Universidad de Antioquia
Departamento de Ingeniera

Análisis de riesgo crediticio

Juan Pablo González Blandón
Juan Felipe Isaza Valencia
Alexis de Jesús Collante Genes
Jorge Antonio Álvarez Sayas

08 de Agosto de 2025

Resumen Ejecutivo

El presente informe describe el desarrollo de un proyecto de análisis de riesgo crediticio utilizando el conjunto de datos *Home Credit Default Risk*, disponible en Kaggle. El objetivo principal fue identificar los factores que influyen en el incumplimiento de crédito, analizar los perfiles de clientes y construir un modelo predictivo que permita anticipar dicho riesgo. Para ello, se realizó un proceso de limpieza, exploración y modelado, aplicando técnicas de análisis de datos y aprendizaje automático, con el fin de proponer mejoras en los procesos de evaluación crediticia.

1. Introducción

En el contexto actual, las instituciones financieras enfrentan el desafío de evaluar de manera precisa el riesgo de incumplimiento de crédito. Este análisis es fundamental para minimizar pérdidas y optimizar la toma de decisiones. El proyecto desarrollado se enfoca en el conjunto de datos *Home Credit Default Risk*, el cual contiene información detallada de clientes, historial de pagos y solicitudes previas de crédito.

A lo largo de este trabajo se busca entender cómo las variables socioeconómicas, demográficas y financieras se relacionan con la probabilidad de incumplimiento. Asimismo, se pretende construir un modelo de predicción capaz de anticipar el comportamiento de nuevos solicitantes, contribuyendo así a una evaluación crediticia más informada y eficiente.

2. Objetivos

2.1. Objetivo general

Analizar los datos del conjunto *Home Credit Default Risk* con el fin de identificar los factores que influyen en el riesgo de incumplimiento de crédito, comprender los perfiles de clientes y construir un modelo predictivo para anticipar dicho riesgo.

2.2. Objetivos específicos

- Identificar las variables más relevantes que influyen en el riesgo de incumplimiento crediticio, prestando atención a factores socioeconómicos como nivel de ingresos, tipo de empleo o educación.
- Analizar perfiles de clientes según su comportamiento crediticio para detectar patrones comunes entre quienes cumplen y quienes incumplen sus pagos.
- Aplicar técnicas de análisis exploratorio de datos (EDA) para limpiar, transformar y visualizar la información.
- Desarrollar un modelo de aprendizaje automático que prediga la probabilidad de incumplimiento en nuevos solicitantes.
- Evaluar el desempeño del modelo mediante métricas como *precision*, *recall* y AUC, optimizando su rendimiento.
- Extraer conclusiones e *insights* que sirvan como apoyo en procesos de evaluación crediticia.

3. Descripción de la base de datos

La base de datos utilizada fue obtenida de la plataforma Kaggle, específicamente del reto *Home Credit Default Risk*. Este conjunto busca mejorar la forma en que las instituciones financieras evalúan el riesgo crediticio, proporcionando información del préstamo actual y del historial financiero de los solicitantes.

La estructura del dataset incluye múltiples tablas relacionadas, siendo `application_train.csv` la principal, con datos personales y financieros junto con el indicador `TARGET` que señala incumplimiento. Existen además tablas como `bureau.csv`, `bureau_balance.csv`, `previous_application.csv`, `POS_CASH_balance.csv`, `installments_payments.csv` y `credit_card_balance.csv`, todas enlazadas mediante las claves `SK_ID_CURR` y `SK_ID_PREV`.

Durante la revisión se detectó la presencia de valores faltantes y estructuras desbalanceadas, lo cual hace necesaria una etapa de limpieza y transformación previa al análisis. La diversidad de productos financieros incluidos (tarjetas, préstamos POS, créditos de consumo) brinda un panorama amplio del comportamiento de los clientes.

4. Metodología

4.1. Adquisición de datos

Los datos fueron descargados directamente desde la página del reto en Kaggle. Se obtuvo un archivo comprimido (`.zip`) que contiene las tablas en formato CSV; dichas tablas se importaron a una instancia MySQL mediante la terminal para facilitar el acceso, la gestión y el procesamiento posterior.

