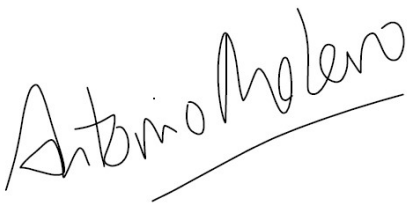
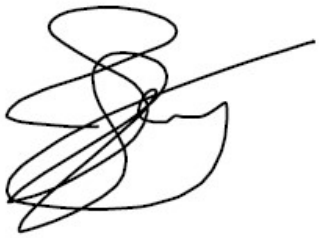
APLICACIÓN DE TÉCNICAS DE ANÁLISIS DE DATOS A LA GESTIÓN ÓPTIMA DE UNA CARTERA DE VALORES

AUTOR:
ANTONIO LUIS MOLERO SENOSIAIN

DIRECTORES:
JUAN LUIS ZAMORA MACHO
JUAN MARTÍNEZ OLONDO





Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

APLICACIÓN DE TÉCNICAS DE ANÁLISIS DE DATOS A LA GESTIÓN ÓPTIMA DE UNA CARTERA DE VALORES

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2017-2018 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Fecha: 24/07/2018

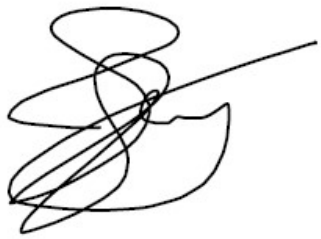
ANTONIO LUIS MOLERO SENOSIAIN

Autorizada la entrega del proyecto

LOS DIRECTORES DEL PROYECTO

Fdo.: Fecha: 24/07/2018

JUAN LUIS ZAMORA MACHO Y JUAN MARTÍNEZ OLONDO



AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. ANTONIO LUIS MOLERO SENOSIAIN DECLARA ser el titular de los derechos de propiedad intelectual de la obra: APLICACI'ON DE T'ECNICAS DE ANÁLISIS DE DATOS A LA GESTIÓN ÓPTIMA DE UNA CARTERA DE VALORES, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.
b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
e) Asignar por defecto a estos trabajos una licencia Creative Commons.
f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente)*.

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

a) Que la Universidad identifique claramente su nombre como autor de la misma
b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
c) Solicitar la retirada de la obra del repositorio por causa justificada.
d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

El autor se compromete a:

a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.
b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.

d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

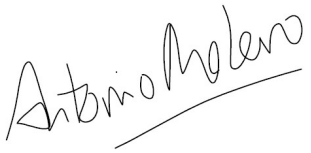
6º. Fines y funcionamiento del Repositorio Institucional.

La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a 24 de JULIO de 2017

ACEPTA



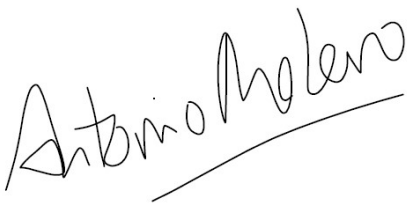
Fdo……

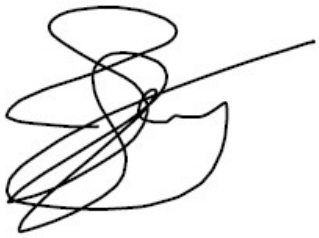
Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:

APLICACIÓN DE TÉCNICAS DE ANÁLISIS DE DATOS A LA GESTIÓN ÓPTIMA DE UNA CARTERA DE VALORES

AUTOR:
ANTONIO LUIS MOLERO SENOSIAIN

DIRECTORES:
JUAN LUIS ZAMORA MACHO
JUAN MARTÍNEZ OLONDO





APLICACIÓN DE TÉCNICAS DE ANÁLISIS DE DATOS A LA GESTIÓN DE UNA CARTERA DE VALORES

Autor: **Molero Senosiain, Antonio Luis.**
Directores: Zamora Macho, Juan Luis.
Martínez Olondo, Juan.
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

1. INTRODUCCIÓN

Desde la creación de los mercados de valores, cazadores de fortunas y académicos han tratado de predecir el comportamiento tanto de los mercados como de las acciones individuales con el doble objetivo de la obtención de beneficios financieros y la obtención de una visión económica a futuro. Al tratarse de un sistema tan complejo que depende de un gran número de variables, el objetivo de predecir su comportamiento es considerablemente ambicioso, y más teniendo en cuenta que los estudios no se terminan de poner de acuerdo en si dicha predicción es si quiera posible. De ahí el gran interés que supone la realización de este proyecto.

Por otro lado, el aprendizaje automático es un campo de la ciencia computacional que estudia y aplica las máquinas capaces de encontrar patrones, generalizar y aprender sin ser explícitamente programadas. En los últimos años, las técnicas relacionadas con Machine Learning han hecho posible analizar de una forma más eficaz las concentraciones masivas de datos, lo que a su vez ha causado que se hayan hecho grandes esfuerzos para intentar predecir el mercado de valores.

Quedando esto presente, la principal motivación de este proyecto es la investigación.

En este proyecto se pueden diferenciar tres objetivos principales:

- Desarrollo de un modelo de clustering que identifique diferentes patrones de cotizaciones
- Diseño de un entorno de simulación para evaluar la calidad de los clusters.
- Análisis de los clusters identificados para la detección de posibles nichos de rentabilidad en función de la tendencia y la volatilidad esperadas.

2. METODOLOGÍA

En primer lugar, el proyecto se ha realizado utilizando Matlab con la extensión de aprendizaje automático.

La metodología seguida para la creación de modelo y su evaluación es la estándar para este tipo de análisis. En primer lugar, se han dividido los datos obtenidos en tres ventanas de diferentes tamaños. La primera ventana se ha utilizado para el entrenamiento del algoritmo y la creación de clusters, y contiene aproximadamente el 50% de los datos. Con la segunda ventana, compuesta por un 30% de los datos aproximadamente, se ha realizado la fase de análisis, que consiste en la modificación de los metaparámetros para la obtención de resultados acordes a lo esperado, es decir, se han modificado el número de clusters y el tamaño de ventana hasta obtener clusters interesantes (con alta rentabilidad y baja variabilidad). Una vez logrados buenos resultados

entrenando el algoritmo con ciertas combinaciones de parámetros y comprobándolo con los datos de análisis, se ha procedido a utilizar la última ventana que contiene el 20% restante de los datos, la de test, para hacer una segunda confirmar que los clusters que en las dos primeras etapas parecían buenos siguen siendo interesantes.

De esta manera, se han utilizado dos técnicas diferentes de clustering (k-means y GMM) y otras dos técnicas de reducción de dimensionalidad de los parámetros.

Por otra parte, para este proyecto se ha estimado oportuna la utilización de los precios de cierre de la acción, esto son, los precios que tienen las diferentes acciones con el cierre de mercado al final del día. Para la elección de estos datos se ha partido de la premisa de que el precio de la acción, como elemento último de representación del mercado, recoge información intrínseca de todos los diferentes factores que afectan al comportamiento del mercado de forma ponderada.

