Escuela de Administración
Escuela de Ingeniería, Ciencia y Tecnología

Maestría en Business Analytics

Modelo de propensión de negocios exitosos en proyectos de edificación de vivienda en una empresa del sector de la construcción

Presentado por:

Paula Daniela Mesa Joven y Natalia Andrea Sarmiento Castillo

Bogotá, D.C. 27 de Mayo de 2023

Escuela de Administración
Escuela de Ingeniería, Ciencia y Tecnología

Maestría en Business Analytics

Modelo de propensión de negocios exitosos en proyectos de edificación de vivienda en una empresa del sector de la construcción

Presentado por:

Paula Daniela Mesa Joven y Natalia Andrea Sarmiento Castillo

Bajo la dirección de:
Ms, c. Carlos Labanda

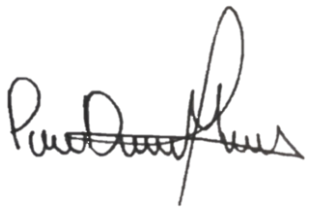
Bogotá, D.C. 27 de Mayo de 2023

Tabla De Contenido

Declaración de Originalidad y Autonomía

Declaramos bajo la gravedad del juramento, que hemos escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por nuestra propia cuenta y que, por lo tanto, su contenido es original.

Declaramos que hemos indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.



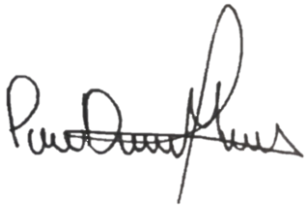
Paula Daniela Mesa Joven

Natalia Andrea Sarmiento Castillo

Firmado en Bogotá, D.C. el 23 de Mayo de 2023

Declaración de Exoneración de Responsabilidad

Declaro(amos) que la responsabilidad intelectual del presente trabajo es exclusivamente de su(s) autor(es). La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.



Paula Daniela Mesa Joven

Natalia Andrea Sarmiento Castillo

Firmado en Bogotá, D.C. el 23 de Mayo de 2023

Lista De Figuras

Lista De Tablas

Resumen Ejecutivo

Para las empresas del sector de la construcción, un negocio es exitoso es aquel que culmina con la escrituración del inmueble y a través de un indicador como la tasa de éxito (total de negocios exitosos con relación al total de solicitantes) se evalúa constantemente si los solicitantes iniciales realizan finalmente la compra del inmueble.

En la actualidad, la compañía estima que la tasa de éxito de un proyecto de vivienda oscila entre el 60%-80%, lo cual tiene gran impacto en la rentabilidad del negocio, debido a que el desistimiento de un solicitante genera reprocesos, disminución de ingresos y aumento de costos.

Desarrollar un modelo de propensión que establezca la probabilidad de éxito del negocio desde el inicio del proceso de venta, facilitará y agilizará la toma de decisiones en la evaluación del perfil de los solicitantes permitiendo enfocar sus recursos de ventas y mercadeo mediante la segmentación de perfiles en función de su probabilidad y garantizar la culminación del negocio hasta la escrituración.

Palabras Clave:

Tasa de éxito, modelo de propensión, probabilidad de éxito, proyectos de vivienda.

Abstract

For companies in construction sector, a successful business is one that culminates in the property deed and through an indicator such as the success rate (successful businesses total in relation to number of applicants total) it is constantly evaluated whether the initials applicants finally make the purchase of the property.

Currently, the company estimates that the success rate of a housing project ranges between 60%-80%, which has a significant impact on the profitability of the business, because the withdrawal of an applicant generates reprocesses, decreases the revenue, and increases costs.

Developing a propensity model that establishes the success probability of the business from the beginning of sales process, will facilitate and speed up decision making in the evaluation of the profile of the applicants, allowing them to focus their sales and marketing resources through the segmentation of profiles in based on its probability and guarantee the completion of the business until the deed.

Keywords:

Success rate, propensity model, success probability, housing projects.

1. Introducción

Teniendo en cuenta que la cifra del PIB del sector de la construcción se encuentra alrededor de un 6,2% en lo que lleva del año 2022, siendo la categoría de edificaciones del 13,3 %1, una de las ramas económicas más importantes del país, surge la necesidad de analizar el comportamiento de la oferta y demanda, ya que no solo aporta económicamente al crecimiento del país, sino que también genera anualmente genera más de 20 mil empleos a nivel nacional. Dicho lo anterior, es imperativo aclarar que el sector ha tenido un aceleramiento importante en los últimos dos años producto del apalancamiento que le ha dado el sector financiero en conjunto con el gobierno nacional, mediante la implementación de nuevas medidas normativas las cuales tiene como fin por medio de la implementación de nuevos subsidios (Jóvenes propietarios, E cobertura, Mi casa ya, subsidios de caja de compensación) y beneficios en las tasas de crédito (Fresh) con el fin incentivar la compra de vivienda nueva.

Sin embargo, a pesar de lo mencionado anteriormente, se evidencia que la tasa de deserción de negocios de vivienda es alta (20%-40%) por lo cual, surge la necesidad de entender cuáles son las principales razones que afectan en la actualidad la concepción de negocios exitosos, con el fin de optimizar recursos, minimizar riesgos en conjunto con el comportamiento del mercado y a fin de responder de forma acertada a las necesidades de este.

De acuerdo con datos tomados del mercado, tal como lo describe la Compañía Trend Group América, la cual, realizó una investigación donde identificó que el tiempo estimado de búsqueda de un inmueble por un hogar colombiano promedio se encontraba en un promedio de 1 a 6 meses antes de la cuarentena; Esta situación cambio en el periodo de pandemia, dado que,

---

[1] (Portafolio, 2022)

se estima que el 53,8% de los compradores potenciales tenía la intensión de aplazar la compra de dicho bien en un lapso de tiempo entre 1 a 2 años; lo cual, hace que el panorama a nivel de ventas de construcción cambie radicalmente y plantea varios retos para las compañías, puesto que surge la necesidad de ofrecer proyectos a largos plazos y por consiguiente implica cambios en las estrategias de Mercadeo a fin de lograr negocios exitosos.2Por consiguiente, se puede afirmar que las empresas del sector de la construcción tienen un gran reto en cuanto a reducción de negocios no exitosos (tasas de desistimiento), ya que al no tener un panorama claro del tipo de cliente ideal de acuerdo al tipo de producto ofrecido y otras variables como ubicación, estrato socioeconómico, ingresos, estado civil entre otros, hacen que las estrategias de marketing no sean lo suficientemente fructíferas y por tanto se empiece a impactar la rentabilidad de los proyectos dada la disminución de ingresos, reprocesos de carácter administrativo y aumentos de costos al tiene inventario muerto, pero que genera unos gatos fijos asociados costos de administración, servicios públicos, impuestos Etc. Una vez se encuentra el bien inmueble entregado a la copropiedad.

[2] (America, 2020)

2. Objetivos

Objetivo General

Identificar los niveles de desistimientos de negocios en proyectos de edificación de vivienda en una empresa del sector de la construcción a través de la generación de modelos analíticos que determinen la probabilidad de éxito de un negocio de acuerdo con sus características evitando afectaciones en el desempeño financiero y procesos administrativos.

Objetivos Específicos

- Integrar y analizar la información histórica de los negocios exitosos y no exitosos de la compañía a partir de la comprensión del entorno empresarial para la identificación de las principales razones de desistimiento, características de clientes y tasas de éxito de los proyectos relacionados.
- Determinar los modelos más apropiados que identifiquen patrones y relaciones que respondan al problema analítico identificado.
- Evaluar el desempeño de los modelos propuestos comparando los resultados obtenidos con métricas de análisis.
- Generar recomendaciones a la organización acerca de la recopilación, Estructura y calidad de los datos.

3. Alcance

Los entregables principales de este proyecto son integrar la información de las tres bases de datos principales incorporando el entendimiento del entorno empresarial y procesos base, la identificación de patrones mediante modelos analíticos seleccionados, la evaluación del desempeño de dichos modelos mediante parámetros previamente establecidos y la creación de la política de aprobación asociada a los hallazgos encontrados.

Para su desarrollo el proyecto se realizará en cinco etapas principales basadas en la metodología ASUM-DM, la cual exponemos en capítulos posteriores. Para cada etapa se especifican los hitos principales los cuales retroalimentaran las etapas siguientes y de ser necesario se llevará a cabo ajustes en el plan inicial de trabajo.

Teniendo en cuenta lo anterior, el proyecto iniciará con el entendimiento del negocio y culminará con la publicación de hallazgos con documentación de la puesta en marcha de la solución, lecciones aprendidas y consideraciones futuras. Las fases siguientes de programación, implementación y prueba de las tareas asociadas con copias de seguridad, pistas de auditoría, registro y archivo las desarrollará la compañía en estudio.

4. Metodología

Para el desarrollo del proyecto, se decide implementar la metodología ASUM-DM (Analytics Solutions Unified Method) de IBM, basada en la metodología CRISP-DM (Cross Standard Process for Data Mining).

La principal diferencia entre estas dos metodologías radica en que a pesar de que ASUM-DM (Analytics Solutions Unified Method) de IBM toma las mismas fases que se plantean en la metodología CRISP-DM, refuerza más la fase de implementación y agrega actividades útiles relacionadas con infraestructura, operaciones y gestión de proyectos[3].

Esta metodología plantea dos ciclos: el ciclo de desarrollo y el ciclo de implementación. Producir un modelo es tan importante como su implementación, por ello ASUM-DM establece unos pasos adicionales que permite mayor efectividad en la entrega de los resultados analíticos coherentes que faciliten la toma de decisiones organizacionales[4].

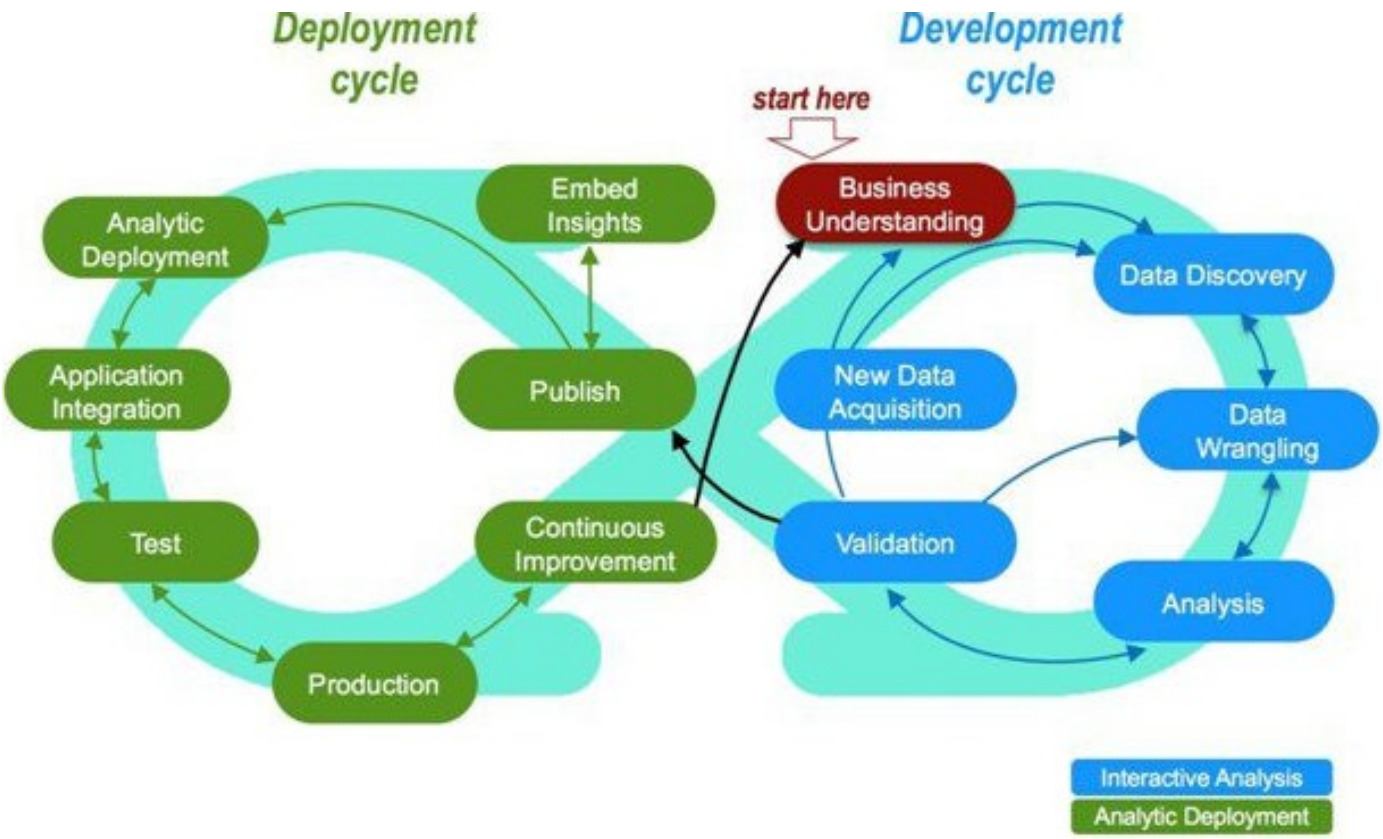
Para el desarrollo del proyecto haremos cada etapa propuesta en la metodología, pero, considerando el alcance definido, este documento presentará los entregables hasta la etapa de publicación en el ciclo de implementación.

---

[3] (IBM, 2022)
[4] (Brethenoux, 2022)

Ilustración 1.

Metodología ASUM-DM, ciclo de implementación y ciclo de desarrollo



Nota: Esta ilustración muestra el flujo de cada una de las etapas de los ciclos de desarrollo e implementación de acuerdo con la metodología ASUM-DM (Brethenoux, 2022).

Tabla 1

Descripción de las etapas y actividades por ciclo en la metodología ASUM-DM

Ciclo	Etapa	Descripción	Actividades	Descripción
Desarrollo	Entendimiento del negocio	Comprender los objetivos y requisitos del proyecto desde una perspectiva comercial, luego convertir este conocimiento en una definición del problema de minería de datos y un plan preliminar diseñado para lograr los objetivos.	Determinar los objetivos de negocio	Comprender a fondo, desde una perspectiva comercial, lo que el cliente realmente quiere lograr.
			Evaluar la situación	Búsqueda de hechos más detallados sobre todos los recursos, restricciones, suposiciones y otros factores que deben considerarse.
			Determinar los objetivos de minería de datos	Establecimiento de los objetivos del proyecto en términos técnicos.
			Crear un plan para el proyecto.	Descripción del plan previsto para lograr los objetivos de minería de datos y, por lo tanto, lograr los objetivos comerciales. El plan debe especificar los pasos que se realizarán durante el resto del proyecto, incluida la selección inicial de herramientas y técnicas.
	Descubrimiento de los datos.	Revisar los datos disponibles para la minería. Implica acceder a los datos y explorarlos para determinar la calidad de los datos. ("Conceptos básicos sobre comprensión de datos - IBM")	Recopilar datos iniciales	Adquisición de los datos. Esta recopilación inicial incluye la carga de datos, que es necesaria para la comprensión de los datos.
			Describir los datos	Examen de las propiedades superficiales de los datos adquiridos e informe sobre los resultados.
			Explorar los datos	Utilizar técnicas de consulta, visualización e informes. Esto incluye relaciones entre atributos, resultados de agregaciones simples, propiedades de subpoblaciones significativas y análisis estadísticos simples.
			Verificar la calidad de los datos	¿Están completos los datos (cubren todos los casos requeridos)? ¿Es correcto o contiene errores y, si hay errores, qué tan comunes son? ¿Hay valores faltantes en los datos? Si es así, ¿cómo se representan, ¿dónde ocurren y qué tan comunes son?

	Validación de los datos	Realizar la preparación de los datos y posterior a ello su validación garantizando que el modelado se realice con los datos correctos.	Seleccionar los datos	Decidir los datos que se utilizarán para el análisis. Los criterios incluyen la relevancia para los objetivos de minería de datos, la calidad y las restricciones técnicas, como los límites en el volumen de datos o los tipos de datos. (“Metodología CRISP-DM para minería de datos - Data Prix”)
			Limpiar los datos	Elevar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de subconjuntos limpios de datos, la inserción de valores predeterminados adecuados o técnicas más ambiciosas, como la estimación de datos faltantes mediante modelado. (“Metodología CRISP-DM para minería de datos - Data Prix”)
			Integrar los datos	Métodos mediante los cuales la información se combina de varias tablas o registros para crear nuevos registros o valores.
			Formatear datos	Modificaciones sintácticas realizadas en los datos que no cambian su significado, pero que pueden ser requeridas por la herramienta de modelado.
			Verificar datos	Validar los datos resultantes de la etapa de preparación. En caso de encontrar errores, retornar a la etapa de preparación de datos. Puede incluir la participación de un experto de negocio.
	Análisis	Ejecutar varios modelos utilizando los parámetros predeterminados y preparar datos para las manipulaciones requeridas por su modelo de elección.	Seleccionar las técnicas de modelamiento	La determinación del modelo más apropiado generalmente se basará en las siguientes consideraciones: Los tipos de datos disponibles para la minería, los objetivos de minería de datos y requisitos específicos de modelado.
			Generar diseño de prueba	Generar un procedimiento o mecanismo para probar la calidad y validez del modelo.
			Construir el modelo	Ejecución de la herramienta de modelado en el conjunto de datos preparado para crear uno o más modelos.

			Evaluar el modelo	Determinación de qué modelo(s) son lo suficientemente precisos o efectivos para ser definitivos.
	Validación	Evaluar los modelos utilizando los criterios de éxito empresarial.	Evaluar resultados	Evaluar los resultados utilizando criterios de éxito empresarial.
			Determinar nuevos pasos	Teniendo en cuenta los resultados se debe determinar si es necesario continuar al ciclo de implementación, redefinir el modelo planteado o ingresar información nueva.
	Nueva adquisición de datos	Agregar datos nuevos o que no se han tenido en cuenta en etapas anteriores.	Incluir nuevos datos o variables	Identificar variables adicionales que sean requeridas para el mejoramiento o precisión del modelo.
Implementación	Publicar	Puesta en marcha la solución y comunicación a la comunidad de usuarios finales y a las partes interesadas que la solución está activa.	Habilitar modelo de soporte para la solución.	Creación y ejecución de pruebas para asegurar que el entorno de producción esté listo para manejar la solución construida. Realizar correcciones y pruebas de regresión, retrocediendo si es necesario.
			Comunicar los detalles de la solución a los usuarios finales e interesados.	Orientar y trasferir el conocimiento a los usuarios finales.
	Publicar hallazgos	Revisar el lanzamiento y recopilar las lecciones aprendidas y los éxitos.	Revisar la preparación operativa y del usuario, y prepararse para la fase de operación.	Aseguramiento de que las lecciones aprendidas y otras documentaciones del proyecto se almacenen de forma centralizada

Nota: Esta tabla muestra la descripción de etapas y actividades a realizar por ciclos de la metodología ASUM-DM (IBM, 2022).

5. Cronograma

A continuación, presentamos el esquema de trabajo propuesto en cinco sprints para el desarrollo del proyecto: Entendimiento del negocio, validación de los datos, análisis del modelo de regresión múltiple, análisis de los modelos de árboles de clasificación y regresión logística e implementación. Para cada uno de los sprints mencionados, se describen los entregables con su respectiva fecha de inicio-fin, como también los Sprint Review y Sprint Retrospective con los cuales se buscará ejercer una adaptación útil, a través de la inspección de los resultados propuestos de cara al valor para el cliente, estipular futuros cambios y asegurar la calidad de los hitos de cada iteración.

Ilustración 2

Cronograma del Proyecto

SPRINT/DESCRIPCIÓN	PROGRESO	INICIO	FIN
SPRINT 1: Entendimiento del negocio y descubrimiento de los datos			
Planing Meeting-Cronograma de proyecto	100%	28-5-22	29-5-22
Objetivos Estratégicos-DOFA-Historias de usuario	100%	28-5-22	1-6-22
Documento descriptivo de las fuentes de información	100%	1-6-22	4-6-22
Exploración inicial de los datos	100%	1-6-22	4-6-22
Data Review Clientes	100%	5-6-22	9-6-22
Análisis de la tasa de éxito	100%	10-6-22	17-6-22
Sprint Review (Director-Product Owner-Scrum team)	100%	18-6-22	18-6-22
Sprint Retrospective	100%	27-6-22	27-6-22

SPRINT 2: Validación de los datos			
Bases de datos actualizadas	100%	12-10-22	12-10-22
Selección de los datos	100%	13-10-22	18-10-22
Limpieza de datos (estimación de datos faltantes, duplicados, outliers)	100%	19-10-22	29-10-22
Sprint Review (Director-Product Owner-Scrum team)	100%	31-10-22	31-10-22
Sprint Retrospective	100%	1-11-22	1-11-22
SPRINT 3: Análisis-Técnicas de modelamiento			
Actualización y Limpieza de datos (estimación de datos faltantes, duplicados, outliers)	100%	3-11-22	6-11-22
Determinación de las técnicas de modelamiento	100%	8-11-22	8-11-22
Análisis de componentes principales	100%	14-11-22	21-11-22
Sprint Review (Director-Product Owner-Scrum team)	100%	12-12-22	12-12-22
Sprint Retrospective	100%	13-12-22	13-12-22
SPRINT 4: Análisis-Árboles de clasificación			
Análisis Gradient Boosting	100%	1-2-23	9-2-23
Análisis Random Forest	100%	10-2-23	18-2-23
Análisis Máquina de Soporte Vectorial	100%	18-2-23	26-2-23
CatBoost/LightGBM/XGBoost	100%	26-2-23	6-3-23
Validación de resultados y determinación de nuevos pasos	100%	7-3-23	7-3-23
Sprint Review (Director-Product Owner-Scrum team)	100%	8-3-23	8-3-23
Sprint Retrospective	100%	9-3-23	9-3-23
SPRINT 5: Implementación			
Proceso de mantenimiento de la solución, lecciones aprendidas y éxitos	100%	11-3-23	21-3-23
Comunicación de hallazgos a usuarios finales	100%	26-3-23	26-3-23
Entrega de informe final	100%	28-3-23	28-3-23
Sprint Review (Director-Product Owner-Scrum team)	100%	13-3-23	13-3-23
Sprint Retrospective	100%	15-3-23	15-3-23

Nota: La ilustración muestra el esquema de trabajo con su respectivo porcentaje de avance y fechas de inicio y fin para cada entregable.

6. Entendimiento del negocio

Una compañía del sector de la construcción con proyectos en Cundinamarca, atlántico, Cali y proyectos de expansión a Medellín que ofrece proyectos de Vivienda tipo VIP (Vivienda de Interés Prioritario), VIS (Vivienda de Interés Social) y NO VIS. desea entender cuáles son los principales factores que afectan la culminación con éxito de negocios de vivienda, dado que al momento no se cuenta con un área analítica en la compañía y se ha identificado que el porcentaje de negocios desistidos se encuentra al rededor del 20%-40%; lo que hace imperativo la aplicación de modelos analíticos que permitan la toma de decisiones ya que se identificó que otras compañías del mercado que presentan mejoras en sus ventas han realizado avances en el entendimiento del negocio.

A pesar de que con los auxilios dados durante la pandemia por el gobierno nacional la compra de vivienda aumento (3030 unidades vendidas y 3215 escrituradas), se evidencio un alto desistimiento, lo cual implica costos extras a nivel comercial. razón por la que se desea mediante el análisis estadístico y la aplicación de modelos analíticos comprender los factores que influyen en la decisión de desistimiento y que estrategias permitirían mitigar el problema, ya que en 2022 comenzaron 13 nuevos proyectos.

Para las empresas del sector de la construcción, un negocio es exitoso es aquel que culmina con la escrituración del inmueble y a través de un indicador como la tasa de éxito (total de negocios exitosos con relación al total de solicitantes) se evalúa constantemente si los solicitantes iniciales realizan finalmente la compra del inmueble.

