



**Universidad
Internacional
de Valencia**

MÁSTER EN BIG DATA Y DATA SCIENCE

**13MBID Metodologías de gestión y diseño de
proyectos Big Data**

Actividad Práctica

Realizado por:

Jonnathan Henry Campoberde Avila

Hugo Elberto Pineda Gutiérrez

Fecha: 29 de noviembre del 2024

Curso 2024 – Ed. Abril

Actividades Prácticas - Aplicando técnicas ágiles para la gestión de proyectos de ciencia de datos.....	3
1. Comprensión del negocio.....	4
1.1. Determinar los objetivos de la Organización.....	4
1.2. Evaluación de la situación.....	4
1.2.1. Recursos de Datos.....	4
1.2.2. Recursos Humanos.....	4
1.2.3. Recursos Técnicos.....	5
1.3. Determinación de los objetivos del proyecto.....	5
1.4. Definir plan del proyecto.....	6
2. Comprensión de los datos.....	9
2.1. Recolección de datos iniciales.....	9
2.2. Descripción de los datos.....	9
2.3. Exploración de datos.....	9
2.4. Verificación de la calidad de los datos.....	17
2.4.1. Definición de objetivos y características de la evaluación inicial.....	17
2.4.1.1. Descripción del uso propuesto.....	17
2.4.1.2. Definición de calidad.....	17
2.4.2. Características que deben cumplir los datos.....	18
2.4.3. Registro de metadatos de cada dataset.....	18
2.4.4. Evaluación inicial de los datos disponibles.....	18
3. Fase de preparación de los datos.....	21
3.1. Selección de datos.....	21
3.2. Limpieza de los datos.....	21
4. Construcción de datos.....	22
4.1. Integración de los datos.....	23
4.2. Formateo de los datos.....	23
5. Modelado.....	32
5.1. Selección de la técnica de modelado.....	32
5.2. Generación del plan de pruebas.....	33
5.2.1. División del Conjunto de Datos.....	33
5.2.2. Ejecución de Instancias de Prueba.....	33
5.2.3. Filtrado Progresivo.....	33
5.2.4. Registro Detallado con MLflow.....	33
5.3. Construcción del Modelo.....	34
5.3.1. Evaluación del modelo.....	34
5.3.1.1. Prueba #1.....	34
5.3.1.2. Prueba #2.....	38
6. Evaluación.....	41
6.1. Evaluación de los resultados.....	41
6.2. Proceso de revisión.....	42

6.3. Determinación de futuras tareas.....	42
7. Despliegue / Implementación.....	44
7.1. Plan de implementación.....	44
7.2. Supervisión y Mantenimiento.....	44
7.3. Informe Final.....	45
7.4. Revisión del proyecto.....	45

Actividades Prácticas - Aplicando técnicas ágiles para la gestión de proyectos de ciencia de datos

El presente documento es una planilla que se utilizará para el desarrollo de la documentación correspondiente a las Actividades Prácticas I y II. El contenido será guiado según las fases y actividades de la metodología CRISP-DM.

Una vez completado con la información correspondiente al proyecto de ciencia de datos y complementado con los reportes de la ejecución de la libreta Jupyter desarrollada se podrán finalizar las tareas del proyecto.

La metodología CRISP-DM cuenta con 6 fases, ver figura 1, que forman un ciclo iterativo, con vistas a lo que se podrá considerar como un proceso iterativo-incremental de desarrollo de soluciones de ciencia de datos para un contexto en particular.

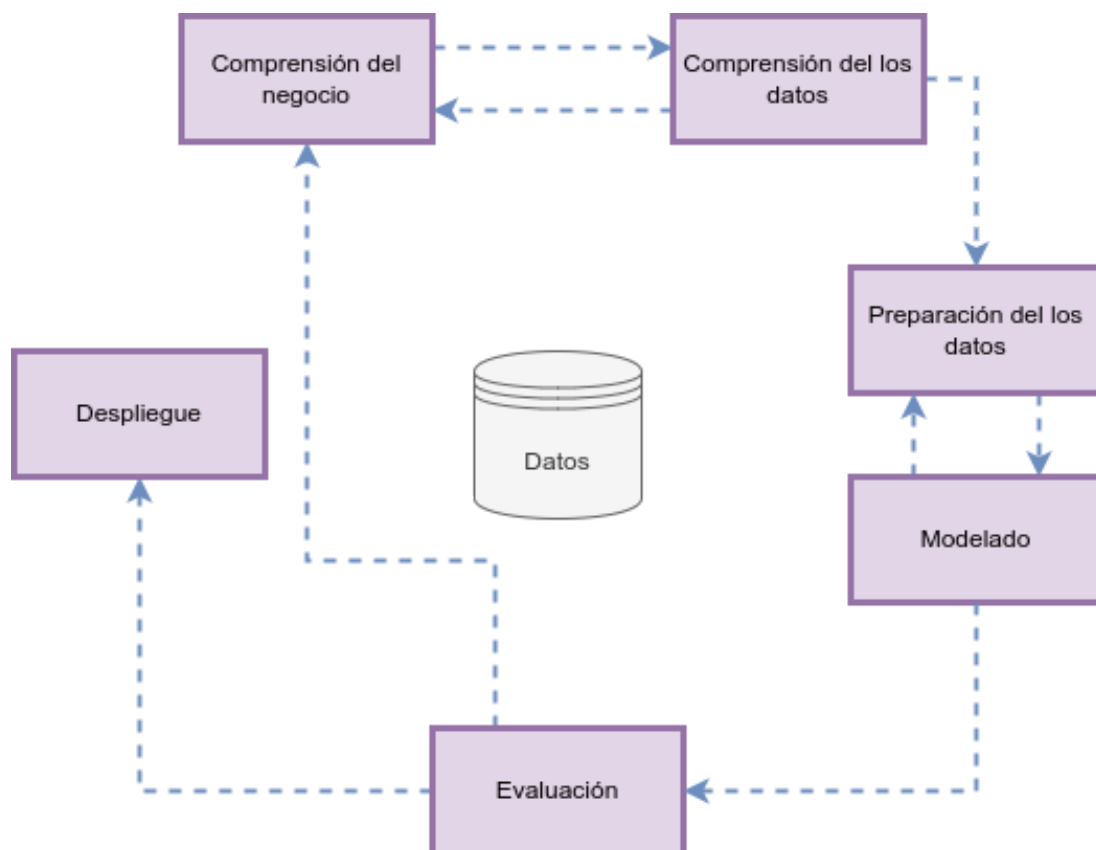


Figura 1. Metodología CRISP-DM

1. Comprensión del negocio

1.1. Determinar los objetivos de la Organización

Las autoridades de una entidad financiera desean obtener conocimiento a partir de su base de datos histórica de créditos otorgados. Para esta tarea, los datos disponibles se agrupan en dos dimensiones:

- **Datos de créditos:** que contienen la información de los créditos solicitados por los clientes y si los mismos han sido considerados en mora en algún momento.
- **Datos de otros productos:** que contienen la información sobre otros productos (en particular tarjetas de crédito) que poseen los clientes con la entidad y un resumen de su actividad y características principales.

1.2. Evaluación de la situación

El proyecto se encuentra en una fase inicial de ejecución y, en base a la evaluación actual, se dispone de los siguientes recursos clave para llevar a cabo las tareas requeridas y garantizar el éxito del proyecto:

1.2.1. Recursos de Datos

- **Datos Históricos de la Entidad:** Se dispone de un conjunto de datos históricos completos que abarcan información sobre el otorgamiento de créditos y productos relacionados con los clientes. Esta base de datos es fundamental para el análisis y modelado de los comportamientos de los clientes, las tasas de aprobación de crédito y otros aspectos clave del negocio.
- **Acceso a Datos Adicionales:** Además de los datos históricos mencionados, el proyecto tiene acceso a información complementaria que podría ser relevante para enriquecer los análisis y otras fuentes que podrían contribuir a obtener una visión más completa del comportamiento de los clientes.

1.2.2. Recursos Humanos

- **Personal Adecuado:** El proyecto cuenta con un equipo multidisciplinario con las habilidades necesarias para llevar a cabo las diversas fases del proyecto. Este equipo incluye especialistas en ciencia de datos, analistas, y desarrolladores, todos con experiencia en la implementación de soluciones basadas en datos. Cada miembro del equipo tiene responsabilidades claras y bien definidas en las tareas del proyecto.
- **Experto en el Dominio:** Se ha designado un experto en el dominio del negocio (sector financiero y de créditos) que está disponible para abordar cualquier duda o desafío que surja durante la ejecución del proyecto. Esta figura es esencial para garantizar que los modelos y análisis sean relevantes, aplicables y alineados con las necesidades del negocio, proporcionando insights clave y validando los resultados desde una perspectiva estratégica.

1.2.3. Recursos Técnicos

- **Herramientas Software Adecuadas:** El proyecto tiene acceso a las herramientas y plataformas de software necesarias para realizar las tareas de análisis de datos, modelado predictivo, y visualización de resultados. Se están utilizando herramientas avanzadas de análisis de datos como Python, plataformas de machine learning (scikit-learn), entre otros. Estas herramientas facilitarán la implementación de los productos resultantes del proyecto.
- **Infraestructura de Hardware:** Se cuenta con una infraestructura de hardware robusta para ejecutar las herramientas de software mencionadas. Esto incluye servidores, almacenamiento adecuado y capacidades computacionales suficientes para manejar los volúmenes de datos, realizar cálculos intensivos y ejecutar modelos complejos. La infraestructura está diseñada para soportar las demandas del proyecto a medida que avanza en sus fases de modelado y análisis.

1.3. Determinación de los objetivos del proyecto

Considerando los datos disponibles el producto a generar es un modelo que permita predecir sobre un conjunto de nuevos créditos otorgados por la entidad la posibilidad de que cada uno de ellos pueda entrar en mora en un futuro.

Como condición necesaria para el uso de los resultados obtenidos en una instancia de producción, se requiere que los mismos posean una efectividad mínima del 80% en el proceso de aprendizaje previo a la predicción a fin de poder reemplazar a los métodos actuales.

1.4. Definir plan del proyecto

El proyecto será gestionado utilizando la herramienta GitHub Projects, la cual permite organizar y supervisar las tareas de manera eficiente mediante tableros visuales. El acceso al tablero del proyecto se encuentra disponible en el siguiente enlace: <https://github.com/users/JonnHenry/projects/2>

La planificación del proyecto se ha estructurado en dos iteraciones, siguiendo la metodología CRISP-DM, para garantizar un enfoque ágil y organizado:

- **Primera Iteración:** Incluye las fases iniciales del proyecto: Comprensión del Negocio, Comprensión de los Datos y Preparación de los Datos. Estas etapas se centran en entender los objetivos del negocio, explorar y evaluar la calidad de los datos disponibles, así como transformarlos para que sean aptos para el análisis.
- **Segunda Iteración:** Comprende las fases restantes de la metodología: Modelado, Evaluación y Despliegue. En esta etapa, se trabajará en la construcción de modelos analíticos, la validación de su desempeño en función de los objetivos definidos, y finalmente, la implementación de los resultados en el entorno productivo o en aplicaciones del negocio.

Este enfoque iterativo no solo permite abordar las tareas de manera secuencial y ordenada, sino que también facilita la retroalimentación y ajustes oportunos al proyecto. Así, se asegura un proceso adaptativo que responde a las necesidades del negocio y los desafíos técnicos que puedan surgir durante su desarrollo.

A continuación, se incluye una captura de pantalla del proyecto en su primera iteración (Sprint Backlog)

