



**Tecnológico
de Monterrey**

Instituto Tecnológico y de Estudios Superiores de Monterrey

Analítica de datos y herramientas de inteligencia artificial II

Actividad 6

Regresión Lineal Múltiple y No Lineal

Integrantes:

Rodrigo Ruiz Teodoro- A01730322
Natalia Cedillo Hernández - A01660022
Elena Nivón Hernández - A01174666
Jarlyn Loza Pacheco - A0176943
José Jaime Ponce de León - A01552256

Profesores:

Rigoberto Cerino Jiménez
Candy Yuridia Alemán Muñoz
Juan Manuel Ahuactzin
Alfredo García Suárez

Training Data Complete

El conjunto de datos contiene información sobre comportamiento histórico de los clientes en función de lo que han observado y proporcionados en el momento de la solicitud de un préstamo.

A continuación se describen de manera detallada las variables del conjunto de datos:

- **Income (int):** Ingresos del usuario.
- **Age (int):** Edad de los usuarios
- **Experience(int):** Años de experiencia profesional.
- **Profesión(string):** Profesión del usuario.
- **Married(string):** Casado o soltero.
- **house_ownership(string):** Dueño, renta o ninguna.
- **car_ownership(string):** Dueño propio de un carro si o no.
- **current_job_years(int):** Años de experiencia en su trabajo actual.
- **current_house_years(int):** Años viviendo en su actual casa.
- **City(string):** Ciudad de residencia.
- **State(string):** Estado de residencia.
- **risk_flag(string):** Ha tenido amonestaciones en el pasado si o no.

Procesamiento de Nulos y Outliers

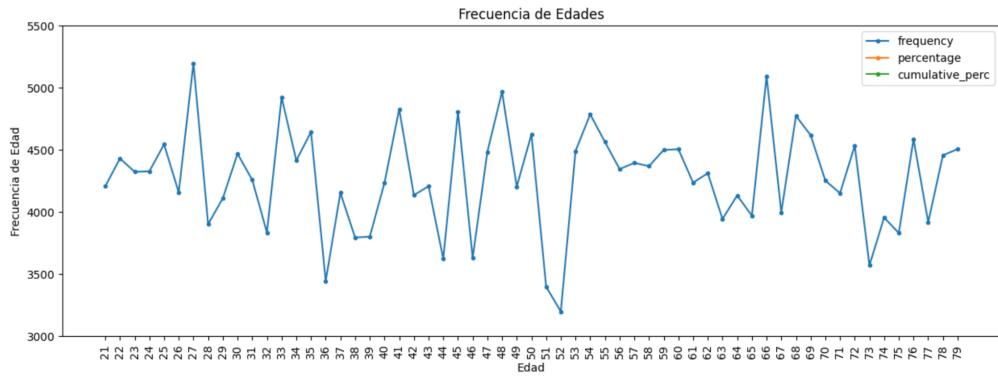
Para la parte de preprocesamiento de los datos se inició con la detección de valores nulos y outliers. Donde, en el caso de los valores nulos se aplicaron conteos de NAs por dataframe y con los outliers se aplicó la identificación de los mismos por rango intercuartílico. Para ambos casos no se identificaron valores de este tipo en la base de datos, lo que implica que el conjunto no requerían ningún proceso de limpieza o tratamiento adicional.

Análisis descriptivo

Dentro de la gráfica ubicada en la *Imagen 1* se logra observar la frecuencia de edades ordenadas de menor a mayor, donde se puede apreciar que la edad con menor frecuencia es

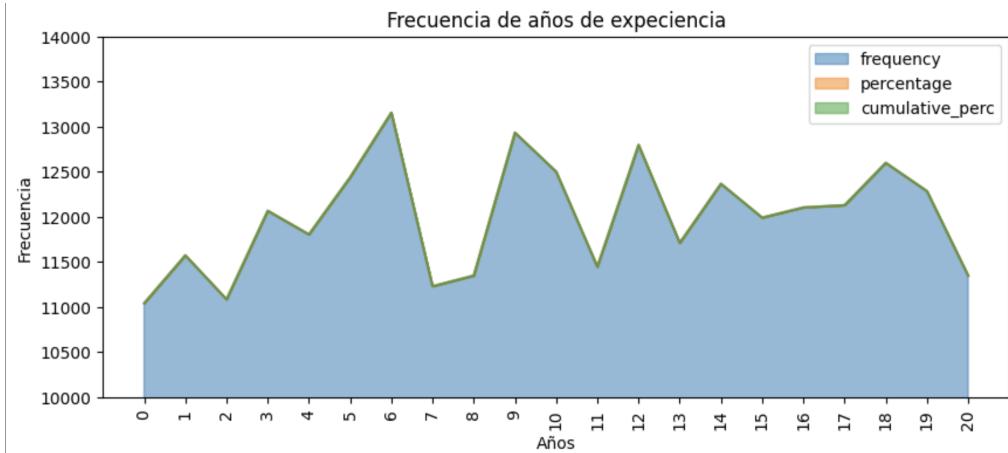
de 52 años, mientras que la edad con mayor frecuencia es 27. Así como muchos picos y variación a lo largo del rango de edades.

Imagen 1. Gráfico de frecuencia de edades



En la *imagen 2*, se logra comprender y analizar la frecuencia con la que se repiten los años de experiencia de trabajo de las personas de nuestra base de datos. Empezando por los años de experiencia que menos se repiten son 0, 2 y 7 mientras que las que más se frecuentan son 6, 9 y 12 años de experiencia.

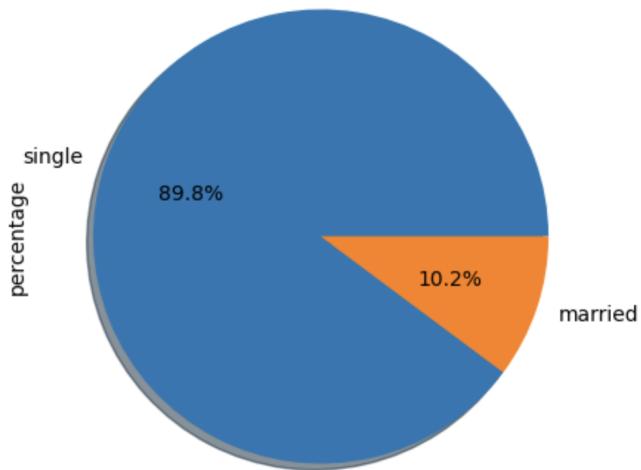
Imagen 2. Gráfico años de experiencia



En la *imagen 3* se muestra una gráfica de pastel donde se puede ver de manera clara que de las personas de nuestra base de datos, la mayoría con un 89.8% son personas solteras, mientras que el resto (10.2%) son personas casadas.

