EDA Memoria técnica

Introducción

El presente análisis exploratorio de datos (EDA) se centra en el **conjunto de datos del censo adulto de Estados Unidos**, que contiene información sobre características demográficas, educativas y laborales de los individuos.

El **objetivo del estudio** es examinar cómo diferentes variables se relacionan con el **nivel de ingresos**, clasificado como más o menos de $50,000 anuales. Este análisis permite:

- Identificar los factores que pueden estar asociados con mayores ingresos.
- Explorar diferencias en ingresos según características personales y ocupacionales.
- Proporcionar información útil para futuras investigaciones o decisiones relacionadas con empleo y salario.

El proceso se desarrollará de manera sistemática, iniciando con la limpieza y estandarización de los datos, seguido por un estudio univariante, bivariante y, cuando corresponda, multivariante, acompañado de visualizaciones claras para facilitar la interpretación.

Descripción del conjunto de datos

El conjunto de datos utilizado en este proyecto corresponde al **Censo Adulto de Estados Unidos**, un dataset ampliamente utilizado en estudios socioeconómicos para analizar la relación entre características individuales y el nivel de ingresos.

Cada observación del conjunto de datos representa a un individuo en edad laboral, y las variables recogen información demográfica, educativa, laboral y económica. El objetivo principal del dataset es clasificar a los individuos según si perciben **ingresos anuales superiores o inferiores a 50,000 dólares**, lo que permite estudiar patrones de desigualdad económica y factores asociados al nivel de ingresos.

El dataset presenta un formato estructurado, con observaciones independientes y variables claramente definidas, lo que lo hace adecuado para la realización de un Análisis Exploratorio de Datos (EDA).

Variable objetivo

La variable objetivo del análisis es income, una variable categórica binaria que clasifica a los individuos en dos grupos:

- **≤ 50K: individuos con ingresos anuales iguales o inferiores a 50,000 dólares.**
- **> 50K: individuos con ingresos anuales superiores a 50,000 dólares.**

Esta variable permite estudiar cómo diferentes características personales, educativas y laborales se relacionan con la probabilidad de pertenecer al grupo de mayores ingresos.

Descripción de las variables

A continuación, se describen las principales variables incluidas en el conjunto de datos, agrupadas por tipo:

Variables demográficas

- **age: edad del individuo en años.**
- **sex: sexo del individuo (Male, Female).**
- **race: raza del individuo. En este análisis se recodificó en tres categorías (White, Black, Other) para mejorar la interpretabilidad.**
- **native_country: país de origen del individuo. Se recodificó como United-States y Other debido a la baja representación de algunos países.**
- **marital_status: estado civil del individuo.**

Estas variables permiten analizar diferencias de ingresos asociadas a características personales y sociales.

Variables educativas

- **education: nivel educativo alcanzado por el individuo (por ejemplo, Bachelors, Masters, HS-grad).**
- **education_num: número de años de educación formal completados.**

Estas variables son clave para evaluar la relación entre formación académica y nivel de ingresos.

Variables laborales

- **workclass: tipo de empleo o sector laboral.**
- **occupation: ocupación desempeñada por el individuo.**
- **relationship: relación del individuo dentro del hogar (por ejemplo, Husband, Not-in-family)**
- **hours_per_week: número de horas trabajadas por semana.**

Estas variables permiten analizar cómo la actividad laboral y la dedicación influyen en el nivel de ingresos.

Variables económicas

- **Capital_gain: ganancias de capital declaradas por el individuo.**
- **capital_loss: pérdidas de capital declaradas por el individuo.**

Debido a la alta concentración de valores cero y la presencia de valores extremos, estas variables se analizaron tanto en su forma original como mediante variables binarias derivadas que indican la presencia o ausencia de ganancias o pérdidas de capital.

Hipótesis

Previo al análisis, se establecen las siguientes **hipótesis y preguntas de investigación**:

Hipótesis principal:

- Las características demográficas, educativas y laborales tienen un impacto significativo sobre el nivel de ingresos de los individuos.

