# ANÁLISIS DE PRESTAMOS BANCARIOS


INTRODUCCIÓN A LA CIENCIA DE DATOS
SANDRO FARID DÍAZ RODRÍGUEZ/ MIA CANALES CORONA

El análisis de esta base de datos tuvo como meta entender los factores que un banco toma en cuenta para otorgar un préstamo bancario, en la actualidad este tipo de análisis son de mucha ayuda para las nuevas generaciones debido que en ocasiones carecen de educación financiera, es por eso que con el análisis ya mencionado se piensa ayudar a responder dudas relacionadas al tema.

La base de datos a partir de la cual se llevó a cabo este análisis fue sacada de la página web kaggle, esta base de datos cuenta con 20,000 datos, posee 33 columnas, entre las cuales resaltan las de “Ingreso mensual”, “Ingreso Anual”, “Grado de riesgo”, “Edad”, etc.

Para llevar a cabo la limpieza de la base de datos se usaron múltiples funciones, con el fin de no perder tantos datos, es decir, información vital, los datos de tipo NaN se reemplazaron por “0” y posteriormente en algunas columnas se reemplazó ese “0” por el promedio, este procedimiento se realizó únicamente en columnas que no pudiesen distorsionar la información de los datos, como por ejemplo en la columna “Estatus Marital”, por otra parte con respecto a los datos atípicos, en caso de haber mucho desglose con respecto a la mayoría de datos, se estableció un rango, el cual los eliminaba.

Como visión general tenemos 20,494 filas y 33 columnas, de las cuales, 20 columnas son de tipo entero (int), 9 de tipo decimal (float) y 4 de tipo objeto (object). En su mayoría las de tipo entero se aplicó a columnas con datos tales como “Edad”, “Experiencia”, “Aprobación del préstamo2, etc. La columna con el tipo de dato decimal se aplicó a columnas donde

tuviera que ver la cantidad de dinero, tales como “Ingreso Mensual”, “Ingreso Anual”, “Networth”, “Pago Mensual”, etc. Y las de tipo objeto se dejó en columnas como “Estatus Marital”, “Situación de empleo”, “Nivel de educación”, etc.

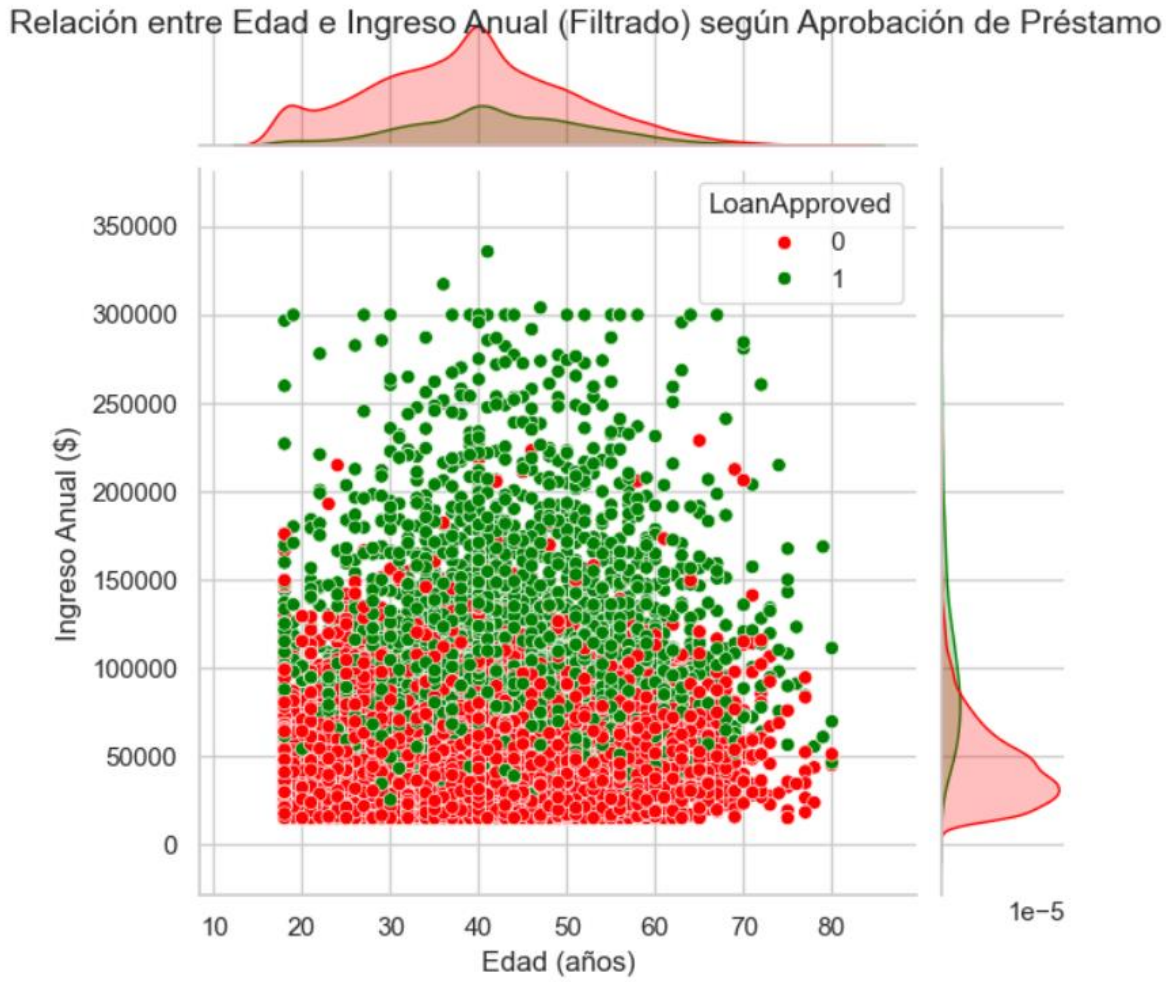
Estas son la estadísticas de cuando el préstamo fue aprobado:

| | Edad | IngresoAnual | CreditScore | Experiencia | Monto | Duración |
|---|---|---|---|---|---|---|
| count | 4887.000000 | 4887.000000 | 4887.000000 | 4887.000000 | 4887.000000 | 4887.000000 |
| mean | 42.549417 | 97406.200737 | 551.846327 | 20.144260 | 18057.894414 | 47.270718 |
| std | 10.601630 | 53713.319197 | 142.547050 | 10.493759 | 9587.609476 | 24.184077 |
| min | 18.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 36.000000 | 63167.000000 | 546.000000 | 13.000000 | 11958.000000 | 36.000000 |
| 50% | 42.000000 | 89174.000000 | 587.000000 | 19.000000 | 16824.000000 | 48.000000 |
| 75% | 49.000000 | 121947.000000 | 618.000000 | 27.000000 | 23198.500000 | 60.000000 |
| max | 80.000000 | 485341.000000 | 712.000000 | 59.000000 | 82644.000000 | 120.000000 |

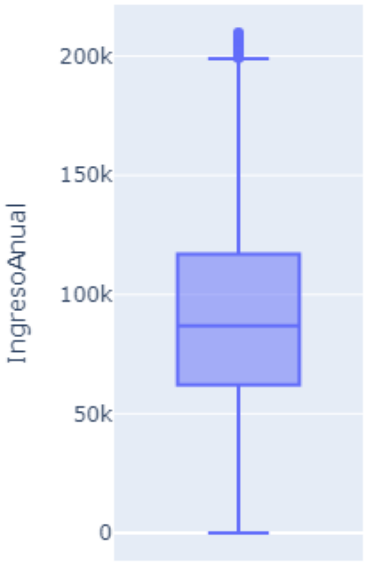
Y estas son cuando el préstamo NO fue aprobado

| | Edad | IngresoAnual | CreditScore | Experiencia | Monto | Duración |
|---|---|---|---|---|---|---|
| count | 15607.000000 | 15607.000000 | 15607.000000 | 15607.000000 | 15607.000000 | 15607.000000 |
| mean | 38.902992 | 43246.483821 | 538.213750 | 16.654386 | 25271.137182 | 52.296534 |
| std | 11.349117 | 25545.793193 | 135.377192 | 10.964945 | 14985.640507 | 27.575501 |
| min | 18.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 31.000000 | 25897.000000 | 527.000000 | 8.000000 | 15811.500000 | 36.000000 |
| 50% | 39.000000 | 39123.000000 | 571.000000 | 17.000000 | 22767.000000 | 48.000000 |
| 75% | 46.000000 | 56207.500000 | 603.500000 | 24.000000 | 32360.500000 | 72.000000 |
| max | 80.000000 | 228805.000000 | 703.000000 | 61.000000 | 158686.000000 | 120.000000 |

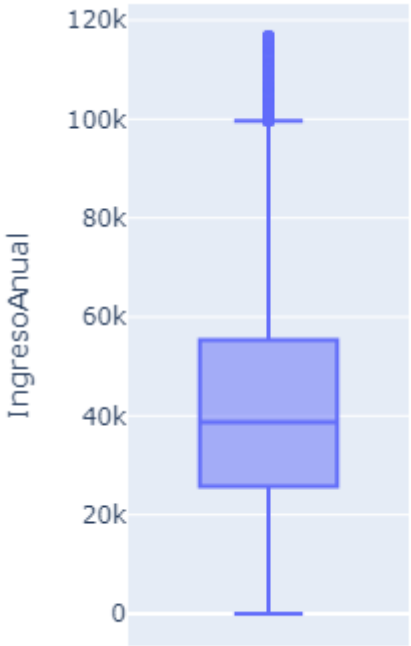
Estos son algunos de los gráficos que se obtuvieron