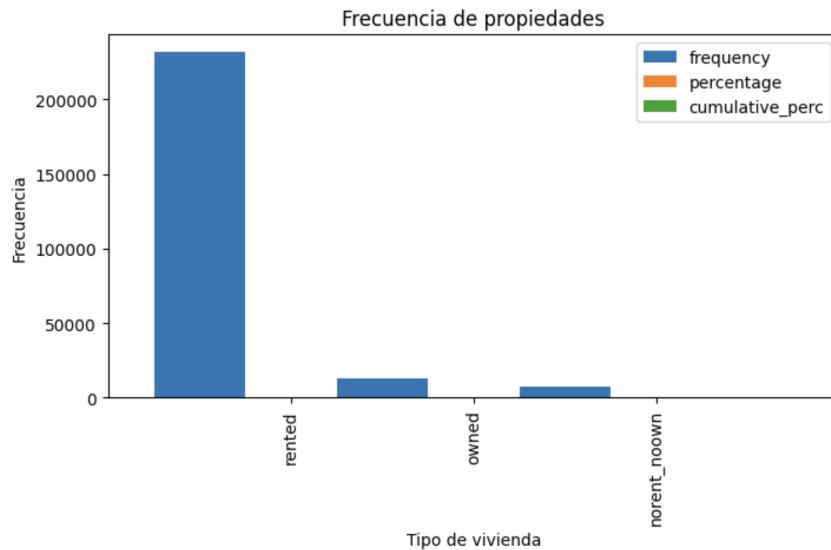
Imagen 3. Porcentaje Married / Single

Porcentaje Married / Single



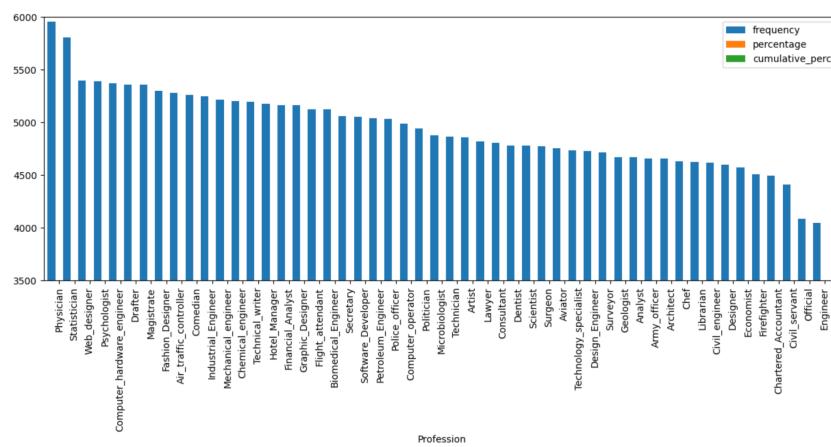
En el siguiente gráfico, es una gráfica de barras que nos muestra la cantidad de personas (frecuencia) que tienen casa propia, que rentan o ninguna de las dos anteriores. Como se puede observar en la *Imagen 4*, Se puede ver como la mayoría de las personas de nuestra base de datos rentan casa, superando los 200,000, mientras que la minoría se encuentra en casa propia o ninguna de las dos (rentada o propia).

Imagen 4. Gráfico de la frecuencia del tipo de vivienda



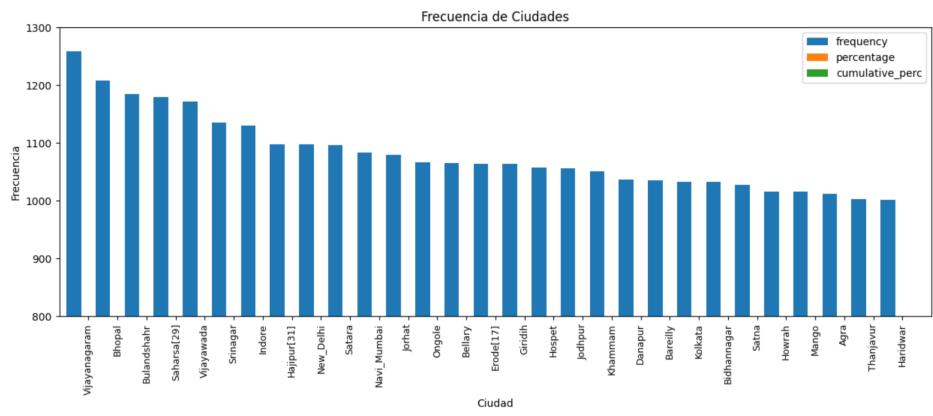
En la *Imagen 5* se muestra la frecuencia del tipo de profesiones entre los usuarios de la base de datos, en este caso podemos observar que los dos trabajos con mayor frecuencia son los fisioterapeutas, con una frecuencia mayor a la de 5500 usuarios. La otra profesión que tiene más cantidad de usuarios es la de estadístico, con una frecuencia mayor a 5500 pero sin llegar a la de fisioterapeutas.

