



Tecnológico de Monterrey

Analítica de datos y herramientas de inteligencia artificial II TI3001C

Reporte Actividad 6

Regresión lineal múltiple y no lineal

Integrantes del equipo:

Jerónimo Bernat Regordosa	A01735591
Estefanía López Ponce	A01654214
José Bryan Zamora Pacheco	A01707585
Manlio F. Rivera Pérez	A01734797

Octubre 12, 2023

1. Preprocesamiento

En el preprocesamiento, se inicia con un proceso de detección de valores nulos en el Data Frame llamado 'TrainingDataComplete'. Para realizar esta tarea, se emplea el comando 'isnull().sum()', el cual calcula la cantidad de valores no válidos por columna. Además, al agregar otro '.sum()', se obtiene el recuento total de valores no válidos en el DataFrame.

Es importante destacar que, tras realizar este proceso, se constató que el conjunto de datos en cuestión no contiene ningún valor nulo, véase en la figura 1.

```
Datos nulos:  
Id           0  
Income        0  
Age           0  
Experience     0  
Married/Single 0  
House_Ownership 0  
Car_Ownership  0  
Profession     0  
CITY           0  
STATE          0  
CURRENT_JOB_YRS 0  
CURRENT_HOUSE_YRS 0  
Risk_Flag       0  
dtype: int64  
  
EL DataFrame cuenta con 0 valores nulos
```

Figura 1. Valores nulos por columna y total del DataFrame

Sucesivamente se llevó a cabo un análisis exhaustivo de los valores atípicos en las variables numéricas con el propósito de identificar y eliminar aquellos que se encuentren fuera de lo que se considera normal en el conjunto de datos. Para realizar esta tarea, se emplea la construcción de diagramas de caja y bigotes (boxplots) para cada variable numérica contenida en el DataFrame.

El resultado de este análisis revela que la única variable que muestra valores atípicos es la denominada "Risk_flag". Sin embargo, al profundizar en el análisis de esta variable, se observa que se trata de una variable binaria, ya que solamente contiene los valores 0 y 1, lo que indica la presencia o ausencia de un riesgo. Por lo tanto, no se considerará la eliminación de valores atípicos en esta variable, ya que su naturaleza binaria no permite una interpretación significativa de los valores atípicos. En resumen, el análisis de las variables numéricas en el DataFrame no arroja evidencia de la presencia de valores atípicos que requieran ser eliminados.

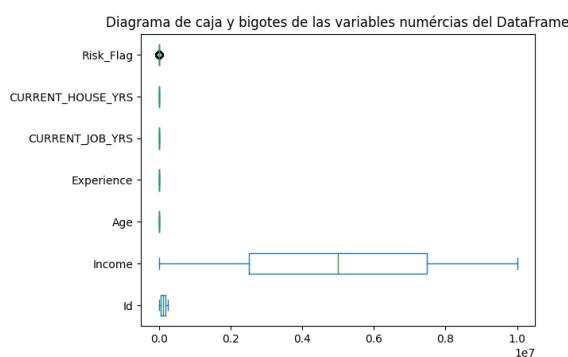


Figura 2. Outliers de las variables del Data Frame

2. Extracción de características

Se llevaron a cabo análisis de frecuencia de datos en las columnas contenidas en el DataFrame. Entre las columnas numéricas sujetas a análisis, se encuentran las siguientes: 'Age', 'Experience', 'CURRENT_JOB_YRS', 'CURRENT_HOUSE_YRS' y 'Risk_Flag'. Además, se realizó un análisis de las columnas categóricas, que incluyen 'Married/Single', 'House_Ownership', 'Profession' y 'CITY'.

En relación a la variable '**Age**', se puede observar que las cinco edades más frecuentes en la base de datos son las siguientes: 27 años, 66 años, 48 años, 33 años y 41 años. Por otro lado, se identifica que las edades menos representadas en la base de datos corresponden a individuos de 44 años, 73 años, 36 años, 51 años y 52 años.

Tabla 1. Mayores frecuencias de la variable Age

Edad	Mayor Frecuencia
27	5197
66	5092
48	4967
33	4921
41	4827

Tabla 2. Menores frecuencias de la variable Age

Edad	Menor Frecuencia
52	3197
51	3397
36	3444
73	3569
44	3622

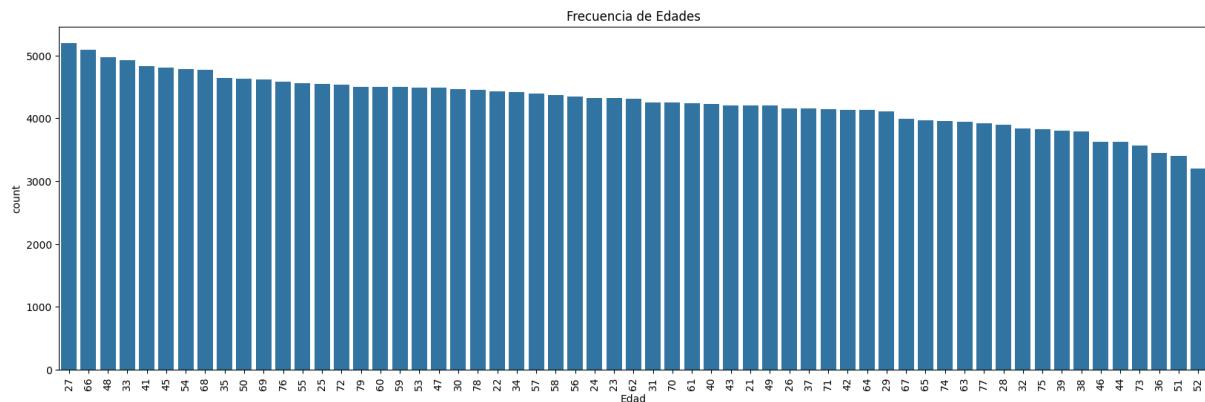


Figura 3. Gráfica de las frecuencias de la variable “Age”

Dentro de la variable "Experience", que representa la cantidad de años de experiencia laboral, se ha observado una distribución de datos en la que algunos valores son más frecuentes que otros. Los años de experiencia más comunes en este conjunto de datos son 6, 9, 12, 18 y 10 años. Esto sugiere que la mayoría de las personas en el grupo de estudio tienen una cantidad de experiencia laboral que se encuentra en este rango.

Tabla 3. Mayores frecuencias de la variable Experience

Años de experiencia	Mayor Frecuencia
6	13,158
9	12,936
12	12,800
18	12,601
10	12,503

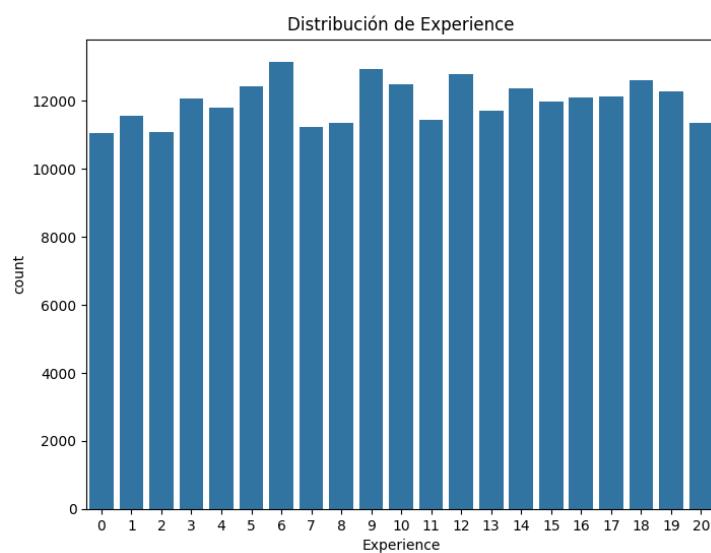


Figura 4. Gráfica las frecuencias de la variable “Experience”

En el contexto de análisis de datos, se ha evaluado la frecuencia de los valores en la columna "**CURRENT_JOB_YRS**", que representa los años que una persona ha estado empleada en el mismo trabajo. Observando la distribución de datos en la siguiente tabla, se ha identificado que los valores más frecuentes son 3, 4, 5 y 6 años de permanencia en el mismo empleo.

Por otro lado, se ha observado que los años de menor frecuencia son 12, 13 y 14 años. Esto indica que hay relativamente pocas personas en el conjunto de datos que han acumulado un tiempo considerable en el mismo empleo, ya que estos períodos de años son menos comunes entre los registros.