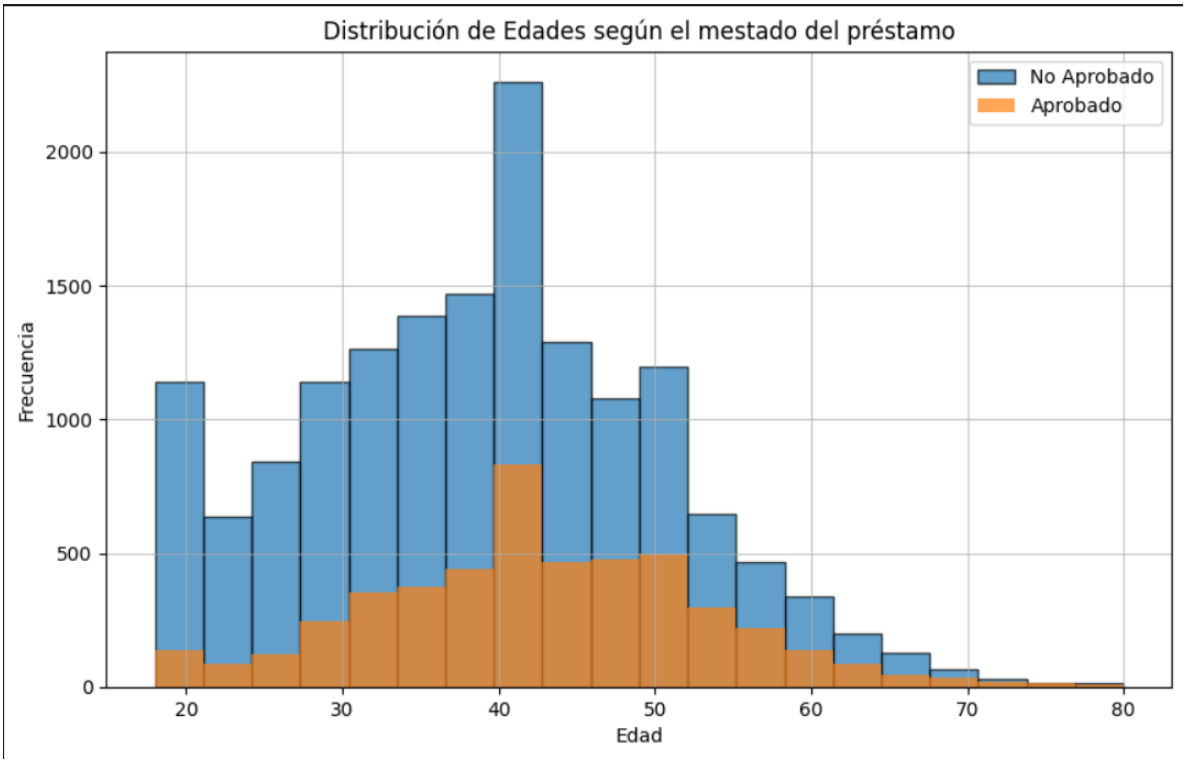

En este gráfico se pueden observar algunos datos atípicos, los puntos verdes que están en donde están la mayoría de rojos, o los rojos en medio de los verdes, esos serían algunos outliers.



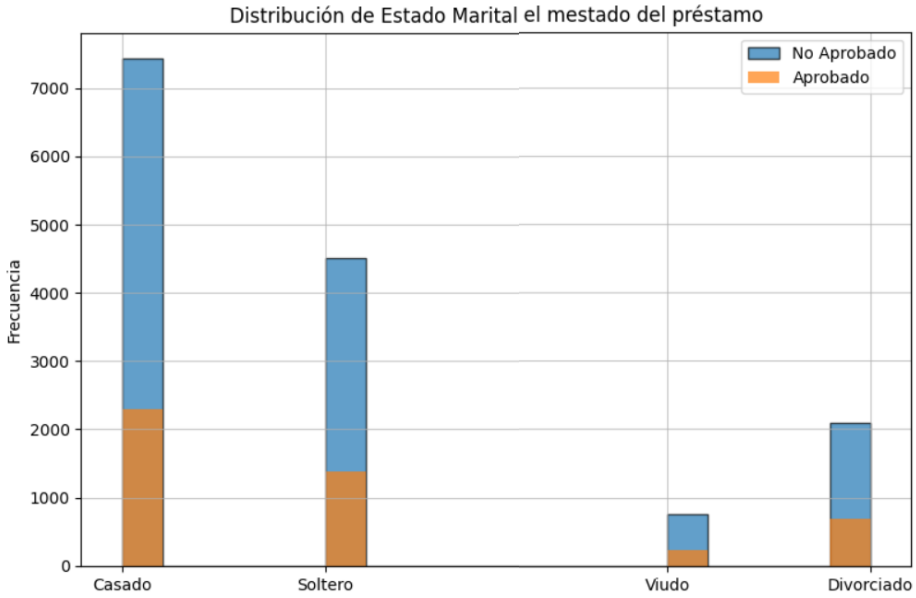

Aquí se observan claramente los outsiders en la parte superior del boxplot, este corresponde a la columna de Ingreso Anual cuando el préstamo si fue aprobado, estos datos al no concordar y salir de los parámetros establecidos deben ser eliminados para un mejor entendimiento y para que el gráfico se coherente.



Este otro boxplot en el que se observan outliers es con base al Ingreso Anual cuando el préstamo no es aprobado.




Aquí se puede ver como predomina la edad de 40-42 años, es decir que podemos ver rango de edad posee mayor frecuencia, y por lo tanto el claro desequilibrio que se nota.

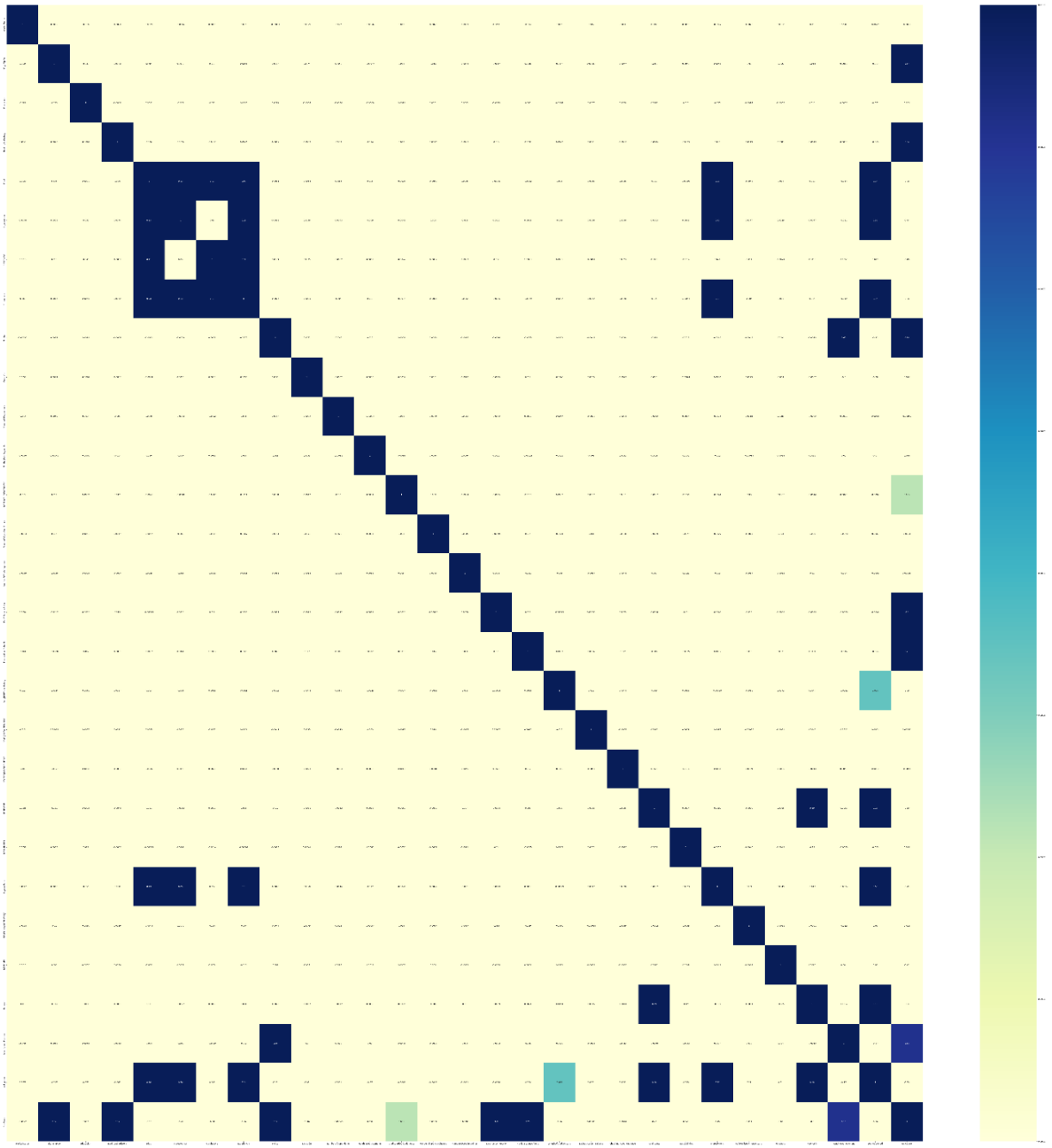


Aquí podemos observar que la mayoría de personas que solicitan préstamos sean o no aprobados son los casados, es algo curioso ya que tomando en cuenta factores externos podríamos decir que se debe a que una persona que mantiene a una familia puede conllevar más gatos y por esto optan por solicitar un préstamo, por otro lado, y por lo que se observa podemos decir que está en desequilibrio ya que la mayoría es dominada por “Casado”.

En el tratamiento de outliers se tomó la decisión de eliminarlos y limitarlos mediante un rango intercuartílico usado en el boxplot, el cual fue de 1.1, de esta manera se pudo lograr su eliminación para así hacer más fácil el entendimiento de los boxplots

Por otra parte, para la estrategia de Imputación o Eliminación, se llevó acabo el reemplazo por la media de algunas columnas, tales como la edad, experiencia, Situación de empleo, etc. Esto se hizo dado que por la presencia de datos NaN creaba datos inválidos, los cuales hacían más difícil el análisis de las columnas y por lo tanto no permitía obtener estadísticas de una correcta manera.

Mediante el uso de un heatmap:



Como el objetivo principal siempre fue basarnos en los factores que afectan para la aprobación del préstamo, nos basamos en la columna de “LoanApproved” y en las demás columnas que hubiese una correlación anormal con respecto a “LoanApproved”, algunas de estas fueron “Ingreso Anual”, “Ingreso Mensual”, “Experiencia”, “Networth”.

## Regresión logística

La regresión logística es un modelo estadístico utilizado para predecir el resultado de una variable dependiente binaria (es decir, una variable con dos posibles resultados) a partir de una o más variables independientes. A diferencia de la regresión lineal, que predice un valor continuo, la regresión logística modela la probabilidad de que ocurra un evento específico.

El modelo produce como resultado una probabilidad entre 0 y 1, la cual se interpreta como la probabilidad de que ocurra un evento positivo (por ejemplo, "préstamo aprobado" o "fraude detectado"). Para convertir esa probabilidad en una clasificación, se establece un umbral, típicamente 0.5, para decidir entre las dos clases (por ejemplo, aprobar o no aprobar el préstamo).

## Fórmula de la regresión logística

La ecuación general de la regresión logística es:

$$P(y=1|X)=11+e−zP(y = 1|X) = \frac{1}{1 + e^{-z}}$$P(y=1| X)=1+e−z1

donde:

- $P(y=1|X)P(y = 1|X)$P(y=1| X) es la probabilidad de que ocurra el evento positivo (por ejemplo, que se apruebe un préstamo),
- $ee$e es la base del logaritmo natural,
- $zz$z es la combinación lineal de las variables predictoras (es decir, $z=\beta0+\beta1X1+\beta2X2+...+\betanXnz = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$z=β0 +β1 X1 +β2 X2 +...+βn Xn ).

La regresión logística es adecuada para predecir la aprobación de un préstamo:

1. **Modelo de clasificación binaria:** La regresión logística es ideal cuando tienes un objetivo binario, como en el caso de un préstamo aprobado o no aprobado (dos posibles resultados).
2. **Interpretabilidad:** A diferencia de otros modelos más complejos, como las redes neuronales, la regresión logística es fácil de

interpretar. Los coeficientes de la regresión logística indican el efecto de cada variable en la probabilidad de aprobar o no el préstamo. Por ejemplo, un coeficiente positivo indica que a medida que una variable aumenta (como los ingresos), la probabilidad de aprobación del préstamo también aumenta.