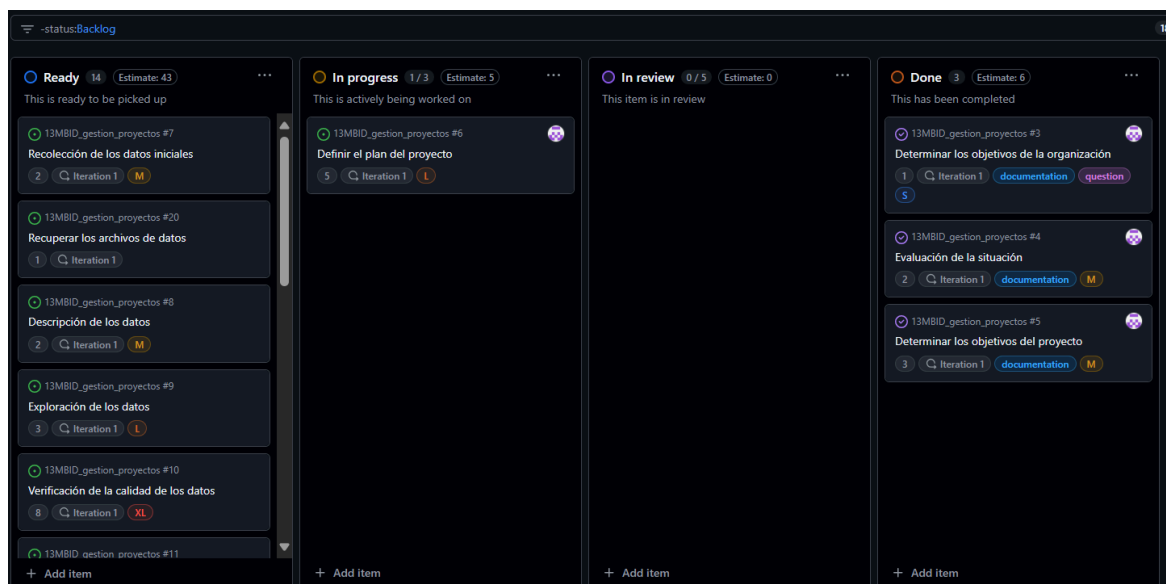


Figura 2. Proyecto en su primera iteración

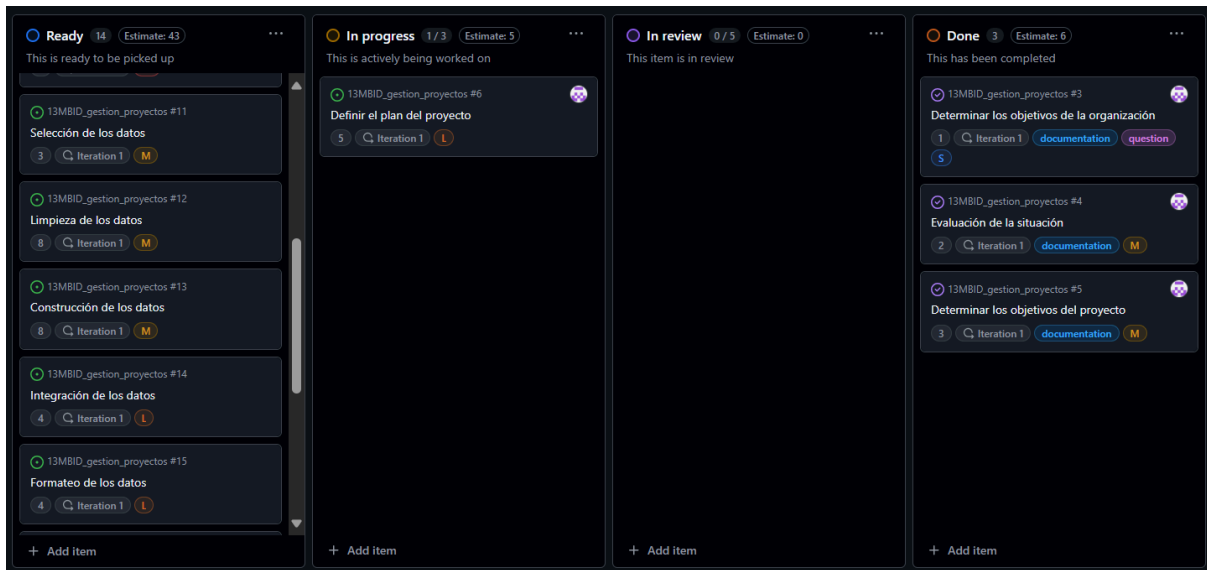


Figura 3. Proyecto en su primera iteración

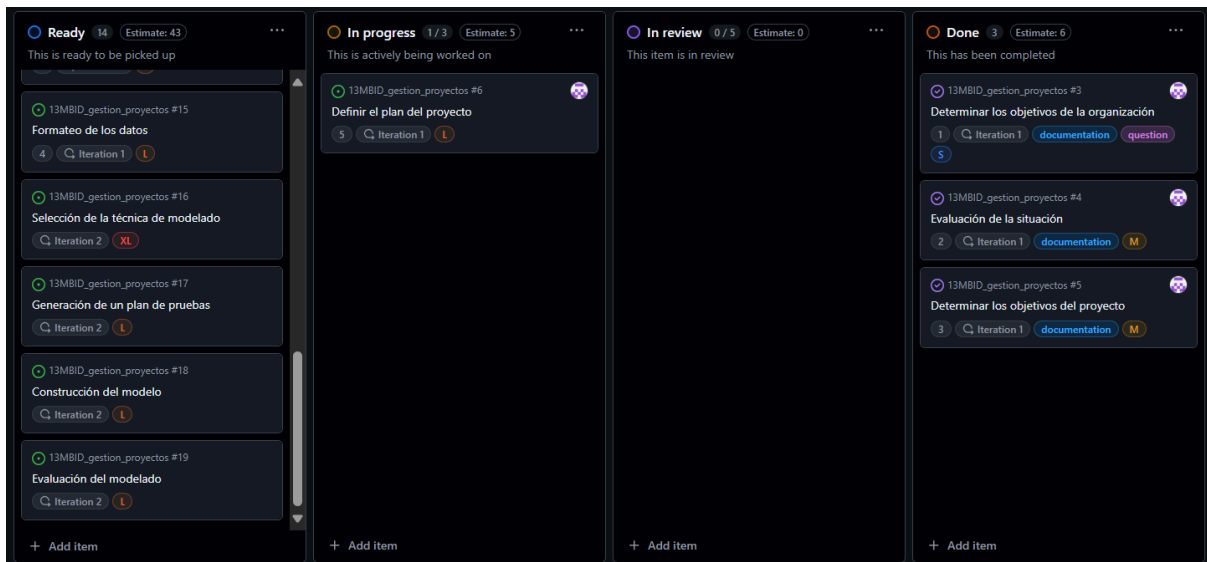


Figura 4. Proyecto en su primera iteración

A continuación, se incluye una captura de pantalla del proyecto en su segunda iteración (Sprint Backlog)

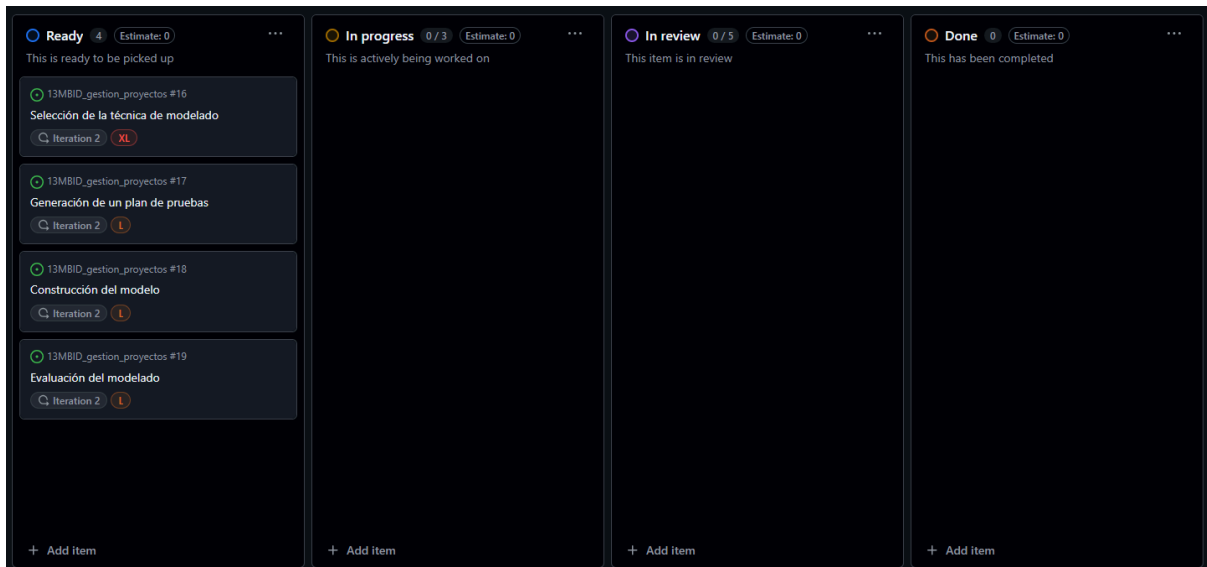


Figura 5. Proyecto en su segunda iteración

Los roles y tareas van a ser asignadas se han distribuido de la siguiente manera entre todos los integrantes de acuerdo a la siguiente tabla

Tabla 1. Roles asignados en el proyecto

Rol	Enfoque Principal	Persona Asignada
Científico de Datos	Calidad de los datos, modelos y análisis predictivos.	Hugo Pineda
Ingeniero de ML	Modelos en producción y rendimiento.	Jonnathan Campoberde
Administrador de Proyecto	Coordinación y gestión del proyecto.	Jonnathan Campoberde
Especialista en Negocio	Contexto y validación del dominio.	Hugo Pineda

2. Comprensión de los datos

2.1. Recolección de datos iniciales

Se cuenta con dos datasets exportados desde los sistemas transaccionales de la organización en formato .csv:

- **Datos de créditos [datos_creditos.csv]:** que contienen la información de los créditos solicitados por los clientes y si los mismos han sido considerados en mora en algún momento.
- **Datos de otros productos [datos_tarjetas.csv]:** que contienen la información sobre otros productos (en particular tarjetas de crédito) que poseen los clientes con la entidad y un resumen de su actividad y características principales.

2.2. Descripción de los datos

Se describen los contenidos de los datasets proporcionados:

Tabla 2. Contenidos de los datasets proporcionados

Dataset	Columnas	Cantidad de filas
datos_creditos.csv	id_cliente Edad importe_solicitado duracion_credito antiguedad_empleado situacion_vivienda ingresos objetivo_credito pct_ingreso tasa_interes estado_credito falta_pago	10127
datos_tarjetas.csv	id_cliente antiguedad_cliente estado_civil estado_cliente gastos_ult_12m genero limite_credito_tc nivel_educativo nivel_tarjeta operaciones_ult_12m personas_a_cargo	10127

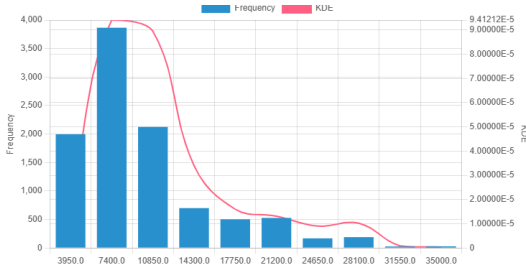
2.3. Exploración de datos

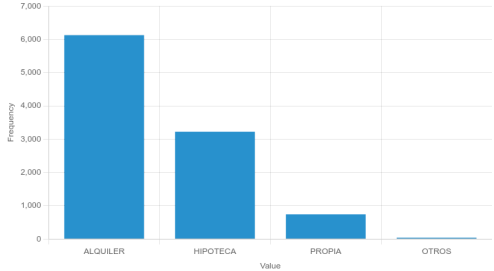
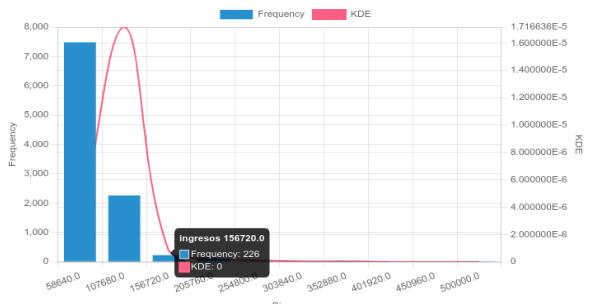
Se realiza un registro de los metadatos de cada dataset en las siguientes tablas:

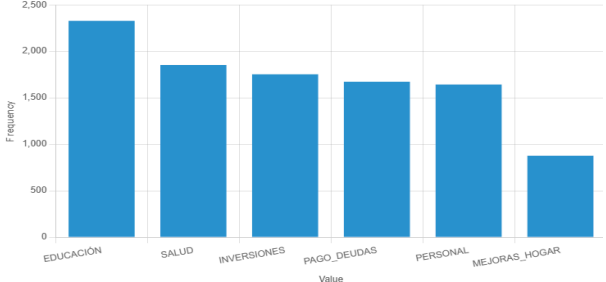
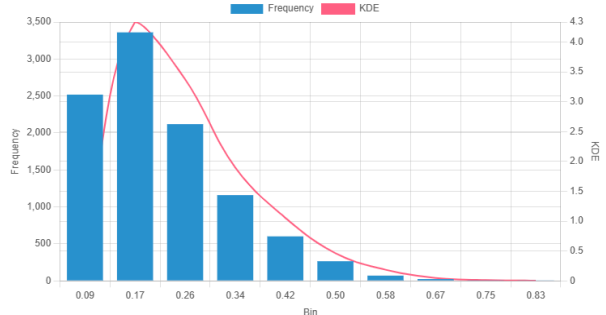
Referencia datasets:

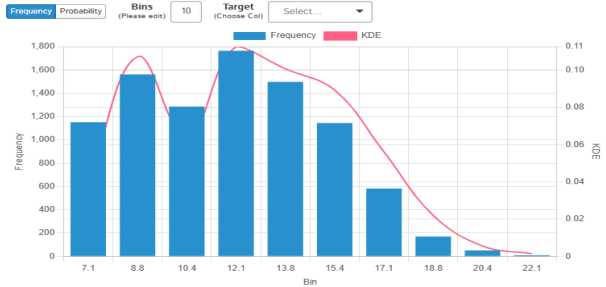
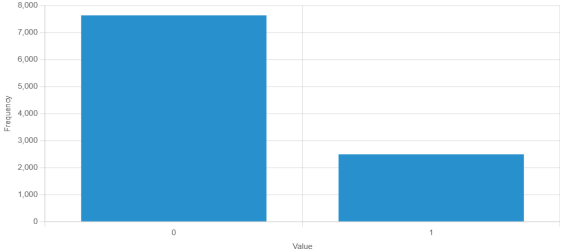
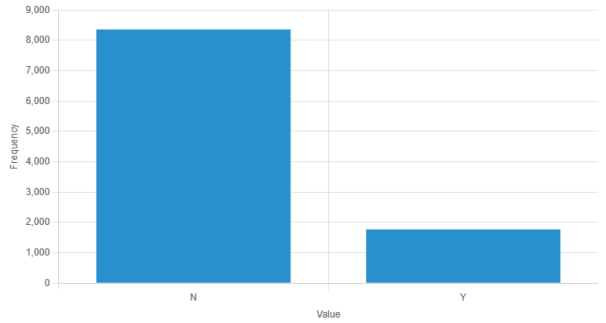
- **[C] = Datos de créditos**
- **[T] = Datos de tarjetas**