Esta información sobre la duración de la permanencia en el mismo empleo proporciona una perspectiva importante sobre la estabilidad laboral de las personas en el conjunto de datos y puede ser útil para comprender mejor sus trayectorias profesionales y sus niveles de experiencia en empleos específicos.

Tabla 4. Mayores frecuencias de la variable CURRENT_JOB_YRS

Años de experiencia	Mayor Frecuencia
3	29,460
4	28,339
5	25,091
6	23,491
7	19,925

Tabla 5. Menores frecuencias de la variable CURRENT_JOB_YRS

Años de experiencia	Menor Frecuencia
14	6,916
13	9,099
12	10,605
0	11,043
2	11,087

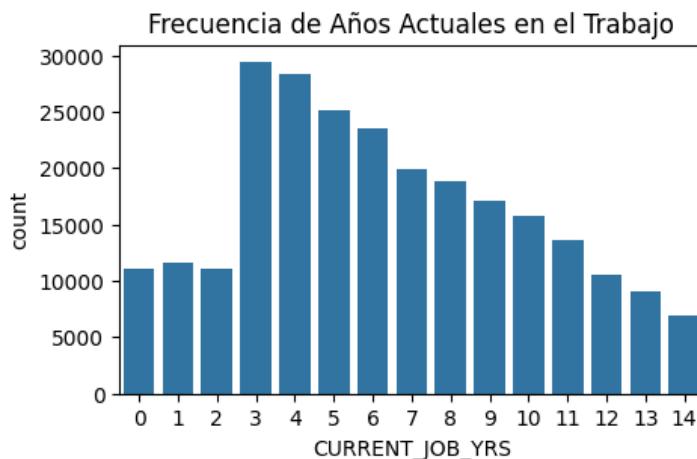


Figura 5. Gráfica de frecuencia de la variable “CURRENT_JOB_YRS”

Del mismo modo, se pretende determinar la frecuencia de los años presentes en la columna '**CURRENT_HOUSE_YRS**'. Al analizar la tabla y el gráfico que hemos elaborado, podemos observar que la distribución de estas frecuencias es notablemente similar.

Tabla 6. Frecuencias de la variable CURRENT_HOUSE_YRS

Años actuales en la casa de actual residencia	Frecuencia
11	51,873
12	51,791
13	50,659
14	49,003
10	48,674

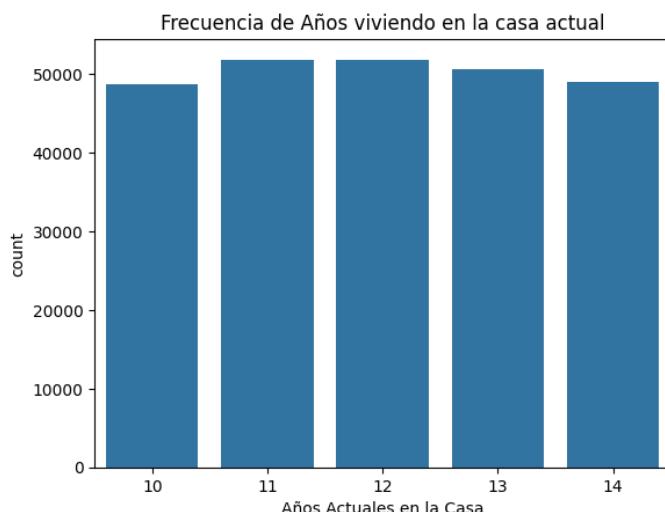


Figura 6. Gráfica de frecuencia de la variable “CURRENT_HOUSE_YRS”

En la columna '**Married/Single**', se constata que el 89,8% de la población incluida en la base de datos pertenece al grupo de individuos solteros, mientras que solamente el 10,2% se encuentra en estado civil de casados.

Tabla 7. Frecuencias de la variable “Single/Married”

Estado civil	Frecuencia
single	226,272
married	25,728

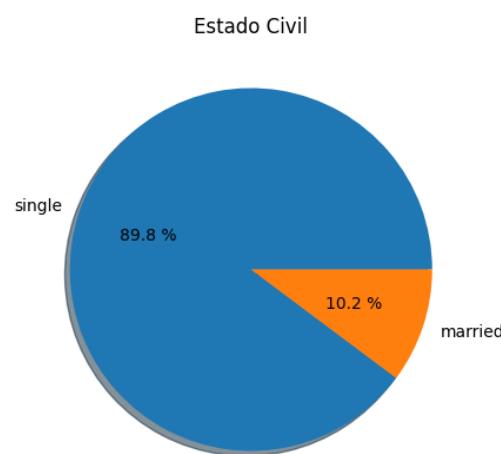


Figura 7. Gráfico de pie del porcentaje de frecuencia de la variable “Single/Married”

En lo que respecta a la columna '**House_Ownership**', se advierte que el 92,0% de los individuos residen en una vivienda en calidad de arrendatarios, mientras que el 2,9% pertenece a aquellos que no están actualmente en régimen de alquiler, y el 5,1% corresponde a personas que son propietarias de la vivienda que ocupan.

Tabla 8. Frecuencias de la variable “House_Ownership”

House Ownership	Frecuencia
rented	231,898
owned	12,918
norent_noown	7,184

Tipos de Propiedad de Vivienda

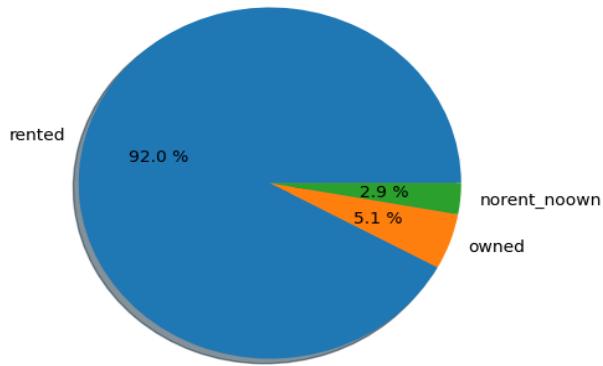


Figura 8. Gráfico de pie del porcentaje de frecuencia de la variable “House_Ownership”

En lo que respecta a la columna '**Profession**', se aplicó un filtro con el propósito de identificar las 20 profesiones más frecuentes en el DataFrame. La figura siguiente muestra las cinco profesiones más comunes, que son: Physician, Statistician, Web Designer, Psychologist y Computer Hardware Engineer.

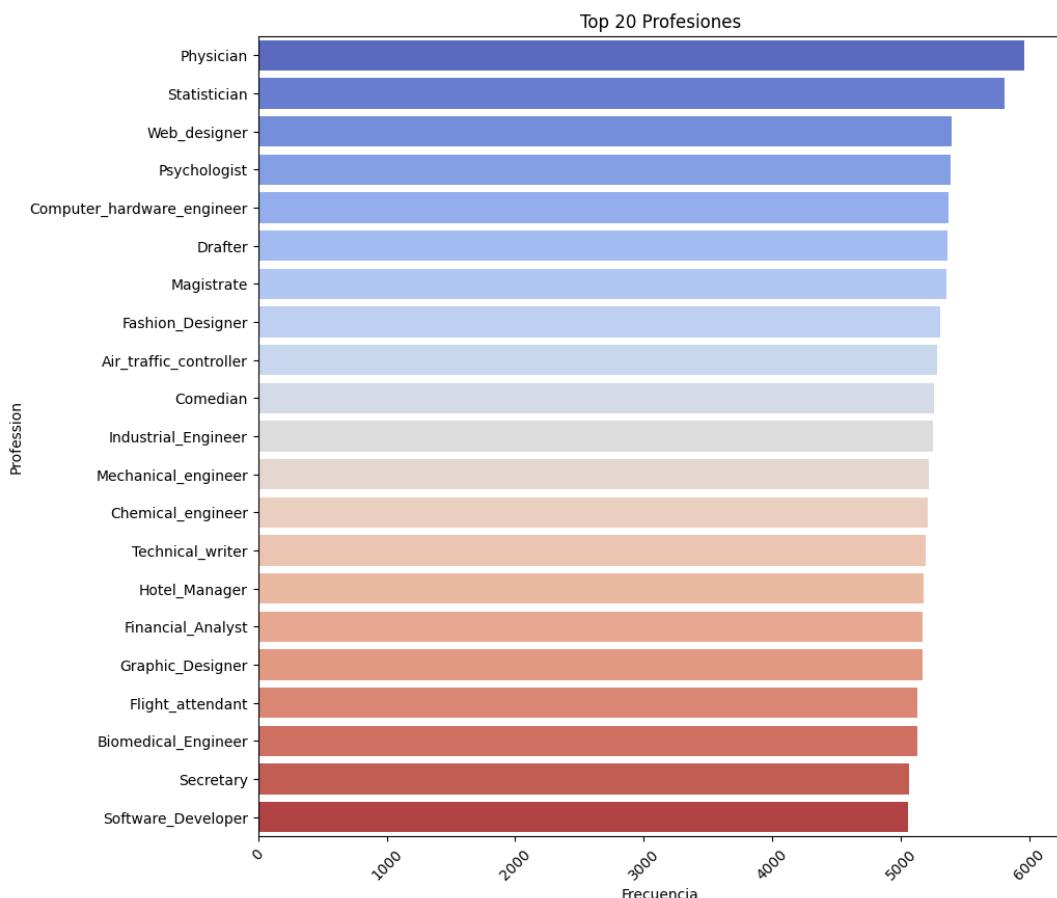


Figura 9. Gráfico de barras de las frecuencias de la variable “Profession”

Este mismo filtro se aplicó a la columna 'CITY' con el objetivo de determinar las 20 ciudades más frecuentemente registradas en el conjunto de datos. Según lo que indica el gráfico, Vijayanagaram, Bhopal y Bulandshahr son las ciudades de origen predominantes de los individuos en la muestra.