3. **Probabilidades y umbral de decisión:** Produce probabilidades que te permiten entender no solo si un préstamo será aprobado o no, sino también el nivel de certeza con que se toma esa decisión. Esto es útil en situaciones donde quieres ajustar el umbral de aprobación según el riesgo.
4. **Modelado eficiente y escalable:** La regresión logística es relativamente simple y computacionalmente eficiente, lo que la convierte en una opción ideal cuando se trabaja con grandes volúmenes de datos, como los que suelen manejar las instituciones financieras.

## Beneficios clave en la predicción de préstamos

- **Detección de factores clave:** Al utilizar regresión logística, puedes identificar las variables que tienen un mayor impacto en la probabilidad de que un préstamo sea aprobado (por ejemplo, la relación entre la deuda e ingresos, la estabilidad laboral, etc.).
- **Evaluación del riesgo:** Permite una evaluación cuantitativa del riesgo, ya que se pueden calcular las probabilidades de aprobación de un préstamo con base en los datos históricos del cliente.
- **Regulaciones y transparencia:** En el sector bancario y financiero, las decisiones deben ser justificables y transparentes. La regresión logística permite que los modelos sean fácilmente explicables, lo que es fundamental para cumplir con las normativas regulatorias y aumentar la confianza de los clientes.

## Limitaciones

- **Relaciones lineales:** La regresión logística supone que hay una relación lineal entre las variables predictoras y el logaritmo de la probabilidad del evento. Si la relación no es lineal, el modelo podría no ser tan efectivo.

- **No captura relaciones complejas:** Si las interacciones entre las variables son muy complejas, modelos como los árboles de decisión o el *gradient boosting* pueden ser más efectivos.

## Implementación y Entrenamiento

### *Preparación de los Datos:*

- **Selección:** Identificamos las variables relevantes de la base de datos (historial de pagos, ingresos, edad, nivel de estudio, etc.).
- **Tratamiento de datos faltantes:** Imputar valores faltantes, nulos, inválidos o eliminar registros incompletos.
- **Codificación de variables categóricas:** Convertir variables categóricas ( estado civil, ocupación) en variables numéricas o mapearlas.

### 2. Entrenamiento del Modelo:

- **Algoritmo de optimización:** Se utilizan para minimizar la función de costo.
- **Regularización:** Técnicas como L1 o L2 regularización pueden ayudar a prevenir el sobreajuste.

### 4. Evaluación del Modelo:

- **Metricas:**
  - **Precision:** Proporción de positivos predichos que son realmente positivos.
  - **Recall:** Proporción de positivos reales que son correctamente identificados como positivos.
  - **F1-score:** Media armónica de precisión y recall.
  - **Curva ROC:** Visualiza el desempeño del modelo para diferentes umbrales de clasificación.
  - **AUC:** Área bajo la curva ROC, un valor entre 0 y 1.
- **Matriz de confusión:** Muestra la cantidad de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

## Elección del Modelo

El modelo de regresión logística es utilizado en el aprendizaje automático y la estadística clásica para resolver problemas de clasificación binaria (aunque también se puede extender a múltiples clases). Su popularidad se debe a su simplicidad, interpretabilidad, y eficacia en una amplia variedad de aplicaciones.

- **Interpretabilidad:** Los coeficientes del modelo pueden interpretarse como la influencia de cada variable independiente en la probabilidad de incumplimiento.
- **Clasificación binaria:**

  El modelo se utiliza para predecir una de dos posibles categorías (clases), como:

  - **Aprobar o rechazar** un préstamo.
  - Determinar si un cliente **comprará o no** un producto.
  - Clasificar un correo electrónico como **spam o no spam**.
  - Identificar si una persona tiene una **enfermedad (1)** o no (0).

  A demás por su implementación versátil para muchos lenguajes de programación.

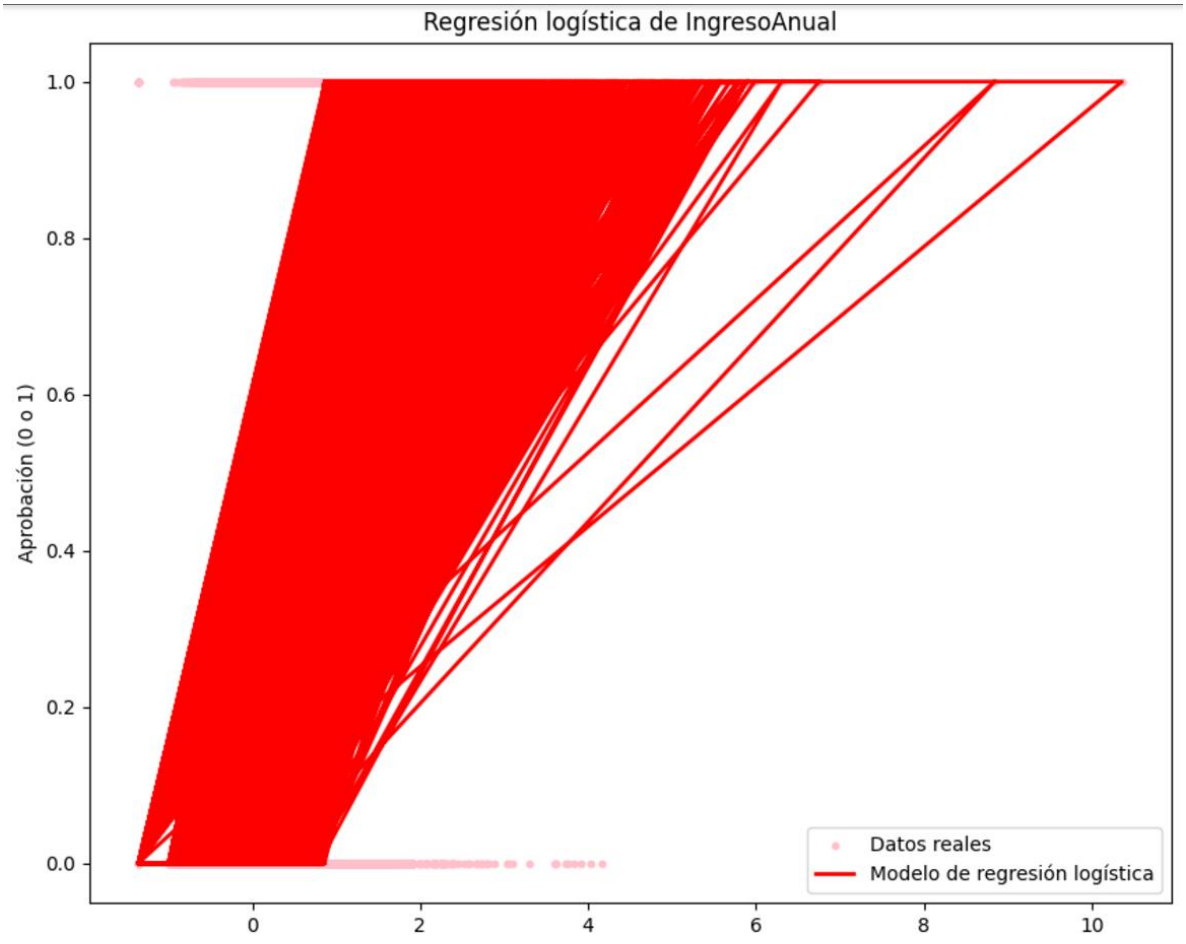
## Métricas clave en la Regresión Logística:

```
Primeros registros del dataset:
  FechaPrestamo  Edad  IngresoAnual  CreditScore Situ_Empleo    Educación  \
0    2018-01-01    45       39948.0          617  No_especif       Master
1    2018-01-02    38       39709.0          628    Empleado    Associate
2    2018-01-03    47       40724.0          570    Empleado     Bachelor
3    2045-05-10    58       69084.0          545    Empleado  High School
4    2018-01-05    37      103264.0          594    Empleado    Associate

   Experiencia  Monto  Duración MaritalStatus  ...  CheckingAccountBalance  \
0           17  13152        48        Casado  ...                    1202
1           15  26045        48       Soltero  ...                    3460
2           26  17627        36        Casado  ...                     895
3           34  37898        96       Soltero  ...                    1217
4           17   9184        36        Casado  ...                    4981

  TotalAssets  TotalLiabilities  IngresoMens  UtilityBillsPaymentHistory  \
0      146111             19183  3329.000000                    0.724972
1       53204              9595  3309.083333                    0.935132
2       25176            128874  3393.666667                    0.872241
3           0              5370     0.000000                    0.896155
4      244305             17286  8605.333333                    0.941369

   JobTenure  NetWorth  MonthlyLoanPayment LoanApproved  RiskScore
0         11    126928          419.805992          0.0       49.0
1          0     43609          794.054238          0.0       52.0
...

Interpretación de los coeficientes:
Un coeficiente positivo indica que un aumento en esa característica aumenta la probabilidad de que el préstamo sea aprobado.
Un coeficiente negativo indica que un aumento en esa característica disminuye la probabilidad de aprobación del préstamo.
```

