

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Modelo para la recomendación de fondos de inversión colectiva para personas naturales en el mercado financiero colombiano

Tomás Vergara Cardona

Universidad Nacional de Colombia
Facultad de Minas, Área Curricular de Sistemas e informática
Medellín, Colombia
2021

Modelo para la recomendación de fondos de inversión colectiva para personas naturales en el mercado financiero colombiano

Tomás Vergara Cardona

Trabajo de grado presentado como requisito para optar por al título de:

Magister en Ingeniería – Ingeniería Analítica

Director:

PhD. Fernán Alonso Villa Garzón

Universidad Nacional de Colombia

Facultad de Minas, Área Curricular de Sistemas e informática

Medellín, Colombia

2021

Resumen

Modelo para la recomendación de fondos de inversión colectiva para personas naturales en el mercado financiero colombiano

Los Fondos de Inversión Colectiva (FIC) son un mecanismo o vehículo de promoción del ahorro e inversión en el mercado de capitales nacional, a través de los cuales varias personas depositan recursos de manera colectiva de acuerdo con lo establecido en el reglamento del respectivo fondo. Un estudio realizado por ASOFIDUCIARIAS en 2018, evidenció que los FIC son un producto sobre el cual hay bajo conocimiento e incluso genera bastante confusión en la población colombiana. Actualmente, la difusión de este mecanismo es escasa; además, la asesoría que brindan las sociedades administradoras está sesgada bajo sus propios productos y no permiten evaluar la oferta total del mercado. Este trabajo propone un modelo de recomendación de FIC que permita de forma autónoma a través de un cuestionario estandarizado de perfilamiento de riesgo conectar a los usuarios con el universo de vehículos disponibles. Para desarrollar este modelo se realiza un proceso de depuración de la información disponible y el cálculo de las rentabilidades y volatilidades de los portafolios para alimentar tres algoritmos de agrupamiento o *clustering*. En este el trabajo se desarrolla un producto de datos para mapear el perfil de riesgo de un inversionista (persona natural) con una recomendación de un conjunto de FIC cuya rentabilidad y volatilidad estén acordes con el grado de aversión al riesgo que se acopla a la persona; el perfilamiento del riesgo del usuario, se realiza mediante el cuestionario “*Identifying your Investor Profile*” estandarizado y desarrollado por *Bank of America Corporation* ver Anexo 1; mientras que, el FIC se recomienda mediante un sistema de recomendación (algoritmo de *clustering*).

Palabras clave: Fondos de inversión colectiva, *clustering*, perfil de riesgo, rentabilidad, volatilidad.

Abstract

Model for the recommendation of collective investment funds for individuals in the Colombian financial market

Collective Investment Funds are an important mechanism for the promotion of savings and investment in the national capital market through which several people deposit resources collectively in accordance with the provisions of the respective fund's regulations. A study conducted by ASOFIDUCIARIAS in 2018 evidenced collective investment funds (CIF) as a product about which there is low knowledge and even generates quite a lot of confusion in the Colombian population. Currently, communication technologies are not used as a means of dissemination of these vehicles and the advice provided by the management companies are biased to their own products and do not allow evaluating the total market offer. This work proposes a collective investment fund recommendation model that allows autonomously, through a standardized risk profiling questionnaire, to connect users with the universe of available vehicles. To develop this model, a process of debugging of the available information, the calculation of returns and volatilities of the portfolios to feed three grouping or clustering algorithms is carried out. In this work a data product is developed to map the risk profile of an investor (natural person) with a recommendation of a set of CIFs whose profitability and volatility are in accordance with the degree of risk aversion of the person; the risk profiling of the user is done through the questionnaire "Identifying your Investor Profile" standardized and developed by Bank of America Corporation (see Annex 1); while the CIF is recommended through a recommendation system (clustering algorithm).

Key words: Mutual funds, grouping algorithms, clustering, risk profile, profitability, volatility.

Tabla de contenidos

Resumen.....	2
Lista de tablas.....	5
Lista de figuras.....	6
Introducción.....	7
1. Entendimiento del problema	10
2. Justificación	13
3. Objetivos.....	16
4. Exploración de la literatura	17
5. Metodología.....	20
6. Productos desarrollados	22
7. Entendimiento de los datos	24
7.1 Dimensión de los datos.....	24
7.2 Caracterización de los datos.....	25
7.3 Limpieza de los datos.....	26
7.4 Caracterización de los fondos de inversión colectiva	27
7.5 Variables derivadas.....	28
7.6 Exploración de las variables por categoría	30
8. Perfil de riesgo del inversionista	33
8.1 Cuestionario de perfil de riesgo.....	34
9. Modelado y Evaluación.....	35
9.1 Métrica de evaluación de los modelos de agrupamiento	35
9.2 Metodología de asignación del perfil de riesgo por grupo	36
9.3 Inventario de modelos disponibles	37
9.3.1 K-Means.....	37
9.3.2 Mini Batch K-Means.....	40
9.3.3 Hierarchical clustering	42
9.4 Resultados.....	44
10. Producto de datos	46
10.1 Esquema de bases de datos.....	47
11. Conclusiones.....	49
12. Referencias.....	52
Anexo 1. Cuestionario de perfilamiento de riesgo, Merrill Edge.....	54

Lista de tablas

Tabla 1. Principales referencias.....	17
Tabla 2. Fases de la metodología CRISP_DM	21
Tabla 3. Estructura de bases de datos rentabilidades FIC	24
Tabla 4. Variables agregadas a la base de datos	25
Tabla 5. Distribución de los fondos por categorías del SIFIC	30
Tabla 6. Ejemplo de la metodología de asignación del perfil de riesgo por clúster	37
Tabla 7. Agrupación por categoría con algoritmo K-Means.....	40
Tabla 8. Agrupación por categoría con algoritmo Mini Batch K-Means.....	42
Tabla 9. Agrupación por categoría con algoritmo Agglomerative Clustering.....	44
Tabla 10. Resultados modelos de agrupamiento.....	44
Tabla 11. Agrupamiento de fondos por clúster con el algoritmo seleccionado.....	45

Lista de figuras

Figura 1. Fases de la metodología CRISP-DM.....	20
Figura 2. Macrocategoría y categorías definidas en el SIFIC.....	28
Figura 3. Gráfico de distribución de la rentabilidad versus volatilidad por categoría.....	31
Figura 4. Gráfico de caja de porcentaje de la rentabilidad por categoría.....	31
Figura 5. Gráfico de cajas de porcentaje de la volatilidad por categoría	32
Figura 6. SEBI Riskometer	33
Figura 7. Gráfico de codo modelo K-Means.....	38
Figura 8. Gráfico de caja de porcentaje por clúster con K-Means.....	39
Figura 9. Gráfico de caja de porcentaje por clúster con Mini Batch K-Means.....	41
Figura 10. Gráfico de caja de porcentaje por clúster con Agglomerative Clustering.....	43
Figura 11. Tablero de visualización del proyecto, vista cuestionario perfilamiento de riesgo.	46
Figura 12. Tablero de visualización del proyecto, vista fondos agrupados por clúster	47
Figura 13. Esquema de bases de datos	48

Introducción

Las tecnologías de la comunicación y la información (TIC) han cambiado la forma en que las industrias en diferentes sectores económicos se relacionan con los consumidores. En el caso del sistema financiero, dichas tecnologías han facilitado la realización de diversos procesos y trámites recurrentes para un usuario habitual de una entidad bancaria, fomentando así mejoras en el nivel de bancarización de la población. A pesar de esto, las TIC no han sido un factor determinante en la profundización del sistema financiero colombiano ni en la educación financiera de la población. Como resalta la Comisión Intersectorial para la Educación Económica y Financiera en la Estrategia Nacional de Educación Económica y Financiera (ENEFF), en Colombia, se estima que el 64% de la población planifica para menos de un mes o no tiene planes financieros, el 58% tiene dificultades para cubrir sus gastos y solamente el 41% de la población de menos de 60 años ha tomado medidas para afrontar todos los gastos de la vejez (CIEEF, 2017).

Una de las dificultades a las que se enfrentan los usuarios (personales naturales) del sistema financiero, al querer llevar a cabo decisiones de ahorro e inversión, es el acceso limitado a una asesoría personalizada que les permita tener un entendimiento de la cantidad y la calidad de los productos disponibles en el mercado, así como la forma en que deben identificar su respectivo perfil de riesgo. Regularmente, para un usuario, no es evidente que el perfil de su riesgo como inversionista se determina en función de la edad, nivel de ingresos, liquidez, patrimonio acumulado, obligaciones financieras, conocimiento financiero y nivel de aversión al riesgo (Frenkel, 2003).

En este sentido, en el trabajo se desarrolla un modelo para vincular el perfil de riesgo de un inversionista, una persona natural, con una recomendación de un conjunto de FIC cuya rentabilidad y volatilidad estén acordes con el grado de aversión al riesgo que se acopla al usuario; las entradas del modelo son el perfilamiento del riesgo del usuario, el cual se obtiene a partir de una encuesta *“Identifying your Investor Profile”* estandarizada y desarrollada por *Bank of America Corporation* ver Anexo 1; mientras que, dado un perfil del usuario se le

recomienda un conjunto de fondos, identificado mediante un sistema de recomendación (*algoritmos de clustering*).

El trabajo explora tres modelos de *clustering* usados ampliamente en tareas de clasificación, como lo son el K-Medias y una variación de este usando *mini-batch*; adicionalmente, se explora el modelo de *hierarchical clustering*. Los tres modelos se comparan utilizando una métrica diseñada respecto a la medida de la oferta de fondos de inversión colectiva en el mercado colombiano y el sistema de categorización de estos desarrollado por ASOFIDUCIARIAS con el apoyo de todas las sociedades administradoras que lo componen. De acuerdo con esta métrica se elige el modelo con el cual se agrupa los fondos en cinco categorías o perfiles de riesgo: Conservador, Moderado Conservador, Moderado, Moderado Agresivo y Agresivo.

Teniendo en cuenta que en el trabajo se aplican técnicas de minería de datos, se hace uso de la metodología CRISP-DM. Esta metodología establece seis etapas en las cuales se desarrolla el ciclo de un producto de datos a través de una serie de actividades: se lleva a cabo la comprensión del negocio o problema abordado, el entendimiento de los datos y su preparación para la etapa del modelado, aplicación de técnicas de aprendizaje de máquinas, evaluación del desempeño de las técnicas aplicadas y un paso a producción en un sistema de información.

Este trabajo está estructurado de la siguiente manera: en el Capítulo 1 y 2 se aborda el entendimiento y la justificación del problema abordado, se plantean las carencias en la educación financiera y la necesidad de explotar los recursos tecnológicos disponibles para suplir carencias de las entidades financieras y de los usuarios. El Capítulo 3 plantea los objetivos del trabajo. En el Capítulo 4 se realiza una exploración de la literatura en materia de técnicas de aprendizaje de máquinas en sistemas de recomendación, referenciación en el perfilamiento de un inversionista y su aversión al riesgo, y el cuestionario utilizado en el trabajo. Los Capítulos 5 y 6 explican la metodología seleccionada para desarrollar el trabajo y los productos desarrollados. El Capítulo 7 plantea el entendimiento de los datos utilizados y del mismo se desprende una explicación de todo el flujo del proceso de actualización y transformación de los datos; finalmente, en este capítulo se explica detalladamente el sistema de categorización de los fondos de inversión colectiva. En el Capítulo 8, se desarrolla una breve explicación del perfil de

riesgo del inversionista y se profundiza en el cuestionario elegido para asignar uno de los perfiles a un inversionista. El Capítulo 9, contiene el inventario de modelos a desarrollar y evaluar en el trabajo y las métricas para la selección del modelo. En el Capítulo 10, se expone el producto de datos definitivo llevado a cabo con el módulo *Dash* del lenguaje de programación Python y el esquema de bases de datos que lo alimenta. Finalmente, el Capítulo 11 se presenta las principales conclusiones del trabajo desarrollado y algunas recomendaciones para seguir contribuyendo a la profundización del sector financiero en Colombia.

1. Entendimiento del problema

En la actualidad en Colombia, la población carece de los conocimientos y habilidades financieras mínimas requeridas para tomar decisiones económicas y financieras de forma responsable. Como resalta la Comisión Intersectorial para la Educación Económica y Financiera en la Estrategia Nacional de Educación Económica y Financiera (ENEEF) en Colombia se estima que el 64% de la población planifica para menos de un mes o no tiene planes financieros, 58% tiene dificultades para cubrir sus gastos y solo el 41% de la población de menos de 60 años ha tomado medidas para afrontar todos los gastos de la vejez (CIEEF, 2017).

Encuestas realizadas por el Banco de Desarrollo de América Latina (CAF) en el 2013 muestran que, aunque el 60% de los adultos colombianos lleva un presupuesto, controla su gasto y ahorra de manera cotidiana, la mayoría de ellos lo hace por fuera del sistema financiero. Esta informalidad en el manejo de los recursos por parte de los colombianos se explica en parte por la falta de educación financiera y también por las deficiencias en la capacidad de penetración de mercado de las entidades financieras. Si bien se resalta en un estudio realizado por la Asociación de Fiduciarias de Colombia que la percepción sobre los productos de ahorro e inversión son más “accesibles en todo sentido” gracias a la facilidad de encontrar más información en medios como internet o la publicidad realizada por los bancos, aún existe la percepción de que dichos contenidos tienden a ser poco claros y detallados (ASOFIDUCIARIA, 2018).

En el diseño de la ENEEF en Colombia se destacan algunas recomendaciones por parte de organismos multilaterales como la OCDE en cuanto a principios, lineamientos y buenas prácticas para mejorar los niveles de Educación Económica y Financiera (EEF) de la población. Para el desarrollo de este trabajo es importante destacar la recomendación que pide combinar diversos canales de difusión con mensajes que sean cortos, claros, contundentes, continuos, razonables y que generen confianza en los sistemas financieros y sus productos. Para cumplir con esta recomendación, en la práctica internacional se han desarrollado páginas de Internet

en los que se consolida y socializa al público en general las diferentes herramientas, contenidos y programas de EEF. Estas páginas deben estar enfocadas en permitirle al usuario consultar de forma digital información relevante respecto de las finanzas personales y del funcionamiento de los servicios financieros (CIEEF, 2017).