Con el objetivo de alimentar al algoritmo con la mayor cantidad de evoluciones de precios de acciones posible, se han elegido un total de 898 empresas, repartidas en los mercados que muestra la siguiente tabla:

Tabla 1. Mercados e índices utilizados para la realización del proyecto

Índice/ Mercado	País
IBEX 35	España
IBEX MEDIUM CAP	
IBEX SMALL CAP	
MERCADO CONTINUO	
FTSE 100	Reino Unido
MILAN	Italia
LISBOA	Lisboa
BRUSELAS	Bélgica
AMSTERDAM	Holanda
DAX 30	Alemania
VIENA	Austria
S&P 500	Estados Unidos
NASDAQ 100	
TORONTO TSX 60	Canadá

Los datos relativos a los diferentes mercados de valores y los datos financieros en general existen con una resolución desde milisegundos a décadas. Cuanto mayor es la resolución, igual que en cualquier otro tipo de sistema dinámico, mayor es la precisión y más detalles se pueden observar, pero a cambio de tener más ruido. Además, la información financiera en alta resolución no es accesible normalmente a no ser que se tenga algún tipo de cuenta premium que supondría un pago periódico. Así, se ha elegido una resolución diaria ya que, aunque puede haber fluctuaciones por eventos puntuales, al tener únicamente en cuenta el precio de cierre se ha entendido que estos cambios inusuales no tendrán un gran peso en la creación del modelo.

Una vez escogida la definición de tiempo a utilizar, se ha elegido el periodo de tiempo en el que se va a trabajar. La utilización de datos muy antiguos podría crear un modelo desfasado en el tiempo, que no tendría suficiente precisión. Por el contrario, cuanto más largo sea el periodo de

tiempo, más datos se tendrán y por lo tanto mayor será la robustez del sistema, siendo de hecho imposible crear el modelo si no se llega a un mínimo de datos.

Para la realización de este proyecto se ha considerado interesante mirar hacia atrás en el tiempo en un periodo de 15 años en los casos en los que ha sido posible (de algunas empresas solo hay datos de los últimos 10 o 12).

Para las distintas fases de entrenamiento, validación y test, se suelen dividir los datos en un 50%, 30% y 20% relativamente. En este caso, además se han querido recoger comportamientos similares, es decir, que cada conjunto de datos tenga una bajada y una recuperación. De esta manera, y para cumplir aproximadamente los porcentajes estándar se ha dividido el tiempo de la siguiente manera:

- Entrenamiento: desde junio de 2.003 hasta mayo de 2.011
- Validación: desde junio 2.011 hasta mayo de 2.015
- Test: desde junio de 2.015 hasta junio de 2.018

Con esta división de los datos, el entrenamiento contiene la crisis financiera de 2.008, la validación el bajón de 2.011 y el test la caída de 2.016, quedando los porcentajes de datos en el tiempo aproximadamente como un 53% para el entrenamiento, 27% para la validación y 20% para el test. Estas proporciones son por lo tanto suficientemente parecidas a las indicativas 50 – 30 - 20 del procedimiento estándar.

La forma en la que se ha diseñado el sistema es relativamente simple: se divide una ventana de un determinado número de días que se puede cambiar como parámetro en dos. Estas dos ventanas se ha escogido que tengan una relación de tamaño uno a uno.

Con el objetivo de homogeneizar los datos para poder realizar un análisis de conjunto en el que se comparen rentabilidades o evolución se ha procedido a una normalización de las ventanas. Dicha normalización se ha realizado dividiendo todos los valores de la ventana por el primer valor de la ventana.

Además, para la clasificación de las ventanas de entrada no se ha trabajado con los precios finales brutos, sino que se han sacado parámetros de cada ventana.

Al utilizar datos históricos financieros, es importante recalcar que la mayoría de los periodos de tiempo tienen una tendencia general creciente ya que la mayoría de los mercados han crecido en el tiempo. Por lo tanto, el modelo tendrá más datos de tendencias alcistas y será normal que unos clusters tengan más miembros que otros, aunque no se debe despreciar ningún tipo de información pues cada cambio en el desarrollo del mercado se puede aprovechar gracias a la gran variedad de productos financieros existentes.

A partir de cada ventana de precios se han sacado treinta parámetros diferentes. Era necesario reducir la dimensión de treinta a dos o tres por dos razones principales:

- Treinta parámetros son demasiados para el algoritmo de clustering, ya que tardaría mucho tiempo en ejecutar y se necesitaría una potencia computacional mayor a la que se tiene para la ejecución de este proyecto

- Entre los parámetros hay correlaciones altas, siendo uno de los mayores problemas de los algoritmos de aprendizaje no supervisado la correlación sin causalidad [1]. Con la reducción de dimensionalidad lo que se consigue entre otras cosas es eliminar las altas correlaciones lineales que existen entre los diferentes parámetros, construyendo de esta manera un modelo más robusto.

  Para realizar esta reducción de dimensionalidad, se ha elegido la técnica de reducción lineal PCA , que consiste en el mapeo de los datos en un espacio de menores dimensiones de tal manera que la varianza de los datos en las dimensiones reducidas sea máxima.

  También se ha probado el método t-SNE, que ofrece mejores resultados al respetar las distancias euclídeas, pero que se ha descartado por la elevada carga computacional que supone.

  Por otro lado, k-means es la metodología que se comenzó utilizando por ser computacionalmente más ligera, aunque GMM es el que ha dado mejores resultados y ha terminado siendo el preferido, ya que con una diferencia computacional baja, el modelo es considerablemente más robusto con respecto a los valores iniciales que se den a los clusters.

  Por otra parte, para que los diferentes experimentos sean rápidos de comparar, se ha elegido un formato de tabla con cinco columnas:

  - Cluster: muestra el número de cluster del que se están mostrando las características en la fila
  - Members: representa el porcentaje de observaciones que pertenecen a ese cluster. Un cluster será tanto más importante cuanto más porcentaje de observaciones represente.
  - Expected growth: rentabilidad media anual de las ventanas de salida cuyas ventanas de entrada se corresponden con el cluster de la fila.
  - External variation: variación de las rentabilidades de las ventanas de salida de cada uno de los clusters. Es una medida de la similaridad entre las rentabilidades de las ventanas de salida de cada cluster. Un cluster será tanto mejor cuanto menor sea esta variación.
  - Internal variation: variación de la evolución de cada uno de los elementos del cluster con respecto a su recta de rentabilidad.

  Cabe añadir que todos los clusters con un valor en el campo External variation mayor de uno han sido descartados, ya que una desviación típica del 100% se ha considerado el máximo para considerar a un cluster nicho de mercado objetivo. Esta decisión se ha tomado en base a la idea de que una desviación típica superior al 100% implicaría la posibilidad de perder dinero. Además, han sido razón de descarte de resultados también una rentabilidad menor al 20% y una representatividad menor del 1%.

3. RESULTADOS

Se han probado los diferentes algoritmos en dieciséis casos diferentes, modificando los metaparámetros tamaño de ventana, que influirá en la longitud del plazo de predicción; y número de clusters, que influirá en el número de clusters que se formen para agrupar los datos y por lo tanto lo similares que sean los datos de un mismo cluster.

De esta manera, se ha encontrado un cluster que ha pasado las tres etapas de entrenamiento, validación y test entre todos los casos estudiados.

4. CONCLUSIONES

En primer lugar, cabe destacar la importancia que supone la capacidad computacional para este proyecto. La ejecución de algoritmos complejos para el análisis de cantidades tan altas de datos (precio de cierre diario de más de 850 empresas durante quince años) permitirá análisis más exhaustivos cuanto mayor sea esta capacidad, pudiéndose aumentar tanto el tamaño de los datos como la complejidad de los algoritmos.