Imagen 5. Gráfico de frecuencia en profesiones



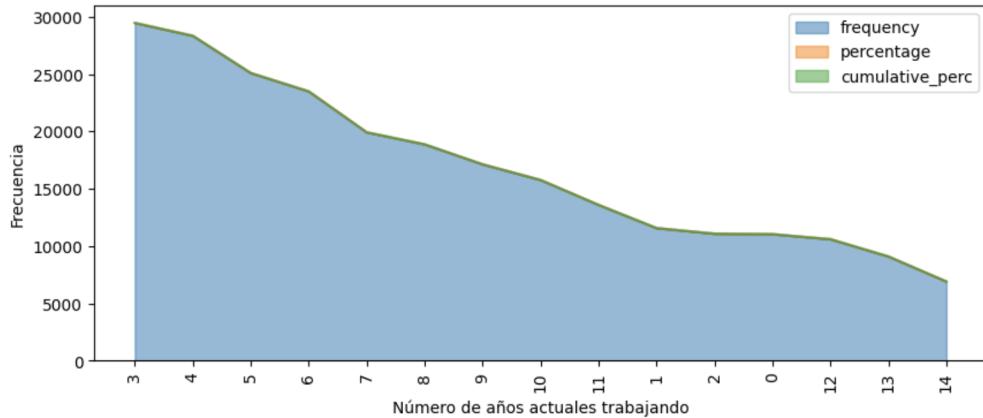
En la *Imagen 6* se puede observar que se trata de la cantidad de veces que se repiten las ciudades (frecuencia), sin embargo, para llegar a esta gráfica se hizo un filtro en esta columna para obtener los valores más relevantes y se mostraron solo las ciudades que repiten más de 1,000 veces. Se puede observar que las ciudades que más ocasiones se repiten son Vijayanagaram, Bhopal y Bulandshahr.

Imagen 6. Gráfico de frecuencia en ciudades



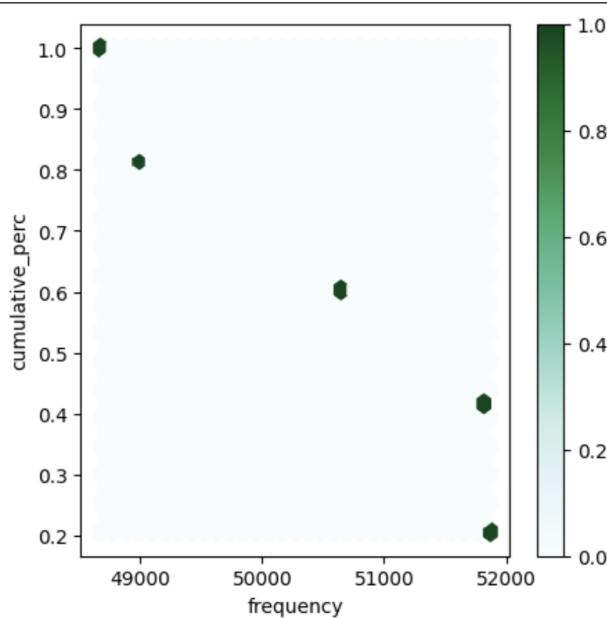
En la *Imagen 7* muestra una gráfica que nos demuestra la frecuencia con la que se repiten la cantidad de años activos de trabajo de las personas que se encuentran en nuestra base de datos. Se puede identificar que la cantidad de años de trabajo que más se frecuentan son 3,4 y 5 años, mientras que los años que menos se frecuentan son 12, 13 y 14 años.

Imagen 7. Gráfico de frecuencia vs años activos de trabajo



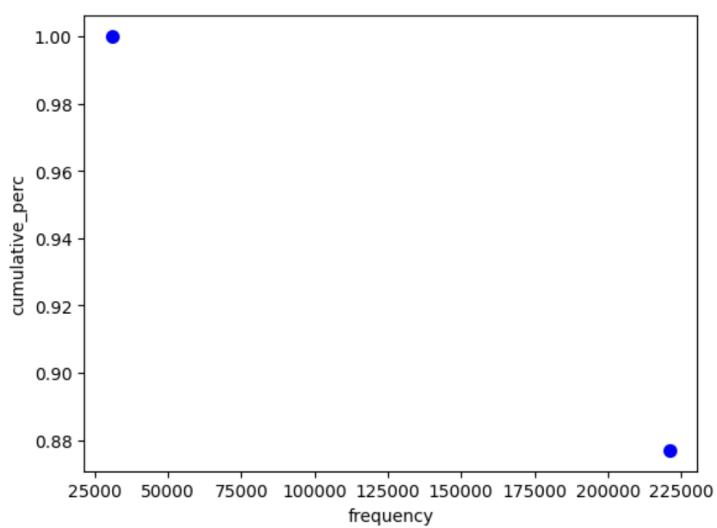
En la *Imagen 8* se puede observar cómo se distribuyen las observaciones en la columna de *años viviendo en su casa actual* en función de su frecuencia acumulada. En este tipo de gráfico, los datos se organizan en orden ascendente o descendente según su valor, y se representa la acumulación de observaciones a medida que se avanzan a lo largo de la escala de valores. Se muestra cómo se aumenta la frecuencia a medida que avanza a lo largo del eje horizontal.

Imagen 8. Gráfico de porcentaje acumulativo vs frecuencia



En el siguiente gráfico de porcentaje acumulativo (*Imagen 9*) se puede observar cómo se distribuye la frecuencia de la columna “Risk Flag” donde se trata sobre si las personas de nuestra base de datos tienen un incumplimiento de pago de sus deudas. En el gráfico se puede observar que el valor que más frecuencia tiene es el valor 0 (No tiene incumplimientos), representando cerca del 0.88 del total de los datos, mientras que el valor 1 es minoría, con solo él .12 restante del total.

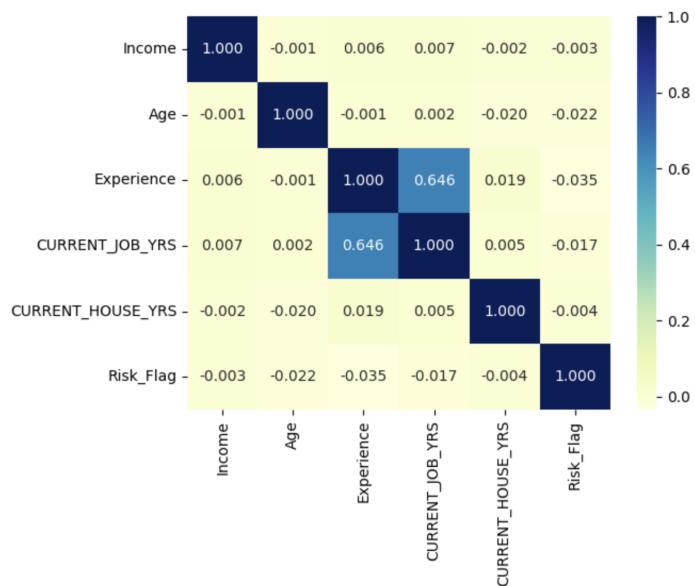
Imagen 9. Gráfico de porcentaje acumulativo vs frecuencia