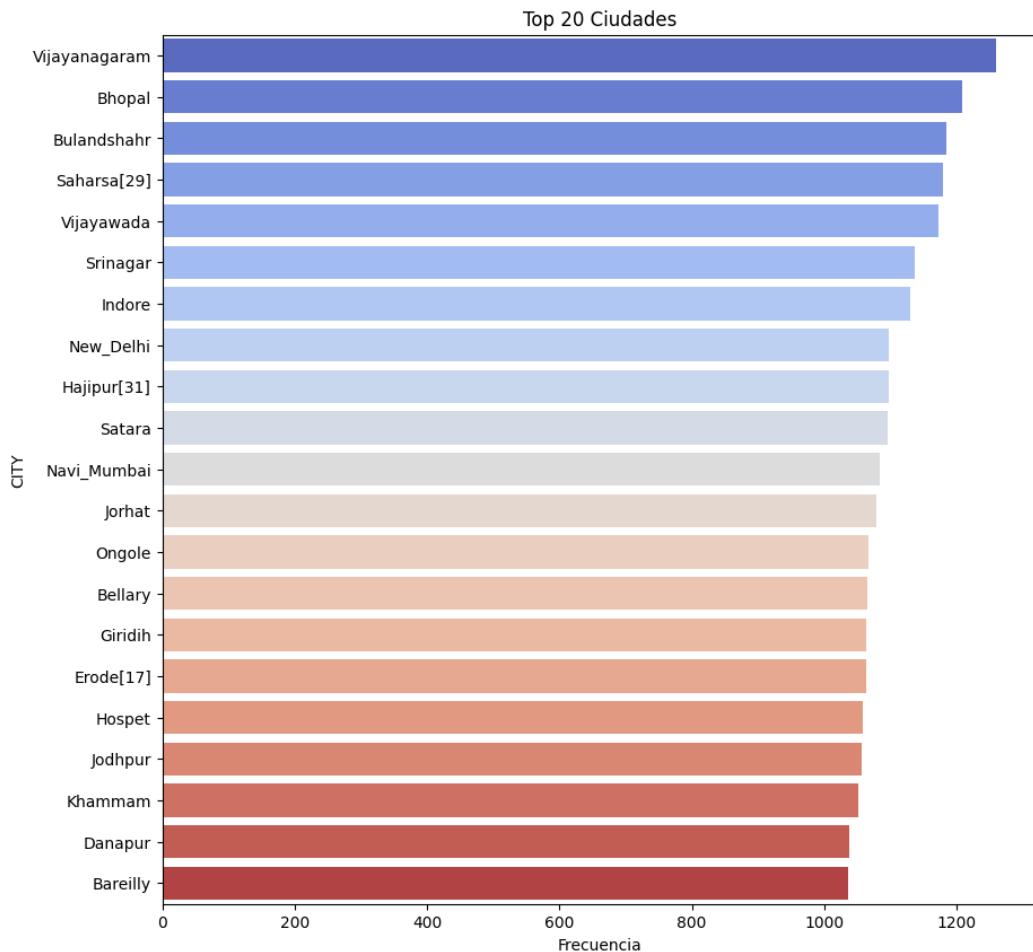


Figura 10. Gráfico de barras de las frecuencias de la variable “CITY”

Por último, en lo que concierne a la columna 'Risk_Flag', se determina la frecuencia de individuos que tienen un historial de incumplimiento en sus deudas, así como aquellos que no lo tienen. Esta información se encuentra representada en la tabla y el gráfico siguientes.

Tabla 9. Frecuencias de la variable “Risk_Flag”

Estatus(0-1)	Frecuencia
0	221,004
1	30,996

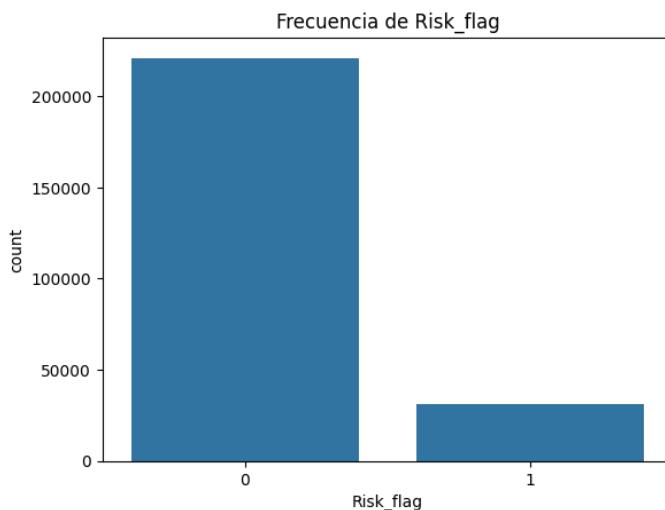


Figura 11. Gráfico de barras de las frecuencias de la variable “Risk_Flag”

3. Análisis de correlación entre variables numéricas

Para llevar a cabo el análisis de correlación entre variables numéricas, es fundamental comenzar con la identificación y depuración de aquellas variables que son de naturaleza cualitativa. En el contexto de la base de datos que estamos manejando, estas variables incluyen: "Married/Single", "House_Ownership", "Car_Ownership", "Profession", "CITY" y "STATE".

Posteriormente, se empleó el método `'.corr()'` para calcular la correlación entre las columnas de la base de datos que estamos analizando. La correlación puede variar en un rango entre -1 y 1, donde -1 indica una correlación negativa perfecta, 1 denota una correlación positiva perfecta y 0 señala la ausencia de correlación. Para facilitar la interpretación, se redondearon todos estos valores numéricos a su valor absoluto utilizando la función `'abs()'`. Esta transformación es conveniente para nuestro propósito, ya que nuestra principal preocupación no radica en la dirección de las correlaciones, es decir, si son positivas o negativas.

Para visualizar la correlación absoluta entre las variables de nuestros datos, hemos empleado la biblioteca Seaborn. Creamos un mapa de calor utilizando la función `'sns.heatmap()'` de esta biblioteca previamente mencionada. Dicha función toma como entrada la matriz de correlación absoluta y permite personalizar la apariencia del mapa de calor.

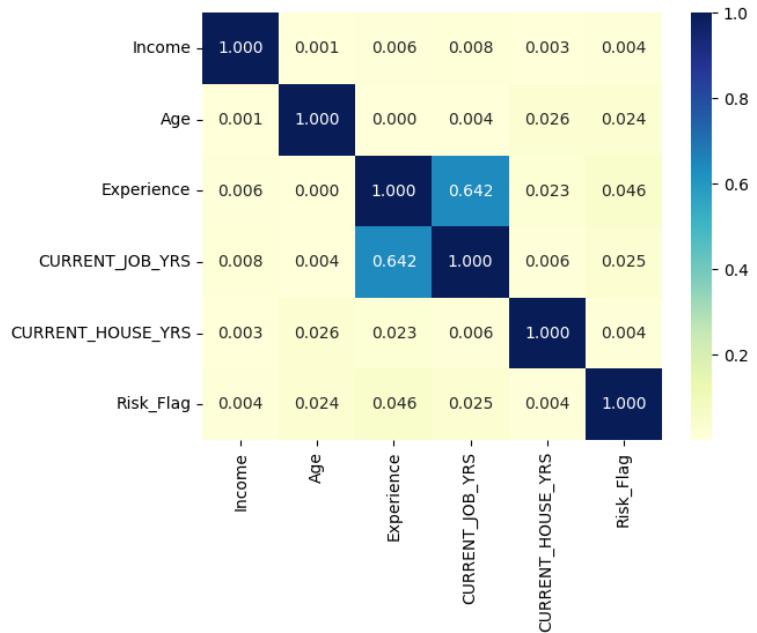


Figura 12. Mapa de calor con las correlaciones entre las variables cuantitativas del Data Frame

A partir de la información proporcionada por el mapa de calor, llegamos a la conclusión de que existe una correlación significativa con la cual podemos trabajar de manera adecuada, y esta es la que se da entre "CURRENT_JOB_YRS" y "Experience". Una vez identificadas las variables con una correlación relevante, seleccionamos "CURRENT_JOB_YRS" como variable independiente y "Experience" como variable dependiente.

