(1-0,05)*0,02*(1-0,2)*0,1 = 0,00144 | 91,7% |
| | Viejo | 0,03*(1-0,82)*0,34*(1-0,92)*0,9 = 0,00013 | 9,3% |

# Ejercicio 2

## Parte A

La implementación del clasificador ingenuo de Bayes se realizó en Go de manera genérica, permitiendo su uso tanto en este ejercicio como en el clasificador de texto.
El clasificador provee funciones de entrenamiento, clasificación y evaluación.
Para poder abstraer el problema a resolver del clasificador, se introdujo el tipo de dato **Example** el cual es un mapa de string a string y representa una instancia de los datos.
De esta forma, el ejemplo (1, 0, 1, 1, 0) se representaría como:

```
Example {
    "scones": "1",
    "cerveza": "0",
    "whisky": "1",
    "avena": "1",
    "futbol": "0"
}
```

Para entrenar el clasificador es necesario una lista de Examples y una lista de clases, cada una correspondiente a la clasificación del ejemplo en la misma posición. Por esto último, se requiere una conversión de los datos previamente.

## Parte B

Para clasificar el ejemplo (1, 0, 1, 1, 0)  el clasificador ingenuo de Bayes implementado en la parte A, fue entrenado con los datos provistos. Una vez entrenado, se utilizó para estimar la clase del ejemplo de la siguiente forma:

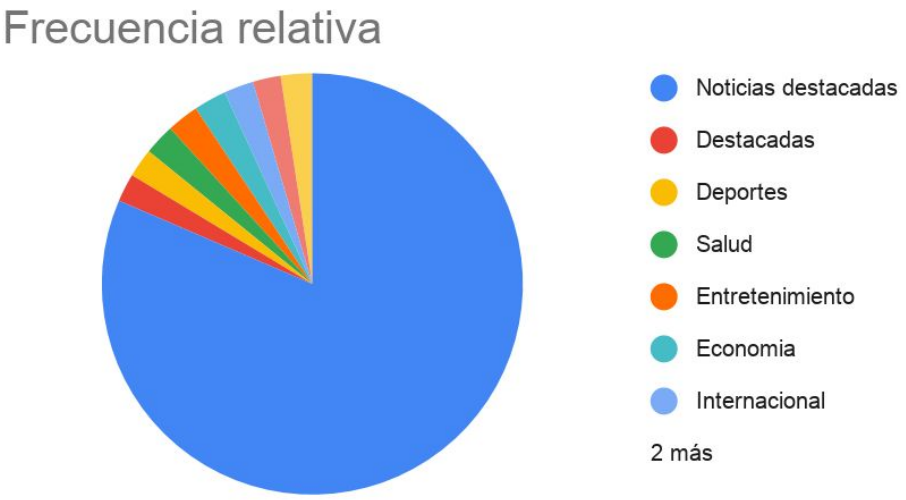
```
> ./britanicos -f datasets/britanicos.csv -p '1,0,1,1,0'
The preferences corresponds to E (0.782538)
```

Como resultado, el clasificador estimó que las preferencias corresponden a una persona **Escocesa** con una probabilidad a posteriori de **78,3%**.