Entre los productos financieros disponibles para el ahorro y la inversión en el mercado colombiano, este trabajo se centra en los fondos de inversión colectiva, definidos por parte del Ministerio de Hacienda y Crédito Público en el decreto 1242 de 2013 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. Estos recursos serán gestionados de manera colectiva para obtener resultados económicos de manera colectiva”*. Estos instrumentos de ahorro e inversión son en esencia de fácil acceso para la población en general, tienden a tener mejores rendimientos que los Certificados de Depósitos a Término (CDT) y pueden ser redimidos en algunos de casos de manera inmediata. Sin embargo, el nivel de utilización doméstica de los FIC es muy bajo frente a los referentes regionales de la industria. Por ejemplo, el valor de los activos administrados por los FIC en Chile representa cerca del 18% del PIB, en Brasil el 50% y Estados Unidos el 150% frente a un 7.3% en Colombia (ANIF, 2016).

Un estudio realizado por ASOFIDUCIARIAS en 2018 evidenció a los FIC como un producto sobre el cual hay bajo conocimiento e incluso genera bastante confusión. Dicho estudio entrega como una de sus recomendaciones: acercarlos más al público en general y desmitificarlos, desarrollando planes de comunicación y promoción más constantes, claros, amigables y contundentes. Adicionalmente, se define un producto ideal como aquel que entre otros atributos tiene una orientación más clara sobre las mejores alternativas, con canales de consulta y administración más amigables (ASOFIDUCIARIA, 2018).

Luego de plasmar los bajos niveles de educación financiera en el país y las carencias de difusión y claridad en algunos de los productos de ahorro e inversión disponibles en el mercado, así como el nivel precario en volumen de recursos invertidos en FIC comparado con países como Chile y Brasil, se plantea el interrogante si en el país existe un sistema de información centralizado que permita al consumidor financiero visualizar el universo de fondos de

inversión colectiva disponibles, de forma amigable, con herramientas basadas en minería de datos y analítica que permitan identificar oportunidades de ahorro e inversión de acuerdo a las expectativas de rentabilidad y perfil de riesgo (ASOFIDUCIARIA, 2017).

Para ayudar a profundizar el nivel de conocimiento financiero en Colombia y con el fin de incentivar el ahorro y la inversión de personas naturales, este trabajo plantea un modelo de recomendación para los fondos de inversión colectiva del mercado colombiano, donde se pueda identificar todo el universo de opciones disponibles y se hagan recomendaciones de acuerdo con el plazo de la inversión y el perfil de riesgo, aportando desde la perspectiva de los canales de difusión a la solidificación de la educación económica y financiera en Colombia.

2. Justificación

Actualmente, en Colombia no existe un sistema de recomendación de Fondos de Inversión Colectiva (FIC) en el que las personas naturales puedan ingresar de forma autónoma a través de las tecnologías de la comunicación. Varios estudios han demostrado que es fundamental categorizar los fondos de inversión con el fin de profundizar el nivel de entendimiento que el usuario final tiene sobre éstos y así mejorar los niveles de ahorro e inversión en la población (ASOFIDUCIARIA, 2017). Bajo estos argumentos es que se considera relevante implementar modelos de recomendación basados en aprendizaje de máquinas que permitan a un usuario vincular su perfil de riesgo con la oferta disponible de mercado.

La industria de la inversión colectiva se considera como uno de los factores claves en el desarrollo de las economías a nivel internacional. Adicionalmente, la industria de los FIC se considera como uno de los principales dinamizadores en el proceso de desintermediación financiera y, por tanto, el desarrollo del mercado de capitales y una herramienta que fomenta la cultura del ahorro. A pesar de la reconocida importancia de este vehículo en la economía y del desarrollo del mercado de capitales en Colombia desde 1990, solo el 3% de la población colombiana invierte en FIC lo cual evidencia que las personas del país desconocen y no confían en estos vehículos (ANIF, 2016). La falta de educación financiera, las barreras para la adquisición y la escasa información disponible dificulta que en el país se alcancen las cifras de pares como Chile (12%) o España (22%) en cuanto a número de personas con FIC como alternativas de ahorro e inversión (ASOFIDUCIARIA, 2017).

En general, los colombianos no tienen el hábito de ahorrar e invertir parte de sus ingresos, por cuestiones culturales o falta de educación financiera, según un estudio del Banco de Desarrollo de América Latina CAF realizado en 2015, el 39% de los colombianos no ahorra y el 37% de los que si lo hacen guardan estos recursos en una alcancía, es decir, afuera del sistema financiero y sin generar ningún rendimiento. Adicionalmente, el 12% de los ahorradores colombianos consigna su dinero en una cuenta de ahorros o corriente, el 9% en “*cadena de ahorro privado*” que no hacen parte del sistema financiero y solo el 6% deja su capital en depósitos a término o CDT. La investigación también encontró que solo el 3% de los ahorradores nacionales invierte

el dinero en busca de mejores ganancias, pero no toda esta pequeña parte lo hace en el sistema financiero ya que también destinan estas inversiones a compra de vivienda y ganado (Zorrilla, 2016).

Fomentar una cultura de ahorro e inversión también puede representar a futuro un alivio para el déficit en el sistema pensional colombiano, de acuerdo con Bosch *et al.* (2015) de continuar con los parámetros del sistema pensional inalterados y una persistencia en la informalidad laboral para 2050 solo el 17% de los adultos mayores de 60 años tendrá una pensión contributiva (Bosch, Berstein, Castellani, Oliveri, & Villa, 2015). Estudios llevados a cabo por diferentes organismos multilaterales sugieren que Colombia migre el sistema pensional actual a un sistema multipilar en el cual existen varios módulos que cumplen funciones diferenciadas y complementarias, el sistema propuesto tiene cuatro pilares: tres contributivos y uno no contributivo (Fedesarrollo, 2018). El pilar no contributivo se determina como “Ahorro Voluntario” el cual, si bien suele implementarse con Fondos de Pensiones Voluntarias (FPV), puede encontrar en vehículos como los FIC una herramienta que ayude a estabilizar el déficit pensional colombiano a futuro.

Planteada la importancia de los fondos de inversión colectiva como herramienta de ahorro e inversión se justifica este trabajo desde la carencia de sistemas de información y modelos que permitan a los usuarios interactuar con la información puesta a disposición por parte de la Superintendencia Financiera de Colombia (SFC). Cada una de las organizaciones que administra los fondos de inversión colectiva es encargada de la distribución y promoción de sus productos, esto se debe a que regulatoriamente no se permite a un tercero promocionar los fondos administrados por las organizaciones autorizadas explícitamente por la ley, lo cual deja sin incentivo a la proliferación de todos los productos disponibles en el mercado.

Los fondos de inversión colectiva son vehículos financieros en los que confluyen los aportes de varios inversionistas, con cargo a los cuales se adquieren activos para estructurar un portafolio colectivo administrado por una sociedad especializada (AMV, 2019). Los fondos de inversión colectiva han sido un importante mecanismo de ahorro e inversión en el mercado de capitales nacional. De acuerdo con la regulación vigente existen dos tipos de FIC: los abiertos y los

cerrados. Por regla general todos los FIC son abiertos a no ser que su reglamento especifique lo contrario, esta característica representa que un inversionista podrá en cualquier momento que lo desee retirar los recursos (con las respectivas pérdidas o utilidades) del portafolio y la sociedad administradora tendrá un determinado número de días hábiles (previamente acordados en el reglamento) para hacer entrega de éstos. Los fondos que por reglamento son cerrados tienen un horizonte de inversión fijo que, al momento de cumplirse, brinda al inversionista la potestad de decidir si retira su inversión o renueva su participación por otro periodo de tiempo. A corte del 2018, habían 97.2 billones de pesos en activos administrados por FIC, 9.9% del PIB (Ministerio de Hacienda y Crédito Público, 2019).

3.Objetivos

Luego de comprender la poca profundidad del sistema financiero colombiano en los niveles de ahorro e inversión de los usuarios en parte por el uso limitado que se ha hecho de las tecnologías de la comunicación y la información, se busca a través de los siguientes objetivos desarrollar un modelo que conecte las necesidades del consumidor con la oferta disponible por parte de las sociedades administradoras de vehículos de inversión colectiva.

3.1 Objetivo General

Desarrollar un modelo para la recomendación y visualización de los fondos de inversión colectiva para personas naturales en el mercado financiero colombiano.

3.2. Objetivos Específicos

3.2.1 Caracterizar las fuentes de información disponible del cierre de los fondos de inversión colectiva reportado por los administradores a través de un proceso de extracción, depuración y categorización tomando como referencia las categorías determinadas por el SIFIC.

3.2.2 Implementar un modelo de clasificación basado en aprendizaje de máquinas que permita relacionar el perfil de riesgo del usuario con una categoría o conjunto de fondos de inversión colectiva.

3.2.3 Validar el modelo de clasificación implementado para la clasificación de fondos de inversión colectiva de acuerdo con el perfil de riesgo del usuario.

4. Exploración de la literatura

Como se hizo referencia en los capítulos anteriores de introducción y justificación, este trabajo explora la posibilidad de desarrollar, mediante algoritmos de aprendizaje de máquinas, un modelo de recomendación de inversiones en fondos de inversión colectiva para personas que no tengan la educación financiera suficiente para identificar su perfil de riesgo y el vehículo que mejor se adapta a sus necesidades.

Se realiza un proceso de exploración de la literatura en Google Scholar®; la búsqueda se realiza en el contexto de investigaciones y construcciones aplicadas al diseño de sistemas de recomendación de inversiones utilizando algoritmos de aprendizaje de máquinas, y el desarrollo de aplicaciones web con Python utilizando el módulo *Dash* de *Plotly*; en la Tabla 1, se relacionan los artículos, informes y libros relacionados. En general, la bibliografía disponible es reciente, dado que la implementación de soluciones analíticas para diferentes industrias viene teniendo un crecimiento importante en los últimos años. También, se resalta que una buena parte de la literatura proviene de autores o universidades de la India.

Titulo	Autores	Año	Tipo de documento
Identifying your Investor Profile	Merril Lynch	2013	Documento
Risk profile and mutual fund scheme selection of investors	Kapil Dev	2016	Articulo
Risk profile and Mutual Fund Scheme Selection of Investors	Kapil Dev	2016	Articulo
The use of machine learning algorithms in recommender systems: A systematic review	Donald Cowan, Paulo Alencar, Ivens Portugal	2018	Articulo
An empirical investigation of investor's behavioral biases on financial decision making	Imran Umer Chhapraa, Muhammad Kashifa, Raja Rehanb, Ashow Baia	2018	Articulo
A study on risk profiling and investment choices of retail investors	Kamini Khanna, Veena Chauhan	2019	Articulo
An Empirical Study on the Determinants of an Investor's Decision in Unit Trust Investment	Sanmugam Annamalah, Murali Raman, Govindan Marthandan and Aravindan Kalisri Logeswaran	2019	Articulo

A Conceptual Framework Proposed for Classifying Investor Behaviour	Ankita Pahwa, Maheshwari Sugam	2020	Artículo
Interactive Dashboards and Data Apps with Plotly and Dash	Elias Dabbas	2021	Libro

Tabla 1. Documentos relacionados

Las fuentes referenciadas en la revisión de la literatura sirven como sustento en las diferentes fases del trabajo. Para la implementación del modelo de recomendación a través de herramientas de aprendizaje de máquinas se tendrá como referencia la revisión sistemática de la literatura llevada a cabo en “*The use of machine learning algorithms in recommender systems: A systematic review*”. Este artículo presenta una revisión sistemática de la literatura que analiza el uso de algoritmos de aprendizaje de máquinas en sistemas de recomendación. El estudio concluye que los algoritmos más implementados para recomendar productos son la red bayesiana y el árbol de decisiones (Cowan, Alencar, & Portugal, 2018).

Respecto al proceso de entender el perfil de riesgo del usuario y las características que se deben priorizar para desarrollar el modelo se basa en el resultado de un cuestionario estandarizado por Merrill Lynch en “*Identifying your Investor Profile*”. Este cuestionario busca a través de nueve preguntas identificar el nivel de tolerancia al riesgo de un inversionista. Las preguntas del cuestionario se enfocan en perfilar el objetivo de inversión, el apetito por volatilidad en las inversiones, la capacidad de asumir pérdidas en el corto plazo con el objetivo de obtener resultados significativos en el largo plazo y el horizonte de tiempo en el cual se desea llevar a cabo la inversión. El sistema de puntuación de las respuestas explicado en la página cinco del cuestionario (ver Anexo 1) posiciona al inversionista en uno de los cinco perfiles de inversión predeterminados: conservador, moderado conservador, moderado, moderado agresivo y agresivo.

La estructura del cuestionario final también deberá tener en cuenta los aspectos relacionados con el área de *Behavioral Finance* y demás referencias, abarcado principalmente en los artículos:

- “*An empirical investigation of investor’s behavioral biases on financial decision making*”, (Umer, Kashifa, Rehanb, & Baia, 2018), este estudio tiene como objetivo determinar el impacto de los sesgos conductuales en la toma de decisiones de inversión en el mercado

de valores de Pakistán. Los resultados principales muestran que algunos comportamientos como el exceso de confianza, los sesgos cognitivos entre otros comportamientos tienen un impacto positivo significativo en las decisiones de inversión.

- “*A study on risk profiling and investment choices of retail investors*” (Khanna & Chauhan, 2019) esta investigación explora la relación entre algunos factores demográficos, el perfil de riesgo y las decisiones de inversión de los inversionistas. El estudio concluye que los inversionistas no siempre se comportan racionalmente y existen sesgos conductuales en sus inversiones.
- “*Risk Profile and Mutual Fund Scheme Selection of Investors*” (Dev, 2016) donde se explica la relación entre el perfil de riesgo del inversionista y el tipo de portafolio que debería escoger de acuerdo con este.

