

Miembros del grupo:

Antonio Manuel Palma Bautista

Gemma Arlandis Esteve



ACTIVIDAD 2

Máster en Big Data y Data Science

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

Fecha: 8 de diciembre de 2023

Curso 2023 - Ed. Abril

Contenido

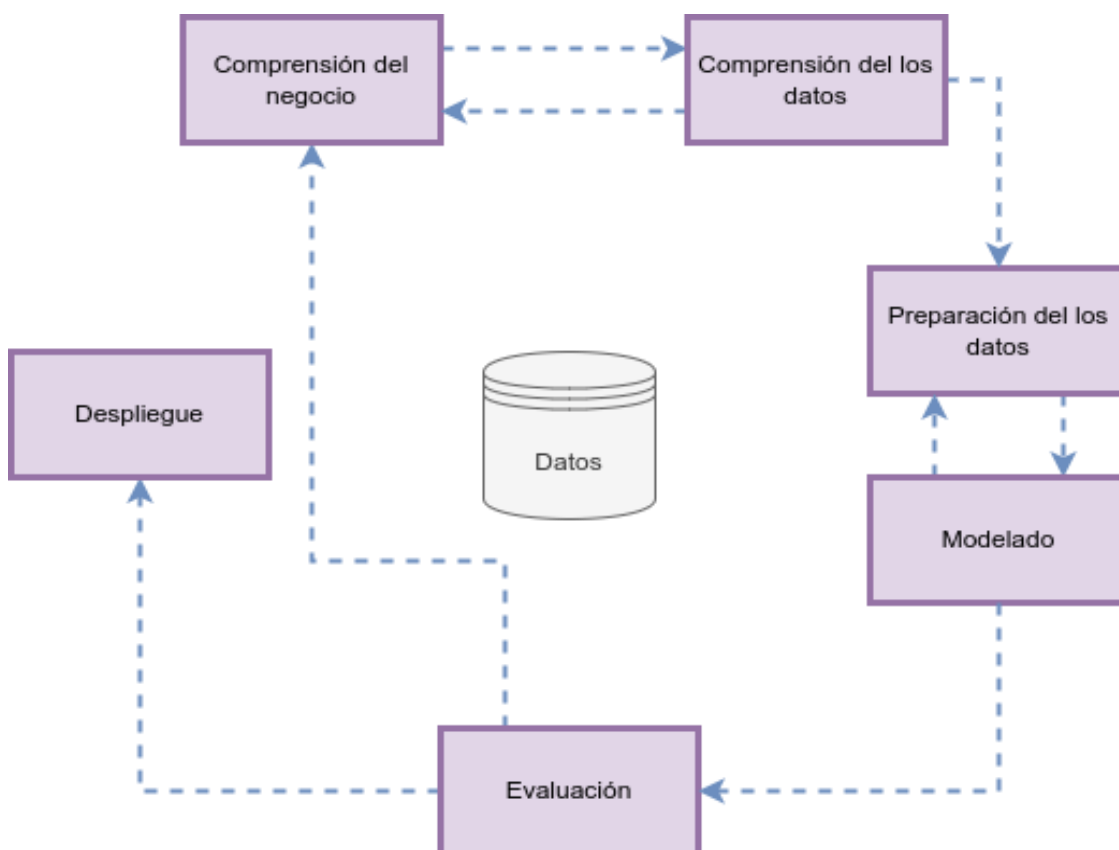
1. Introducción	4
2. Comprensión del negocio.....	5
Determinar los objetivos de la Organización	5
Evaluación de la situación.....	5
Determinación de los objetivos del proyecto	5
Definir plan del proyecto (tareas, recursos, etc).....	6
3. Comprensión de los datos	8
Recolección de datos iniciales	8
Descripción de los datos	8
Exploración de datos	9
Verificación de la calidad de los datos.....	18
4. Fase de preparación de los datos	19
Selección de datos.....	19
Limpieza de los datos.....	19
Integración de los datos.....	20
Construcción / Transformación de datos.....	20
Formateo de los datos	22
5. Modelado	30
Selección de la técnica de modelado.....	30
Generación del plan de pruebas	30
Construcción del Modelo.....	30
Evaluación del modelo.....	31
• Prueba #1	31
• Prueba #2	33
• Prueba #3	34
6. Evaluación.....	36
Evaluación de los resultados.....	36
Proceso de revisión	36
Determinación de futuras tareas.....	36
7. Despliegue/Implementación.....	37
Plan de implementación.....	37
Supervisión y Mantenimiento	37

Informe Final.....	37
Revisión del proyecto.....	37
Anexo 1. Modelo de memoria de trabajo para evaluación de calidad de datos	39
Definición de objetivos y características de la evaluación inicial.....	39
Descripción del uso propuesto	39
Definición de calidad.....	39
Características que deben cumplir los datos	40
Registro de metadatos de cada dataset.....	40
Evaluación inicial de los datos disponibles.....	40
Resultados de los análisis	41
Dimensión: Completitud.....	41
Dimensión: Exactitud.....	41
Dimensión: Consistencia.....	44
Identificación de mejoras aplicables.....	45

1. Introducción

El objetivo principal de la actividad es aplicar métodos de gestión ágiles al desarrollo del proyecto de ciencia de datos en cuestión.

Se utiliza la metodología CRISP-DM, que cuenta con 6 fases, véase figura 1. Estas fases 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.



2. Comprensión del negocio

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. El objetivo principal será predecir si un crédito determinado podría pasar a ser considerado en mora (default). 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.

Evaluación de la situación

Se cuenta con los siguientes recursos para la ejecución del proyecto:

- Los datos históricos de créditos otorgados por la entidad y los datos de otros productos que tienen los mismos clientes contratados con la entidad.
- Se cuenta con el personal para la realización de las tareas en cuestión. Incluso se cuenta con la colaboración inicial de expertos en el dominio para aclaraciones sobre los datos.
- Se cuenta con las herramientas y hardware necesarios para la ejecución de las actividades del proyecto.

Se deja a disposición el acceso al repositorio utilizado para el desarrollo de las actividades aquí descritas: <https://github.com/ampalmabautista/13MBID.git> .

Determinación de los objetivos del proyecto

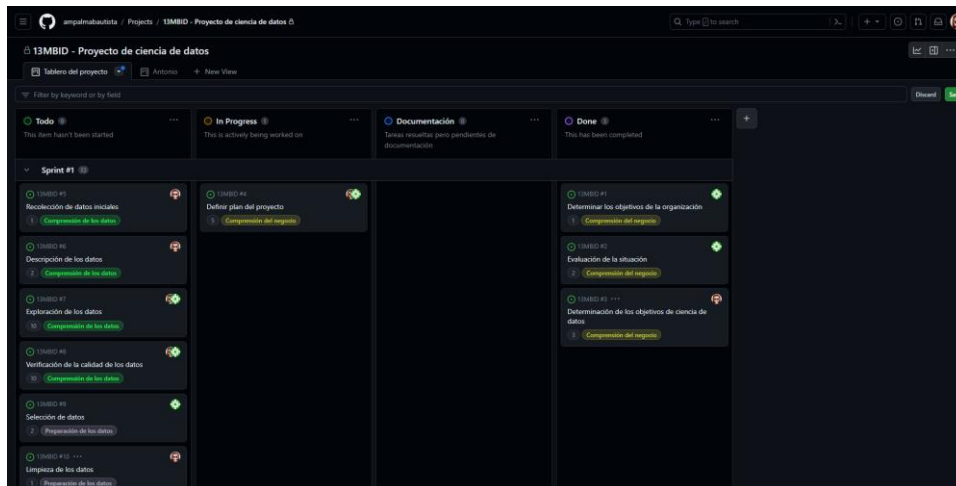
Desarrollar un producto de datos que permita **predecir** sobre un conjunto de nuevos créditos otorgados por la entidad la posibilidad de que cada uno de ellos pueda entrar en mora en un futuro.

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.

Definir plan del proyecto (tareas, recursos, etc)

El proyecto, implementado bajo la herramienta Projects vinculada al repositorio de GitHub antes mencionado se puede encontrar en el siguiente enlace:

<https://github.com/users/ampalmabautista/projects/1/views/2>

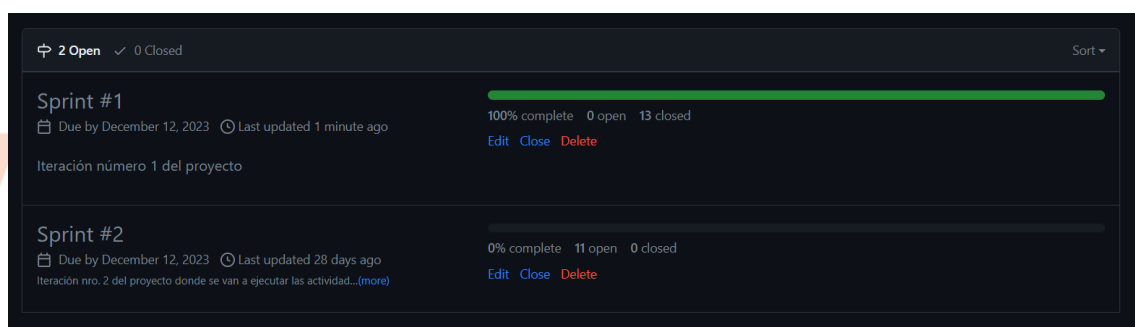


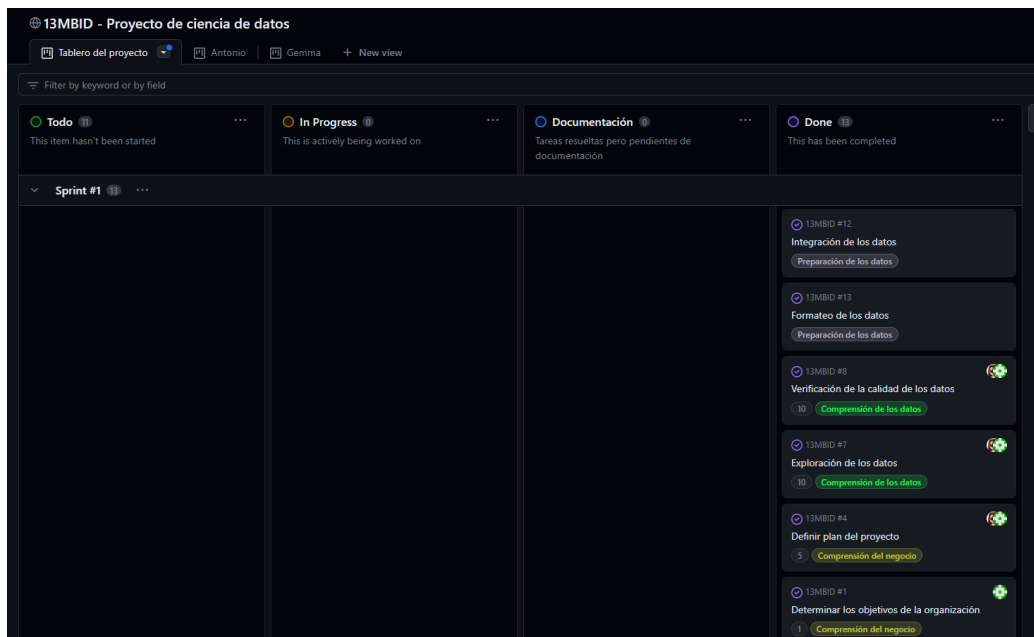
Los roles del proyecto se distribuyen entre los dos miembros del grupo de trabajo de la siguiente forma:

- Product owner: Antonio Palma (será el responsable de representar las necesidades y prioridades del cliente, definiendo y manteniendo el Product Backlog).
- Scrum Master: Gemma Arlandis (será la facilitadora del proceso Scrum, asegurando que sea aplicado de forma correcta y eliminando los obstáculos que impidan que el equipo logre sus objetivos).

Equipo de trabajo: Antonio Palma realizará las funciones de ingeniero de datos y analista de datos y Gemma Arlandis las de científica de datos y analista de negocio, conformando ambos el equipo multidisciplinario y auto-organizado que será responsable de entregar el producto.

Una vez finalizadas las tareas de las fases: *Comprensión del Negocio*, *Comprensión de los Datos* y *Preparación de los Datos* se ha cerrado el Sprint #1:



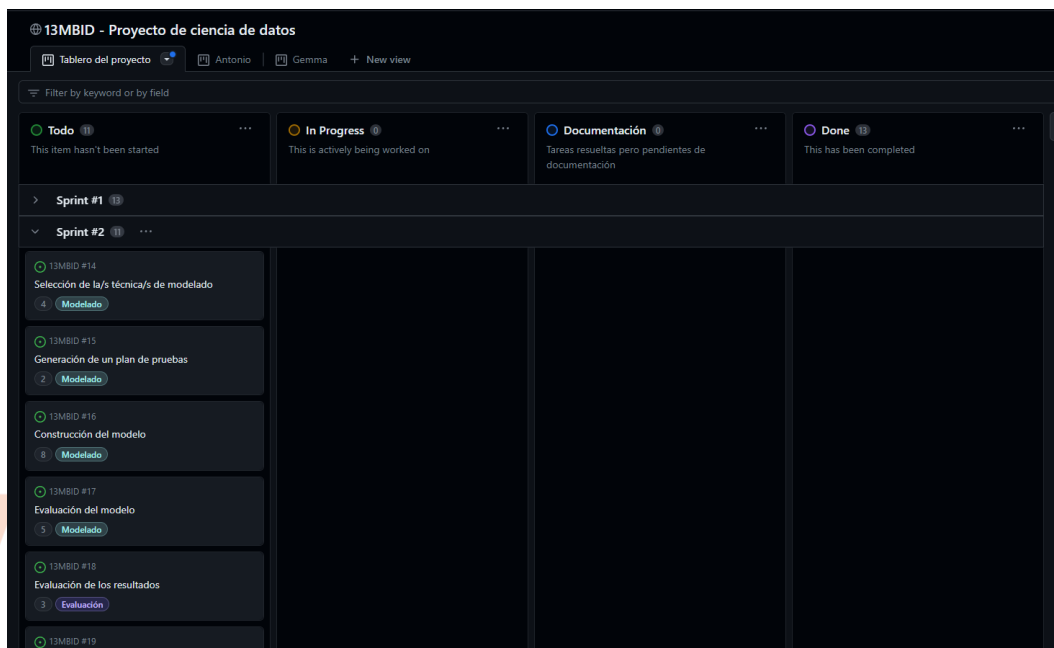


Inicio del Sprint #2

En esta 2da iteración se van a ejecutar las operaciones de las siguientes fases de la metodología CRISP-DM:

- Modelado
- Evaluación
- Despliegue

El sprint backlog de la iteración ha quedado conformado de la siguiente manera:



3. Comprensión de los datos

Recolección de datos iniciales

Para el proyecto, se cuenta con los datos necesarios agrupados en dos archivos:

- **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.

Los datos han sido verificados con respecto a su origen y se han agregado al esquema de versionado a utilizar.

Descripción de los datos

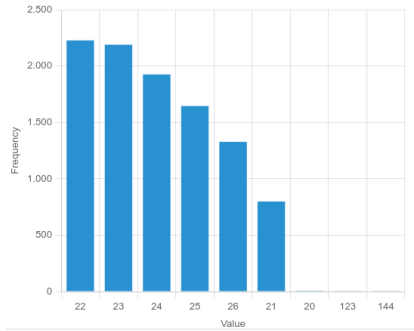
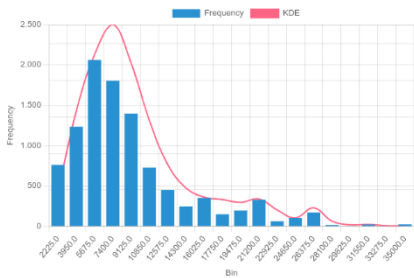
Se describen las propiedades principales de cada dataset:

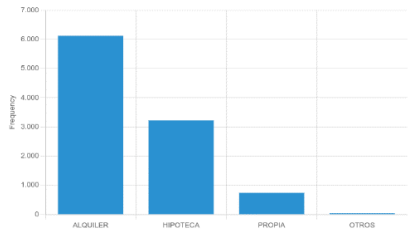
Dataset	Columnas / Atributos	Cantidad de filas
Datos_creditos	id_cliente edad importe_solicitado duracion_credito antiguedad_empleado situacion_vivienda ingresos objetivo_credito pct_ingreso tasa_interes estado_credito falta_pago	10127
Datos_tarjetas	id_cliente antiguedad_cliente estado_civil estado_cliente gastos_ult_12m genero limite_credito_tc nivel_educativo nivel_tarjeta operaciones_ult_12m personas_a_cargo	10127

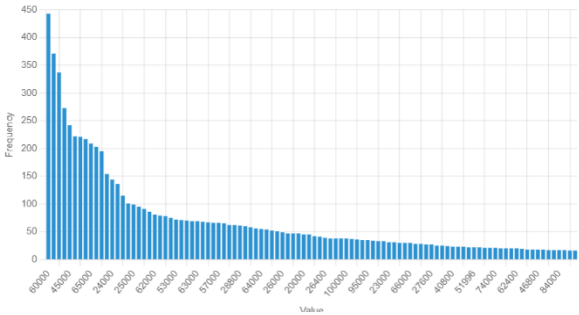
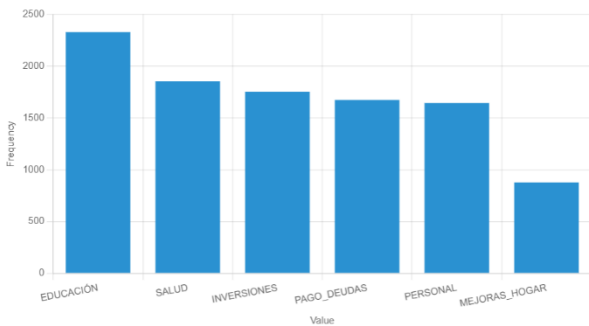
Exploración de datos

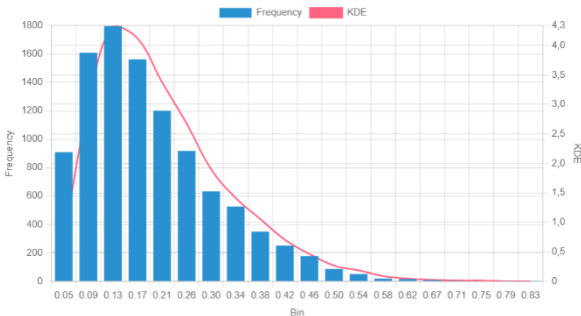
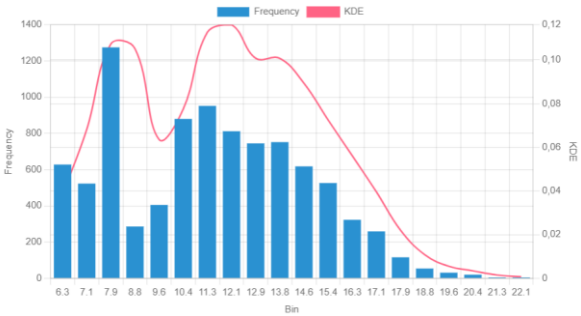
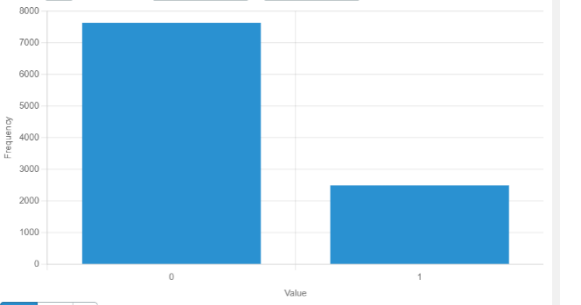
Se describen los metadatos de cada dataset:

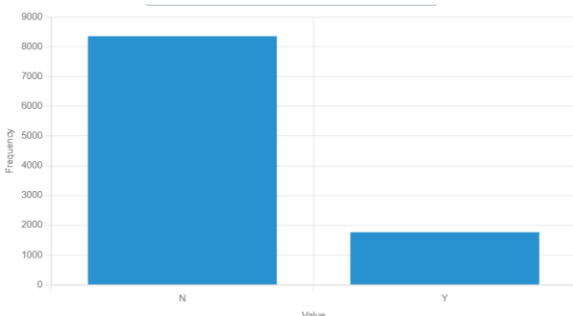
- Datos_creditos.csv

Columna	Tipo de datos	Observaciones
id_cliente	Float64 (numérico)	Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0
edad	Int64 (numérico)	Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0 
importe_solicitado	Int64 (numérico)	Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0 <i>Estadísticas de los valores:</i> max:35,000 mean:8,138.7331 median:6,500 min:500 mode:5,000 
duracion_credito	Int64 (numérico)	Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0

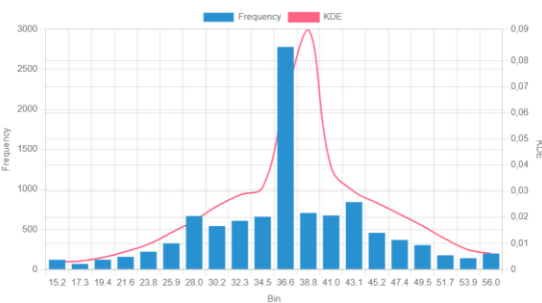
		<p><i>Valores únicos:</i> 2 (3404), 3(3364), 4(3359)</p> <table> <tr> <th>duracion_credito</th><th>Frequency</th><th>Percent</th></tr> <tr> <td>2</td><td>3,404</td><td>33.61%</td></tr> <tr> <td>3</td><td>3,364</td><td>33.22%</td></tr> <tr> <td>4</td><td>3,359</td><td>33.17%</td></tr> <tr> <td>TOTAL</td><td>10,127</td><td>100.00%</td></tr> </table>	duracion_credito	Frequency	Percent	2	3,404	33.61%	3	3,364	33.22%	4	3,359	33.17%	TOTAL	10,127	100.00%
duracion_credito	Frequency	Percent															
2	3,404	33.61%															
3	3,364	33.22%															
4	3,359	33.17%															
TOTAL	10,127	100.00%															
antiguedad_emplead o	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):9,790 Count (missing):337 % Missing:3.33</p> <p><i>Estadísticas de los valores:</i> max:123 mean:3.9385 median:4 min:-122</p>															
situacion_vivienda	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): ALQUILER - 6,125 - 60.48% HIPOTECA - 3,223 - 31.83% PROPIA - 741 - 7.32% OTROS - 38 - 0.38%</p> 															
ingresos	Int64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores:</i> max:500,000 mean:50,381.8976 median:46,000 min:9,600 mode: 60,000</p>															

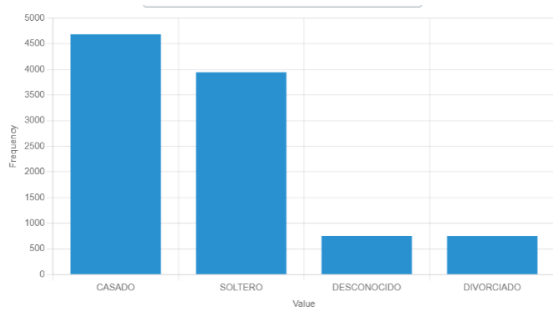
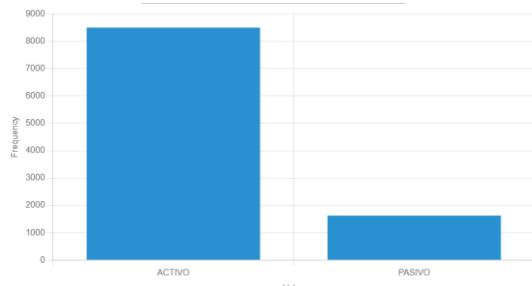
		
objetivo_credito	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): EDUCACIÓN - 2328 – 22.99% SALUD - 1753 – 18.30% INVERSIONES – 1753 – 17.31% PAGO_DEUDAS – 1673 – 16.52% PERSONAL – 1643 – 16.22% MEJORAS_HOGAR – 877 – 8.66%</p> 
pct_ingreso	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores:</i> max:0.83 mean:0.1772 median:0.15 min:0.01</p>

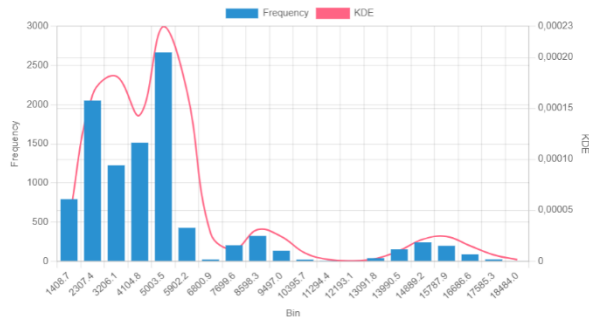
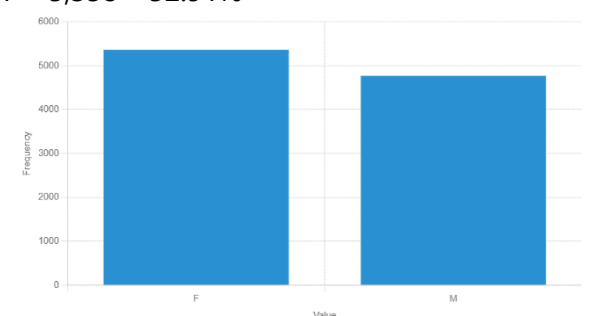
		
tasa_interes	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan): 9,215 Count (missing): 912 % Missing: 9.01</p> <p><i>Estadísticas de los valores:</i> max:22.11 mean:10.9794 median:10.99 min:5.42</p> 
estado_credito	Int64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): 0 – 7,635 – 75.36% 1 – 2,492 – 24.61%</p> 

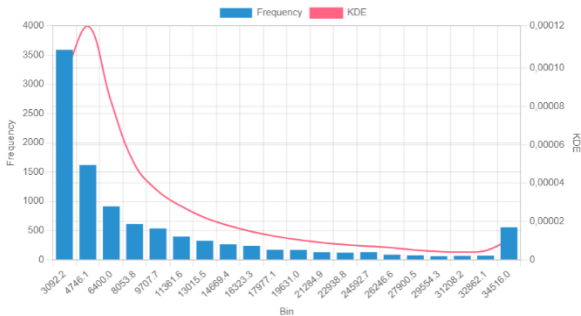
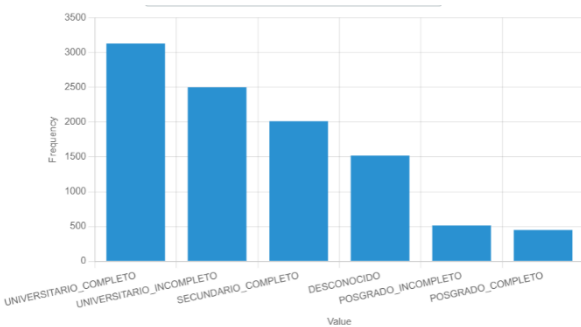
falta_pago	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): Y – 8,359 – 82.54% N – 1,769 – 17.46%</p> 
------------	------------------	--

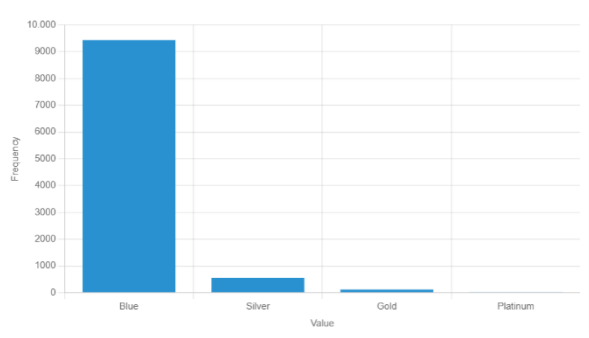

- Datos_tarjetas

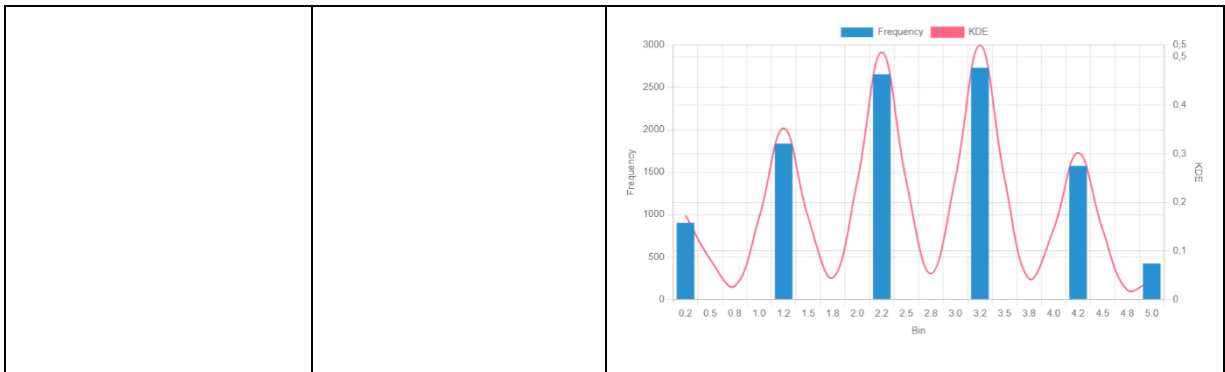
Columna	Tipo de datos	Observaciones
id_cliente	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p>
antigüedad_cliente	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores:</i> max:56 mean:35.9284 median:36 min:13</p> 

estado_civil	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): CASADO – 4,687 – 46.28% SOLTERO – 3,943 – 38.94% DESCONOCIDO – 749 – 7.40% DIVORCIADO – 748 – 7.39%</p> 
estado_cliente	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): ACTIVO – 8,500 – 83.93% PASIVO – 1,627 – 16.07%</p> 
gastos_ult_12m	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores:</i> max:18,484 mean:4,404.0863 median:3,899 min:510</p>

