Entrega Final Proyecto Integrador

Juliana Ochoa Ramírez - Código: 201910048228
Mateo Graciano Londoño - Código: 201910010228
Alejandro Palacio Vásquez - Código: 201910049228
Camila Mejía Quintero - Código: 201910015228
Juan Esteban Torres Marulanda - Código: 201910052228
Andrés Franco Zapata - Código: 201910043228
Programa: Maestría en Ciencia de los Datos y Analítica
Docentes:
Olga Lucía Quintero
Édison Valencia Díaz
Andrés Ramírez Hassan
16 de noviembre de 2019

***Resumen*—Este trabajo presenta una metodología que busca abordar la problemática de asignación de activos dentro de un portafolio de inversión, con el fin, de vencer sistemáticamente el Benchmark de referencia. Dicha metodología utiliza conceptos de aprendizaje no supervisado y supervisado, buscando replicar indicadores de análisis técnico sobre cada uno de los activos financieros que componen el portafolio para poder definir la mejor alternativa de inversión.**

1. JUSTIFICACIÓN

En los mercados financieros, las entidades encargadas de la administración de activos se enfrentan a la importante tarea de conformar los portafolios de inversión administrados, esta actividad implica grandes retos considerando el amplio universo de activos disponibles en los mercados de valores [1].

Uno de los vehículos de inversión colectiva más reconocidos en la industria financiera a nivel mundial son los fondos mutuos, en los cuales los administradores de activos buscan reunir recursos de diferentes inversionistas con el fin de ejecutar la estrategia de inversión del respectivo fondo.

Existen dos enfoques utilizados por los administradores para la conformación de los portafolios de inversión, el primero de ellos es el enfoque fundamental y el más utilizado en la actualidad por los administradores de fondos, donde la base de sus decisiones se soportan en las expectativas que tenga la empresa emisora del respectivo activo, teniendo en cuenta elementos como el equipo directivo, producto, competencia y sector en el cual se encuentra la empresa, dejando de lado métodos analíticos más rigurosos.

Este primer enfoque tiene un problema en particular, relacionado con la baja escalabilidad que generan los altos costos debido a las contrataciones adicionales necesarias para conformar un equipo de trabajo lo suficientemente robusto, que permita cubrir la mayor cantidad de instrumentos financieros posibles y lograr llegar a todos los activos al nivel de profundidad deseado. Por consiguiente, este tipo de estructuras incrementan los costos de administración del fondo impactando directamente el desempeño en el largo plazo.

El segundo enfoque es el cuantitativo, el cual se soporta en análisis técnico sobre la información histórica de los activos [2]. A diferencia del enfoque fundamental, los administradores de activos encuentran en la aplicación de métodos analíticos sofisticados ventajas como, poder analizar electrónicamente miles de activos financieros de forma simultánea, tener un mayor control de los costos asociados, y así obtener una mayor capacidad de gestión y acceso a diferentes tipos de instrumentos sin incurrir en mayores costos.

En este trabajo se propone una metodología basada en la capacidad de aprendizaje de modelos supervisados, utilizando como información de entrada diferentes indicadores de análisis técnico sobre cada una de los activos financieros disponibles, con el fin de poder determinar la mejor alternativa de inversión, de acuerdo con un Benchmark definido.

2. REVISIÓN LITERATURA

Pronosticar el comportamiento de los mercados de valores para poder desarrollar estrategias de inversión rentables han sido foco de estudio a lo largo del tiempo y si bien, ha ganado fuerza últimamente con los desarrollos tecnológicos que soportan los algoritmos y estrategias de inversión, los primeros desarrollos en este frente datan casi unos 60 años atrás.

En su libro de 1967 [3], Thorp y Kassouf desarrollan lo que se conoce como un sistema bursátil científico. Este es sin duda uno de los libros más influyentes de todos los tiempos en el mundo de las finanzas pues plantearon métodos que desataron la revolución cuantitativa de las mismas, combinando conceptos de probabilidad y estadística con su conocimiento profundo del mercado, dieron pie a lo que hoy se realiza en finanzas cuantitativas modernas. Hoy en día y como se mencionó, el análisis cuantitativo se ha venido utilizando cada vez más a raíz de los significativos avances tecnológicos, que han permitido a los profesionales aplicar técnicas cuantitativas complejas para obtener altos rendimientos de manera más eficiente y frecuente. A continuación, se presentan algunos estudios relacionados con los modelos de aprendizaje automático y sus aplicaciones en las inversiones del mercado de capitales.

En [4], los autores propusieron un método de aprendizaje profundo (Deep Learning) para la predicción de movimientos en el mercado de valores asociados a estrategias derivados de eventos. Una estrategia impulsada por eventos es un tipo de estrategia de inversiones que aprovecha la sobre o subvaloración temporal de acciones, que se puede presentar antes o después de un evento corporativo, entiendo este último como cualquier proceso de fusiones y adquisiciones, cambio de estructura organizacional o nombramientos de altos cargos. En primer lugar, los eventos se extraen del texto de las noticias y se representan como vectores densos (Text Mining) y se entrenan utilizando una Neural Tensor Network. Luego, utilizan una red neuronal convolucional profunda para modelar las influencias de los eventos a corto y largo plazo en los movimientos del precio de las acciones. Los obtenidos muestran que este modelo es capaz de lograr un 6 % aproximado de mejoras en la predicción del índice SP 500 (índice accionario de Estados Unidos) y la predicción de las acciones individuales.