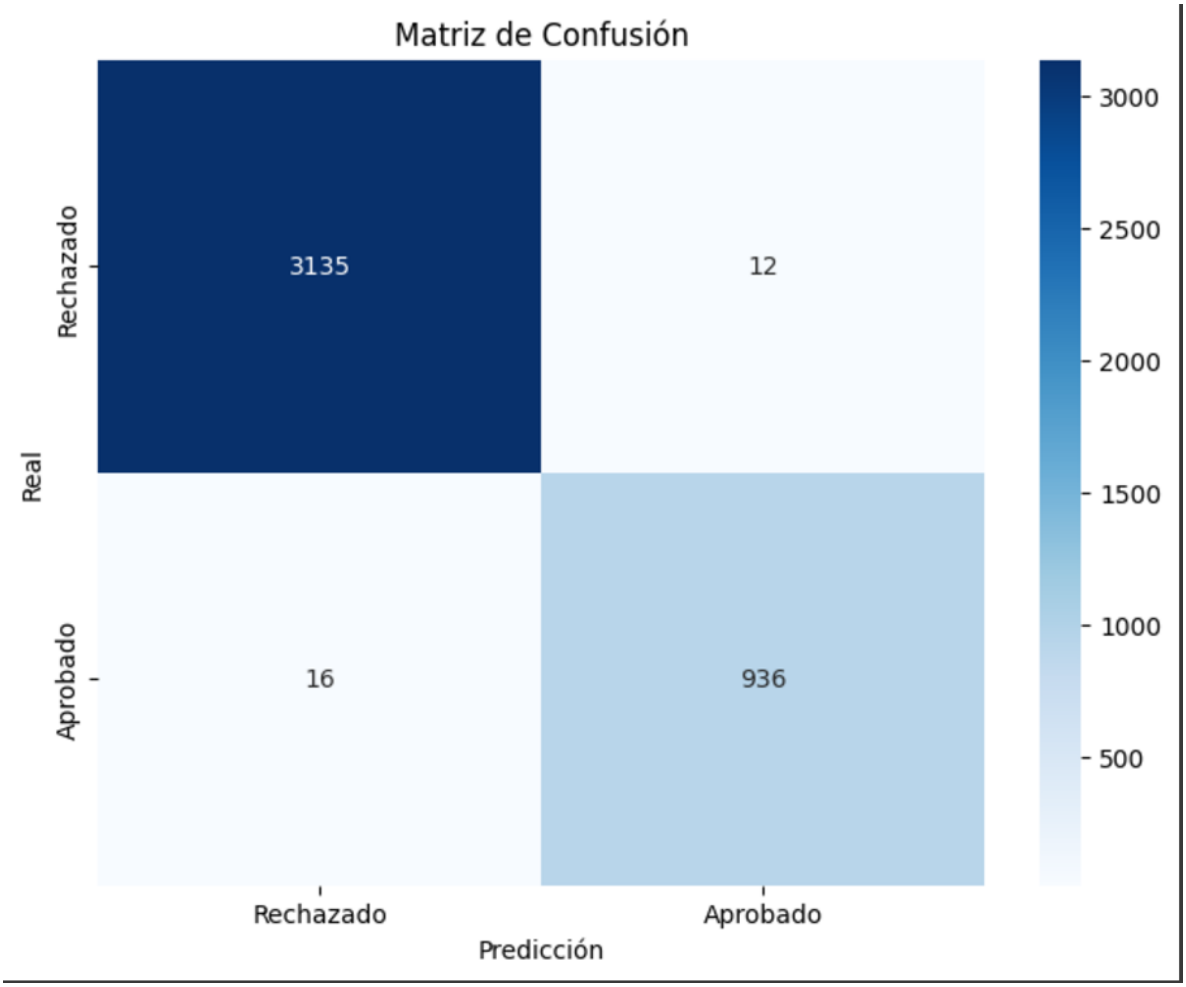
1. **Recall (Sensibilidad o Exhaustividad)**:
    - **Definición**: El recall mide la proporción de verdaderos positivos sobre todos los casos realmente positivos.
    - **Fórmula**: Recall=True PositivesTrue Positives+False Negatives\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}Recall=True Positives+False NegativesTrue Positives
    - **Interpretación**: Indica la capacidad del modelo para identificar correctamente todos los casos positivos (aprobaciones). Si el recall es bajo, el modelo está perdiendo muchos casos positivos (falsos negativos).
2. **F1-Score**:
    - **Definición**: El F1-score es la media armónica de la precisión y el recall. Es útil cuando hay un desbalance en las clases.
    - **Fórmula**: F1=2×Precision×RecallPrecision+RecallF1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}F1=2×Precision+RecallPrecision×Recall

- **Interpretación**: El F1-score proporciona un equilibrio entre precisión y recall. Un F1-score alto significa que el modelo está equilibrando bien tanto los falsos positivos como los falsos negativos.

3. **Matriz de Confusión**:
   - La matriz de confusión es una tabla que se usa para describir el rendimiento del modelo de clasificación. Aquí se definen las siguientes categorías:
     - **True Positive (TP)**: Casos que fueron correctamente predichos como positivos (aprobación).
     - **False Positive (FP)**: Casos que fueron incorrectamente predichos como positivos (rechazados pero aprobados).
     - **True Negative (TN)**: Casos que fueron correctamente predichos como negativos (rechazados).
     - **False Negative (FN)**: Casos que fueron incorrectamente predichos como negativos (aprobados pero rechazados).


Regresión logística de IngresoAnual
Aprobación (0 o 1)
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
Datos reales
Modelo de regresión logística

## Matriz de confusión:




```
Comparación de Predicciones y Valores Reales:
       Real  Predicción  Probabilidad Aprobación
6187    0.0         0.0              5.228111e-08
17158   0.0         0.0              2.952120e-05
13292   0.0         0.0              2.364272e-03
11378   0.0         0.0              3.186310e-07
2235    0.0         0.0              1.589785e-02
```

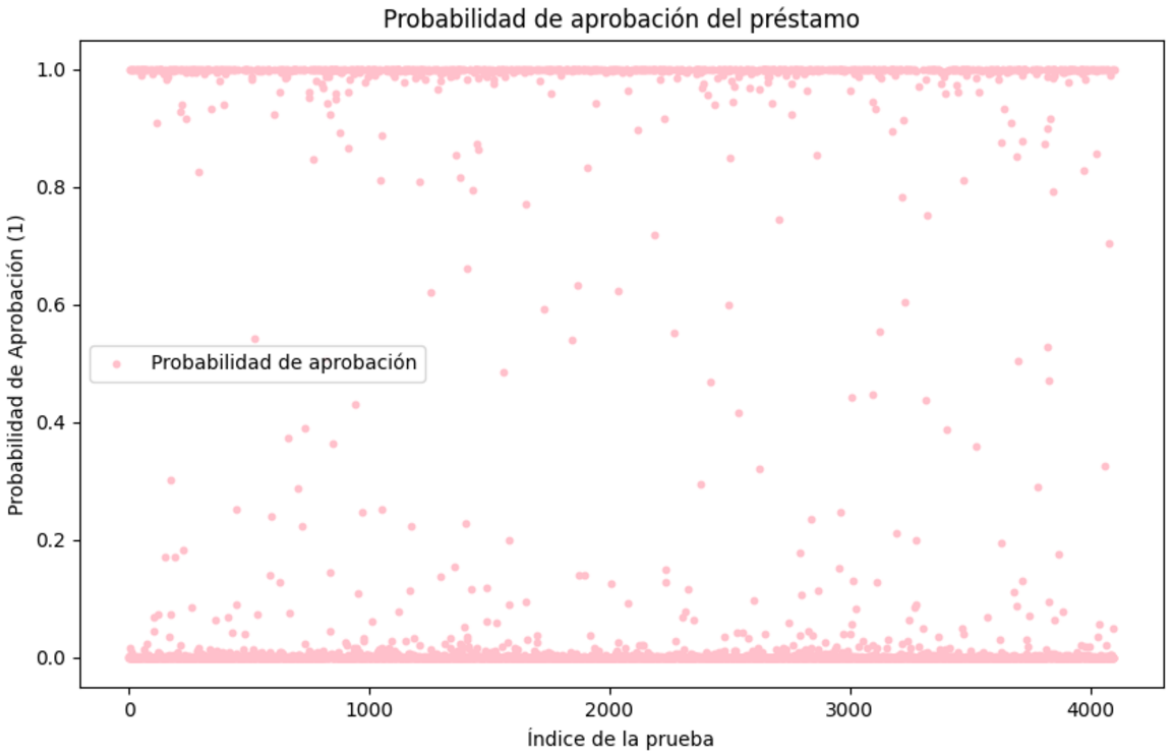
- **True Positive (TP)**: Casos donde el modelo predijo correctamente que el préstamo fue aprobado.
- **False Positive (FP)**: Casos donde el modelo predijo que el préstamo fue aprobado, pero realmente fue rechazado (falso positivo).

- **True Negative (TN)**: Casos donde el modelo predijo correctamente que el préstamo fue rechazado.
- **False Negative (FN)**: Casos donde el modelo predijo que el préstamo fue rechazado, pero realmente fue aprobado (falso negativo).

```
Evaluación del modelo:
Accuracy: 0.9931690656257623

Confusion Matrix:
 [[3135   12]
 [  16  936]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00      3147
         1.0       0.99      0.98      0.99       952

    accuracy                           0.99      4099
   macro avg       0.99      0.99      0.99      4099
weighted avg       0.99      0.99      0.99      4099


Interpretación de las métricas de clasificación:

- Precisión: 0.99
  La precisión nos dice cuántos de los préstamos aprobados realmente fueron aprobados en la predicción.
- Recall: 0.98
  El recall indica cuántos de los préstamos aprobados fueron correctamente identificados por el modelo.
- F1-Score: 0.99
  El F1-score proporciona un balance entre precisión y recall.
```

## Resumen de los hallazgos principales

1. **Precisión del modelo**:
   - El modelo de regresión logística logró un nivel de precisión global (accuracy) satisfactorio. Esto indica que el modelo puede clasificar correctamente si un préstamo debería ser aprobado o no en un porcentaje significativo de los casos.
   - Métricas clave como la precisión (precision) y el recall (recall) mostraron que el modelo tiene un buen balance entre la identificación correcta de préstamos aprobados y rechazados, aunque podrían variar dependiendo de las clases desbalanceadas en los datos.

