Avance 1. Análisis exploratorio de datos

A01150742 José Ovalle Alvarado

A01794879 Carlos de Jesús Méndez Tornero

A01104775 Alejandro Munguía Salazar

Proyecto Integrador | Fecha: 04/05/2025

# Introducción

La gestión eficiente de datos es uno de los pilares fundamentales para el desarrollo de soluciones basadas en inteligencia artificial. En esta fase se realiza un análisis exploratorio de datos (EDA) con el objetivo de preparar y comprender el conjunto de datos que servirá de base para el modelo de predicción de morosidad en la plataforma MiCoto. Además, se incluye un análisis complementario relacionado con la funcionalidad del chatbot que proporcionará asistencia contextual y automatizada a los usuarios.

El objetivo principal es conocer en profundidad la estructura del conjunto de datos, evaluar su calidad, y detectar tanto problemas como oportunidades que guíen la construcción de modelos predictivos y herramientas inteligentes dentro de la plataforma.

El EDA permite comprender el comportamiento general de las variables, identificar relaciones relevantes entre atributos, y detectar anomalías, valores extremos o faltantes que puedan afectar el desempeño de los modelos de aprendizaje automático. Asimismo, esta fase ofrece una visión integral del conjunto de datos mediante visualizaciones y estadísticas descriptivas, que permiten validar supuestos, generar hipótesis, y tomar decisiones informadas sobre las transformaciones necesarias para el preprocesamiento.

Durante este avance, se abordarán diversos aspectos como la detección de valores nulos o ausentes, la identificación de outliers, el análisis de la distribución de variables numéricas y categóricas, la evaluación de la cardinalidad de atributos, el estudio de relaciones multivariadas, y la presencia de tendencias temporales. También se evaluará si existe un desbalance en las clases de la variable objetivo (por ejemplo, casos de morosidad vs. cumplimiento), lo cual puede influir significativamente en el diseño y evaluación de modelos supervisados.

En caso de ser necesario se aplicarán técnicas de limpieza y transformación de datos justificadas por los hallazgos encontrados, con el fin de preparar un dataset robusto y confiable para las siguientes etapas del proyecto.

### Antecedentes

Icono

El contenido generado por IA puede ser incorrecto.

MiCoto.mx es una empresa que ofrece una solución tecnológica en modalidad SaaS (Software as a Service), especializada en la administración integral de condominios residenciales. Fundada con el objetivo de transformar digitalmente la operación cotidiana de administradores y residentes, la plataforma se ha posicionado como una herramienta clave en ciudades como Guadalajara, así como en otras regiones del país como Quintana Roo y el Estado de México.

A través de su plataforma web y móvil, MiCoto permite registrar pagos, emitir recibos, enviar notificaciones, coordinar mantenimientos, y centralizar información financiera y operativa de los condominios. Sin embargo, y pese a contar con una base sólida de operaciones, la empresa enfrenta actualmente retos estratégicos en términos de datos e inteligencia artificial:

Gestión de la morosidad: La falta de mecanismos predictivos para anticipar el incumplimiento en el pago de cuotas de mantenimiento representa una limitación operativa. Actualmente, los administradores reaccionan de manera tardía ante la morosidad, lo que genera problemas de liquidez y conflictos internos.

Acceso inteligente a la información: El sistema actual exige navegación manual para consultar reportes o información histórica. Esto limita la experiencia del usuario y ralentiza la toma de decisiones, especialmente para residentes sin conocimientos técnicos.

Frente a este escenario, surge este proyecto, cuyo propósito es evolucionar la plataforma actual hacia una solución analítica e inteligente. El proyecto busca integrar modelos de aprendizaje supervisado para predecir la morosidad, procesamiento de lenguaje natural (PLN) para permitir consultas conversacionales sobre la base de datos y chatbots para mejorar la interacción con los usuarios.

El análisis exploratorio de datos presentado en este avance se basa en un conjunto de datos estructurado que incluye historiales de pago por unidad habitacional, fechas de corte, montos, frecuencia de transacciones y configuración de políticas de cobro específicas por condominio. También se dispone de información de interacciones en la plataforma que servirá para entender mejor los patrones de uso y consulta.