		
Genero	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): M – 4,769 – 47.09% F – 5,358 – 52.91%</p> 
limite_credito_tc	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores:</i> max:34,516 mean:8,631.9537 median:4,549 min:1,438.3</p>

		
nivel_educativo	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): UNIVERSITARIO_COMPLETO – 3,128 – 30.89% UNIVERSITARIO_INCOMPLETO – 2,500 – 24.69% SECUNDARIO_COMPLETO – 2,013 – 19.88% DESCONOCIDO – 1,519 – 15.00% POSGRADO_INCOMPLETO – 516 – 5.10% DESCONOCIDO – 451 – 4.45%</p> 
nivel_tarjeta	String (nominal)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): BLUE – 9,436 – 93.18% SILVER – 555 – 5.48% GOLD – 116 – 1.15% PLATINUM – 20– 0.20%</p>

		
operaciones_ult_12m	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores:</i> max:139 mean: 64.8587 median:67 min:10</p> 
personas_a_cargo	Float64 (numérico)	<p>Total Rows:10,127 Count (non-nan):10,127 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores:</i> max:5 mean: 2,3462 median:2 min:0</p>



Verificación de la calidad de los datos

El desarrollo de las tareas de esta actividad se encuentra detallado en el [Anexo 1](#).

4. Fase de preparación de los datos

Selección de datos

En función de los resultados del análisis de calidad de los datos ejecutado en la fase anterior, se realizaron los siguientes filtros a nivel de columnas en los datasets:

- Dataset: datos_creditos – Columnas eliminadas:
 - 'completitud_fila', 'situacion_vivienda_ok', 'objetivo_credito_ok', 'estado_credito_ok' y 'falta_pago_ok'. En ambos casos se trata de columnas agregadas como elementos auxiliares del proceso de verificación de calidad de datos.
 - 'tasa_interes'. En este caso se ha eliminado este atributo porque presentaba una proporción de nulos mayor al límite establecido para el dataset.
- Dataset: datos_tarjetas – Columnas eliminadas:
 - 'nivel_tarjeta'. Se elimina el atributo porque tiene una alta correlación con el valor del límite de crédito disponible en el producto para el cliente. Se considera mejor procesar el valor del límite de crédito que el valor categórico del nivel de tarjeta.

Limpieza de los datos

En esta actividad se aplican filtros a nivel de filas en los datasets del escenario para subsanar o corregir valores fuera de rango en los siguientes atributos:

Dataset	Atributo	Filtro aplicado	Observaciones
datos_creditos	'edad'	El valor del atributo no puede ser mayor a 90.	Cantidad de filas filtradas por esta operación: 4
	'antigüedad_empleado'	El valor del atributo no puede ser mayor a 50 ni nulo.	Cantidad de filas filtradas por esta operación: 339
	'regla_pct_ingresos'	El valor del atributo no debe ser 'err' ya que implica un no cumplimiento de la regla de negocio definida.	Cantidad de filas filtradas por esta operación: 14
	'regla_duracion_credito'	El valor del atributo no debe ser 'err' ya que implica un no cumplimiento de la regla de negocio definida.	Cantidad de filas filtradas por esta operación: 5

Para los datos_tarjeta se ha considerado no realizar el filtro de datos ya que no hay datos nulos ni fuera de rango.

Integración de los datos

A partir de los datos originales relativos al problema a resolver:

- Datos de créditos (dataset: datos_creditos)
- Datos de productos financieros de los clientes (dataset: datos_tarjetas)

Se ha realizado una operación de unión en función de los valores del campo 'id_cliente', con los siguientes resultados:

- Cantidad de columnas del dataset integrado: 29.
- Cantidad de filas del dataset integrado: 9765.

Construcción / Transformación de datos

Se documentan a continuación las transformaciones aplicadas en el dataset integrado:

Atributo	Transformación aplicada
'estado_civil'	Cambios realizados para mejor lectura de los datos: <ul style="list-style-type: none"> • 'CASADO' : 'C', • 'SOLTERO' : 'S', • 'DESCONOCIDO' : 'N', • 'DIVORCIADO' : 'D',
'estado_credito'	Cambios realizados para modificar el tipo de datos de la columna: <ul style="list-style-type: none"> • 0: 'C', • 1 : 'P',
'edad'	Valores numéricos convertidos a nominales aplicando rangos (<i>etiqueta – rango de valores</i>): <ul style="list-style-type: none"> • 'menor_25' : [0, 24], • '25_a_30' : [25, 50*] <p>(*) Se coloca como valor superior del rango para evitar perder datos.</p>
'antiguedad_empleado'	Valores numéricos convertidos a nominales aplicando rangos (<i>etiqueta – rango de valores</i>): <ul style="list-style-type: none"> • 'menor_5' : [0, 4], • '5_a_10' : [5, 10] • 'mayor_10' : [10, 50*]

	<ul style="list-style-type: none"> 'NA' : Valor agregado a modo de reemplazo de los valores nulos del atributo. <p>(*) Se coloca como valor superior del rango para evitar perder datos.</p>
'pct_ingreso'	<p>Valores numéricos convertidos a nominales aplicando rangos (<i>etiqueta – rango de valores</i>):</p> <ul style="list-style-type: none"> 'hasta_20' : [0, 0.19], '20_a_40' : [0.20, 0.39] '40_a_60' : [0.40, 0.59] 'mayor_60' : [0.60, 0.99*] <p>(*) Se coloca como valor superior del rango para evitar perder datos.</p>
'ingresos'	<p>Valores numéricos convertidos a nominales aplicando rangos (<i>etiqueta – rango de valores</i>):</p> <ul style="list-style-type: none"> 'hasta_20k' : [0, 19999], '20k_a_50k' : [20000, 49999] '50k_a_100k' : [50000, 99999] 'mayor_100k' : [100000, 999999*] <p>(*) Se coloca como valor superior del rango para evitar perder datos.</p>
'antigüedad_cliente'	<p>Valores numéricos convertidos a nominales aplicando rangos (<i>etiqueta – rango de valores</i>):</p> <ul style="list-style-type: none"> 'menor_2y' : [0, 23], '2y_a_4y' : [24, 47] 'mayor_4y' : [48, 100*] <p>(*) Se coloca como valor superior del rango para evitar perder datos.</p>
'limite_credito_tc'	<p>Valores numéricos convertidos a nominales aplicando rangos (<i>etiqueta – rango de valores</i>):</p> <ul style="list-style-type: none"> 'menor_3k' : [0, 2999], '3k_a_5k' : [3000, 4999] '5k_a_10k' : [5000, 9999] 'mayor_10k' : [10000, 100000*] <p>(*) Se coloca como valor superior del rango para evitar perder datos.</p>

'gastos_ult_12m'	<p>Valores numéricos convertidos a nominales aplicando rangos (<i>etiqueta – rango de valores</i>):</p> <ul style="list-style-type: none"> • 'menor_1k' : [0, 999], • '1k_a_4k' : [1000, 3999] • '4k_a_6k' : [4000, 5999] • '6k_a_8k' : [6000, 7999] • '8k_a_10k' : [8000, 9999] • 'mayor_10k' : [10000, 100000*] <p>(*) Se coloca como valor superior del rango para evitar perder datos.</p>
'operaciones_ult_12m'	<p>Valores numéricos convertidos a nominales aplicando rangos (<i>etiqueta – rango de valores</i>):</p> <ul style="list-style-type: none"> • menor_15' : [0, 14], • '15_a_30' : [15, 29] • '30_a_50' : [30, 49] • '50_a_75' : [50, 74] • '75_a_100' : [75, 99] • 'mayor_100' : [100, 1000*] <p>(*) Se coloca como valor superior del rango para evitar perder datos.</p>

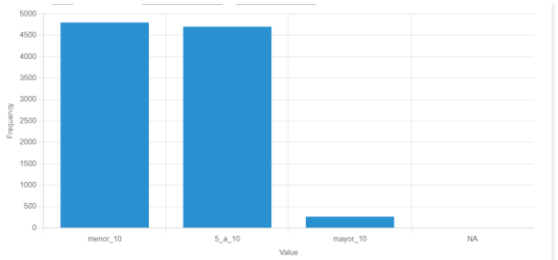
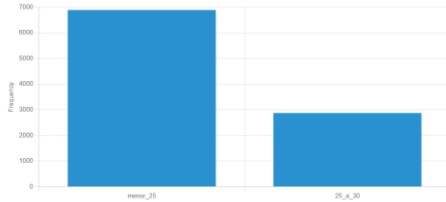
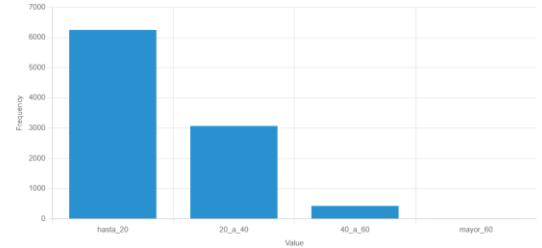
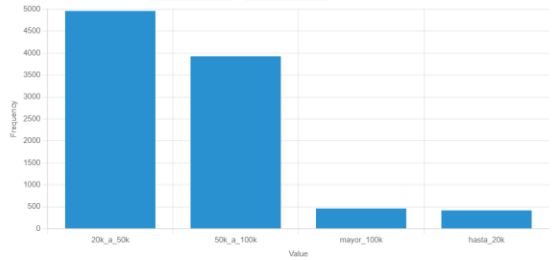
Formateo de los datos