Se importaron las tablas provistas por el reto, entre ellas: `application_train.csv`, `application_test.csv`, `bureau.csv`, `bureau_balance.csv`, `previous_application.csv`, `POS_CASH_balance.csv`, `installments_payments.csv` y `credit_card_balance.csv`.

Estructura de almacenamiento (Bronze / Silver / Gold)

- **Bronze:** Datos crudos tal cual como fueron entregados por Kaggle (CSV originales). Se mantienen sin modificaciones para preservar la fuente original.
- **Silver:** Datos procesados y limpios (imputación de valores faltantes, conversión de tipos, normalizaciones y tratamiento de outliers). Estas tablas son las que se utilizan para el análisis exploratorio y el modelado.
- **Gold:** Tablas agregadas y preparadas específicamente para el consumo por dashboards y por los modelos de ML (con features finales, joins precomputados y agregados relevantes).

A modo de ejemplo, la importación masiva de un CSV a MySQL mediante la terminal puede realizarse con un comando similar al siguiente:

```
LOAD DATA LOCAL INFILE '/ruta/a/application_train.csv'
INTO TABLE application_train
FIELDS TERMINATED BY ',' OPTIONALLY ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
```

4.2. Limpieza de datos

La fase de limpieza de datos fue realizada de manera colaborativa por todos los integrantes del equipo. El objetivo principal fue asegurar la calidad, consistencia y coherencia de la información antes de proceder con el análisis y modelado.

Entre las acciones realizadas se destacan:

- **Conversión de tipos de datos:** Ajuste de cada columna a su tipo de dato correcto (numérico, fecha, cadena, categórico, etc.) para evitar errores en el procesamiento.
- **Ajuste de montos monetarios:** Aproximación de valores a dos cifras decimales, siguiendo el estándar monetario europeo utilizado en el reto.
- **Eliminación de columnas redundantes:** Se eliminaron aquellas columnas cuyo valor podía obtenerse a partir de otras o que no aportaban información adicional.
- **Tratamiento de valores atípicos (outliers):**
  - En muchos casos, los outliers numéricos fueron reemplazados por `0` para evitar sesgos en el análisis.
  - Otros outliers fueron evaluados como indicadores de clientes potencialmente riesgosos, registrando esta información como parte del análisis de comportamiento.
- **Eliminación de columnas sin valor informativo:** Variables irrelevantes o sin variabilidad fueron eliminadas.
- **Estandarización de nombres de columnas:** Renombrado de campos para mantener la convención de `SK_ID_CURR` (ID de cliente) y `SK_ID_PREV` (ID de crédito previo) en todas las tablas.

Todas las tablas resultantes de esta etapa fueron almacenadas en la capa **Silver**, quedando listas para su uso en el análisis exploratorio.

4.3. Análisis exploratorio de datos (EDA)

La etapa de análisis exploratorio se desarrolló directamente sobre la base de datos en MySQL, trabajando principalmente con las tablas en la capa **Silver** y generando consultas avanzadas para extraer variables clave que luego fueron integradas en la capa **Gold**. El objetivo fue comprender el comportamiento crediticio de los clientes y derivar métricas que permitieran modelar el riesgo de incumplimiento.

Integración y análisis de `bureau` La tabla `bureau` contiene el historial de créditos que los clientes tienen con otras instituciones financieras, enlazada a `bureau_balance` que detalla el estado mensual de esos créditos.

- Se identificaron los estados de crédito más frecuentes (`Closed`, `Active`, `Sold`, `Bad debt`), detectando que la mayoría de registros corresponden a créditos cerrados, pero con una proporción relevante de cuentas activas.
- Se calcularon métricas agregadas por cliente, como el número total de créditos activos, la suma de saldos y la antigüedad promedio de los créditos.
- Se analizó la distribución de días de atraso (`DAYS_CREDIT` y `CREDIT_DAY_OVERDUE`) para detectar patrones de morosidad interinstitucional.

Procesamiento de `POS_CASH_balance` Esta tabla registra información mensual de contratos tipo POS y de créditos al consumo.

- Se calcularon proporciones de estados contractuales (`Active`, `Completed`, `Signed`) para cada cliente.
- Se generó una métrica de estabilidad crediticia a partir del conteo de cambios de estado a lo largo del tiempo.
- Se evaluó la relación entre el número de meses con estado `Active` y la probabilidad de incumplimiento.