Regresión lineal simple

El gráfico de calor con coeficientes de correlación (Imagen 10) nos ayuda a mostrar, de manera visual, la fuerza de correlación de las variables en el conjunto de datos. En este gráfico, se observa que las variables "Experience" y "Current Job Years" tienen una fuerte correlación positiva, con un coeficiente de correlación de 0.646, que es cercano a 1. Es por eso que se hace una modelo matemático donde se toma la variable independiente "Current_Job_Years" va a predecir por nuestra variable independiente que es "Experience"

Imagen 10. Gráfico de calor con coeficientes de correlación



Mejor modelo de regresión lineal:

Variable dependiente = 'Experience'

Variable independiente = ' Current_Job_Yrs'

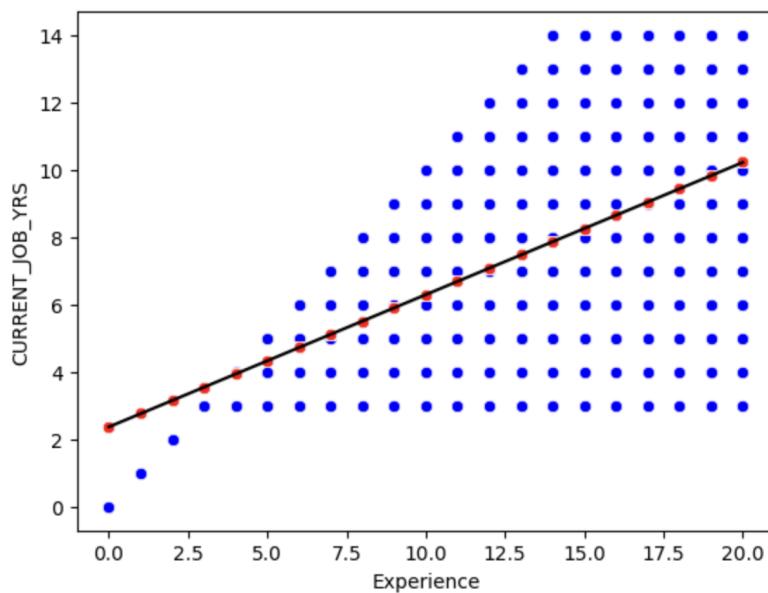
Modelo matemático:

$$0.39255587 x + 2.37517222$$

De acuerdo al análisis de regresión entre 'Curren_Job_Yrs' y 'Experience' revela que aproximadamente el 41.74% de la variabilidad en 'Experience' puede ser predecida por 'Curren_Job_Yrs', como lo indica el coeficiente de determinación (R^2). Existe una

correlación moderada positiva, con un coeficiente de correlación (r) de aproximadamente 0.6461, lo que significa que a medida que 'Curren_Job_Yrs' aumenta, 'Experience' tiende a aumentar, aunque la relación no es perfectamente lineal.

Imagen 11. Gráfica comparativa total real de 'Experience' y total predicción 'Experience' con la variable 'Curren_Job_Yrs'



Coeficiente de determinación es 0.4174420012200907

Coeficiente de correlación es 0.6460975168038419

Regresion lineal multiple

En la tabla que se muestra a continuación, se presentan las principales métricas y resultados del análisis de modelos de regresión lineal múltiple.

Variable dependiente	Variables independientes	Modelo matemático	Coeficiente de determinación	Coeficiente de correlación
Income	Age, Experience, Current job yrs, Current house yrs	y= -119.60460679x1+1 570.49372642x2,	6.2314950952213 62e-05	0.007893981 945267777

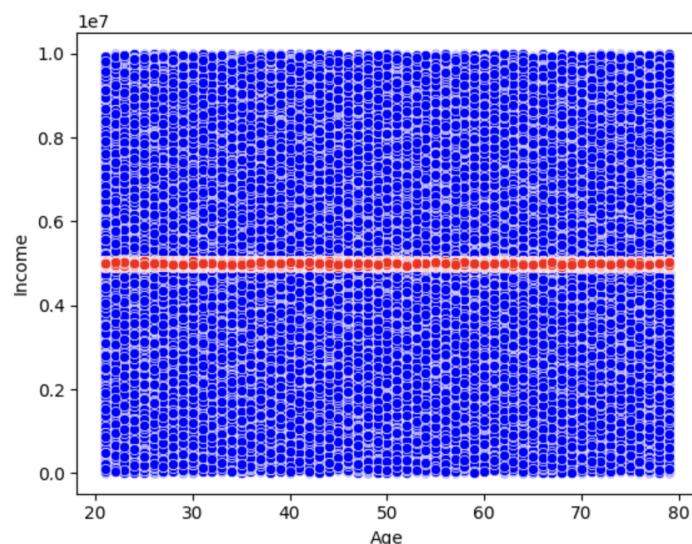
		3901.45155326x3 + -5146.0166579x4 + 5024283.389913392		
Age	Experience, Current job yrs, Current house yrs,Income	y= -119.60460679X1 + 1570.49372642X2 + 3901.45155326X3+ -5146.0166579X4 + 5024283.389913392	0.0004192541358 00462	0.020475696 222606497
Experience	Age, Current job yrs, Current house yrs,Income	y= -119.60460679x1 + 1570.49372642x2 + 3901.45155326x3 + -5146.0166579 x4 + 5024283.389913392	0.4177012934333 134	0.646298145 9305863
Current job yrs	Income,Age,Expe rience, Current house yrs	y= -119.60460679x1 + 1570.49372642x2 + 3901.45155326x3 + -5146.0166579 x4 + 5024283.389913392	0.4175082512923 082	0.646148784 1761433
Current house yrs	Income,Age, Experience, Current job yrs	y= -119.60460679x1 + 1570.49372642x2 + 3901.45155326x3 + -5146.0166579 x4 + 5024283.389913392	0.0008688150174 576137	0.029475668 227499332
Risk Flag	Income,Age, Experience,	y= -119.60460679x1 + 1570.49372642x2	0.0017448954344 837508	0.041771945 54343562

	Current job yrs, Current house yrs	$+ 3901.45155326x3$ $-5146.0166579x4 +$ 5024283.389913392		
--	---------------------------------------	---	--	--

El análisis de regresión entre 'Income' y 'Age' muestra que el coeficiente de determinación es extremadamente bajo, aproximadamente 0.0000623. Esto significa que prácticamente no hay capacidad del modelo para explicar la variabilidad en 'Income' basándose en 'Age'.

