UNIVERSIDAD DE COLIMA

Diplomado profesional en línea

ANALÍTICA Y CIENCIA DE DATOS (Excel, SQL, R, PYTHON, POWER BI Y TABLEAU)

Proyecto Final Informe - Analitica y ciencia de datos

Presentan:

Equipo 12

Francisco Javier Perez Cruz

Pedro Oswaldo Espíritu García

Colima, Col; 12 de Octubre de 2024

ÍNDICE

1. Selección y Justificación de la Base de Datos

La base de datos elegida se rescató del repositorio de GitHub del Dr. Gastón Sánchez, quien explica a grandes rasgos que es un proyecto que se enfoca en evaluar el historial crediticio o, como se le conoce en inglés, Credit Scoring, esta base de datos contiene valores de comportamiento financiero en encuestados de varias edades y se compone de 4,455 filas y 14 columnas, que se explicarán con más detalle en el siguiente punto (2. Preparación y Limpieza de los Datos).

Como interés personal y profesional, se puede decir que esta elección fue hecha ya que el manejo y los hábitos financieros se pueden analizar en un contexto laboral real, especialmente si se trabaja en una institución bancaria o si las aspiraciones personales van orientadas al manejo de grandes volúmenes de datos. Con los campos incluidos, se espera realizar una limpieza de datos profunda y un análisis completo que nos ayudará a comprender y destacar lo más importante de esta base de datos.

2. Preparación y Limpieza de los Datos

La base de datos con la que se trabajó, que se encuentra en formato .csv, cuenta de forma cruda con un total de 4,455 filas y 14 columnas. Cada columna es una variable, y estas pueden ser tanto categóricas como numéricas. En su primera versión, todas las variables son numéricas; sin embargo, dependiendo de su descripción, a cada valor se le asignó un concepto, lo cual se demuestra en el siguiente listado que muestra el significado de cada variable y a qué tipo corresponde:

1. Status: estatus del crédito (variable categórica)

- 1 = Bueno
- 2 = Malo

2. Seniority: antigüedad en el trabajo (variable numérica)

3. Home: tipo de propiedad de la casa (variable categórica)

- 1 = Renta
- 2 = Propia
- 3 = Privada
- 4 = No contestaron
- 5 = De sus padres
- 6= Otro

4. Time: tiempo del préstamo solicitado (variable numérica)

5. Age: edad del cliente (variable numérica)

6. Marital: situación conyugal (variable categórica)

- 1 = Solter@
- 2 = Casad@
- 3 = Viud@
- 4 = Separad@
- 5 = Divorciad@

7. Records: existencia de registros (variable categórica)

- 1 = No hay registros
- 2 = Sí hay registros

8. Job: tipo de empleo (variable categórica)

- 1 = Fijo
- 2 = Medio tiempo
- 3 = Freelancer
- 4 = Otro

9. Expenses: Gasto (variable numérica)

10. Income: Ingreso (variable numérica)

11. Assets: Cantidad de activos (variable numérica)

12. Debt: Deuda (variable numérica)

13. Amount: cantidad solicitada de préstamo (variable numérica)

14. Price: precio del bien (variable numérica)

Procedimos a manipular la base de datos con el software de ***Microsoft Excel***, empezando con la limpieza de los datos, nos percatamos de que hacía falta una columna para darle una identificación a la persona encuestada, por lo tanto, para facilitar el recuento de las respuestas, agregamos una columna llamada "ID Encuestado", la cual, lógicamente, se arrastró a lo largo de las filas, obteniendo que se encuestaron 4,455 personas.

Otro dato espurio que encontramos en las variables categóricas fue el número 0, que no correspondía a ningún factor de los mencionados anteriormente. Al analizar esto, nos dimos cuenta de que este 0 se asignaba a las personas que omitían la pregunta o para quienes ninguna respuesta encajaba, por lo que decidieron no responder, en estos casos, se utilizó la función de Excel “buscar y reemplazar”, donde los 0 se reemplazaron con "NA" de "No Aplica".