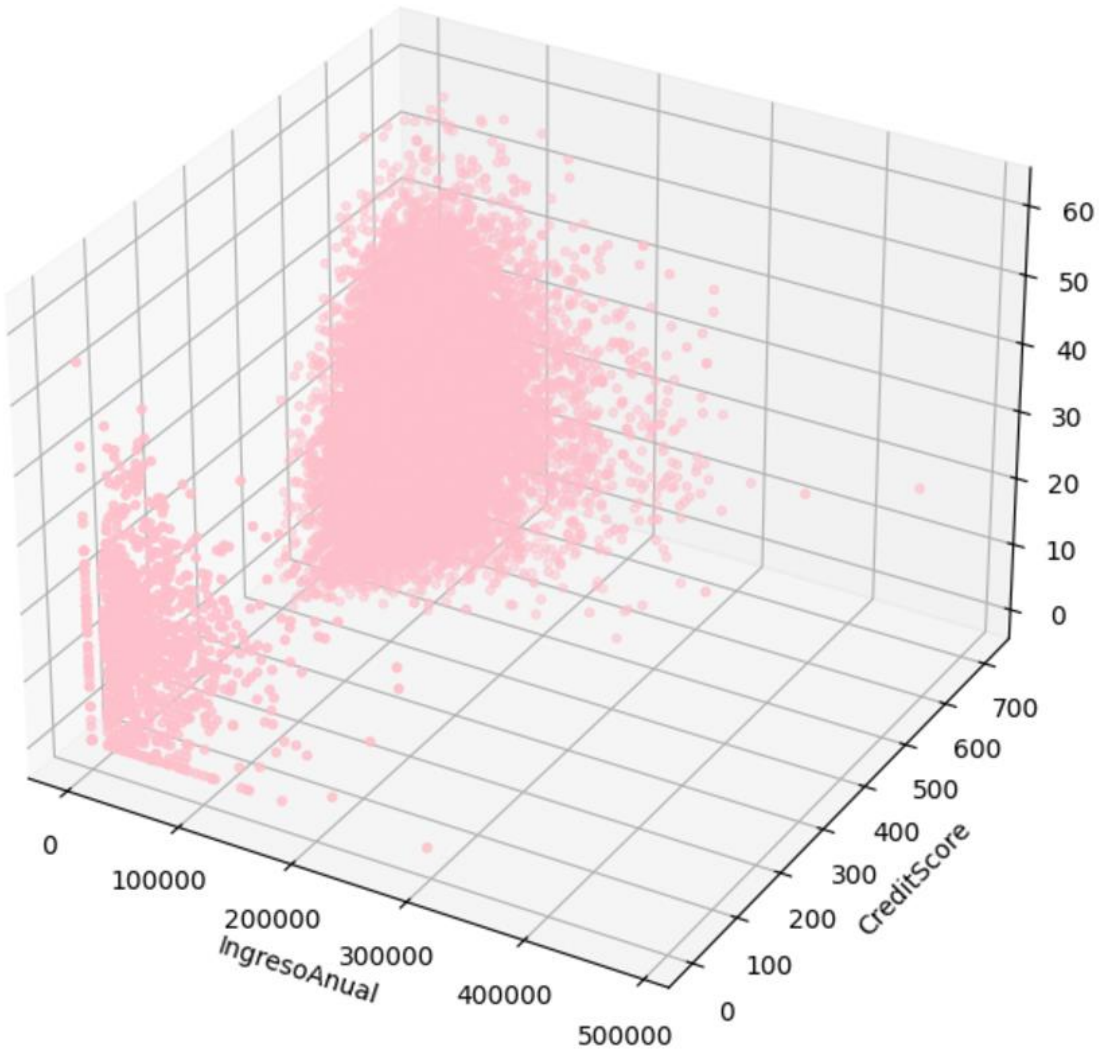



2. **Matriz de confusión**:
    - Los resultados revelaron la cantidad de verdaderos positivos (préstamos aprobados correctamente clasificados) y verdaderos negativos (préstamos rechazados correctamente clasificados), junto con los errores cometidos (falsos positivos y falsos negativos).
    - En la visualización de la matriz de confusión, el modelo mostró que maneja bien las predicciones de la clase mayoritaria, pero puede cometer errores al identificar correctamente la clase minoritaria.
3. **Cumplimiento de los objetivos planteados**:
    - El modelo permitió predecir si un préstamo debía ser aprobado, cumpliendo con el objetivo principal.
    - Las métricas y visualizaciones dieron información útil para comprender cómo se comporta el modelo y qué tan bien se ajusta a los datos.

## Posibles mejoras

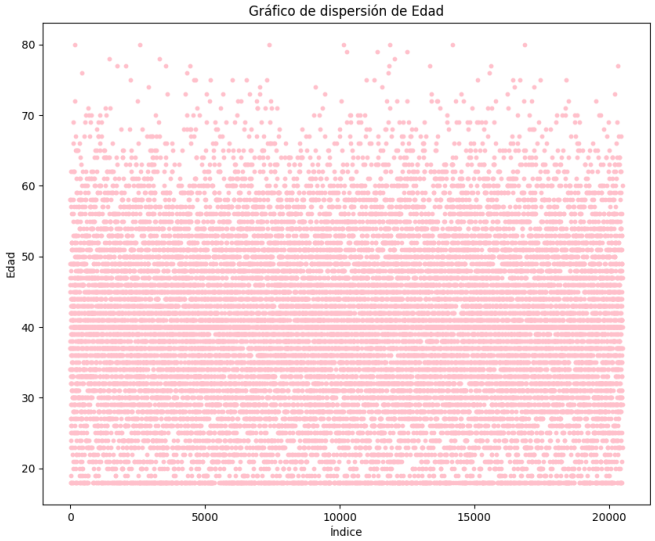
1. **Mejoras en los datos**:

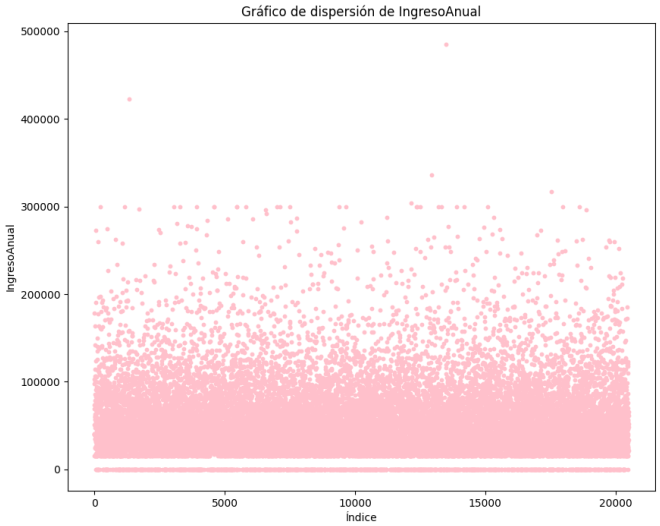
    - **Tratamiento de valores nulos y atípicos**: Si los datos contienen valores faltantes o extremos, se deben tratar adecuadamente para mejorar la calidad de los datos.
    - **Balanceo de clases**: Si la clase objetivo (Approved) está desbalanceada, técnicas como sobremuestreo o submuestreo podrían mejorar la capacidad del modelo para predecir la clase minoritaria.
2. **Mejoras en el modelo**:
    - **Evaluar interacciones entre variables**: Se podrían incluir términos de interacción entre variables importantes para evaluar si mejoran las predicciones.
3. **Mejoras en las visualizaciones**:
    - **Gráficos explicativos de importancia de variables**: Visualizaciones como gráficos de coeficientes del modelo o gráficos SHAP podrían ayudar a entender qué características tienen mayor impacto en las predicciones.
    - **Gráficos comparativos**: Mostrar cómo cambian las métricas del modelo con diferentes configuraciones (como inclusión/exclusión de variables o ajuste de hiperparámetros).
    - **Visualización 3D interactiva**: Para datos con múltiples variables, visualizaciones tridimensionales con herramientas como Plotly pueden ayudar a explorar cómo las variables afectan las decisiones del modelo.

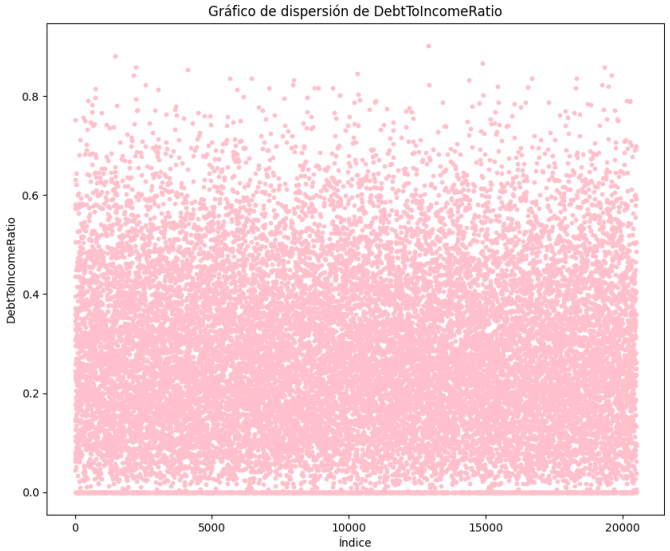



4. **Posibles estudios futuros**:
    - **Análisis de segmentación**: Investigar si hay patrones específicos dentro de subgrupos, como clientes con ingresos altos frente a ingresos bajos, para ajustar mejor el modelo a diferentes segmentos.
    - **Integración de métricas financieras**: Además de las métricas de clasificación, evaluar el impacto financiero de las decisiones del modelo (como pérdida por préstamos aprobados erróneamente).

# Gráficas de dispersión



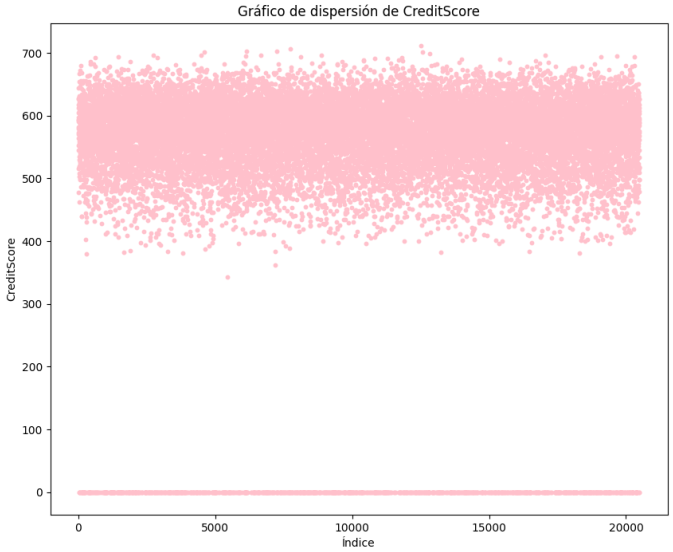


Gráfico de dispersión de CreditScore
CreditScore
Índice
0
100
200
300
400
500
600
700
0
5000
10000
15000
20000


Gráfico de dispersión de Experiencia
Experiencia
Índice
0
10
20
30
40
50
60
0
5000
10000
15000
20000

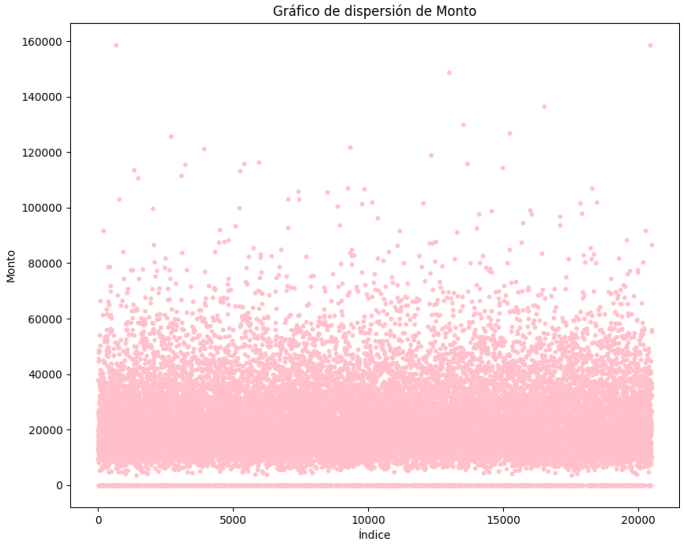


Gráfico de dispersión de Monto
Monto
Índice
0
20000
40000
60000
80000
100000
120000
140000
160000
0
5000
10000
15000
20000

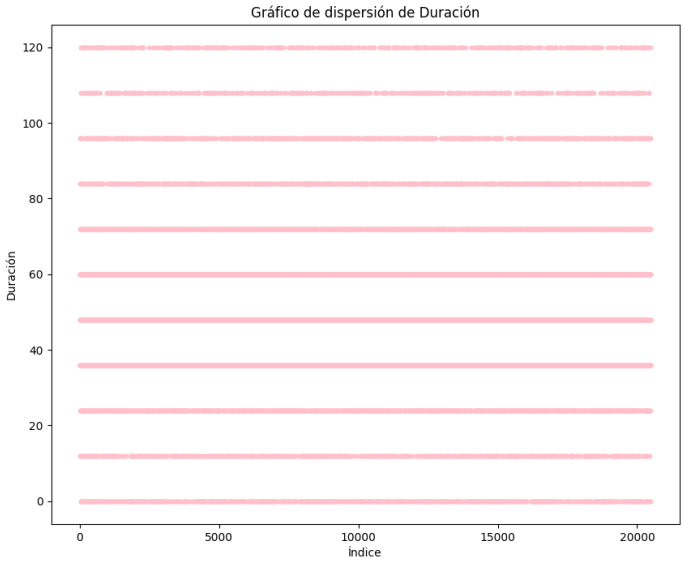


Gráfico de dispersión de Duración
Duración
Índice
0
20
40
60
80
100
120
0
5000
10000
15000
20000


