Universidad Tecnológica de Panamá

Facultad de Ingeniería de Sistemas Computacionales

Facultad de Ingeniería Industrial

Maestría en Analítica de Datos

Materia

Modelos Predictivos

Proyecto Final

Predicción de Salarios de Carreras Afines a las Ciencias de Datos

Profesor

Juan Marcos Castillo, PhD

Estudiante

Gabriel Ah Chu

Cédula

8-791-250

Fecha

9 de abril de 2025

Índice

Introducción

El presente documento tiene como objetivo presentar el detalle del Proyecto Final de la materia Modelos Predictivos llamado “Predicción de Salarios de Carreras Afines a las Ciencias de Datos”. Este proyecto tiene como propósito principal aplicar técnicas avanzadas de análisis predictivo con la finalidad de estimar los salarios esperados en distintas posiciones relacionadas con las ciencias de datos. Además, busca ofrecer un marco referencial útil y orientador para la toma de decisiones profesionales entre los estudiantes.

Justificación

El análisis realizado tiene como propósito informar y orientar a los estudiantes de la clase acerca de los rangos salariales típicos que podrían obtener en posiciones relacionadas con las ciencias de datos. La relevancia del estudio radica en proporcionar herramientas objetivas y cuantitativas que faciliten una adecuada valoración del mercado laboral en este campo. Esto permitirá que los estudiantes puedan tomar decisiones fundamentadas acerca de su futuro profesional y reconocer adecuadamente el valor económico de sus conocimientos y habilidades en el contexto laboral actual.

Antecedentes

Actualmente, las profesiones relacionadas con las ciencias de datos han experimentado un crecimiento significativo a nivel global, impulsado principalmente por la necesidad creciente de las organizaciones de transformar grandes volúmenes de datos en conocimiento estratégico útil para la toma de decisiones. Este auge ha generado una demanda considerable por profesionales especializados, incentivando así la creación de programas académicos y formaciones técnicas enfocadas específicamente en preparar recursos humanos capacitados en técnicas de análisis avanzado, machine learning, inteligencia artificial y manejo de datos. Por ende, la valoración salarial de estos profesionales se ha convertido en un aspecto de gran interés para estudiantes y profesionales en esta área.

Definición del Problema

Si bien es evidente la alta demanda laboral existente para profesionales en áreas vinculadas a las ciencias de datos, surge la incertidumbre sobre el nivel salarial que un especialista debería esperar, particularmente en contextos locales como Panamá, donde aún no existen estudios suficientemente detallados. Por lo tanto, este proyecto busca responder a esta necesidad mediante la aplicación de métodos predictivos avanzados a una base de datos salarial proveniente de contextos internacionales, con la intención de proporcionar un referente comparativo válido. El objetivo final es ayudar a los estudiantes a entender mejor cuál podría ser su remuneración justa y

así tomar decisiones más informadas respecto a la aceptación o rechazo de futuras ofertas laborales, evitando la infravaloración de sus competencias profesionales.

Análisis Predictivo

Determinación de base de datos

La base de datos utilizada se obtuvo de Kaggle, bajo el título "The AI, ML, Data Science Salary (2020-2025)", y contiene información detallada sobre los salarios anuales de profesionales que trabajan en áreas relacionadas con la ciencia de datos y la inteligencia artificial.

Este conjunto de datos cuenta con 88,584 registros y 11 columnas, abarcando información desde el año 2020 hasta el primer trimestre de 2025. Cada registro representa el salario anual de un profesional durante un año específico.

Se escogió esta base de datos porque incluye información relevante sobre salarios en distintos puestos vinculados con la ciencia de datos, lo que permite explorar tendencias, realizar estimaciones de remuneración futuras y comprender mejor la dinámica del mercado laboral en este campo.

Una de las principales ventajas de este conjunto de datos es que ya está preprocesado, por lo que no presenta valores nulos y cuenta con una transformación monetaria que facilita la comparación justa e imparcial de los salarios.

Preprocesamiento y Limpieza

El dataset viene bastante procesado y requiere poca limpieza. No contiene valores nulos, existe algunos valores atípicos que luego de analizarlos se decide incluirlos en el análisis.