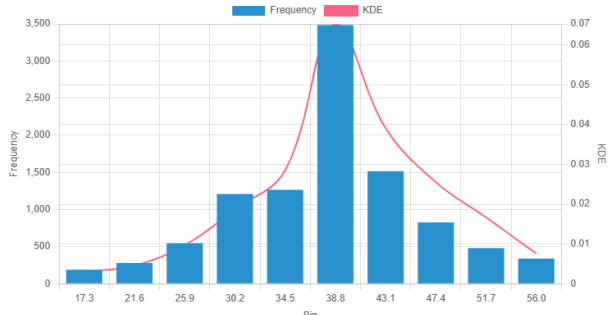
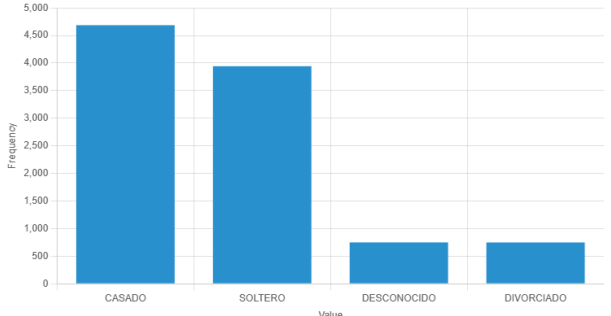
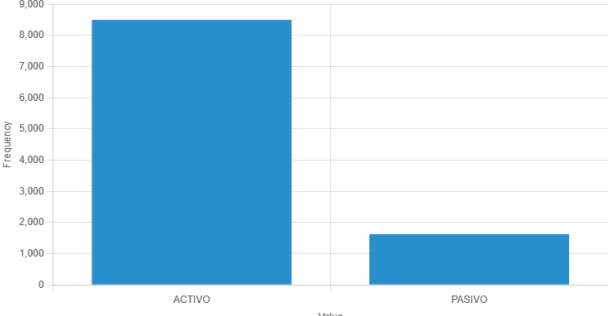
Tabla 3. Exploración de los datos proporcionados

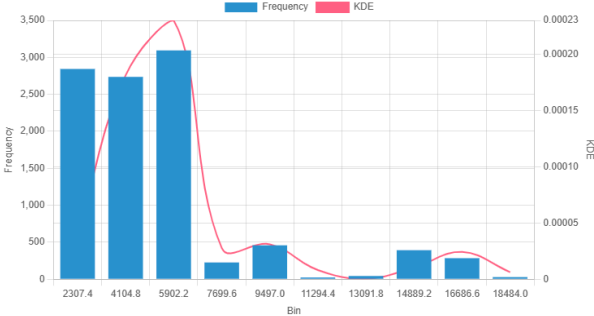
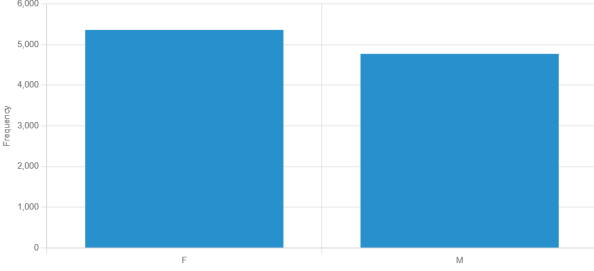
Dataset	Atributo / Columna	Tipo de dato	Observaciones
C	id_cliente	Float64 (Numérico)	Valores completos:10,127 Valores nulos: 0 (0%)
C	edad	Int64 (Numérico)	Valores completos:10,127 Valores nulos: 0 (0%) Distribución de valores: <ul style="list-style-type: none"> • 25%: 22 • 50%: 23 • 75%: 25 • Máximo: 144 • Media: 23,5727 • Mediana: 23 • Mínimo: 20 • Moda: 22
C	importe_solicitado	Int64 (Numérico)	Valores completos:10,127 Valores nulos: 0 (0%) Distribución de valores: <ul style="list-style-type: none"> • 25%: 4,425 • 50%: 6,500 • 75%: 10,000 • Máximo: 35,000 • Media: 8,138.7331 • Mediana: 6,500 • Mínimo: 500 • Moda: 5,000 
C	duracion_credito	Int64 (Numérico)	Valores completos:10,127 Valores nulos: 0 (0%) Distribución de valores: <ul style="list-style-type: none"> • 25%: 2 • 50%: 3 • 75%: 4 • Máximo: 4 • Media: 2.9956 • Mediana: 3 • Mínimo: 2 • Moda: 2
C	antiguedad_empleado	Float64 (Numérico)	Total de filas: 10,127 Conteo (no nulo): 9,790 Conteo (faltantes): 337

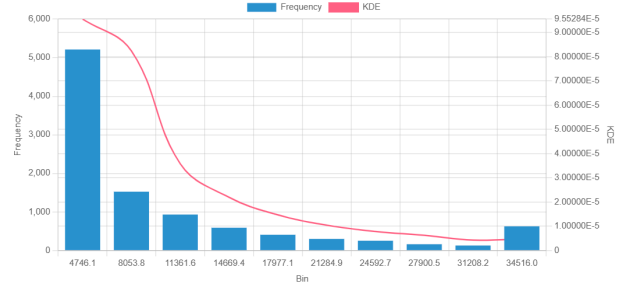
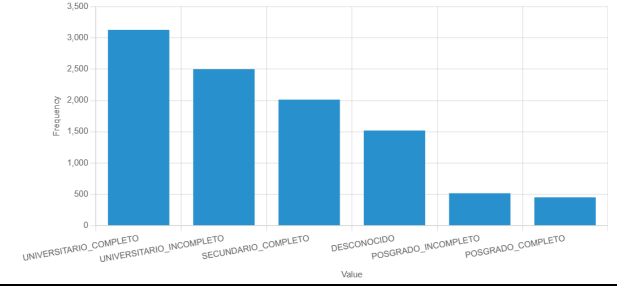
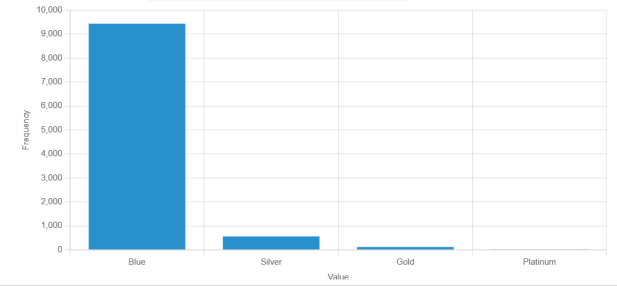
			<p>% Faltantes: 3.33%</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> • 25%: 2 • 50% (mediana): 4 • 75%: 6 • Máximo: 123 • Media: 3.9385 • Mediana: 4 • Mínimo: 0
C	situacion_vivienda	String (Categórico)	<p>Valores completos:10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores ("VALOR" - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • ALQUILER – 6125 (60.48%) • HIPOTECA – 3223 (31.83%) • PROPIA – 741 (7.32%) • OTROS – 38 (0.38%) 
C	ingresos	int64 (Numérico)	<p>Valores completos:10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> • 25%: 33,600 • 50%: 46,000 • 75%: 59,790.5 • Máximo: 500,000 • Media: 50,381.8976 • Mediana: 46,000 • Mínimo: 9,600 • Moda: 60,000 
C	objetivo_credito	String (Categórico)	<p>Valores completos: 10,127 Valores nulos: 0 (0%)</p>

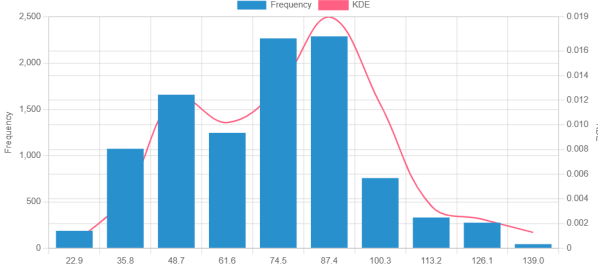
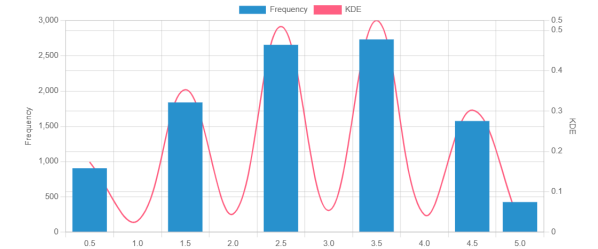
			<p>Distribución de valores ("VALOR" - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • EDUCACIÓN (2328) • SALUD (1853) • INVERSIONES (1753) • PAGO_DEUDAS (1673) • PERSONAL (1643) • MEJORAS_HOGAR (877) 
C	pct_ingreso	Float64 (Numérico)	<p>Valores completos: 10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> • 25%: 0,1 • 50%: 0,15 • 75%: 0,24 • Máximo: 0,83 • Media: 0,1772 • Mediana: 0,15 • Mínimo: 0,01 
C	tasa_interes	Float64 (Numérico)	<p>Total de filas: 10,127 Conteo (no nulo): 9,215 Conteo (faltantes): 912 % Faltantes: 9.01%</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> • 25%: 7.9 • 50%: 10.99 • 75%: 13.43 • Máximo: 22.11 • Media: 10.9794 • Mediana: 10.99 • Mínimo: 5.42

			
C	estado_credito	int64 (Numérico)	<p>Valores completos: 10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> • 25%: 0 • 50%: 0 • 75%: 0 • Máximo: 1 • Media: 0.2461 • Mediana: 0 • Mínimo: 0 • Moda: 0 
C	falta_pago	String (Categórico)	<p>Valores completos: 10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores ("VALOR" - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • N (8359) • Y (1768) 
T	id_cliente	Float64 (Numérico)	<p>Valores completos: 10,127 Valores nulos: 0 (0%)</p>
T	antigüedad_cliente	Float64 (Numérico)	<p>Valores completos: 10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> • 25%: 31 • 50%: 36

			<ul style="list-style-type: none"> • 75%: 40 • Máximo: 56 • Media: 35.9284 • Mediana: 36 • Mínimo: 13 
T	estado_civil	String (Categórico)	<p>Valores completos: 10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • CASADO (4687) • SOLTERO (3943) • DESCONOCIDO (749) • DIVORCIADO (748) 
T	estado_cliente	String (Categórico)	<p>Valores completos: 10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • ACTIVO (8500) • PASIVO (1627) 

T	gastos_ult_12m	Float64 (Numérico)	<p>Valores completos: 10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores: 25%: 2,155.5 50%: 3,899 75%: 4,741 Máximo: 18,484 Media: 4,404.0863 Mediana: 3,899 Mínimo: 510</p> 
T	genero	String (Categórico)	<p>Valores completos:10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> • F (5358) • M (4769) 
T	limite_credito_tc	Float64 (Numérico)	<p>Valores completos:10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores: 25%: 2,555 50%: 4,549 75%: 11,067.5 Máximo: 34,516 Media: 8,631.9537 Mediana: 4,549 Mínimo: 1,438.3</p>

			
T	nivel_educativo	String (Categórico)	<p>Valores completos:10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> ● UNIVERSITARIO_COMPLETO (3128) ● UNIVERSITARIO_INCOMPLETO (2500) ● SECUNDARIO_COMPLETO (2013) ● DESCONOCIDO (1519) ● POSGRADO_INCOMPLETO (516) ● POSGRADO_COMPLETO (451) 
T	nivel_tarjeta	String (Categórico)	<p>Valores completos:10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> ● Blue (9436) ● Silver (555) ● Gold (116) ● Platinum (20) 
T	operaciones_ult_12m	Float64 (Numérico)	<p>Valores completos:10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> ● 25%: 45 ● 50%: 67 ● 75%: 81 ● Máximo: 139 ● Media: 64.8587

			<ul style="list-style-type: none"> ● Mediana: 67 ● Mínimo: 10 
T	personas_a_cargo	Float64 (Numérico)	<p>Valores completos:10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> ● 25%:1 ● 50%:2 ● 75%:3 ● Máximo:5 ● Media:2.3462 ● Mediana:2 ● Mínimo:0 

2.4. Verificación de la calidad de los datos

2.4.1. Definición de objetivos y características de la evaluación inicial

2.4.1.1. Descripción del uso propuesto

Las autoridades de una entidad financiera desean obtener conocimiento a partir de su base de datos histórica de créditos otorgados.

Según este objetivo, se desarrolla un proyecto de ciencia de datos para desarrollar un producto de datos que sea una propuesta de solución para este escenario.

En este contexto, se requiere realizar un análisis de calidad de los datos disponibles para dar cumplimiento a lo establecido en la fase Comprensión de los Datos de la metodología CRISP-DM con la que se está gestionando el mencionado proyecto.

2.4.1.2. Definición de calidad

Para asegurar la calidad de los datos se va a analizar los siguientes atributos de calidad:

Tabla 4. Atributos de calidad de los datos

Atributo	Observaciones
Exactitud	Grado en el que los datos de un atributo representan un valor verdadero.
Compleitud	Grado en el que los datos de un registro tienen valores asociados a cada una de sus columnas y el dataset en general aplica el mismo criterio para todas sus filas.
Consistencia	Grado en el cual los datos son coherentes con otros datos del contexto y con los conjuntos de datos disponibles para este proyecto.

Cada una de las dimensiones definidas por los atributos antes listados, será relacionada con una o más características a analizar a fin de establecer la calidad de los datos disponibles:

Tabla 5. Dimensiones a analizar para la calidad de datos

Atributo	Características que analizar
Exactitud	Cumplimiento de reglas de formateo. Cumplimiento de reglas del negocio.
Compleitud	Compleitud de registros y del dataset.
Consistencia	Unicidad en atributo clave. Cumplimiento de integridad referencial.

2.4.2. Características que deben cumplir los datos

Los datos, para ser útiles y efectivos en el análisis, deben cumplir con una serie de características clave que se muestran en la tabla 6.