Otro error que encontramos fue en las columnas de Income, Expenses y Assets, donde ocurría lo mismo: algunos encuestados, por razones personales, decidieron

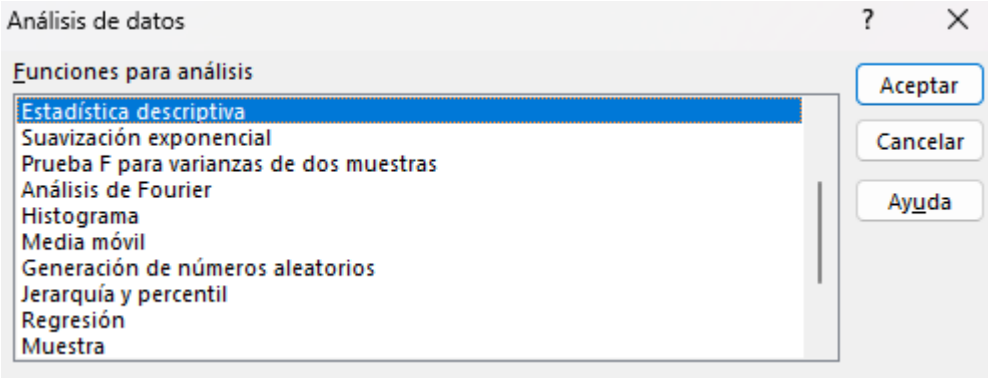
no hacer pública esta información, a la cual el autor asignó el valor de 999999, para que estos valores no interfirieran más adelante con el cálculo de la estadística descriptiva, se les dio el valor de 0.

Continuando con la limpieza de la base de datos, para facilitar su comprensión, se reemplazaron las variables categóricas a las que se les había asignado un valor numérico, estas variables ya se habían identificado previamente, siendo las que constan de factores: Status, Home, Marital, Records y Job.

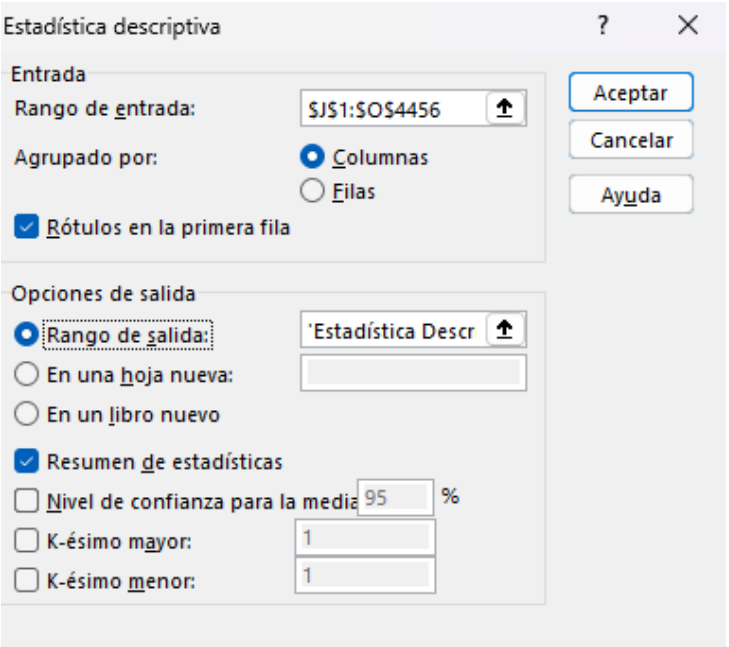
3. Análisis Exploratorio de Datos (EDA)

Para el análisis exploratorio también utilizamos Microsoft Excel. La razón por la que elegimos este software es que nos pareció el más completo debido a la gran cantidad de funciones que contiene.

Una vez limpios los datos, se guardó un archivo llamado CleanCreditScoring. Sobre este archivo, comenzamos a trabajar con la función *"Análisis de datos"*, específicamente con la herramienta de *"Estadística descriptiva"*.



Esta función se utilizó con la siguiente configuración para que nos arrojara las medidas de tendencia en otra página que le nombramos *"Estadística Descriptiva"*, el rango de entrada fueron las variables numéricas de Expenses, Income, Assets, Debt, Amount, Price y Age.