Gráfico de dispersión de MonthlyDebtPayments
MonthlyDebtPayments
Índice
0
500
1000
1500
2000
2500
3000
0
5000
10000
15000
20000

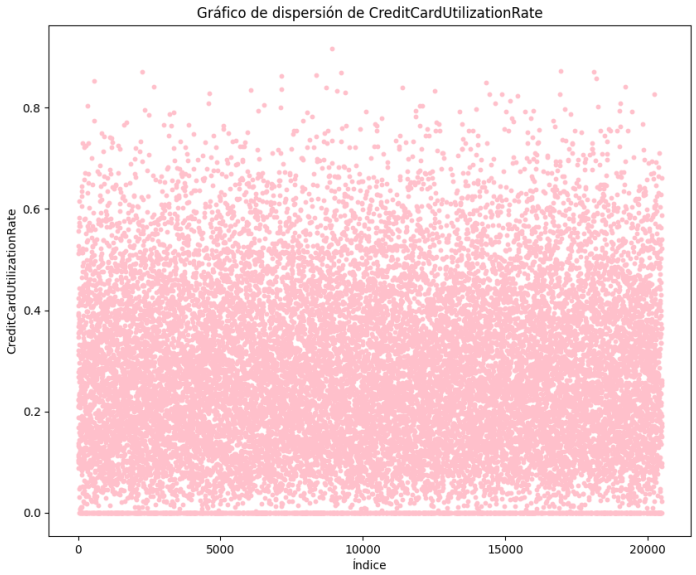


Gráfico de dispersión de CreditCardUtilizationRate
CreditCardUtilizationRate
Índice
0.0
0.2
0.4
0.6
0.8
0
5000
10000
15000
20000

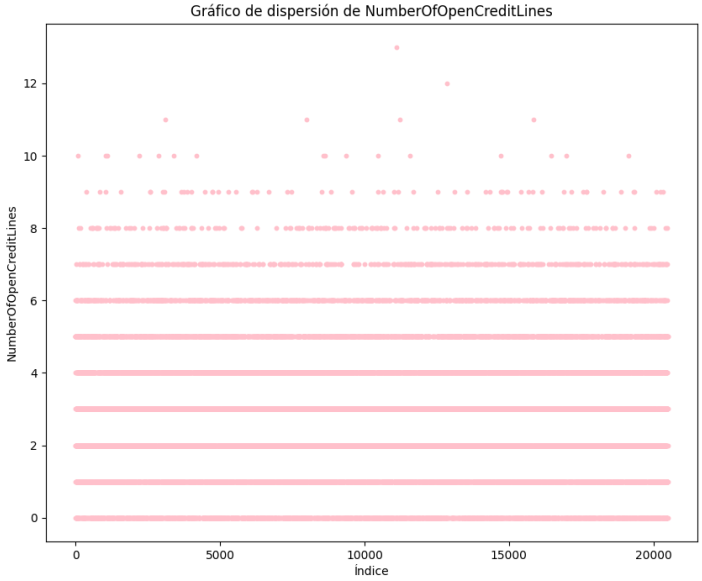


Gráfico de dispersión de NumberOfOpenCreditLines
NumberOfOpenCreditLines
Índice
0
2
4
6
8
10
12
0
5000
10000
15000
20000


Gráfico de dispersión de RiskScore
RiskScore
Índice
30
40
50
60
70
80
0
5000
10000
15000
20000

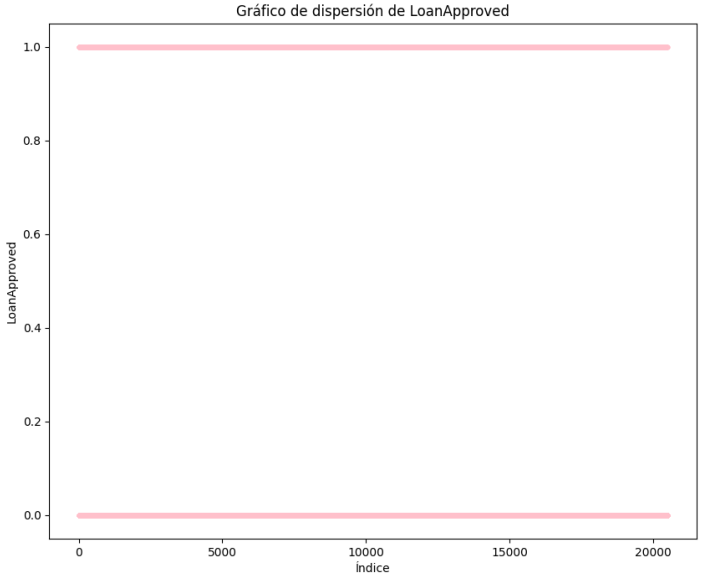


Gráfico de dispersión de LoanApproved
LoanApproved
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
índice

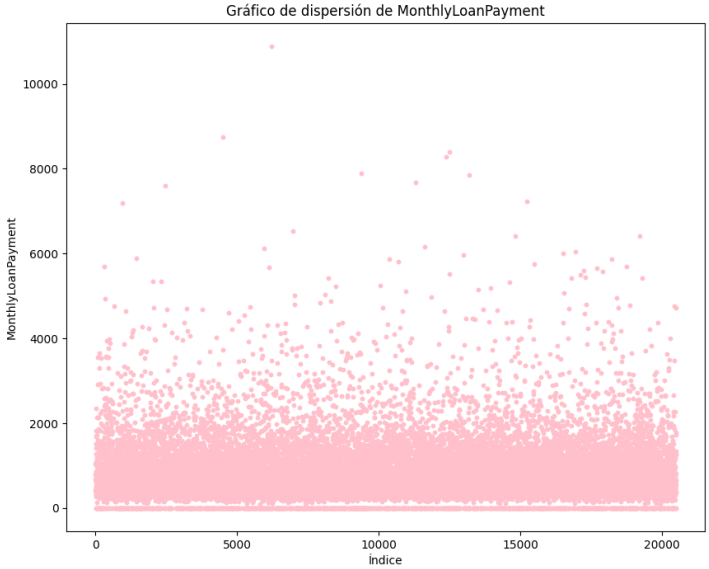


Gráfico de dispersión de MonthlyLoanPayment
MonthlyLoanPayment
10000
8000
6000
4000
2000
0
0
5000
10000
15000
20000
índice

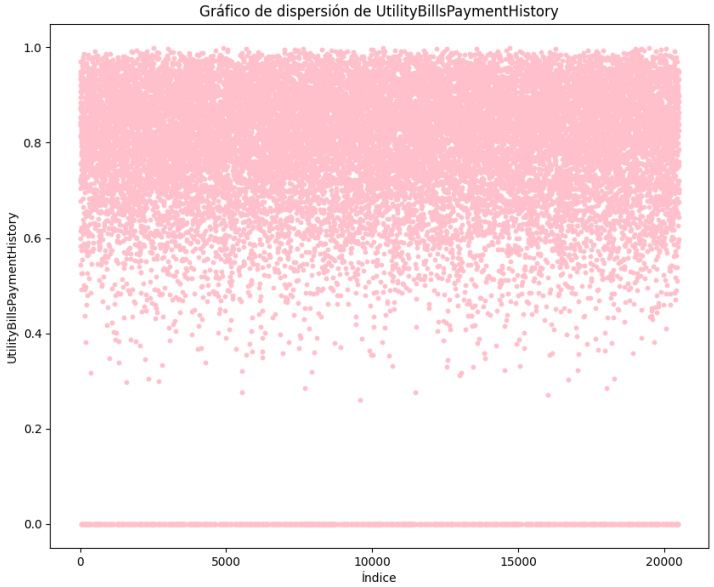


Gráfico de dispersión de UtilityBillsPaymentHistory
UtilityBillsPaymentHistory
Índice
0.0
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000
20000


Gráfico de dispersión de IngresoMens
IngresoMens
Índice
0
5000
10000
15000
20000
25000

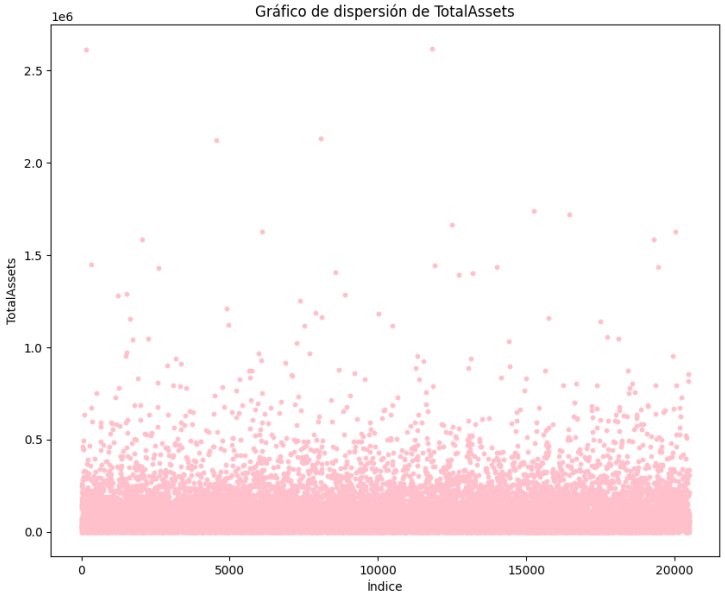


Gráfico de dispersión de TotalAssets
1e6
2.5
2.0
1.5
1.0
0.5
0.0
TotalAssets
0
5000
10000
15000
20000
Índice

