Actividad 3: Modelización predictiva

Enunciado

Análisis Estadístico - Semestre 2025.1

Índice

Introducción

En esta actividad utilizaremos un conjunto de datos reales procedentes de una muestra de la **Encuesta Cuatrienal de Estructura Salarial 2022**.

- **Fuente de los datos:** Instituto Nacional de Estadística (INE).[1]
- **Disponibilidad:** Datos de acceso libre, difundidos bajo la licencia de uso de la información del INE, que permite su descarga y reutilización con fines no comerciales.

Este dataset nos proporciona una radiografía detallada de los salarios en España y contiene las variables clave para analizar los determinantes del salario y la brecha de género. Nota conceptual: aunque en este documento hablamos de "brecha de género", el fichero de datos solo dispone de la variable `SEXO`, basada en el sexo registrado (hombre/mujer). En consecuencia, nuestro análisis se limita a comparar salarios entre hombres y mujeres, y no recoge la diversidad completa de identidades y expresiones de género.

El objetivo principal de este estudio es doble: 1. Averiguar cuáles son los factores que más influyen a la hora de determinar el **salario bruto anual** de una persona trabajadora (Regresión Lineal). 2. Entender qué factores aumentan la **probabilidad** de tener un "salario alto" (Regresión Logística).

Este tipo de análisis es especialmente relevante para el estudio de las desigualdades salariales y puede ser útil para informar la toma de decisiones en políticas públicas y gestión de personas. En particular, se relaciona con varios Objetivos de Desarrollo Sostenible (ODS):

- **ODS 5 (Igualdad de Género):** La brecha salarial entre hombres y mujeres es uno de los focos principales de la actividad.
- **ODS 10 (Reducción de las Desigualdades):** También se analizan diferencias salariales asociadas al origen (nacionalidad).
- **ODS 8 (Trabajo Decente):** Se trabaja con factores estructurales que influyen en el salario, como la jornada o el sector, aunque no se abordan todos los aspectos de este ODS.

Variables Clave del Estudio (Selección):

- RETRINOIN: Salario bruto anual excluyendo importes derivados de incapacidades temporales (Numérica)
- SEXO: Variable recogida por el INE basada en el sexo registrado (no el género). En esta muestra tiene dos categorías: 1 = Hombre y 6 = Mujer. Esta variable no recoge identidades no binarias ni otras expresiones de género. En este estudio se utiliza como una *proxy* limitada del género, únicamente para analizar la diferencia salarial entre hombres y mujeres. Cuando hablemos de "brecha de género", nos referiremos estrictamente a esta diferencia, reconociendo las limitaciones conceptuales del dataset.
- TIPOPAIS: Nacionalidad administrativa (Categórica: 1=Española, 2=Extranjera)
- ANOANTI: Años de antigüedad (Numérica)
- JAP: Jornada Anual Pactada (número de horas anuales) (Numérica).
- ANOS2: Grupos de Edad (Categórica)
- ESTU: Nivel de estudios (Categórica)
- TIPOJOR: Tipo de jornada (Categórica: 1=Tiempo completo, 2=Tiempo parcial)
- TIPOCON: Duración del contrato (Categórica: 1=Indefinido, 2=Determinado)
- CNACE: Sector de actividad (Categórica)
- CNO1: Categoría ocupacional (Categórica)
- ESTRATO2: Tamaño de la empresa (Categórica)

[1] https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177025&menu=resultados&idp=1254735976596#tabs-1254736195110

- CONTROL: Propiedad o control (Categórica: 1=Público, 2=Privado)

En toda la actividad utilizaremos un nivel de confianza del 95 %. La muestra proporcionada no contiene valores perdidos (NAs), por lo que no es necesario aplicar procedimientos de imputación ni eliminar observaciones antes del análisis.