Análisis de `previous_application` Esta tabla contiene solicitudes previas de crédito con Home Credit, aprobadas o rechazadas.

- Se clasificaron los contratos según el resultado final (`Approved`, `Refused`, `Canceled`, `Unused offer`).
- Se analizaron los montos solicitados vs. montos aprobados para detectar discrepancias y patrones de rechazo en solicitudes de alto valor.
- Se calculó la tasa de aprobación por canal de solicitud (`NAME_TYPE_SUITE` y `CHANNEL_TYPE`), identificando canales presenciales como los de mayor efectividad.
- Se observó variabilidad en las tasas de aprobación por día de la semana, siendo fines de semana los días con mayor aceptación.

Variables derivadas clave Como resultado del EDA, se generaron variables que capturan el comportamiento crediticio de forma resumida:

- `FRAC_LATE_INSTALLMENTS`: proporción de cuotas pagadas con retraso.
- `MAX_DAYS_LATE`: mayor atraso registrado para un cliente.
- `AVG_UTILIZATION_RATIO_TDC`: uso promedio de la línea de crédito en tarjetas.
- `BUREAU_ACTIVE_COUNT`: número de créditos activos en otras entidades.
- `POS_CASH_ACTIVE_MONTHS`: meses en estado activo para contratos POS.
- `APPROVAL_RATE_PREV`: tasa de aprobación histórica en solicitudes previas.

Consolidación en la capa Gold Todas las métricas anteriores se unificaron mediante `MERGE` sobre la clave `SK_ID_CURR o incluyendo segun el caso SK_ID_PREV`, generando tablas **Gold** optimizadas para su uso en el dashboard y el modelado de riesgo. Estas tablas, ya limpias y enriquecidas, permitieron acelerar las fases posteriores y reducir la complejidad de procesamiento en tiempo real.

5. Visualizaciones del Dashboard y Análisis

En esta sección se presentan las principales visualizaciones desarrolladas en el dashboard interactivo, construido a partir de las tablas **Gold**. Cada gráfico se acompaña de una breve descripción e interpretación de los resultados.

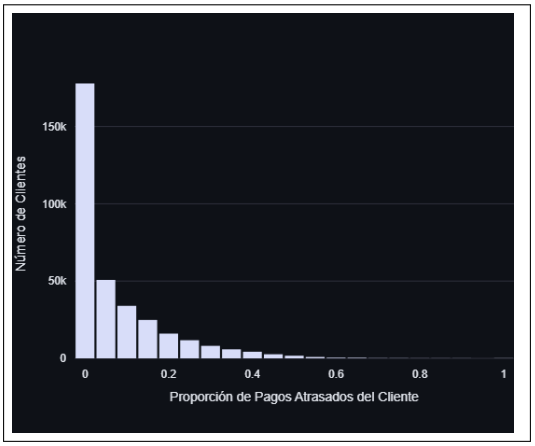

Figura 1. Proporción de pagos atrasados

Histograma que muestra la fracción de pagos atrasados por cliente. El alto pico en `0.0` refleja que la mayoría de clientes cumplen puntualmente con sus obligaciones. Sin embargo, la cola derecha del gráfico indica la presencia de clientes con atrasos recurrentes, siendo los valores cercanos a 1.0 representativos de comportamiento crónico de morosidad. Esta distribución evidencia que, aunque el grueso de la cartera es de bajo riesgo, existe un segmento minoritario de alto riesgo que podría impactar en la estabilidad crediticia.