Los resultados del proyecto han demostrado la existencia de los nichos de mercados de los que se hablaba en los objetivos, y su escasez en los diferentes casos estudiados es una muestra de las dificultades que supone la predicción del mercado de valores tal y como muestran muchas publicaciones, que no terminan de ponerse de acuerdo ni si quiera en la posibilidad de esta predicción. [2] [3]

Además, los requisitos de representatividad, rentabilidad y variación de los resultados para dar un cluster como bueno han sido bastante conservadores, por lo que un inversor con un perfil de riesgo más agresivo podría haber encontrado, a partir de los mismos resultados, más nichos de inversión debido a la subjetividad del asunto.

5. FUTUROS DESARROLLOS

- Utilización de algoritmos más pesados si aumenta la capacidad computacional de los recursos.
- Búsqueda de parámetros y datos de diferente naturaleza a los utilizados.
- Creación de una herramienta de inversión.
- Adición de datos no utilizados en este proyecto

6. REFERENCIAS

[1] D. Leinweber, "Stupid Data Miner Tricks, overfitting the S&P 500," [Online]. Available: https://www.researchgate.net/publication/247907373_Stupid_Data_Miner_Tricks_Overfitting_the_SP_500. [Accessed 11 07 2018].

[2] A. W. L. y. A. C. MacKinlay, "Stock Market Prices Do Not Follow Random Walks: Evidence From a Simple Specification Test," 02 1987. [Online]. Available: https://www.nber.org/papers/w2168. [Accessed 28 06 2018].

[3] "Random Walk Theory," [Online]. Available: https://www.investopedia.com/terms/r/randomwalktheory.asp. [Accessed 28 06 2018].

DATA ANALYSIS TECHNIQUES APPLICATION FOR THE OPTIMAL MANAGEMENT OF A STOCK PORTFOLIO

Author: **Molero Senosiain, Antonio Luis.**
Directors: Zamora Macho, Juan Luis.
Martínez Olondo, Juan.
Collaborating Entity: ICAI – Universidad Pontificia Comillas

ABSTRACT OF THE PROJECT

1. INTRODUCTION

Since the creation of the stock markets, fortune hunters and researchers have tried to predict the behavior of markets and individual stocks with the double objective of obtaining financial benefits and a future economic sight. As it is such a complex system, that depends on large amounts of variables, the objective of predicting its behavior is considerably ambitious, even more taking into account that experts do not agree on the predictability of the market. Thus, the great interest that this project arises.

On the one hand, machine learning is a field of computer science that studies and applies machines that are able to find patterns, generalize and learn without having been explicitly programmed. In the last years, a lot of techniques related with Machine Learning have made possible to effectively find massive data amounts agglomerations, which at the same time has provoked an increase on the efforts of trying to predict the stock market behavior.

Being this said, the main motivation of this project is the investigation work behind it.

Three main objectives can be differentiated in this project:

- Development of a clustering model to identify different evolution patterns.
- Design of a simulation environment in order to analyze the quality of the clusters.
- Clusters analysis in order to detect possible investment niches depending on their profitability and volatility.

En este proyecto se pueden diferenciar tres objetivos principales:

- Desarrollo de un modelo de clustering que identifique diferentes patrones de cotizaciones
- Diseño de un entorno de simulación para evaluar la calidad de los clusters.
- Análisis de los clusters identificados para la detección de posibles nichos de rentabilidad en función de la tendencia y la volatilidad esperadas.

2. METODOLOGY

The project has been performed using Matlab and its machine learning packages.

The methodology used in order to create and evaluate the model is the standard for this type of analysis. Firstly, the data has been divided in three different data sets. The different sets created have been training set, validation set and test set, and they have different sizes. The training set contains about a 50% of the data, the validation set around a 30%, and the test set the left 20%.

Training phase consists on the creation of the clusters from the training set of data; validation phase is about trying different metaparameters combinations in order to analyze the results of the training phase; while in the test phase, the interesting clusters of the validation phase are analyzed in order to confirm their quality.

The metaparameters modified for the different cases have been the window length, which is a measure of the prediction horizon, and the number of clusters.

In the project, two different clustering techniques have been used: k-means and GMM; and another two different dimensionality reduction methods: PCA and t-SNE.

On the other hand, for this project only the closing prices have been used as data, based on the assumption that prices have all the information and the weighted actuation of the different variables of the market on them.

In order to feed the algorithms with the largest amount of data possible (to increase the system's resilience), data from a total of 898 different companies has been used. These companies are in the different stock markets that the table below shows:

Table 1. Stock markets and indexes used for the Project realization

Índice/ Mercado	País
IBEX 35	España
IBEX MEDIUM CAP	
IBEX SMALL CAP	
MERCADO CONTINUO	
FTSE 100	Reino Unido
MILAN	Italia
LISBOA	Lisboa
BRUSELAS	Bélgica
AMSTERDAM	Holanda
DAX 30	Alemania
VIENA	Austria
S&P 500	Estados Unidos
NASDAQ 100	
TORONTO TSX 60	Canadá

All the data concerning the different stock markets, and financial data in general, exist from milliseconds to decades resolution. The higher the resolution, the higher the information they have, but the higher the noise too. In addition, very detailed financial information is not usually available for free on the internet, being necessary to have some kind of premium account, paying a periodical fee. This being said, a diary resolution has been chosen as it has been considered accurate enough to have important information in it but not too much in order to avoid intra-daily atypical behaviors.

After having the time resolution defined, the next step has been to choose the time scale of the data. Using very old data could create a non-valid model due to the development that happens in the market over time, but a very short period of time would not give enough data in order to create a resilient system.

The period of time of the data used has been determined as 15 years from the moment of the study, although

some companies do not have this large amount of time data.

For the different training, validation and test phases, an attempt of having similar market behaviors in each of the data sets has been done. Therefore, time has been divided following the next time frames:

- Training set: from June 2003 to May 2011
- Validation set: from June 2011 to May 2015
- Test set: from June 2015 to June 2018

This data split makes each data set to have a growth and a falling period, having the training set the 2008 financial crisis, the validation set the 2011 fall and the test set the 2016 fall, and being the different percentages similar to de standard ones named before (50-30-20): 53% for the training set, 27% for the validation set and 20% for the test set.

The system has been designed in a relative simple way: windows with a specific days-length are divided in two of the same size.

In order to homogenize the data to be able to perform an overall analysis where profitability and evolution can be done, a normalization has been done by dividing every element of each window by the first price of the window.

In addition, clustering of the different windows has not been done over the raw prices data but over different parameters calculated from it.

When using historical financial data, it is important to highlight that most of the time the market has grown, and it will be normal if some clusters have more members than others. Non information should be despised in this analysis because the best investing opportunities usually appear in changing situations and there are a lot of different financial products in the market.

From each of the prices windows, thirty parameters have been calculated, and it has been necessary to reduce the dimension of the data because of two main reasons:

- Thirty parameters are a lot for the clustering algorithm to compute, because it would take a long time and computing capacity for it.
- There are a lot of high correlations between parameters, and using the dimensionality reduction can avoid correlation without causality, which is one of the most common problems in machine learning. [1]

To perform the dimensionality reduction, the used technique has been PCA. PCA is a linear dimensionality reduction algorithm that maps the different data in a space with less dimensions while maximizing variance between the different elements.