A tener en cuenta para realizar la actividad:

- Es necesario entregar el archivo `.Rmd` y el archivo de salida (`PDF` o `html`). El archivo de salida debe incluir: el enunciado de cada apartado, el código R y el resultado de la ejecución (paso a paso).
- Para facilitar la corrección, los archivos deben entregarse por separado en el aula virtual. Es decir, hay que subir cada archivo por separado. Verificad que estén colgados correctamente.
- Se debe respetar la misma numeración de los apartados que en el enunciado.
- No se pueden incluir listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificultaría la revisión. Para verificar el código, usad funciones como `head` y `tail`, que muestran pocas líneas.
- Se valorará la precisión de los términos utilizados y el uso correcto de la terminología estadística.
- Es necesario explicar el resultado y argumentar la respuesta obtenida de forma clara.
- Se valorará la concisión. No se trata de realizar explicaciones muy largas o documentos excesivamente extensos.
- El código presentado debe ser de **elaboración propia**. No está permitido el uso de herramientas de inteligencia artificial generativa; su uso se considera una conducta irregular.
- Se aceptarán entregas fuera de plazo con una penalización de 1 punto por día de retraso. No se aceptarán tras la publicación de la solución.
- Para dudas técnicas sobre R, tenéis a vuestra disposición el aula Laboratorio en Python y R.

---

1. Regresión Lineal

En esta primera parte, construiremos un modelo para predecir el salario bruto anual (`RETRINOIN`).

1.1. Carga y preparación de los datos

Cargad el fichero de datos `salarios_INE_2022_sample.csv` y revisad su estructura. Identificad las variables categóricas codificadas (como `SEXO` o `ESTU`) y convertidlas a factor, incluidas `CNACE` y `CNO1`, después de recodificarlas en los tres niveles indicados a continuación.

Asignad etiquetas descriptivas a los niveles de los factores para facilitar la interpretación posterior (ej. 1 corresponde a "Hombre" y 6 a "Mujer"). Podéis consultar el diccionario de variables anexo (`dr_EES_2022.xlsx`) para explorar mejor los códigos y descripciones.

Para las variables `CNACE` y `CNO1`, consultad las tablas `TCNAE` y `TCNO`, respectivamente, en la última hoja del Excel (`Tablas2`). El objetivo es crear las variables `CNACE_grp` y `CNO1_grp`. Para la primera, categorizad `CNACE` en tres niveles (etiquetándolos según las descripciones): 1: empiezan por B, C, D, E, F; 2: empiezan por O, P, Q; 3: el resto. Para la segunda, categorizad `CNO1` según el rango en el código de ocupación: 1: A0–C0; 2: D0–J0; 3: el resto.

Por simplicidad, en esta Actividad trataremos todas las variables categóricas como factores nominales, incluidas aquellas que tienen un orden natural, como los grupos de edad (`ANOS2`) o el nivel de estudios ('ESTU').

1.2. Estudio de correlación lineal

Calculad la matriz de correlación lineal entre todas las variables cuantitativas del estudio e interpretadla.

1.3. Estudio comparativo del salario

Representad gráficamente la variación del salario (`RETRINOIN`) según el sexo y según la nacionalidad (dos gráficos en total). Dado que la distribución del salario es muy asimétrica, debéis complementar la visualización o bien con un summary() o bien aplicando una transformación en el eje Y (solo para la visualización). Interpretad los resultados.

1.4. Generación de los conjuntos de entrenamiento y de test

Ahora separaremos el conjunto de datos en dos partes: el conjunto de entrenamiento (`train`), que incluirá un 80 % de los datos, y el conjunto de prueba (`test`), que incluirá el 20 % restante. En los siguientes apartados, ajustaremos el modelo de regresión lineal con el conjunto de entrenamiento y evaluaremos el rendimiento con el conjunto de prueba. Fijad una semilla (`set.seed(42)`) para asegurar la reproducibilidad.

1.5. Estimación de modelos de regresión lineales simples con variable cualitativa

Queremos analizar si existen desigualdades estructurales asociadas a características demográficas, sin "contaminarlas" todavía con otras variables. A esto lo llamamos analizar la **brecha salarial bruta**.

Estimad dos modelos lineales simples por separado: 1. `RETRINOIN` en función de `SEXO`. 2. `RETRINOIN` en función de `TIPOPAIS` (Nacionalidad).

