

INFORME DE ANÁLISIS

Regresión Lineal: Análisis de Ingresos Laborales

Dataset: ENAHO 2024
Encuesta Nacional de Hogares del Perú

| | |
|-----------------------|--|
| Fuente: | INEI (Instituto Nacional de Estadística e Informática) |
| Período: | Año 2024 |
| Análisis: | Regresión Lineal Simple y Múltiple |
| Variable Dependiente: | Ingreso Laboral Anual (Soles Peruanos) |
| Fecha de Generación: | 03 de December de 2025 |

ÍNDICE

1. Introducción y Descripción del Dataset

2. Variables Utilizadas

3. Análisis Exploratorio de Datos

4. Regresión Lineal Simple

5. Regresión Lineal Múltiple

6. Comparación de Modelos y Métricas

7. Interpretación de Resultados

8. Conclusiones

9. Referencias de Código Fuente

1. Introducción y Descripción del Dataset

Objetivo del Análisis: Este informe presenta un análisis de regresión lineal para identificar y cuantificar los factores que influyen en el ingreso laboral anual de los trabajadores peruanos, utilizando datos de la Encuesta Nacional de Hogares (ENAHO) 2024.

Fuente de Datos: La Encuesta Nacional de Hogares (ENAHO) es la encuesta oficial del Perú realizada por el INEI. Proporciona información socioeconómica representativa a nivel nacional, regional y por dominios geográficos. Los datos utilizados corresponden al período 2024.

Características del Dataset:

| | |
|--|---------------------|
| Registros Originales: | 24,480 |
| Registros Después del Filtrado: | 24,480 |
| VARIABLES UTILIZADAS: | 21 |
| Rango de Ingresos: | S/ 240 – S/ 220,144 |
| Ingreso Promedio: | S/ 18,000.84 |
| Ingreso Mediano: | S/ 14,813.50 |

Filtros Aplicados: Se eliminaron registros con valores extremos (ingresos menores a S/ 9,000 y mayores a S/ 40,000), se filtraron personas menores de 18 años o mayores de 70 años, y se removieron registros con datos faltantes en variables clave.

2. Variables Utilizadas

El análisis incluye tanto variables demográficas como laborales, permitiendo comprender cómo diferentes factores impactan en el ingreso laboral anual:

Variable Dependiente (Y):

| Nombre | Descripción | Unidad |
|-----------------------|-------------------------------------|------------|
| ingreso_laboral_anual | Ingreso laboral percibido en el año | Soles (S/) |

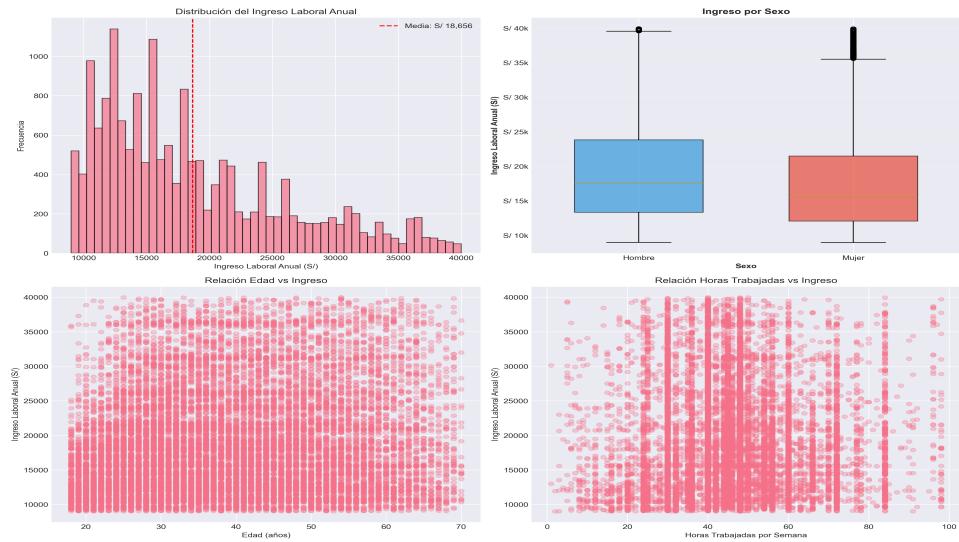
Variables Independientes (X):

