

AVANCES EN LA EJECUCION DE LA METODOLOGIA CRISP-DM PARA EL MODELO ANALITICO DE LA COOPERATIVA DE CREDITO Y AHORRO UNIMOS

1. Diseño Metodológico

En el segmento de Diseño Metodológico, se traza la estructura metodológica que orientará la ejecución de este proyecto de consultoría. Se ha seleccionado la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) por su enfoque riguroso y estructurado, que es altamente pertinente en el ámbito de la consultoría para proyectos de minería de datos.

La metodología CRISP-DM se distingue por su ciclo iterativo compuesto por seis fases: Entendimiento del Negocio, Entendimiento de los Datos, Preparación de los Datos, Modelado, Evaluación y Despliegue. Sin embargo, en el contexto de este proyecto, se abordarán las etapas hasta la Evaluación, dado que la fase Despliegue no se encuentran dentro del alcance del presente proyecto.

A través de esta estructura metodológica, se aspira a asegurar una gestión eficaz del proyecto, proporcionando a la vez una comprensión clara y detallada de los procedimientos y técnicas que se aplicarán en el transcurso del proyecto de consultoría.

2. Entendimiento del negocio

En este apartado se describe el contexto en el cual será aplicado el proyecto, de acuerdo con la metodología CRISP-DM, en esta fase se busca entender el proyecto desde la perspectiva del negocio.

2.1.1 introducción al Contexto Empresarial

Establecida en abril de 2004 y con casi 20 años en el mercado, la Cooperativa de Ahorro y Crédito Unimos nació de una colaboración e iniciativa de algunos colaboradores de la Caja de Compensación Familiar Compensar. El propósito superior de la entidad es impactar positivamente en cada momento de la vida del asociado a través servicios financieros que mejoren la calidad de vida. Se enorgullece de proporcionar una propuesta económica atractiva y con un enfoque distintivo en la gestión de recursos, garantizando beneficios sociales para toda su comunidad (Cooperativa de ahorro y crédito UNIMOS, 2023).

El objetivo Retador de la cooperativa es consolidarse como entidad referente y competitiva en la oferta de servicios que generan valor agregado a sus asociados mediante el servicio oportuno, de calidad y eficaz, que los posicione como una de las mejores cooperativas que existen en el país, todo esto a través del fomento del ahorro e inversión, brindando distintas opciones de crédito y construyendo un bienestar solidario entre sus asociados.

Actualmente, sirve a más de 30.000 asociados y ha establecido más de 300 convenios empresariales, ofreciendo una variedad de productos de ahorro, crédito y bienestar (Cooperativa de ahorro y crédito UNIMOS, 2023)

2.1.2 Identificación de Problemas y Oportunidades

Para la Cooperativa UNIMOS la prospectación de bases es un procedimiento crucial que les permite identificar y cultivar relaciones con los asociados actuales y potenciales. Sin embargo, los procesos actuales de prospectación y otorgamiento de créditos enfrentan desafíos, tales como:

- Centralización del proceso de prospectación en una sola persona.
- Falta de herramientas automatizadas para un análisis eficiente y moderno.
- La complejidad en la aprobación masiva de créditos.

Estos desafíos ralentizan la respuesta y limitan la capacidad de atender a todos los asociados, lo cual, a largo plazo, puede afectar el crecimiento sostenible de la cooperativa.

Por tales motivos y como se había mencionado el objetivo general es, diseñar un modelo analítico mediante algoritmos de aprendizaje automático que optimice la prospección de los asociados y mejore la toma de decisiones para otorgar cupos de crédito.

De igual manera, para cumplir este objetivo, se hace relevante recordar los objetivos específicos que se plantearon para el proyecto:

- Realizar un análisis exploratorio de datos para definir las variables relevantes para el modelo analítico y predictivo.
- Diseñar un modelo analítico de aprendizaje automático que optimice la asignación de cupos de crédito de forma masiva.
- Evaluar la eficacia del modelo analítico comparando los resultados obtenidos con los procesos anteriores, corroborando el cumplimiento de los criterios de otorgamiento de crédito.

2.1.3 Evaluación de la Situación Actual

Actualmente, la cooperativa lleva a cabo un proceso manual de manejo de información para la creación de bases de prospección, una tarea que recae principalmente en un individuo. Esta persona también contribuye a la gestión de los créditos que llegan por el mercado natural, es decir, las solicitudes presentadas por los asociados. Junto con su equipo, procesan aproximadamente 600 solicitudes de crédito cada mes.

A pesar de que el sistema de información actual permite contar con una base de datos inicial de 32,000 registros, solo se retienen entre 1,500 a 3,200 registros tras varios procesos de cruce y validación de la información. Estos registros contienen variables como el ID del registro, datos personales y sociodemográficos del asociado, información y comportamientos financieros de

los miembros de la cooperativa, entre otras variables relevantes para el análisis y el proceso de prospectación llevado a cabo por la cooperativa.

Adicionalmente, se incorpora información de entidades externas como TransUnion, la cual proporciona un perfil de riesgo del asociado mediante una calificación conocida como score el cual es consolidado a través de métodos estadísticos y analíticos.

Esta información procesada se maneja de acuerdo con las políticas internas de la cooperativa y el manual de otorgamiento de crédito. Todo el conjunto de datos se gestiona utilizando la aplicación Microsoft Office Excel, que se emplea para consolidar, cruzar, validar y entregar la información.

La centralización de la tarea de prospectación y la falta de herramientas automatizadas han generado retrasos e ineficiencias en la prospectación, y, como resultado, en la asignación oportuna de créditos.