En la actualidad, la compañía estima que la tasa de éxito de un proyecto de vivienda oscila entre el 60%-80%, lo cual tiene gran impacto en la rentabilidad del negocio, debido a que

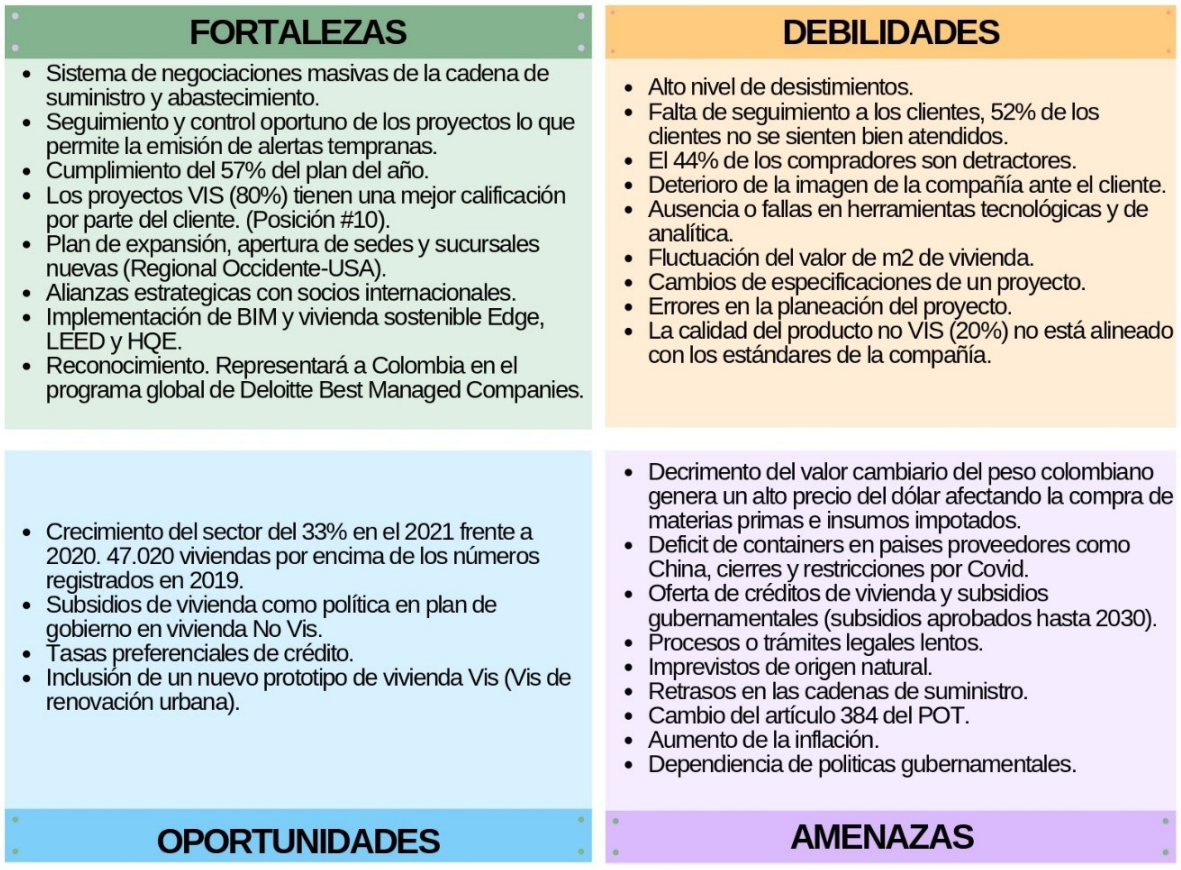
el desistimiento de un solicitante genera reprocesos, disminución de ingresos y aumento de costos.

6.1.Evaluación De La Situación Actual

Para estudiar el modelo de negocio, identificar los aspectos internos y externos positivos y negativos, se optó por la implementación de la herramienta DOFA para documentar las necesidades o problemáticas de la empresa.

Ilustración 3

Análisis DOFA



Nota: Esta ilustración muestra el análisis de debilidades, oportunidades, fortalezas y amenazas de la empresa estudio.

6.2.Evaluación Del Riesgo Para El Negocio

Como en 2020 el índice de confianza del consumidor y la intención de compra de vivienda llegaron a los niveles mínimos registrados, y el indicador de cartera vencida para crédito constructor en bancos ha presentado los valores más altos en los últimos 20 años y que no se veían desde la crisis del -UPAC- (Unidad de Poder Adquisitivo Constante) ha representado un reto para vender vivienda y culminar los negocios con éxito, ya que el riesgo inherente al desistimiento o bajas ventas en proyectos ha aumentado en los años, lo que deja importantes consecuencias para la compañía.

6.2.1. Riesgo Técnico

Implica la reformulación del proyecto dadas la pocas ventas o índices altos de desistimiento, estos cambios pueden dar a nivel de arquitectura con el fin de ofrecer un producto más atractivo, pero a su vez implican cambios en el personal requerido, tipos de contratos, materiales que pueden afectar directamente la ejecución del proyecto, así como diversos trámites y licencias. Sin embargo, estos riesgos no afectan al proyecto en más de un 10%.

6.2.2. Riesgo Financiero

La falta de ventas repercute directamente en disminución de los ingresos, los cuales, son utilizados para pagar los costos directos, costos indirectos, costos de honorarios, costos del lote, costos de urbanismo y costos financieros. Es por esto que, cuando se llega a la etapa final de proyecto, y el nivel de ventas no es suficiente para cubrir los costos, se dice que el cierre financiero del proyecto depende de la venta de inventario y puede surgir la amenaza de que los ingresos recaudados a la fecha no sean suficientes para cubrir los costos tales como el crédito constructor y el tener estas unidades en stock hace que los costos fijos del proyecto sigan

creciendo hasta el punto de que no es posible pagarlos con las ventas a la fecha e incluso las que generará el proyecto en el futuro, por lo cual, el proyecto podría generar perdidas para la compañía.

6.2.2.1. Preventas.

Para el esquema fiduciario de preventas, la fiduciaria solo desembolsa los recursos recaudados de los compradores una vez se alcanza el punto de equilibrio (unidades vendidas y avance obra) por lo cual, una baja velocidad de ventas puede representar problemas de liquidez.

6.2.2.2. Crédito Constructor.

Este riesgo va amarrado directamente la velocidad de ventas y a la velocidad de entregas, ya que determinan la duración de un proyecto y por tanto los costos financieros atados al crédito constructor.

A continuación, se explican las etapas del crédito constructor y como estas impactan directamente el correcto desarrollo del proyecto:

6.2.2.3. Estudio.

El banco otorga a la constructora un cupo de endeudamiento atado a los costos del proyecto reportados por el constructor, que, por políticas bancarias, no supera el 80 % (Asobancaria, 2021).

6.2.2.4. Desembolsos.

se dan una vez cumplidas las condiciones pactadas por el banco y de acuerdo con el avance de obra, tales como el nivel de preventas, monto de recaudos, Punto de equilibrio comercial, cierre financiero, entre otros.

6.2.2.5. Cancelación.

Se da siempre y cuando el cupo aprobado se encuentre desembolsado y la obra culminada.

Según el proceso mencionado, es importante aclarar que los recursos para pagar el crédito constructor provienen de créditos hipotecarios o leasing habitacionales tomados por los clientes, que para el caso de estudio resulta en negocios exitosos, dado que la modalidad de pago de los inmuebles suele darse un 30% de cuota inicial (Recaudos y subsidios) y un 70% por medio de crédito constructor; por lo que, el valor del crédito se causa conforme se escrituren los inmuebles.

6.2.2.6. Escrituración.

Retrasos en la escrituración también pueden darse a causa del desistimiento del cliente por motivos de negación del subsidio o crédito constructor (por un estudio de crédito negativo, reportes en centrales, informalidad de ingresos, capacidad de pago insuficiente, entre otras), lo que resulta en una situación crítica para el proyecto, dado que, en esta etapa, se vuelve complicado la venta de esta unidad nuevamente puesto que, el nuevo comprador no tendrá el tiempo suficiente para pagar la cuota inicial de forma diferida, sino que, deberá cancelar el 100% del valor del inmueble en un periodo corto de tiempo.

Lo mencionado anteriormente, puede incrementar los costos fijos y financieros al retrasarse los recursos provenientes de los créditos hipotecarios y operaciones de leasing habitacional, ocasionando un desfase en el presupuesto que lleva a pérdidas económicas.

6.3.Objetivos Organizacionales

Para la empresa en estudio, uno de los objetivos estratégicos es el posicionamiento de la marca. Para lograrlo buscan un mejoramiento en la recordación de la marca y atención al cliente. Lo anterior no solo mediante el excelente servicio de la fuerza comercial y el conocimiento de los clientes para brindarles servicios según sus necesidades, aumentando la captación y evitando fuga de solicitantes o clientes en cada etapa del embudo de ventas.

Por otro lado, la empresa manifiesta necesario generar proyectos exitosos, proyectos que generan la menor pérdida posible para de esta manera lograr no solo sanidad financiera sino también una mayor expansión en la región.

Teniendo en cuenta lo anterior, identificamos los objetivos organizacionales y los indicadores de negocio relacionados:

6.3.1. Objetivo Organizacional 1.

Ofrecer productos y servicios que respondan a las necesidades y expectativas de los clientes de forma sobresaliente.

KPI´s relacionados al objetivo.

- Posicionar marca estar en el top of mind.
- Disminuir el nivel de desistimientos.
- Incrementar Net Promoter Score.

6.3.2. Objetivo Organizacional 2.

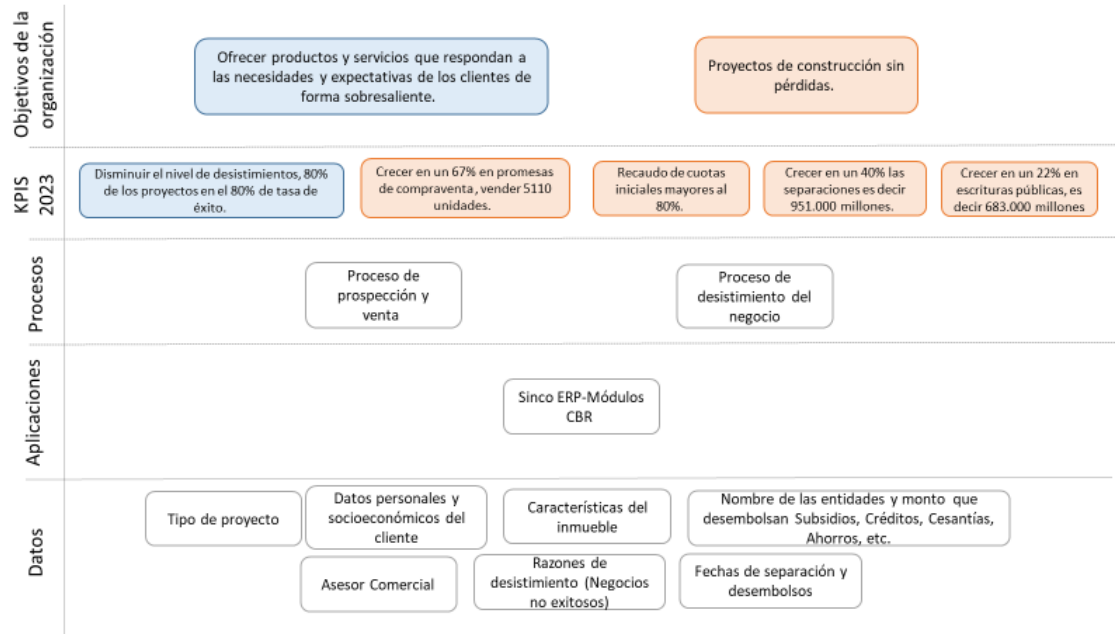
Proyectos de construcción sin pérdidas.

KPI´s relacionados al objetivo.

- 80% de los proyectos en el 80% de tasa de éxito.
- Recaudo de cuotas iniciales mayores al 80%.
- Generar estructuras contractuales equilibradas asegurando negocios sostenibles para las partes.
- Crecer en un 67% en promesas de compraventa es decir que debemos lograr vender 5110 unidades.
- Crecer en un 40% las separaciones es decir 951.000 millones.
- Crecer en un 22% en escrituras públicas, es decir 683.000 millones.

6.3.3. Arquitectura Empresarial

Ilustración 4
TOGAF



Nota: Esta ilustración muestra el esquema de arquitectura empresarial en cuatro dimensiones: objetivos organizacionales, procesos, aplicaciones y datos.

6.3.4. Proceso De Prospección Y Venta

Aplica para todos los proyectos comercializados por la compañía, desde la entrega del negocio aprobado por parte de la sala de ventas hasta la escrituración del negocio.

6.3.4.1. Áreas Involucradas.

- Planeación
- Comercial
- Relaciones Comerciales.
- Relaciones Corporativas.
- Financiero
- Administración

6.3.4.2. Condiciones Generales.

Al momento de estructurar los planes de pago pactar las cuotas mensuales dentro los días 1 al 20 con el fin de evitar moras en los negocios y a su vez generar un negocio con calificación de alto riesgo

Para la firma de la Promesa de Compraventa es necesario que el cliente cuente con los siguientes documentos con el fin de revisar el plan de pagos y poder realizar los ajustes necesarios si hay lugar a ello:

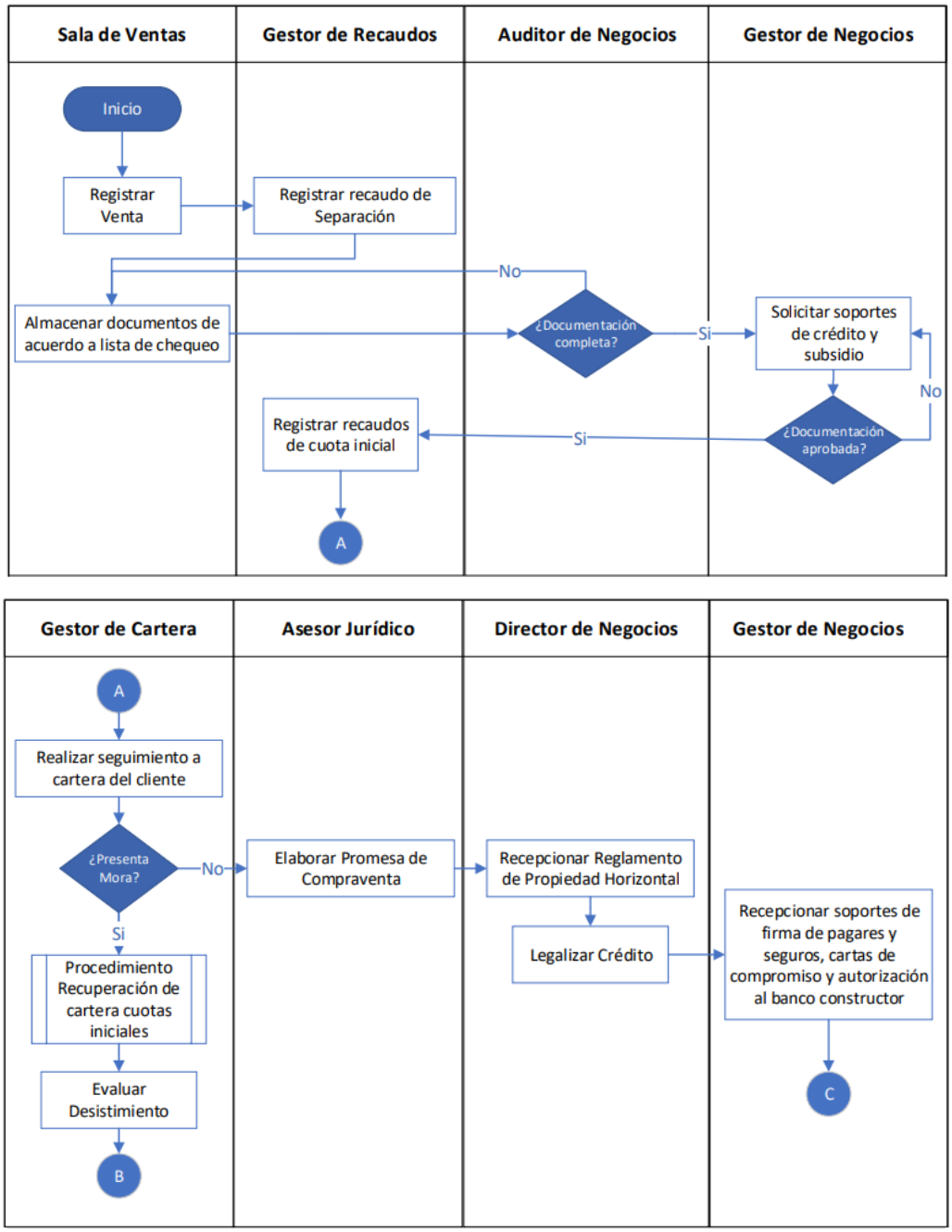
- Carta de aprobación del crédito (si aplica)
- Carta de aprobación de subsidio de vivienda (si aplica)
- Soporte de las consignaciones realizadas
- Certificación de cesantías, cuentas AFC, CDT, entre otros (si aplica)

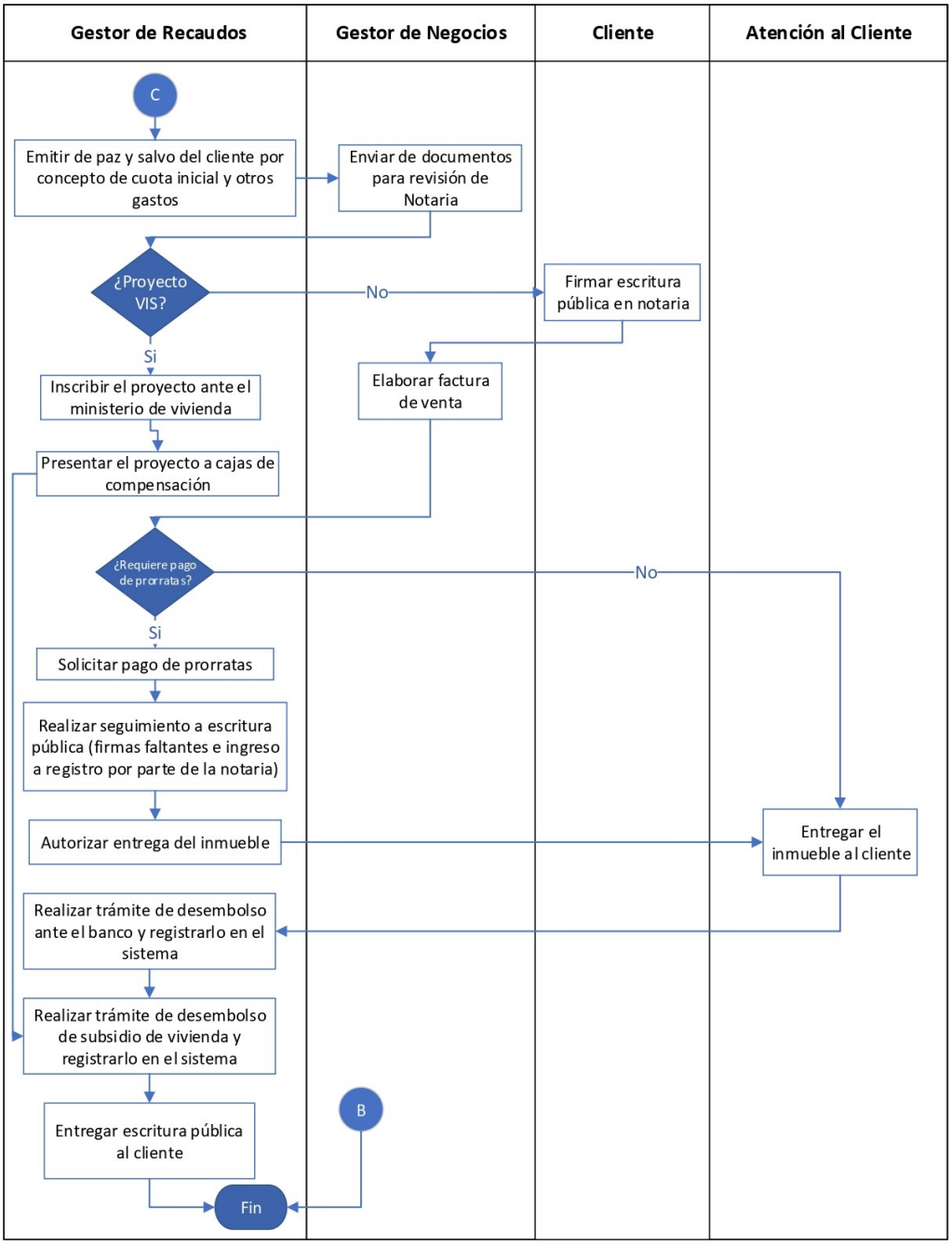
Para la entrega de los inmuebles es necesario contar con la liberación por parte de la obra

6.3.4.3. Diagrama Del Proceso.

As Is.

Ilustración 5
AS IS ventas



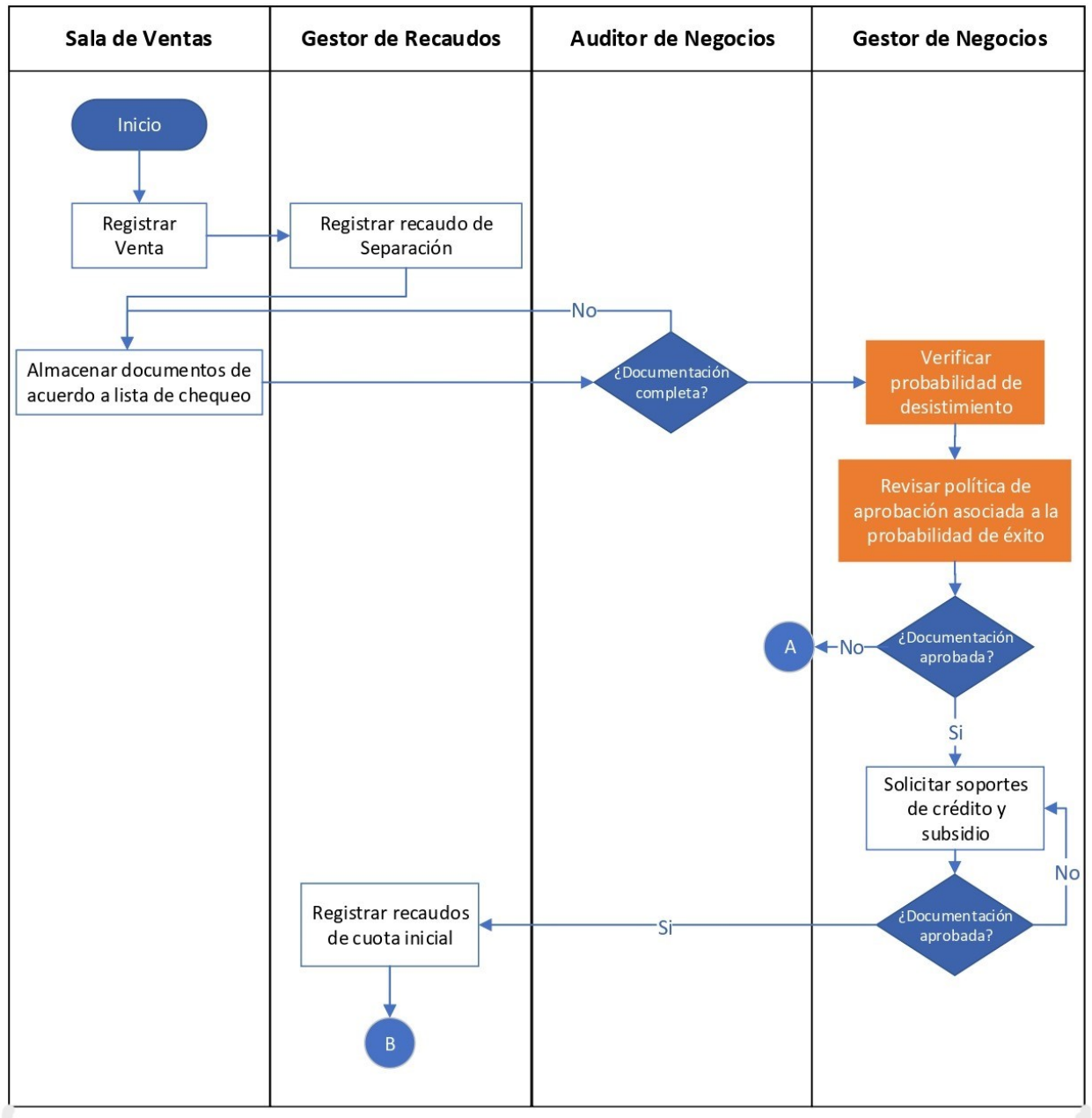


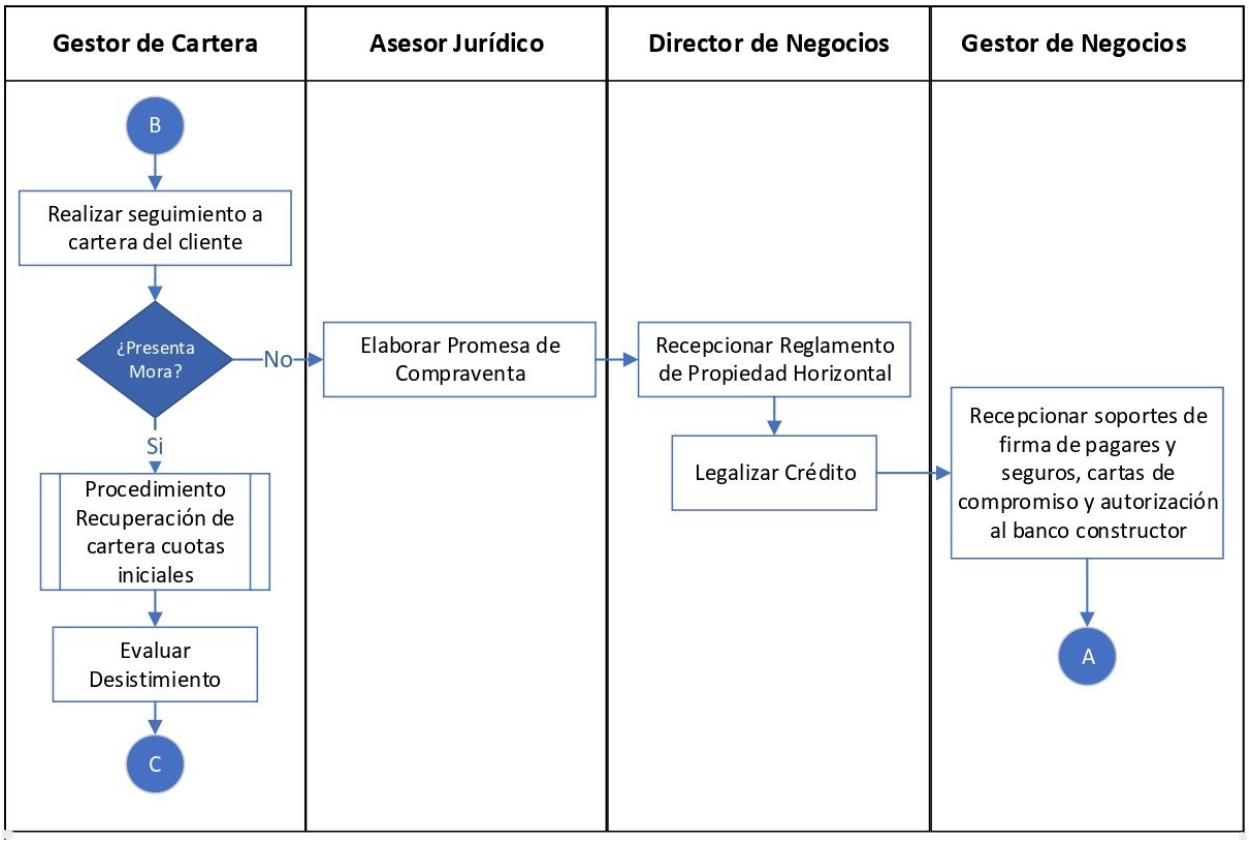
Nota: Esta ilustración muestra el esquema del proceso de venta de un inmueble en la compañía de estudio desde el registro de la venta hasta la entrega del inmueble.