An attempt of using t-SNE algorithm has been done, but it is too computationally demanding to be performed with the available equipment for the project.

On the other hand, k-means is the clustering algorithm that was used at the beginning of the project, but over time it was switched to GMM for being less dependent on the initial conditions by increasing computational capacity needed by a little.

In order to compare and read the different results of the experiments, a table format with five columns has been chosen:

- Cluster: identification number of the cluster
- Memers: percentage of the observations contained in the cluster. A cluster will be more important if this number is high.
- Expected growth: mean of the expected normalized growths of the elements of the cluster.
- External variation: volatility of the profitability of the different elements of the same cluster. It measures similarity between members of the same cluster. A cluster will be better if this variation is low.
- Internal variation: measures the volatility of the price behavior compared with the profitability straight line.

Every cluster with an External variation value higher than 1 has been discarded, same as clusters with expected growth lower than 20% in absolute numbers and clusters with a membership percentage lower than 1%.

3. RESULTS

The algorithm has been run sixteen different times with different metaparameters configurations. The metaparameters that have changed their values are the window length, that is related with the prediction horizon, and the number of clusters, which determines how big groups will be and hence how similar elements in it will be.

This being said, a cluster that has "passed" the three phases of training validation and test has been found in all the sixteen studied cases.

4. CONCLUSIONS

Firstly, it is important to highlight the importance of the computational capacity for this project. The execution of complex algorithms for the large amounts of data that has been used in the project will be better with more powerful computers, being possible to increase the size of the data and the complexity of the algorithms.

The project results have proved the existence of the investing niches named in the initial objectives of the project, and the low number of them that have been found in the analysis is a clear signal of how complex is to predict the stock market behavior. This complexity is such that a lot of researches do not agree whether it is possible to predict the market or not. [2] [3]

In addition, the necessary conditions for a cluster to be considered good in membership percentage, profitability and volatilities have been set with a conservative mind. This means that an investor with a more aggressive risk-profile will probably find more interesting the same numerical results that have been discarded in the results analysis of this project.

5. FUTURE DEVELOPMENT

- Utilization of more complex algorithm if computational capacity resources increase.
- Use of parameters and data from a different nature.
- Development of an investing tool.

- Increase in the amount and type of data used.

6. REFERENCES

[1] D. Leinweber, "Stupid Data Miner Tricks, overfitting the S&P 500," [Online]. Available: https://www.researchgate.net/publication/247907373_Stupid_Data_Miner_Tricks_Overfitting_the_SP_500. [Accessed 11 07 2018].

[2] A. W. L. y. A. C. MacKinlay, "Stock Market Prices Do Not Follow Random Walks: Evidence From a Simple Specification Test," 02 1987. [Online]. Available: https://www.nber.org/papers/w2168. [Accessed 28 06 2018].

[3] "Random Walk Theory," [Online]. Available: https://www.investopedia.com/terms/r/randomwalktheory.asp. [Accessed 28 06 2018].

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1 INTRODUCCIÓN

1.1 MOTIVACIÓN Y OBJETIVOS DEL PROYECTO

Desde la creación de los mercados de valores, cazadores de fortunas y académicos han tratado de predecir el comportamiento tanto de los mercados como de las acciones individuales con el doble objetivo de la obtención de beneficios financieros y la obtención de una visión económica a futuro. Al tratarse de un sistema tan complejo que depende de un gran número de variables, el objetivo de predecir su comportamiento es considerablemente ambicioso, y más teniendo en cuenta que los estudios no se terminan de poner de acuerdo en si dicha predicción es si quiera posible. De ahí el gran interés que supone la realización de este proyecto.

Por otro lado, el aprendizaje automático es un campo de la ciencia computacional que estudia y aplica las máquinas capaces de encontrar patrones, generalizar y aprender sin ser explícitamente programadas. En los últimos años, las técnicas relacionadas con Machine Learning han hecho posible analizar de una forma más eficaz las concentraciones masivas de datos, lo que a su vez ha causado que se hayan hecho grandes esfuerzos para intentar predecir el mercado de valores, sin embargo, aún hay muchos mercados, técnicas de aprendizaje automático y combinaciones de parámetros por probar. En el caso de este proyecto, y como se verá más adelante, el estudio no se centrará en un mercado de valores específico sino en la utilización de diferentes técnicas y parámetros para el análisis de datos y la búsqueda de patrones de comportamiento.

Quedando esto presente, la principal motivación de este proyecto es la investigación.

En este proyecto se pueden diferenciar tres objetivos principales:

- Desarrollo de un modelo de clustering que identifique diferentes patrones de cotizaciones
- Diseño de un entorno de simulación para evaluar la calidad de los clusters.
- Análisis de los *clusters* identificados para la detección de posibles nichos de rentabilidad en función de la tendencia y la volatilidad esperadas

1.2 METODOLOGÍA DE TRABAJO

En todo momento se trabajará con precios de cierre de acciones con alto nivel de liquidez, quedando la decisión de compraventa determinada antes de la apertura de la sesión de mercado en la que se debe ejecutar. Además, como el horizonte de inversión de este proyecto

es de meses (como mínimo de semanas), hay elementos como la profundidad del mercado que no tendrán importancia en el análisis.

Además, se deberán completar las siguientes tareas, según el cronograma inicial:

- Estudio del análisis técnico de los mercados financieros y del funcionamiento de la Bolsa
- Estudio de posibles técnicas de análisis de datos y aprendizaje automático
- Adaptación de los datos para el análisis de clusters (parametrización y reducción de dimensionalidad)
- Diseño de las estrategias de clustering
- Análisis y comparativa de resultados
- Mejoras finales y optimización de metaparámetros
- Redacción de la memoria

1.3 Recursos

El recurso principal que se utilizará para la realización del proyecto es la herramienta de software matemático MATLAB. Además, para obtener los valores históricos de los activos a estudiar se utilizarán datos del sitio web Yahoo Finance.

1.4 Estructura

Este trabajo de fin de máster está dividido en cinco capítulos: introducción, antecedentes, metodología, experimentos y conclusión. El capítulo de antecedentes explica la teoría utilizada en los experimentos y muestra una visión general de los trabajos previos realizados en este campo. En el capítulo de metodología se explican cómo son los datos utilizados y la forma de implementación de los algoritmos. En el capítulo de resultados se presentan los mismos y se explica cómo interpretarlos y qué significan. El capítulo de conclusiones, como su propio nombre indica, contendrá conclusiones sacadas de los resultados obtenidos.

2 ANTECEDENTES

Este capítulo describe la teoría de finanzas y aprendizaje automático que compone la base de los experimentos. La sección 2.1 explica el mercado de valores, qué le afecta y qué retos hay que superar para predecir su comportamiento. El apartado 2.2 describe el aprendizaje automático, su potencial y sus desventajas. La sección 2.3 muestra una vista general de los modelos de aprendizaje automático utilizados, y el apartado 2.4 muestra un breve resumen del estado del arte relacionado con la predicción del comportamiento de acciones con aprendizaje automático.