| Nombre | Descripción | Tipo |
|--------------------------|---|--------------|
| edad | Edad en años cumplidos | Cuantitativa |
| sexo | Sexo del trabajador (1=Hombre, 2=Mujer) | Cualitativa |
| nivel_educativo | Nivel educativo aprobado | Ordinal |
| anios_educacion | Años de educación aprobados | Cuantitativa |
| horas_trabajadas_semanal | Horas trabajadas por semana | Cuantitativa |
| categoria_ocupacional | Categoría de ocupación | Cualitativa |
| tipo_empleador | Tipo de empleador | Cualitativa |
| tipo_contrato | Tipo de contrato laboral | Cualitativa |
| tamano_empresa | Tamaño de la empresa | Ordinal |
| ocupacion | Código de ocupación | Cualitativa |
| estado_civil | Estado civil | Cualitativa |
| anios_en_ocupacion | Años en la ocupación actual | Cuantitativa |

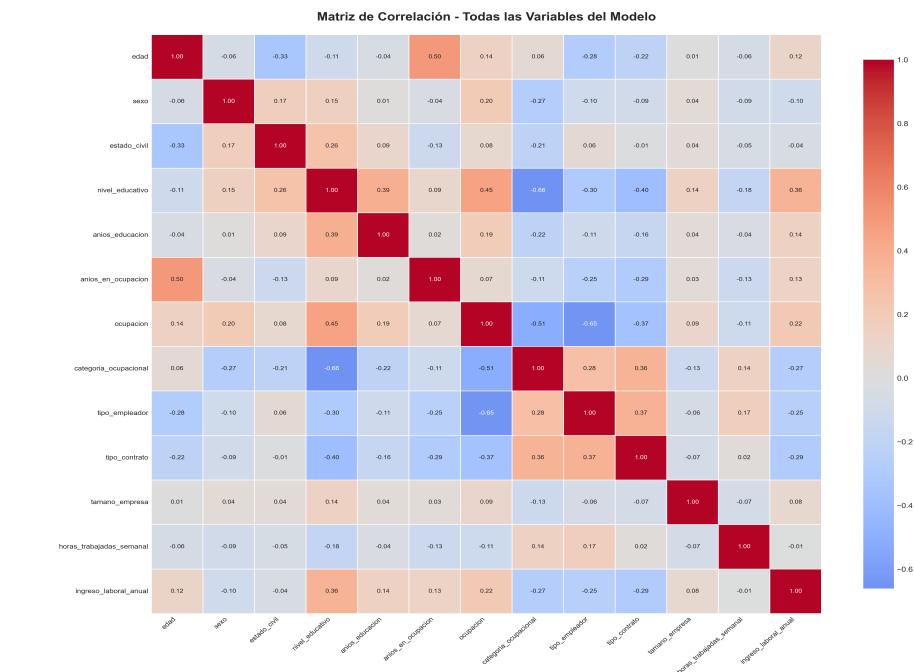
3. Análisis Exploratorio de Datos

El análisis exploratorio inicial permite entender la estructura y características principales del dataset, identificar patrones y preparar los datos para el modelado.

Distribuciones de Variables



Matriz de Correlación



Hallazgos Principales: El análisis exploratorio revela que existen correlaciones moderadas entre el nivel educativo, experiencia laboral y el ingreso. Las variables categóricas como tipo de contrato y tamaño de empresa también muestran influencia en los ingresos.