Modelo matemático

$$1.06339513x + 3.3490225813747347$$

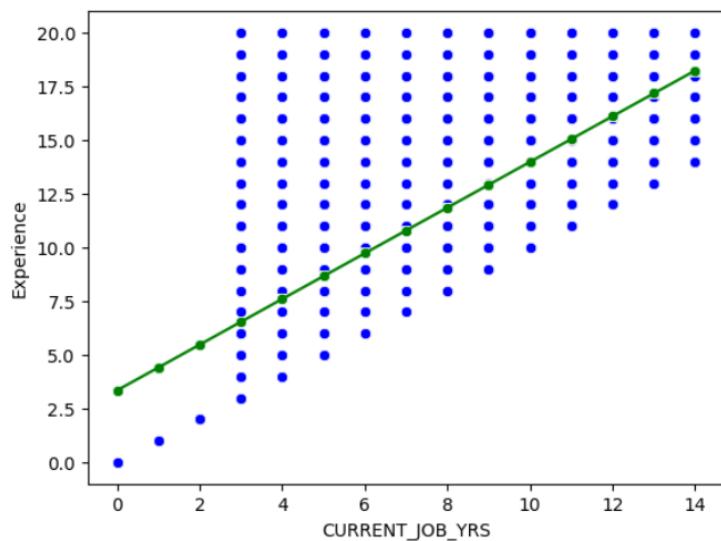


Figura 13. Predicciones de la variable “Experience” con el modelo de regresión lineal simple vs datos

Por último, se obtuvieron los coeficientes de determinación y de correlación de nuestro modelo

Coeficiente de determinación

0.417442

Coeficiente de correlación

0.64609

4. Modelo de regresión lineal múltiple

El DataFrame incluye un total de 12 variables, divididas equitativamente entre categorías y valores numéricos. En esta sección, se llevará a cabo el desarrollo de un modelo de regresión lineal múltiple para cada variable cuantitativa presente en el DataFrame. Luego, se procederá a comparar los coeficientes obtenidos en estos modelos con los coeficientes obtenidos en el mapa de calor.

El objetivo es determinar si los coeficientes de correlación en los modelos de regresión lineal múltiple son superiores, lo que indicaría que estos modelos son más efectivos que los modelos de regresión lineal simple. Para seleccionar las variables apropiadas para cada modelo, se utilizará el mapa de calor de la figura 12, que muestra los valores absolutos de los coeficientes de correlación entre todas las variables cuantitativas disponibles.

4.1. Income

Analizando la figura #, podemos observar que las variables que presentan más correlación con "Income" son "CURRENT_JOB_YRS" y "Experience", por lo que utilizaremos ambas para crear un modelo de regresión lineal múltiple.

Modelo matemático

$$y = 3922.76017175x_1 + 1539.35024741x_2 + 4956746.905125832$$

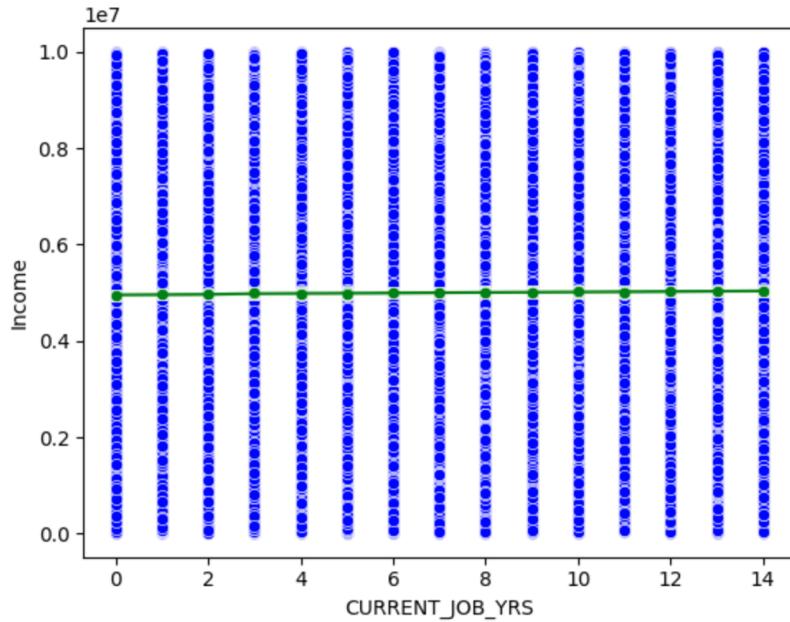


Figura 14. Predicciones de la variable “Income” con el modelo de regresión lineal múltiple vs datos

Donde

- x1: CURRENT_JOB_YRS
- x2: Experience

Coeficientes de determinación (R^2)

5.5629e-05

Coeficiente de correlación

0.0074

El coeficiente de correlación del modelo de regresión lineal múltiple generado aumentó ligeramente, siendo este de 0.0074, en comparación del coeficiente de correlación más alto en el mapa de calor que fue del 0.0070. Ambos valores son extremadamente bajos, pero el modelo lineal múltiple tuvo un mejor valor.

4.2. Age

Analizando el mapa de calor podemos observar que las variables que presentan más correlación con "Age" son "CURRENT_HOUSE_YRS" y "Risk_Flag", por lo que utilizaremos ambas para crear un modelo de regresión lineal múltiple.

Modelo matemático

$$y = -0.2467444x_1 - 1.13769054x_2 + 53.05439573135235$$

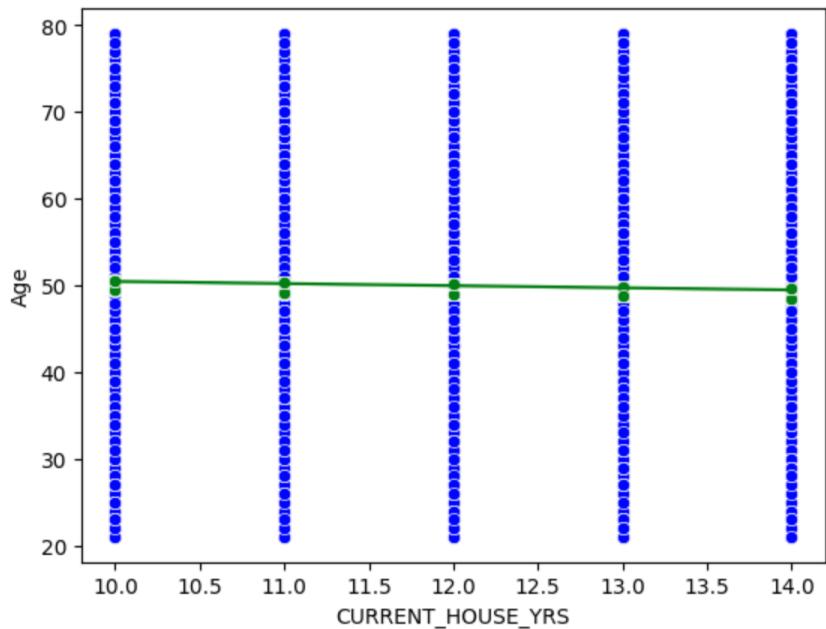


Figura 15. Predicciones de la variable “Age” con el modelo de regresión lineal múltiple vs datos

Donde:

- x1: CURRENT_HOUSE_YRS
- x2: Risk_Flag

Coefficientes de determinación (R^2)

0.00088

Coeficiente de correlación

0.02974

El coeficiente de correlación del modelo de regresión lineal múltiple generado aumentó ligeramente, siendo este de 0.0297, en comparación del coeficiente de correlación más alto en el mapa de calor que fue del 0.0218. Ambos valores son extremadamente bajos, pero el modelo lineal múltiple tuvo un mejor valor.