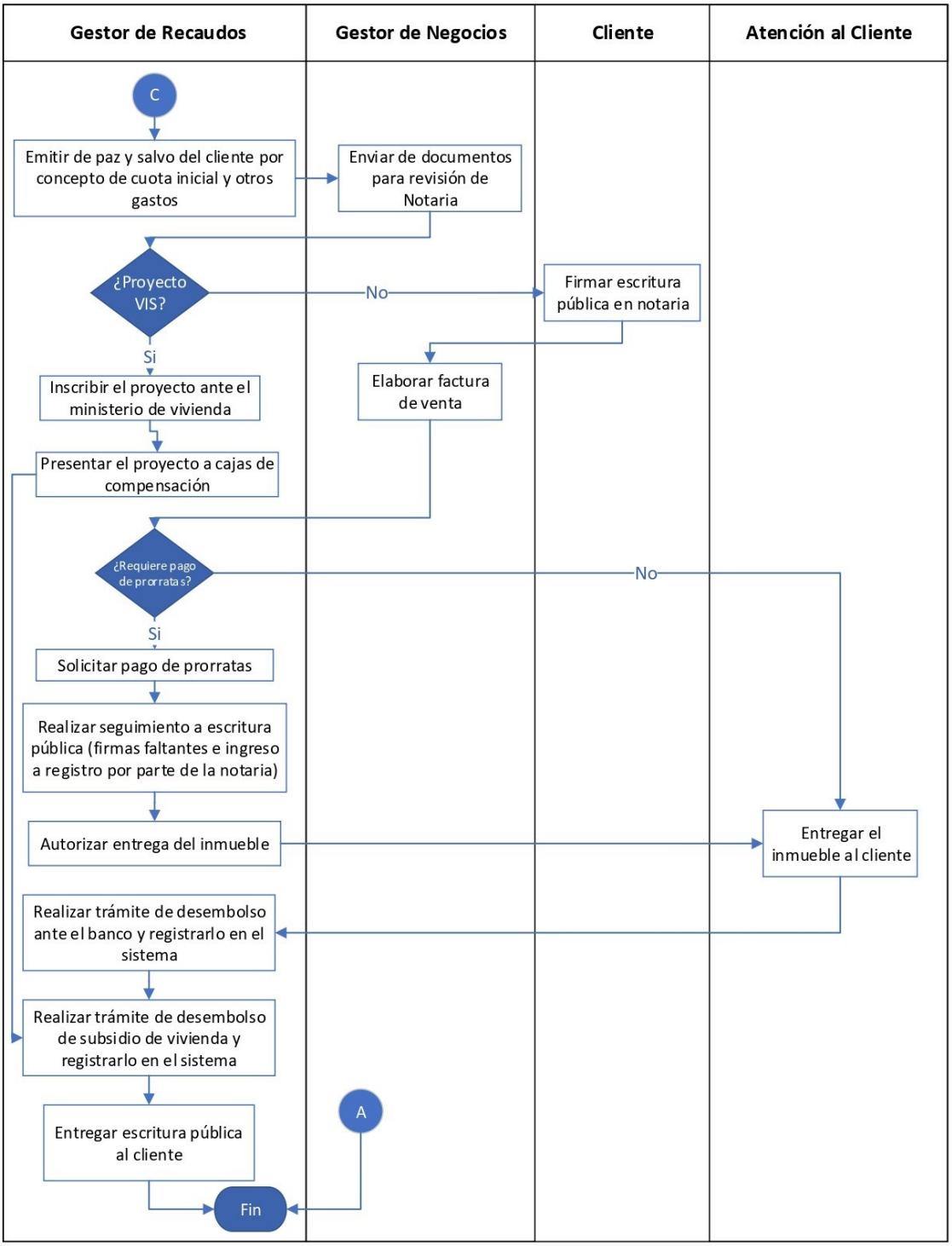
To Be.

Ilustración 6

TO BE ventas




Gestor de Cartera
Asesor Jurídico
Director de Negocios
Gestor de Negocios
B
Realizar seguimiento a cartera del cliente
¿Presenta Mora?
No
Si
Elaborar Promesa de Compraventa
Recepcionar Reglamento de Propiedad Horizontal
Legalizar Crédito
Recepcionar soportes de firma de pagares y seguros, cartas de compromiso y autorización al banco constructor
A
Procedimiento Recuperación de cartera cuotas iniciales
Evaluar Desistimiento
C



Nota: Esta ilustración muestra el esquema del proceso de venta de un inmueble en la compañía de estudio desde el registro de la venta hasta la entrega del inmueble teniendo en cuenta una mejora al proceso.

Teniendo en cuenta que el objetivo del proyecto es determinar la probabilidad de éxito en negocios de vivienda con el fin de mitigar o poder emitir alertas tempranas a las áreas que intervienen en los procesos y que puede causar sobrecostos, se plantea una actividad previa a la aprobación de la documentación enviada por el cliente, la cual consiste en la verificación del gestor de negocios de la probabilidad de desistimiento del cliente por medio de la aplicación del modelo propuesto, el cual emitirá un score que deberá evaluar de acuerdo a la política de aprobación y/o vinculación asociada a la propensión de éxito de negocios de vivienda para el proyecto.

6.3.5. Proceso De Desistimiento

Cliente que manifiesta desistir por voluntad propia, desistido unilateral. Los casos se presentan en mesas de trabajo en el mes en curso antes de pasarlos a acta final.

El acta final, se socializa por correo electrónico los treinta o treinta y uno del mes y se enviara por Docusign para toma de firmas de los miembros del comité.

Así mismo el acta se libera de Sipro los primeros dos días hábiles del mes.

Los tipos de desistidos que se pueden presentar en las diferentes etapas del trámite del negocio (desde su inicio hasta la escrituración de este), son los siguientes:

Voluntad Del Cliente. El cliente manifiesta su voluntad de desistir del negocio por un motivo especifico, quien deberá los siguientes documentos según sea el caso:

Incapacidad Económica. Soportes que lo demuestren

Pérdida De Empleo. Última certificación laboral con vigencia no mayor a 30 días.

Enfermedad. Copia de la historia clínica

Desistimiento Unilateral. Prodesa toma la decisión de Desistir el negocio por motivos tales como Incumplimiento en pagos o tramites.

Desistimiento Sala de Ventas. Aquellos negocios que la sala de ventas no logra legalizar semanalmente.

Mensualmente se realiza un Comité de Desistidos, en donde se revisan todos los negocios recibidos para desistir con su respectiva justificación y se toma una decisión de estos (posibilidad de traslado, replanteo del negocio, desistido con o sin arras).

Áreas Del Negocio: Comercial, Negocios, Cartera.

Comercial. Presenta ventas aquellos negocios que no llegaron a un feliz término y no firmaron documentos de cierre. (presentan toda la documentación de desistimiento)

Negocios. Presentan los negocios que ya estén recibidos por el área (presentan toda la documentación de desistimiento)

Cartera. Presenta los negocios que estén cerrados (presentan toda la documentación de desistimiento)

6.3.5.1. Requisitos.

Contar con la documentación completa en acta final, unidades que no estén con documentos completos pasa a mesa de trabajo para presentarse en la siguiente acta, importante verificar que están vinculados, y con pagos aplicados.

Duración de la operación es de 45 días hábiles, una vez sale el acta del sistema.

6.3.5.2. Soportes.

Se debe tener en cuenta los soportes según fiduciaria.

Sin Arras (Se Le Devuelven Los Recursos Al Cliente).

- Estado de cuenta
- Soporte de consignación para venta no ingresada a SIPRO
- Cedula de ciudadanía para venta no ingresada a SIPRO
- Solicitud de desistimiento por parte de la cliente firmada
- Certificación bancaria (Que la cuenta este activa y pertenezca al comprador)
- Soporte Vinculación Efectiva

Con Arras (No Se Le Devuelven Los Recursos Al Cliente).

- Formulario rojo Davivienda
- Comunicado donde se le informa el cobro de la sanción
- Para los casos donde el cliente es moroso, se requiere soporte de envió carta 3
- Soporte de envió del comunicado
- Certificación bancaria (Que la cuenta este activa y pertenezca al comprador)
- Soporte vinculación
- Estado de Cuenta

Fallecimiento Titular Del Negocio.

- Estado de cuenta
- Certificado de defunción
- Cedula del fallecido
- Juicio de sucesión
- Cedulas de los beneficiarios

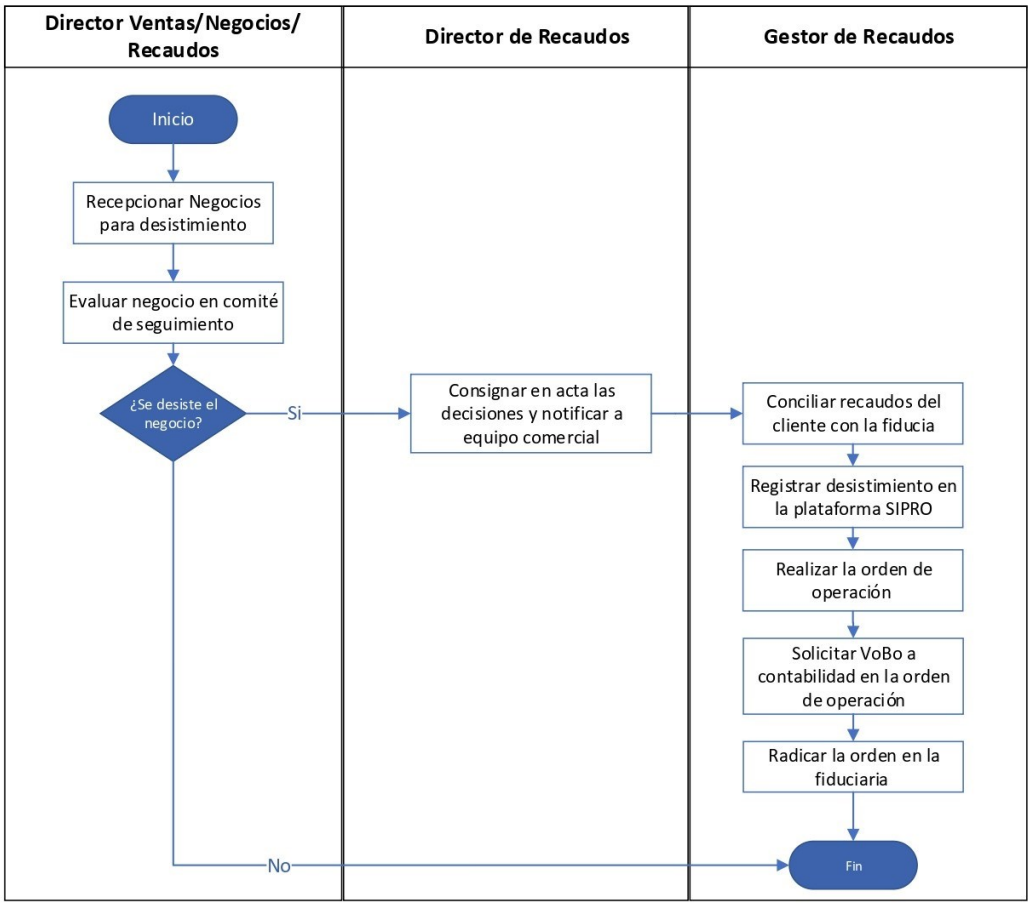
- Certificación Bancaria
- Formulario rojo
- Carta de solicitud a los beneficiaros designados en el juicio de sucesión indicando: banco, número de cuenta y tipo de cuenta, está carta de solicitud debe venir diligenciada con reconocimiento de huella.

Nota: La unidad queda liberada en sipro y la fiduciaria *si la compra fue durante la emergencia sanitaria se debe anexar soporte de correo de aceptación de vinculación del cliente.

AS IS.

Ilustración 7

AS IS desistimientos

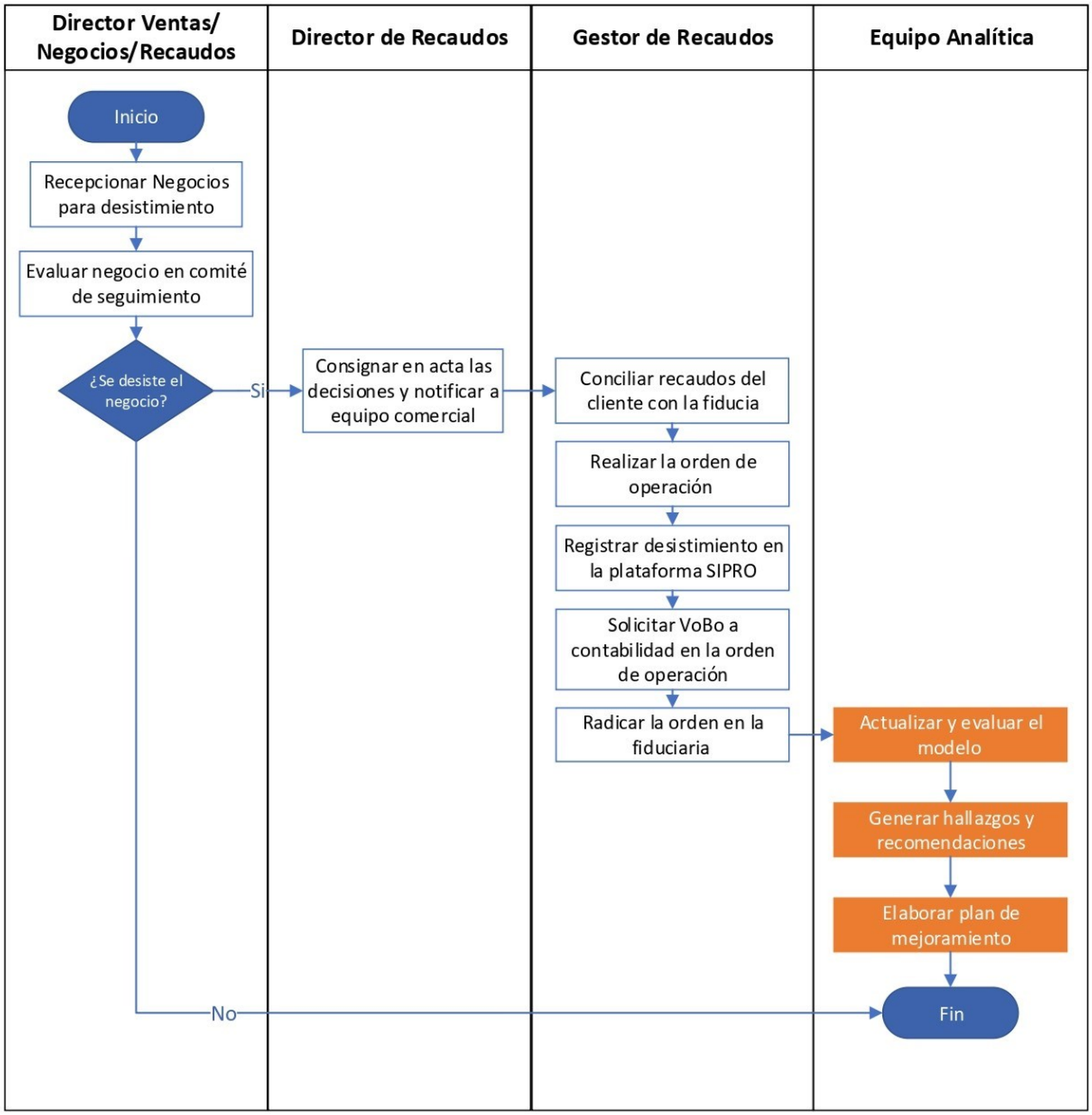


Nota: Esta ilustración muestra el esquema del proceso de desistimiento de un inmueble en la compañía de estudio.

TO BE.

Ilustración 8

TO BE desistimientos



Nota: Esta ilustración muestra el esquema del proceso de desistimiento de un inmueble en la compañía de estudio con la mejora de proceso planteada a partir del modelo.

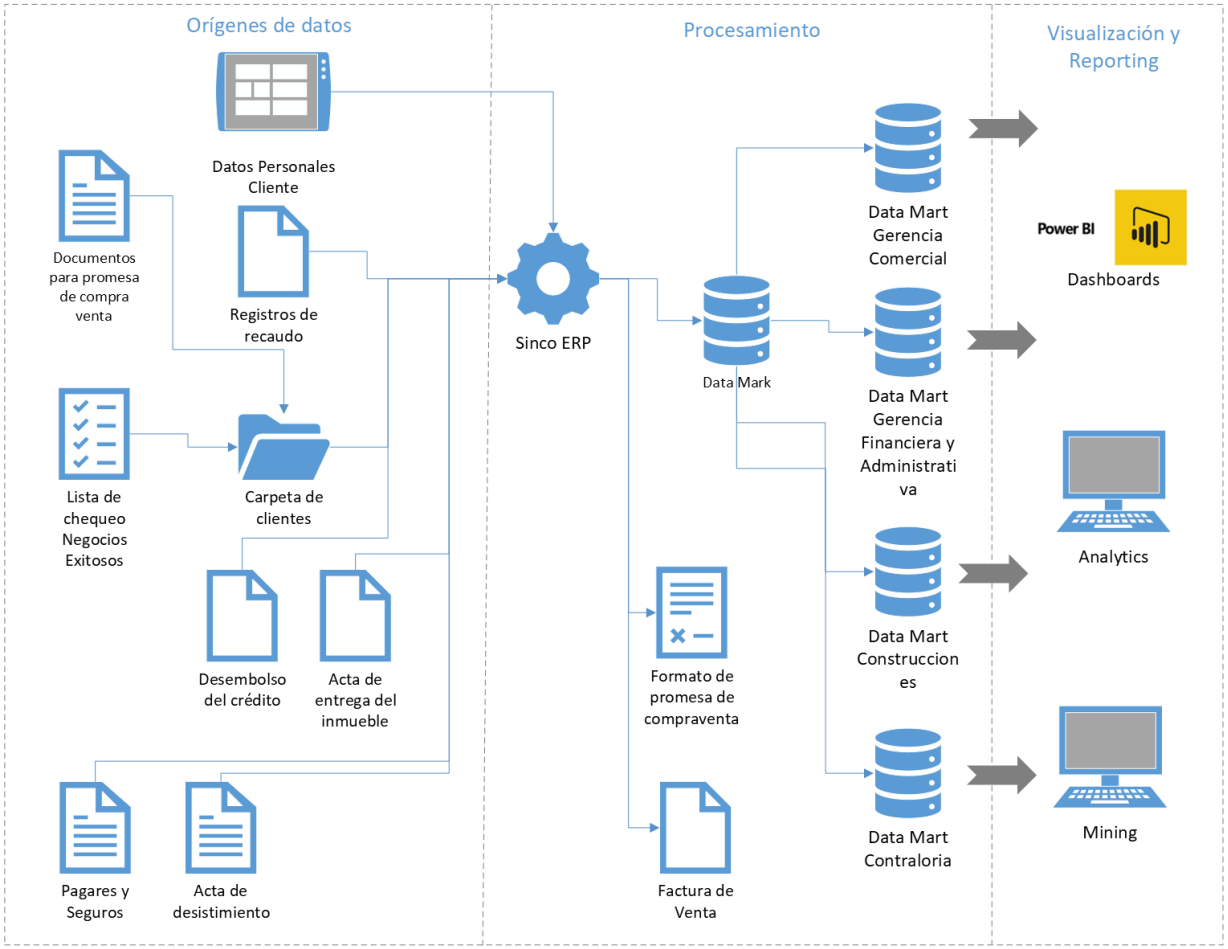
Para este proceso se plantea la actualización del modelo para emitir alertas tempranas, elaborar planes de acción y evaluar la política actual para el equipo de ventas.

6.3.6. Arquitectura de Datos

Teniendo en cuenta la definición de los procesos TO BE, mostramos a continuación la administración de los recursos de información en un estado ideal en la organización.

Ilustración 9

Arquitectura de Datos



Nota: Esta ilustración muestra el esquema del proceso de datos en la compañía desde el origen hasta la visualización.

6.3.7. Objetivo De Minería De Datos

6.3.8. Formulación Del Problema De Negocio

Un alto nivel de desistimientos generado por la falta de herramientas de analítica para el conocimiento de los clientes y solicitantes tiene un alto impacto negativo en los resultados financieros de los proyectos de vivienda afectando el recaudo, generando reprocesos administrativos y gastos operacionales.

6.3.9. Objetivo Analítico

A partir de la información comercial de la compañía, identificar los niveles de desistimiento a través de modelos analíticos que determinen la probabilidad de éxito de un negocio de acuerdo con sus características evitando afectaciones en el desempeño financiero y procesos administrativos.

6.3.10. Diseño De La Solución Y Plan De Trabajo Del Proyecto

Partiendo del análisis de los factores internos y externos junto con los objetivos organizacionales anteriormente expuestos, para el desarrollo de la especificación de requisitos de la empresa en estudio, se determina con el equipo los siguientes requerimientos:

Tabla 2
Historias de usuario

Requerimientos	Criterios de aceptación
Estimación de la tasa de éxito por segmentos.	• La tasa de éxito por área construida, ciudad, entidad de subsidio, valor del subsidio, entidad de crédito, valor del crédito, tipo de negocio y medio publicitario.

Clasificación de los motivos de desistimiento de los clientes.	• Motivos deben pasar de comentarios por parte del cliente a una variable categorizada.
Visualización de la información para análisis univariable con el fin de encontrar en primera instancia las variables que pueden aportar información significativa a los modelos y excluir aquellas que no generan diferencias claras en las poblaciones evaluadas.	• La segmentación de la información para análisis debe realizarse por macroproyecto y proyecto. • Debe incorporar como mínimo la información demográfica de los negocios que finalizan en escrituración. • Las variables para usar en el modelo serán aquellas que muestren la mayor correlación con la variable objetivo (Análisis de componentes principales).
Determinación de modelos que describan y clasifiquen los datos para la identificación de negocios con alta probabilidad de éxito.	• Generar árboles de clasificación Gradient Boosting, Random Forest, SVM y reglas del modelo para la efectividad de las variables objetivo. • Establecer una política de aprobación y/o vinculación asociada a la propensión de éxito.

Nota: Esta tabla muestra las historias de usuario y criterios de aceptación para cada una.

Tabla 3

Etapas plan de trabajo inicial del proyecto

ETAPA	DESCRIPCIÓN
ETAPA 1: Entendimiento del negocio y descubrimiento de los datos	
Objetivos Estratégicos-DOFA-Historias de usuario	Comprender los objetivos y requisitos del proyecto desde una perspectiva comercial, luego convertir este conocimiento en una definición del problema de minería de datos y un plan preliminar diseñado para lograr los objetivos.
Documento descriptivo de las fuentes de información	
Exploración inicial de los datos	
Análisis de la tasa de éxito	
ETAPA 2: Validación de los datos	
Bases de datos actualizadas	Revisar los datos disponibles para la minería. Implica acceder a los datos y explorarlos para determinar la calidad

	de los datos. (“Conceptos básicos sobre comprensión de datos - IBM”)
Selección de los datos	Realizar la preparación de los datos y posterior a ello su validación garantizando que el modelado se realice con los datos correctos.
Limpieza de datos (estimación de datos faltantes, duplicados, outliers)	
ETAPA 3: Análisis-Técnicas de modelamiento	
Actualización y Limpieza de datos (estimación de datos faltantes, duplicados, outliers)	Agregar datos nuevos o que no se han tenido en cuenta en etapas anteriores.
Determinación de las técnicas de modelamiento	La determinación del modelo más apropiado generalmente se basará en las siguientes consideraciones: Los tipos de datos disponibles para la minería, los objetivos de minería de datos y requisitos específicos de modelado.
Análisis de componentes principales	Descripción de las relaciones entre las variables. Las filas corresponden a los individuos y las columnas variables de tipo continuo.
Análisis de correspondencia simple	Comparación de los perfiles de fila y columna de dos variables categóricas.
ETAPA 4: Análisis-árboles de clasificación y regresión logística	
Análisis Gradient Boosting	Ejecutar varios modelos utilizando los parámetros predeterminados y preparar datos para las manipulaciones requeridas por su modelo de elección.
Análisis Random Forest	
Análisis SVM	
Análisis XGBoost	
Análisis CatBoost	
Análisis LightGBM	
Validación de resultados y determinación de nuevos pasos	Evaluar los modelos utilizando los criterios de éxito empresarial.
ETAPA 5: Implementación	
Proceso de mantenimiento de la solución, lecciones aprendidas y éxitos	Puesta en marcha la solución y comunicación a la comunidad de usuarios finales y a las partes interesadas que la solución está activa.
Comunicación de hallazgos a usuarios finales	
Entrega de informe final	

Nota: Esta tabla muestra las etapas con su respectiva descripción del plan inicial de trabajo para el logro de los objetivos comerciales.