Preguntas específicas:

- ¿Un mayor nivel educativo está relacionado con una mayor probabilidad de ganar más de $50,000 anuales?
- ¿Qué efecto tienen la edad y las horas trabajadas por semana sobre los ingresos?
- ¿Existen diferencias en los ingresos según sexo, estado civil u ocupación?
- Entre quienes reportan capital-gain, ¿qué factores se asocian con mayores ganancias?

Estas hipótesis servirán como guía para el análisis exploratorio y permitirán evaluar cómo distintos factores se relacionan con el nivel de ingresos.

Preparación y estandarización de los datos

El conjunto de datos utilizado corresponde al censo adulto de Estados Unidos.
Como paso previo al análisis exploratorio, se cargó el dataset limpio proporcionado y se realizó una estandarización de los nombres de las columnas, transformándolos a formato snake_case, en minúsculas y sin caracteres especiales.

Este proceso no modifica la información contenida en los datos, pero mejora la legibilidad del código, reduce posibles errores y facilita análisis posteriores.

Distribución de la variable objetivo

La variable objetivo del estudio es income, que clasifica a los individuos según si perciben ingresos anuales mayores o menores o iguales a $50,000.

El análisis de su distribución muestra que aproximadamente el 75.9% de los individuos pertenecen al grupo de ingresos ≤ $50,000, mientras que el 24.1% superan dicho umbral.

Este resultado indica la existencia de un desequilibrio de clases, que debe tenerse en cuenta en la interpretación de los resultados del análisis exploratorio y en posibles fases posteriores del proyecto.

Análisis univariante de variables categóricas

Se realizó un análisis univariante de las variables categóricas con el objetivo de comprender la distribución de sus categorías, identificar valores poco representados y detectar posibles necesidades de recodificación.

Las visualizaciones mostraron que algunas variables contenían un elevado número de categorías con frecuencias muy bajas, lo que dificultaba la interpretación y podía introducir ruido en análisis posteriores.

Como resultado de este análisis, se tomaron las siguientes decisiones:

- La variable native.country se transformó en una variable binaria, agrupando los valores en United-States y Other, debido a la baja representación del resto de países.
- La variable race se recodificó en tres categorías (White, Black y Other) con el fin de mejorar la claridad del análisis sin pérdida significativa de información.

Estas transformaciones permitieron simplificar la estructura de los datos y facilitar su interpretación posterior.

Análisis univariante de variables numéricas

Para las variables numéricas se calcularon estadísticas descriptivas y se representaron histogramas con estimación de densidad (KDE) y diagramas de caja (boxplots).

Este análisis puso de manifiesto que la mayoría de las variables numéricas presentan distribuciones asimétricas y valores atípicos, especialmente aquellas relacionadas con ganancias y pérdidas de capital. Asimismo, se observó una alta concentración de valores en determinados rangos, lo que dificultaba su análisis directo.

Creación de nuevas variable

Con el fin de mejorar la interpretación de los datos y facilitar análisis posteriores, se crearon nuevas variables derivadas:

- binned_hours_per_week: creada para agrupar las horas trabajadas por semana en intervalos, reduciendo el impacto de valores extremos.
- has_capital_gain y has_capital_loss: variables binarias creadas a partir de capital_gain y capital_loss, respectivamente, debido a la alta concentración de valores cero y la presencia de pocos valores extremadamente elevados.

Estas transformaciones permiten analizar la presencia o ausencia de ganancias y pérdidas de capital de forma más robusta.

Test chi cuadrado: Nivel Educativo vs Ingresos

El test chi-cuadrado muestra una asociación estadísticamente significativa entre el nivel educativo y el nivel de ingresos ($\chi^2 = 4428.40$, $p < 0.001$), lo que permite rechazar la hipótesis de independencia entre ambas variables. Este resultado confirma que el nivel educativo está fuertemente relacionado con la probabilidad de percibir ingresos superiores a 50,000 dólares anuales.

