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 usades 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:

PAC2

Una vez importados los datos:

Problema 1 (30 puntos)

Trabajaremos con el entorno de trabajo (`work_setting`) y el nivel de experiencia (`experience_level`).

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

Solución (1a)

```
# Contamos cuántos registros son Hybrid y calculamos la proporción sobre el total
n_hybrid <- sum(dades$work_setting == "Hybrid", na.rm = TRUE)
p_hybrid <- n_hybrid / n_total

cat("Número total de registros:", n_total, "")
```

```
## Número total de registros: 9355
```

```
cat("Trabajadores en modalidad 'Hybrid':", n_hybrid, "")
```

```
## Trabajadores en modalidad 'Hybrid': 191
```

```r
cat(sprintf("Probabilidad P(Hybrid) = %0.6f", p_hybrid))
```

```
## Probabilidad P(Hybrid) = 0.020417
```

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

Solución (1b)

```r
# Subconjunto de empleados que trabajan en Hybrid
sub_hybrid <- subset(dades, work_setting == "Hybrid")
n_hybrid_tot <- nrow(sub_hybrid)

n_mid_in_hybrid <- sum(sub_hybrid$experience_level == "Mid-level",
                       na.rm = TRUE)
p_mid_given_hybrid <- n_mid_in_hybrid / n_hybrid_tot

cat("Trabajadores 'Hybrid':", n_hybrid_tot, "")
```

```
## Trabajadores 'Hybrid': 191
```

```r
cat("De estos, 'Mid-level':", n_mid_in_hybrid, "")
```

```
## De estos, 'Mid-level': 69
```

```r
cat(sprintf("Probabilidad P(Mid-level | Hybrid) = %0.6f", p_mid_given_hybrid))
```

```
## Probabilidad P(Mid-level | Hybrid) = 0.361257
```

Problema 2 (70 puntos)

Ahora consideraremos la relación entre **año de registro** (`work_year`) y **tipo de empleo** (`employment_type`).

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

Solución (2a)

```r
# Tabla de frecuencias absolutas
Taula_WE <- table(dades$work_year, dades$employment_type)
Taula_WE
```

```
##
##        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
```

```
# Para comodidad, también mostramos la tabla de proporciones sobre el total
round(prop.table(Taula_WE), 6)
```

```
##
##        Contract Freelance Full-time Part-time
##   2020 0.000321  0.000107  0.006948  0.000214
##   2021 0.000321  0.000321  0.019989  0.000428
##   2022 0.000428  0.000321  0.173276  0.000641
##   2023 0.000962  0.000428  0.794976  0.000321
```

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

Solución (2b)

```
num_joint <- sum(dades$work_year == 2023 & dades$employment_type ==
                   "Full-time", na.rm = TRUE)
p_joint <- num_joint / n_total

cat("Número de casos Full-time", num_joint, "")
```

```
## Número de casos Full-time 7437
```

```
cat(sprintf("Probabilidad conjunta = %0.6f", p_joint))
```

```
## Probabilidad conjunta = 0.794976
```

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

Solución (2c)

```
sub_freelance <- subset(dades, employment_type == "Freelance")

p_2023_given_freelance <- sum(sub_freelance$work_year == 2023, na.rm = TRUE)/
    nrow(sub_freelance)
cat("Total Freelance:", nrow(sub_freelance), "")
```

```
## Total Freelance: 11
```

```
cat(sprintf("P(2023 | Freelance) = %0.6f", p_2023_given_freelance))
```

```
## P(2023 | Freelance) = 0.363636
```

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

Solución (2d)

```r
sub_2022 <- subset(dades, work_year == 2022)

p_contract_given_2022 <- sum(sub_2022$employment_type == "Contract",
                             na.rm = TRUE) / nrow(sub_2022)
cat("Total registros 2022:", nrow(sub_2022), "")
```

```
## Total registros 2022: 1634
```

```r
cat(sprintf("P(Contract | 2022) = %0.6f", p_contract_given_2022))
```

```
## P(Contract | 2022) = 0.002448
```

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

Solución (2e)

```r
# Probabilidades marginales y conjunta
p_A  <- mean(dades$employment_type == "Full-time")
p_B  <- mean(dades$work_year == 2023)
p_AB <- mean(dades$employment_type == "Full-time" & dades$work_year == 2023)

cat(sprintf("P(A = Full-time) = %0.6f", p_A))
```

```
## P(A = Full-time) = 0.995190
```

```r
cat(sprintf("P(B = año 2023) = %0.6f", p_B))
```

```
## P(B = año 2023) = 0.796686
```

```r
cat(sprintf("P(A intersección B) = %0.6f", p_AB))
```

```
## P(A intersección B) = 0.794976
```

```r
# Criterio de independencia (con tolerancia numérica)
prod_AB <- p_A * p_B
diff <- abs(p_AB - prod_AB)
cat(sprintf("diferencia = %0.8f", diff))
```

```
## diferencia = 0.00212195
```

```r
if (is.na(p_A) || is.na(p_B) || is.na(p_AB)) {
  cat("No se puede evaluar la independencia por valores no disponibles.
")
} else if (diff < 1e-8) {
  cat("Conclusión: con igualdad numérica, A y B ",
      "se pueden considerar independientes.
", sep = "")
} else {
```

```
  cat("Conclusión: como P(A intersección B) distinto de P(A) por P(B), ",
      "\n**no** son independientes.", sep = "")
}
```

```
## Conclusión: como P(A intersección B) distinto de P(A) por P(B),
## **no** son independientes.
```