Como se mencionó, la administración de portafolios desde el enfoque cuantitativo se basa en el análisis técnico como instrumento para la toma de decisiones de inversión, ayuda a predecir los precios futuros de las acciones y también a vislumbrar de forma anticipada la posibilidad de decidir en cuanto a la compra-retención-venta de la acción. Esta alternativa de gestión se basa en el movimiento histórico de los precios y los utiliza para predecir sus futuros movimientos identificando si existen alguna reversión de tendencia para formular la estrategia de inversión. En [5], se presenta una profundización de análisis técnico en acciones de compañías pertenecientes a Nifty 50 (índice accionario de India). Los indicadores utilizados en el análisis incluyen medias móviles, RSI, bandas de Bollinger y MACD. En este, se concluye que una buena estrategia de inversión debe combinar un poco de ambos mundos (fundamental como técnico) pero que el mercado de valores estudiado ofrecía una gran oportunidad tanto a corto como a largo plazo para inversionistas que trabajaran con el RSI y MACD, pues son indicador que ofrecen señales contundentes sobre la dirección en la que se dirige la empresa a invertir y ayuda a identificar si existe situación de sobreventa, sobrecompra o de reversiones de tendencias.

Entrando más en materia y buscando la conexión de los indicadores de análisis técnicos descritos anteriormente con los modelos de aprendizaje automático, en [6], presentan un sistema de predicción y negociación de precios de acciones basado en redes neuronales que utiliza indicadores de análisis técnico. El modelo desarrollado por los autores primero trabaja con los datos de la serie de tiempo financiera y los convierte en una serie de señales de activación de compra-venta-retención utilizando los indicadores de análisis técnico. Luego, entrenan un modelo de red neuronal Perceptrón Multicapa (MLP) sobre los precios diarios de acciones entre 1997 y 2007 para todas las acciones pertenecientes al Dow30 y, por último, el modelo entrenado se prueba con datos de 2007 a 2017. Los resultados obtenidos son muy interesantes pues indican que, al seleccionar los indicadores de análisis técnico más apropiados, el modelo es capaz de obtener resultados comparables con la estrategia de comprar y mantener en la mayoría de los casos.

Para concluir en esta sección, las técnicas de inteligencia artificial para sistemas financieros se están volviendo muy populares, en especial los modelos de aprendizaje profundo que comienzan a recibir más atención dada la conexión con el procesamiento de imágenes. En [7], se propone un nuevo modelo de transacciones algorítmico, llamado CNN-TA, que utiliza una red neuronal convolucional basada en las propiedades de procesamiento de imágenes. Las series de tiempo financieras se convierten en imágenes utilizando 15 indicadores del análisis técnico diferentes, cada uno con diferentes configuraciones de parámetros y cada instancia de indicadores genera datos para un período de 15 días. Como resultado, se obtienen imágenes de tamaño 15x15 que son etiquetadas como Comprar, Vender o Retener, según las “colinas” y los “valles” de la serie de tiempo original. Al final, los autores indican que el modelo desarrollado presenta buenos resultados para acciones y ETF.

3. Descripción del Problema

Como se presentó en la sección anterior, para la administración de portafolios existen dos reconocidas alternativas de gestión, en primer lugar, la gestión pasiva, en la cual el administrador de portafolios debe replicar el desempeño de un índice bursátil que hace las veces de Benchmark de mercado, es decir, replicar el mismo portafolio construido por el índice sin tomar decisiones propias que busquen diferenciarse.

En segundo lugar, existe la gestión activa, en la cual el administrador de portafolios busca conformar un portafolio similar al índice de referencia, pero con la posibilidad de otorgar diferentes pesos a las inversiones del índice, con el fin de obtener un mejor desempeño en términos de rentabilidad que el índice de referencia.

Dentro de las diferentes alternativas de inversión en el mercado financiero existen diferentes posibilidades como son, el mercado de renta fija, de renta variable y el de alternativos. En el mundo de la renta variable existen diferentes Benchmark que permiten replicar el mercado.

En este caso, se decidió crear un propio Benchmark permitiendo generar replicabilidad en el tiempo. El Benchmark que se buscará vencer con la estrategia de inversión propuesta es uno conformado por los ETF de diferentes mercados, pero con un peso igual dentro del portafolio que se conforme llamado naive equally weighted (EW).

Es de interés entonces entrenar un algoritmo que de acuerdo con los precios de los activos y toda la información que el análisis técnico entrega, proponga alternativas de inversión de forma distinta buscando, por medio de una gestión activa, ganarle al Benchmark.

4. Objetivo General

Generar una estrategia de inversión que permita mejorar la gestión activa de portafolios que realizan los administradores de portafolio.

5. Objetivo Especifico

- Realizar agrupación en clústeres, a través de algoritmos no supervisados, en ETF representativos de los mercados asiáticos, europeos y de Estados Unidos.

- Evaluar diferentes metodologías de algoritmos supervisados en los clústeres encontrados para obtener un mejor retorno sobre el Benchmark (Equally Weigthed) medido desde diferentes medidas de desempeño para la construcción de portafolios.
- Comparar el funcionamiento de los modelos con altas y bajas dimensionalidades, realizando procesos de selección de variables.

6. FUENTES DE DATOS

Los datos fueron recopilados en el laboratorio financiero de la universidad EAFIT usando el paquete RBLAPI y se componen de las cotizaciones diarias de los Exchange Trade Founds (ETF) de tres diferentes mercados:

- Europa: compuesto por 10 países
- Asia: compuesto por 13 países
- Estados Unidos: compuesto por 10 sectores

Particularmente para todos exceptuando Estados Unidos, se tomaron los países referentes, mientras que en Estados Unidos se tomaron los sectores más representativos como lo son, por ejemplo: consumo, energía y financiero. Todo esto se presenta en las siguientes tablas:

Figura 1: Composición ETF Asia

Asia		
Asia	ISHARES MSCI ALL COUNTRY ASI	AAXJ US Equity
Hong Kong	ISHARES MSCI HONG KONG ETF	EWH US Equity
South Korea	ISHARES MSCI SOUTH KOREA ETF	EWY US Equity
Taiwan	ISHARES MSCI TAIWAN ETF	EWT US Equity
Australia	ISHARES MSCI AUSTRALIA ETF	EWA US Equity
Malaysia	ISHARES MSCI MALAYSIA ETF	EWM US Equity
Thailand	ISHARES MSCI THAILAND ETF	THD US Equity
Singapore	ISHARES MSCI SINGAPORE ETF	EWS US Equity
Indonesia	ISHARES MSCI INDONESIA ETF	EIDO US Equity
Philippines	ISHARES MSCI PHILIPPINES ETF	EPHE US Equity
China	ISHARES MSCI CHINA ETF	MCHI US Equity
India	ISHARES MSCI INDIA ETF	INDA US Equity
Pakistan	GLOBAL X MSCI PAKISTAN ETF	PAK US Equity

Figura 2: Composición ETF Europa

Europa		
Sweden	iShares MSCI Sweden ETF	EWD US Equity
Germany	iShares MSCI Germany ETF	EWG US Equity
Switzerland	iShares MSCI Switzerland ETF	EWL US Equity
Netherlands	iShares MSCI Netherlands ETF	EWN US Equity
Spain	iShares MSCI Spain ETF	EWP US Equity
France	iShares MSCI France ETF	EWQ US Equity
United Kingdom	iShares MSCI United Kingdom ETF	EWU US Equity
Eurozone	iShares MSCI Eurozone ETF	EZU US Equity
Europe	iShares Europe ETF	IEV US Equity
Italy	iShares MSCI Italy ETF	EWI US Equity

Figura 3: Composición ETF Estados Unidods

Estados Unidos - Sectores		
Basic Materials	MATERIALS SELECT SECTOR SPDR	XLB US Equity
Communications	COMM SERV SELECT SECTOR SPDR	XLC US Equity
Consumer, Cyclical	CONSUMER DISCRETIONARY SELT	XLY US Equity
Consumer, Non-cyclical	CONSUMER STAPLES SPDR	XLP US Equity
Energy	ENERGY SELECT SECTOR SPDR	XLE US Equity
Financial	FINANCIAL SELECT SECTOR SPDR	XLF US Equity
Industrial	INDUSTRIAL SELECT SECT SPDR	XLI US Equity
Technology	TECHNOLOGY SELECT SECT SPDR	XLK US Equity
Utilities	UTILITIES SELECT SECTOR SPDR	XLU US Equity
U.S.	SPDR S&P 500 ETF TRUST	SPY US Equity

Como inicialmente se desea realizar sobre la información seleccionada una clasificación no supervisada, se decide usar cada uno de los días para reconstruir nuevas características. Lo anterior se realiza a partir de la historia de cada uno de los ETF, donde se extraen diferentes características de la serie temporal de cada uno como lo son: la media de los datos, la varianza, diferentes percentiles, entre otros. En total se quedan reflejados 794 variables para cada uno de los ETF.

7. METODOLOGÍA

7.1. *Clustering: Encontrando activos similares*

En una primera etapa, sobre los ETF descritos anteriormente se aplicaron modelos de aprendizaje no supervisado [8] con el fin de explorar y comprender los datos, agrupando y clasificando las observaciones.

En general, es de interés explorar en una primera instancia los datos con el fin de encontrar el número óptimo de grupos para hacer la clasificación, sin embargo, esto puede ser costoso en términos computacionales y no ser lo óptimo para conjuntos con grandes volúmenes de información. Para esta etapa se trabajaron con los siguientes algoritmos de clasificación:

7.1.1. ***Substractive Clustering****:* Substractive Clustering o Agrupación Sustractiva es un algoritmo de agrupación basado en la búsqueda de los nodos o centros dentro del conjunto original de datos. Este algoritmo comienza considerando cada observación en el conjunto de datos como un potencial centro de clúster y el potencial de cada observación es una función de su distancia a las demás observaciones restantes en el conjunto de datos. En consecuencia, un objeto con muchos objetos cercanos (es decir, con una alta densidad de objetos circundantes) tendrá un alto valor potencial para ser centro de clúster.

El algoritmo funciona con un parámetro r que es un radio positivo que busca condicionar el efecto de las observaciones sobre los posibles centros de agrupación en la medida de densidad. Particularmente se utiliza para definir un vecino alrededor de cada posible centro para medir su valor potencial.

7.1.2. ***K-Means****:* K-means clustering es el más representativo de los algoritmos no supervisados, se asumen que es hard dado que a diferencia de los anteriores se le debe ingresar el número de clúster que se quieren usar para hacer la clusterización.

Este algoritmo es estocástico y heurístico, dado que la función de optimización no se garantiza que converja al óptimo, por lo que es sensible a las primeras particiones que hace el algoritmo, pero se soluciona corriéndolo varias veces.

El algoritmo a partir de un conjunto inicial de centroides comienza a iterar entre la asignación que hace de cada uno de los elementos y recalcula el centroide como el centroide de las observaciones del grupo.

7.1.3. ***Fuzzy C-means****:* Fuzzy C-Means (FCM) es un método de agrupación que permite que un grupo de datos pertenezca a dos o más clústeres. Este método fue desarrollado por Dunn en 1973 y se utiliza con frecuencia en el reconocimiento de patrones.

En este algoritmo los datos se asignan o vinculan a cada grupo por medio de una función de pertenencia que resulta ser la representación del comportamiento difuso de este algoritmo. Para hacer eso, se debe construir una matriz llamada U cuyos valores son números entre 0 y 1, y representan el grado de pertenencia entre los datos y los centros de los grupos.

7.1.4. ***Aglomerativo****:* Agglomerative clustering trabaja inicialmente considerando cada elemento como un único clúster. Luego los dos clústeres que son más cercanos de acuerdo con la distancia que se defina, la combina dentro de un único clúster Este procedimiento es iterativo.

