A green and black logo

AI-generated content may be incorrect.A logo of a university

AI-generated content may be incorrect.Universidad Tecnológica de Panamá

A logo of a university

AI-generated content may be incorrect.

Maestría en Analítica de Datos

Materia: Modelos Predictivos

Profesor: Juan Marcos Castillo, PhD

Estudiante: Irma Martínez

Título: Reporte- Predicción de Nivel de Ingreso

Fecha: 21/07/2025

1. **Introducción**

Conocer qué factores pueden influir en cuánto gana una persona es algo que nos ayuda a entender mejor las condiciones sociales y económicas en las que vivimos. En este estudio, se buscar relacionar cómo aspectos como el tipo de empleo, el estado civil o el nivel de educación están relacionados con los ingresos que recibe un individuo.

Para esto, usamos una base de datos, Adult *Income Dataset*, que contiene información recolectada en el Censo de Estados Unidos de 1994 y está disponible en la biblioteca UCI de aprendizaje automático. En este conjunto, hay 15 variables que describen características socioeconómicas de las personas, además de una variable que divide a los ingresos en dos grupos: quienes ganan $50,000 o menos y quienes ganan más que eso.

El primer paso fue revisar bien los datos para asegurarnos de que su calidad estuviera apta, aplicando métodos de análisis exploratorio y estadística descriptiva. Luego, seleccionamos las variables que parecían tener una relación con los ingresos. Por último, aplicamos técnicas de predicción para clasificar a las personas según su nivel de ingresos, utilizando algoritmos que están hechos para este tipo de análisis.

El propósito principal es entender mejor los patrones que afectan los ingresos y crear modelos que puedan servir para apoyar decisiones relacionadas con políticas sociales y económicas, buscando así contribuir a un panorama más justo.

**2. Justificación**

El ingreso de una persona define en muchos casos su acceso a oportunidades, su bienestar y su futuro. Hay elementos en la vida que pueden marcar una gran diferencia en cuánto gana y en sus posibilidades de progresar. Entender como se conectan estos factores no es solo una cuestión estadística: es una manera de entender realidades humanas que están presente día a día.

Comprender cómo ciertos factores están relaciones con el ingreso de una persona es fundamental para la estructuración de estrategias orientadas a la igualdad de oportunidades y el desarrollo social. Esto, permite identificar factores claves que impactan en la distribución de la riqueza y las oportunidades económicas dentro de una población. Se relacionan con el ingreso de una persona es fundamental para el análisis y la formulación de estrategias orientadas a la igualdad de oportunidades y el desarrollo social.

Esta relación permite identificar factores clave que influyen en la distribución de la riqueza y las oportunidades económicas dentro de una población. Además, proporciona información valiosa para la toma de decisiones por parte de gobiernos, instituciones educativas y el sector privado, permitiéndoles diseñar políticas, programas o intervenciones laborales más eficaces.

A nivel personal, estos conocimientos ofrecen a las personas herramientas para tomar decisiones informadas sobre su educación, trayectoria profesional y desarrollo personal.

Desde un punto de vista técnico, identificar las variables que mejor predicen los ingresos también contribuye a la construcción de modelos analíticos robustos, útiles para diversas aplicaciones en el ámbito económico, social y académico

**3. Antecedentes**

El Adult Income Dataset ha sido utilizado en múltiples estudios como referencia para comparar algoritmos de clasificación. Estudios anteriores han aplicado modelos como regresión logística, árboles de decisión, Naives Bayes, redes neuronales y random forest para predecir la variable objetivo: ingreso mayor a 50 mil dólares anuales o menor igual que esta cantidad.

**4. Definición del problema**

Problema: Tomando en cuenta ciertas variables, ¿se podría predecir si una persona gana más de $50,000 anuales en función de sus características demográficas y laborales?

Variable de salida: ingreso (binaria: >50K o <=50K)

Variables predictoras:

- edad  
- educación  
- ocupación  
- estado civil  
- raza  
- género  
- horas por semana trabajadas  
- País de nacimiento

**5. Análisis Predictivo**

**5.1 Determinación de la base de datos**

El conjunto de datos fue obtenido mediante la UCI Machine Learning que es un extracto de una encuesta del censo de los estados unidos, 1994.

**5.1.1. Características del dataset**

* **Número de instancias:** 48,842
* **Número de atributos predictivos:** 14
* **Tipo de problema:** Clasificación binaria
* **Variable objetivo:** income (<=50K o >50K)
* **Tipo de datos:** Mixto (numérico y categórico)

**5.1.2. Atributos**

Los principales atributos son:

**Atributo Tipo Descripción**

* **age** Numérico Edad de la persona
* **workclass** Categórico Tipo de empleo (privado, estatal, etc.)
* **fnlwgt** Numérico Peso estadístico de la muestra
* **education** Categórico Nivel de educación
* **education**-num Numérico Equivalente numérico del nivel educativo
* **marital-status** Categórico Estado civil
* **occupation** Categórico Ocupación
* **relationship** Categórico Relación familiar
* **race** Categórico Raza
* **sex** Categórico Género
* **capital-gain** Numérico Ganancia de capital
* **capital-loss** Numérico Pérdida de capital
* **hours-per-week** Numérico Horas trabajadas a la semana
* **native-country** Categórico País de origen
* **income** Categórico (target) Nivel de ingreso (<=50K, >50K)

**5.2 Preprocesamiento y limpieza**

* Se eliminaron los espacios innecesarios en las cadenas de texto.
* Se reemplazaron los valores desconocidos representados por "?" con NaN.
* Se eliminaron registros con valores faltantes para garantizar integridad.
* La variable objetivo income fue transformada a formato binario:

<=50K → 0 >50K → 1

A screen shot of a computer code

AI-generated content may be incorrect.