4.3. Experience

Analizando el mapa de calor podemos observar que "Experience" presenta una correlación media/alta con "CURRENT_JOB_YRS", siendo la más alta entre todas las variables del dataframe, con un valor de 0.646. Para el modelo lineal múltiple contemplaremos esta variable, al igual que las variables: "CURRENT_HOUSE_YRS" y "Risk_flag", las cuales presentan correlaciones bajas pero podrían ayudar a conseguir un mejor modelo.

Modelo matemático

$$y = 1.06260028x_1 + 0.06752352x_2 - 0.42978313 + 2.5967870708885386$$

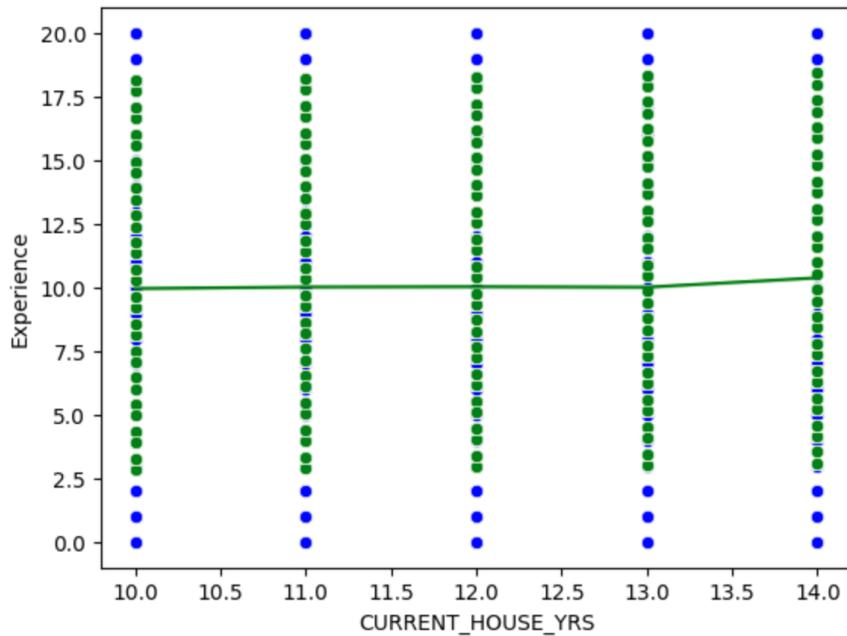


Figura 16. Predicciones de la variable “Experience” con el modelo de regresión lineal múltiple vs datos

Donde

- x1: CURRENT_JOB_YRS
- x2: CURRENT_HOUSE_YRS
- x3: Risk_Flag

Coefficientes de determinación (R^2)

0.4182

Coeficiente de correlación

0.6467

El coeficiente de correlación del modelo de regresión lineal múltiple generado experimentó un leve aumento, alcanzando un valor de 0.6467, en comparación con el coeficiente de correlación más alto previamente registrado, que se situaba en 0.6461. Ambos valores representan una correlación media/alta, pero el objetivo de este análisis es determinar si hay un incremento al emplear un modelo lineal múltiple y, en este caso, se observó una ligera mejora.

4.4. CURRENT_JOB_YRS

Las variables "CURRENT_JOB_YRS" y "Experience" son las que presentan mayor correlación dentro del dataframe, siendo esta de 0.64. Por lo que para el modelo utilizaremos la variable "Expereince", al igual que la variable "Risk_flag", que es la segunda con más correlación con la variable "CURRENT_JOB_YRS".

Modelo matemático

$$y = 0.3926685x_1 + 0.05962746x_2 + 2.3667022068255905$$

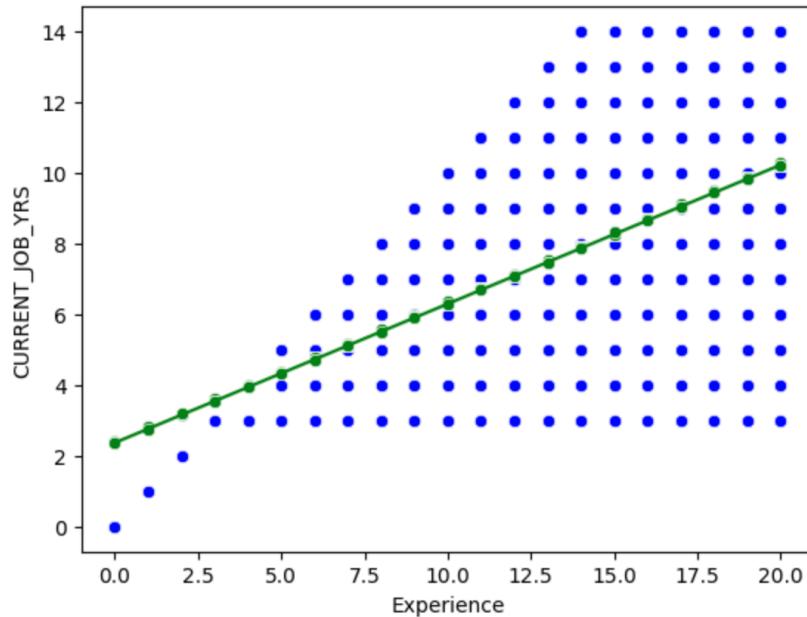


Figura 17. Predicciones de la variable “CURRENT_JOB_YRS” con el modelo de regresión lineal múltiple vs datos

Donde

- x1: Experience
- x2: Risk_Flag

Coefficientes de determinación (R^2)

0.4174

Coefficiente de correlación

0.64611

El coeficiente de correlación del modelo de regresión lineal múltiple generado experimentó un incremento muy leve, alcanzando un valor de 0.64611, en comparación con el coeficiente de correlación más alto previamente registrado, que se situaba en 0.64610. Estos valores son sumamente similares, dado que la otra variable independiente incorporada en el modelo presentaba una correlación muy baja, lo que limitó su impacto en el resultado. No obstante, se observó una mejora mínima en el coeficiente, aunque prácticamente insignificante.

4.5. CURRENT_HOUSE_YRS

Analizando el mapa de calor podemos observar que las variables que presentan más correlación con "CURRENT_HOUSE_YRS" son "Age", "Experience" y "CURRENT_JOB_YRS", por lo que utilizaremos esas tres para crear un modelo de regresión lineal múltiple.

Modelo matemático

$$y = -0.00164616x_1 + 0.00631643x_2 - 0.00463972 + 12.045715902117509$$

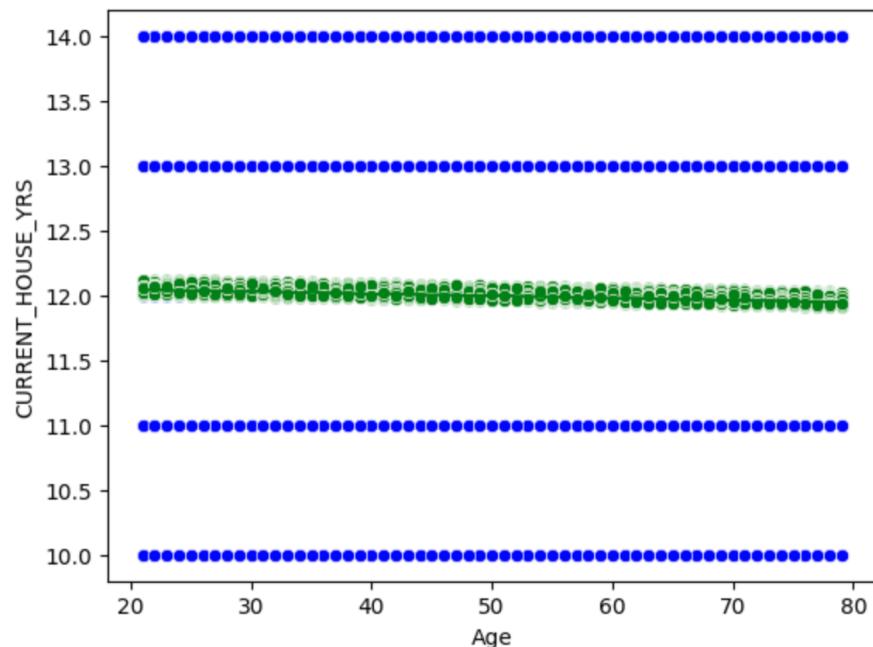


Figura 18. Predicciones de la variable “CURRENT_HOUSE_YRS” con el modelo de regresión lineal múltiple vs datos

Donde

- x1: Age
- x2: Experience
- x2: CURRENT_JOB_YRS

Coeficientes de determinación (R^2)

0.00086

Coeficiente de correlación

0.02936

El coeficiente de correlación del modelo de regresión lineal múltiple generado experimentó un aumento, alcanzando un valor de 0.0293, en contraste con el coeficiente de correlación más alto previamente registrado, que se situaba en 0.0193. A pesar de que ambos valores son sumamente bajos, el objetivo de este análisis es evaluar si la implementación de un modelo

lineal múltiple resulta en un incremento de la correlación. En este caso, se observó una mejora mínima en la correlación.