# 2. Variables relacionadas a los pagos y tratamiento.

* Número de registros: 3896
* Número de columnas: 15
* Los datos no contienen valores faltantes explícitos, salvo en la columna `fecha\_pago`, que tiene fechas inválidas o faltantes debido a la falta de un pago reflejado en esa transacción.

Tipos de datos por columna:

|  |  |  |
| --- | --- | --- |
| Variable | Tipo | Descripción |
| id: | object | Número de indentificación codificado. |
| casa | object | Cadena de texto que contiene el Coto y número de casa asociado al id. |
| coto | object | Cadena de Texto asociadao al número de Coto, del id. |
| tipo | object | Cadena texto correspondiente al tipo de pago. |
| status | object | Cadena de texto que corresponde a la condición del pago. |
| monto | int64 | Entero, monto del pago o del pendiente. |
| concepto\_original: | object | Cadena de texto que hace referencia al mes que el mantenimiento está pagado |
| hasta el mes | datetime64[ns] | Fecha hasta el cual está pagado el mantenimiento. |
| fecha\_creacion | datetime64[ns] | Fecha de creación del registro. |
| fecha\_pago: | datetime64[ns] | Fecha de pago reportada. |
| info\_recibo: | object | Cadena de texto con la información o evidencia del pago. |
| Estatus pago: | object | Cadena de texto para determinar que se haya registrado pago. |
| Moroso: | object | Cadena de texto para determinar si la cuenta casa está vigente o vencida. |
| Numero de moras maxima: | int64 | Entero relacionado con el número máximo de moras en la casa |
| moras: | int64 | Entero relacionado con el conteo de moras por casa. |

# 2.1 Análisis univariado.

## Variable: id

Valores únicos: 3896

Valores nulos: 0 (0.00%)

Valores más frecuentes:

6644ed063cc8e 1  
6592b7529113d 1  
64c8f3c345ce3 1  
64f1d2436ce53 1  
65195f42a0fab 1  
65423dc2bf614 1  
6569d8d33ac7f 1  
6592b75290e77 1  
65bb95d3347f0 1  
65e1d1530f443 1

## Variable: casa

Valores únicos: 344

Valores nulos: 0 (0.00%)

Valores más frecuentes:

Coto 1 Casa 20: 17  
Coto 1 Casa 104: 16  
Coto 1 Casa 133: 15  
Coto 1 Casa 43: 14  
Coto 1 Casa 144: 14  
Coto 1 Casa 39: 14  
Coto 1 Casa 198: 14  
Coto 1 Casa 179: 14  
Coto 1 Casa 64: 14  
Coto 1 Casa 56: 13

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

## Variable: coto

Valores únicos: 3

Valores nulos: 0 (0.00%)

Valores más frecuentes:

Coto 1: 2679  
Coto 2: 1216  
Casa 49: 1

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

## Variable: tipo

Valores nulos: 0 (0.00%)

Valores más frecuentes:

Mantenimiento: 3836  
afavor: 60

Gráfico

El contenido generado por IA puede ser incorrecto.

## Variable: status

Valores únicos: 5

Valores nulos: 0 (0.00%)

Valores más frecuentes:

PAGADO: 3642  
PENDIENTE: 101  
ATRASADO: 88  
CANCELADO: 64  
CONVENIO: 1

Imagen que contiene Gráfico

El contenido generado por IA puede ser incorrecto.

## Variable: monto

Valores únicos: 19

Valores nulos: 0 (0.00%)

Estadísticas descriptivas:

Count: 3896.000000  
mean: 1286.255133  
std : 669.580457  
min: 250.000000  
25%: 800.000000  
50%: 1400.000000  
75%: 1600.000000  
max: 19100.000000

Gráfico

El contenido generado por IA puede ser incorrecto.

Tabla

El contenido generado por IA puede ser incorrecto.

