DETECCIÓN DE OPERACIONES SOSPECHOSAS DE LAVADO DE ACTIVOS EN EL SISTEMA FINANCIERO, USANDO VARIABLES NO TRANSACCIONALES, MÁQUINAS DE SOPORTE VECTORIAL Y ÁRBOLES DE CLASIFICACIÓN.

Marlon Efraín Gracia Granados

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
MEDELLÍN
2016

DETECCIÓN DE OPERACIONES SOSPECHOSAS DE LAVADO DE ACTIVOS EN EL SISTEMA FINANCIERO, USANDO VARIABLES NO TRANSACCIONALES, MÁQUINAS DE SOPORTE VECTORIAL Y ÁRBOLES DE CLASIFICACIÓN.

Marlon Efraín Gracia Granados

Trabajo de investigación para optar al título de Maestría en Ingeniería

Directora

Ana Lucía Pérez Patiño

Doctora en Ingeniería

Universidad Nacional Sede Medellín

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN

2016

RESUMEN

El lavado de activos es un delito que mueve anualmente entre el 2% y el 5% del producto interno bruto mundial, este delito trae consigo varios efectos sociales y económicos, como la desconfianza en el sistema financiero, inflación de los precios de ciertos bienes y el fortalecimiento económico de las bandas criminales. El delito de lavado de activos se concentra principalmente en el sector financiero, a pesar de que existen múltiples sistemas para identificar y reportar las operaciones sospechosas. Sin embargo, el problema radica en que los sistemas de detección pueden producir altas tasas de falsos positivos, debido a que las operaciones reportadas como inusuales realmente son operaciones legales.

La literatura disponible muestra múltiples aproximaciones para la detección de operaciones sospechosas de lavado de activos que intentan obtener bajas tasas de falsos positivos, sin embargo, la mayoría de soluciones propuestas no incluyen datos de anomalías reales, esto por la dificultad de obtener este dato de las entidades financieras, por lo que sus modelos son entrenados con datos sintéticos para simular los datos anómalos, por lo que al momento de su implementación los resultados presentan limitaciones. Adicionalmente, ninguna de estas propuestas ha explorado la utilización de variables no transaccionales, con las cuales se podrían tener mejores tasas de falsos positivos.

Como resultado de esta investigación, se propone un modelo de detección de operaciones sospechosas de lavado de activos que incorpore, además de las variables transaccionales convencionales, variables no transaccionales para obtener mejores tasas de falsos positivos y permitir su validación y comparación con las otras técnicas en un conjunto de datos completamente real.

Palabras clave: Detección de anomalías, Lavado de dinero, Falsos positivos, Operaciones sospechosas.

ABSTRACT

Money laundering is a crime that generates great concern internationally, this due to the large amounts of money that annually moves worldwide and the social and economic impact that it generates. In the financial sector which is where this crime occurs mostly, anti-money laundering systems has been implemented in order to identify and report suspicious transactions of money laundering, the problem is that these detection systems can produce high false positive rates, e.g. many operations reported as unusual when actually they are legal operations, this causes a lot of time lost in the investigation of operations that really are not unusual and can hinder the detection and reporting system.

Doing a literature review, proposals were found for the detection of suspicious transactions of money laundering that try to produce low false positives rates. The problem is that many of these solutions do not include data from real anomalies, so at the time of its actual implementation outcomes may not be desired. Additionally none of the proposals have explored the use of non-transactional variables, which could have better false positives rates.

This investigation aims to design a model of detection of suspicious transactions of money laundering that incorporates in addition to conventional transactional variables, non-transactional variables with which you can get better false positives rates and can be tested and compared with other techniques on a set of completely real data.

CONTENIDO

LISTADO DE FIGURAS

LISTADO DE TABLAS

1. INTRODUCCIÓN

En este trabajo de investigación se realizó un estudio sobre algunos métodos de detección de anomalías utilizados actualmente para la detección de operaciones sospechosas de lavado de activos en las entidades financieras, con miras a obtener varias métricas que permitan determinar cuál de estos métodos de detección posee las mejores propiedades, sus ventajas, desventajas y con esa información poder crear un modelo de detección con unas mejores métricas y así poder realizar una contribución al estado del arte de los modelos de detección de operaciones sospechosas de lavado de activos. Este primer capítulo busca dar un contexto de la problemática y delimitar el objetivo a abordar en la investigación, detallando porqué es un problema, sus efectos (sociales y económicos) y la importancia de poder mitigar estos, finalmente este capítulo concluye mostrando la estructura del documento.

1.1 Contexto de la problemática

El lavado de activos es una actividad por la cual se ingresa dinero obtenido de actividades criminales como tráfico de drogas, tráfico de personas, venta de armas y corrupción al sistema financiero, el fin es darle apariencia de legalidad[1]. Este es un delito que mueve entre el 2% y el 5% del producto interno bruto mundial [2][3][4]. Los esfuerzos a nivel internacional van desde mitigar el delito [5], hasta la conformación de entidades que ayuden a la prevención, detección y la generación de recomendaciones a entidades financieras que implementan sistemas anti lavado de dinero para identificar, investigar y reportar transacciones sospechosas a las autoridades competentes [6][7].

En 1989, el G7 creó el FATF/GAFI (Financial Action Task Force / Grupo de Acción Financiera Internacional) con la intención de generar políticas anti lavado[5] y la primera recomendación dada por este grupo, es la criminalización del lavado de activos denotando como delito según la convención de Viena y la convención de Palermo[6].

En el caso colombiano, el lavado de activos está tipificado como delito en el artículo 323 del código penal[8] y considera que se incurre en él al adquirir, resguardar, invertir o transportar los bienes con origen mediato o inmediato del tráfico de migrantes, la trata de personas, la extorsión, el tráfico de armas, el tráfico de estupefacientes, entre otros.

El lavado de activos es un delito que normalmente se lleva a cabo en tres etapas, a saber: i) colocación (Etapa en la cual ocurre el ingreso de los capitales ilícitos al sistema financiero), ii) transformación (Etapa en la cual mediante un conjunto de transacciones se trata de esconder el origen de los recursos), iii) Integración (Etapa en la cual se devuelven los fondos al sector real para ser utilizado por los criminales)[9].

En las entidades financieras, se tienen implementados sistemas ALD (Anti Lavado de Dinero) para detectar y evitar las operaciones de lavado de activos que se realicen en la institución. El problema es que la detección es una tarea complicada, bien sea por la habilidad de los delincuentes para evitar la detección o porque existen operaciones inusuales originadas con dineros completamente lícitos que surgen por ganancias ocasionales; por todo lo anterior en las alertas detectadas por el sistema ALD se pueden generar falsos positivos (operaciones reportadas como sospechosas de ser empleadas para lavar activos cuando provienen realmente de dineros lícitos) y de falsos negativos (operaciones de lavado de activos que no fueron reconocidos por el sistema de detección)[7].

Los efectos negativos que trae consigo el delito de lavado de activos, afecta a quienes sufren el delito fuente de manera directa, por ejemplo: el sector real, el sector público, el sector financiero y la sociedad en general[2][10] debido a que es un delito que además de darle apariencia de legalidad al dinero ilícito, también ayuda a darle cierta legalidad al crimen como tal[6]. Este tipo de delito puede distorsionar las inversiones e incrementar los precios de los bienes, alterando así la demanda y los precios de los productos[9].

Finalmente, es importante mencionar que otro de los efectos negativos del lavado de activos y uno de los más peligrosos, es el hecho de que es un delito que puede ser un multiplicador de actividades criminales, debido a que los criminales ganan un mayor

poder económico[2] con el cual pueden autofinanciarse, diversificar sus actividades y expandirse a otros lugares[5].

1.2 Definición del problema

Si bien el lavado de activos como tal es un problema en la sociedad, no es este el problema que se quiere abordar en esta investigación. El problema que se desea estudiar y mitigar es la generación de altas cifras de falsos positivos al momento de detectar operaciones sospechosas de lavado de activos en las instituciones financieras, las cuales según una encuesta realizada a varios encargados de administrar los sistemas ALD de diferentes entidades financieras se pudo observar que estas cifras de falsos positivos pueden llegar incluso a valores cercanos al 100% [11].

Las implicaciones que tiene este problema son varias, por ejemplo,obtener altas cifras de falsos positivos en los sistemas ALD de una entidad financiera, ocasiona que se pierda tiempo analizando operaciones lícitas como si fueran sospechosas, reduciendo así la capacidad del sistema para analizar operaciones que realmente son sospechosas. De esta manera más operaciones sospechosas pueden pasar inadvertidas y materializar el delito de lavado de activo generando a su vez todos los efectos sociales y económicos negativos detallados en el punto anterior. Adicionalmente, podría generar pérdida de la reputación de la institución financiera, la cual no reportó a tiempo estas operaciones y recibir multas por parte de las entidades encargadas de regular que el sistema de prevención y detección de lavado de activos de las entidades financieras funcione correctamente[12][13].

Para mitigar este problema, hoy por hoy se crean más modelos de detección de lavado de activos. Estos modelos utilizan técnicas robustas que buscan optimizar las diferentes métricas que indican que tan buenos son para detectar las operaciones sospechosas de lavados de activos.Estos modelos son generados a partir de técnicas estadísticas y matemáticas los cuales buscan identificar patrones anómalos en un conjunto de datos.

Actualmente es difícil identificar cuál de los modelos propuestos es el mejor, esto debido a que cada uno de los modelos realizados utilizan conjuntos de datos diferentes, donde las características de los datos son diferentes, las métricas que miden también los son y finalmente y más importante cómo los datos para anomalías en su mayoría son generados, la manera cómo son generados varían también entre método y método.

Una de las mayores limitaciones que existen en los modelos de detección de operaciones sospechosas de lavado de activos que se han implementado hasta el día de hoy, es la falta de anomalías reales con las cuales se puedan entrenar los modelos. Esto porque los modelos generados a partir de datos sintéticos que en teoría simulan las anomalías, cuando son aplicados a datos reales no obtienen los mejores resultados. Lo anterior porque en estos datos reales las diferencias entre operaciones inusuales y no inusuales no es tan marcadas como por lo general lo son en los datos sintéticos generados.

El punto central de este trabajo de investigación, radica en realizar una implementación de cada una de las técnicas estudiadas con un conjunto de datos reales, realizar mediciones que dan cuenta acerca de las principales características del modelo a la hora de predecir datos, y a partir de estas métricas realizar una comparación para establecer sobre los modelos estudiados con cual modelo se obtiene los mejores resultados, además de las bondades y desventajas de la utilización de cada uno de estos modelos y con esta información poder generar un nuevo modelo de detección capaz de superar las métricas de los otros modelos.

La hipótesis que se quiere validar al realizar la presente investigación, es que utilizando datos no transaccionales, como lo son las características de quien realiza la operación más que los datos de la operación misma, se puede obtener un modelo con mejores capacidades de detección de inusualidades.

Las contribuciones que por medio de este trabajo se esperan lograr, son la comparación entre modelos existentes y el poder establecer cuáles de estos según las métricas comúnmente utilizadas es el mejor. Adicionalmente se desea poder generar un nuevo modelo que supere las métricas que se obtienen actualmente.

Este documento tiene la siguiente estructura: después de la introducción se aborda la problemática a estudiar, a continuación el estado del arte de la detección de lavado de activos, posteriormente se hace el diseño de experimento que se seguirá, luego viene la implementación y comparación de todos los modelos de detección revisados, seguido por la implementación del nuevo modelo de detección, finalizando con las conclusiones y trabajo futuro.

2. PROBLEMÁTICA

El lavado de activos es un problema global, el cual a medida que ha crecido la globalización este también lo hace pues a medida que se facilitan las operaciones a nivel internacional este delito se aprovecha de ello para de igual manera generar mayores ingresos a partir de él[1].

Los efectos que este delito genera en la población son diversos, en el ámbito macroeconómicos puede llegar a "contaminar" el dinero legítimo por sospecha de provenir de una jurisdicción donde se conozcan varios casos de lavado de activos[14].Puede ocasionar también la alteración de los flujos de inversión y ahorros, debido a que se pueden llegar a dar inversiones cuyo objetivo sea disfrazar el origen de dinero más que para obtener rentabilidad a partir de esas inversiones lo cual genera que el dinero no sea invertido de la manera más óptima[2].

Siguiendo la misma línea anterior, el lavado de activo puede llevar también a una "subida artificial" de precios, esto debido a que bienes como los inmuebles pueden ser utilizados para lavar dinero al comprarlo por valores mayores a su valor comercial. Esto ocasiona que como se oferta por esta clase de bienes un valor mayor al usual, eventualmente el precio demandado por estos crece corrompiendo así la economía y haciendo mas difícil la adquisición de esta clase de producto a las demás personas[2].

Pero el lavado de activos no solo afecta a la población desde la perspectiva macro, el lavado de activos también tiene efecto negativo directo sobre la población, un ejemplo son todas aquellas personas que son víctimas directas de los delitos fuentes de lavado de activos, es decir, las personas que son utilizadas para la trata de blancas, las personas que son extorsionadas, las poblaciones afectadas por la venta de armas. Incluso la sociedad como un todo se puede ver afectada, esto porque el lavado de activo suele traer consigo actividades como el soborno que van permeando en la sociedad y pueden llegar a crear una cultura de crimen[2][9].

Adicional a lo mencionado anteriormente, el lavado de activos también puede llegar a producir los siguientes efectos negativos sobre la población, distorsión de estadísticas

macroeconómicas, afectar la reputación del sistema financiero, aumentar la corrupción y el soborno, contaminar las instituciones políticas, aumentar el terrorismo y aumentar las actividades delictivas[2][9].

La prevención y detección de las operaciones sospechosas de lavado de activos se realiza mayormente en las instituciones financieras pues es allí donde usualmente se realizan las fases del lavado de activo (colocación, transformación e integración)[15], para lograr esta tarea usualmente se utilizan sistemas ALD (Anti lavado de dinero) los cuales son plataformas que automáticamente permiten la clasificación de operaciones entre normales y sospechosas[7], estos sistemas se basan usualmente de algunos umbrales que se parametrizan en su sistema y que comparan estadísticas básicas como medias y medianas para determinar si una operación se encuentra en los "rangos normales" y es clasificada como una operación no sospechosa o si por otro lado supera estos umbrales y debe ser clasificada como una operación inusual[16].

Luego de la clasificación realizada por estos sistemas usualmente son los trabajadores de la instituciones financieras quienes realizan un trabajo de investigación para determinar si efectivamente la operación reportada como sospechosa si es realmente una operación inusual, de serlo procede a realizar el reporte respectivo a la autoridad competente y de no serlo, archiva la operación como un falso positivo[5][6].

Los sistemas clásicos ALD en ocasiones suelen ser muy simples en su estructura y por tanto son muy susceptibles a errores, la clásica orientación a definir inusualidades por medio de percentiles, medias y mediana, genera modelos pocos robustos y que pueden llegar a generar altas tasas de fallos[16].

Para dar cuenta del nivel de confianza y la calidad de los resultados arrojados por los diversos sistemas ALD de varias entidades financieras alrededor del mundo, una encuesta fue realizada por las compañías Dow Jones Risks & Compliance y ACAMS, por medio de esta encuesta se entrevistóa más de 600 encargados de los sistemas de ALD en varias instituciones globalmente[11].

Los resultados de esta encuesta cuando se le pregunto a los participantes acerca de el porcentaje de falsos positivos que existe en sus instituciones se puede observar en la Figura 1.