2.1.4 Objetivo de la minería de datos

El objetivo de este proyecto es diseñar un modelo predictivo basado en datos históricos y características de los asociados, que sea capaz de procesar de manera ágil los datos obtenidos previamente del proceso de ETL e identificar con alta precisión a los candidatos ideales para la prospectación y asignación de cupos de crédito.

Este modelo deberá considerar variables relevantes como el score, la capacidad de pago, datos socio demográficos, comportamiento financiero previo, entre otras variables que se definen en el manual de otorgamiento de crédito interno de la cooperativa, y que proporcionan una puntuación o codificación de las variables que guíe la decisión de otorgar o no un cupo de crédito.

2.1.5 Criterios de éxito de la minería de datos

Se considera que el modelo analítico ha tenido éxito cuando los hallazgos o agrupaciones obtenidos permitan:

- Definir con mayor precisión los asociados sujetos de crédito que cumplan los criterios del otorgamiento de crédito, donde el porcentaje de cumplimiento de las variables sea superior al 75%.
- Identificar los perfiles de riesgo o medio bajos de acuerdo con la codificación que se da a las variables, según el manual de otorgamiento de crédito interno de la cooperativa.
- Que, al evaluar los asociados seleccionados de manera individual, el resultado en cuanto al perfil de riesgo y cupo de crédito sea igual o muy cercano al arrojado por la prospección que se viene realizando de manera manual.

2.1.6 Planificación de la Metodología CRISP-DM

El proyecto se segmentará en diversas fases con el objetivo de organizarlo adecuadamente y calcular el tiempo que tomará completarlo (ver Tabla 1)

Es relevante destacar que la metodología CRISP-DM tiene un enfoque cíclico. Esto implica que, a medida que el proyecto progresa y surgen nuevos datos o se encuentran obstáculos no previstos, puede ser necesario regresar a etapas previas. Por ende, el tiempo total de ejecución del proyecto podría superar la suma de las estimaciones iniciales para cada fase

Tabla 1

Planificación de la Metodología

Fase	Tareas Principales	Tiempo Estimado
Comprensión del Negocio	Consulta con expertos en el dominio	4 semanas
	Definición del problema	
	Identificación de objetivos	
Comprensión de los Datos	Recopilación de datos	3 semanas
	Análisis exploratorio de datos (EDA)	
	Verificación de la calidad de los datos	
Preparación de los Datos	Limpieza de datos	3 semanas
	Selección de características	
	Transformación de variables	
Modelado	Selección de técnicas de modelado	4 semanas
	Construcción de modelos	
	Calibración de parámetros	

Fuente: Elaboración propia

3. Entendimiento de los datos

En este apartado, se lleva a cabo la recopilación, descripción, exploración y verificación de la calidad de los datos disponibles. Estas acciones facilitan la realización de un Análisis Exploratorio de Datos (EDA, por sus siglas en inglés), permitiendo comprender las características, la distribución y las relaciones entre las variables.

Para la exploración y posteriormente la limpieza de los datos, que se abordará en otro apartado, se utilizan las herramientas Google Colaboratory o Deepnote. Estas herramientas permiten emplear el lenguaje de programación Python, junto con librerías especializadas como Pandas, para el análisis de datos.

3.1 Recolección de datos

Los datos utilizados para este proyecto provienen de la extracción de la información del sistema transaccional de la cooperativa UNIMOS y complementada con información de otras fuentes como TransUnions. El conjunto de datos es exportado en un archivo estructurado de MS Excel con previa autorización de la entidad.

El conjunto de datos comprende 1.398 registros de asociados activos, recopilando información personal y transaccional correspondiente a los meses de julio a septiembre de 2023. La extracción de esta información se realizó el 1 de octubre de 2023.

Este conjunto incluye diversos tipos de información:

- Sociodemográfica
- Comportamiento crediticio
- Situación financiera
- Capacidad de pago
- Solvencia
- Información de centrales de riesgo
- Comportamiento de pago
- Garantías

Estas categorías de información están definidas conforme al manual de otorgamiento de crédito de la cooperativa. Esta información es esencial y se utiliza durante el proceso de prospectación que se lleva a cabo para evaluar la viabilidad de otorgar créditos a los asociados.

Las características del archivo se encuentran en la Tabla 2.

Tabla 2

Descripción del origen los de datos

Nombre Archivo:	CREDITO JULIO A SEPT.xlsx
Cantidad de Registros:	1,398
Cantidad de Variables:	90



Fuente: Elaboración propia

3.2 Descripción de los datos

Describir los datos implica reconocer los atributos incluidos en el archivo, los cuales serán utilizados para el proceso de minería de datos. El archivo en cuestión alberga datos clasificados de forma categórica y numérica. Como se mencionó anteriormente, el archivo consta de 1.398 registros y 90 columnas o variables; de estas, 28 son categóricas y 62 son numéricas.

Los tipos de datos, junto con la descripción de las variables categóricas, variables numéricas y valores nulos, se ilustran en las figuras 1, 2, 3 y 4, respectivamente. Es importante mencionar que, debido a la cantidad de variables, la visualización de los resultados en pantalla está limitada por el programa utilizado.