## Variable: concepto\_original

Valores únicos: 40

Valores nulos: 0 (0.00%)

Valores más frecuentes:

may-24: 344  
ene-24: 343  
jun-24: 343  
abr-24: 343  
feb-24: 343  
mar-24: 343  
dic-23: 343  
nov-23: 343  
oct-23: 343  
sep-23: 343

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

## Variable: hasta el mes

Valores únicos: 15

Valores nulos: 0 (0.00%)

Rango de fechas:

De 2023-08-01 00:00:00 a 2025-02-01 00:00:00

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

## Variable: fecha\_creacion

Valores únicos: 137

Valores nulos: 0 (0.00%)

Rango de fechas:

De 2023-08-01 06:00:00 a 2024-06-01 12:00:02

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

## Variable: fecha\_pago

Valores únicos: 2879

Valores nulos: 253 (6.49%)

Rango de fechas:

De 2023-08-15 06:00:00 a 2024-06-17 01:50:32

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

## Variable: info\_recibo

Información básica:

Tipo de dato: object

Valores únicos: 12

Valores nulos: 0 (0.00%)

## Variable: estatus pago

Valores únicos: 2

Valores nulos: 0 (0.00%)

Valores más frecuentes:

PAGADO 3643  
SIN PAGO 253

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

## Variable: moroso

Valores únicos: 2

Valores nulos: 0 (0.00%)

Valores más frecuentes:

VIGENTE: 3707  
VENCIDO: 189

Gráfico

El contenido generado por IA puede ser incorrecto.

## Variable: numero de moras maxima

Valores únicos: 3

Valores nulos: 0 (0.00%)

Estadísticas descriptivas:

Count: 3896.000000  
mean: 0.537218  
std : 0.864446  
min: 0.000000  
25%: 0.000000  
50%: 0.000000  
75%: 1.000000  
max: 2.000000

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

Gráfico

El contenido generado por IA puede ser incorrecto.

## Variable: moras

Valores únicos: 3

Valores nulos: 0 (0.00%)

Estadísticas descriptivas:

count: 3896.000000  
mean: 0.071099  
std: 0.333537  
min: 0.000000  
25%: 0.000000  
50%: 0.000000  
75%: 0.000000  
max: 2.000000

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

Interfaz de usuario gráfica, Aplicación

El contenido generado por IA puede ser incorrecto.

## Días hasta el pago

Estadísticas descriptivas de días hasta el pago:

count: 3643.000000  
mean: 15.563547  
std: 14.814393  
min: 0.000000  
25%: 10.000000  
50%: 14.000000  
75%: 15.000000  
max: 107.000000

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

## Análisis por Tipo de Pago

Distribución por tipo de pago:

mantenimiento: 3836  
afavor: 60

Gráfico

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

## Evolución Temporal de Montos

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

## Análisis por Casa

Top 10 casas con más pagos:

Coto 1 Casa 20: 17  
Coto 1 Casa 104: 16  
Coto 1 Casa 133: 15  
Coto 1 Casa 43: 14  
Coto 1 Casa 144: 14  
Coto 1 Casa 39: 14  
Coto 1 Casa 198: 14  
Coto 1 Casa 179: 14  
Coto 1 Casa 64: 14  
Coto 1 Casa 56: 13

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

## Análisis por Coto

Distribución por coto:

Coto 1: 2679  
Coto 2: 1216  
Casa 49: 1

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

# Análisis univariado.

# Correlación

Gráfico, Gráfico de rectángulos

El contenido generado por IA puede ser incorrecto.

Al considerar la variable “moroso”, es decir si es vigente o vencido y siendo esta dependiente condicional de estatus y del pago, hay una correlación alta entre ellas.

# 2.2 Análisis bivariado: Tipo vs Moroso

Gráfico, Gráfico en cascada

El contenido generado por IA puede ser incorrecto.

## Distribución de la variable objetivo: Moroso

VIGENTE: 3707 (95.15%)

VENCIDO: 189 (4.85%)

## Distribución general

|  |  |  |
| --- | --- | --- |
| Cuenta total |  |  |
| ATRASADO | 88 | 2% |
| CANCELADO | 64 | 2% |
| CONVENIO | 1 | 0% |
| PAGADO | 3642 | 93% |
| PENDIENTE | 101 | 3% |
| Total general | 3896 | 100% |

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

El análisis general del estado de las transacciones revela que el 93% de las cuentas registradas se encuentran **pagadas**, lo que demuestra un alto nivel de cumplimiento por parte de los usuarios o propietarios del conjunto habitacional. Este porcentaje representa un desempeño positivo en términos de cobranza, evidenciando que la mayoría de los residentes están al corriente con sus obligaciones.