Test de Mann-Whitney: Edad vs Ingresos

El test de Mann–Whitney indica diferencias significativas en la distribución de la edad entre los grupos de ingresos, confirmando la asociación positiva observada en el análisis bivariante

Análisis bivariante de variables categóricas y el ingreso

Se realizó un análisis bivariante entre las variables categóricas y la variable objetivo income, utilizando gráficos de barras segmentados y tablas de contingencia normalizadas.

Los resultados muestran que las variables relacionadas con la educación, el estado civil, la ocupación y la estructura familiar presentan diferencias claras en la distribución del ingreso. En particular, se observa que:

- Los niveles educativos más altos concentran una mayor proporción de individuos con ingresos superiores a $50,000.
- Las personas casadas presentan una probabilidad significativamente mayor de pertenecer al grupo de mayores ingresos.
- Las ocupaciones más cualificadas muestran mayores proporciones de ingresos elevados.
- La categoría Husband dentro de la variable relationship destaca claramente por una mayor presencia de ingresos altos.

Por el contrario, variables demográficas generales como raza o país de origen, tras las recodificaciones realizadas, muestran un impacto más limitado cuando se analizan de forma aislada.

Análisis bivariante de variables numéricas e Ingreso

El análisis bivariante de las variables numéricas se realizó mediante diagramas de caja, comparando las distribuciones entre los dos grupos de ingresos.

Los resultados indican que:

- La edad presenta una asociación positiva con el nivel de ingresos, observándose medianas más elevadas en el grupo de mayores ingresos.
- El número de años de educación formal (education_num) muestra una diferencia clara entre ambos grupos, confirmando su relevancia en la explicación del ingreso.

  Información interesante: hay personas con alta educación ganando poco (valores atípicos superiores en <=50K) y personas con baja educación ganando bien (valores atípicos inferiores en >50K) . Esto sugiere que la educación es importante pero no es el único factor. (En este caso usamos un lineplot en la presentación).

- Las horas trabajadas por semana presentan una mediana similar en ambos grupos, aunque el grupo de mayores ingresos muestra una mayor dispersión hacia valores altos.

- Las variables relacionadas con capital presentan distribuciones altamente asimétricas, con valores positivos concentrados principalmente en el grupo de mayores ingresos.

Análisis multivariante

Como aproximación inicial al análisis multivariante, se estudió la matriz de correlación entre las variables numéricas.

Los resultados muestran que no existen correlaciones lineales fuertes entre las variables, lo que indica una baja multicolinealidad y sugiere que cada variable aporta información complementaria al análisis.

Asimismo, el análisis conjunto de variables como nivel educativo y horas trabajadas por semana refuerza la idea de que el nivel de ingresos es el resultado de la combinación de múltiples factores, siendo la educación un elemento especialmente relevante.

Siendo la educación un elemento relevante en determinar los ingresos, estudiamos como se relacionan los componentes sociales de genero y grupo racial con el nivel educativo

.

Composición de la muestra: La muestra presenta un claro desequilibrio demográfico. Los hombres representan el 66.92% (21,775 personas) frente al 33.08% de mujeres (10,762 personas). En términos raciales, la población blanca domina con un 85.43% (27,795 personas), seguida de población negra con 9.60% (3,122) y otros grupos con apenas 4.98% (1,620).

Nivel educativo por grupos: En una escala educativa de 1 a 16, los promedios oscilan entre 9.42 y 10.47 años, mostrando diferencias moderadas pero significativas entre grupos:

- Población blanca: Niveles similares entre hombres (10.14) y mujeres (10.13), sin brecha de género apreciable.
- Población negra: Presenta los niveles educativos más bajos del estudio. Las mujeres (9.55) superan ligeramente a los hombres (9.42), invirtiendo el patrón tradicional.
- Otros grupos: Exhiben la mayor brecha de género, con hombres alcanzando el nivel más alto (10.47) y mujeres uno de los más bajos (9.96) del dataset.

la interacción entre género y raza revela dinámicas educativas diferenciadas. La brecha de género no es constante: favorece a mujeres en población negra, es neutral en población blanca, y favorece marcadamente a hombres en otros grupos raciales.