2.1 EL MERCADO DE VALORES

Según la RAE (Real Academia Española de la lengua), una acción se define como: "Título valor que representa una parte alícuota en el capital de una sociedad mercantil y que da derecho a una parte proporcional en el reparto de beneficios y a la cuota patrimonial correspondiente en la disolución de la sociedad" [1]; o dicho de forma más simple, una acción es un título emitido por una empresa que representa una parte de la empresa. Un mercado bursátil es una agrupación de compradores y vendedores de acciones y otros productos financieros en un lugar determinado donde se producen las transacciones.

2.1.1 La Bolsa

Una Bolsa es un mercado de valores en el que corredores y agentes de bolsa compran, venden o intercambian productos financieros que están listados públicamente. Tradicionalmente, las bolsas eran sitios físicos, normalmente conocidos como el parqué, donde los corredores y agentes de bolsa cambiaban acciones por otras acciones o por dinero. En la actualidad, casi todas las bolsas son instituciones que facilitan operar de manera electrónica y prácticamente instantánea según el principio de subasta. Al cumplirse este principio y tratarse de transacciones prácticamente instantáneas que se pueden hacer desde cualquier lugar del mundo, las bolsas en la actualidad están regidas únicamente por la ley de oferta y la demanda.

Hay cientos de bolsas en todo el mundo. Algunas bolsas, como la bolsa española, incluyen una lista de acciones de empresas de un determinado país o región (IBEX 35), mientras que otras como el NASDAQ o la NYSE están más especializadas en cierto tipo de empresas e industrias. Aunque la mayoría de las transacciones se realizan de forma electrónica, algunas bolsas como la de Nueva York aún permiten la realización de transacciones en el parqué.

2.1.2 Compra, venta y beneficio

Cuando alguien compra cualquier número de acciones de una empresa, se dice que ha entrado en el mercado. Si la acción adquirida cambia de valor, el valor real de la inversión cambia con

él: si el precio de la acción aumenta y el inversor vende, habrá ganado dinero, y lo contrario ocurrirá si el inversor vende a un precio menor que el de compra. La ganancia o pérdida de dinero se consolida una vez la acción es vendida. Cuando el inversor cree que el valor de la acción va a aumentar en el tiempo, invertirá con una posición en largo, mientras que se posicionará en corto si cree que el valor de la acción va a decrecer. Hay otras formas de obtención de beneficios en la bolsa, pero las posiciones a largo principalmente y a corto después son las más importantes para este trabajo de fin de máster.

El mercado de valores no es un sistema de suma neta cero, sino que a lo largo del tiempo ha tenido una tendencia creciente en su valor, con un aumento superior a la inflación, con nuevo dinero entrando en el mercado diariamente. Lo que esto significa teóricamente es que, si un inversor comprase un gran número de acciones repartidas de forma aleatoria entre las distintas empresas que coticen en bolsa y las mantuviese en el tiempo, con el pasar de los años estaría generando beneficios.

Un índice es un número que representa el comportamiento de un grupo de acciones. El valor del índice se construye con una fórmula matemática que típicamente consiste en la media ponderada del valor de las acciones con su peso en el valor total del mercado. Los índices son una buena forma de medir el comportamiento de determinados mercados de valores, ya que representan el comportamiento general del mercado. Al ser los índices una representación del mercado, el objetivo del inversor que actúa en un mercado determinado es vencer al índice, o lo que es lo mismo, conseguir una rentabilidad mayor que la de la media ponderada del mercado.

2.1.3 Factores que afectan los cambios de precio de las acciones

A lo largo del tiempo se han desarrollado un gran número de estudios intentando explicar el comportamiento de la bolsa a través de las principales fuerzas o factores que motivan los cambios de precios, y ninguno ha sido capaz de determinar la evolución del mercado con este enfoque. Algunos de los factores estudiados y que explican parcialmente el comportamiento de los precios pueden ser la macroeconomía, los efectos psicológicos, la política, las noticias o el estado financiero de la empresa en ese momento.

David M. Cutler intenta determinar en un artículo llamado “What moves stock prices” qué factores afectan a los cambios en el mercado y en qué medida. El artículo está centrado en cambios a corto plazo, y lo primero que examina son los efectos de las noticias. La conclusión principal del documento es que el conjunto de noticias de macroeconomía es capaz de explicar más de la tercera parte de la variabilidad del comportamiento de los precios. En el mismo

documento, Cutler estudió los efectos de los acontecimientos políticos y otro tipo de noticias, concluyendo que también suponen parte de las fuerzas que modifican los precios, aunque menos importante que las noticias macroeconómicas. [2]

Otros factores que moldean los precios del mercado son las tasas de cambio entre monedas, (sobre todo para compañías internacionales) y las fluctuaciones en los precios de materias primas como el petróleo o el aluminio. Estos son unos pocos ejemplos de factores que pueden afectar de una manera considerable el valor de una compañía, y han sido mostrados con el objetivo de hacer entender al lector que responder a la pregunta qué fuerzas moldean el mercado de valores no tiene una respuesta fácil y mucho menos simple.

Los métodos utilizados en este trabajo de fin de máster, tal y como se explicará más adelante en el documento no intentan entender las razones que hay detrás de los cambios debido a su dificultad y a que no haya sido demostrado que el efecto en el precio pueda ser explicado por una serie de factores. Es por esto por lo que se ha elegido el clustering como técnica de análisis principal, ya que permite conocer con cierta probabilidad el comportamiento que van a tener los precios sin necesidad de comprender las causas que los hacen cambiar.

2.1.4 Predictibilidad

Como se ha comentado en los apartados anteriores de este capítulo, algunos investigadores consideran que el precio de las acciones no se puede predecir según el comportamiento histórico. Una de las teorías más generalizadas es la del paseo aleatorio ("Random Walk Theory"), que defiende que los cambios en los precios de las acciones tienen la misma distribución y son independientes unos de otros, por lo que el movimiento o la tendencia pasadas del precio de una acción no pueden ser utilizados para predecir los movimientos del futuro. Simplificando, esta teoría defiende que los precios de las acciones toman caminos aleatorios e independientes. [3]

Sin embargo, la mayoría de los expertos coinciden en que el comportamiento de los precios no es aleatorio, tal y como defiende el famoso artículo "Stock prices do not follow random walks". En este documento se descarta la teoría de los paseos aleatorios para el período estudiado (1962-1985), y para todos los subperiodos para un conjunto de índices y carteras de inversión. [4] Los autores por lo tanto defienden que hay una evidencia empírica suficiente para afirmar que el precio de las acciones es predecible hasta cierto tiempo.

A partir de esta conclusión, se ha hecho la suposición de que el comportamiento histórico de las acciones contiene información suficiente para conocer cómo se va a comportar el

mercado, al menos parcialmente. Además, como defiende Eugene F. Fama en su artículo "The Behavior of Stock-Market Prices", el comportamiento de los precios en el tiempo se repite históricamente formando patrones. Esta idea es defendida por la mayoría de chartistas y defensores del análisis técnico. [5]

Entonces, ¿es predecible el comportamiento del mercado de valores? Los defensores de la teoría del camino aleatorio y los investigadores que afirman que el mercado es predecible al menos parcialmente afirman haber aportado pruebas empíricas suficientes como para defender su posición. Es por esto por lo que la respuesta más acertada a la pregunta probablemente sea quizás. Sí es un hecho que un gran número de gente piensa que el comportamiento es predecible, y quizá ese hecho añada predictibilidad al sistema. Por ejemplo, la revista Business Week recomienda una acción por semana que suele aumentar de forma acelerada y anormal, probablemente como consecuencia de la publicación de la predicción de crecimiento. [6]

Aunque la totalidad del mercado de valores no sea predecible, se pueden encontrar y crear estrategias que den beneficios basadas en el comportamiento histórico. Un ejemplo sería el modelo anteriormente nombrado de Business Week, donde el hecho de publicar un futuro comportamiento provoca que el precio se comporte tal y como ha sido descrito. Otra estrategia podría ser la compraventa de acciones basada en los cambios de precio de una materia prima como el petróleo o la reacción más rápida que la mayoría de los inversores a una noticia macroeconómica.