Sin embargo, un 3% de las transacciones está clasificado como **pendiente**, lo que indica posibles rezagos que podrían derivar en morosidad si no se gestionan adecuadamente. El 2% que aparece como **atrasado** es particularmente importante, ya que estas cuentas podrían estar en riesgo de impago crónico. Las categorías **cancelado** (2%) y **convenio** (0%) tienen una participación marginal, pero su existencia es útil para observar comportamientos financieros alternativos o mecanismos de regularización.

Aunque el escenario general es favorable, el segmento de transacciones pendientes y atrasadas debe ser vigilado cuidadosamente, ya que representa focos de riesgo que pueden comprometer el desempeño financiero en el mediano plazo.

## Vencidos vs Vigentes

|  |  |  |
| --- | --- | --- |
| Vencidos vs vigentes | |  |
| VENCIDO | 189 | 5% |
| VIGENTE | 3707 | 95% |
| Total general | 3896 | 100% |

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

En cuanto al estado de vigencia de las cuentas, se observa que **el 95% de las transacciones se encuentran vigentes**, lo que significa que han sido pagadas a tiempo o aún están dentro del periodo permitido para su liquidación. Solo **el 5% se encuentra vencido**, lo que representa 189 registros.

Aunque este porcentaje pueda parecer bajo, representa una población concreta de casos con alta probabilidad de entrar en morosidad prolongada si no se interviene. Por tanto, estos casos deben considerarse prioritarios para análisis de comportamiento y diseño de estrategias preventivas.

La morosidad en términos de vencimiento es contenida, pero los 189 casos vencidos representan una muestra ideal para analizar factores de riesgo y entrenar modelos predictivos de comportamiento futuro.

## Historial de moras acumuladas

|  |  |  |
| --- | --- | --- |
| Moras |  |  |
| 0 | 2774 | 71% |
| 1 | 151 | 4% |
| 2 | 971 | 25% |
| Total general | 3896 | 100% |

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Analizando el historial de moras, se observa que el 71% de las transacciones nunca ha presentado mora alguna, lo cual es un dato positivo y muestra un historial de cumplimiento estable para la mayoría de los casos.

Sin embargo, un **29% ha presentado al menos una mora**, con una proporción considerable del 25% acumulando **dos moras**. Esta situación no solo evidencia la existencia de incumplimientos repetidos, sino que además alerta sobre la presencia de un grupo significativo de usuarios que podrían estar en una dinámica de pagos intermitentes o estructuralmente problemáticos.

Aunque la mayoría cumple regularmente, hay una cuarta parte de los registros que ha caído dos veces en mora. Esto representa una señal de alerta para generar programas de seguimiento personalizado, políticas de recordatorio preventivo o medidas correctivas específicas.

## Distribución del número de moras entre los casos vencidos

|  |  |  |
| --- | --- | --- |
| Distribución moras | |  |
| 1 | 13 | 7% |
| 2 | 176 | 93% |
| Total general | 189 | 100% |

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Cuando se estudia exclusivamente a los casos vencidos (189 registros), se encuentra que **el 93% de ellos tiene dos moras acumuladas**, mientras que solo el 7% presenta una sola mora. Este hallazgo refuerza una hipótesis clave para el análisis predictivo: **la acumulación de dos moras parece ser el umbral a partir del cual una cuenta se vuelve significativamente riesgosa**.

Esto sugiere que el sistema de alertas o intervención debe activarse con urgencia al momento en que una cuenta acumula su **segunda mora**.

La acumulación de múltiples moras entre los vencidos confirma que el deterioro del comportamiento de pago ocurre rápidamente después de la primera falta. Por lo tanto, es crucial diseñar alertas que identifiquen y prioricen estos casos antes de llegar a ese punto.