Se enfoca en agrupar objetos que están a distancias cercanas y separarlo de objetos que su distancia es alta es decir que no son similares.

7.1.5. ***Espectral****:* La agrupación espectral es un método flexible, que se basa en la teoría de gráficos, donde el enfoque se utiliza para identificar comunidades de nodos en un gráfico en función de los bordes que los conectan.

Este método utiliza información de los valores propios (espectro) de matrices especiales construidas a partir del gráfico o el conjunto de datos, por lo que resulta ser una metodología de agrupación que recoge varios de los elementos que hemos visto a lo largo de los cursos, pasando por teoría de grafos, definición de nodos, aristas y matrices de adyacencia, hasta llegar a la descomposición espectral de matrices e identificación de valores y vectores propios para poder realizar la identificación de comunidades y de clústeres dentro de los datos.

Dado lo anterior, la aproximación para la definición del mejor número de grupos es la siguiente:

1. Utilizamos el algoritmo de substractive Clustering, trabajando con diferentes parámetros de Radio, h para analizar qué número de grupos identifica el algoritmo. Ese parámetro resultante, es en general, el dato más complejo de obtener en el momento de aplicar cualquier algoritmo de agrupación por lo que nos basamos del obtenido en este algoritmo como base para los siguientes.
2. Luego corremos los demás algoritmos de agrupación que expusimos: K-Means, Fuzzy C-Means, Aglomerativo y Espectral, partiendo del número de grupos que obtuvimos en el primer paso.
3. Posterior a esto, evaluamos la consistencia de los algoritmos, analizando cuáles elementos comparten el mismo grupo en los diferentes algoritmos, calificando con 0 aquel que no comparte ninguno y 4 el que los comparte todos.
4. Dado lo anterior, se analiza cómo los datos quedan clasificados dentro de los grupos y se analiza si existe alguna relación lógica para ello.

7.2. Creación de motor de decisión

Se utilizaron diferentes indicadores técnicos para así generar una estrategia basada en decisiones de pares de activos que nos permita obtener un mejor retorno que el Benchmark definido como Equally Weigthed.

Para ello la idea es crear para cada pareja de activos (i y j) un activo ficticio $F^{i,j}$ que definimos de la siguiente forma:

$$F_t^{i,j} = \frac{A_t^i}{A_t^j}$$

Figura 4: Activos reales

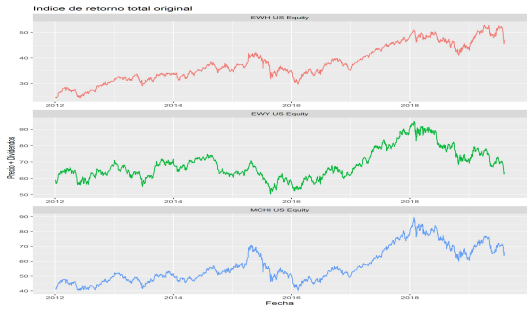
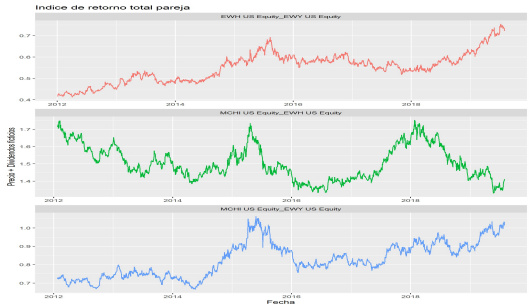


Figura 5: Activos ficticios



De esta forma es posible calcularle todos los indicadores técnicos a $F^{i,j}$.

El problema entonces lo convertimos en predecir el signo del retorno de $F^{i,j}$ k periodos en el futuro con la información que tenemos en t, por eso creamos el $label_t^{i,j}$ con base en lo siguiente:

$$label_t^{i,j} = \begin{cases} i & \frac{F_{t+k}^{i,j}}{F_t^{i,j}} \geq 1 \\ j & \text{En otro caso} \end{cases} \tag{1}$$

De la ecuación $label_t^{i,j}$ se puede inferir que la etiqueta de $F_t^{i,j}$ es i cuando el activo i tiene un mejor retorno en los siguientes k periodos.

De este modo el problema de clasificación se convierte en hallar una función G tal que:

$$G(Tech(F_t^{i,j}) + \epsilon) = label_t^{i,j} \tag{2}$$

En $G(Tech(F_t^{i,j}) + \epsilon)$ aún necesitamos definir bien cual es la familia G y además qué es $Tech$.

$Tech$ por su parte es una función que construye diferentes indicadores técnicos con base en la información de $F^{i,j}$ hasta t, los indicadores técnicos que se utilizaron fueron los siguientes:

- Media móvil corto plazo (14 periodos)
- Media móvil mediano plazo (50 periodos)
- Media móvil largo plazo (200 periodos)
- Media exponencial corto plazo (14 periodos)
- Momentum (2 periodos)
- MACD 12,26,9
- RSI (14 periodos)
- CCI (20 periodos)
- VHF (28 periodos)

Por su parte para definir la forma de G se probaron los siguientes modelos reconocidos de clasificación de machine learning [9]:

- LogisticRegression
- Random Forest
- Suport Vector Machine

Con base en los indicadores técnicos seleccionados, se procedió con el entrenamiento de los modelos de aprendizaje supervisado para cada uno de los cluster identificados.

Sin embargo, adicionalmente se proponen los siguientes 3 métodos para el entrenamiento de los algoritmos, incluyendo técnicas de extracción de nuevas características, reducción de dimensionalidad a través de selección de variables y componentes principales.

7.2.1. ***Aumentando dimensionalidad modelo dinámico****:* Esta alternativa busca extraer nuevas características a partir de las series de los retornos diarios de los activos sintéticos $F^{i,j}$, a diferencia de utilizar solo los indicadores técnicos de la propuesta inicial. Para este ejercicio se utilizó la librería tsfresh en python, la cual permite extraer múltiples características a partir de una serie de tiempo dada.