Figura 1. Encuesta realizada sobre tasas de falsos positivos. Tomado de: [Dow jones y ACAMS, 2011]

Se puede observar varios puntos interesantes a partir de la grafica expuesta por la encuesta. Primero que todo se puede observar que existe un grupo en la población que no se encuentra en capacidad de responder el porcentaje de falsos positivos que su sistema produce, lo cual da una idea de lo difícil que puede resultar el establecer si una operación reportada por un sistema de detección de operaciones sospechosas si es inusual o no. Otra observación importante que puede resultar luego de mirar los datos obtenidos, es como el 44% de la población encuestada contesto que posee unas tasas de falsos positivos superior al 50% e incluso un 14% informo que el 100% de las operaciones reportadas son falsos positivos.

Lo anterior sirve para dar una idea del problema que surge al intentar mitigar el problema principal, el lavado de activos. Dado que al intentar tener un sistema ALD que ayude a detectar las operaciones de lavado de activo en una institución financiera se puede llegar a generar un alto número de falsos positivos que trae con sigo nuevas implicaciones y varios efectos negativos.

Entre los principales efectos negativos que se tiene al producirse altas tasas de falsos positivos se encuentra el tiempo que se pierde por parte de los investigadores en revisar la operación, la frustración y la pérdida de la confianza en el sistema que genera las alertas, para la institución financiera representa perdida de dinero al tener a un conjunto de investigadores revisando operaciones en las cuales los dineros provienen de fuentes lícitas y pérdida de capacidad investigativa pues al enfocar esfuerzos en analizar operaciones que en realidad no son sospechosas se pierde capacidad para analizar aquellas transacciones que si lo son [11].

Adicionalmente a los efectos directos anteriormente mencionados, al no poder detectar correctamente las operaciones que fueron utilizadas para disfrazar el origen ilegal de ciertos dineros, el lavado de activos puede cumplir sus fases de colocación, transformación e integración completamente, por lo que se pueden materializan los efectos negativos que fueron mencionados al principio de este capítulo.

Por todo lo anterior, la problemática que sedesea abordar y sobre la cual se desea realizar una contribución para ayudar a mitigar sus efectos es la siguiente. ¿Cómo se puede reducir la generación de falsos positivos al momento de detectar operaciones sospechosas de lavado de activos en el sistema financiero?

Revisando la literatura existente al día de hoy, notamos que son varias las aproximaciones que se han realizado para utilizar técnicas más robustas para detectar operaciones sospechosas de lavado de activos[17][18], pero también hemos podido observar que dada la dificultad para poder obtener un conjunto real de datos anómalos para poder entrenar los modelos de detección varios investigadores se han visto obligados a entrenar sus técnicas con datos sintéticos[19]. Pero al tomar este camino dos problemas pueden ocurrir, primero que estos datos sintéticos no representen de una manera correcta el comportamiento de los datos sospechosos que se presentan en una institución financiera y por tanto el modelo entrenado al utilizarlo con operaciones reales no ofrezca buenos resultados. Otro inconveniente que puede suceder es que la forma como se genera estos conjuntos varié drásticamente entre investigador e investigador, por tanto no permita que sean comparables modelos diferentes propuestos por investigadores diferentes.

Adicionalmente se pudo notar en varios estudios revisados que las variables no transaccionales han sido poco utilizadas, es decir variables como la edad, los ingresos de la persona, su ocupación, sus egresos y demás información que no es propia de la transacción realizada sino de quien la realiza. Se considera que esta información puede llegar a ser muy relevante al momento de la construcción de un modelo de detección más afinado, puesto que esas variables pueden dar una idea acerca del perfil de la persona quien realiza las operaciones y se puede llegar a determinar si las operaciones realizadas son coherentes con el perfil de la persona quien la realiza, siendo esta otra manera de determinar inusualidades en las operaciones estudiadas.

Por todo lo anterior, la hipótesis de investigación que se desea contrastar es: usar datos no transaccionales en un modelo de detección de operaciones sospechosas de lavado de activos puede ayudar a reducir las tasas de falsos positivos.

3. ESTADO DEL ARTE DE LA DETECCIÓN DE OPERACIONES SOSPECHOSAS DE LAVADO DE ACTIVOS

Para el problema de la detección y prevención del lavado de activos en las entidades financieras existen múltiples soluciones planteadas, desde controles y políticas guiadas por las recomendaciones del FATF[6] como topes para ciertos tipos de transacciones, documentación adicional para ciertos grupos de personas y la caracterización del cliente, hasta la investigación de operaciones sospechosas generadas por el sistemas de detección de inusualidades y la realización de reportes a las entidades pertinentes. Estas soluciones de detección de inusualidades se apoyan en diversos métodos estadísticos y matemáticos para encontrar comportamientos atípicos, como son las técnicas de redes neuronales[20], las máquinas de soporte vectorial[21], los modelos de Markov ocultos[22], el sequence matching[23], las redes bayesianas dinámicas[24] y los algoritmos de clasificación[25].

Rohit y Patel[18] crearon una categorización de varios tipos de enfoques que se pueden seguir a la hora de detectar estas operaciones de lavado, a saber: clasificación, detección de datos atípicos, visualización de los datos, técnicas de regresión, predicción y finalmente agrupamiento.

En este capítulo se detalla algunos de los modelos utilizados en la literatura para afrontar el problema de la detección de operaciones sospechosas de lavado de activos en el sistema financiero por medio de la utilización de técnicas estadísticas y matemáticas que ayuden a reducir las tasas de falsos positivos y aumentar las tasas de detección.

3.1 Métodos estadísticos y matemáticos aplicados a la detección de lavado de activos

3.1.1 Redes bayesianas dinámicas

Una de las aplicaciones revisadas para la detección de lavado de activos es la realizada por Raza y Haider[24], ellos combinaron múltiples métodos para la identificación de anomalías.

Esta solución parte desde una etapa de agrupación, la cual es realizada por medio de un algoritmo de agrupamiento difuso y que busca formar grupos de clientes diferentes donde el comportamiento transaccional de las personas que pertenecen a un mismo grupo sea similar y que por otro lado difiere de las personas de los otros grupos. Las variables que se tienen en cuenta para conformar estas agrupaciones son los montos transados por los clientes, las frecuencias de sus transacciones y el tiempo que ocurría entre dos transacciones consecutivas.

El paso a seguir en la solución es realizar, para cada uno de los grupos generados, una red bayesiana dinámica, donde se tuviera por cada registro las últimas N transacciones realizadas por el cliente con el detalle de las tres variables que se utilizaron para conformar la red, el monto de la transacción, el medio de pago (Cheque, efectivo, traslado virtual) y el periodo donde ocurrió la transacción (Inicio, mitad o fin de mes).

Con las redes bayesianas dinámicas ya generadas entra en juego la última etapa de la solución planteada, la etapa de identificación de anomalías, en la cual con las N-1 transacciones previas se calcula la probabilidad de ocurrencia de cada una de las tres variables planteadas para la transacción N y se corrobora con la operación que realmente ocurrió.

A partir de esta comparación se obtienen dos datos, un rango el cual expresa según las probabilidades posteriores, qué posición ocuparía el valor real de cada variable según el pronosticado, es decir, 1 si según estas probabilidades el valor real de la variable era el más probable, 2 si era el segundo más probable y así para cada variable. El segundo

dato que se obtiene es la entropía de cada variable de la operación la cual se calcula de la siguiente manera.

$$Entropia = \sum_{1}^{k} \frac{plog_2(p)}{log_2(k)}$$

Siendo k el número de estados de la variable a la que se le calcula la entropía y p la probabilidad de que la transacción N tome el valor del estado que se está iterando.

Con estos dos valores se construye un índice propio denominado AIRE (Anomaly Index using Rank and Entropy) con el cual se calcula el grado de anomalía de la operación, el cual si supera cierto umbral predefinido es posteriormente marcado como una operación inusual. La manera como se calcula este índice es la siguiente.

$$AIRE_i = \frac{r - 1 + e * (.5 + .5k - r)}{k - 1}$$

Donde r es el rango, e la entropía y k el número de estados de la variable i, posteriormente los indicadores que de momento están a nivel de cada variable son ponderados por medio de la siguiente fórmula.

$$AIRE = \sum_{1}^{i} W_i * AIRE_i$$

Donde W representa los pesos de ponderación que son definidos subjetivamente por quien realiza el método, consolidando un único indicador AIRE para la transacción el cual será comparado contra un umbral para determinar si la transacción es categorizada o no como inusual.

Al final, para corroborar los resultados obtenidos, se revisa la tasa de predicción del modelo, comparando para cada variable el valor real contra el valor más probable según el modelo y contando el porcentaje de veces donde estos dos valores coinciden. El obtener buenos porcentajes de predicción sugiere para los clientes que vayan en contra

de varias de estas predicciones, una alta probabilidad de que estén presentando un comportamiento inusual, debido a que sus transacciones no son consecuentes con el conjunto de transacciones realizadas previamente.

3.1.2 Redes neuronales de base radial

Otra solución planteada para detectar anomalías en transacciones en busca de lavado de activos es la propuesta por Lin-Tao, Na-Ji y Jiu-Long[20], la cual utiliza una red neuronal de base radial, el objetivo de esta red es: dado un conjunto de variables de entradas obtener una clasificación en dos posibles grupos (normal o sospechoso),las variables de entrada definidas son la frecuencia de retiros, la frecuencia de los ingresos y los montos totales transados, estas variables de entrada son pre procesadas antes de ser utilizadas propiamente en el algoritmo, esto realizando una diferencia normalizada con lo cual se garantiza que cada variable esté en un rango de 0 a 1, lo cual se realiza para que las variables sean comparables y de entrada no tienen pesos diferentes, para esto se utilizó una diferencia normalizada.

$$\varepsilon_i = max_i(X_{ij}) - min_i(X_{ij})$$

Donde en $\varepsilon_i$representa la diferencia normalizada (diferencia entre el valor más grande y más pequeño de cada atributo i) y X_{ij} representa el valor del atributo i en el dato j.
Posteriormente cada valor de cada variable se normaliza de la siguiente manera.

$$X'_{ij} = \frac{X_{ij} - min_i(X_{ij})}{\varepsilon_i}$$

Donde X'_{ij} será el nuevo valor el cual estará normalizado y solo tendrá valores en el rango [0,1].
Estas variables de entrada ya pre procesadas son mapeadas a la capa oculta por medio de una función de base radial que utiliza la función gaussiana y posteriormente el paso de esta capa oculta a la capa de salida que determina la clasificación final es realizado por medio de una función lineal que utiliza unos pesos que se calculan en el algoritmo, la ecuación con la que se puede resumir lo anterior es la siguiente.

$$Y_i = W_0 + \sum_{i=1}^{m} W_j \emptyset(\|x - C_i\|)$$

Donde Y es la salida en la cual se indica si la operación es sospechosa o no, W es el conjunto de pesos, X son los datos de entrada, C los centros de las capas ocultas, m el número de capas ocultas y ∅ es la función gaussiana, la cual a su vez se define de la siguiente manera.

$$\emptyset(\|x - C_i\|) = \begin{cases} \exp\left(-\frac{(\|x - C_i\|)^2}{\sigma_i^2}\right) i = 1, 2, \dots, m \\ 1\ i = 0 \end{cases}$$

En la anterior ecuación σ representa el ancho de la base radial de la función gaussiana.
En total el algoritmo requiere calcular tres parámetros para que la red neuronal de base radial pueda funcionar correctamente, los centros y los anchos de la base radial los cuales son requeridos para ser utilizados en la función de base radial, y los pesos que son utilizados para realizar el paso de la capa oculta a la capa de salida.
El método que se utiliza para calcular los centros es un clúster APC-III, este algoritmo de agrupamiento requiere, un parámetro α sea suministrado; posteriormente los autores indican que encontraron mejores resultados cuando dicho parámetro es inferior a 1.04, el resultado que se obtiene de este algoritmo es el conjunto de centros de cada capa oculta que son requeridos para la utilización de la función de base radial. Para el cálculo del ancho de la base radial de cada una de las capas ocultas son utilizados los centros previamente obtenidos, los cuales están en función de la distancia máxima entre el centro de cada capa y datos de entrenamiento y el número de capas ocultas que posee la red neuronal. Por último, para el cálculo del último parámetro que son los pesos con los cuales se realiza el mapeo de la capa oculta a la capa de salida, se utiliza un algoritmo RLS.

Para realizar la prueba de la solución planteada se utilizó un conjunto de transacciones reales que se obtuvieron de una entidad financiera y adicionalmente se añadieron transacciones sospechosas ficticias debido a la dificultad de obtener transacciones que claramente fueran casos reales de lavado de activos. Con este conjunto de datos, primero se escoge una muestra que contenga datos normales y datos sospechosos y se entrena el modelo hasta que se tenga un error por debajo de 0.01 y posteriormente se escoge otra muestra para probar los resultados del modelo obtenido, adicionalmente con

ese mismo conjunto de datos se obtienen los resultados para otras dos técnicas, una máquina de soporte vectorial y un proceso de detección de datos atípicos.

El criterio de comparación utilizado entre estas tres técnicas es la tasa de detección y la tasa de falsos positivos. Con los resultados obtenidos se realiza la comparación, donde se obtiene que para la técnica de máquina de soporte vectorial probada la tasa de detección es aproximadamente 30% y la tasa de falsos positivos 10%, para la técnica de detección de datos atípicos se obtuvo una tasa de detección aproximadamente del 40% y una tasa de falsos positivos aproximada de 10% y finalmente la técnica propuesta por los autores obtuvo una tasa de detección superior al 80% y una tasa de falsos positivos inferior al 5%.

3.1.3 Máquinas de soporte vectorial

Las máquinas de soporte vectorial también han sido utilizadas en la detección de operaciones sospechosas de lavado de activos, un ejemplo de aplicación de esta técnica en este ámbito fue realizado por Tang y Yin[21], ellos realizaron dos propuestas para lograr la detección de inusualidades.

La primera propuesta consistía en la versión supervisada de la técnica, propiamente el algoritmo C-SVM, que al ser un método supervisado cuenta con una primera muestra inicial a partir de la cual se entrenará el modelo y que contiene una marca sobre cuáles operaciones son inusuales y cuáles operaciones no los son, a partir de estos datos y con el conjunto de variables de entrada se crea un hiperplano que divide de manera óptima estos puntos tratando de conservar separadas las transacciones normales y las sospechosas, la manera como se construya este hiperplano dependerá de los atributos del conjunto de datos, pudiendo ser este linealmente separable, linealmente inseparable o no lineal.

Para la segunda propuesta se realizó la versión no supervisada de la técnica por medio del algoritmo SVM de una clase, con el cual se busca con un conjunto de datos que ya no contiene ninguna marca, encontrar una función que separe los datos que contienen características muy similares que se espera sean la mayoría de los datos de los otros datos que serían marcados como inusuales, para encontrar dicha función primero por medio de una función kernel se transforma el conjunto de datos de entrada a una dimensión más alta en la cual se encontrará el hiperplano que permita la división del conjunto de datos en dos conjuntos, uno que posee un gran número de datos y serán catalogados como transacciones normales y otro grupo que posee el resto de registros los cuales serán catalogados como inusuales.

En la función kernel utilizada en la segunda propuesta los autores realizaron una modificación, en vez de utilizar la clásica función gaussiana como función kernel, hicieron una optimización en la distancia que es requerida en dicha función gaussiana, no calculándola de manera normal si no utilizando una variante denominada HVDM

(Métrica para la diferencia de valores heterogéneos) con la cual se busca que si el conjunto de datos contiene datos heterogéneos la distancia aún pueda ser calculada de manera que la implementación del algoritmo SVM no posea inconvenientes.

Luego de que se definieron estas dos propuestas, finalmente la segunda fue la utilizada para realizar pruebas, el conjunto de datos fue una mezcla de una gran cantidad de información real obtenida de una entidad financiera y un pequeño conjunto de transacciones artificiales creadas para ser reconocidas como inusualidades, esto para asegurarse de que el conjunto de datos que se utilizaban en la prueba si contuviera un conjunto de datos sospechosos.