Figura 1. Tipos de datos

df.dtypes		df.dtypes.value_counts()	
			
N°	int64	int64	47
Id	int64	object	24
NUMERO IDENTIFICACION	int64	float64	15
IDENTIFICACION	object	datetime64[ns]	4
NOMBRES APELLIDOS	object	dtype: int64	
...	...		
NIT	float64		
Perfil de Riesgo Score 01 de Agosto	object		
APROBADO-DESEMBOLSADO	object		
Valor Desembolsado	int64		
NEGADO	object		
Length: 90, dtype: object			

Fuente: Elaboración propia

Figura 2. Descripciones variables numéricas

df.describe()

	N° float64	Id float64	NUMERO IDENTIFI...	EDAD float64	ESTRATO float64	ANTIGUEDAD float...	PERSONAS A CAR...	CAN
count	1398	1398	1398	1398	1398	1397	1398	
mean	699.9470672	6959.634478	615896608.6	37.59012876	2.693133047	5.503937008	0.8054363376	
std	404.140542	1226.763281	489783401.9	10.3672763	0.8312757559	6.268767144	0.8821293801	
min	1	5366	360850	18	0	0	0	
25%	350.25	5694.25	52955208.5	30	2	1	0	
50%	699.5	7798.5	1012357443	36	3	3	1	
75%	1049.75	8118.75	1026570711	44	3	7	1	
max	1399	8457	1233894843	93	6	56	6	

8 rows, showing 10 per page

Fuente: Elaboración propia

Figura 3. Descripciones variables categóricas

This code has been hidden. [Show it.](#)

	IDENTIFICACION o...	NOMBRES APELLI...	NIVEL EDUCATIVO o...	OCUPACION object	DEPARTAMENTO ...	CIUDAD_RESIDEN...	PROFESION object
count	1398	1398	1398	1398	1398	1398	1391
unique	2	1292	8	4	55	92	934
top	CEDULA DE CIUD...	JORGE HERNAN ...	Técnico	EMPLEADO	BOGOTA	BOGOTA D.C	AUXILIAR DE ENF.
freq	1397	5	455	1338	944	1185	57

4 rows, showing 25 per page

Fuente: Elaboración propia

Figura 4. Descripción valores nulos

```

DESTINACION_CREDITO      17
PROFESION                  7
CANTIDAD_TDC_CIFIN        2
INICIO_ACTIVIDAD_ECONOMICA 1
NIT                        1
..
SALDO_CODEUDOR_CIFIN      0
RESULTADO_PRECISION        0
CLEAR_SCORE                0
CLEAR_SCORE_EXPERIENCIA    0
NEGADO                     0
Length: 90, dtype: int64

```

Fuente: Elaboración propia

En la Figura 5, se presentan los valores contenidos en las variables: edad, nivel educativo, tipo de vivienda, personas a cargo y estado civil, que fueron seleccionadas como ejemplo.

Figura 5, Contenido de las variables

```

✓
Valores únicos para EDAD: [25 33 23 36 27 61 42 59 29 37 64 38 30 48 31 26 41 35 45 55 44 46 39 32
34 53 54 43 49 28 40 80 51 52 24 50 47 19 22 21 57 56 93 68 58 60 20 62
66 67 70 65 18 78 73 75 76 69]

Valores únicos para TIPO CONTRATO: ['INDEFINIDO' 'INDEPENDIENTE' 'PRESTACION DE SERVICIOS' 'FIJO'
'PENSIONADO' 'LABOR CONTRATADA' 'TEMPORAL']

Valores únicos para NIVEL EDUCATIVO: ['Técnico' 'Universitario' 'Postgrado' 'Bachillerato' 'Tecnológico'
'Primaria' 'Ninguno' 'Postgrado']

Valores únicos para TIPO VIVIENDA: ['FAMILIAR' 'PROPIA' 'ARRENDADA']

Valores únicos para PERSONAS A CARGO: [0 2 1 6 3 5 4]

Valores únicos para ESTADO CIVIL: ['SOLTERO ' 'SOLTERO' 'CASADO' 'SEPARADO' 'UNION LIBRE' 'DIVORCIADO'
'VIUDO' '1' 'FUSAGASUGA' 'EL ROSAL']

```

Fuente: Elaboración propia

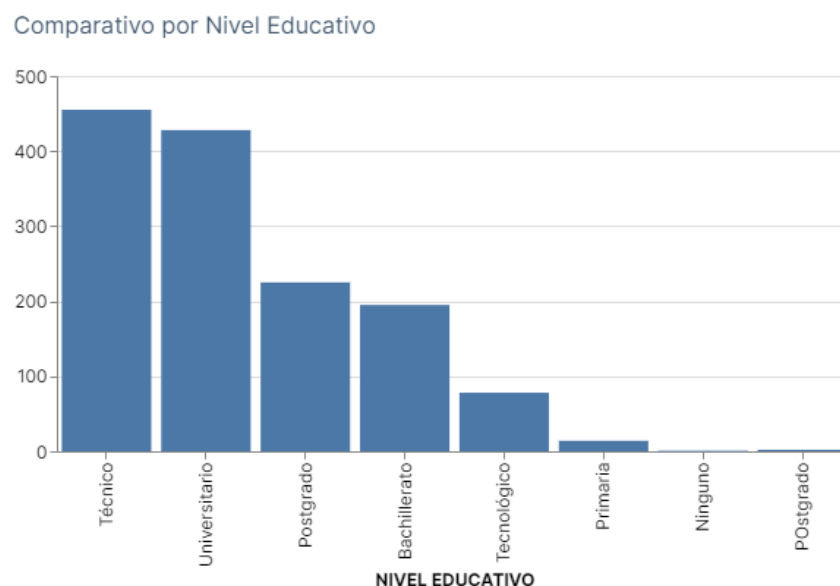
Esto se realiza con la finalidad de identificar valores repetidos o errores en la escritura. La limpieza de los datos, necesaria para corregir estas inconsistencias, se abordará más adelante.