## Cuenta por coto

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cuenta total por coto | Casa 49 | Coto 1 | Coto 2 | Total general | Coto 1 | Coto 2 |
| ATRASADO |  | 1 | 87 | 88 | 0% | 7% |
| CANCELADO |  | 58 | 6 | 64 | 2% | 0% |
| CONVENIO |  | 1 |  | 1 | 0% | 0% |
| PAGADO | 1 | 2606 | 1035 | 3642 | 97% | 85% |
| PENDIENTE |  | 13 | 88 | 101 | 0% | 7% |
| Total general | 1 | 2679 | 1216 | 3896 | 100% | 100% |

Imagen que contiene Gráfico

El contenido generado por IA puede ser incorrecto.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Vencidos vs vigentes por coto | Casa 49 | Coto 1 | Coto 2 | Total general | Coto 1 | Coto 2 |
| VENCIDO |  | 14 | 175 | 189 | 1% | 14% |
| VIGENTE | 1 | 2665 | 1041 | 3707 | 99% | 86% |
| Total general | 1 | 2679 | 1216 | 3896 | 100% | 100% |

El análisis por ubicación geográfica muestra diferencias relevantes entre los cotos. **Coto 1** destaca con un 97% de cuentas pagadas y solo 14 casos vencidos, lo que lo posiciona como el conjunto más estable y con menor perfil de riesgo.

En contraste, **Coto 2** refleja una realidad distinta: solo el 85% de las cuentas están pagadas, mientras que el 14% está vencido. Además, este coto concentra **el 93% de las cuentas con moras múltiples**, lo que lo convierte en el foco principal para intervenciones, políticas de seguimiento e incluso análisis de contexto social o económico.

Coto 1 es un conjunto ejemplar en términos de cumplimiento, mientras que Coto 2 requiere atención inmediata. Las acciones correctivas y estrategias de cobranza deben enfocarse principalmente en esta segunda zona.

## Distribución de moras por coto

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Moras por coto | Casa 49 | Coto 1 | Coto 2 | Total general | Coto 1 | Coto 2 |
| 0 | 1 | 2665 | 1041 | 3707 | 99% | 86% |
| 1 |  | 13 | 88 | 101 | 0% | 7% |
| 2 |  | 1 | 87 | 88 | 0% | 7% |
| Total general | 1 | 2679 | 1216 | 3896 | 100% | 100% |

De los 88 casos con dos moras:

* **87** pertenecen a Coto 2.
* Solo **1** a Coto 1.

De los 101 casos con una mora:

* **88** están en Coto 2.
* **13** en Coto 1.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

Esto confirma que **Coto 2 no solo concentra los casos vencidos, sino que también agrupa casi la totalidad de las moras**, tanto simples como múltiples. Esta concentración debe tomarse en cuenta para diseñar intervenciones focalizadas y estrategias de seguimiento más personalizadas.

La morosidad no está distribuida uniformemente entre los cotos. Coto 2 es responsable del 98% de los casos con problemas de pago, lo que refuerza la necesidad de segmentar los modelos de predicción por zona y aplicar criterios diferenciados.

## Relación entre moras y número de viviendas afectadas

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cuantos en cuantas moras por coto | Coto 1 | Coto 2 | Total general | Coto 1 | Coto 2 |
| 1 | 13 | 88 | 101 | 93% | 50% |
| 2 | 1 | 87 | 88 | 7% | 50% |
| Total general | 14 | 175 | 189 | 100% | 100% |

En Coto 2:

* El 50% de las casas con moras tiene dos moras.
* El otro 50% tiene una mora.

En cambio, en Coto 1:

* El 93% de las casas vencidas tiene solo una mora.
* Solo el 7% ha acumulado dos moras.

Este dato revela que la morosidad en **Coto 1** es más leve y potencialmente recuperable, mientras que en **Coto 2** hay mayor recurrencia de incumplimiento.

Las estrategias de gestión de morosidad deben ser proporcionales a la gravedad del incumplimiento. En Coto 1, medidas preventivas pueden ser suficientes, mientras que en Coto 2 se requerirán esquemas más estructurados de cobranza y seguimiento.

## Desequilibrio en las clases de la variable objetivo

Sí. La variable objetivo Moroso presenta un **claro desbalance de clases**:

* Aproximadamente **70–80% de los registros son “NO” morosos**, mientras solo el **20–30% son “SI”**.
* Este desbalance podría hacer que los modelos predictivos tiendan a **clasificar incorrectamente los casos positivos**, ya que el modelo puede obtener una precisión alta simplemente prediciendo la clase mayoritaria.

Para abordar esto, se deben considerar estrategias como:

* **Resampling**: sobremuestreo de la clase minoritaria (ej. SMOTE) o submuestreo de la mayoritaria.
* **Ajuste de pesos en el modelo**: muchos algoritmos permiten balancear el impacto de cada clase mediante class\_weight.