2.1.5 Accesibilidad de los datos

Si se quiere utilizar cualquier técnica de análisis de datos, lo primero y esencial es tener datos, y preferiblemente en gran cantidad por ser la robustez de las conclusiones proporcional al número de datos utilizados. Afortunadamente, en internet hay grandes cantidades de datos relacionados con los precios de las acciones. Estos datos aparecen en todo tipo de estructuras, tamaños y detalle. Para este trabajo se han utilizado datos obtenidos del sitio web Yahoo Finance. Los más importantes en este estudio han sido precios de cierre, volúmenes de mercado y tiempo, y a partir de éstos se han calculado otros datos complementarios e índices que serán explicados en mayor detalle más adelante en el documento.

2.1.6 El ruido en finanzas

El ruido en los datos de finanzas recogidos puede deberse a diferentes factores. Puede haber por ejemplo incertidumbre sobre acontecimientos futuros, tecnología, tendencias o reacciones que no se correspondan con la respuesta esperada. Sea cual sea la causa del ruido

en los datos, presenta imperfecciones que hacen que el mercado no sea completamente eficiente, y es difícil de modelar. El ruido siempre está y estará presente en los datos financieros, y es una de las razones por la que no se han desarrollado teorías académicas ni modelos prácticos que expliquen completamente los diferentes comportamientos del mercado. [7]

2.1.7 Análisis técnico de acciones y tendencias

El análisis técnico de acciones y tendencias ha sido usado durante cientos de años. En Europa, José de la Vega (sefardita de origen español) utilizó las primeras técnicas de análisis para predecir el comportamiento del mercado en Holanda en el siglo XVII [8]. En Asia, Homma Munehisa creó un sistema que evolucionó más adelante a los diagramas de velas en el siglo XVIII [9], utilizados por un gran número de analistas técnicos en la actualidad.

La principal idea detrás de este tipo de análisis es que el precio del mercado es un reflejo de toda la información que podría afectar al mercado. De este modo, no hay necesidad de hacer otros tipos de análisis como el fundamental o el económico de cada compañía ni desarrollar otro tipo de análisis porque no van a añadir información que no esté ya contenida en el precio de mercado. Los analistas técnicos piensan que los precios se mueven en tendencias, y que las tendencias tienden a repetirse en la historia debido a la psicología del mercado. [10]

Otra asunción importante es la de que la historia se repite. Se pueden encontrar en el histórico de datos patrones que tienen una alta probabilidad de repetirse por haber respondido el mercado de una determinada forma a las mismas entradas. Un ejemplo de esto podría ser el paso de los inversores de un estado de optimismo a uno de miedo, donde un determinado comportamiento se repetirá, haciendo que los inversores comiencen a vender, provocando así una caída de los precios de los valores del mercado.

Hay varios tipos de análisis técnico, siendo los dos más famosos el análisis de gráficos y el de indicadores técnicos.

2.1.7.1 Análisis de gráficos

El estudio de patrones gráficos es una forma subjetiva de análisis que depende del analista. De esta manera, el analista buscará puntos de apoyo y resistencia para determinar si el precio de la acción se encuentra en un máximo o en un mínimo y de esa manera decidir si se debe comprar o vender la acción.

Principalmente se utilizan cuatro tipos diferentes de gráficos, dependiendo de la información que se tenga y de lo que se quiera conseguir: gráficos de líneas, gráficos de barras, gráficos de velas y gráficos de punto y figura.

- Gráficos de líneas: son el tipo de gráfico más básico y representa los precios de cierre unidos por una línea durante un período de tiempo. Aunque este tipo de gráfico no aporta mucha información sobre los movimientos del precio intradía, muchos inversores consideran que es el precio más importante del día, por encima del de apertura, del máximo o del mínimo. Además, estos gráficos hacen más fácil la detección de tendencias por haber menos "ruido" comparado con otros gráficos.

  Estos gráficos normalmente vienen acompañados de un histograma en la parte inferior que contiene los datos correspondientes al volumen.

Ilustración 1. Ejemplo de gráfico de línea [11]

- Gráficos de barras: incluyen más información que los gráficos de línea. Estos gráficos están formados por líneas verticales que representan el rengo de valores en los que se ha encontrado el precio durante un periodo de tiempo, con dos pequeñas marcas horizontales que muestran el precio de apertura (a la izquierda) y el precio de cierre (a la derecha). Si el precio de apertura es más bajo que el de cierre, es decir, si después del día la acción ha aumentado su precio, la barra será representada de color negro, mientras que, si ha ocurrido lo contrario, la figura será roja. Igual que en el gráfico de barras, en el inferior de la figura se representa el volumen de cada periodo mediante un histograma.

Ilustración 2. Ejemplo de gráfico de barras [11]

- Gráficos de velas: tienen su origen en Japón hace unos trescientos años y en la actualidad son muy utilizados por los inversores y analistas. Igual que los gráficos de barras, los gráficos de velas tienen una línea vertical que muestra el rango de precios para un determinado periodo, y su color (rojo o negro) depende de si en el periodo el valor ha aumentado o disminuido. Lo que hace a este gráfico diferente del de barras es la existencia de un rectángulo o una barra más ancha que representa la diferencia entre los precios de apertura y cierre.

  Los periodos de pérdida típicamente tendrán un cuerpo de vela rojo o negro, mientras que los periodos de crecimiento tendrán un color blanco o vacío en el cuerpo de la vela. Si en un periodo no hay diferencia entre los precios de apertura y cierre, entonces esa representación no tendrá rectángulo alguno. Además, igual que los anteriores tipos de gráfico, el volumen de cada periodo está representado en el inferior de la figura mediante un histograma.

Ilustración 3. Ejemplo de gráfico de velas [11]

- Gráficos de punto y figura: este tipo de gráfico no es muy conocido ni utilizado por la mayoría de los inversores, y su origen se remonta a los primeros analistas técnicos. El gráfico refleja los movimientos del precio sin tener en cuenta el tiempo o el volumen, algo que permite eliminar ruido que pueda distorsionar las tendencias presentes en la evolución del precio. Además, este tipo de gráficos ayuda a eliminar el sesgo que la representación en el tiempo produce en el diagrama, facilitando el análisis.

  Los diagramas de punto y figura están formados por una serie de Xs y Os. Las Xs representan una tendencia alza en el precio, mientras que las Os se utilizan para las tendencias bajistas. Además, se pueden encontrar números y letras, que representan los meses para dar una idea aproximada al analista de gráficos. Cada rectángulo del gráfico representa una escala de precios, que se ajusta según el precio de la acción: cuanto más alto es el precio de la acción, más valor representa cada rectángulo. Lo más común es encontrar gráficos donde cada rectángulo represente una unidad de moneda, es decir, un dólar o un euro por ejemplo.