4.6. Risk_Flag

Analizando el mapa de calor podemos observar que las variables que presentan más correlación con "Risk_Flag" son "Age" y "Experience", por lo que utilizaremos ambas para crear un modelo de regresión lineal múltiple.

Modelo matemático

$$y = -0.00042052x_1 - 0.00189028x_2 + 0.16306900612184838$$

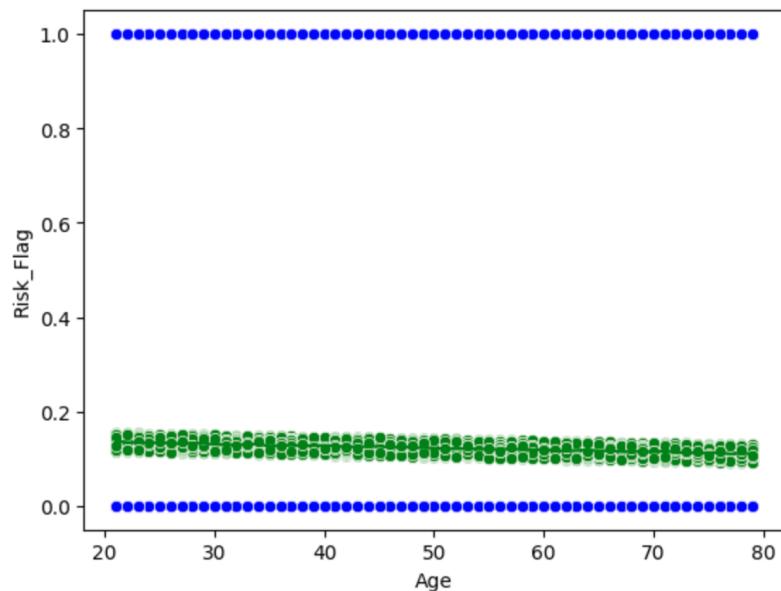


Figura 19. Predicciones de la variable “Risk_Flag” con el modelo de regresión lineal múltiple vs datos

Donde

- x1: Age
- x2: Experience

Coeficientes de determinación (R^2)

0.00166

Coeficiente de correlación

0.04085

El coeficiente de correlación del modelo de regresión lineal múltiple generado experimentó un ligero aumento, alcanzando un valor de 0.04085, en contraste con el coeficiente de correlación más alto registrado previamente, que se situaba en 0.0345.

En conclusión de esta sección podemos decir que en los modelos de regresión lineal múltiple que incluyeron al menos dos variables independientes, se observó un coeficiente de correlación superior en comparación con el valor más alto registrado en el mapa de calor de las variables del conjunto de datos. Aunque en todos los casos esta mejora fue mínima, esto se debió a que las variables utilizadas mostraron correlaciones extremadamente bajas entre sí, lo que limitaba su impacto en el modelo. En resumen, podemos concluir que los modelos de regresión lineal múltiple siempre arrojaron un coeficiente de correlación superior.

5. Modelos de regresión no lineal

Así como en la sección anterior, se llevará a cabo el desarrollo de un modelo de regresión no lineal para cada variable cuantitativa presente en el DataFrame. Luego, se procederá a comparar los coeficientes obtenidos en estos modelos con los coeficientes obtenidos en los modelos de regresión lineal simple y múltiple.

De igual manera, el objetivo es determinar si los coeficientes de correlación en estos modelos son superiores, lo que indicaría que estos modelos son más efectivos que los modelos de regresión lineal simple o múltiple.

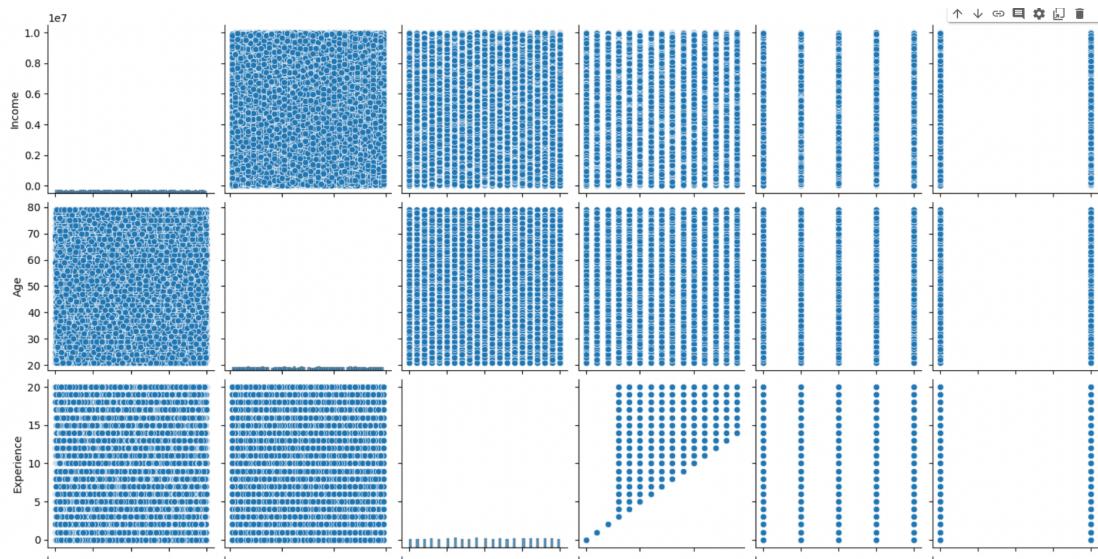


Figura 20. Gráficas de dispersiones de todas las variables. Parte 1.

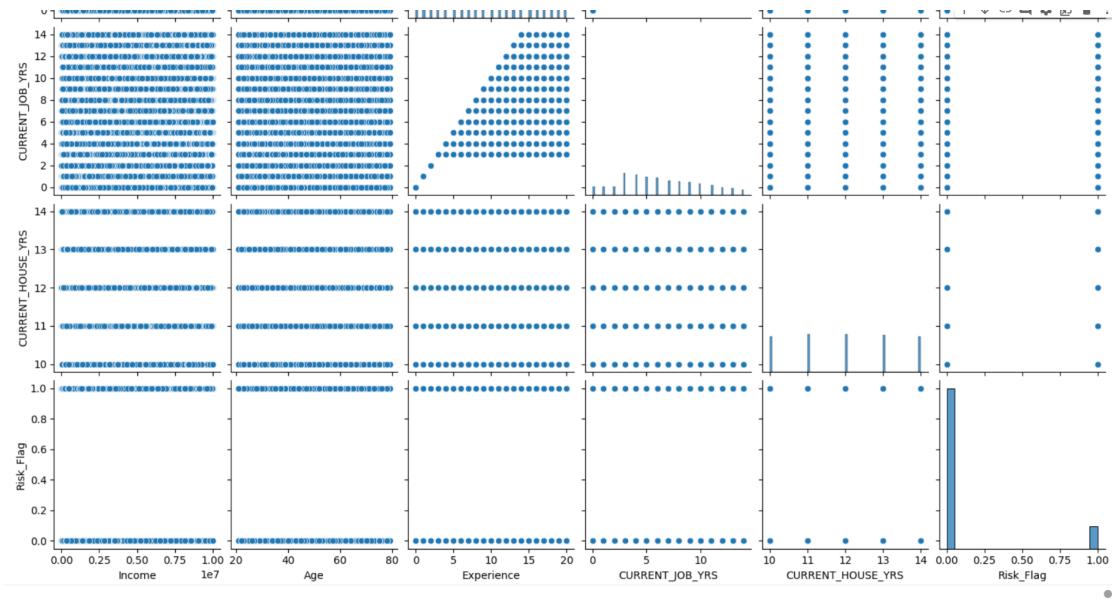


Figura 21. Gráficas de dispersiones de todas las variables. Parte 1.