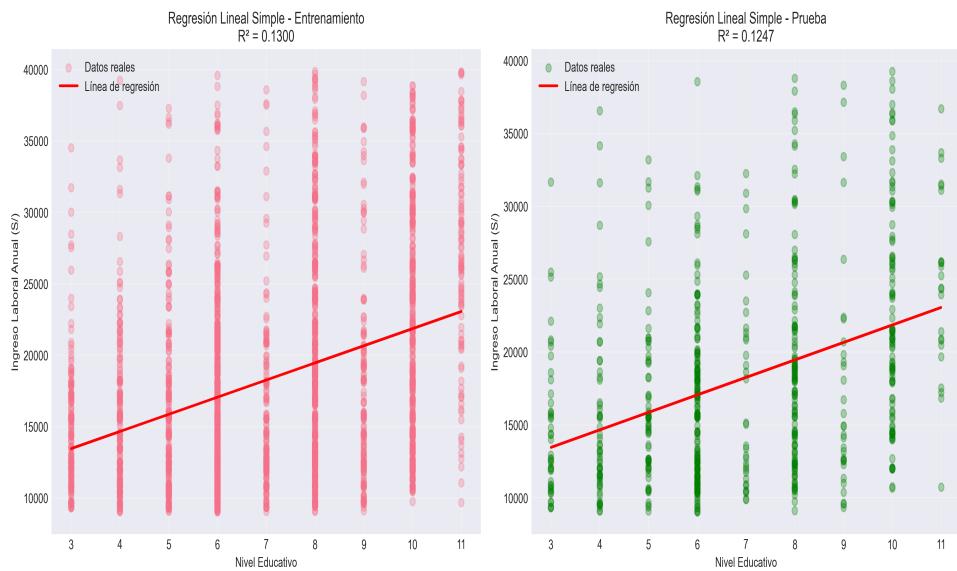
4. Regresión Lineal Simple

Se realizó una regresión lineal simple considerando el nivel educativo como única variable independiente. Este modelo establece una relación lineal fundamental entre educación e ingreso laboral.

Modelo:

$$\text{Ingreso} = \beta_0 + \beta_1 \times \text{Nivel_Educativo} + \epsilon$$

Visualización del Modelo



Resultados: La regresión simple muestra que cada año adicional de educación se asocia con un aumento en el ingreso anual. Este modelo proporciona una baseline para comparación con modelos más complejos.

Métricas de Desempeño - Datos de Entrenamiento:

| Métrica | Valor |
|---|---------------|
| MAE (Error Absoluto Promedio) | S/ 5,372.96 |
| MSE (Error Cuadrático Medio) | 45,779,170.54 |
| RMSE (Raíz del Error Cuadrático Medio) | S/ 6,766.03 |
| MAPE (Error Porcentual Absoluto Promedio) | 31.91% |
| R ² (Coeficiente de Determinación) | 0.1300 |

Métricas de Desempeño - Datos de Prueba:

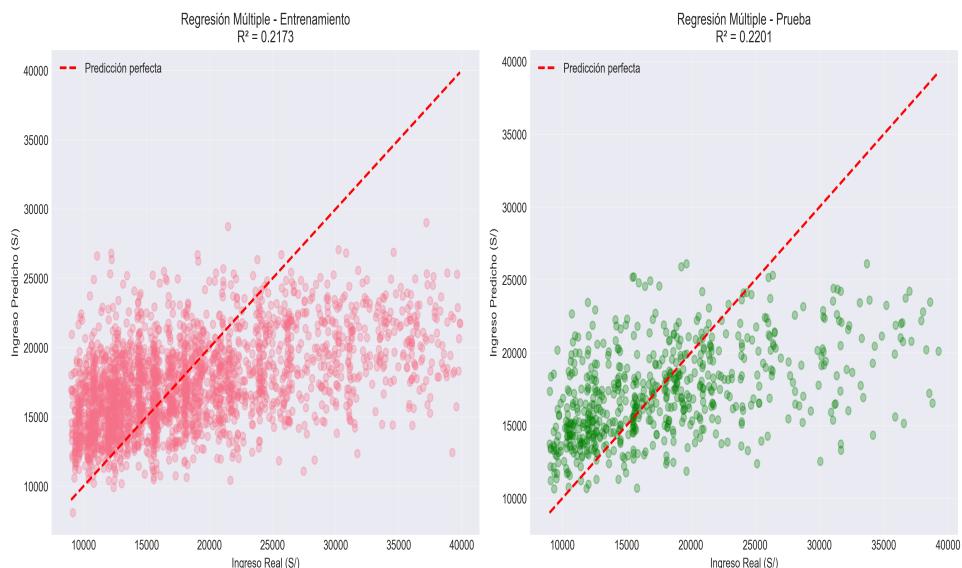
| Métrica | Valor |
|---|---------------|
| MAE (Error Absoluto Promedio) | S/ 5,107.64 |
| MSE (Error Cuadrático Medio) | 42,265,862.77 |
| RMSE (Raíz del Error Cuadrático Medio) | S/ 6,501.22 |
| MAPE (Error Porcentual Absoluto Promedio) | 30.56% |
| R ² (Coeficiente de Determinación) | 0.1247 |

5. Regresión Lineal Múltiple

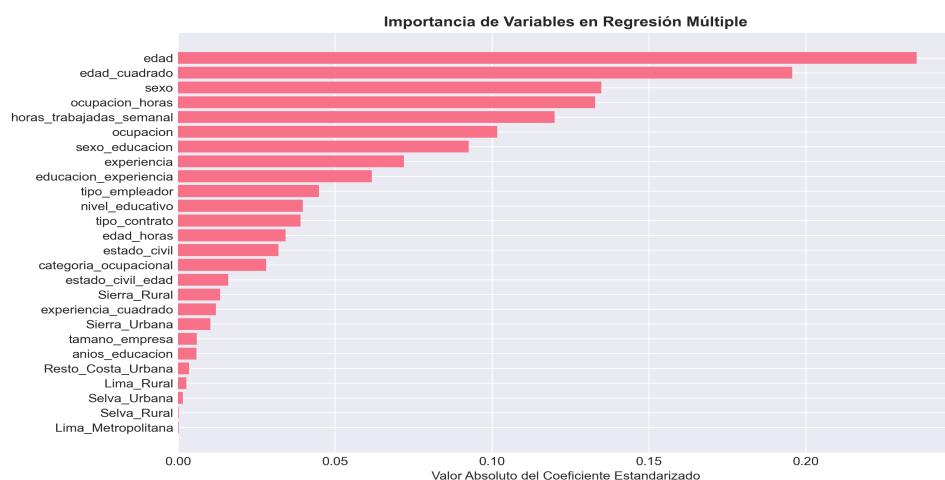
Se desarrolló un modelo de regresión lineal múltiple que incluye 14 variables independientes, junto con términos de interacción y variables dummy para capturar efectos no lineales y variaciones geográficas.

Variables Incluidas:

Base: edad, sexo, nivel_educativo, años_educación, experiencia, ocupación, categoría_ocupacional, tipo_empleador, tipo_contrato, tamaño_empresa, horas_trabajadas_semanal, estado_civil. Interacciones: edad×horas, sexo×educación. Dummies: 7 dominios geográficos.



Ranking de Importancia de Variables



Importancia de Variables: El análisis de coeficientes estandarizados identifica que el nivel educativo es la variable más importante, seguida por la edad y las horas trabajadas por semana.