Para poder ejecutar el modelo planteado finalmente era requerido obtener dos parámetros, el parámetro de penalización por clasificación incorrecta y el parámetro de factor de control, los cuales fueron hallados por ensayo y error debido a que los autores de la técnica no encontraron un método de escogencia de parámetros lo suficientemente bueno. El criterio para verificar los resultados obtenidos nuevamente fueron la tasa de detección y la tasa de falsos positivos, los cuales obtuvieron valores de 69% y 3.4% respectivamente en sus mejores versiones.

3.1.4 Clúster de dos fases

Otra propuesta realizada en el campo de la detección de anomalías es la dada por M.F. Jiang, S.S. Tseng y C.M. Su[25] la cual no fue aplicada propiamente en el campo de detección de lavado de activos, pero que fácilmente se podría aplicar en ese ámbito.

Esta propuesta plantea la solución al problema de detección de anomalías en dos fases. En una primera fase explican cómo realizar un proceso de agrupación por medio del algoritmo modificado de K-medias, el cual con la modificación realizada incorpora una heurística que permite al algoritmo, si encuentra patrones muy atípicos que no corresponde a patrones ya incluidos en los k grupos ya formados, establecer ese patrón como un nuevo grupo, uniendo en caso de que se hayan generado k+1 grupos los grupos más cercanos, con esta simple regla en mente los resultados obtenidos respecto al algoritmo tradicional de K-medias es significativamente diferente tal como dejan ver en sus ejemplos los autores, logrando obtener grupos de pocos datos los cuales distan mucho del comportamiento del resto de grupos y los cuales se catalogaron posteriormente como inusuales.

En la segunda fase de la propuesta realizada, los autores desarrollan un proceso de encontrar anomalías, esto se realiza partiendo de los centros de los grupos encontrados en el anterior paso, con la distancia entre cada par de los K centros se construye un árbol de expansión mínima, luego se elimina la arista más grande de este árbol generado obteniendo dos nuevos árboles, el árbol que posea un menor número de datos será considerado el árbol inusual y todos los datos que contengan serán marcados como inusualidades, si el árbol marcado como inusual posee más de un vértice el proceso se podría seguir repitiendo hasta que sólo queden el número de vértices deseados.

A pesar de la sencillez del algoritmo planteado, éste presenta muy buenos resultados, los autores en su artículo presentan un conjunto de ejemplos en los cuales se aplicó el algoritmo y en ellos se evidenció que realmente la técnica es capaz de identificar grupos anómalos, por ejemplo plantearon el caso de cómo identificar spam basados en el log de un servidor de correos observando las características inusuales de estos correos al compararlos con los demás correos, realizando el proceso de agrupación y detección de anomalías planteado en su solución.

3.1.5 Esperanza-Maximización

La técnica de Esperanza-Maximización también ha sido utilizada para la detección del lavado de activos. En su artículo Chen et al.[26] muestran cómo a través de un proceso iterativo de Esperanza-Maximización se logra obtener una clasificación de un conjunto de transacciones ya sean diarias, semanales o mensuales en los grupos anómalos o normales, esto por medio de esta técnica que tiene un enfoque no supervisado y que se puede enmarcar entre las técnicas de clasificación.

El algoritmo es un proceso iterativo entre dos pasos, uno en el cual se estima la verosimilitud de que el dato pertenezca a cada uno de los grupos (normal o inusual), esto se realiza por medio de la siguiente ecuación de probabilidad.

$$P(C_j|x) = \frac{|\Sigma_j(t)|^{-\frac{1}{2}} e^j P_j(t)}{\sum_{k=1}^{M} |\Sigma_j(t)|^{-\frac{1}{2}} e^j P_j(t)}$$

Donde C_j indica el grupo final j, P_j la media de cada grupo, M el total de grupos y Σ_j indica la covarianza del grupo final j en el tiempo t.

Luego de estimar esta probabilidad se procede al segundo paso, el cual consiste en maximizar los parámetros de media y covarianza de tal manera que la próxima iteración obtenga mejores resultados, esto se hace por medio de las siguientes ecuaciones.

$$\mu_j(t+1) = \frac{\sum_{K=1}^{N} P(C_j|x_k) x_k}{\sum_{K=1}^{N} P(C_j|x_k)}$$

$$\Sigma_j(t+1) = \frac{\sum_{K=1}^{N} P(C_j|x_k)\,(x_k - \mu_j(t))^2}{\sum_{K=1}^{N} P(C_j|x_k)}$$

$$P_j(t+1) = \frac{1}{N} \sum_{k=1}^{N} P(C_j|x_k)$$

Y se vuelve a realizar la estimación con los nuevos parámetros para así repetir el ciclo, el cual se detendrá sólo cuando se lleve a un nivel de convergencia entre las estimaciones o se alcance el número de iteraciones máximo fijados previamente, que en el artículo se tomó como 200.

Este algoritmo fue ejecutado sobre un conjunto de datos obtenido de un banco local de Malasia, en el cual ya se tenían marcadas las operaciones como normales o inusuales y un conjunto de variables de esas transacciones como son el monto de entrada o de salida y la frecuencia de dichas transacciones. Como el método es no supervisado, al momento de la ejecución no se tenía en cuenta la marcación de inusualidad o no, permitiendo que el algoritmo clasificara las operaciones entre los dos grupos planteados (normales e inusuales) y luego validando con la clasificación que realmente poseían, obteniendo a partir de estos resultados las tasas de falsos positivos y de detección.

Los resultados obtenidos a partir de este algoritmo fueron bastante positivos, comparado con un método de agrupación no supervisado más utilizado como es el k-medias con el cual los autores realizaron también pruebas, determinando al final mayores tasas de detección y menor número de falsos positivos para el método de Esperanza-Maximización.

3.1.6 Sequence Matching

La aplicación del método de Sequence Matching a la detección de lavado de activos fue llevada a cabo por Liu et al.[23], en este artículo el autor detalla las fases en las cuales divide el algoritmo para su aplicación a las transacciones del cliente con fin de detectar las secuencias de operaciones sospechosas que éste realice, este algoritmo posee un enfoque no supervisado pues no requiere conocer previamente casos reales de anomalías y de datos normales para entrenar el modelo.

El objetivo de este algoritmo no es detectar una operación puntual que sea inusual, sino detectar un conjunto de transacciones que vistas como un todo sean denominadas como inusuales, definiendo esa inusualidad como la diferencia con las transacciones que ese mismo cliente ha realizado en la historia o la diferencia que ese cliente posee con otros clientes con sus mismas características.

Las fases en las cuales el autor dividió el proceso de detección fueron las siguientes:

1. Adquisición de datos y conversión: en esta etapa se obtienen las transacciones por grupos de clientes que posean el mismo comportamiento y se definen las características que serán utilizadas para calcular la diferencia entre diversas secuencias (montos de entrada y salida, frecuencias) además del ancho de la ventana que será utilizado posteriormente, con el cual se indica cuántos elementos poseerá cada secuencia.

2. Selección de las secuencias de alto riesgo: como el proceso de comparar cada una de las secuencias posibles en las transacciones a evaluar, contra el histórico de transacciones del cliente y el grupo al que pertenecen es una labor que puede consumir altos tiempos de cómputo, el autor decidió en esta etapa evaluar el riesgo individual de cada una de las transacciones ocurridas y solo sobre las transacciones con más alto riesgo, obtener la secuencia de transacciones que están antes y después de ella, teniendo en cuenta el tamaño de ventana definido en la etapa anterior, para determinar cuáles son las de alto riesgo se definieron unos umbrales sobre cada una de las características también mencionadas en la etapa anterior y las transacciones de alto riesgo serán las que

por lo menos una de sus características supere uno de sus umbrales definidos. En esta etapa, adicionalmente, se define el conjunto de secuencias de referencias que serán aquellas con las que las secuencias de alto riesgo se compararán para determinar la inusualidad o no de las mismas, estas secuencias de referencias son definidas por el autor como las secuencias del histórico de transacciones pertenecientes a la cuenta riesgosa (sin tener en cuenta la secuencia a evaluar) sumado a las secuencias de las cuentas que pertenecen al mismo grupo de la cuenta a evaluar.

3. Cálculo de similaridad: en esta etapa se desea calcular la distancia promedio entre la secuencia obtenida en la etapa anterior y su conjunto de secuencias de referencias. El primer paso para realizar esto es estandarizar tanto la secuencia que contiene la posible inusualidad y su conjunto de secuencias de referencia, esto para eliminar posibles ruidos que existan en las transacciones, luego de esto es realizado el cálculo de similaridad basada en la distancia euclidiana la cual tiene en cuenta la posibilidad de que las dos secuencias a evaluar posean diferentes tamaños, luego de realizada esta operación para cada par de secuencias se obtiene el promedio de esta distancia obteniendo así la medida de similaridad.

4. Clasificación de la secuencia: en esta etapa final lo que se hace es definir un umbral con el cual se comparará la medida de similaridad calculada en la etapa anterior, clasificando las secuencias con menor similaridad respecto a su conjunto de referencias como secuencias inusuales.

Los resultados obtenidos de este algoritmo fueron evaluados teniendo en cuenta los criterios de especificidad y sensibilidad, los cuales son calculados a partir de las tasas de falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos tal como se muestra a continuación.

$$Sensibilidad = \frac{numeroverdaderospositivos}{numeroverdaderospositivos + numerofalsosnegativos}$$

$$especificidad = \frac{numerodeverdaderosnegativos}{numerodeverdaderosnegativos + numerofalsospositivos}$$

Con el objetivo de maximizar estas dos cantidades, los umbrales definidos en la cuarta etapa fueron variados para determinar con cuál umbral se obtenían mejores resultados.

3.2 Métricas para evaluar el desempeño de los modelos

En cuanto a las métricas que se obtienen a partir de cada una de las técnicas implementadas y a partir de las cuales se determina la calidad de estas son muchos los indicadores planteados [27][28][29], las más comunes son las que se pueden obtener a partir de una tabla de contingencia como la siguiente.

		Verdadera clasificación	
		Inusual	Normal
Clasificación planteada	Inusual	VP	FP
	Normal	FN	VN

Donde:

VP: Verdaderos positivos o número de registros correctamente identificados como inusuales.

VN: Verdaderos negativos o número de registros correctamente identificados como normales.

FP: Falsos positivos o número de registros que se clasificaron como inusuales pero realmente son normales.

FN: Falsos negativos o número de registros que se clasificaron como normales pero en realidad son inusuales.

A partir de esta tabla de contingencia son varias las métricas que se obtienen, siendo las más comunes:

Error: Métrica que mide el porcentaje de malas clasificaciones realizadas, se obtiene a partir de los valores obtenidos de la siguiente manera.

$$Error = 1 - \frac{VP + VN}{VP + FP + VN + FP}$$

Tasa de detección: Métrica que mide el porcentaje de operaciones inusuales que se identificaron correctamente, se obtiene a partir de los valores obtenidos de la siguiente manera.

$$Tasa\ de\ deteccion = \frac{VP}{VP + FN}$$

Tasa de falsos positivos: Métrica que mide el porcentaje de que tantas operaciones son reportadas equivocadamente como inusuales, se obtiene a partir de los valores obtenidos de la siguiente manera.

$$Tasa\ falsos\ positivos = \frac{FP}{FP + VN}$$

Precisión: Métrica que mide el porcentaje de entre lo reportado como inusual que tan efectivo fue dicha clasificación, se obtiene a partir de los valores obtenidos de la siguiente manera.

$$Precisión = \frac{VP}{VP + FP}$$

Como cada una de estas métricas busca mostrar cualidades diferentes, la idea que se busca en esta tesis es obtener todas estas métricas y compararlas simultáneamente para conocer de una técnica cuales métricas intenta optimizar, cuáles son sus ventajas y desventajas.

3.3Conclusiones revisión inicial

Con base en la revisión detallada de las principales técnicas utilizadas en la detección de operaciones sospechosas de lavado de activos, se construyó la Tabla 1, en la cual se detalla y se comparan algunas de las características a tener en cuenta del modelo de detección.

TÉCNICA	Aprendizaje supervisado	Aprendizaje no supervisado	Evalúa tasas de detección	Anomalías reales	Utiliza datos no transaccionales
Esperanza-Maximización [17]		✓	✓	✓	
Clúster de dos fases [15]		✓			
Redes Bayesianas dinámicas[14]		✓	✓		
Sequence Matching [13]		✓	✓	✓	
Red neuronal de base radial [10]	✓		✓		
Máquinas de soporte vectorial [11]	✓	✓	✓		

Tabla 1. Comparación inicial modelos de detección.

Cuando se hace referencia a datos no transaccionales se refieren a información propia de quienes intervienen en la transacción y no en la transacción como tal, como la ocupación, ingresos, patrimonio, edad, etc.

A partir de la revisión realizada también se pudieron identificar algunas brechas entre las cuales se encuentran:

- La mayoría de soluciones revisadas, al momento de realizar las pruebas de sus modelos, toman información real para representar los casos normales pero para las anomalías a detectar generan datos sintéticos.

- No se ha realizado una comparación que parta de un mismo conjunto de datos para probar la efectividad de cada uno de estos modelos, por lo que no es posible calificar el mejor modelo para la reducción de falsos positivos.

- Las soluciones propuestas solo están tomando en cuenta información transaccional y se está desaprovechando la información existente sobre quien realiza la

transacción y que las entidades financieras poseen debido a las políticas Know Your Customer.

El esquema de trabajo que se seguirá para abordar la creación de un nuevo modelo de detección de operaciones sospechosas de lavado de activos que logre mejores métricas que los modelos que se revisaron se exponen en la Figura 2.

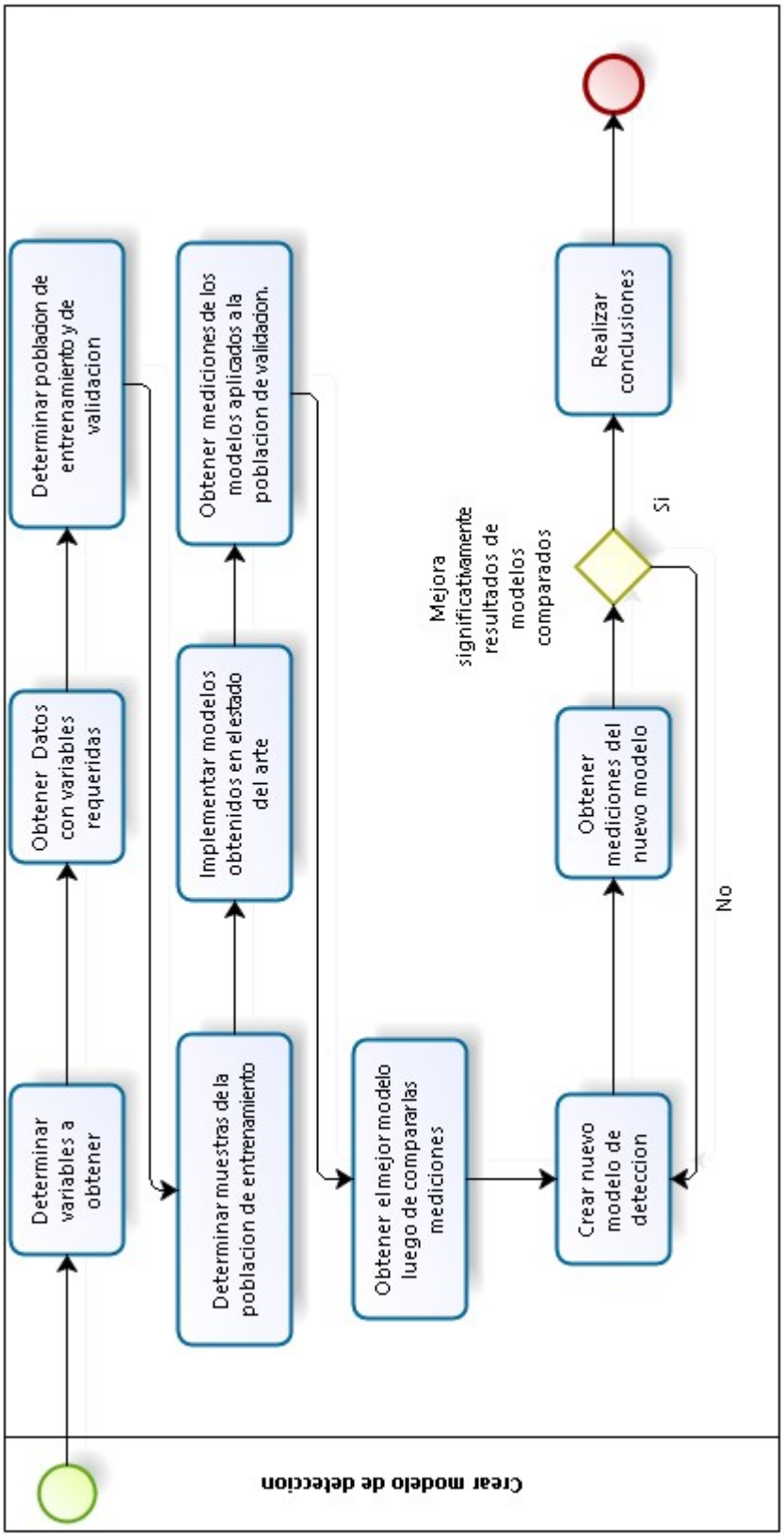

Figura 2. Esquema creación de nuevo modelo de detección

4. DISEÑO DEL EXPERIMENTO

El objetivo de este capítulo es detallar como se realizo el diseño del experimento, es decir, como se eligieron las variables a utilizar, cual es el número total de observaciones con el que se cuenta y como se dividirá estas observaciones en las poblaciones de entrenamiento y de validación. Adicionalmente a la hora de implementar los modelos de detección como se tomara la muestra final de entrenamiento a partir de la gran población de entrenamiento existente.

