PAC2 (Otoño 2025)

En esta actividad no está permitido el uso de herramientas de inteligencia artificial. En el plan docente y en la web sobre integridad académica y plagio de la UOC encontraréis información sobre qué se considera conducta irregular en la evaluación y las consecuencias que puede tener.

Introducción

Consideramos la tabla de datos `jobs_in_data.csv` que proporciona información sobre salarios y trabajos en el ámbito de la Ciencia de Datos.

Contiene, entre otros, los siguientes campos:

`work_year`: el año en que se registraron los datos. Este campo indica el contexto temporal de los datos, importante para comprender las tendencias salariales a lo largo del tiempo.

`job_title`: el título específico del puesto de trabajo, como "Científico de datos", "Ingeniero de datos" o "Analista de datos", entre otros. Esta columna es crucial para comprender la distribución salarial entre diversos roles especializados dentro del campo de datos.

`job_category`: clasificación del puesto de trabajo en categorías más amplias para facilitar el análisis. Esto podría incluir áreas como "Análisis de datos", "Aprendizaje automático", "Ingeniería de datos", etc.

`salary`: el salario bruto anual convertido a dólares estadounidenses (USD). Esta conversión uniforme ayuda en las comparaciones y análisis salariales globales.

`Employee_residence`: el país de residencia del empleado. Este dato se puede utilizar para explorar las diferencias salariales geográficas y las variaciones del coste de vida.

`experience_level`: clasifica el nivel de experiencia profesional del empleado. Las categorías comunes pueden incluir "Nivel de entrada", "Nivel medio", "Senior" y "Ejecutivo", lo que proporciona información sobre cómo la experiencia influye en el salario en roles relacionados con datos.

`employment_type`: especifica el tipo de empleo, como "tiempo completo", "tiempo parcial", "contrato", etc. Esto ayuda a analizar cómo los diferentes acuerdos laborales afectan las estructuras salariales.

`work_setting`: el entorno de trabajo, como "Remoto", "Presencial" o "Híbrido". Esta columna refleja el impacto de los entornos laborales en los niveles salariales en la industria de datos.

`company_location`: el país donde está ubicada la empresa. Ayuda a analizar cómo la ubicación de la empresa afecta las estructuras salariales.

`company_size`: el tamaño de la empresa empleadora, a menudo categorizada en tamaño pequeño (S), mediano (M) y grande (L). Esto permite analizar cómo el tamaño de la empresa influye en el salario.

Para importar los datos podemos usar la siguiente instrucción:

```
dades <- read.csv("jobs_in_data.csv")
n_total <- nrow(dades)  # Número total de registros
```

Es necesario entregar la práctica en forma de archivo PDF (exportando el resultado final a PDF, por ejemplo) **únicamente** en esta misma tarea.

Indicar las fórmulas usadas del tipo $P(A|B), P(A \cap B)$, etc.

Os puede ser útil la función `table` para tabular los datos.

Consultad las actividades resueltas de probabilidad del reto 2.

NOMBRE: José Carlos López Henestrosa

PAC2

Una vez importados los datos:

Problema 1 (30 puntos)

a) Calculad la probabilidad de que un trabajador trabaje en modalidad híbrida (`work_setting == "Hybrid"`). *(15 puntos)*

```
# 1) Calcular P(Hybrid)
n_hybrid <- sum(dades$work_setting == "Hybrid")
prob <- n_hybrid / n_total

# 2) Mostrar resultado formateado como porcentaje
sprintf("%0.4f%%", prob * 100)
```

```
## [1] "2.0417%"
```

b) Calculad la probabilidad de que un trabajador en modalidad híbrida tenga un nivel de experiencia Medio (`experience_level == "Mid-level"`). *(15 puntos)*

```
# 1) Calcular P(Mid-level | Hybrid)
n_mid_level_hybrid <- with(
  dades,
  sum(experience_level == "Mid-level" & work_setting == "Hybrid")
)
prob <- n_mid_level_hybrid / n_hybrid

# 2) Mostrar resultado formateado como porcentaje
sprintf("%0.4f%%", prob * 100)
```

```
## [1] "36.1257%"
```

Problema 2 (70 puntos)

a) Construid la tabla de contingencia número de trabajadores por año (`work_year`) y tipo de empleo (`employment_type`). *(15 puntos)*

```
contingency_table <- with(dades, table(work_year, employment_type))
contingency_table
```

```
##           employment_type
## work_year Contract Freelance Full-time Part-time
##      2020        3         1        65         2
##      2021        3         3       187         4
##      2022        4         3      1621         6
##      2023        9         4      7437         3
```

b) Calculad la probabilidad de que un trabajador sea del año 2023 y de tipo 'Full-time'. *(10 puntos)*

```
# 1) Calcular P(2023 ∩ Full-time)
n_2023_full_time <- with(
  dades,
  sum(work_year == 2023 & employment_type == "Full-time")
)
prob <- n_2023_full_time / n_total

# 2) Mostrar resultado formateado como porcentaje
sprintf("%0.4f%%", prob * 100)
```

```
## [1] "79.4976%"
```

c) Calculad la probabilidad de que un trabajador `Freelance` esté registrado en el año 2023. *(15 puntos)*

```
# 1) Calcular P(2023 | Freelance)
dt_freelance <- dades[dades$employment_type == "Freelance", , drop = FALSE]
n_freelance <- nrow(dt_freelance)
n_2023 <- sum(dt_freelance$work_year == 2023)
prob <- n_2023 / n_freelance

# 2) Mostrar resultado formateado como porcentaje
sprintf("%0.4f%%", prob * 100)
```

```
## [1] "36.3636%"
```

d) Calculad la probabilidad de que un trabajador registrado en 2022 tenga tipo de empleo Contract. *(15 puntos)*

```
# 1) Calcular P(Contract | 2022)
dt_2022 <- dades[dades$work_year == 2022, , drop = FALSE]
n_2022 <- nrow(dt_2022)
n_contract <- sum(dt_2022$employment_type == "Contract")
prob <- n_contract / n_2022

# 2) Mostrar resultado formateado como porcentaje
sprintf("%0.4f%%", prob * 100)
```

```
## [1] "0.2448%"
```

e) ¿Son independientes los eventos *A*: "ser `Full-time`" y *B*: "estar registrado en el año 2023"? *(15 puntos)*

Dos sucesos son independientes si la probabilidad de que ocurra uno de ellos no se ve afectada por la posibilidad de que ocurra el otro. Una de las condiciones que cumplen los sucesos independientes es que la probabilidad de que ocurra uno no se ve modificada por el hecho de que ocurra el otro. Esto se refleja en la siguiente expresión:

$$P(A|B) = P(A)$$

Dicho esto, procedemos a comprobar la independencia de ambos eventos en R:

```
# 1) Inicializar variables
A <- dades$employment_type == "Full-time"
B <- dades$work_year == 2023

n_AB <- sum(A & B)
n_B <- sum(B)

# 2) Calcular P[A (Full-time) | B (2023)]
prob_AB = n_AB / n_B

# 3) Calcular P[A (Full-time)]
prob_A <- mean(A)

# 4) Mostrar resultado
sprintf("P(A|B) = %.4f  ≠  P(A) = %.4f", prob_AB, prob_A)
```

```
## [1] "P(A|B) = 0.9979  ≠  P(A) = 0.9952"
```

Como podemos comprobar, $P(A|B) \neq P(A)$, por lo que los eventos A y B **no son independientes**.