Tsfresh tiene 63 métodos de caracterización de series de tiempo, que computa un total de 794 características, también tiene una selección de variables basado en pruebas de hipótesis, identificando las características que son estadísticamente significativas. [10]

De esta forma para cada pareja de activos de cada uno de los cluster, el retorno de los activos ficticios, se extrajeron para cada día 794 características a partir de la información de los retornos diarios de los 50 días previos. Una vez consolidada la información de todos los activos sintéticos para cada cluster se procede a realizar una selección de las variables más relevantes.

Para cada uno de los cluster se procede a entrenar los siguientes modelos variando los parámetros principales:

- LogisticRegression
- GradientBoostingClassifier
- Suport Vector Machine
- Random Forest

A diferencia de la primer propuesta, para cada año se selecciona el mejor modelo de los 4 evaluados utilizando validación cruzada, de esta forma el algoritmo de entrenamiento es dinámico en la selección del mejor modelo de aprendizaje, por consiguiente cada cluster puede tener en el tiempo total de validación diferentes modelos según su comportamiento.

7.2.2. ***Bajando la dimensionalidad: Selección de variables****:* Para selección de variables se usa el modelo Lasso (Least Absolute Shrinkage and Selection Operator). Como método de regresión regularizada, este modelo genera un análisis de regresión que realiza selección de variables y regularización para mejorar la exactitud e interpretabilidad del modelo estadístico producido por este [11].

Donde, en la fórmula, se tiene integrada la restricción que tiene la función para la penalización de los coeficientes de la función por medio del parámetro λ. Para lograr de esta forma estabilizar las estimaciones y predicciones y, por ende, realizar la selección de variables.

Lasso reduce la variabilidad de las estimaciones por la reducción de los coeficientes y al mismo tiempo produce modelos más interpretables y simples por la reducción de algunos coeficientes a cero.

$$\hat{\beta}^{lasso} = argmin_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{K} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{K} |\beta_j| \right\} \tag{3}$$

Utilizando Lasso, es posible obtener un modelo con buena precisión y que sea interpretable, pero este método también tiene varias limitaciones como lo son las siguientes:

- En el caso $P > n$ Lasso selecciona a lo sumo n variables antes de saturarse, debido a la naturaleza del problema de optimización convexa y esto podría ser una limitación para un método de selección de variables. Además, Lasso no está bien definido a menos que el límite de la norma L1 de los coeficientes sea menor que un cierto valor.
- Si existe un grupo de variables entre las cuales las correlaciones por parejas son muy altas, entonces Lasso tiende a seleccionar sólo una variable del grupo, sin importarle cuál de ellas selecciona.
- Para el caso $n > P$, si existe una alta correlación entre los predictores, se ha observado que, en general, la predicción a través de regresión Ridge resulta más óptima que la obtenida a través de Lasso.

Para este modelo frecuentista se utilizó $cross\ validation$ para calibrar el parámetro λ determinando el nivel de $shrinkage$ a aplicar a los $betas$.

7.2.3. ***Bajando la dimensionalidad: Analisis de componentes principales****:* La idea en este caso fue bastante sencilla, analisis de componentes principales, PCA por sus siglas en ingles, es una metodologia comunmente utilizada para bajar la dimemnción de los datos intentando utilizar menos caracteristicas de las iniciales, es decir, la idea es seleccionar menos caracteristicas pero no perder información util.

La forma en la que se implementó fue aplicando la transformación lineal a los datos de train y test pero el analisis solo se hacia sobre los datos de entrenamiento. El numero de componentes utilizados fue el menor numero que explicara al menos el 90 % de la varianza de los datos.

7.3. *Creación del portafolio activo*

Para el caso en el que tenemos un portafolio de sólo dos activos tener dicha función G es suficiente para ganarle a un portafolio EW (En caso de que G esté bien calibrada). Para el caso en el que se tengan k activos, se van a tener $kC2$ parejas, por lo que toca hacer algo para convertir las diferentes clasificaciones en un único portafolio.

La propuesta metodológica para crear el portafolio es crear un sistema de votaciones, además una zona de indecisión de los algoritmos en la que no se tomen apuestas. De esta forma definimos la formación de los portafolios de la siguiente manera:

$$votos_t^{i,j} = \begin{cases} i,i & G(Tech(F_t^{i,j})) \geq 0{,}6 \\ j,j & G(Tech(F_t^{i,j})) \leq 0{,}4 \\ i,j & \text{En otro caso} \end{cases} \quad (4)$$

Con base en todos los resultados de las votaciones todas las parejas i y j, llegamos finalmente al portafolio definido de la siguiente forma:

$$w_t^i = \frac{\text{votos} \quad i}{2 * kC2} \quad (5)$$

En modo de ejemplo consideremos los siguientes resultados con $k = 3$:

- $G(Tech(F_t^{1,2})) = 0{,}7$
- $G(Tech(F_t^{2,3})) = 0{,}56$
- $G(Tech(F_t^{3,1})) = 0{,}35$

De este modo el vector de votos sería [1,1,2,3,1,1] y el portafolio propuesto para el periodo t sería: $[\frac{4}{6}, \frac{1}{6}, \frac{1}{6}]$

7.4. *Backtest*

La evaluación de los algoritmos en este caso no se hizo con métricas usuales en problemas de clasificación como lo son el accuracy, el AUC o recall. En este problema lo que realmente buscamos hacer es tener un sistema que de un mejor retorno que el benchmark, para evaluar esto lo que hacemos es una simulación del retorno que obtendriamos utilizando el portafolio propuesto haciendo rebalanceos todos semanales (los miercoles). La forma en la que hicimos esto fue la siguiente:

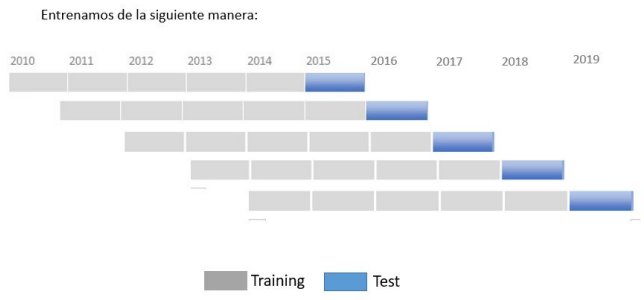


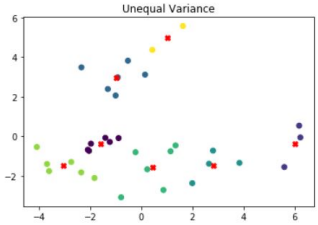
Figura 6: Train - Test

8. RESULTADOS

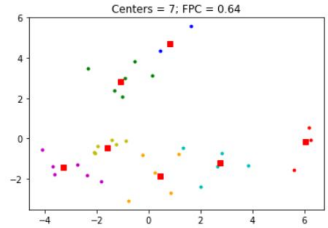
8.1. *Resultado Clustering: encontrando activos similares*

Después de evaluar diferentes opciones de cluster se encontró que con K=7, los grupos obtenidos tienen cierta lógica desde el punto de vista financiero, donde se dividen de la siguiente manera:

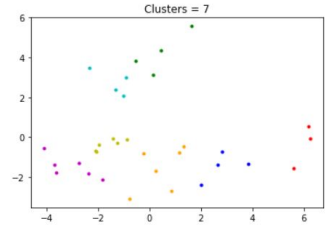
- Todos los elementos de Europa están juntos, lo cual hace pensar que es muy difícil buscar diversificación en un portafolio solo con acciones de Europa.
- A pesar de que los países de Asia que fueron seleccionados son considerados en general como Asia emergente, los resultados muestran que hay un par de subgrupos.
- Todo Estados Unidos se comporta de forma muy parecida, eso lo vemos en la consistencia de los clusters de los diferentes sectores. Según los resultados, parece que la única forma de tener algún sector que de un poco de diversificación seria invertir en 3 grandes bloques.
- Australia y Pakistán tienen una dinámica muy diferente al resto de ETF's de la muestra.



(a) K-Means



(b) Fuzzy-C-Means



(c) Spectral

Figura 7: Clusters ETFs PCA

Cuando evaluamos el resultado de los cluster construidos por los diferentes algoritmos, encontramos que los grupos estimados son consistentes en los tres modelos de clasificación.

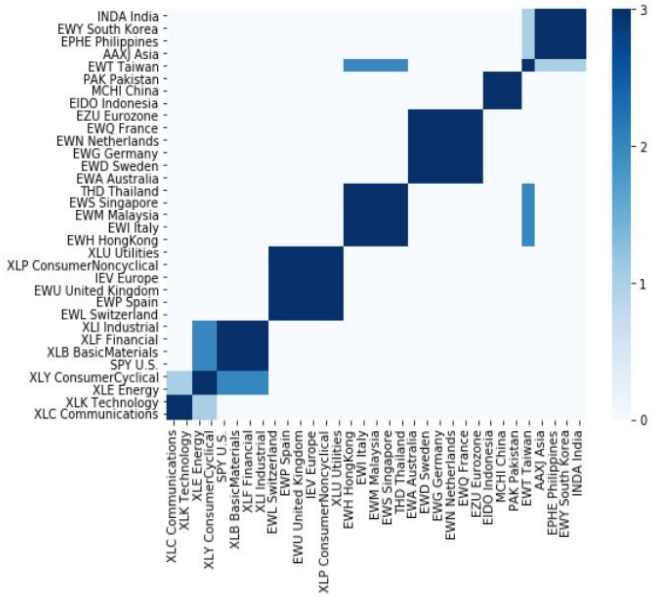


Figura 8: Resultados Consistencia ETF's k=7

De lo anterior, dado que dos de los cluster son poco informativos, se decidió que solo 5 de los 7 grupos financieramente tienen sentido. Adicionalmente, se hizo una selección con los ETF's mas representativos del mercado, dando como resultado:

- **Cluster1 Asia Emergente:** China, Hong Kong y Korea.
- **Cluster2 Europa:** Francia, Alemania y United Kingdom.
- **Cluster3 EEUU Sectores 1:** Industrial, Materials, Financial y Energy.
- **Cluster4 Asia Subdesarrollado:** India, Tailandia y Malasia.
- **Cluster5 EEUU Sectores 2:** Standar Poor's, Technologyy Consumer Discretionary.
- **ClusterAll:** Se toma una ETF's de cada cluster anterior que son, Estados Unidos SP y sector financiero, India, Francia y China.

8.2. *Modelos de decisión*

Una vez seleccionados los cluster se procede a ejecutar para cada uno de ellos las 4 técnicas propuestas y sus modelos de aprendizaje seleccionados:

Indicadores técnicos puros:

- LogisticRegression
- Suport Vector Machine
- Random Forest

Indicadores técnicos con selección de variables:

- LogisticRegression
- Suport Vector Machine
- Random Forest

Indicadores técnicos aplicando PCA:

- LogisticRegression
- Suport Vector Machine
- Random Forest

Modelo dinámico con extracción de nuevas características y selección de mejor modelo anual:

- LogisticRegression, GradientBoostingClassifier, Suport Vector Machine, Random Forest.

De esta forma, en total para cada cluster se entrenaron 10 modelos, con un periodo de entrenamiento de 5 años y ejecución en validación durante el año siguiente antes de actualizar nuevamente los datos de entrenamiento eliminando el año mas antiguo e ingresando año mas reciente.