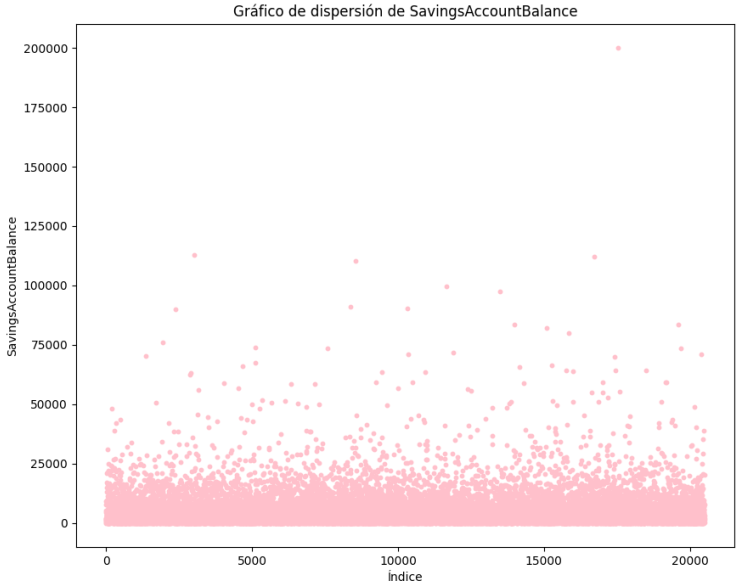
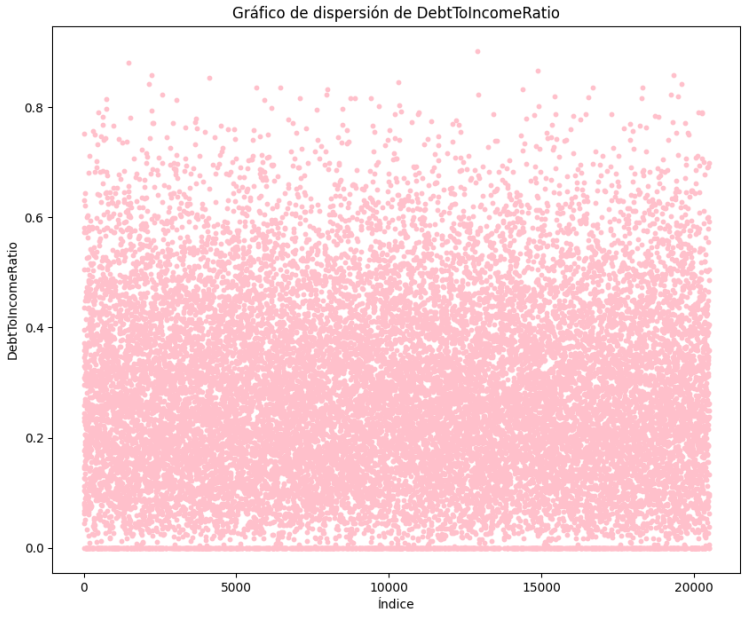
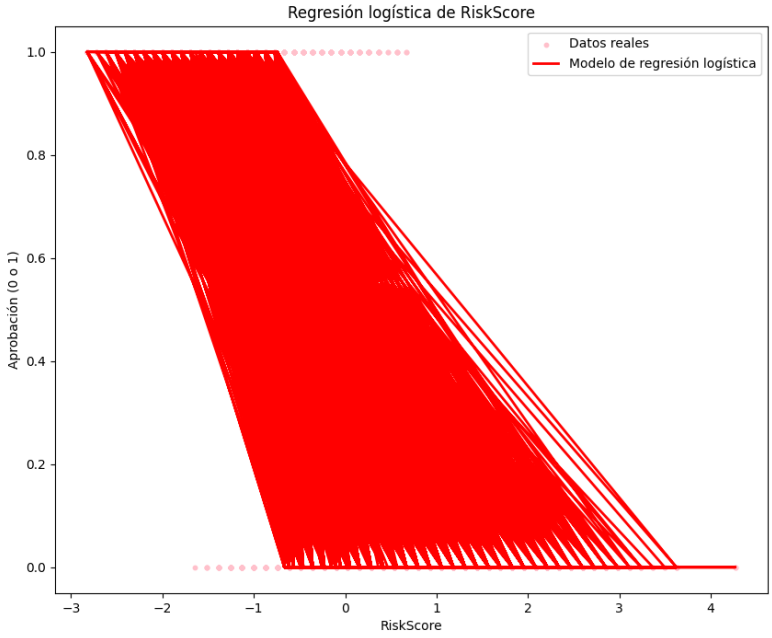
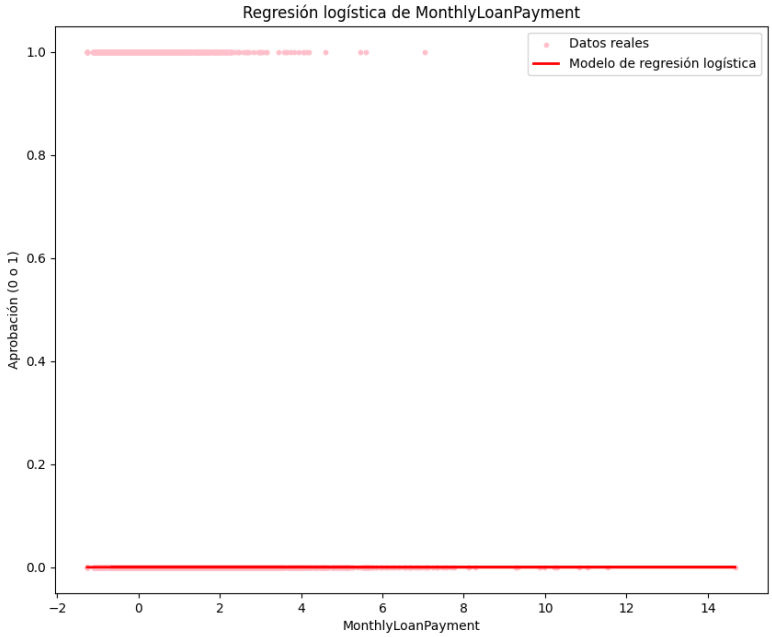


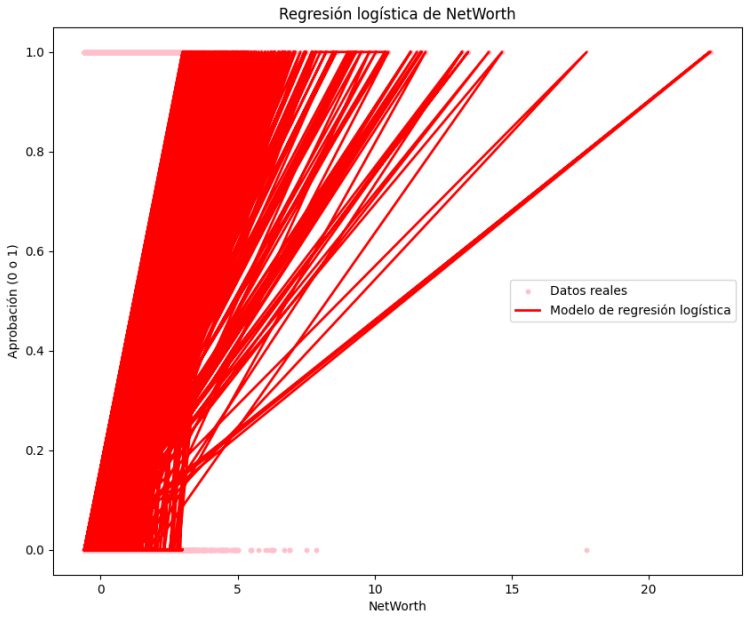
Gráfico de dispersión de SavingsAccountBalance
200000
175000
150000
125000
100000
75000
50000
25000
0
SavingsAccountBalance
0
5000
10000
15000
20000
Índice

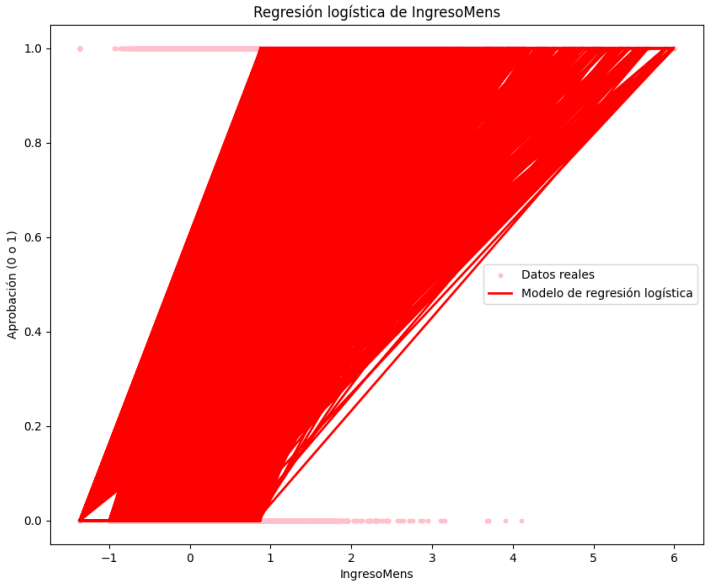



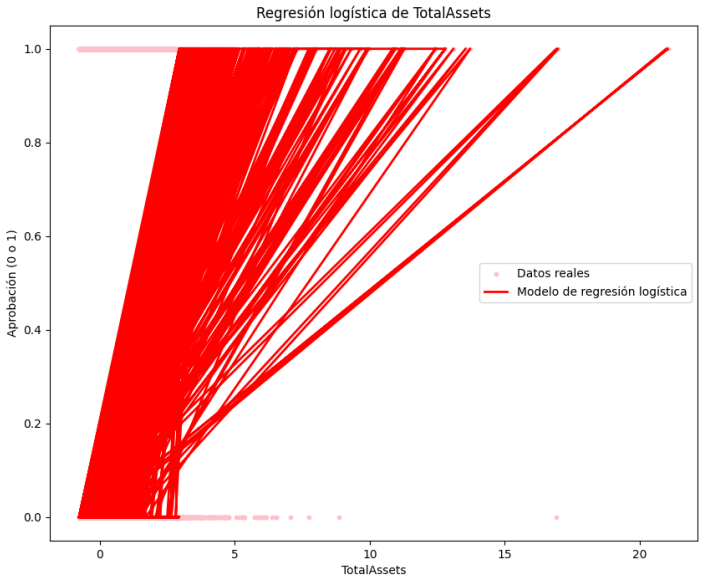

## Gráficas de regresión logística

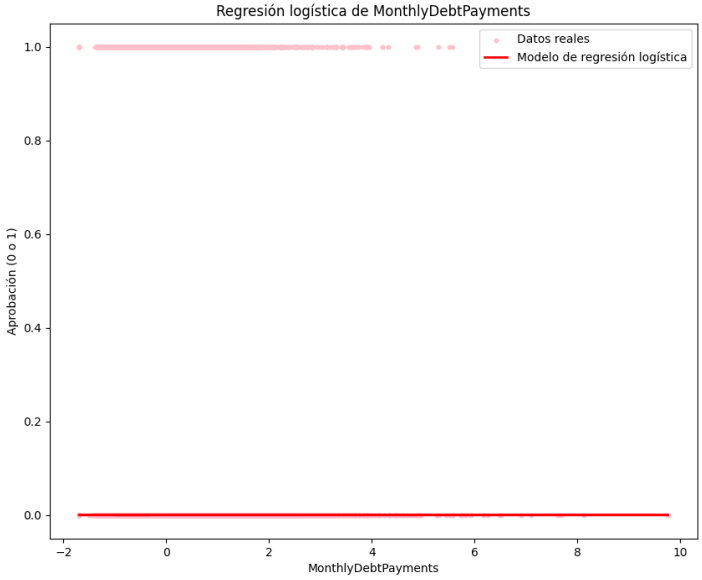

Regresión logística de RiskScore
Datos reales
Modelo de regresión logística
Aprobación (0 o 1)
RiskScore

