



UNIVERSIDAD NACIONAL DE COLOMBIA

# Modelo predictivo para riesgo de liquidez de una entidad fiduciaria usando minería de datos

**Diana Patricia León Sánchez**

Universidad Nacional de Colombia  
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial  
Bogotá, D.C., Colombia  
2015



# Modelo predictivo para riesgo de liquidez de una entidad fiduciaria usando minería de datos

**Diana Patricia León Sánchez**

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:  
**Magister en Ingeniería Industrial**

Director(a):  
Elizabeth León Guzmán, Ph.D.

Línea de Investigación:  
Nombrar la línea de investigación en la que enmarca la tesis o trabajo de investigación  
Grupo de Investigación:  
Nombrar el grupo en caso que sea posible

Universidad Nacional de Colombia  
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial  
Bogotá, D.C., Colombia  
2015



A Dios, mi familia y mi novio, que me apoyaron y aconsejaron en la realización de este trabajo.



# Agradecimientos

A mis padres, a mi familia, amigos y compañeros del Grupo de Investigación MIDAS que me colaboraron en la realización del trabajo y en la revisión de las aplicaciones realizadas en Minería de Datos. Así mismo, a la Superintendencia Financiera de Colombia por su colaboración en la resolución de inquietudes referentes al suministro de información financiera de los fondos de inversión.

Especialmente le agradezco a la Profesora Elizabeth León Guzmán por su paciencia, orientación y apoyo incondicional para la culminación de este trabajo





## Resumen

Este trabajo presenta un modelo predictivo para riesgo de liquidez basado en técnicas de minería de datos para una entidad fiduciaria. Se presentan los principales avances realizados desde la minería de datos en la gestión de riesgos financieros y otros métodos estadísticos empleados actualmente para riesgo de liquidez. El modelo en la fase de preprocesamiento analiza una serie histórica de registros de adiciones y retiros de una cartera colectiva y de otros factores macroeconómicos que inciden sobre el comportamiento del nivel de liquidez. Para la parte de predicción utiliza técnicas como redes neuronales y Naive Bayes. El modelo es validado con datos reales del producto fiduciario, los cuales son publicados por Asofiduciarias, Superfinanciera y la misma entidad.

**1) Minería de datos y 2) riesgo de liquidez.**

## Abstract

This paper presents a predictive model for liquidity risk based on data mining techniques for trustee. Major progress since data mining in financial risk management and other statistical methods currently used for liquidity risk are presented. The model at the preprocessing phase analyzes a series of historical records additions and withdrawals from a pooled portfolio and other macroeconomic factors that influence the behavior of liquidity. For the prediction part uses techniques such as neural networks and Nave Bayes. The model is validated with real data of the Trust product, which are published by Asofiduciarias, Superfinanciera and the same entity.

**Keywords: 1) Data Mining and 2) liquidity risk**

# Contenido

<b>Agradecimientos</b>	<b>vii</b>
<b>Resumen</b>	<b>ix</b>
<b>Introducción</b>	<b>xii</b>
<b>1. Marco Teórico y Estado del Arte</b>	<b>1</b>
1.1. Relevancia del riesgo de liquidez en entidades financieras . . . . .	1
1.2. Caracterización del sector fiduciario en Colombia . . . . .	2
1.2.1. Caracterización del producto carteras colectivas en Colombia . . . . .	2
1.3. Medición del riesgo de liquidez . . . . .	4
1.3.1. Metodología de medición de la Superintendencia Financiera de Colombia para entidades fiduciarias . . . . .	4
1.3.2. Medición del riesgo de liquidez a partir del gap de liquidez . . . . .	5
1.4. Técnicas de clasificación y predicción . . . . .	6
1.4.1. Series de tiempo . . . . .	6
1.4.2. Técnicas de regresión . . . . .	7
1.4.3. Técnicas de clasificación . . . . .	9
1.5. Aplicaciones de minería de datos en riesgo financiero . . . . .	15
1.5.1. Riesgo de mercado . . . . .	15
1.5.2. Riesgo de lavado de activos y financiación del terrorismo (LA/FT) en detección de fraude . . . . .	17
1.5.3. Detección de riesgos operativos . . . . .	18
1.5.4. Aplicaciones en riesgo de crédito . . . . .	19
1.5.5. Riesgo de solvencia y liquidez en flujo de caja . . . . .	19
1.6. Resumen del capítulo . . . . .	21
<b>2. Análisis descriptivo y Pre-procesamiento</b>	<b>22</b>
2.1. Conjunto de datos . . . . .	22
2.1.1. Análisis de tendencias . . . . .	25
2.2. Medidas estadísticas de las variables . . . . .	31
2.3. Preprocesamiento . . . . .	32
2.3.1. Detección de valores atípicos y limpieza de datos . . . . .	32
2.3.2. Transformación de datos . . . . .	36

---

2.4. Selección de Atributos para el Modelo de Clasificación . . . . .	37
2.5. Matriz de correlación . . . . .	37
2.6. Resumen del capítulo . . . . .	39
<b>3. Modelo predictivo de Riesgo de Liquidez</b>	<b>42</b>
3.1. Modelo General de Predicción de Riesgo de Liquidez . . . . .	42
3.2. Predicción del MRO a partir de la serie de tiempo usando Redes Neuronales	43
3.2.1. Preprocesamiento para series de tiempo . . . . .	44
3.2.2. Arquitectura de la Red Neuronal . . . . .	44
3.2.3. Experimentos con Redes Neuronales . . . . .	45
3.2.4. Resultados entrenamiento de RNA . . . . .	46
3.2.5. Análisis comparativo de resultados . . . . .	48
3.3. Predicción del MRO mediante árboles de decisión . . . . .	49
3.3.1. Preprocesamiento y Selección de atributos del modelo . . . . .	49
3.3.2. Árbol de decisión . . . . .	50
3.4. Integración de los módulos del modelo para predicción . . . . .	51
3.5. Resumen . . . . .	53
<b>4. Conclusiones y Trabajos Futuros</b>	<b>54</b>
4.1. Conclusiones . . . . .	54
4.2. Trabajos futuros . . . . .	55

## Introducción

La gestión de riesgos consiste esencialmente en tomar decisiones con base en las expectativas de beneficios futuros o las posibilidades de pérdidas, según el riesgo asumido. En la actualidad la globalización de los mercados financieros, las complejas estructuras transaccionales y el alto nivel de apalancamiento han generado mayor volatilidad de las inversiones realizadas en el mercado y un nivel de exposición equivalente a la materialización de eventos de riesgo financiero. Este grado mayor de exposición a la variabilidad del mercado implica adoptar instrumentos de medición y controles adecuados para mitigar los riesgos financieros a los que están expuestos los inversionistas, instituciones financieras, agentes gubernamentales y otras entidades comerciales. Sin embargo, esta labor se ha vuelto cada vez más compleja debido al gran volumen de información almacenado en sus bases de datos, análisis que se ha convertido en una tarea titánica para la toma de decisiones.

Desde finales del Siglo XX esta problemática ha despertado el interés de entidades financieras e instituciones gubernamentales como consecuencia de los efectos generados por las principales crisis financieras, entre los que se encuentra la quiebra de importantes instituciones financieras y la exposición al riesgo financiero. Los efectos generados por la crisis financiera, las grandes pérdidas y la desestabilización económica, conllevo en el año de 1974 a la conformación del comité Basilea para la regulación bancaria y prácticas de supervisión de bancos centrales. Este primer acercamiento, favoreció la formulación de recomendaciones de carácter regulatorio en instituciones de supervisión y vigilancia de entidades financieras a nivel internacional y del mercado doméstico[?, ?]. En la figura0-1 (Clases de riesgo financiero) se presentan los principales riesgos financieros.



Figura 0-1: Clases de riesgo financiero

Dentro del listado de riesgos financieros que inciden sobre el desempeño de una entidad, se encuentra el riesgo de liquidez, ”entendido como la contingencia de no poder cumplir de

manera plena y oportuna con las obligaciones de pago en las fechas correspondientes, debido a la insuficiencia de recursos líquidos o a la necesidad de asumir costos inusuales de fondeo”, el cual ha centrado la atención y seguimiento por parte del sector financiero, especialmente, por los efectos negativos que ha tenido sobre el mercado. En Colombia, entre los años 2002 y 2006 el sistema financiero experimentó dificultades, particularmente, en el precio de mercado de los títulos de deuda pública y por ende en la administración del portafolio de inversionistas particulares, quienes asumen un nivel de riesgo diferente, de acuerdo a sus preferencias de rentabilidad, dichos perfiles son: conservador, moderado y alto (o dinámico). Un perfil conservador normalmente no aceptará pérdidas de capital en ningún momento, un perfil moderado tolerará pérdidas temporales, mientras que un inversionista con un perfil de riesgo alto estará dispuesto a perder capital incluso al final de la inversión [39].

Los gobiernos locales y las comunidades de integración regional han adoptado estas recomendaciones desde su fundación hasta hoy en día, buscando adoptar mecanismos de medición cada vez más eficientes que mitiguen el nivel de exposición al riesgo financiero y faciliten la toma de decisiones. Es así como en la última década se ha pasado de técnicas estadísticas sencillas y se ha centrado la atención en la minería de datos, que constituye un proceso promisorio para lograr este propósito, dada su capacidad para manejar altos volúmenes de datos, permitiendo además estimar el comportamiento de la demanda, así como la toma de decisiones y oportunidades que aseguren la competitividad en los mercados.

Muchos trabajos han realizado procesos de análisis de riesgo financiero, suscitando la generación de modelos predictivos y descriptivos, enfocados principalmente en técnicas estadísticas y en aprendizaje de máquina [15], y las principales aplicaciones surgieron a partir de la necesidad de:

- Realizar un pronóstico de series de tiempo multidimensional con alto nivel de ruido y tener en cuenta criterios específicos de eficiencia como el nivel de rendimiento de las inversiones a partir de un nivel de riesgo determinado, así como la predicción de los mismos.
- Efectuar pronde multi-resolución coordinada en una medida de tiempo (minutos, d, semanas, meses y años)
- Incorporar un flujo de señales de alerta como datos de entrada para la generación de modelos preventivos (por ejemplo el caso de Enron, los efectos del 11 de septiembre, entre otros)
- Ser capaz de explicar el pronóstico y predicción generados por el modelo, para la toma de decisiones. Así como identificar patrones de comportamiento en el corto plazo e incorporar el impacto de medidas regulatorias de mercado [25]

Aunque son varios los estudios realizados son pocas las aplicaciones realizadas en países en vía de desarrollo y que también requieren tener modelos que se ajusten a las condiciones del mercado local, especialmente, en materia de riesgo de liquidez.

Para el desarrollo de este trabajo se empleará la metodología KDD [4] para el desarrollo del modelo de riesgo de liquidez para una cartera colectiva de una entidad fiduciaria, enfocándose en la aplicación de técnicas de clasificación y predicción de minería de datos en la identificación de patrones o señales que indiquen potenciales retiros de dinero.

El análisis de grandes bases de datos transaccionales y la comprensión del comportamiento futuro de las series de tiempo financieras se ha transformado en una tarea cada vez más difícil [1] y se ha convertido en uno de los principales problemas del sector financiero, especialmente en los periodos de crisis donde resulta más urgente contar con elementos de juicio [2] y tomar decisiones de inversiones oportunas. La adopción de instrumentos y herramientas que permitan identificar y predecir patrones de comportamiento de series de tiempo financieras, evitar incurrir en riesgos innecesarios, aprovechar oportunidades de mercado a priori, y que a su vez, faciliten la toma de decisiones, ha congregado diversas técnicas y disciplinas que permitan tener una mejor aproximación a esta problemática y anticipe posibles quiebras financieras.

En la actualidad es posible pronosticar series de tiempo financieras a partir de modelos estadísticos y técnicas de inteligencia artificial. El primer grupo contempla modelos como el ARIMA, GARCH, Box-Jenkins y mixtos, y el segundo grupo incluye técnicas de aprendizaje de máquina, redes neuronales y algoritmos genéticos [15].

Estudios realizados en la última década han aplicado redes neuronales a series de tiempo financieras, obteniendo resultados favorables que demuestran que las redes neuronales superan modelos de regresión lineal, especialmente en el caso de las series de tiempo generadas por el mercado de valores al ser considerados como mercados complejos, no lineales, dinámicos y caóticos. [37] y [38] No obstante las aplicaciones en series de tiempo de riesgo de liquidez en mercados financieros como el colombiano son limitadas, y emplean técnicas basadas en supuestos que en ocasiones pueden distorsionar los resultados.

El ejercicio de pronosticar el monto de recursos de capital que debe mantener en efectivo una compañía para pagar todas sus obligaciones a tiempo, se ha convertido en una difícil tarea, particularmente en el caso de las entidades fiduciarias, quienes no tienen un valor estándar similar al encaje bancario para cada uno de los negocios administrados por contratos de fiducia.

En Colombia, existen diferentes productos fiduciarios para la administración eficiente de recursos de capital, tal es el caso de las carteras colectivas, donde son depositados los recursos

del fideicomitente (cliente), e invertidos en el mercado a través de títulos de renta fija (ej. CDTs) y/o variables (acciones); sin embargo existen diferentes tipos de carteras colectivas: abiertas, escalonadas y cerradas, las primeras se caracterizan porque la redención de las participaciones (retiros) se podrá realizar en cualquier momento. Las segundas se caracterizan porque la redención de las participaciones sólo se puede realizar una vez transcurran los plazos que se hayan determinado previamente en el reglamento del producto. Las terceras se caracterizan porque la redención de la totalidad de las participaciones sólo se puede realizar al final del plazo previsto para la duración de la cartera colectiva. [13]

Conforme a lo señalado se ha visto la necesidad de adoptar o desarrollar herramientas que se anticipen a un retiro masivo de recursos, especialmente para carteras colectivas abiertas, que pongan en riesgo la obligación que se tiene con otros clientes y comprometa la perdurabilidad como entidad fiduciaria en el mercado, dada su incapacidad para responder ante todos los pagos demandados por sus fideicomitentes (clientes), provocando un problema de iliquidez y posible insolvencia, lo que puede transgredir las puertas de la organización y trasladarse al mercado, generando un efecto conocido como "bola de nieve" que puede afectar la estabilidad del sistema financiero. [14]

Grandes entidades financieras han encontrado en la minería de datos, una fuente de aplicación para estimar el comportamiento de la demanda, la medición y control de riesgos financieros, dado el alto volumen de datos que manejan tanto las subsidiarias como la casa matriz y teniendo en cuenta la complejidad que representa el análisis de la información y la necesidad que tienen los gerentes de las organizaciones para tomar decisiones ágiles y oportunas, que les permita ser competitivos en el mercado, y dada la necesidad de estimar el nivel de riesgo de liquidez al que se encuentran expuestos en un mercado altamente cambiante, especialmente, en periodos de crisis financiera, resulta pertinente desarrollar un instrumento de medición y control de este riesgo, que genere señales de alerta y que contemple el mayor número de variables exógenas a la entidad, cuyo grado de incidencia sea crucial para determinar el comportamiento del mercado.

Dados los resultados presentados por otras entidades financieras a nivel internacional, y las ventajas que han obtenido respecto al resto de sus competidores, al contar con información clara y precisa, que le permite a la alta dirección tomar decisiones oportunas y obtener una mejor participación del mercado, al anticiparse a las necesidades de sus clientes (merchandising), riesgos operativos que conlleven a pérdidas de capital, y por ende, determinar el grado de exposición ante eventos que le impidan a la entidad atender sus obligaciones, y en ocasiones de crisis, llevarlas a la quiebra; motivo por el cual se considera que el resultado de este estudio, representado en el desarrollo de un modelo de riesgo, será una herramienta muy útil para entidades fiduciarias, especialmente, en Colombia, donde no hay un método que pronostique en productos fiduciarios eventos de riesgo de liquidez, y queden expuestos

a sanciones por parte de la Superintendencia Financiera de Colombia (SFC) o la quiebra.

El propósito de este trabajo es suministrar un modelo para la predicción del riesgo de liquidez de una cartera colectiva, utilizando técnicas de minería de datos, centrandó nuestra atención en el uso de técnicas de redes neuronales "Multi Layer Perceptron" con backpropagation, feed-forward y alimentada con Moving Average, "Non linear Autoregressive Exogenous Model - NARX" para el Máximo Retiro Observado (MRO), cuyos resultados serán comparados mediante el Error Mínimo Cuadrático (EMC) frente al resultado obtenido por modelos de riesgo tradicionales, y un árbol de decisión, que busca crear un modelo de clasificación para el riesgo de liquidez al incluir variables macroeconómicas y otros factores que inciden sobre dicha variable, que pueden incidir en su comportamiento.

## Objetivos

### General

Diseñar un modelo de predicción basado en minería de datos para el control del riesgo de liquidez en entidades fiduciarias.

### Específicos

- Estudiar metodologías adoptadas por entidades financieras para control y gestión de riesgo de liquidez
- Determinar las técnicas de minería de datos que harán parte del modelo, tomando como base la recopilación y el análisis del estado del arte.
- Definir el proceso de *KDD* con la técnica seleccionada y con los datos de liquidez de una entidad fiduciaria
- Adoptar y validar el modelo usando una herramienta de software de minería de datos.

## Metodología

Para la realización de este trabajo se tuvo en cuenta la metodología Knowledge Discovery in Databases (KDD) [3] y [78] para identificar la tendencia del Máximo Retiro Observado (MRO) y su respectiva predicción, a partir del comportamiento histórico de su serie y de las variables exógenas que inciden en su pronóstico, lo que permitió el reconocimiento de patrones



y tendencias de riesgo de liquidez que no son identificados a simple vista. A continuación se describen cada una de las etapas del KDD desarrolladas en este trabajo para el desarrollo del modelo de riesgo de liquidez basado en técnicas de minería de datos:

- Selección de datos. Se seleccionó el Máximo Retiro Observado (MRO) como la variable objetivo y los atributos que serán incorporados en el modelo de riesgo de liquidez, basado en técnicas de minería de datos.
- Limpieza de datos. Se identificaron los outliers a partir de los diagramas de cajas y bigotes y los ruidos en la información, con el fin de evitar que distorcionen los resultados obtenidos [21].
- Integración de datos. Se incluyeron variables macroeconómicas publicadas por el Banco de la República, el DANE y otros organismos gubernamentales, junto con la información de la cartera colectiva publicada por la SFC y por Asofiduciarias.
- Transformación. Se calcularon las variaciones diarias del MRO para identificar la tendencia que sigue la serie de datos. [8]. Ver la figura 0-2
- Minería de datos. Se aplicaron redes neuronales para predicción del valor del MRO y árboles de clasificación para la generación de reglas.
- Interpretación de resultados. Busca comprender los resultados del análisis y sus implicaciones, y si se requiere retroceder en pasos anteriores para ajustar el modelo.

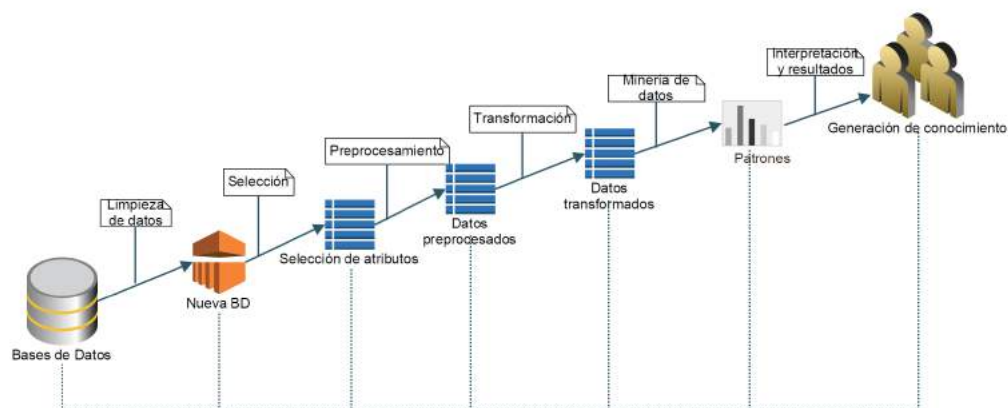


Figura 0-2: Proceso de Minería de Datos

## Limitaciones

Mediante la aplicación de técnicas de minería de datos se busca validar la hipótesis de que estas técnicas permiten clasificar y pronosticar el comportamiento futuro del máximo retiro de una cartera colectiva abierta.

Los siguientes aspectos restringen el alcance de los resultados:

Los registros históricos de variables financieras se ven influenciados por la inflación y por la estacionalidad del comportamiento de la economía, lo que obliga a realizar un tratamiento especial a dichos registros.

El acceso únicamente a información de carácter público limita el alcance del uso de los modelos, ya que existen datos que son de uso interno de cada compañía que no están disponibles, y que contienen información importante para la caracterización de los clientes; sin embargo, con la información financiera y del movimiento diario de las carteras colectivas, publicados por la Superintendencia Financiera y Asofiduciarias, junto con la publicación de los registros diarios, semanales y mensuales de varias variables macroeconómicas del Banco de la República y otras entidades se logró construir una propuesta de modelo de riesgo de liquidez adecuada y de gran aporte para el sector financiero.

## Aportes

La contribución de este trabajo de grado incluye lo siguiente:

- Revisión del estado del arte respecto a las técnicas empleadas en la predicción de series de tiempo del sector financiero
- Construcción de un modelo de caracterización y predicción para riesgo de liquidez de una cartera colectiva, basado en arboles de decisión y redes neuronales.
- Elaboración de una publicación producto de la investigación realizada en técnicas de minería de datos, aplicadas a riesgo de liquidez.

