# UNIVERSIDAD TORCUATO DI TELLA – ESCUELA DE NEGOCIOS

## MASTER EN MANAGEMENT + ANALYTICS

## Árboles de Decisión: Predicciones como alertas del sistema de prevención de lavado de activos en una entidad financiera

---


**RESUMEN**

En los últimos años, el blanqueo de capitales se ha ido sofisticando cada vez más, desafiando tanto a las naciones como a las entidades financieras a reestructurar sus sistemas de prevención y por lo tanto a aplicar nuevas metodologías para combatir las actividades ilícitas relacionadas con esta temática. El objetivo de este trabajo consistió en analizar el sistema actual de monitoreo de alertas correspondiente a un banco público de Argentina, y proponer un cambio en el mismo a través del desarrollo de distintos algoritmos de aprendizaje automático basados en técnicas de árboles de decisión (árbol simple, Gradient Boosting y Random Forest), para luego comparar los resultados de estos modelos con los costos que conlleva mantener este sistema de monitoreo en la actualidad. Si bien los resultados obtenidos por los modelos aplicados no llegan a mejorar el sistema actual de la entidad, el trabajo realizado sirve para profundizar distintas alternativas basadas en aprendizaje automático.



**Alumno:** Santiago Piñeiro

**Tutor:** Alejandro Nakab



**Fecha de entrega:** 31 de mayo del 2022

UNIVERSIDAD TORCUATO DI TELLA

# UNIVERSIDAD TORCUATO DI TELLA – ESCUELA DE NEGOCIOS

## MASTER EN MANAGEMENT + ANALYTICS

## Decision Trees: Predictions as alerts of the money laundering prevention system in a financial institution.

---


**ABSTRACT**

In recent years, there has been a constant sophistication of money laundering, which has challenged both nations and financial institutions to restructure their prevention systems and, therefore, to apply new methods to fight unlawful activities related to this issue. The aim of this piece of work has been to analyze the current alert monitoring system belonging to a public bank in Argentina and to propose changes through the development of different machine learning algorithms based on decision tree techniques (simple tree, Gradient Boosting and Random Forest) to compare then the results of these models to the costs of maintaining these monitoring systems nowadays. Even though the results obtained by the models do not improve the entity's current system, the work performed here is useful to examine different alternatives based on machine learning.



Student: Santiago Piñeiro

Tutor: Alejandro Nakab


Date: May 31st, 2022

# Tabla de contenido

# Introducción

El lavado de dinero se refiere al proceso utilizado para convertir e introducir en la economía formal los fondos generados por actividades de origen ilícito en activos de apariencia licita. El mismo es un conjunto de actos o actividades, efectuadas por personas físicas o jurídicas, cuya finalidad es ocultar o disimular el origen ilícito de bienes o recursos monetarios provenientes de actos criminales. Es decir, es un mecanismo mediante el cual una persona o una organización criminal que comete un delito (narcotráfico, corrupción, trata de personas, etc.) busca ocultar, disimular y/o encubrir el dinero conseguido de su actividad ilícita intentando en ese proceso dar, a esos fondos, apariencia de haber sido obtenidos legalmente.

Para llevar a cabo esta modalidad delictiva, los lavadores de dinero deben comenzar por la etapa de colocación, la misma consiste en la colocación del efectivo en instituciones que aceptan depósitos o en su mezcla con legítimos resultados de una empresa. Una vez colocado el dinero en las instituciones, los delincuentes realizan la estratificación de este, esto es la transferencia del dinero depositado entre cuentas de distintos bancos por medio de una serie de complejas transacciones destinadas a ocultar de donde provienen los fondos, llegando a la última etapa que consiste en la incorporación del dinero ilícito a empresas aparentando que el mismo ha sido generado por actividades corrientes de las mismas. Para combatir contra esta modalidad delictiva, se adoptó un cambio de paradigma en donde los Estados, los supervisores y los sujetos obligados, debieron modificar sus sistemas basados en reglas hacia otros cuyo enfoque se focaliza en los riesgos (Harold Koster, 2019), los cuales a través de técnicas de aprendizaje automático y de data mining, logran detectar distintas maniobras que pueden llegar a relacionarse con el lavado de dinero. En nuestro país, hace unos pocos años que distintas entidades financieras poseen sistemas de monitoreo de transacciones de manera centralizada, utilizadas por personas expertas en la materia, y en la brevedad, son menos las que utilizan técnicas de machine learning. En este trabajo, me enfocaré en analizar el sistema de prevención de un banco público del país debido a que, por su magnitud, la gran cantidad de transacciones que se realizan por el mismo y el alcance a nivel país que representa, se puede considerar el más importante de Argentina. El mismo para atendar el gran número de alertas que se generan, presenta un sistema de monitoreo descentralizado de las transacciones, haciendo que, para cumplir con las exigencias impuestas por los organismos de control, muchas de estas alertas que surgen por posibles maniobras de lavado de activos sean atendidas por personas

no profesionales en la materia. A su vez esta entidad, al momento, no utiliza modelos de aprendizaje automático, ni técnicas de data mining para facilitar la tarea de la unidad de prevención de lavado de activos (UPLA) de este banco.

El objetivo de este trabajo consiste en realizar un modelo de aprendizaje automático que sirva para reducir la gran cantidad de alertas que se generan cada año por presuntas maniobras de lavado de activos, esto sería que el modelo prediga que alerta debería ser cerrada y por lo tanto no ser analizada por los equipos de la Unidad de Prevención de Lavado de Activos de la entidad financiera, y que alertas generadas por el sistema de monitoreo deberían ser analizadas por estos. La implementación de un modelo de machine learning, sería sumamente beneficioso para la entidad, debido a que el mismo podría ser capaz de adaptarse de forma independiente cuando se exponen a nuevos datos y aprender de los cálculos anteriores para poder detectar actividades de blanqueo de capitales. Muchos de los algoritmos de aprendizaje automático están diseñados para captar patrones complejos como las relaciones no lineales entre una variable dependiente y las variables explicativas (Yan Zhang - Peter Trubey. 2018). Si la implementación de este modelo es eficiente, esto podría ayudar a la entidad bancaria a cambiar su sistema de monitoreo descentralizado a uno centralizado y analizado solamente por los recursos expertos en el tema, que es lo que se utiliza en la industria financiera para detectar las maniobras de lavado de activos.

Para realizar el objetivo de este trabajo, se utilizará como modelo en primer lugar un árbol de decisión simple. Un árbol de decisión es una estructura de árbol que intenta separar los registros dados en subgrupos mutuamente excluyentes. Para ello, partiendo del nodo raíz, cada nodo se divide en nodos hijos de forma binaria o multi división en función del método utilizado, basado en el valor del atributo (variable de entrada) que mejor separa los registros dados. Los registros de un nodo se separan recursivamente en nodos hijos hasta que no haya ninguna división que suponga una diferencia estadística en la distribución de los registros del nodo o el número de registros de un nodo sea demasiado pequeño. El mismo, ha sido elegido por su fácil implementación y comprensión en la importancia de que variables han sido utilizadas para lograr los resultados aparte de su implementación beneficiosa en otras ramas similares como la detección de fraudes en tarjetas de crédito (Y. Sahin and E. Duman, 2011). Tras analizar la performance del árbol simple de decisión realizado y ver si este cumple o no con el objetivo de eficiencia comparándolo con el sistema actual del banco. En el caso que cumpla, se realizarán distintas alternativas de modelos de machine learning junto con el agregado de nuevas variables relacionadas con la prevención de lavado de

activos, con el objetivo, segundo en este caso, de encontrar el modelo que obtenga la mejor performance en la predicción de si una alerta debe ser justificada y cerrada, o analizada por los equipos de la unidad de prevención de lavado de activos del banco, comparándolo con el árbol de decisión básico realizado. Por otro lado, si el modelo básico no logra cumplir con las medidas de performance establecidas, se llevarán a cabo distintas alternativas de modelos, sumando las nuevas variables mencionadas anteriormente y se compararán entre estos nuevos modelos para determinar cuál de los mismos es el que mejor predicción tienen de acuerdo con las métricas de evaluación utilizadas, para luego analizar si esta adopción es mejor que el método actual que posee la entidad financiera. A su vez dentro de los modelos que logren superar las métricas de evaluación utilizadas en la etapa de validación, se analizará que porcentajes de aciertos obtuvieron estos sobre las alertas ingresadas UPLA que representan un mayor riesgo en lavado de activos.

Con respecto a los resultados, los mismos han sido planteados y establecidos de acuerdo con métricas de evaluación como el recall o sensibilidad de la clase minoritaria, en primera medida y luego bajo los determinados valores de evaluación del recall general del modelo y del accuracy general de este. A su vez como métrica auxiliar para ayudar a la comprensión de los resultados obtenidos se ha utilizado la precisión obtenida en la clase minoritaria y general del modelo. Dentro de los resultados obtenidos, al validarlos, los modelos desarrollados en algunos casos han podido cumplir con estos criterios de evaluación mencionados, en especial los modelos de árboles de decisión y Random Forest. Pero al evaluar estos con los datos utilizados para el testeo, este tipo de modelos no logran mejorar el sistema actual, incluso con las variables nuevas adheridas, debido a que los mismos se decidían por cerrar muchas alertas que deberían ser analizadas por los equipos de la UPLA, y como contrapartida, que estos equipos analicen pocas alertas. Por lo tanto, se debería encontrar otro tipo de modelos o incluir más variables que puedan ser de utilidad a estos en la predicción, con el objetivo de poder reemplazar al sistema descentralizado actual.

# Marco conceptual y lavado de dinero

## Contexto mundial y nacional sobre la prevención de lavado de dinero en entidades financieras

Si bien las maniobras de lavado de activos toman conocimiento a partir de las décadas del 1920 - 1930, tal es así que la denominación de "lavado" se corresponde a cuando el mafioso Al Capone blanqueaba sus capitales provenientes de las actividades ilegales que realizaba, a través de servicios de lavanderías para insertarlo en el sistema como dinero lícito. Estas actividades ilícitas fueron en aumento con el correr de los años, y en especial la venta de drogas, teniendo como principales actores los carteles provenientes de Colombia. Los mismos durante las décadas del 1970 – 1980 realizaban depósitos en los bancos sin ningún tipo de control por parte de estos últimos sobre el origen del dinero recibido.

Es por lo mencionado en el párrafo anterior que, a partir de final de la década de 1980, la importancia en la prevención de lavado de dinero a nivel mundial ha ido en aumento, esto fue a través de la Convención contra el tráfico ilícito de estupefacientes y sustancias psicotrópicas (llamada Convención de Viena), en donde se redactó el primer documento internacional en el que los Estados se obligaron en términos jurídicamente vinculantes a aprobar una legislación interna en la que se previera la imposición de penas a quienes trataran de dar apariencia de licitud a capitales procedentes de actividades ilegales (Eduardo Fabián Caparrós, 2018).

Del resultado de esta convención, en la cumbre del G-7 celebrada en París en el año 1989, para aunar los esfuerzos por combatir contra este delito, Los participantes de esta cumbre deciden crear el GAFI (Grupo de Acción Financiera). El GAFI (integrado por 37 jurisdicciones y 2 organizaciones regionales) fue creado con el propósito de combatir a nivel global los delitos de lavado de activos y posteriormente el de financiamiento del terrorismo (FT).

Este organismo, ha publicado 40 recomendaciones, las cuales han logrado mediante una red de distintas instituciones, locales como internacionales, sentar las bases y parámetros para el control y prevención de estos delitos. Sobre estas 40 recomendaciones, las que toman una importante relevancia dentro del sistema financiero son las siguientes:

*"Los países deben identificar, evaluar y entender sus riesgos de lavado de activos/financiamiento del terrorismo, y deben tomar acción, incluyendo la designación de una autoridad o mecanismo para coordinar acciones para evaluar los riesgos, y aplicar recursos encaminados a asegurar que se mitiguen eficazmente los riesgos. Con base en esa evaluación, los países deben aplicar un enfoque basado en riesgo (EBR) a fin de asegurar que las medidas para prevenir o mitigar el lavado de activos y el financiamiento del terrorismo sean proporcionales a los riesgos identificados. Este enfoque debe constituir un fundamento esencial para la asignación eficaz de recursos en todo el régimen antilavado de activos y contra el financiamiento del terrorismo (ALA/CFT) y la implementación de medidas basadas en riesgo en todas las Recomendaciones del GAFI. Cuando los países identifiquen riesgos mayores, éstos deben asegurar que sus respectivos regímenes ALA/CFT aborden adecuadamente tales riesgos. Cuando los países identifiquen riesgos menores, éstos pueden optar por permitir medidas simplificadas para algunas Recomendaciones del GAFI bajo determinadas condiciones".* (GAFILAT)

*"Los países deben exigir a las instituciones financieras y actividades y profesiones no financieras designadas (APNFD) que identifiquen, evalúen y tomen una acción eficaz para mitigar sus riesgos de lavado de activos, financiamiento del terrorismo y financiamiento de la proliferación.*" (GAFILAT)

En estas citaciones toma la relevancia de un cambio en el sistema de prevención de los países, y por lo tanto la adecuación de las normas locales, Ley 25.246, Resolución UIF 30e/2017 principalmente. Esta recomendación, tal como dice el segundo párrafo de la citación, generó un impacto en las instituciones financieras. Las cuales han sido afectadas por medio de esta resolución, a través de la modificación de un enfoque meramente de cumplimiento, a uno nuevo, enfocado en los riesgos internos y externos de cada institución.

*"Los países deben contar con políticas ALA/CFT/CFP a escala nacional, que tomen en cuenta los riesgos identificados, las cuales deben ser sometidas a revisión periódicamente, y deben designar a una autoridad o contar con un mecanismo de coordinación o de otro tipo que sea responsable de dichas políticas."* (GAFILAT)

En el párrafo anterior, amplía lo mencionado en las citas anteriores, estableciendo que la matriz de riesgos internos y externos, realizada por cada uno de los sujetos obligados, la Unidad de Información Financiera debe ser analizada para generar una matriz de riesgo de lavado a de activos a nivel nacional que deberá trabajar mancomunadamente junto con otros sectores del Estado y los mismos sujetos obligados en la prevención de lavado de dinero.

*"Los países deben asegurar que, las autoridades que hacen las políticas, la Unidad de Inteligencia Financiera (UIF), las autoridades del orden público, los supervisores y otras autoridades competentes relevantes, tanto a nivel de formulación de políticas como operativo, cuenten con mecanismos eficaces establecidos que les permita cooperar y, cuando corresponda, entablar entre sí una coordinación e intercambio de información a nivel interno en el desarrollo e implementación de políticas y actividades para combatir el lavado de activos, el financiamiento del terrorismo y el financiamiento de la proliferación de armas de destrucción masiva. Esto debe incluir cooperación y coordinación entre las autoridades relevantes para garantizar la compatibilidad de los requisitos ALA//CFT/CFP con las normas de Protección de Datos y Privacidad y otras disposiciones similares (i.e. datos de seguridad y de localización)."* (GAFILAT)

Siguiendo con los explicado anteriormente, El Estado a través de la UIF como coordinadora, y otros entes gubernamentales como la A.F.I.P. y el B.C.R.A., realizan junto con los sujetos obligados (bancos, escribanos, contadores) políticas de intercambio de información y cooperación entre estos últimos y los entes gubernamentales con el fin de prevenir el lavado de activos en el país.