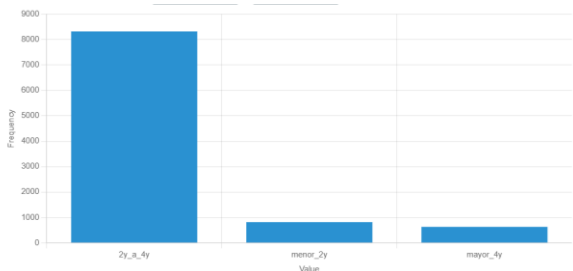
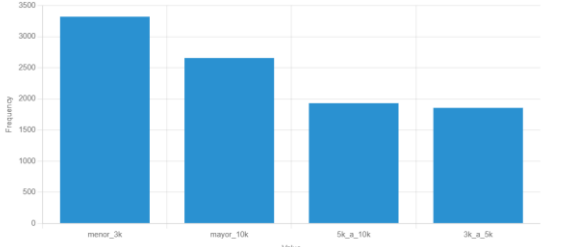
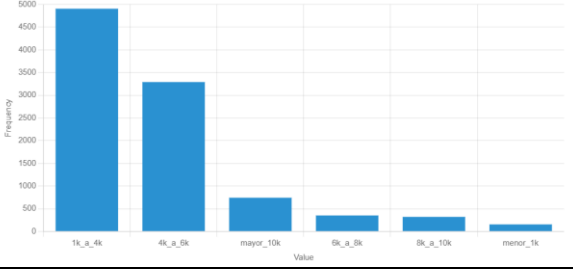
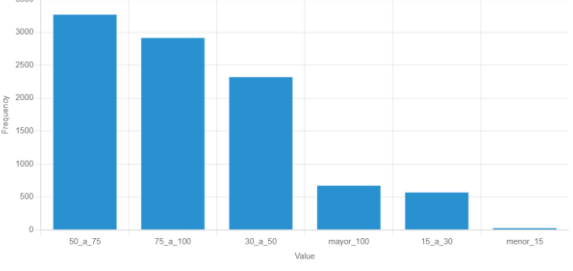
Se han especificado las operaciones en el apartado anterior. Como resultado se ha obtenido un nuevo conjunto de datos, denominado *datos_finales.csv* que tiene las siguientes dimensiones y características:

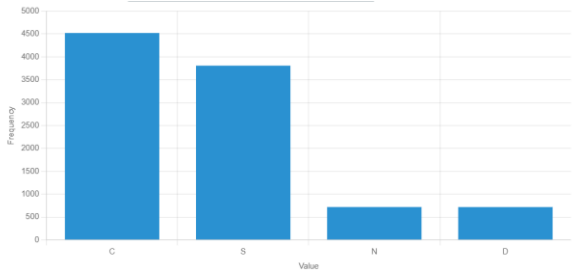
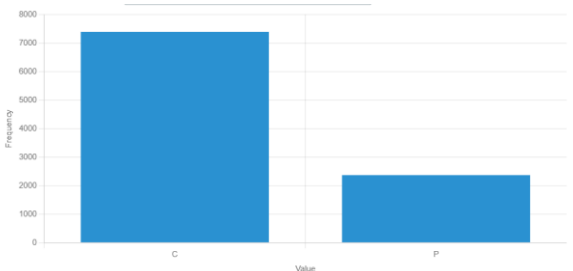
- Cantidad de columnas: 26
 - Se ha eliminado la columna '**id_cliente**' dado que la misma ya ha sido utilizada en la integración de ambos datasets originales.
- Cantidad de filas: 9765

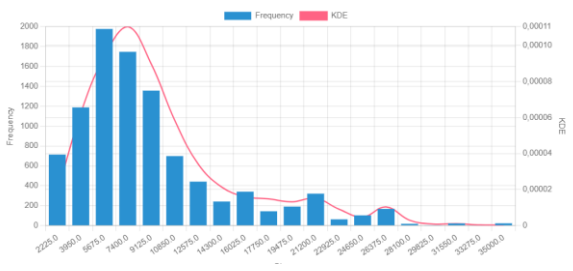
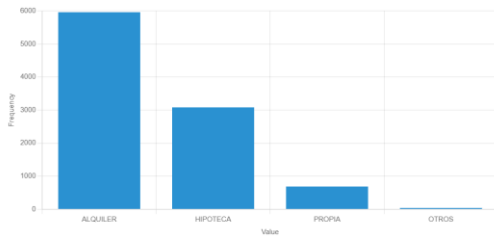
Se listan a continuación los metadatos del conjunto:

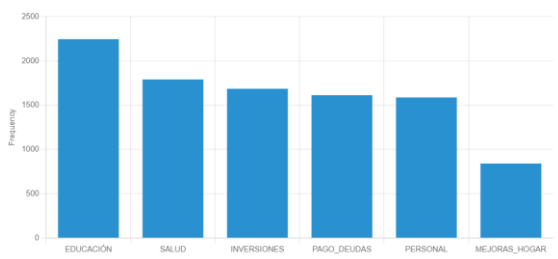
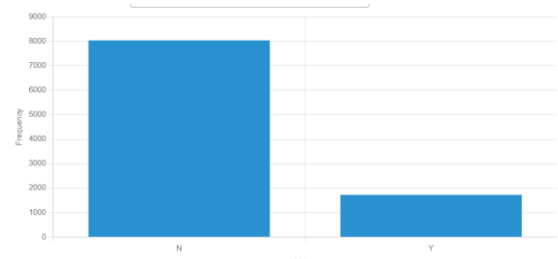
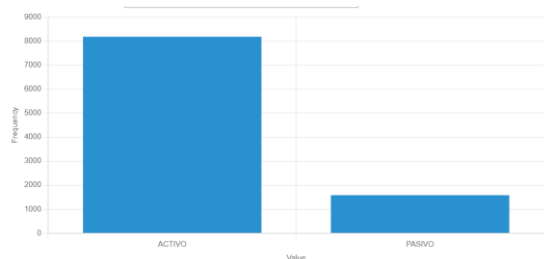
Columna	Tipo de datos	Observaciones
antigüedad_empleado	category	<p>Total Rows: 9765 Count (non-nan): 9,765 Count (missing): 0 % Missing: 0</p>

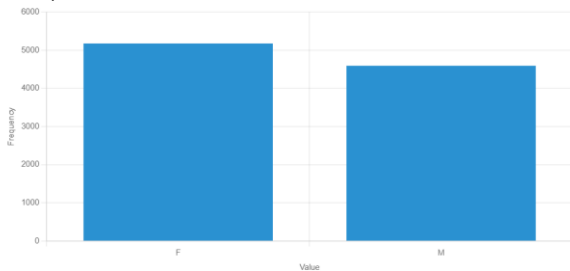
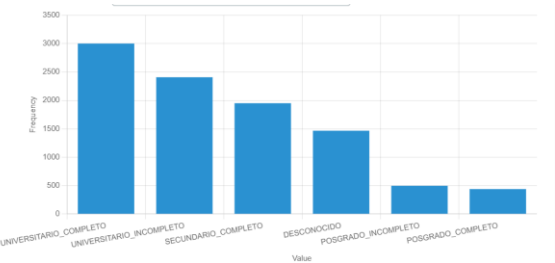
		
edad	category	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> 
pct_ingreso	category	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> 
ingresos	category	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> 
antiguedad_cliente	category	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p>

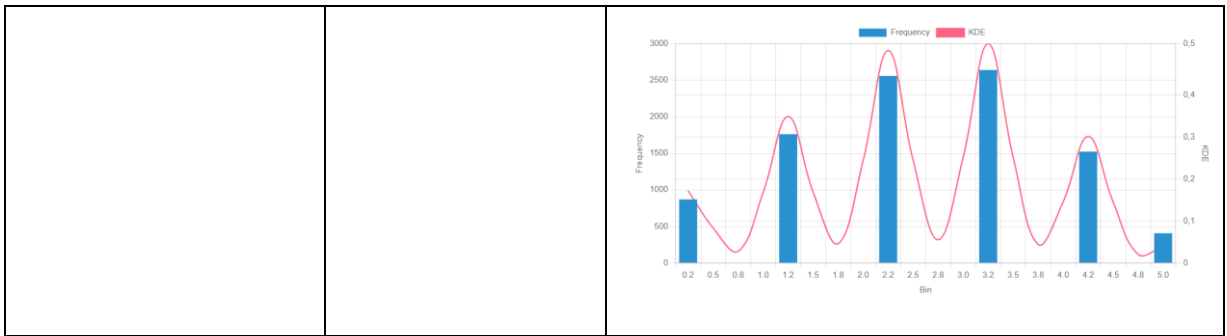
		
limite_credito_tc	category	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> 
gastos_ult_12m	category	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> 
operaciones_ult_12m	category	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> 