## Estructura del Documento de Tesis

El trabajo consta de ocho secciones. El capítulo introductorio realiza una presentación de la problemática del riesgo de liquidez, limitaciones y alcances, y el propósito u objetivo de realizar este trabajo.

El primer capítulo presenta el estado del arte en riesgo de liquidez, las técnicas aplicadas para su medición y las características principales del negocio fiduciario y de su producto "carteras colectivas". Así como la adopción de técnicas de minería de datos en riesgos financieros, centrandó su atención en las técnicas aplicadas a riesgo financiero, tomándolas como base para desarrollar la aplicación que se realizará para riesgo de liquidez desde minería de datos.

El segundo capítulo realiza un análisis descriptivo de la base de datos y el preprocesamiento realizado, describiendo en detalle, las actividades realizadas en esta etapa: selección de datos, limpieza, transformación y discretización de las variables, paso previo y determinante en el KDD, y base fundamental para la selección de la(s) técnica(s) que se aplican en minería de datos.

El tercer capítulo describe el diseño del modelo de minería de datos compuesto por técnicas como redes neuronales y árboles de decisión. Las primeras técnicas se aplicaron sobre la serie de tiempo del MRO y las segundas sobre el conjunto de datos que incluye la discretización del MRO y las variables macroeconómicas que inciden sobre su comportamiento, con el fin de extraer reglas y patrones de comportamiento del riesgo de liquidez. Al final del capítulo se describe el proceso de entrenamiento y validación realizado sobre el conjunto de datos para cada una de las técnicas de minería de datos aplicadas. A su vez, sus resultados fueron comparados con otras técnicas utilizados en la actualidad por el sector financiero para riesgo de liquidez (ARIMA, regresión logística, entre otras).

El cuarto capítulo se centra en la presentación de las conclusiones obtenidas y posibles trabajos futuros y el quinto capítulo hace referencia a las referencias bibliográficas utilizadas para el desarrollo de este trabajo.

# 1 Marco Teórico y Estado del Arte

En este capítulo se presentan las bases teóricas para el desarrollo del modelo de liquidez basado en técnicas de minería de datos. Así mismo, se incluye una revisión de la literatura relacionada con la aplicación de técnicas y herramientas empleadas para la medición y pronóstico del riesgo de liquidez, y las técnicas de minería de datos que se han adoptado en otros riesgos financieros, dado que la implementación de minería de datos en riesgo de liquidez se encuentra todavía en sus inicios porque su exigencia ha sido relativamente reciente. Se hace énfasis en técnicas de minería de datos como redes neuronales y árboles de decisión, ya que son la base de este trabajo.

## 1.1. Relevancia del riesgo de liquidez en entidades financieras

El riesgo de liquidez se puede definir como "la contingencia de no poder cumplir plenamente, de manera oportuna y eficiente los flujos de caja esperados e inesperados, vigentes y futuros, sin afectar el curso de las operaciones diarias o la condición financiera de la entidad. Esta contingencia (riesgo de liquidez de fondeo) se manifiesta en la insuficiencia de activos líquidos disponibles para ello y/o en la necesidad de asumir costos inusuales de fondeo" [?] [50] y [51]

Resulta crucial para las instituciones financieras e inversionistas conocer el comportamiento de los activos financieros que componen un fondo de inversión o cartera colectiva, con el fin de prever acontecimientos futuros para evitar la materialización de eventos de riesgo y obtener mayores beneficios de inversión que representan una ventaja competitiva en el mercado. [15]

En este sentido, la información juega un papel muy importante, tal como lo afirma el estudio realizado por the Economist Intelligence Unit (EIU), titulado "la mayoría de los ejecutivos toman decisiones económicas poco fiables". Esta afirmación es preocupante para las instituciones e inversionistas, y es una de las conclusiones que se desprende del estudio realizado, en el que se entrevistó a más de 150 ejecutivos, y se descubrió que más de la mitad de ellos consideran que la falta de información los lleva a tomar decisiones inciertas e inadecuadas [16]. Esto se debe a contar con un bajo porcentaje de registros completos en la base de datos

de cada entidad, son de baja calidad e implican largos procesos que retrasan la generación oportuna de información. A esta situación se suma la carencia de un canal de comunicación eficiente al interior y exterior de la organización. [16]

Este hecho se ha trasladado de la esfera corporativa a la economía global, donde el tratamiento inadecuado de la información no ha permitido generar señales en tiempo real del comportamiento de los mercados financieros y han llevado a las entidades a la quiebra a casusa del riesgo de liquidez, donde la variabilidad de los resultados es impredecible en situaciones donde el control y el monitoreo están ausentes o parcialmente aplicados [16]. Asociado a pérdidas por errores operativos (materialización de riesgos operativos), riesgos de mercado (incremento en las tasas de interés, volatilidad en la tasa de cambio, entre otros).

## **1.2. Caracterización del sector fiduciario en Colombia**

Las sociedades fiduciarias son entidades de servicios financieros, constituidas como sociedades anónimas, sujetas a la inspección y a la vigilancia de la Superintendencia Financiera de Colombia (SFC), cuyo objeto es la realización de todos aquellos actos de confianza en virtud de los cuales una persona (fideicomitente) entrega a otra (fiduciario) unos bienes determinados, transfiriéndole o no la propiedad de los mismos, con el objeto de que éste cumpla con ellos una actividad específica, bien sea en provecho del mismo fideicomitente o de un tercero (administración de bienes).

Las fiduciarias administran dos tipos principales de productos: carteras colectivas y negocios fiduciarios. El primero se define como todo mecanismo o vehículo de captación o administración de sumas de dinero u otros activos, integrado con el aporte de un número plural de personas, recursos que serán gestionados de manera colectiva para obtener resultados económicos también colectivos. Estas se clasifican en carteras colectivas abiertas (redención de las participaciones (retiros) se podrá realizar en cualquier momento), cerradas (se caracterizan porque la redención de la totalidad de las participaciones sólo se puede realizar al final del plazo previsto para la duración de la cartera colectiva) y escalonadas (la redención de las participaciones sólo se puede realizar una vez transcurran los plazos que se hayan determinado previamente en el reglamento) [3] (ver Figura1-1 ).

### **1.2.1. Caracterización del producto carteras colectivas en Colombia**

Las carteras colectivas anteriormente se les conocida como fondos de valores, y a partir del año 2007, mediante el Decreto 2175, se les denominó como "carteras colectivas", las cuales son administradas únicamente por Sociedades comisionistas de bolsa de valores, sociedades fiduciarias y sociedades administradoras de inversión. La función principal de las carteras

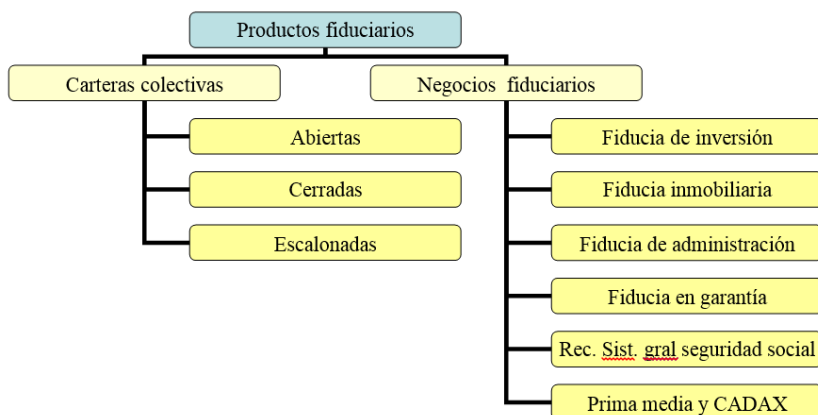


Figura 1-1: Clasificación de productos fiduciarios, según la Superintendencia Financiera de Colombia

colectivas es servir de vehículo de inversión a través del cual se captan o administra dinero u otros activos de diversos inversionistas, que son administrados de manera colectiva [13], facilitando la movilización de recursos a través de acciones, Títulos de Tesorería del Estado *TES*, Certificados de Depósito a Término *CDT*, y otros activos. Posteriormente mediante el Decreto 2555 de 2010 se, establece (Tabla 1-1) se presenta la clasificación de las carteras colectivas en razón a la redención de las participaciones de sus inversionistas vinculados.

Tipo de carácter	Características	Dec. 2555 de 2010
Abierta	Redención de participaciones en cualquier momento, sin perjuicio de que el reglamento establezca un pacto de permanencia mínima, con el cobro de una penalidad	Artículo 3.1.2.1.3.
Cerrada	La redención total de las participaciones sólo puede realizarse al final del plazo previsto para la duración de la cartera colectiva, sin perjuicio de la posibilidad de redención parcial y anticipada y la distribución de mayor valor de la participación	Artículo 3.1.2.1.5.
Escalonada	La redención de las participaciones se realizar una vez transcurran los plazos determinados previamente en el reglamento. El plazo mínimo de redención de sus participaciones no puede ser inferior a treinta días comunes	Artículo 3.1.2.1.4.

Tabla 1-1: Características de Fondos de inversión / cartera colectiva.

Dentro de los principales beneficios señalados por entidades participes de Asofiduciarias, se encuentran: la eficiencia en la administración de las inversiones, disminución de costos,

seguridad en el manejo de los recursos que cuentan con el respaldo del conocimiento y la experiencia del gestor o administrador de los recursos, quien contribuye a determinar el mejor destino de inversión conforme al perfil de riesgo del inversionista. Así mismo, este tipo de instrumento financiero permite la diversificación de los riesgos asociados a la inversión.

### 1.3. Medición del riesgo de liquidez

Históricamente los pioneros en el desarrollo de herramientas de medición de riesgo de liquidez y crediticio han sido las entidades bancarias, quienes cuentan con un marco normativo estructurado para su operación, instrumentos y controles que les permiten tener un mayor control [28]; sin embargo, existen otras entidades financieras diferentes a los bancos, como sociedades fiduciarias, establecimientos de crédito, fondos de inversión, entre otras, quienes han visto reflejado en sus estados financieros y en su operación los efectos del riesgo de liquidez, al ser captadoras de recursos para inversión. Este efecto ha obligado al regulador a extender su política de vigilancia y control de riesgo de liquidez.

Con esta normativa mencionada surgen modelos y técnicas para controlar, hacer seguimiento o evaluación de fenómenos físicos o socio-económicos e inclusión de nuevos actores que forman parte del conjunto del mercado, siendo aspectos incorporados en la metodología particular desarrollada por cada entidad financiera.

#### 1.3.1. Metodología de medición de la Superintendencia Financiera de Colombia para entidades fiduciarias

La SFC establece la necesidad de adoptar un modelo para medir y prever el riesgo de liquidez de una cartera colectiva con la estimación del M<sup>TM</sup>aximo Retiro Probable (MRP) [3], a partir de los siguientes modelos sugeridos por el regulador son: VaR de liquidez paramétrico (VaR-P), VaR de liquidez por simulación histórica (LaR-H) y modelos de regresión lineal o autorregresivos (ARIMA, modelos ARIMA estacionales, modelos de función de transferencia y modelos de heterocedasticidad condicional (GARCH)), y con el cual puede calcular su correspondiente índice de riesgo de liquidez (IRL), el cual se valida mediante pruebas de backtesting. (Ver Figura 1-2)

La diferencia principal en el cálculo del MRP a partir de VaR de liquidez paramétrico (LaR-P), por simulación histórica y modelos lineales regresivos o autorregresivos, radica en que el primero busca estimar el Valor en Riesgo, bajo el supuesto de que la variable movimiento (diferencia entre [Constituciones + Adiciones] y [Cancelaciones + Retiros]) de una cartera colectiva, para un horizonte de tiempo) sigue una distribución normal; mientras que el segundo, no asume ninguna distribución ni supuesto sobre el comportamiento de los datos

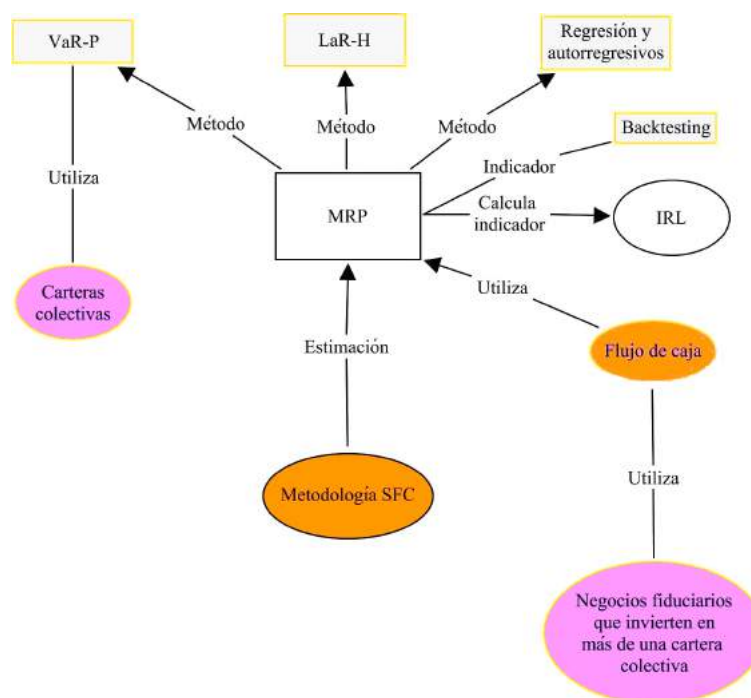


Figura 1-2: Mapa conceptual de Metodologías, Medidas y productos para la medición y control del Riesgo de liquidez

[3]. El tercer modelo requiere identificar factores, estructuras de correlación temporal y de varianza incondicional para determinar la distribución que siguen los datos [2].

### 1.3.2. Medición del riesgo de liquidez a partir del gap de liquidez

Una herramienta clave para estimar el riesgo de liquidez está basada en un análisis de descalces (o gap) entre los flujos entrantes y salientes en diferentes bandas temporales. Se cuantifican los descalces para cada banda y se les suma para obtener un descalce acumulado neto. La posición o gap de liquidez de la entidad suele expresarse a través del ratio de descalce acumulado neto sobre total de depósitos (u otro pasivo), al cual se le aplican límites para este ratio en términos porcentuales [43].

Las políticas y procesos relativos al riesgo de liquidez incluyen una medición continua de esta posición para un horizonte temporal corto, lo cual permite un monitoreo de los requerimientos netos de fondos, facilitándole a la entidad financiera tomar decisiones sobre cómo manejar ciertos rubros, por ejemplo activos con liquidez variable, pasivos con vencimientos no determinados (depósitos a la vista) [43]. En un estudio realizado por el Banco Central de Argentina, se analizó la adopción de recomendaciones, las técnicas y herramientas utilizadas por cada ente supervisor, para medir y controlar el riesgo de liquidez, en el que se manifiesta



una buena y sana gestión del riesgo de liquidez, dada las particulares características con que éste se presenta en las distintas entidades.

Este estudio señala que los organismos encargados de establecer las mejores prácticas en la materia, a la fecha de realizado el estudio, no han establecido una regulación específica ni cuantitativa sino que se han limitado a establecer principios generales que pueden utilizarse como guía en la administración del riesgo. A su vez, la mayoría de los países analizados se han hecho eco de estas recomendaciones, ya sea en la regulación o los manuales de supervisión, en algunos casos dando libertad a las entidades para la aplicación de métodos internos de medición de este riesgo, en tanto que cumplan con los principios allí establecidos. Algunos países crearon una guía para las entidades que aún no han avanzado en la materia y, en otros casos, se consideró importante instaurar una regulación cuantitativa [43].

## **1.4. Técnicas de clasificación y predicción**

Las técnicas utilizadas para predecir desde minería de datos parten de la selección del atributo que se desea proyectar. El atributo usual a predecir es conocido como clase o variable dependiente, mientras que los atributos empleados para realizar predicción se les conoce como variables independientes [7] y [9]. Entre las principales técnicas predictivas está el árbol de decisión y redes neuronales [10], siendo las primeras las más sencillas en términos de técnicas de aprendizaje, lo que constituye un conjunto de condiciones estructuradas dentro de una estructura jerárquica, empleadas para detectar reglas de negocio traducidas en un método de orden natural.

La predicción de eventos se ha consolidado como un aspecto de gran interés para los gerentes de grandes compañías, encargados de la toma de decisiones, quienes se apoyan en otras metodologías que facilitan el direccionamiento estratégico de las organizaciones, en temas financieros, operativos, tecnológicos, entre otros. Aspectos que han tenido gran acogida en el merchandising, a través de la determinación de hábitos de compra, especialmente, en grandes superficies, donde el número de productos, marcas y preferencias de los clientes, son diferentes, y si a ello se añade, el lugar donde opera el supermercado, la economía del país, cambios o volatilidad de la tasa de interés. Entre las técnicas de minería de datos que existen actualmente las más utilizadas son:

### **1.4.1. Series de tiempo**

Las series de tiempo en la minería de datos permiten buscar patrones a partir de grandes cantidades de datos, lo que la ha convertido en un campo de gran importancia y utilidad, siendo los mercados financieros una de las actividades que genera gran cantidad de datos

[11]. Esta técnica se utiliza a partir del comportamiento histórico de los datos, que permite modelar los componentes básicos de la serie, y así se logra hacer predicciones [12]. Este tipo de datos es frecuente en el análisis económico, de ahí el interés por su estudio en carreras universitarias relacionadas con el mundo de la economía y la empresa [5].

El análisis de series temporales permite la descripción de una serie temporal y la predicción de la evolución futura de la misma. El análisis clásico de series temporales se basa en un modelo de descomposición que considera a toda serie temporal integrada por cuatro componentes [31]:

- Tendencia. Representa el movimiento general a largo plazo de una serie temporal.
- Ciclo. Corresponde a las oscilaciones de la variable con una amplitud superior a un año.
- Estacionalidad: oscilaciones de la variable aleatoria de una periodicidad igual o inferior a un año y que se reproducen de manera similar cada año.
- Residuo: movimiento no periódico.

### 1.4.2. Técnicas de regresión

Los métodos de predicción lineales buscan modelar el comportamiento de una serie temporal mediante una función lineal. Entre los que se destacan los modelos ARMA y ARIMA y árboles de decisión.

La predicción de series de tiempo financieras ha adquirido gran atención, especialmente desde la econometría, que han empezado a dedicar mayores esfuerzos a este tipo de información. Este tipo de análisis intenta descubrir la relación entre variables ya sean independientes o dependientes. Por ejemplo, si las ventas son una variable independiente, entonces el beneficio puede ser una variable de pendiente. Mediante el uso de datos históricos de ambas ventas y beneficios, las técnicas lineales o no lineales de regresión pueden producir una curva que permita la predicción de beneficios en el futuro [14].

La metodología Box-Jenkins es una clase de técnicas de regresión y es el principio común en los modelos ARMA y ARIMA, los cuales emplean procesos estocásticos para el análisis de series temporales. Estos modelos han permitido obtener buenas aproximaciones, pero se debe tener cuidado con su aplicación, dado que es un proceso iterativo que depende de la experiencia y criterio del experto que prueba el modelo [15]. A continuación se describen los principales pasos que se deben seguir para construir ambos modelos:

- a) Identificación del modelo para validar la hipótesis principal
- b) Estimación de parámetros
- c) Validación del modelo
- d) Predicción

### ARIMA (p,d,q)

Es un modelo de regresión lineal múltiple integrado de promedio móvil, desarrollado por Box y Jenkins, que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción de series hacia el futuro, es decir, que las estimaciones futuras vienen explicadas por los datos pasados y no por variables independientes. El modelo ARIMA (p,d,q) corresponde a modelos autorregresivos (AR), Integrados (I) y de Medias Móviles (MA) que se representa en la Ecuación 1-1

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^p \phi_i \Delta^d Y_{(t-i)} - \sum_{i=1}^q \phi_i \epsilon_{(t-i)} + \epsilon_t \quad (1-1)$$

En donde d corresponde a las diferencias que son necesarias para convertir la serie original en estacionaria,  $\phi_0, \dots, \phi_p$ , son los parámetros pertenecientes a la parte autorregresiva del modelo,  $\phi_1, \dots, \phi_p$ , los parámetros pertenecientes a la parte de medias móviles del modelo,  $\phi_0$ , es una constante, y  $\epsilon_t$ , es el término de error (llamados también innovaciones).  $Y_t$  denota la serie temporal que depende de una constante C, linealmente de valores pasados de la misma variable y linealmente de una ponderación de errores de ajuste pasados.

Así, la serie de tiempo debe cumplir con los siguientes principios: estacionalidad (media y varianza de los datos analizados sean constantes en el tiempo), la no presencia de una raíz unitaria y la no significancia de autocorrelación muestral (PAC). Seguido de la revisión de la normalidad del residuo, la estacionalidad de los parámetros y la presencia de un coeficiente de desigualdad de Theil. [33] y [45].

### Modelos ARCH y GARCH

La aplicación de estos modelos se basa en modelar los cambios de varianza (o volatilidad); sin embargo no llevan a un mejor pronóstico puntual de la serie, pero pueden llevar a una mejor estimación de la varianza y permiten tener intervalos de predicción más confiables cuya mejor medición del riesgo, motivo por el cual tiene gran aplicación en algunas series de tiempo económicas y financieras.