El coeficiente de correlación, es igualmente muy bajo, aproximadamente 0.0079, lo que sugiere una correlación extremadamente débil entre 'Income' y 'Age'.

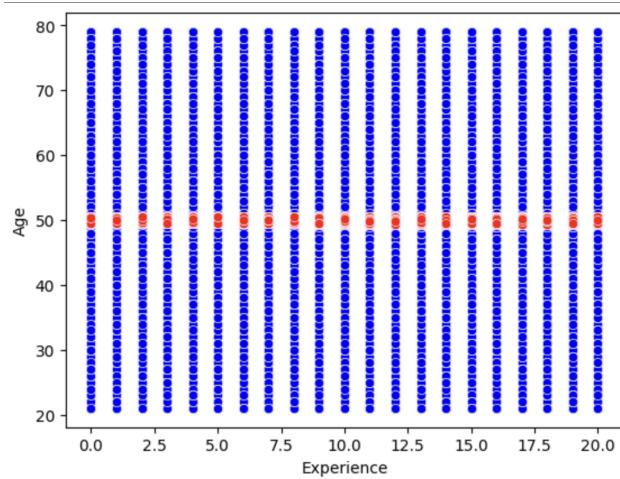
Imagen 12. Gráfica comparativa total real de 'Income' y total predicción 'Income' con la variable 'Age'



En la gráfica que se muestra en la Imagen 13, se presenta una representación visual que compara dos conjuntos de datos esenciales: "Total Real" y "Total Predicción" para la variable "Age" (Edad) en relación con la variable "Experience" (Experiencia). El objetivo principal de esta visualización es evaluar el rendimiento del modelo utilizado para predecir la edad en función de la experiencia laboral.

Se puede observar que existe una notable variación entre las predicciones (línea roja) y los valores reales (línea azul). Esta variabilidad sugiere que el modelo tiene dificultades para capturar y predecir con precisión las edades en función de la experiencia laboral.

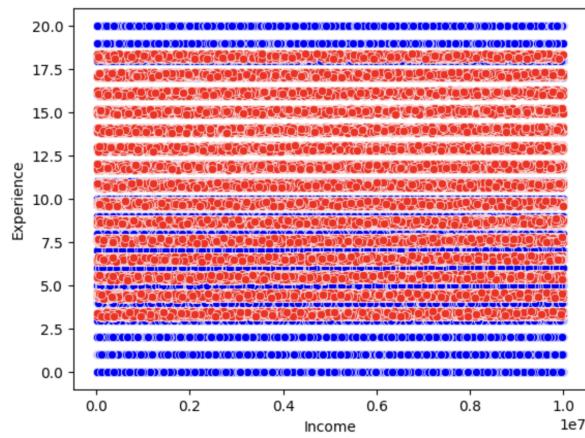
Imagen 13. Gráfica comparativa total real de 'Age' y total predicción 'Age' con la variable 'Experience'



Los coeficientes de determinación y de correlación indican la calidad del ajuste del modelo de regresión entre 'Experience' e 'Income'. Esto significa que alrededor del 41.77% de la variabilidad en 'Experience' puede ser explicada por 'Income'. Por otro lado, el coeficiente de correlación es aproximadamente 0.6463, lo que indica que hay una correlación moderada positiva entre 'Income' y 'Experience'.

Debido a que los valores de determinación y correlación son moderados y significativos, es probable que haya una relación útil entre 'Income' y 'Experience'.

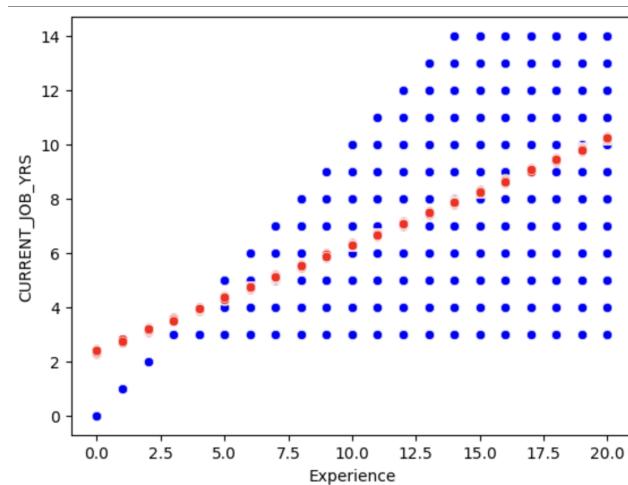
Imagen 14. Gráfica comparativa total real de 'Experience' y total predicción 'Experience' con la variable 'Income'



En esta gráfica (Imagen 15) que se incluye en el reporte, se presenta una visualización que compara dos conjuntos de datos clave: "Total Real" y "Total Predicción" para la variable "Current Job Yrs" (Años Actuales en el Trabajo), con respecto a la variable "Experience" (Experiencia). Esta visualización es resultado del uso de un modelo de regresión lineal múltiple para comprender la relación entre la experiencia laboral y la cantidad de años que las personas llevan en su trabajo actual.

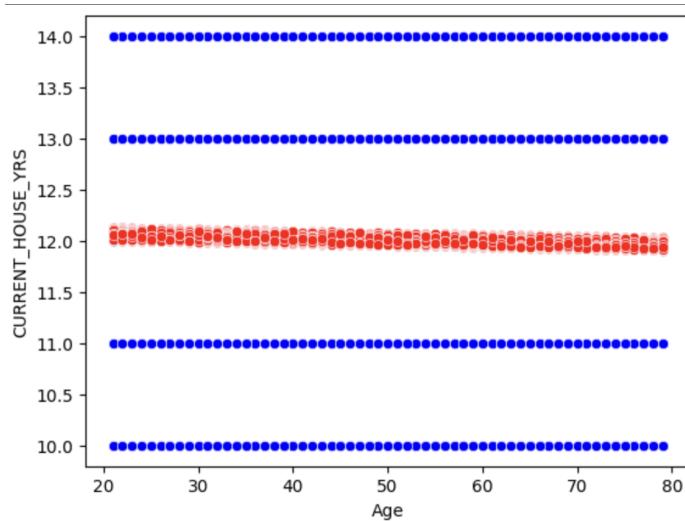
Se puede observar que la línea roja de predicción atraviesa por puntos que se encuentran dentro de la región azul que representa los datos reales. Este patrón es indicativo de una correlación positiva entre la variable "Experience" y "Current Job Yrs". El modelo de regresión lineal múltiple ha logrado capturar y reflejar la tendencia general que a medida que la experiencia laboral aumenta, también aumenta el tiempo que las personas permanecen en sus trabajos actuales.