6.3.11. Determinación De Criterios Iniciales:

Selección De Las Técnicas De Modelamiento.

Para la selección de las técnicas de modelamiento, se tuvo en cuenta los tipos de datos disponibles para la minería (los cuales se describirán en los próximos capítulos) y los objetivos de minería de datos junto con los requerimientos de los usuarios finales.

Tabla 4
Selección preliminar de técnicas de modelamiento

Necesidad para el desarrollo del objetivo analítico	Modelo seleccionado
Predicción de la probabilidad de éxito (Entre 0 y 1) de la variable binaria dependiente culminación exitosa en términos de las variables cualitativas y cuantitativas relacionadas con las características del cliente y el negocio.	Árbol de clasificación Máquina de soporte vectorial Gradient Boosting Random Forest CatBoost XGBoost LightGBM
Clasificación y reconocimiento de patrones e identificar los puntos de corte de las variables para establecer perfiles.	ACP

Nota: Esta tabla muestra el modelo seleccionado de acuerdo con las necesidades para el desarrollo del objetivo analítico.

Antes de la implementación de los modelos, es importante la integración y modificación de las bases de datos. Se requiere integrar todas las bases de datos conservando el orden de las columnas y generar la culminación exitosa donde se clasificará con 1 a los negocios que terminan la compra del inmueble hasta la escrituración y 0 a aquellos que culminen no se escribió generando reprocesos administrativos.

Para todos los modelos a utilizar, se requiere la división de los datos en conjuntos de entrenamiento y prueba (70 entrenamiento - 30 prueba). La información que se tiene es

suficiente para la ejecución del modelo, sin embargo, se requiere la inclusión de información adicional de proyectos y clientes nuevos.

Generación De Modelos.

En este punto experimentaremos con los modelos mencionados en el apartado anterior, los cuales al ser comparados nos permitirán llegar a las primeras conclusiones. Se registrarán los cambios y datos utilizados para cada uno.

Configuración de parámetros: se incluye las variables a usar en el modelo serán aquellas que muestren la mayor relación con la variable objetivo. Se toma la decisión a partir de los análisis de componentes principales y correspondencias múltiples.

Descripción del modelo: se determinará a partir de cada modelo si es posible la generación de conclusiones significativas y oportunidades.

Descripciones de resultados de modelos: En caso de presentar inconvenientes en la ejecución, se retrocederá a la etapa anterior y se validará la calidad y pertinencia de los datos.

Evaluación Del Modelo.

Para la evaluación de cada modelo, se definirán métricas de desempeño. Estas medidas nos permitirán comparar los modelos y a partir de las métricas tomar decisiones sobre ellos, algunas de las medidas de desempeño que se han considerado para la evaluación son: área bajo la curva ROC-AUC (sensibilidad y especificidad) y las matrices de confusión.

A partir de estos resultados, como se mencionó, tendremos dos opciones: realizar ajustes o la selección de un modelo diferente asociado al objetivo analítico.

7. Descubrimiento De Los Datos.

7.1.Recopilación Y Descripción De Los Datos

Para el desarrollo del proyecto, se utilizarán cuatro bases de datos suministrados por la compañía:

7.1.1. Negocios No Exitosos

La tabla contiene información de negocios de noviembre de 2018 a septiembre de 2022, donde está la información demográfica e información del inmueble de todos aquellos negocios en los que hubo desistimiento del solicitante.

7.1.2. Negocios Exitosos

La tabla contiene información de negocios de enero de 2010 a diciembre de 2022, donde se encuentra la información demográfica e información del inmueble de todos aquellos negocios en los que no hubo desistimiento del solicitante.

7.1.3. Información Compradores Culminación Exitosa

La tabla contiene información demográfica de los clientes que culminaron el proceso de venta.

7.1.4. Información Compradores Culminación No Exitosa

La tabla contiene información demográfica de los clientes que no culminaron el proceso de venta.

A continuación, se describen los campos pertenecientes a cada una de las fuentes de información suministradas por parte de la compañía en estudio para la exploración de datos:

Tabla 5
Negocios no Exitosos

Campo	Definición	Tipo de variable
Macroproyecto	Nombre del macroproyecto al cual pertenece el inmueble.	Cualitativa Nominal
Proyecto	Nombre del proyecto de al cual pertenece el inmueble	Cualitativa Nominal
Id Venta	Identificador único de consecutivo de negocio.	Cualitativa Nominal
CodUnidad	Código asignado a la unidad residencial	Cualitativa Nominal
Agrupación	Número o Letra asignado de Bloque/Torre/Apartamento/Casa del inmueble	Cualitativa Nominal
Área Construida	Área privada más áreas compartidas con otros apartamentos.	Cuantitativa Continua
Área Privada	Área habitable del inmueble.	Cuantitativa Continua
Comprador Principal	Cédula del comprador principal del inmueble.	Cualitativa Nominal
Ciudad	Ciudad de residencia del comprador principal del inmueble	Cualitativa Nominal
Vendedor	Nombre del asesor comercial que acompañó la venta del inmueble.	Cualitativa Nominal
Ent Subsidio	Nombre de la entidad que desembolsa subsidio.	Cualitativa Nominal
Vr Subsidio	Dinero del subsidio desembolsado para la compra de la vivienda.	Cuantitativa Discreta
Ent Crédito	Nombre de la entidad que desembolsa el crédito.	Cualitativa Nominal
Vr Crédito	Dinero del crédito desembolsado para la compra de la vivienda.	Cuantitativa Discreta
Ent Cesantías	Nombre de la entidad que desembolsa cesantías.	Cualitativa Nominal
Vr Cesantías	Dinero de cesantías desembolsado para la compra de la vivienda.	Cuantitativa Discreta
Ent Ahorro	Nombre de la entidad que desembolsa el ahorro.	Cualitativa Nominal
Vr Ahorro	Dinero de ahorro desembolsado para la compra de la vivienda.	Cuantitativa Discreta

Campo	Definición	Tipo de variable
Ent Subsidio Concurrente	Nombre de la entidad que desembolsa el subsidio concurrente (integración de las ayudas para compra de vivienda).	Cualitativa Nominal
Vr Subsidio Concurrente	Dinero de subsidio concurrente desembolsado para la compra de la vivienda.	Cuantitativa Discreta
Fecha Subsidio Concurrente	Fecha en la que se desembolsa el subsidio concurrente.	Cuantitativa Continua
Tipo Venta	Tipo de pago con el que se efectuó la venta (Contado, Crédito, Crédito terceros, leasing)	Cualitativa Nominal
Vr Recaudo	Dinero de total desembolsado para la compra de la vivienda.	Cuantitativa Discreta
Vr Agrupación Incluida la reforma	Valor total del inmueble incluido mayores valores tales como matrices de acabados.	Cuantitativa Discreta
Vr Venta	Valor de venta de la vivienda.	Cuantitativa Discreta
Vr Arras	Dinero pactado en contrato de arras.	Cuantitativa Discreta
Fecha Venta	Fecha en que se realiza la venta del inmueble.	Cuantitativa Continua
Fecha Desistimiento	Fecha en que el solicitante desistió del negocio	Cuantitativa Continua
Motivo Desistimiento	Razón del desistimiento manifestado por el cliente para dar fin al negocio.	Cualitativa Nominal
Vr Devolución	Dinero devuelto al cliente luego del manifiesto de desistimiento por parte de este.	Cuantitativa Discreta
DOEF-C	Desistido efectivo	Cuantitativa Continua

Nota: Esta tabla muestra las variables de la base de datos de negocios no exitosos, su definición y tipo. Total, de registros 3247.

Tabla 6
Negocios Exitosos

Campo	Definición	Tipo de variable
Macroproyecto	Nombre del macroproyecto al cual pertenece el inmueble.	Cualitativa Nominal
Proyecto	Nombre del proyecto de al cual pertenece el inmueble	Cualitativa Nominal
Id Venta	Identificador único de consecutivo de negocio	Cualitativa Nominal
CodUnidad	Código asignado a la unidad residencial	Cualitativa Nominal
Agrupación	Número o Letra asignado de Bloque/Torre/Apartamento/Casa del inmueble	Cualitativa Nominal
Área Construida	Área privada más áreas compartidas con otros apartamentos.	Cuantitativa Continua
Área Privada	Área habitable del inmueble.	Cuantitativa Continua
Comprador Principal	Cédula del comprador principal del inmueble.	Cualitativa Nominal
Ciudad	Ciudad de residencia del comprador principal del inmueble	Cualitativa Nominal
Vendedor	Nombre del asesor comercial que acompañó la venta del inmueble.	Cualitativa Nominal
Ent Subsidio	Nombre de la entidad que desembolsa subsidio.	Cualitativa Nominal
Vr Subsidio	Dinero del subsidio desembolsado para la compra de la vivienda.	Cuantitativa Discreta
Ent Crédito	Nombre de la entidad que desembolsa el crédito.	Cualitativa Nominal
Vr Crédito	Dinero del crédito desembolsado para la compra de la vivienda.	Cuantitativa Discreta
Ent Cesantías	Nombre de la entidad que desembolsa cesantías.	Cualitativa Nominal
Vr Cesantías	Dinero de cesantías desembolsado para la compra de la vivienda.	Cuantitativa Discreta
Ent Ahorro	Nombre de la entidad que desembolsa el ahorro.	Cualitativa Nominal
Vr Ahorro	Dinero de ahorro desembolsado para la compra de la vivienda.	Cuantitativa Discreta

Campo	Definición	Tipo de variable
Ent Subsidio Concurrente	Nombre de la entidad que desembolsa el subsidio concurrente (integración de las ayudas para compra de vivienda).	Cualitativa Nominal
Vr Subsidio Concurrente	Dinero de subsidio concurrente desembolsado para la compra de la vivienda.	Cuantitativa Discreta
Tipo Venta	Tipo de pago con el que se efectuó la venta (Contado, Crédito, Crédito terceros, leasing)	Cualitativa Nominal
Vr Recaudo	Dinero de total desembolsado para la compra de la vivienda.	Cuantitativa Discreta
Vr Agrupación Incluida la reforma	Valor total del inmueble incluido mayores valores tales como matrices de acabados.	Cuantitativa Discreta
Fecha de separación	Fecha en la que se realiza separación del inmueble.	Cuantitativa Continua
ODSE-C	Fecha en la que se realizó la orden de separación.	Cuantitativa Continua

Nota: Esta tabla muestra las variables de la base de datos de negocios exitosos, su definición y tipo. Total, de registros 27276.

Como se puede observar, el esquema de las bases de datos correspondientes a negocios exitosos y no exitosos es prácticamente el mismo, lo que facilitará la integración y entendimiento de la información.

Para la información de compradores con culminación exitosa y no exitosa, los campos suministrados para cada una de las tablas son los mismos que explicamos a continuación:

Tabla 7
Información compradores culminación exitosa y no exitosa

Campo	Definición	Tipo de variable
Id Comprador	Identificador único de consecutivo de cliente	Cualitativa Nominal
Doc. Comprador	ID del comprador principal del inmueble.	Cualitativa Nominal
Comprador Ciudad Residencia	Ciudad de residencia del comprador principal del inmueble.	Cualitativa Nominal

Campo	Definición	Tipo de variable
Comprador País Residencia	País de residencia del comprador principal del inmueble.	Cualitativa Nominal
Comprador Personas Cargo	Número de personas que dependen económicamente del comprador.	Cuantitativa Discreta
Comprador Ingresos Mensuales	Ingresos regulares recibidos del comprador de manera mensual.	Cuantitativa Discreta
Comprador Salario	Salario mensual del comprador principal.	Cuantitativa Discreta
Comprador Estado Civil	Estado civil del comprador.	Cualitativa Nominal
Comprador Fecha Nacimiento	Fecha de nacimiento del comprador principal	Cuantitativa Continua
CompradorEdad	Edad del comprador inicial	Cuantitativa Discreta
CompradorNumeroHijos	Número de hijos del comprador.	Cuantitativa Discreta
Comprador Ocupación	Tipo de trabajo que desarrolla el comprador.	Cualitativa Nominal
Comprador Cargo	Si es empleado, cargo que desempeña en la compañía para la cual trabaja el comprador.	Cualitativa Nominal
Comprador Profesión	Actividad laboral que realiza el comprador principal	Cualitativa Nominal
Comprador Nivel Académico	Título académico reciente del comprador principal.	Cualitativa Ordinal
Comprador Tipo Vivienda	Tipo de vivienda en la que vive actualmente.	Cualitativa Nominal
Comprador Empleador	Si es empleado, nombre del empleador.	Cualitativa Nominal
Comprador Entidad Caja Compensación	Nombre de la entidad de caja de compensación a la que pertenece.	Cualitativa Nominal
Comprador Valor Caja Compensación	Dinero declarado por parte del cliente de ahorro en su caja de compensación que destinará para la compra de la vivienda.	Cuantitativa Discreta
Comprador Tiempo Permanencia Vivienda	Tiempo en años de permanencia del comprador en su último inmueble.	Cuantitativa Continua
Comprador Ciudad Oficina	Ciudad en la que se encuentra su oficina.	Cualitativa Nominal
Comprador Tipo Contrato	Tipo de contrato del comprador principal.	Cualitativa Nominal

Campo	Definición	Tipo de variable
Comprador Tipo Negocio	Tipo de negocio en el que trabaja actualmente el comprador principal.	Cualitativa Nominal
Comprador Tiempo Actividad	Tiempo en años de ejecución de la actividad económica del comprador inicial.	Cuantitativa Continua

Nota: Esta tabla muestra las variables de la base de datos Información compradores culminación exitosa y no exitosa, su definición y tipo. Total de registros compradores culminación exitosa 25.957 y 8.251 registros para compradores culminación no exitosa.

7.2. Exploración de los datos

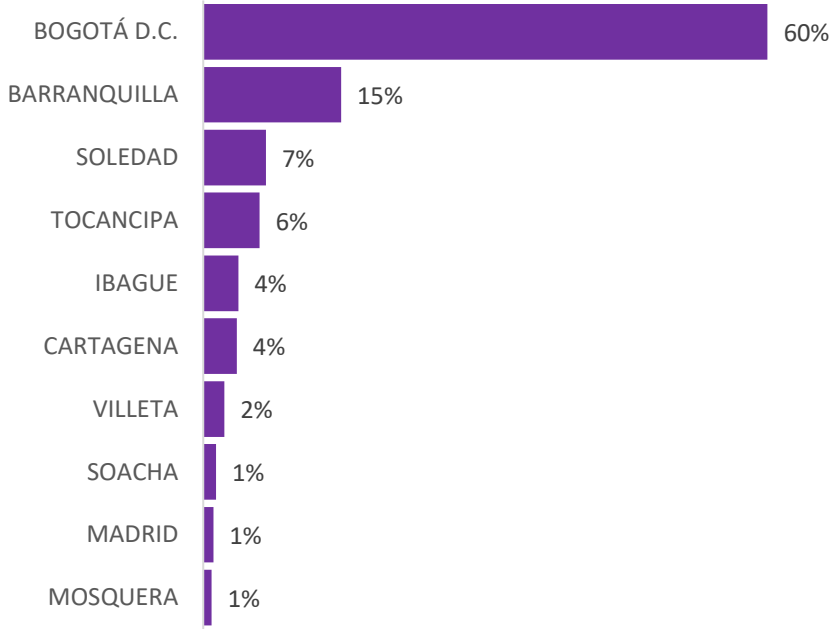
Se hace una revisión detallada de la información y se analizan cada variable en detalle para luego validar las primeras relaciones entre las diferentes variables suministradas.

7.2.1. Data Review Clientes Con Culminación Exitosa

De la base de datos "Información adicional negocios exitosos, se toma la información demográfica de los clientes que culminan el proceso de venta con la empresa. Se realiza la clasificación bajo tres criterios: localización, entorno familiar, personal y económico.

Ilustración 10

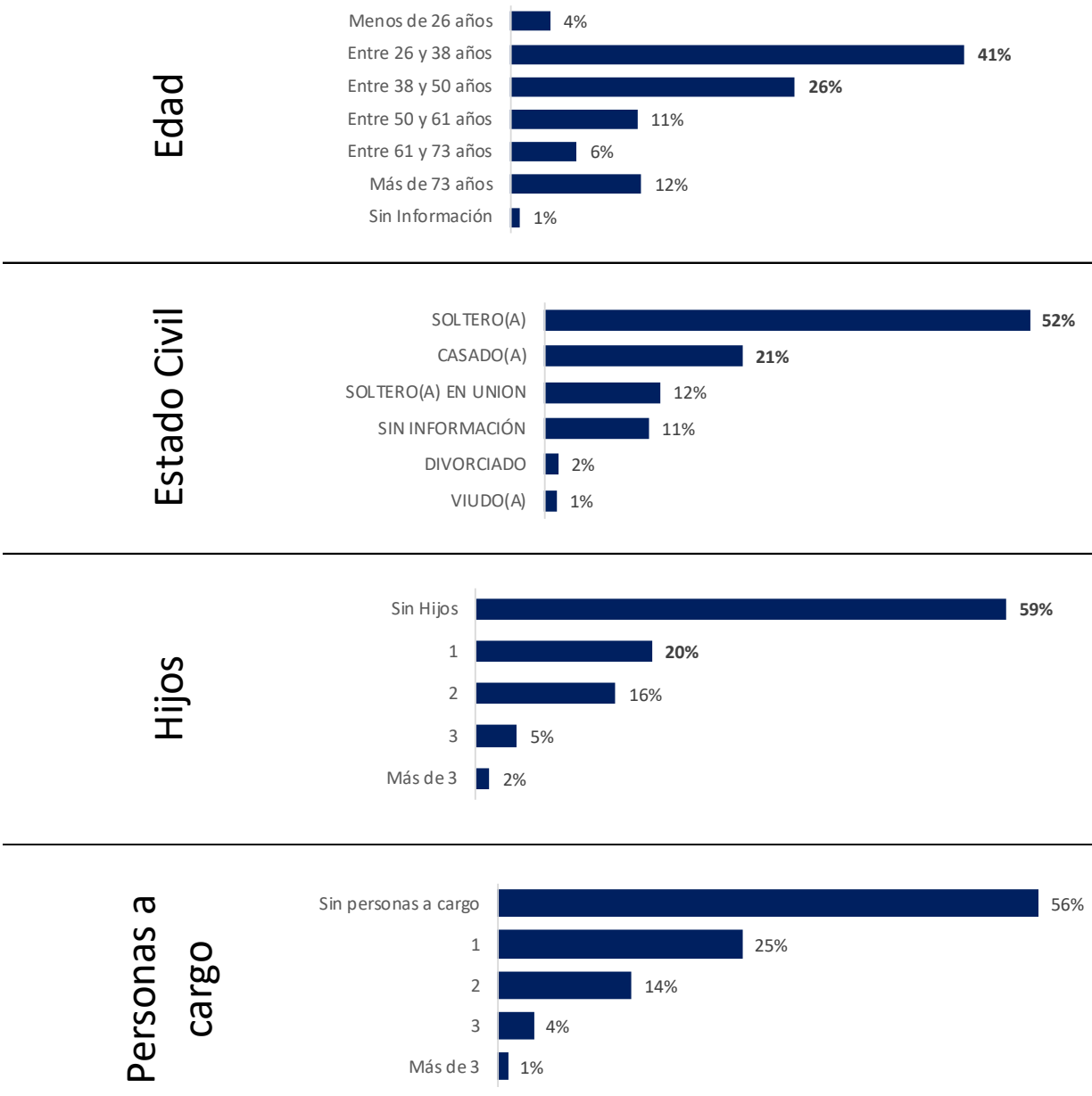
Características del cliente con culminación de negocio exitoso. Variable de clasificación: Localización.



Nota: El gráfico representa por ciudad/municipio el porcentaje de clientes residentes ordenadas de mayor a mejor.

Ilustración 11

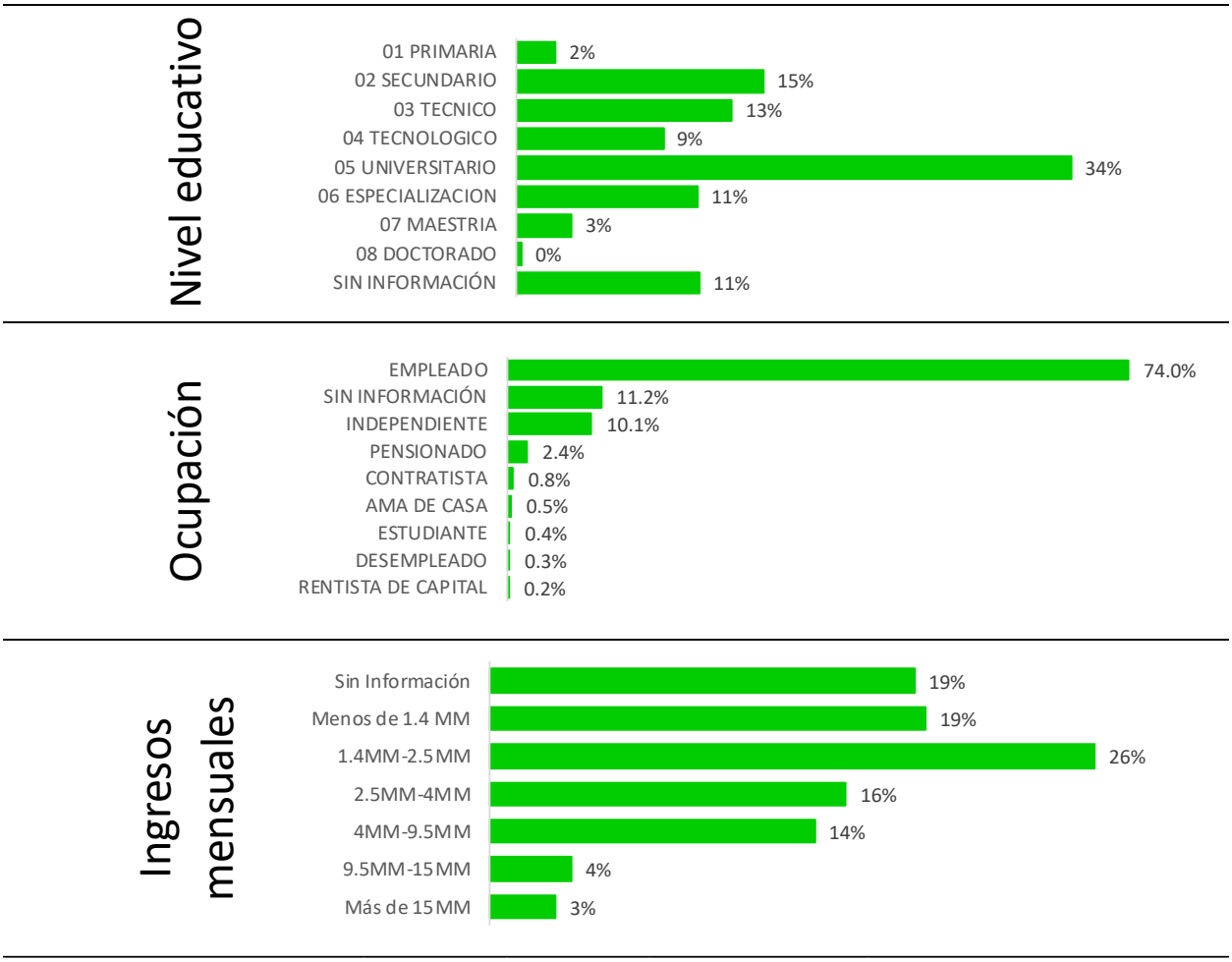
Características del cliente con culminación de negocio exitoso. Variable de clasificación: Entorno familiar y personal.



Nota: Los gráficos están relacionados con aspectos familiares y personales del cliente como la edad, el estado civil, hijos y personas a cargo.

Ilustración 12

Características del cliente con culminación de negocio exitoso. Variable de clasificación: Entorno Económico.



Nota: Los gráficos están relacionados con aspectos económicos del cliente como el nivel educativo, la ocupación y los ingresos mensuales.

Según el resultado de los gráficos, observamos que los clientes de la empresa están principalmente en Bogotá, Barranquilla y municipios aledaños a estas dos ciudades.