El modelo ARCH es un modelo autorregresivo condicional heteroscedástico que busca modelar la no homoscedasticidad de la serie, representado mediante la Ecuación 1-2 [33]:

$$\delta_2 = y + (\Delta_x 2_t - 1) \quad (1-2)$$

Dado que la varianza no es constante depende del valor más reciente de la serie y no incluye dentro de su formulación el término de error lo que implica la no definición de un proceso estocástico.

El modelo GARCH o ARCH generalizado de orden (p,q) depende de los últimos p valores y se denota mediante la Ecuación 1-3:

$$\mu_i^2 = y + \sum_p \quad (1-3)$$

### 1.4.3. Técnicas de clasificación

#### árboles de decisión

Un árbol de regresión es una herramienta de predicción para representar modelos de clasificación y regresión, y permiten establecer relaciones entre variables de entrada y salida (generalmente es categórica) [11]. El punto de encuentro entre la relación de variables, se le conoce como nodo, el cual representa una decisión que, a su vez, genera reglas para la clasificación de un conjunto de datos. Los árboles de decisión son de fácil interpretación y admiten atributos de tipo discreto y continuo [10 y 12].

También puede ser interpretado como una "regresión lineal por partes", ya que las hojas de estos árboles contienen distintas rectas de regresión, donde cada nodo representa interno establece una condición o un conjunto de condiciones sobre uno o varios atributos, representando en cada rama saliente el cumplimiento de una de las condiciones. La clasificación se lleva a cabo recorriendo el árbol desde la raíz hasta una de las hojas, siguiendo el camino determinado por el cumplimiento de las condiciones. Por tanto, recorrerá el árbol hasta alcanzar una determinada hoja y la recta de regresión que ésta contenga será el modelo de predicción usado para estimar el atributo deseado. Los problemas de clasificación donde los valores de los atributos son etiquetas han sido resueltos mediante algoritmos de aprendizaje que generan árboles de decisión los cuales han resultado ser robustos, eficientes y simples [29].

En la última década los algoritmos de aprendizaje que generan árboles de decisión han evolucionado para predecir valores de atributos cuando éstos son numéricos, dando paso a los árboles de regresión que son árboles de decisión pero difieren de éstos en que en las hojas del árbol en lugar de tener clases tienen valores, el algoritmo más popular que construye árboles de regresión es conocido como CART [30], en el cual el valor de la predicción en cada hoja es el valor medio de todos los ejemplos de un conjunto de entrenamiento que alcanzan esa

hoja, en el cual el valor de la predicción en cada hoja es el valor medio de todos los ejemplos de un conjunto de entrenamiento que alcanzan esa hoja.

## Redes neuronales

Esta técnica de inteligencia artificial que simula las técnicas empleadas por el cerebro humano para el procesamiento de órdenes, estableciendo redes de comunicación entre variables (neuronas o nodos). Su principal objetivo es el discernimiento de patrones de comportamiento de un gran número de variables y la realización de operaciones de cálculo simultáneo, que permiten generar un sistema que aprende de experiencias anteriores. Así mismo permite el uso de patrones difusos, reajustarse a nuevas condiciones, adaptarse e identificar ruidos en el conjunto de datos o entradas irrelevantes [14]. Una de las características principales de las redes neuronales es su capacidad de trabajar con datos incompletos, incluso paradójicos [12]

Las redes neuronales han tenido un gran éxito en aplicaciones de aprendizaje supervisado y no supervisado. Existe una gran variedad de arquitecturas de redes neuronales, sin embargo las más utilizadas y que han tenido éxito en su aplicación a riesgo financiero ha sido las redes neuronales "Multi Perceptron" (MLP) [41]. Esta técnica supervisada permite extraer información a partir de las relaciones de entrada-salida, que son almacenadas en ecuaciones matemáticas para pronosticar o para la toma de decisiones.

## Estructura de la red neuronal artificial (RNA)

Las RNA están compuestas por unidades de procesamiento simples llamadas neuronas, que se interconectan entre sí para dar lugar a la red, las cuales pueden ir desde la entrada a una neurona, como de una neurona hacia la salida, o de una neurona a otra. Así mismo, la RNA puede presentar bucles o retroalimentación de la información a través de la RNA y retorna a los nodos iniciales como una entrada más. La estructura más común es la multicapa sin retroalimentación, conocida como "feed-forward" y consta de:

- Capa de entrada: consta de los valores de entrada que se introducen en la red para pasar la información a las siguientes capas a través de las conexiones existentes entre ellas. En esta capa tiene unidades que proporcionan información de las variables de entrada.
- Capa(s) oculta(s): Capas intermedias entre la entrada y la salida, y se encargan de establecer las relaciones entre las variables de entrada. Entre mayor sea el número de capas ocultas, mayor complejidad tendrá la RNA, siempre y cuando existan funciones de activación no lineales, ya que si sólo se usan funciones de activación lineales el

número de capas ocultas es irrelevante porque siempre habrá una red equivalente con una sola capa oculta.

La información transita por las conexiones entre capas de neurona a neurona y se evalúa mediante los pesos correspondientes, los cuales se ajustarán durante la fase de entrenamiento para conseguir una mayor adaptación.

### **Tipos de redes neuronales artificiales**

Existe una gran variedad de RNA que se han desarrollado de acuerdo con las necesidades del usuario o analista; sin embargo existe cuatro modelos básicos de RNA a partir de las cuales se desarrollan modelos específicos:

- Redes de propagación hacia adelante (feedforward)
- Redes recurrentes
- Redes estocásticas
- Redes modulares - MLP

### **Redes de propagación hacia adelante (feedforward)**

Es el tipo de RNA más sencillo, donde la información viaja a través de las interconexiones entre neuronas desde la capa de entrada hasta la de salida hacia adelante (sin bucles). Puede estar conformado por una capa oculta (perceptrón simple) o por varias capas ocultas (perceptrón multicapa). Este tipo de RNA resulta ser válido para la predicción de series temporales, al cumplir con el teorema de aproximación universal de las redes neuronales cualquier función que tome un conjunto de números reales y devuelva otro conjunto de números reales se puede aproximar con una precisión arbitrariamente buena por un perceptrón multicapa con una sola capa oculta. [45]

Adicionalmente, existen otros tipos de modelos de RNA, entre los que se encuentran: Red ADALINE (Red Lineal Adaptativa), que adapta sus parámetros en función de las entradas más recientes utilizando la señal de polarización, el peso y una función de suma. La Red con función de base radial (RBF) es un tipo de interpolador multidimensional que emplea funciones de base radial en cada neurona. Por último, los mapas de Kohonen que utilizan un método de aprendizaje no supervisado donde un conjunto de neuronas clasifican un conjunto de datos.

### Redes recurrentes (RNN): Red NARX

La información se transmite desde las neuronas de entrada hacia las de salida con bucles entre medias, donde la información se retroalimenta a neuronas previas, lo que facilita que el proceso no solo dependa de las variables actuales sino que los resultados obtenidos por la red se convierten en entradas o insumos para el sistema. Entre las redes más relevantes de este tipo se encuentran: red Elman, la Hopfield y NARX.

Una RNN es la red tipo NARX (Red Neuronal No Lineal Autorregresiva con Entradas Exógenas). Esta red consta de interconexiones de retroalimentación desde la salida hasta las neuronas de entrada [60 y 61], que puede ser representado mediante la Ecuación 1-4:

$$y(n+1) = f[y(n), \hat{a}, y(n-d_y+1); u(n), u(n-1), \hat{a}, u(n-d_u+1)] \quad (1-4)$$

Donde  $u(n) \in \mathbb{R}$  y  $y(n) \in \mathbb{R}$ , corresponden a las entradas y salidas del modelo, respectivamente, mientras que  $d_u \hat{a} 1 y d_y \hat{a} 1, d_u \hat{a} d_y$ , son la memoria de entrada y la memoria de salida, respectivamente. Esta última utiliza la función sigmoide como elemento de activación para los nodos de la capa oculta y de salida [6], cuya arquitectura se presenta en la Figura 1-3

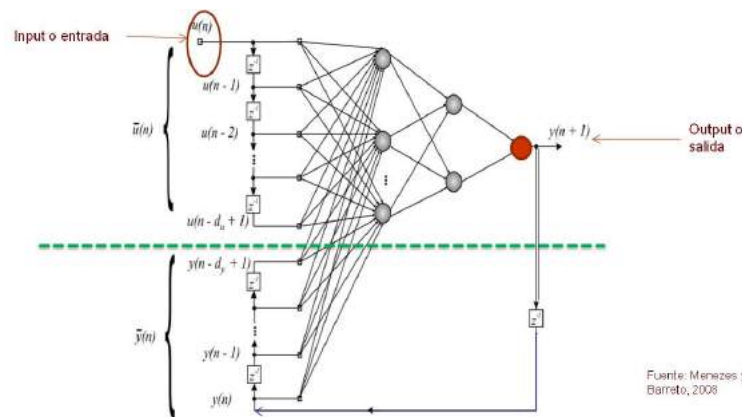


Figura 1-3: Arquitectura de la red neuronal NARX

El modelo NARX está basado en el modelo lineal ARX de difundido uso en series de tiempo. La principal ventaja de la red NARX sobre una red feed-forward es que permite ganar grados de libertad al incluir la predicción de un período como dato en períodos siguientes, resumiendo información de las variables exógenas, permitiendo incluir menos rezagos de las mismas y resumiendo el número de parámetros "a estimar" [40].

## Redes estocásticas

Este tipo de red introduce variaciones aleatorias a la red para sortear mínimos locales que se pueda encontrar durante la fase de entrenamiento. Parte del principio utilizado por los modelos termodinámicos (recocido simulado) donde la red se asemeja a un cuerpo que tiene muchas partículas (neuronas) y cada una con una temperatura aleatoria inestable, pero todo el cuerpo se conserva en equilibrio, buscando alcanzar el estado de mínima energía. Este tipo de red se aplica principalmente al reconocimiento de patrones, procesamiento de conocimiento y resolución de problemas de optimización.

Este modelo de "recocido simulado" menciona que en sus inicios cuando la temperatura es alta, la red es muy inestable y abarca gran parte del espacio de estados posibles. A medida que se va perdiendo temperatura, la red pierde movilidad y tiende a converger hacia un mínimo. Las redes estocásticas más conocidas son: máquina de Boltzmann y la máquina de Cauchy.

## Red neuronal Multilayer Perceptron - MLP

La red neuronal MLP se compone de varias capas y nodos de procesamiento, y el estado de un nodo en un momento determinado está dado por su activación, que por lo general es un número real entre cero y 1, o en el intervalo  $(-1,1)$ . La capa de entrada de la red contiene nodos y las activaciones incorporan valores que representan las características del dominio del problema, tal como se presenta en la Figura 1-4.

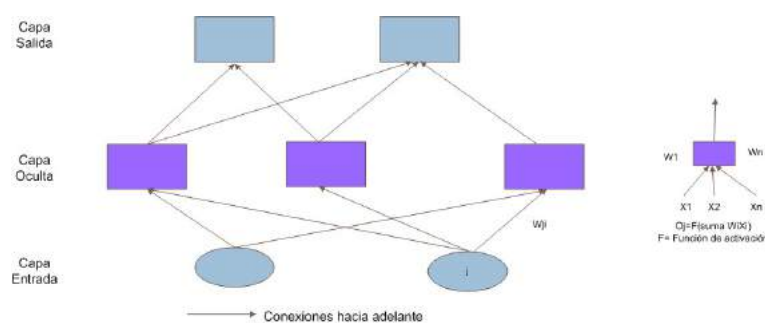


Figura 1-4: Modelo computacional de red neuronal

Los nodos en la capa de salida representan las decisiones tomadas por la red, y entre los nodos de entrada y de salida existe un número de capas ocultas, estas capas permiten transformar el espacio de entrada en un espacio más ajustado para hacer discriminaciones lineales de los nodos de salida.

Esta técnica de red neuronal inicia determinando valores de activación para los nodos de



entrada, los cuales representan un dominio del problema. El proceso continúa a través de las conexiones a los nodos de las capas ocultas y de salida, y el resultado obtenido por la red está determinado por la activación resultante de los nodos de salida, representado por la Ecuación 1-5

$$net(i) = \sum_i w_{ij} a_j + \delta(i) \quad (1-5)$$

Donde,  $w_{ij}$  corresponde al peso del nodo  $j$  al nodo  $i$ ,  $a_j$  es el valor de activación de entrada del nodo y  $\delta(i)$  es el sesgo (bias) para el nodo  $i$ , el cual se puede considerar como un nodo alto o bajo, siendo este el valor de activación antes de que reciba alguna señal de un nodo de red. La activación de un nodo oculto o de salida está determinada por la función de transferencia (algunas veces llamada función de activación).

El aprendizaje de una red neuronal está definido por la topología de la red, la función de activación usada para los nodos de la capa oculta y de salida, los parámetros asociados con las conexiones de la red (i.e pesos, bias). Las reglas generadas por las redes neuronales no generan un resultado definitivo, si no existe un análisis por parte de un experto en la problemática a tratar. Por lo general las redes tienen muchos parámetros con valores reales, los cuales codifican las relaciones entre las características de entrada y las clases de salida.

Por otra parte, en redes multi-capas, estos parámetros pueden representar relaciones no lineales entre las características de entrada y los valores de salida, lo que resulta difícil determinar el efecto de una determinada característica en el valor objetivo.

Estas relaciones no lineales representadas por las unidades ocultas en una red, combinan múltiples características creando un modelo de dependencias entre las características. Comprender las unidades ocultas a menudo es difícil debido a que tienen representaciones distribuidas de aprendizaje. [41]

Las redes neuronales se aplican principalmente de tres maneras:

- Clasificación de los elementos de una muestra o población.
- Las redes neuronales se emplean para la resolución de problemas de clasificación, mediante aprendizaje supervisado y no supervisado. Para el primer caso, los elementos de la muestra o población se asignan a dos o más categorías, las cuales son establecidas con anterioridad; mientras que para el caso del método no supervisado, los elementos de la muestra o población se asignan a dos o más categorías que dependen del número, las características y de los parámetros especificados en la red neuronal artificial.

- La aproximación de funciones
- Memoria asociativa. Establece que la red neuronal artificial otorga un patrón de salida previamente memorizado a partir de variables independientes, lo cual puede emplearse para el procesamiento de información difusa, errónea, incompleta o ruidosa, y para el reconocimiento sonoro de palabras o para el tratamiento visual de caracteres.

### **Reglas de asociación**

Es un método empleado para descubrir relaciones interesantes entre variables pertenecientes a grandes bases de datos. Estas reglas fueron utilizadas por Piatetsky-Shapiro, quien presentó la generación de reglas de asociación en bases de datos que utilizaban diferentes medidas de interés. Posteriormente, Agrawal et al. [14] y [32] Introducen las reglas de asociación para descubrir patrones entre productos registrados en una base de datos del sistema de ventas de los supermercados (POS). Así mismo, la generación de reglas de asociación en grandes conjuntos de datos ha permitido determinar el nivel de correlación entre atributos, contribuyendo al desarrollo de modelos en el área de finanzas, marketing y en ventas al por menor. A continuación se hará un breve recorrido por los métodos más usados en la literatura para la predicción de series temporales, centrando la atención en aquellos que han sido usados para la predicción del comportamiento del mercado bursátil y financiero.

## **1.5. Aplicaciones de minería de datos en riesgo financiero**

Aplicaciones como la estimación de la probabilidad de retiro de clientes, detección de transacciones inusuales de lavado de activos, fraude, modelos predictivos para riesgo de mercado, a partir de la construcción de medidas de dispersión de datos a partir de series de tiempo y la asignaciones de probabilidades de ocurrencia, han permitido la identificación de correlaciones significativas, patrones o tendencias estadísticas. En contraste, el uso de redes neuronales artificiales ha permitido realizar predicciones sobre el comportamiento de series financieras, tal como lo señala Cruz en su investigación, quien obtuvo un mejor desempeño en comparación con un modelo ARIMA utilizado como referencia, ya que las redes neuronales son capaces de identificar relaciones no lineales [15].

### **1.5.1. Riesgo de mercado**

Las principales aplicaciones de Minería de Datos respecto a riesgo de mercado, se han centrado en el pronóstico del mercado de valores caracterizado por ser un sistema dinámico, complejo y no estacional, realizando la estimación de variaciones en la tasa de cambio, quiebras bancarias, la gestión de futuros, entre otros aspectos. Este pronóstico ha permitido

encontrar tendencias de mercado, crear estrategias de inversión, y ser un importante apoyo en la toma de decisiones. [2]

En la literatura se han aplicado diferentes algoritmos para la medición y predicción del riesgo de mercado. Tahseen y Syed presentan una variación de métrica difusa de series de tiempo, en el que se utiliza el enfoque heurístico para definir la frecuencia de densidad y se tiene en cuenta un factor de predicción para pronosticar la tendencia del precio de las acciones. En este trabajo se demuestra que el método propuesto es más preciso respecto a otras técnicas de lógica difusa en series de tiempo. [32]

La aplicación realizada por Muh-Cherng et al., combina el método de regla de filtrado con los árboles de decisión, con el fin de predecir los puntos adecuados de negociación para que el inversionista compre o venda acciones en el mercado de valores. Este ejercicio se realizó con títulos del mercado Taiwanés y Nasdaq. El resultado obtenido demostró que la propuesta de negociación supera los dos métodos presentados [15 y 17].

Estos métodos se han aplicado conjuntamente con series de tiempo, como es el caso del trabajo de Linares [19]. A partir de 214.781 instancias de datos históricos de acciones transadas en la bolsa de Nueva York, compuestas por cuatro atributos (fecha, precio de apertura, de cierre, precio máximo y mínimo), predice comportamientos bursátiles significativos.

El autor desarrolló cuatro fases de experimentación, como son la clasificación, validación de candlesticks, catálogo de patrones y selección de activos. En la primera fase se valida el sistema clasificador difuso propuesto en el proceso de reconocimiento de formas básicas (candlesticks), empleando dos grupos de experimentos:

- Grupo 1: se utilizó el clasificador difuso para un conjunto reducido de reglas y para el conjunto completo de reglas, con el fin de identificar sobre todo el conjunto de datos, el porcentaje de instancias no clasificadas, es decir etiquetadas con la clase undefined. El conjunto reducido tiene las reglas para clasificar Marubozu, Long, Doji y Spinning Top.
- Grupo 2: Emplea el sistema clasificador difuso para reconocer formas básicas en un subconjunto de datos y luego validar la clasificación contra las clases reales (etiquetadas manualmente).

De acuerdo con los resultados obtenidos, se encuentra que de todas las reglas obtenidas durante la experimentación, se presentan valores altos de entropía.

### **1.5.2. Riesgo de lavado de activos y financiación del terrorismo (LA/FT) en detección de fraude**

La detección de operaciones financieras fraudulentas presenta múltiples desafíos debido a la sofisticación de los instrumentos empleados por sus perpetradores.

Estudios realizados para la detección de operaciones de lavado de activos y fraude, como el realizado por Ngai y Hu quienes a partir de la revisión bibliográfica de artículos y documentos publicados entre 1997 y 2008, clasificaron el fraude financiero en cuatro categorías: fraude bancario, de seguros, fraude de valores y materias primas, y otros fraudes relacionados. El análisis de estas categorías de fraude fue realizada a partir de seis técnicas de minería de datos: clasificación, regresión, agrupación, predicción, detección de anomalías, y visualización. El resultado del análisis realizado con base en estas técnicas, señaló que han sido aplicadas en su mayoría a la detección de fraude en seguros, fraude corporativo y con tarjetas de crédito y un escaso análisis respecto al fraude hipotecario, en lavado de activos y en títulos valor.

Esta investigación concluye que las técnicas de mayor aplicación para análisis de fraude son los modelos logísticos, redes neuronales, redes bayesianas, árboles de decisión, las cuales aportan distintas soluciones a los principales problemas en la detección y clasificación de información fraudulenta. [36]

El primero consiste en la combinación de expertos y herramientas que reportan alarmas de eventos inusuales, entre las que se encuentran las listas de control (OFAC, BOE, CIFIN), y el uso de sistemas avanzados. Estos se basan en la aplicación de reglas construidas a partir de sentencias SQL, definidas con la ayuda de expertos para la identificación de transacciones que por lo general realizan depósitos en una cuenta por periodos cortos de tiempo. Así mismo, se emplean técnicas estadísticas de clasificación y regresión de datos históricos para detectar comportamientos anómalos respecto al comportamiento histórico de las transacciones que se realizan en un periodo de tiempo determinado. [22]

El segundo mecanismo se basa en modelos que permiten la identificación de eventos sospechosos, entre los que se encuentran: modelos de datos inusuales basados en técnicas de análisis de clustering y la detección de outliers; y los modelos de relaciones inexplicables, el cual pretende identificar relaciones no obvias entre variables como sexo, edad, profesión, nacionalidad, que no tiene un comportamiento normal, para el cual emplean técnicas de clustering, para encontrar grupos sospechosos y reglas de asociación [22]. En la tabla 1 se resumen las principales técnicas de minería de datos.