Finalmente, para el desarrollo del producto de datos, es decir, la interfaz que permita al usuario relacionarse con el modelo de recomendación y las demás visualizaciones implementadas se tomara como referencia el libro desarrollado por Elias Dabbas “*Interactive Dashboards and Data Apps with Plotly and Dash*” (Dabbas, 2021).

5. Metodología

En el desarrollo de este trabajo se usa un enfoque cuantitativo de diseño experimental ya que se realiza manipulación, experimentos y pruebas con los datos, se evalúan modelos iterando diferentes parámetros y se hará un esfuerzo significativo en la distribución o *Deployment*. La metodología de minería de datos seleccionada será CRISP-DM (*Cross Industry Standard Process for Data Mining*) la cual consiste en un modelo jerárquico de procesos, constituido por un conjunto de tareas organizadas en cuatro niveles de abstracción, que van desde el nivel más general hasta los casos más específicos. CRISP-DM organiza el desarrollo de un proyecto de explotación de datos, en una serie de seis fases, que constituyen el primer nivel de abstracción (Peralta, 2014). La Figura 1 y la Tabla 2 muestran el detalle de la metodología implementada y los capítulos donde se aplican cada una de las fases.

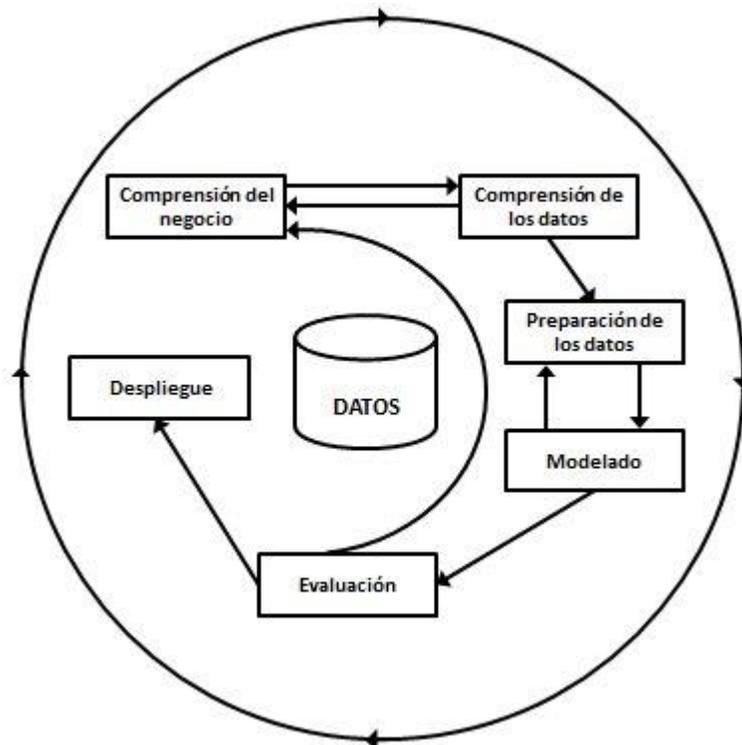


Figura 1. Fases de la metodología CRISP-DM (Peralta, 2014).

Fase	Definición
Comprensión del negocio	Se basa en el entendimiento de los objetivos del proyecto y la comprensión de los requerimientos de este desde el punto de vista del negocio. En los Capítulos 1, 2 y 3 se plantea la pregunta de investigación del problema, la justificación del trabajo y los objetivos a cumplir en el mismo.
Comprensión de los datos	Comienza con la recolección inicial de los datos con el propósito de familiarizarse con los mismos. El Capítulo 7, numeral 7.1 y 7.2 se explica de manera detallada la dimensión de los datos y las transformaciones necesarias para alimentar todo el proceso.
Preparación de los datos	Abarca todas aquellas actividades destinadas a la construcción del conjunto de datos finales sobre los cuales se aplicarán los modelos de caracterización. El Capítulo 7, numeral 7.3, 7.4, 7.5 y 7.6 muestran el detalle del procesamiento de los datos previo a implementar los modelos.
Modelado	Se seleccionan y aplican las técnicas de modelado más apropiadas para el proyecto en cuestión, calibrando sus parámetros a valores óptimos. El Capítulo 9 presenta los modelos evaluados y sus respectivos resultados.
Evaluación	Involucra la evaluación del proyecto y revisión de los pasos ejecutados en relación con los objetivos del negocio. Esta fase permite identificar si se ha omitido algún objetivo del proyecto y si los resultados serán implementados en la fase final. El Capítulo 9 presenta los modelos evaluados y sus respectivos resultados.
Despliegue	Permite al cliente o usuario final del proyecto conocer los resultados. El Capítulo 10 explica el detalle del modelo de visualización del proyecto.

Tabla 2. Fases de la metodología CRISP-DM (Peralta, 2014)

6.Productos desarrollados

La comercialización de los fondos de inversión colectiva en Colombia se desarrolla principalmente de forma individual por cada uno de los gestores, los cuales únicamente brindan asesoría acerca de sus productos y no permiten al usuario del sistema financiero identificar el universo de productos disponibles en el mercado y los diferentes niveles de riesgo y rentabilidad a los que se puede exponer para lograr sus metas de ahorro e inversión. Con la implementación de este trabajo se desarrolla un acercamiento inicial de lo que podría considerarse un sistema de asesoría virtual donde un usuario conecte sus necesidades con la oferta de instrumentos de ahorro e inversión del sistema financiero colombiano.

Es importante resaltar que este trabajo es con fines académicos y que no tiene como fin reemplazar en el corto plazo con los modelos de agrupamiento y perfilamiento de riesgo, el deber de asesoría profesional que exige la regulación colombiana en la comercialización y recomendación de inversiones. Para recomendar a un potencial inversionista si un determinado fondo le es o no conveniente, el distribuidor o asesor debe establecer su perfil de riesgo y necesidades de acuerdo con la información que este le suministre, adicional a esto, la asesoría profesional solo puede ser prestada a través de un profesional debidamente certificado por el Autorregulador del Mercado de Valores (AMV) e inscrito en el Registro Nacional de Profesionales del Mercado de Valores (RNPMV), dicho requisito es necesario para cualquiera que sea el medio que se emplee para la distribución de los FIC (AMV, 2019).

Al desarrollar este trabajo se han obtenido los resultados presentados a continuación:

- Implementación de una base de datos a través de una actualización periódica de la información pública correspondiente a los cierres diarios de los fondos de inversión colectiva transmitidos por las sociedades administradoras a la Superintendencia Financiera de Colombia (SFC). Esta implementación permite identificar cada una de las participaciones que componen los fondos y asociarlas con su respectiva categoría de mercado.
- Cálculo de las métricas de retornos y volatilidad en marcos de tiempo similares que permitan una comparación justa de la relación riesgo/volatilidad de los fondos analizados.
- Encuesta de perfilamiento de riesgo, implementada en Python, que permita a un usuario identificar su nivel de aversión al riesgo y el tipo de activos en lo que debería invertir de acuerdo con este.

- Implementación de modelo de agrupamiento que permita identificar de acuerdo con la relación riesgo/rentabilidad de cada portafolio resultados similares en la gestión de los administradores del fondo y por lo tanto permitan que un usuario identifique con mayor facilidad que tipos de FIC existen y que oportunidades de inversión hay disponibles.
- Implementación de la vinculación del resultado de la encuesta de perfilamiento de riesgo con los grupos obtenidos en el modelo de agrupación de forma que el usuario pueda visualizar de forma automática los fondos que mejor se acoplan a su nivel de aversión al riesgo.

7. Entendimiento de los datos

7.1 Dimensión de los datos

A continuación, se indica la estructura de la información disponible. Los datos presentados en la Tabla 3 son transmitidos periódicamente por las entidades financieras autorizadas para la administración de los fondos de inversión colectiva (Sociedades Comisionistas de Bolsa, Sociedades Fiduciarias y Sociedades Administradoras de Inversión) a la Superintendencia Financiera de Colombia y luego publicados a través de diferentes plataformas. Para este trabajo se utilizará la API disponible en <https://www.datos.gov.co>.

Nombre de Columna	Tipo
FECHA_CORTE	Fecha y hora
TIPO_ENTIDAD	Número
NOMBRE_TIPO_ENTIDAD	Texto simple
CODIGO_ENTIDAD	Número
NOMBRE_ENTIDAD	Texto simple
TIPO_NEGOCIO	Número
NOMBRE_TIPO_PATRIMONIO	Texto simple
SUBTIPO_NEGOCIO	Número
NOMBRE_SUBTIPO_PATRIMONIO	Texto simple
CODIGO_NEGOCIO	Número
NOMBRE_PATRIMONIO	Texto simple
PRINCIPAL_COMPARTIMENTO	Número
TIPO_PARTICIPACION	Número
RENDIMIENTOS_ABONADOS	Número
PRECIERRE_FONDO_DIA_T	Número
NUMERO_UNIDADES_FONDO_CIERRE_OPER_DIA_T_ANTERIOR	Número
VALOR_UNIDAD_OPERACIONES_DIA_T	Número
APORTES_RECIBIDOS	Número
RETIROS_REDENCIONES	Número
ANULACIONES	Número
VALOR_FONDO_CIERRE_DIA_T	Número
NUMERO_INVERSIONISTAS	Número
RENTABILIDAD_DIARIA	Número

RENTABILIDAD_MENSUAL	Número
RENTABILIDAD_SEMESTRAL	Número
RENTABILIDAD_ANUAL	Número

Tabla 3. Estructura base de datos rentabilidades FIC

7.2 Caracterización de los datos

Luego de desarrollar la conexión vía Python utilizando el módulo *Socrata* de *sodapy* que permite a través de sentencias llamar los datos acotando la información por fecha y por fondo. Del proceso de descarga se excluyen desde el principio los Fondos de Capital Privado, ya que este tipo de fondos se caracterizan por ser cerrados a un mediano o largo plazo y tienen como objetivo obtener retornos principalmente de las inversiones en empresas y proyectos que no cotizan en bolsa (URF, 2018).

Los administradores de los FIC diariamente actualizan la base de datos en <https://www.datos.gov.co>. Una vez se tienen los datos actualizados, se procede a agregar una variable que permita diferenciar cada uno de los fondos de forma única y otra variable adicional que, al interior de cada fondo, permita identificar cada uno de los diferentes tipos de participaciones que lo componen. Bajo un mismo reglamento y plan de inversiones, los fondos de inversión colectivas pueden tener diferentes tipos de participaciones creadas en conformidad con los tipos de inversionistas que vinculan (ASOFIDUCIARIAS, 2018), cada tipo de participación otorga deberes y derechos diferentes a los inversionistas que los adquieran, como por ejemplo el monto de comisiones que cobra el gestor por la administración del portafolio, es por esto que en un mismo fondo, las participaciones tienen valores de unidad diferentes y por lo tanto diferentes rentabilidades en el mismo periodo de tiempo. Para distinguir tanto fondo como participaciones la Tabla 4 muestra las dos variables:

Nombre de Columna	Tipo
CODIGO_FIC	Texto simple
CODIGO_CLASE	Texto simple

Tabla 4. Variables agregadas a la base de datos

Estas variables son una combinación de otras variables que vienen de la base original. Para la variable CODIGO_FIC se combinan: TIPO_ENTIDAD, CODIGO_ENTIDAD y CODIGO_NEGOCIO. Para construir CODIGO_CLASE se toma la misma combinación anterior y se adiciona al final TIPO_PARTICIPACION.

7.3 Limpieza de los datos

Una vez se tienen todos los datos y las variables de identificación adicionales, se procede a explorar la completitud de estos. El modelo de agrupamiento a implementar en este trabajo, todas las métricas se derivan del valor de unidad, en la base de datos la variable VALOR_UNIDAD_OPERACIONES_DIA_T, el valor de unidad se define en la Circular Básica Contable y Financiera (Circular Externa 100 de 1995) como el valor de los aportes de los inversionistas y representan cuotas partes del valor patrimonial del respectivo fondo. El mayor valor de la unidad representa los rendimientos que se han obtenido. En términos generales, el valor de un FIC es el resultado de los rendimientos más los aportes y demás ingresos, menos los retiros y gastos de este, este resultado se divide entre el número de unidades del fondo, para así determinar el valor de unidad. Las demás variables corresponden a información descriptiva de cada fondo y aunque dichas características pueden influir en la gestión del fondo no afectan el cálculo del valor de la unidad y por lo tanto no intervienen en el cálculo de la rentabilidad y la volatilidad del portafolio.

Entonces, la limpieza de los datos se lleva a cabo identificando los fondos cuyo valor de unidad no cuenta con actualizaciones recientes, definiendo el concepto de reciente para este trabajo como más de quince días calendario anteriores a la fecha en la que se actualice la información. Para llevar a cabo el procesamiento de modelamiento también se considera que se necesitan fondos que cuenten con la historia suficiente para tener datos consolidados de rentabilidad y volatilidad robustos y muestren el comportamiento del fondo en los últimos tres años, para esto se filtran de la base original los fondos que cuentan con más de 1095 registros.

Finalmente, se decide excluir los fondos que en la última actualización disponible tengan menos de diez inversionistas, esto debido a que muchos de estos fondos corresponden a portafolios en proceso de liquidación, este proceso de liquidación puede generar ruido en cuanto a rentabilidades y volatilidad, lo cual genera datos atípicos que generan ruido en el resultado de los algoritmos de agrupamiento.

Se resalta que por características propias de la base de datos en el proceso de limpieza de estos no se cuenta con el problema de los valores faltantes, ya que cuando un fondo es cerrado o liquidado simplemente se dejan de actualizar sus datos y al llevar a cabo los filtros previamente mencionados estos fondos quedan descartados del proceso de actualización de la información.

Además, cabe destacar que la variable valor de unidad no cuenta con datos atípicos, dado que el valor de unidad es una característica propia de cada fondo, que varía de acuerdo con los resultados de cada portafolio y que en su nacimiento tiene un valor a discreción del administrador y que es indiferente su nivel a las rentabilidades y volatilidades obtenidas. Por ejemplo, un fondo puede nacer con un valor de unidad de 10.000 y otro fondo con el mismo plan de inversiones y un reglamento similar puede empezar con un valor de unidad de 100 y no habrá diferencias en cómo se miden los resultados.