Los resultados obtenidos ya con el formato correspondiente y plasmados en una tabla, son los siguientes:

Medida	*Expenses*	*Income*	*Assets*	*Debt*	*Amount*	*Price*	*Age*
Media	$ 55.57	$ 129.57	$ 5,346.43	$ 341.56	$ 1,039.02	$ 1,462.88	37.07766554
Error típico	0.29	1.30	172.67	18.63	7.11	9.41	0.16
Mediana	$ 51.00	$ 119.00	$ 3,000.00	$ -	$ 1,000.00	$ 1,400.00	36.00
Moda	35.00	0.00	0.00	0.00	1000.00	1500.00	28.00
Desviación estándar	19.52	86.78	11525.17	1243.53	474.54	628.09	10.98
Varianza de la muestra	380.87	7531.54	132829557.79	1546372.05	225191.07	394496.94	120.67
Curtosis	1.41	9.83	181.50	149.89	4.59	26.36	-0.62
Coeficiente de asimetría	1.02	1.96	10.25	9.37	1.15	3.00	0.48
Rango	145.00	959.00	300000.00	30000.00	4900.00	11035.00	50.00
Mínimo	$ 35.00	$ -	$ -	$ -	$ 100.00	$ 105.00	18.00
Máximo	$ 180.00	$ 959.00	$ 300,000.00	$ 30,000.00	$ 5,000.00	$ 11,140.00	68.00
Suma	$ 247,559.00	$ 577,242.00	$ 23,818,337.00	$ 1,521,663.00	$ 4,628,842.00	$ 6,517,111.00	165181.00
Cuenta	4455.00	4455.00	4455.00	4455.00	4455.00	4455.00	4455.00

Dentro de nuestro análisis, podemos destacar que la media de ingreso fue de $129.57. La edad mínima de los encuestados fue de 18 años y la máxima de 68, lo que nos indica que hubo un rango considerable de edades entre los encuestados. También se observa que la media de gastos fue de $55.57, la de deuda fue de $341.56, y el valor de los bienes tiene un mínimo de $105 y un máximo de $11,140.

Sin embargo, estos datos por sí solos no nos brindan mucha información ni un trasfondo que resulte interesante para un análisis, por lo cual procedimos a realizar ***"Tablas Dinámicas"***, las cuales más adelante nos ayudaron a decidir cuáles plasmar en nuestro dashboard.

A continuación se presentan las tablas dinámicas que se realizaron y una breve interpretación de estas:

Cantidad solicitada de préstamo	
Tipo de trabajo	**Cantidad solicitada**
FIJO	$2,872,309.00
FREELANCER	$1,160,540.00
MEDIO TIEMPO	$428,990.00
NA	$2,500.00
OTRO	$164,503.00
Total general	**$4,628,842.00**

Total de gastos según tipo de trabajo	
Tipo de trabajo	**Total de Gasto**
FIJO	$ 155,694.00
FREELANCER	$ 60,387.00
MEDIO TIEMPO	$ 22,783.00
NA	$ 80.00
OTRO	$ 8,615.00
Total general	**$ 247,559.00**

Total de ingreso según trabajo	
Tipo de trabajo	**Total de Ingreso**
FIJO	$ 394,018.00
FREELANCER	$ 120,035.00
MEDIO TIEMPO	$ 46,932.00
NA	$ -
OTRO	$ 16,257.00
Total general	**$ 577,242.00**

Esta tabla muestra la cantidad solicitada de préstamo según el tipo de trabajo. Esta información nos ayudó a detectar que las personas con un trabajo fijo son las que más dinero solicitaron en préstamos, y también coincide en que tienen la mayor suma de gastos, así como la que más ingresos generan.