Adicionalmente existe un tercer grupo conocido como modelos de características generales,

el cual busca predecir futuros ingresos de operaciones sospechosas, para este caso se emplean técnicas de regresión, árboles de decisión y redes neuronales. (Tabla 1-2) [22]

<b>Tarea</b>	<b>Meta</b>	<b>Técnica de Minería</b>
Encontrar datos inusuales	Detectar registros con valores anormales, múltiples ocurrencias de valores y relaciones entre registros	Análisis de anomalías
Identificar relaciones inexplicables	Determinar perfiles, registros duplicados, registros con referencias de valores anormales, relaciones indirectas entre registros y registros con combinaciones de valores anormales	Análisis de cluster, anomalías y relaciones de asociación
Características generales	Encontrar criterios como reglas y calificación de transacciones sospechosas	Modelos predictivos

**Tabla 1-2:** Técnicas de minería de datos para el lavado de activos

### 1.5.3. Detección de riesgos operativos

La administración efectiva y la identificación correcta de riesgos, resulta ser una tarea complicada en muchas organizaciones [18]; no obstante las instituciones gubernamentales y educativas, se han preocupado por encontrar herramientas que les permitan identificar el nivel de exposición al riesgo. En Turquía se desarrolló un modelo para determinar el perfil financiero para la detección de riesgo operacional [23].

El estudio dirigido por la organización regional industrial de Ankara en el año 2008, se desarrolló en tres fases para medir e identificar factores de riesgo operacional que afectan el correcto desarrollo de las firmas. En la primera fase se tuvo en cuenta el proceso empleado para determinar el nivel de desarrollo de las empresas, cuya base de datos se obtuvo de los estados financieros reportados por las empresas entre los años 1994 y 2006 al Banco Central de Turquía [23]. A partir de los cuales se determinaron los principales ratios financieros. Posteriormente en la segunda fase del estudio se obtuvieron las variables de riesgo operacional de las entidades analizadas.

En la tercera fase, se obtuvieron datos cualitativos y cuantitativos analizados a través de minería de datos de los estados financieros y ratios, estimados en las dos primeras fases empleó un algoritmo de árbol de decisión respecto al total de ratios y riesgos identificados, dado su facilidad de comprensión visual y de interpretación. Teniendo en cuenta el proceso de Knowledge Discovery in Database (KDD) [24]:

Este algoritmo permitió crear perfiles de empresas (seleccionadas mediante un muestreo aleatorio estratificado con un nivel de confianza del 95 % y su respectiva clasificación de acuerdo a su desempeño financiero y así se pudo determinar como señal de alerta aquellos ratios financieros con mayor dificultad financiera para las entidades analizadas.

#### **1.5.4. Aplicaciones en riesgo de crédito**

La mayoría de los estudios realizados para riesgo de crédito se centran en la construcción de un modelo capaz de determinar si se concede o no un crédito a clientes nuevos, es por ello que el estudio realizado por Chen y Xiang va más allá, y propone un modelo híbrido basado en minería de datos. Este modelo se realizó en dos fases, la primera etapa de agrupación permitió la formación de grupos homogéneos y elimino los elementos aislados, mediante el uso de k-means. En la fase de clasificación se realiza un proceso de reasignación de etiquetas para la construcción del modelo de clasificación.

Los resultados experimentales basados en los datos de crédito de un banco local en China, demostró que al elegir un punto de corte, se obtiene mayor precisión en la clasificación de créditos buenos y malos, y de esta forma, es posible seleccionar las estrategias de gestión de riesgos de acuerdo a las características de cada clase. [34]

En el trabajo realizado por Baesens y Setiono se analiza el riesgo de crédito mediante el uso de redes neuronales y la extracción de reglas a partir de los algoritmos Trepan Y Nef-class, han permitido la captura de conocimiento y facilita la toma de decisiones y permite identificar rápidamente los créditos otorgables de las solicitudes de crédito no aceptadas. [35]

Así mismo, se incluyó la aplicación de tablas de decisión DT<sup>TM</sup>s que buscan la generación de conocimiento a partir de la representación gráfica y genera reglas de extracción, que son validadas por un experto. A partir de las técnicas y procesos de generación de conocimiento se concluyó que las redes neuronales y el algoritmo Trepan fueron capaces de extraer conjuntos de reglas precisas y comprensibles, facilitando la consulta y su uso en la práctica empresarial cotidiana. [35]

#### **1.5.5. Riesgo de solvencia y liquidez en flujo de caja**

Estudios realizados por Sergio Cabrales y aplicaciones al área de riesgos para el monitoreo y control del riesgo de liquidez, se desarrolla una metodología estructurada para la cuantificación y monitoreo del riesgo de liquidez en instituciones financieras, utilizando Simulación Montecarlo, Modelo Ho-lee para tasas de interés, árboles binomiales y opciones reales, espe-

cialmente, ante la necesidad de proteger el patrimonio de las instituciones financieras, ante eventuales riesgos propios de la actividad financiera, reflejados a través de la variabilidad de resultados financieros, y con el fin de lograr una eficiente administración del riesgo [31].

Cabe mencionar el aporte realizado por Mesbah y Toony, quienes realizaron una aplicación de algoritmos genéticos en minería de datos, para generar reglas de clasificación a partir de la información generada por los estados financieros, el flujo de caja e información financiera de compañías públicas de Estados Unidos. A partir de estas técnicas de aplicación, se logró descubrir interesantes reglas de clasificación no obvias, anomalías presentes en los estados financieros, concluyendo que las inversiones generadas para el periodo de tiempo analizado no eran tan adecuadas al negocio de las compañías, debido a la baja correlación de las reglas de clasificación identificadas, sin embargo los autores sugieren que existe un amplio campo de estudio todavía inexplorado. [37]

Los aportes realizados por las investigaciones señaladas, contribuyeron a la creación de estudios en Latinoamérica. Estudiantes del departamento de ingeniería industrial de la Universidad de Chile, quienes retomaron los principios formulados por el estudio realizado por Muh et al. para aplicar técnicas de minería de datos con el fin de medir el riesgo crediticio de microempresarios chilenos. Se determinó si un cliente que presenta características  $X$  va a caer en periodo de default o incumplimiento en el corto plazo y no devolverá íntegramente el crédito otorgado [17]. En particular, se buscó encontrar el vector  $X$  que permitiera predecir la probabilidad de ocurrencia de un fenómeno binario y, en este caso si el cliente falla en la devolución del crédito, con un margen de error razonable. Matemáticamente queda expresado en la Ecuación Ecuación 1-6:

$$p(y = falla_x) = f(x) \tag{1-6}$$

Donde  $p(y=falla-x)$  corresponde a la probabilidad que la empresa caracterizada porque  $X$  no pueda cumplir sus compromisos financieros y  $f(X)$  corresponde a una función matemática que aproxima la probabilidad a partir de los datos disponibles. [26] y se incluyeron como datos de entrada las bases de datos internas y externas, variables e indicadores derivados (indicadores financieros de ingresos y deuda).

A partir de la realización de un preprocesamiento de los datos, se analizó el nivel de contribución de cada una de las variables en la fase de clasificación, mediante un análisis multivariado realizado sobre el total de las variables. Aquellas variables que no contribuyeran significativamente se utilizaron en un árbol de decisión sin poda, en el que se incluyeron las posibles relaciones entre variables que presentan algún tipo de comportamiento. A su vez se seleccionó

como técnica de clasificación, la regresión logística para modelación de riesgo de crédito, debido a su buen desempeño predictivo.

De acuerdo a los resultados obtenidos por este modelo, se encontró que aunque las técnicas de medición para el conjunto de datos analizado siguen entregando resultados favorables, pero en la selección y construcción de variables resulta ser confusa su determinación, situación que es resuelta con minería de datos en el preprocesamiento de los datos. [27]

## **1.6. Resumen del capítulo**

En este capítulo se presentó el marco teórico sobre el cual se desarrolló y diseño el modelo de predicción de riesgo de liquidez para una entidad fiduciaria. En principio se planteó el problema que presentan actualmente las entidades financieras para medir y predecir el riesgo financiero y las técnicas que han sido empleadas por dichas instituciones para la gestión del riesgo, sus limitaciones y alcance que han tenido en el tratamiento de grandes bases de datos o registros transaccionales. Para ello se efectuó una recopilación de la literatura más importante sobre el tema. Finalmente se aborda el tema del proceso de aprendizaje, realizado a partir de series de tiempo y redes neurales, y la utilidad de los árboles de decisión en la caracterización de los aspectos más importantes, y que pueden incidir en el comportamiento futuro del riesgo financiero de una organización.



## 2 Análisis descriptivo y Pre-procesamiento

Este capítulo presenta el análisis descriptivo de los datos transaccionales de una cartera colectiva de una sociedad fiduciaria, la caracterización de los atributos o variables y testing estadísticos que son utilizados para validar las distribuciones de los datos, que serán tenidas en cuenta en el siguiente capítulo para el modelo de minería de datos propuesto. Adicionalmente se describen los métodos de limpieza y transformación de datos por medio de los cuales se excluyeron valores atípicos y ruidos que distorsionaban la distribución de los datos.

En esta fase, se seguirá lo mencionado por la metodología del KDD (Figura 0-2), en lo que respecta al análisis exploratorio siguiendo los cuatro pasos de este proceso sobre la base de datos: preparación, selección, pre-procesamiento y transformación de datos, el cual permitirá conocer la distribución y comportamiento de los datos, identificación de datos atípicos, ruidos, frecuencias, y la identificación de variables o atributos potencialmente relevantes para el modelo a estimar desde minería de datos que será presentado en el desarrollo del trabajo.

### 2.1. Conjunto de datos

La información utilizada para el desarrollo de este trabajo fue obtenida de fuentes de información abierta, publicada en las páginas de internet de la Superintendencia Financiera de Colombia, Asofiduciarias, Banco de la República e informes periódicos de la cartera colectiva objeto de estudio.

Se cuenta con un registro histórico de 571 registros diarios, 17 variables o atributos y una clase, que corresponde al máximo retiro observado (MRO), exceptuando fines de semana y días festivos del periodo comprendido entre los años 2009 y 2011. Las variables corresponden a información propia del fondo de inversión y de las principales variables macroeconómicas.

Esta información se complementó con: a) la ficha técnica de la cartera colectiva, que incluye el número de unidades en circulación, valor de la unidad, el número de suscriptores; y con b) el comportamiento histórico de indicadores macroeconómicos que son considerados por expertos del mercado como determinantes en el comportamiento de una cartera colectiva

[46], motivo por el cual se tuvieron en cuenta las variables descritas en la (Tabla 2-6).

<b>Nombre de la variable</b>	<b>Descripción de la variable</b>
MRO	Máximo Retiro Observado
VFCt	Valor del fondo al cierre del día antes de rendimientos
Rentabilidad diaria	Índice que mide la relación entre la utilidad y la inversión que se utilizaron para obtener dichas utilidades.
Coltes	Títulos de Deuda Pública - TES. Son títulos de renta fija expedidos por el Gobierno Nacional para obtener financiación o apalancamiento financiero [47].
Coltes CP	Títulos de Deuda Pública de corto plazo
Coltes LP	Títulos de Deuda Pública de largo plazo
Coltes UVR	Títulos de Deuda Pública indexados a la tasa de la unidad
IGBC	índice General de la Bolsa de Valores de Colombia
COLCAP	Es un indicador que refleja las variaciones de los precios de las 20 acciones más líquidas de la Bolsa de Valores de Colombia (BVC), donde el valor de la Capitalización Bursátil Ajustada de cada compañía determina su nivel de ponderación.
TRM	Tasa representativa del mercado (TRM), es la cantidad de pesos colombianos por un dólar de los Estados Unidos. La Superfinanciera la calcula con base en las operaciones de compra y venta de divisas entre intermediarios financieros que transan en el mercado cambiario colombiano, con cumplimiento el mismo día cuando se realiza la negociación de las divisas [47].
IPC	índice de Precios al Consumidor (IPC). Es un indicador sobre el cual se acumulan a partir de un periodo base las variaciones promedio de los precios de los bienes y servicios consumidos por los hogares del país en un periodo de tiempo. [49]

Nombre de la variable	Descripción de la variable
PIB	El Producto Interno Bruto (PIB), mide el resultado final de la actividad productiva (unidades de producción) de los residentes del país. Se mide desde el punto de vista del valor agregado, de la demanda final o las utilidades finales de los bienes y servicios y de los ingresos primarios distribuidos por las unidades de producción residentes. [49]
DTF	Tasa de referencia (DTF). Promedio de las tasas de captación de los depósitos a 90 días de establecimientos bancarios, corporaciones financieras y compañías de financiamiento comercial [47]
TCC	Tasa de interés estimada a partir del promedio ponderado semanal por monto, de las tasas promedios de captación diarias de los CDT's a 90 días, pagadas por las corporaciones financieras privadas. Es calculada por el Banco de la República con la información suministrada por la Superintendencia Bancaria
IBR	índice Bancario de Referencia (IBR). Tasa de interés de referencia de corto plazo en pesos colombianos, que representa el precio al que los bancos captan recursos en el mercado monetario [47]
Tasa desempleo	Expresa el nivel de desocupación entre la población económicamente activa [48].
ICC	índice de confianza del consumidor (ICC). Promedio simple de los balances de preguntas que cuestionan la opinión de los hogares, tanto de las condiciones actuales como de perspectivas a futuro del país.
Día	Día de la semana hábil en el que se realiza transacciones (lunes, martes, miércoles, jueves y viernes)
Semana	Número de la semana del mes al que pertenece la transacción realizada (MRO)

**Tabla 2-1:** Descripción de las variables para el modelo de minería de datos

### 2.1.1. Análisis de tendencias

Se presenta un alto grado de volatilidad de la serie del MRO durante el periodo de análisis (2009-2011), lo que sugiere revisar el comportamiento diario e intentar encontrar un patrón de comportamiento. El primer paso es la representación gráfica de las frecuencias de los movimientos diarios (adiciones, retiros) de la cartera colectiva estudiada, mediante el uso de un histograma con el fin de identificar gráficamente cuáles son los días en los cuales se presentan mayor número de operaciones, y con ello, identificar patrones de comportamiento de la serie de los retiros observados (ver Figura 2-1 ).

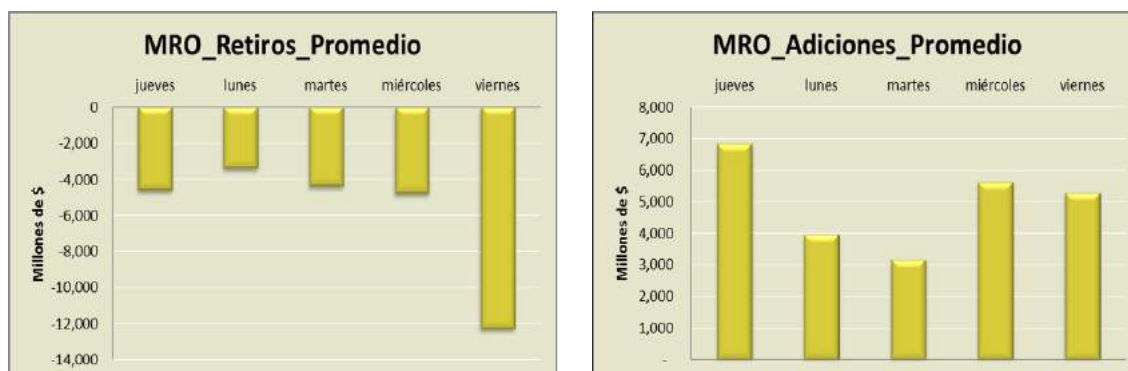
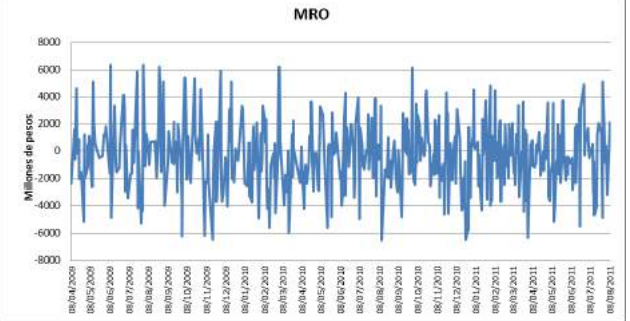

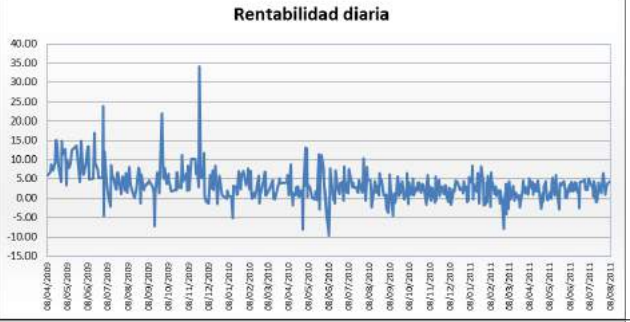
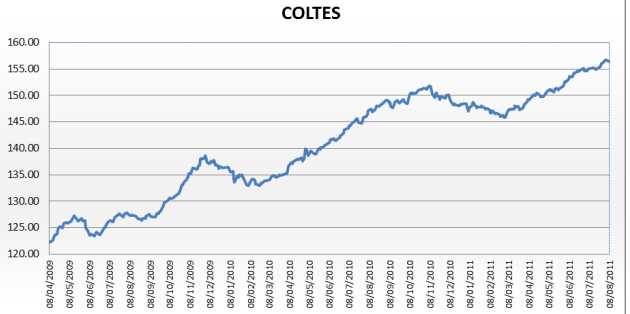




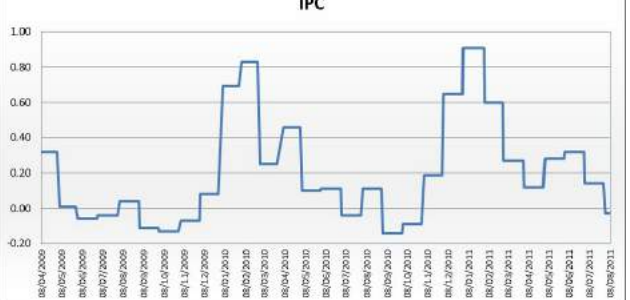
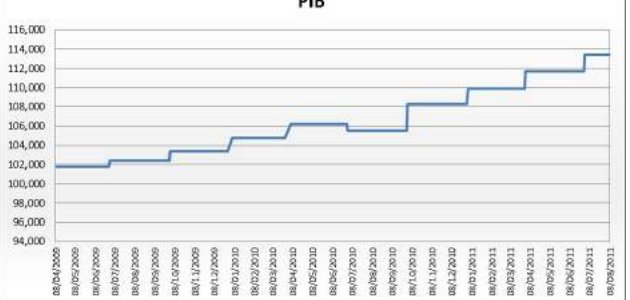
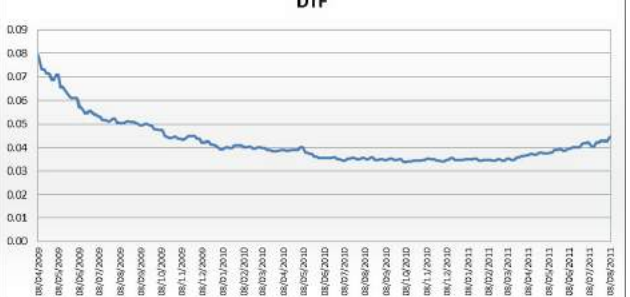
Figura 2-1: Histograma de Retiros semanales del MRO

Según los datos presentados en la Figura 2-1, los días en los cuales mayor cantidad de retiros se realizan los viernes, seguidos por los jueves y miércoles; mientras que en el día lunes se realizan menos operaciones, a las presentadas en otros días de la semana; mientras que los días en los cuales mayor cantidad de adiciones se realizan los jueves, seguidos por los miércoles y viernes, pero el día martes se realizan menos operaciones.

A continuación se presenta el análisis descriptivo de cada una de las variables consideradas para el modelo de minería de datos en la Tabla 2-2

Variable	Tendencia
Máximo retiro observado	 <p><b>MRO</b></p> <p>Millones de pesos</p>
Valor del fondo al cierre del día (VFct)	 <p><b>Valor del fondo al cierre del día (VFct)</b></p> <p>Millones de pesos</p>
Rentabilidad diaria	 <p><b>Rentabilidad diaria</b></p>
Tasa de Títulos de Deuda Pública (COLTES)	 <p><b>COLTES</b></p>

Variable	Tendencia
CP	<p>The chart 'COLTES CP' displays a line graph with the y-axis ranging from 120.00 to 155.00 and the x-axis showing dates from 08/04/2009 to 08/08/2011. The data shows a steady increase with some minor fluctuations, starting at approximately 120.00 and ending near 150.00.</p>
COLTES LP	<p>The chart 'COLTES LP' displays a line graph with the y-axis ranging from 120.00 to 170.00 and the x-axis showing dates from 08/04/2009 to 08/08/2011. The data shows a general upward trend with significant volatility, starting at approximately 120.00 and ending near 165.00.</p>
COLTES UVR	<p>The chart 'COLTES UVR' displays a line graph with the y-axis ranging from 120.00 to 155.00 and the x-axis showing dates from 08/04/2009 to 08/08/2011. The data shows a consistent upward trend with minor fluctuations, starting at approximately 120.00 and ending near 155.00.</p>
IGBC	<p>The chart 'IGBC' displays a line graph with the y-axis ranging from 7,500.00 to 17,500.00 and the x-axis showing dates from 08/04/2009 to 08/08/2011. The data shows a strong upward trend with significant volatility, starting at approximately 8,500.00 and ending near 13,500.00.</p>

Variable	Tendencia
COLCAP	 <p>The COLCAP chart shows a fluctuating but generally increasing trend. It starts at approximately 900.00 in early 2009, rises to about 1,300.00 by mid-2009, reaches a peak of nearly 2,000.00 in late 2010, and then declines to around 1,500.00 by early 2011.</p>
TRM	 <p>The TRM chart displays a downward trend with significant volatility. It begins at approximately 2,400 in early 2009, drops to about 2,000 by mid-2009, and continues to decline with fluctuations, reaching a low of around 1,700 by early 2011.</p>
IPC	 <p>The IPC chart shows a step-like pattern. It starts at approximately 0.30 in early 2009, drops to near 0.00, and then exhibits several sharp increases, peaking at about 0.90 in late 2010, before ending around 0.10 in early 2011.</p>
PIB	 <p>The PIB chart illustrates a steady, step-wise increase. It starts at approximately 100,000 in early 2009 and shows consistent growth, reaching about 114,000 by early 2011.</p>
DTF	 <p>The DTF chart shows a general downward trend. It starts at approximately 0.08 in early 2009 and gradually declines to about 0.04 by early 2011, with some minor fluctuations.</p>

Las variables analizadas presentan un comportamiento estacionario, como la serie del valor del fondo, quien presenta una tendencia decreciente de largo plazo, y se hace más evidente en abril de 2011, pero su tendencia cambia y se estabiliza en lo restante del año 2011. Mientras que la rentabilidad diaria presenta una tendencia relativamente constante, sin embargo se han presentado crecimientos superiores a la media e inferiores. A finales del mes de noviembre de 2009 se registró una rentabilidad cercana al 34 % respecto al día anterior, y en junio de 2010 se registró un pérdida del 9 %.