Tabla 6. Características de aceptación de los datos

Dimensión	Característica	Granularidad	Umbral de aceptación
Compleitud	Compleitud a nivel de filas	Filas	20%
	Compleitud a nivel del dataset	Dataset	10%
Exactitud	Cumplimiento de reglas de formateo	Dataset	10%
	Cumplimiento de reglas de valores	Filas	0%
	Cumplimiento de reglas de negocio	Dataset	10%
Consistencia	Unicidad en atributos clave	Dataset	0%
	Integridad referencial	Dataset	10%

2.4.3. Registro de metadatos de cada dataset

Los metadatos ya han sido registrados en los pasos previos motivo por el cual no se va a repetir en esta sección.

2.4.4. Evaluación inicial de los datos disponibles

Se inicia expresando la definición de las métricas aplicables para la medición de las características mencionadas en la sección anterior.

Tabla 7. Métricas para calidad de datos

Identificador	Descripción	Forma de realizar la medición	Umbral de aceptación
completitud_f	Complejidad a nivel de filas	atributos_vacios / total_atributos	20%
completitud_d	Complejidad a nivel del dataset	filas_con_vacios / total_filas	10%
formato_valido	Cumplimiento de reglas de formateo	filas_no_cumplen_formato / total_filas	10%
valores_ajustados	Cumplimiento de reglas de valores	filas_fuera_rango / total_filas	0%
valores_errores	Cumplimiento de reglas de negocio	filas_claves_duplicadas / total_filas	10%
claves_unicas	Unicidad en atributos clave	filas_con_problemas_relacion / total_filas	0%
integridad_referencial	Integridad referencial	filas_con_errores / total_filas	10%

Una vez aplicados los cálculos descritos en la tabla anterior se obtendrán los valores necesarios para realizar la evaluación de calidad de los datos en sí, los resultados se registran en las siguientes tablas.

Tabla 8. Resultados de la dimensión completitud

Indicador	Umbral	Resultados obtenidos	Evaluación
completitud_f	20%	Filas que incumplen el umbral de nulos en columnas [completitud_f]: - datos créditos: 0 (0.0)%	OK
completitud_d	10%	Filas que presentan nulos en el dataset [completitud_d]: - datos_credits: 1225 (12.1)%	No Cumplimiento

Tabla 9. Resultados de la dimensión exactitud

Indicador	Umbral	Resultados obtenidos	Evaluación
formato_valido	10%	No se encuentran atributos que registren un formato particular, por lo tanto, no se requiere el análisis.	OK
valores_ajustados	0%		No cumplimiento
Atributo: "edad"		Cantidad de filas con valores fuera de rango en atributo edad: 4 (0.04%)	No cumplimiento
Atributo: "importe_solicitado"		Cantidad de filas con valores fuera de rango en el atributo importe_solicitado: 0 (0%)	OK
Atributo: "duracion_credito"		Cantidad de filas con valores fuera de rango en atributo edad: 0 (0%)	OK
Atributo: "antiguedad_empleado"		Cantidad de filas con valores fuera de rango en el atributo antiguedad_empleado: 339 (3.35 %)	No cumplimiento

Atributo: "situacion_vivienda"	Cantidad de filas con valores fuera de rango en el atributo situacion_vivienda: 0 (0%)	OK
Atributo: "objetivo_credito"	Cantidad de filas con valores fuera de rango en el atributo objetivo_credito: 0 (0%)	OK
Atributo: "ingresos"	Cantidad de filas con valores fuera de rango en el atributo ingresos: 0 (0%)	OK
Atributo: "pct_ingreso"	Cantidad de filas con valores fuera de rango en el atributo pct_ingreso: 0 (0%)	OK
Atributo: "tasa_interes"	Cantidad de filas con valores fuera de rango en el atributo tasa_interes: 912 (9.01 %)	No cumplimiento
Atributo: "estado_credito"	Cantidad de filas con valores fuera de rango en el atributo estado_credito: 0 (0%)	OK
Atributo: "falta_pago"	Cantidad de filas con valores fuera de rango en el atributo falta_pago: 0 (0%)	OK
Atributo: "antiguedad_cliente"	Cantidad de filas con valores fuera de rango en el atributo antiguedad_cliente: 0 (0%)	OK
Atributo: "estado_civil"	Cantidad de filas con valores fuera de rango en el atributo gastos_ult_12m: 0 (0%)	OK
Atributo: "estado_cliente"	Cantidad de filas con valores fuera de rango en el atributo estado_cliente: 0 (0%)	OK
Atributo: "gastos_ult_12m"	Cantidad de filas con valores fuera de rango en el atributo gastos_ult_12m: 0 (0%)	OK
Atributo: "limite_credito_tc"	Cantidad de filas con valores fuera de rango en el atributo limite_credito_tc: 0 (0%)	OK
Atributo: "operaciones_ult_12m"	Cantidad de filas con valores fuera de rango en el atributo limite_credito_tc: 0 (0%)	OK
Atributo: "genero"	Cantidad de filas con valores fuera de rango en el atributo genero: 0 (0%)	OK
Atributo: "personas_a_cargo"	Cantidad de filas con valores fuera de rango en el atributo personas_a_cargo: 0 (0%)	OK
Atributo: "nivel_educativo"	Cantidad de filas con valores fuera de rango en el atributo nivel_educativo: 0 (0%)	OK
Atributo: "nivel_tarjeta"	Cantidad de filas con valores fuera de rango en el atributo nivel_tarjeta: 0 (0%)	OK
valores_errores	10%	OK
Regla de negocio #1	Casos de problemas según Regla de Negocio #1: 15 (0.15%)	OK
Regla de negocio #2	Casos de problemas según Regla de Negocio #2: 7 (0.07%)	OK

Tabla 10. Resultados de la dimensión consistencia

Indicador	Umbral	Resultados obtenidos	Evaluación
claves_unicas	0%	Cantidad de errores detectados: 0(0%)	OK
integridad_referencial	10%	Casos de problemas de integridad referencial: 0(0 %)	OK

3. Fase de preparación de los datos

3.1. Selección de datos

En función de los resultados del análisis de calidad de datos ejecutado, se ha determinado remover las siguientes columnas de cada dataset:

- **Datos_creditos:** no se realizan modificaciones al dataset.
- **Datos_tarjetas:** se elimina la columna “nivel_tarjeta” por su alta correlación con el monto del límite de la misma.

3.2. Limpieza de los datos

En esta actividad se aplicaron filtros a nivel de filas de los datasets disponibles con base en las recomendaciones del experto en el dominio participante del proyecto y los resultados del análisis de calidad ejecutado previamente:

Tabla 11. Filtros en la limpieza de los datos

Dataset	Atributo	Filtro aplicado	Observaciones
Datos_creditos	“edad”	Se eliminan las filas cuyo valor de edad sea > 90	Cantidad de filas filtradas: 4
	“antigüedad_empleado”	Se eliminan las filas que presentan valores > 100	Cantidad de filas filtradas: 2
	“regla_pct_ingresos”	Se eliminan las filas que presentan errores de cumplimiento de esta regla del negocio	Cantidad de filas filtradas: 15
	“regla_pct_ingresos_importe”	Se eliminan las filas que presentan errores de cumplimiento de esta regla del negocio	Cantidad de filas filtradas: 7
Datos_tarjetas			No se han encontrado filas a filtrar

4. Construcción de datos

Se detallan a continuación las operaciones de transformación de datos aplicadas sobre los datasets del escenario:

Tabla 12. Transformaciones realizadas sobre los datos

Atributo	Transformación aplicada
"estado_civil"	Cambios para mejorar la lectura de los datos: <ul style="list-style-type: none"> • 'CASADO': 'C', • 'SOLTERO': 'S', • 'DESCONOCIDO': 'N', • 'DIVORCIADO': 'D',
"estado_credito"	Cambios para mejorar la lectura de los datos: <ul style="list-style-type: none"> • 0: 'C' (completados) • 1: 'P' (pendientes)
"edad"	Valores numéricos fueron discretizados aplicando los siguientes rangos (etiqueta – rango establecido): <ul style="list-style-type: none"> • 'menor_25': (0, 25) • '25_a_30': [25, 50)
"antigüedad_empleado"	Valores numéricos fueron discretizados aplicando los siguientes rangos (etiqueta – rango establecido): <ul style="list-style-type: none"> • 'menor_5': [0, 5) • '5_a_10': [5, 10] • 'mayor_10': [11, 50]
"pct_ingreso"	Valores numéricos fueron discretizados aplicando los siguientes rangos (etiqueta – rango establecido): <ul style="list-style-type: none"> • 'hasta_20': [0.00, 0.20) • '20_a_40': [0.20, 0.40) • '40_a_60': [0.40, 0.60) • 'mayor_60': [0.60, 0.99)
"ingresos"	Valores numéricos fueron discretizados aplicando los siguientes rangos (etiqueta – rango establecido): <ul style="list-style-type: none"> • 'hasta_20k': [0, 20000) • '20k_a_50k': [20000, 50000) • '50k_a_100k': [50000, 100000) • 'mayor_100k': [100000, 1000000)
"tasa_interes"	Valores numéricos fueron discretizados aplicando los siguientes rangos (etiqueta – rango establecido): <ul style="list-style-type: none"> • 'hasta_7p': [0, 7) • '7p_a_15p': [7, 15) • '15p_a_20p': [15, 20) • 'mayor_20p': [20, 100)
"antigüedad_cliente"	Valores numéricos fueron discretizados aplicando los siguientes rangos (etiqueta – rango establecido): <ul style="list-style-type: none"> • 'menor_2y': [0, 24) • '2y_a_4y': [24, 48) • 'mayor_4y': [48, 100)
"limite_credito_tc"	Valores numéricos fueron discretizados aplicando los siguientes rangos (etiqueta – rango establecido): <ul style="list-style-type: none"> • 'menor_3k': [0, 3000) • '3k_a_5k': [3000, 5000) • '5k_a_10k': [5000, 10000) • 'mayor_10k': [10000, 100000)

"gastos_ult_12m"	Valores numéricos fueron discretizados aplicando los siguientes rangos (etiqueta – rango establecido): 'menor_2k': [0, 2000) '2k_a_4k': [2000, 4000) '4k_a_6k': [4000, 6000) '6k_a_8k': [6000, 8000) '8k_a_10k': [8000, 10000) 'mayor_10k': [10000, 100000)
"operaciones_ult_12m"	Valores numéricos fueron discretizados aplicando los siguientes rangos (etiqueta – rango establecido): 'menor_15': [0, 15) '15_a_30': [15, 30) '30_a_50': [30, 50) '50_a_75': [50, 75) '75_a_100': [75, 100) 'mayor_100': [100, 1000)

4.1. Integración de los datos

Se ha llevado a cabo la integración de ambos conjuntos de datos utilizando el atributo "id_cliente" como clave de unión. El resultado de esta operación ha sido un nuevo conjunto de datos denominado "datos_integrados".

El resultado de estas operaciones fue el siguiente:

- **Cantidad de filas:** 9765
- **Cantidad de columnas:** 23

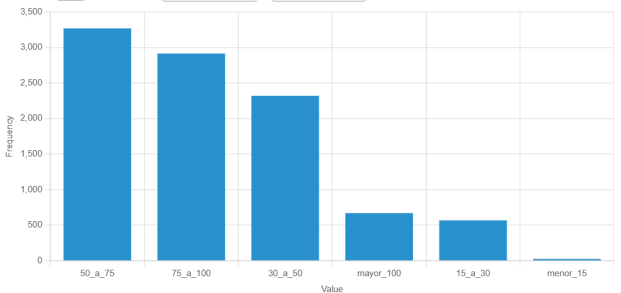
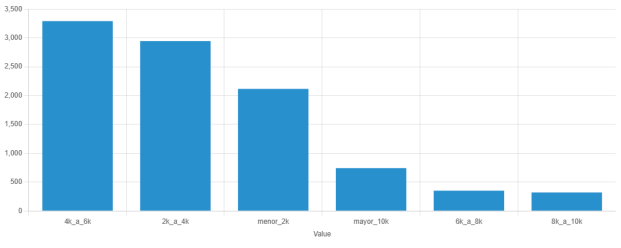

4.2. Formateo de los datos

En este apartado se registran los metadatos del dataset resultante de las operaciones ejecutadas en esta fase de la metodología. Además, se detallan los cambios de tipos de datos y la eliminación del atributo "id_cliente" al no ser necesario una vez integrado el dataset.

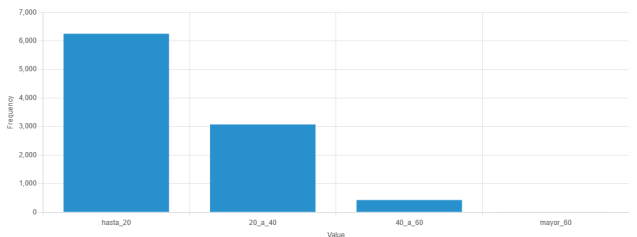
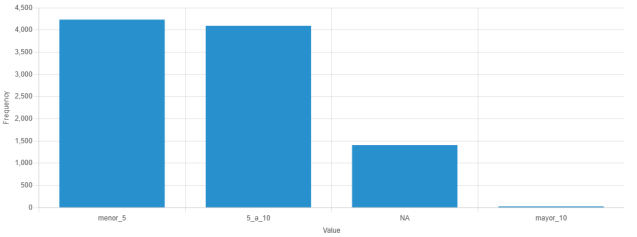
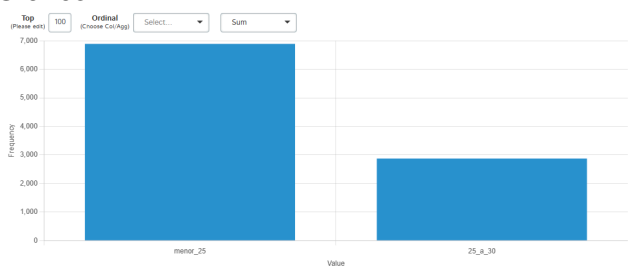
Estos cambios se han registrado en un nuevo archivo con el nombre "datos_final.csv".