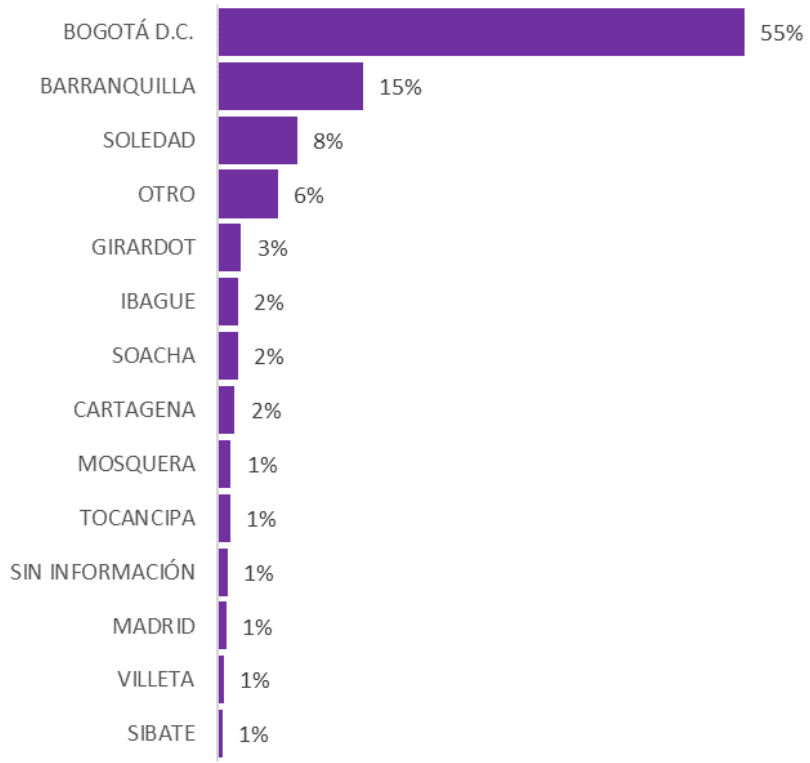
En cuanto a las características personales y familiares, la población entre los 26 y 38 años, solteros, sin hijos o personas a cargo son los principales clientes de negocios exitosos.

En cuanto al entorno económico, se caracterizan por culminar sus estudios universitarios, técnicos y secundarios, empleados con ingresos promedio entre 1.4 y 2.5 millones de pesos.

7.2.2. Data Review Clientes Con Culminación No Exitosa

Ilustración 13

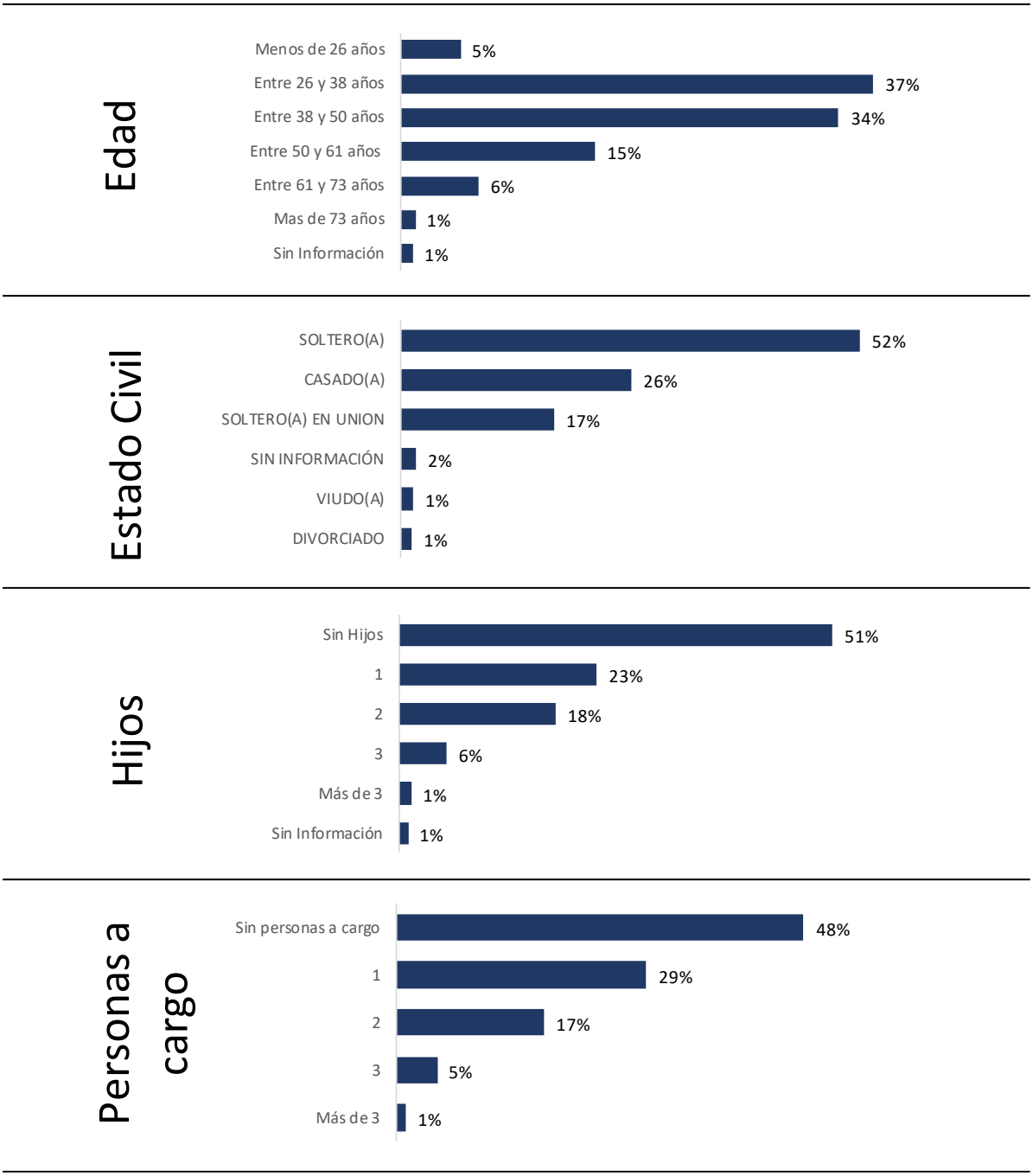
Características del cliente con culminación de negocio no exitoso. Variable de clasificación: Localización.



Nota: El gráfico representa por ciudad/municipio el porcentaje de clientes residentes ordenadas de mayor a mejor.

Ilustración 14.

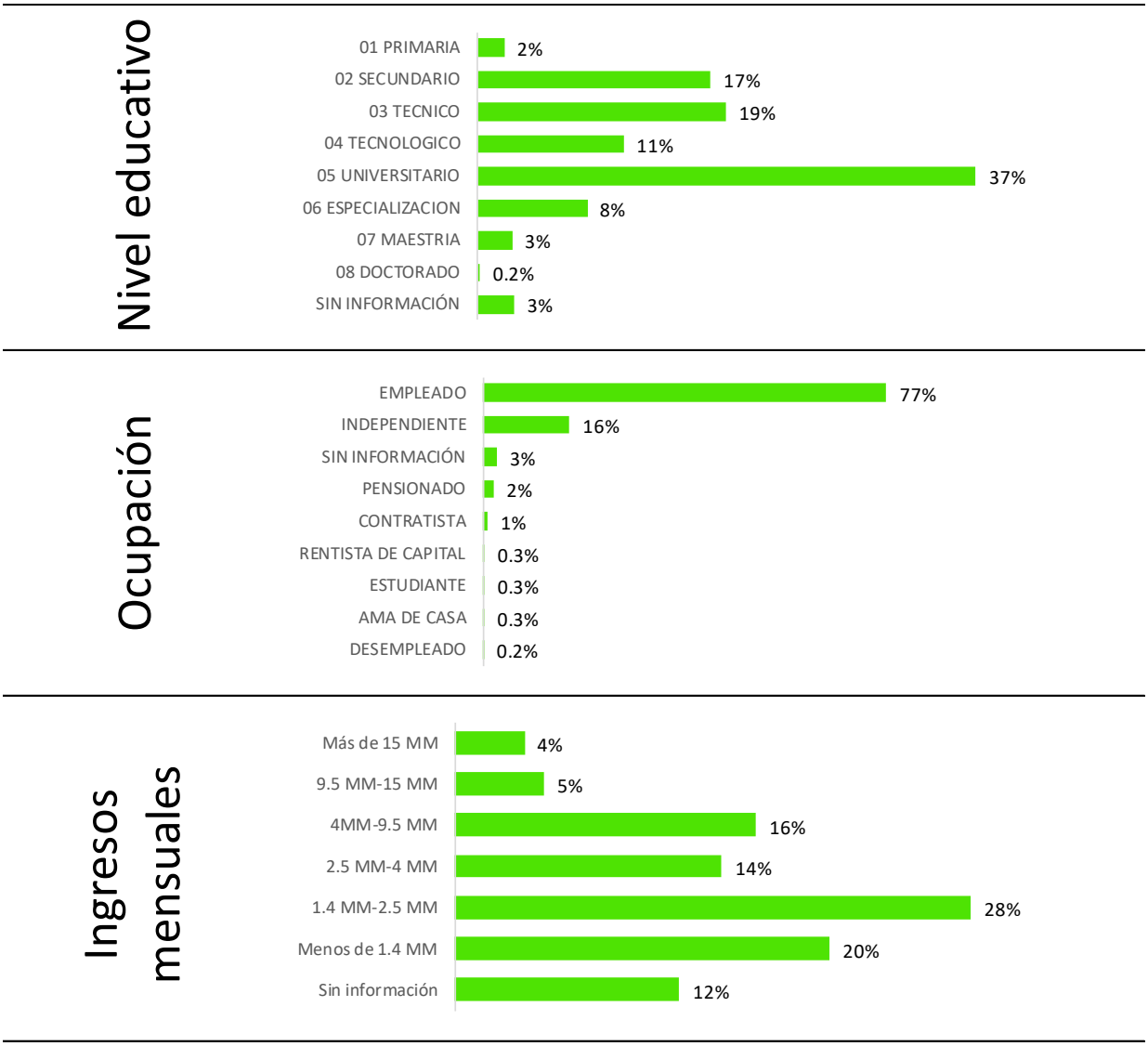
Características del cliente con culminación de negocio no exitoso. Variable de clasificación: Entorno familiar y personal.



Nota: Los gráficos están relacionados con aspectos familiares y personales del cliente como la edad, el estado civil, hijos y personas a cargo.

Ilustración 15

Características del cliente con culminación de negocio no exitoso. Variable de clasificación: Entorno Económico.



Nota: Los gráficos están relacionados con aspectos económicos del cliente como el nivel educativo, la ocupación y los ingresos mensuales.

Según el resultado de los gráficos, observamos que los clientes de la empresa están principalmente en Bogotá, Barranquilla y el municipio de Soledad. A diferencia de los clientes con culminación exitosa podemos observar que aparecen otros municipios como Girardot y Sibaté en Cundinamarca.

En cuanto a las características personales y familiares, la población entre los 26 y 38 años, solteros, sin hijos o personas a cargo son los principales clientes de negocios no exitosos. En comparación con el comportamiento de clientes que culminan el negocio aumenta la participación de clientes entre los 38 y 50 años, al igual que el número de personas a cargo entre 1 y 2.

En cuanto al entorno económico, presentan comportamiento muy similar a los clientes que culminan el negocio, se caracterizan por tener un título universitario, técnicos y secundarios y ser empleados con ingresos promedio entre 1.4 y 2.5 millones de pesos.

7.2.3. Identificación De Las Razones De Desistimiento:

La compañía cuenta con información relacionada a las razones de desistimiento del negocio suministrada por el solicitante, sin embargo, la información se almacenaba en el campo "Motivo de desistimiento" de la tabla 1. Negocios No Exitosos con el comentario textual del cliente captado por el asesor de ventas. Teniendo en cuenta lo anterior, es necesario tomar la información histórica y categorizar las razones para facilitar un posterior análisis de estas.

A continuación, relacionamos la categorización de las razones de desistimiento identificadas de acuerdo con la información histórica:

Grupo A. Causas asociadas a cambio en las condiciones financieras del solicitante.

Cambio de condiciones laborales

Disminución de ingresos

Pérdida de empleo

Sobre endeudamiento

Grupo B. Causas asociadas al crédito.

Crédito aprobado por menor valor

Crédito negado

Crédito no radicado

Reportes data crédito y/o Asobancaria

Grupo C. Causas asociadas al subsidio

Subsidio aprobado por menor valor

Subsidio negado

Subsidio no radicado

Grupo D. Causas asociadas a razones personales del solicitante.

Divorcio

Enfermedad

Fallecimiento

Incumplimiento en tramites – cliente

Pérdida de interés en el proyecto

Traslado de ciudad/país de residencia

Traslado de proyecto

Grupo E. Causas asociadas a cartera.

Cartera morosa

Grupo F. Causas asociadas al proceso de venta.

Cliente no firma documentos

Venta mal perfilada por ingresos o reportes

Grupo G. Causas asociadas a Sarlaft.

Listas restrictivas

Grupo H. Causas asociadas a aceptación de condiciones por parte del solicitante.

No acepta condiciones/no firma ep

No acepta condiciones/no firma pcv

Resciliación de ep

Grupo I. Causas asociadas a cambios en el proyecto.

Cambio de diseño o especificaciones

Cambio de fechas

Incremento de precio

Incumplimiento en trámites-Empresa

Negocio no legalizado

Al realizar la categorización de las causas de desistimiento y se calcula el porcentaje de participación de cada causa en la base de datos de negocios no exitosos. Con el fin de encontrar el impacto sobre el total de la información, se calcula el porcentaje de participación en el total de registros (negocios exitosos y no exitosos).

Tabla 8

Causas de desistimiento

Causa de desistimiento	% Negocios No exitosos	% Total Negocios
E. Cartera morosa	19%	4%
A. Pérdida de empleo	13%	2%
A. Disminución de ingresos	13%	2%
I. Cambio de fechas	8%	2%
Crédito negado	7%	1%
D. Pérdida de interés en el proyecto	6%	1%
Crédito no radicado	5%	1%
D. Traslado de proyecto	4%	1%
Negocio no legalizado	4%	1%
D. Incumplimiento en trámites - cliente	3%	1%
Cambio de diseño o especificaciones	2%	0.32%
D. Traslado de ciudad/país de residencia	2%	0.31%
Cliente no firma documentos	2%	0.31%
D. Enfermedad	2%	0.30%
D. Fallecimiento	1%	0.27%
H. No acepta condiciones/no firma pcv	1%	0.24%
Subsidio negado	1%	0.22%
I.Incumplimiento en tramites -empresa	1%	0.18%
A. Cambio de condiciones laborales	1%	0.17%
F. Venta mal perfilada por ingresos o reportes	1%	0.17%
C.Subsidio no radicado	1%	0.11%
A. Sobre endeudamiento	1%	0.10%
D. Divorcio	1%	0.10%
B. Crédito aprobado por menor valor	0.45%	0.08%
H. No acepta condiciones/no firma ep	0.31%	0.06%
C.Subsidio aprobado por menor valor	0.18%	0.03%
I.Incremento de precio	0.13%	0.02%
B. Reportes data crédito y/o Asobancaria	0.09%	0.02%
H. Resciliación de ep	0.09%	0.02%
G. Listas restrictivas	0.04%	0.01%

Nota: Esta tabla muestra las principales causas de desistimiento y el porcentaje correspondiente para los negocios no exitosos y el total de registros de negocios.

Las causales principales de desistimiento se relacionan con la financiación y el respaldo de las obligaciones del solicitante ante la constructora, encontramos aspectos como la cartera morosa, pérdida de empleo, disminución de ingresos, crédito negado y crédito no radicado.

Adicionalmente, observamos que también se encuentran otros aspectos relacionados con el seguimiento al cliente y procesos administrativos de la compañía como cambios en las fechas promesa de entregas, pérdida de interés en el proyecto, traslado del proyecto y documentación sin entrega que impide la legalización.

7.2.3. Análisis Tasa De Éxito

Considerándose un negocio exitoso, como aquel que llega a escrituración, se procedió al realizar el cálculo de la tasa de éxito de un negocio de acuerdo con los intervalos definidos previamente del área privada, dado que constituyen la promesa de valor del producto entregado y ofrecido en cada una de las cuidades donde se encuentran los proyectos, evidenciando una tasa de éxito del 79%.

Tabla 9

Tasa de éxito por Rango de Área Construida

Rango de área	% Personas	Tasa de Éxito
Entre 56 y 62 m2	10%	84%
Entre 63 y 70 m2	15%	84%
Entre 71 y 80 m2	4%	87%
Entre 81 y 100 m2	2%	83%
Más de 100 m2	3%	53%
Menos de 55 m2	53%	80%
Sin Información	13%	78%
Total general	**100%**	**79%**

Nota: Esta tabla muestra el número de personas, negocios finalizados y porcentaje de tasa de éxito por rango de área construida.

Al evaluar la variable correspondiente al área construida, se observa que el 53% de la población se concentra en áreas de menos de 55 m2, mientras que solo un 2% se encuentra en proyectos con áreas superiores a los 80 m2, estos valores se sustentan en el Core Business de la compañía, el cual, es la construcción de vivienda VIS (Vivienda de interés Social).

Tabla 10

Tasa de éxito por Ciudad

Ciudad	% Personas	Tasa de Éxito
BOGOTÁ D.C.	55%	83%
BARRANQUILLA	14%	77%
SOLEDAD	8%	66%
TOCANCIPA	5%	88%
IBAGUE	4%	77%
CARTAGENA	3%	86%
VILLETA	2%	88%
SOACHA	1%	79%
MADRID	1%	78%
MOSQUERA	1%	74%
OTRAS CIUDADES	7%	71%
Total	**100%**	**79%**

Nota: Esta tabla muestra el número de personas, negocios finalizados y porcentaje de tasa de éxito por Ciudad.

En cuanto a la distribución de vivienda en el país, según los datos, se dice que la compañía tiene presencia en 140 municipios y ciudades, de los que destaca Bogotá con un 55 % de las personas que acceden a proyectos en la compañía (Negocios Exitosos y No exitosos), seguida por Barranquilla con un 14 %.

Tabla 11

Tasa de éxito por Entidad de Subsidio

Entidad De Subsidio	% Personas	Tasa de Éxito
CAFAM	5%	84%
CAJA DE COMPENSACION FAMILIAR CAJACOPI ATLANTICO	0%	80%

Entidad De Subsidio	% Personas	Tasa de Éxito
CAJA DE COMPENSACIÓN FAMILIAR COMFAMILIAR DEL ATLA	0%	73%
CAJA DE COMPENSACIÓN FAMILIAR DE BARRANQUILLA COMB	0%	93%
CAJA DE COMPENSACION FAMILIAR DE CASANARE	0%	50%
CAJA DE COMPENSACION FAMILIAR DE LA GUAJIRA	0%	50%
CAJA PROMOTORA DE VIVIENDA MILITAR	0%	87%
COLSUBSIDIO	13%	80%
COMFACUNDI	0%	92%
COMFENALCO TOLIMA	0%	100%
COMPENSAR	10%	84%
FONVIVIENDA	22%	82%
MUNICIPIO DE TOCANCIPA	0%	80%
SIN ASIGNAR	48%	78%
Total	**7%**	**80%**

Nota: Esta tabla muestra el número de personas, negocios finalizados y porcentaje de tasa de éxito por Entidad de subsidio.

Al realizar un análisis del desembolso de subsidios, se identificó que el 48% de los subsidios no se encuentran asignados; siendo las entidades con mayor tasa de éxito en el otorgamiento Comfandi, Caja de compensación familiar de barranquilla, Compensar y Cafam.

Tabla 12

Tasa de éxito por Rango de Valor de Subsidio

Rango De Subsidio	% Personas	Tasa de Éxito
Entre 17 a 22 MM	11%	84%
Entre 22 a 25 MM	8%	77%
Entre 25 a 27 MM	8%	81%
Mayor a 27 MM	15%	78%
Menor a 17 MM	10%	83%
Sin asignar	48%	80%
Total	**100%**	**80%**

Nota: Esta tabla muestra el número de personas, negocios finalizados y porcentaje de tasa de éxito por Valor del subsidio.

Respecto al valor de los subsidios no se observa un ordenamiento claro que indique una relación directa con el éxito del negocio. El mejor desempeño se asocia con los otorgamientos que se encuentran entre 17MM y 22MM y las tasas de éxito más bajas a subsidios mayores a 27 MM.

Tabla 13

Tasa de éxito por Entidad de Crédito

Entidad Bancaria	% Personas	Tasa de Éxito
BANCO DAVIVIENDA	45%	87%
SIN ASIGNAR	15%	73%
BANCOLOMBIA S.A.	8%	73%
FONDO NACIONAL DEL AHORRO	7%	79%
BANCO CAJA SOCIAL	6%	86%
BANCO DE BOGOTA	6%	65%
CREDIFAMILIA	3%	67%
BANCO COLPATRIA	2%	89%
BBVA COLOMBIA	1%	77%
LAHIPOTECARIA DE COLOMBIA S.A.	1%	63%
BANCO AV VILLAS S.A.	1%	77%
BANCO DE OCCIDENTE	1%	86%
COLSUBSIDIO	0%	76%
CONFIAR COOPERATIVA FINANCIERA	0%	87%
BANCO POPULAR S.A.	0%	80%
BANCOOMEVA	0%	68%
COMPENSAR	0%	96%
ITAU CORPBANCA COLOMBIA S.A.	0%	88%
BANCOMPARTIR	0%	100%
OTROS BANCOS	1%	89%
Total	**100%**	**80%**

Nota: Esta tabla muestra el número de personas, negocios finalizados y porcentaje de tasa de éxito por Entidad de Crédito.

A pesar de que el principal financiador de los créditos solicitados por los clientes es Davivienda, la tasa de éxito con esta entidad es del 87% mientras que las tasas de éxito más altas se presentan con Bancompartir y Compensar.

Tabla 14

Tasa de éxito por Rango de Valor de Crédito

Rango De Crédito	% Personas	Tasa de Éxito
25MM-50MM	13%	84%
50MM-62MM	11%	81%
62MM-75MM	15%	88%
75MM-100MM	19%	85%
Más de 100MM	21%	79%
Menos de 25MM	6%	79%
Sin Información	15%	66%
Total	**100%**	**80%**

Nota: Esta tabla muestra el número de personas, negocios finalizados y porcentaje de tasa de éxito por valor del crédito.

La hipótesis de que la tasa de éxito más alta es de los montos aprobados, que oscilan entre los 62MM a 75MM, considerando que los valores financiables por los bancos según el tipo de negocio están entre un 60%-70% del valor de inmueble, sugiere que se puede referir principalmente a crédito en vivienda VIS.

Tabla 15

Tasa de éxito por Tipo de Negocio

Tipo de Venta	% Personas	Tasa de Éxito
Contado	10%	92%
Crédito	51%	80%
Crédito Terceros	38%	78%
Leasing	0%	0%
Fondos	1%	100%
Total	**100%**	**70%**

Nota: Esta tabla muestra el número de personas, negocios finalizados y porcentaje de tasa de éxito por tipo de negocio.

De acuerdo con la distribución de los tipos de ventas, se puede analizar que la mayoría de los clientes financian parte del valor de sus inmuebles con crédito directo con el banco constructor (51%) y con otros bancos (38%). Se observa que los negocios pagados de contado

tienen la mayor probabilidad de éxito (92%) y la totalidad de los negocios financiados por fondos son exitosos.

7.3. Calidad De Los Datos

Para verificar la calidad de los datos usamos como referencia la guía de fundamentos para la gestión de datos DAMA-DMBOK, donde describe las dimensiones de la calidad de los datos. Se utilizará aquellas dimensiones que apliquen al proyecto y conjunto de datos, implementando las métricas de medición que facilite la toma de decisiones para la selección definitiva de los datos.

Previo a la ejecución del ejercicio se unificaron las cuatro tablas suministradas a través de la creación de una columna llamada Id único compuesto por columnas de la base original como código de la unidad y cédula del comprador generando así una llave primaria que nos permitiera la integración de la información.

7.3.1 Completitud

La completitud hace referencia a que todas las filas del conjunto de datos tengan valores asignados[5]. Para realizar su medición se toma cada uno de los campos y se analiza cuántos registros tienen información con relación al total.

Para determinar la completitud de la tabla, se analiza si los registros de las columnas requeridas se encuentran completos con relación al total.

A continuación, mostramos el resultado obtenido:

---

[5] (DAMA internacional, 2010)

Tabla 16

Indicador de completitud total y por campo de la base de datos.

Campo	Completitud
Macroproyecto	83.4%
Proyecto	83.4%
Id Venta	83.4%
CodUnidad	83.4%
Agrupación	83.4%
Área Construida	83.4%
Área Privada	83.4%
Estado	83.4%
Ciudad	83.3%
Vendedor	**53.0%**
Ent Subsidio	83.4%
Vr Subsidio	83.4%
Ent Crédito	83.3%
Vr Crédito	83.4%
Ent Cesantías	83.3%
Vr Cesantías	83.4%
Ent Ahorro	83.4%
Vr Ahorro	83.4%
Ent Subsidio Concurrente	80.9%
Vr Subsidio Concurrente	80.9%
Tipo Venta	83.4%
Vr Recaudo	83.4%
Vr Agrupación Incluida la reforma	83.4%
Fecha Venta	83.2%
Id Comprador	99.2%
CompradorCiudadResidencia	99.0%
CompradorPaisResidencia	99.0%
CompradorPersonasCargo	99.2%
CompradorIngresosMensuales	99.2%
CompradorSalario	99.2%
CompradorEstadoCivil	93.9%
CompradorFechaNacimiento	94.2%
CompradorEdad	94.2%
CompradorNumeroHijos	95.7%
Comprador Ocupación	92.9%
Comprador Cargo	75.9%

Comprador Profesión	87.0%
Comprador Nivel Académico	92.9%
Comprador Tipo Vivienda	93.8%
Comprador Empleador	77.0%
Comprador Entidad Caja Compensación	**15.7%**
Comprador Valor Caja Compensación	95.9%
Comprador Tiempo Permanencia Vivienda	92.7%
Comprador Ciudad Oficina	85.2%
Comprador Tipo Contrato	64.4%
Comprador Tipo Negocio	**14.1%**
Comprador Tiempo Actividad	**12.7%**
Total	0.1%

Nota: Esta tabla muestra el porcentaje de completitud por campo de la base de datos.