El grupo de índices del mercado de renta fija, COLTES (Títulos de deuda pública), COLTES CP (Títulos de deuda pública de corto plazo), COLTES LP (Títulos de deuda pública de largo plazo) y COLTES UVR (Títulos de deuda pública indexados a la Unidad de Valor Real), presentan una tendencia creciente entre los años 2009 y 2011 del 16 %, aproximadamente.

El IGBC y el COLCAP, índices generales del mercado de valores colombiano, presentan una tendencia creciente de largo plazo durante todo el periodo de análisis, con un repunte importante en noviembre de 2010. Posteriormente, se registró un leve decrecimiento.

La Tasa Representativa del Mercado (TRM) sigue una tendencia decreciente a lo largo del periodo de análisis; sin embargo para el primer semestre del año 2010 (1S10), se observa un cambio de tendencia, registrando un leve crecimiento, el cual se repite en el siguiente año, aunque con menor fuerza.

El índice de Precios al Consumidor presenta un comportamiento ciclico con cambios de tendencia en periodos de tiempo similares, que se repiten para el siguiente periodo con la publicación de este índice por parte del DANE, lo que genera un ajuste en los principales indicadores de la canasta básica.

El Producto Interno Bruto entre los años 2009 y 2011 presentó un comportamiento creciente, al pasar de 101,752 mil millones a 112,752 mil millones. La serie sigue un comportamiento ascendente durante el periodo de análisis.

La DTF presentó un comportamiento a la baja como consecuencia de la política monetaria expansiva adoptada por el gobierno en el año 2008, y que tuvo repercusión en 2009. Su estabilización vino dada a partir de la mitad del año 2010, la cual presenta un leve crecimiento a mediados del año 2011. La serie del TCC tuvo una tendencia decreciente entre el año 2009, y cambia su comportamiento al permanecer constante el resto del tiempo.

El indicador bancario de referencia para Colombia presentó una tendencia decreciente durante el año 2009 y se estabilizó en lo corrido del año 2010, gracias al incremento en la demanda



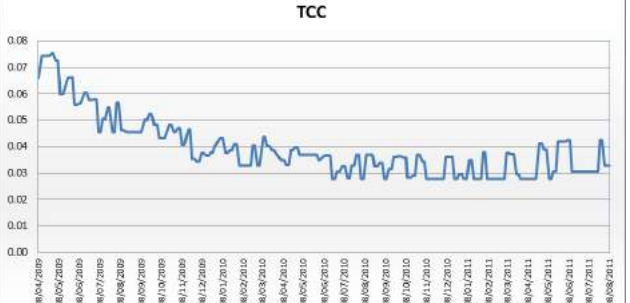



Variable	Tendencia
TCC	 <p>The chart for TCC shows a general downward trend. It starts at approximately 0.075 in early 2009, drops to about 0.05 by mid-2009, and continues to decline with some fluctuations to around 0.035 by late 2011.</p>
IBR	 <p>The chart for IBR shows a downward trend. It starts at approximately 0.07 in early 2009, drops to about 0.045 by mid-2009, and continues to decline to around 0.03 by late 2011.</p>
Tasa de Desempleo	 <p>The chart for Tasa de Desempleo shows fluctuations. It starts at approximately 0.54 in early 2009, peaks at about 0.565 in mid-2009, and then fluctuates between 0.53 and 0.57 throughout the period.</p>
ICC	 <p>The chart for ICC shows an upward trend. It starts at approximately -0.15 in early 2009, rises to about 0.05 by mid-2009, and continues to rise to around 0.25 by late 2011.</p>

Tabla 2-2: Comportamiento tendencial de las variables

de productos financieros indexados a la IBR.

La tasa de desempleo se incremento gradualmente desde el año 2010, principalmente por el aumento de la tasa de informalidad y reducción del empleo formal, lo que ha llevado a la población en edad productiva a buscar nuevas fuentes de recursos paa su propio sustento. El índice de precios al consumidor ha seguido una tendencia al alza entre los años 2009 y 2011, lo que refleja un aumento del nivel de precios.

## **2.2. Medidas estadísticas de las variables**

Se utilizarán test estadísticos que permitan identificar la distribución que siguen los datos por cada variable, su comportamiento a partir de sus principales medidas estadísticas. En la Tabla **2-3** se presentan las principales estadísticas calculadas para cada una de las variables seleccionadas. Las variables analizadas no tienen valores perdidos, porque son reportadas diariamente por la entidad financiera para cerrar el portafolio de inversión, y las variables macroeconómicas seleccionadas para la construcción del modelo de minería de datos, son calculadas y publicadas por lo general por entidades gubernamentales, quienes reportan periódicamente dicha información.

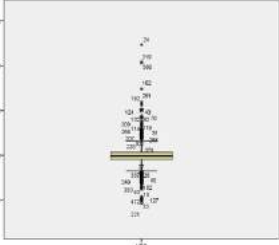
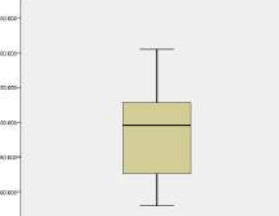
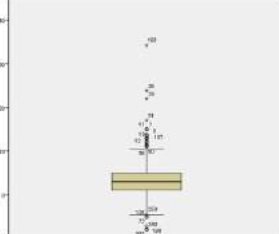
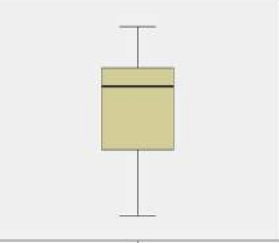
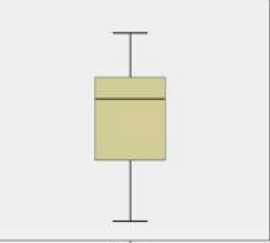
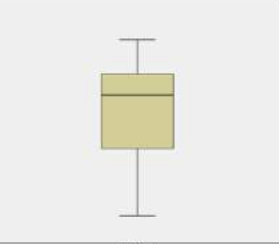
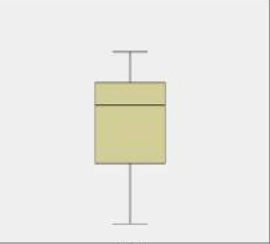
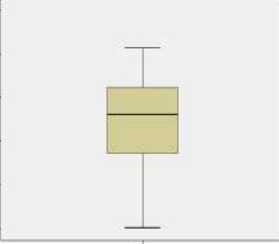
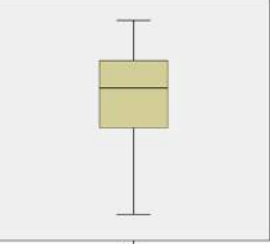
Tabla 2-3: Medidas estadísticas de las variables

Variable	Mínimo	Máximo	Media		Desviación estándar
	Estadístico	Estadístico	Estadístico	Error estándar	Estadístico
VFct	279748099977	504553828507	382296089378	2592987756	55431746991
Rent.diaria	-10	34	3,39	0,189	4,049
Coltes	122,28	156,66	141,6041	0,442	9,446
Coltes CP	122,4	150,1	138,6004	0,358	7,658
Coltes LP	121,63	164,37	144,7607	0,549	11,744
Coltes UVR	121,26	152,19	139,0597	0,412	8,817
IGBC	8041,74	16278,88	12867,6226	93,557	2000,023
COLCAP	878,58	1942,37	1516,4101	12,027	257,104
TRM	1748	2417	1919,28	5,751	122,943
IPC	-0,14	0,91	0,221	0,014	0,300
PIB	101752	113384	106724,101	160,280	3426,404
DTF	0,03	0,08	0,041	0,000	0,009
TCC	0,03	0,08	0,0381	0,001	0,011
IBR	0,03	0,07	0,0355	0,000	0,008
Tasa desemp.	0,53	0,57	0,5524	0,001	0,011
ICC	-0,2	0,39	0,1673	0,006	0,122
MRO	-6526925000	6366000000	-360960363	118565064	2534631575

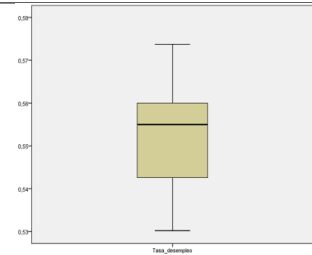
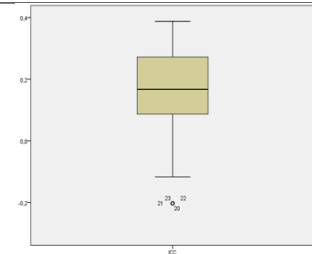
## 2.3. Preprocesamiento

### 2.3.1. Detección de valores atípicos y limpieza de datos

En esta fase no se encontraron datos faltantes ni incompletos porque son datos publicados y validados por instituciones gubernamentales, para el caso de las variables macroeconómicas, y se presentó el mismo caso, para el MRO, el valor del fondo (VFct) y la rentabilidad diaria del fondo, ya que es información transmitida y validada por la SFC. No obstante, se identificaron valores atípicos que distorsionaban la distribución de los datos, mediante el uso de diagramas de cajas y bigotes, los cuales fueron descartados de la serie de datos de cada variable para evitar ruido en los resultados, pasando de un conjunto de datos de 571 a 520 registros. En la Tabla 2-4 se presentan los resultados obtenidos con Spss (software estadístico) para la identificación de valores atípicos por cada variable.

Variable	Valores atípicos																																							
MRO		<table border="1"> <thead> <tr> <th colspan="3">Valores extremos</th> </tr> <tr> <th></th> <th>Número del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">MRO Mayores</td> <td>1</td> <td>24</td> <td>61592</td> </tr> <tr> <td>2</td> <td>566</td> <td>52057</td> </tr> <tr> <td>3</td> <td>219</td> <td>51454</td> </tr> <tr> <td>4</td> <td>102</td> <td>37020</td> </tr> <tr> <td>5</td> <td>291</td> <td>29680</td> </tr> <tr> <td rowspan="5">Menores</td> <td>1</td> <td>221</td> <td>-26925</td> </tr> <tr> <td>2</td> <td>33</td> <td>-25722</td> </tr> <tr> <td>3</td> <td>127</td> <td>-25508</td> </tr> <tr> <td>4</td> <td>10</td> <td>-25070</td> </tr> <tr> <td>5</td> <td>101</td> <td>-24473</td> </tr> </tbody> </table>	Valores extremos				Número del caso	Valor	MRO Mayores	1	24	61592	2	566	52057	3	219	51454	4	102	37020	5	291	29680	Menores	1	221	-26925	2	33	-25722	3	127	-25508	4	10	-25070	5	101	-24473
Valores extremos																																								
	Número del caso	Valor																																						
MRO Mayores	1	24	61592																																					
	2	566	52057																																					
	3	219	51454																																					
	4	102	37020																																					
	5	291	29680																																					
Menores	1	221	-26925																																					
	2	33	-25722																																					
	3	127	-25508																																					
	4	10	-25070																																					
	5	101	-24473																																					
VFCt		<table border="1"> <thead> <tr> <th colspan="3">Valores extremos</th> </tr> <tr> <th></th> <th>Número del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">Valor de fondo al cierre del día (VFCt) - columna de precios</td> <td>Mayor 1</td> <td>81</td> <td>5.E+11</td> </tr> <tr> <td>2</td> <td>92</td> <td>5.E+11</td> </tr> <tr> <td>3</td> <td>92</td> <td>5.E+11</td> </tr> <tr> <td>4</td> <td>89</td> <td>5.E+11</td> </tr> <tr> <td>5</td> <td>80</td> <td>5.E+11</td> </tr> <tr> <td rowspan="5">Menor</td> <td>1</td> <td>454</td> <td>3.E+11</td> </tr> <tr> <td>2</td> <td>455</td> <td>3.E+11</td> </tr> <tr> <td>3</td> <td>456</td> <td>3.E+11</td> </tr> <tr> <td>4</td> <td>451</td> <td>3.E+11</td> </tr> <tr> <td>5</td> <td>457</td> <td>3.E+11</td> </tr> </tbody> </table>	Valores extremos				Número del caso	Valor	Valor de fondo al cierre del día (VFCt) - columna de precios	Mayor 1	81	5.E+11	2	92	5.E+11	3	92	5.E+11	4	89	5.E+11	5	80	5.E+11	Menor	1	454	3.E+11	2	455	3.E+11	3	456	3.E+11	4	451	3.E+11	5	457	3.E+11
Valores extremos																																								
	Número del caso	Valor																																						
Valor de fondo al cierre del día (VFCt) - columna de precios	Mayor 1	81	5.E+11																																					
	2	92	5.E+11																																					
	3	92	5.E+11																																					
	4	89	5.E+11																																					
	5	80	5.E+11																																					
Menor	1	454	3.E+11																																					
	2	455	3.E+11																																					
	3	456	3.E+11																																					
	4	451	3.E+11																																					
	5	457	3.E+11																																					
Rentabilidad diaria		<table border="1"> <thead> <tr> <th colspan="3">Valores extremos</th> </tr> <tr> <th></th> <th>Número del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">Rentabilidad diaria Mayor</td> <td>1</td> <td>183</td> <td>34</td> </tr> <tr> <td>2</td> <td>36</td> <td>24</td> </tr> <tr> <td>3</td> <td>78</td> <td>22</td> </tr> <tr> <td>4</td> <td>31</td> <td>-17</td> </tr> <tr> <td>5</td> <td>7</td> <td>18</td> </tr> <tr> <td rowspan="5">Menor</td> <td>1</td> <td>198</td> <td>-18</td> </tr> <tr> <td>2</td> <td>180</td> <td>-8</td> </tr> <tr> <td>3</td> <td>382</td> <td>-8</td> </tr> <tr> <td>4</td> <td>72</td> <td>-7</td> </tr> <tr> <td>5</td> <td>126</td> <td>-8</td> </tr> </tbody> </table>	Valores extremos				Número del caso	Valor	Rentabilidad diaria Mayor	1	183	34	2	36	24	3	78	22	4	31	-17	5	7	18	Menor	1	198	-18	2	180	-8	3	382	-8	4	72	-7	5	126	-8
Valores extremos																																								
	Número del caso	Valor																																						
Rentabilidad diaria Mayor	1	183	34																																					
	2	36	24																																					
	3	78	22																																					
	4	31	-17																																					
	5	7	18																																					
Menor	1	198	-18																																					
	2	180	-8																																					
	3	382	-8																																					
	4	72	-7																																					
	5	126	-8																																					
Títulos de Deuda Pública - COLTES																																								
																																								
IGBC y COLCAP																																								

Variable	Valores atípicos																																									
PIB		<table border="1"> <thead> <tr> <th colspan="3">Valores extremos</th> </tr> <tr> <th></th> <th></th> <th>Numero del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">PIB</td> <td rowspan="5">Mayores</td> <td>1</td> <td>440</td> <td>113384.00</td> </tr> <tr> <td>2</td> <td>441</td> <td>113384.00</td> </tr> <tr> <td>3</td> <td>442</td> <td>113384.00</td> </tr> <tr> <td>4</td> <td>443</td> <td>113384.00</td> </tr> <tr> <td>5</td> <td>444</td> <td>113384.00</td> </tr> <tr> <td rowspan="5">Menores</td> <td>1</td> <td>35</td> <td>101752.00</td> </tr> <tr> <td>2</td> <td>34</td> <td>101752.00</td> </tr> <tr> <td>3</td> <td>33</td> <td>101752.00</td> </tr> <tr> <td>4</td> <td>32</td> <td>101752.00</td> </tr> <tr> <td>5</td> <td>31</td> <td>101752.00</td> </tr> </tbody> </table>	Valores extremos					Numero del caso	Valor	PIB	Mayores	1	440	113384.00	2	441	113384.00	3	442	113384.00	4	443	113384.00	5	444	113384.00	Menores	1	35	101752.00	2	34	101752.00	3	33	101752.00	4	32	101752.00	5	31	101752.00
Valores extremos																																										
		Numero del caso	Valor																																							
PIB	Mayores	1	440	113384.00																																						
		2	441	113384.00																																						
		3	442	113384.00																																						
		4	443	113384.00																																						
		5	444	113384.00																																						
Menores	1	35	101752.00																																							
	2	34	101752.00																																							
	3	33	101752.00																																							
	4	32	101752.00																																							
	5	31	101752.00																																							
IPC		<table border="1"> <thead> <tr> <th colspan="3">Valores extremos</th> </tr> <tr> <th></th> <th></th> <th>Numero del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">IPC</td> <td rowspan="5">Mayores</td> <td>1</td> <td>322</td> <td>.91</td> </tr> <tr> <td>2</td> <td>323</td> <td>.91</td> </tr> <tr> <td>3</td> <td>324</td> <td>.91</td> </tr> <tr> <td>4</td> <td>325</td> <td>.91</td> </tr> <tr> <td>5</td> <td>326</td> <td>.91</td> </tr> <tr> <td rowspan="5">Menores</td> <td>1</td> <td>296</td> <td>-.14</td> </tr> <tr> <td>2</td> <td>295</td> <td>-.14</td> </tr> <tr> <td>3</td> <td>294</td> <td>-.14</td> </tr> <tr> <td>4</td> <td>293</td> <td>-.14</td> </tr> <tr> <td>5</td> <td>292</td> <td>-.14</td> </tr> </tbody> </table>	Valores extremos					Numero del caso	Valor	IPC	Mayores	1	322	.91	2	323	.91	3	324	.91	4	325	.91	5	326	.91	Menores	1	296	-.14	2	295	-.14	3	294	-.14	4	293	-.14	5	292	-.14
Valores extremos																																										
		Numero del caso	Valor																																							
IPC	Mayores	1	322	.91																																						
		2	323	.91																																						
		3	324	.91																																						
		4	325	.91																																						
		5	326	.91																																						
Menores	1	296	-.14																																							
	2	295	-.14																																							
	3	294	-.14																																							
	4	293	-.14																																							
	5	292	-.14																																							
DTF		<table border="1"> <thead> <tr> <th colspan="3">Valores extremos</th> </tr> <tr> <th></th> <th></th> <th>Numero del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">DTF</td> <td rowspan="5">Mayores</td> <td>1</td> <td>1</td> <td>.08</td> </tr> <tr> <td>2</td> <td>2</td> <td>.07</td> </tr> <tr> <td>3</td> <td>3</td> <td>.07</td> </tr> <tr> <td>4</td> <td>4</td> <td>.07</td> </tr> <tr> <td>5</td> <td>5</td> <td>.07</td> </tr> <tr> <td rowspan="5">Menores</td> <td>1</td> <td>272</td> <td>.03</td> </tr> <tr> <td>2</td> <td>271</td> <td>.03</td> </tr> <tr> <td>3</td> <td>270</td> <td>.03</td> </tr> <tr> <td>4</td> <td>269</td> <td>.03</td> </tr> <tr> <td>5</td> <td>268</td> <td>.03</td> </tr> </tbody> </table>	Valores extremos					Numero del caso	Valor	DTF	Mayores	1	1	.08	2	2	.07	3	3	.07	4	4	.07	5	5	.07	Menores	1	272	.03	2	271	.03	3	270	.03	4	269	.03	5	268	.03
Valores extremos																																										
		Numero del caso	Valor																																							
DTF	Mayores	1	1	.08																																						
		2	2	.07																																						
		3	3	.07																																						
		4	4	.07																																						
		5	5	.07																																						
Menores	1	272	.03																																							
	2	271	.03																																							
	3	270	.03																																							
	4	269	.03																																							
	5	268	.03																																							
TCC		<table border="1"> <thead> <tr> <th colspan="3">Valores extremos</th> </tr> <tr> <th></th> <th></th> <th>Numero del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">TCC</td> <td rowspan="5">Mayores</td> <td>1</td> <td>10</td> <td>.08</td> </tr> <tr> <td>2</td> <td>11</td> <td>.08</td> </tr> <tr> <td>3</td> <td>12</td> <td>.08</td> </tr> <tr> <td>4</td> <td>2</td> <td>.07</td> </tr> <tr> <td>5</td> <td>3</td> <td>.07</td> </tr> <tr> <td rowspan="5">Menores</td> <td>1</td> <td>409</td> <td>.03</td> </tr> <tr> <td>2</td> <td>408</td> <td>.03</td> </tr> <tr> <td>3</td> <td>407</td> <td>.03</td> </tr> <tr> <td>4</td> <td>406</td> <td>.03</td> </tr> <tr> <td>5</td> <td>396</td> <td>.03</td> </tr> </tbody> </table>	Valores extremos					Numero del caso	Valor	TCC	Mayores	1	10	.08	2	11	.08	3	12	.08	4	2	.07	5	3	.07	Menores	1	409	.03	2	408	.03	3	407	.03	4	406	.03	5	396	.03
Valores extremos																																										
		Numero del caso	Valor																																							
TCC	Mayores	1	10	.08																																						
		2	11	.08																																						
		3	12	.08																																						
		4	2	.07																																						
		5	3	.07																																						
Menores	1	409	.03																																							
	2	408	.03																																							
	3	407	.03																																							
	4	406	.03																																							
	5	396	.03																																							
IBR		<table border="1"> <thead> <tr> <th colspan="3">Valores extremos</th> </tr> <tr> <th></th> <th></th> <th>Numero del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">IBR</td> <td rowspan="5">Mayores</td> <td>1</td> <td>6</td> <td>.07</td> </tr> <tr> <td>2</td> <td>8</td> <td>.07</td> </tr> <tr> <td>3</td> <td>7</td> <td>.07</td> </tr> <tr> <td>4</td> <td>5</td> <td>.07</td> </tr> <tr> <td>5</td> <td>3</td> <td>.07</td> </tr> <tr> <td rowspan="5">Menores</td> <td>1</td> <td>121</td> <td>.03</td> </tr> <tr> <td>2</td> <td>120</td> <td>.03</td> </tr> <tr> <td>3</td> <td>289</td> <td>.03</td> </tr> <tr> <td>4</td> <td>338</td> <td>.03</td> </tr> <tr> <td>5</td> <td>320</td> <td>.03</td> </tr> </tbody> </table>	Valores extremos					Numero del caso	Valor	IBR	Mayores	1	6	.07	2	8	.07	3	7	.07	4	5	.07	5	3	.07	Menores	1	121	.03	2	120	.03	3	289	.03	4	338	.03	5	320	.03
Valores extremos																																										
		Numero del caso	Valor																																							
IBR	Mayores	1	6	.07																																						
		2	8	.07																																						
		3	7	.07																																						
		4	5	.07																																						
		5	3	.07																																						
Menores	1	121	.03																																							
	2	120	.03																																							
	3	289	.03																																							
	4	338	.03																																							
	5	320	.03																																							
PIB		<table border="1"> <thead> <tr> <th colspan="3">Valores extremos</th> </tr> <tr> <th></th> <th></th> <th>Numero del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">PIB</td> <td rowspan="5">Mayores</td> <td>1</td> <td>440</td> <td>113384.00</td> </tr> <tr> <td>2</td> <td>441</td> <td>113384.00</td> </tr> <tr> <td>3</td> <td>442</td> <td>113384.00</td> </tr> <tr> <td>4</td> <td>443</td> <td>113384.00</td> </tr> <tr> <td>5</td> <td>444</td> <td>113384.00</td> </tr> <tr> <td rowspan="5">Menores</td> <td>1</td> <td>35</td> <td>101752.00</td> </tr> <tr> <td>2</td> <td>34</td> <td>101752.00</td> </tr> <tr> <td>3</td> <td>33</td> <td>101752.00</td> </tr> <tr> <td>4</td> <td>32</td> <td>101752.00</td> </tr> <tr> <td>5</td> <td>31</td> <td>101752.00</td> </tr> </tbody> </table>	Valores extremos					Numero del caso	Valor	PIB	Mayores	1	440	113384.00	2	441	113384.00	3	442	113384.00	4	443	113384.00	5	444	113384.00	Menores	1	35	101752.00	2	34	101752.00	3	33	101752.00	4	32	101752.00	5	31	101752.00
Valores extremos																																										
		Numero del caso	Valor																																							
PIB	Mayores	1	440	113384.00																																						
		2	441	113384.00																																						
		3	442	113384.00																																						
		4	443	113384.00																																						
		5	444	113384.00																																						
Menores	1	35	101752.00																																							
	2	34	101752.00																																							
	3	33	101752.00																																							
	4	32	101752.00																																							
	5	31	101752.00																																							