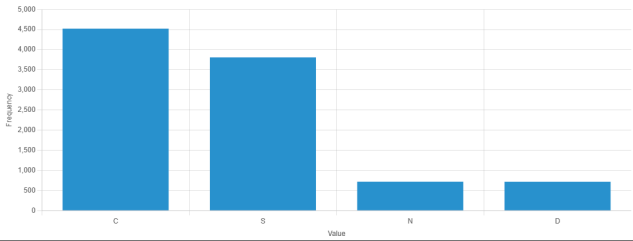
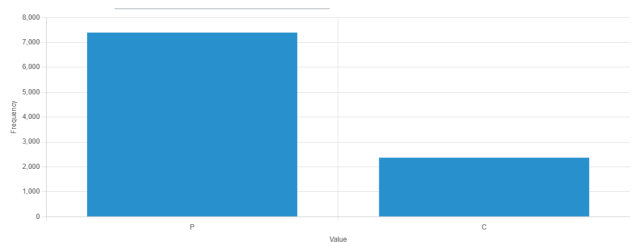
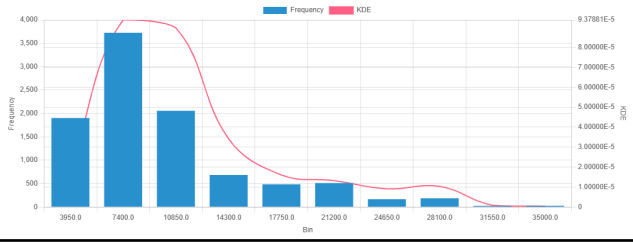
Tabla 13. Metadatos del dataset resultante

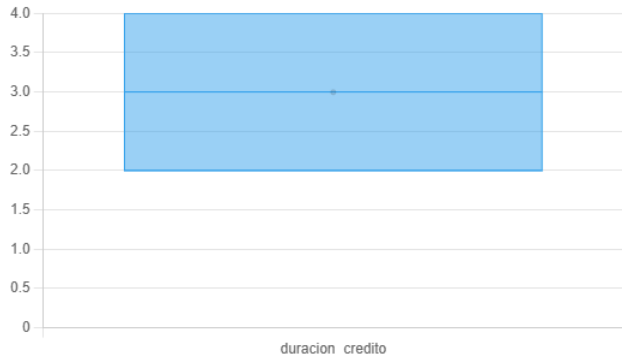
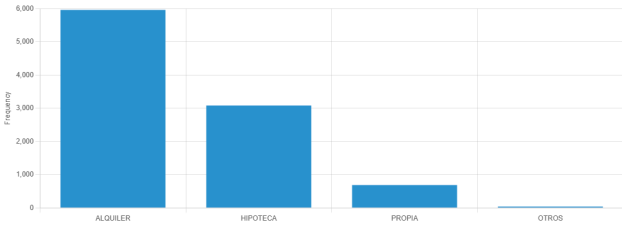
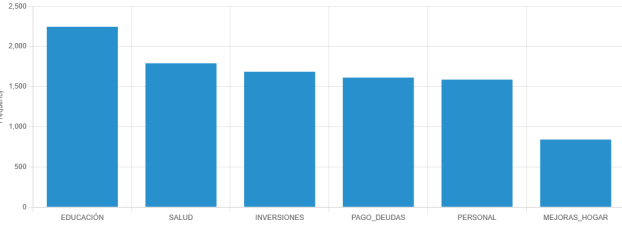
Atributo / Columna	Tipo de datos	Observaciones
operaciones_ult_12m	String	Valores completos: 9,765 Valores nulos: 0 (0%) Distribución de valores ("VALOR" - Cantidad de ocurrencias (%)): <ul style="list-style-type: none"> ● 50_a_75 (3269) ● 75_a_100 (2915) ● 30_a_50 (2320) ● mayor_100 (670) ● 15_a_30 (567) ● menor_15 (24) Gráfico:

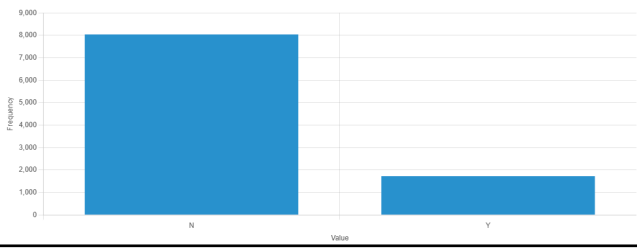
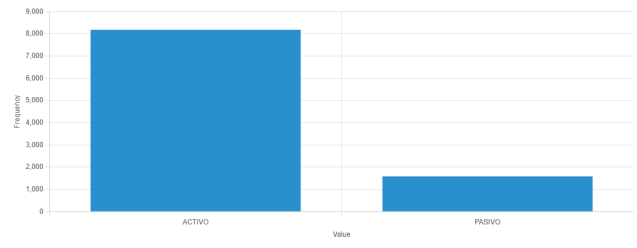
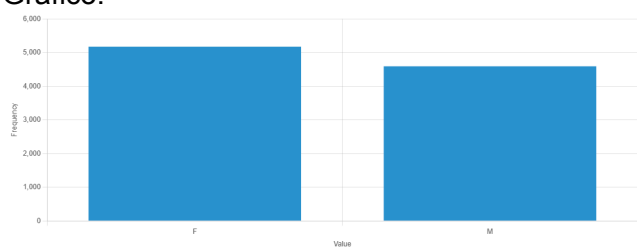
		<p>Top (Please edit) 100 Ordinal (Choose Col/Agg) Select... Sum</p>  <p>Frequency</p> <p>Value</p>
gastos_ult_12m	String	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • 4k_a_6k (3291) • 2k_a_4k (2945) • menor_2k (2116) • mayor_10k (742) • 6k_a_8k (351) • 8k_a_10k (320) <p>Gráfico:</p>  <p>Frequency</p> <p>Value</p>
limite_credito_tc	String	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • menor_3k (3321) • mayor_10k (2657) • 5k_a_10k (1931) • 3k_a_5k (1856) <p>Gráfico:</p>  <p>Frequency</p> <p>Value</p>
antigüedad_cliente	String	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • 2y_a_4y (8319)

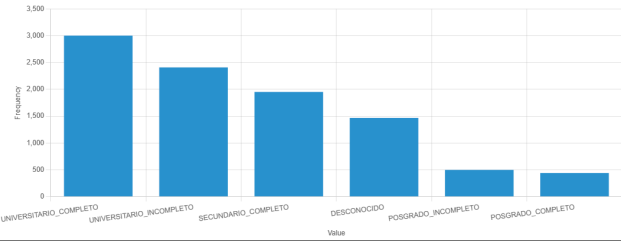
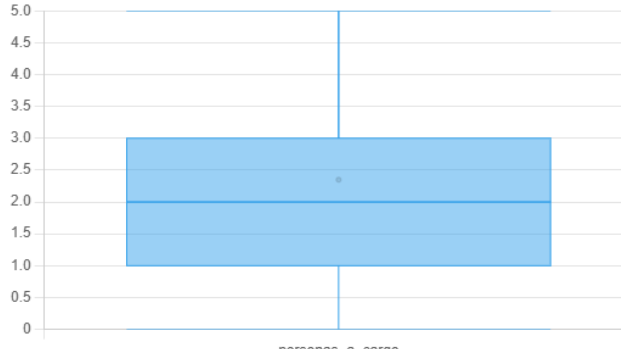
		<ul style="list-style-type: none"> menor_2y (816) mayor_4y (630) <p>Gráfico</p>
tasa_interes	String	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> 7p_a_15p (6809) hasta_7p (1060) 15p_a_20p (992) mayor_20p (17) <p>Gráfico:</p>
ingresos	String	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> 20k_a_50k (4961) 50k_a_100k (3927) mayor_100k (460) hasta_20k (417) <p>Gráfico:</p>
pct_ingreso	String	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> hasta_20 (6253) 20_a_40 (3074) 40_a_60 (426) mayor_60 (12)

		<p>Gráfico:</p> 
antigüedad_emplea do	String	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • menor_5 (4235) • 5_a_10 (4095) • NA (1408) • mayor_10 (27) <p>Gráfico:</p> 
edad	String	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • menor_25 (6891) • 25_a_30 (2874) <p>Gráfico:</p> 
estado_civil_N	String	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • C (4522) • S (3808) • N (718) • D (717) <p>Gráfico:</p>

		
estado_credito_N	String	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • P (7395) • C (2370) <p>Gráfico:</p> 
importe_solicitado	Int64 (Numérico)	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> • 25%: 4,500 • 50%: 6,500 • 75%: 10,000 • Máximo: 35,000 • Media: 8,171.7947 • Mediana: 6,500 • Mínimo: 500 • Moda: 5,000 
duracion_credito	Int64 (Numérico)	<p>Valores completos: 9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> • 25%: 2 • 50%: 3 • 75%: 4 • Máximo: 4 • Media: 2.9961 • Mediana: 3 • Mínimo: 2 • Moda: 2 <p>Gráfico:</p>

		 <p>duracion_credito</p>
situacion_vivienda	String (Categórico)	<p>Valores completos:9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores ("VALOR" - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> ● ALQUILER (5960) ● HIPOTECA (3081) ● PROPIA (686) ● OTROS (38) <p>Gráfico:</p> 
objetivo_credito	String	<p>Valores completos:9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores ("VALOR" - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> ● EDUCACIÓN (2246) ● SALUD (1791) ● INVERSIONES (1686) ● PAGO_DEUDAS (1613) ● PERSONAL (1588) ● MEJORAS_HOGAR (841) <p>Gráfico:</p> 
falta_pago	String	<p>Valores completos:9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores ("VALOR" - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> ● N (8039)

		<ul style="list-style-type: none"> • Y (1726) <p>Gráfico:</p> 
estado_cliente	String	<p>Valores completos:9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • ACTIVO (8182) • PASIVO (1583) <p>Gráfico:</p> 
genero	String	<p>Valores completos:9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • F (5174) • M (4591) <p>Gráfico:</p> 
nivel_educativo	String	<p>Valores completos:9,765 Valores nulos: 0 (0%)</p> <p>Distribución de valores (“VALOR” - Cantidad de ocurrencias (%)):</p> <ul style="list-style-type: none"> • UNIVERSITARIO_COMPLETO (3001) • UNIVERSITARIO_INCOMPLETO (2408) • SECUNDARIO_COMPLETO (1951) • DESCONOCIDO (1467) • POSGRADO_INCOMPLETO (498) • POSGRADO_COMPLETO (440) <p>Gráfico:</p>

		
personas_a_cargo	Float64 (Numérico)	<p>Valores completos:10,127 Valores nulos: 0 (0%)</p> <p>Distribución de valores:</p> <ul style="list-style-type: none"> • 25%:1 • 50%:2 • 75%:3 • Máximo:5 • Media:2.3489 • Mediana:2 • Mínimo:0 <p>Gráfico:</p> 

Con la ejecución de estas actividades, se da por concluido el primer sprint del proyecto. A continuación, se adjuntan capturas de pantalla que documentan el registro de las tareas completadas en la herramienta de seguimiento (GitHub Projects):

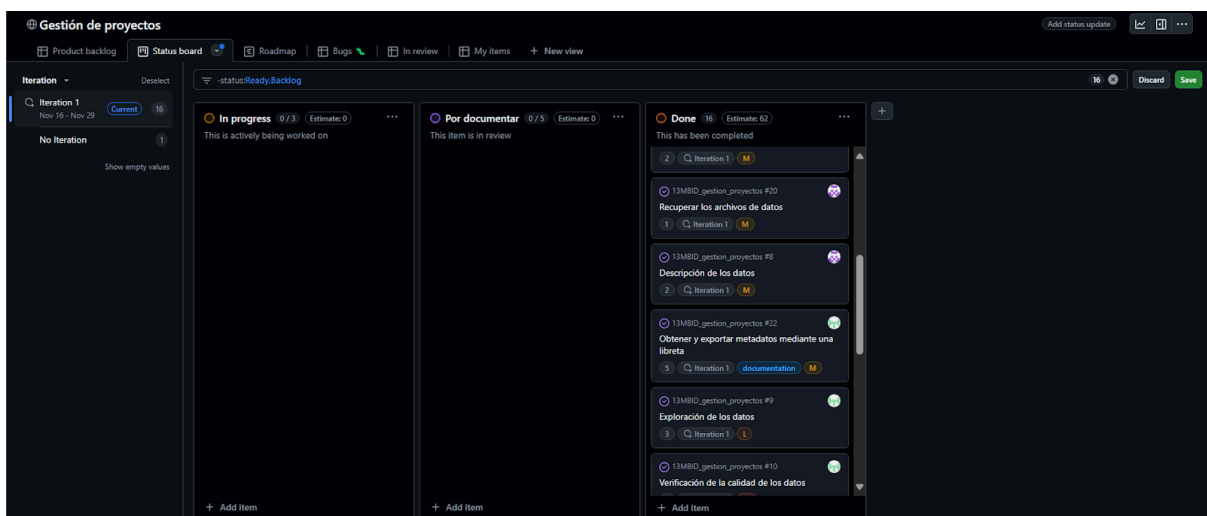


Figura 5. Finalización de la primera iteración del proyecto

Se adjunta el enlace del repositorio público para que pueda ser accedido: https://github.com/JonnHenry/13MBID_gestion_proyectos.git, adicionalmente se adjunta la captura de la invitación al repositorio.