3.3 Exploración de los datos

Con el objetivo de comprender mejor los datos de la base, se lleva a cabo un análisis exploratorio. Para esta actividad, el empleo de gráficas y tablas facilita la identificación de ciertas características de los datos de los asociados.

Inicialmente, se grafican algunas de las variables indicadas en el manual de otorgamiento de crédito de la cooperativa, tales como: edad, tipo de contrato, tipo de vivienda, nivel educativo y plazo solicitado en meses.

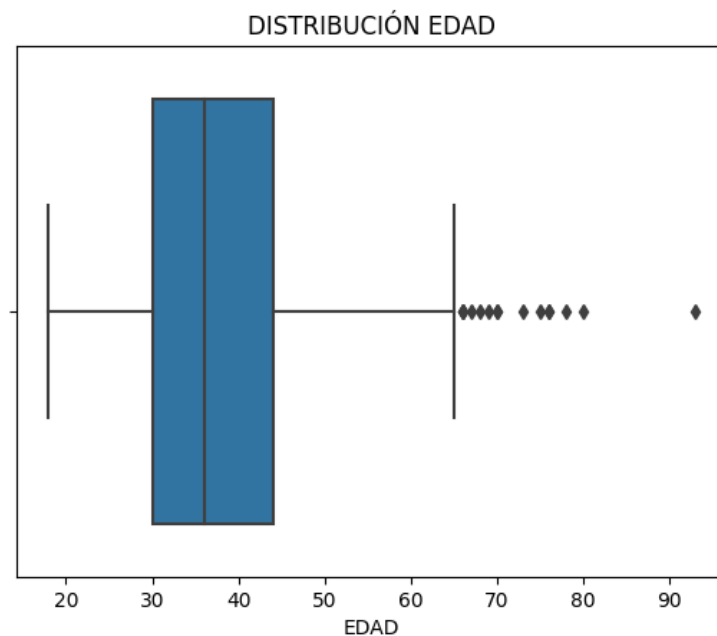
Figura 6. Grafica Comparativo por Nivel Educativo



Fuente: Elaboración propia

La gráfica de la Figura 6 muestra que la mayoría de los asociados de la cooperativa de ahorro y crédito tienen educación técnica o universitaria, lo que sugiere una base de miembros con un nivel educativo relativamente alto. Hay una menor representación de miembros con estudios de posgrado y aún menos con educación secundaria o primaria, y solo unos pocos no tienen educación formal. Además, la gráfica permite evidenciar inconsistencias en la escritura de los datos, como es el caso de las categorías "postgrado" y "pOstgrado", lo cual deberá ser tenido en cuenta durante la limpieza de los datos.

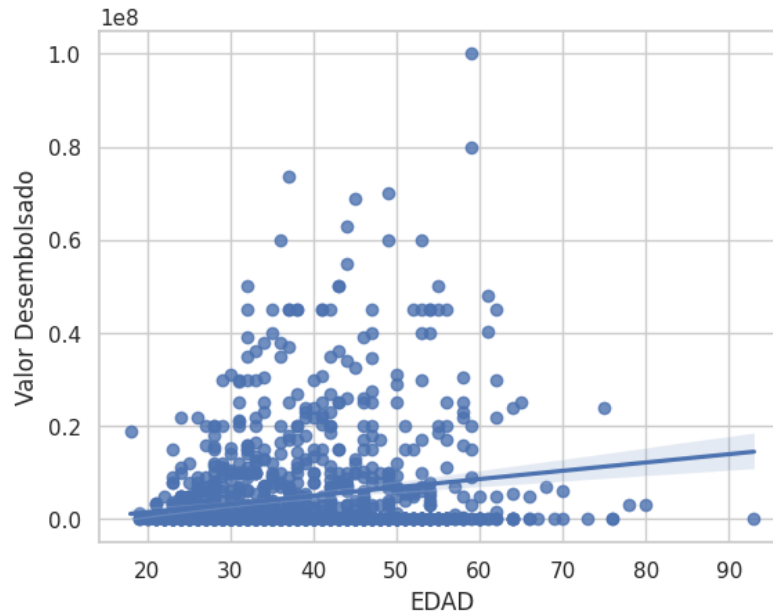
Figura 7. Grafica de caja y bigotes Edad



Fuente: Elaboración propia

En la Figura 7, el diagrama de caja y bigotes muestra que la mediana de edad del grupo estudiado está alrededor de los 36 años, con la mayoría de las edades concentradas entre los 35 y los 45 años, lo que podría representar el rango intercuartílico (la diferencia entre el primer y tercer cuartil). La distribución de edad se extiende desde aproximadamente los 20 años hasta más allá de los 70 años, con algunos valores atípicos que indican la presencia de individuos significativamente mayores por encima de los 90 años. El cual se debe validar antes de uso posible uso en el modelo.