4.1 Variables a utilizar

Antes de implementar los modelos a comparar y a desarrollar es necesario obtener los datos a partir de los cuales estos modelos serán entrenados y validados. Para esto es necesario primero recopilar cuáles serán las variables mínimas necesarias para llevar a cabo los modelos estudiados de la manera más fiel posible.

Luego de analizar las seis técnicas que se compararan, se obtuvo un listado de cuáles variables se requiere obtener y para cuales modelos se necesitan, esa información puede ser observada en la Tabla 2.

Variable	Tipo	Modelo				
		Red bayesiana dinámica	Red neuronal de base radial	Esperanza maximización	Sequence Matching	Máquina de soporte vectorial
Monto ingresos a la cuenta	Transaccional	X	X	X	X	X
Monto salidas de la cuenta	Transaccional	X	X	X		
Frecuencia ingresos a la cuenta	Transaccional	X	X	X		
Frecuencia salidas a la cuenta	Transaccional	X	X	X		
Tipo de transacción	Transaccional	X			X	
Periodo de la operación	Transaccional	X			X	
Tamaño de la compañía	Cliente				X	
Tiempo entre dos transacciones	Transaccional	X				
Persona Sospechosa	Cliente	X	X	X	X	X

Tabla 2. Variables requeridas en los modelos a comparar.

En la tabla anterior se puede observar que el modelo de detección utilizando un clúster de dos fases no se encuentra, esto se debe porque este es el único que en su respectivo artículo no es utilizado propiamente en la detección de lavado de activos y simplemente se detalla su funcionamiento.

Adicional a las variables anteriormente mencionadas es también necesario determinar qué otras variables es interesante obtener con miras a desarrollar un nuevo modelo, partiendo de la premisa observada en el estado del arte de la subutilización de variables no transaccionales, es decir variables propias del cliente, las cuales permiten tener un mejor conocimiento del cliente y así poder realizar un mejor perfilamiento de que es normal y que no lo es para un determinado tipo de cliente.

Teniendo en cuenta lo anterior un nuevo listado de variables adicionales fueron propuestas y se pueden observar en la Tabla 3.

VARIABLE	TIPO
Edad	Cliente
Ingresos	Cliente
Egresos	Cliente
Patrimonios	Cliente
Ventas	Cliente
Segmento del cliente	Cliente
Tipo de persona	Cliente
Ocupación	Cliente
Actividad	Cliente
Monto giros enviados	Transaccional
Monto giros recibidos	Transaccional
Monto créditos desembolsados	Transaccional

Tabla 3. Variables adicionales a obtener.

Como se puede observar la gran mayoría de estas variables son propias del cliente y no de la transacción, con lo cual se espera obtener a través de un mayor conocimiento de quien realiza la operación una mejoría en las tasas de falsos positivos y negativos.

4.2 Obtención de los datos

Luego de tener nuestro listado definitivo de variables necesarias para implementar los modelos estudiados en el estado del arte y las que se consideran pueden ser utilizadas en el desarrollo del nuevo modelo, se considera que para tener mejores resultados es necesario utilizar datos reales y no datos generados sintéticamente, por lo cual la información de estas variables serán construidas a partir de la información suministrada por una entidad financiera la cual será anonimizada para no incurrir en problemas legales.

Es importante tener también en cuenta que las técnicas y las variables anteriormente listadas implican un grado de agregación distinto, lo cual también debe afectar el conjunto de datos que se obtendrá, existiendo en el conjunto de técnicas dos de ellas que toman los datos transaccionales uno por uno pues son modelos que se basan en la secuencia en la cual se realizan las operaciones, estas técnicas son la de sequence matching y la de redes bayesianas dinámicas, esto implica que el conjunto de datos que se utilice para el entrenamiento y validación sean de transacciones desagregadas, donde se detallan fechas, tipos de operación y medios de pago de la operación.

Por otro lado para las demás técnicas es conveniente tener un mayor grado de agregación, pues con una vista más agrupada de los datos la información obtenida resulta de mayor utilidad, las cuatro técnicas restantes utilizan esta aproximación, la técnica propia que se desarrollará tendrá el enfoque que tenga la mejor técnica luego de la comparación.

Por todo lo anterior, dos conjuntos de datos fueron obtenidos a partir de la información suministrada por la entidad financiera local, una información agregada que contiene 7'524.989 registros donde cada uno de ellos detalla la información transaccional agregada en un año de los clientes junto a la información propia del cliente la cual no se requiere agregar.

El segundo conjunto de datos contiene 156'483.844 registros y corresponde a la información desagregada en la cual cada registro es una transacción realizada y en la cual se indica información como fecha, cuantía, tipo de operación y medio de pago utilizado, además de la información propia del cliente.

La estructura final del archivo consolidado puede ser observada en la Tabla 4, donde se encuentra el tipo de dato, una breve explicación y los valores esperados en esa variable.

VARIABLE	**TIPO**	**EXPLICACIÓN**	**VALORES ESPERADOS**
Id	Numérico	Identificador de la persona. En vista que los datos están anonimizados este id corresponde a un consecutivo	Número entre 1 y 7'524.989
Tipo persona	Alfanumérico	Indica si la persona es natural o jurídica	N (para personas naturales) J (Para personas jurídicas)
Edad	Numérico	Edad para las personas naturales, tiempo de constitución para las personas jurídicas	Número entero no negativo
Ocupación	Alfanumérico	Ocupación de la persona natural	Una de las siguientes: Ama de Casa, Ganadero, Agricultor, Profesional Independiente, Jubilado, Desempleado con Ingresos, Empleado, Comerciante, Independiente, Estudiante y Otro
Actividad	Numérico	Actividad laboral según el CIIU	Código CIIU de la actividad: 10, 1399, 145
Ingresos	Numérico	Ingresos mensuales de la persona	Número no negativo
Ventas	Numérico	Ventas anuales de la persona	Número no negativo
Egresos	Numérico	Egresos mensuales de la persona	Número no negativo
Patrimonio	Numérico	Patrimonio de la persona	Número no negativo
Código segmento	Numérico	Segmento asignado por la entidad a la persona	Número que identifica el segmento de la persona.
Monto entrada	Numérico	Monto total que entra a las cuentas de la persona en el año observado	Número no negativo
Monto salida	Numérico	Monto total que sale de las cuentas de la persona en el año observado	Número no negativo
Frecuencia entrada	Numérico	Número total de operaciones de entrada a las cuentas de la persona en	Número no negativo

		el año observado	
Frecuencia salida	Numérico	Número total de operaciones de salida de las cuentas de la persona en el año observado	Número no negativo
Monto giros enviados	Numérico	Monto total de dinero enviado por la persona por medio de giros en el año observado	Número no negativo
Monto giros recibidos	Numérico	Monto total de dinero recibido por la persona por medio de giros en el año observado	Número no negativo
Monto créditos desembolsados	Numérico	Monto total dinero que se desembolsó en las cuentas de la persona en el año observado	Número no negativo
Sospechoso	Numérico	Variable que indica si la persona según la entidad ha sido catalogada como sospechosa de lavar activos	0 si no es sospechoso 1 si es sospechoso

Tabla 4. Variables conjunto de datos agregados.

Por otro lado la estructura del conjunto de datos desagregados puede observarse en la Tabla 5, al igual que la anterior tabla en esta también se detalla el tipo de variable, una pequeña descripción y los valores esperados para cada variable.

VARIABLE	TIPO	EXPLICACIÓN	VALORES ESPERADOS
Id	Numérico	Identificador de la persona, en vista que los datos están anonimizados este id corresponde a un consecutivo	Número entre 1 y 7'524.989
Periodo	Alfanumérico	Indica en que parte del mes la operación fue realizada	Inicio (primeros 10 días) Medio (del 10 al 20 del mes) Fin (del 21 al final del mes)
Medio	Alfanumérico	Detalla el medio de pago de la operación	Puede tomar los valores, efectivo, electrónico y cheque.
Débito/Crédito	Alfanumérico	Indica si la operación fue un débito (Salida de dinero) o un crédito (Entrada de dinero)	D para débito C para crédito
Monto	Numérico	Cuantía de la operación	Número no negativo
Fecha	Numérico	Fecha de la operación	Fecha de la operación en formato aaaammdd
Días diferencia	Numérico	Número de días desde la operación	Número no negativo

		anterior y esta	
Ingresos	Numérico	Ingresos mensuales de la persona	Número no negativo
Ventas	Numérico	Ventas anuales de la persona	Número no negativo
Egresos	Numérico	Egresos mensuales de la persona	Número no negativo
Patrimonio	Numérico	Patrimonio de la persona	Número no negativo
Tipo persona	Alfanumérico	Indica si la persona es natural o jurídica	N (para personas naturales) J (Para personas jurídicas)
Actividad	Numérico	Actividad laboral según el CIIU	Código CIIU de la actividad: 10, 1399, 145
Edad	Numérico	Edad para las personas naturales, tiempo de constitución para las personas jurídicas	Número entero no negativo
Ocupación	Alfanumérico	Ocupación de la persona natural	Una de las siguientes: Ama de Casa, Ganadero, Agricultor, Profesional Independiente, Jubilado, Desempleado con Ingresos, Empleado, Comerciante, Independiente, Estudiante y Otro
Código segmento	Numérico	Segmento asignado por la entidad a la persona	Número que identifica el segmento de la persona.
Sospechoso	Numérico	Variable que indica si la persona según la entidad ha sido catalogada como sospechosa de lavar activos	0 si no es sospechoso 1 si es sospechoso

Tabla 5. Variables conjunto de datos desagregados.

4.3 Obtención de poblaciones de entrenamiento y de validación

Una vez obtenidos los dos conjuntos de datos, es ahora necesario determinar cuál será la manera en la cual se utilizarán estos datos para realizar el entrenamiento de cada uno de los modelos y luego realizar la comparación, todo esto para poder garantizar que los modelos son entrenados con un conjunto significativo de datos y que las comparaciones de estos se hace sobre un conjunto significativo de datos lo cual ayuda a aumentar la validez de las conclusiones.

La metodología que se empleara para el manejo de datos, tanto para el conjunto de datos desagregado como para el conjunto de datos consolidado, será tomar el conjunto de datos respectivo como un universo el cual se dividirá en dos conjuntos excluyentes entre ellos, el conjunto de entrenamiento y el conjunto de validación, de estos dos conjuntos el primero será utilizado como otro universo de datos del cual se obtendrán las muestras para cada uno de los modelos a implementar, las cuales su tamaño dependerá de las características propias de cada modelo. El segundo conjunto, el de validación, serán los datos sobre los cuales se probarán los seis modelos a comparar y posteriormente también se probará el modelo generado, de tal manera que las métricas obtenidas como falsos positivos y falsos negativos sean tomados del mismo conjunto y las conclusiones que se puedan dar sobre su efectividad pueden tener un grado mayor de credibilidad.

Algo que se debe tener en cuenta con los conjuntos de información obtenidos es que dado que la variable respuesta que se tiene y se desea predecir en cada uno de los modelos, la cual indica si la persona realiza operaciones sospechosas o no, es una variable que en la gran mayoría de registros tomará el valor de 0 o no sospechoso. Esto es importante al momento de generar las poblaciones de entrenamiento y validación, debido a que el conjunto de datos será de naturaleza desbalanceada y puede producir un comportamiento no deseado en los modelos a realizar, porque el modelo puede asumir que al haber tan pocas ocurrencias de los eventos sospechosos puede marcar todos los eventos como no sospechosos y aun así obtener buenas métricas.

Lo otro que se debe tener en cuenta es que los conjuntos de datos que se están trabajando son significativamente grandes tanto para los datos desagregados (156 millones de registros) y para los datos consolidados (7 millones de registros) por lo que

el porcentaje de datos que se escogerá para realizar el entrenamiento y la validación no será tan influyente en los resultados, es decir, se podrían esperar resultados bastante similares si se escoge el 40% de los datos para validación y el 60% para entrenamiento, o si se escogen 40% de los datos para entrenamiento y el 60% para validación.

Si bien conocedores del tema han dado algunas recomendaciones de cómo repartir las poblaciones entre población de entrenamiento y poblaciones de validación [30],en nuestro caso dado el gran número de transacciones que se tiene permite mayor libertad en la escogencia de cómo distribuir el universo de datos en estas dos poblaciones. Pero teniendo en cuenta que se quiere aprovechar los recursos y probar los resultados en un conjunto significativo de datos, se escoge el acercamiento de 40% de datos para entrenamiento y 60% para la validación.

Haciendo referencia nuevamente al problema del desbalanceo entre operaciones sospechosas y no sospechosas, esto afectará la manera en la cual se hará la división entre población de entrenamiento y población de validación[31], por tanto se debe realizar una partición aleatoria pero que respete dicha proporción, es decir si el 1% de las operaciones son sospechosas y el 99% son no sospechosas en el conjunto de datos original, esta proporción deberá ser mantenida en el 40% de la población que se tome para entrenamiento y en el 60% de la población que se deje para validación.

Finalmente el problema del desbalanceo de la población también debe ser tratado al realizar la muestra, pues como se muestra en varios estudios[32][33], esto tiene gran incidencia en el comportamiento de los modelos y los resultados que estos ofrecen, por este motivo en los modelos que requiera tomar muestras del universo de entrenamiento la estrategia que se tomará será realizar un balanceo por debajo con la muestra[34], es decir, se toma el total de los datos de operaciones sospechosas en la muestra y se toma al azar un número igual de operaciones no sospechosas del resto del universo de datos de entrenamiento, así se obtendrá una muestra balanceada donde el 50% será sospechoso y el otro 50% no lo será.

Con todo lo abordado en este capitulo ya nos encontramos en la capacidad de empezar el entrenamiento y validación de cada uno de los modelos de detección a estudiar, esto

debido a que ya se posee el conjunto de datos con las variables necesarias para probar cada uno ellos.