Imagen 15. Gráfica comparativa total real de 'Current_Job_Yrs' y total predicción 'Current_Job_Yrs' con la variable 'Experience'



El análisis de regresión entre 'Current_House_Yrs' y 'Age' muestra que el coeficiente de determinación es extremadamente bajo, (aproximadamente 0.0008688). Esto indica que apenas el 0.08688% de la variabilidad en 'Current_House_Yrs' puede ser explicada por 'Age'. El coeficiente de correlación también es muy bajo, aproximadamente 0.0295. Esto indica que hay una correlación muy débil entre 'Current_House_Yrs' y 'Age', siendo que la relación entre estas variables es prácticamente inexistente en términos de una relación lineal.

Imagen 16. Gráfica comparativa total real de ‘Current_House_Yrs’ y total predicción ‘Current_House_Yrs’ con la variable ‘Age’

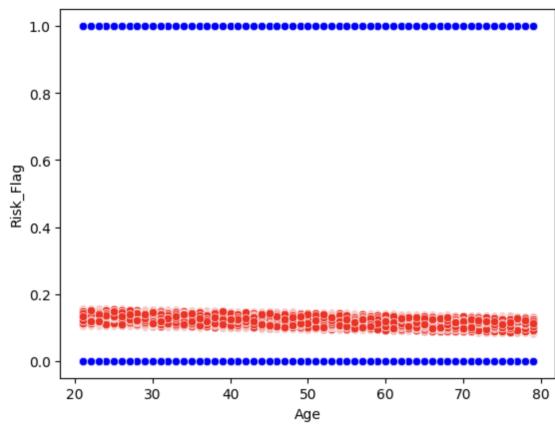


La gráfica que se muestra en la Imagen 17 que se presenta en el reporte compara dos conjuntos de datos: "Total Real" y "Total Predicción" para la variable 'Risk_Flag', en relación con la variable 'Age' (Edad). Esta visualización se basa en un modelo de regresión lineal múltiple y tiene como objetivo examinar la relación entre la edad y la probabilidad de que ocurra una Risk Flag en los datos.

Se puede observar que la línea roja de predicción no coincide con las líneas azules que representan los datos reales. Esto sugiere que el modelo de regresión lineal múltiple no logra visualizar una correlación fuerte entre la variable 'Age' y 'Risk_Flag'. Los puntos de datos de predicción y los datos reales no se alinean de manera significativa, lo que indica que el modelo no es efectivo para predecir “Risk Flag” basada únicamente en la edad.

Esta falta de coincidencia entre las predicciones y los datos reales se justifica por los bajos valores de coeficiente de determinación y correlación. Estos valores bajos indican que existe una correlación mínima entre la edad y la probabilidad de una “Risk_Flag”

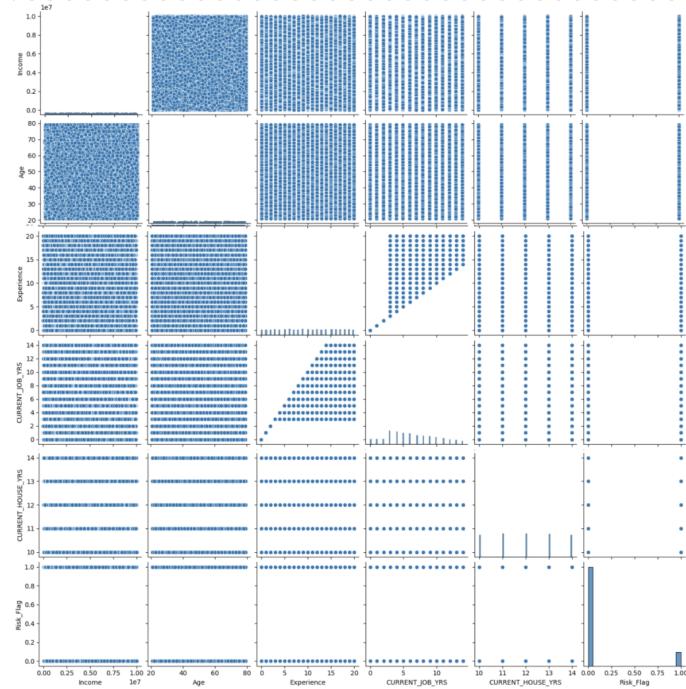
Imagen 17. Gráfica comparativa total real de ‘Risk_Flag’ y total predicción ‘Risk_Flag’ con la variable ‘Age’



Propuesta de modelo no lineal

A continuación, se presenta el gráfico de dispersión de todas las posibles combinaciones de variables numéricas. Estos gráficos ayudan a identificar tendencias, correlaciones y patrones en los datos, lo que puede ser fundamental para la toma de decisiones y el análisis de datos.

Imagen 18. Gráfico de dispersión de todas las posibles combinaciones de variables numéricas



A continuación, se muestra una tabla que indica que la experiencia laboral (Experience) y la edad (Age) están fuertemente correlacionadas, lo que sugiere que a medida que las personas envejecen,