Para cada uno de los modelos:

- A partir de la salida del modelo, indicad cuál es la **diferencia salarial media estimada** entre los grupos y cuál sería el salario medio del grupo más desfavorable.
- Indicad si estas diferencias son **estadísticamente significativas** y justificad la respuesta citando el p-valor correspondiente.
- Interpretad los valores de R^2 ajustado y explicad qué nos dice sobre la capacidad explicativa del modelo.

1.6. Estimación del modelo de regresión lineal múltiple con predictores cuantitativos

Se sugiere que el salario debería depender principalmente de la cantidad de trabajo y la experiencia. Para comprobarlo, estimad un modelo utilizando **únicamente las variables cuantitativas** explicativas del dataset: la antigüedad (`ANOANTI`) y la jornada anual pactada (`JAP`). Interpretad los coeficientes y los valores de R^2 ajustado.

1.7. Estimación del modelo de regresión lineal múltiple con predictores cuantitativos y cualitativos

Estimad el modelo completo añadiendo las variables cualitativas (`SEXO`, `TIPOPAIS`, `ESTU`, etc.) al modelo cuantitativo anterior.

Una vez aplicado el modelo y en base a los resultados, decidid cuáles de las variables explicativas propuestas hasta el momento deben mantenerse en el modelo de regresión lineal. Para tomar estas decisiones no utilicéis métodos de selección automática. Considerad:

- la significatividad estadística de las variables,
- el valor del coeficiente de determinación ajustado,
- y la posible existencia de problemas de multicolinealidad.

Para evaluar la multicolinealidad, podéis utilizar medidas como el Factor de Inflación de la Varianza (VIF) mediante la función `vif()`. Como referencia orientativa, un VIF = 1 indica ausencia de colinealidad y valores superiores a 5 pueden considerarse potencialmente problemáticos. No es necesario un estudio exhaustivo, pero sí justificar si alguna variable podría ser redundante.

Sobre el modelo final (eliminando, si es preciso, las variables redundantes), observad qué ha pasado con los coeficientes de `SEXO` y `TIPOPAIS` respecto al apartado 1.5. Al controlar por otras variables (obteniendo una **brecha salarial ajustada**), ¿qué ha sucedido? ¿Qué explicación le podríamos dar?

Tened en cuenta que puede haber más de una especificación razonable de modelo final, siempre que las decisiones estén bien justificadas.

1.8. Diagnosis del modelo

Realizad una diagnosis gráfica del modelo múltiple escogido. Generad y analizad un gráfico de residuos (de valores observados menos los predichos por el modelo) vs. valores ajustados, para ver la homocedasticidad, y un gráfico cuantil-cuantil (Q-Q plot) que compara los residuos del modelo con los valores de una variable que se distribuye normalmente. Interpretad los resultados.

1.9. Predicción del modelo

Utilizad el modelo múltiple para predecir los salarios en la tabla `test`. Representad gráficamente los valores predichos frente los valores observados. Evaluad la precisión del modelo mediante la raíz cuadrada del error cuadrático medio (RMSE). Interpretad brevemente el resultado.

---

2. Regresión Logística

En esta segunda parte, crearemos un modelo para clasificar a los empleados en dos grupos: salario alto o no.

2.1. Preparación de los datos (Binarización)

Se quiere estudiar qué factores influyen en la probabilidad de tener un salario alto.

- Para ello, cread una nueva variable binaria (dicotómica) llamada `SalarioAlto`.
- **Definición del umbral:** Para este estudio, usaremos la **mediana global** de `RETRINOIN`(no la mediana por subgrupos) como punto de corte. Esta elección permite trabajar con un dataset aproximadamente equilibrado (50/50) y facilita la evaluación del modelo.
- **Codificación:** Cread la variable `SalarioAlto`, que tomará el valor **1** (evento de interés, "Alto") si `RETRINOIN` es **superior a la mediana** del salario de la muestra, y el valor **0** ("No Alto", categoría de referencia) en caso contrario. Convertid esta nueva variable a factor.
- Para evitar problemas de colinealidad, no usaremos la variable `RETRINOIN` original en los modelos de regresión logística binaria (pero sí mantendremos el resto de predictores del modelo lineal múltiple).
- Reutilizad la misma partición de `train`/`test` definida en el Apartado 1.4, aplicada ahora al conjunto de datos que incluye la nueva variable `SalarioAlto`.