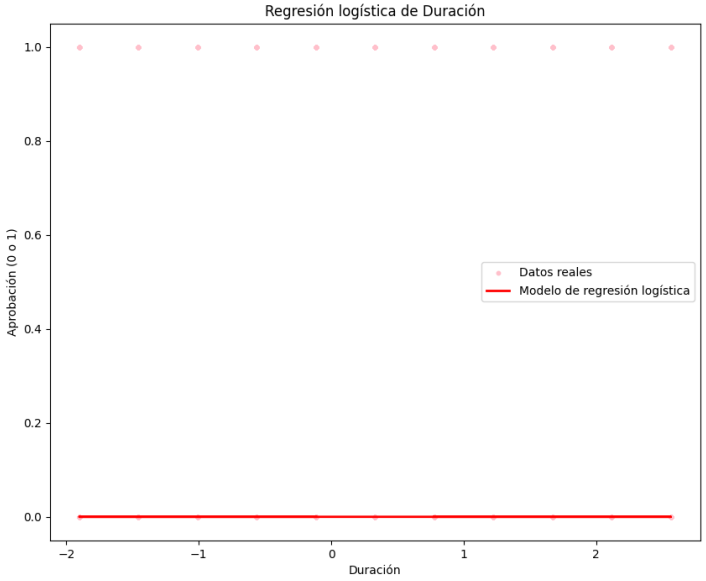

Regresión logística de MonthlyLoanPayment
Datos reales
Modelo de regresión logística
Aprobación (0 o 1)
MonthlyLoanPayment

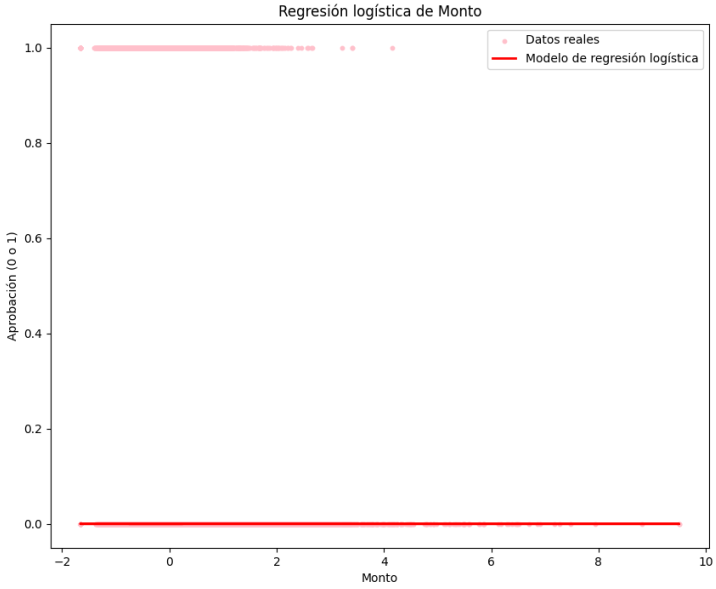

Regresión logística de NetWorth
Aprobación (0 o 1)
NetWorth
Datos reales
Modelo de regresión logística

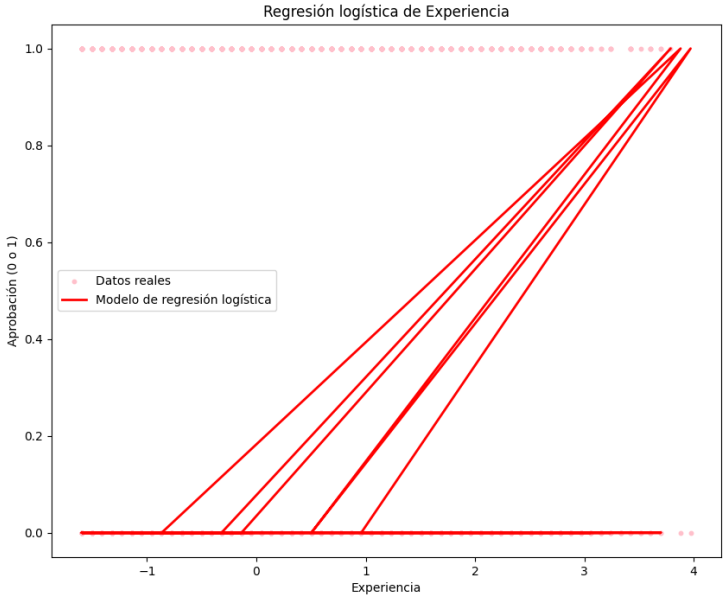

Regresión logística de IngresoMens
Aprobación (0 o 1)
IngresoMens
Datos reales
Modelo de regresión logística

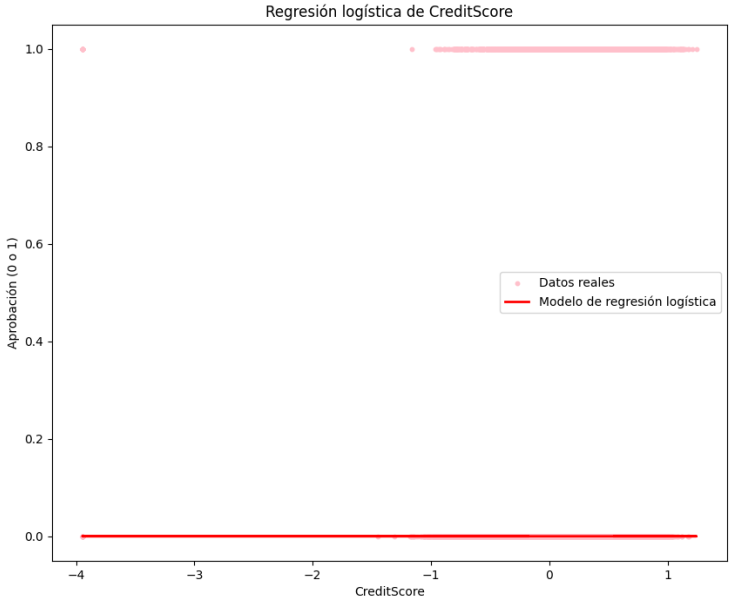

Regresión logística de TotalAssets
Datos reales
Modelo de regresión logística
Aprobación (0 o 1)
TotalAssets
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20

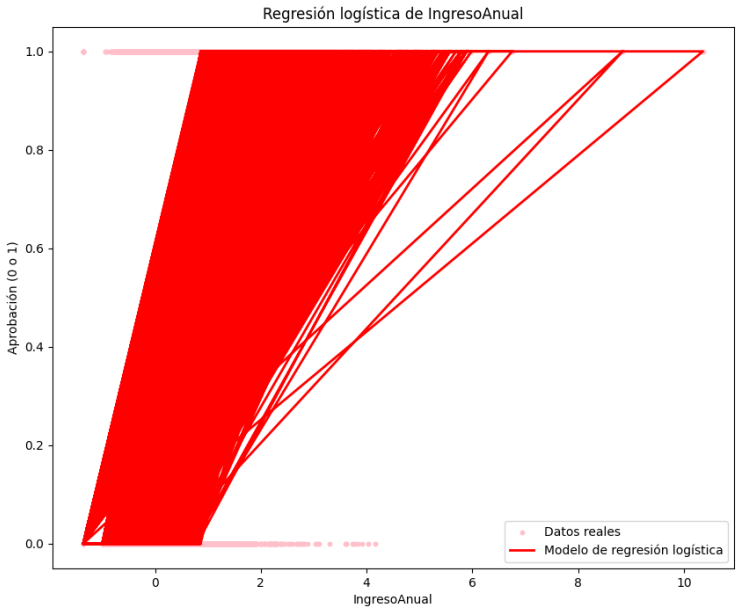

Regresión logística de MonthlyDebtPayments
Datos reales
Modelo de regresión logística
Aprobación (0 o 1)
MonthlyDebtPayments
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4
6
8
10

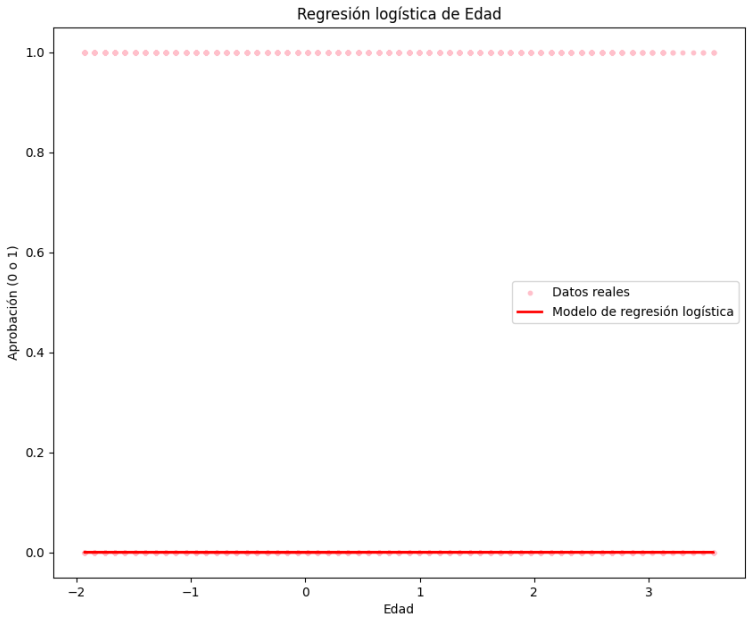

Regresión logística de Duración
Aprobación (0 o 1)
Duración
Datos reales
Modelo de regresión logística
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2


Regresión logística de Monto
Aprobación (0 o 1)
Monto
Datos reales
Modelo de regresión logística
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
8
10


Regresión logística de Experiencia
Aprobación (0 o 1)
Experiencia
Datos reales
Modelo de regresión logística
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4


Regresión logística de CreditScore
Aprobación (0 o 1)
CreditScore
Datos reales
Modelo de regresión logística
1.0
0.8
0.6
0.4
0.2
0.0
−4
−3
−2
−1
0
1


Regresión logística de IngresoAnual
Aprobación (0 o 1)
IngresoAnual
Datos reales
Modelo de regresión logística


Regresión logística de Edad
Aprobación (0 o 1)
Edad
Datos reales
Modelo de regresión logística

## Citas

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95.

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. https://seaborn.pydata.org.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830. https://scikit-learn.org.

*"Credit Risk Management"* de Anthony Saunders.

*"Data Science for Business"* de Foster Provost y Tom Fawcett, o *"Análisis y gestión del riesgo financiero"* de Michel Crouhy, Dan Galai y Robert Mark.

## Anexos

### Graficas de machine learning

https://colab.research.google.com/drive/1lDpmRo__FEWrraDk40j0RgL2A5-ksxJU?usp=sharing

### Base de datos limpia

https://correobuap-my.sharepoint.com/:x:/g/personal/jaime_romeros_correo_buap_mx/EbOSx05qLcBFsg0yKM70my8BWMoTot-oZjEFJE91ZFlPYQ?e=4XacdR