Los resultados más bajos lo presentan los campos relacionados con el nombre del vendedor que realizó la venta, y aspectos relacionados con información del cliente, como la caja de compensación a la que pertenece, el tipo de actividad económica en la que trabaja y el tiempo que lleva ejerciendo la actividad.

En vista de lo anterior, es necesario tener presente al limpiar retirar estas columnas porque afectan los resultados de completitud general de la base de datos y no aportan información relevante en el estudio. Al retirar los campos mencionados anteriormente mejora el comportamiento del indicador pasando de un 0.1% a 48%.

7.3.2 Validez

La validez hace referencia a que los registros cumplan con los criterios de formato, rango y tipo.

Para ello, se deben definir los requerimientos del campo y posteriormente validar los registros que cumplen con la condición con relación a los registros totales de la tabla.

Es importante aclarar que en el sistema de registro de información la mayoría de los campos se encuentran estandarizadas las opciones de respuesta para evitar errores de este tipo, por ende, en la mayoría de campos no se presentan inconvenientes de formato excepto el campo de tiempo de permanencia en la vivienda debido a que es un campo no estandarizado donde se encuentra que en el 95% de los campos hay respuestas tipo texto donde se ingresa la información en meses cuando la regla indica que debe indicar la información en años.

Reglas

a. ID Venta, Código de la unidad, Áreas (construida y privada), Valores en dinero (Subsidio, Crédito, Cesantías, Ahorros, Valor del recaudo, Valor de la agrupación incluida la reforma, Ingreso mensual, Salario), Id Comprador, Personas a cargo, Número de hijos y tiempo de permanencia en años únicamente en formato numérico.
b. Fecha de venta y Fecha de nacimiento del comprador en formato de fecha dd/mm/aa

Tabla 17
Indicador de validez

Campos	Validez
ID Venta	100%
CodUnidad	100%
Área Construida	100%
Área Privada	100%
Vr Subsidio	100%
Vr Crédito	100%
Vr Cesantías	100%
Vr Ahorro	100%
Vr Subsidio Concurrente	100%
Vr Recaudo	100%
Vr Agrupación Incluida la reforma	100%
Fecha Venta	100%
Id Comprador	100%
Comprador Personas Cargo	100%
Comprador Ingresos Mensuales	100%
Comprador Fecha Nacimiento	100%

Comprador Salario	100%
Comprador Numero Hijos	99.99%
Comprador Tiempo Permanencia Vivienda	**5%**

Nota: Esta tabla muestra el porcentaje de validez por campo de la base de datos.

7.3.3 Unicidad

La unicidad hace referencia a la cantidad de registros únicos en la llave primaria. Su cálculo consiste en el número de registros únicos en relación con el total de registros.

Tabla 18

Indicador de unicidad

Campos	**Unicidad**
Id_Unico	100%

Nota: Esta tabla muestra el porcentaje de unicidad del campo ID Venta en la base de datos.

Se debe validar en la base de datos la razón por la que hay un 8% de registros duplicados, aspecto que se debe considerar en la etapa de limpieza.

7.3.4 Precisión

La precisión se refiere a los registros que cumplen con la descripción de la realidad, igual que en apartados anteriores se define para cada tabla los requerimientos mínimos y se estima el nivel de cumplimiento, calculado como registros precisos en relación con los registros totales.

Reglas

a. Valor del recaudo, Valor de la agrupación incluida la reforma, Cuota inicial, Área construida y Área privada no deben tener valores en ceros o números negativos.
b. Fechas de venta inferiores al año 2000.
c. Número de hijos o personas a cargo mayores de 50.
d. Fecha de nacimiento del comprador superior al año 2005 e inferior al año 1923.
e. Edades superiores a 100 años e inferiores a 18 años.

Tabla 19
Indicador de precisión

Campos	Precisión
Área construida	94%
Área privada	81%
Vr Subsidio	100%
Vr Crédito	100%
Vr Cesantías	100%
Vr Ahorro	100%
Vr Subsidio Concurrente	100%
Vr Recaudo	97%
Vr Agrupación Incluida la reforma	97%
Fecha de Venta	100%
CompradorPersonasCargo	99.99%
CompradorIngresosMensuales	100%
CompradorSalario	100%
CompradorFechaNacimiento	99%
CompradorEdad	99%
CompradorNumeroHijos	100%
Comprador Tiempo Permanencia Vivienda	99.3%

Nota: Esta tabla muestra el indicador de precisión por campo en la base de datos.

Campos de ingreso manual no se incluyeron en el análisis de precisión porque los campos no tienen reglas para el ingreso de la información. Para la etapa siguiente de limpieza es importante entender que al no estar estandarizado y ser de ingreso manual se presentan errores de digitación aumentando la cantidad de registros únicos que pueden afectar el modelamiento.

8. Validación De Los Datos

8.1. Selección De Los Datos

La selección de los datos se da en cuenta la calidad y la utilidad de los campos de cada base de datos considerando el análisis del negocio.

8.1.1. Base De Datos Negocios Exitosos:

Tabla 20
Selección de datos

Campo	Estado	Justificación
Macroproyecto	Activo	Campo significativo para el análisis dado que indica el proyecto macro en el cual el cliente adquirido el inmueble
Proyecto	Activo	Campo significativo para el análisis dado que indica el proyecto en el cual el cliente adquiere el inmueble
Id Venta	Inactivo	Campo no significativo para el análisis de la información.
CodUnidad	Inactivo	Campo no significativo para el análisis dado que indica la nomenclatura del inmueble, sin embargo, se utilizará para la creación de otro campo que indique el tipo de vivienda (casa o apartamento)
Agrupación	Inactivo	Campo no significativo para el análisis dado que indica la nomenclatura del inmueble, sin embargo, se utilizará para la creación de otros campos que indiquen el bloque/torre y número asignado de apartamento.
Área Construida	Activo	Campo significativo para el análisis dado que indica el área construida del inmueble y permite cuantificar rangos.
Área Privada	Inactivo	Campo significativo para el análisis dado que se utilizará el campo área construida.
Comprador Principal Cédula	Inactivo	Campo no significativo para el análisis de la información.
Ciudad	Activo	Campo significativo para el análisis dado que presenta la ubicación del proyecto.

Vendedor	Inactivo	Se considera un dato importante que puede impactar el desistimiento, sin embargo, sólo tiene el 53% de completitud y es un campo que no se encuentra estandarizado
Ent Subsidio	Activo	Campo significativo para el análisis dado que presenta parte del costo del negocio necesaria para realizar la evaluación del cierre financiero y puede impactar de forma directa la tasa de desistimiento
Vr Subsidio	Activo	Campo significativo para el análisis dado que presenta parte del costo del negocio necesaria para realizar la evaluación del cierre financiero y puede impactar de forma directa la tasa de desistimiento
Ent Crédito	Activo	Campo significativo para el análisis dado que presenta parte del costo del negocio necesaria para realizar la evaluación del cierre financiero y puede impactar de forma directa la tasa de desistimiento
Vr Crédito	Activo	Campo significativo para el análisis dado que presenta parte del costo del negocio necesaria para realizar la evaluación del cierre financiero y puede impactar de forma directa la tasa de desistimiento
Ent Cesantías	Activo	Campo significativo para el análisis dado que presenta parte del costo del negocio necesaria para realizar la evaluación del cierre financiero y puede impactar de forma directa la tasa de desistimiento
Vr Cesantías	Activo	Campo significativo para el análisis dado que presenta parte del costo del negocio necesaria para realizar la evaluación del cierre financiero y puede impactar de forma directa la tasa de desistimiento
Ent Ahorro	Activo	Campo significativo para el análisis dado que presenta parte del costo del negocio necesaria para realizar la evaluación del cierre financiero y puede impactar de forma directa la tasa de desistimiento
Vr Ahorro	Activo	Campo significativo para el análisis dado que presenta parte del costo del negocio necesaria para realizar la evaluación del cierre financiero y puede impactar de forma directa la tasa de desistimiento
Ent Subsidio Concurrente	Activo	Campo significativo para el análisis dado que presenta parte del costo del negocio necesaria para realizar la evaluación del cierre financiero y puede impactar de forma directa la tasa de desistimiento

Vr Subsidio Concurrente	Activo	Campo significativo para el análisis dado que presenta parte del costo del negocio necesaria para realizar la evaluación del cierre financiero y puede impactar de forma directa la tasa de desistimiento
Tipo Venta	Activo	Campo significativo para el análisis dado que presenta la información del tipo de negocio pactado con el cliente (contado, leasing, crédito, crédito tercero u fondos.
Vr Recaudo	Inactivo	Campo no significativo para el análisis dado que presenta la información del valor total pago por el cliente para el inmueble.
Vr Agrupación Incluida la reforma	Activo	Campo significativo para el análisis dado que presenta la información del precio total pactado para el negocio.
Fecha Venta	Activo	Campo significativo para el análisis dado que presenta la información acerca de la fecha de venta del inmueble.
Id Comprador	Inactivo	Campo no significativo para el análisis dado que cumple la misma función de Id Venta, y éste último cuenta con mejor calidad.
CompradorCiudadResidencia	Activo	Campo significativo para el análisis dado que presenta la localización del cliente.
CompradorPaisResidencia	Inactivo	Campo no significativo para el análisis dado que presenta errores de validez no corresponde la ciudad con el país específicamente en países del extranjero.
CompradorPersonasCargo	Activo	Campo significativo para el análisis dado que presenta la cantidad de personas a cargo por parte del cliente.
CompradorSalario	Activo	Campo significativo para el análisis sin embargo requiere transformación debido a que trae la misma información del campo ingresos mensuales, es necesario sumar los valores de las dos columnas para generar un registro único.
CompradorIngresosMensuales	Inactivo	Campo no significativo para el análisis dado que se duplica con el campo salario.
CompradorEstadoCivil	Activo	Campo significativo para el análisis dado que presenta el estado civil del comprador principal.
CompradorFechaNacimiento	Inactivo	Campo no significativo para el análisis dado que se tiene el campo edad.
CompradorEdad	Activo	Campo significativo, refleja la edad del comprador principal.

CompradorNumeroHijos	Activo	Campo significativo para el análisis dado que presenta el número de hijos del comprador principal.
Comprador Ocupación	Activo	Campo significativo para el análisis dado que presenta la ocupación del comprador principal.
Comprador Cargo	Inactivo	Campo no significativo para el análisis dado que no se encuentra estandarizado generando un alto porcentaje de invalidez por errores en la digitalización.
Comprador Profesión	Inactivo	Campo no significativo para el análisis dado que no se encuentra estandarizado generando un alto porcentaje de invalidez por errores en la digitalización.
Comprador Nivel Académico	Activo	Campo significativo para el análisis dado que presenta el nivel académico del comprador principal.
Comprador Tipo Vivienda	Activo	Campo significativo para el análisis dado que presenta el tipo de vivienda en que se encuentra el comprador principal.
Comprador Empleador	Inactivo	Campo no significativo para el análisis dado que no se encuentra estandarizado generando un alto porcentaje de invalidez por errores en la digitalización.
Comprador Entidad Caja Compensación	Inactivo	Campo no cuenta con la calidad requerida para ser utilizado en el análisis.
Comprador Valor Caja Compensación	Inactivo	Campo no cuenta con la calidad requerida para ser utilizado en el análisis.
Comprador Tiempo Permanencia Vivienda	Inactivo	Campo no significativo para el análisis dado que presenta el tiempo en años de permanencia en la vivienda en que se encuentra el comprador principal.
Comprador Ciudad Oficina	Inactivo	Campo no se considera relevante en el análisis, se encuentra la información en el campo Comprador ciudad de residencia.
Comprador Tipo Contrato	Activo	Campo significativo para el análisis dado que presenta el tiempo en años de permanencia en la vivienda en que se encuentra el comprador principal.
Comprador Tipo Negocio	Inactivo	Campo no cuenta con la calidad requerida para inclusión en el modelo.

Comprador Tiempo Actividad	Inactivo	Campo no cuenta con la calidad requerida para inclusión en el modelo.

Nota: Esta tabla muestra los campos fueron eliminados y aquellos que se deciden mantener en la base de datos para el análisis.

8.2. Integración Y Formateo De Los Datos

Se recibieron 4 bases de datos con 36709 registros entre negocios exitosos y no exitosos, compradores exitosos y no exitosos, que se convierten en una base única con llaves en Python y, tras eliminar filas vacías, dan un total de datos de 30122.

A continuación, se describen los pasos llevados a cabo para la unificación de las bases de datos:

a. De las bases correspondientes a negocios y clientes (Exitosos y No Exitosos) se decidió eliminar las columnas que no generan valor en la base de datos.

b. Una vez eliminadas las columnas mencionadas, se escoge la columna ID único para realizar una llave entre las bases de datos de negocio y clientes tanto para desistidos como para compradores.

c. se elimina la columna ID único y las filas vacías de las bases después de realizar las llaves (Compradores y Desistidos).

d. Se realiza un apilamiento de la base de desistidos y compradores.

Como se dijo antes, para integrar las bases de datos se utiliza la columna ID único, que surge a partir de la cédula del comprador principal y el código de la unidad. Por seguridad de la información debido a que contiene el número de cédula del cliente se eliminará para evitar exposición de los clientes de la compañía en estudio.

A continuación, se enuncian las consideraciones a priori producto del análisis de la calidad de la información antes de iniciar de la limpieza de la información:

a. Se elimina información de locales y parqueaderos debido a que presentan el valor del inmueble con valores en ceros.
b. Se unifica la columna de Ingresos y Salarios debido a que presentan la misma información.
c. Se eliminaron alrededor de 6.150 registros donde no se encontraba información del tipo de negocio como tampoco las características del proyecto adquirido.
d. Se crea una columna nueva que indique a partir de la fecha el año en que se generó la venta o desistimiento.
e. Se crea una columna nueva bajo el nombre "Estado" que identificará si el negocio corresponde al estado final del negocio desistido (no finalizado) o vendido (finalizado).
f. Se crean tres columnas adicionales con información como el tipo de inmueble (apartamento o casa), bloque o torre y número de apartamento.
g. Antes de realizar la intervención a la base de datos, se analiza la información de tasa de éxito por año, encontrando que antes del año 2020 no se realizó registro de información de negocios desistidos. Teniendo en cuenta lo anterior, se elimina de la base de datos todo negocio realizado antes del 2020 como también aquellos en los que no se tenga información de la fecha de venta o desistimiento.

Año	% Éxito	Registros
2000	100%	12
2010	100%	55
2011	100%	772
2012	100%	1713
2013	100%	2197
2014	100%	2026
2015	100%	1522

2016	100%	1969
2017	100%	2107
2018	100%	2381
2019	100%	2755
2020	94%	2557
2021	93%	3919
2022	92%	3470
Sin Información	3%	2667
Total	**89%**	**30122**

8.3. Limpieza De Los Datos

8.3.1. Limpieza De Base De Datos Negocios No Exitosos

8.3.1.1. Análisis Descriptivo.

Considerando lo anterior, obtenemos una base de datos de 29 columnas y 10226 filas, revisadas para entender el tipo de variable a la que pertenece y aplicar la correcta limpieza a la base. Los valores correspondientes en dinero, edades, personas a cargo, entre otros; son de tipo numérico (decimal [float]) y entero ([int]).

Se corrige en primera instancia el formato de aquellas variables que no se encuentran con el tipo de datos que requerimos asignar.

Ilustración 16
Análisis Descriptivo

```
Tipo                                object
Piso.1                              object
Area                                 int64
Estado                              object
Ciudad                              object
Ent Subsidio                        object
Vr Subsidio                          int64
Ent Crédito                         object
Vr Credito                           int64
Ent Cesantías                       object
Vr Cesantias                         int64
Ent Ahorro                          object
Vr Ahorro                            int64
Ent Subsidio Concurrente            object
Vr Subsidio Concurrente              int64
Tipo Venta                          object
Vr Recaudo                           int64
Vr Agrupacion Incluida la reforma    int64
Año Venta                           object
Año escritura/desistimiento         object
Tasa Interes                       float64
CompradorPersonasCargo               int64
Sueldo                               int64
CompradorEstadoCivil                object
Edad                                 int64
CompradorNumeroHijos                 int64
CompradorOcupacion                  object
CompradorNivelAcademico             object
CompradorTipoVivienda               object
dtype: object
```

Nota: Esta ilustración, muestra el nombre de la variable y el tipo.

En vista de lo anterior, se describen las variables tipo texto para entender las filas vacías, datos únicos, moda y frecuencia de cada columna.

Tabla 21

Descripción de las variables tipo texto

Variable	Conteo	Datos Únicos	Moda	Frecuencia
Tipo	10226	2	Apto	10027
Piso	10209	21	3	1742
Estado	10226	2	Vendida	9541
Ciudad	10210	199	BOGOTÁ D.C.	5327
Ent Subsidio	10226	17	Sin Asignar	4174
Ent Crédito	10224	48	BANCO DAVIVIENDA	4219
Ent Cesantías	10220	17	Sin Asignar	7087
Ent Ahorro	10224	49	Sin Asignar	9725
Ent Subsidio Concurrente	9770	4	Sin Asignar	8740
Tipo Venta	10226	7	Crédito	4864
Año Venta	10226	3	2021	4043
Año escritura/desistimiento	3838	3	2022	2267
CompradorEstadoCivil	9454	6	SOLTERO(A) SIN UNION MARITAL DE HECHO	6882
CompradorOcupacion	9456	8	EMPLEADO	8145
CompradorNivelAcademico	9455	8	05 UNIVERSITARIO	3897
CompradorTipoVivienda	9455	3	FAMILIAR	5913
CompradorTipoContrato	7442	6	04 DE APRENDIZAJE	5056

Nota: Esta tabla muestra las variables tipo texto con su respectiva moda y frecuencia.

Debido a que algunos de los datos categóricos cuentan con una cantidad significativa de datos únicos, se toma la decisión de aplicar la técnica de frecuency encoding en aquellas variables cuyos datos únicos fueran superiores a ocho con el objetivo de disminuir la dimensionalidad de los datos en la etapa de modelamiento al generar variables dummies.

Se realiza una descripción de las variables de tipo numérico, donde se mostrará: conteo de datos, media, desviación estándar, valor mínimo, cuartil 1 (25%), mediana cuartil 2 (50%), cuartil 3 (75%) y valor máximo para cada una de las variables.

Tabla 22

Descripción de las variables Numéricas

Variable	Conteo	Media	Desviación Estándar	Valor Mínimo	25%	50%	75%	Valor Máximo
Area	10226	67	22	38	55	69	71	236
Vr Subsidio	10226	$ 15,974,680	$ 13,898,368	$ -	$ -	$ 20,000,000	$ 30,000,000	$ 109,931,646
Vr Credito	10226	$ 110,418,639	$ 98,389,450	$ -	$ 57,928,999	$ 92,220,000	$ 120,000,000	$ 1,384,000,000
Vr Cesantias	10226	$ 2,405,360	$ 7,011,620	$ -	$ -	$ -	$ 1,800,000	$ 194,000,000
Vr Ahorro	10226	$ 599,549	$ 4,824,436	$ -	$ -	$ -	$ -	$ 214,465,600
Vr Subsidio Concurrente	9770	$ 2,039,960	$ 5,969,971	$ -	$ -	$ -	$ -	$ 36,750,000
Vr Recaudo	10226	$ 60,240,189	$ 103,483,008	$ -	$ 4,508,125	$ 16,154,150	$ 79,661,555	$ 1,260,000,000
Vr Agrupacion Incluida la reforma	10226	$ 185,806,422	$ 139,177,113	$ 1	$ 114,920,000	$ 151,200,000	$ 185,150,000	$ 1,730,000,000
Tasa Interes	3838	11	2	8	9	10	14	15
CompradorPersonasCargo	9963	201450	20107606	0	0	0	1	2007037224
Sueldo	9963	$ 5,481,747.79	$ 72,622,058.94	$ -	$ 1,370,000.00	$ 1,950,000.00	$ 3,900,000.00	$ 5,304,833,000.00
Edad	9426	38	11	18	29	35	44	86
CompradorNumeroHijos	9963	1	1	0	0	0	1	7

Nota: Esta tabla muestra las variables cuantitativas, con la información de la media, desviación estándar y cuartiles.

A simple vista, la información muestra que se requiere tratamiento para aquellas variables que están presentando información que no corresponde al comportamiento habitual del negocio. Aspectos como los valores faltantes y outliers pueden estar afectando el comportamiento medio de algunas de las variables, es por esto por lo que en la siguiente sección se explicará con mayor detalle el tratamiento a aquellas variables sin perder de vista el comportamiento y lógica de negocio.

8.3.1.2. Tratamiento De Datos.

8.3.1.3. Tratamiento De Datos Duplicados.

Se realiza verificación en la base de datos de datos duplicados, encontrando que en esta base de datos no se presenta duplicidad de la información.

8.3.1.4. Tratamiento De Datos Nulos.

Como se mostró en el capítulo de calidad, para cada variable se determinó la cantidad de datos nulos por variable. Aquellos en los que no se tenía información del negocio se eliminaron de manera definitiva. En el caso del área donde se encontraron errores de validez de la

información porque se presentaban valores en ceros, se decide determinar el área a partir del cruce de información entre el proyecto y el valor del inmueble permitiendo estimar con mayor exactitud las áreas.

Ilustración 17.
Porcentaje de campos vacíos por columna

Tasa Interes	62.468218
Año escritura/desistimiento	62.468218
CompradorTipoContrato	27.224721
Edad	7.823196
CompradorEstadoCivil	7.549384
CompradorNivelAcademico	7.539605
CompradorTipoVivienda	7.539605
CompradorOcupacion	7.529826
Ent Subsidio Concurrente	4.459222
Vr Subsidio Concurrente	4.459222
CompradorNumeroHijos	2.571876
Sueldo	2.571876
CompradorPersonasCargo	2.571876
Piso.1	0.166243
Ciudad	0.156464
Ent Cesantías	0.058674
Ent Crédito	0.019558
Ent Ahorro	0.019558

Nota: Esta ilustración muestra el porcentaje de valores vacíos por columna.

Como se observa en la ilustración, la información donde se presenta la mayor cantidad de campos vacíos es en las columnas tasa de interés y año de escritura/desistimiento. El que tengan exactamente el mismo porcentaje se debe a que la tasa de interés es una variable importante en la decisión de compra de un cliente y se incluye actualmente en la base de datos, es un campo calculado a partir de la información publicada por el banco de la república tomando la columna de fecha en la cual se realizó la escrituración. Teniendo en cuenta lo anterior, se decide para la columna correspondiente al año agregar la etiqueta “Sin información” y mantener los valores vacíos de la columna tasa de interés.

En cuanto a la columna relacionada con el tipo de contrato del cliente, que presenta un 27 % de campos vacíos, se decide no reemplazarlo por el valor que más se repite (moda) ya que, al

reemplazarlos, se afirmaría que alrededor del 77% presenta este tipo de contrato y por recomendación de la empresa, que afirma que esto no corresponde a la realidad, no es habitual que su tipo de cliente presente este tipo de contrato por lo que se decide eliminar la columna para estudios posteriores.

En el caso de los valores restantes debido a que son inferiores al 10%, se toma la decisión de para los valores categóricos reemplazar los campos vacíos por la moda y para los valores numéricos utilizar la técnica de k-means donde el valor perdido se estima a través de la media de los valores de los vecinos más cercanos excepto en el campo correspondiente al valor del subsidio concurrente donde se asigna la mediana correspondiente.

8.3.1.5. Tratamiento De Outliers.

En cuanto a la limpieza de datos atípicos, para cada una de las variables de tipo numérico se realiza un diagrama de cajas para validar la existencia de datos atípicos. Para realizar su limpieza, optamos por acotarlos a través de la estimación de los cuartiles y rango intercuartílico.

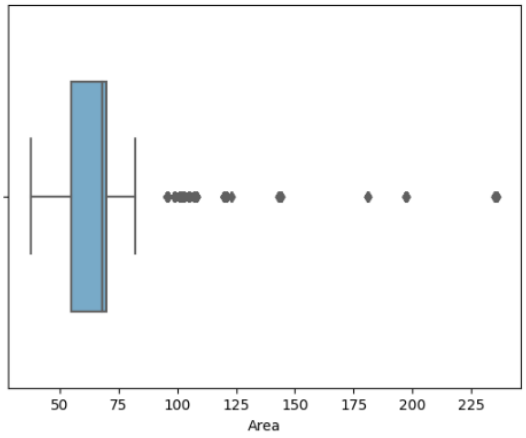
Para ello, primero identificamos los cuartiles de cada una de las variables junto con el rango intercuartílico (IQR = Q3-Q1). El outlier estará dado por aquel valor que esté fuera del rango Q1-1.5*IQR, por debajo; y Q3+1.5*IQR, por arriba.