2.2. Estimación del modelo de regresión logística

Usando el conjunto `train` (80 %), estimad un modelo de regresión logística donde la variable dependiente sea `SalarioAlto` y las explicativas sean `SEXO`, `ANOANTI`, `JAP`, `ESTU`, `TIPOCON`, `CNACE_grp`, `CNO1_grp` y `TIPOPAIS`. Mostrad el resumen del modelo e interpretadlo brevemente. ¿Eliminaríais alguna variable del modelo? Justificad vuestra decisión.

Para los siguientes apartados, considerad como modelo logístico de referencia este modelo inicial, aunque en vuestra discusión hayáis propuesto modificarlo.

2.3. Cálculo de las OR (Odds-Ratio)

Calculad las OR para las variables más relevantes del modelo e interpretadlas. Resumid cuáles de las variables pueden considerarse factores de riesgo o de protección.

2.4. Matriz de confusión

Evaluad la precisión del modelo sobre el conjunto `test`. Asumid un punto de corte de probabilidad de 0.5. Para ello:

- Calculad la matriz de confusión (considerando `SalarioAlto = 1` ("Alto") como la clase positiva) e interpretadla brevemente.
- Calculad e interpretad la **Exactitud (Accuracy)**, la **Sensibilidad (Sensitivity)** y la **Especificidad (Specificity)** del modelo.

2.5. Predicción

Identificad los 5 perfiles de empleados del conjunto `test` con la probabilidad predicha más alta de ser `SalarioAlto`. Mostrad en una tabla sus características principales y su probabilidad predicha.

2.6. Bondad del ajuste y curva ROC

Evaluad la eficacia global del modelo.

- Calculad el p-valor del test Chi-cuadrado (diferencia de devianzas nula y residual) usando `pchisq()`. ¿Es el modelo globalmente significativo?
- Dibujad la curva ROC y calculad el área bajo la curva (AUC). Podéis usar la librería pROC. Discutid el resultado.

3. Resumen ejecutivo y Reflexión Ética

3.1. Tabla Resumen

Resumid las conclusiones del estudio en una tabla, indicando la pregunta de investigación, los resultados estadísticos clave y una breve conclusión.

3.2. Conclusiones y Reflexión Ética

Redactad una breve conclusión dirigida a una audiencia no técnica (máximo 2500 caracteres totales) en la que se responda de forma crítica a las siguientes cuestiones. Tened presente que, en esta actividad, la "brecha de género" se ha medido únicamente a partir de la variable `SEXO` (sexo registrado hombre/mujer), que utilizamos como proxy limitada del género.

- Basándoos en vuestros resultados, ¿qué ha pasado con la brecha salarial de género? Comparad el resultado del modelo lineal simple (Apartado 1.5) con el del modelo múltiple (Apartado 1.7). ¿Qué podéis concluir sobre la **brecha salarial "bruta" vs. la "ajustada"**?
- ¿Qué limitaciones tienen los datos y los modelos utilizados (variables no observadas, posibles sesgos, supuestos del modelo…)? ¿Cómo afectan estas limitaciones a la interpretación ética de la brecha salarial?

Puntuación de la actividad

- Apartado 1.1 (5 %)
- Apartado 1.2, 1.3, 1.4 (10 %)
- Apartado 1.5 (5 %)
- Apartado 1.6 (5 %)
- Apartado 1.7 (10 %)
- Apartado 1.8 (5 %)
- Apartado 1.9 (5 %)
- Apartado 2.1 y 2.2 (10 %)
- Apartado 2.3 (5 %)
- Apartado 2.4 (10 %)
- Apartado 2.5 (5 %)
- Apartado 2.6 (10 %)
- Apartado 3.1 (5 %)
- Apartado 3.2 (5 %)
- Calidad del informe dinámico (5 %)