Proyecto Análisis Modelado Optimización Datos

Análisis y Desarrollo de Software

Data set: Caracterización del Empleo Publico

Nicolas Quintana

Yeisson Romero

Juan Marín

Christian Joven

Sebastian Lizarazo

Ficha 3066474

Sena

Esteban Hernández

15/12/2025

1. **introducción**

Este proyecto tiene como objetivo realizar un análisis del empleo público orientado a la estimación del salario mensual promedio, mediante la aplicación de técnicas de analítica de datos e inteligencia artificial. Para asegurar un proceso estructurado y reproducible, se adoptó la metodología ASUM-DM, la cual guía el desarrollo del proyecto desde la comprensión del problema hasta la evaluación y comunicación de resultados.

Durante el desarrollo del proyecto se llevó a cabo un análisis exploratorio de datos (EDA) que permitió evaluar la calidad de la información, identificar patrones y detectar posibles anomalías. Posteriormente, se aplicaron procesos de limpieza y transformación de los datos con el fin de prepararlos adecuadamente para la etapa de modelado. Finalmente, se implementó un modelo basado en redes neuronales, el cual fue optimizado a partir de una primera iteración, evidenciando mejoras en su desempeño mediante el uso de métricas cuantitativas.

2. **Metodología ASUM-DM**

La metodología ASUM-DM se utilizó como marco estructural del proyecto, permitiendo organizar de manera sistemática cada una de las etapas del análisis de datos. Esta metodología facilitó la planificación del trabajo desde la comprensión del problema hasta la evaluación y comunicación de los resultados, asegurando coherencia, orden y reproducibilidad.

El proyecto inició con la definición del problema y la selección del conjunto de datos, seguida por las fases de comprensión y preparación de la información, donde se llevaron a cabo procesos de exploración, limpieza y transformación de los datos. Posteriormente, se estructuró la etapa de modelado de forma iterativa, permitiendo la evaluación y optimización del modelo seleccionado. Todo el proceso fue documentado y gestionado a través de un repositorio en GitHub, garantizando la trazabilidad y el trabajo colaborativo.

- **Comprensión del problema,** donde se definieron los objetivos y el alcance del análisis.
- **Comprensión de los datos,** orientada a la selección y revisión inicial del conjunto de datos.
- **Preparación de los datos,** que incluyó procesos de limpieza, transformación y normalización.
- **Modelado,** en la cual se construyó y evaluó una primera versión del modelo.
- **Evaluación y comunicación,** enfocadas en analizar el desempeño del modelo y presentar los resultados de manera clara.

3. Comprensión del problema

El presente proyecto tiene como objetivo analizar el comportamiento del salario mensual promedio dentro del empleo público, a partir de un conjunto de datos oficiales que caracterizan distintas variables asociadas al sector.

Debido a la complejidad del fenómeno analizado y a la posible presencia de relaciones no lineales entre las variables, se decidió emplear un modelo de inteligencia artificial, específicamente una red neuronal, como herramienta principal para el análisis predictivo.

La finalidad del modelo es identificar patrones relevantes y evaluar su capacidad para estimar el salario mensual promedio, permitiendo así apoyar procesos de análisis y toma de decisiones basadas en datos.

4. Comprensión de los datos

El data set utilizado corresponde a la Caracterización del Empleo Público, el cual contiene información agregada relacionada con variables sociodemográficas, institucionales y salariales.

Durante esta fase se realizó una exploración inicial de los datos con el fin de:

- Identificar el tipo de variables (numéricas y categóricas).
- Detectar valores faltantes.
- Analizar la distribución de los datos.
- Reconocer posibles valores atípicos.
- Evaluar la escala de las variables.

5. Análisis Exploratorio de Datos (EDA)

A partir del análisis exploratorio se obtuvieron los siguientes hallazgos relevantes:

- El salario mensual promedio presenta una alta variabilidad entre los registros.
- Existen variables categóricas que requieren procesos de codificación para su uso en modelos predictivos.
- Se identificaron valores faltantes que debieron ser tratados para evitar sesgos en el análisis.
- Las variables presentan escalas muy diferentes, lo que hace necesaria la normalización de los datos.
- Se observaron valores atípicos asociados a registros de gran escala.

- El comportamiento del salario sugiere relaciones no lineales con algunas variables, justificando el uso de redes neuronales.
- Estos hallazgos fundamentaron las decisiones tomadas en las fases posteriores del proyecto.

Preparación de los datos

En esta fase se realizaron procesos de limpieza y transformación, entre los que se incluyen:

- Eliminación o imputación de valores faltantes.
- Codificación de variables categóricas.
- Normalización de las variables numéricas.
- Separación del conjunto de datos en entrenamiento y prueba.

Adicionalmente, se identificó que la variable objetivo (salario mensual promedio) presentaba una magnitud considerablemente mayor en comparación con las variables de entrada.

6. Modelo Seleccionado

Dado el carácter no lineal de las relaciones identificadas durante el análisis exploratorio de los datos, se seleccionó una red neuronal como modelo principal del proyecto. Este tipo de modelo resulta adecuado para capturar patrones complejos y relaciones no lineales entre las variables, las cuales difícilmente pueden ser representadas de forma precisa mediante modelos tradicionales.