Variable	Valores atípicos																																
Tasa de desempleo		<table border="1"> <thead> <tr> <th colspan="4">Valores extremos</th> </tr> <tr> <th colspan="2"></th> <th>Numero del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">Tasa de desempleo</td> <td rowspan="5">Mayores</td> <td>1</td> <td>267</td> </tr> <tr> <td>2</td> <td>268</td> </tr> <tr> <td>3</td> <td>269</td> </tr> <tr> <td>4</td> <td>270</td> </tr> <tr> <td>5</td> <td>271</td> </tr> <tr> <td rowspan="5">Menores</td> <td>1</td> <td>134</td> </tr> <tr> <td>2</td> <td>133</td> </tr> <tr> <td>3</td> <td>132</td> </tr> <tr> <td>4</td> <td>131</td> </tr> <tr> <td>5</td> <td>130</td> </tr> </tbody> </table>	Valores extremos						Numero del caso	Valor	Tasa de desempleo	Mayores	1	267	2	268	3	269	4	270	5	271	Menores	1	134	2	133	3	132	4	131	5	130
Valores extremos																																	
		Numero del caso	Valor																														
Tasa de desempleo	Mayores	1	267																														
		2	268																														
		3	269																														
		4	270																														
		5	271																														
Menores	1	134																															
	2	133																															
	3	132																															
	4	131																															
	5	130																															
ICC		<table border="1"> <thead> <tr> <th colspan="4">Valores extremos</th> </tr> <tr> <th colspan="2"></th> <th>Numero del caso</th> <th>Valor</th> </tr> </thead> <tbody> <tr> <td rowspan="5">ICC</td> <td rowspan="5">Mayores</td> <td>1</td> <td>230</td> </tr> <tr> <td>2</td> <td>231</td> </tr> <tr> <td>3</td> <td>232</td> </tr> <tr> <td>4</td> <td>233</td> </tr> <tr> <td>5</td> <td>234</td> </tr> <tr> <td rowspan="5">Menores</td> <td>1</td> <td>23</td> </tr> <tr> <td>2</td> <td>22</td> </tr> <tr> <td>3</td> <td>21</td> </tr> <tr> <td>4</td> <td>20</td> </tr> <tr> <td>5</td> <td>19</td> </tr> </tbody> </table>	Valores extremos						Numero del caso	Valor	ICC	Mayores	1	230	2	231	3	232	4	233	5	234	Menores	1	23	2	22	3	21	4	20	5	19
Valores extremos																																	
		Numero del caso	Valor																														
ICC	Mayores	1	230																														
		2	231																														
		3	232																														
		4	233																														
		5	234																														
Menores	1	23																															
	2	22																															
	3	21																															
	4	20																															
	5	19																															

**Tabla 2-4:** Diagramas de cajas y bigotes de las variables

El diagrama de cajas y bigotes del MRO de la **2-4** muestra la simetría de los datos, donde es posible identificar los puntos cercanos a la media de la serie y aquellos valores extremos, distantes de la media, ubicados dentro del percentil 90 (10%); sin embargo, también se estimaron valores extremos a través de Spss para eliminarlos de la serie y evitar que el ruido que generan puedan distorsionar los resultados.

El valor del fondo diario arrojó 10 casos de valores extremos que se ubicaron dentro del 10% de los valores más distantes respecto a la media; mientras que la rentabilidad diaria del fondo, arrojó 14 valores atípicos que se ubican por encima y por debajo de los bigotes de la caja, considerados como atípicos. El valor medio de los COLTES, COLTES CP, COLTES LP y COLTES UVR entre los años 2009-2011 fue de 141.6, 138.6, 144.76, 139.06, respectivamente. Los límites superior e inferior del bigote no registran valores ni por encima ni por debajo, es decir, que no se presentan valores atípicos.

El valor medio del IGBC y el COLCAP entre los años 2009-2011 fue de 1267.62 y 1516.41, respectivamente. Los límites superior e inferior del bigote no registran valores atípicos. En el diagrama de cajas y bigotes de la variable PIB permite identificar que la mayoría de los datos de la serie se concentran alrededor de la media, y no se identifican puntos extremos.

El valor medio de la serie de la TRM es de 1919.28 y a partir del diagrama de cajas y bigotes se identifican datos atípicos únicamente sobre el límite superior del bigote. El IPC tiene datos atípicos únicamente sobre el límite superior del bigote (valor medio: 0.22).

Se observa la presencia de valores atípicos de la DTF únicamente sobre el límite superior del bigote (valor medio: 0.041); mientras que los datos atípicos del TCC se ubican sobre el límite superior del bigote (valor medio: 0.038). El IBR tiene valores atípicos sobre el límite superior del bigote (valor medio: 0.036). La tasa de desempleo, y no se identifica la presencia de datos atípicos ni por el límite superior ni inferior del bigote. Se observa que el valor medio de la serie es de 0.552.

La tasa de desempleo no presenta datos atípicos ni por el límite superior ni inferior del bigote (valor medio: 0.167); mientras que el ICC muestra la presencia de valores atípicos sobre el límite inferior del bigote.

Un segundo paso, fue la aplicación de la prueba de normalidad Kolmogorov-Smimov para identificar el tipo de distribución que sigue las variables y se identificó que se distribuyen diferente a una normal, al no superar el 0.05 de significancia, en las pruebas de normalidad, lo que servirá de base para la construcción del modelo de minería de datos, presentado en el siguiente capítulo.

### 2.3.2. Transformación de datos

Debido a que las variables del modelo se encuentran en diferentes escalas (porcentajes, precios, otros), se requirió normalizar la base de datos, utilizando la Ecuación 2-1 en una escala de 0 a 1.

$$z = \frac{x - x1}{s} \quad (2-1)$$

Donde, Z: dato estandarizado o normalizado X = valor nominal del dato a estandarizar X1 = media aritmética del conjunto de datos S = desviación estándar.

La normalización del MRO y de las demás variables se realiza con el fin de transformar una variable aleatoria que tiene alguna distribución en una nueva variable aleatoria con distribución normal o aproximadamente normal, estandarizar los resultados y hacerlos comparables. A continuación se presenta el resultado obtenido a partir de la aplicación del proceso de normalización sobre las variables de entrada del modelo, lo cual al ser comparado con las gráficas de comportamiento de cada variable, evidencia el cambio presentado.

#### Variación del MRO

La transformación de características o descubrimiento de características [52] y [53], donde se generan nuevas a partir de las ya existentes. En esta parte se calculó la variación diaria

de cada variable y se convirtieron en las nuevas variables a ser utilizadas, debido a que la serie de precios por sí sola no recoge la tendencia histórica de la variable. De un total de 571 registros iniciales, quedaron 422 datos por variable después de realizar limpieza y transformación a cada una de ellas. En la tabla se presentan las variables transformadas.

Adicionalmente, se construyeron atributos a partir de otros (i.e. Ingresos â“ egresos, saldo neto), lo que elimino información redundante de la base de datos. Así mismo, no se presentaron datos faltantes porque esta información es reportada generada diariamente mediante el cierre del portafolio de inversiones, con el cual se negocia en el mercado de valores.

## 2.4. Selección de Atributos para el Modelo de Clasificación

La selección de variables o atributos consta de la identificación de características relevantes para el modelo. Se parte del supuesto de que los datos contienen características redundantes o irrelevantes, es decir, que no proporcionan información adicional a la aportada por las variables seleccionadas [54]. Dentro de las ventajas que otorga las técnicas empleadas para seleccionar atributos se encuentran las siguientes: mejorar el entendimiento e interpretación del modelo, reducción del tiempo de entrenamiento y generalización mejorada debido a la minimización del sobre-ajuste.

Para este ejercicio se hizo uso de la técnica de ranking de características â“ filtro de correlación, respecto a su clase (MRO) para seleccionar las variables que se utilizaran en el modelo de minería de datos. Este algoritmo disponible en Weka (software de minería de datos) encuentra los pesos de atributos continuos a partir de su correlación con el atributo de la clase continua. Dichos resultados se combinaron con los resultados arrojados por la matriz de correlación, presentada a continuación.

## 2.5. Matriz de correlación

A continuación se presenta el nivel de correlación de las variables que serán evaluadas más adelante para hacer parte del modelo de minería de datos, con el fin de identificar aquellas variables que estén altamente correlacionadas con otras, y que puedan distorsionar los resultados.

De acuerdo con la matriz de correlación de Tabla 2-5, existe una correlación positiva superior al 70 % entre las variables COLTES, COLTES CP, COLTES LP y COLTES UVR, lo que se



traduce, como colinealidad o dependencia entre ella, y dado que la tasa de estos títulos de deuda pública COLTES reúne de manera general las otras variables correlacionadas, se optó por descartarlas y conservar solo el COLTES. También se presenta una correlación positiva entre el IGBC y el COLCAP, (índices del mercado de valores), y se optó por el primer índice por contener al segundo.

Existe una correlación negativa entre el IGBC y la TRM, pero en este caso la segunda no esta contenida en la primera variable, lo que indica es que mientras el IGBC crece positivamente, la TRM decrece, son inversamente proporcionales; motivo por el cual no se descarta ninguna de las dos variables. Este mismo resultado se obtuvo a partir del la técnica de ranking de características mencionada anteriormente. En la Tabla 2-5 se presentan las variables seleccionadas:

<b>Variable seleccionada</b>	<b>Descripción</b>
MRO	Máximo Retiro Observado
VFCt	Valor del fondo al cierre del día antes de rendimientos
Rentabilidad diaria	Índice que mide la relación entre la utilidad y la inversión que se utilizaron para obtener dichas utilidades.
Coltes	Títulos de Deuda Pública - TES. Son títulos de renta fija expedidos por el Gobierno Nacional para obtener financiación o apalancamiento financiero [47].
IGBC	índice General de la Bolsa de Valores de Colombia
TRM	Tasa representativa del mercado (TRM), es la cantidad de pesos colombianos por un dólar de los Estados Unidos. La Superfinanciera la calcula con base en las operaciones de compra y venta de divisas entre intermediarios financieros que transan en el mercado cambiario colombiano, con cumplimiento el mismo día cuando se realiza la negociación de las divisas [47].
IPC	índice de Precios al Consumidor (IPC). Es un indicador sobre el cual se acumulan a partir de un periodo base las variaciones promedio de los precios de los bienes y servicios consumidos por los hogares del país en un periodo de tiempo. [49]
PIB	El Producto Interno Bruto (PIB), mide el resultado final de la actividad productiva (unidades de producción) de los residentes del país. Se mide desde el punto de vista del valor agregado, de la demanda final o las utilizaciones finales de los bienes y servicios y de los ingresos primarios distribuidos por las unidades de producción residentes. [49]

## 2.6. Resumen del capítulo

En este capítulo se presentó el análisis descriptivo, preprocesamiento y selección de variables para el modelo de minería de datos, mediante el uso de gráficos de tendencias, estadísticas, y la identificación de datos atípicos con diagramas de cajas y bigotes.

Adicionalmente, se calculó la matriz de correlación entre variables y se hizo uso de la técnica de ranking de características - filtro por correlación de Weka para la selección de variables, lo que permitió conformar el conjunto de datos final, que se tiene en cuenta para el modelo de minería de datos.

Tabla 2-5: Matriz de correlación

	VFCt	Ren dia	Coltes	Coltes CP	Coltes LP	Coltes UVR	IGBC	colcap	TRM	IPC	PIB	DTF	TCC	IBR	Tdes	ICC	MRO
VFCt	1	0.25	-0.73	-0.75	-0.70	-0.78	-0.67	-0.66	0.52	-0.35	-0.88	0.35	0.44	0.12	-0.33	-0.25	0.01
Ren dia	0.25	1	-0.36	-0.37	-0.34	-0.38	-0.41	-0.41	0.40	-0.16	-0.34	0.45	0.41	0.40	-0.18	-0.37	0.05
Coltes	-0.73	-0.36	1	0.99	0.99	0.98	0.92	0.93	-0.84	0.18	0.89	-0.77	-0.76	-0.57	0.67	0.73	-0.02
Coltes CP	-0.75	-0.37	0.99	1	0.99	0.99	0.93	0.94	-0.84	0.21	0.91	-0.78	-0.78	-0.57	0.65	0.72	-0.03
Coltes LP	-0.70	-0.34	0.99	0.99	1	0.96	0.91	0.92	-0.83	0.15	0.86	-0.75	-0.75	-0.54	0.69	0.74	-0.02
Coltes UVR	-0.78	-0.78	0.98	0.99	0.97	1	0.92	0.93	-0.84	0.29	0.93	-0.76	-0.77	-0.54	0.57	0.69	-0.03
IGBC	-0.67	-0.41	0.92	0.93	0.91	0.92	1	0.99	-0.81	0.29	0.81	-0.84	-0.81	-0.69	0.59	0.71	-0.02
colcap	-0.66	-0.41	0.93	0.94	0.92	0.27	0.99	1	-0.83	0.28	0.81	-0.86	-0.82	-0.70	0.62	0.73	-0.02
TRM	0.52	0.40	-0.84	-0.84	-0.83	-0.37	-0.81	-0.83	1	-0.03	-0.73	0.83	0.78	0.67	-0.56	-0.79	0.02
IPC	-0.35	-0.16	0.18	0.21	0.15	0.30	0.30	0.28	-0.03	1	0.34	-0.25	-0.28	-0.21	-0.27	-0.06	-0.07
PIB	-0.88	-0.34	0.89	0.91	0.86	0.93	0.81	0.81	-0.73	0.34	1	-0.60	-0.64	-0.35	0.49	0.48	-0.02
DTF	0.35	0.45	-0.76	-0.78	-0.75	-0.76	-0.84	-0.86	0.83	-0.25	-0.60	1	0.90	0.93	-0.44	-0.81	0.05
TCC	0.44	0.41	-0.76	-0.78	-0.75	-0.77	-0.81	-0.82	0.78	-0.28	-0.64	0.90	1	0.83	-0.43	-0.73	0.04
IBR	0.12	0.40	-0.55	-0.57	-0.54	-0.54	-0.69	-0.70	0.67	-0.21	-0.35	0.93	0.83	1	-0.29	-0.72	0.03
Tdes	-0.33	-0.18	0.67	0.65	0.69	0.57	0.59	0.62	-0.56	-0.27	0.49	-0.44	-0.43	-0.29	1	0.50	-0.03
ICC	-0.25	-0.37	0.73	0.72	0.74	0.69	0.71	0.73	-0.79	-0.60	0.48	-0.81	-0.73	-0.72	0.50	1	-0.01
MRO	0.01	0.05	-0.02	-0.03	-0.23	-0.29	-0.02	-0.02	0.02	-0.07	-0.02	0.05	0.04	0.03	-0.03	-0.01	1

Nombre de la variable	Descripción de la variable
DTF	Tasa de referencia (DTF). Promedio de las tasas de captación de los depósitos a 90 días de establecimientos bancarios, corporaciones financieras y compañías de financiamiento comercial [47]
TCC	Tasa de interés estimada a partir del promedio ponderado semanal por monto, de las tasas promedios de captación diarias de los CDT's a 90 días, pagadas por las corporaciones financieras privadas. Es calculada por el Banco de la República con la información suministrada por la Superintendencia Bancaria
IBR	índice Bancario de Referencia (IBR). Tasa de interés de referencia de corto plazo en pesos colombianos, que representa el precio al que los bancos captan recursos en el mercado monetario [47]
Tasa desempleo	Expresa el nivel de desocupación entre la población económicamente activa [48].
ICC	índice de confianza del consumidor (ICC). Promedio simple de los balances de preguntas que cuestionan la opinión de los hogares, tanto de las condiciones actuales como de perspectivas a futuro del país.
Día	Día de la semana hábil en el que se realiza transacciones (lunes, martes, miércoles, jueves y viernes)
Semana	Número de la semana del mes al que pertenece la transacción realizada (MRO)

**Tabla 2-6:** Descripción de las variables para el modelo de minería de datos

# 3 Modelo predictivo de Riesgo de Liquidez

En este capítulo se presenta el modelo propuesto de minería de datos para predecir riesgo de liquidez (MRO), basado en redes neuronales y árboles de decisión. De conformidad con lo mencionado por el proceso KDD, en su etapa de descubrimiento de conocimiento se seleccionaron los mejores resultados obtenidos para el modelo de riesgo de liquidez. Se tomó la decisión de utilizar redes neuronales por el éxito que han tenido en la predicción de series financieras y árboles de clasificación por su aplicación en riesgo de crédito, los cuales se toman como base para construir el modelo.

Los resultados obtenidos son validados con base en la experiencia del analista en riesgo de liquidez (autor de este documento), lo que facilita su interpretación y análisis frente a otros modelos tradicionales. El entrenamiento y la validación de las redes neuronales se realizaron en lenguaje de programación "Python", utilizando las librerías de neurolab y ffinet. Mientras que para el árbol de decisión se utilizó el set de algoritmos de Weka (J48).