7.4 Caracterización de los fondos de inversión colectiva

Con el propósito de organizar y hacer comparable la información de la industria de Fondos de Inversión Colectiva en Colombia, los gremios que representan a los administradores ASOFIDUCIARIAS y ASOBOLSA con el acompañamiento técnico de LVA Índices, decidieron en el 2018 crear un sistema de categorización cuyo objetivo principal es crear grupos homólogos de FIC con objetivos de inversión comparables, estableciendo criterios que sirvan de guía a las sociedades administradoras para que la clasificación de los respectivos fondos que administre sea coherente con la forma en que deben promocionarse dichos vehículos a los diferentes grupos de interés, entre ellos, las personas naturales (SIFIC, 2018).

Se aclara que no todos los administradores se acogieron a cumplir con los criterios necesarios para entrar en el sistema de categorización y, de los que lo hicieron, no la totalidad de sus vehículos hacen parte de alguna categoría. Entonces, en el presente trabajo existe la categoría “Otros Fondos” que se define por SIFIC y la categoría “Sin Categoría” que se asignó a los fondos disponibles en la base de datos que no se encuentran en ninguna categoría del SIFIC.

En la Figura 2, se observan las macro categorías y categorías definidas en el SIFIC. En la macro categoría Renta Fija se encuentra gran parte de los FIC disponibles en el mercado, los que pertenecen a las categorías de liquidez, corto plazo y mediano plazo se asocian a un menor nivel de riesgo y los de largo plazo y alto rendimiento tienden a tener mayor volatilidad. En los fondos balanceados se encuentran los tres niveles de riesgo vía portafolios que combinan diferentes

tipos de activos. Los fondos inmobiliarios suelen ser fondos cerrados, de baja volatilidad y uno o varios dividendos al año. Los fondos accionarios y bursátiles se consideran portafolios de alto riesgo ya que se invierten totalmente en activos de renta variable nacional o internacional.

MACROCATEGORÍA	CATEGORÍA
 Fondos de Renta Fija	Nacional de Liquidez
	Nacional para Entidades Públicas
	Nacional de Corto Plazo
	Nacional de Mediano Plazo
	Nacional de Largo Plazo
	Nacional de Alto Rendimiento
	Internacional
 Fondos Balanceados	Menor Riesgo
	Moderados
	Mayor Riesgo
 Fondos Accionarios	Nacionales
	Internacionales
 Fondos Bursátiles	Accionarios Beta
	Accionarios Smart Beta
 Fondos Inmobiliarios	de Desarrollo Nacional
	de Rentas Nacionales
 Otros Fondos	Fondos de Activos de Crédito
	Otros Fondos Accionarios
	Otros Fondos

Figura 2. Macrocategoría y categorías definidas en el SIFIC (SIFIC, 2018).

7.5 Variables derivadas

La base de datos permite identificar para cada uno de los fondos el valor de unidad a una fecha de corte determinada. Como se mencionó anteriormente el valor de unidad representa la totalidad de los resultados de la gestión de los recursos del portafolio por parte de la sociedad administradora menos los costos y gastos asociados, principalmente el valor de la comisión que diferencia las participaciones. La Circular Básica Contable y Financiera (Circular Externa 100 de 1995) en su anexo XI Valoración de las Carteras Colectivas define de forma estándar la metodología con la que se debe calcular el valor de la unidad de un FIC. Adicionalmente, define la forma en que se debe calcular la rentabilidad efectiva anual para un periodo determinado:

$$Rp(x, y) = \left(\frac{VUO_y}{VUO_x} \right)^{(365/n)} - 1$$

Ecuación 1. Rentabilidad efectiva anual

Donde:

$Rp(x, y)$ = Rentabilidad efectiva anual para el periodo comprendido entre los días x e y.

VUO_y = Valor de la unidad para las operaciones del último día del periodo de cálculo.

VUO_x = Valor de la unidad para las operaciones del primer día del periodo de cálculo.

n = Número de días durante el lapso x e y.

En los modelos financieros el riesgo es una variable fundamental. El riesgo de una inversión se asocia a la amplitud de las fluctuaciones de los rendimientos en el tiempo, esto ha llevado a caracterizarla a través de una medida de dispersión de las variables, siendo la más frecuentemente utilizada la desviación típica o estándar (Robles Fernández, 1999). De esta forma se define la variable de volatilidad como la desviación estándar de un año de rentabilidades del valor de unidad de cada una de las participaciones de los fondos seleccionados:

$$\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{X})^2}{N}}$$

Ecuación 2. Desviación estándar (volatilidad)

Donde:

σ = Desviación estándar

N = Numero de observaciones

\bar{X} = Rentabilidad promedio del periodo observado

x_i = Observación numero i de la variable X

Al incluir la rentabilidad efectiva anual y la volatilidad para un periodo de tiempo específico, se tienen los inputs necesarios para alimentar un modelo de agrupamiento que permita diferenciar cada uno de los fondos de inversión colectiva en un grupo respectivo. Para este trabajo se asume que la rentabilidad y la volatilidad son las características que definen totalmente el compartimiento del portafolio de un fondo y a partir de las cuales se puede hacer una conexión directa con el perfil de riesgo del inversionista. Se descartan otras características

de los fondos como el valor de este, el número de inversionistas y la comisión cobrada ya que estas dependen principalmente de la red comercial del administrador.

7.6 Exploración de las variables por categoría

Una vez llevadas a cabo las fases anteriores y creadas todas las variables necesarias para el modelamiento, se procede con la descripción y el análisis de los datos para un mejor entendimiento de su comportamiento. Primero, se verifica como se distribuyen las categorías asociadas para todos los fondos que componen la base de datos. En la Tabla 5, se puede observar cómo el 69,25% de los fondos de la base de datos no está aún adherido al sistema de categorización elaborado por SIFIC.

Categoría	Número de Fondos	% Fondos Totales
Accionario Internacional	3	0,75%
Accionario Nacional	11	2,75%
Activo de Crédito	9	2,25%
Alto Rendimiento Nacional	6	1,50%
Balanceado Mayor Riesgo	2	0,50%
Balanceado Menor Riesgo	4	1,00%
Balanceado Moderado	3	0,75%
Bursátiles Accionarios Beta	2	0,50%
Inmobiliario de Desarrollo Nacional	3	0,75%
Inmobiliario de Renta Nacional	6	1,50%
Otros Fondos Bursátiles	3	0,75%
Renta Fija Internacional	7	1,75%
Renta Fija Nacional para Entidades Publicas	18	4,50%
Renta Fija Pesos de Corto Plazo	12	3,00%
Renta Fija Pesos de Largo Plazo	7	1,75%
Renta Fija Pesos de Liquidez	18	4,50%
Renta Fija Pesos de Mediano Plazo	9	2,25%
Sin Categoría	277	69,25%

Tabla 5. Distribución de los fondos por categoría del SIFIC

De la Figura 3, se observa la distribución de cada uno de los fondos por categoría en el plano de rentabilidad versus volatilidad y cómo por lo general existe una cercanía entre los fondos que

componen cada una. Se destaca que gran parte de los fondos se ubican en la parte inferior izquierda, es decir, baja rentabilidad y volatilidad que se asocia a los fondos de renta fija.

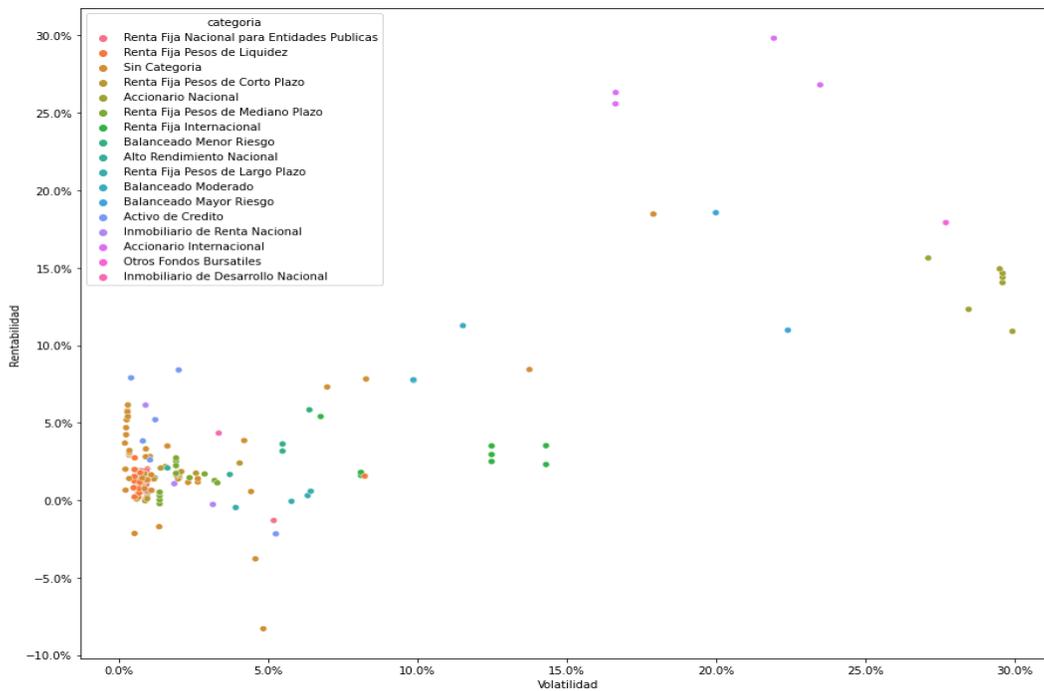


Figura 3. Gráfico de distribución de la rentabilidad versus volatilidad por categoría.

En la Figura 4 y 5 de cajas de porcentaje de la rentabilidad y volatilidad respectivamente por categoría muestra como en la mayor parte de las categorías no hay una distancia relevante entre el primer y el tercer cuartil y nos son comunes los datos atípicos en comparación con los fondos “Sin Categoría”.

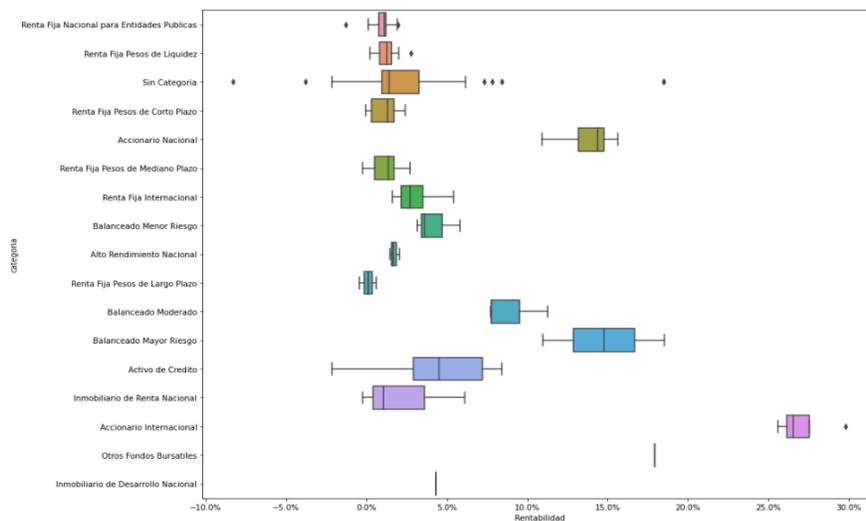


Figura 4. Grafico de cajas de porcentaje de la rentabilidad por categoría.

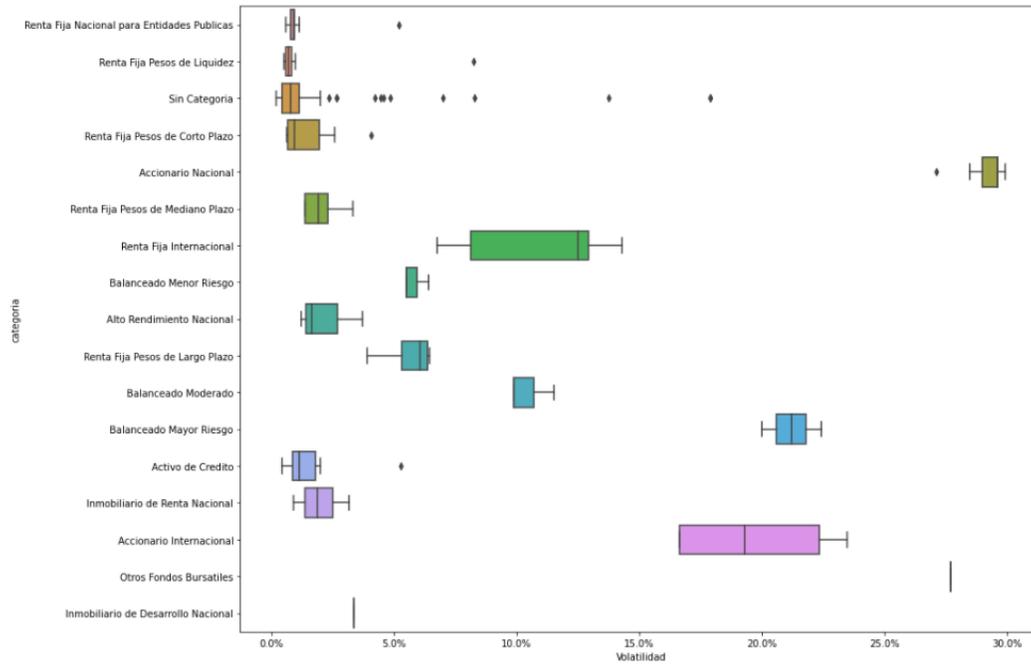


Figura 5. Gráfico de cajas de porcentaje de la volatilidad por categoría.

8. Perfil de riesgo del inversionista

Al momento de realizar una inversión, un inversionista informado o un profesional que esté llevando a cabo una asesoría debe tener en cuenta el perfil de riesgo del inversionista. Este perfil se puede entender como la suma de dos partes: el horizonte de tiempo y la tolerancia al riesgo. El horizonte de tiempo indica el plazo durante el cual el inversionista planea sostener una inversión y la tolerancia al riesgo se relaciona como la capacidad que tiene el inversionista de absorber las pérdidas de su portafolio (Dev, 2016).