A continuación, mostramos el comportamiento de cada variable junto con los cuartiles y rangos estimados para cada variable:

Ilustración 18

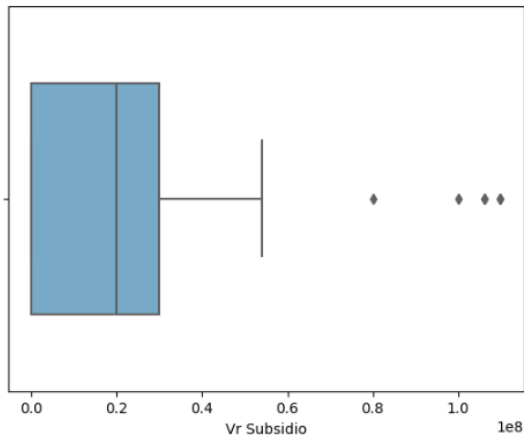
Diagrama de cajas por variable

Área



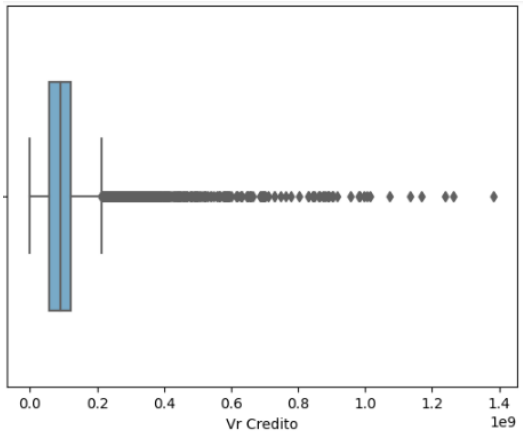

```
Q1= 55.0
Q3= 70.0
IQR= 15.0
Valor outlier por encima= 92.5
Valor outlier por debajo= 32.5
Mediana= 68.0
```

Valor del subsidio



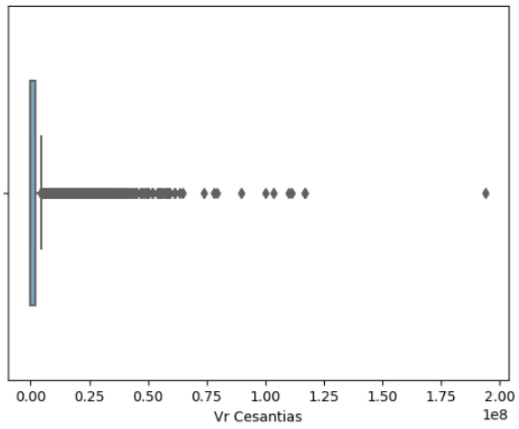

```
Q1= 0.0
Q3= 30000000.0
IQR= 30000000.0
Valor outlier por encima= 75000000.0
Valor outlier por debajo= -45000000.0
Mediana= 20000000.0
```

Valor del crédito



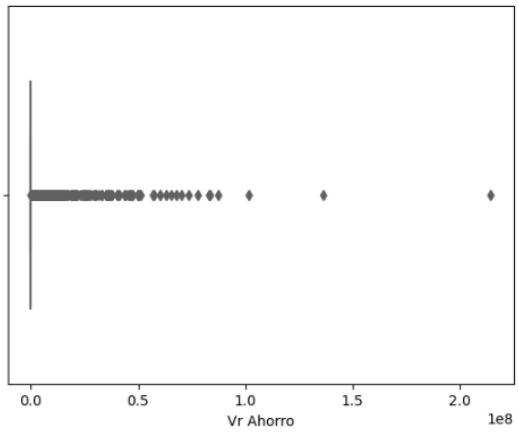

```
Q1= 57928999.0
Q3= 120000000.0
IQR= 62071001.0
Valor outlier por encima= 213106501.5
Valor outlier por debajo= -35177502.5
Mediana= 92220000.0
```

Valor Cesantías



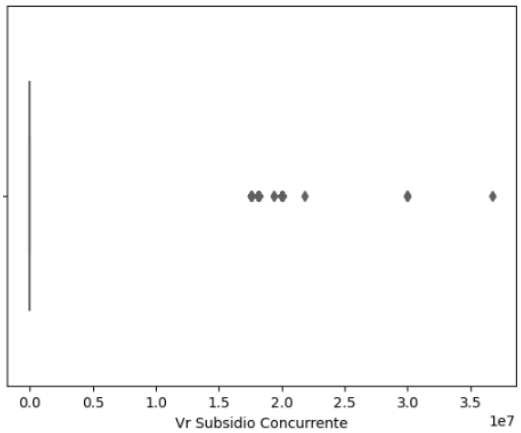

```
Q1= 0.0
Q3= 1800000.0
IQR= 1800000.0
Valor outlier por encima= 4500000.0
Valor outlier por debajo= -2700000.0
Mediana= 0.0
```

Valor de Ahorro



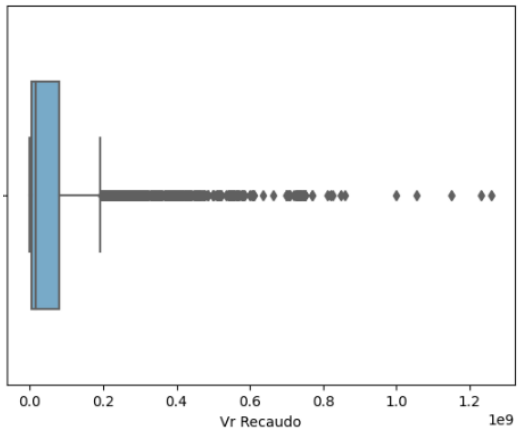

```
Q1= 0.0
Q3= 0.0
IQR= 0.0
Valor outlier por encima= 0.0
Valor outlier por debajo= 0.0
Mediana= 0.0
```

Valor de subsidio Concurrente



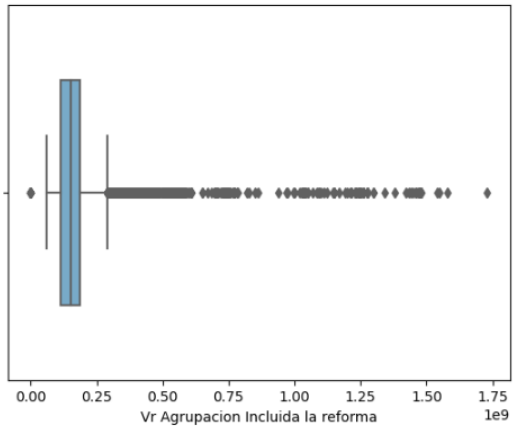

```
Q1= 0.0
Q3= 0.0
IQR= 0.0
Valor outlier por encima= 0.0
Valor outlier por debajo= 0.0
Mediana= 0.0
```

Valor del recaudo



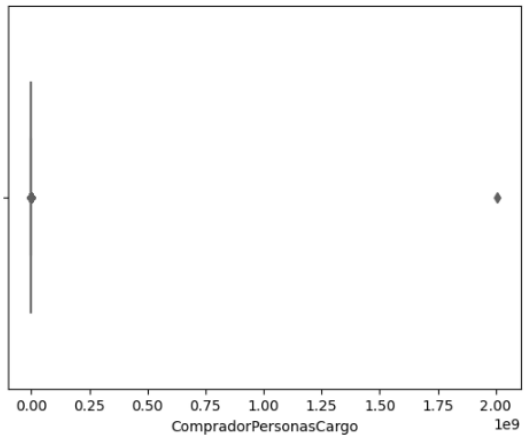

```
Q1= 4508125.0
Q3= 79661555.25
IQR= 75153430.25
Valor outlier por encima= 192391700.625
Valor outlier por debajo= -108222020.375
Mediana= 16154150.0
```

Valor agrupación incluida la reforma



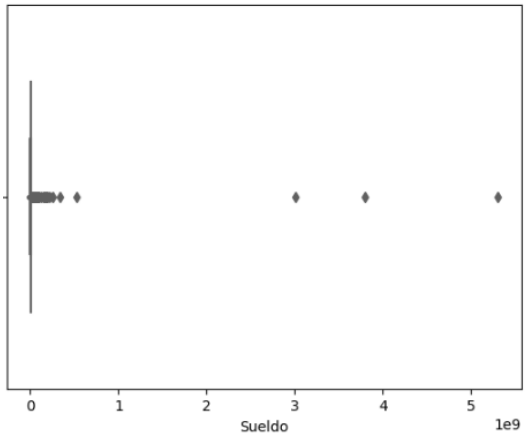

```
Q1= 114920000.0
Q3= 185150000.0
IQR= 70230000.0
Valor outlier por encima= 290495000.0
Valor outlier por debajo= 9575000.0
Mediana= 151200000.0
```

Personas a cargo



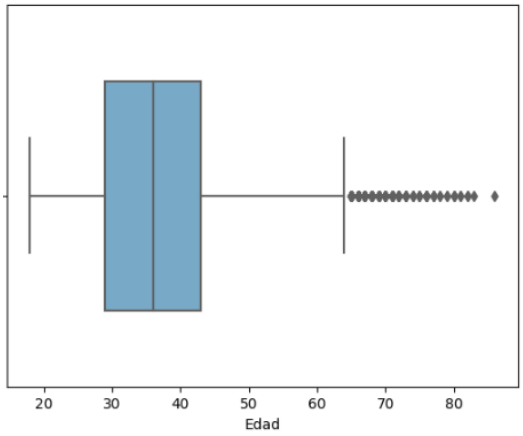

```
Q1= 0.0
Q3= 1.0
IQR= 1.0
Valor outlier por encima= 2.5
Valor outlier por debajo= -1.5
Mediana= 0.0
```

Salario



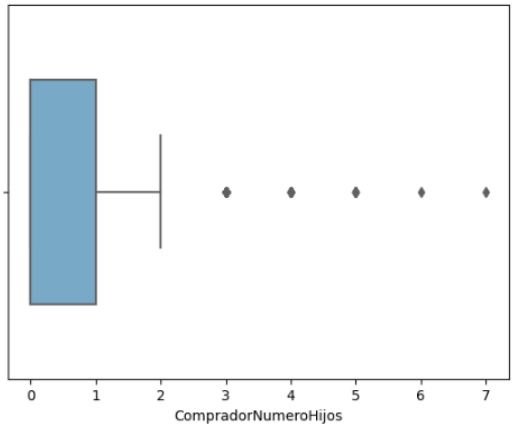

```
Q1= 1400000.0
Q3= 4200000.0
IQR= 2800000.0
Valor outlier por encima= 8400000.0
Valor outlier por debajo= -2800000.0
Mediana= 2000000.0
```

Edad




```
Q1= 29.0
Q3= 43.0
IQR= 14.0
Valor outlier por encima= 64.0
Valor outlier por debajo= 8.0
Mediana= 36.0
```

Número de hijos




```
Q1= 0.0
Q3= 1.0
IQR= 1.0
Valor outlier por encima= 2.5
Valor outlier por debajo= -1.5
Mediana= 0.0
```

Nota: Esta ilustración muestra el comportamiento de los outliers en cada una de las variables de estudio.

De acuerdo con los gráficos, observamos que en el caso de la variable área y valor de la agrupación incluida la reforma no se realiza tratamiento debido a que aquellos valores que se encuentran por encima del valor outlier por encima puede modificar el comportamiento del segmento casa generando sesgo en la implementación del modelo.

En el caso de las variables relacionadas con los, cesantías y crédito no es necesario realizar el tratamiento porque la variabilidad en cuanto al desembolso es bastante elevada y al tratarlos llevaríamos muchos datos a un valor en particular.

Para las variables de cesantías, ahorro y subsidio recurrente se decide no realizar tratamiento debido a que se presenta una cantidad de datos significativa con valores de cero y que de acuerdo con la compañía de estudio corresponde a un escenario completamente habitual debido a que no es frecuente el desembolso de dinero por estos medios.

En las variables relacionadas con entorno del cliente, como el salario, no tendrá tratamiento porque se presenta una situación similar al área y valor de la vivienda entre más alto, es el salario del cliente y si se realiza tratamiento podríamos presentar sesgo al momento de la interpretación y aplicación del modelo. Para otras variables como el valor del subsidio, personas dependientes, edad e hijos se realizará tratamiento a través de la sustitución de los valores atípicos por el valor outlier por encima el cual se observó en la **¡Error! No se encuentra el origen de la referencia.**.

9. Análisis

9.1. Análisis De Componentes Principales

La aplicación del método de análisis de componentes principales (ACP) permitirá analizar las estructuras subyacentes que se encuentran en las bases de datos, por esta razón, se realizará el análisis de la base de datos unificada en capítulos anteriores.

En las bases de datos se reciben 10227 registros entre negocios exitosos y no exitosos, de los que 9540 clientes iniciaron el trámite de separación del inmueble entre 2020 y noviembre del 2022.

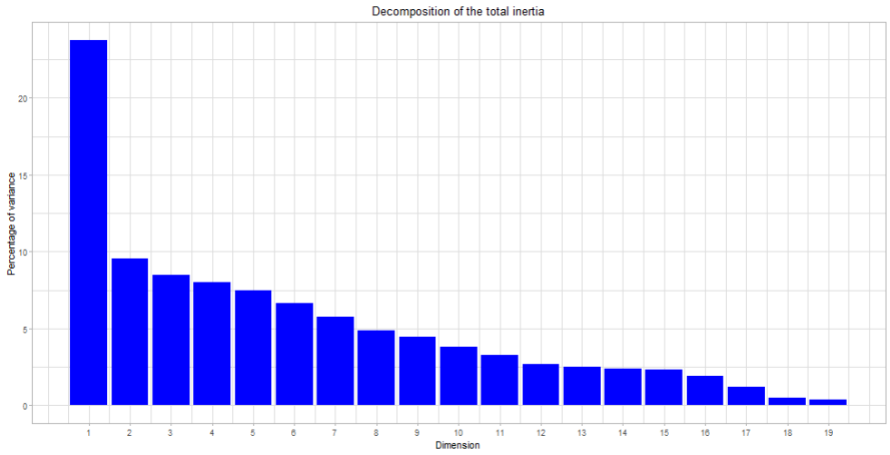
Se decide utilizar el método ACP debido a que en la fila se encuentra la información por cliente y en las columnas variables de tipo continuo que se han medido sobre los individuos (clientes).

El objetivo, es comparar a los individuos entre sí. Las gráficas nos permitirán observar la forma de la nube de individuos lo que permitirá detectar patrones entre ellos como también describir las relaciones entre las variables.

A continuación, se presenta un análisis del porcentaje de varianza entre las dimensiones, del cual se decide escoger 2 dimensiones (dim 1 y dim 2) dado que presenta la mayor diferencia de acuerdo con la descomposición del total de la inercia dentro del conjunto de datos.

Ilustración 19

Descomposición de inercia.



Nota: Esta ilustración muestra el porcentaje de varianza entre dimensiones.

Tabla 23

Porcentaje de varianza

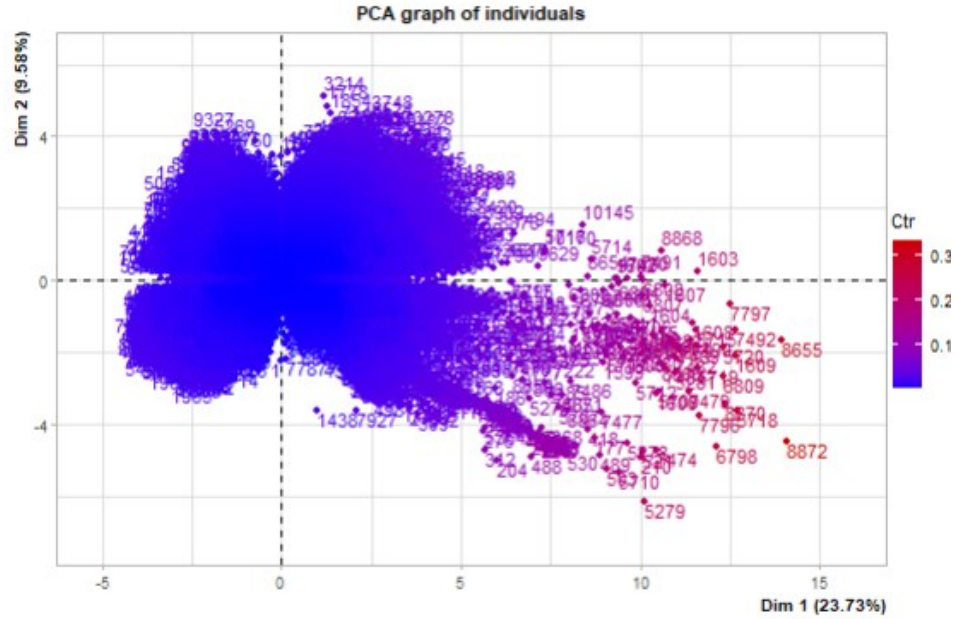
Eigenvalue	Porcentaje de varianza	Porcentaje acumulado de varianza
4,51	23,73	23,73
1,82	9,58	33,3
1,61	8,49	41,79
1,52	7,99	49,78
1,42	7,47	57,25
1,26	6,65	63,9
1,09	5,75	69,66
0,92	4,86	74,52
0,84	4,43	78,95
0,73	3,82	82,78
0,62	3,27	86,04
0,51	2,71	88,75
0,48	2,52	91,27
0,45	2,38	93,65
0,44	2,31	95,96
0,37	1,93	97,89
0,23	1,22	99,11
0,1	0,5	99,61
0,07	0,39	100

Nota: Esta tabla muestra el porcentaje de la varianza y porcentaje de varianza acumulada por dimensión.

El plano factorial que se muestra a continuación representa los datos en un 33% aproximadamente, por lo cual se advierte de una pérdida significativa de la información inicial suministrada.

Ilustración 20

Plano factorial



Nota: Esta ilustración muestra el comportamiento de contribución de los individuos en la dimensión 1 y 2.

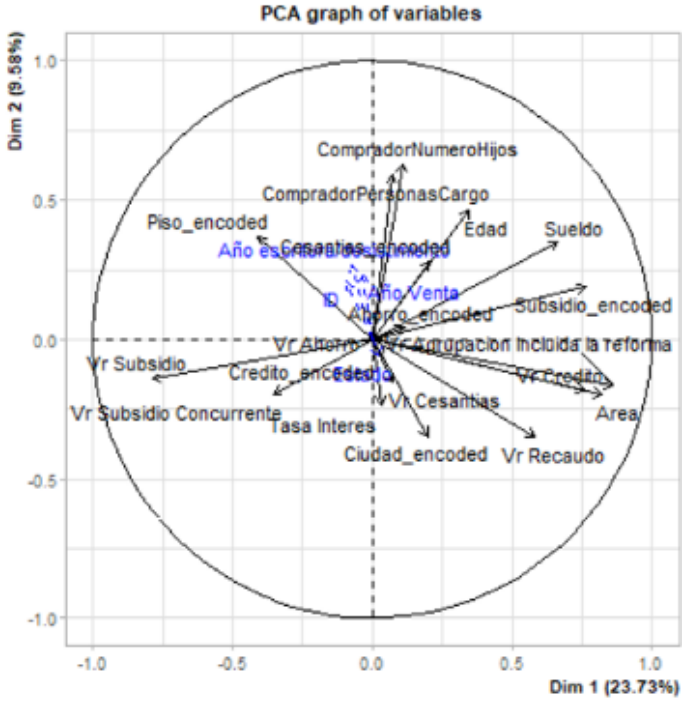
Según la gráfica anterior, los individuos con mayor contribución son los que están en la parte inferior derecha del plano.

Luego, se realiza la gráfica del círculo de correlaciones, del que se evidencia como ciertas variables del modelo se correlacionan entre sí. Por ejemplo, la variable personas a cargo tiene una importante correlación con la variable número de hijos y una contribución al modelo. Se

presenta una situación similar entre las variables área y valor de la agrupación incluida la reforma.

Ilustración 21

Circulo de correlaciones



Nota: Esta ilustración muestra el gráfico de correlación de las variables en estudio en la dimensión 1 y 2.

9.1.1. Contribución De Las Variables En Cada Componente

A continuación, se muestra la tabla de contribución por cada uno de los componentes del examen para las dimensiones representadas (Dim1 y Dim2):

Tabla 24

Tabla de contribución por cada componente

Variable	Dim.1	Dim.2	Total
Vr Agrupación Incluida la reforma	16.4	1.5	17.9

Área	15.0	2.2	17.2
Vr Subsidio	13.7	1.1	14.8
Vr Crédito	12.9	1.8	14.7
Sueldo	9.8	6.6	16.4
Vr Recaudo	7.5	6.8	14.3
Vr Subsidio Concurrente	2.8	2.2	5.0
Edad	2.7	11.8	14.5
CompradorNumeroHijos	0.3	21.9	22.2
CompradorPersonasCargo	0.1	19.1	19.2
Vr Cesantías	0.1	1.3	1.4
Vr Ahorro	0.0	0.0	0.0

Nota: Esta tabla muestra la contribución de cada una de las variables de estudio.

Considerando la tabla anterior, se identifica que la variable con mayor variabilidad y por mayor contribución, en la dimensión 1 es el valor de la agrupación incluida la vivienda y el área, mientras que en la dimensión 2 es el número de hijos y el número de personas a cargo.

9.2 Técnicas De Modelamiento

A continuación, mostramos en la siguiente tabla cada una de las técnicas de modelamiento seleccionadas basándose en las siguientes consideraciones: los datos obtenidos, el objetivo analítico y requisitos específicos del modelo.

Tabla 25
Descripción de Modelos

(ArcGis Pro, 2023)	Descripción
Arboles de Clasificación	Sistemas de clasificación que predicen o clasifican características futuras basándose en un conjunto de reglas de decisión.
Gradient Boosting	Técnica de Machine Learning para el análisis de la regresión y para problemas de clasificación estadística, la cual, produce un modelo predictivo en forma de un conjunto de modelos de predicción débiles, típicamente árboles de decisión. Construye el modelo de forma escalonada como lo hacen otros métodos de Boosting, y los generaliza permitiendo la optimización arbitraria de una función de pérdida diferenciable.
Random Forest	Conjunto de árboles de decisión combinados con bagging, lo que permite que distintos árboles vean distintas porciones de

	los datos. Ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor.
Máquina de soporte Vectorial	Conjunto de algoritmos de Machine Learning, relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra.
XGBoost	Método de Machine Learning para clasificación y regresión que utiliza la herramienta AutoML . XGBoost es la abreviatura de aumento de gradiente extremo. Este método se basa en árboles de decisión y mejora otros métodos, como el Radom Forest y el Gradient boost. Funciona bien con conjuntos de datos grandes y complicados mediante el uso de varios métodos de optimización.
LightGBM	método de Gradient Boost que utiliza la herramienta AutoML y se basa en árboles de decisión, se puede utilizar tanto para la clasificación como para la regresión, este método crea árboles de decisión que crecen en hojas, lo que significa que, dada una condición, solo se divide una sola hoja, según la ganancia. Los datos se agrupan en contenedores con un histograma de la distribución. Los contenedores, en lugar de cada punto de datos, se utilizan para iterar, calcular la ganancia y dividir los datos.

Nota: Esta tabla muestra una breve descripción de las técnicas de modelamiento seleccionadas.[6]

9.3 Construcción De Los Modelos:

La herramienta seleccionada para realizar las técnicas de modelamiento fue Colaboratory, producto de Google que permite la ejecución de código Python, es libre, maneja una excelente visualización y es apta para el volumen de información que se tiene.

Teniendo en cuenta la revisión de los datos descrita en capítulos anteriores, encontramos un desbalance muy importante entre los negocios exitosos y los negocios no exitosos. Los negocios sin éxito representan el 6 % de los registros de la base de datos.

[6] (ArcGis Pro, 2023)

Considerando lo anterior, se toma la decisión para la selección final de variables que incluirán en los modelos los resultados de contribución del ACP para validar la consistencia de la información en cada variable. Aquellas variables donde se encuentren registros con valores de cero superiores al 50% de los datos totales de negocios en estado de desistimiento no se incluirán en la etapa de modelado.

A partir de los criterios expuestos anteriormente, las variables que se ingresarán al modelo son: Piso, Área, Ciudad, Tipo de Venta, Valor de la agrupación incluida la reforma, CompradorPersonasaCargo, Sueldo, CompradorEstadoCivil, Edad, CompradorNumeroHijos, CompradorOcupacion, CompradorNivelAcademico, y CompradorTipoVivienda.

Posterior a la determinación de las variables se selecciona la información para entrenamiento y prueba a través de código en Python se divide la base de datos en dos lo que quiere decir que cada una tendrá aproximadamente 5100 registros.

Esta división se realiza para evitar que, en el proceso de entrenamiento, los modelos se ajusten a los datos. Tener información de prueba permitirá verificar el error de una segunda muestra que no se relaciona con la información del entrenamiento.

La variable por predecir es de tipo categórico y corresponde al estado del negocio “Desistido” o “No desistido”, para la aplicación del modelo se ingresará 1 si el negocio culminó y 0 si el negocio no culminó.