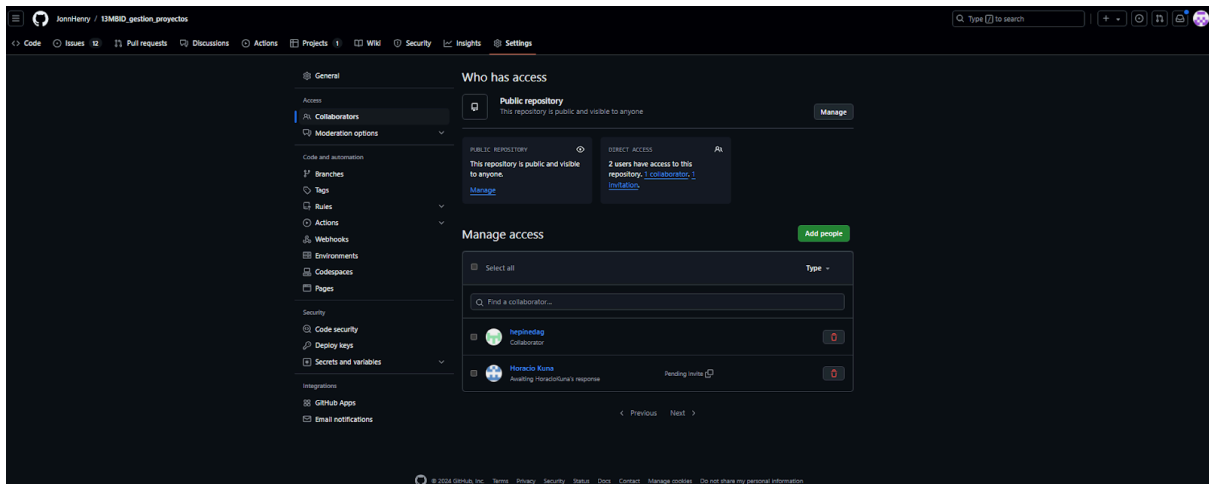


Figura 6. Invitación realizada para la revisión del proyecto

Se ha definido la 2da. Iteración del proyecto en ejecución, para esta acción se ha utilizado nuevamente la plataforma GitHub Projects en el mismo proyecto de la iteración anterior. En este caso, las fases de la metodología CRISP-DM abarcadas son:

- Modelado
- Evaluación
- Despliegue

A continuación se muestran las capturas de pantalla sobre la definición de la iteración 2:

Gestión de proyectos									
Filter by keyword or by field									
Title	Status	Assignees	Size	Estimate	Linked pull requests	Iteration			
No Priority 13 Estimate: 47									
1 Selección de la técnica de modelado #16	Ready		XL	3		Iteration 2			
2 Generación de un plan de pruebas #17	Ready		L	5		Iteration 2			
3 Construcción del modelo #18	Ready		L	5		Iteration 2			
4 Generar la libreta para la experimentación #32	Ready		XL	5		Iteration 2			
5 Ejecutar la experimentación #33	Ready		L	3		Iteration 2			
6 Evaluación del modelo #19	Ready		L	5		Iteration 2			
7 Evaluación de los resultados #25	Ready		L	3		Iteration 2			
8 Proceso de revisión de los resultados #26	Ready		L	3		Iteration 2			
9 Determinación de futuras tareas #27	Ready		XL	5		Iteration 2			
10 Plan de implantación #28	Ready		M	2		Iteration 2			
11 Supervisión y mantenimiento #29	Ready		S	1		Iteration 2			
12 Redacción del informe final #30	Ready		XL	5		Iteration 2			
13 Revisión del proyecto #31	Ready		M	2		Iteration 2			

Figura 6. Proyecto en su primera iteración

5. Modelado

5.1. Selección de la técnica de modelado

En función de los objetivos del proyecto, se registran las decisiones de interés con respecto a la selección de la técnica de modelado para obtener el producto solicitado.

Las técnicas seleccionadas en primera instancia son:

- Métodos de la familia TDIDT (Árboles de Decisión)
- Métodos de regresión
- Métodos de ensamblado

5.2. Generación del plan de pruebas

Para el plan de pruebas, se han definido las siguientes acciones y lineamientos clave que garantizarán un enfoque estructurado y orientado a los objetivos del proyecto. Este plan tiene como propósito optimizar los resultados obtenidos en términos de efectividad y documentar detalladamente el proceso experimental:

5.2.1. División del Conjunto de Datos

Se dividirá el conjunto de datos disponible en proporciones estándar para crear los datasets requeridos:

- Datos de entrenamiento: 75% del total, utilizado para ajustar los modelos.
- Datos de prueba: 25% del total, reservado para evaluar el rendimiento del modelo.

Esta división asegura una validación robusta y minimiza el riesgo de sobreajuste.

5.2.2. Ejecución de Instancias de Prueba

Se llevarán a cabo tres (3) instancias de prueba, diseñadas para iterar sobre las técnicas y parámetros utilizados, con el objetivo de alcanzar un nivel de efectividad mínimo del 80% establecido por los objetivos del proyecto. Durante cada instancia:

- Se experimentará con diversas técnicas de modelado y enfoques algorítmicos.
- Los parámetros específicos de cada ejecución serán registrados para permitir su replicabilidad.
- Los resultados serán evaluados con métricas clave que reflejen la efectividad y la calidad del modelo.

5.2.3. Filtrado Progresivo

A medida que avancen las instancias de prueba:

- Las técnicas que muestren un rendimiento inferior al promedio serán descartadas para concentrar recursos en enfoques más prometedores.
- Se realizará un análisis comparativo de las métricas obtenidas para identificar patrones de mejora.

5.2.4. Registro Detallado con MLflow

Toda la experimentación será documentada rigurosamente utilizando la librería mlflow. Esto incluye:

- El registro de los parámetros de configuración, métricas y artefactos generados por cada modelo.
- La generación de un historial completo de ejecuciones para facilitar la trazabilidad y el análisis posterior.

5.3. Construcción del Modelo

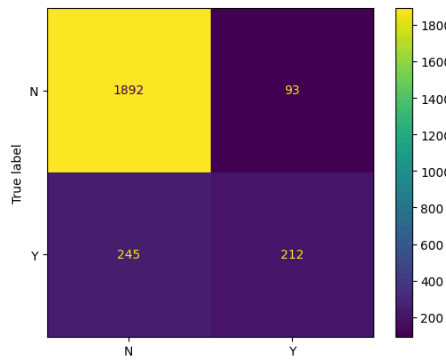
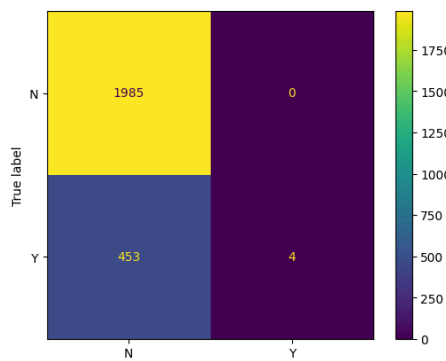
En esta actividad se hará uso de varias librerías de Python con el objetivo de implementar modelos de predicción de manera eficiente y sencilla. Estas herramientas permitirán realizar un análisis detallado del entorno de trabajo previamente definido, facilitando la preparación, el procesamiento y la interpretación de los datos. Además, se explorarán diferentes enfoques para optimizar los modelos y evaluar su desempeño. El repositorio que almacena el código y la documentación de esta experimentación, incluyendo scripts y notebooks relacionados, se encuentra disponible en: https://github.com/JonnHenry/13MBID_gestion_proyectos

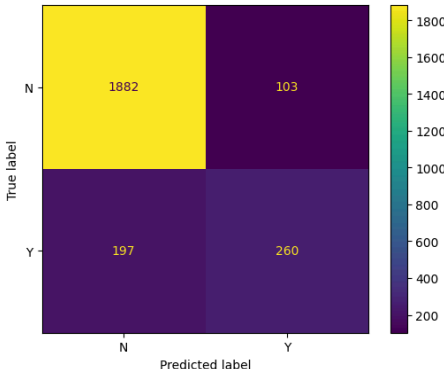
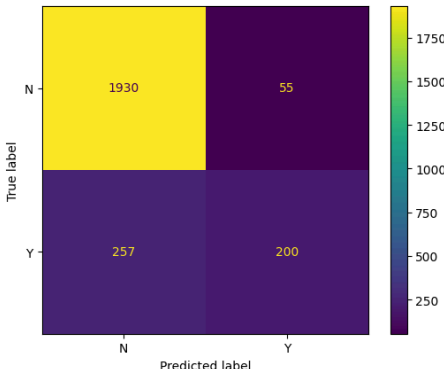
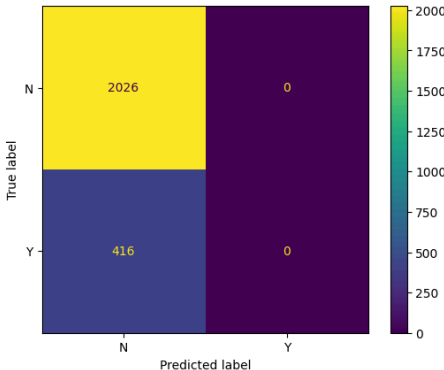
5.3.1. Evaluación del modelo

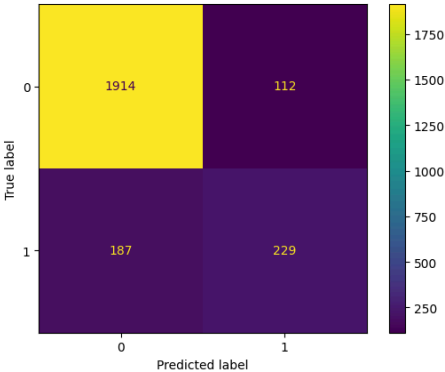
Se documenta en esta sección la ejecución de cada instancia de prueba con las técnicas empleadas y sus resultados obtenidos:

5.3.1.1. Prueba #1

Tabla 14. Técnicas de modelado y sus resultados en la primera prueba

Técnica utilizada	Parametrización	Resultados obtenidos
LogisticRegression	C: 1.0 class_weight: None dual: False fit_intercept: True intercept_scaling: 1 l1_ratio: None max_iter: 100 multi_class: deprecated n_jobs: None penalty: l2 random_state: None solver: liblinear tol: 0.0001 verbose: 0 warm_start: False	Rendimiento obtenido: 0.8616 Validación cruzada: 0.8704 Matriz de confusión: 
KNeighborsClassifier	algorithm: ball_tree leaf_size: 25 metric: minkowski metric_params: None n_jobs: None n_neighbors: 50 p: 2 weights: uniform	Rendimiento obtenido: 0.8145 Validación cruzada: 0.8268 Matriz de confusión: 
DecisionTreeClassifier	ccp_alpha: 0.0	Rendimiento obtenido: 0.8771

	<div>class_weight: None criterion: entropy max_depth: 3 max_features: None max_leaf_nodes: None min_impurity_decrease: 0.0 min_samples_leaf: 1 min_samples_split: 10 min_weight_fraction_leaf: 0.0 monotonic_cst: None random_state: None splitter: best</div>	<div>Validación cruzada: 0.8812</div> <div>[1] Árbol obtenido</div> <div>Matriz de confusión:</div> <div><table><thead><tr><th></th><th>N</th><th>Y</th></tr></thead><tbody><tr><th>N</th><td>1882</td><td>103</td></tr><tr><th>Y</th><td>197</td><td>260</td></tr></tbody></table></div>		N	Y	N	1882	103	Y	197	260
	N	Y									
N	1882	103									
Y	197	260									
RandomForestClassifier	<div>bootstrap: True ccp_alpha: 0.0 class_weight: None criterion: gini max_depth: None max_features: sqrt max_leaf_nodes: None max_samples: None min_impurity_decrease: 0.0 min_samples_leaf: 1 min_samples_split: 2 min_weight_fraction_leaf: 0.0 monotonic_cst: None n_estimators: 10 n_jobs: None oob_score: False random_state: None verbose: 0 warm_start: False</div>	<div>Rendimiento obtenido: 0.8722 Validación cruzada: 0.876</div> <div>Matriz de confusión:</div> <div><table><thead><tr><th></th><th>N</th><th>Y</th></tr></thead><tbody><tr><th>N</th><td>1930</td><td>55</td></tr><tr><th>Y</th><td>257</td><td>200</td></tr></tbody></table></div>		N	Y	N	1930	55	Y	257	200
	N	Y									
N	1930	55									
Y	257	200									
Support Vector Machine	Vector <div>break_ties: False C: 1 cache_size: 200 class_weight: None coef0: 0.0 decision_function_shape: ovr degree: 3 gamma: scale kernel: rbf max_iter: -1 probability: False random_state: None shrinking: True tol: 0.001</div>	<div>Rendimiento obtenido: 0.8296 Validación cruzada: 0.8211</div> <div>Matriz de confusión:</div> <div><table><thead><tr><th></th><th>N</th><th>Y</th></tr></thead><tbody><tr><th>N</th><td>2026</td><td>0</td></tr><tr><th>Y</th><td>416</td><td>0</td></tr></tbody></table></div>		N	Y	N	2026	0	Y	416	0
	N	Y									
N	2026	0									
Y	416	0									

	verbose: False										
XGBClassifier	base_score: None booster: None colsample_bylevel: None colsample_bynode: None colsample_bytree: None custom_metric: None device: None early_stopping_rounds: None eval_metric: None gamma: None grow_policy: None interaction_constraints: None learning_rate: None maximize: None max_bin: None max_cat_threshold: None max_cat_to_onehot: None max_delta_step: None max_depth: None max_leaves: None min_child_weight: None monotone_constraints: None multi_strategy: None num_boost_round: 100 num_parallel_tree: None n_jobs: None objective: binary:logistic random_state: 42 reg_alpha: None reg_lambda: None sampling_method: None scale_pos_weight: None subsample: None tree_method: None validate_parameters: None verbose_eval: True verbosity: None	Rendimiento obtenido: 0.8776 Validación cruzada: 0.8729 Matriz de confusión:  <table border="1"> <thead> <tr> <th></th> <th>Predicted 0</th> <th>Predicted 1</th> </tr> </thead> <tbody> <tr> <th>True 0</th> <td>1914</td> <td>112</td> </tr> <tr> <th>True 1</th> <td>187</td> <td>229</td> </tr> </tbody> </table>		Predicted 0	Predicted 1	True 0	1914	112	True 1	187	229
	Predicted 0	Predicted 1									
True 0	1914	112									
True 1	187	229									