*"Los países deben establecer una Unidad de Inteligencia Financiera (UIF) que sirva como un centro nacional para la recepción y análisis de: (a) reportes de transacciones sospechosas; y (b) otra información relevante al lavado de activos, delitos determinantes asociados y el financiamiento del terrorismo, y para la comunicación de los resultados de ese análisis. La UIF debe ser capaz de obtener información adicional de los sujetos obligados, y debe tener acceso oportuno a la información financiera, administrativa y del orden público que requiera para desempeñar sus funciones apropiadamente."* (GAFILAT)

Por medio de esta recomendación, se da lugar a la creación de las Unidades de Información Financieras (UIFs) de cada país incluido en este grupo y a través de la ley nacional 25.246 en su artículo 5, se da a la creación de la Unidad de Información Financiera (UIF) en nuestro país. La misma está integrada por nueve miembros, representados por un presidente, un vicepresidente y un consejo asesor de siete vocales conformado por funcionarios y expertos en la materia de otros organismos nacionales. En nuestro país es la encargada del análisis, el tratamiento y la transmisión de información a los efectos de prevenir e impedir el delito de lavado de activos que provengan del tráfico de drogas, armas, asociaciones ilícitas, para financiar actos terroristas, etc.

A su vez, la UIF a través de resoluciones, principalmente la resolución 30e/2017, dicta reglas a cumplir para los distintos sujetos obligados, la cual involucra a la institución financiera en cuestión,

en cuanto al análisis y segmentación de los clientes; al riesgo en materia de lavado de dinero que pueden generar estos en la institución; regímenes informativos y posibles sanciones por falta de un eficaz análisis y posterior reporte a tiempo, en el caso que se requiera.

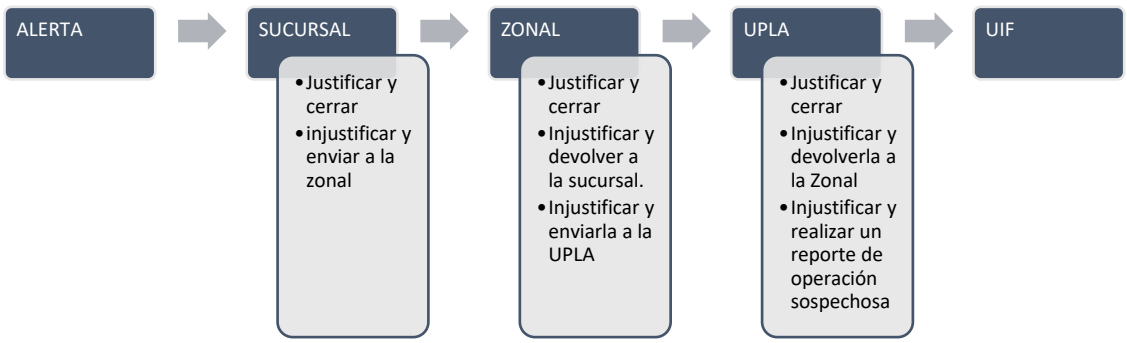
## Sistema de monitoreo de la entidad financiera

Desde el año 2015, y ante las nuevas exigencias establecidas por los organismos reguladores sumado al gran volumen de operaciones que se realizan a través de este, se toma la decisión de implementar un sistema automático de alertas para poder detectar posibles maniobras de lavado de dinero. Este sistema se comprende de un set de alertas, las cuales han sido establecidas de acuerdo con los distintos tipos de transacciones, canales y zonas en donde se realice la operación, por ejemplo, alertas por depósitos en efectivo, por transferencias recibidas, tanto nacionales como internacionales.

Las alertas son generadas, cuando los clientes realizan alguna operación dentro del banco, y la misma supera varios umbrales preestablecidos. Estos umbrales son dinámicos y dependen de varios factores como el tipo de maniobra que se realice (operación de efectivo o transferencias, por ejemplo), y de acuerdo con perfil transaccional del cliente, riesgo de lavado de activos que presente tanto este como la zona en donde se realizó la operación.

Una vez que el sistema genera una alerta, la misma se separa en tres fases para su tratamiento:

**Gráfico 1: Etapas del sistema de monitoreo de la entidad financiera**




La primera fase ocurre en la sucursal en donde el cliente o la persona que realiza la operación se encuentra radicado. Esta fase cuenta con dos instancias, la primera de estas es el análisis por una primera persona de la sucursal en donde a través del sistema de monitoreo analizan la misma y realizan una búsqueda exhaustiva de información relacionada a la operatoria alertada, solicitando en algunos casos documentación a la persona que realizó o recibió dicha operación, y verifican si el monto involucrado de la transacción se coindice con el perfil económico y transaccional del cliente.

En el caso que el monto se encuentre justificado con el perfil económico y transaccional del cliente, esta persona procederá a justificar la alerta generada por el sistema, ya que no se encontraron indicios de posibles maniobras de lavado de activos. En este caso la alerta sigue hacia una segunda persona de la misma sucursal, superior en jerarquía a la primera, en donde puede validar el análisis realizado por la primera persona y por lo tanto cerrar la alerta generada por el sistema de monitoreo. La otra opción que puede suceder es que esta segunda persona, no valide el análisis efectuado por la primera y por lo tanto injustifique la alerta y la misma vuelva a la primera persona de la sucursal para que rehaga su análisis y vuelva a determinar si la alerta se puede justificar o no.

Si, por el contrario, la primera persona de la sucursal, al analizar la operación realizada en conjunto con el perfil económico y transaccional y la documentación aportada por la persona involucrada en la operación, no puede establecer una justificación de dicha operatoria, procede a injustificar la

alerta, la cual recae en la segunda persona mencionada en el párrafo anterior. Esta segunda persona cuenta nuevamente con dos opciones, justificar o no la alerta. Si la justifica, la misma se cierra. En cambio, si no procede a justificarla, esta llega a otra instancia y por lo tanto, a la segunda fase.

La segunda fase que puede seguir una alerta recae en una persona perteneciente a una zonal de la institución financiera. Las zonales dentro de la institución financiera son los primeros medios de control que tienen las sucursales, las mismas se encuentran distribuidas geográficamente a través del territorio nacional y abarcan a un número de sucursales de acuerdo con la distancia que se encuentren estas últimas y la zonal. La persona en la zonal puede seguir con tres alternativas:

- Justificar la alerta proveniente de la sucursal y por lo tanto cerrar la misma dentro del sistema de monitoreo.
- Injustificar la alerta y devolverla a la sucursal, para mejorar el análisis realizado sobre la misma, volviendo a la fase uno explicada anteriormente.
- No justificar la alerta recibida por la sucursal y por lo tanto enviarla a la última instancia y tercera fase del tratamiento.

La tercera y última instancia, se comprende cuando la alerta llega a la Unidad de Prevención de Lavado de Dinero de esta institución financiera, la cual se llamará UPLA. Dentro de la UPLA, se encuentran personas especializadas en temas relacionados a la prevención de lavado de dinero, las cuales analizan las alertas no justificadas que provienen de la segunda fase mencionada. Estas personas al analizar dicha alerta pueden tomar las siguientes tres acciones:

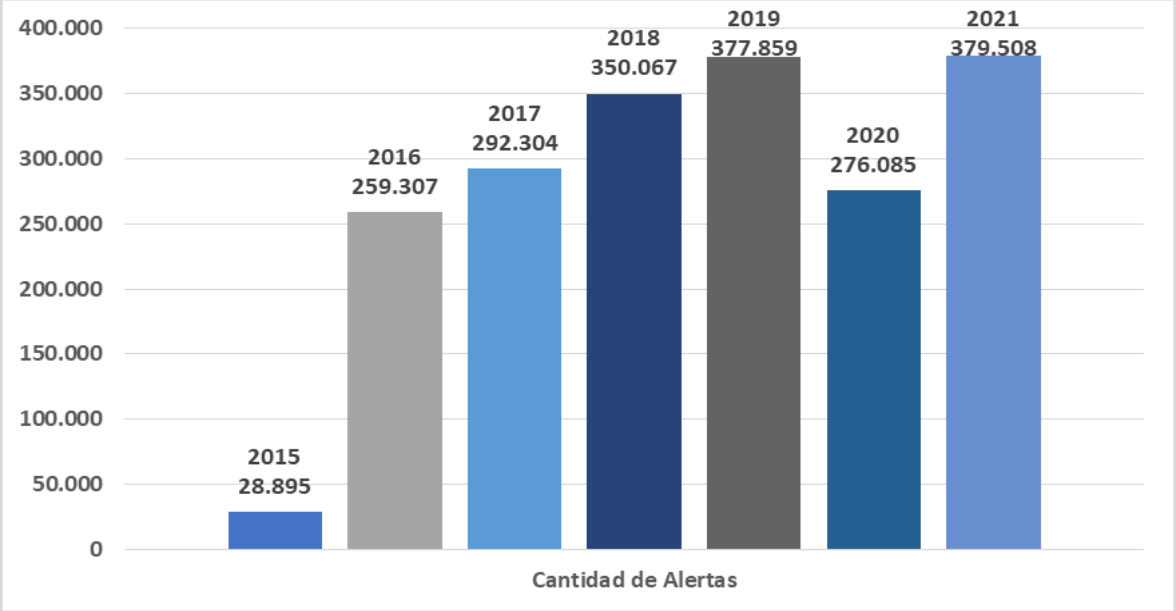
- Justificar la alerta proveniente de las zonal, y por consecuencia cerrar la misma dentro del sistema de monitoreo.
- Injustificar la alerta y devolverla nuevamente a la zonal, la cual podrá tomar las acciones descriptas en la fase dos.
- No justificar la alerta proveniente de la zonal, y por lo tanto realizar un reporte de operación sospechosa, el cual es enviado a la Unidad de Información Financiera (UIF), para su posterior análisis y determinación de si la operatoria reportada por el analista de la UPLA, en representación de la institución financiera, se corresponde con una maniobra de lavado de activos y/o financiamiento del terrorismo.

## El problema

En los últimos años, las operaciones que se realizan por medio de esta entidad financiera han ido en aumento. Si bien el sistema de monitoreo implementado en el año 2015 trajo soluciones para monitorear el gran volumen de operaciones que se cursan a través de esta entidad, el mismo presenta algunas limitaciones.

La primera resulta en la gran cantidad de alertas que se generan año tras año. En el siguiente gráfico se desagrega esta información.

**Gráfico 2: Cantidad de alertas generadas por el sistema de monitoreo. Periodo: 2015 a 2021**




Esto significa que para atender la gran cantidad de alertas que se generan en cada año y sumado a que la institución financiera tiene un plazo de 150 días desde que se realizó la operación que generó una alerta para realizar el proceso mencionado en el apartado 2.2., la misma para poder cumplir con la normativa local y no caer en posibles sanciones, dispuso tener un sistema descentralizado para atender estas alertas en plazo y proteger al banco de posibles maniobras de lavado de dinero que puedan cursarse a través del mismo.

Para mantener este sistema se necesita tener 2 personas en cada sucursal y una en cada zonal de la entidad. Actualmente esta entidad cuenta con 640 sucursales y 40 zonales repartidas por todo el

país, involucrando a 1.320 recursos humanos para cumplir con dichas tareas. Estos recursos tienen un rol indirecto o secundario en la prevención de lavado de activos, debido a que se encargan de otras tareas operativas no relacionadas con la prevención de lavado de dinero. Al número de personas involucradas que se mencionaron anteriormente, se le suman los recursos cuyas tareas, si se encuentran avocadas principalmente al análisis de posibles operaciones sospechosas de lavado de activos. Estos recursos en la actualidad son un total de 20 personas, los cuales se encuentran en la UPLA de esta entidad.

De acuerdo con lo mencionado en el apartado anterior, este sistema descentralizado genera en el banco dos limitaciones:

- La primera limitación surge de la gran cantidad de recursos que el banco debe disponer, aunque sea de manera indirecta, para poder cumplir con las tareas de prevención de lavado de dinero. Esta limitación ocasiona que estos recursos, no desempeñen otras tareas operativas o comerciales para la entidad, generando una pérdida de posibles futuros ingresos económicos para la misma.

Esta primera limitación se podría solucionar con un cambio en el método del sistema de monitoreo actual que tiene esta entidad. Es decir, reemplazar el sistema de monitoreo descentralizado que hoy en día tiene la entidad, por un sistema de monitoreo centralizado. Esto sería las alertas que se generen en la entidad sean solamente atendidas por los recursos que se encuentran hoy en la Unidad de Prevención de Lavado de Activos. Por lo tanto, se pasaría de 1.340 recursos que utiliza el banco para cumplir con la prevención de lavado de activos a solamente 20 recursos.

Realizar esta modificación, a priori, resultaría beneficioso para la entidad no solo por la cantidad de personas avocadas a la tarea de atender las alertas que el sistema genera, sino también a la gran experiencia que poseen las personas que trabajan de manera directa en la prevención de lavado de activos por sobre los recursos involucrados de manera indirecta, y así obtener mejores decisiones sobre las alertas, mejorando el sistema de prevención de la entidad.

Si bien, con la modificación del sistema actual a uno centralizado, el mismo podría eliminar a la brevedad las instancias de sucursales y de zonales. Por la gran cantidad de alertas que el sistema genera, sería difícil para el personal en la UPLA poder analizarlas, por lo que el sistema centralizado de alertas debería ser más eficiente para que el número de alertas sea menor, sin perder aquellas alertas que son propicias de generar un reporte de operación sospechosa a la UIF.

De esta manera, en este trabajo se propone un nuevo método, el cual se explicará en la propuesta de trabajo, con el fin de generar un sistema de monitoreo más eficiente para reducir la gran cantidad de alertas justificadas sin perder aquellas alertas que deban ser injustificadas.

## Propuesta de Trabajo

En este trabajo se propone un nuevo método para generar un sistema de alertas más eficiente para reducir la gran cantidad de alertas justificadas. Para ello lo que se realizará es la implementación de un método basado en análisis de datos que pueda predecir de la mejor manera posible que alertas generadas por el sistema de monitoreo actual deberían ser justificadas y por lo tanto cerradas, y por el otro lado, cuáles deberían ser injustificadas y por esta razón ser analizadas por los recursos de la UPLA, siguiendo lo descripto en la tercera fase del sistema de monitoreo.

El método por utilizar serán modelos de datos supervisados de clasificación por medio de aprendizaje automático. Los mismos son de la familia de los árboles de decisión. La elección de estos modelos de clasificación se debe a las ventajas que presentan sobre otros modelos supervisados, entre las mas importantes se encuentran la facilidad de comprensión e interpretabilidad de estos modelos; pueden manejar atributos de entradas tanto nominales como numéricos; pueden trabajar con conjunto de datos que pueden tener valores nulos o erróneos y, por lo tanto, no requieren un preprocesamiento de los datos demasiado exigente. Dentro de las desventajas que presentan estos modelos se encuentran el overfitting si el número de variables predictivas es alto; la creación de un árbol completamente diferente ante un pequeño cambio en los datos de entrada.

Para poder realizar esta propuesta de solución, se llevarán a cabo dos objetivos. El primero será el de crear un modelo de datos supervisado de clasificación por medio de aprendizaje automático, el cual será un árbol de decisión simple y comparar si el mismo de acuerdo con determinadas métricas de evaluación cumple con el objetivo de ser mejor que el método descentralizado que se lleva hoy en día en la entidad financiera.

En cuanto al segundo objetivo para esta propuesta de trabajo, el mismo dependerá de la resolución favorable o desfavorable del primer objetivo planteado. Si se cumple el primer objetivo, el motivo del segundo objetivo será el de realizar nuevos modelos de machine learning y compararlos con el modelo de árbol simple que se realizó en el primer objetivo. A sí mismo, en los nuevos modelos, se

realizarán una optimización de parámetros, junto con el agregado de nuevas variables correspondientes al riesgo de lavado de activos y movimientos transaccionales de los clientes que hayan generado alertas, para analizar si estas nuevas variables mejoran la predicción de los modelos. En el caso que el primer objetivo no se haya cumplido, este segundo constará en encontrar el modelo, con la optimización de parámetros y agregado de nuevas variables mencionada anteriormente, que contenga la mejor performance de acuerdo con las métricas de evaluación utilizadas.