## 3.1. Modelo General de Predicción de Riesgo de Liquidez

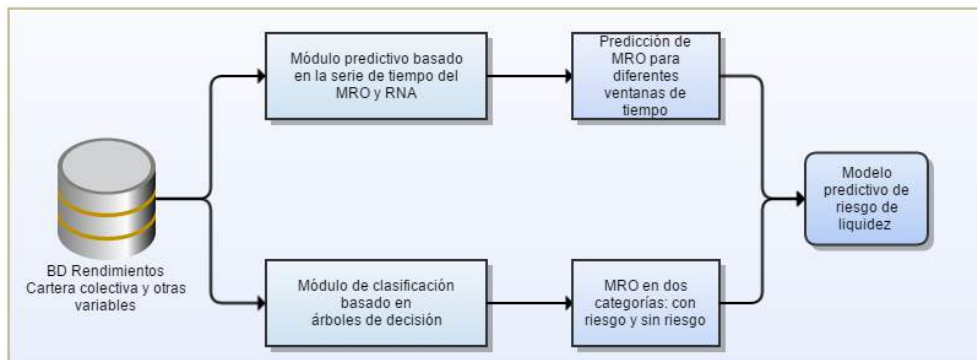


Figura 3-1: Modelo predictivo de riesgo de liquidez

En la Figura 3-1 se presenta el modelo general de predicción de riesgo de liquidez, el cual esta compuesto por dos módulos: el modelo predictivo y el modelo descriptivo o de clasificación

de riesgo de liquidez. El módulo predictivo pretende predecir el valor del MRO usando la serie histórica y redes neuronales. El módulo de clasificación pretende predecir la categoría del MRO ("sin riesgo" y "con riesgo"), basado en un árbol de decisión y dada una serie de valores de las variables macroeconómicas correspondientes a un día.

Se optó por utilizar estas técnicas de minería de datos porque con las redes neuronales se busca predecir el valor y con el árbol de clasificación se predice su comportamiento en función de las variables macroeconómicas. Adicionalmente, su aplicación en riesgos financieros, tal como se mencionó en el estado del arte, ha sido exitosa, lo que permite la creación de un modelo propio e innovador para la administración del riesgo de liquidez de una cartera colectiva, convirtiéndose en una herramienta esencial para la toma de decisiones.

### 3.2. Predicción del MRO a partir de la serie de tiempo usando Redes Neuronales

La primera parte del modelo predictivo presentado en la Figura 3-1, permite la predicción del valor del MRO, basado en la serie histórica por dos años y medio y utilizando redes neuronales (ver Figura 3-2). Se optó por utilizar redes neuronales porque en la predicción de series económicas han obtenido mejores resultados frente a otras técnicas de proyección (mencionado en el estado del arte), especialmente, al permitir utilizar series no lineales, lo que caracteriza el sector financiero y determina el desempeño del modelo.

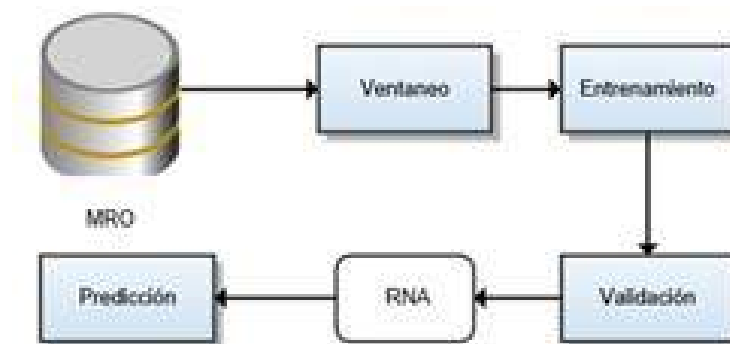


Figura 3-2: Predicción del MRO con RNA

Para definir la arquitectura y topología de la red neuronal (RNA) a ser parte del modelo, se seleccionaron tres arquitecturas de red: red neuronal multiperceptrón, red neuronal basada en la media móvil y red neuronal autorregresiva tipo NARX. Se realizaron experimentos para comparar la precisión en la predicción y adicionalmente se comparó con un modelo

econométrico ARIMA.

### 3.2.1. Preprocesamiento para series de tiempo

Para utilizar la serie del MRO en RNA, es indispensable definir el rango máximo de observaciones, y así seleccionar la mejor RNA. En este caso, se consideraron 30 días como rango máximo de observaciones para la predicción y no un periodo superior, ya que la existencia de cambios estructurales en la serie, pueden forzar a que el modelo no asimile correctamente la dinámica que exhibiría la serie real, y en consecuencia, se podría presentar un gran número de errores en el pronóstico [44].

Se construyeron diferentes conjuntos de datos correspondientes a las ventanas de tiempo seleccionadas (5 en total), discretizando el MRO en dos categorías: Con riesgo y Sin riesgo, con el fin de predecir el comportamiento o tendencia de esta variable en relación con las variables macroeconómicas seleccionadas, que inciden en su comportamiento.

Es importante mencionar, que dentro del preprocesamiento se comenzó vinculando la serie del MRO con el día de la semana para identificar cuándo se presentaban los retiros más altos y con base en ello, identificar un patrón de comportamiento de la series, sin embargo no se obtuvieron buenos resultados, y no pudo ser utilizado en el modelo.

### 3.2.2. Arquitectura de la Red Neuronal

- Configuración: se empleó solo una variable de entrada (neuronas de entrada), 1-2 capas ocultas (hidden layer) por cada experimento realizado y diferentes niveles de retardo (r) para obtener un mejor resultado predictivo de la serie de tiempo.
- Vector de salida o respuesta (output layer o capa de salida). Se generó una neurona de salida en cada experimento.
- Con el propósito de realizar el aprendizaje de los algoritmos, se particionó la serie del MRO aleatoriamente sin reemplazo, para el entrenamiento y la validación (90 % y 10 %, respectivamente).

Cada una de las redes neuronales creadas se reentrenaron varias veces, con el fin de obtener el mejor resultado posible, el cual fue seleccionado a partir del menor error mínimo cuadrático obtenido por cada ventana de tiempo.

- Las épocas varian entre 2 y 10.

### 3.2.3. Experimentos con Redes Neuronales

Se probaron diferentes arquitecturas sobre el orden de entrada, orden de salida, pesos de salida, pesos de entrada y tasa de aprendizaje, empleando 1 y 2 capas ocultas, y una sola variable de entrada, se entrenó sobre el 90 % del conjunto de los datos de la serie, con diferentes ventanas de tiempo para diferentes configuraciones de los dos tipos de red neuronal comparadas: Red neuronal de media móvil y Red neuronal backpropagation, cuya predicción se comparó con los datos reales del MRO.

#### Configuración

Se realizaron aproximadamente cincuenta ensayos, buscando obtener la mejor configuración de los parámetros de las redes neuronales elaboradas para la generación de un modelo de riesgo de liquidez. El entrenamiento de la red neuronal se realizó bajo los siguientes parámetros para encontrar la topología con mejor desempeño:

- Orden de entrada: 1
- Capa(s) oculta(s): 1 a 2
- Orden de salida: 1 a 5
- Rata de aprendizaje: se definió en un rango de (0.3 a 0.4), dando como parámetro óptimo 0.3
- Épocas: 2 a 10. Se logró el óptimo en 10 épocas.
- Función de propagación (excitación): uso de señales de entrada positiva (exitatoria) ó de transferencia: sigmoide (generó salidas en intervalo: 0,1) y tangente hiperbólica (generó salidas en intervalo: -1,1). La primera función mostró un mejor desempeño en la activación de los nodos de la capa oculta.
- Ventanas de tiempo: 3,5,8,15 y 30 días
- Error: para la RNA tipo NARX se convierte en una neurona de entrada en la corrida  $t+1$ ; mientras que para las RNA tipo FNN y MM el error no se toma en cuenta para la corrida  $t+1$ .

Validación: se optó por realizar validación cruzada, cambiando el número de folds, entre 10 y 30 "folding". El mejor resultado se obtuvo con 10 "folding", utilizando el lenguaje de programación de "Python".



### 3.2.4. Resultados entrenamiento de RNA

Los resultados obtenidos con diferentes ventanas de tiempo para los tres tipos de arquitecturas de red frente al valor real del Máximo Retiro Observado, se presentan en la (Figura 3-3) y 30 (Figura 3-4). Los mejores resultados se obtuvieron a partir de la ventana 15

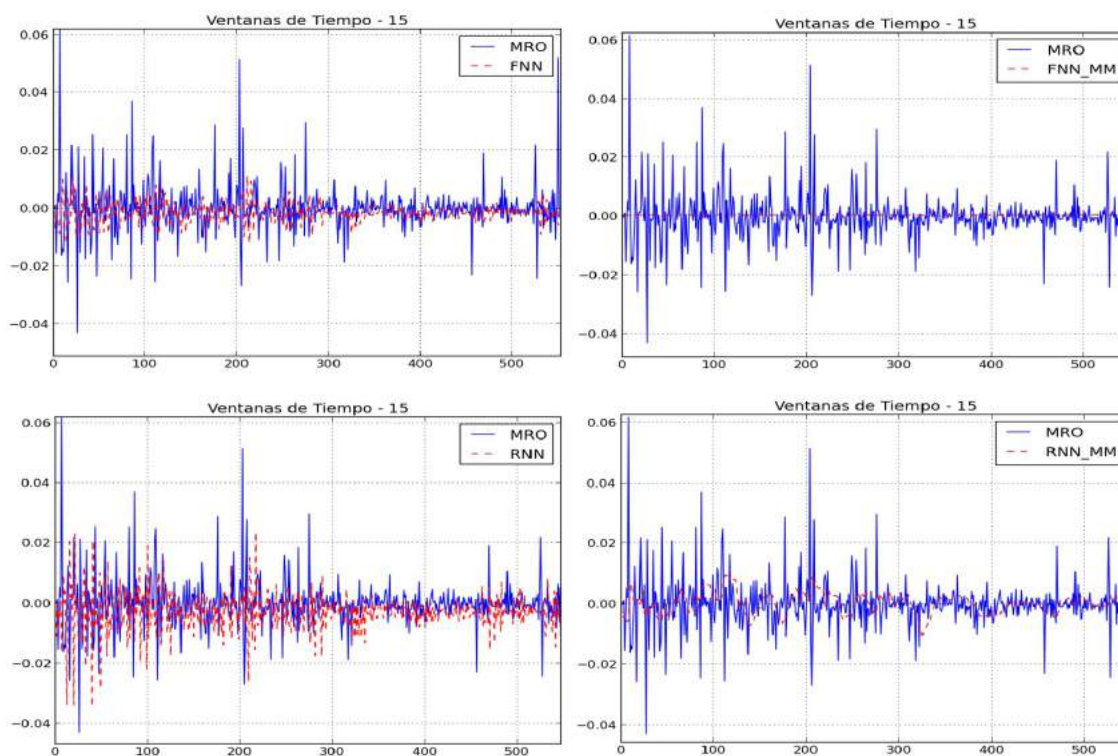


Figura 3-3: Predicción del MRO con RNA

Se encontró que entre más alta sea la tasa de aprendizaje, más elevado es el error obtenido por cada corrida, motivo por el cual se llevaron a cabo varias corridas con el fin de encontrar la tasa óptima de aprendizaje y evitar sobre entrenamiento del modelo. A partir de los parámetros definidos en la configuración de la RNA, se obtuvieron los resultados presentados en la Tabla 3-1 para las ventanas de tiempo seleccionadas.

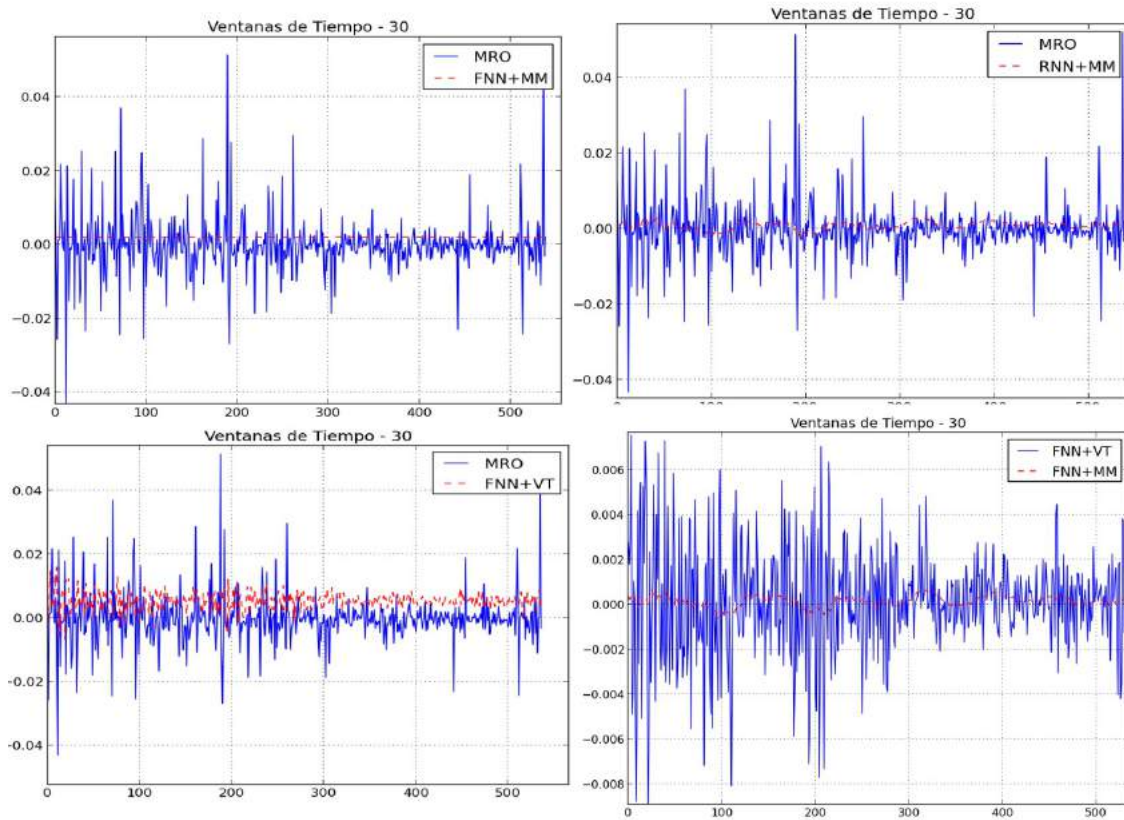


Figura 3-4: Predicción del MRO con RNA

Tipo de RNA	Ventana de tiempo	de	Error Cuadrático Mínimo	Real vs. Proyectado en Millones de pesos
Feed-Forward (FNN)	15		8.67198840352e-05	43,037.92
Feed-Forward (FNN)- MM	15		6.91290765857e-05	43,039.76
Feed-Forward (FNN)	30		6.50032086958e-05	43,040.20
Feed-Forward (FNN)- MM	30		6.15065009272e-05	43,040.56

Tabla 3-1: Resultados obtenidos para el MRO por RNA

Tipo de RNA	Ventana de tiempo	Error Medio Absoluto	Real vs. Proyectado (millones pesos)
FNN	15	0.00008670	43,038
FNN- MM	15	0.00006910	43,040
RNN	15	0.00012600	43,034
RNN - MM	15	0.00008570	42,038
Arima	15	0.000444	43,05
FNN	30	0.00011100	43,035
FNN- MM	30	0.00006500	43,040
RNN	30	0.00006280	43,040
RNN - MM	30	0.00006150	43,041
Arima	30	0.000456	43,047

**Tabla 3-2:** Resultados obtenidos para el MRO por RNA y ARIMA

En la Tabla 3-1 y en la Figura 3-3 y Figura 3-4 se observa que para el conjunto de entrenamiento y validación (90 % y 10 %, respectivamente), la RNA autorregresiva tipo NARX, tiene un mejor nivel de predicción en los saltos de la serie, y un error mínimo cuadrático bajo, lo que corrobora que para series financieras este tipo de RNA predice con un buen resultado el comportamiento de la serie. Aunque se obtuvieron mejores resultados en la predicción del valor del MRO, la RNA es capaz de predecir adecuadamente su tendencia, convirtiéndose en una valiosa herramienta de análisis para la toma de decisiones en el corto y mediano plazo. Lo que se corrobora con otros estudios realizados con series de tiempo financieras, como el movimiento de las acciones en el mercado de valores, que buscan reconocer patrones de comportamiento, tendencias y la capacidad de predecir comportamientos de sistemas complejos [42].

### 3.2.5. Análisis comparativo de resultados

Los resultados obtenidos por los tres tipos de RNA se compararon con los resultados obtenidos con el modelo ARIMA, siendo uno de los más utilizados para predecir el riesgo de liquidez en el sector financiero colombiano. Dicho modelo se calculó en Eviews (software econométrico), para el cual se efectuaron pruebas de normalidad - Jaque Bera, se transformo la serie de tiempo tomando los logaritmos y corrigiendo la estacionalidad y pruebas de ruido blanco.

En la Tabla 3-2 se presentan los resultados obtenidos por las RNA y el modelo ARIMA.

En la Tabla 3-2 se observa que la RNN tipo NARX obtuvo un mejor resultado en la predic-

ción del valor del MRO frente al valor real, respecto al obtenido por el modelo ARIMA. En un periodo de 30 días el EMC y la diferencia entre el valor proyectado (0.000065 y 43,040 millones, respectivamente) y el valor real, frente a los resultados obtenidos para ventanas de tiempo menores. Este patrón se evidencia en las otras dos arquitecturas de red y en el modelo ARIMA.

### 3.3. Predicción del MRO mediante árboles de decisión

Las técnicas de árboles de decisión se utilizan para el descubrimiento de reglas y patrones de comportamiento, los cuales se convierten en insumo para la generación del modelo de predicción, y se parte del problema de determinar el riesgo de liquidez y sus implicaciones en el proceso de toma de decisiones. En este caso se seleccionó un árbol de clasificación para predecir el comportamiento del MRO, dada su efectividad en riesgo de crédito.

#### 3.3.1. Preprocesamiento y Selección de atributos del modelo

Dentro del preprocesamiento se transformaron las variables del conjunto de datos para evitar el efecto de la estacionalidad de las variables macroeconómicas, y como el MRO es un valor numérico que se pretende usar para hacer clasificación, se decidió discretizar en dos clases (con riesgo y sin riesgo) para identificar el comportamiento que presenta el MRO en relación con otras variables macroeconómicas, y complementar la predicción hecha en el primer módulo del modelo de riesgo de liquidez.

Debido a que dentro del conjunto de datos, existen variables, como el PIB, TCC e ICC que son publicadas por el DANE mensualmente, sin embargo es una cifra que tiene vigencia durante todo el mes, por lo cual se incluyó dentro del análisis, pero se dejó fijo como variación el cambio mensual. Se tomó la decisión de conservarlas como parte del modelo, por la incidencia que ejercen sobre la inversión y el ahorro en términos económicos. El incremento o decrecimiento por cada sector de la economía, para el caso particular del PIB, incide en el crecimiento o no de las tasas de colocación y por ende en la disposición que tendrían los inversionistas en ahorrar o destinar sus recursos al gasto, lo que de forma indirecta incide en la decisión que toma el inversionista de una cartera colectiva de conservar o no sus recursos en una cartera colectiva, o gastarlos en cubrir necesidades básicas.

Se tuvo en cuenta los resultados obtenidos en el análisis descriptivo y estadístico del capítulo 2, en el que se seleccionaron atributos a partir de la matriz de correlación y la metodología de ranking de características de WEKA. De un total de 18 variables y una clase, se seleccionaron 14 variables y una clase (MRO) tal como se presentan en la Tabla **3-3**.

Atributo	Clase
Día	Categorico
Semana	Categorico
Var VFct	Numerico
Var Rentab d	Numerico
Var COLTES	Numerico
Var IGBC	Numerico
Var TRM	Numerico
Var IPC	Numerico
Var PIB	Numerico
Var DTF	Numerico
Var TCC	Numerico
Var IBR	Numerico
Var i desem	Numerico
Var ICC	Numerico
MRO	Categorico - clase

**Tabla 3-3:** Resultados obtenidos para el MRO por RNA y ARIMA

Las variables excluidas de la selección de atributos se descartaron por la correlación que existe entre ellas, el caso particular del COLTES, quien tiene una alta correlación con COLTES CP, COLTES LP y COLTES UVR. Así mismo, sucedió con el IGBC y el COLCAP, y dado que el primero contiene al segundo índice burátil, se optó por tomar en cuenta el primero para el modelo de clasificación.

### 3.3.2. Árbol de decisión

Con el fin de encontrar reglas y patrones de comportamiento se optó por aplicar el árbol de decisión J48, implementado en Weka. Este árbol fue configurado utilizando la opción de poda y con un nivel de confianza del 0.25 y 10 "folding". La evaluación del desempeño del módulo de clasificación se realizó mediante el uso de la matriz de confusión, la cual provee información de las instancias correcta e incorrectamente clasificadas. A su vez, se utilizaron medidas como la precisión y accuracy para establecer la relación entre instancias correctamente clasificadas y el total de las proyecciones realizadas.

El conjunto de datos utilizado no presenta desbalanceo de clase, lo que facilita su análisis. El 57.58 % representan la etiqueta "sin riesgo" del MRO y el 42.41 % corresponden a la etiqueta "con riesgo". Se obtuvo un nivel de precisión del 89.34 % de instancias correctamente clasificadas, y se obtuvieron las reglas que indican si se presenta o no retiro de la cartera

colectiva. (ver Figura 3-5).