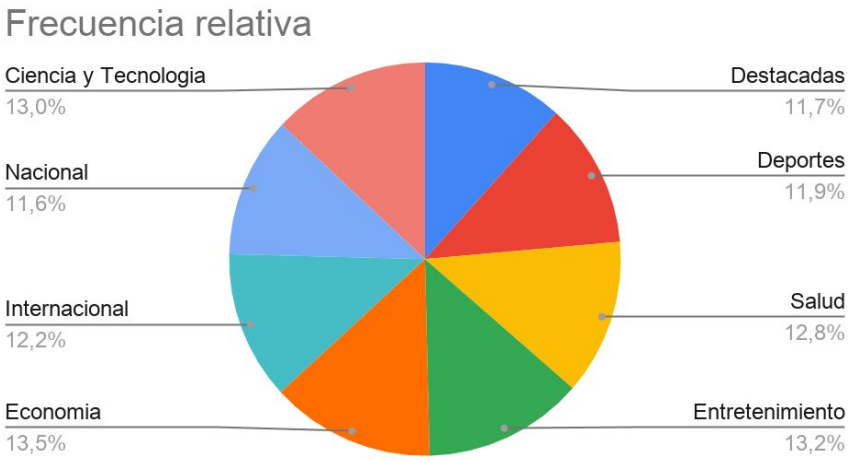
# Ejercicio 3

## Parte A

El clasificador de texto implementado permite clasificar títulos de noticias utilizando el dataset provisto por Ariel Aizemberg.
Se realizó el análisis exploratorio de los datos para tener una mejor idea de la composición de los mismos. Como resultado se obtuvo que, de los títulos, una gran parte pertenecía a la categoría “Noticias destacadas”.




Luego, para ver cuales eran las características de los títulos pertenecientes a la categoría “Noticias destacadas”, se seleccionó el 0,01% de ellas (aproximadamente 100) de manera aleatoria para su posterior análisis. De este conjunto, se observó que estos títulos no habían sido categorizados en base al tema de la noticia sino, presuntamente, a la relevancia que tenían en el momento de la noticia. Esto implica que, esos datos, para el uso que se les va a dar, no aportan nada, pues solo serían noticias sin categorizar.
Por esto último se removió completamente esta categoría, resultando en un conjunto de datos con una distribución más uniforme, como se puede observar en la siguiente figura.



Luego, para la evaluación se utilizó la validación cruzada en 10 partes, por lo que se tomó un conjunto de entrenamiento del 10% del tamaño del conjunto de datos.
Se observó, como era esperado, que las noticias clasificadas como “Destacadas” son confundidas con las pertenecientes a las categorías “Internacional” y “Nacional”.

| Promedio | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Ciencia y Tecnologia | Deportes | Destacadas | Economia | Entretenimiento | Internacional | Nacional | Salud |
| **Ciencia y Tecnologia** | 329 | 0 | 0 | 1 | 2 | 0 | 0 | 0 |
| **Deportes** | 1 | 300 | 1 | 1 | 4 | 1 | 1 | 1 |
| **Destacadas** | 1 | 23 | 171 | 5 | 7 | 45 | 56 | 1 |
| **Economia** | 3 | 1 | 2 | 334 | 1 | 2 | 1 | 1 |
| **Entretenimiento** | 1 | 1 | 1 | 0 | 331 | 1 | 1 | 0 |
| **Internacional** | 2 | 1 | 16 | 2 | 3 | 283 | 3 | 2 |
| **Nacional** | 1 | 5 | 44 | 2 | 2 | 3 | 245 | 1 |
| **Salud** | 0 | 1 | 1 | 0 | 1 | 2 | 1 | 323 |

Al igual que se hizo para la categoría “Noticias destacadas”, se observaron títulos seleccionados al azar de esta categoría y se llegó a la misma conclusión, no están clasificados en base al tema de la noticia. Por esta última razón, también se optó por excluir esa categoría del conjunto de datos.

# Parte B

Para obtener la matriz de confusión se evaluó el clasificador con una validación cruzada en 10 partes y se calculó el promedio de las 10.

| Promedio | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Ciencia y Tecnologia | Deportes | Economia | Entretenimiento | Internacional | Nacional | Salud |
| **Ciencia y Tecnologia** | 330 | 1 | 1 | 2 | 0 | 0 | 0 |
| **Deportes** | 0 | 301 | 1 | 4 | 1 | 1 | 1 |
| **Economia** | 3 | 1 | 335 | 1 | 2 | 2 | 1 |
| **Entretenimiento** | 1 | 1 | 0 | 332 | 1 | 2 | 0 |
| **Internacional** | 1 | 2 | 3 | 4 | 294 | 7 | 2 |
| **Nacional** | 0 | 5 | 2 | 2 | 3 | 288 | 1 |
| **Salud** | 0 | 1 | 0 | 1 | 2 | 1 | 324 |

Se puede observar que eliminando las categorías que sólo aportaban ruido a la clasificación, la confusión entre categorías se redujo ampliamente, pudiéndose considerar casi nula.