Del modelo de machine learning resultante con mejor performance, se analizará el mismo contra el método actual de la institución para resolver si resulta conveniente la sustitución del modelo actual y por lo tanto la implementación del nuevo modelo, o si por el momento, de acuerdo con los resultados predictivos, no es conveniente realizar dicha modificación en el sistema de prevención del banco.

# Materiales y Métodos

## Datos

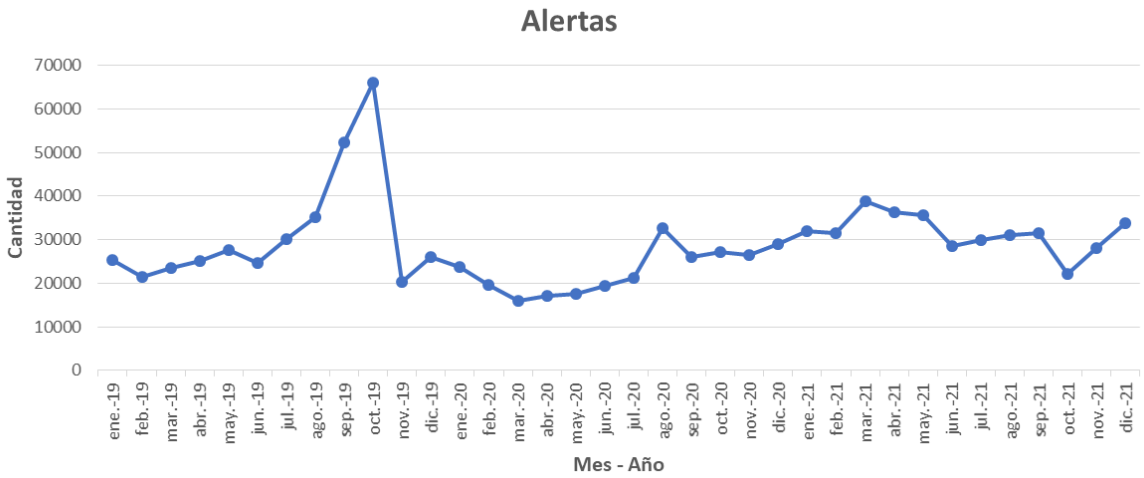
Los datos utilizados para realizar la propuesta de trabajo provienen de distintas fuentes de la entidad financiera. Las mismas son generadas por sistemas de variada índole que tiene la misma, entre ellos se encuentran el sistema de monitoreo de alertas que tiene el banco; base de clientes de la entidad; y las generadas propiamente por la Unidad de Prevención de Lavado de Dinero.

Para este trabajo, debido a que algunas de estas fuentes se han creado a principios del año 2019, se determinó utilizar los datos desde enero 2019 hasta diciembre 2021. En el anexo 1 de este documento, se explican y describen las fuentes utilizadas.

## Análisis descriptivo de los datos

Tal como se explicó en la sección Datos, se realizará una descripción de los datos utilizados entre los periodos de enero 2019 a diciembre 2021 para una mayor comprensión de estos.
Como un primer análisis de estos datos, en el gráfico 2 se visualizará la cantidad de alertas generadas mes a mes.

**Gráfico 3: Cantidad de alertas generadas mensualmente. Periodo enero 2019 a diciembre 2021**



Lo mismo realizaremos con la cantidad de clientes, los cuales a través del gráfico 3 se visualizará la cantidad de estos involucrados en las alertas.

**Gráfico 4: Cantidad de clientes vinculados sobre las alertas generadas mensualmente. Periodo enero 2019 a diciembre 2021**


De estos dos gráficos, podemos analizar que no se encuentra una estacionalidad marcada en este periodo. De los mismos, podemos hacer una clara interpretación de dos tramos.

El primer tramo abarca los meses de agosto 2019 a noviembre 2019, y el mismo se explica a través de dos momentos:

El primer momento resulta luego de las elecciones primarias o PASO de agosto del 2019. elecciones en donde los resultados obtenidos por el partido político gobernante no fueron auspiciosos, por lo cual se genera una fuerte devaluación del peso argentino frente al dólar. Como medida para frenar la devaluación de la moneda argentina, el gobierno realiza un “cepo” o restricción a la compra de dólares estadounidenses de U$S 10.000 por mes.

Esta medida restrictiva a la compra de dólares por parte del poder gobernante ocasiona que las personas que quieran adquirir los mismos, realicen una operación de compraventa de bonos a través del sistema financiero, denominado “Rulo Financiero”. Esta acción, en el sistema de monitoreo del Banco, cuando los clientes vendían los bonos y compraban dólares, generó un aumento considerable en la cantidad de alertas, principalmente de las alertas de cambio de moneda extranjera.

**Cuadro 1: Incidencia alerta de Cambio de Moneda Extranjera. Periodo junio 2019 a noviembre 2019**

| Mes | Cantidad de alertas | Alertas Cambio de Moneda Extranjera | Incidencia |
|---|---|---|---|
| Junio del 2019 | 24.723 | 6.804 | 27,52% |
| Julio del 2019 | 30.249 | 8.671 | 28,67% |
| Agosto del 2019 | 35.126 | 12.455 | 35,46% |
| Septiembre del 2019 | 52.234 | 26.633 | 50,99% |
| Octubre del 2019 | 65.960 | 35.369 | 53,62% |
| Noviembre del 2019 | 20.374 | 652 | 3,20% |

De acuerdo con lo que se observa del cuadro 1, existe una fuerte correlación entre lo explicado en el párrafo anterior y el aumento de las alertas de Cambio de Moneda Extranjera en el sistema de monitoreo.

Esta correlación continua como se puede ver con la cantidad de este tipo de alertas para el mes de noviembre 2019, y en donde se explica el segundo momento de este primer tramo analizado, la cual se reduce de manera drástica por los motivos de la derrota de final de octubre del mismo año, del partido gobernante y la modificación más agresiva de la medida restrictiva que se encontraba funcionando, la cual llevo a pasar de comprar 10.000 dólares mensuales a solamente 200 dólares.

El segundo tramo que se puede analizar es en el periodo marzo 2020 a julio del mismo año. En este tramo, ocurre la pandemia por el virus SARS-C0V-2. En el inicio de la pandemia, el gobierno argentino dispuso medidas de aislamiento social, preventivo y obligatorio (ASPO), que coinciden con este periodo, las cuales generan una caída en la actividad económica del país y por lo tanto en una disminución de las operaciones en efectivo que se cursan en la entidad bancaria. Esto ocasiona a su vez una disminución en la cantidad de alertas de este tipo de operaciones que se generan en el sistema de monitoreo de la entidad. Si bien durante este periodo señalado las operaciones relacionadas a través de manera electrónica (transferencias) han aumentado de manera considerable, el riesgo de estas operaciones hacia la entidad resulta menor que el ocasionado por operaciones en donde se utiliza el efectivo, teniendo como consecuencia esta disminución de las alertas.

Este segundo tramo finaliza con las flexibilizaciones realizadas por el gobierno sobre las medidas tomadas para mantener el ASPO durante los meses de marzo 2020 a junio 2020. Estas flexibilizaciones, representaron un aumento de actividad económica y por lo tanto la intensificación de las operaciones en la entidad y de las alertas de monitoreo de esta, tal como se puede observar en el cuadro 3, en donde el aumento considerable por parte de operaciones originadas de manera electrónica ocasiona una mayor creación de alertas de este tipo.

**Cuadro 2: Cantidad de alertas. Periodo enero 2020 a septiembre 2020**

| Mes | Cantidad de alertas |
|---|---|
| Enero del 2020 | 23.650 |
| Febrero del 2020 | 19.593 |
| Marzo del 2020 | 15.958 |
| Abril del 2020 | 17.017 |
| Mayo del 2020 | 17.667 |
| Junio del 2020 | 19.353 |
| Julio del 2020 | 21.226 |
| Agosto del 2020 | 32.625 |
| Septiembre del 2020 | 26.119 |

**Cuadro 3: Cantidad de alertas generadas sobre transacciones de efectivo y transferencias. Periodo enero 2020 a septiembre 2020**

| Mes | Operaciones al crédito realizadas | Operaciones al crédito en efectivo | % de incidencia | Cantidad de alertas relacionadas con el uso de efectivo | Operaciones al crédito en transferencias | % de incidencia | Cantidad de alertas relacionadas con el uso de transferencias |
|---|---|---|---|---|---|---|---|
| Enero del 2020 | 1.024.137 | 704.215 | 68,76% | 11.407 | 319.922 | 31,24% | 2.409 |
| Febrero del 2020 | 816.115 | 550.660 | 67,47% | 8.661 | 265.455 | 32,53% | 2.163 |
| Marzo del 2020 | 717.516 | 483.032 | 67,32% | 6.489 | 234.484 | 32,68% | 2.117 |
| Abril del 2020 | 751.814 | 277.075 | 36,85% | 5.173 | 474.739 | 63,15% | 3.485 |
| Mayo del 2020 | 929.372 | 318.491 | 34,27% | 5.368 | 610.881 | 65,73% | 3.541 |
| Junio del 2020 | 1.162.022 | 372.283 | 32,04% | 6.138 | 789.739 | 67,96% | 3.762 |
| Julio del 2020 | 1.303.289 | 362.372 | 27,80% | 6.019 | 940.917 | 72,20% | 4.637 |
| Agosto del 2020 | 1.355.180 | 363.742 | 26,84% | 6.081 | 991.438 | 73,16% | 5.239 |
| Septiembre del 2020 | 1.233.375 | 361.729 | 29,33% | 6.007 | 871.646 | 70,67% | 4.428 |

## Clientes

Con respecto a los clientes de la entidad que se encuentran en la base de datos utilizada durante enero 2019 a diciembre 2021, se puede realizar los siguientes comentarios al respecto.

A nivel general, la proporción de los clientes que se encuentran en la base de datos es de 72,30% para las personas humanas y de un 27,70% para personas jurídicas. Esta proporción se revierte cuando se analiza la proporción del monto alertado por tipo de persona, dando como resultado un 87.69% de este monto correspondiendo a la persona jurídica y un 12.31% para personas humanas.

**Gráfico 5: Proporción de clientes. Periodo enero 2019 a diciembre 2021**

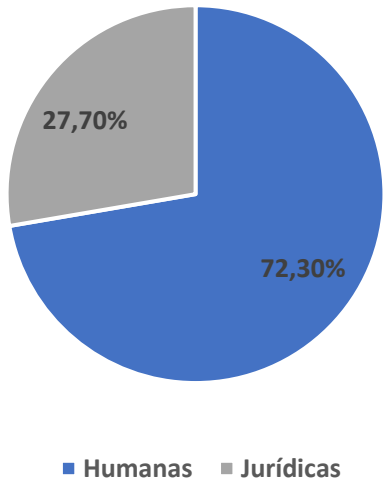


**Gráfico 6: Proporción del monto alertado por tipo de clientes. Periodo enero 2019 a diciembre 2021**

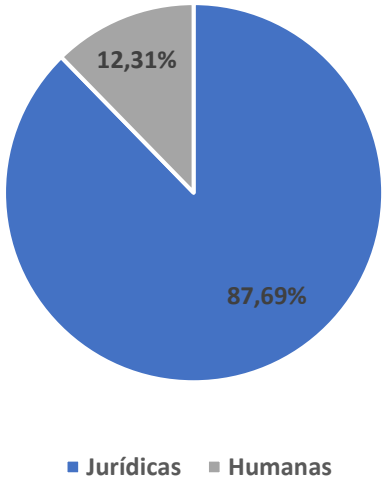
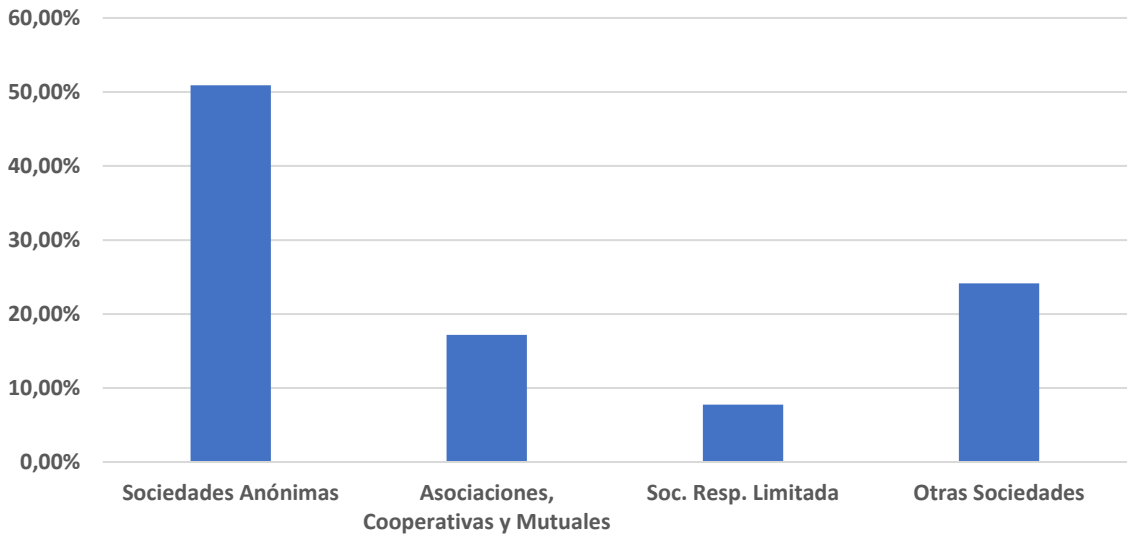
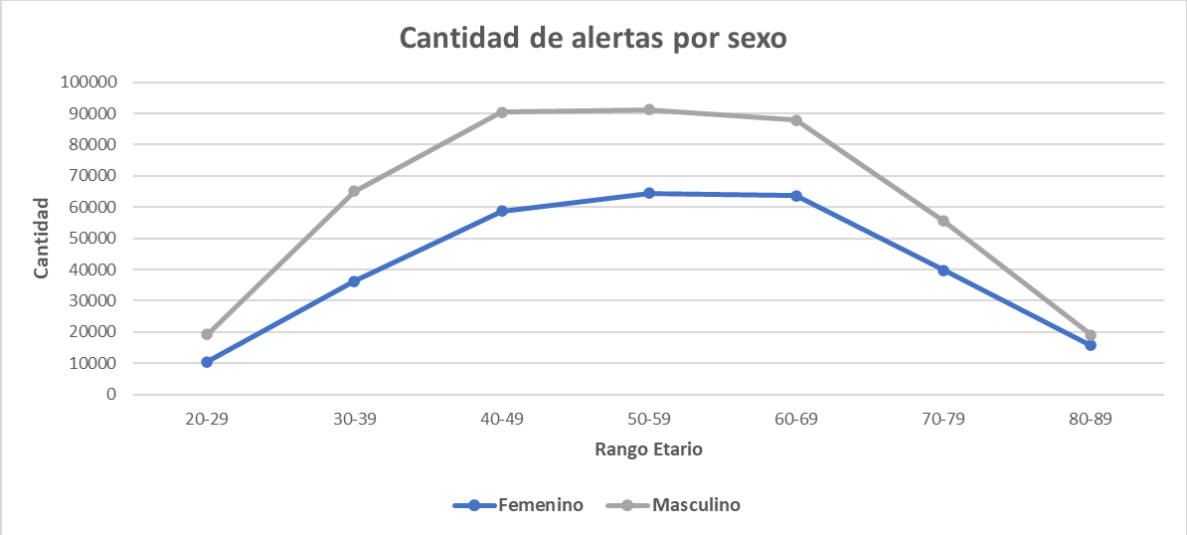



**Gráfico 7: Proporción del monto alertado por tipo de societario clientes jurídicos. Periodo enero 2019 a diciembre 2021**




A través del gráfico 7, podemos observar que las tres primeras formas jurídicas, representan más del 75% del monto alertado dentro de los más de 25 tipos societarios que posee la entidad para identificar a los clientes jurídicos. Esto nos lleva a poner el foco de atención en estas tres tipo de sociedades, pero en especial en las asociaciones, cooperativas y mutuales, debido a que las mismas por sus actividades económicas y tratamiento impositivos, resultan de posibles indicios de operaciones sospechosas.