Figura 8. Grafica de Dispersión edad vs valor Desembolsado



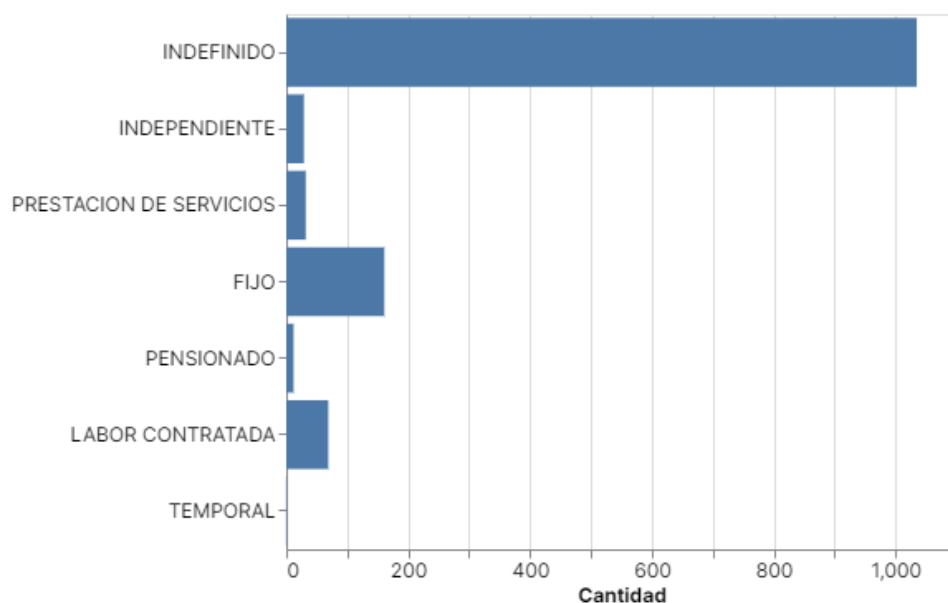
Fuente: Elaboración propia

En la Figura 8, el diagrama de dispersión muestra una leve correlación positiva entre la edad y el valor desembolsado, indicando que puede haber un ligero aumento en el valor desembolsado a medida que las personas envejecen, aunque la tendencia no es pronunciada.

La mayoría de los desembolsos se concentran en el extremo inferior del rango monetario, especialmente entre las edades más jóvenes. Se observan valores atípicos, principalmente entre los 30 y 60 años, donde algunos individuos reciben cantidades significativamente mayores. La densidad de los puntos es más alta en las edades más jóvenes y los valores desembolsados bajos, lo que sugiere que las personas más jóvenes son las que más frecuentemente reciben menores cantidades de dinero.

Figura 9. Grafica por tipo de contrato

Analisis por Tipo de Contrato



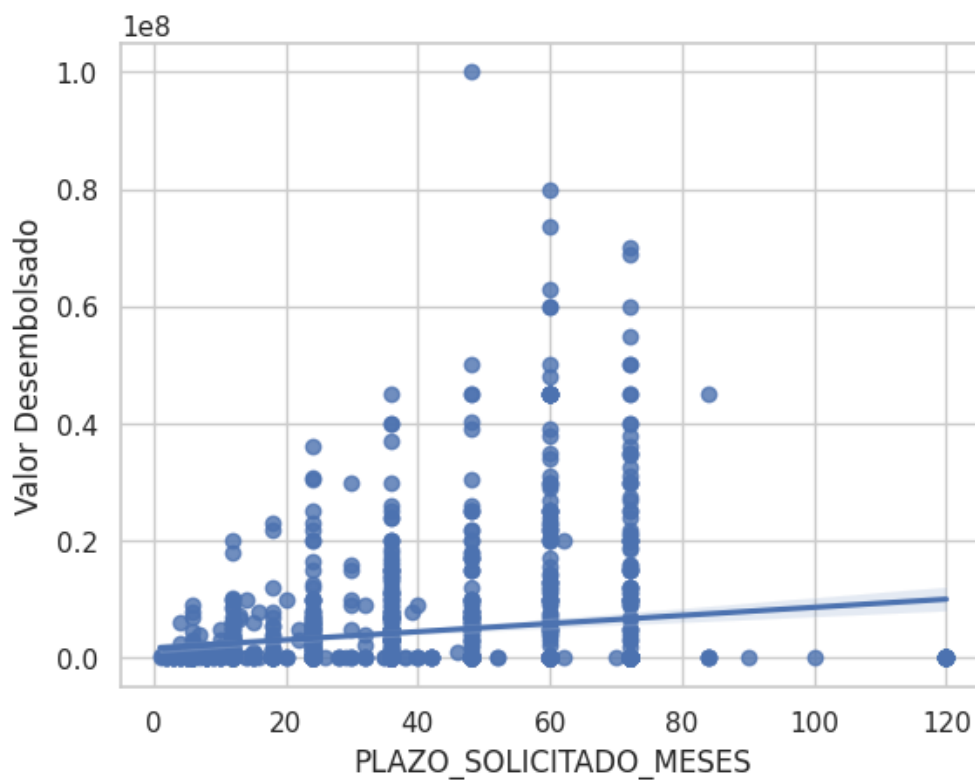
Fuente: elaboración propia

"La Figura 9 ilustra de manera clara que la barra correspondiente al tipo de contrato "indefinido" tiene la mayor concentración de datos, sugiriendo que los asociados a la cooperativa gozan de una estabilidad laboral en sus respectivas entidades. Posteriormente, las barras que representan los contratos a término fijo y "Labor Contratada" también exhiben una concentración notable de datos, aunque en una medida menor en comparación con los contratos indefinidos.

Los datos están clasificados en siete categorías, a saber: contrato por tiempo indefinido, trabajadores independientes, contrato por prestación de servicios, contrato fijo, pensionados, trabajadores con labor contratada y trabajadores

con contrato temporal. Estas categorías representan las diferentes modalidades de vínculos laborales que tienen los individuos dentro de la muestra analizada.