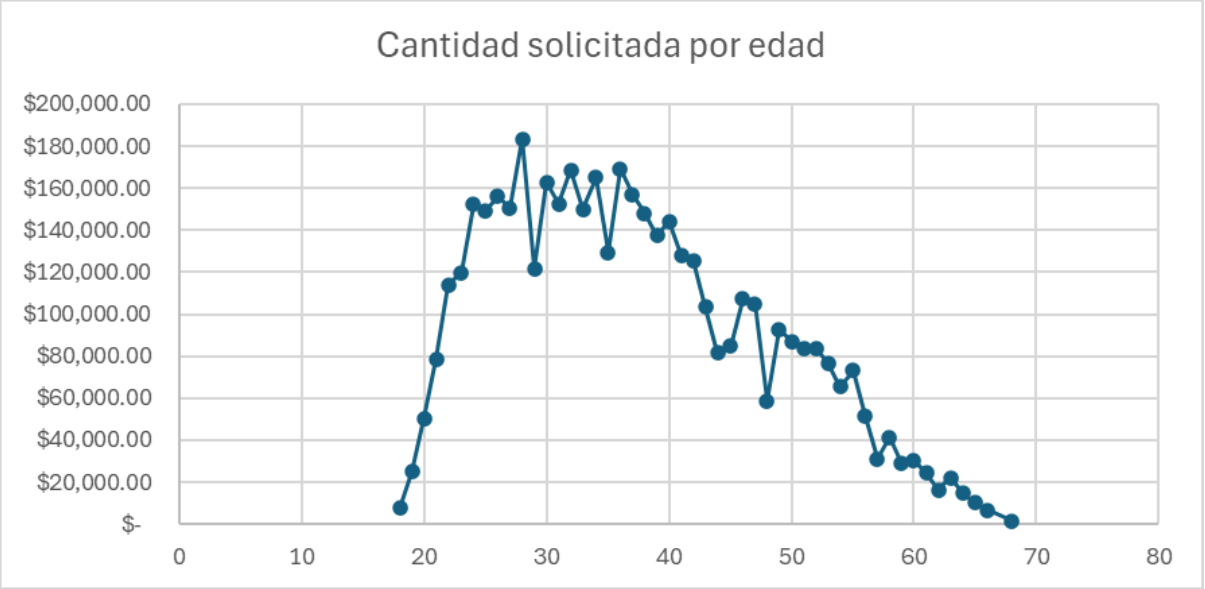
Historial crediticio según tipo de trabajo

Cuenta de Status	Etiquetas	
Tipo de trabajo	BUENO	MALO
FIJO	2226	580
FREELANCER	691	333
MEDIO TIEMPO	181	271
NA		2
OTRO	103	68
Total general	**3201**	**1254**

Historial crediticio de acuerdo a vivienda

Cuenta de Status	Etiquetas	
Tipo de viviend	BUENO	MALO
NA	2	4
NO DIJO	11	9
OTRO	173	146
PADRES	550	233
PRIVADA	163	84
PROPIA	1717	390
RENTADA	585	388
Total general	**3201**	**1254**

En estas tablas se agrupó el historial crediticio según tipo de trabajo y vivienda, en cuanto a trabajo los que mayor tienen un mal historial crediticio son los que cuentan con un tipo de trabajo fijo, y en vivienda los que tienen una casa propia seguido de los que rentan.



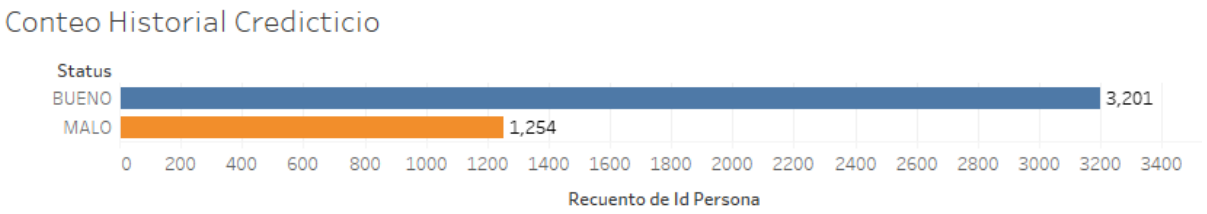
Por último, se realizó un gráfico de dispersión para identificar qué edad es la que más cantidad de dinero solicitó en préstamos siendo esta la edad de 28.

Todo esto se guardó en el archivo titulado CreditScoring Dinámico.

4. Visualización de Datos

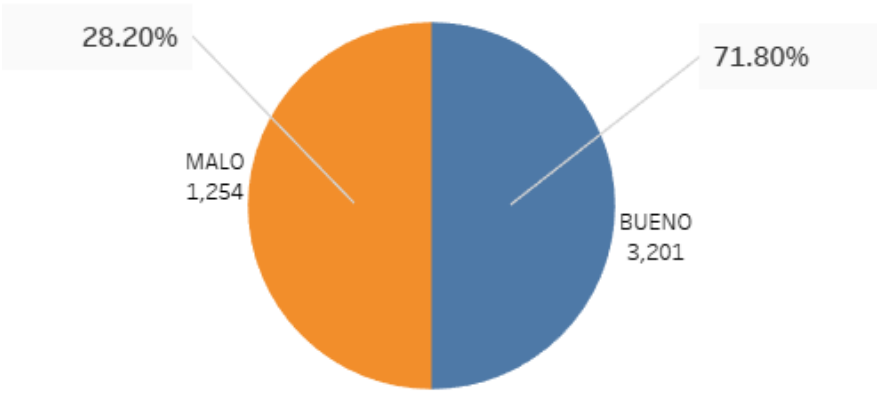
Para la visualización de los datos se hizo uso del Software ***Tableau*** para su correcta representación. Se elaboraron 4 gráficas distintas que representan a nuestro parecer, los puntos más importantes a considerar de la base de datos.

