Corporación Universitaria del Huila
Facultad de Ingeniería
Programa Ingeniería de Sistemas
Integrantes: Wilkyn Julián Vargas Bahamón, Johan Ariel Verjan Carrillo
Docente: Ing. Julián Andrés Quimbayo Castro

1. Conclusiones

El análisis realizado se desarrolló a partir del dataset Adult Income, el cual contiene más de 32 000 registros con características sociodemográficas y laborales. El objetivo fue estudiar cómo influyen la edad, el nivel educativo y el sexo biológico en la probabilidad de obtener ingresos superiores a 50 000 USD al año. Para ello, se siguieron once pasos que incluyeron la carga del dataset, la limpieza de valores faltantes, la traducción de categorías, la estandarización de variables, el análisis exploratorio, la construcción de visualizaciones, la normalización de los datos y la preparación del modelo. Dentro de este proceso también fue necesario aplicar One-Hot Encoding para transformar las variables categóricas en un formato numérico. Aunque esto incrementó la cantidad de columnas del dataset, su uso era indispensable porque garantiza que la regresión lineal maneje correctamente las categorías sin crear relaciones numéricas falsas entre ellas, como ocurriría con métodos como LabelEncoder. Gracias a esta técnica, cada categoría se representó de manera independiente y el modelado estadístico mantuvo su coherencia. En conjunto, estos pasos permitieron preparar adecuadamente la información y garantizar que las conclusiones se fundamentaran en datos limpios y estructurados.

Entre los hallazgos principales, las visualizaciones y estadísticas revelaron patrones consistentes en la relación entre las variables sociodemográficas y los ingresos. El nivel educativo mostró una relación positiva con mayores ingresos: personas con licenciatura, maestría o doctorado presentan una probabilidad considerablemente más alta de superar los 50 000 USD anuales. La edad también evidenció un patrón creciente, donde las personas adultas con mayor experiencia laboral tienen mejores oportunidades salariales. El estado civil y la ocupación fueron otros factores clave: las personas casadas y quienes trabajan en ocupaciones ejecutivas, directivas o profesionales presentan ingresos superiores, mientras que las personas solteras o en empleos operativos tienden a formar parte del grupo de menores ingresos. Asimismo, se identificaron brechas importantes por sexo biológico, donde los hombres presentan mayores niveles de ingreso en la mayoría de ocupaciones.

ACTIVIDAD EDA

Respecto a la pregunta de investigación, los resultados permiten afirmar que la edad, el nivel educativo y el sexo biológico sí influyen en la probabilidad de alcanzar ingresos superiores a 50 000 USD al año. La educación es el factor más determinante, seguida de la edad, mientras que el sexo biológico presenta un efecto menor pero estadísticamente visible en las correlaciones y tendencias observadas. El modelo de regresión lineal, aunque limitado, confirmó estas relaciones a través de los coeficientes obtenidos y coincidió con las conclusiones del análisis exploratorio.

La interpretación de estos resultados muestra que las variables estudiadas funcionan como indicadores sociales y laborales que reflejan dinámicas económicas dentro del dataset. Una mayor formación académica suele abrir acceso a empleos mejor remunerados, y la edad refleja la experiencia acumulada que contribuye a aumentar los ingresos a lo largo del tiempo. Las diferencias por sexo biológico evidencian desigualdades presentes en el mercado laboral, reforzadas por factores como el tipo de ocupación, el rol familiar y la estabilidad en el empleo. En conjunto, el análisis resalta la importancia de las condiciones sociodemográficas en la determinación de los ingresos salariales.

No obstante, el estudio presenta varias limitaciones. En primer lugar, el dataset contiene una desproporción entre clases, con una baja representación de personas que ganan más de 50 000 USD, lo que afecta la capacidad predictiva del modelo. La regresión lineal, por su naturaleza, no es el método más adecuado para clasificación binaria, ya que predice valores continuos y no categorías, generando alta dispersión en las predicciones. Además, aunque el One-Hot Encoding era necesario, incrementó notablemente la dimensionalidad del dataset, lo cual puede introducir ruido y volver el modelo más pesado de entrenar. Otro límite es que varias variables presentan distribuciones no normales o valores extremos, lo cual afecta la estabilidad y el ajuste del modelo. Finalmente, el dataset refleja una realidad específica y podría no generalizarse a otras poblaciones o contextos.

A pesar de estas limitaciones, el análisis ofrece recomendaciones valiosas para futuros trabajos. Para mejorar el rendimiento predictivo sería conveniente utilizar modelos diseñados para clasificación, como regresión logística, Random Forest o XGBoost, que manejan mejor relaciones no lineales y múltiples variables categóricas. También se recomienda aplicar técnicas de balanceo como SMOTE para mejorar la detección de personas con ingresos altos. De igual forma, sería beneficioso integrar variables más detalladas relacionadas con la experiencia laboral, el sector económico o el tipo de contrato, así como explorar interacciones entre variables (por ejemplo, educación por sexo u ocupación por edad) para capturar mejor la relación entre las características sociodemográficas y los ingresos.

En conclusión, este estudio permitió comprender de manera clara y fundamentada cómo factores como la edad, la educación y el sexo influyen en la probabilidad de obtener ingresos altos, demostrando la relevancia de estas variables dentro de la estructura salarial representada en el dataset y destacando las posibilidades de análisis y mejora para investigaciones futuras.