Una guía que indica cómo identificar los diferentes niveles de riesgo en los que se puede clasificar a un inversionista es el propuesto por la *Securities and Exchange Board of India* (SEBI) en 2015 a través de una circular que obliga a los administradores de fondos de inversión a marcar sus productos, como se puede apreciar en la Figura 6.

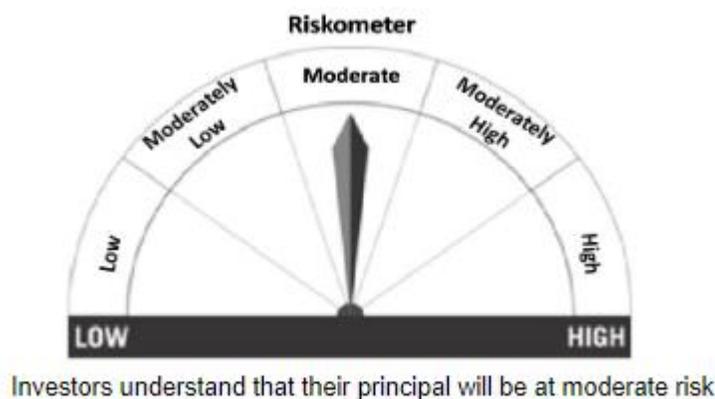


Figura 6. SEBI Riskometer (Dev, 2016)

Estos niveles de riesgo van desde el bajo hasta el alto, haciendo una distinción de moderadamente bajo, moderado y moderadamente alto. El nivel bajo se asocia con inversiones de muy corto plazo gestionados principalmente con fondos de liquidez, el nivel moderadamente bajo se asocia con inversiones de uno a tres años y se gestiona con títulos de renta fija con duraciones cortas o medias, para las inversiones moderadas ya se comienzan a incluir activos diferentes a renta fija y se orienta principalmente a inversionistas que quieren generar retornos considerables pero están preocupados por la seguridad de sus recursos, el perfil de riesgo moderadamente alto aplica a inversionistas con un horizonte de tiempo alto y se gestiona a través de portafolios balanceados con alta orientación a activos de renta variable, finalmente,

las inversiones de riesgo alto son también de tiempo largo y tienden a tener una alta probabilidad de pérdidas pero un alto potencial de generar ganancias excepcionales en el largo plazo (Dev, 2016).

En este trabajo sigue una estructura similar a la propuesta por el SEBI, cambiando el nombre bajo por conservador y alto por agresivo, quedando de la siguiente manera:

- Conservador
- Moderadamente conservador
- Moderado
- Moderadamente agresivo
- Agresivo

La definición de los perfiles de riesgo por parte del SEBI permite conectar a través de la encuesta de perfilamiento propuesta por *Merril Edge* con las personas naturales que por falta de educación financiera desconozcan cual es la decisión de inversión apropiada de acuerdo con su idiosincrasia o sus características demográficas. Ahora, utilizando los datos de rentabilidad y volatilidad se puede alimentar un algoritmo de agrupamiento que permita conectar con los fondos de inversión colectiva disponibles en el mercado.

8.1 Cuestionario de perfil de riesgo

En la búsqueda de conectar o vincular las necesidades de un usuario o inversionista con los perfiles de riesgo propuestos, este trabajo toma como ejemplo la metodología planteada por la firma *Merril Edge* para determinar la tolerancia al riesgo de un inversionista usando un cuestionario de nueve puntos. El objetivo es que el cuestionario se realice individual y autónomamente, evaluando aspectos como los objetivos de la inversión, máxima pérdida esperada, selección de portafolios, horizonte de tiempo y monto a invertir. Los detalles del cuestionario pueden observarse en el Anexo 1, *Identifying Your Risk Profile*.

9. Modelado y Evaluación

La fase de modelado se lleva a cabo utilizando Python, donde se usa algoritmos de *clustering* de la librería *scikit-learn* para la aplicación de transformaciones y aplicaciones de aprendizaje de máquinas para técnicas de aprendizaje supervisado y no supervisado. En este Capítulo se exponen los modelos de agrupamiento seleccionados para evaluar el nivel de coincidencia del algoritmo con las categorías previamente establecidas por el SIFIC.

El agrupamiento o *clustering* es una técnica que divide un set de datos en un número específico de grupos de características similares entre ellos y disimiles comparados con los datos de otro grupo (Abu Abbas, 2008). Los algoritmos de agrupamiento se dividen en dos grupos, los jerárquicos y los particionales (Zhao & Karypis, 2002). Se considera un modelo particional (*K-Means*) y uno jerárquico (*Agglomerative clustering*); adicionalmente, se evalúa una variación del modelo particional *K-Means* que utiliza pequeños lotes de información para reducir el tiempo computacional buscando optimizar la misma función objetivo.

9.1 Métrica de evaluación de los modelos de agrupamiento

Con el fin de validar el modelo, es decir, medir la certeza de los algoritmos de agrupamiento sobre los fondos de inversión colectiva, se crea una métrica que determine el número de clústeres por categoría del SIFIC. El objetivo es que el algoritmo seleccionado tenga la más alta semejanza posible con el sistema de categorización establecido por los administradores. Este criterio de evaluación se define como la sumatoria del número de clústeres asignados a una misma categoría sobre el número de fondos que componen dicha categoría:

$$\text{Criterio Evaluacion} = \sum \frac{\# \text{Clústeres}}{\# \text{Fondos}}$$

Ecuación 3. Criterio de evaluación certeza agrupamiento.

Este criterio que estará entre 0 y 1 prefiere a los algoritmos que en una misma categoría les asignen a los fondos que la componen un solo grupo, es decir, si una categoría tiene tres fondos el número ideal sería 0.333, ya que para los tres fondos de similares características el algoritmo solo asigno una categoría. Por el contrario, el peor escenario posible sería un criterio de 1,

donde a los tres fondos mencionados el algoritmo les asigne tres grupos diferentes. De esta manera, el valor total del criterio de evaluación será la suma de cada una de las 17 categorías existentes y se seleccionará el algoritmo que presente el mínimo valor.

9.2 Metodología de asignación del perfil de riesgo por grupo

Como resultado de los algoritmos de agrupamiento se tienen cinco grupos (en la sección 9.3 se explica por qué son cinco) asignados a cada uno de los fondos acuerdo a la rentabilidad y la volatilidad de sus respectivos portafolios en los últimos años. Cada uno de estos grupos tiene unas características similares de rentabilidad y volatilidad y se hace necesario conectarlas con los perfiles de riesgo mencionado en el Capítulo 8. Para esto se crea una variable adicional en la base de datos consolidada (contiene las rentabilidades y volatilidades históricas de cada participación de los FIC) con las rentabilidades y volatilidades de cada fondo. Esta variable no sigue una lógica financiera ya que su objetivo es ponderar los grupos resultantes con un indicador que combine las otras variables que alimentan el set de datos. Este indicador (Ecuación 3) no hace parte de los datos que alimentan los algoritmos de agrupamiento y solo se usa para asignar los perfiles de riesgo a cada clúster.

$$\text{Risk Return Ratio} = (|\text{Rentabilidad}| * 0.5) + (\text{Volatilidad} * 0.5)$$

Ecuación 4. Risk return ratio

Cuando se obtiene el grupo que le corresponde a cada fondo, se procede a agrupar cada uno de los grupos calculando las medias aritméticas de la rentabilidad, la volatilidad y el *risk ratio* creado. Esta agrupación da como resultado una matriz con cinco entradas, esta matriz se ordena de forma ascendente en el criterio *risk return ratio* (RR) de forma que se le asigne el perfil de riesgo Conservador al que quede como el primer grupo y el agresivo para el último como se muestra en la Tabla 6.

Cluster	Fondo	categoria	Rentabilidad	Volatilidad	RR	Perfil
0	152	9	0.016091	0.010878	0.013804	Conservador
2	18	8	0.017552	0.061089	0.048000	Moderado Conservador
4	9	3	0.055556	0.123515	0.089536	Moderado
1	8	3	0.218794	0.196214	0.207504	Moderado Agresivo
3	8	2	0.143522	0.289421	0.216472	Agresivo

Tabla 6. Ejemplo de la metodología de asignación del perfil de riesgo por clúster.

De la Tabla 6, se evidencia que los portafolios con las menores rentabilidades y volatilidades gracias a la jerarquía asignada por la variable *risk return ratio* permite que el perfil de riesgo conservador sea asignado a ese conjunto de fondos con menor rentabilidad y volatilidad, casi siempre asociados a las estrategias más conservadoras. De igual forma, al utilizar el valor absoluto de la rentabilidad en la Ecuación 3, se evita que los fondos con rentabilidades muy negativas queden en el primer lugar, ya que las rentabilidades de gran magnitud (negativa o positiva) suelen estar asociadas a estrategias de inversión agresivas.

9.3 Inventario de modelos disponibles

9.3.1 K-Means

El algoritmo de *K-Means* es un modelo clásico de agrupamiento perteneciente a la categoría de algoritmos particionales. El algoritmo sigue una forma simple de definir una base de datos en k grupos fijados a priori. La idea principal es definir k centroides (uno para cada grupo) y luego tomar cada punto de la base de datos y situarlo en la clase del centroide más cercano, luego se recalculan los centroides de cada grupo minimizando la distancia entre datos y centroide hasta que ya no haya cambios en los grupos de un paso al siguiente (Pascual, Pla, & Sánchez, 2007).

Para la estimación de los grupos se utiliza la función *KMeans* del módulo *sklearn.cluster*. El modelo tiene como parámetro el número de centroides o clústeres, este valor se determina usando el gráfico de codo en la Figura 7 que permite identificar el número de centroides (k)

óptimos. El criterio del codo mide las distancias al cuadrado de cada uno de los datos con su centroide, iterando el número de centroides gráficamente podemos visualizar como a partir de cierto número de centroides el valor de la distancia con su centro tiene un cambio marginal, se aplana la curva, el punto de inflexión de dicho grafico se considera el numero óptimo de centroides (Bholowalia & Kumar, 2014). En la Figura 6, se observa como a partir de $k = 5$, el valor de la suma de las distancias con el centroide de su grupo no tiene cambios significativos y la curva tiende a aplanarse.

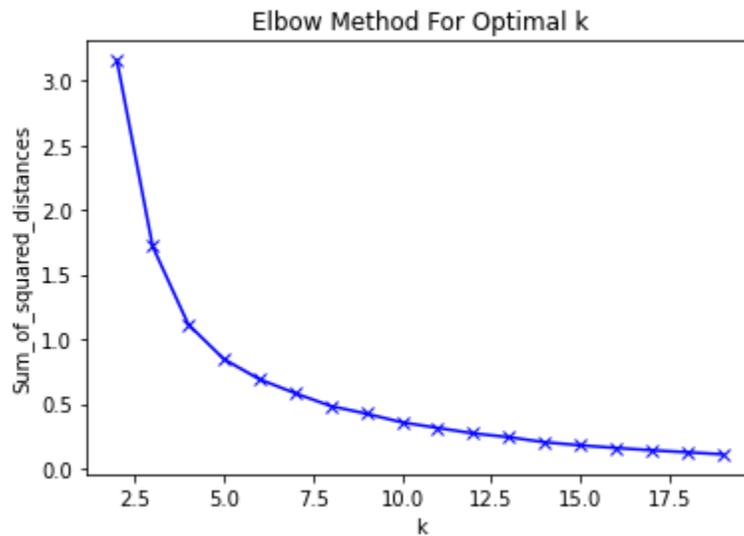


Figura 7. Gráfico de codo modelo K-Means

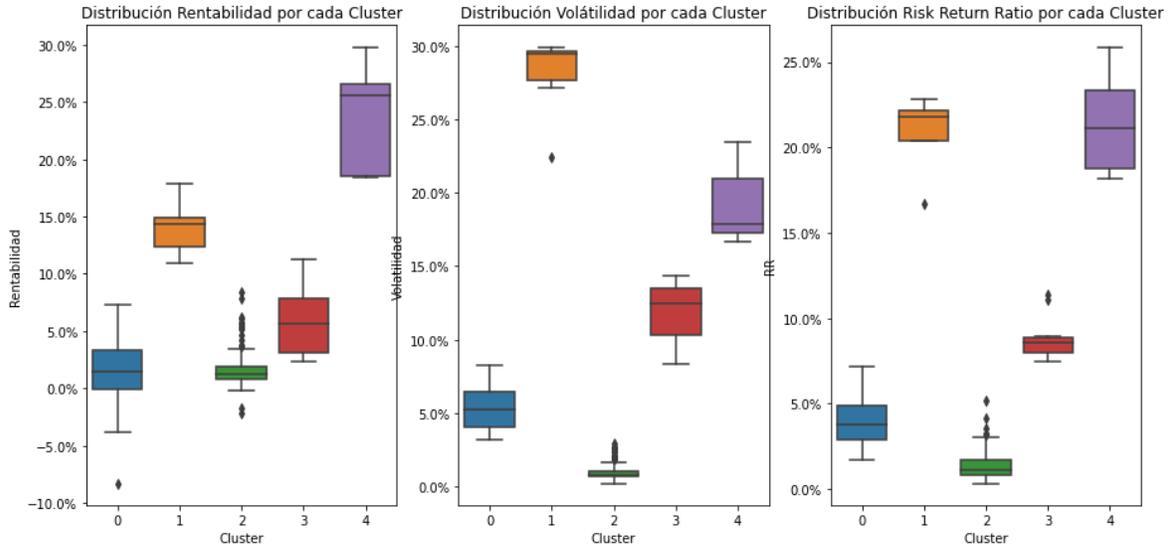


Figura 8. Gráfico de caja de porcentaje por Clúster con K-Means

En la Figura 8, se evidencia como los clústeres determinados por el algoritmo de *K-Means* distribuyen los valores de rentabilidad, volatilidad y *Risk Return Ratio* en un gráfico de caja de porcentajes. Se observa que en general los grupos resultantes se diferencian en todas las variables y hacen una buena diferenciación en entre los cuartiles uno al tres. Para el caso del clúster 2, que en su rentabilidad tiene similitud en sus datos con los clústeres 0 y 3 muestra una clara diferencia en la volatilidad y el *risk return ratio*. Además, de la Tabla 7 se observa el valor de la métrica de evaluación para el algoritmo *K-Means* con cinco centroides es de 6.695 en la sumatoria de la columna P.