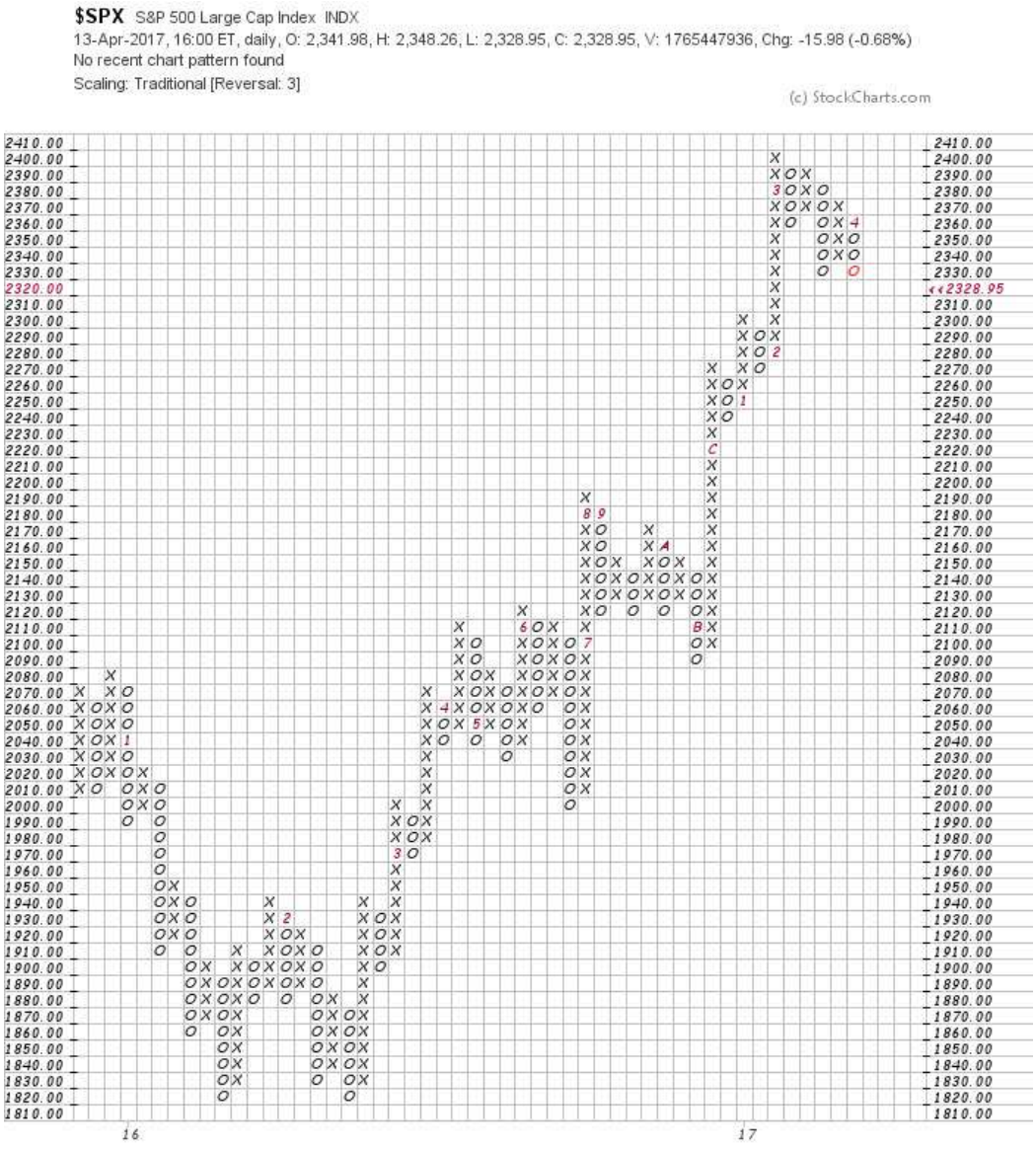


Ilustración 4. Ejemplo de gráfico de punto y figura [11]

Otro punto a tener en cuenta sobre este tipo de gráfico es el criterio de cambio de tendencia, que el analista debe fijar y que tradicionalmente es de tres. El criterio de cambio de tendencia representa cuánto tiene que cambiar el precio de la acción para considerar que la tendencia ha pasado a la contraria, es decir, cuánto tiene que subir un precio para que una tendencia bajista se pueda considerar alcista y viceversa. Dicho de otra forma, el analista debe decidir cuándo una columna de Xs se puede convertir en una columna de Os y al revés. Cuando aparece un cambio de tendencia, el nuevo comportamiento se representa una columna hacia la derecha, de tal manera que a una columna de Xs siempre le seguirá una de Os.

Cada analista tiene normalmente un tipo de gráfico que utiliza, y sobre él trata de encontrar patrones con el objetivo de predecir futuras tendencias. Existen multitud de patrones estudiados en el análisis de gráficos, siendo los más importantes: hombro-cabeza-hombro, taza y asa, doble máximo o mínimo, los triángulos y banderas y banderines.

- Hombro-cabeza-hombro: este patrón indica un cambio de tendencia una vez se ha completado. Esta figura consiste en tres picos, siendo el del medio el más alto (cabeza) y los otros dos más bajos e iguales (hombros). Los picos exteriores están unidos por una línea de tendencia (la línea de cuello), que ayuda a entender la futura tendencia tras el patrón, y que marca el paso de crecimiento a decrecimiento. También existe la figura simétrica, llamada hombro-cabeza-hombro invertida y que consiste en tres mínimos de los cuales el del medio es el que menor valor tiene. Esta segunda figura es señal de un cambio de tendencia de bajista a alcista.

Ilustración 5. Ejemplo del patrón hombro-cabeza-hombro [12]

- Taza y asa (Cup and Handle): es una continuación de optimismo en el mercado en la que una tendencia alcista se frena, pero continúa una vez completado el patrón. La "taza" debería tener forma de U, recordando a un bol, en lugar de una V con iguales alturas a ambos lados de la taza. El "asa" se forma en la parte derecha de la taza y recuerda a un gráfico de bandera o banderín. Una vez el asa está terminada, el precio alcanzará nuevos máximos, siguiendo la tendencia anterior.

Ilustración 6. Ejemplo de patrón Taza y asa [12]

- El doble máximo y el doble mínimo son fáciles de reconocer y unos de los patrones más fiables que hay, siendo el preferido por muchos analistas de gráficos. Cuando en una tendencia creciente hay dos máximos consecutivos con el mismo valor aproximado, eso implica que el nivel de resistencia no se va a romper en esa tendencia y por lo tanto habrá un cambio de tendencia a decreciente. En un patrón de doble mínimo pasará lo contrario, en una tendencia decreciente se repetirán dos mínimos, significando el final de esa tendencia y el comienzo del crecimiento.

Ilustración 7. Ejemplo de patrón Doble máximo [12]

- Los triángulos son unos de los tipos de patrones más populares debido a la frecuencia con la que se encuentran comparado con otras formas. Los tres tipos de triángulo más comunes son los simétricos, los ascendentes y los descendentes, y pueden durar desde un par de semanas a varios meses.