**5.3 Análisis descriptivo**

Se aplicaron técnicas estadísticas y visualizaciones con matplotlib y seaborn para identificar tendencias y relaciones relevantes:

* Gráfico de caja entre edad e ingreso
* Histograma de hours-per-week
* Conteo cruzado de educación vs ingreso
* Mapa de calor de correlaciones entre variables numéricas

Estas visualizaciones permitieron detectar relaciones como:

* A mayor edad y educación, mayor probabilidad de altos ingresos.
* Personas con más horas de trabajo semanal tienen ingresos superiores.

A screenshot of a graph

AI-generated content may be incorrect.

**5.4 Selección de variables**

Se transformaron las variables categóricas en variables dummy mediante pd.get\_dummies.

Se aplicaron dos métodos de selección de características:

* SelectKBest con chi-cuadrado para encontrar variables relevantes estadísticamente.
* Importancia de características del modelo RandomForestClassifier.

Las 10 variables más relevantes de ambos métodos se combinaron para formar el conjunto de predictores finales.  
Las variables seleccionadas combinadas incluyeron atributos como education-num, marital-status, occupation, hours-per-week, entre otras.

A computer screen shot of a program code

AI-generated content may be incorrect.

**5.5 Selección de modelos**

**5.5.1. Separación de datos**

Se dividió el dataset en conjuntos de entrenamiento y prueba usando train\_test\_split, con una proporción 70/30 y estratificación para preservar la proporción de clases.

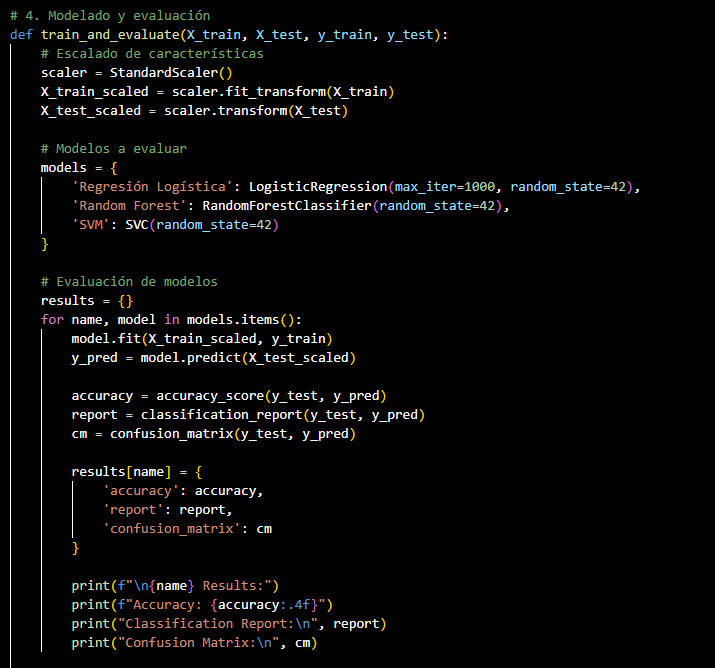
A computer code on a black background

AI-generated content may be incorrect.

**5.5.2 Entrenamiento de modelos**

Se entrenaron 3 algoritmos:

* Regresión Logística
* Random Forest
* SVM



Todos los modelos fueron ajustados sobre los datos escalados (StandardScaler), y evaluados mediante las métricas:

* Accuracy
* Matriz de confusión
* Classification Report (precisión, recall, F1-score)

**7. Evaluación comparativa**

Se compararon los modelos gráficamente en términos de su accuracy. El modelo de Random Forest obtuvo el mejor rendimiento general, seguido de cerca por la Regresión Logística.

A blue rectangular object with text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**6. Conclusiones**

* Los modelos desarrollados demostraron ser eficaces para predecir el nivel de ingresos con un buen nivel de precisión.
* Las variables educativas, laborales y patrimoniales tienen alto poder explicativo en la predicción.
* Random Forest mostró un excelente equilibrio entre rendimiento y facilidad de interpretación.
* El preprocesamiento y la selección adecuada de variables fueron factores cla

**7. Recomendaciones y futuros estudios**

- Explorar técnicas de optimización de hiperparámetros (GridSearchCV).  
- Probar algoritmos avanzados como XGBoost o redes neuronales.  
- Analizar el impacto del sesgo en los datos respecto a variables sensibles como raza o género.

- Utilizar técnicas de validación cruzada para mejorar la generalización del modelo.

**8. Bibliografía**

- Kohavi, R., & Becker, B. (1996). UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/adult  
- Scikit-learn documentation: <https://scikit-learn.org/stable/>

- Scikit-learn: Machine learning in Python. Journal of machine learning research, 12, 2825-2830.