Para un mejor entendimiento del dataset, se transforman las categorías de las siguientes columnas:

experience_level	
Categoría Actual	**Nueva Categoría**
EN	Junior
MI	Intermedio
SE	Experto
EX	Director

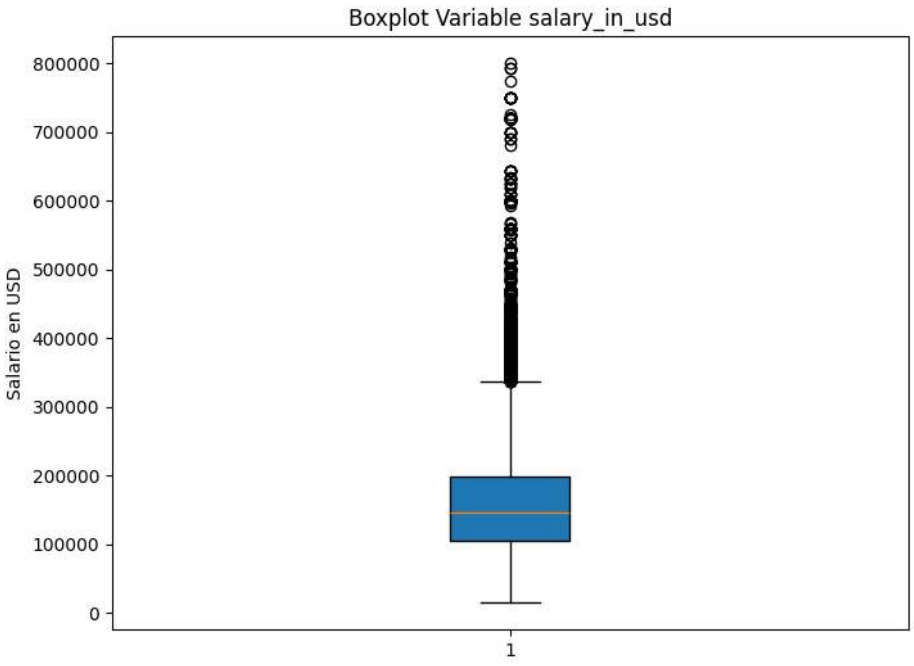
employment_type	
Categoría Actual	**Nueva Categoría**
PT	Tiempo Parcial
FT	Tiempo Completo
CT	Contrato
FL	Freelance

remote_work_ratio	
Categoría Actual	**Nueva Categoría**
0	En Sitio
50	Híbrido
100	Remoto

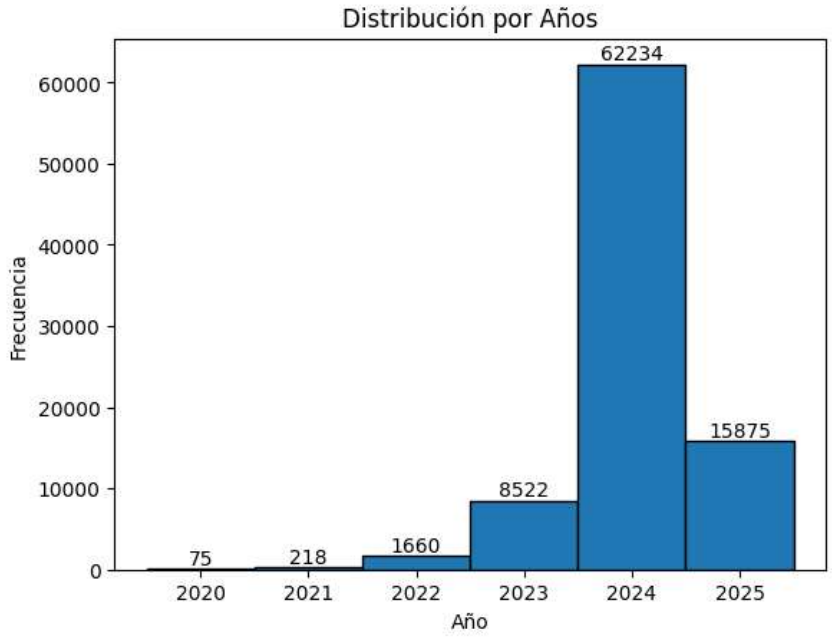
Luego de esto el dataset queda solo con 2 variables numéricas, work_year que representa el año del salario y salary_in_usd que representa el salario en dólares estadounidenses y es la columna que se busca predecir.