# Parte C

| Promedio | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Ciencia y Tecnologia | Deportes | Economia | Entretenimiento | Internacional | Nacional | Salud |
| **TP** | 330 | 301 | 335 | 333 | 294 | 288 | 324 |
| **FP** | 6 | 10 | 8 | 14 | 9 | 13 | 6 |
| **TN** | 1930 | 1950 | 1918 | 1918 | 1948 | 1955 | 1935 |
| **FN** | 5 | 9 | 9 | 5 | 18 | 14 | 5 |
| **Accuracy** | 0,9960 | 0,9920 | 0,9930 | 0,9920 | 0,9880 | 0,9880 | 0,9950 |
| **Precision** | 0,9830 | 0,9670 | 0,9780 | 0,9610 | 0,9690 | 0,9560 | 0,9820 |
| **Recall** | 0,1460 | 0,1340 | 0,1490 | 0,1480 | 0,1310 | 0,1280 | 0,1430 |
| **F1Score** | 0,2540 | 0,2350 | 0,2580 | 0,2560 | 0,2310 | 0,2260 | 0,2500 |
| **TPRate** | 0,9860 | 0,9720 | 0,9740 | 0,9850 | 0,9420 | 0,9520 | 0,9850 |
| **FPRate** | 0,0030 | 0,0050 | 0,0040 | 0,0070 | 0,0050 | 0,0060 | 0,0030 |

Para calcular las métricas de evaluación se partió de la matriz de confusión y se obtuvieron los valores para cada clase ***i*** de la siguiente forma:

- Verdadero positivo (TP): cantidad de aciertos, es decir, ejemplos de clase ***i*** clasificados como clase ***i***.
- Falso positivo (FP): cantidad de ejemplos de clase distinta a ***i*** clasificados como clase ***i***.
- Verdadero negativo (TN): cantidad de ejemplos de clase distinta a ***i*** que no fueron clasificados como clase ***i***.
- Falso negativo (FN): cantidad de ejemplos de clase ***i*** que no fueron clasificados como clase ***i***.

<table>
<tr><td></td><td>Clase 1</td><td>Clase 2</td><td>Clase 3</td><td>Clase N</td></tr>
<tr><td>Clase 1</td><td>TN</td><td>FP</td><td colspan="2">TN</td></tr>
<tr><td>Clase 2</td><td>FN</td><td>TP</td><td colspan="2">FN</td></tr>
<tr><td>Clase 3</td><td rowspan="2">TN</td><td rowspan="2">FP</td><td colspan="2" rowspan="2">TN</td></tr>
<tr><td>Clase N</td></tr>
</table>

Ejemplo del cálculo de las métricas para la clase 2

Al igual que para la matriz de confusión, se calcularon las métricas para cada una de las 10 partes de la validación cruzada y se presenta el promedio de ellas.

# Ejercicio 4

Para resolver los ejercicios se confeccionaron las siguientes tablas de probabilidades condicionales a partir del conjunto de datos provisto.

| Rank | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| *15,25%* | *37,75%* | *30,25%* | *16,75%* |

| GPA / Rank | | |
|---|---|---|
| Rank | >=3 | <3 |
| 1 | *86,89%* | *13,11%* |
| 2 | *82,78%* | *17,22%* |
| 3 | *83,47%* | *16,53%* |
| 4 | *80,60%* | *19,40%* |

| GRE / Rank | | |
|---|---|---|
| Rank | >=500 | <500 |
| 1 | *81,97%* | *18,03%* |
| 2 | *81,46%* | *18,54%* |
| 3 | *79,34%* | *20,66%* |
| 4 | *79,10%* | *20,90%* |