## Implicaciones para modelado y gestión de riesgo

Con base en estos hallazgos, se identifican múltiples oportunidades para mejorar la gestión de cartera vencida:

* **Segmentación por coto:** Los modelos predictivos deben considerar la ubicación (Coto 1 o Coto 2) como una variable crítica, debido a las diferencias de comportamiento claramente observables.
* **Poca información.**
* **Variables clave para modelos de IA:**
  + Número de moras históricas
  + Estado actual de la transacción
  + Coto
  + Vencido/Vigente
  + Historial de pagos
* **Diseño de alertas:** Activar sistemas de alerta automática cuando una cuenta pase de 0 a 1 mora y con prioridad al llegar a 2 moras.

1. **Transformaciones necesarias:** Las variables categóricas pueden ser codificadas (e.g., mora\_0, mora\_1, mora\_2), y algunas transformaciones logarítmicas o categóricas pueden ser útiles para manejar la asimetría en la distribución.
2. **Acciones correctivas específicas por coto:** Diseñar estrategias personalizadas para Coto 2, como visitas domiciliarias, convenios de pago y mayor frecuencia en recordatorios.

# 2.3 Conclusiones

* No hay valores faltantes críticos, salvo en 'fecha\_pago'. Esto podría indicar transacciones aún no saldadas.
* 'monto' y 'moras' presentan distribuciones sesgadas, con algunos valores extremos. Variables como 'tipo', 'status' y 'Moroso' muestran patrones interesantes para predicción.
* Existe un desbalance moderado en la variable 'Moroso', que debe considerarse en el modelado.
* Se recomienda transformar variables categóricas en dummies y normalizar variables numéricas.
* La mayoría de las columnas están bien estructuradas y tipificadas; sin embargo, se identificaron valores **faltantes en la columna fecha\_pago**, lo que sugiere que ciertas transacciones aún no han sido saldadas o no tienen un registro de pago actualizado. Este comportamiento puede ser útil para identificar morosidad potencial futura.
* No se encontraron valores nulos explícitos en otras columnas, pero sí hay **valores categóricos ambiguos** o con etiquetas parecidas que podrían requerir estandarización.
* **monto** muestra una **distribución sesgada a la derecha**, con presencia de algunos valores atípicos o extraordinariamente altos. Esto sugiere que hay transacciones ocasionales con montos muy superiores al promedio. Este comportamiento puede impactar negativamente en modelos sensibles a la escala.
* Las variables **moras** y **numero de moras maxima** presentan **una concentración alta en valores bajos (0 o 1)**, pero también algunos valores altos, lo que indica que aunque la mayoría de los residentes pagan a tiempo, hay un grupo consistente con historial recurrente de moras.
* El campo **tipo** tiene dos categorías dominantes: "afavor" y "mantenimiento". Su análisis cruzado con la variable Moroso sugiere que los movimientos "afavor" casi no presentan morosidad, mientras que los de "mantenimiento" sí, lo que los convierte en una característica valiosa para la predicción.
* Existe una **correlación positiva entre numero de moras maxima y moras**, lo cual es esperable, pero también confirma que ambas variables están relacionadas directamente con el comportamiento de pago a lo largo del tiempo.
* Aunque **monto** no está fuertemente correlacionado con las moras, sí es importante monitorearlo para evaluar la severidad de las deudas.
* El análisis de **frecuencia entre status y moroso** sugiere que los estados "pendiente" y "atrasado" son los principales predictores de morosidad, mientras que estados como "pagado" o "condonado" no muestran morosidad, por definición.
* Hay un **desequilibrio moderado en las clases** de la variable Moroso, donde la mayoría de los registros corresponden a "NO" (no moroso). Este **desbalance de clases** deberá ser abordado durante el entrenamiento de modelos predictivos, usando técnicas como **SMOTE, ponderación de clases o muestreo estratificado**.
* Variables como **coto y casa** tienen una cardinalidad muy alta. Aunque son importantes para la identificación individual de cada vivienda, pueden **introducir ruido o sobreajuste** si se codifican directamente. Se recomienda excluir casa como predictor o bien codificar coto en función de su relación con la morosidad promedio.
* Las variables categóricas como tipo, status, Estatus pago y Moroso son **candidatas ideales para codificación dummy (one-hot encoding)**, al ser de baja cardinalidad y contener información predictiva importante.
* Aunque no se analizó en profundidad, las variables **fecha\_creacion y fecha\_pago** permiten explorar patrones de morosidad a lo largo del tiempo, como estacionalidad en pagos atrasados. Esta línea de análisis puede ser fundamental para pronósticos temporales o decisiones estratégicas (por ejemplo, mayor morosidad en ciertos meses del año).
* Las siguientes variables muestran un **alto potencial predictivo para un modelo de clasificación de morosidad**:
  + tipo
  + status
  + monto
  + Numero de moras maxima
  + moras
  + Estatus pago
* Variables como casa deben usarse solo como identificadores, no como predictores.
* Se recomienda separar el análisis en dos enfoques: uno **cross-sectional** (por registro) y otro **longitudinal por casa**, para capturar patrones persistentes.