5.1. Income

En este análisis, se llevó a cabo un ajuste de regresión no lineal para estudiar la relación entre el ingreso (Income) y la experiencia laboral (Experience). El objetivo era determinar si la función inversa, representada por la ecuación "y = 1/a*x," podría utilizarse para modelar esta relación. El proceso se dividió en varias etapas que se describen a continuación.

Función

$$y = 1/a*x$$

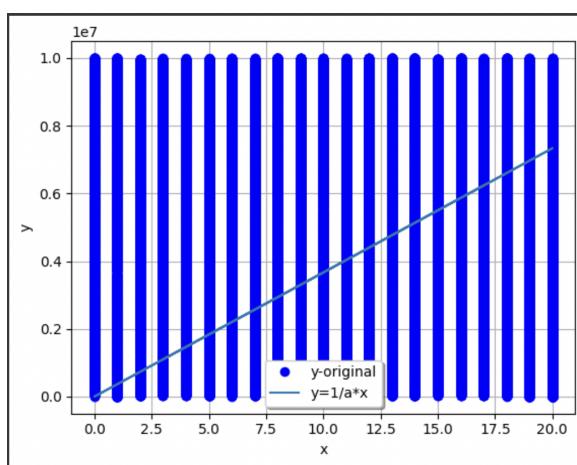


Figura 22. Predicciones de la variable “Income” con el modelo de regresión No lineal

Donde

- X = Variable Independiente: Experience
- Y= Variable Dependiente: Income

Coeficientes de determinación (R^2)

0.7787267413363248

Coeficiente de determinación

0.8824549514486985

El modelo de regresión no lineal basado en una función inversa se ajustó a los datos de ingreso y experiencia. Es fundamental considerar que el uso de una función inversa puede ser apropiado en ciertos contextos, pero la elección del modelo debe basarse en la naturaleza de los datos y la relación subyacente entre las variables. En este caso se observa que el coeficiente de determinación es alto, muy cercano al 1, lo que nos indica que el modelo si es eficaz.

5.2. Age

En este análisis, se llevó a cabo un ajuste de regresión no lineal para estudiar la relación entre la edad (Age) y el número de años en la casa actual (CURRENT_HOUSE_YRS). El objetivo fue determinar si una función de ajuste basada en una ecuación de cociente entre polinomios podría utilizarse para modelar esta relación. El proceso se dividió en varias etapas que se describen a continuación.

Función

$$y = (a*x^{**1} + b)/ c*x^{**2}$$

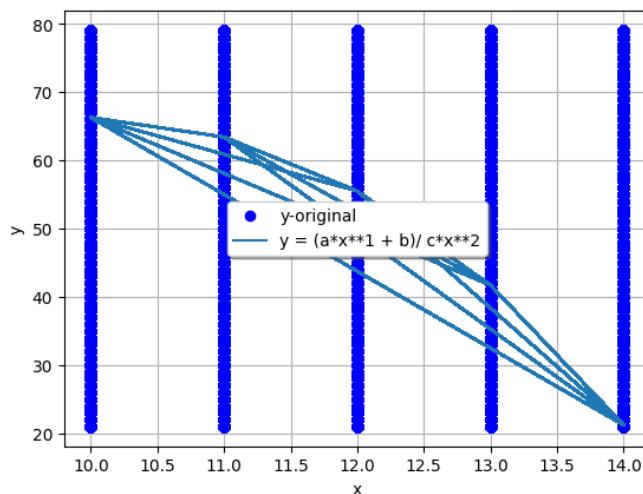


Figura 23. Predicciones de la variable “Age” con el modelo de regresión No lineal

Donde

- X = Variable Independiente: CURRENT_HOUSE_YRS

- Y= Variable Dependiente: Age

Coeficientes de determinación (R^2)

0.8845262555588107

Coeficiente de determinación

0.940492560076267

El modelo de regresión no lineal basado en una función de cociente entre polinomios se ajustó a los datos de edad y años en la casa actual. Este es el modelo con el coeficiente de determinación más alto.

5.3. Experience

En este análisis, se llevó a cabo un ajuste de regresión no lineal para investigar la relación entre la experiencia laboral (Experience) y el número de años en el trabajo actual (CURRENT_JOB_YRS). El objetivo fue determinar si una función de ajuste basada en una ecuación que involucra el valor absoluto y términos lineales podría utilizarse para modelar esta relación. El proceso se dividió en varias etapas que se describen a continuación.

Función

$$y = a * \text{np.abs}(x) + b * x + c \text{ (Función Valor absoluto)}$$

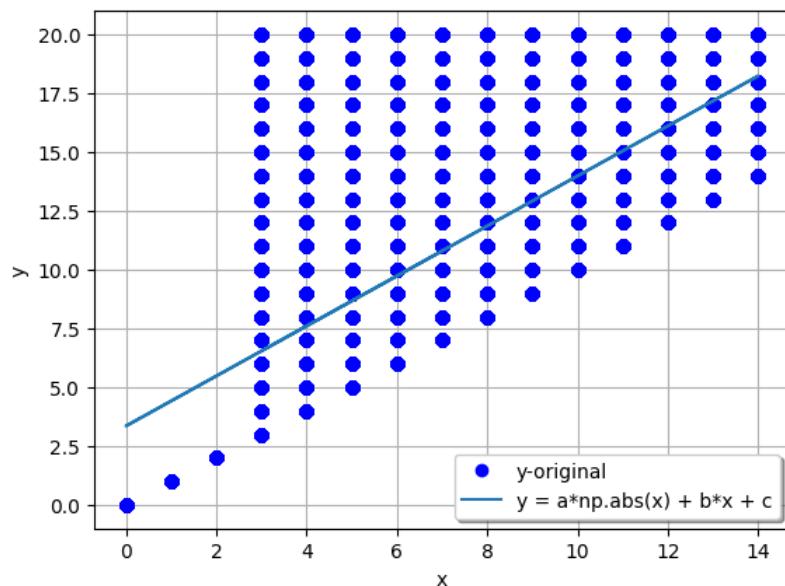


Figura 24. Predicciones de la variable “Age” con el modelo de regresión No lineal

Donde

- X = Variable Independiente: CURRENT_JOB_YRS
- Y= Variable Dependiente: Experience

Coeficientes de determinación (R^2)

0.4174420012200908

Coeficiente de determinación

0.6460975168038419

El modelo de regresión no lineal basado en una función que combina términos de valor absoluto y lineales se ajustó a los datos de experiencia y años en el trabajo actual.

5.4. CURRENT_JOB_YRS

Este análisis se centra en el ajuste de un modelo de regresión no lineal para examinar la relación entre el número de años en el trabajo actual (CURRENT_JOB_YRS) y el número de años en la casa actual (CURRENT_HOUSE_YRS). El objetivo es determinar si una función de ajuste basada en una ecuación que incluye términos polinómicos y una función de cociente entre polinomios puede utilizarse para modelar esta relación.

Función

$$y = (a*x^{**1} + b)/ c*x^{**17} \text{ (Función cociente entre polinomios)}$$

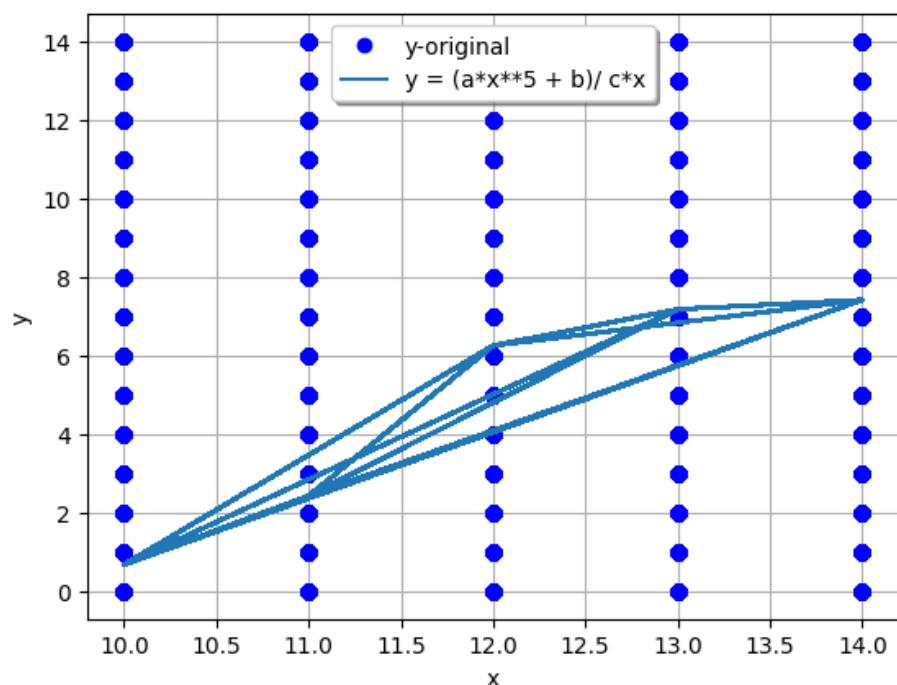


Figura 25. Predicciones de la variable “CURRENT_JOB_YRS” con el modelo de regresión No lineal

Donde

- X = Variable Independiente: CURRENT_HOUSE_YRS
- Y= Variable Dependiente: CURRENT_JOB_YRS

Coeficientes de determinación (R^2)

0.7298514492149053

Coeficiente de determinación

0.854313437337202

El modelo de regresión no lineal basado en una función de cociente entre polinomios y términos de potencia se ajustó a los datos de años en el trabajo actual y años en la casa actual.