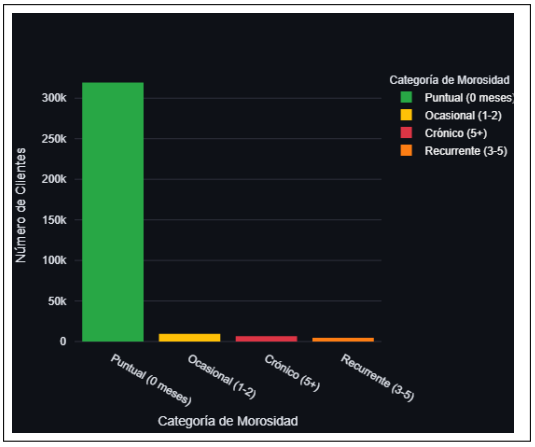

Figura 2. Distribución por categoría de morosidad

Gráfico de barras que clasifica a los clientes según el número de meses en mora. El grupo `Puntual (0 meses)` concentra la gran mayoría, indicando un alto nivel de cumplimiento. La categoría `Ocasional (1-2)` refleja casos esporádicos, posiblemente asociados a olvidos o problemas temporales de liquidez. En contraste, los segmentos `Recurrente (3-5)` y `Crónico (5+)` representan un riesgo elevado, dado que evidencian patrones persistentes de impago. La clara concentración en el grupo puntual evidencia una cartera predominantemente sana, aunque la existencia de un núcleo reducido pero significativo de morosos crónicos requiere vigilancia activa.

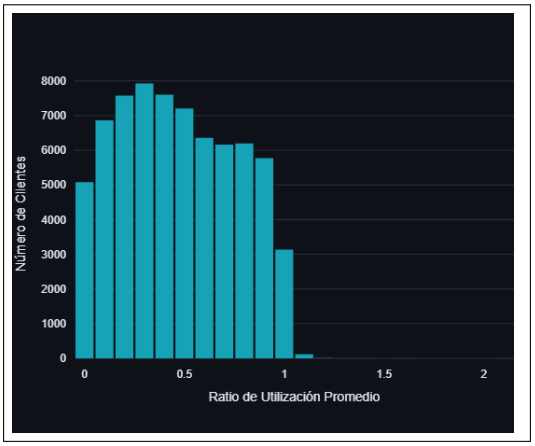

Figura 3. Distribución del ratio de utilización de crédito

Histograma que muestra la proporción promedio del límite de crédito utilizada por los clientes. Se observa una concentración mayoritaria en valores inferiores al `70%` (`0.7`), lo que indica un uso moderado y controlado del crédito. Los clientes con ratios cercanos o superiores a `1.0` representan un mayor estrés financiero y un riesgo potencial de impago, dado que están utilizando todo su límite o más. El sesgo hacia la izquierda sugiere que la mayoría mantiene un comportamiento de utilización saludable, aunque el segmento de alta utilización merece seguimiento para prevenir incumplimientos.

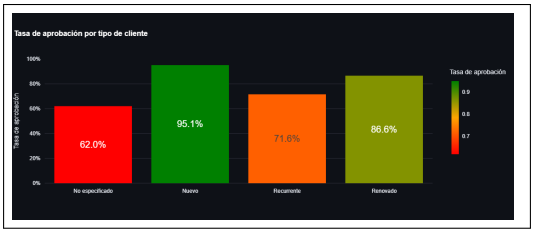

Figura 4. Tasa de aprobación por tipo de cliente

Clientes no especificados tienen la tasa de aprobación más baja, posiblemente por problemas de calidad de datos o menor confiabilidad. Los clientes nuevos presentan la tasa más alta, lo que sugiere políticas de entrada flexibles o evaluación optimista. Los clientes recurrentes muestran una tasa moderada, indicando mayor escrutinio sobre su historial, y los clientes renovados mantienen una tasa alta, reflejando confianza basada en experiencia previa. Estos patrones pueden guiar ajustes en la segmentación y evaluación de riesgo, especialmente en el tratamiento de clientes sin clasificación clara y en la validación del desempeño de nuevos perfiles.

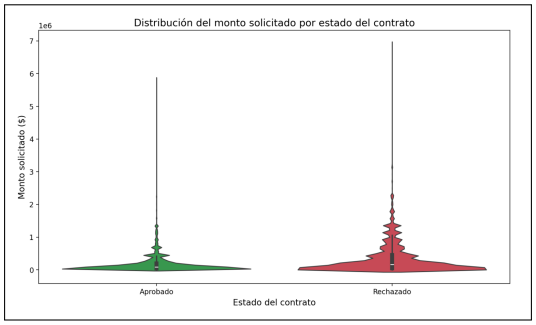

Figura 5. Distribución del monto solicitado por estado del contrato

Contratos aprobados se concentran principalmente en montos bajos, con una mediana significativamente menor.