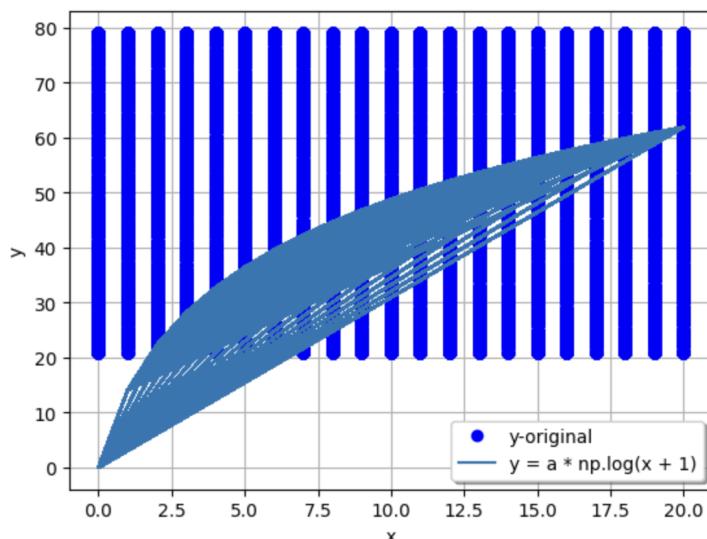
acumulan más experiencia de manera constante. Sin embargo, la relación entre los ingresos (Income) y la experiencia laboral es más compleja, ya que el modelo logarítmico utilizado no se ajusta bien a los datos. La permanencia en el trabajo actual (Current Job Years) muestra una relación positiva con la experiencia laboral y una correlación moderada. Los años en la vivienda actual (Current House Years) y la edad también están relacionados, aunque el modelo logarítmico no se ajusta adecuadamente a los datos. Por último, la columna (Risk Flag) tiene una correlación muy baja con la experiencia laboral, lo que sugiere que la experiencia no es un fuerte predictor del riesgo en este contexto. Estos insights resaltan la importancia de comprender las relaciones entre las variables y la necesidad de considerar modelos adecuados para el análisis de datos.

Variable dependiente	Variable independiente	Función y modelo empleado	Parámetros	r^2	Coeficiente de correlación
Income	Experience	Modelo Logarítmico a * np.log(x + 1)	a=2.03501779e +06 b=1.00000000 e+00 c=1.00000000e +00	-0.3383	0.581
Age	Experience	Modelo Logarítmico a * np.log(x + 1)	a=20.3199303 b= 1.0 c= 1.0	-0.9793	0.989
Experience	Curren_job_yrs	Modelo potencial a * (x**b)	a= 3.78864277 b= 0.56387016 c= 1.0	0.4498	0.670
Curren_job_yrs	Age	Valor absoluto a* x +b	a=0.39255587 b= 2.37517222	0.4174	0.64
Curren_hous_e_yrs	Age	Modelo Logarítmico a * np.log(x + 1)	a= 3.07344713 b= 1.0	-0.6896	0.830

Risk_flag	Experience	Modelo Logarítmico $a * np.log(x + 1)$	$a=0.04824648$	-0.0247	0.157
-----------	------------	---	----------------	---------	-------

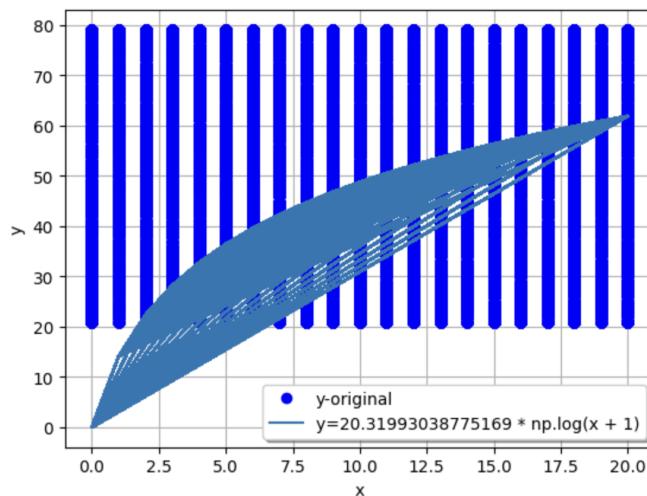
Ahora bien, el gráfico de "Income" vs la predicción basada en el modelo logarítmico de "Experience" (Imagen 19) muestra una correlación moderada entre la experiencia laboral y los ingresos reales, indicando que, en general, a medida que la experiencia aumenta, los ingresos tienden a aumentar. Sin embargo, el modelo logarítmico empleado no es adecuado para describir esta relación, ya que el coeficiente de determinación (r^2) es negativo, lo que sugiere que no se ajusta bien a los datos.

Imagen 19. Gráfico real vs predicción de 'Income'



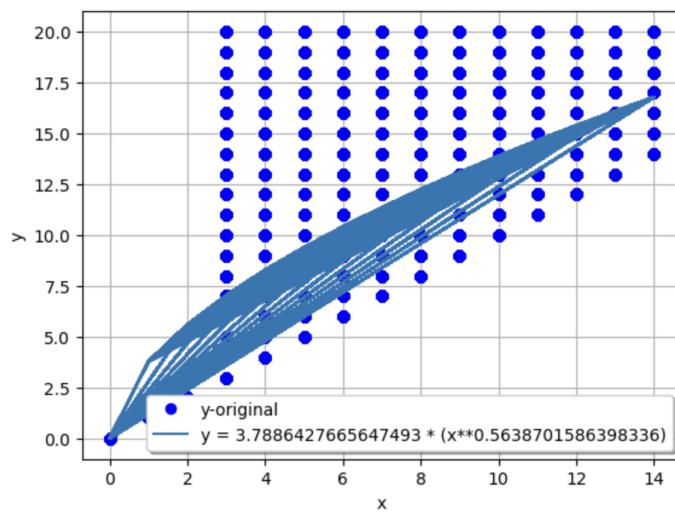
El gráfico de "Age" en comparación con la predicción basada en el modelo logarítmico de "Experience" (Imagen 20) revela una relación inversa muy fuerte y altamente significativa entre estas dos variables. Con un coeficiente de correlación de 0.9840 y un coeficiente de determinación (r^2) igualmente alto, la experiencia laboral se correlaciona de manera positiva y sólida con el aumento de la edad. Esto indica que la experiencia laboral es un indicador confiable y robusto de la edad, y el modelo logarítmico empleado capta de manera efectiva esta relación.