Para evaluar los resultados de las técnicas y determinar los mejores modelos, se definió como rango de validación para medir el desempeño de los portafolios el periodo comprendido entre los años 2015 y 2019. Luego se realizaron 2 análisis de acuerdo con las métricas comúnmente utilizadas en la gestión de portafolios.

8.2.1. ***Evaluación con el alpha****:* En primer lugar para cada cluster evaluamos la generación de $alpha$ de los portafolios construidos por cada uno de los modelos entrenados frente al portafolio benchmark $naive$ durante el periodo completo de validación, con el fin de determinar para cada cluster que modelo arrojo el mejor resultado en términos de rentabilidad. Los resultados mostraron que los mejores modelos fueron los siguientes:

- **Cluster 1:** Suport Vector Machine radial
- **Cluster 2:** PCA Random Forest
- **Cluster 3:** Dinámico anual
- **Cluster 4:** PCA Random Forest
- **Cluster 5:** Variable selection Random Forest
- **Cluster All:** Variable selection Random Forest

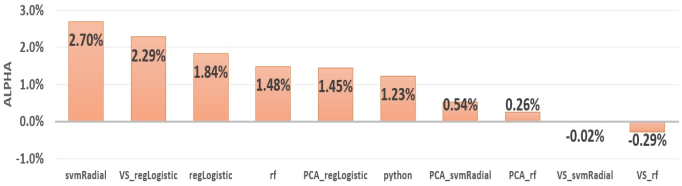


Figura 9: Resultados alpha cluster1

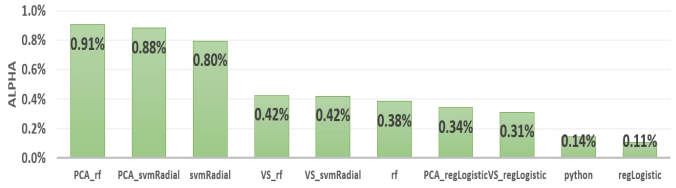


Figura 10: Resultados alpha cluster2

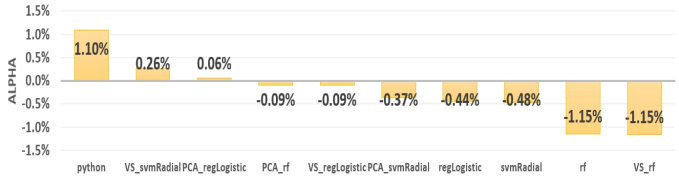


Figura 11: Resultados alpha cluster3

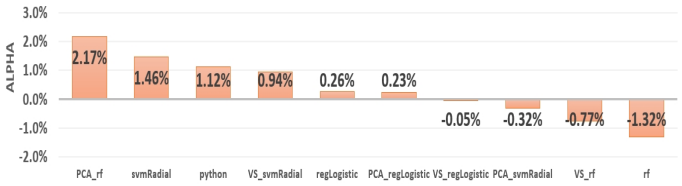


Figura 12: Resultados alpha cluster4

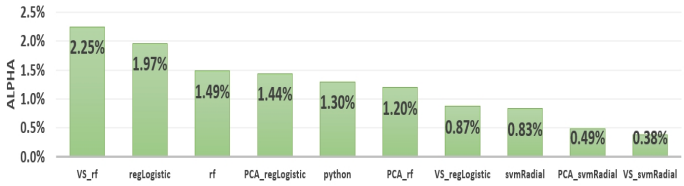


Figura 13: Resultados alpha cluster5

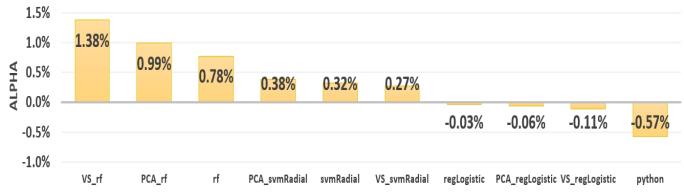


Figura 14: Resultados alpha clusterAll

A continuación se presenta la selección de los mejores modelos anuales del modelo dinámico.

	Cluster1	Clsuter2	Cluster3	Cluster4	Cluster5	ClusterAll
2015	SVM	SVM	SVM	Lineal	Lineal	SVM
2016	RF	Lineal	SVM	SVM	Lineal	RF
2017	Lineal	Lineal	SVM	RF	Lineal	SVM
2018	Lineal	Lineal	SVM	SVM	SVM	SVM
2019	SVM	Lineal	SVM	SVM	Lineal	SVM

Figura 15: Modelo dinámico-modelo anual

8.2.2. ***Consistencia****:* En segundo lugar, se definió una metodología de votación que permitiera determinar la consistencia de cada uno de los 10 modelos, considerando su desempeño en todos los cluster y en cada uno de los años evaluados, a diferencia de la primer medición, la cual solo premia el resultado final y no la estabilidad en el tiempo y capacidad de generalización.

Para la votación se calcularon para cada uno de los modelos, en los diferentes clusters, las siguientes métricas relacionadas con la gestión activa de portafolios:

- alpha
- sharpe ratio fig:Sharpe
- Information ratio fig:Informatio

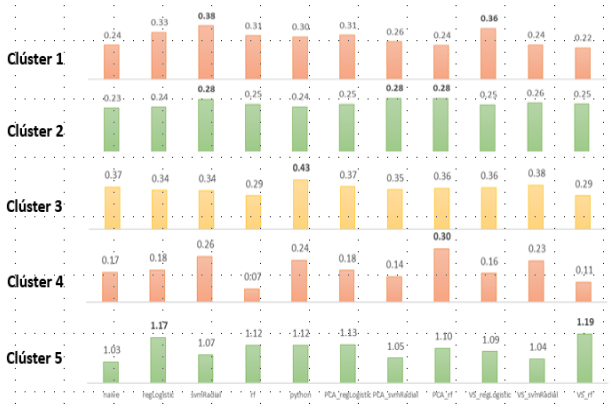


Figura 16: Sharpe Ratio

Figura 17: Informatio Ratio

En total cada modelo tiene la posibilidad de obtener un total de 18 puntos, es decir el numero de clusters por el numero de metricas.