	Fondo	Cluster	Rentabilidad	Volatilidad	RR	P
categoria						
Accionario Internacional	4	1	0.271377	0.196835	0.234106	0.250000
Accionario Nacional	7	1	0.138418	0.291190	0.214804	0.142857
Activo de Credito	6	2	0.042954	0.018028	0.034108	0.333333
Alto Rendimiento Nacional	3	2	0.017315	0.021957	0.019636	0.666667
Balanceado Mayor Riesgo	2	2	0.147716	0.212083	0.179900	1.000000
Balanceado Menor Riesgo	3	1	0.042133	0.057934	0.050033	0.333333
Balanceado Moderado	3	1	0.089234	0.104291	0.096763	0.333333
Inmobiliario de Desarrollo Nacional	1	1	0.043300	0.033574	0.038437	1.000000
Inmobiliario de Renta Nacional	3	2	0.023120	0.019802	0.022364	0.666667
Otros Fondos Bursatiles	1	1	0.179248	0.277043	0.228146	1.000000
Renta Fija Internacional	8	2	0.029506	0.111402	0.070454	0.250000
Renta Fija Nacional para Entidades Publicas	17	2	0.009393	0.010922	0.010926	0.117647
Renta Fija Pesos de Corto Plazo	25	2	0.010850	0.012995	0.011937	0.080000
Renta Fija Pesos de Largo Plazo	4	1	0.000859	0.056239	0.029902	0.250000
Renta Fija Pesos de Liquidez	31	2	0.012502	0.009644	0.011073	0.064516
Renta Fija Pesos de Mediano Plazo	14	2	0.012429	0.020321	0.016534	0.142857
Sin Categoria	63	4	0.025123	0.019985	0.025083	0.063492

Tabla 7. Agrupación por categoría con algoritmo K-Means

9.3.2 Mini Batch K-Means

El algoritmo Mini Batch K-Means es una variante del algoritmo K-Means que usa pequeños lotes de información para reducir el tiempo computacional de convergencia del algoritmo mientras optimiza la misma función objetivo. Los pequeños lotes son muestras de la base de datos total tomadas aleatoriamente en cada iteración. En comparación con otras variaciones del algoritmo K-Means que buscan disminuir el tiempo de convergencia, el algoritmo *Mini-Batch K-Means* generalmente produce resultados que pueden llegar a ser un poco peores que el algoritmo estándar (Sculley, 2010).

Para la estimación de los grupos se utiliza la función *MiniBatchKMeans* del módulo *sklearn.cluster*. El modelo tiene como parámetros el número de centroides o clústeres, que quedara igual al establecido en el algoritmo *K-Means*, el número máximo de iteraciones sobre el set de datos completo antes de detenerse (*max_iter*) que se establece en diez y el tamaño de las muestras de los mini lotes (*batch_size*) igual a seis para este ejercicio.

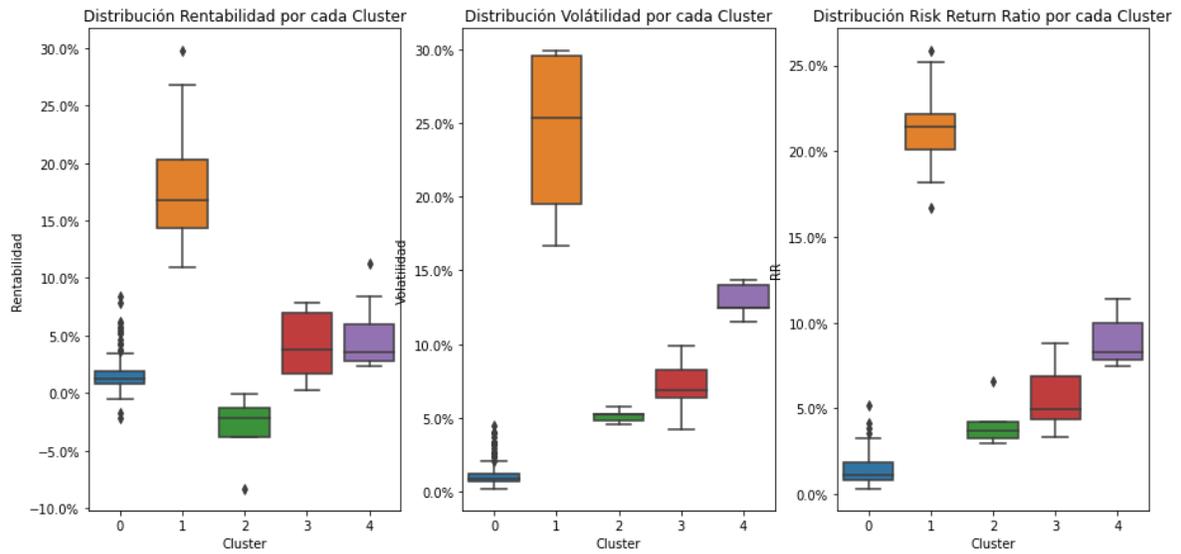


Figura 9. Caja de porcentajes por Clúster con Mini Batch K-Means

En la Figura 9, se evidencia como los clústeres determinados por el algoritmo de *Mini Batch K-Means* distribuye los valores de rentabilidad, volatilidad y *Risk Return Ratio* en un gráfico de caja de porcentajes. Se observa que en general los grupos resultantes se diferencian en todas las variables y en general hacen una buena diferenciación en entre los cuartiles uno al tres. Para el caso del clúster 4 que en su rentabilidad tiene similitud en sus datos con el clúster 3 muestra una clara diferencia en la volatilidad y el *risk return ratio*. De la Tabla 8, se observa el valor de la métrica de evaluación para el algoritmo *K-Means* con cinco centroides es de 6.266 en la sumatoria de la columna P.

categoria	Fondo	Cluster	Rentabilidad	Volatilidad	RR	P
Accionario Internacional	4	1	0.271377	0.196835	0.234106	0.250000
Accionario Nacional	7	1	0.138418	0.291190	0.214804	0.142857
Activo de Credito	6	2	0.042954	0.018028	0.034108	0.333333
Alto Rendimiento Nacional	3	1	0.017315	0.021957	0.019636	0.333333
Balanceado Mayor Riesgo	2	1	0.147716	0.212083	0.179900	0.500000
Balanceado Menor Riesgo	3	1	0.042133	0.057934	0.050033	0.333333
Balanceado Moderado	3	2	0.089234	0.104291	0.096763	0.666667
Inmobiliario de Desarrollo Nacional	1	1	0.043300	0.033574	0.038437	1.000000
Inmobiliario de Renta Nacional	3	1	0.023120	0.019802	0.022364	0.333333
Otros Fondos Bursatiles	1	1	0.179248	0.277043	0.228146	1.000000
Renta Fija Internacional	8	2	0.029506	0.111402	0.070454	0.250000
Renta Fija Nacional para Entidades Publicas	17	2	0.009393	0.010922	0.010926	0.117647
Renta Fija Pesos de Corto Plazo	25	1	0.010850	0.012995	0.011937	0.040000
Renta Fija Pesos de Largo Plazo	4	3	0.000859	0.056239	0.029902	0.750000
Renta Fija Pesos de Liquidez	31	2	0.012502	0.009644	0.011073	0.064516
Renta Fija Pesos de Mediano Plazo	14	1	0.012429	0.020321	0.016534	0.071429
Sin Categoria	63	5	0.025123	0.019985	0.025083	0.079365

Tabla 8. Agrupación por categoría con algoritmo K-Means

9.3.3 Hierarchical clustering

Los algoritmos de agrupamiento pertenecientes a la categoría de los jerárquicos son aquellos en los que se va particionando el conjunto de datos por niveles, de modo tal que en cada nivel generalmente, se unen o se dividen dos grupos del nivel anterior, según si es un algoritmo aglomerativo o divisivo (Pascual, Pla, & Sánchez, 2007). En este trabajo se utiliza un algoritmo de tipo aglomerativo por lo cual el proceso de agrupación consiste en comenzar asignando cada fondo como su propio clúster y a partir de ahí de forma iterativa agrupar pares de clústeres hasta converger a un criterio determinado (Zhao & Karypis, 2002).

Para la estimación de los grupos se utiliza la función *AgglomerativeClustering* del módulo *sklearn.cluster*. El modelo tiene como parámetros el número de centroides o clústeres, que quedara igual al establecido en el algoritmo *K-Means*.

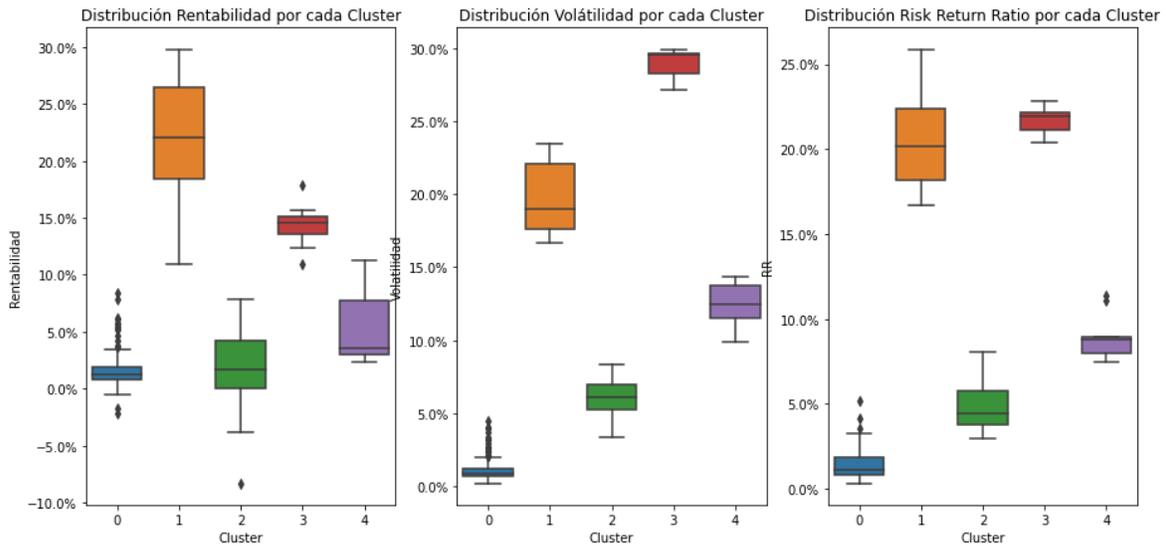


Figura 10. Caja de porcentajes por Clúster con Agglomerative Clustering

En la Figura 10, se evidencia como los clústeres determinados por el algoritmo de *K-Means* distribuye los valores de rentabilidad, volatilidad y *Risk Return Ratio* en un gráfico de caja de porcentajes. Se observa que en general los grupos resultantes se diferencian en todas las variables y en general hacen una buena diferenciación en entre los cuartiles uno al tres. Para el caso del clúster 2 que en su rentabilidad tiene similitud en sus datos con los clústeres 0 y 4 muestra una clara diferencia en la volatilidad y el *risk return ratio*. De la Tabla 9, se observa el valor de la métrica de evaluación para el algoritmo *Agglomerative Clustering* con cinco centroides es de 5.666 en la sumatoria de la columna P.

categoria	Fondo	Cluster	Rentabilidad	Volatilidad	RR	P
Accionario Internacional	4	1	0.271377	0.196835	0.234106	0.250000
Accionario Nacional	7	1	0.138418	0.291190	0.214804	0.142857
Activo de Credito	6	2	0.042954	0.018028	0.034108	0.333333
Alto Rendimiento Nacional	3	1	0.017315	0.021957	0.019636	0.333333
Balanceado Mayor Riesgo	2	1	0.147716	0.212083	0.179900	0.500000
Balanceado Menor Riesgo	3	1	0.042133	0.057934	0.050033	0.333333
Balanceado Moderado	3	1	0.089234	0.104291	0.096763	0.333333
Inmobiliario de Desarrollo Nacional	1	1	0.043300	0.033574	0.038437	1.000000
Inmobiliario de Renta Nacional	3	1	0.023120	0.019802	0.022364	0.333333
Otros Fondos Bursatiles	1	1	0.179248	0.277043	0.228146	1.000000
Renta Fija Internacional	8	2	0.029506	0.111402	0.070454	0.250000
Renta Fija Nacional para Entidades Publicas	17	2	0.009393	0.010922	0.010926	0.117647
Renta Fija Pesos de Corto Plazo	25	1	0.010850	0.012995	0.011937	0.040000
Renta Fija Pesos de Largo Plazo	4	2	0.000859	0.056239	0.029902	0.500000
Renta Fija Pesos de Liquidez	31	2	0.012502	0.009644	0.011073	0.064516
Renta Fija Pesos de Mediano Plazo	14	1	0.012429	0.020321	0.016534	0.071429
Sin Categoria	63	4	0.025123	0.019985	0.025083	0.063492

Tabla 9. Agrupación por categoría con algoritmo Agglomerative Clustering

9.4 Resultados

Después explicar las métricas de evaluación de los modelos, la metodología mediante la cual se asignaron los perfiles de riesgo por clúster, el número óptimo de grupos y de presentar los modelos disponibles, los resultados se consolidan en la Tabla 10:

Modelo	Número de centroides	Métrica de evaluación
K-Means	5	6.695
Mini Batch K-Means	5	6.266
Agglomerative Clustering	5	5.666

Tabla 10. Resultados Modelos Agrupamiento

El modelo que mayor semejanza tiene con la categorización de SIFIC y por lo tanto que minimiza la métrica de evaluación es el que recurre a la agrupación jerárquica de tipo aglomerativo. También se resalta como la variación del algoritmo particional *K-Means* que usa pequeños lotes muestra un mejor resultado que el algoritmo original. La Tabla 11, muestra cuantos fondos y categorías se agrupan en cada uno de los clústeres seleccionados con *Agglomerative Clustering* así como a rentabilidad, volatilidad y *risk return ratio* promedio.