Análisis Descriptivo

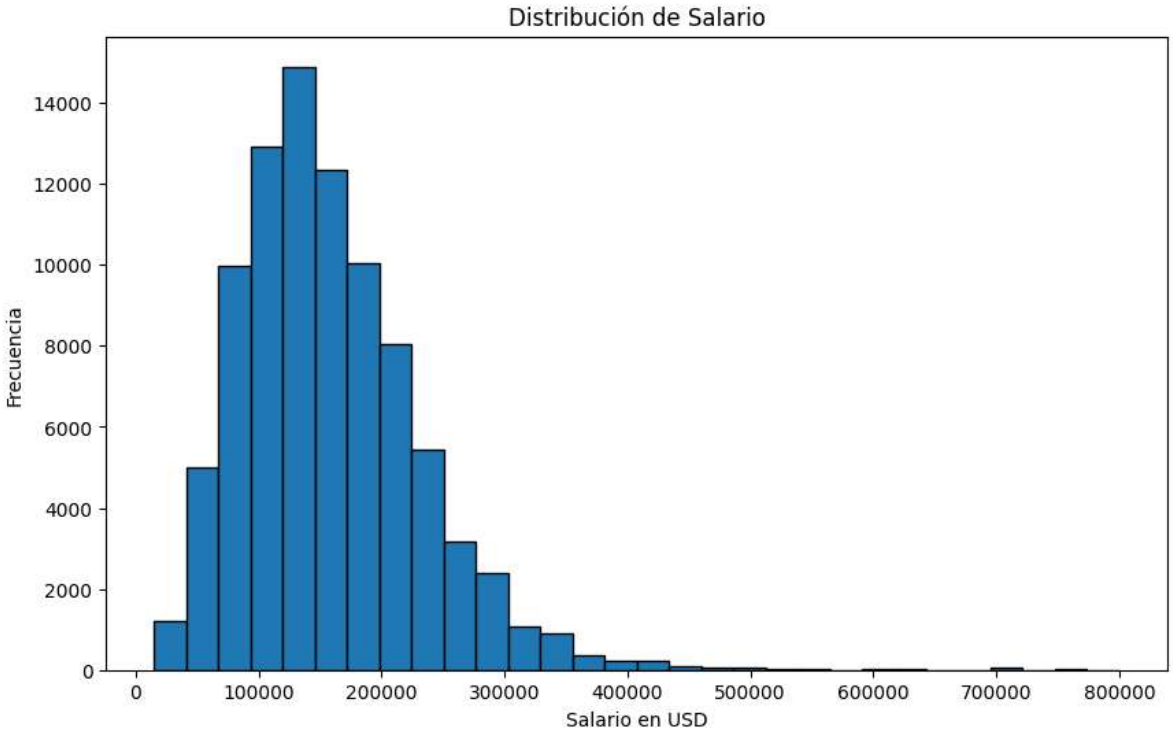
Se encuentran 1750 valores atípicos en la columna de salary_in_usd. Tras un análisis, se determinó mantener dichos valores en el conjunto de datos, puesto que reflejan puestos de trabajo con alta especialización o experiencia, y por tanto, resultan representativos de la realidad del mercado laboral para perfiles muy calificados.



Se realiza histograma para ver la distribución por años. Se nota claramente que en el 2024 hubo un incremento en la cantidad de empleos:

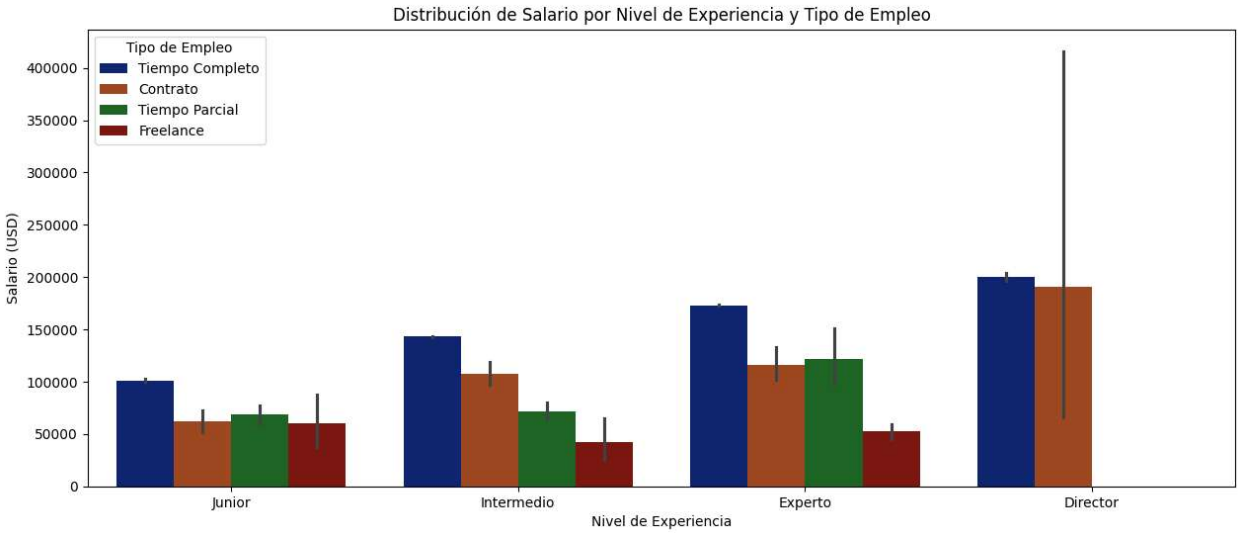


La distribución de salario incluyendo los valores atípicos muestran una asimetría positiva, ubicando la mayor cantidad de salarios alrededor de los 150 mil dólares:

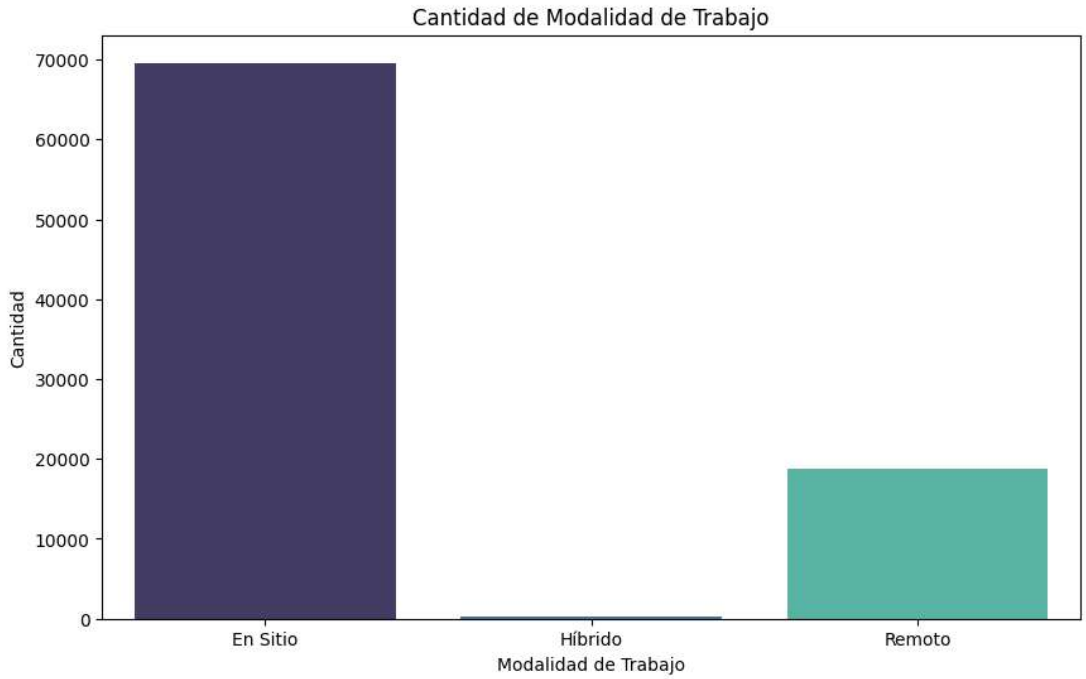


En la siguiente gráfica se aprecia cómo varían los salarios según el nivel de experiencia (Junior, Intermedio, Experto, Director) y el tipo de empleo (Tiempo Completo, Contrato, Tiempo Parcial y Freelance). Los empleos de Tiempo Completo registran los salarios promedio más altos, y se observa una tendencia al incremento salarial conforme aumenta el nivel de experiencia.

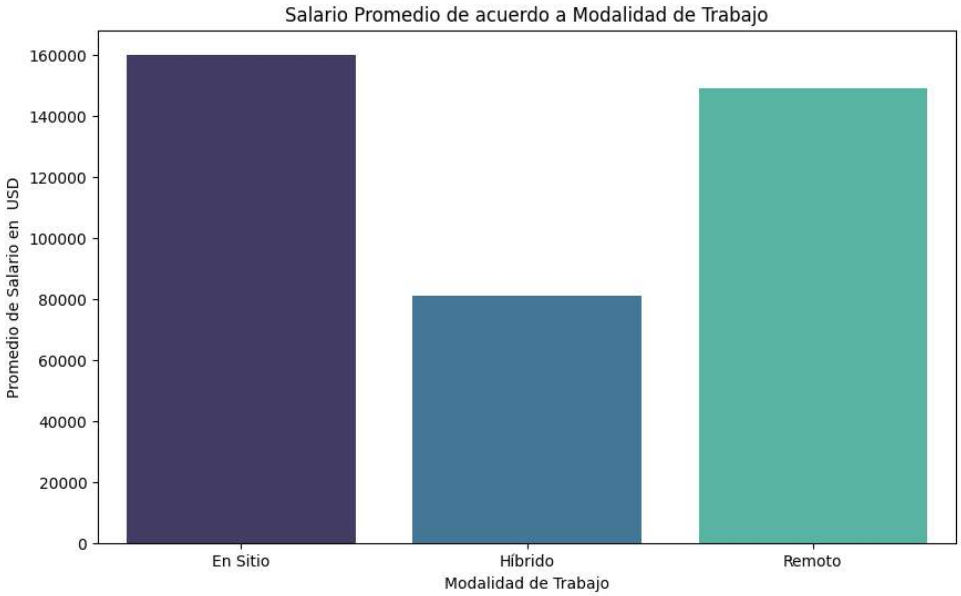
Además, destaca la categoría Director bajo contrato, donde se evidencia mayor variabilidad en los salarios, lo que podría deberse a la diversidad de roles y responsabilidades asociadas a este tipo de contrataciones.