5. COMPARACIÓN ENTRE MODELOS

Antes de poder construir un nuevo modelo de detección de lavado de activos, es importante realizar una comparación de los modelos de detección contemplados en el estado del arte, esto con dos objetivos, primero establecer cuál es el modelo referente en ámbitos de detección de inusualidades con miras a determinar operaciones sospechosas de lavado de activos, por otro lado aprender las bondades de cada uno de los modelos estudiados con el fin de poder usar dichas bondades en el nuevo modelo creado. El objetivo de este capítulo es entonces realizar la implementación y comparación de los métodos que se obtuvieron en el estado del arte y determinar según las métricas también allí expuestas cuál de ellos posee mejores cualidades.

Como se detallo en la sección anterior, para que la comparación tenga mayor fundamento el conjunto de validación será el mismo para cada uno de los modelos, al igual que las métricas utilizadas para calificar los modelos.

5.1 Redes bayesianas dinámicas.

La implementación de esta técnica está separada en tres etapas, una primera etapa es la de segmentación en la cual cada uno de los individuos es asignado a un grupo con el cual comparte características comunes y con el cual se pueden realizar más fácilmente comparaciones. La segunda etapa corresponde a la de identificación de la Red Bayesiana Dinámica de cada uno de los grupos en los que se segmento la población. La tercera y última etapa es la de identificación de anomalías, la cual se hace con base en la Red Bayesiana identificada en la etapa anterior y adicionalmente con una métrica especial creada por los autores de esta técnica de detección, denominada AIRE (Índice de anomalía utilizando rango y entropía).

Para la primera parte de la técnica, dado que el conjunto de datos entregados ya poseía una segmentación dada por la propia entidad, la cual es quien conoce a sus clientes y

puede realizar de mejor manera dicha clasificación, será la segmentación que se utilizara en esta técnica.

Para la segunda parte se requiere para cada uno de los K segmentos en que se divide la población obtener la red bayesiana dinámica que la representa, teniendo en cuenta tres variables en tres periodos de tiempos diferentes y adyacentes, tal como lo muestra la Figura 3.

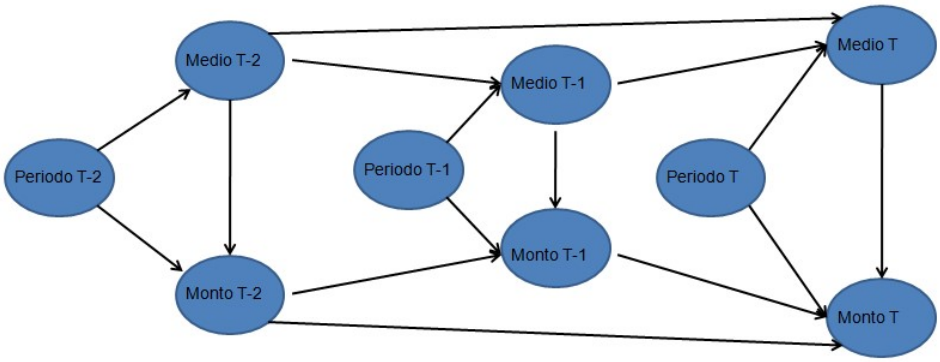

Figura 3.Red bayesiana dinámica para detección de lavado de activos

Lo que esta red bayesiana nos indica es que dado las dos últimas transacciones (T-2 y T-1) en términos de periodo donde ocurrió en el mes (Inicio, Mitad o Fin), medio de pago por el cual se llevó a cabo la transacción (Efectivo, Cheque o Electrónico) y el monto, es posible decir la probabilidad que la transacción que se está realizando actualmente (en el tiempo T). También indica la relación entre las variables anteriormente mencionadas, por ejemplo el medio de pago del periodo T-2 tiene influencia sobre el medio de pago del periodo T-1 y ambos tienen influencia sobre el medio de pago del periodo actual T.

Lo que se obtiene al final de esta red bayesiana siempre será una probabilidad, dado el comportamiento en las dos últimas transacciones cual era la probabilidad de realizar los movimientos obtenidos en este periodo actual, es decir, se obtiene una calificación de que tan usual o inusual fue el realizar los movimientos del periodo actual dado el comportamiento inmediatamente anterior, teniendo como base para esta probabilidad el comportamiento de todo el segmento al cual pertenece la persona.

Antes de comenzar es muy importante realizar una transformación de datos, de tal manera que se puedan categorizar cada una de las variables que se requiere en el algoritmo, el periodo, el medio de pago y el monto. Para los dos primeros la categorización se especificó en el propio artículo, Inicio, Mitad y Fin para el periodo, Efectivo, Cheque y Electrónico para el medio de pago, pero para el monto también se requiere una categorización.

El método elegido fue clasificar los montos en cada grupo según umbrales que se obtengan de cada población, para poder categorizar cada monto de transacción como Bajos, Medios o Altos. Los umbrales elegidos fueron la mediana y el percentil 90, si el monto era inferior a la mediana se clasifica como bajo, si era superior a la mediana pero inferior al percentil 90 se clasificaba como medio y si supera el percentil 90 se clasifica como alto.

Finalmente para terminar la etapa de preprocesamiento, es importante asegurar que en cada registro se encuentre la operación a estudiar si fue inusual o no, con su periodo, medio de pago y monto ya categorizados, adicionalmente debe tener estas mismas tres variables de las dos operaciones que le anteceden.

Es decir, si una persona realizó 5 operaciones: operación 1, operación 2, operación 3, operación 4, operación 5, cada una de estas con un monto, período y medio de pago, la persona en cuestión deberá poseer tres registros en la base de datos de la manera como lo indica la Tabla 6.

Operación T-2	**Operación T-1**	**Operación T**
Operación 1	Operación 2	Operación 3
Operación 2	Operación 3	Operación 4
Operación 3	Operación 4	Operación 5

Tabla 6. Ejemplo de representación en base de datos de operaciones consecutivas

Dada la información de la anterior tabla y de la red bayesiana dinámica del segmento al cual pertenezca la persona se obtendrá una probabilidad de cada una de las operaciones del periodo T, dada la información de las operaciones predecesoras T-2 y T-1.

Para la tercera y última parte de la técnica, se calcula el indicador AIRE de cada una de las transacciones, el cual depende del rango y la entropía de la operación a calificar, y las cuales son determinadas según la probabilidad de ocurrencia de los eventos según la red bayesiana. Para cada individuo se obtiene la transacción con mayor AIRE y es ese indicador el que se compara contra un umbral obtenido a partir de la calificación de toda la población, para establecer cuáles son los individuos que poseen un AIRE más inusual y por tanto serán marcados como personas con operaciones sospechosas de lavado de activos.

El total de muestras utilizadas para realizar el entrenamiento del modelo y obtener la red bayesiana dinámica de cada segmento y el umbral sobre el cual se separa individuos normales de sospechosos, será el total de datos disponibles en el entrenamiento, esto se hace de esta manera para poder construir una red bayesiana dinámica más completa. Esta red tiene muchas combinaciones de probabilidades condicionales, dado que tiene varias variables, con varios estados en varios periodos de tiempos, las cuales afectan la probabilidad de ocurrencia de las demás.

Una vez obtenidas las redes bayesianas de cada segmento y el umbral sobre el cual se determinarán las inusualidades, se procede a ejecutar este modelo sobre el conjunto de validación. Los resultados obtenidos se encuentran en la Tabla 7.

Modelo	Red Bayesiana Dinámica
Número de muestras	2.957.028
Error	0,89%
Verdaderos positivos	197
Verdaderos negativos	2.930.627
Falsos positivos	19.034
Falsos negativos	7.170
Tasa de detección	2,67%
Tasa de falsos positivos	0,65%
Precisión	1,02%

Tabla 7. Resultados red bayesiana dinámica

Las conclusiones que se obtuvieron al realizar la implementación de esta técnica fueron las siguientes:

- Si bien las tasas de error y de falsos positivos de la técnica son muy bajas, la precisión y la tasa de detección también lo son, lo cual resta aplicabilidad a la técnica en un ambiente real.

- La fortaleza de esta técnica como se muestra en la tabla resumen, radica más en la identificación de registros que efectivamente no son sospechosos, por lo cual sigue siendo una técnica viable si se planea utilizar como un primer paso de una técnica más compleja y cuya función sea reducir el tamaño de la población a identificar como sospechosa, filtrando a una gran cantidad de personas que muy probablemente serán clasificadas como no sospechosas.

- Una consideración especial que se debe tener en cuenta al realizar esta técnica es que opera sobre datos desagregados, es decir, el insumo principal es la operación puntual que ha realizado una persona, además de las dos últimas transacciones realizadas por el individuo, por tanto los individuos que hayan realizado menos de 3 operaciones en el periodo de estudio no podrían ser analizados con esta técnica.

5.2 Redes neuronales de base radial

Esta técnica busca por medio de una red neuronal de base radial, obtener las probabilidades de que una operación sea inusual o normal, y la opción que más probabilidad posea será la clasificación que se dará a dicha operación.

Para lograr la implementación de esta técnica primero se debe seleccionar la muestra con la que se entrenará la red neuronal, teniendo en cuenta que la red conforma dinámicamente el número de grupos y centros que se utilizaran en la capa oculta, se recomienda que esta muestra de entrenamiento no sea demasiado numerosa, adicionalmente para que el comportamiento inusual como el normal queden bien representado también se debería optar por una muestra balanceada. Por todo lo anterior, del gran subconjunto de muestra de entrenamiento se escogen 10.000 individuos aleatoriamente para realizar el entrenamiento de la red neuronal de base radial, en el cual 5.000 individuos fueron clasificados como inusuales y 5000 como normales.

Luego de haber escogido el conjunto de individuos con el cual se llevará a cabo el entrenamiento de la red, se requiere que se realice una normalización de los datos, como se mencionó en el estado del arte, para cada variable a utilizar se utilizarán las siguientes formulas para llevar a cabo la normalización.

$$\varepsilon_i = max_i(X_{ij}) - min_i(X_{ij})$$

$$X'_{ij} = \frac{X_{ij} - min_i(X_{ij})}{\varepsilon_i}$$

De esta manera cada una de las variables se moverá entre el rango [0 - 1].

Tal como proponen los autores del modelo las variables que se utilizaran para realizar la detección de inusualidad serán, el número de operaciones de entrada y de salida en las cuentas del cliente en el periodo a estudiar, adicionalmente el monto total que sumaron esas operaciones de entradas y salidas.

Ya con estas definiciones previas, se procede a la construcción de la red neuronal de base radial, el modelo propuesto por los autores es el que se muestra en la Figura 4.

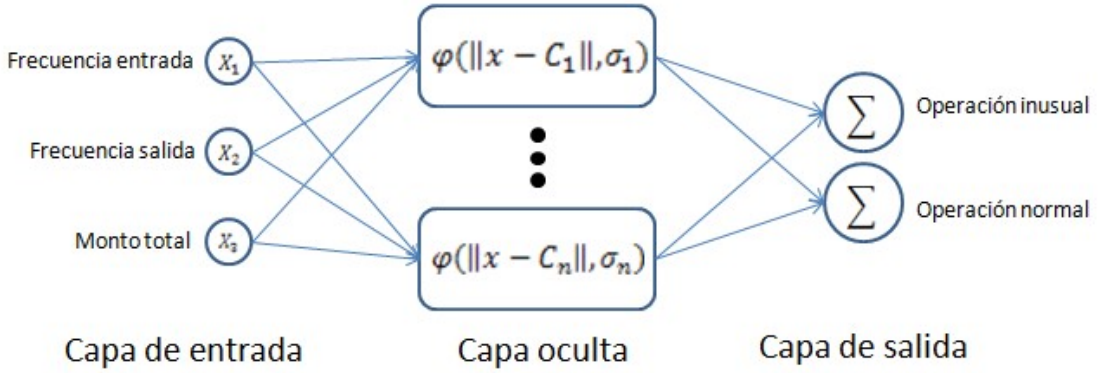

Figura 4. Red neuronal de base radial para identificar anomalías

La anterior figura lo que nos indica es que dada las tres variables de entrada normalizadas que se definieron, se mapean en las n capas ocultas que se conformarán posteriormente por medio de una función gaussiana y finalmente todas estas capas ocultas serán ponderadas por medio de una función lineal, al final se obtendrán dos variables que contendrán la probabilidad de que el individuo estudiado sea inusual o normal.

Para poder ejecutar el paso de la capa de entrada a la capa de salida pasando por las capas ocultas, es requerido conocer múltiples parámetros que serán obtenidos a partir de la muestra de entrenamiento, los parámetros requeridos son, número de capas ocultas, los centros de cada una de las capas ocultas, el ancho gaussiano y los pesos con lo cual se mapeara a la capa de salida.

Para obtener el número de capas ocultas a desarrollar y los centros de estas capas ocultas, se utiliza un algoritmo denominado APC-III Clúster, la idea detrás de este algoritmo es primero establecer un parámetro de distancia promedio entre los datos, posteriormente partir de un centro aleatorio y recorrer los datos, si la distancia de estos al clúster más cercano es superior al parámetro de distancia promedio establecido anteriormente, ese dato pasa a convertirse en un el centro de un nuevo clúster que solo contendrá a este dato, si por otra parte el dato pasa a integrarse a un clúster existente, se re calcula el nuevo centro de dicho clúster. Al final del algoritmo se tendrá un número

de grupos que depende completamente de las características de los datos con que se entrenó y también poseerá el centro de cada uno de estos grupos.

El parámetro de ancho gaussiano es calculado a partir de los grupos creados anteriormente, al calcular cual es la distancia máxima entre cualquier individuo y su clúster más lejano y dividir esta distancia por una constante que depende del número de grupos conformados.

Para obtener los últimos parámetros que son los pesos con los cuales se ponderan cada una de las salidas obtenidas en las capas ocultas y a partir de las cuales se obtendrán la probabilidad final que determinará la clasificación del individuo, se requiere un algoritmo con el cual se pueda resolver la optimización lineal que implica la escogencia de estos pesos, los autores al ver que el método del gradiente si bien es una posible solución es muy lento, proponen resolver este problema por medio del algoritmo RLS, el cual es un método iterativo a partir del cual se busca llegar a una tasa de error inferior a una tasa de tolerancia definida previamente.

Desarrolladas todas las técnicas con las cuales se obtendrán todos los parámetros requeridos para el entrenamiento y la posterior utilización del algoritmo en la validación de individuos, se aplica sobre la muestra de entrenamiento definida, obteniendo el número de capas ocultas, los centros de estas capas, el ancho gaussiano y los pesos que se utilizan para ponderar las capas hacia las dos salidas definidas.

Finalmente con la red neuronal de base radial ya entrenada, se realiza la clasificación entre individuos inusuales y normales de la población de validación y se obtiene los resultados que se muestran en la Tabla 8.

Modelo	Red neuronal de base radial
Número de muestras	4.514.994
Error	52,47%
Verdaderos positivos	18.117
Verdaderos negativos	2.127.981
Falsos positivos	2.363.117
Falsos negativos	5.779
Tasa de detección	75,82%
Tasa de falsos positivos	52,62%
Precisión	0,76%

Tabla 8. Resultados red neuronal de base radial

Las conclusiones que se obtuvieron al realizar la implementación de esta técnica fueron las siguientes:

- El gran esfuerzo de esta técnica radica en el entrenamiento de las capas ocultas, la obtención del parámetro del ancho gaussiano y en los pesos requeridos para mapear la capa oculta a la capa de salida, pero una vez hecho ese esfuerzo con la población de entrenamiento, aplicar la técnica a datos nuevos es mucho más sencillos, pues basta para cada dato de entrada, realizar el mapeo con la parametrización existente a la capa oculta y obtener la parametrización con la agrupación realizada en la capa de salida.