Esto sugiere que los créditos de menor cuantía tienen mayor probabilidad de ser aprobados. Los contratos rechazados presentan mayor dispersión, una mediana más alta y valores extremos frecuentes, lo que indica que solicitudes por montos elevados tienden a ser rechazadas. La distribución visual muestra que los rechazos se agrupan en rangos intermedios y altos, mientras que los aprobados predominan en montos bajos. Este comportamiento sugiere la necesidad de establecer umbrales más definidos para los montos solicitados y revisar los criterios de aprobación en función del riesgo asociado a créditos de mayor valor.

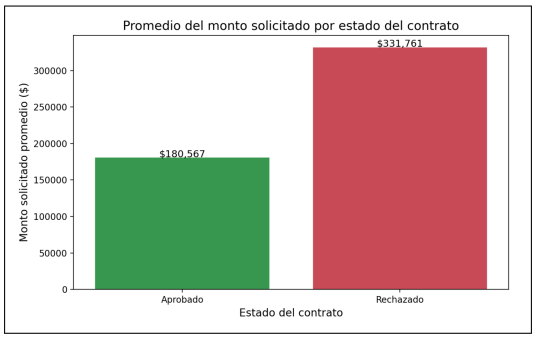

Figura 6. Promedio del monto solicitado por estado del contrato

Contratos aprobados tienen un monto solicitado promedio de 180,567, lo que indica que las solicitudes más modestas tienen mayor probabilidad de ser aceptadas. En contraste, los contratos rechazados presentan un promedio mucho más alto, de 331,761, sugiriendo que los montos elevados están más asociados al rechazo, posiblemente por el riesgo financiero que implican. La diferencia entre ambos promedios refleja una política conservadora, donde los créditos grandes enfrentan mayor escrutinio. Este patrón puede servir para ajustar los criterios de evaluación, estableciendo límites más claros o segmentando las solicitudes por rangos de monto para optimizar el proceso de aprobación.

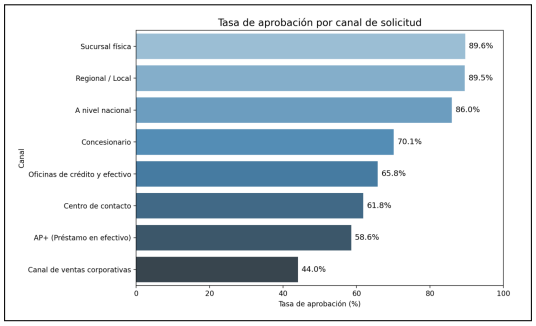

Figura 7. Tasa de aprobación por canal de solicitud

Canales presenciales como Sucursal física y Regional / Local lideran con tasas superiores, confirmando su efectividad en la aprobación de solicitudes. Por otro lado, los canales no presenciales, como Centro de contacto y AP+ (Préstamo en efectivo), muestran tasas más bajas, posiblemente por la falta de interacción directa. Por último, el canal de ventas corporativas presenta la tasa más baja, lo que podría deberse a criterios más exigentes o perfiles de cliente con mayor riesgo. Identificar los canales con mayor efectividad permite optimizar estrategias comerciales, asignar recursos de forma más eficiente y diseñar campañas enfocadas en los canales con mayor potencial de aprobación.

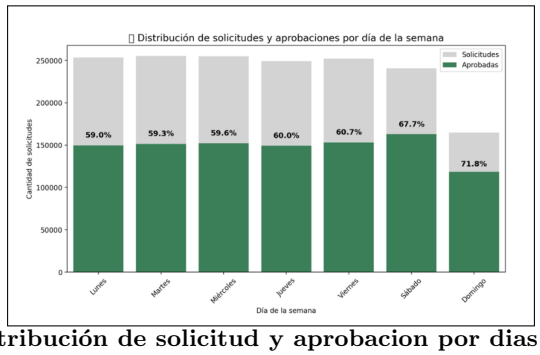

Figura 8. Distribución de solicitud y aprobacion por dias de la semana

Fines de semana presentan las tasas más altas, lo que sugiere decisiones más favorables en esos días; mientras, los días hábiles mantienen tasas más estables, con un leve repunte el viernes, posiblemente por mayor rigurosidad en la evaluación durante la semana. Aunque la mayoría de las solicitudes se concentran entre lunes y viernes, los fines de semana muestran una mayor proporción de aprobaciones, lo que podría reflejar un cambio en el perfil del solicitante o en la política de evaluación. Estos patrones pueden servir para ajustar la estrategia de atención, redistribuir recursos operativos y diseñar campañas que aprovechen los días con mayor probabilidad de aprobación.