Figura 10. Grafica plazo solicitado vs valor desembolsado



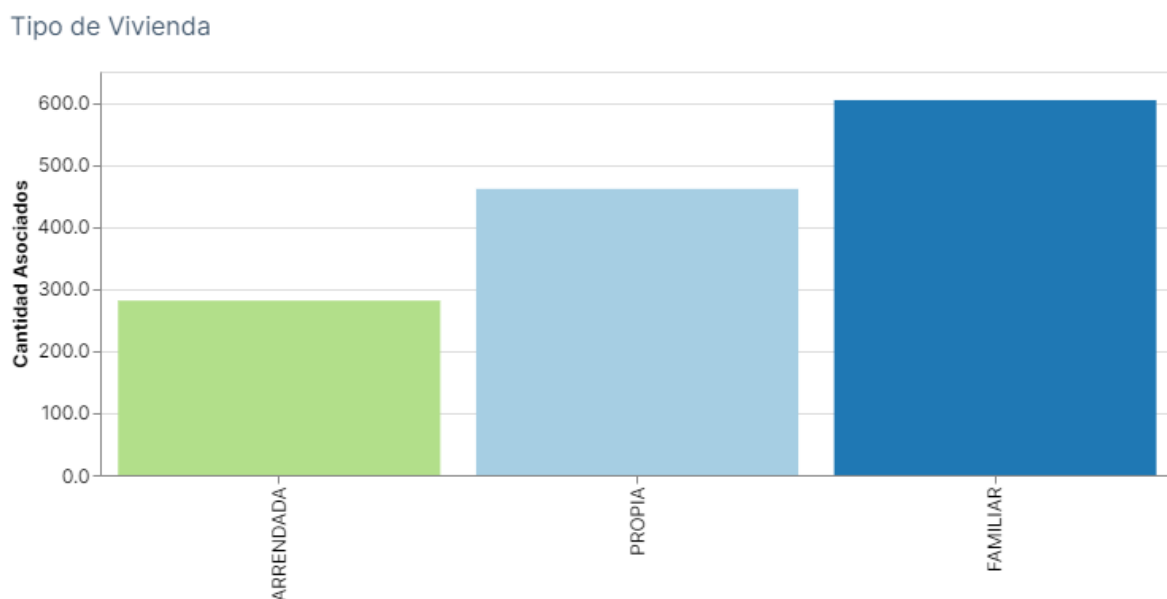
Fuente: elaboración propia

La Figura 10 ilustra cómo la distribución de puntos en el gráfico revela una concentración significativa de préstamos con plazos cortos, destacando una

amplia variabilidad en los montos desembolsados. Se observa que, a medida que el plazo del préstamo se extiende, tanto la frecuencia de los préstamos como los montos desembolsados disminuyen.

Esto podría sugerir una tendencia hacia la concesión de préstamos menores a medida que el plazo se alarga. Se detectan picos de concentración que podrían corresponder a plazos estándar de préstamos. Los montos más elevados parecen ser casi exclusivos de plazos más cortos, y la dispersión de los montos se estrecha con plazos más largos, lo que podría implicar una política de préstamos más conservadora para plazos extendidos. La distribución general no es uniforme, sugiriendo que existen múltiples factores que influyen el valor desembolsado además del plazo del préstamo.

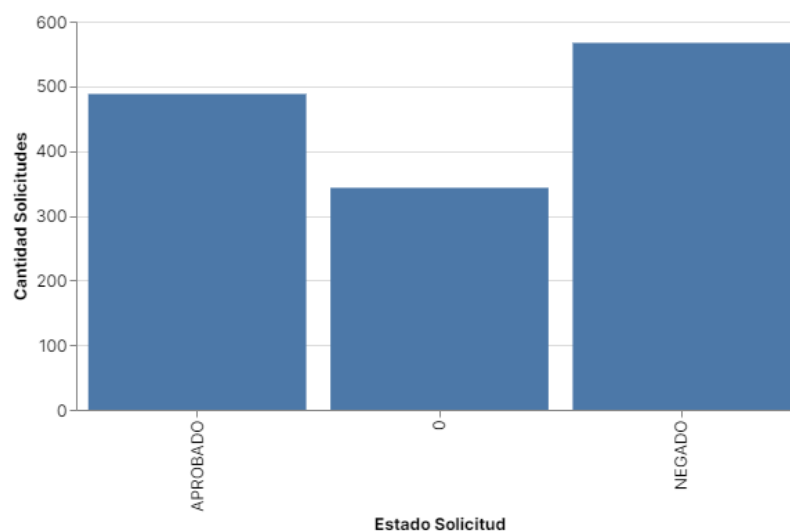
Figura 11. Grafica Tipo de vivienda



Fuente: elaboración propia

La Figura 11 despliega un gráfico de barras que ilustra la distribución de los tipos de vivienda entre los asociados. Es evidente que la mayoría de los asociados residen en viviendas de tipo familiar, seguidas por las viviendas propias, y en menor proporción, las viviendas arrendadas. Los datos están clasificados en tres categorías: arrendada, propia y familiar, lo que facilita la comparación entre estos grupos.