- Si bien la tasa de detección es muy alta, superior al 75%, el error y la tasa de falsos positivos también lo es, superando más del 50%, adicionalmente la precisión final obtenida también es demasiado baja, inferior al 1%. Por todo esto la aplicabilidad de la técnica pierde muchos puntos en un ambiente real, esto debido a que la técnica está dando como inusual a aproximadamente el 50% de la población, en la cual realmente sólo logra acertar menos al 1%.

5.3 Máquinas de soporte vectorial

Para la implementación de esta técnica se utilizó el mismo software utilizado por los autores del modelo llamado libsvm, este software recibe un conjunto de parámetros con el cual se determina el tipo de técnica svm a implementar y los argumentos que estos necesitan.

El primer paso es determinar el conjunto de datos que será utilizado para entrenar el modelo, dado la complejidad de la técnica se sugiere que el conjunto de entrenamiento no sea muy alto, adicionalmente para mantener la muestra balanceada y que el modelo pueda aprender correctamente el comportamiento de los datos atípicos y normales, se eligen al azar una muestra de 5.000 individuos normales y 5.000 individuos anómalos.

Luego de determinado la población con la cual se entrenará el modelo, se elige las variables que se le entregarán al modelo para que determine el hiperplano que se utilizará para clasificar las nuevas muestras como inusuales, tal como lo sugiere el artículo las tres variables a utilizar serán el número de operaciones que ingresan a la cuenta de la persona en el periodo determinado, el número de operaciones que salen de la cuenta y por último el monto total que suman dichas operaciones de entrada y salida.

Tal como se realizó en el artículo se entrenan dos modelos diferentes, la máquina de soporte vectorial en su variante supervisada, es decir, conociendo de antemano los resultados a predecir y entregándoselos al modelo para que a partir de este dato realice el entrenamiento del modelo y genere el hiperplano correspondiente, el algoritmo que utilizaron los autores para implementar este modelo fue el C-SVM. Por otro lado los autores también utilizaron la versión no supervisada del algoritmo, es decir, sin pasarle los datos que indican si el individuo es inusual o sospechoso, siendo el modelo el encargado de determinar a partir de las características de la muestra de entrenamiento cuáles son los individuos "extraños" y clasificándolas por aparte, el algoritmo que se utilizó para implementar este modelo fue el SVM de una clase.

Antes de pasarle los datos al modelo, tal como lo indica el artículo también se requiere realizar una transformación de los datos, aunque en el artículo se contempla una manera donde se calculan distancias para datos heterogéneos, es decir, que contengan tanto datos cuantitativos y cualitativas denominada distancia HVDM, dado que el conjunto de variables a utilizar solo es cuantitativa la transformación requerida será la misma para todas las variables e implica dividir cada dato por cuatro veces su desviación estándar.

Finalmente, antes de poder empezar a entrenar los dos modelos a realizar, se requiere entregar al modelo los parámetros que necesita, el primero indica un factor de control y es requerido por los dos modelos y el segundo indica el factor de castigo por mala clasificación y solo es utilizado por el C-SVM. Los parámetros con los cuales se entrenara los dos modelos son las mismas tres combinaciones sugeridas por los autores del modelo.

Por todo lo anterior, sobre la muestra de entrenamiento escogida y transformada se realizaron un total de seis modelos diferentes, tres sobre el modelo C-SVM y tres sobre el modelo de SVM de una clase, en los cuales cada uno se hacía sobre cada una de las combinaciones propuestas por los autores.

Una vez entrenados los seis modelos se obtuvieron las métricas para determinar con cuál de ellos la muestra de entrenamiento se obtenían los mejores resultados, obteniendo que la mejor combinación se daba cuando el factor de control es 0.5 y el castigo por mala clasificación es 50 para el modelo C-SVM, por tanto, este es el modelo que se utilizara para realizar la validación.

Una vez aplicado el modelo entrenado que obtuvo mejores resultados sobre la población de validación se obtuvieron los resultados que se muestran en la Tabla 9.

Modelo	Máquina de soporte vectorial
Número de muestras	4514994
Error	2.61%
Verdaderos positivos	13860
Verdaderos negativos	4383436
Falsos positivos	107662
Falsos negativos	10036
Tasa de detección	58.00%
Tasa de falsos positivos	2.40%
Precisión	11.41%

Tabla 9. Resultados maquina de soporte vectorial

Las conclusiones que se obtuvieron al realizar la implementación de esta técnica fueron las siguientes:

- En términos generales esta técnica es muy interesante, el porcentaje de error es bastante bajo, inferior al 3%, una tasa de detección significativamente alta, 58% y una tasa de falsos positivos baja, inferior al 3% también. Finalmente con una precisión del 11.41% es una técnica que podría aplicarse en un ambiente real, pero tiene el inconveniente que el número de registros que clasifica como inusual aún es un poco grande para la capacidad que puede tener una entidad financiera para atender periódicamente.

- Una de las principales características que se pueden observar de esta técnica al observar las métricas finales, es que es una técnica muy buena determinando cuáles individuos son normales, por lo cual es un gran candidato a ser utilizada como el primer paso en una técnica más compleja, y actuar como un filtro el cual permita dado un gran conjunto de datos, elimine un gran número de estos datos los cuales con muy alta probabilidad serán normales, quitando todo este ruido de los datos una técnica posterior podría tomar esta información como insumo y mejorar las pocas falencias que esta técnica presenta.

5.4 Clúster de dos fases

Esta es una técnica que opera en dos fases, primero genera grupos con los individuos según sus características, posteriormente selecciona los grupos anómalos por medio de un árbol de expansión mínima. Pero para que la técnica pueda ser utilizada de una mejor manera, requiere que los individuos sean agrupados previamente, esto debido a que si se evalúa la población entera en el algoritmo dará como resultado que el algoritmo separará poblaciones diferentes mas no necesariamente sospechosas, por ejemplo, separando personas naturales de empresas. Para evitar este escenario se recomienda que los individuos a clasificar como inusuales o normales sean comparables, para esto se utilizó la segmentación dada por la entidad financiera y por cada uno de los K segmentos los cuales son a priori comparables, se realizó la técnica. Al final el listado completo de personas clasificadas como inusuales sería la agrupación de los inusuales detectados en cada uno de los K grupos.

Como este fue un modelo que no fue propiamente aplicado al lavado de activos, las variables que se utilizaran en el algoritmo son las que se utilizan mayormente en los modelos de detección de inusualidades de lavado de activos, el número de operaciones de entrada a las cuentas del individuo, el número de operaciones de salida de las cuentas, el monto de entrada y el monto de salida, todo esto en el periodo estudiado.

Antes de aplicar el algoritmo y teniendo en cuenta que la conformación de grupos depende completamente de las distancia entre los datos, es altamente recomendable que las variables a utilizar estén normalizadas, porque de lo contrario como las variables tienen escalas diferentes tendrían pesos diferentes y la evaluación de distancias quedaría distorsionada, por eso cada variable es estandarizada al restarle la media y dividirla por su desviación estándar.

Una vez definidas cuáles son las variables de entrada y que estas hayan sido estandarizadas, se procede a realizar el entrenamiento para cada uno de los K grupos que se tienen. Para este algoritmo no se requiere obtener una sub-muestra de entrenamiento sino que se utilizara toda la población de entrenamiento, de manera que

se pueda establecer los parámetros del algoritmo de manera más precisa antes de aplicarlo a la población de validación.

Como se dijo anteriormente el algoritmo posee dos fases, de ahí su nombre, en la primera se requiere clasificar los individuos en grupos, pero esta clasificación es hecha por medio de un algoritmo especial, el cual no busca que los datos queden igualmente distribuidos sino que si hay datos que por sus características son muy diferentes a los demás, así lo refleje el grupo al que sea asignado. Para lograr esto los autores realizaron una modificación al muy conocido algoritmo de clasificación de K-medias, en el cual si bien se sigue partiendo de un número de grupos K el cual siempre se mantendrá, se agrega un heurístico que busca permitir que se conformen grupos de poco tamaño y alejados de los demás grupos, que posteriormente serán identificados como anómalos.

El algoritmo de K-medias modificado que se aplica es el siguiente, partir de K elementos aleatoriamente escogidos de todo el conjunto de datos a clasificar, estos serán los K centros de los K grupos iniciales, posteriormente uno a uno los datos restantes serán evaluados y serán asignados a los K grupos disponibles, evaluando la distancia del individuo a cada uno de los centros y asignándole al grupo cuyo centro sea el más cercano re calculando el nuevo centro cuando un elemento ingresa a un grupo, pero antes de eso también se evalúa la mínima distancia que existe entre cada par de centro y se compara si la distancia más pequeña entre el individuo evaluado y el centro más cercano es más grande que la distancia más pequeña entre cualquier par de centros, el elemento evaluado se vuelve el solo un nuevo grupo, pero como siempre se debe mantener los K grupos, entonces los dos grupos más cercanos se fusionan en uno solo y re calculando su nuevo centro. Esos pasos se siguen hasta que ya no quede ningún individuo por evaluar, de manera que al final se tienen a todos los individuos de la población clasificado en uno de los K grupos, donde se pueden encontrar grupos muy pequeños y muy alejados de los demás, pero como esto no se puede realizar por simple observación, ahí es donde entra la segunda fase del algoritmo.

En la segunda fase del algoritmo, se toman los K centros formados en el algoritmo anterior y a partir de esos centros se crea un árbol de expansión mínima, por lo cual ahora se tiene un árbol donde los vértices son cada uno de los K centros y las aristas

poseen pesos que son las distancias entre cada par de vértices, la siguiente que se realiza es eliminar de este árbol la arista con mayor distancia, dejando así dos componentes conexas en lo que antes era un árbol, ahora se realiza un recuento de individuos en cada una de las componentes conexas, sumando los elementos que posee cada uno de los grupos cuyos centros quedaron en cada componente. Al final los individuos que se marcan como inusuales serán todos aquellos que pertenezcan a los grupos de la componente conexa con menores individuos.

Como se puede observar en el algoritmo el único parámetro que se requiere es el K número de grupos que se conformará, para obtener este parámetro se realizó con la población de entrenamiento la clasificación y evaluación de métricas iterando el parámetro K y se encontró que el mejor resultado para una población con las características que se tenía era un K de seis.

Aplicando el mismo algoritmo a la población de validación con el parámetro de K igual a seis, se obtuvieron las métricas que se muestran en la Tabla 10.

Modelo	Clúster de dos fases
Número de muestras	4514994
Error	0.530%
Verdaderos positivos	7
Verdaderos negativos	4491045
Falsos positivos	53
Falsos negativos	23889
Tasa de detección	0.029%
Tasa de falsos positivos	0.001%
Precisión	11.67%

Tabla 10. Resultados clúster de dos fases

Las conclusiones que se obtuvieron al realizar la implementación de esta técnica fueron las siguientes:

- La técnica posee unos puntos muy positivos, una tasa de error muy baja, inferior al 1%, una tasa de falsos positivos supremamente baja del 0.001% y una precisión respetable, del 11.67%. Pero por otro lado la tasa de detección de la técnica también es significativamente baja, de tan solo el 0.029%. A pesar de todo lo anterior esta técnica sigue siendo bastante aplicable en el sector financiero, debido a que arroja muy pocos registros como inusuales por lo que se podrían revisar rápidamente y entre ellos posee una precisión considerable, es decir el esfuerzo invertido no es mucho al revisar pocos registros y el obtener varios aciertos dentro de este pequeño conjunto de datos hace que el esfuerzo valga la pena.

- Otra gran ventaja que posee la técnica es la facilidad de implementación y la rapidez del algoritmo.

- A pesar de lo anterior el que la tasa de detección sea tan baja implica que si se desea utilizar esta técnica en el sector financiero, esta sea una técnica complementaria y no la técnica principal puesto que podría dejar la empresa muy vulnerable a individuos que estén lavando activos y que no sean detectados nunca por ella.

5.5 Esperanza-Maximización

Esta método es una técnica que se cataloga como no supervisada, esto debido a que no requiere entrenar el modelo previamente con un conjunto que ya posea la clasificación entre normal e inusual.

Lo que busca la técnica es inferir a partir de unos datos observados el comportamiento de otras variables que no se pueden medir directamente, todo esto mediante cálculos iterativos de la verosimilitud, realizando optimizaciones paso a paso hasta que se llegue a cierta estabilidad en los valores obtenidos en la verosimilitud o por otro lado alcanzando el máximo de iteraciones configuradas en el algoritmo.

Lo primero que se debe determinar son las variables que utilizará el algoritmo, tal como lo sugieren los autores, las variables a utilizar serán el número de operaciones de entrada y de salida de las cuentas del individuo en el periodo de estudio, y también el monto de entrada y de salida en ese mismo periodo.

Dado que esta técnica depende mucho en las distancias para determinar los grupos y posteriormente las clasificaciones entre inusual y normal, los datos son estandarizados para que todas las variables puedan tener el mismo peso a la hora de calcular la distancia entre dos individuos.

Para implementar la técnica se utilizó un modelo de mezclas gaussianas que internamente utiliza el algoritmo de Esperanza-Maximización para conformar los grupos que devuelve, la técnica requiere unos parámetros como el número de grupos a formar, número máximo de iteraciones a realizar, tolerancia mínima de la técnica, tipo de covarianza que se tendrá y si es compartida o no.

En el artículo se detalla que se probaron varios tamaños de grupos para evaluar con cual se obtenía mejores resultados, adicionalmente en la implementación que se realizó para esta comparación, también se comparó el resultado de las métricas finales si se tomaba

una covarianza que se restringía a que solo podía ser diagonal o si se tomaba una covarianza sin esta restricción también conocido como covarianza completa.

Esta es una técnica que requiere el número de datos utilizadas en el entrenamiento tengan características de distribución similares a las de entrenamiento, por este motivo no es requerido una submuestra de esta población de entrenamiento sino que se toma toda.

La respuesta que ofrece el algoritmo a la entrada dada que son las variables y los parámetros del mismo, son los datos asociado a los K grupos configurados en la parametrización, el paso a seguir para determinar los elementos inusuales dentro de ese conjunto es determinar el grupo con menor número de elementos y esos elementos serán los que deberían ser categorizados como inusuales.

Se realiza el experimento con las ocho diferentes combinaciones a probar, cuatro posibles números de grupos (2, 3, 4 y 5) y dos posibles tipos de covarianzas (diagonal o completa), los otros parámetros enviados al algoritmo son un número máximo de iteraciones de 1000, una tolerancia obtenida a partir de la función chi-cuadrado y un que la covarianza será compartida.

Según los resultados obtenidos en la población de entrenamiento, se pudo observar que los mejores resultados se obtenían con cuatro grupos y un tipo de covarianza completa. Con estos parámetros ya definidos se vuele a implementar el algoritmo pero esta vez sobre la población de validación, obteniendo los resultados que se detallan en la Tabla 11.

Modelo	Esperanza - Maximización
Número de muestras	4514994
Error	0.53%
Verdaderos positivos	10
Verdaderos negativos	4491015
Falsos positivos	83
Falsos negativos	23886
Tasa de detección	0.04%
Tasa de falsos positivos	0.002%
Precisión	10.75%

Tabla 11. Resultados esperanza - maximización

Las conclusiones que se obtuvieron al realizar la implementación de esta técnica fueron las siguientes

- La técnica tiene grandes bondades como lo es la producción de pocos errores como lo puede evidenciar una tasa de error muy baja, inferior al 1% y una tasa de falsos positivos del 0.002%, por otro lado se tiene una tasa de detección que es muy pequeña, de tan solo el 0.04% pero aun con esto la precisión que se tiene es considerable, del 10.75%. Con todo lo anterior lo que se puede concluir es que esta al igual que la técnica de clúster de dos fases puede ser una técnica que se utilice en el sector financiero, esto debido al bajo número de registros producidos del gran conjunto de datos que originalmente se tenía, y que aun así se logra identificar individuos anómalos, los cuales podrían ser fácilmente investigados en su totalidad. Pero al igual que se comentó en el clúster de dos fases, es una técnica que puede ser secundaria en un sistema que tenga técnicas más robustas y que logre identificar por medio de otros métodos más individuos sospechosos, pues si solo se tuviese esta técnica el número de personas sospechosas que podrían pasar desapercibidas sería muy grande.