A partir de la técnica de validación cruzada, se realiza 60 iteraciones para la identificación óptima de los parámetros de los árboles de decisión como máxima profundidad, mínimo de muestras por hoja y mínimo de muestras para dividir. Se fijo un valor de 5 para la

máxima profundidad, 7 para el mínimo de muestras por hoja y 2 para el mínimo de muestras para dividir.

9.4 Evaluación de los modelos

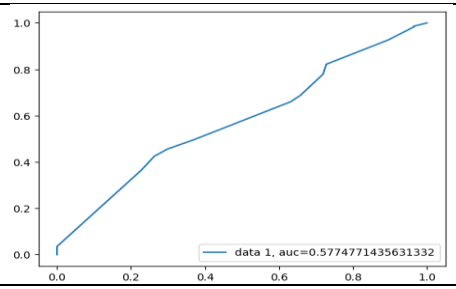
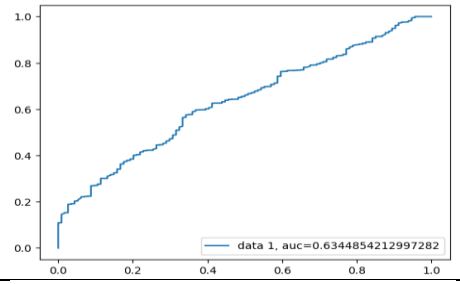
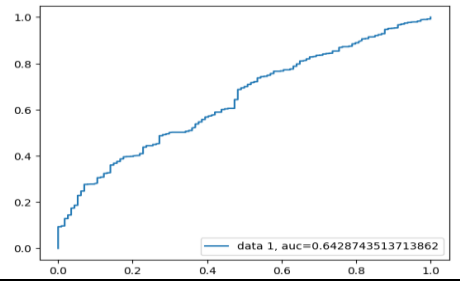
Para el entrenamiento del modelo, se utilizan dos medidas de desempeño: La curva ROC y el área bajo la curva (AUC).

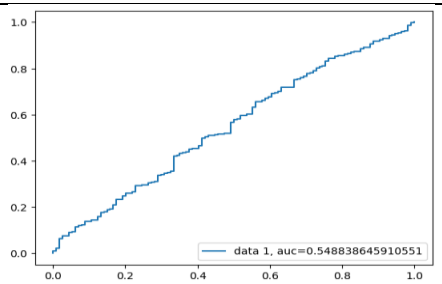
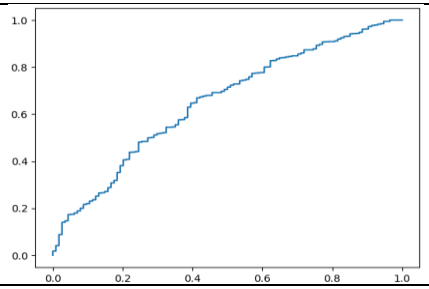
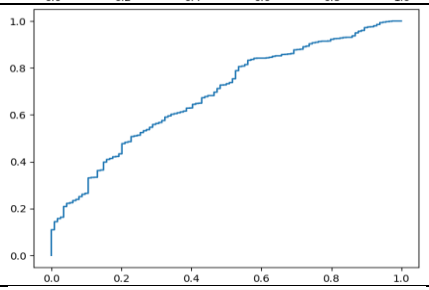
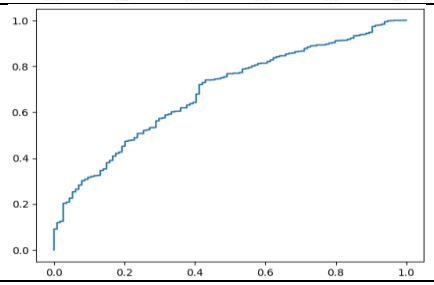
Se toma la decisión de utilizar estas dos medidas debido a que la curva ROC, es una representación visual del comportamiento del rendimiento del modelo donde el eje x corresponde a la tasa de falsos positivos y el eje y a la tasa de verdaderos positivos. La visualización de la curva es útil en la evaluación porque representa la compensación entre errores y beneficios que un clasificador hace entre dos clases. Por otro lado, el área bajo la curva (AUC) es un valor escalar entre 0 y 1 que representa el rendimiento esperado de la curva ROC (Fawcett, 2006).

A continuación, mostramos el resultado obtenido de curva ROC y área bajo la curva (AUC) de cada una de las técnicas de modelamiento seleccionadas:

Tabla 26

Medidas de desempeño datos de entrenamiento

Técnica de Modelamiento	ROC	AUC (%)
Árbol de clasificación		57.74
Gradient Boosting		63.44
Random Forest		64.28

Máquina de Soporte Vectorial		54.88
XGBoost		65.15
LightGBM		68.57
CatBoost		68.82

Nota: Esta tabla muestra el resultado de cada una de las medidas de desempeño de entrenamiento del modelo.

Según lo anterior, la técnica de modelado con peor rendimiento es máquina de soporte vectorial, difícilmente supera el 55 % por ende, no presenta una mejor capacidad para posterior implementación. Por otro lado, los modelos que presentan el mejor rendimiento son LightGBM y CatBoost con un valor de 68.57 y 68.82% respectivamente, indicando que hay una mejor capacidad que el azar para clasificar correctamente, sin embargo, aún tiene opción de mejoramiento. A continuación, se realizará la validación del modelo a través de la prueba para obtener una evaluación más confiable del rendimiento del modelo.

9.5 Validación de los modelos

La validación de los modelos se realiza a través de la técnica de validación cruzada con un k-folds igual a 5 y se estima a partir de los resultados de los subconjuntos el promedio de las medidas: AUC, precisión, recuperación (recall), F1 Score y exactitud (accuracy).

Se decide realizar la técnica de validación cruzada porque permite la utilización de más datos de prueba (Microsoft, 2022) , es un método que nos permite entrenar y probar el modelo varias veces. Para nuestro conjunto de datos, divide la base en 5 subconjuntos de manera aleatoria y en cada iteración utiliza un subconjunto como validación mientras que los cuatro restantes como entrenamiento.

Como se mencionó anteriormente, se decide revisar otras medidas como:

Precisión: Capacidad del clasificador de no etiquetar una muestra negativa como positiva.

Ecuación 1
Ecuación de Precisión

$$\frac{Verdaderos\ Positivos}{(Verdaderos\ positivos + Falsos\ positivos)}$$

Recuperación (Recall): Capacidad del modelo para encontrar todas las muestras positivas.

Ecuación 2
Ecuación de Recuperación

$$\frac{Verdaderos\ Positivos}{(Verdaderos\ positivos + Falsos\ negativos)}$$

F1 Score: Media armónica ponderada de la precisión y la recuperación (Scikit-learn, 2023).

Exactitud (Accuracy): Precisión del conjunto. Proporción de instancias clasificadas correctamente por el modelo en relación con el total de instancias (Scikit-learn, 2023).

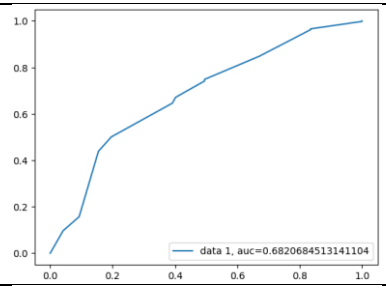
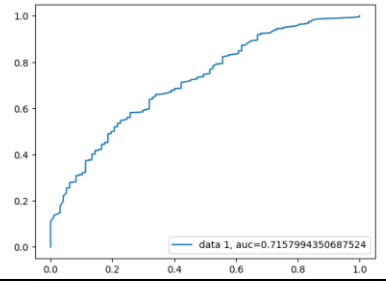
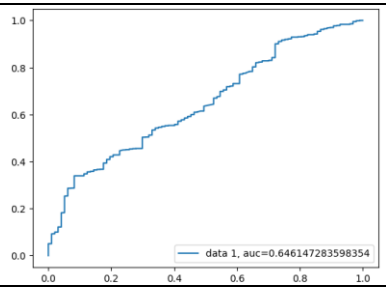
Ecuación 3
Ecuación de Exactitud

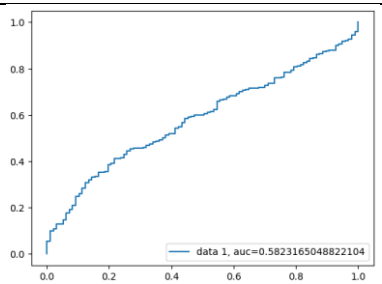
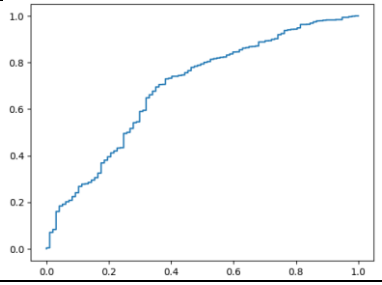
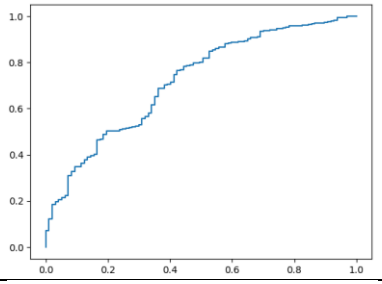
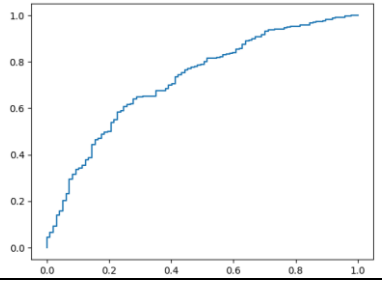
$$\frac{(Verdaderos\ Positivos + Verdaderos\ negativos}{(Total\ de\ registros)}$$

A continuación, mostramos en la siguiente tabla los resultados de cada una de las métricas definidas anteriormente para cada técnica de modelado:

Tabla 27

Medidas de desempeño datos de prueba

		Validación cruzada (K-folds=5)				
Técnica de Modelamiento	**ROC**	**AUC**	**PRECISIÓN**	**RECALL**	**F1**	**ACCURACY**
Árbol de clasificación		52	93	85	89	81
Gradient Boosting		63	93	93	92	87
Random Forest		60	93	98	96	92

Máquina de Soporte Vectorial		53	93	100	96	93
XGBoost		63	93	92	92	86
LightGBM		64	93	93	93	87
CatBoost		63	93	93	92	87

Nota: Esta tabla muestra el resultado de cada una de las medidas de desempeño de prueba del modelo.

Teniendo los resultados de la prueba, continúa siendo mejor el resultado de la técnica de modelamiento de LightGBM en cuanto al AUC. Sin embargo, al comparar los resultados entre cada técnica no se presentan diferencias significativas entre sí además frente a los resultados del entrenamiento el rendimiento disminuye por lo que aparentemente los modelos son un poco menos consistentes cuando se evalúa en diferentes divisiones de datos o en conjuntos de prueba más desafiantes.

En cuanto al recall, en la validación cruzada el resultado es alto en la mayoría de los modelos con valores entre 92 y 100. Esto indica que los modelos tienen una buena capacidad para identificar correctamente las instancias positivas lo que impacta a su vez los resultados de F1 Score.

El que el AUC sea más bajo frente a las otras métricas propuestas se debe a que en principio no se basan en umbrales de probabilidad sino directamente en las predicciones del modelo y al encontrarse en desequilibrio la base de datos puede estar enfocándose en detectar las instancias positivas.

A pesar de que los modelos mencionados tienen una capacidad mejor que el azar para clasificar correctamente las muestras, consideramos que es indiscutiblemente necesaria la adquisición de más información relacionada con negocios no exitosos para continuar con el ciclo de implementación.

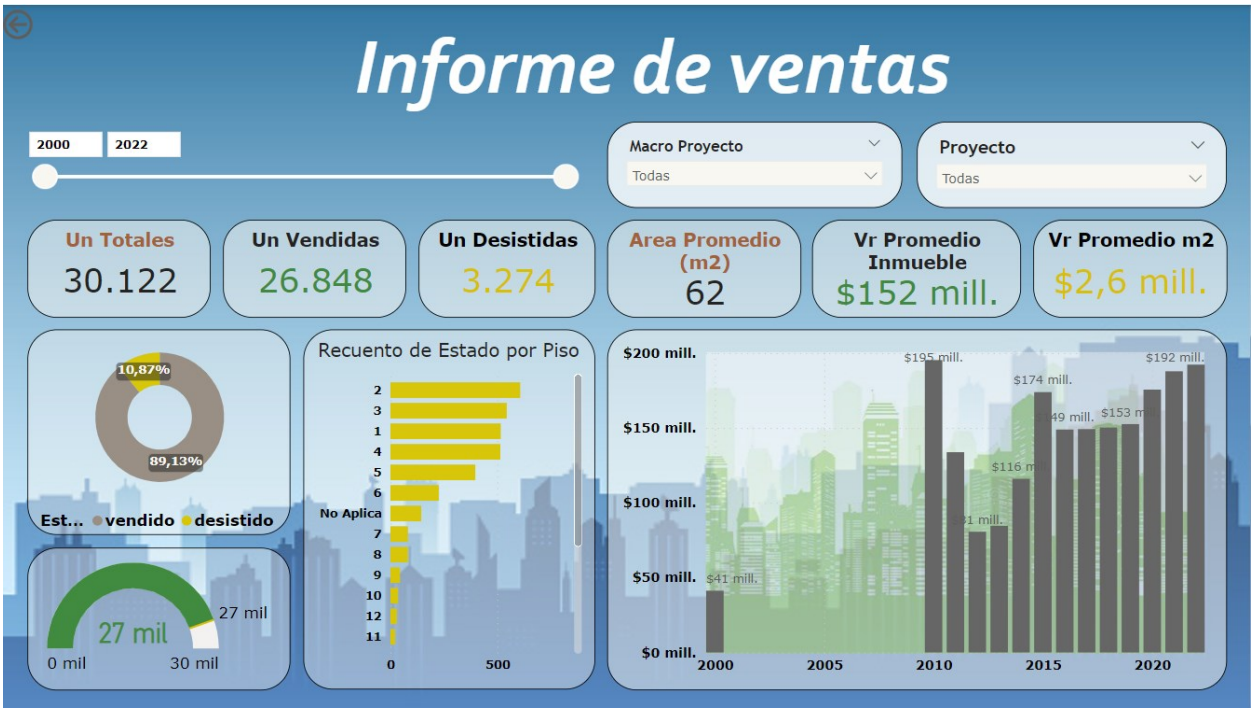
10. Visualización

A continuación, se muestra el modelo de visualización en Power BI, con el fin de generar valor en la compañía por medio de un modelo que permita convertir datos en información significativa, mediante la aplicación de tablas y gráficos de forma dinámica, de modo tal, que facilite la toma de decisiones estratégicas mediante la detección de tendencias y el análisis de los datos en tiempo real.

De acuerdo, con lo anterior se presenta un informe en Power BI con 3 pestañas: Ventas, Cliente y negocio.

Ventas: Muestra la información correspondiente a las unidades vendidas y desistidos para los proyectos de la compañía, así como valores promedio de m2.

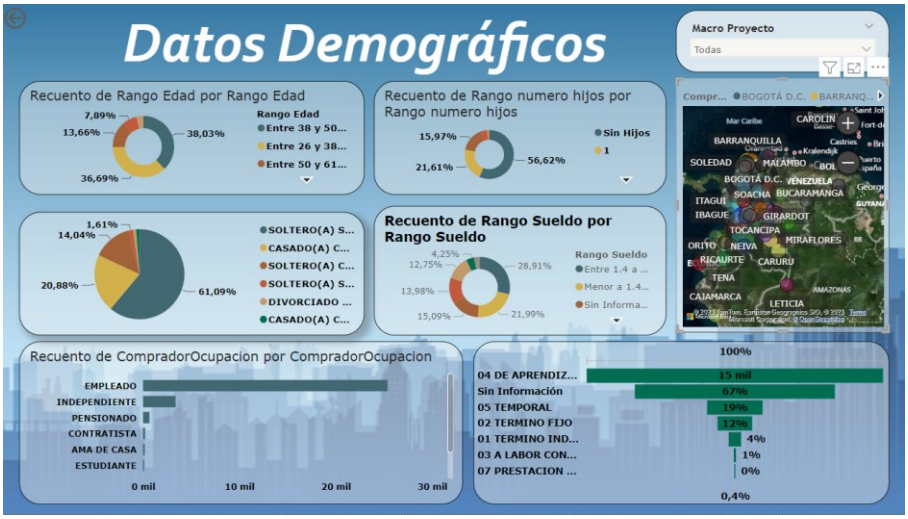
Ilustración 22
Informe de ventas Power BI



Cliente: Muestra la información demográfica de los clientes que logran culminar una venta exitosos para los diferentes proyectos de la compañía, presenta datos como edad, número de hijos, Salario, Tipo de contrato y ciudad.

Ilustración 23

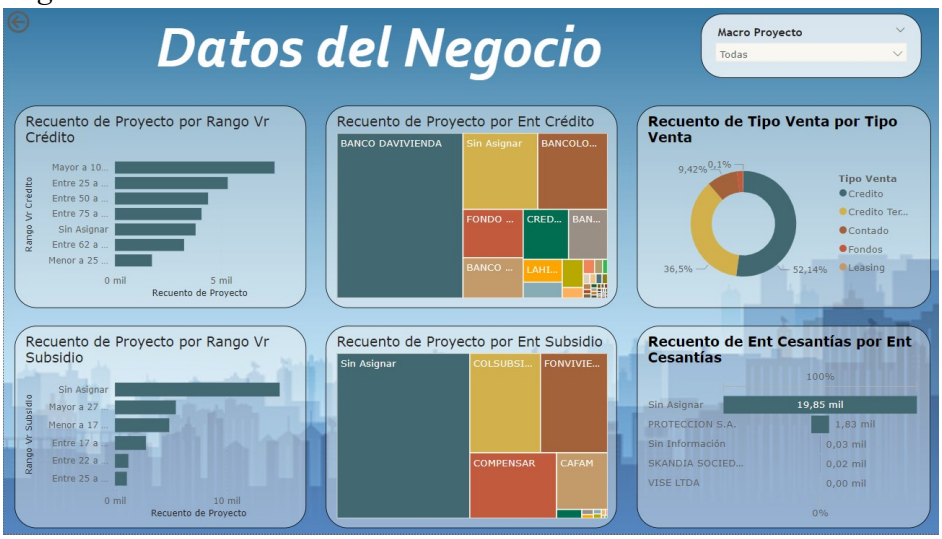
Informe de Clientes Power BI



Negocio: Muestra la información de los negocios exitosos para los diferentes proyectos de la compañía, presenta datos como tipo de venta, valor de crédito, entidad de crédito, valor de subsidio y entidad de subsidio.

Ilustración 24

Informe del Negocio Power BI



11. Recomendaciones a la Organización

Teniendo en cuenta que de los datos totales presentados por la empresa (36709) solo se utilizó un 30% de la información recibida en los modelos por presentar una baja calidad de los datos, después de haber realizado varias limpiezas y técnicas para mejorar la información, es preciso decir que se requiere realizar acciones para el mejoramiento de la obtención de la información antes de realizar la implementación de los modelos en la compañía, por lo cual y con el fin de generar valor se proponen las siguientes acciones para mejorar la calidad de las bases de datos:

A continuación, se enuncian los campos que requieren acciones de mejora, ya que son de utilidad para los modelos desarrollados:

Área Construida: Esta variable presenta campos vacíos y diferencias de áreas en el mismo proyecto, que deben corregirse, y se encuentran áreas mayores a 1000 m2, considerados errores según el tipo de proyecto.

Ciudad y país de Residencia: Se evidencia que este campo no tiene restricciones de formato, se encontraron valores numéricos, nombres, apellidos, ciudades, municipios y países que no existen o no pertenecen a Colombia.

Personas a Cargo y Número de hijos: Este campo presenta un número considerable de valores nulos, también presentan letras y números mayores a 50, lo cual no es coherente.

Salario: En este campo se encontraron valores negativos, 0 y mayores a 100 millones, por lo cual se consideran errores, entiendo los datos del tipo de negocio.

Estado Civil, Profesión, Nivel académico, Tipo de contrato, Tipo de Negocio: se presentan gran cantidad de valores vacíos y se requiere estandarizar.

Fecha de nacimiento: El campo presenta valores vacíos y fechas de nacimiento que dan una edad de más de 100 años, las cuales deben revisarse.

Por eso se requiere una socialización con las áreas de negocios y comercial para capacitar al personal encargado, para que la información del cliente sea verídica y se encuentre diligenciada para volver a alimentar los modelos con datos reales y fiables en un periodo de 1 año.

Por otro parte, se requiere estandarizar la información clave para evaluar las variables mencionadas y otras que pueden ser de gran impacto para la compañía a futuro, como la tasa de interés bancaria expedida por el Banco de la Republica.

12. Conclusiones

Al realizar el proceso de la calidad de los datos, se identificaron aquellas variables que afectarían a futuro los resultados de modelamiento. Sin embargo, se debe estipular con la compañía los requerimientos de precisión de los campos de entrada manual como vendedor, ocupación y profesión para posteriormente determinar si es posible la inclusión de estos campos en los modelos siguientes de aplicación.

De los análisis de la base de datos exitosos, se evidencia que la edad se relaciona con el nivel de ingresos y la presencia de hijos; conocer los 3 grupos de edad presentados en la población de compradores afectan u influyen directamente en la compra de vivienda para identificar la estrategia de mercado según el proyecto que mejor acogida tenga para cada población objetivos y así optimizar los costos de la compañía.

Por otro lado, se evidencia como la formación profesional no tiene un impacto directo en los ingresos mensuales mientras que el hecho de ser empleado u independiente si, por lo cual y con el fin de evitar problemas en los proyectos por aumentos de los desistimientos se debería tener en cuenta esta variable dadas las condiciones del mercado durante la pandemia las cuales generan una inestabilidad laboral; lo cual hace imperativo entender como el factor salarial se ve afectado de acuerdo al departamento.

Finalmente , se puede concluir que de los 8 modelos utilizados para la predicción de la variable objetivo (Estado) no pueden ser admitidos para la etapa de producción dado que las pruebas realizadas no tienen un margen significativo de precisión(Alrededor de 64), por lo cual, antes de aplicar cualquier modelo es necesario que la compañía haga un refinamiento de la data y lleve un proceso de estandarización y mejoramiento de la recolección de datos, con el fin de que a futuro los datos recolectada sea lo suficiente fiables para ser probados en un modelo

analítico que permita predecir con más exactitud la probabilidad de éxito y/o desistimiento de un negocio de vivienda.

Referencias

America, F. M.–T. (23 de 06 de 2020). *Thrend Group America*. Obtenido de Intención de Compra de vivienda “en tiempos de pandemia” en Colombia: https://www.trendgroupamerica.com/intencion-de-compra-de-vivienda-en-tiempos-de-pandemia-en-colombia/

ArcGis Pro. (18 de 05 de 2023). *ArcGis Pro*. Obtenido de https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm

Brethenoux, E. (10 de 11 de 2022). *KD nuggets*. Obtenido de https://www.kdnuggets.com/2017/02/analytics-grease-monkeys.html

DAMA internacional. (2010). *DAMA Guía de fundamentos para la gestión de datos.* Technics Publications.

Dolan, S., Valle, R., Jackson, S., & Schuler, R. (2007). *La Gestión de los Recursos Humanos* (Vol. 3). Editorial Mc Graw Hill.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 861-874.

IBM. (10 de 11 de 2022). *IBM ANALYTICS SOLUTIONS UNIFIED METHOD (ASUM)*. Obtenido de http://i2t.icesi.edu.co/ASUM-DM_External/index.htm#cognos.external.asum-DM_Teaser/tasks/sps_compile_business_background_ABF3C4C6.html?proc=_0eKIHlt6EeW_y7k3h2HTng&path=_0eKIHlt6EeW_y7k3h2HTng,_0eHEyVt6EeW_y7k3h2HTng,_0eEBg1t6EeW_y7k3h2HTng,_0eEojlt6EeW_y7

Jericó, & Pilar. (2000). *La gestión del talento: del talento individual al talento organizativo.* Madrid.: Prentice Hall.

Kennedy, W. (2006). *So What? who Cares? why You?* Ottawa, Canadá: Wendykennedy.

Kotler , P., & Gertner, D. (2007). *Marketing Internacional de lugares y destinos.* Méxio: Pearson.

Madie, D. (2019). Growth Wheel Tool Kit. Copenhagen.

Microsoft. (26 de Septiembre de 2022). *Learn Microsoft*. Obtenido de Learn Microsoft: https://learn.microsoft.com/es-es/azure/machine-learning/component-reference/cross-validate-model

Neck, H., Neck, C., & Murray, E. (2018). *Entrepreneurship: The Practice and Mindset.* Londres: SAGE.

Portafolio. (22 de 06 de 2022). *El Publímetro*. Obtenido de Edificaciones impulsan el PIB de construcción en este año: https://www.portafolio.co/economia/edificaciones-impulsan-el-pib-de-construccion-en-este-ano-566580

Porter, M. (2015). *Estrategia Competitiva: Técnicas para el análisis de los sectores industriales y de la competencia.* México: Grupo Editorial Patria.

Roberto, D. (2004). *Fundamentos de marketing.* Buenos aires: Ediciones granica.

Scikit-learn. (Mayo de 2023). *Scikit learn*. Obtenido de Scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

Scikit-learn. (Mayo de 2023). *Scikit learn*. Obtenido de Scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html