**Gráfico 8: Cantidad de alertas por rango etario sexo masculino y femenino. Periodo enero 2019 a diciembre 2021**




En el gráfico 8 podemos notar que no existen diferencias significativas entre los rangos etarios de ambos sexos. De ambos gráficos podemos ver que los rangos etarios entre 40 y 79 años son los que más alertas generan dentro del sistema de monitoreo de la entidad bancaria. A su vez, de este análisis se puede interpretar que es coherente que estos rangos etarios sean los que más alertas generan, en especial de 40 a 65 años, debido a que su nivel socioeconómico y a las actividades que realizan estos sujetos, generan por lo general, mayores recursos económicos que los rangos etarios más jóvenes (de 20 años a 39 años) y de los más longevos (mayores a 79 años).

# Metodología

## Modelos

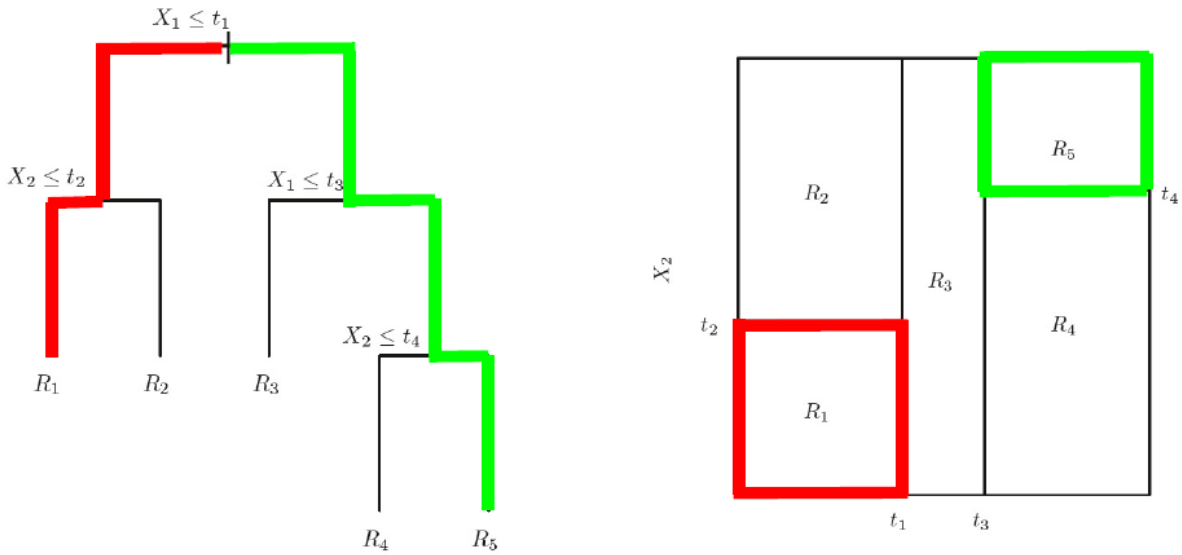
Para este trabajo se han utilizado distintas técnicas de aprendizaje supervisado basadas en árboles de decisión.

Entre las distintas técnicas sobre árboles de decisión utilizadas, se han optado por los siguientes tres modelos:

### Árbol de decisión simple

Esta técnica es la más básica dentro de la familia de los árboles, el mismo se basa en crear un árbol invertido con un nodo raíz, nodos internos y nodos hoja. El algoritmo no es paramétrico y puede manejar de manera eficiente conjuntos de datos grandes y complicados sin imponer una estructura paramétrica complicada. El mismo dentro del trabajo realizado se ha implementado como modelo base o benchmark, para luego compararlo y analizar los resultados que arroje las otras técnicas de árboles de decisión utilizadas contra este. Para facilitar esta explicación se agrega un gráfico con la actividad que realiza el árbol de decisión simple.

**Gráfico 9: Funcionamiento del árbol de decisión simple**



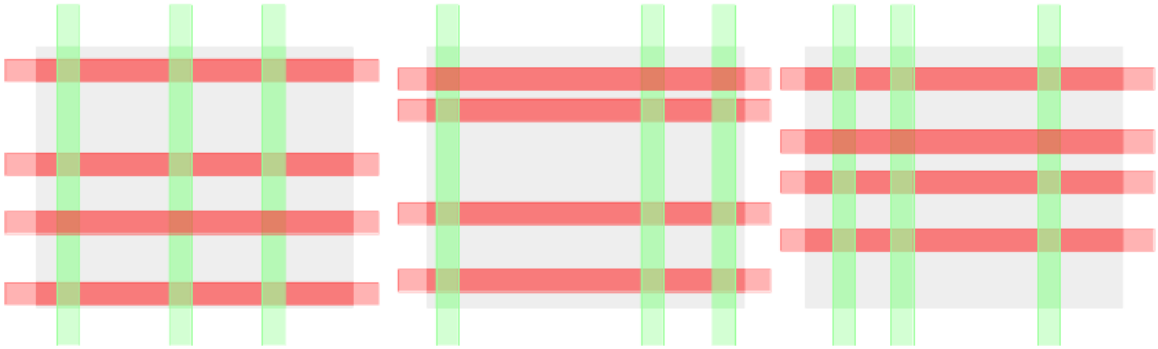
## Gradient Boosting

Esta técnica implica crear muchos árboles de decisión simples a partir de los datos de entrenamiento. Entre las características que presentan estos árboles es que son modelos predictivos débiles, en donde cada árbol se crea de manera secuencial y utiliza los errores de su predecesor para mejorar el aprendizaje.  es por ello por lo que se dice que estos tipos de algoritmos "aprenden despacio" (James, Witten, Hastie, Tibshirani, 2017), obteniendo como resultado un gran poder de predicción en los mismos.

## Random Forest

Los bosques aleatorios o Random Forest (Breiman, 2001) son una modificación sustancial del bagging que construye una gran colección de árboles descorrelacionados y luego los promedia. En muchos problemas, el rendimiento de los bosques aleatorios es muy similar al de los árboles que utilizas boosting pero con la ventaja que son más sencillos de entrenar y ajustar reduciendo la varianza del modelo entrenado. (Breiman, 2001). Para una mejor explicación se presenta un gráfico de como el algoritmo de Random Forest decorrelaciona los árboles realizados.

**Gráfico 10: decorrelación de los árboles en Random Forest**



Este gráfico presenta el trabajo que realiza Random Forest dentro del data set, en donde se presentan tres cuadrantes, representando a tres árboles simples, en los cuales cada uno elige distintas observaciones (líneas rojas) y distintas variables (líneas verdes) generando una decorrelación entre estos.

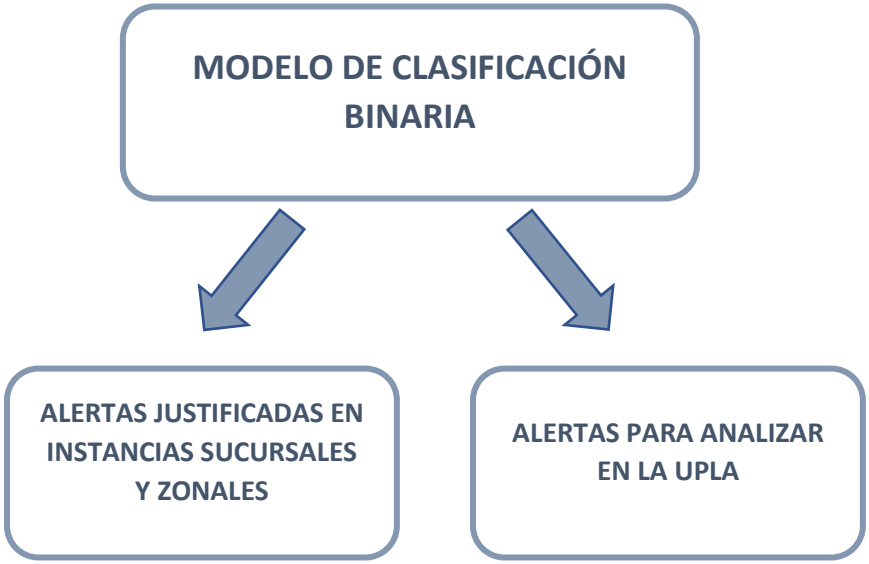
Estos tres modelos de árboles se encuentran en la librería scikit-learn. A su vez, se aclara que se ha querido implementar otros dos tipos de técnicas de machine learning. Las mismas han sido Linear Support Vector Classification (Linear SVC) de la misma librería y XGBoost pero por inconvenientes en el poder de cómputo no ha sido posible realizar estos modelos.

## Explicación del objetivo y tratamiento a realizar con los modelos descriptos

De acuerdo con lo explicado en la propuesta de trabajo, los objetivos que van a tener que resolver estos modelos es la predicción de que alertas generadas por el sistema de monitoreo de la entidad bancaria deben se justificadas y por lo tanto cerradas, y cuales ser resueltas por los analistas de la unidad de prevención de lavado de activos. Está última predicción que deberán realizar los distintos modelos posee una mayor importancia que justificar las alertas, debido a los riesgos que pueden generar una alerta mal justificada a la entidad bancaria. Los mismos pueden llegar a ser posibles multas al banco por parte de los organismos de control, pero como un mayor daño pasible al mismo sería un daño a la imagen de la entidad, teniendo como consecuencia menores inversiones y beneficios comerciales y económicos. Como casos sobre multas aplicadas a entidades financieras, se encuentran en el marco internacional el banco británico HSBC y el banco Standar Bank, y por su parte de manera nacional los bancos BBVA Frances y el banco Macro.

Sobre este aspecto, para analizar la capacidad de los modelos de detectar las alertas que presentan un mayor riesgo en lavado de activos y evitar lo mencionado en el párrafo anterior, se evaluará en la etapa de test cuantas alertas de este riesgo el sistema predijo que deben ser tratadas por los equipos de la UPLA sobre las alertas de mayor riesgo que han sido analizadas por estos.

**Gráfico 11: Explicación gráfica del modelo a realizar**




Es por ello que para entrenar los modelos propuestos y analizar los resultados de los mismos, se realizarán 4 subetapas de las tres técnicas de árboles descriptas, variando la comparación de cada modelo realizado con el benchmark o modelo de referencia. Esto sería utilizando el árbol de decisión simple en el caso que el mismo cumpla con el primer objetivo propuesto o el modelo actual de la entidad financiera si no se supera con el primer objetivo.

La primera subetapa se comprende en realizar dos modelos de árboles descriptos con las técnicas no utilizadas al momento y con los hiper parámetros sin optimizar o por default. Con los resultados que arrojen las predicciones de estos, se hará una comparación entre estas y el modelo de referencia como se mencionó anteriormente. Y en el caso que al ser evaluados obtengan una mejor performance que el modelo de referencia, se los utilizará como punto de partida para compararlos con los modelos de las siguientes subetapas.

La segunda subetapa se comprende en la optimización de determinados hiper parámetros para cada uno de los árboles de decisión de la primera subetapa y su comparación con el modelo benchmark.

Con respecto a la tercera subetapa, lo que se añadirá a los modelos de la primera subetapa son nuevas variables para analizar si las mismas ayudan a estos a mejorar las predicciones de los modelos de las subetapas anteriores, y tal sea el caso, se los confrontará con el modelo de referencia.

Por último, la cuarta subetapa comprende la optimización de los hiper parámetros de cada modelo con las nuevas variables agregadas en la subetapa tres y su comparación con el modelo de referencia que corresponda.

Terminada esta etapa, se analizará la eficacia de los modelos que hayan superado la etapa de validación en predecir de acuerdo con las alertas que hayan llegado a la UPLA, cuales de estás presentan un mayor riesgo de lavado de activos.

## Optimización de hiper parámetros

Para encontrar los mejores hiper parámetros, la librería scikit-learn permite utilizar las estrategias grid search y random search. La estrategia grid search consiste en realizar cada combinación posible con cada valor fijado en los hiper parámetros. La gran desventaja de utilizar esta estrategia es el costo computacional que requiere realizarla. Por su parte, la estrategia random search consiste en utilizar aleatoriamente los valores de cada hiper parámetro. De esta forma quedan seleccionados los hiper parámetros correspondientes para un posible modelo. Repitiendo esto varias veces, quedan armados distintos modelos a evaluar. La ventaja de utilizar esta estrategia es la de mejorar el costo computacional y, además, se ha demostrado que random search es una estrategia de igual o incluso de mayor eficacia para encontrar la optimización de estos valores (Bengio, Bergstra, 2012). Por lo consiguiente se decidió por utilizar esta última estrategia explicada optimizando los siguientes parámetros de los modelos utilizados:

**Cuadro 4: Árbol de decisión simple. Rango de hiper parámetros**

| Hiper parámetro | Rango |
|---|---|
| Criterion | [gini, entropy] |
| Splitter | [best, random] |
| Max_depth | [2 : 15] |
| Class_weight | [balanced] |
| Min_samples_Split | [2 : 15] |
| Min_samples_leaf | [1 : 5] |
| Max_features | [auto, sqrt, log2] |

**Cuadro 5: Gradient Boosting. Rango de hiper parámetros**

| Hiper parámetro | Rango |
|---|---|
| N_estimators | [100 : 500] |
| Learning_rate | [0.1 : 0.3] |
| Max_depth | [3 : 8] |
| Min_samples_Split | [2 : 8] |
| Min_samples_leaf | [1:5] |
| Loss | [deviance, exponential] |

**Cuadro 6: Random Forest. Rango de hiper parámetros**

| Hiper parámetro | Rango |
|---|---|
| Criterion | [gini, entropy] |
| N_estimators | [100 : 1000] |
| Max_depth | [4 : 20] |
| Class_weight | [balanced] |
| Min_samples_Split | [2 : 10] |
| Min_samples_leaf | [1 : 10] |
| Max_features | [auto, sqrt] |

En la subsección dos y cuatro, tal como se comentó antes, se utilizó la estrategia random search realizando 15 configuraciones distintas para cada técnica de árbol utilizada. Esta estrategia representa realizar 45 modelos para cada una de estas subsecciones, sumando a los 6 modelos de las subsecciones uno y tres llevados a cabo, da como resultado 96 modelos distintos para comparar y analizar cuál de estos es el que representa la mejor predicción sobre los datos utilizados.

## Métrica utilizada para la evaluación de los modelos

Para la selección de la métrica de evaluación a utilizar, se analizaron la proporción de las clases a predecir para cada año de las alertas generadas por el sistema.

**Cuadro 7: Proporción de clases por año. Periodo enero 2019 a diciembre 2021**

| Estado | Porcentaje |
|---|---|
| **Año 2019** | |
| Justificada Sucursal/Zonal | 96,74% |
| Analizar en la UPLA | 3,26% |
| **Año 2020** | |
| Justificada Sucursal/Zonal | 88,07% |
| Analizar en la UPLA | 11,93% |
| **Año 2021** | |
| Justificada Sucursal/Zonal | 92,88% |
| Analizar en la UPLA | 7,12% |

Al tratarse de datos muy desbalanceados, si como métrica de evaluación para analizar y comparar los distintos modelos de clasificación realizados utilizaríamos le medida de Accurracy, la misma puede resultar un impacto negativo en el análisis debido a que si elegimos aquellos modelos con el mejor Acurracy, lo más probable es que estos sean los que por minimizar el error, predicen con mucha más frecuencia la clase mayoritaria. Esto en este trabajo, generaría que el modelo prediga en la mayoría de las veces que la alerta tiene que ser justificada en las instancias sucursales y zonales y por lo tanto no debería ser analizada por los equipos de la UPLA. Lo cual podría generar daños en la imagen y perjuicios económicos y comerciales a la entidad.