[1] Árbol obtenido:

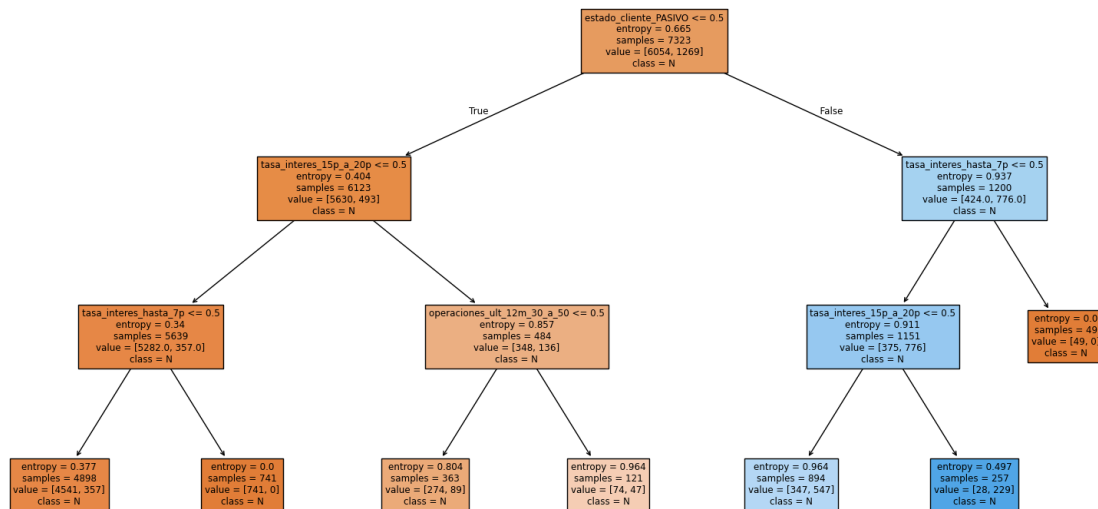


Figura 7. Árbol de decisiones obtenido en la primera iteración

Comparación de los resultados obtenidos en esta instancia de pruebas, basada en las métricas de rendimiento almacenadas en MLflow, con un enfoque particular en las siguientes métricas: `training_accuracy_score`, `training_f1_score`, `training_recall_score` y `training_roc_auc`. Estas métricas proporcionan una visión detallada del desempeño del modelo en términos de precisión, capacidad de clasificación, recuperación y área bajo la curva ROC durante el entrenamiento.

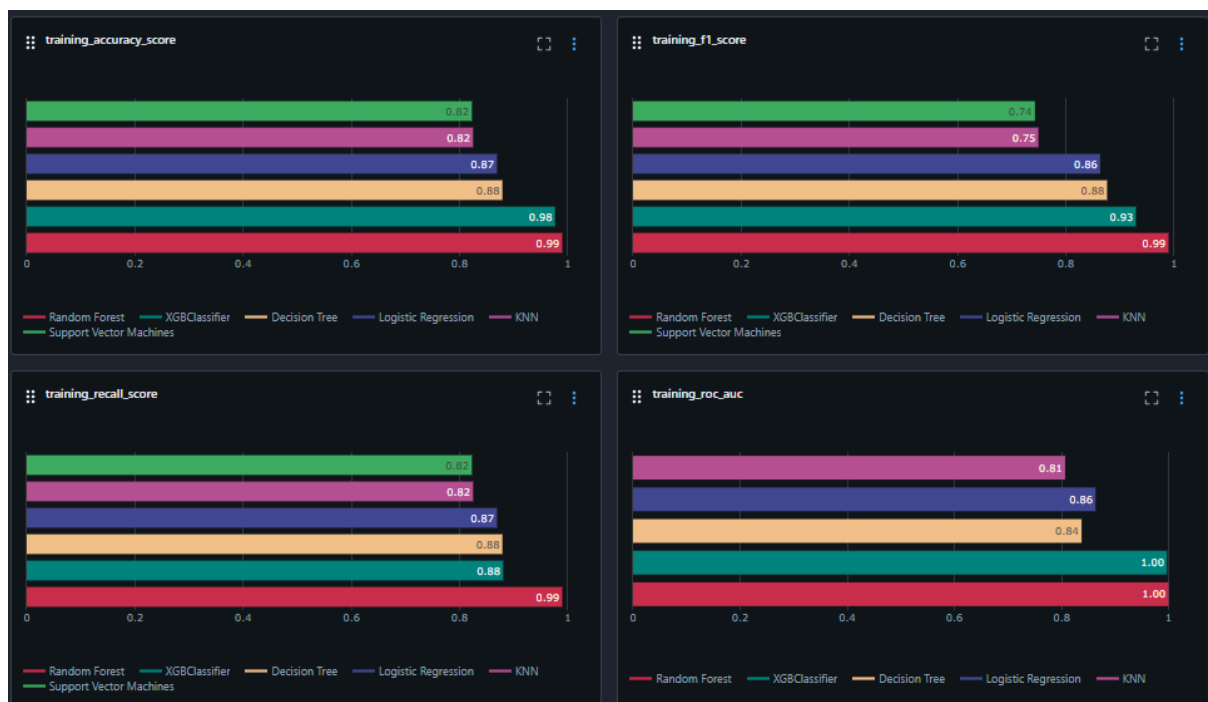
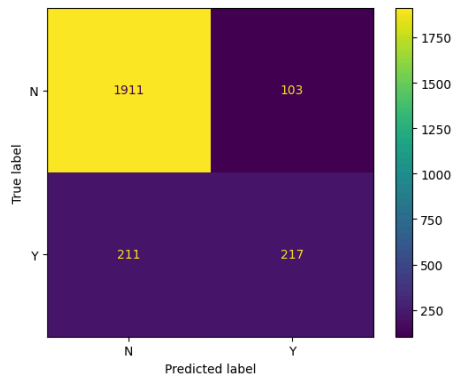
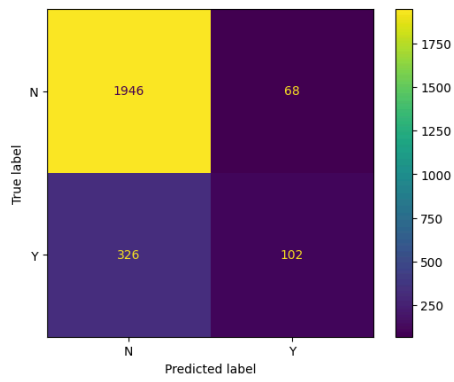


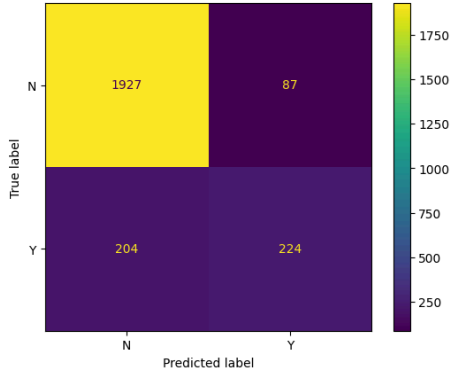
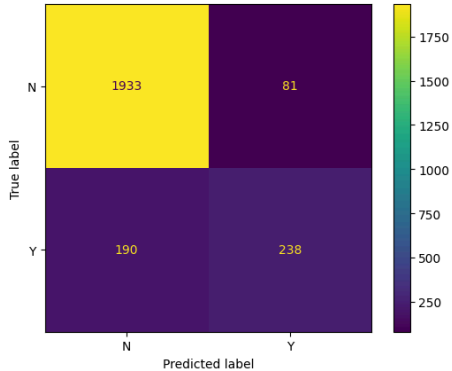
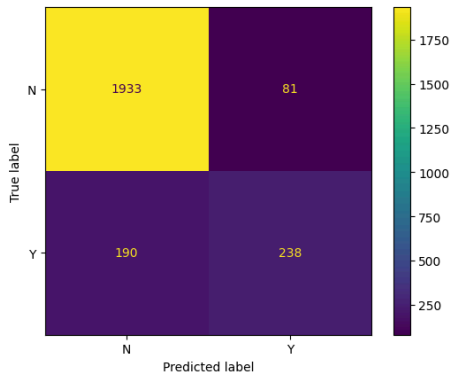
Figura 8. Resultados obtenidos en el primer experimento

5.3.1.2. Prueba #2

Basándonos en los resultados obtenidos en la instancia anterior, se decide no proceder con el análisis de la técnica SVM con optimización de hiperparámetros debido a su bajo rendimiento. En su lugar, se ajustan los parámetros de las técnicas restantes con el objetivo de optimizar sus resultados, empleando los enfoques de RandomizedSearchCV y GridSearchCV.

Tabla 15. Técnicas de modelado y sus resultados en la segunda prueba

Técnica utilizada	Parametrización	Resultados obtenidos
Logistic Regression	C: 2.848404943774676 class_weight: None dual: False fit_intercept: True intercept_scaling: 1 l1_ratio: 0.1 max_iter: 100 multi_class: deprecated n_jobs: None penalty: l2 random_state: None solver: liblinear tol: 0.0001 verbose: 0 warm_start: False	Rendimiento obtenido: 0.8714 Validación cruzada: 0.871 Matriz de confusión: 
KNeighborsClassifier	algorithm: ball_tree leaf_size: 12 metric: minkowski metric_params: None n_jobs: None n_neighbors: 6 p: 1 weights: distance	Rendimiento obtenido: 0.8387 Validación cruzada: 0.8412 Matriz de confusión: 
DecisionTreeClassifier	ccp_alpha: 0.0 class_weight: None criterion: entropy max_depth: 7 max_features: None max_leaf_nodes: None min_impurity_decrease: 0.0 min_samples_leaf: 18 min_samples_split: 2 min_weight_fraction_leaf: 0.0	Rendimiento obtenido: 0.8808 Validación cruzada: 0.8815 [2] Árbol obtenido Matriz de confusión:

	monotonic_cst: None random_state: None splitter: best	
RandomForestClassifier	bootstrap: False ccp_alpha: 0.0 class_weight: None criterion: gini max_depth: 10 max_features: sqrt max_leaf_nodes: None max_samples: None min_impurity_decrease: 0.0 min_samples_leaf: 1 min_samples_split: 7 min_weight_fraction_leaf: 0.0 monotonic_cst: None n_estimators: 30 n_jobs: None oob_score: False random_state: None verbose: 0 warm_start: False	Rendimiento obtenido: 0.889 Validación cruzada: 0.8824 Matriz de confusión: 
XGBClassifier	objective: binary:logistic base_score: None booster: None callbacks: None colsample_bylevel: None colsample_bynode: None colsample_bytree: 0.6 device: None early_stopping_rounds: None enable_categorical: False eval_metric: None feature_types: None gamma: 0.05 grow_policy: None importance_type: None interaction_constraints: None learning_rate: 0.01 max_bin: None max_cat_threshold: None	Rendimiento obtenido: 0.889 Validación cruzada: 0.8806 Matriz de confusión: 

	max_cat_to_onehot: None max_delta_step: None max_depth: 5 max_leaves: None min_child_weight: 5 missing: nan monotone_constraints: None multi_strategy: None n_estimators: 550 n_jobs: None num_parallel_tree: None random_state: None reg_alpha: None reg_lambda: None sampling_method: None scale_pos_weight: None subsample: 0.8 tree_method: None validate_parameters: None verbosity: None	
--	---	--

[2] Árbol obtenido:

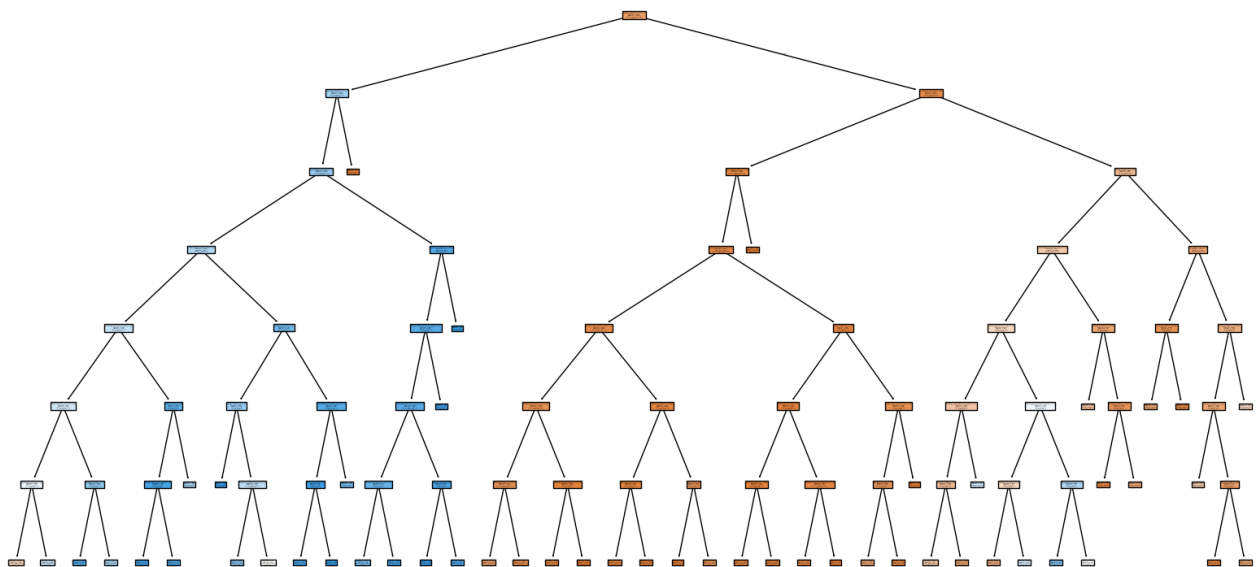


Figura 9. Árbol de decisiones obtenido en la primera iteración



Figura 10. Resultados obtenidos en el segundo experimento

6. Evaluación

6.1. Evaluación de los resultados

En base a los resultados obtenidos del plan de pruebas, se ha seleccionado la técnica de ensamblado de modelos **RandomForest** para realizar las predicciones sobre nuevos créditos. Esta elección se justifica por la alta efectividad observada en las distintas instancias de evaluación, alcanzando un 92% en las métricas `training_accuracy_score`, `training_f1_score` y `training_recall_score`, y un 96% en la métrica `training_roc_auc` en los datos de entrenamiento. En los datos de evaluación, el modelo muestra un excelente desempeño para la clase N, con una precisión de 0.91, recall de 0.96 y f1-score de 0.93, lo que indica una alta eficacia en la predicción de casos negativos. Para la clase Y, aunque los resultados son más bajos, con una precisión de 0.75 y un f1-score de 0.64, sigue demostrando un buen rendimiento.