| Admision / Rank, GRE, GPA | | | | |
|---|---|---|---|---|
| Rank | GRE | GPA | NO | SI |
| 1 | >=500 | >=3 | *44,68%* | *55,32%* |
| 1 | >=500 | <3 | *0,00%* | *100,00%* |
| 1 | <500 | >=3 | *50,00%* | *50,00%* |
| 1 | <500 | <3 | *80,00%* | *20,00%* |
| 2 | >=500 | >=3 | *57,69%* | *42,31%* |
| 2 | >=500 | <3 | *84,21%* | *15,79%* |
| 2 | <500 | >=3 | *80,95%* | *19,05%* |
| 2 | <500 | <3 | *57,14%* | *42,86%* |
| 3 | >=500 | >=3 | *75,29%* | *24,71%* |
| 3 | >=500 | <3 | *63,64%* | *36,36%* |
| 3 | <500 | >=3 | *81,25%* | *18,75%* |
| 3 | <500 | <3 | *100,00%* | *0,00%* |

## Parte A

La probabilidad de que una persona que proviene de una escuela de rango 1 no haya sido admitida en la universidad es

$$P(admit = 0/rank = 1) = \frac{P(admit = 0, rank = 1)}{P(rank = 1)}$$

Por otro lado,

$$P(admit = 0, rank = 1) = \sum_{GRE \in \{\geq 500, <500\}} \left( \sum_{GPA \in \{\geq 3, <3\}} P(admit = 1, rank = 1, GRE, GPA) \right)$$

Donde cada sumando, por teorema de la factorización de la probabilidad sería

$$P(rank = 1) \cdot P(GRE/rank = 1) \cdot P(GPA/rank = 1) \cdot P(admit = 0/rank = 1, GPE, GPA)$$

Utilizando los datos de las tablas, se puede calcular la probabilidad de la siguiente manera

| | | |
|---|---|---|
| P(admit=0, rank=1, gre=1, gpa=1) | P(rank=1)*P(gre=1/rank=1)*P(gpa=1/rank=1)*P(admit=0/rank=1, gre=1, gpa=1) | 4,85% |
| P(admit=0, rank=1, gre=1, gpa=0) | P(rank=1)*P(gre=1/rank=1)*P(gpa=0/rank=1)*P(admit=0/rank=1, gre=1, gpa=0) | 0,00% |
| P(admit=0, rank=1, gre=0, gpa=1) | P(rank=1)*P(gre=0/rank=1)*P(gpa=1/rank=1)*P(admit=0/rank=1, gre=0, gpa=1) | 1,19% |
| P(admit=0, rank=1, gre=0, gpa=0) | P(rank=1)*P(gre=0/rank=1)*P(gpa=0/rank=1)*P(admit=0/rank=1, gre=0, gpa=0) | 0,29% |
| P(admit=0/rank=1) | P(admit=0/rank=1)/P(rank=1) | 41,55% |

***Nota:*** *gre=1 significa gre>=500, al igual que gpa=1 significa gpa>=3*

# Parte B

La probabilidad de que una persona que fue a una escuela de rango 2, que tiene GRE = 450 y GPA = 3.5 sea admitida en la universidad es

$$P(admit = 1/rank = 2, GRE < 500, GPA \geq 3)$$

con lo cual se puede obtener directamente de la tabla **Admisión/Rank, GRE, GPA** siendo el resultado 19,05%.

# Parte C

En este ejercicio el proceso de aprendizaje fue solo de tipo paramétrico ya que la estructura (red) ya estaba dada.
Este consistió en obtener las frecuencias relativas del nodo raíz (rank) contando las ocurrencias de cada rango de escuelas en el conjunto de datos. Luego se realizó lo mismo con los demás nodos que tienen padre (GPA, GRE y admit) pero obteniendo las probabilidades condicionales dados sus padres. Con estos datos se confeccionaron las tablas las cuales son el conocimiento adquirido.