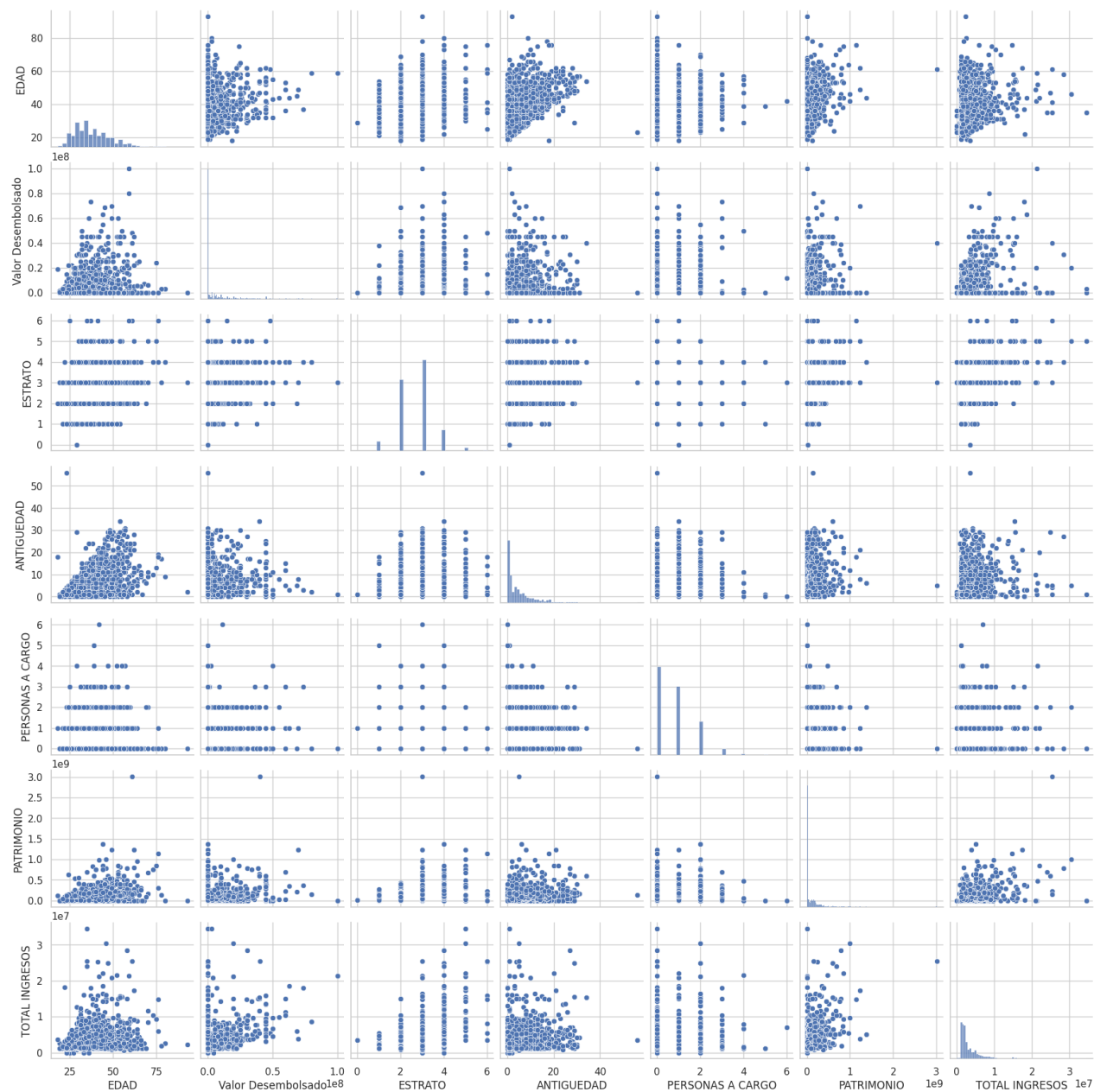
Figura 12. Grafica estado solicitud



Fuente: elaboración propia

En la Figura 12, se evidencia que, aunque el número de solicitudes pendientes es el menor, las solicitudes negadas superan a las aprobadas. Esto puede indicar un proceso de evaluación riguroso o un contexto en el que se presentan más solicitudes que no cumplen con los criterios necesarios para ser aprobadas. Además, la presencia de un valor numérico en la categoría del eje x sugiere que esta variable en la base de datos requiere un proceso de limpieza de datos.

Figura 13. Grafica de Dispersión



Fuente: elaboración propia

La Figura 13 presenta una matriz de dispersión que ilustra la relación entre varias variables, incluyendo el valor desembolsado, la edad, el total de ingresos, el

estrato, la antigüedad laboral y las personas a cargo. En la matriz, se manifiestan patrones de correlación lineal en algunos casos, y distribuciones agrupadas o discretas en otros.

Conclusiones:

Con base en lo expuesto desde la descripción de la base de datos y en los análisis realizados mediante las gráficas, se concluye que los datos presentes en la base requieren pasar por un proceso de limpieza y reducción de variables.

Además, se identificaron algunos valores atípicos tanto en la cantidad como en la categorización de algunas variables. Según la metodología CRISP-DM, estos hallazgos nos conducen a la etapa de "Preparación de Datos", con el objetivo de optimizar la base y mejorar la calidad de los datos.

4.Preparación De Los Datos

Tras el análisis de los datos que componen nuestra base, en esta fase se llevarán a cabo los procesos necesarios de selección, limpieza, incorporación de nuevos datos y formateo requerido, para que estos puedan ser implementados adecuadamente en el modelo.

4.1 Selección de los datos

En la sección previa, se identificó que la base de datos incluye 90 variables. Para determinar la relevancia de cada una, se llevó a cabo una revisión con el experto en crédito de la cooperativa. Este análisis subrayó la necesidad de ajustar la cantidad de variables.

La selección de los datos más pertinentes para el modelo se alinea con las directrices establecidas en el manual de otorgamiento de crédito de la cooperativa, que sirve como guía principal para la persona a cargo de la prospección.