-Una gran ventaja que posee esta técnica es el hecho que es un método no supervisado, por tanto para las entidades financieras que no tengan el insumo de personas identificadas como sospechosas para entrenar otros modelos, siempre podrán utilizar estas variantes que no requiere de tales insumos.

5.6 Sequence Matching

Este método busca encontrar secuencias de transacciones que en su conjunto, al compararla con otras transacciones de individuos con cualidades similares, sean consideradas sospechosas.

Las variables que se utilizaran para este algoritmo son las propuestas por los autores, las cuales agrupan el comportamiento diario de cada uno de los individuos en el periodo a estudiar, las variables son el número de operaciones de entrada, número de operaciones de salida, total de operaciones, monto de entrada, monto de salida y monto total ingresado a todas las cuentas de cada individuo.

Esta técnica está estructurada en cuatro fases, en la primera etapa se obtienen los datos, se clasifica a los individuos en segmentos según su comportamiento, se forman las secuencias de transacciones y finalmente se obtienen las características transaccionales de cada uno de los segmentos.

Para la clasificación de los individuos en segmentos se utiliza nuevamente la clasificación dada por la entidad financiera dado el conocimiento más especializado que tiene la entidad sobre sus clientes. Las secuencias se forman tomando operaciones consecutivas realizadas por el mismo individuo, el número de estas operaciones que se deben tomar según los autores no debería ser muy grande, esto debido al gran número de cálculos que implican un tamaño de secuencia muy grande, por tal motivo el tamaño que se tomó para esta implementación fue de solo 3.

Para finalizar esta primera etapa se requiere a partir de los grupos conformados obtener los umbrales de cada una de las variables definidas, la manera como se obtuvieron dichos umbrales fue por medio de percentiles, con ello para cada uno de los K grupos definidos se obtendrán los seis umbrales correspondientes a las seis variables estudiadas.

Para la segunda fase del algoritmo se requiere obtener las secuencias riesgosas de todo el universo de secuencias existentes y que se obtuvieron en la primera fase, el criterio

con el cual se seleccionaron estas secuencias riesgosas es obtener las transacciones que por lo menos en una variable supera el umbral definido en la fase anterior, a partir de esta transacción se toma la secuencia formada por ella, la transacción previa y posterior realizada por el mismo individuo.

También en esta etapa se debe obtener un listado de secuencias de referencia para cada uno de los segmentos, dada la gran cantidad de individuos que pertenecen a cada uno de los grupos y el número de secuencias totales que existen en los mismos, se decidió que este listado de operaciones de referencias a lo sumo llegará a 1000, esto debido a que posteriormente cada una de las n transacciones determinadas como riesgosas serán comparadas con cada una de estas m transacciones de referencia, si no se controla el número de operaciones de referencia el número de comparaciones n*m podría llegar a ser muy complejo de realizar y los resultados obtenidos a partir de esa comparación muy probablemente serán muy similares al de un conjunto de referencia no tan grande, dado que lo que se toma al final es un promedio.

En la tercera etapa del algoritmo se realiza la comparación de similaridad entre cada una de las secuencias riesgosas obtenidas en la etapa anterior y todas las secuencias de referencias también seleccionadas anteriormente, pero antes de realizar dicha comparación se requiere normalizar las secuencias tanto las riesgosas como las de referencias, esta normalización se hace a nivel de cada variable, asegurando que cada una de ellas quede en el rango de 0 a 1.

Para realizar la comparación de similaridad se debe tener en cuenta varias cosas, se debe tener una función con la cual se pueda realizar una comparación de secuencias con igual longitud, la función elegida fue la distancia euclidiana, también se debe tener una función que esté en la capacidad de medir la similaridad entre secuencias de diferente tamaño, para esto los autores propusieron una técnica para tal fin.

Una vez definido cómo se hará las comparaciones y que los datos estén completamente normalizados, se procede a realizar la comparación de cada secuencia riesgosa con todas las secuencias de referencia obteniendo el promedio del resultado de la comparación entre cada par de secuencias.

En la cuarta y última etapa del algoritmo ya teniendo todas los promedios de cada una de las secuencias riesgosas se obtiene un umbral a partir del cual se realizará la clasificación entre operación inusual o normal, para la elección de este umbral se toman diferentes valores y se aplica sobre la población de entrenamiento y se observa con cual se obtienen mejores resultados.

Con toda la información obtenida en la fase de implementación para la población de entrenamiento como umbrales por variable y umbrales de clasificación final, se realiza la implementación del algoritmo sobre la población de validación y se obtiene los resultados que se pueden observar en la Tabla 12.

Modelo	Sequence Matching
Número de muestras	3437420
Error	1.60%
Verdaderos positivos	383
Verdaderos negativos	3381921
Falsos positivos	47257
Falsos negativos	7859
Tasa de detección	4.64%
Tasa de falsos positivos	1.38%
Precisión	0.80%

Tabla 12. Resultados sequence matching

Las conclusiones que se obtuvieron al realizar la implementación de esta técnica fueron las siguientes

- Con esta técnica se pudo obtener un porcentaje de error significativamente bajo, del 1.6% y también una tasa de falsos positivos baja, del 1.38%, pero por otro lado se tuvo una tasa de detección no muy buena, de tan solo un 4.64% y más importante una precisión muy baja del 0.8%, con todos estos datos la técnica parece poco atractiva de utilizar en el sector financiero.

- Aunque la técnica tiene ciertas bondades al momento de identificar operaciones normales, no es muy recomendable tampoco utilizarla como primer paso de una técnica más compleja en donde esta cumpla un rol de filtro, esto debido a que si bien se eliminarían muchas operaciones normales quitando ruido de los datos para una siguiente técnica, lamentablemente un gran número de individuos que si son sospechosos también serían filtrados.

5.7 Conclusiones de la comparación

Luego de realizadas todas las implementaciones y obtenidas todas las métricas de cada método se tiene la información presentada en la Tabla 13.

Modelo	Red Bayesiana Dinámica	Red neuronal de base radial	Máquina de soporte vectorial	Clúster de dos fases	Esperanza - Maximización	Sequence Matching
Número de muestras	2957028	4514994	4514994	4514994	4514994	3437420
Error	0.89%	52.47%	2.61%	0.53%	0.53%	1.60%
Verdaderos positivos	197	18117	13860	7	10	383
Verdaderos negativos	2930627	2127981	4383436	4491045	4491015	3381921
Falsos positivos	19034	2363117	107662	53	83	47257
Falsos negativos	7170	5779	10036	23889	23886	7859
Tasa de detección	2.67%	75.82%	58.00%	0.03%	0.04%	4.64%
Tasa de falsos positivos	0.65%	52.62%	2.40%	0.001%	0.002%	1.38%
Precisión	1.02%	0.76%	11.41%	11.67%	10.75%	0.80%

Tabla 13. Comparación resultados modelos de detección

Observando la información, se podría concluir que los mejores resultados entre las técnicas implementados están en las técnicas de Máquina de soporte vectorial, clúster de dos fases y esperanza maximización, en los cuales se obtienen precisiones similares, tasas de falsos positivos bajas y también tasas de error significativamente bajas.

Pero como se comento en cada uno de los métodos, es la tasa de detección la que nos da un indicio de la aplicabilidad de la técnica o por lo menos si la técnica puede hacer parte del método principal de detección en la entidad financiera, esto porque si la tasa de detección es muy baja, como lo es en el caso de clúster de dos fases y esperanza maximización, es muy riesgoso utilizar la técnica como método principal, debido a que son muchos los individuos sospechosos que pueden llegar a no ser detectados por el sistema.

Por todo lo anterior y teniendo en cuenta las cifras obtenidas se podría concluir que entre los modelos aquí comparados la máquina de soporte vectorial es quien obtiene unos mejores resultados y puede ser la técnica base a partir de la cual se construya un nuevo modelo de detección con unos mejores resultados.

6. IMPLEMENTACIÓN NUEVO MODELO DE DETECCIÓN

Teniendo en cuenta los resultados obtenido en la comparación de modelos de detección y el aprendizaje obtenido acerca de las ventajas y desventajas de los distintos modelos, en este capítulo se realizara un modelo que sea capaz de ofrecer unos mejores resultados a la hora de determinar si un individuo es sospechoso o no de lavar activos.

Como se dijo en el estado del arte, existe mucha información acerca de quien realiza la transacción que con los modelos existentes se puede estar desperdiciando, atributos como la edad, ingresos, egresos, patrimonio, ocupación y el tipo de persona de quien realiza la operación, todos estos datos pueden ayudar a dar un mejor perfil de quien hace la operación, pues puede que el nivel de riesgo de una transacción varíe notablemente dependa de quien realice dichos movimientos, sus características y la diferencia con otros que pueden tener características similares.

Se cuenta con una gran ventaja al poseer un conjunto real de operaciones sospechosas y un gran conjunto de individuos de entrenamiento, pues gracias a esto se podrían conformar con estos datos varias modelos que intenten abstraer la relación que existe entre las variables predictivas y las respuestas.

La idea detrás del modelo que se plantea es utilizar las bondades del modelo el cual se determinó poseía mejores resultados en la comparación, las máquinas de soporte vectorial, y utilizarlo como el paso uno del nuevo modelo, encargándose este primer modelo de servir como filtro, descartando un gran número de operaciones que con alta probabilidad resultan siendo no sospechosas reduciendo así la cantidad de individuos a aplicarle la segunda parte de la técnica la cual se encargará de refinar los resultados en búsqueda de un menor conjunto de salida con una precisión mucho más alta.

La técnica que se propone en la segunda etapa del modelo dada sus características de aprendizaje es un árbol de clasificación, el cual pese a su simpleza es una poderosa

herramienta de aprendizaje de patrones, y con el cual se puede llegar a refinar los resultados y obtener una muy buena precisión en la clasificación final.

Antes de implementar las dos fases propuestas, la de filtrado (Máquina de soporte vectorial) y la de refinamiento (Árbol de clasificación), es importante poder realizar una etapa cero en la cual se selecciona del conjunto total de variables, las variables que serán utilizadas como entrada tanto en la máquina de soporte vectorial como en el árbol de clasificación, esta fase cero se realizará sobre una submuestra de la población de entrenamiento, la cual se obtiene de manera aleatoria.

Los algoritmos utilizados para realizar la selección de variables a partir de la submuestra obtenida fueron los métodos Backward, Forward y Stepwise, bajo el criterio AIC. Con estas tres técnicas la idea es partir de un punto inicial y según la técnica agregar o eliminar variables según como se afecte el criterio de decisión, en este caso el AIC, de tal manera que siempre se agregue o se elimine la variable que mejore significativamente el indicador, cuando ningún otro ingreso a egreso de variables al modelo cambian significativamente el criterio de decisión se toma ese conjunto de variables como el modelo final.

Las anteriores técnicas podrían llegar a dar resultados diferentes, por eso mismo y para cubrir todas las posibilidades serán realizadas las tres y se observará las diferentes variables propuestas y a partir de estas se escogen las variables que serán enviadas a los modelos para realizar la labor de detección.

Luego de ejecutado los tres algoritmos en la submuestra seleccionada, los resultados sobre cuáles variables incluir fueron idénticos, siendo las variables elegidas en orden de importancia las siguientes:

Edad, Ocupación 1, Ocupación 2, Ocupación 3, Tipo de persona, Ocupación 4, Ocupación 5, Ocupación 6, Ventas, Ocupación 7 e Ingresos.

Se pudo observar el gran número de variables que resultaron significativas para este conjunto de individuos de entrenamiento, destacando que la edad resultó siendo la variable más significativa y que un gran número de ocupaciones resultaron en este

conjunto, un punto importante a tener en cuenta es que adicional a estas variables se incluirán entre las variables que se le pasaran a los modelos, las variables que originalmente utilizaban las máquinas de soporte vectorial, el número de operaciones de entrada y salida a las cuentas del cliente y el monto total que sumaban estas operaciones, esto porque quedó demostrado el buen resultado que se obtenían con estas variables, por tal motivo el total de variables que se utilizaran en el modelo asciende a 14.

Ya con las selección de variables, se comienza la implementación de las máquinas de soporte vectorial donde nuevamente se utilizara el software libsvm, pero antes de realizar nuevamente el entrenamiento y predicción por medio de esta librería, se implementa la misma estandarización propuesta por los autores de la técnica de máquina de soporte vectorial, es decir a cada variable transformarla dividiéndola por cuatro veces su desviación estándar.

Una vez la preparación de las variables se encuentren listas se aplica la técnica de máquina de soporte vectorial de tipo C-SVM sobre la muestra de la población de entrenamiento, con los parámetros con los cuales se determinó que funcionaba mejor el algoritmo, es decir, el parámetro de factor de control tomando el valor de 0.5 y el castigo por mala clasificación tomando el valor de 50.

Una vez entrenado el modelo se procede a realizar la predicción sobre la propia población completa de entrenamiento, todo esto para obtener el grupo de individuos que esta primera técnica determinó que era sospechosa, con esto se obtiene que se pasa de tener 3.009.995 individuos donde se tienen 15.930 individuos sospechosos, a tener 1.023.570 donde se tiene 12.012 individuos sospechosos, es decir, se logra eliminar el 60% de la población manteniendo más del 75% de la población sospechosa.

Con este nuevo conjunto de 1.023.570 individuos se procede a entrenar un árbol de clasificación, indicando las mismas 14 variables como entrada y siendo la variable sospechoso la variable de clasificación, este árbol se genera y se guarda para ser posteriormente utilizado.

Ahora que la fase de entrenamiento ha sido terminada y se tienen ya los modelos entrenados para realizar la clasificación por máquina de soporte vectorial y la clasificación por árbol de regresión, se puede aplicar el modelo completo a la población de validación.

Para poder aplicar el modelo de máquina de soporte vectorial entrenado, primero se transforma los datos de entrada dividiéndolos por cuatro veces su desviación estándar, luego se aplica la predicción con base al modelo que se tiene, con esto se pasa en la primera fase de tener 4.514.994 individuos con 23.896 marcados como sospechosos, a tener 1.457.571 con 17.708 individuos marcados como sospechosos, reduciendo para el conjunto de validación el 67% de la población, manteniendo el 74% de los individuos sospechosos.

Finalmente, a esa nueva población de 1.457.571 individuos que en principio fue marcado por las máquinas de soporte vectorial como sospechosos, se le aplica la clasificación por medio del árbol que se obtuvo en la fase previa, con lo cual se obtuvo que este árbol dio como sospechosos a un total de 9.572 individuo de los cuales realmente eran sospechosos 1.756.

Al obtener los datos globales de cómo se comportan los indicadores de este nuevo método se obtiene la Tabla 14.

Modelo	Nuevo modelo
Número de muestras	4514994
Error	0.66%
Verdaderos positivos	1756
Verdaderos negativos	4483282
Falsos positivos	7816
Falsos negativos	22140
Tasa de detección	7.35%
Tasa de falsos positivos	0.17%
Precisión	18.34%

Tabla 14. Resultados nuevo modelo de detección

Para obtener una visión más global de los resultados del modelo se puede realizar de una vez la comparación con los modelos previamente generados, tal como la muestra la Tabla 15.