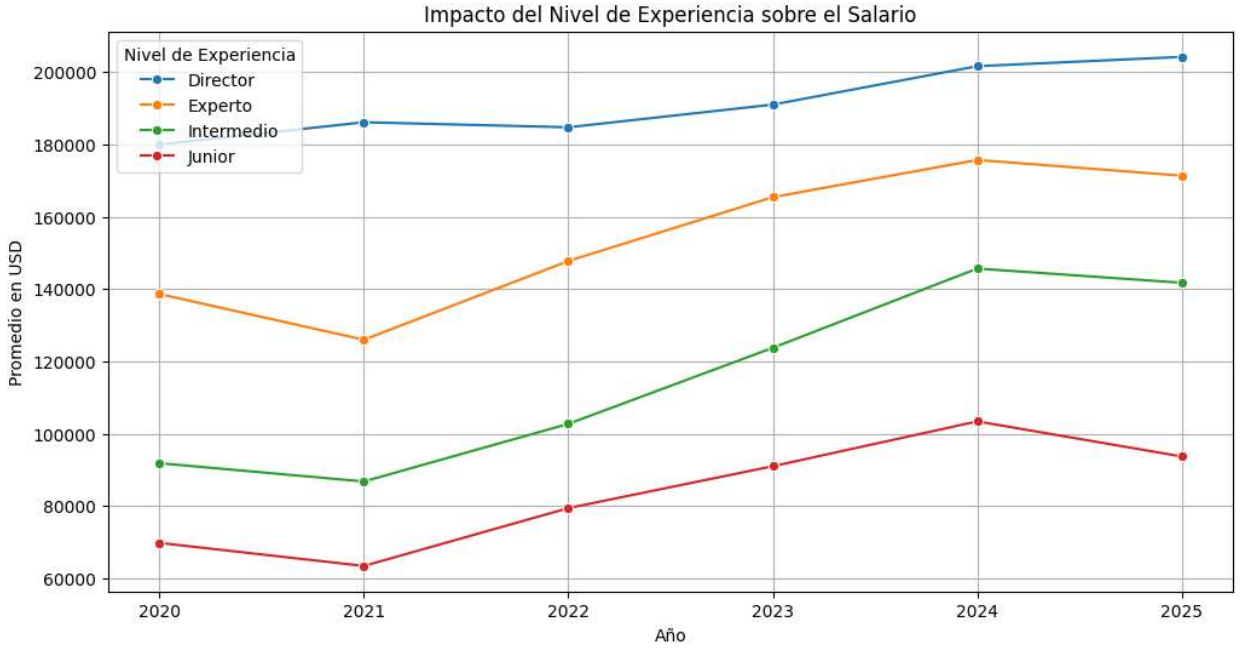
La modalidad de trabajo más popular es en sitio, seguida por trabajo remoto. La modalidad híbrida es muy rara en este dataset.



El salario promedio de acuerdo con la modalidad de trabajo es mayor para “En Sitio” que para el resto de las otras categorías.



La siguiente gráfica ilustra cómo varían los salarios promedio según el nivel de experiencia (Junior, Intermedio, Experto y Director) a lo largo de los años 2020 a 2025. Se observa un incremento generalizado en los salarios para todos los niveles conforme avanza el tiempo. Se muestra que los incrementos conservan su orden jerárquico. Esto sugiere que la experiencia sigue siendo un factor determinante para la evolución salarial a lo largo de los años.



Selección de Variables

Se realiza ANOVA con hipótesis nula de que los niveles de experiencia no influyen en el salario.

ANOVA (Salario vs Nivel de Experiencia				
	sum_sq	df	F	PR(>F)
experience_level	4.91E+13	3	3372.699	0
Residual	4.30E+14	88580		