Por lo tanto, las métrica que se utilizará para medir y analizar los modelos será el recall o sensibilidad. Esta métrica se encuentra cuando se realiza la matriz de confusión y la misma informa la cantidad de aciertos que el modelo logra identificar.

**Cuadro 8: Cálculo métrica de Recall**

$$\text{RECALL:} \quad \frac{TP}{TP + FN}$$

En donde “TP” se refiere a los valores denominados verdaderos positivos, los cuales son aquellos que el modelo predijo que son verdaderos y su predicción fue correcta. Mientras que “FN”, se refiere a los valores en donde el modelo predijo que eran falsos y su predicción fue incorrecta.

En resumen, a un mayor recall, y en especial en la clase minoritaria, mejor será nuestro modelo prediciendo que alertas deberían llegar a los equipos de análisis de la UPLA.

Como métrica de evaluación auxiliar, que ayude a una mejor comprensión de nuestro modelo, se utilizará la métrica de precisión de la clase minoritaria y del modelo en general. La misma informa sobre la efectividad del modelo en identificar que alertas deben ser justificadas y cuales enviadas a la UPLA.

**Cuadro 9: Cálculo métrica de precisión**

$$\text{PRECISIÓN:} \quad \frac{TP}{TP + FP}$$

En donde “TP” se refiere a los valores denominados verdaderos positivos, los cuales son aquellos que el modelo predijo que son verdaderos y su predicción fue acertada. Mientras que “FP”, se refiere a los valores en donde el modelo predijo que eran positivos y su predicción no fue correcta.

## Ventana de tiempo y conjuntos de entrenamiento, validación y testeo

Como datos de entrenamiento se utilizarán los datos disponibles sobre las alertas de los años 2019 y 2020. A su vez se separarán de manera aleatoria una porción de estos datos correspondientes a estos años para utilizarlos en la validación de cada modelo. También se aclara que en la subsección dos y cuatro, al optimizar los hiper parámetros los mismos fueron validados y para que no ocurra el error en caer en un data leakage, se ha utilizado dentro de la estrategia de random search, el método de cross validation. En las subsecciones dos y cuatro (en donde se realiza la estrategia mencionada), se elegirán los 7 mejores modelos que obtengan el mejor recall en la fase de validación.

Los modelos en donde no se ha realizado el random search para optimizar los hiper parámetros de estos, junto con los 7 mejores modelos de cada técnica en donde sí se utilizó la optimización de estos, serán testeados con los datos correspondientes a las alertas del año 2021 para predecir que alertas tienen que ser justificadas y por lo tanto cerradas, en instancias anteriores a la UPLA y cuales deben llegar a esta unidad y ser analizadas por los recursos de esta.

Por último, se evaluarán en otra muestra de test aquellos modelos que hayan superado los criterios de evaluación propuestos en la etapa de validación, para analizar el rendimiento de estos. Estos datos se corresponderán con las alertas generadas por el sistema de monitoreo en el periodo enero – mayo 2022.

## Ingeniería de variables

A nivel general, debido a que los modelos de árboles de clasificación y al encontrarse con un data set con muchas variables categóricas, para no eliminar esta información, se realizó la técnica de one hot encoding.

Para las subetapas tres y cuatro, al data set original se le han agregado variables pertenecientes a la Unidad de prevención de lavado de dinero de la entidad. Las mismas contienen información referida al riesgo y demás factores que presenta el cliente en materia de lavado de dinero hacia la entidad. El objetivo de sumar estas nuevas variables es analizar si las mismas ayudan a mejorar la predicción, y en particular el recall, de los distintos árboles utilizados.

Estas variables agregadas las podemos dividir en dos grupos, referidas al riesgo y referidas a la operatoria transaccional. Las variables que representan el riesgo de lavado de activos son cinco, y las mismas hacen referencia a la probabilidad e impacto que puede ocasionar las operaciones que el cliente efectué en la entidad. Estas dos variables al sumarse determinar el riesgo de lavado de dinero que puede realizar el cliente en contra del banco.

**Cuadro 10: Variables relacionadas al riesgo de lavado de activos**

| Variable | Detalle |
|---|---|
| Probabilidad | Es la posibilidad de que ocurra un evento de lavado de activos. |
| Impacto | Es el resultado de un suceso de lavado de activos. |
| Atributo sensible | Es el motivo principal que ocasiona un evento de lavado de activos. |
| Riesgo en lavado de dinero anterior | Comprende el riesgo anterior a riesgo actual del cliente al momento de haberse generado la alerta. |
| Riesgo en lavado de dinero | Comprende el riesgo actual del cliente al momento de haberse generado la alerta. |

En cuanto las variables agregadas referidas a la operatoria transaccional, las mismas en principio fueron cuatro, las cuales contenían:

- La cantidad de operaciones en efectivo que realizó el cliente en el mes anterior de haberse generado la alerta en el sistema de monitoreo.

- La suma total de las operaciones en efectivo que realizó el cliente en el mes anterior de haberse generado la alerta.
- La cantidad de operaciones de crédito que realizó el cliente en el mes anterior de haberse generado la alerta en el sistema de monitoreo.
- La suma total de las operaciones de crédito que realizó el cliente en el mes anterior de haberse generado la alerta.

A la vez, una estrategia adoptada para ayudar a los distintos modelos realizados a mejorar su poder de predicción fue la de realizar la media de los últimos 3, 6, 9 y 12 meses de estas cuatro variables. Esta estrategia es importante debido a que ayuda a las limitaciones que presentan estos modelos basados en árboles para realizar agregaciones por sí mismos (Chen, T., & He, T., 2015). En otras palabras, al agregar en el data set estas nuevas variables, el árbol ya cuenta con ellas desde el principio y no debe realizar una cierta cantidad de cortes para llegar a ese mismo calculo.

Por lo tanto, al data set original, utilizado en las subetapas uno y dos, se le han agregado un total de cinco variables categóricas relacionadas al riesgo y 20 variables numéricas relacionadas a la operatoria transaccional del cliente.

# Resultados

Como primera medida y para comparar de manera económica si le es conveniente o no a la institución financiera modificar su modelo actual de monitoreo de alertas, se adjuntan dos cuadros estimativos con el costo económico que le provocaría al banco ser multado por la Unidad de Información Financiera, tomando como referencia multas aplicadas por este ente y en materia internacional por otros organismos de control de prevención de lavado de activos hacia bancos locales e internacionales. Con respecto al otro cuadro, el mismo estimará el costo diario por parte de los recursos avocados al lavado de activos de la entidad, para mantener el modelo de monitoreo de las alertas descentralizado.

**Cuadro 11: Multas aplicadas a bancos por lavado de activos**

| Banco | Fecha de la multa | Monto de la multa aplicada | | Valores expresados en pesos al 31/12/2021 | |
|---|---|---|---|---|---|
| HSBC | 11/12/2012 | USD | 1.900.000.000,00 | $ | 109.938.411.189,16 |
| Standar Bank | 23/01/2014 | £ | 7.600.000,00 | $ | 1.068.938.877,87 |
| BBVA Frances | 14/10/2010 | $ | 39.393.072,00 | $ | 558.121.487,35 |
| Macro | 10/06/2011 | $ | 1.430.000,00 | $ | 19.381.386,53 |

**Cuadro 12: Costo por día en recursos humanos por mantener el sistema descentralizado del banco**

| Cargo | Cantidad | Haber Mensual | Haber diario | Haber Hora | Costo de la prevención de lavado de activos por día |
|---|---|---|---|---|---|
| Analista Sucursal | 640 | $ 250.000,00 | $ 8.333,33 | $ 1.041,67 | $ 666.666,67 |
| Supervisor Sucursal | 640 | $ 350.000,00 | $ 11.666,67 | $ 1.458,33 | $ 933.333,33 |
| Supervisor Zonal | 40 | $ 400.000,00 | $ 13.333,33 | $ 1.666,67 | $ 66.666,67 |
| **Total** | | | | | **$ 1.666.666,67** |

En lo que respecta al Cuadro 9, los valores actualizados a diciembre del 2021 para las multas hacia los bancos argentinos (BBVA Frances y Macro), se tomó como medida la inflación de acuerdo con el índice de precios al consumidor desde el momento en que se realizó la multa hasta la fecha tomada como corte de la medición. En cuanto a las multas de los otros dos bancos, se realizó el mismo procedimiento, pero anteriormente se convirtió el monto de la moneda de origen de cada multa a pesos argentinos al tipo de cambio de cuando se aplicó esta. Sobre el cuadro 10, se calculó un estimativo del haber mensual, diario y por hora, de cada recurso que participa para que el sistema de

monitoreo descentralizado pueda funcionar. Por eso, se estimó que para que el mismo pueda funcionar, se debe destinar una hora por día de cada recurso dando como total la suma de $1.666.666,67. Al llevar este importe diario a uno anual, el mismo sería de $608.333.333,33. A su vez, analizando los montos de las multas en pesos actualizados a diciembre del 2021, centrándonos en los bancos multados en la argentina y sumando ambos importes, estos no alcanzan al costo anual estimado para mantener el sistema descentralizado de monitoreo.

Continuando con el análisis del cuadro 9, las multas que han recibido estos dos bancos argentinos son de los años 2010 para BBVA Banco Frances y 2011 para el banco Macro, por lo que también podría estimarse una baja probabilidad de que las alertas cerradas en el sistema de monitoreo sean pasibles de multas por la Unidad de Información Financiera.

Del análisis realizado, también se le adiciona que durante el año 2021 los equipos de la unidad de prevención de lavado de activos han recibido 26.762 alertas para ser analizadas, de las cuales 25.085 han sido justificadas por estos. Sobre este número de alertas justificadas en esta instancia, 14.328 alertas corresponden con clientes catalogados por la unidad en riesgo bajo y medio en lavado de activos. El motivo de esta segmentación en las alertas tratadas por la unidad radica en que las alertas de clientes de riesgo alto que han llegado a la unidad presentan un control previo mucho más exhaustivo, de acuerdo con las normativas nacionales y recomendaciones de organismos internacionales (GAFILAT) y por lo tanto se tiene un conocimiento mayor del cliente alertado que con los clientes de riesgo medio y bajo. Como último, el monto promedio de estas 14.328 alertas es de $2.991.958,48.

Sobre estos datos explicados en los párrafos anteriores, se determinará el valor mínimo establecido de la métrica de evaluación del recall de la clase minoritaria para tomar la decisión si, de acuerdo con el primer objetivo de esta propuesta de trabajo, el árbol de decisión simple puede sustituir al modelo actual, y también servirá como métrica de evaluación en los demás modelos (segundo objetivo) que se generen, conforme con el resultado del primer objetivo. También se tomarán como valores mínimos de evaluación para decidir si el modelo propuesto sirve para cambiar al sistema de monitoreo descentralizado actual el accuracy y el recall en general del modelo. Estas dos métricas tienen una condición para que el modelo sea elegido, la cual es la siguiente:

| | | | | | | |
|---|---|---|---|---|---|---|
| **Valor Mínimo recall clase minoritaria** | **<** | **Valor recall clase minoritaria del modelo** | **<** | **Valor recall del modelo** | **<** | **Valor accuracy del modelo** |

Para el cálculo del valor mínimo del recall de la clase minoritaria se estimó, como se menciona anteriormente, una baja probabilidad de que el banco sea pasible de recibir una multa por parte del organismo regulador. Por esto, la probabilidad para la ocurrencia de este caso fue del 5%. Esta probabilidad se multiplicó por las 14.328 alertas de clientes catalogados como riesgo bajo y medio cerradas por en la unidad de prevención de lavado de activos, dando como resultado 716 alertas pasibles de que el banco sea multado en un escenario con baja probabilidad. Por otro parte para calcular la cantidad de alertas que corresponden con el costo anualizado de mantener el sistema descentralizado, el mismo se realizó en base a una división entre este costo anual y el monto promedio de las alertas cerradas por la UPLA de los clientes de riesgo bajo y medio. Este cálculo da como resultado una relación de 203 alertas. Por lo tanto, haciendo la relación de estos dos resultados (relación costo anual sistema descentralizado - monto promedio alertado, y relación de alertas pasibles de que el banco sea multado en un escenario con baja probabilidad), podemos estimar el error o trade-off de la clase minoritaria. El mismo arroja un resultado de 28,38%, dando como valor mínimo para el recall o sensibilidad de la clase minoritaria de 71,62%. Para una mejor interpretación de lo explicado, se adjunta esta información en los siguientes tres cuadros.

**Cuadro 13: Estimación del valor mínimo para el recall de la clase minoritaria**

| | | |
|---|---|---|
| Cantidad de alertas clientes riesgo bajo y medio justificadas en UPLA | | 14.328 |
| Probabilidad estimada en que la entidad sea multada por estas alertas | | 5% |
| Relación de alertas pasibles de que el banco sea multado en un escenario con baja probabilidad | | **716** |

| | | |
|---|---|---|
| Costo anual sistema descentralizado | $ | 608.333.333,33 |
| Monto promedio de las alertas clientes riesgo bajo y medio justificadas | $ | 2.991.958,48 |
| Relación en alertas entre el costo anual sistema descentralizado - monto promedio alertado | | **203** |

| | |
|---|---|
| Estimación del error en el recall de la clase minoritaria (203/716) | **28,38%** |
| Recall clase minoritaria mínimo esperado | **71,62%** |

## Resolución del primer objetivo

Para la resolución del primer objetivo, de acuerdo con lo mencionado antes, se ha entrenado un modelo árbol de decisión simple. Como primera medida, el mismo fue entrenado con datos correspondientes a alertas de los años 2019 y 2020. En esta etapa los resultados que arrojaron las métricas de evaluación en la etapa de validación fueron alcanzados. A la vez se analizando la métrica de precisión del modelo podemos decir que el modelo tiene una muy buena precisión en general. Los resultados auspiciosos arrojados por este modelo, no se han podido replicar en la etapa de test, cuando el mismo ha sido testeado con las alertas del año 2021, ya que no ha alcanzado el valor mínimo establecido para la métrica de recall de la clase minoritaria, como también podemos notar que la métrica de precisión baja considerablemente. Sus resultados son expuestos en el siguiente cuadro.

**Cuadro 14: Métricas modelo árbol de decisión simple**

| Modelo | Accuracy Validación del modelo | Recall Validación Modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|
| Árbol de decisión simple | 96% | 85% | 72% | 91% | 66% | 38% |

| Modelo | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|
| Árbol de decisión simple | 80% | 56% | 65% | 35% |

Luego de realizar el testeo del primer modelo se llega a la conclusión que este tipo de modelo, sin adherir nuevas variables y no optimizando sus hiper parámetros, no logra mejorar el sistema actual. Por lo tanto, se hacen los modelos de Gradient Boosting y Random Forest y a resolver el segundo objetivo planteado.

## Resolución del segundo objetivo

Con respecto al segundo objetivo, el mismo será el de encontrar entre los distintos modelos que se lleven a cabo, aquel que cumpla con valores mínimos aceptados de acuerdo con las métricas de evaluación para establecer un cambio en el sistema actual de monitoreo de la entidad. Para ello los modelos que se realicen se compararán con estas métricas y no contra el modelo de árbol de decisión simple que se realizó para analizar el primer objetivo.

### Subetapa 1: Modelos de árboles con el data set original y sin optimización de parámetros

Se realizaron dos modelos básicos de las técnicas de Gradient Boostiing y Random Forest con los siguientes resultados.