Modelo	Red Bayesiana Dinámica	Red neuronal de base radial	Máquina de soporte vectorial	Clúster de dos fases	Esperanza - Maximización	Sequence Matching	Nuevo Modelo
Número de muestras	2957028	4514994	4514994	4514994	4514994	3437420	4514994
Error	0.89%	52.47%	2.61%	0.53%	0.53%	1.60%	0.66%
Verdaderos positivos	197	18117	13860	7	10	383	1756
Verdaderos negativos	2930627	2127981	4383436	4491045	4491015	3381921	4483282
Falsos positivos	19034	2363117	107662	53	83	47257	7816
Falsos negativos	7170	5779	10036	23889	23886	7859	22140
Tasa de detección	2.67%	75.82%	58.00%	0.03%	0.04%	4.64%	7.35%
Tasa de falsos positivos	0.65%	52.62%	2.40%	0.001%	0.002%	1.38%	0.17%
Precisión	1.02%	0.76%	11.41%	11.67%	10.75%	0.80%	18.34%

Tabla 15.Comparación final modelos existentes y nuevo modelo

Como se puede ver, comparativamente son muchas las bondades de este nuevo modelo sobre todos los modelos estudiados y comparados anteriormente, la precisión obtenida del 18.34% mejora en un 57.16% más la mejor precisión obtenida hasta ahora que era la del clúster de dos fases, las tasas de error y de falsos positivos siguen siendo significativamente muy bajas, del 0.66% y 0.17% respectivamente, pero lo más importante del método es la cantidad de registros que al final reporta como sospechosos, tan solo 9.572 individuos, teniendo en cuenta el gran número de individuos del que se parte originalmente, 4.514.994, el que se pueda tener un método que clasifique a una

cantidad no tan significativa como sospechosa teniendo una precisión bastante buena según los indicadores que se tenían de referencia es bastante provechoso, esto porque este número de registro da la posibilidad de realizar una investigación minuciosa de cada uno de los individuos marcados como sospechosos por el modelo, sabiendo que gran parte del esfuerzo realizado no se perderá y que pueden ser muchos los reportes realizados a los entes superiores.

Para finalizar, visto como un todo el proceso que se llevó a cabo para poder generar este nuevo modelo de detección es el que se detalla en la Figura 5.

Aquí se puede ver el conjunto de pasos a realizar si se desea implementar este nuevo modelo, contemplando las fases de estudio y adquisición de las variables que serán potencialmente utilizadas por el modelo, las cuales pueden variar según la región y la entidad sobre la cual se desee implementar el modelo.

También se detalla el proceso de separación de poblaciones de entrenamiento y validación, y cuáles son las diferentes acciones que se realizan sobre cada una de estas poblaciones, siendo la población de entrenamiento utilizadas para obtener los modelos base de máquinas de soporte vectorial y árbol de clasificación que serán utilizadas posteriormente por la población de validación para al final poder clasificar los individuos sospechosos y no sospechosos.

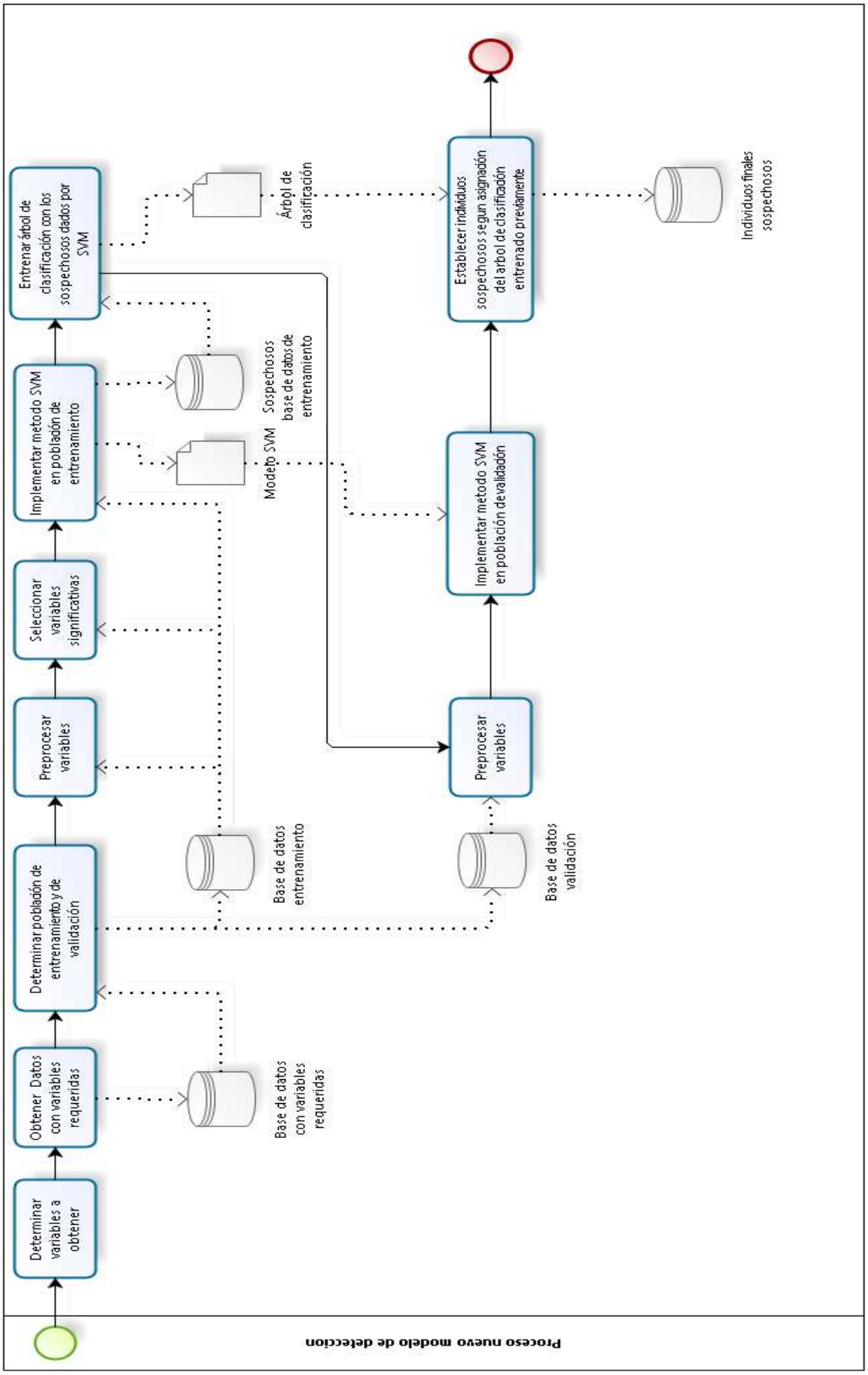

Figura 5.Proceso de construcción del nuevo modelo de detección

7. CONCLUSIONES Y TRABAJO FUTURO

Por medio de este trabajo se han podido obtener varios resultados positivos en el ámbito de la detección de lavado de activos y la financiación del terrorismo, el primer resultado que se obtuvo fue en el análisis inicial del estado del arte, donde se pudo determinar las brechas existentes en los métodos actuales y se pudo empezar a plantear un plan de acción para cerrar dichas brechas.

Se pudo determinar que en las metodologías actuales la información propia del cliente ha sido subutilizada y se planteo la hipótesis que el poder utilizarlas como variables complementarias a las transaccionales podría ayudar a reducir las tasas de falsos positivos.

Una contribución que se pudo realizar por medio de esta investigación fue la comparación de las técnicas analizadas con un conjunto real de datos, esto debido a que cada una de las técnicas cuando fueron propuestas por sus autores fueron entrenadas y evaluadas con conjuntos diferentes de datos, por lo que aunque las métricas que se reflejaran en los artículos fueran similares una verdadera comparación no puede ser realizada debido a esas grandes diferencias entre los conjunto de datos y el cómo fueron generados, pues si bien en la gran mayoría de artículos los datos de transacciones no sospechosas fueron tomado de conjuntos de datos reales, los datos de las transacciones sospechosas en su mayoría fueron generados sintéticamente por medio de diversas funciones, pero en el experimento que se llevo a cabo en esta investigación, estas anomalías también fueron tomadas a partir de datos reales y todas las técnicas estudiadas fueron comparadas con el mismo conjunto de datos.

El comparar con las mismas métricas y tener varias de ellas y no una sola también se considera como algo positivo en la investigación, esto debido que el tener varias métricas da cuenta de las diferentes bondades que puede tener una técnica de detección y la aplicabilidad que puede darse a cada una de estas según sus resultados.

Luego de la comparación se pudo ver como con el conjunto de datos estudiado las maquinas de soporte vectorial obtuvieron comparando varios criterios un mejor

resultado, por lo cual fue tomada como técnica base para construir el nuevo modelo de detección, siendo importante también todo el conocimiento adquirido al implementar las demás modelos de detección.

Al final el resultado entregado es un nuevo modelo de detección que estuvo en capacidad de obtener unos mejores resultados que las técnicas previamente estudiadas, combinando las bondades de la máquina de soporte vectorial en una primera etapa como método que filtraba operaciones normales y utilizando un segundo modelo, un árbol de clasificación, para tomar la salida entregada por la máquina de soporte vectorial y refinarla para obtener unos mejores resultados que son mas aplicables en el sistema financiero.

Algo importante de esta investigación es que la metodología del cómo se obtuvo el nuevo modelo se dejo muy especificada, de tal manera que si se desea replicar este modelo para utilizarla con otro conjunto de datos y otro conjunto de variables, pueda realizarse sin ningún problema.

El método acá planteado a pesar de su simpleza obtuvo muy buenos resultados superando notablemente a los métodos estudiados, así que se espera que en un futuro este trabajo pueda servir como base y a partir de la combinación de modelos diferentes con diferentes características como se realizo en esta investigación se puedan obtener aun mejores resultados.

BIBLIOGRAFÍA

[1] B. Buchanan, «Money laundering—a global obstacle», *Res. Int. Bus. Finance*, vol. 18, n.º 1, pp. 115-127, abr. 2004.

[2] B. Unger *et al.*, «The amounts and effects of money laundering», *Minist. Finance Hague*, 2006.

[3] UNODC, *ESTIMATING ILLICIT FINANCIAL FLOWS RESULTING FROM DRUG TRAFFICKING AND OTHER TRANSNATIONAL ORGANIZED CRIMES*. 2011.

[4] J. Walker y B. Unger, «Measuring global money laundering: the Walker gravity model», *Rev. Law Econ.*, vol. 5, n.º 2, pp. 821-853, 2009.

[5] M. Levi, «Money Laundering and Its Regulation», *Ann. Am. Acad. Pol. Soc. Sci.*, vol. 582, n.º 1, pp. 181-194, ene. 2002.

[6] F. A. T. Force, «The forty recommendations», 2003.

[7] R. Menon y S. Kuman, «Understanding the role of technology in anti-money laundering compliance», *Infosys Technol. Ltd*, vol. 1, pp. 2-4, 2005.

[8] Colombia, «Codigo Penal Colombiano». 2000.

[9] B. Bartlett, «The negative effects of money laundering on economic development», *Asian Dev. Bank Reg. Tech. Assist. Proj. No*, vol. 5967, 2002.

[10] P. J. Quirk, «Money laundering: muddying the macroeconomy», *Finance Dev.*, vol. 34, pp. 7-9, 1997.

[11] «AML Survey Results from Dow Jones Risk & Compliance & ACAMS». .

[12] N. Mackrell, «Economic consequences of money laundering», *Money Laund. 21st Century Risks Countermeas.*, pp. 29-35, 1996.

[13] A. Chong y F. Lopez-De-Silanes, «Money Laundering and Its Regulation», *Econ. Polit.*, vol. 27, n.º 1, pp. 78-123, mar. 2015.

[14] P. J. Quirk, «Macroeconomic implications of money laundering», *Trends Organ. Crime*, vol. 2, n.º 3, pp. 10-14, mar. 1997.

[15] D. Masciandaro y U. Filotto, «Money laundering regulation and bank compliance costs: What do your customers know? Economics and the Italian experience», *J. Money Laund. Control*, vol. 5, n.º 2, pp. 133-145, 2001.

[16] N. A. L. Khac y M. T. Kechadi, «Application of Data Mining for Anti-money Laundering Detection: A Case Study», en *2010 IEEE International Conference on Data Mining Workshops*, 2010, pp. 577-584.

[17] Z. Gao y M. Ye, «A framework for data mining-based anti-money laundering research», *J. Money Laund. Control*, vol. 10, n.º 2, pp. 170-179, 2007.

[18] K. D. Rohit y D. B. Patel, «Review On Detection of Suspicious Transaction In Anti-Money Laundering Using Data Mining Framework», *Int. J. Innov. Res. Sci. Technol.*, vol. 1, n.º 8, pp. 129-133, 2015.

[19] E. A. Lopez-Rojas y S. Axelsson, «Money laundering detection using synthetic data», en *The 27th annual workshop of the Swedish Artificial Intelligence Society (SAIS); 14-15 May 2012; Örebro; Sweden*, 2012, pp. 33-40.

[20] L.-T. Lv, N. Ji, y J.-L. Zhang, «A RBF neural network model for anti-money laundering», en *International Conference on Wavelet Analysis and Pattern Recognition, 2008. ICWAPR '08*, 2008, vol. 1, pp. 209-215.

[21] J. Tang y J. Yin, «Developing an intelligent data discriminating system of anti-money laundering based on SVM», en *Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005*, 2005, vol. 6, p. 3453-3457 Vol. 6.

[22] E. Dorj y E. Altangerel, «Anomaly detection approach using Hidden Markov Model», en *2013 8th International Forum on Strategic Technology (IFOST)*, 2013, vol. 2, pp. 141-144.
[23] X. Liu, P. Zhang, y D. Zeng, «Sequence matching for suspicious activity detection in anti-money laundering», en *Intelligence and Security Informatics*, Springer, 2008, pp. 50-61.
[24] S. Raza y S. Haider, «Suspicious activity reporting using dynamic bayesian networks», *Procedia Comput. Sci.*, vol. 3, pp. 987-991, 2011.
[25] M. F. Jiang, S. S. Tseng, y C. M. Su, «Two-phase clustering process for outliers detection», *Pattern Recognit. Lett.*, vol. 22, n.º 6–7, pp. 691-700, may 2001.
[26] Z. Chen, L. D. V. Khoa, A. Nazir, E. N. Teoh, y E. K. Karupiah, «Exploration of the effectiveness of expectation maximization algorithm for suspicious transaction detection in anti-money laundering», en *2014 IEEE Conference on Open Systems (ICOS)*, 2014, pp. 145-149.
[27] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, y H. Nielsen, «Assessing the accuracy of prediction algorithms for classification: an overview», *Bioinformatics*, vol. 16, n.º 5, pp. 412-424, 2000.
[28] C. Liu, P. M. Berry, T. P. Dawson, y R. G. Pearson, «Selecting thresholds of occurrence in the prediction of species distributions», *Ecography*, vol. 28, n.º 3, pp. 385-393, 2005.
[29] T. Fawcett, «An introduction to ROC analysis», *Pattern Recognit. Lett.*, vol. 27, n.º 8, pp. 861-874, 2006.
[30] J. Friedman, T. Hastie, y R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin, 2001.
[31] Q. Wei y R. L. Dunbrack Jr, «The role of balanced training and testing data sets for binary classifiers in bioinformatics», *PloS One*, vol. 8, n.º 7, p. e67863, 2013.
[32] N. Japkowicz, «Learning from imbalanced data sets: a comparison of various strategies», en *AAAI workshop on learning from imbalanced data sets*, 2000, vol. 68, pp. 10-15.
[33] N. V. Chawla, «C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure», en *Proceedings of the ICML*, 2003, vol. 3.
[34] C. Drummond y R. C. Holte, «C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling», en *Workshop on learning from imbalanced datasets II*, 2003, vol. 11.