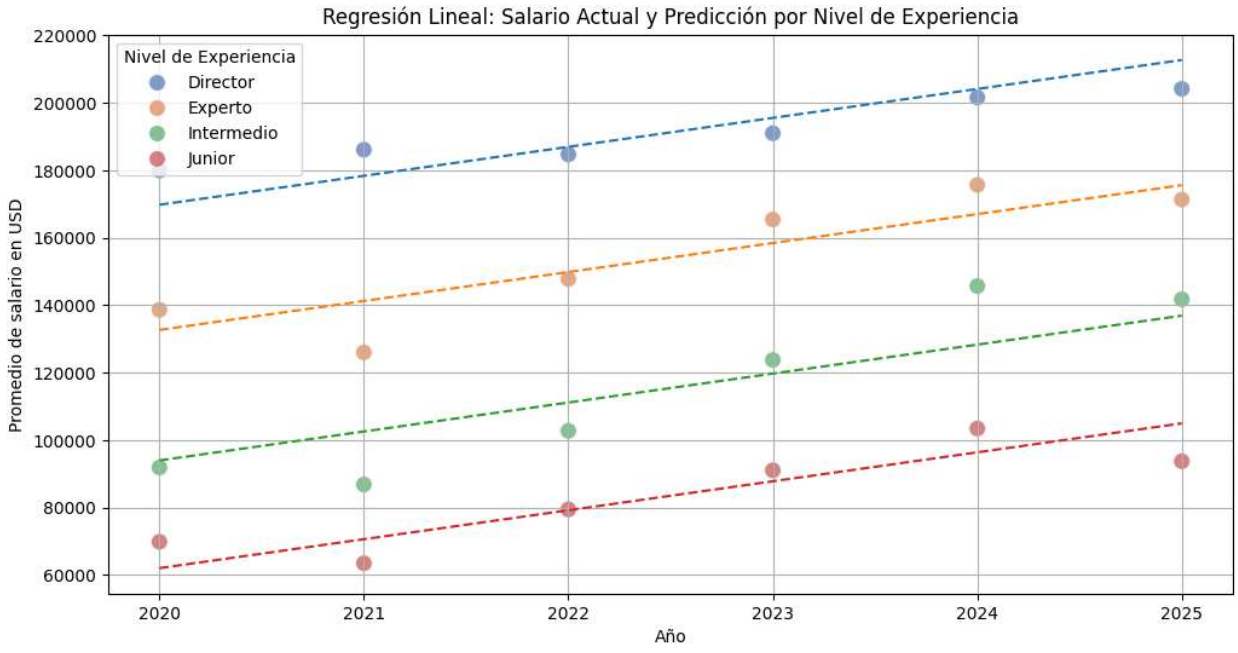
Se rechaza la hipótesis nula por tener un p-value bastante bajo, se determina que el nivel de experiencia sí afecta el salario.

Se utilizarán todas las variables para el modelo predictivo.

Selección de Modelos

Regresión Lineal

El primer modelo que se realiza es la Regresión Lineal, para esto se calcula el promedio de los salarios por nivel de experiencia por año y se realiza la regresión. Esto da por resultado lo siguiente:



MSE: 68610789.28

R^2: 0.9644

MAD: 7011.73

MAPE: 5.86%

Sin embargo, esta aproximación tiene desventajas ya que se pierde granularidad al trabajar con promedios y se omiten factores relevantes como el tamaño de la compañía, cargo específico, etc.

Regresión Lineal Múltiple

Al correr regression lineal con múltiples variables, se determina se consiguen niveles poco confiables en la evaluación:

R^2: 0.1797

MSE: 4,444,640,125.3912

MAD: 42,600.3456

MAPE: 34.2179%

Se considera que el tipo de datos es no lineal, por lo tanto, requiere de modelos más avanzados.

Random Forest Regression

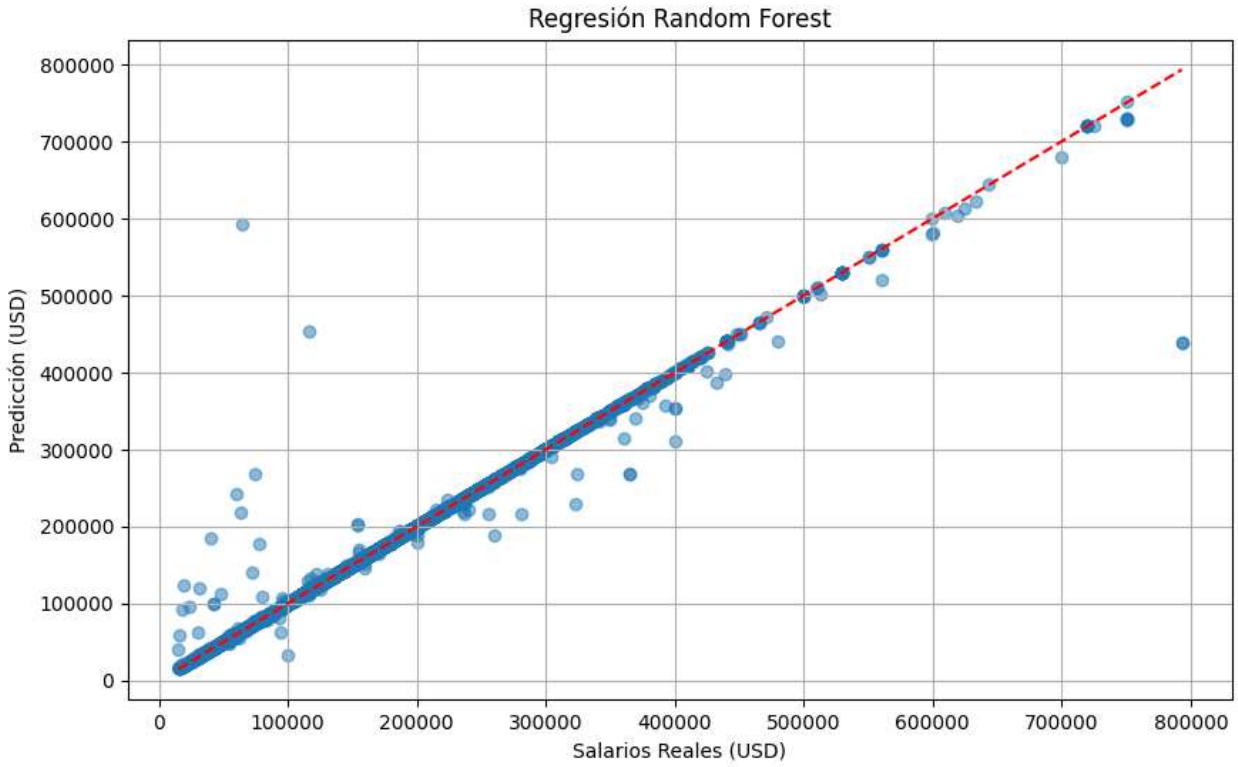
Random Forest es un método de aprendizaje que consiste en entrenar múltiples árboles de decisión de manera independiente para realizar predicciones (clasificación o regresión). El resultado final se obtiene combinando las predicciones de todos los árboles (por promedio o votación). Esta "combinación" reduce la varianza del modelo y mitiga el sobreajuste típico de un solo árbol de decisión, logrando así una mejor capacidad de generalización. Es especialmente robusto ante datos con valores atípicos y puede manejar conjuntos de datos de alta dimensión de manera efectiva. Se utilizaron todas las variables disponibles y el modelo fue capaz de generar buenos resultados:

R^2: 0.9906

MSE: 50,877,351.3014

MAD: 297.9034

MAPE: 0.3648%



XGBoost Regression

El XGBoost (eXtreme Gradient Boosting) es un método de Machine Learning basado en la técnica de gradient boosting que construye múltiples árboles de decisión de manera secuencial. A diferencia de otros algoritmos similares, XGBoost incorpora optimizaciones a nivel de software y hardware para operar de forma extremadamente eficiente en grandes conjuntos de datos, además de ofrecer varios hiperparámetros que permiten controlar con detalle el proceso de entrenamiento, evitando sobreajuste (overfitting) y maximizando el rendimiento del modelo.

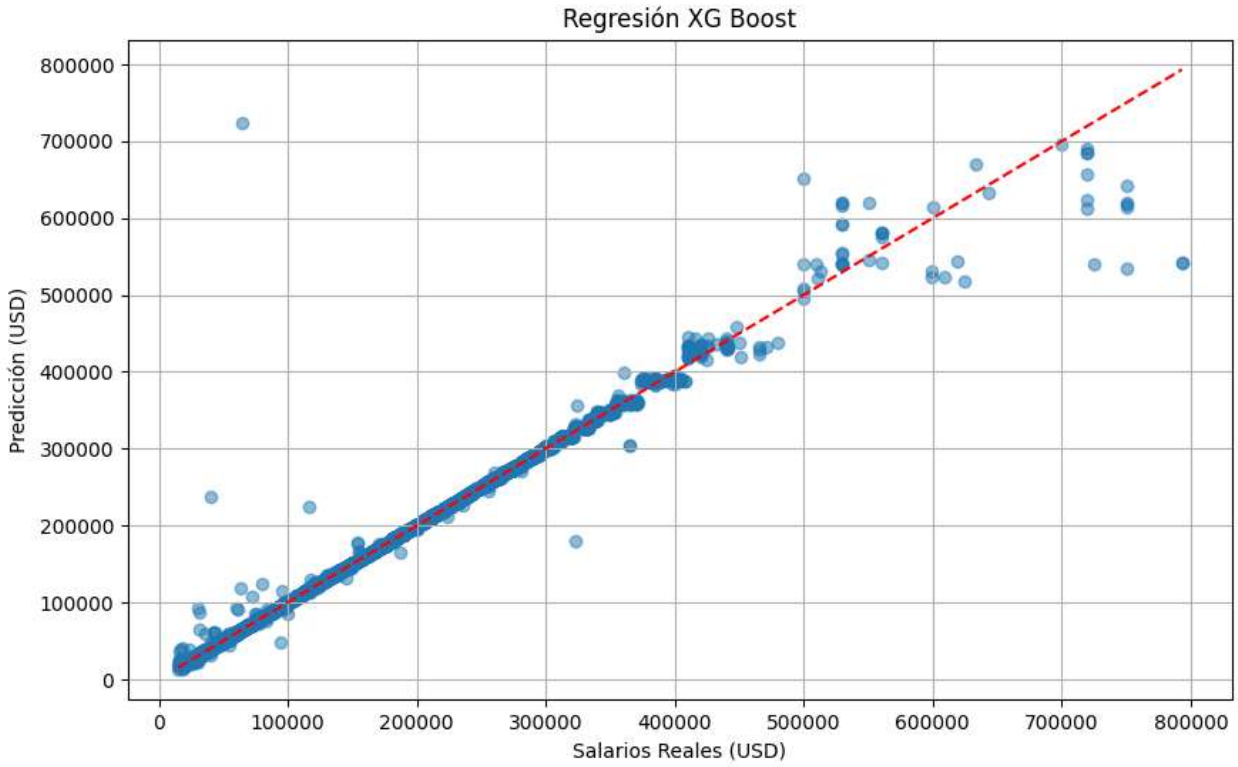
En nuestro caso utilizamos 100 árboles de decisión para llegar al modelo:

R^2: 0.9898

MSE: 55,386,132.0000

MAD: 826.7100

MAPE: 0.6326%



Conclusiones

El conjunto de datos presenta un número significativo de variables categóricas, por lo que fue necesario recurrir a distintos modelos de Machine Learning para capturar de manera adecuada la complejidad y relaciones subyacentes. Para ello, se realizó la codificación de cada variable categórica, facilitando su tratamiento numérico y permitiendo evaluar matemáticamente su influencia en el modelo predictivo.

Tras el entrenamiento y validación de los modelos, se observaron los siguientes resultados:

Evaluación de Modelos Predictivos de Machine Learning				
	MSE	**R^2**	**MAD**	**MAPE**
Regresión Lineal (promedios anuales)	68610789.28	0.9644	7011.73	5.86%
Regresión Lineal	4,444,640,125.39	0.1797	42,600.35	34.22%
Random Forest Regression	50,877,351.30	0.9906	297.9034	0.36%
XG Boost Regression	55,386,132.00	0.9898	826.71	0.63%

1. Random Forest Regression se ubica como el mejor modelo, con un R^2=0.9906, un MSE relativamente bajo y una MAPE de tan solo 0.36%. Esto indica su gran capacidad para manejar múltiples variables y capturar relaciones no lineales, ofreciendo una excelente precisión en la predicción de salarios.
2. XGBoost Regression también muestra un alto desempeño (R^2=0.9898), evidenciando su eficacia como modelo de conjunto y su habilidad para manejar datos complejos.
3. Regresión Lineal (basada en promedios anuales), si bien exhibe un desempeño aceptable (R^2=0.9644), su flexibilidad es limitada al no captar del todo la complejidad inherente al conjunto de datos.
4. Regresión Lineal múltiple resultó ser el modelo con el peor rendimiento (R^2=0.1797), lo que era previsible dado que los datos carecen de una relación netamente lineal y presentan características categóricas múltiples que dificultan su ajuste.

Los resultados muestran que los métodos de ensemble (Random Forest y XGBoost) se adaptan mejor a la naturaleza compleja y diversa de los datos de salarios, gracias a su habilidad para manejar múltiples variables categóricas y capturar patrones no lineales. Por el contrario, los métodos lineales, especialmente en su versión múltiple sin agregación, se ven limitados cuando las relaciones entre variables no responden a la linealidad.

Recomendaciones y futuros estudios

Análisis y Selección de Variables

Se sugiere realizar un análisis exhaustivo de relevancia de características para identificar y descartar aquellas variables que aporten poca o ninguna información. Con esto, se busca reducir la complejidad del modelo y optimizar el rendimiento, tanto en términos de precisión como de eficiencia computacional.

Agrupación de Categorías

Dada la gran cantidad de variables categóricas, es recomendable agrupar o combinar categorías que sean similares. Al disminuir el número de categorías, se evita la explosión dimensional tras la codificación y se reduce la carga de memoria, lo que puede hacer que el procesamiento sea más ágil y requiera menos recursos de hardware.

Uso de la Nube y Consideraciones de Seguridad

Una alternativa para manejar grandes volúmenes de datos y modelos complejos es la ejecución en plataformas en la nube, las cuales ofrecen escalabilidad y flexibilidad. Sin embargo, se deben ponderar aspectos de costo y las implicaciones de seguridad, especialmente en situaciones donde los datos son confidenciales o se encuentran sujetos a normativas de protección de la información.

Líneas de Investigación Futuras

- Optimización de Modelos: Profundizar en técnicas de feature engineering, así como en algoritmos de selección y reducción dimensional para mejorar la interpretabilidad y el rendimiento de los modelos.
- Nuevos Modelos de Ensamble y Deep Learning: Explorar otros métodos, como Gradient Boosting o redes neuronales, para comparar su desempeño y robustez frente a los datos disponibles.
- Integración de Factores Externos: Incluir variables económicas (por ejemplo, inflación o costo de vida según la región) para enriquecer el análisis y obtener una visión más completa de los factores que influyen en los salarios.

Bibliografía

Pattern Classification – Richard O.Duda, Peter E. Hart, David G. Stork

The AI, ML, Data Science Salary (2020- 2025)

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

https://xgboost.readthedocs.io/en/release_3.0.0/

Anexos

https://github.com/egabach/m_predictivos