**Cuadro 15: Métricas modelo Gradient Boosting subetapa 1**

| Modelo | Accuracy Validación del modelo | Recall Validación Modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|
| Gradient Boosting | 95% | 66% | 32% | 94% | 60% | 19% |

| Modelo | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|
| Gradient Boosting | 74% | 53% | 65% | 31% |

**Cuadro 16: Métricas modelo Random Forest subetapa 1**

| Modelo | Accuracy Validación del modelo | Recall Validación Modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|
| Random Forest | 96% | 76% | 53% | 93% | 63% | 27% |

| Modelo | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|
| Random Forest | 54% | 18% | 60% | 35% |

Los dos modelos realizados en esta subetapa, correspondientes a las técnicas mencionadas anteriormente, han obtenido tanto en la etapa de validación del modelo como en la etapa de test una performance más baja que el modelo de árbol de decisión simple realizado en el primer objetivo. Analizando la métrica de precisión e ambos modelos, se nota que el modelo de Gradient

Boosting presenta una mayor porcentaje que el modelo de Random Forest debido a que selecciona una menor cantidad alertas a enviar a la UPLA que este último modelo mencionado. Por lo tanto, los mismos, han demostrado que hay que analizar otro tipo de variantes para sustituir al modelo actual del banco.

## Subetapa 2: Modelo de árboles con el data set original y optimización de parámetros

En esta subetapa, se visualizan los resultados obtenidos luego de realizar un random search de 15 interacciones para cada técnica de modelo de árboles utilizada, con el objetivo de optimizar los hiper parámetros de cada modelo. Los mismos fueron los siguientes, mostrando los primeros 7 modelos de cada técnica que habían obtenido la mejor performance en la etapa de validación.

**Cuadro 17: Métricas modelo árbol de decisión simple con optimización de parámetros subetapa 2**

| Modelo | Splitter | Min samples split | Min samples leaf | Max feauture s | Max depth | Criterion | Class weight | Accuracy Validación del modelo | Recall Validación del modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | best | 15 | 3 | auto | 25 | gini | balanced | 82% | 70% | 57% | 78% | 62% | 43% |
| 2 | random | 15 | 4 | sqrt | 15 | entropy | balanced | 60% | 73% | 89% | 58% | 61% | 66% |
| 3 | random | 2 | 1 | auto | 15 | gini | balanced | 64% | 65% | 65% | 69% | 58% | 45% |
| 4 | random | 2 | 3 | sqrt | 15 | entropy | balanced | 79% | 63% | 45% | 78% | 62% | 43% |
| 5 | best | 15 | 3 | sqrt | 15 | gini | balanced | 77% | 72% | 67% | 59% | 61% | 64% |
| 6 | best | 2 | 3 | auto | 15 | entropy | balanced | 63% | 68% | 74% | 68% | 61% | 53% |
| 7 | random | 2 | 1 | auto | 8 | entropy | balanced | 27% | 58% | 93% | 63% | 61% | 59% |

| Modelo | Splitter | Min samples split | Min samples leaf | Max feauture s | Max depth | Criterion | Class weight | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisió n test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | best | 15 | 3 | auto | 25 | gini | balanced | 55% | 35% | 47% | 33% |
| 2 | random | 15 | 4 | sqrt | 15 | entropy | balanced | 59% | 18% | 45% | 46% |
| 3 | random | 2 | 1 | auto | 15 | gini | balanced | 46% | 45% | 39% | 33% |
| 4 | random | 2 | 3 | sqrt | 15 | entropy | balanced | 47% | 33% | 46% | 32% |
| 5 | best | 15 | 3 | sqrt | 15 | gini | balanced | 60% | 49% | 47% | 45% |
| 6 | best | 2 | 3 | auto | 15 | entropy | balanced | 56% | 59% | 47% | 33% |
| 7 | random | 2 | 1 | auto | 8 | entropy | balanced | 70% | 13% | 47% | 38% |

## Cuadro 18: Métricas modelo Gradient Boosting con optimización de parámetros subetapa 2

| Modelo | N_estimators | Min samples split | Min samples leaf | Max depth | Loss | Learning_rate | Accuracy Validación del modelo | Recall Validación del modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 500 | 8 | 1 | 8 | deviance | 0,1 | 96% | 79% | 59% | 93% | 65% | 33% |
| 2 | 500 | 8 | 3 | 5 | deviance | 0,2 | 96% | 79% | 59% | 93% | 65% | 33% |
| 3 | 300 | 5 | 1 | 3 | deviance | 0,3 | 96% | 77% | 54% | 93% | 65% | 32% |
| 4 | 100 | 2 | 1 | 8 | exponential | 0,2 | 96% | 72% | 45% | 94% | 63% | 28% |
| 5 | 100 | 8 | 2 | 3 | deviance | 0,2 | 96% | 72% | 44% | 94% | 62% | 26% |
| 6 | 200 | 2 | 4 | 5 | exponential | 0,1 | 96% | 68% | 36% | 94% | 61% | 23% |
| 7 | 100 | 5 | 2 | 3 | exponential | 0,1 | 95% | 63% | 27% | 94% | 57% | 13% |

| Modelo | N_estimators | Min samples split | Min samples leaf | Max depth | Loss | Learning_rate | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 500 | 8 | 1 | 8 | deviance | 0,1 | 86% | 70% | 74% | 55% |
| 2 | 500 | 8 | 3 | 5 | deviance | 0,2 | 86% | 69% | 75% | 53% |
| 3 | 300 | 5 | 1 | 3 | deviance | 0,3 | 84% | 65% | 74% | 53% |
| 4 | 100 | 2 | 1 | 8 | exponential | 0,2 | 79% | 72% | 70% | 46% |
| 5 | 100 | 8 | 2 | 3 | deviance | 0,2 | 78% | 72% | 67% | 40% |
| 6 | 200 | 2 | 4 | 5 | exponential | 0,1 | 70% | 68% | 64% | 38% |
| 7 | 100 | 5 | 2 | 3 | exponential | 0,1 | 67% | 63% | 55% | 9% |

## Cuadro 19: Métricas modelo Random Forest con optimización de parámetros subetapa 2

| Modelo | N_estimators | Min_samples_split | Min_samples_leaf | Max_features | Max_depth | Criterion | Class_weight | Accuracy Validación del modelo | Recall Validación del | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1000 | 3 | 1 | auto | 8 | gini | balanced | 81% | 78% | 74% | 71% | 66% | 60% |
| 2 | 1000 | 5 | 3 | sqrt | 8 | gini | balanced | 80% | 78% | 74% | 71% | 66% | 60% |
| 3 | 100 | 8 | 4 | sqrt | 8 | entropy | balanced | 80% | 77% | 74% | 70% | 66% | 61% |
| 4 | 1000 | 3 | 3 | auto | 4 | gini | balanced | 81% | 78% | 74% | 68% | 65% | 61% |
| 5 | 300 | 10 | 4 | sqrt | 8 | entropy | balanced | 81% | 77% | 74% | 71% | 66% | 60% |
| 6 | 200 | 10 | 10 | auto | 8 | gini | balanced | 81% | 78% | 74% | 72% | 66% | 59% |
| 7 | 300 | 10 | 3 | sqrt | 15 | gini | balanced | 81% | 80% | 79% | 76% | 67% | 57% |

| Modelo | N_estimators | Min_samples_split | Min_samples_leaf | Max_features | Max_depth | Criterion | Class_weight | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1000 | 3 | 1 | auto | 8 | gini | balanced | 63% | 20% | 55% | 14% |
| 2 | 1000 | 5 | 3 | sqrt | 8 | gini | balanced | 63% | 21% | 55% | 14% |
| 3 | 100 | 8 | 4 | sqrt | 8 | entropy | balanced | 64% | 21% | 55% | 14% |
| 4 | 1000 | 3 | 3 | auto | 4 | gini | balanced | 62% | 19% | 55% | 14% |
| 5 | 300 | 10 | 4 | sqrt | 8 | entropy | balanced | 63% | 20% | 54% | 13% |
| 6 | 200 | 10 | 10 | auto | 8 | gini | balanced | 63% | 21% | 55% | 14% |
| 7 | 300 | 10 | 3 | sqrt | 15 | gini | balanced | 63% | 21% | 56% | 16% |

Los resultados de esta subetapa también nos demuestran que estos modelos, aun optimizando los hiper parámetros no logran superar el modelo actual. En los modelos realizados de árbol de decisión simple, se puede ver que en la etapa de validación los modelos 2, 6 y 7 de la tabla si bien los mismos superan el valor mínimo establecido para la clase minoritaria, estos no llegan a cumplir con las otras métricas debido a que el resultado es menor que la métrica de recall de dicha clase. Esta apreciación es porque estos tres modelos se han inclinado por predecir muchas alertas de las que deberían como injustificadas y por lo tanto que se traten por los equipos de la UPLA. A si mismo ninguno de los modelos de esta técnica en la etapa de testeo logra superar las métricas de performance establecidas. Por la técnica de Gradient Boosting, los modelos que mejor performance tuvieron, ninguno logra superar en la etapa de validación y en la de testeo, las métricas mínimas. En cuanto a los modelos de Random Forest, todos los modelos del cuadro en la etapa de validación cumple con los establecido para proponer una sustitución del modelo actual de monitoreo de alertas de la entidad, pero al pasar a la evaluación de estos con los datos de testeo, estos quedan lejos de superar la métrica evaluación del recall de la clase minoritaria.

A su vez, analizando la métrica de precisión de los modelos de estas tres variantes, se puede notar que los modelos de la técnica de Gradient Boosting obtienen mejores resultados. La ocurrencia de estos mejores porcentajes se debe a que en esta técnica en la librería utilizada no cuenta dentro de sus hiper parámetros con la propiedad de balancear las clases, y al tratarse de un data set con un elevado porcentaje de una clase y otro porcentaje relativamente menor de otra clase, estos modelos predicen de manera más habitual la clase mayoritaria, prediciendo muy pocas alertas de la clase minoritaria. Por eso que la precisión de estos ha sido bastante buena. Por el contrario, los modelos de las otras dos técnicas utilizadas al presentar este hiper parámetro generan un mejor porcentaje de Recall pero la precisión no supera en casi todos los modelos el 50% de efectividad.

## Subetapa 3: Modelo de árboles con agregado de las variables de riesgo y operatoria transaccional, sin optimización de parámetros

En esta subetapa del segundo objetivo en donde agregamos al data set original las variables de riesgo y operatoria transaccional para analizar si las mismas logran ayudar a mejorar las predicciones de los modelos de árboles realizados. el resultado de estos no llega a ser lo suficientemente buenos para tomar la decisión de cambiar el sistema de monitoreo actual de la entidad financiera.

**Cuadro 20: Métricas modelo árbol de decisión simple subetapa 3**

| Modelo | Accuracy Validación del modelo | Recall Validación Modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|
| Árbol de decisión simple | 96% | 84% | 71% | 90% | 66% | 39% |

| Modelo | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|
| Árbol de decisión simple | 80% | 56% | 65% | 41% |

**Cuadro 21: Métricas modelo Gradient Boosting subetapa 3**

| Modelo | Accuracy Validación del modelo | Recall Validación Modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|
| Gradient Boosting | 95% | 66% | 32% | 94% | 59% | 18% |

| Modelo | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|
| Gradient Boosting | 74% | 53% | 64% | 30% |

**Cuadro 22: Métricas modelo Random Forest subetapa 3**

| Modelo | Accuracy Validación del modelo | Recall Validación Modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|
| Random Forest | 96% | 72% | 45% | 94% | 60% | 21% |

| Modelo | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|
| Random Forest | 54% | 14% | 55% | 27% |

## Subetapa 4: Modelo de árboles con agregado de las variables de riesgo y operatoria transaccional, con optimización de parámetros

En la última subetapa del segundo objetivo, al optimizar los hiper parámetros de los distintos modelos de árboles realizados, se puede notar una mejoría en las métricas de performance de estos. Incluso esta mejora se ve reflejada en los modelos de Random Forest, lo cual seis de los siete modelos en la etapa de validación superan los criterios de evaluación establecidos. Si bien estos resultados fueron alentadores en los indicadores del recall y el accuracy, en la métrica de precisión los modelos no han logrado una mejora importante en la performance de estos en comparación con los modelos realizados en la subetapa 2.

A su vez probar estos modelos en la etapa de test, la performance de estos no ha podido superar los valores de evaluación y por lo tanto se llegó a la conclusión que este tipo de modelos, con la optimización de hiper parámetros y la adhesión de nuevas variables, no logran ser lo bastante buenos para querer tomar la decisión de cambiar al sistema actual del banco.

**Cuadro 23: Métricas modelo árbol de decisión simple con optimización de parámetros**

| Modelo | Splitter | Min_samples_split | Min_samples_leaf | Max_features | Max_depth | Criterion | Class_weight | Accuracy Validación del modelo | Recall Validación del modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | random | 2 | 2 | sqrt | 8 | gini | balanced | 54% | 60% | 68% | 94% | 57% | 15% |
| 2 | random | 10 | 1 | auto | 8 | gini | balanced | 27% | 53% | 84% | 66% | 61% | 55% |
| 3 | best | 8 | 5 | sqrt | 8 | gini | balanced | 85% | 59% | 29% | 71% | 61% | 49% |
| 4 | best | 15 | 3 | log2 | 15 | entropy | balanced | 91% | 64% | 32% | 11% | 51% | 97% |
| 5 | best | 15 | 2 | auto | 8 | entropy | balanced | 76% | 66% | 53% | 80% | 59% | 34% |
| 6 | best | 3 | 5 | auto | 8 | entropy | balanced | 82% | 58% | 30% | 83% | 87% | 26% |
| 7 | random | 3 | 4 | sqrt | 2 | gini | balanced | 93% | 51% | 3% | 72% | 53% | 32% |

| Modelo | Splitter | Min samples split | Min samples leaf | Max feautures | Max depth | Criterion | Class weight | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | random | 2 | 2 | sqrt | 8 | gini | balanced | 46% | 50% | 38% | 50% |
| 2 | random | 10 | 1 | auto | 8 | gini | balanced | 40% | 18% | 45% | 46% |
| 3 | best | 8 | 5 | sqrt | 8 | gini | balanced | 46% | 45% | 46% | 33% |
| 4 | best | 15 | 3 | log2 | 15 | entropy | balanced | 47% | 33% | 38% | 9% |
| 5 | best | 15 | 2 | auto | 8 | entropy | balanced | 47% | 49% | 42% | 36% |
| 6 | best | 3 | 5 | auto | 8 | entropy | balanced | 56% | 59% | 33% | 45% |
| 7 | random | 3 | 4 | sqrt | 2 | gini | balanced | 49% | 25% | 44% | 35% |

## Cuadro 24: Métricas modelo Gradient Boosting con optimización de parámetros

| Modelo | N_estimators | Min samples split | Min samples leaf | Max depth | Loss | Learning_rate | Accuracy Validación del modelo | Recall Validación del modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 500 | 8 | 1 | 8 | deviance | 0,1 | 97% | 79% | 60% | 93% | 65% | 32% |
| 2 | 500 | 8 | 1 | 5 | deviance | 0,2 | 96% | 78% | 57% | 93% | 65% | 33% |
| 3 | 500 | 5 | 3 | 3 | deviance | 0,3 | 96% | 76% | 53% | 93% | 65% | 32% |
| 4 | 100 | 2 | 1 | 8 | exponential | 0,2 | 92% | 72% | 43% | 94% | 63% | 27% |
| 5 | 200 | 8 | 2 | 3 | deviance | 0,2 | 96% | 72% | 42% | 94% | 64% | 28% |
| 6 | 200 | 2 | 4 | 5 | exponential | 0,1 | 95% | 67% | 35% | 93% | 60% | 22% |
| 7 | 100 | 5 | 2 | 3 | exponential | 0,1 | 95% | 63% | 25% | 94% | 57% | 13% |

| Modelo | N_estimators | Min samples split | Min samples leaf | Max depth | Loss | Learning_rate | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 500 | 8 | 1 | 8 | deviance | 0,1 | 86% | 68% | 74% | 55% |
| 2 | 500 | 8 | 1 | 5 | deviance | 0,2 | 86% | 69% | 75% | 53% |
| 3 | 500 | 5 | 3 | 3 | deviance | 0,3 | 83% | 65% | 74% | 53% |
| 4 | 100 | 2 | 1 | 8 | exponential | 0,2 | 79% | 70% | 70% | 49% |
| 5 | 200 | 8 | 2 | 3 | deviance | 0,2 | 78% | 72% | 64% | 44% |
| 6 | 200 | 2 | 4 | 5 | exponential | 0,1 | 68% | 65% | 66% | 38% |
| 7 | 100 | 5 | 2 | 3 | exponential | 0,1 | 67% | 60% | 55% | 12% |