estado_civil	String (nominal)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): C – 4,522 – 46.31% S – 3,808 – 39.00% N – 718 – 7.35% D – 717 – 7.34%</p> 
estado_credito	String (nominal)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): C – 7,395 – 75.73% P – 2,370 – 24.27%</p> 
importe_solicitado	Int64 (numérico)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores:</i> max:35,000 mean:8,171.7947 median:6,500 min:500 mode:5,000</p>

																	
duracion_credito	Int64 (numérico)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0 <i>Valores únicos:</i> 2 (3278), 3(3247), 4(3240)</p> <table border="1"> <thead> <tr> <th>duracion_credito</th><th>Frequency</th><th>Percent</th></tr> </thead> <tbody> <tr> <td>2</td><td>3,278</td><td>33.57%</td></tr> <tr> <td>3</td><td>3,247</td><td>33.25%</td></tr> <tr> <td>4</td><td>3,240</td><td>33.18%</td></tr> <tr> <td>TOTAL</td><td>9,765</td><td>100.00%</td></tr> </tbody> </table>	duracion_credito	Frequency	Percent	2	3,278	33.57%	3	3,247	33.25%	4	3,240	33.18%	TOTAL	9,765	100.00%
duracion_credito	Frequency	Percent															
2	3,278	33.57%															
3	3,247	33.25%															
4	3,240	33.18%															
TOTAL	9,765	100.00%															
situacion_vivienda	String (nominal)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad – %</i>): ALQUILER – 5,960 – 61.03% HIPOTECA – 3,081 – 31.55% PROPIA – 686 – 7.03% OTROS – 38 – 0.39%</p> 															
objetivo_credito	String (nominal)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad – %</i>): EDUCACIÓN - 2246 – 23.00% SALUD - 1791 – 18.34% INVERSIONES – 1686 – 17.27% PAGO_DEUDAS – 1613 – 16.52% PERSONAL – 1588 – 16.26%</p>															

		<p>MEJORAS_HOGAR – 841 – 8.61%</p> 
falta_pago	String (nominal)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): Y – 8,039 – 82.32% N – 1726 – 17.68%</p> 
estado_cliente	String (nominal)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): ACTIVO – 8,182 – 83,79% PASIVO – 1,583 – 16,21%</p> 
genero	String (nominal)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p>

		<p>Distribución de valores (<i>valor – cantidad - %</i>): M – 4,591 – 47.01% F – 5,174 – 52.99%</p> 
nivel_educativo	String (nominal)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> <p>Distribución de valores (<i>valor – cantidad - %</i>): UNIVERSITARIO_COMPLETO – 3,001 – 30.73% UNIVERSITARIO_INCOMPLETO – 2,408 – 24.66% SECUNDARIO_COMPLETO – 1,951 – 19.98% DESCONOCIDO – 1,4679 – 15.02% POSGRADO_INCOMPLETO – 498 – 5.10% DESCONOCIDO – 440 – 4.51%</p> 
personas_a_cargo	Float64 (numérico)	<p>Total Rows: 9765 Count (non-nan): 9765 Count (missing):0 % Missing:0</p> <p><i>Estadísticas de los valores:</i> max:5 mean: 2,3489 median:2 min:0</p>



5. Modelado

Selección de la técnica de modelado

Con base en los objetivos del proyecto, se podrán utilizar diferentes técnicas para obtener el modelo que permita realizar la predicción de la situación de nuevos créditos. Por ejemplo, se listan algunas técnicas:

- Regresión Logística
- Métodos KNN
- Árboles de decisión
- Métodos de ensamblado de modelos (RandomForest)
- Métodos de refuerzo de gradiente: GradientBoosting

Generación del plan de pruebas

En primer lugar, se va a realizar una distribución de filas del *dataset* final integrado que ha resultado de la ejecución de las tareas de la fase de Preparación de los Datos, esta operación se realiza a fin de dar cumplimiento a las buenas prácticas planteadas en las industria y bibliografía del área:

- Datos para entrenamiento de las técnicas: 75%
- Datos para prueba de los resultados obtenidos: 25%

En segunda instancia, los lineamientos para la ejecución de las pruebas serán:

- Para cada modelo/técnica se van a documentar sus parámetros de ejecución y la efectividad obtenida en el proceso de entrenamiento al utilizar los datos de prueba.
- Se van a ejecutar tres (3) iteraciones de prueba para seleccionar progresivamente las técnicas con mejores resultados (efectividad) y seleccionar así la que será utilizada para la predicción de los datos nuevos del escenario.
- Se mostrarán los resultados obtenidos en cada iteración para cada técnica, en concreto el rendimiento obtenido y la matriz de confusión, para facilitar la selección de la mejor técnica para la predicción.

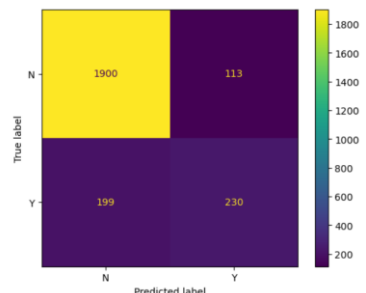
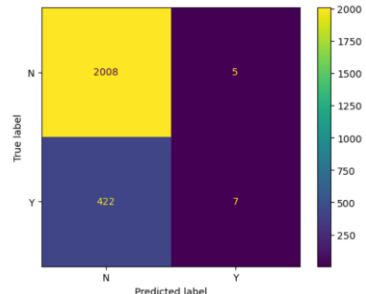
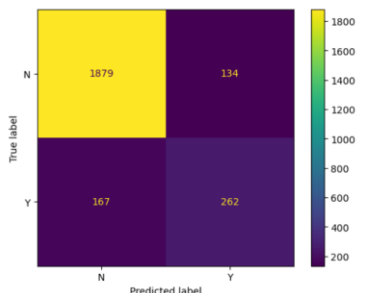
Construcción del Modelo

En esta actividad se van a utilizar diferentes librerías de Python orientadas a la generación de modelos de predicción automática. El código de tales acciones puede encontrar en el repositorio de GitHub enlazado en el presente documento.

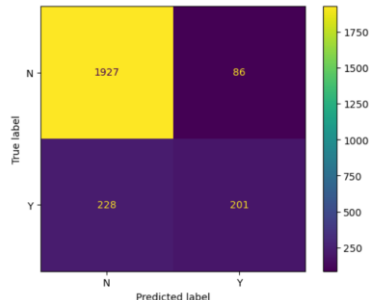
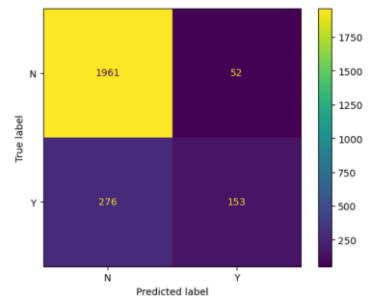
Para la generación de los modelos documentados se utilizó la librería *sci-kit learn* en su versión 1.3.1.

Evaluación del modelo

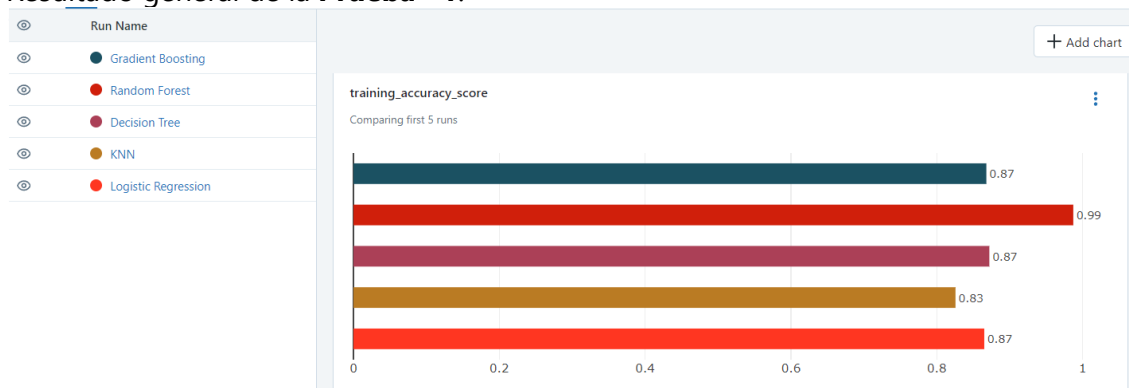
- Prueba #1

Técnica utilizada	Parámetros ¹	Resultados obtenidos
Logistic Regression	C: 1.0 class_weight: None dual: False fit_intercept: True intercept_scaling: 1 l1_ratio: None max_iter: 100 multi_class: auto n_jobs: None penalty: l2 random_state: None solver: liblinear tol: 0.0001 verbose: 0 warm_start: False	Rendimiento obtenido: 0.8722358722358723 Matriz de confusión: 
KNN	algorithm: ball_tree leaf_size: 25 metric: minkowski metric_params: None n_jobs: None n_neighbors: 50 p: 2 weights: uniform	Rendimiento obtenido: 0.8251433251433251 Matriz de confusión: 
Arboles de decisión (TDIDT)	ccp_alpha: 0.0 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 random_state: 0 splitter: best	Rendimiento obtenido: 0.8767403767403767 Matriz de confusión: 

¹ Se **destacan** los parámetros que se han modificado con respecto a su valor por defecto.