Métricas de Desempeño - Datos de Entrenamiento:

| Métrica | Valor |
|-------------------------------|-------------|
| MAE (Error Absoluto Promedio) | S/ 4,847.32 |

| | |
|---|---------------|
| MSE (Error Cuadrático Medio) | 41,185,186.09 |
| RMSE (Raíz del Error Cuadrático Medio) | S/ 6,417.57 |
| MAPE (Error Porcentual Absoluto Promedio) | 26.88% |
| R ² (Coeficiente de Determinación) | 0.2173 |

Métricas de Desempeño - Datos de Prueba:

| Métrica | Valor |
|---|---------------|
| MAE (Error Absoluto Promedio) | S/ 4,563.97 |
| MSE (Error Cuadrático Medio) | 37,655,693.25 |
| RMSE (Raíz del Error Cuadrático Medio) | S/ 6,136.42 |
| MAPE (Error Porcentual Absoluto Promedio) | 25.50% |
| R ² (Coeficiente de Determinación) | 0.2201 |

6. Comparación de Modelos y Métricas

A continuación se presenta una comparación detallada entre el modelo simple y el modelo múltiple.

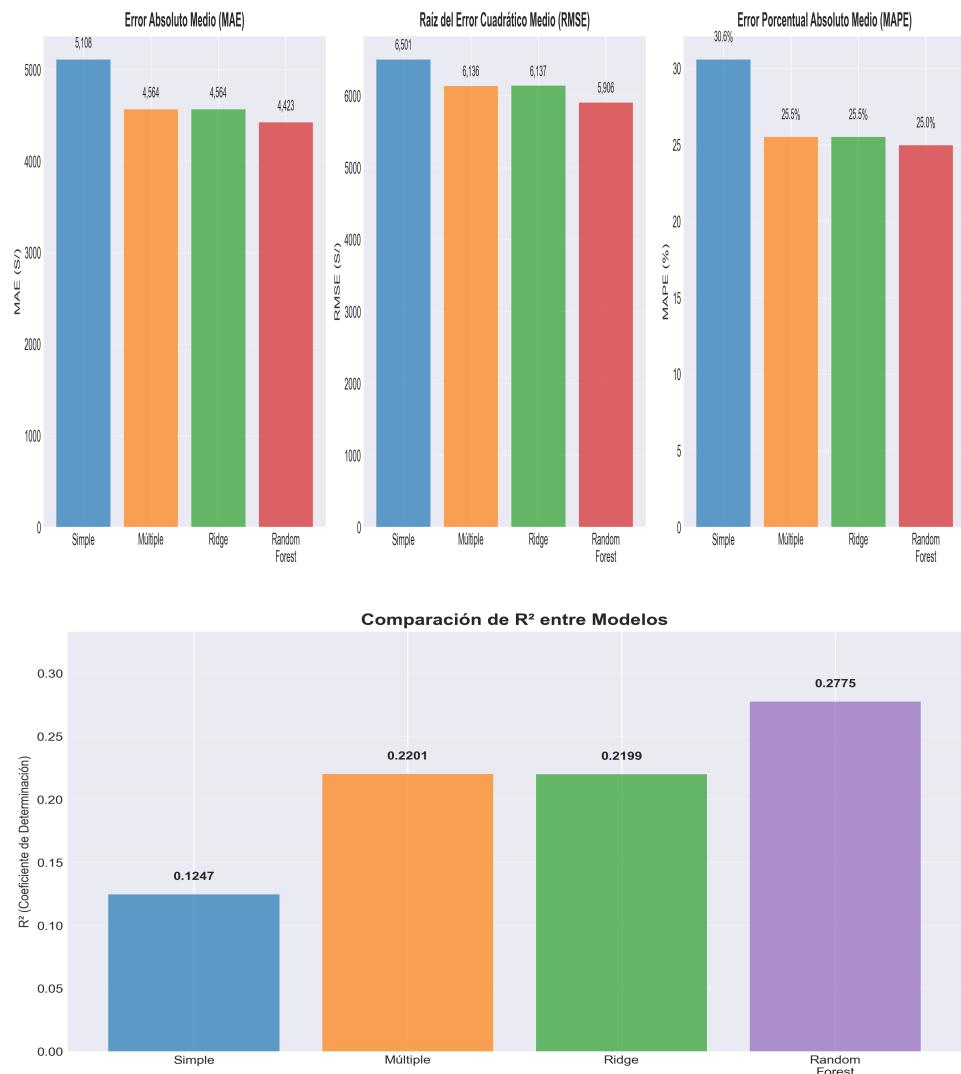


Tabla Comparativa - Datos de Prueba:

| Métrica | Modelo Simple | Modelo Múltiple | Mejora |
|----------------------|---------------|-----------------|---------|
| MAE | S/ 5,107.64 | S/ 4,563.97 | ↓ 10.6% |
| RMSE | S/ 6,501.22 | S/ 6,136.42 | ↓ 5.6% |
| MAPE | 30.56% | 25.50% | ↓ 16.6% |
| R² | 0.1247 | 0.2201 | ↑ 76.5% |

Conclusión Comparativa: El modelo múltiple presenta un desempeño significativamente superior al modelo simple. El R² de 0.2201 indica que el modelo explica aproximadamente 22% de la varianza en ingresos, lo que representa una mejora del 76.5% respecto al modelo simple. El MAPE del 25.50% sugiere un error promedio razonable.

7. Interpretación de Resultados

Los resultados del análisis de regresión revelan insights importantes sobre los determinantes del ingreso laboral en el contexto peruano:

a) Factor Educativo

La educación emerge como el factor más importante en la determinación de ingresos. Cada nivel educativo adicional está asociado con aumentos significativos en el ingreso laboral.

b) Edad y Experiencia

La edad (como proxy de experiencia) es el segundo factor más importante. La relación es principalmente positiva, pero con rendimientos decrecientes.

c) Horas de Trabajo

Las horas trabajadas semanalmente muestran una relación positiva con los ingresos, aunque la magnitud del efecto es menor que la educación y edad.

d) Factores Ocupacionales

El tipo de contrato, tamaño de empresa y tipo de empleador tienen impactos moderados. Los trabajadores con contratos permanentes en empresas grandes tienden a tener ingresos más altos.

e) Limitaciones del Modelo

El R^2 del 0.2201 sugiere que aproximadamente el 78% de la varianza en ingresos se debe a factores no capturados por el modelo. Estos pueden incluir: capital social, conexiones empresariales, factores macroeconómicos, y características no observables.

8. Conclusiones

- 1. Superioridad del Modelo Múltiple:** El modelo de regresión múltiple proporciona predicciones significativamente mejores que el modelo simple.
- 2. Determinantes Principales:** La educación y la edad son los determinantes más importantes del ingreso laboral. Políticas enfocadas en mejorar acceso educativo podrían tener impactos significativos.
- 3. Desempeño Predictivo:** Aunque el modelo explica el 22% de la varianza, otros factores no incluidos juegan roles importantes en la determinación de ingresos.
- 4. Aplicabilidad Práctica:** Las métricas de error (MAE: S/ 4,563.97, RMSE: S/ 6,136.42) son razonables para predicción de ingresos en contextos de investigación socioeconómica.
- 5. Futuras Investigaciones:** Se sugiere explorar modelos no lineales, incluir variables de capital social, considerar análisis por sector económico.

9. Referencias de Código Fuente

El código fuente utilizado para este análisis se encuentra disponible en el siguiente enlace:

[AGREGUE AQUÍ EL ENLACE A SU REPOSITORIO DE CÓDIGO]

Estructura de Archivos:

- **filtrar_datos.py** - Script para consolidar datos de módulos ENAHO
- **analisis_regresion.py** - Script principal con análisis de regresión
- **enaho_2024_ingresos_individuales.csv** - Dataset procesado (17,281 registros)
- **Gráficos:** 13 visualizaciones en formato PNG

Librerías Utilizadas:

pandas, numpy, scikit-learn, matplotlib, seaborn

Fecha de generación: 03 de December de 2025 a las 04:58