## Cuadro 25: Métricas modelo Random Forest con optimización de parámetros

| Modelo | N_estimators | Min_samples_split | Min_samples_leaf | Max_features | Max_depth | Criterion | Class_weight | Accuracy Validación del modelo | Recall Validación del modelo | Recall Validación clase 1 | Accuracy test del modelo | Recall test del modelo | Recall test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 300 | 10 | 1 | sqrt | 20 | entropy | balanced | 86% | 82% | 76% | 83% | 66% | 47% |
| 2 | 200 | 5 | 3 | auto | 20 | entropy | balanced | 85% | 81% | 77% | 82% | 67% | 48% |
| 3 | 300 | 10 | 4 | sqrt | 20 | gini | balanced | 85% | 81% | 77% | 82% | 66% | 48% |
| 4 | 200 | 2 | 3 | sqrt | 15 | gini | balanced | 84% | 80% | 75% | 79% | 66% | 51% |
| 5 | 500 | 10 | 4 | sqrt | 15 | entropy | balanced | 83% | 80% | 77% | 78% | 66% | 53% |
| 6 | 500 | 10 | 1 | auto | 8 | gini | balanced | 82% | 78% | 73% | 79% | 66% | 50% |
| 7 | 1000 | 8 | 2 | auto | 8 | entropy | balanced | 83% | 77% | 69% | 73% | 65% | 55% |

| Modelo | N_estimators | Min_samples_split | Min_samples_leaf | Max_features | Max_depth | Criterion | Class_weight | Precisión Validación del modelo | Precisión Validación clase 1 | Precisión test del modelo | Precisión test clase 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 300 | 10 | 1 | sqrt | 20 | entropy | balanced | 64% | 20% | 58% | 20% |
| 2 | 200 | 5 | 3 | auto | 20 | entropy | balanced | 64% | 19% | 58% | 19% |
| 3 | 300 | 10 | 4 | sqrt | 20 | gini | balanced | 63% | 21% | 57% | 19% |
| 4 | 200 | 2 | 3 | sqrt | 15 | gini | balanced | 64% | 19% | 56% | 17% |
| 5 | 500 | 10 | 4 | sqrt | 15 | entropy | balanced | 63% | 20% | 56% | 17% |
| 6 | 500 | 10 | 1 | auto | 8 | gini | balanced | 63% | 21% | 56% | 17% |
| 7 | 1000 | 8 | 2 | auto | 8 | entropy | balanced | 64% | 20% | 56% | 18% |

## Eficacia de los modelos que superaron la etapa de validación en identificar las alertas de mayor riesgo en lavado de activos enviadas a la UPLA

Luego de haber terminado las distintas subetapas para encontrar un modelo que sea pasible de cambiar el sistema actual de monitoreo del banco, en esta subsección se analizará la eficacia en predecir que alertas enviadas a analizar por los equipos de la UPLA representan un mayor riesgo en lavado de activos.

Para ello se seleccionaron los modelos que cumplieron con los criterios de evaluación en la etapa de validación. los mismos han sido trece (uno de la primera subetapa, 7 de la segunda subetapa y 6 de la cuarta subetapa).

**Cuadro 26: Eficacia de los modelos en detectar las alertas de mayor riesgo**

| Modelo | Eficacia |
|---|---|
| Árbol de decisión simple SE nro 1 | 49,25% |
| Random Forest Modelo 1 SE nro 2 | 61,67% |
| Random Forest Modelo 2 SE nro 2 | 61,55% |
| Random Forest Modelo 3 SE nro 2 | 62,24% |
| Random Forest Modelo 4 SE nro 2 | 61,99% |
| Random Forest Modelo 5 SE nro 2 | 61,41% |
| Random Forest Modelo 6 SE nro 2 | 61,49% |
| Random Forest Modelo 7 SE nro 2 | 59,56% |
| Random Forest Modelo 1 SE nro 4 | 57,43% |
| Random Forest Modelo 2 SE nro 4 | 57,13% |
| Random Forest Modelo 3 SE nro 4 | 57,12% |
| Random Forest Modelo 4 SE nro 4 | 60,25% |
| Random Forest Modelo 5 SE nro 4 | 61,04% |
| Random Forest Modelo 6 SE nro 4 | 61,29% |

El resultado de este arrojó que los modelos de Random Forest obtuvieron una mejor eficacia enviar las alertas de mayor riesgo a analizar en los equipos de la UPLA. Estos obtuvieron en general valores por encima del 60% de eficacia. Mientras que el primer modelo realizado obtuvo una eficacia menor al 50%. La diferencia se sostiene en que los modelos de la técnica Random Forest han predicho un mayor número de alertas a analizar por los equipos UPLA (aproximadamente 60.000 alertas por año contra 20.000 del modelo de árbol de decisión). Si bien la eficacia de los modelos de Random Forest es mejor, el gran volumen de alertas que estos predicen que deben ser analizadas por los equipos de la UPLA generaría hoy en día un cuello de botella en esta última etapa del sistema de prevención de alertas, ocasionado que complicaciones en la atención de las alertas y posibles sanciones por el ente regulador a no cumplir con el plazo establecidos de 150 días en el caso que se deba injustificar una alerta y enviar a la UIF.

## Testeo de los modelos en otras muestras

Tras haber realizado las etapas en la parte de resultados y demostrar que estos modelos con la adhesión de distintas variables no han podido cumplir con el objetivo propuesto de modificar el sistema actual del banco, se procedió a analizar los modelos que superaron las métricas de evaluación establecidas con nuevos datos. Estos corresponden con el período enero a mayo del 2022, dentro de estos se encuentran 176.959 alertas con un desbalance marcado de las clases como los anteriores data sets (2,8% para la clase minoritaria y 97,2% la clase mayoritaria). Los resultados de los modelos en esta nueva muestra fueron los siguientes:

**Cuadro 27: Métricas modelo Árbol de decisión simple subetapa 1. Datos out of sample**

| Modelo | Accuracy test del modelo | Recall test Modelo | Recall test clase 1 | precisión test del modelo | precisión test clase 1 |
|---|---|---|---|---|---|
| Árbol de decisión simple | 91% | 65% | 38% | 56% | 14% |

**Cuadro 28: Métricas modelos Random Forest subetapa 2. Datos out of sample**

| Modelo | Accuracy test del modelo | Recall test Modelo | Recall test clase 1 | precisión test del modelo | precisión test clase 1 |
|---|---|---|---|---|---|
| 1 | 61% | 63% | 65% | 51% | 5% |
| 2 | 61% | 63% | 65% | 52% | 5% |
| 3 | 61% | 63% | 65% | 52% | 5% |
| 4 | 57% | 62% | 67% | 51% | 4% |
| 5 | 61% | 63% | 65% | 52% | 5% |
| 6 | 63% | 63% | 65% | 52% | 5% |
| 7 | 65% | 64% | 61% | 52% | 5% |

**Cuadro 29: Métricas modelos Random Forest subetapa 4. Datos out of sample**

| Modelo | Accuracy test del modelo | Recall test Modelo | Recall test clase 1 | precisión test del modelo | precisión test clase 1 |
|---|---|---|---|---|---|
| 1 | 84% | 62% | 37% | 53% | 7% |
| 2 | 83% | 62% | 39% | 53% | 7% |
| 3 | 80% | 61% | 42% | 52% | 6% |
| 4 | 74% | 61% | 48% | 52% | 5% |
| 5 | 73% | 61% | 49% | 52% | 5% |
| 6 | 67% | 61% | 56% | 51% | 5% |

Los resultados obtenidos por estos modelos demuestran que han sido más bajos que los logrados por la muestra de alertas del año 2021. Incluso se puede observar que la sensibilidad de estos modelos, en especial los modelos con la adhesión de las variables generadas por la UPLA, resulto menor a la esperada. Analizando la métrica de precisión, también se visualizan valores muy bajos en todos los modelos, debiéndose a, como se explicó anteriormente sobre estas dos técnicas, al utilizar el hiper parámetro de “Class_Weight”, los modelos predicen que muchas alertas deben ser analizadas por los equipos de la unidad de prevención de lavado de activos, ocasionando que la precisión de estos se vea afectada, en especial a intentar predecir la clase minoritaria.

# Conclusiones

El presente trabajo se ha centrado en encontrar una nueva alternativa al modelo de prevención de lavado de activos actual de monitoreo de la entidad financiera analizada. Los resultados de las demostrado que estos tipos de modelos con las variables utilizadas no logran mejorar el sistema descentralizado actual de prevención de este banco público. El presente trabajo es relevante para el banco y otras instituciones financieras como punto de partida para mejorar el sistema de prevención y sus alertas. Entre los puntos a considerar para continuar con el análisis con el fin de mejorar el sistema de prevención, se mencionan los siguientes:

Como primera medida, modificar el rango de tiempo en cómo se han tomado los datos para entrenar los modelos y luego testearlos. En este trabajo se optó por entrenar los modelos con los datos de los años 2019 y 2020 en conjunto y testear estos con datos del año 2021 y los primeros cinco meses del año 2022 como muestra de testeo alternativa. Se podría probar con utilizar los mismos de manera mensual, validarlos con los datos del mes siguiente y luego testearlos con el de mes subsiguiente (ejemplo: Datos enero 2020, validación con datos de febrero 2020 y testeo con datos de marzo 2020).

En cuanto a los modelos desarrollados, los mismos se basaron en distintas técnicas de árboles de decisión. Dentro de estos, los modelos basados en árboles de decisión simple y Random Forest han sido los que mejores resultados obtuvieron, no siendo el caso de la técnica Gradient Boosting. Los resultados desfavorables que se obtuvieron en este modelo fueron por el problema de las clases, muy desbalanceadas, en los datos. Está técnica al contrario de las otras dos, no tiene un hiper parámetro para controlar este problema de clases desbalanceadas (Hiper parámetro: “Class_Weight”). Como mejora se podría implementar un tratamiento previo en estos datos para que los modelos de Gradient Boosting puedan mejorar, tales como la técnica Synthetic Minority Oversampling Technique o SMOTE (Alberto Fernández, 2018).

Otra medida importante a analizar con detenimiento es la importancia que tuvieron las nuevas variables agregadas y si las mismas ayudaron a los modelos utilizados en mejorar la predicción de estos. Por lo tanto, como mejora sería la de modificar estas nuevas variables con otras medidas estadísticas (aquí solo se utilizó la media en distintos periodos) y analizar si las mismas sirven para mejorar métricas de evaluación utilizadas en los modelos.

Otro punto por considerar sería la implementación de otros algoritmos de clasificación que no estén relacionados con los modelos de árboles. Entre los modelos que podrían ser tenidos en cuenta y que han sido testeados tanto en la prevención de lavado de activos como en problemas similares como la prevención de fraude, y con buenos resultados son los de Support Vector Machines (Eweoya, Adebiyi, Azeta y Olufunmilola Amosu, 2019). Otros modelos que podrían analizarse en futuros trabajos y de con un nivel de dificultad mayor son los de redes neuronales (Desrousseaux, Bernard, Mariage. 2021) y modelos de aprendizaje no supervisado (Sain, Puri. 2018).

# Anexo 1: Fuentes utilizadas

**Alertas generadas por el sistema de monitoreo**

Esta fuente contiene los datos de las alertas generadas de manera mensual. En esta base aparecen los datos que identifican a la persona que realizó la operación, junto con los datos identificatorios de la alerta generada. Esta fuente contiene 12 columnas, las cuales se detallan:

- ID Cliente
- Cliente
- CUIL/CUIT
- Tipo Persona
- Sucursal
- Zonal
- Estado
- Acción
- Tipo de Alerta
- Número de Alerta
- Fecha Alerta
- Monto Alerta

A continuación, se muestran las primeras filas, junto con las columnas descriptas.[1]

[1] Los datos de identificación de los clientes fueron modificados para preservar la confidencialidad de estos.

**Tabla 1: Muestra de la base Alertas generadas por el sistema de monitoreo noviembre 2021**

| ID Cliente | Cliente | CUIL/CUIT | Tipo Persona | Sucursal | Zonal | Estado | Accion | Tipo de Alerta | Nro de Alerta | Fecha Alerta |
|---|---|---|---|---|---|---|---|---|---|---|
| 14440 | XXXXXX | XXXXXXXXX | F | BALVANERA | CONGRESO | Pendiente Analista Sucursal | Analizar en sucursal | USU.BN2 - Constitución de Plazo fijo en efectivo | 5732713 | 2021-11-30 00:00:00.000 |
| 81939 | XXXXXX | XXXXXXXXX | F | HURLINGHAM | LINIERS | Justificada Sucursal | NULL | USU.BN2 - Constitución de Plazo fijo en efectivo | 5732714 | 2021-11-30 00:00:00.000 |
| 78760 | XXXXXX | XXXXXXXXX | F | PALERMO | PALERMO | Pendiente Analista Sucursal | Analizar en sucursal | USU.BN2 - Constitución de Plazo fijo en efectivo | 5732715 | 2021-11-30 00:00:00.000 |
| 69143 | XXXXXX | XXXXXXXXX | F | LA PLATA | LA PLATA | Justificada Sucursal | NULL | USU.BN2 - Constitución de Plazo fijo en efectivo | 5732716 | 2021-11-30 00:00:00.000 |
| 4532 | XXXXXX | XXXXXXXXX | F | AVENIDA GAONA | PALERMO | Pendiente Analista Sucursal | Analizar en sucursal | TRF.L02 - Transferencias nacionales recibidas | 5713371 | 2021-11-30 00:00:00.000 |
| 80754 | XXXXXX | XXXXXXXXX | F | RIO CUARTO | RIO CUARTO | Pendiente Analista Sucursal | Analizar en sucursal | USU.BN1 - Operaciones realizadas con dinero en efectivo | 5725983 | 2021-11-30 00:00:00.000 |
| 78270 | XXXXXX | XXXXXXXXX | F | C.BUSTOS IFFLINGER | VILLA MARIA | Justificada Sucursal | NULL | TRF.L02 - Transferencias nacionales recibidas | 5713372 | 2021-11-30 00:00:00.000 |
| 6260 | XXXXXX | XXXXXXXXX | F | PARANA | PARANA | Justificada Sucursal | NULL | USU.BN1 - Operaciones realizadas con dinero en efectivo | 5725984 | 2021-11-30 00:00:00.000 |
| 82020 | XXXXXX | XXXXXXXXX | F | DAIREAUX | TRENQUE LAUQUEN | Justificada Sucursal | NULL | USU.BN2 - Constitución de Plazo fijo en efectivo | 5732717 | 2021-11-30 00:00:00.000 |
| 65189 | XXXXXX | XXXXXXXXX | F | URDINARRAIN | PARANA | Pendiente Analista Sucursal | Analizar en sucursal | TRF.L02 - Transferencias nacionales recibidas | 5713373 | 2021-11-30 00:00:00.000 |
| 69935 | XXXXXX | XXXXXXXXX | F | CAMPANA | SAN ISIDRO | Justificada Sucursal | NULL | USU.BN2 - Constitución de Plazo fijo en efectivo | 5732718 | 2021-11-30 00:00:00.000 |
| 47448 | XXXXXX | XXXXXXXXX | F | COMODORO RIVADAVIA | COMODORO RIVADAVIA | Pendiente Analista Sucursal | Analizar en sucursal | TRF.L02 - Transferencias nacionales recibidas | 5713374 | 2021-11-30 00:00:00.000 |
| 32023 | XXXXXX | XXXXXXXXX | F | AVENIDA LA PLATA | FLORES | Justificada Sucursal | NULL | USU.BN2 - Constitución de Plazo fijo en efectivo | 5732719 | 2021-11-30 00:00:00.000 |
| 95335 | XXXXXX | XXXXXXXXX | F | ALTA GRACIA | V.CARLOS PAZ | Pendiente Supervisor Sucursal | Resolver en Sucursal | TRF.L02 - Transferencias nacionales recibidas | 5713375 | 2021-11-30 00:00:00.000 |
| 14740 | XXXXXX | XXXXXXXXX | F | ALTA GRACIA | V.CARLOS PAZ | Justificada Sucursal | NULL | USU.BN6 - Pago de Cheques en Efectivo por Cliente | 5702830 | 2021-11-09 00:00:00.000 |
| 59829 | XXXXXX | XXXXXXXXX | F | ALTA GRACIA | V.CARLOS PAZ | Justificada Sucursal | NULL | USU.BN5 - Cobro de Cheques en Efectivo por Beneficiario | 5702958 | 2021-11-09 00:00:00.000 |
| 77150 | XXXXXX | XXXXXXXXX | F | LA CARLOTA | RIO CUARTO | Justificada Sucursal | NULL | TRF.L02 - Transferencias nacionales recibidas | 5713376 | 2021-11-30 00:00:00.000 |
| 39135 | XXXXXX | XXXXXXXXX | F | 5A. SECCION | MENDOZA ESTE | Justificada Sucursal | NULL | USU.BN2 - Constitución de Plazo fijo en efectivo | 5732720 | 2021-11-30 00:00:00.000 |