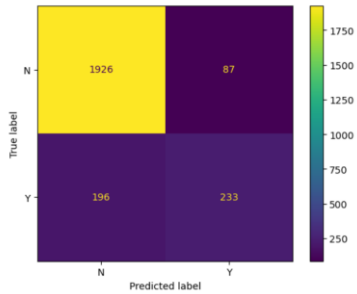
RandomForest	<p> Bootstrap: True ccp_alpha: 0.0 class_weight: None criterion: gini max_depth: None max_features: sqrt max_leaf_nodes: None max_samples: None min_impurity_decrease: 0.0 min_samples_leaf: 1 min_samples_split: 2 min_weight_fraction_leaf: 0.0 n_estimators: 10 n_jobs: None oob_score: False random_state: 0 verbose: 0 warm_start: False </p>	<p>Rendimiento obtenido: 0.8714168714168714</p> <p>Matriz de confusión:</p>  <table border="1"> <thead> <tr> <th></th> <th>Predicted N</th> <th>Predicted Y</th> </tr> </thead> <tbody> <tr> <th>True N</th> <td>1927</td> <td>86</td> </tr> <tr> <th>True Y</th> <td>228</td> <td>201</td> </tr> </tbody> </table>		Predicted N	Predicted Y	True N	1927	86	True Y	228	201
	Predicted N	Predicted Y									
True N	1927	86									
True Y	228	201									
Gradient Boosting	<p> ccp_alpha: 0.0 criterion: friedman_mse init: None learning_rate: 0.1 loss: log_loss max_depth: 3 max_features: None max_leaf_nodes: None min_impurity_decrease: 0.0 min_samples_leaf: 1 min_samples_split: 2 min_weight_fraction_leaf: 0.0 n_estimators: 10 n_iter_no_change: None random_state: 0 subsample: 1.0 tol: 0.0001 validation_fraction: 0.1 verbose: 0 warm_start: False </p>	<p>Rendimiento obtenido: 0.8656838656838657</p> <p>Matriz de confusión:</p>  <table border="1"> <thead> <tr> <th></th> <th>Predicted N</th> <th>Predicted Y</th> </tr> </thead> <tbody> <tr> <th>True N</th> <td>1961</td> <td>52</td> </tr> <tr> <th>True Y</th> <td>276</td> <td>153</td> </tr> </tbody> </table>		Predicted N	Predicted Y	True N	1961	52	True Y	276	153
	Predicted N	Predicted Y									
True N	1961	52									
True Y	276	153									

Resultado general de la Prueba #1:

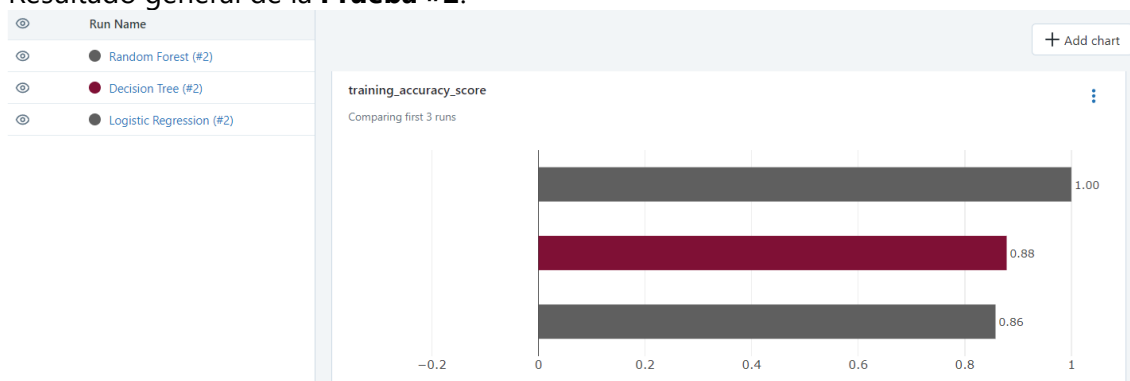


• Prueba #2

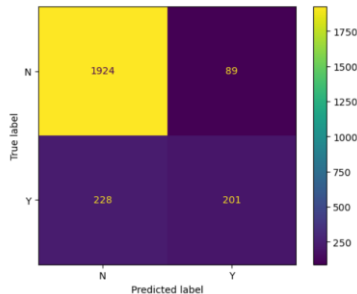
Técnica utilizada	Parámetros	Resultados obtenidos
Logistic Regression	C: 1.0 class_weight: None dual: False fit_intercept: True intercept_scaling: 1 l1_ratio: None max_iter: 100 multi_class: auto n_jobs: None penalty: l2 random_state: None solver: lbfgs tol: 0.0001 verbose: 0 warm_start: False	Rendimiento obtenido: 0.8660933660933661 Matriz de confusión:
Decision Tree	ccp_alpha: 0.0 class_weight: None criterion: entropy max_depth: 5 max_features: None max_leaf_nodes: None min_impurity_decrease: 0.0 min_samples_leaf: 1 min_samples_split: 15 min_weight_fraction_leaf: 0.0 random_state: None splitter: best	Rendimiento obtenido: 0.8767403767403767 Matriz de confusión:
RandomForest	Bootstrap: True ccp_alpha: 0.0 class_weight: None	Rendimiento obtenido: 0.8841113841113841

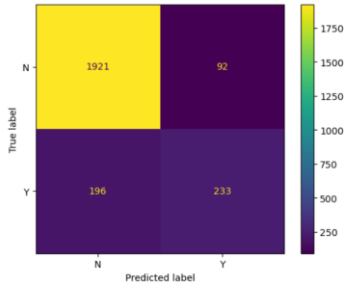
	criterion: entropy max_depth: None max_features: sqrt max_leaf_nodes: None max_samples: None min_impurity_decrease: 0.0 min_samples_leaf: 1 min_samples_split: 2 min_weight_fraction_leaf: 0.0 n_estimators: 100 n_jobs: None oob_score: False random_state: 0 verbose: 0 warm_start: False	Matriz de confusión: 
--	--	---

Resultado general de la **Prueba #2**:

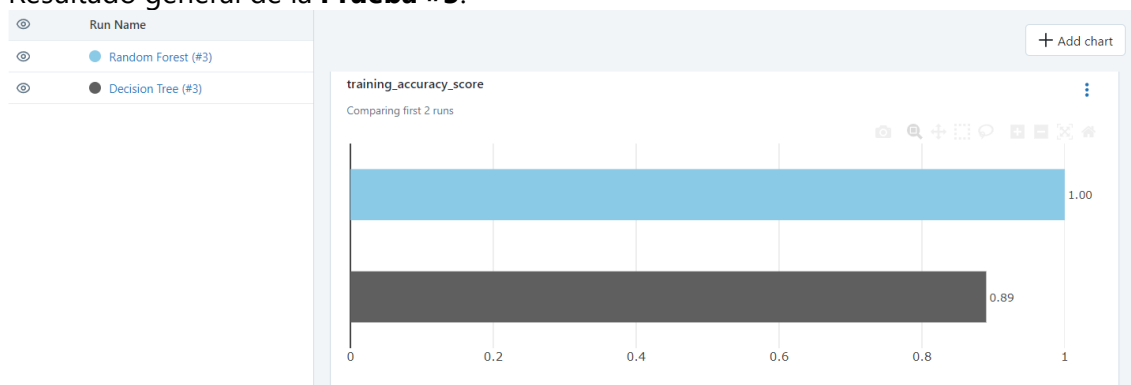


• Prueba #3

Técnica utilizada	Parámetros	Resultados obtenidos
Arboles de decisión (TDIDT)	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: 1 min_samples_split: 20 min_weight_fraction_leaf: 0.0 random_state: 0 splitter: best	Rendimiento obtenido: 0.8701883701883701 Matriz de confusión: 
RandomForest	Bootstrap: True ccp_alpha: 0.0 class_weight: None	Rendimiento obtenido: 0.8820638820638821

	<p>criterion: log_loss 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 n_estimators: 150 n_jobs: None oob_score: False random_state: 0 verbose: 0 warm_start: False</p>	<p>Matriz de confusión:</p> 
--	--	---

Resultado general de la **Prueba #3**:



6. Evaluación

Evaluación de los resultados

En función de los resultados obtenidos a partir del plan de pruebas planteado se ha seleccionado la técnica *Random Forest* para realizar la predicción sobre los datos correspondientes a nuevos créditos. Esto se ha debido a la efectividad obtenida en las diferentes instancias de evaluación que ha alcanzado el valor de 88.21%.

Proceso de revisión

En función de los datos obtenidos, la técnica seleccionada en la actividad anterior va a ser ejecutada sobre el *dataset* "**datos_nuevos.csv**" que se corresponde a nuevos créditos otorgados por la entidad solicitante del proyecto.

En estos datos se tendrá que predecir el valor del atributo "**falta_pago**" para indicar si un crédito en particular podría o no entrar en mora.

Los resultados se documentan en el apartado de "**Informe Final**" del presente documento.

Determinación de futuras tareas

Como tareas a implementar en un futuro, dada la continuidad del proyecto se pueden mencionar:

- Incorporar más atributos con respecto a la situación socioeconómica del cliente, por ejemplo: los ingresos totales del hogar, las edades de las personas que tiene a cargo, si alguna de ellas tiene algún tipo de dificultad agregada que pudiera indicar una erogación en particular, entre otros.
- Visualizar el árbol de decisión, por ejemplo con la librería *dtreeviz* de Python, para poder visualizar en detalle la estructura y las decisiones tomadas por el árbol, para de tal forma entender mejor el modelo.

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

- Mejorar aspectos de calidad de datos con respecto a la presencia de valores nulos en diferentes atributos.
- Añadir nuevas reglas para mejorar la limpieza de los datos, con los nuevos atributos que se incorporarán con respecto a la situación socioeconómica del cliente.

7. Despliegue/Implementación

Plan de implementación

Las autoridades de la entidad financiera han determinado que el modelo generado sea utilizado como herramienta de asesoramiento al sector de la entidad que se dedica al monitoreo de créditos. Además, se ha dispuesto realizar actualizaciones periódicas del modelo con nuevos datos que vayan siendo recolectados a lo largo del año.

Por otro lado, se ha establecido que se van a evaluar diferentes alternativas para incorporar más información socioeconómica de los clientes, aún si esto implica utilizar fuentes externas de datos.