A continuación, se enlistan las variables esenciales según el manual:

- Nivel educativo
- Edad
- Tipo de contrato
- Tipo de vivienda
- Personas a cargo
- Antigüedad laboral (años)
- Antigüedad en Unimos (meses)
- Riesgo Empresa
- Capacidad de descuento
- Capacidad de pago
- Plazo de crédito
- Razón Gastos Totales
- Razón Gastos Financieros
- Razón Endeudamiento Total
- Puntaje Scoring.

4.2 Limpieza de los Datos

Para este proyecto, se identificó que los datos de los asociados listados en la Tabla 3 no son relevantes para el modelo o presentan redundancia. Un ejemplo claro es la duplicidad entre la 'Fecha de nacimiento' y la 'Edad'. Por este motivo, se procedió a eliminar dichas variables de la base de datos.

Tabla 3

Datos no contemplados para la creación del modelo

VARIABLES		
ALIVIO_FINANCIERO	IDENTIFICACION	SALDO_CDAT
ARRENDAMIENTOS	INGRESO_COMISIONES	SALDO_CODEUDOR_CIFIN
CANT_MORA_CODEUDOR	INGRESOS ACT ECONOMICA	SALDO_CREDITO_UNIMOS
CANTIDAD_PROD_MORA_CIFN	INICIO ACTIVIDAD ECONOMICA	SALDO_CUPO_ROTATIVO
CANTIDAD_PRODUCTOS_CIFN	INMUEBLES	SALDO_INICIAL_COMPRADA
CANTIDAD_TDC_CIFIN	LINEA_CREDITO_SOLICITUD	SALDO_MORA_CIFIN

CIUDAD_RESIDENCIA	MONTO COMPRA	SALDO_MORA_CODEUDOR
CLEAR_SCORE_EXPERIENCIA	MONTO_SOLICITADO	SALDO_TDC_MILES
COBERTURA_GARANTIA_%	Nº	SALDO_TOTAL_CIFIN
COMPRA CARTERA	NEGADO	SOLICITANTE
CRED_DES_DUPL	NIT	TIPO_GARANTIA
CUOTA_MES_COMPRADA	NOMBRE_EMPRESA	TOTAL ACTIVO
CUOTA_MORA_CIFIN	NOMBRES APELLIDOS	TOTAL GASTOS
CUOTA_MORA_CODEUDOR	NUMERO IDENTIFICACION	TOTAL INGRESOS
CUOTA_SOLICITUD	OBLIGACIONES FINANCIERAS	TOTAL PASIVO
CUOTA_TDC_MILES	OCUPACION	VALOR_APORTES_VENCIDOS
CUOTA_TOTAL_CIFIN	OTROS ACTIVOS	VALOR_CUOTA_CODEUDOR
CUPO_TDC_MILES	OTROS GASTOS	VALOR_GARANTIA
DEPARTAMENTO NACIMIENTO	OTROS INGRESOS	VALOR_INICIAL_CODEUROR
DESCUENTOS NOMINA	OTROS PASIVOS	VALOR_INICIAL_TOTAL
DESTINACION_CREDITO	PAGO DEUDAS	VALOR_MORA
ESTADO CIVIL	PATRIMONIO	VEHICULOS
ESTRATO	PROFESION	
FECHA	SALDO_AH_CONTRACTUAL	
FECHA NACIMIENTO	SALDO_AH_PERMANENTE	
FECHA VINCULACION UNIMOS	SALDO_APORTES	

Fuente: Elaboración propia

Además, se eliminaron los registros que no aportaban valor significativo a la información. Esto incluyó aquellos registros donde la columna 'APROBADO-DESEMBOLSADO' mostraba valores de 0. Los valores nulos o faltantes se identificaron en las siguientes variables:

- DESTINACION_CREDITO
- PROFESIÓN
- CANTIDAD_TDC_CIFIN
- INICIO ACTIVIDAD ECONÓMICA
- NIT
- CANTIDAD_PROD_MORA_CIFIN
- VALOR_INICIAL_CODEUROR
- INGRESO_COMISIONES

- NOMBRE_EMPRESA
- CUOTA_TDC_MILES
- ANTIGÜEDAD

Para la variable ANTIGÜEDAD, se eliminó solo un registro. Las demás columnas mencionadas forman parte de la Tabla 3, por lo que su eliminación no afectó la base de datos.

4.3 incorporación de nuevos datos

Se identificó que ciertas variables mencionadas en el manual de otorgamiento de crédito no estaban calculadas directamente en la base de datos. Por tanto, se desarrolló el código necesario en Python y pandas para generar estas variables. Las variables recién incorporadas son:

- Capacidad de pago
- Plazo de crédito
- Razón Gastos Totales
- Razón Gastos Financieros
- Razón Endeudamiento Total

4.4 Formateo Requerido

En esta etapa del proyecto, se ha observado que las variables seleccionadas en la base de datos son de naturaleza cualitativa. Sin embargo, para su integración en el modelo, es necesario codificarlas según las escalas numéricas ordinales del Manual de Otorgamiento de Crédito de la cooperativa, que van de 1 a 5, representando el nivel de riesgo asignado por la cooperativa a cada variable.

Este nivel de riesgo ha sido definido por el consejo de administración y respaldado por pruebas de desempeño, conocidas como Backtesting, para los modelos de otorgamiento de crédito en la cooperativa.

Para alcanzar este objetivo, se aplicó un proceso de imputación de datos, reemplazando los valores existentes en cada variable por aquellos especificados en el manual. Este método asegura la coherencia de las variables, mejorando así su aplicación efectiva en el modelo.