Cluster	Fondo	categoria	Rentabilidad	Volatilidad	RR	Perfil
0	152	9	0.016091	0.010878	0.013804	Conservador
2	18	8	0.017552	0.061089	0.048000	Moderado Conservador
4	9	3	0.055556	0.123515	0.089536	Moderado
1	8	3	0.218794	0.196214	0.207504	Moderado Agresivo
3	8	2	0.143522	0.289421	0.216472	Agresivo

Tabla 11. Agrupación de fondos por clúster con el algoritmo seleccionado.

10. Producto de datos

Los modelos usados se integran en un producto de datos; en este sentido, se implementa una visualización de los resultados finales del presente trabajo mediante un tablero desarrollado con la librería *Dash* de *Plotly* (<https://plotly.com/dash/>), un *framework* desarrollado con el objetivo crear aplicaciones alojadas en la web. *Dash* se usa como la herramienta para desarrollar la visualización del proyecto, dado que es una librería de libre uso y permite crear *widgets* para interactuar con la información. Esto permite que un usuario pueda leer y contestar el cuestionario de perfilamiento de riesgo y obtener a través de una función desarrollada en Python el resultado de este, en la Figura 10, se observa cómo es la visualización del cuestionario en el tablero.

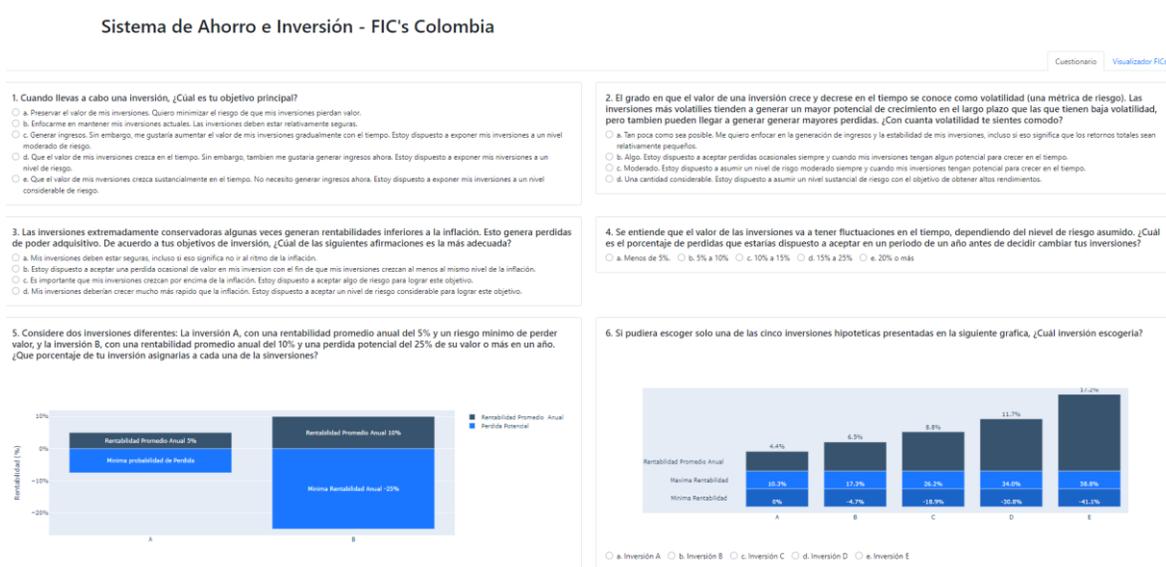


Figura 11. Tablero de visualización del proyecto, vista cuestionario perfilamiento de riesgo.

Una vez se obtiene el resultado del cuestionario en la esquina superior derecha se selecciona la vista de visualización de los fondos con un selector de los perfiles de riesgo. Esta vista, como se muestra en la Figura 12, contiene una tabla con los principales indicadores asociados al fondo y dos gráficos en la parte inferior. La grafica de la izquierda es un *scatter plot* con la distribución de los fondos y su respectiva rentabilidad y volatilidad. En la gráfica derecha se observa un gráfico de línea que permite ver como se hubiera comportado una inversión de cien pesos en

un fondo específico (el fondo que se grafica cambia al mover el puntero del *mouse* sobre el *scatter plot*).

Sistema de Ahorro e Inversión - FIC's Colombia

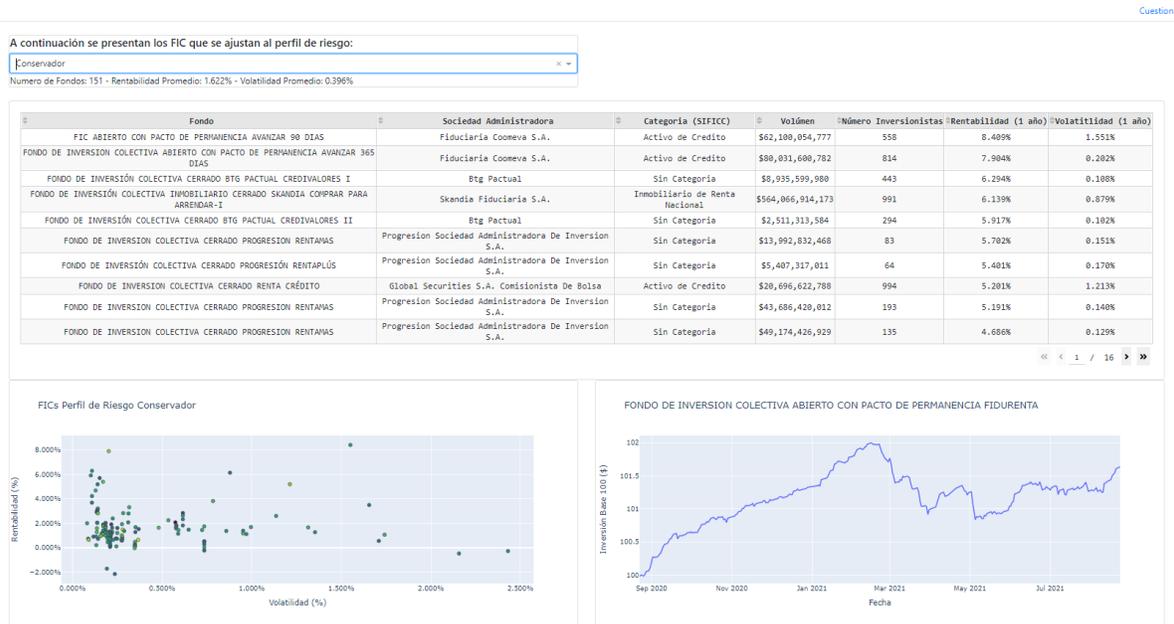


Figura 12. Tablero de visualización del proyecto, vista de los fondos agrupados por clúster.

10.1 Esquema de bases de datos

El modelo de visualización de resultados usa un esquema de bases de datos que a su vez se alimenta de las diferentes partes del proceso. En la Figura 13, se observa el esquema de bases de datos. Como se explica en el Capítulo 7, la base de datos inicial “BD_FICS” se alimenta de la fuente original, la API de datos abiertos del gobierno colombiano. La tabla “Maestro” se deriva de asociar los códigos de los fondos con las categorías del SIFIC. Usando el valor de la unidad histórico de cada participación del fondo se calculan los valores de la tabla “Base_Modelo” que alimentan los modelos de agrupamiento. Finalmente, la tabla “BD_Clusters” se alimenta del resultado final del agrupamiento con el modelo que mejor métrica de evaluación presenta.

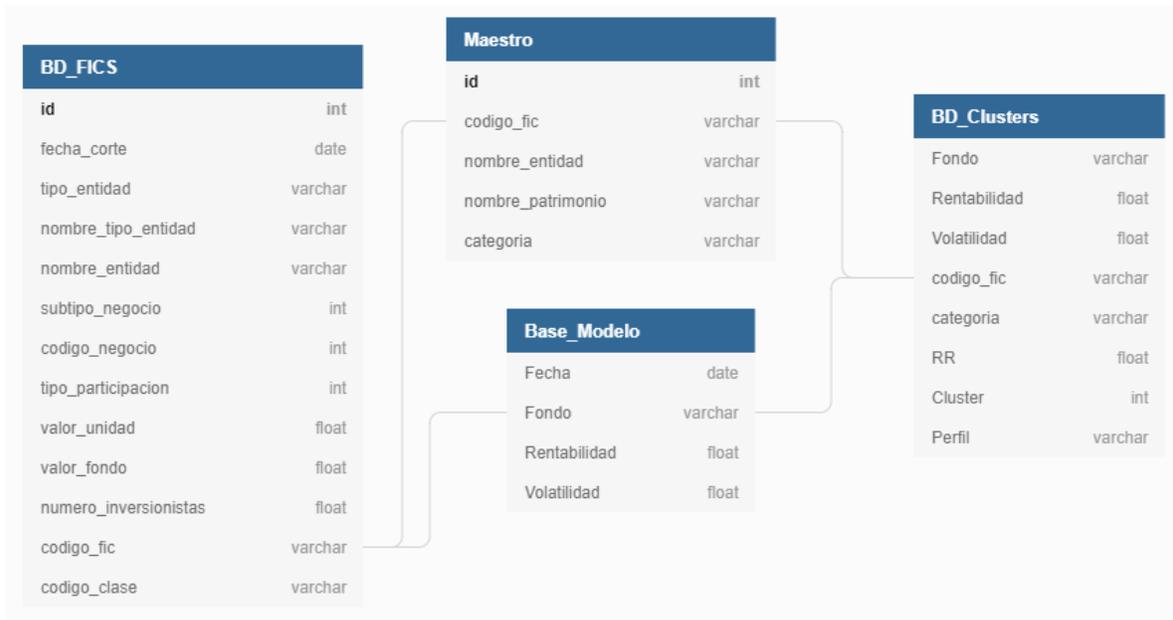


Figura 13. Esquema de bases de datos

11. Conclusiones

En este trabajo se ha desarrollado una secuencia de procedimientos que permiten acceder y estructurar de forma autónoma la información disponible acerca de los fondos de inversión colectiva del sistema financiero colombiano, democratizando el acceso a la asesoría financiera y buscando diversificar las fuentes de información que incentivan el ahorro y la inversión en el país. Con este trabajo se han definido las bases para un sistema de información web que usa aprendizaje de máquinas y permite interacción de un inversionista interesado, inicialmente personas naturales, con el universo de FIC disponibles en el mercado de acuerdo con las necesidades y características propias de cada uno de los usuarios.

Como aspirante al título de Magister en ingeniería analítica se ha integrado todo el flujo de trabajo de un proyecto de minería de datos, desde la administración de bases de datos hasta el sistema de visualización de un producto de datos completo, pasando por el modelamiento y la evaluación de algoritmos de aprendizaje de máquinas. Este proceso se desarrolla sobre los fondos de inversión colectiva en Colombia; al respecto, no se ha evidenciado un trabajo robusto en la literatura, este trabajo es pionero en esto, contribuyendo con la profundización del mercado de capitales colombiano y los niveles de ahorro e inversión de la población.

Objetivo general: Desarrollar un modelo para la recomendación y visualización de los fondos de inversión colectiva para personas naturales en el mercado financiero colombiano.

El trabajo ha permitido llegar a desarrollar un producto de datos que permita a una persona natural usuaria del mercado financiero colombiano conocer su nivel de aversión al riesgo y el perfil que mejor se adecua a sus necesidades de ahorro e inversión, a través de un modelo de recomendación basado en algoritmos de agrupamiento este perfil se conecta con los fondos de inversión colectiva que dadas sus características de rentabilidad y volatilidad. El producto final también incluye un módulo de visualización que como se evidencia en el Capítulo 10 permite al usuario identificar a través de gráficas y tablas como se han comportado los diferentes fondos y donde se ubican en un gráfico que relaciona rentabilidad y volatilidad.

A continuación, se relaciona por cada objetivo específico cuál ha sido el resultado obtenido.

Objetivo 1: Caracterizar las fuentes de información disponible del cierre de los fondos de inversión colectiva reportado por los administradores a través de un proceso de extracción, depuración y categorización tomando como referencia las categorías determinadas por el SIFIC.

Gracias a este trabajo se logró establecer un flujo de trabajo utilizando Python que permite extraer la información disponible en <https://www.datos.gov.co/> relacionada con los cierres diarios de los fondos de inversión colectiva del mercado colombiano. Estos fondos fueron caracterizados de acuerdo con los códigos internos que los identifican en la Superintendencia Financiera de Colombia y el código de cada una de sus respectivas participaciones.

Los fondos fueron asociados a una categoría de acuerdo con el sistema de categorización establecido por ASOFIDUCIARIAS, ASOBOLSA y el acompañamiento técnico de LVA Índices, resumidos en el SIFIC (<https://sificcolombia.com/>).

Los resultados de este objetivo están relacionados con el Capítulo 7: Entendimiento de los datos, donde se explica el proceso de caracterización, limpieza, exploración y creación de variables derivadas de los datos.

Objetivo 2: Implementar un modelo de clasificación basado en aprendizaje de máquinas que permita relacionar el perfil de riesgo del usuario con una categoría o conjunto de fondos de inversión colectiva.

En el presente trabajo se evaluaron modelos de agrupamiento o *clustering* donde se evaluaron modelos jerárquicos (*Agglomerative Clustering*) y particionales (*K-Means* y *Mini Batch K-Means*) utilizando las rentabilidades y las volatilidades del valor de unidad histórico de los fondos de inversión colectiva. Los resultados de estos modelos fueron contrastados con el sistema de categorización del SIFIC y se estableció un criterio de evaluación que permitió seleccionar el algoritmo de *Agglomerative Clustering* como el que mejor agrupa los datos con un valor de 5.666 versus valores superiores de los algoritmos de *K-Means* (6.695) y *Mini Batch K-Means* (6.266).