Supervisión y Mantenimiento

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

- Monitoreo de la efectividad de los resultados del producto de datos contra la realidad. Tal vez incorporando esta evaluación a partir del trabajo con usuarios expertos en el dominio particular.
- Contabilización de los accesos a la herramienta por parte de los usuarios definidos para la misma.
- Implementación de un sistema de mejora continua mediante encuestas y comentarios de los usuarios para evaluar la satisfacción e identificar posibles mejoras o recopilar sugerencias para futuras evoluciones.

Informe Final

Se presentan los resultados de la aplicación del modelo generado con la técnica que ha presentado el mejor resultado en las diferentes iteraciones de pruebas aplicado sobre los datos de créditos nuevos otorgados por la entidad.

Los resultados obtenidos se describen a continuación:

	Cantidad	Porcentaje
Créditos que podrían presentar mora	Y: 27	24.11%
Créditos que podrían no presentar mora	N: 85	75.89%
Total	112	100%

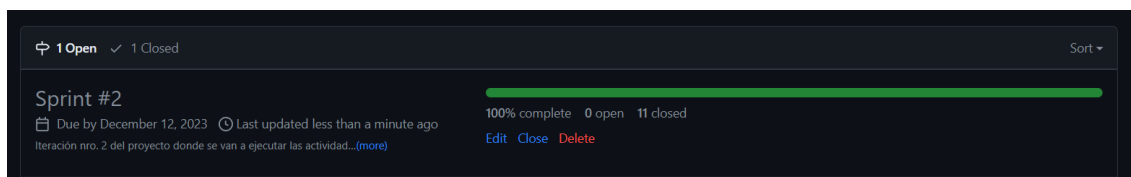
Revisión del proyecto

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

- Incorporar herramientas que brinden mayor soporte a la interacción entre los expertos del dominio y el equipo de trabajo para evitar esperas que limitan la velocidad del desarrollo del producto.

- Implementar puntos de retrospectiva al final de cada iteración para identificar las lecciones aprendidas y las posibles mejoras en la metodología, así como proponer cambios para optimizar la eficiencia del equipo en futuras iteraciones.
- Estudiar la posibilidad de establecer un sistema de evaluación de la calidad de los datos durante cada fase, para garantizar la integridad, precisión y consistencia de los datos utilizados.

A partir de la ejecución de esta actividad se da por finalizado el Sprint #2 del proyecto.



Anexo 1. Modelo de memoria de trabajo para evaluación de calidad de datos

Definición de objetivos y características de la evaluación inicial

Descripción del uso propuesto

Las autoridades de una entidad financiera desean obtener conocimiento a partir de su base de datos histórica de créditos otorgados. El objetivo principal será predecir si un crédito determinado podría pasar a ser considerado en mora (default).

En función de este objetivo se desarrolla un proyecto de ciencia de datos a fin de desarrollar un producto de datos que constituya 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.

Definición de calidad

Se van a analizar los siguientes atributos de calidad:

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:

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

Características que deben cumplir 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%

Registro de metadatos de cada dataset

Tarea resuelta en las actividades anteriores.

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.

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

Resultados de los análisis

Dimensión: Completitud

Dataset Crédito

Identificador	Umbral de aceptación	Resultados obtenidos	Evaluación
completitud_f	20%	Filas que incumplen el umbral de nulos en columnas [completitud_f]: 0 (0.0)%	Ok
completitud_d	10%	Filas que presentan nulos en el dataset [completitud_d]: 1225 (12.1)%	No cumplimiento

Dataset Tarjetas

Identificador	Umbral de aceptación	Resultados obtenidos	Evaluación
completitud_f	20%	Filas que incumplen el umbral de nulos en columnas [completitud_f]: 0 (0.0)%	Ok
completitud_d	10%	Filas que incumplen el umbral de nulos en columnas [completitud_f]: 0 (0.0)%	Ok

Dimensión: Exactitud

Dataset Crédito

Identificador	Umbral de aceptación	Resultados obtenidos	Evaluación
formato_valido	10%	<u>No</u> se encuentran atributos con formato específico	Ok
valores_ajustados	0%		Evaluación: no cumplimiento (3/11)
Atributo: "edad"		Cantidad de filas con valores fuera de rango en atributo edad (%): 4 (0.04 %)	No cumplimiento

Atributo "situacion_vivienda"	Cantidad de filas con valores fuera de rango en atributo situacion_vivienda (%): 0 (0.0 %)	Ok
Atributo "importe_solicitado"	Cantidad de filas con valores fuera de rango en atributo importe_solicitado (%): 0 (0.0 %)	Ok
Atributo "duración_credito"	Cantidad de filas con valores fuera de rango en atributo duración_credito (%): 0 (0.0 %)	Ok
Atributo "antigüedad_empleado"	Cantidad de filas con errores de rango en atributo antigüedad_empleado (%): 339 (3.35%)	No cumplimiento
Atributo "ingresos"	Cantidad de filas con valores fuera de rango en atributo ingresos (%): 0 (0.0 %)	Ok
Atributo "objetivo_credito"	Cantidad de filas con valores fuera de rango en atributo objetivo_credito(%): 0 (0.0 %)	Ok
Atributo "pct_ingreso"	Cantidad de filas con valores fuera de rango en atributo pct_ingreso (%): 0 (0.0 %)	Ok
Atributo "tasa_interes"	Cantidad de filas con errores de rango en atributo tasa_interes(%): 912 (9.01%)	No cumplimiento
Atributo "estado_credito"	Cantidad de filas con valores fuera de rango en atributo estado_credito (%): 0 (0.0 %)	Ok
Atributo "falta_pago"	Cantidad de filas con valores fuera de rango	Ok

		en atributo falta_pago (%) : 0 (0.0 %)	
valores_errores (reglas del negocio)	10%		Evaluación: ok (2/2)
Regla 1: Para aquellos casos en que los créditos constituyan un porcentaje de los ingresos del cliente mayor al 50% sus ingresos deberán ser mayores a 20.000.		Cantidad de filas que no cumplen la regla: 15 (0.15 %)	Ok
Regla 2: Para aquellos créditos cuya duración sea la mínima permitida el porcentaje de los ingresos del cliente (con respecto al importe solicitado) no podrá exceder el 60% salvo en los casos de los que sea propietario de su vivienda.		Cantidad de filas que no cumplen la regla: 7 (0.07 %)	Ok

Dataset Tarjetas

Identificador	Umbral de aceptación	Resultados obtenidos	Evaluación
formato_valido	10%	No se encuentran atributos con formato específico	Ok
valores_ajustados	0%		Evaluación: ok (10/10)
Atributo: "antigüedad_cliente"		Cantidad de filas con valores fuera de rango en atributo antigüedad_cliente (%) : 0 (0.0 %)	Ok
Atributo "estado_civil"		Cantidad de filas con valores fuera de rango en atributo estado_civil (%) : 0 (0.0 %)	Ok
Atributo "estado_cliente"		Cantidad de filas con valores fuera de rango en atributo estado_cliente (%) : 0 (0.0 %)	Ok
Atributo "gastos_ult_12m"		Cantidad de filas con valores fuera de rango en atributo duración_credito (%) : 0 (0.0 %)	Ok
Atributo "genero"		Cantidad de filas con valores fuera de rango en atributo genero (%) : 0 (0.0 %)	Ok

Atributo "limite_credito_tc"	Cantidad de filas con valores fuera de rango en atributo limite_credito_tc (%): 0 (0.0 %)	Ok
Atributo "nivel_educativo"	Cantidad de filas con valores fuera de rango en atributo nivel_educativo (%): 0 (0.0 %)	Ok
Atributo "nivel_tarjeta"	Cantidad de filas con valores fuera de rango en atributo nivel_tarjeta (%): 0 (0.0 %)	Ok
Atributo "operaciones_ult_12m"	Cantidad de filas con valores fuera de rango en atributo operaciones_ult_12m (%): 0 (0.0 %)	Ok
Atributo "personas_a_cargo"	Cantidad de filas con valores fuera de rango en atributo personas_a_cargo (%): 0 (0.0 %)	Ok

Dimensión: Consistencia

Identificador	Umbral de aceptación	Resultados obtenidos	Evaluación
claves_unicas	0%		Ok
Dataset: datos_creditos		Antes del análisis de duplicados: 10127 - Después del filtrado de duplicados: 10127 <i>No se detectaron claves duplicadas</i>	Ok
Dataset: datos_tarjetas		Antes del análisis de duplicados: 10127 - Después del filtrado de duplicados: 10127 <i>No se detectaron claves duplicadas</i>	Ok
integridad_referencial	10%	- Filas del dataset creditos (inicial): 10127	Ok

		- Filas del dataset tarjetas (inicial): 10127 - Errores detectados en la operación de unión: 0 - Filas del dataset unificado: 10127	
--	--	--	--

Identificación de mejoras aplicables

En función del análisis realizado se pueden establecer las siguientes mejoras aplicables sobre la calidad de los datos analizados:

- En el análisis realizado se han detectado cuatro valores fuera de rango en el atributo edad (*datos_creditos*), lo cual podría tratarse de un error de carga o de cálculo de datos.
- También se han detectado filas con errores de rango en el atributo antigüedad_empleado (*datos_creditos*), entre ellos 337 valores nulos y dos valores fuera de rango, que podrían deberse a errores de carga o de cálculo de los datos. Se recomienda confirmar la naturaleza de estos nulos, al poder tratarse de clientes sin trabajo por cuenta ajena.
- Igualmente se ha detectado un porcentaje elevado de nulos, casi un millar, en el atributo tasa_interes (*datos_creditos*), por un posible error de carga o de cálculo de los datos.

Por lo tanto, en estos tres casos se recomienda revisar y, si fuera necesario, ajustar los controles o validaciones en los sistemas de origen de tal conjunto de datos.