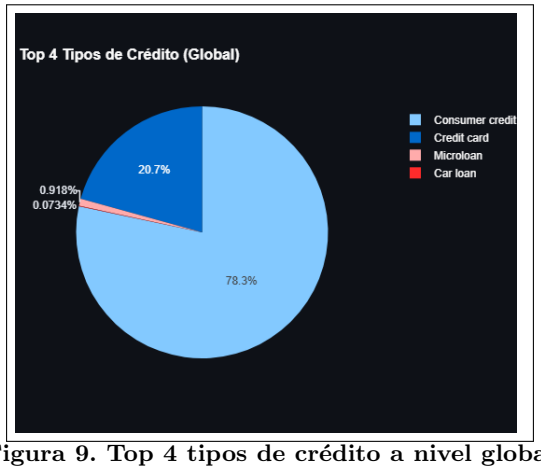

Figura 9. Top 4 tipos de crédito a nivel global

Gráfico circular muestra que el crédito de consumo domina el portafolio con un 78.3 %, seguido por las tarjetas de crédito con un 20.7 %. Los microcréditos (0.918 %) y préstamos de auto (0.0734 %) representan una fracción mínima del total, evidenciando que la oferta está fuertemente concentrada en productos de consumo masivo. Esta concentración sugiere una estrategia centrada en créditos de amplio alcance, aunque implica dependencia de un solo segmento.

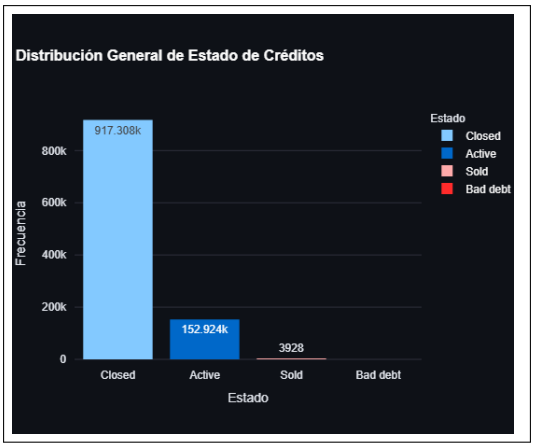

Figura 10. Distribución general del estado de créditos

Gráfico de barras muestra que la gran mayoría de créditos se encuentran cerrados (917,308), seguidos por créditos activos (152,924). Las categorías de créditos vendidos (3,928) y deudas incobrables son marginales. Esta distribución indica un alto nivel de ciclo completo en las operaciones, con baja incidencia de morosidad severa, pero también con un volumen moderado de cartera activa que representa oportunidades de gestión y retención de clientes.

6. Modelos Predictivos

Para la etapa de modelado se construyeron dos modelos de *Random Forest* con el fin de clasificar el riesgo crediticio de dos segmentos de clientes:

1. **Clientes no registrados**: sin historial financiero previo con la entidad, pero con información sociográfica y de comportamiento general.

2. **Clientes registrados**: con historial financiero registrado y mayor detalle de transacciones y comportamiento crediticio.

En ambos casos, el flujo metodológico fue idéntico, diferenciándose únicamente en las variables de entrada utilizadas para cada segmento. El proceso general fue el siguiente:

1. **Preparación de datos**: a partir de la capa *gold* del pipeline de datos, se seleccionaron las características relevantes para cada segmento. Para clientes no registrados, la información se basó en variables de segmentación sociográfica y comportamental. Para clientes registrados, las variables incluyeron principalmente datos financieros y de historial crediticio.

2. **Segmentación inicial (K-Means)**: se aplicó un modelo de *K-Means* para agrupar a los clientes en clústeres homogéneos, con el fin de explorar patrones y potencialmente mejorar la discriminación del modelo posterior. Este paso ayudó a identificar perfiles de clientes con características similares.

3. **Entrenamiento del modelo**: se implementó un clasificador *Random Forest* para predecir el riesgo crediticio. Este algoritmo fue elegido por su capacidad para manejar conjuntos de datos con variables heterogéneas y por ofrecer interpretabilidad a través de la importancia de variables.