5.5. CURRENT_HOUSE_YRS

Este análisis se enfoca en el ajuste de un modelo de regresión no lineal para estudiar la relación entre el número de años en la casa actual (CURRENT_HOUSE_YRS) y la edad (Age). El objetivo es determinar si una función de ajuste basada en una ecuación polinómica inversa podría utilizarse para modelar esta relación.

Función

$$y = a/b*x^{**2} + c*x \text{ (Función polinomial inversa)}$$

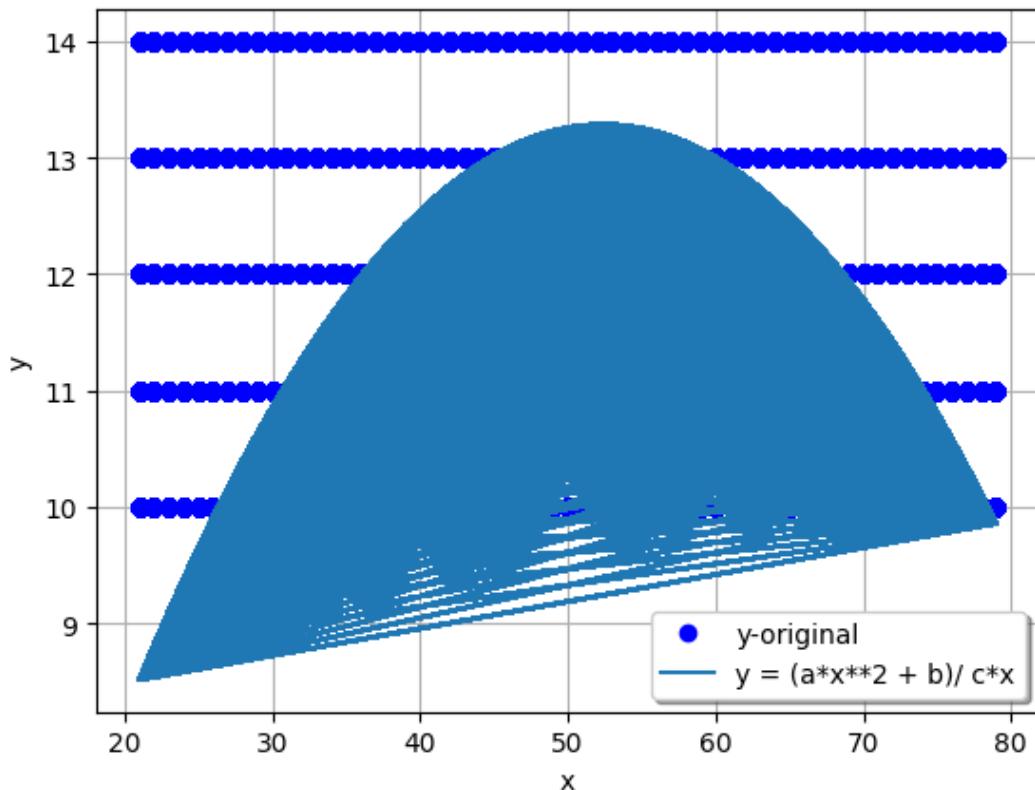


Figura 26. Predicciones de la variable “CURRENT_HOUSE_YRS” con el modelo de regresión No lineal

Donde

- X = Variable Independiente: Age
- Y= Variable Dependiente: CURRENT_HOUSE_YRS

Coeficientes de determinación (R^2)

0.9329962788057777

Coeficiente de determinación

0.965917325036557

El modelo de regresión no lineal basado en una función de ecuación polinómica inversa se ajustó a los datos de años en la casa actual y edad.

5.6. Risk_Flag

En este análisis, se realizó un ajuste de regresión no lineal para estudiar la relación entre la variable Risk_Flag y el ingreso (Income). El objetivo es determinar si una función de ajuste basada en una ecuación polinómica inversa puede utilizarse para modelar esta relación.

Función

$$y = a/b*x^{**2} + c*x \text{ (Función polinomial inversa)}$$

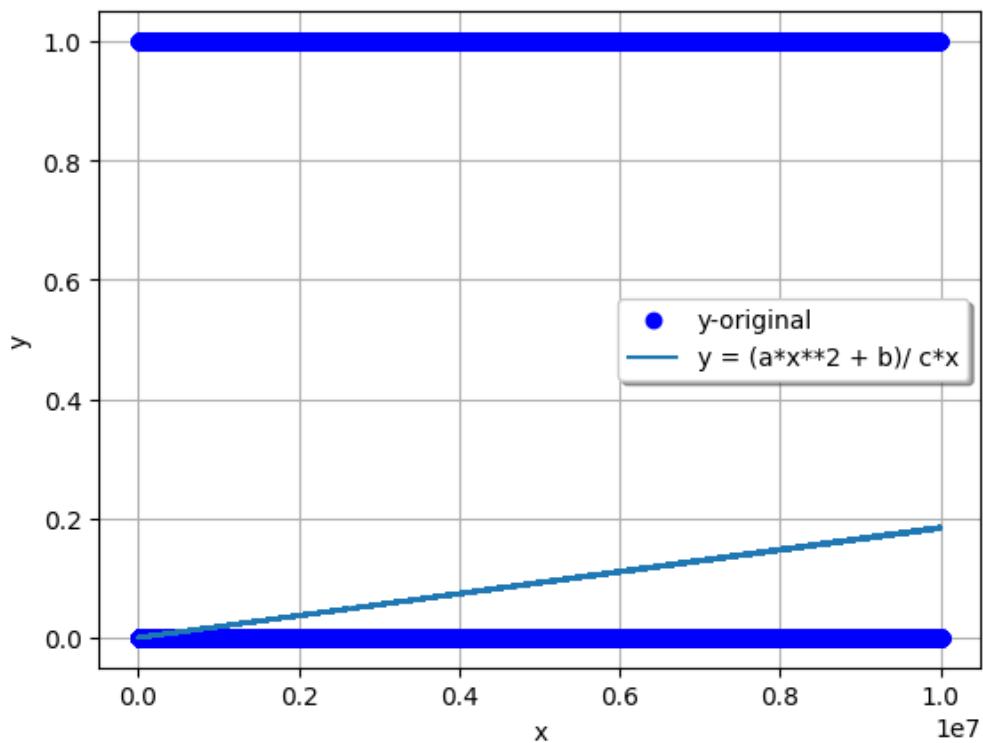


Figura 27. Predicciones de la variable “Risk_Flag” con el modelo de regresión No lineal

Donde

- X = Variable Independiente: Income
- Y= Variable Dependiente: Risk_Flag

Coeficientes de determinación (R^2)

0.035937180474464814

Coeficiente de determinación

0.18957104334382088

El modelo de regresión no lineal basado en una función de ecuación polinómica inversa se ajustó a los datos de Risk_Flag e Income.

Resultados

En esta sección del informe, se generará una tabla, donde se llevará a cabo una comparación de los coeficientes de correlación de los modelos con el objetivo de determinar cuáles de ellos obtuvieron los resultados más favorables.

Tabla 10. Comparación de los coeficientes de correlación obtenidos en los modelos creados

Variable Dependiente	Coeficiente de correlación		
	Regresión Lineal Simple	Regresión Lineal Múltiple	Regresión No Lineal
Income	0.0070	0.0074	0.8824
Age	0.0218	0.0297	0.9404
Experience	0.6461	0.6467	0.6460
CURRENT_JOB_YRS	0.64610	0.64611	0.8543
CURRENT_HOUSE_YRS	0.0193	0.0293	0.9659
Risk_Flag	0.0345	0.0408	0.1895