# 3. Análisis Exploratorio – Chatbot de Asistencia Virtual

## 3.1 Objetivo

El objetivo principal del chatbot es ofrecer una herramienta conversacional que permita a residentes y administradores de condominios resolver dudas comunes y obtener información confiable de forma inmediata. Para ello, se realizó un EDA sobre el comportamiento de los mensajes enviados por los usuarios y se analizaron los contenidos disponibles en el manual de usuario y el documento de funcionalidades dictadas.

## 3.2 Fuentes de datos

Para este análisis se integraron tres fuentes principales:

* **mensajes\_65fe1e0b76945.csv**: archivo exportado desde la plataforma que contiene mensajes escritos por usuarios.
* **Manual de Usuario**: documento PDF oficial que describe el funcionamiento de MiCoto.
* **Funcionalidad MiCoto (dictado)**: archivo PDF con descripción actualizada de funcionalidades no reflejadas en el manual.

## 3.3 Análisis de Mensajes

Se extrajo el contenido textual de los mensajes utilizando expresiones regulares y conversión de estructuras JSON anidadas. Con esto, se consolidaron las conversaciones en una sola columna llamada mensaje\_texto, permitiendo el análisis textual unificado. Las variables clave fueron:

* asunto: tema general del mensaje
* usuario\_uno: autor del mensaje (residente o administrador)
* fecha\_creacion: fecha del mensaje

Se aplicó un análisis de frecuencia sobre los “asuntos” y se generó una nube de palabras con las expresiones más comunes.

A close-up of words

AI-generated content may be incorrect.

**Figura 1.** Nube de palabras basada en los mensajes enviados por usuarios.

A graph of a bar graph

AI-generated content may be incorrect.

**Figura 2.** Asuntos más frecuentes identificados en el archivo de mensajes.

## 3.4 Análisis de Manual y Dictado

A partir del Manual de Usuario y el documento dictado, se realizó un análisis de frases clave con técnicas básicas de NLP para detectar intenciones relacionadas con verbos de acción (consultar, cambiar, enviar, apartar, registrar, descargar, etc.).

Ambas fuentes fueron procesadas con PyMuPDF, permitiendo la extracción del texto y su segmentación en frases útiles para alimentar futuras intenciones del chatbot.

## 3.5 Intenciones frecuentes identificadas

A continuación, se presenta una tabla con ejemplos de frases clave representativas de cada módulo, extraídas del análisis de los mensajes reales, el manual y el dictado:

|  |  |  |
| --- | --- | --- |
| **Módulo** | **Intención identificada** | **Frase ejemplo extraída** |
| Mi Cuenta | Cambiar contraseña | ¿Cómo cambio mi contraseña? |
|  | Actualizar datos | ¿Dónde puedo actualizar mi teléfono? |
|  | Subir constancia fiscal | ¿Dónde se sube la constancia fiscal? |
| Finanzas | Consultar estado de cuenta | ¿Dónde veo mi saldo o mis adeudos? |
|  | Registrar pagos | ¿Cómo informo que ya pagué? |
|  | Generar factura | ¿Dónde agrego mis datos fiscales? |
| Áreas comunes | Apartar áreas | ¿Cómo aparto el salón de eventos? |
|  | Cancelar reservación | Quiero cancelar una reservación |
| Mensajes | Enviar ticket | Envié un mensaje y no me han respondido |
|  | Ver mensajes del administrador | ¿Dónde veo lo que escribió el administrador? |
| Documentos | Consultar actas | ¿Dónde están las actas de la asamblea? |
|  | Consultar reglamento | ¿Dónde está el reglamento del condominio? |