4.1 Gráfica N°1 - Conteo total de historial crediticio



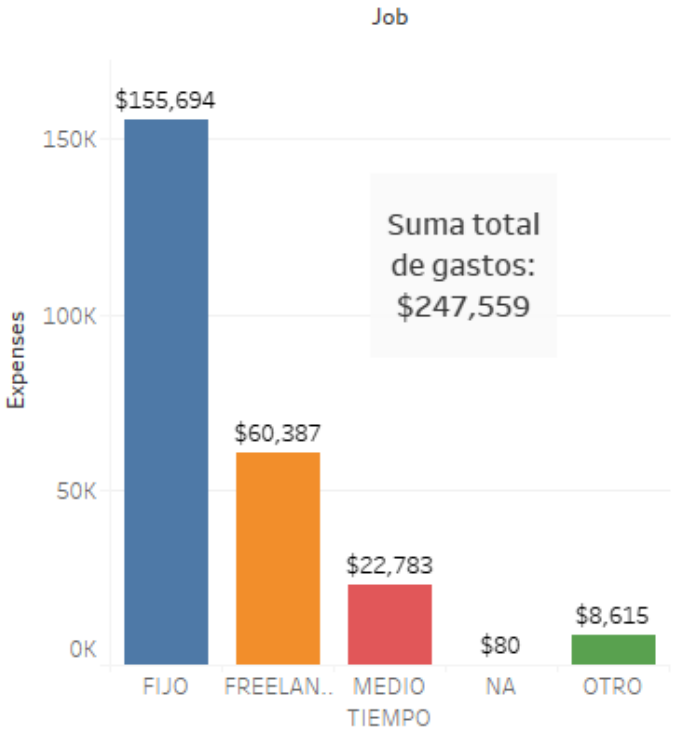
En esta gráfica podemos apreciar el total de encuestados en esta base datos y cuál es su prevalencia de acuerdo su estado crediticio. Podemos observar que un total de 3,201 encuestados tienen un historial bueno, al contrario 1,254 cuentan con un historial crediticio malo. Aquí pudimos deducir que en el total de encuestados es más la mayoría que cuenta con un historial bueno.

4.2 Gráfica N°2 - Porcentaje de Historial Crediticio



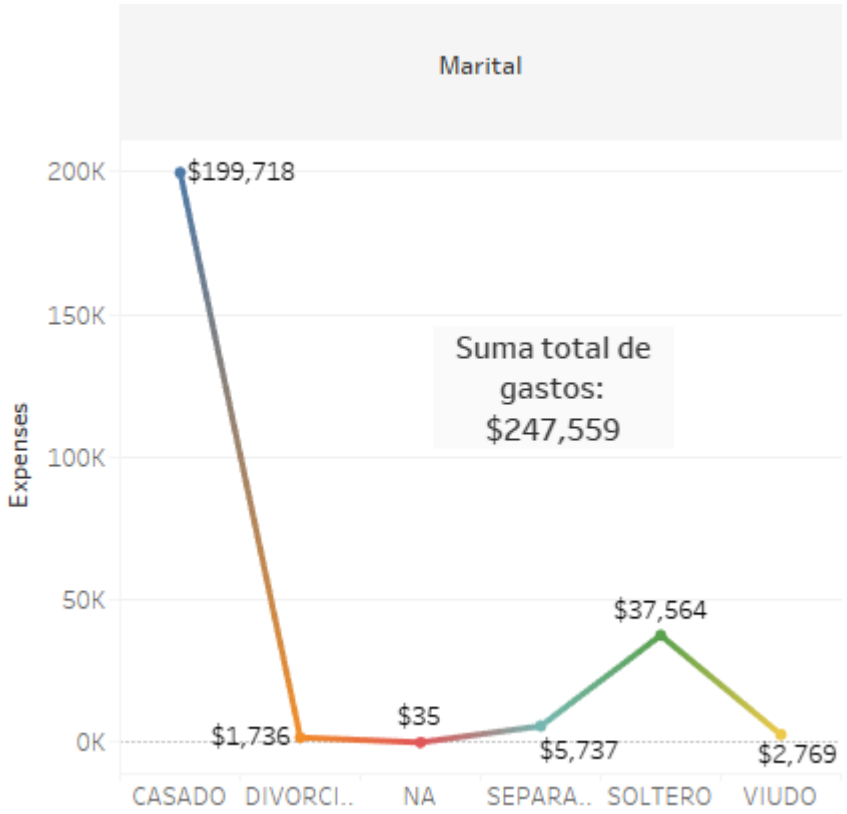
Esta gráfica, como su nombre lo dice, muestra el porcentaje del total de encuestados, podemos observar que del total, un 71.80% tiene buen historial crediticio, y un 28.20% cuenta con un mal historial.