**Base de datos de los clientes registrados en el banco**

Esta fuente contiene los datos identificatorios de los clientes registrados en el banco. La misma es generada de manera mensual y contiene datos descriptivos, geográficos, impositivos y de los productos que tiene el cliente en dicha entidad. Esta fuente contiene 144 columnas, entre las cuales se encuentran:

- ID Cliente
- Nacionalidad
- Edad
- Tipo de persona
- Sucursal, Zonal y Regional en donde se encuentra radicado
- Variables relacionadas a la actividad de los clientes
- Variables impositivas de los clientes
- Datos demográficos de los clientes
- Cantidad de productos de cada clientes

A continuación, se muestran las primeras filas, junto con las columnas descriptas.[2]

[2] Los datos de identificación de los clientes fueron modificados para preservar la confidencialidad de estos.

**Tabla 2: Muestra de la base de datos de los clientes registrados en el banco noviembre 2021**

| ID Cliente | Tipo Persona | SUCURSAL | ZONAL | REGIONAL_UPLA | FORMA_JURIDICA | CODIGO_ACTIVIDAD | POSICION_IVA | SECTOR | TOTAL_SUMA_PRODUCTOS |
|---|---|---|---|---|---|---|---|---|---|
| 8984549 | F | BELEN | S.F.V. CATAMARCA | SALTA | 0 - EMP.UNIP. | 821 - PERS.FIS | 4 - CONSUMIDOR FINAL | 1 - SECTOR PRIVADO NO FINANCIERO | 11 |
| 8984727 | F | 28 DE NOVIEMBRE | COMODORO RIVADAVIA | SUR | 0 - EMP.UNIP. | 821 - PERS.FIS | 4 - CONSUMIDOR FINAL | 1 - SECTOR PRIVADO NO FINANCIERO | 5 |
| 8984979 | F | MATTALDI | RIO CUARTO | CORDOBA | 0 - EMP.UNIP. | 151 - AGRIC-GANAD | 1 - RESPONSABLE INSCRIPTO | 1 - SECTOR PRIVADO NO FINANCIERO | 2 |
| 8985076 | F | BARRANQUERAS | RESISTENCIA | CORRIENTES | 0 - EMP.UNIP. | 821 - PERS.FIS | 4 - CONSUMIDOR FINAL | 1 - SECTOR PRIVADO NO FINANCIERO | 9 |
| 8985098 | F | ROJAS | PERGAMINO | JUNIN | 0 - EMP.UNIP. | 821 - PERS.FIS | 4 - CONSUMIDOR FINAL | 1 - SECTOR PRIVADO NO FINANCIERO | 7 |
| 8989904 | J | BARRIO SAN VICENTE | CORDOBA | CORDOBA | 1 - SOCIEDAD ANONIMA | 352 - MET.NO FERROSO | 1 - RESPONSABLE INSCRIPTO | 1 - SECTOR PRIVADO NO FINANCIERO | 7 |
| 8991030 | J | CARLOS CALVO | CONGRESO | CONGRESO | 1 - SOCIEDAD ANONIMA | 372 - MAQ.HERRAM. | 1 - RESPONSABLE INSCRIPTO | 1 - SECTOR PRIVADO NO FINANCIERO | 3 |
| 8985471 | F | CIPOLLETTI | NEUQUEN | MENDOZA | 0 - EMP.UNIP. | 821 - PERS.FIS | 4 - CONSUMIDOR FINAL | 1 - SECTOR PRIVADO NO FINANCIERO | 6 |
| 8985778 | F | GENERAL PICO | SANTA ROSA | JUNIN | 0 - EMP.UNIP. | 657 - MENOR-OTROS | 1 - RESPONSABLE INSCRIPTO | 1 - SECTOR PRIVADO NO FINANCIERO | 4 |
| 8985793 | F | VILLA ORTUZAR | PALERMO | PALERMO | 0 - EMP.UNIP. | 822 - JUBIL-PENS | 4 - CONSUMIDOR FINAL | 1 - SECTOR PRIVADO NO FINANCIERO | 7 |
| 8985794 | F | B NAV PTO BELGRANO | BAHIA BLANCA | SUR | 0 - EMP.UNIP. | 821 - PERS.FIS | 4 - CONSUMIDOR FINAL | 1 - SECTOR PRIVADO NO FINANCIERO | 2 |
| 8985903 | F | TRES ARROYOS | AZUL | SUR | 0 - EMP.UNIP. | 821 - PERS.FIS | 4 - CONSUMIDOR FINAL | 1 - SECTOR PRIVADO NO FINANCIERO | 1 |

**Base de datos sobre el riesgo de lavado para la UPLA**

Esta fuente contiene los datos relacionados al riesgo en materia de lavado de activos que puede generar el cliente en la entidad. Esta base es generada de forma cuatrimestral y aparecen los datos de probabilidad, impacto y riesgo de cada cliente registrado por el banco. Las mismas variables pueden contener solamente 4 valores (bajo, medio, alto y null en caso de no haberse evaluado al cliente). Esta fuente contiene 10 columnas, las cuales se detallan:

- ID Cliente
- Cliente
- CUIL/CUIT
- Tipo Persona
- Probabilidad
- Impacto
- Riesgo actual
- Periodo actual
- Riesgo anterior
- Periodo anterior

A continuación, se muestran las primeras filas, junto con las columnas descriptas.[3]

[3] Los datos de identificación de los clientes fueron modificados para preservar la confidencialidad de estos.

**Tabla 3: Muestra de la base de datos sobre el riesgo de lavado para la UPLA agosto 2021**

| ID Cliente | Cliente | CUIL/CUIT | Tipo Persona | PROBABILIDAD | IMPACTO | RIESGO_ACTUAL | PERIODO_ACTUAL | RIESGO_ANTERIOR | PERIODO_ANTERIOR |
|---|---|---|---|---|---|---|---|---|---|
| 8501510 | XXXXXXX | XXXXXXXX | F | MEDIO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8501811 | XXXXXXX | XXXXXXXX | F | BAJO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8501824 | XXXXXXX | XXXXXXXX | F | BAJO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8501992 | XXXXXXX | XXXXXXXX | F | BAJO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502025 | XXXXXXX | XXXXXXXX | F | BAJO | MEDIO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502049 | XXXXXXX | XXXXXXXX | F | BAJO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502062 | XXXXXXX | XXXXXXXX | F | BAJO | MEDIO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502150 | XXXXXXX | XXXXXXXX | F | BAJO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502156 | XXXXXXX | XXXXXXXX | F | BAJO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502289 | XXXXXXX | XXXXXXXX | F | BAJO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502349 | XXXXXXX | XXXXXXXX | F | ALTO | BAJO | MEDIO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502437 | XXXXXXX | XXXXXXXX | F | MEDIO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502475 | XXXXXXX | XXXXXXXX | F | MEDIO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502569 | XXXXXXX | XXXXXXXX | F | BAJO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502899 | XXXXXXX | XXXXXXXX | F | ALTO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502913 | XXXXXXX | XXXXXXXX | F | ALTO | BAJO | MEDIO | 31/08/2021 | BAJO | 30/04/2021 |
| 8502997 | XXXXXXX | XXXXXXXX | F | MEDIO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8503121 | XXXXXXX | XXXXXXXX | F | BAJO | MEDIO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8503220 | XXXXXXX | XXXXXXXX | F | BAJO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |
| 8503374 | XXXXXXX | XXXXXXXX | F | BAJO | BAJO | BAJO | 31/08/2021 | BAJO | 30/04/2021 |

**Base de datos sobre movimientos en efectivo**

Esta fuente contiene la cantidad y montos mensuales en efectivo que realizaron los clientes en el banco. Esta fuente contiene 6 columnas, las cuales se detallan:

- ID Cliente
- Cliente
- Cuil/Cuit
- Tipo de persona
- monto de los movimientos en efectivo por cliente
- cantidad de movimientos en efectivo por cliente

A continuación, se muestran las primeras filas, junto con las columnas descriptas.[4]

Tabla 4: Muestra de la base de datos sobre movimientos en efectivo noviembre 2021

| ID Cliente | Cliente | CUIL/CUIT | Tipo Persona | Monto Movimientos en efectivo | Cantidad de Movimientos en efectivo |
|---|---|---|---|---|---|
| 8984549 | xxxxxxx | xxxxxxxxxx | F | $ - | 0 |
| 8984727 | xxxxxxx | xxxxxxxxxx | F | $ 1.556,45 | 1 |
| 8984979 | xxxxxxx | xxxxxxxxxx | F | $ - | 0 |
| 8985076 | xxxxxxx | xxxxxxxxxx | F | $ - | 0 |
| 8985098 | xxxxxxx | xxxxxxxxxx | F | $ - | 0 |
| 8985302 | xxxxxxx | xxxxxxxxxx | F | $ - | 0 |
| 8985408 | xxxxxxx | xxxxxxxxxx | F | $ - | 0 |
| 8985471 | xxxxxxx | xxxxxxxxxx | F | $ - | 0 |
| 8985778 | xxxxxxx | xxxxxxxxxx | F | $ - | 0 |
| 8985793 | xxxxxxx | xxxxxxxxxx | F | $ 138.500,00 | 4 |

[4] Los datos de identificación de los clientes fueron modificados para preservar la confidencialidad de estos.

**Base de datos sobre movimientos al crédito**

Esta fuente contiene la cantidad y montos mensuales de los créditos que realizaron y/o recibieron los clientes en el banco. Esta fuente contiene 6 columnas, las cuales se detallan:

- ID Cliente
- Cliente
- Cuil/Cuit
- Tipo de persona
- monto de los créditos realizado y/o recibidos por cliente
- cantidad de créditos realizado y/o recibidos por cliente

A continuación, se muestran las primeras filas, junto con las columnas descriptas.[5]

Tabla 5: Muestra de la base de datos sobre movimientos al crédito noviembre 2021

| ID Cliente | Cliente | CUIL/CUIT | Tipo Persona | Total creditos | Cantidad creditos |
|---|---|---|---|---|---|
| 3471408 | xxxxxxx | xxxxxxxxxx | F | $ 7.423.342,58 | 4 |
| 3507981 | xxxxxxx | xxxxxxxxxx | F | $ 4.844.434,94 | 7 |
| 3511208 | xxxxxxx | xxxxxxxxxx | J | $ 10.000.000,00 | 7 |
| 3600416 | xxxxxxx | xxxxxxxxxx | J | $ 353.176,53 | 21 |
| 18657119 | xxxxxxx | xxxxxxxxxx | J | $ 13.139.232,88 | 4 |
| 6881643 | xxxxxxx | xxxxxxxxxx | J | $ 1.307.863,25 | 3 |
| 24739394 | xxxxxxx | xxxxxxxxxx | F | $ 4.239.361,36 | 1 |
| 3614988 | xxxxxxx | xxxxxxxxxx | F | $ 24.357.203,47 | 3 |
| 3628133 | xxxxxxx | xxxxxxxxxx | F | $ 28.669.542,06 | 13 |
| 3722119 | xxxxxxx | xxxxxxxxxx | F | $ 7.446.915,00 | 1 |

[5] Los datos de identificación de los clientes fueron modificados para preservar la confidencialidad de estos.

# Bibliografía


- GAFILAT. Las recomendaciones del GAFI- estándares internacionales sobre la lucha contra el lavado de activos, el financiamiento del terrorismo y la proliferación de armas de destrucción masiva. Recomendación 1: “Evaluación de riesgos y aplicación de un enfoque basado en riesgo”.
- GAFILAT. Las recomendaciones del GAFI- estándares internacionales sobre la lucha contra el lavado de activos, el financiamiento del terrorismo y la proliferación de armas de destrucción masiva. Recomendación 2: “Cooperación y coordinación nacional”.
- GAFILAT. Las recomendaciones del GAFI- estándares internacionales sobre la lucha contra el lavado de activos, el financiamiento del terrorismo y la proliferación de armas de destrucción masiva. Recomendación 29: “Unidades de inteligencia financiera”.
- Harold Koster, 2019. Towards better implementation of the European Union’s anti-money laundering and countering the financing of terrorism framework.
- Yan Zhang - Peter Trubey. 2018. Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection.
- Y. Sahin and E. Duman, 2011. Detecting Credit Card Fraud by Decision Trees and Support Vector Machines.
- Eduardo Fabián Caparrós, 2018. Combate al Lavado de Activos desde el Sistema Judicial
- Veronica Smink. BBC News Mundo, Buenos Aires. 12/08/2019. PASO: el dólar se dispara en Argentina tras las elecciones primarias en las que arrasó el kirchnerismo
- Infobae. 10/09/2019. Qué es "el rulo" del dólar oficial, la operación que permite ganar 5% en pocos minutos
- Matías Becerra. Cronista. 28/10/2019. Endurecen el cepo: baja a u$s 200 el tope a las compras mensuales

- James G, Witten D, Hastie T, Tibshirani R. 2017. An Introduction to Statistical Learning. 8 edición. Springer.
- Breiman. 2001. Random Forest.
- Bengio Y, Bergstra J. 2012. Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research.
- Chen, T., & He, T. (2015, August). Higgs boson discovery with boosted trees. In NIPS 2014 workshop on high-energy physics and machine learning (pp. 69-80).
- BBC, 2012. Multa récord al HSBC por posibilitar lavado de dinero.
- Michael Kilmes. International Business Times 2014. FCA Fines Standard Bank £7.6m for Slack Anti-Money Laundering Controls.
- La Política Online, 2010. Multan al BBVA Banco Francés con $39 millones por presunto lavado.
- El Cronista, 2011. Multan al Macro por no denunciar operaciones sospechosas de lavado de dinero.
- Alberto Fernández, 2018. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary.
- Eweoya, Adebiyi, Azeta y Olufunmilola Amosu, 2019. Fraud prediction in loan default using support vector machine.
- Desrousseaux, Bernard, Mariage. 2021. Profiling Money Laundering with Neural Networks: A Case Study on Environmental Crime Detection.
- Sain, Puri. 2018. Detection of money laundering accounts using data mining techniques.