La consolidación de las tres fuentes permitió clasificar las intenciones frecuentes según módulos:

* **Mi Cuenta**: cambiar contraseña, actualizar datos, subir constancia fiscal
* **Finanzas**: consultar estado de cuenta, registrar pagos, generar factura
* **Áreas comunes**: apartar áreas, cancelar reservación
* **Mensajes**: enviar ticket, ver mensajes del administrador
* **Documentos**: consultar actas, reglamento

## 3.6 Visualizaciones complementarias del análisis de intenciones

Con el objetivo de reforzar el análisis de intenciones, se generaron visualizaciones adicionales a partir de las frases clave detectadas en el Manual de Usuario y el Dictado de Funcionalidades:

* **Nube de palabras exclusiva de frases clave**: Se creó una nube de palabras a partir de frases que contienen verbos de acción como “consultar”, “registrar”, “cambiar”, entre otros. Esta nube refleja de forma visual la frecuencia de temas e intenciones expresadas directamente en los documentos guía de la plataforma.

A close-up of words

AI-generated content may be incorrect.

**Figura 3.** Nube de palabras generada exclusivamente con frases clave extraídas del manual y dictado.

* **Frecuencia de palabras en frases clave**: Se realizó un análisis cuantitativo de las palabras más comunes dentro de las frases clave, lo que permite visualizar de forma más precisa las acciones más recurrentes en el uso de la plataforma.

A graph of different colored bars

AI-generated content may be incorrect.

**Figura 4.** Palabras más frecuentes en frases que representan intenciones clave del usuario.

* **Bigramas más frecuentes**: Se identificaron combinaciones de dos palabras más comunes (bigrama) dentro de las frases clave. Estas combinaciones como “estado cuenta” o “cambiar contraseña” sirven como guía directa para alimentar la base de entrenamiento del chatbot.

A graph with different colors

AI-generated content may be incorrect.

**Figura 5.** Bigrama más frecuentes en las frases de intención, útiles para entrenamiento y categorización.

## 3.6 Conclusiones del EDA del Chatbot

* Las consultas de los usuarios se centran en operaciones básicas y frecuentes que pueden ser anticipadas y estructuradas.
* La coincidencia entre lo expresado por los usuarios en mensajes reales y lo documentado en el manual permite construir un marco de referencia confiable para las intenciones del chatbot.
* La integración del chatbot debe ir acompañada de accesos visibles y claros dentro de la aplicación para incentivar su uso y reducir fricciones en la experiencia de usuario.

Este análisis permitirá construir una base sólida de intenciones y respuestas esperadas, organizadas por categoría, que podrá integrarse con herramientas de NLP para entrenar un chatbot funcional dentro del ecosistema de MiCoto.

# Fuentes:

Hernández-Sampieri, R., & Mendoza, C. (2023). Metodología de la investigación: Las rutas de la investigación cuantitativa, cualitativa y mixta. McGraw-Hill. <https://0-bc-vitalsource-com.biblioteca-ils.tec.mx/tenants/BIB_TECDEMTY/libraries?bookmeta_vbid=9786071520326>

Mukhiya, S., & Ahmed, U. (2020). Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. Packt Publishing.

Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A., & Plöd, M. (2023). CRISP-ML(Q): The ML Lifecycle Process. INNOQ. <https://ml-ops.org/content/crisp-ml>

**MiCoto.** (2024). Manual de Usuario de la plataforma MiCoto. Documento interno (PDF).

**MiCoto.** (2024). Descripción funcional de MiCoto (dictado). Documento interno (PDF).

**MiCoto.** (2024). Base de datos de mensajes de usuarios (archivo CSV interno). Exportado desde la plataforma MiCoto para fines de análisis.

**MiCoto.mx.** (s.f.). *¿Quiénes somos?* Recuperado el 01/5/2025 de: <https://micoto.mx/#quienes-somos>