4.3 Grafica N°3 - Total de gastos según el trabajo



Aquí se observa un punto muy interesante, se llegó a la deducción de las personas que cuentan con un trabajo y salario fijo, son las que más gastan dinero, después siguen las personas con un trabajo de freelancer, esto se puede deducir ya que este tipo de trabajos el salario puede variar y no siempre puede ser el mismo, por ende, los gastos son menores, también las personas con empleo de medio tiempo suelen gastar menos, ya que estas personas perciben un salario menos a si su trabajo fuera de tiempo completo.

4.4 Grafica N°4 - Total de gastos en situación marital



Esta gráfica, al igual que la anterior, muestra un dato muy interesante y hasta cierto punto esperado, ya que pudimos deducir que las personas que se encuentran en situación marital casad@ son las que más gastan dinero, las razones pueden varias, pero la más fiable es debido a que juntando los dos salarios las parejas se pueden dar más lujos y hacer más gastos. Después están los solteros, que no gastan mucho como los casados pero tampoco gastan a alguien viudo o separado.

5. Interpretación y Conclusiones

5.1 Resultados del Análisis

El análisis realizado sobre la base de datos de historial crediticio ha permitido obtener hallazgos significativos que reflejan patrones y tendencias en el comportamiento financiero de los encuestados. En cuanto a historial crediticio, un 71.80% de los encuestados cuenta con un historial crediticio bueno, mientras que un 28.20% tiene un historial malo. Esta diferencia sugiere que, en general, la mayoría de los encuestados tiene un manejo financiero positivo.

Los trabajadores con empleos fijos son los que más gastan, seguidos por aquellos con trabajos freelance y de medio tiempo. Esto indica que la estabilidad laboral puede estar correlacionada con mayores gastos, posiblemente debido a una mayor capacidad adquisitiva.

Los encuestados casados presentan los gastos más altos, lo que podría deberse a la acumulación de ingresos en pareja y a la posibilidad de compartir costos. Por otro lado, los solteros y viudos tienden a tener menores gastos.

La edad que más solicita préstamos es la de 28 años, lo que puede indicar que este grupo está en un momento crucial para adquirir bienes, como viviendas o vehículos.

Los resultados muestran que la mayoría de los encuestados tienen un buen historial crediticio, lo cual es alentador para instituciones financieras al considerar el otorgamiento de préstamos, sin embargo, la existencia de un porcentaje considerable con historial malo destaca la importancia de seguir educando a la población sobre el manejo adecuado de las finanzas personales.

Además, los patrones de gasto observados según el tipo de empleo y situación marital refuerzan la idea de que la estabilidad financiera y la situación personal influyen significativamente en el comportamiento económico de los individuos.

5.2 Conclusiones

Para finalizar solo nos queda por agregar que este tipo de actividades y proyectos son muy útiles, ya que pones a prueba tus conocimientos y habilidades obtenidas durante el diplomado. Podemos concluir con que esta actividad es de suma

importancia para un futuro, ya que en él se realizó y se trabajó en diferentes aspectos tecnológicos y se adquirieron nuevas habilidades de suma importancia.

Como recomendaciones basadas en los resultados tenemos que es crucial implementar programas de educación financiera dirigidos a aquellos con historial malo, enfocándose en la gestión de deudas y la importancia del ahorro.

En resumen, el análisis ha permitido resaltar la relación entre el historial crediticio, los ingresos y los gastos, la mayoría de los encuestados muestra un comportamiento financiero responsable, pero también existe una porción de la población que presenta dificultades en este aspecto, lo que merece atención.