Imagen 20. Gráfico real vs predicción de 'Age'



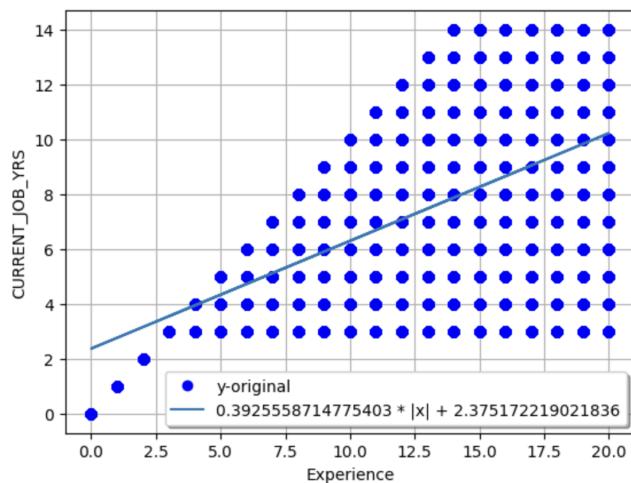
El gráfico de "Experience" contra la variable "Current Job Years" (Imagen 21) muestra una correlación positiva entre la experiencia y la duración en el trabajo actual. El coeficiente de correlación es moderado (0.6707), indicando que, en general, a medida que aumenta la experiencia laboral, los empleados tienden a permanecer más tiempo en sus trabajos actuales. El coeficiente de determinación (r^2) de 0.4499 sugiere que puede capturar la relación entre la experiencia y los años en el trabajo actual.

Imagen 21. Gráfico real vs predicción de 'Experience'



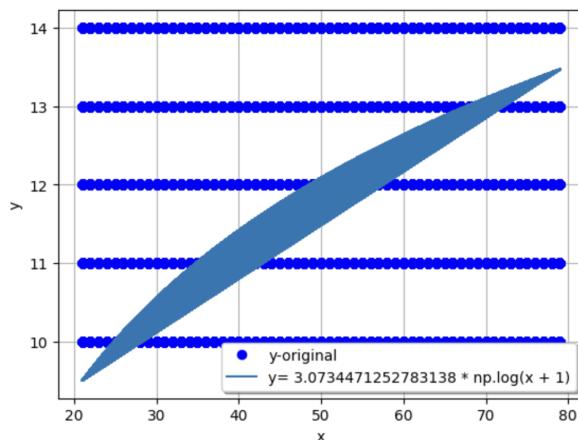
El gráfico de "Current Job Years" en comparación con la predicción basada en el modelo de valor absoluto de "Age" (Imagen 22) muestra una correlación moderada positiva entre la edad y los años en el trabajo actual. Con un coeficiente de correlación es 0.64, lo que indica que existe una relación, pero no es necesariamente lineal. El coeficiente de determinación (r^2) de 0.4174 indica que el modelo de valor absoluto se ajusta a los datos, aunque no de manera perfecta.

Imagen 22. Gráfico real vs predicción de 'Curren_job_yrs'



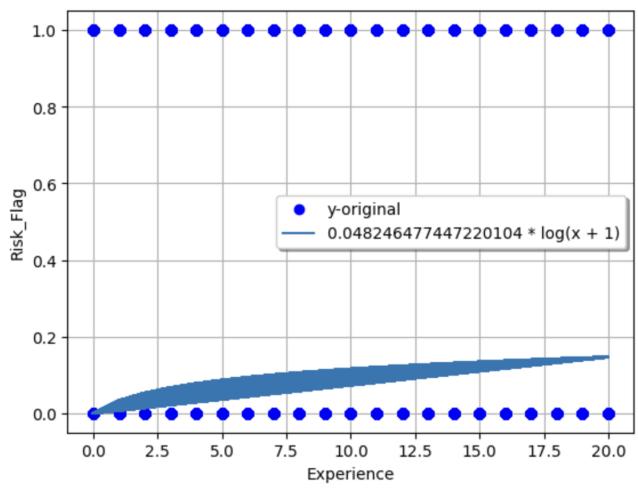
El siguiente gráfico de "Current House Years" realiza una predicción basada en la función logarítmica de "Age" (Imagen 22) en donde se obtuvo un coeficiente de correlación muy alto (0.8369), lo que indica una fuerte relación, el modelo logarítmico se ajusta muy bien a los datos. Esto sugiere que la edad es un indicador confiable de la duración en la vivienda actual, y que a medida que una persona envejece, es probable que pase menos tiempo en su vivienda actual.

Imagen 22. Gráfico real vs predicción de 'Curren_house_yrs'



El gráfico de "Risk_Flag" se hace una predicción basada en una función logarítmica de "Experience" (Imagen 23) mostrando una correlación muy débil y no significativa entre la experiencia laboral y "Risk_Flag". Tanto el coeficiente de correlación (0.1483) como el coeficiente de determinación (r^2 de -0.02199) indican que el modelo logarítmico empleado no se ajusta adecuadamente a los datos.

Imagen 23. Gráfico real vs predicción de 'Risk_Flag'



Conclusión

Luego de realizar un análisis exhaustivo de los resultados y gráficas generadas a partir de diversos modelos de regresión, incluyendo la regresión lineal simple, regresión lineal múltiple y regresión no lineal, hemos identificado que el mejor modelo que se ajusta a nuestros datos es el modelo de regresión no lineal utilizando la función logarítmica. Este modelo se enfoca en predecir la variable "Age" utilizando "Experience" como variable independiente, con los coeficientes $a=20.3199303$, $b=1.0$, $c=1.0$, representados en la fórmula $a * np.log(x + 1)$.

$$a * np.log(x + 1)$$

Los resultados obtenidos de este modelo revelan un coeficiente de determinación (R^2) de -0.9682630433 y un valor de clasificación de 0.984 . Estos valores son notoriamente altos, lo que indica una evaluación significativa entre las variables "Age" y "Experience". Este alto valor refuerza el planteamiento de que la experiencia laboral está fuertemente relacionada con la edad de las personas, y el modelo logarítmico proporciona una representación precisa de esta relación.

Además, al aplicar el mismo modelo no lineal logarítmico observamos una compensación destacable entre la variable "Current House Yrs" y "Age", con un valor de compensación de 0.83 , cuando los coeficientes $a=3.07344713$ y $b=1,0$. Esto sugiere que la variable "Current House Yrs" también está influenciada por la edad de las personas, y el modelo logarítmico es efectivo para capturar esta relación.

En resumen, el enfoque de regresión no lineal utilizando el modelo logarítmico ha demostrado ser el más adecuado para describir las relaciones entre las variables en estudio. Los altos valores de R^2 y calificación respaldan la solidez del modelo y su capacidad para predecir de manera precisa la edad y su influencia en otras variables, proporcionando una valiosa herramienta para comprender y tomar decisiones basadas en estos datos.