4. **Evaluación**: el rendimiento de cada modelo fue evaluado mediante la matriz de confusión y el análisis de las variables más significativas en la decisión del clasificador.

6.1. Resultados para Clientes No Registrados

La Figura 1 muestra la matriz de confusión y las variables más relevantes para el modelo entrenado con información sociográfica y de comportamiento.

Figura 1: Resultados del modelo *Random Forest* para clientes no registrados.

6.2. Resultados para Clientes Registrados

La Figura 2 presenta la matriz de confusión y la importancia de las variables para el modelo entrenado con información financiera detallada.

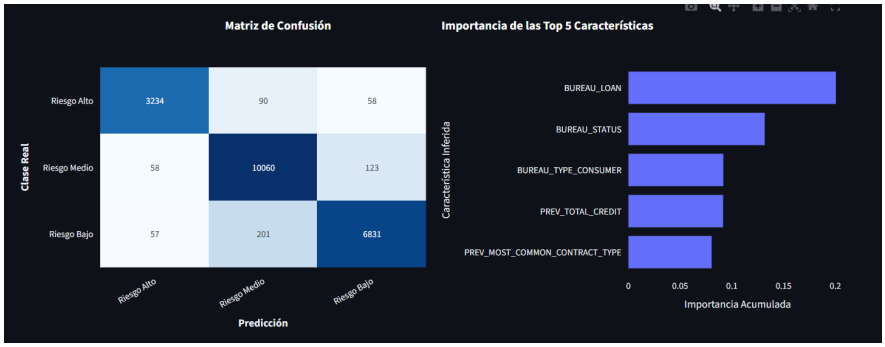

Figura 2: Resultados del modelo *Random Forest* para clientes registrados.

7. Conclusiones

1. Se identificaron variables clave como `AMT_INCOME`, `NAME_FAMILY_STATUS`, `NAME_HOUSING_TYPE` y `FLAG_OWN_CAR` para clientes no registrados. Para clientes registrados, destacaron `BUREAU_LOAN_AMOUNT`, `BUREAU_STATUS`, `PREV_TOTAL_CREDIT` y `PREV_MOST_COMMON_CONTRACT_TYPE`. Estas variables reflejan tanto el perfil socioeconómico como el comportamiento financiero, siendo determinantes en la predicción del riesgo.

2. Se observó que la mayoría de los clientes presentan comportamiento puntual, pero existe un segmento crónico con morosidad persistente. Los clientes con alta utilización de crédito y historial de rechazo en solicitudes previas tienden a mostrar mayor riesgo. La segmentación con K-Means permitió agrupar perfiles homogéneos, facilitando el análisis de patrones de incumplimiento.

3. Se realizó una limpieza exhaustiva de los datos, incluyendo conversión de tipos, tratamiento de outliers y eliminación de columnas irrelevantes. Se derivaron variables como `FRAC_LATE_INSTALLMENTS`, `MAX_DAYS_LATE` y `AVG_UTILIZATION_RATIO_TDC`, que sintetizan el comportamiento crediticio. Las visualizaciones revelaron insights clave, como la relación entre montos solicitados y tasas de aprobación.

4. Se entrenaron dos modelos Random Forest, uno para clientes registrados y otro para no registrados. Ambos modelos lograron clasificar el riesgo en tres niveles (alto, medio, bajo), con buena capacidad discriminativa. La importancia de variables permitió interpretar el modelo y entender los factores que influyen en la decisión.

8. Repositorio y Documentación Técnica

Para quienes deseen profundizar en el procedimiento técnico, el flujo de trabajo y la codificación realizada durante el desarrollo del proyecto, se ha documentado todo el proceso en una serie de notebooks disponibles en GitHub.

Estos notebooks incluyen:

- La importación y estructuración de las bases de datos en MySQL.
- El proceso de limpieza, transformación y generación de variables derivadas.
- Consultas SQL utilizadas en el análisis exploratorio.
- La implementación de modelos de aprendizaje automático, incluyendo Random Forest y K-Means.

- Evaluación de modelos y análisis de importancia de variables.

El repositorio está organizado por secciones temáticas, lo que permite seguir el paso a paso desde la adquisición de datos hasta la generación de modelos predictivos.

Enlace al repositorio:
https://github.com/JuanGonzalez47/credit-risk-analysis-project