6.2. Proceso de revisión

Según los datos obtenidos en las etapas previas, la técnica seleccionada en el apartado anterior será aplicada al conjunto de datos denominado "**datos_nuevos.csv**", el cual corresponde a los registros de los nuevos créditos otorgados por la entidad financiera. Este conjunto de datos incluye información relevante de los nuevos solicitantes y sus respectivos créditos, que se utilizará para predecir la probabilidad de incumplimiento en los pagos. El objetivo principal de esta predicción es estimar el valor del atributo "**falta_pago**", que indica si cada crédito tiene el potencial de entrar en mora (es decir, si se producirá un cese de pagos en el futuro).

La aplicación de la técnica elegida anteriormente permitirá identificar patrones y tendencias en los datos que podrían predecir el comportamiento de los nuevos créditos en relación con el riesgo de impago. Los resultados obtenidos a partir de esta predicción serán documentados detalladamente en el apartado de **"Informe Final"** de este documento, donde se incluirán las métricas de evaluación, análisis de los resultados y recomendaciones para la entidad financiera.

6.3. Determinación de futuras tareas

Como tareas a implementar en la continuidad del proyecto, donde se podrían realizar futuras iteraciones para mejorar la precisión del modelo:

- Optimización y Retroalimentación, se deben realizar ajustes en los modelos si los resultados de la evaluación anterior no son satisfactorios. Es importante incorporar nuevos atributos que permitan tener más información de los solicitantes como pueden ser ingresos totales del hogar, edades de las personas a cargo, entre otros. Evaluar la robustez de los nuevos modelos entrenados con los datos y analizar los resultados obtenidos para descubrir nuevos patrones o insights que puedan mejorar la interpretación de los datos y el modelado.
- Innovación y Expansión, explorar nuevas tecnologías y enfoques emergentes que puedan optimizar el modelo, como el deep learning o técnicas de aprendizaje no supervisado. También se debe evaluar la expansión de la solución, integrando más fuentes de datos o extendiendo su aplicabilidad a otras áreas del negocio.

Para la próxima iteración del proyecto, se propone ejecutar las siguientes tareas:

- Mejorar la preparación de los datos mediante el uso de técnicas avanzadas de limpieza y transformación. Por ejemplo, implementar algoritmos de imputación más sofisticados para datos faltantes o probar técnicas de normalización y escalado de datos más robustas. Además, se podrían explorar nuevos enfoques de selección de características, como el uso de métodos automáticos de reducción de dimensionalidad (como PCA) o la eliminación de multicolinealidad, para optimizar el rendimiento del modelo.
- Explorar técnicas de modelado no supervisado como clustering o análisis de componentes principales (PCA) para obtener insights adicionales sobre los datos. Estas técnicas pueden revelar patrones ocultos que no son evidentes con los modelos supervisados, lo que puede enriquecer la interpretación y aportar valor a las decisiones del negocio.

7. Despliegue / Implementación

7.1. Plan de implementación

Las autoridades de la entidad financiera han decidido que el modelo desarrollado se utilice como una herramienta clave para el asesoramiento en el sector de monitoreo de créditos de la entidad. Asimismo, se ha dispuesto realizar actualizaciones periódicas del modelo, incorporando nuevos datos que puedan ser recolectados a lo largo del próximo año, con el fin de mantener su relevancia y precisión.

Por otro lado, se ha establecido la evaluación de diversas alternativas para integrar información socioeconómica adicional de los clientes, incluso si esto requiere la utilización de fuentes externas de datos. Este enfoque busca enriquecer el modelo y mejorar la calidad de los análisis realizados.

7.2. Supervisión y Mantenimiento

Una vez que el producto se encuentre en uso por parte de los usuarios finales, se propone realizar alguna de las siguientes acciones:

- Monitoreo y Mantenimiento, después de la implementación, el foco debe estar en el monitoreo continuo del modelo en producción. Esto incluye la recolección de métricas, el seguimiento de cambios en los datos y la evaluación constante del rendimiento del modelo. Además, es fundamental mantener una comunicación con las partes interesadas del negocio para asegurar que los resultados estén alineados con los objetivos estratégicos y realizar ajustes si es necesario. También se debe automatizar el proceso y mejorar la escalabilidad de la solución.
- Se llevará a cabo un seguimiento exhaustivo de la duración y frecuencia de las interacciones de los usuarios con la herramienta. Esto permitirá identificar patrones de uso, como qué funcionalidades son más accesadas, qué áreas requieren más tiempo de uso y cuáles son las rutas más comunes que siguen los usuarios. Estos datos son cruciales para mejorar la interfaz y optimizar la experiencia del usuario.
- A través del monitoreo de los accesos y las interacciones de los usuarios, se identificarán posibles problemas de usabilidad, como tiempos de carga excesivos, pasos innecesarios en los flujos de trabajo o secciones de la herramienta que causan confusión. Con esta información, se podrán realizar ajustes para hacer la herramienta más intuitiva y eficiente.
- Además del monitoreo técnico, se incluirá una evaluación cualitativa mediante encuestas, comentarios o valoraciones de los usuarios para medir su nivel de satisfacción con la herramienta. Este feedback será clave para identificar áreas de mejora desde la perspectiva del usuario final y asegurar que la herramienta cumpla con sus expectativas y necesidades.

7.3. Informe Final

Se presentan los resultados de la aplicación del modelo desarrollado utilizando la técnica seleccionada en las fases previas, evaluando su efectividad. La implementación se ha llevado a cabo sobre un conjunto de datos no etiquetados en términos de morosidad, lo que permite medir el desempeño del modelo en la identificación y predicción de este fenómeno sin la necesidad de etiquetas previas.

	Cantidad	Porcentaje
Créditos que podrían presentar mora	22	19.64
Créditos que podrían no presentar mora	90	80.36
Total	112 clientes	100 %

7.4. Revisión del proyecto

Una vez finalizada la presente iteración de la metodología CRIS.DM para el proyecto en curso, se reconocen como mejoras aplicables:

- Una mejora esencial es la integración de herramientas que faciliten la colaboración en tiempo real entre los expertos del dominio y el equipo de trabajo. Plataformas como Jupyter Notebooks compartidos o Google Colab permiten que los analistas de datos y expertos trabajen conjuntamente, compartiendo código, resultados y visualizaciones sin la necesidad de esperar entre iteraciones. Esto también fomenta la retroalimentación inmediata, acelerando los ciclos de prueba y ajuste del modelo.
- La implementación de pipelines automáticos que gestionen el proceso de transformación, validación y modelado de datos es una mejora común que puede reducir el tiempo de espera y errores humanos. Utilizar herramientas como Apache Airflow o Kubeflow puede garantizar que los procesos de carga de datos, limpieza y modelado se realicen de manera eficiente y sin interrupciones, manteniendo al equipo enfocado en tareas de alto valor.
- El monitoreo proactivo del rendimiento del modelo y la calidad de los datos es crucial para evitar problemas inesperados. Incorporar sistemas de monitoreo como Grafana o Prometheus puede permitir detectar caídas de rendimiento, cambios en los datos o anomalías en tiempo real. Esto es especialmente útil en producción para que los expertos en el dominio puedan actuar rápidamente antes de que se conviertan en problemas significativos.
- Es fundamental fomentar la capacitación continua del equipo, en especial en metodologías avanzadas de análisis de datos y aprendizaje automático. Esto garantiza que el equipo no solo esté al tanto de las últimas tendencias y mejores prácticas, sino que también sea capaz de adaptarse rápidamente a los nuevos desafíos que surjan durante el proyecto.

- Además de las herramientas colaborativas, es importante establecer un flujo de retroalimentación constante con los stakeholders para asegurar que los modelos desarrollados estén alineados con los objetivos del negocio. Reuniones periódicas con el equipo de dominio pueden ayudar a priorizar las características y mejoras del modelo, así como ajustar las expectativas a medida que surgen nuevos datos o requisitos.

Con la ejecución de estas actividades, se da por concluido el segundo sprint del proyecto. A continuación, se adjuntan capturas de pantalla que documentan el registro de las tareas completadas en la herramienta de seguimiento (GitHub Projects):

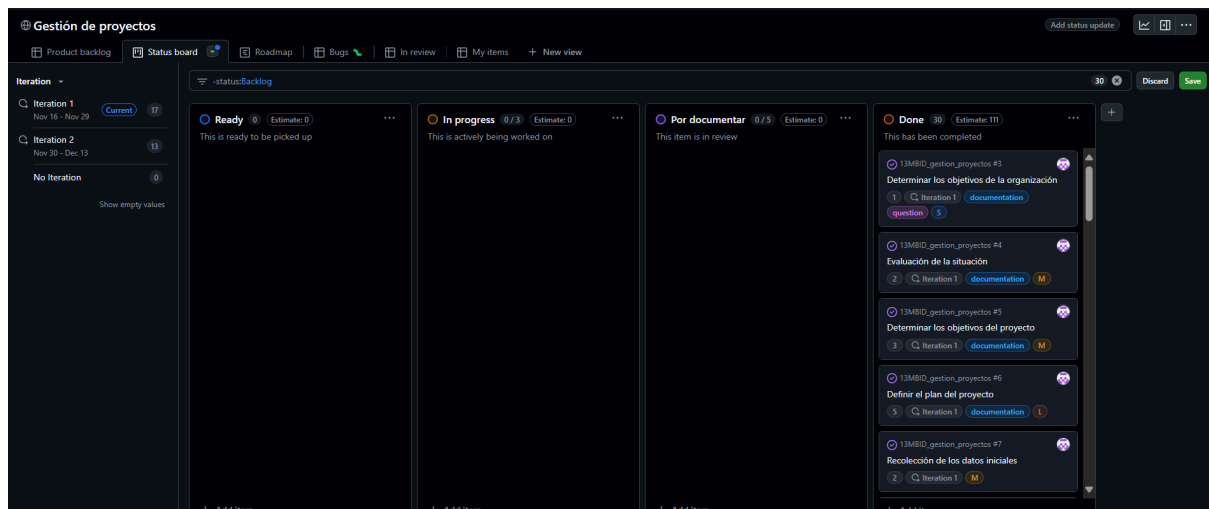


Figura 11. Finalización de la segunda iteración del proyecto

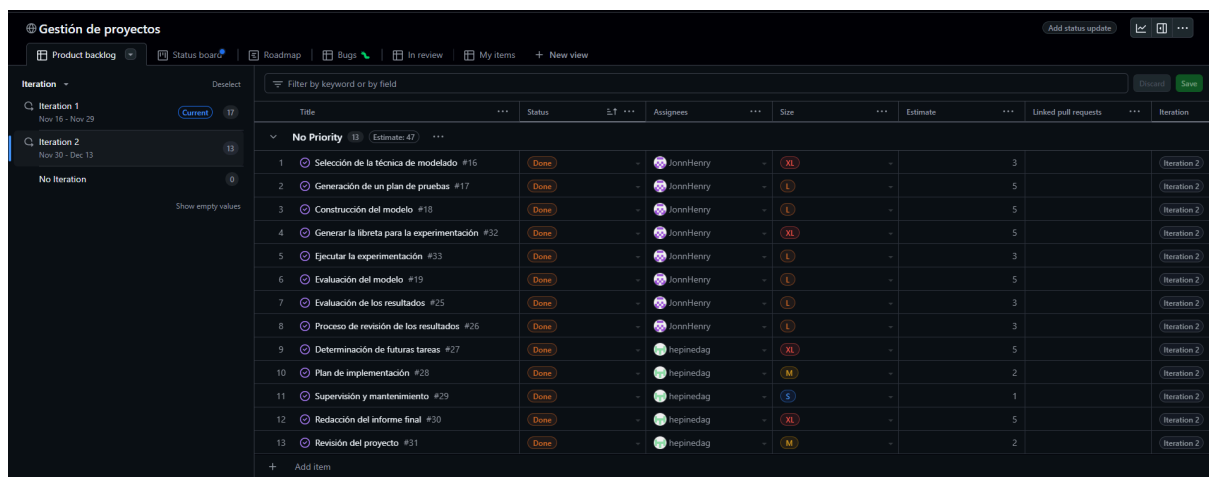


Figura 12. Finalización de la segunda iteración del proyecto vista del product backlog

Se adjunta el enlace del repositorio público para que pueda ser accedido: https://github.com/JonnHenry/13MBID_gestion_proyectos.git