Los resultados de este objetivo están relacionados con el Capítulo 9: Modelado y Evaluación, donde se explican la métrica de evaluación de los modelos, la metodología para asignar los perfiles y el inventario de los modelos evaluados.

Objetivo 3: Validar el modelo de clasificación implementado para la clasificación de fondos de inversión colectiva de acuerdo con el perfil de riesgo del usuario.

La validación del modelo seleccionado se lleva a cabo cuando el modelo de agrupamiento de las variables relevantes de los FIC coincide con los perfiles de riesgo definidos en (Dev, 2016) y los

que se obtienen con el cuestionario de perfilamiento de Merrill Edge (ver anexo 1). Los grupos o clústeres resultantes del modelo fueron asociados a uno de los cinco perfiles de riesgo establecidos a través de una variable definida como *risk return ratio* que combina los valores promedio de rentabilidad y volatilidad, esta variable asocia los grupos con menores magnitudes a los perfiles de baja tolerancia al riesgo (Conservador) hasta los de mayor magnitud con un nivel de alta tolerancia al riesgo (Agresivo) pasando por todos los perfiles intermedios.

Los resultados de este objetivo están relacionados con el Capítulo 10: Producto de datos, donde se evidencia como tras completar el cuestionario de perfilamiento el usuario visualiza cuales son los fondos que mejor se adaptan a su perfil de riesgo.

Finalmente, como recomendaciones futuras para desarrollar a partir de este trabajo, se destaca la evaluación de otros algoritmos de agrupamiento, así como la evaluación de otras metodologías de agrupamiento y clasificación que incluya redes neuronales (*Deep learning*). Adicionalmente, una versión posterior del modelo de recomendación que mejoraría ampliamente la funcionalidad de asesoría y perfilamiento de riesgo se podría construir a partir de una base de datos sobre los inversionistas, incluyendo además de factores demográficos datos que permitan clasificar factores socio económicos como los ingresos y la capacidad de ahorro de forma que no solo se agrupen los fondos de inversión colectiva por rentabilidad y volatilidad si no que se logre clasificar por grupos al universo de inversionistas (personas naturales, personas jurídicas, inversionistas institucionales).

12. Referencias

- Abu Abbas, O. (2008). Comparisons Between Data Clustering Algorithms. *the international Arab Journal of information technology*, 320-325.
- Afi; ASOFIDUCIARIA. (2017). *Transformación y evolución de la industria FIC en Colombia: recomendaciones basadas en el caso español*. Bogotá: Corficolombiana.
- AMV. (2019). *Guía de Estudio Fondos de inversión Colectiva - Directivo*. Bogotá.
- ANIF. (2016). *Fondos de Inversión Colectiva en Colombia: Evolución y Oportunidades*. Bogotá.
- ASOFIDUCIARIA. (2018). *Nivel de Conocimiento Fondos de Inversión Colectiva*. Bogotá.
- ASOFIDUCIARIAS. (2018). *Todo lo que debe saber sobre los fondos de inversión colectiva - FICs*. Bogotá: Programa de educación Financiera.
- Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, 17-24.
- Bosch, M., Berstein, S., Castellani, F., Oliveri, M. L., & Villa, J. M. (2015). *Diagnóstico del sistema previsional colombiano y opciones de reforma*. BID.
- Cano, C. G., Esguerra, M. d., García, N., Rueda, J. L., & Velasco, A. M. (2013). *Acceso a servicios financieros en Colombia*. Bogota: Banco de la República.
- CIEEF. (2017). *Estrategia Nacional de Educación Económica y Financiera de Colombia*. Bogotá.
- Cowan, D., Alencar, P., & Portugal, I. (2018). *The use of machine learning algorithms in recommender systems: A systematic review*.
- Dabbas, E. (2021). *Interactive Dashboards and Data Apps with Plotly and Dash*. Birmingham: Packt Publishing.
- Dev, K. (2016). Risk Profile and Mutual Fund Scheme Selection of Investors. *International Journal of Research and Analytical Reviews*, 208-214.
- Fedesarrollo. (2018). *La reforma pensional que Colombia requiere*. Informe Mensual del Mercado Laboral.
- Frenkel, R. (2003). Globalizacion y crisis financiera en America Latina. *Revista de la CEPAL*, 41-54.
- Khanna, K., & Chauhan, V. (2019). *A Study on Risk Profiling and Investment Choices of Retail Investors*.
- Ministerio de Hacienda y Crédito Público. (2019). *Misión del Mercado de Capitales*. Bogotá.
- Pascual, D., Pla, F., & Sánchez, S. (2007). *Algoritmos de agrupamiento*.

- Peralta, C. F. (2014). Proceso de Conceptualización del Entendimiento del Negocio para Proyectos de Explotación de Información. *Revista Latinoamericana de Ingeniería de Software*, 273-306.
- Robles Fernández, M. D. (1999). *Medidas de Volatilidad*. Madrid: Universidad Europea - CEES.
- Sculley, D. (2010). *Web-Scale K-Mans Clustering*. Pittsburgh: Proceedings of the 19th international conference on World wide web.
- SIFIC. (2018). *Metodología de Categorización de los Fondos de Inversión Colectiva*. Bogotá.
- Umer, I., Kashifa, C. M., Rehanb, R., & Baia, A. (2018). *An empirical investigation of Investor's Behavioral Biases on Financial Decision Making*.
- URF. (2018). *urf.gov.co*. Obtenido de http://www.urf.gov.co/webcenter/ShowProperty?nodeld=/ConexionContent/WCC_CLUSTER-114209
- Zhao, Y., & Karypis, G. (2002). *Comparison of Agglomerative and Partitional Document Clustering Algorithms*. Minneapolis: University of Minnesota.
- Zorrila, J. P. (16 de Febrero de 2016). *¿En que ahorran e invierten los colombianos?* Obtenido de <https://www.portafolio.co/>: <https://www.portafolio.co/mis-finanzas/ahorro/ahorran-e-invierten-colombianos-116496>

Anexo 1. Cuestionario de perfilamiento de riesgo, Merrill Edge



Identifying your Investor Profile

Name _____ Date _____

Determining risk tolerance

One of the first steps in developing an investment strategy is to identify your tolerance for risk as an investor, referred to as your Investor Profile. It can depend on the goals you are investing toward, as well as your personality, in making investment decisions.

Merrill Edge® has identified investor profiles that generally coincide with the ways in which investors characterize themselves, their objectives, and their feelings about risk. Each investor profile—Conservative, Moderately Conservative, Moderate, Moderately Aggressive and Aggressive—has an associated asset allocation based on your overall risk tolerance.

Review the investor profile descriptions on page 5 and the investor profile models on page 6 to identify the Investor Profile that best describes your investment objectives. Or you can complete the self-evaluation to help identify your personal investment preferences. This evaluation will measure your reaction to market volatility and help you identify your investment objectives.

Investor Profile self-evaluation

Please read and answer the following questions. Then, use the scoring process to identify the Investor Profile that corresponds to your feelings about investing. A description of each profile is provided on page 5. Your Investor Profile will be based on all of your responses collectively, with no single question being the determining factor.

<p>1 When you invest money, what is your primary goal?</p> <ul style="list-style-type: none">a. Preserve the value of my investments. I want to minimize the risk of my investments losing value.b. Emphasize current income. My investments should be relatively safe.c. Generate current income. However, I would also like to build the value of my investments gradually over time. I am willing to expose my investments to a moderate level of risk.d. Have the value of my investments grow over time. However, I would also like to generate some current income. I am willing to expose my investments to a fair level of risk.e. Have the value of my investments grow substantially over time. I do not need to generate current income. I am willing to expose my investments to a considerable level of risk.	<p>2 The degree to which the value of an investment increases and decreases is called volatility (one measure of risk). More volatile investments generally offer greater long-term growth potential than less volatile investments, but they may produce greater losses. How much volatility are you comfortable with?</p> <ul style="list-style-type: none">a. As little as possible. I want to focus on current income and stability of value even if it means that my total returns are relatively small.b. Some. I am willing to accept occasional losses in value as long as my investments have some potential for growth over time.c. Moderate. I am willing to take moderate risk as long as my investments have a greater potential for growth over time.d. A considerable amount. I am willing to take a substantial risk in pursuit of higher total returns.
--	---

1

3 Extremely conservative investments sometimes earn less than the inflation rate. This may result in the loss of purchasing power. With respect to your investment objectives, which of the following is most true?

- a. My investments should be safe, even if it means my returns do not keep pace with inflation.
- b. I am willing to risk an occasional loss of investment value so that my investments may grow at about the same rate as inflation over time.
- c. It is important that my investments grow somewhat faster than inflation. I am willing to accept some risk to achieve this goal.
- d. My investments should grow much faster than inflation. I am willing to accept considerable risk to achieve this goal.

4 You understand the value of investments will fluctuate over time, depending on the amount of risk taken. What is the approximate loss in any one-year period that you would be willing to accept before deciding to change your investments?

- a. Less than (5%)
- b. (5%) to (10%)
- c. (10%) to (15%)
- d. (15%) to (25%)
- e. (25%) or more

5 Consider two different investments: Investment A, which provides an average annual return of 5% with a minimal risk of loss of value, and Investment B, which provides an average annual return of 10% and a potential loss of 25% or more in any year. How would you divide your investment dollars?



- a. 100% in Investment A and 0% in Investment B
- b. 80% in Investment A and 20% in Investment B
- c. 50% in Investment A and 50% in Investment B
- d. 20% in Investment A and 80% in Investment B
- e. 0% in Investment A and 100% in Investment B

6 If you could choose only one of the five hypothetical portfolios characterized below, which would you select?



- a. Portfolio A
- b. Portfolio B
- c. Portfolio C
- d. Portfolio D
- e. Portfolio E

Note: These hypothetical portfolios do not correspond to the Merrill Lynch Investor Profile models. "A" represents the least risk, whereas "E" represents the highest risk. No time frame is provided for the swing between the lowest annual return and highest annual return possibilities, and this graph is used to determine risk tolerance only.

7 When is the earliest you anticipate needing all or a substantial portion of your investment assets?

- a. Short term — 0 to 2 years
- b. Medium term — More than 2 but less than 5 years
- c. Long term — 5 years or more

8 Are your total investment assets less than \$1 million?

- a. Yes
- b. No

9 If your response to Question 8 is No, is the amount you are investing less than 10% of your total investment assets?

- a. Yes
- b. No

Determination of Investor Profile model

STEP 1 For easy reference, list your answer to each question in the box above the question number. Then, circle the number of points awarded for your response to each question, add the points and record your total points in the box to the right.

My answers												Total points
Questions	1	2	3	4	5	6						
a	1	1	1	1	1	1						
b	3	3	3	3	3	3						
c	5	7	7	5	5	5						
d	7	9	9	7	7	7						
e	9			9	9	9						

STEP 2 List your answers to questions 7-9 in the boxes above the question numbers, and circle your response below.

My answers				
Questions	7	8	9	
a	Short	Yes	Yes	
b	Medium	No	No	
c	Long			

STEP 3 Find the matrix that corresponds to your total points. Read down and across, using your responses to the selected questions, to identify your Investor Profile.

My Investor Profile

If your total points are 6-15 and:

Response to Question 3	Response to Question 7		
	a - SHORT	b - MEDIUM	c - LONG
a	Conservative	Conservative	Conservative
b, c or d	Conservative	Conservative	Moderately Conservative

If your total points are 16-25 and:

Response to Question 1	Response to Question 7		
	a - SHORT	b - MEDIUM	c - LONG
a	Conservative	Conservative	Moderately Conservative
b, c or d	Moderately Conservative	Moderately Conservative	Moderately Conservative
e	Moderately Conservative	Moderately Conservative	Moderate

If your total points are 26-34 and:

Response to Question 1	Response to Question 7		
	a - SHORT	b - MEDIUM	c - LONG
a or b	Moderately Conservative	Moderately Conservative	Moderate
c	Moderately Conservative	Moderate	Moderate
d or e	Moderate	Moderate	Moderately Aggressive

If your total points are 35-44 and:

Response to Question 8	Response to Question 7		
	a - SHORT	b - MEDIUM	c - LONG
a - Yes	Moderate	Moderate	Moderately Aggressive
b - No	Proceed to next chart		

Response to Question 9	Response to Question 7		
	a - SHORT	b - MEDIUM	c - LONG
a - Yes	Moderate	Moderately Aggressive	Aggressive
b - No	Moderate	Moderate	Moderately Aggressive

If your total points are 45-54 and:

Response to Question 7		
a - SHORT	b - MEDIUM	c - LONG
Moderate	Aggressive	Aggressive

Investor Profile sample scoring

The scoring procedure involves four steps:

- 1 Totaling your points per your responses to 1–6.
- 2 Identifying your responses to 7–9.
- 3 Selecting the correct scoring matrix per your total points.
- 4 Responding to the questions in the matrix per your previous answers.

If you have difficulty understanding this procedure, refer to the sample test below.

Time frame to invest: More than two but less than five years

My answers							Total points
Questions	1	2	3	4	5	6	36
a	1	1	1	1	1	1	
b	3	3	3	3	3	3	
c	5	7	7	5	5	5	
d	7	9	9	7	7	7	
e	9			9	9	9	

My answers			
Questions	7	8	9
a	Short	Yes	Yes
b	Medium	No	No
c	Long		

If your total points are 35–44 and: ←

Response to Question 8	Response to Question 7		
	a – SHORT	b – MEDIUM	c – LONG
a – Yes	Moderate	Moderate	Moderately Aggressive
b – No	Proceed to next chart →		

Response to Question 9	Response to Question 7		
	a – SHORT	b – MEDIUM	c – LONG
a – Yes	Moderate	Moderately Aggressive	Aggressive
b – No	Moderate	Moderate	Moderately Aggressive

The investor scored 36 points and therefore selects the 35–44 scoring matrix to identify her Investor Profile. She pencils in her response to Question 8 on the matrix, reads across and follows the directions to proceed to the next matrix. On the next matrix, she fills in her responses to Questions 9 and 7, and reads across and down from her answers to find her Profile—Moderate.