	Con riesgo	Sin riesgo
Con riesgo	158	21
Sin Riesgo	22	221

Figura 3-5: Matriz de confusión

A continuación se presentan las reglas generadas por el algoritmo j48 de clasificación para el modelo multivariado seleccionado (Ver Figura 3-6):

- Si la variación de saldo del fondo  $\leq -0.000415$ , no se dan retiros
- Si la variación de saldo del fondo  $> 0.000095$ , se genera riesgo de retiro
- Si la variación de saldo del fondo  $\leq 0.000095$  y la variación de la tasa de desempleo es  $\leq -0.006045$ , se presenta riesgo de retiro
- Si la variación de saldo del fondo  $> 0.016487$  y la variación de la DTF es  $> 0.0025$ , se presenta riesgo de retiro
- Si la variación de saldo del fondo  $> 0.016487$  y la variación de la DTF es  $\geq 0.0025$ , no se presenta riesgo de retiro

Analizando los resultados anteriores, dado el conjunto de datos de validación, se observa que obtuvo un nivel de precisión alto (89.34%), se presenta un margen de error en el entrenamiento del modelo, lo que sugiere que sea complementado con otras técnicas que aporten mayor información en la toma de decisiones, como se sugiere en este trabajo, al buscar integrar los resultados del modelo de clasificación con el modelo predictivo a partir de RNA.

Vale la pena mencionar que en entrenamientos anteriores no se obtuvieron resultados satisfactorios (precisión superior al 80%) al crear más de dos etiquetas por la clase del MRO, para este caso se habían creado las etiquetas de: riesgo alto, medio, bajo y sin riesgo, discretizando la variable a partir de la construcción de rangos por la distribución de la serie.

### 3.4. Integración de los módulos del modelo para predicción

Se realizó una integración de los resultados de los dos módulos del modelo predictivo de riesgo de liquidez para la conformación del modelo general. En primera instancia, se tomo

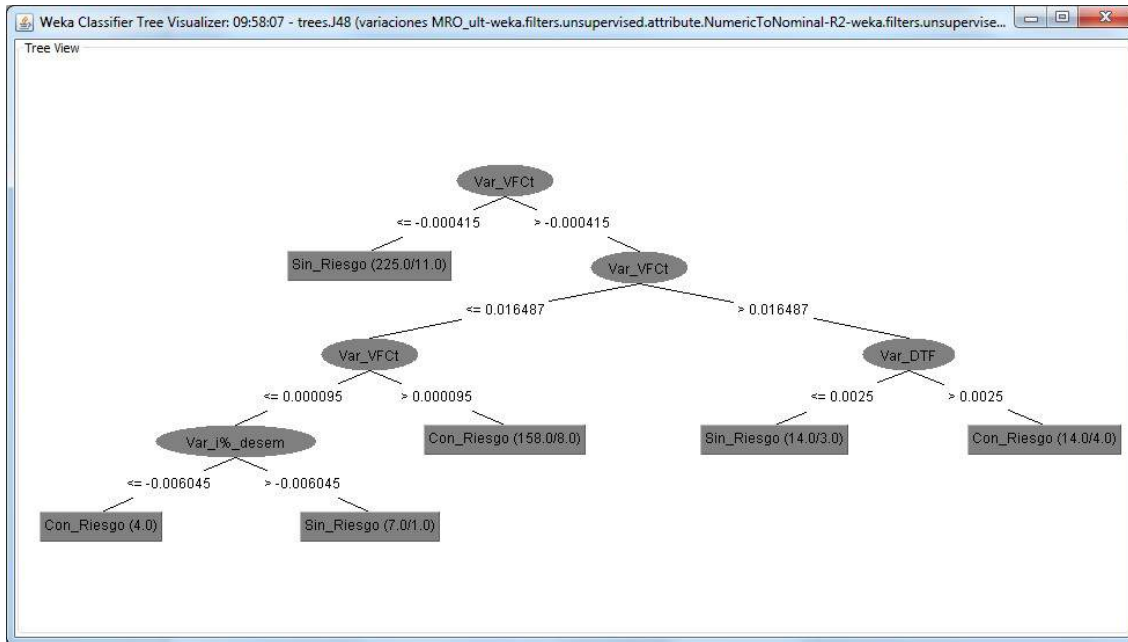


Figura 3-6: árbol de clasificación para el MRO

el conjunto de datos históricos utilizados por la RNA para validación en la predección del MRO, y se paso por el árbol de clasificación para complementar e integrar los resultados obtenidos por cada módulo del modelo. Se utilizó la misma configuración del módulo de clasificación para el testing del modelo, y se construyó la matriz de confusión a partir del testing (ver Tabla 3-4)

Tabla 3-4: Matriz de confusión del modelo de riesgo de liquidez

	true Con_riesgo	true Sin_riesgo	class precision
pred. Con_riesgo	4	0	100.00 %
pred. Sin_riesgo	1	6	85.71 %
class recall	80.00 %	100.00 %	

En la Tabla 3-4 para la matriz de confusión se tomó un listado aleatorio de 10 registros para mostrar el resultado del testing, como ejemplo, dada la extensión del conjunto de datos para el testing. Se obtuvo un resultado positivo (85.71 % de precisión), lo que permite concluir que los dos módulos seleccionados para la construcción del modelo de riesgo de liquidez se complementan y brindan mayores herramientas para la toma de decisiones del administrador de portafolio, en la identificación de potenciales retiros de dinero de su cartera colectiva (ver Tabla 3-5).

**Tabla 3-5:** Testing del modelo de riesgo de liquidez

<b>MRO_Real</b>	<b>MRO_predicc</b>	<b>J48_MRO</b>	<b>Acertabilidad</b>
-5,162,545,000	-99,433,470	Con_riesgo	Si
-4,454,031,000	-1,373,102,836	Con_riesgo	Si
-665,474,000	-444,981,313	Con_riesgo	Si
-1,522,432,000	-2,153,166,968	Con_riesgo	Si
1,962,247,000	-501,243,334	Con_riesgo	No
-1,665,754,000	-2,060,697,379	Con_riesgo	Si
-2,348,092,000	-911,301,359	Con_riesgo	Si
-641,036,000	-853,283,715	Con_riesgo	Si
-185,884,000	-689,920,978	Con_riesgo	Si
-2,135,017,000	-1,277,918,939	Con_riesgo	Si

### 3.5. Resumen

En este capítulo se presentó el modelo predictivo para riesgo de liquidez de una cartera colectiva colombiana, el cual se compone de dos módulos: el modelo de predicción del valor del MRO usando RNA y el modelo de clasificación. Se presentaron las validaciones computacionales a las que fueron sometidos los dos módulos del modelo: predicción del valor del MRO usando RNA y la caracterización del MRO mediante árbol de decisión.

Para el modelo de predicción usando la serie de tiempo del MRO, se experimentó con tres arquitecturas de RNA, y su desempeño fue validado a partir del EMC y la diferencia entre el valor real y el proyectado. Mientras que el modelo de clasificación utilizando un árbol de decisión J48 fue elaborado mediante la incorporación de variables macroeconómicas y la discretización del MRO en dos etiquetas de clase (con riesgo y sin riesgo).

Posteriormente, se realizó la integración de los dos módulos del modelo propuesto, y se llevó a cabo el testing del modelo con 10 registros que formaron parte del conjunto de validación utilizado en la construcción de la red neuronal. Se obtuvo un nivel de precisión del 85.71 %, lo que permitió concluir que la integración de las técnicas de minería de datos seleccionadas fueron adecuadas a lo previsto en la propuesta del trabajo de grado.



# 4 Conclusiones y Trabajos Futuros

## 4.1. Conclusiones

- Este trabajo propuso un modelo predictivo para riesgo de liquidez de una cartera colectiva colombiana, como alternativa para el análisis del alto volumen transaccional y el impacto que tienen las principales variables macroeconómicas sobre su comportamiento, basado en técnicas de minería de datos. El modelo se compone de dos partes para la predicción del Máximo Retiro Observado –MRO (determinante en el riesgo de liquidez): predicción del MRO a partir de redes neuronales y árboles de decisión. En la primera parte, se utilizan diferentes ventanas de tiempo para estimar el precio del MRO respecto a sus registros transaccionales históricos. La segunda parte predice el riesgo de liquidez mediante la combinación de variables macroeconómicas, lo que permite al administrador de portafolio tomar mejores decisiones y la discretización del MRO para predecir salida de dinero de la cartera colectiva.
- A partir del preprocesamiento de los datos, se determinó que para el tratamiento de datos económicos se requiere traer a precios constantes si se toma un histórico de varios años. En este caso, se logró incorporar la tendencia de la serie del MRO mediante la variación diaria de precios, lo que permitió obtener un mejor diagnóstico.
- En la primera y segunda parte de la construcción del modelo, se realizó un análisis exploratorio, donde se identificaron valores atípicos, la distribución de los datos, principales medidas de tendencia central, mediante el uso de diagramas de cajas y bigotes. Adicionalmente, para la segunda fase del modelo fue necesario normalizar los datos de cada una de las variables, para dejarlos en los mismos términos, y resultó útil el uso de matrices de correlación para la selección de atributos.
- En el proceso de entrenamiento y validación de las redes neuronales, utilizadas para predecir el precio del MRO, fue necesario realizar 50 corridas para las RNA tipo FNN y Moving Average. Para el caso de la RNA tipo NARX se realizaron el mismo número de corridas, pero la ventaja que ha otorgado en otros estudios de variables económicas es que el error generado en cada corrida, se convierte en una neurona de entrada para la siguiente, lo que le permitió a la RNA incorporar el error en su entrenamiento y obtener el menor error mínimo cuadrático. Adicionalmente, se identificó que para una

ventana de 30 días, se obtiene una mejor predicción del retiro, frente a los resultados generados por el modelo ARIMA.

- La aplicación y validación se efectuó sobre un conjunto de datos reales y tomando como base la experiencia que posee el autor de este trabajo en riesgos financieros se validaron los resultados obtenidos; no obstante sería interesante validar su aplicación en otro tipo de carteras colectivas (cerrada y escalonada).
- Las ventajas que se identificaron mediante el uso de técnicas de minería de datos para la generación de conocimiento y patrones de comportamiento fue la reducción del costo, incremento de los ingresos y de la capacidad de respuesta.
- Este tipo de trabajos se pueden aplicar en otros campos de ingeniería industrial (logística y transporte, innovación, entre otros), dado el buen desempeño predictivo de los modelos computacionales basados en técnicas de minería de datos.
- De acuerdo con los resultados obtenidos en este trabajo, se puede concluir que el modelo construido a partir de técnicas de minería de datos se convierte en una herramienta importante para el proceso de toma de decisiones, especialmente, por la identificación de patrones, tendencias y comportamientos del entorno económico sobre el MRO.

## 4.2. Trabajos futuros

De acuerdo con los resultados presentados en el capítulo 3, correspondientes a la red neuronal, Se presentan como una serie de aspectos que se podrían realizar en un futuro para emprender investigaciones similares o fortalecer la investigación realizada. Como trabajo futuro quedan los siguientes:

- Búsqueda o construcción de algoritmos para el tratamiento de datos que presenten desbalanceo en la clase y que es un problema usual en el manejo de información financiera
- A partir del ejercicio planteado con las redes neuronales es importante que se busque la extracción de reglas para comprender el comportamiento de las salidas generadas por la RNA
- Revisar la implementación de técnicas tradicionales de econometría que se pueden potencializar con la RNA y que le permiten tener mayor dominio de sus atributos, al no exigir el principio de "ceterisparibus" o efecto parcial, el cual exige que los atributos que explican la clase permanezcan constantes, mientras que la variable o clase es el único atributo que puede ser dinámico.
- Evaluar la implementación del modelo planteado desde redes neuronales y árboles de decisión en otros productos financieros, donde se desee predecir el riesgo de liquidez, y

extraer reglas para la generación de alertas para el administrador del producto financiero.

- Otra de las posibilidades futuras de esta investigación consiste en la incorporación de otras variables en los criterios de las técnicas de minería de datos. Estos incluyen nuevas variables que reflejar la información futuro y los que reflejan los impactos de otros mercados bursátiles.

# Bibliografía

- [1] LÓPEZ SANABRIA, Pablo. Administración de Riesgos Financieros. III Coloquio predoctoral iberoamericano, 2005
- [2] JORION, Philippe. Value at Risk: the New Benchmark for Managing Financial Risk. 2 edition. McGraw-Hill, 2000.
- [3] LARRIETA, María de los Ángeles, y SANTILLÁN, Angélica María. Minería de datos: Concepto, características, estructura y aplicaciones. Journal-Universidad Autónoma de México, 2007.
- [4] Fayyad U., Piatetsky-Shapiro G., Smyth P. y Uthurusamy P. Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996.
- [5] BORIS K, Evgenii V. Data Mining for Financial Applications, the Data Mining and Knowledge Discovery Handbook, 2005.
- [6] Menezes, J., & Barreto, G. 2008. Long-term time series prediction with the NARX network: an empirical evaluation. Neuro computing 71, 3335–3343.
- [7] TAN, Pang-Ning; STEINBACH Michael and KUMAR, Vipin. Introduction to Data Mining, (First Edition).Addison-Wesley Longman Publishing Co. Inc., Boston, MA, USA, 2005.
- [8] CHEN, Ming-Syan; HAN, Jiawei, and YU, Philip S. Data mining: An over view from a database perspective. IEEE Trans. on Know and Data Eng., 8(6):866-883, 1996.
- [9] MAIMON, Odded and ROKACH, Lior. Data Mining and Knowledge Discovery Handbook. Springer, 2005.

- [10] J.P. and Fortes. Characterization and Parallelization of Decision-Tree introduction. Journal of Parallel and Distributed Computing, Vol 61, (Jan 2001); p. 322-349.
- [11] Stephens, C.R., & Sukumar, R. An Introduction to Data Mining. The Handbook of Market Research. Sage Publications. 2007; p. 455-486
- [12] OLSON, D. y DENLE, D. Parte 2. Data mining methods and tools. Berlin: Springer, 2008; p. 39-144
- [13] JAB, Michael; LINOFF, Gordon S. Data Mining Techniques, for Marketing Sales, and Customer Relationship Management, WILEY, 2004.
- [14] HAJIZADEH, Ehsan; DAVARI ARDAKANI, Hamed and SHAHRABI, Jamal. Application of data mining techniques in stock markets: A survey. Journal of Economics and International Finance Vol. 2(7).Industrial Engineering Department, Amirkabir University of Technology, Tehran, Iran.(July 2010); p. 109-118.
- [15] CRUZ - Iván. Pronósticos en el mercado de derivados utilizando redes neuronales y modelos ARIMA: una aplicación al Cete de 91 días en el Mex Der. Facultad de Contaduría y Administración. Universidad Nacional Autónoma de México, 2007.
- [16] Business Intelligence: Putting Enterprise Data To Work” de Dan Briody, miembro de The Economist Intelligence Unit, 2007.
- [17] CHERNG W, Muh; YU L, Sheng; HSIN L., Chia. An effective application of decision tree to stock trading, Expert Systems with Applications, 2006.
- [18] BASEL COMMITTEE ON BANKING SUPERVISION. History of the Basel Committee and its Membership, 2009.
- [19] LINARES VÁSQUEZ, Mario. Desarrollo de una herramienta computacional, basada en minería de datos, para Soporte al proceso de selección de portafolio. Artículo - Maestría en Ingeniería de sistemas y computación. Universidad Nacional de Colombia, 2006.
- [21] PAN, Jian-Xin; FUNG, Wing-Kam and FANG, Kai-Tai. Multiple outlier detection in multivariate data using projection pursuit techniques. Journal of Statistical Planning and Inference. Vol.83, 2000; p. 153-167.

- [22] SANTAMARÍA RUÍZ, Wilfredy y LEON GUZMAN, Elizabeth. Modelo de Detección de Fraude basado en el descubrimiento simbólico de reglas de clasificación extraídas de una red neuronal. Tesis de grado para optar por el título de Magíster en Ingeniería de Sistemas y Computación. Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia; 2010.
- [23] ORGULBAN. Ali. Financial Profiling for Detecting Operational Risk by Data Mining. Ankara; 2008.
- [24] FAYYAD, U. M. Data mining and knowledge discovery: making sense out of data. IEEE Expert, October, 1996; p. 20-25.
- [25] C.L Wilson, Self Organizing Neural Networks for trading common Stocks, in Proc IEEE Worl Congr. Neural Networks; 1994.
- [26] KOYUNCUGIL, Ali Serhan and OZGULBAS, Nermin. Financial Profilling for Detecting Operational Risk by Data Mining. World Academy of Science, Engineering and Technology 46; 2008.
- [27] BRAVO, Cristián y otros. Experiencias Prácticas en la Medición de Riesgo Crediticio de Microempresarios utilizando Modelos de Credit Scoring. Revista de Ingeniería de Sistemas. Volumen XXIV; junio de2010.
- [28] CABRALES, Sergio. Diseño de una metodología para la identificación y la medición del riesgo operativo en Instituciones Financieras. Universidad de los Andes. Departamento de Ingeniería Industrial. Bogotá; 2008.
- [29] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
- [30] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Wadsworth, Belmont CA, 1984.
- [31] AJ, Tahseen and MAB. A, Syed refined fuzzy time series model for stock market forecasting. PHYSICA A, Physica.
- [32] AGRAWAL R, IMILENSKI T. Swami A. Mining Association rules between sets of items in large database. In procceding of the ACM SIGMOD international conference on management of data; 1993.
- [33] MARTINEZ, Carlos. Uso de técnicas de procesamiento de datos e inteligencia artificial (lógica difusa) en la clasificación/predicción del riesgo bancario. Caso de

estudio: la banca comercial. Universidad de los Andes, Facultad de Ingeniería. Mérida, Venezuela; 2007.

[34] BAESENS Bart, SETIONO Rudy, MUES Christophe and VANTHIENEN Jan. Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. Management Science, Vol.49, No.3; 2003.

[35] E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Journal of the Decision Support Systems; 2010.

[36] SALEH MESBAH, Elkaffas and AHMED A. Toony. Applications of Genetic Programming in Data Mining. Journal of World Academy of Science, Engineering and Technology 17; 2006.

[37] TRIPPI Robert R., LEE Jae L. Artificial Intelligence in Finance & Investing, Ch 10, IRWIN, 1996.

[38] BANSAL Arun, KAUFFMAN Robert J., WEITZ Rob R. Comparing the Modeling Performance of Regression and Neural Networks as Data Quality Varies: A Business Value Approach, Journal of Management Information Systems, Vol 10. p. 11-32; 1993.

[39] Harrison, D, Personal Financial Planning: Theory and Practice, Financial Times Prentice Hall, Harlow, 2005

[40] Lasfer, Assia. Performance analysis of artificial neural networks in forecasting financial time series. Thesis of Master of Science in Engineering Systems Management, American University of Sharjah

[41] M. W. Craven, extracting comprehensible models from trained neural networks, Ph.D. dissertation, University of Wisconsin-Madison, 1996.

[42] Raquel Garetá, Luis M. Romeo, Antonia Gil, Alberto Monzón. Las redes neuronales como herramienta para predecir el precio de la energía eléctrica. Centro de Investigación de Recursos y Consumos Energéticos (CIRCE). Universidad de Zaragoza. Centro Politécnico Superior. María de Luna, 3, 50018 Zaragoza

[43] Delfiner, Miguel. La administración del riesgo de liquidez en el sistema financiero argentino. Banco Central de la República de Argentina, pp 9-15, agosto de 2007.

- [44] Predicción de series temporales con redes neuronales: una aplicación a la inflación colombiana. Revista Colombiana de Estadística, volumen 29 No.1. pp 77 a 92. Junio de 2006.
- [45] Kröse, Ben and Van Der Smagt, Patrick. "An Introduction to Neural Networks", The University of Amsterdam, 1996.
- [46] López Antón, Manuel. Requerimientos de información de los Analistas Financieros y Capacidad de predicción. Universidad Autónoma de Madrid, Septiembre 2010.
- [47] Banco de la República. Series estadísticas – tasas de captación. <http://www.banrep.gov.co/es/tasas-captacion> (Consulta: 30 sept. 2014)
- [48] Cepal. Diseminación de Estadísticas y Censos. <http://celade.cepal.org/redatam/index.es.html> (Consulta: 30 sept. 2014)
- [49] Departamento Administrativo Nacional de Estadística – DANE. [https://www.dane.gov.co/files/faqs/faq\\_ipc.pdf](https://www.dane.gov.co/files/faqs/faq_ipc.pdf) (Consulta: 30 sept. 2014)
- [50] Superintendencia Financiera de Colombia. Circular Externa 041 de 2002, Cap. XI. Fondos Comunes Ordinarios Administrados Por Las Sociedades Fiduciarias, Septiembre de 2002.
- [51] Superintendencia Financiera de Colombia. Circular Básica Contable, cap. VI. Reglas Relativas al Sistema de Administración del Riesgo de Liquidez. Noviembre de 2013.
- [52] Srinivasa K G \*, Venugopal K R 1 and L M Patnaik. Feature Extraction using Fuzzy C - Means Clustering for Data Mining Systems Data Mining IJCSNS International Journal of Computer Science and Network Security Volumen 6. 2006.
- [53] PechenizkiyM. The impact of feature extraction on the performance of a classifier: kNN, Naïve Bayes and C4.5AI'05 Proceedings of the 18th Canadian Society conference on Advances in Artificial Intelligence: Pages 268-279, 2005
- [54] Liu H.Motoda Hiroshi. Setiono R. y Zhao Zheng. Feature Selection: An Ever Evolving Frontier in Data Mining.Journal ofMachine Learning Research - Proceedings Track. 2010.