De esta forma el criterio para otorgar un voto a favor de uno de los modelos, consiste en que el modelo debe obtener un indicador superior o favorable frente al benchmark $naive$. En caso de empate entre algunos de los modelos, se utiliza como item de desempate el information ratio que recoge el efecto del tracking error como medida de variabilidad del $alpha$. Los resultados fueron los siguientesfig:consistencia:

	regLogisti	svmRadia	rf	python	PCA_regLogis	PCA_svmRadi	PCA_rf	VS_regLogist	VS_svmRadi	VS_rf
Cluster 1	18	15	15	15	18	13	9	18	6	9
Cluster 2	15	12	13	11	15	13	15	12	12	11
Cluster 3	3	6	3	12	12	6	6	6	9	5
Cluster 4	9	9	9	9	9	3	11	5	12	7
Cluster 5	15	11	11	11	14	15	15	15	9	14
Mejor Modelo	60	53	51	58	68	50	56	56	48	46

Mejor Modelo - se desempata por IR
Igual calificación

Figura 18: Consistenca

Estos resultados se pueden ver en este repositorio[1]

[1] https://github.com/camilaMejia/proyectoIntegrador2/tree/mateo

9. SOFTWARE UTILIZADO

El software utilizado es R en su versión 3.6.1 y Python 3.7

10. CONCLUSIONES Y RECOMENDACIONES

Los resultados muestran que la metodología propuesta puede terminar en la construcción de portafolios que vencen a un benchmark.

Se evidencia que en todos los cluster se obtuvieron modelos capaces de vencer al benchmark equally weighted, lo cual no es tan sencillo de lograr en la práctica, considerando que se tuvieron en cuenta portafolios conformados por mercados muy diferentes como el Europeo, Asiático y Americano. Puntualmente una firma administradora de activos requiere de un equipo humano bastante robusto y bien calificado para lograr abarcar todos los mercados, lo cual es bastante costoso, e inclusive no garantiza obtener resultados consistentes.

En términos de alpha, en los cluster 2 y 5 el 100 % de los modelos lograron vencer al benchmark, en el cluster 1 el 80 % de los modelos, los cluster 4 y All el 60 % de los modelos, y solo en el cluster 3 menos del 50 % de los modelos le ganaron al benchmark; evidenciando en general un buen comportamiento de los 10 modelos implementados en los diferentes clusters.

La selección del mejor modelo para los cluster puede variar en función de la métrica seleccionada, por lo cual es importante que el administrador defina su objetivo; es decir, si desea optimizar la generación de alpha únicamente o ser más consistente en el tiempo en términos del information ratio, es decir un alpha positivo que no tenga mucha volatilidad.

En términos de consistencia, el modelo que tuvo el mejor comportamiento anual consolidando los resultados de todos los clusters, fue el de regresión logística entrenado a partir de las componentes principales de los indicadores técnicos originales. Por consiguiente, si tenemos un nuevo cluster de activos, recomendamos utilizar este modelo para medir la capacidad inicial de aprendizaje.

Igualmente el análisis de consistencia permite observar que los modelos de regresión logística y el dinámico tienen buen comportamiento en términos de generalización en el tiempo, lo cual es muy relevante a la hora de ranquear a los mejores gestores.

Dentro de las posibles mejoras al trabajo se encuentra una mejor estimación de los rangos para tomar decisión en los votos, hoy el 40 % y 60 % fueron seleccionados sin una base teórica. Otra posible mejora es una mejor selección de los años de entrenamiento y de test, es posible que sea óptimo utilizar 2 años de entrenamiento y re-entrenar cada 3 meses, no cada 5 años.

Adicionalmente observamos que los modelos tienen un buen comportamiento cuando se realiza el ejercicio de clustering no supervisado, es decir, cuando existe una similitud entre los activos, por tanto se recomienda realizar más pruebas sobre portafolios que incluyan activos de diferentes clusters.

Probar los modelos contra un benchmark que no sea el equally weighted para evaluar la capacidad de generalización.

REFERENCIAS

[1] B. K. Marcus, "Investments," 1989.

[2] D. P. Brown and R. H. Jennings, "On technical analysis," *The Review of Financial Studies*, vol. 2, no. 4, pp. 527–551, 1989.

[3] E. O. Thorp and S. T. Kassouf, "Beat the market: A scientific stock market system," 1967.

[4] T. L. Xiao Ding, Yue Zhang and J. Duan, "Deep learning for event-driven stock prediction," *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[5] S. C. Pushpa BV and M. Hegde, "Investment decision making using technical analysis: A study on select stocks in indian stock market," *IOSR Journal of Business and Management*, vol. 19, no. 9, pp. 24–33, 2017.

[6] A. M. O. Omer Berat Sezer and E. Dogdu, "An artificial neural network-based stock trading system using technical analysis and big data framework," 2017.

[7] O. B. Sezer and A. M. Ozbayoglu, "Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach," 2018.

[8] R. Xu and D. C. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks, Institute of Electrical and Electronics Engineers (IEEE)*, 2005.

[9] S. Kotsiantis, "Supervised machine learning: a review of classification techniques," *Emergin Artificial Intelligence Applications y Computer Engineering*, 2007.

[10] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package)," *Neurocomputing*, vol. 307, pp. 72–77, 2018.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of Royal Statistical Society: Series B (methodology)*, vol. 67, no. 1, pp. 91–108., 1996.