Además, las redes neuronales ofrecen una alta capacidad de adaptación a conjuntos de datos con múltiples variables y permiten modelar interacciones complejas entre ellas. Por estas razones, la elección de este modelo se consideró coherente con el objetivo del proyecto y con las características del conjunto de datos analizado.

7. Modelado – Red neuronal (primera iteración)

Se construyó una primera versión de una red neuronal con una arquitectura básica, con el objetivo de establecer una línea base de desempeño.

El modelo fue entrenado utilizando las variables de entrada normalizadas y se evaluó mediante las métricas RMSE y MAE.

Sin embargo, los resultados iniciales evidenciaron errores elevados, lo cual indicó dificultades del modelo para aprender adecuadamente debido a la escala de la variable objetivo.

8. Optimización del modelo

A partir de los resultados obtenidos en la primera iteración, se implementó una optimización clara y justificada, que incluyó los siguientes ajustes:

Qué ajustes se realizaron

- Normalización de la variable objetivo (salario mensual promedio).
- Incremento en el número de neuronas.
- Aumento del número de épocas de entrenamiento.

Cómo se implementaron

- Se aplicó un proceso de escalado adicional a la variable objetivo utilizando técnicas de normalización.
- Se ajustó la arquitectura de la red neuronal para mejorar su capacidad de aprendizaje.
- Se entrenó nuevamente el modelo bajo estas condiciones.

Por qué estos ajustes mejoraron el modelo

La normalización de la variable objetivo permitió estabilizar el proceso de entrenamiento, reduciendo la magnitud de los errores y facilitando que la red neuronal identificara patrones relevantes en los datos.

9. Evaluación y comparación de resultados

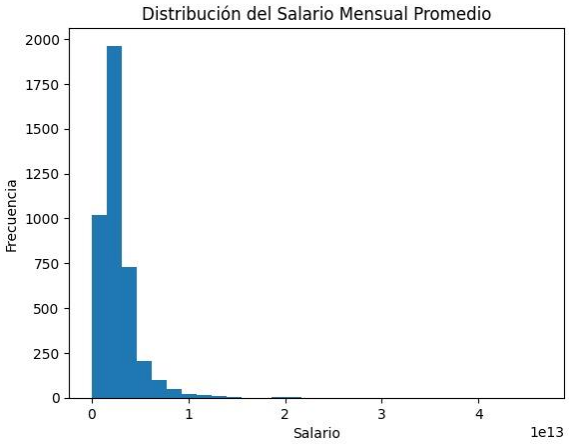
Para evaluar el desempeño de los modelos se utilizaron las métricas RMSE y MAE.

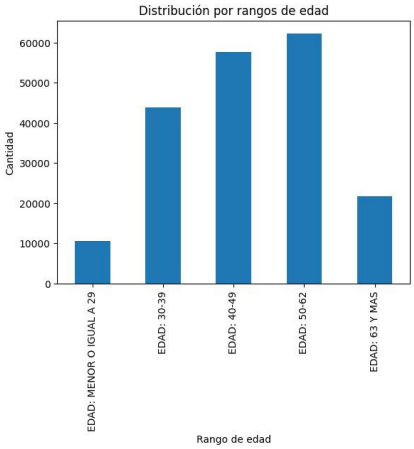
La comparación entre la versión inicial y la versión optimizada evidenció una reducción significativa de los errores, lo que demuestra que los ajustes implementados mejoraron considerablemente el desempeño del modelo.

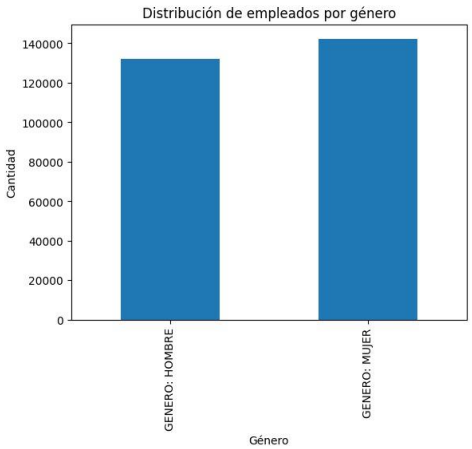
Este proceso de optimización quedó documentado mediante una comparación directa de métricas, cumpliendo con los requerimientos establecidos.

10. Conclusiones de análisis de dato

- El salario mensual promedio presenta una alta variabilidad entre registros.
- Se identificaron variables categóricas que requerían codificación previa al modelado.
- Existen valores faltantes que debían ser tratados para evitar sesgos.
- Algunas variables presentan escalas muy diferentes, lo que justifica la normalización.
- Se detectaron posibles valores atípicos asociados a registros agregados de gran escala.
- El comportamiento del salario sugiere una relación no lineal con varias variables, lo que motivó el uso de redes neuronales.







11. Conclusiones

- La metodología ASUM-DM permitió estructurar de manera ordenada el proceso de análisis.
- La normalización de los datos fue un factor clave para el éxito del modelo.
- Las redes neuronales resultaron adecuadas para modelar relaciones complejas en el empleo público.
- El proceso de optimización mejoró significativamente el desempeño predictivo.
- El proyecto evidencia la importancia de la iteración en analítica de datos.