Los triángulos simétricos ocurren cuando dos líneas de tendencia convergen la una hacia la otra, los triángulos ascendentes se dan cuando la tendencia superior es horizontal y la inferior ascendente, y los descendentes cuando la tendencia inferior es horizontal y la superior descendente. La magnitud de los cambios que aparecen después de cada triángulo es normalmente proporcional al lado vertical del triángulo, y puede ocurrir en ambas direcciones con independencia del tipo de triángulo.

Ilustración 8. Ejemplo de patrón triangular simétrico [12]

- Los patrones de banderas y banderines (flags and pennants) son la continuación a corto plazo de la consolidación que sigue a un movimiento brusco de precio en la tendencia existente. Se caracterizan por tener un pequeño patrón rectangular con pendiente contraria a la tendencia, mientras que los banderines son pequeños triángulos pequeños

Ilustración 9. Ejemplo de patrón de banderín

El estudio de gráficos es la parte más importante del análisis técnico. Se utilizan gráficos porque resulta más fácil para las personas localizar tendencias si están representadas de forma visual. Durante la realización de este proyecto, y como se verá más adelante, aunque las ideas que hay detrás del análisis son similares a las del análisis de gráficos, no se analizarán gráficos como tal debido a que las máquinas son capaces de localizar patrones y tendencias a partir de los datos "en crudo" o de parámetros sacados a partir de los datos.

2.1.7.2 Indicadores técnicos

Esta forma de análisis consiste en aplicar diferentes fórmulas matemáticas, normalmente relacionadas con la estadística, para unir la información que proporcionan los precios y los volúmenes. La principal ventaja de este método es que todos los indicadores se pueden calcular de forma cuantitativa, disminuyendo la subjetividad del estudio. [13]

Cuatro de los indicadores más importantes del análisis técnico son las medias móviles, las medias móviles convergentes y divergentes, el índice de fuerza relativa (RSI del inglés Relative Strength Index) y el estocástico bajo (slow stochastic).

En este proyecto se han utilizado diferentes indicadores técnicos debido a la posibilidad de realizar cálculos cuantitativos sobre ellos, cualidad muy interesante a la hora de aplicar técnicas de análisis de datos.

2.1.7.2.1 Medias móviles

Son el indicador más conocido y utilizado del análisis técnico. Se utilizan para suavizar la evolución del mercado y así hacer más fácil al analista la detección de tendencias. De esta manera, una media móvil es el resultado de calcular la media de una serie de valores históricos. Por ejemplo, una media móvil de los precios de cierre de diez días, será la suma de los precios de cierre de los últimos diez días entre diez. De esta manera, la media móvil del día siguiente será la de los últimos diez días, que incluirán nueve de los que se utilizaron para calcular la del día anterior (los nueve más recientes) y uno nuevo que es el del propio día. [14] Este cálculo queda recogido en la siguiente fórmula:

$$MM(n) = \frac{1}{n}\sum_{i=1}^{n} p_{t-i} \tag{2.1}$$

Donde MM es la media móvil, n el número de precios históricos utilizados para calcular la media móvil y t el instante de tiempo en el que se quiere calcular la media móvil.

2.1.7.2.2 Índice de fuerza relativa o RSI (Relative Strength Index)

El índice de fuerza relativa es un indicador de momento que mide la magnitud de los cambios de precio más recientes para la detección de casos de sobrecompra y sobreventa en el precio de una acción.

$$RSI = 100 - \frac{100}{1 + RS} \tag{2.2}$$

Donde RS es el ratio de promedio de subidas entre promedio de bajadas en un determinado periodo de tiempo, que normalmente es de catorce días de mercado.

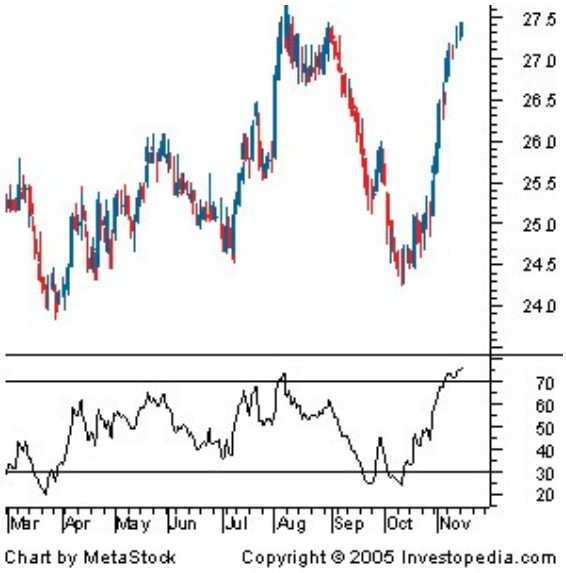


Ilustración 10. Ejemplo gráfico del indicador de Índice de Fuerza Relativa

Este indicador aporta una visión relativa de la fuerza de los comportamientos recientes, dando al analista señales de que es buen momento para comprar si la acción está sobrevendida y de vender si está sobrecomprada.

2.1.7.2.3 Bandas de Bollinger

Las bandas de Bollinger llevan el nombre de su creador, John Bollinger, y muestran una distancia de dos desviaciones típicas desde una media móvil. Son muy populares en el análisis técnico, ya que un gran número de analistas cree que cuanto más cerca de la banda superior se encuentra un precio, más sobrecomprado está el mercado, y cuanto más cerca de la banda inferior, más sobrevendido. John Bollinger hizo una serie de reglas que seguir para el correcto entendimiento de las bandas de Bollinger, pero en este caso no aplican por tratarse de un indicador financiero introducido en un sistema de aprendizaje no supervisado. [15]

Dicho esto, las bandas de Bollinger se calculan de la siguiente manera:

$$BB = MM \mp 2 \times \sigma \tag{2.3}$$

Siendo BB las bandas de Bollinger, MM la media móvil de un periodo y σ la desviación típica del mismo periodo.

Aproximadamente el 90% del precio de la acción se mantiene dentro de las bandas, y cualquier ruptura por encima o por debajo implica un evento importante. La ruptura no es una señal de compra o de venta, sino que no da pistas sobre la dirección y futuro comportamiento del precio tal y como muchos analistas creen de manera errónea.

La información que los analistas suelen sacar de las bandas de Bollinger es el nivel de volatilidad que la acción va a tener en el futuro. De esta forma, si las bandas de Bollinger se encuentran cerca de la media móvil, es decir, si la desviación típica de un periodo es baja, es más probable que sea seguida por un periodo de desviación típica alta y viceversa.

2.2 APRENDIZAJE AUTOMÁTICO

Las empresas generan cada vez más datos, y extraer información valiosa de ellos puede significar una importante ventaja competitiva. El aprendizaje automático, o Machine Learning en inglés, es una rama de la inteligencia artificial que permite a las máquinas aprender cosas que no les han sido programadas explícitamente. Se puede decir que esta materia comenzó a finales de los años 1950, cuando el informático Arthur Samuel creó un programa para jugar a las damas que contenía un sencillo algoritmo para optimizar los movimientos con el objetivo de ganar. El aprendizaje automático es una técnica aplicable a prácticamente cualquier campo de la actualidad [16].

Hablar de aprendizaje en este contexto es sinónimo de identificación de patrones concretos escondidos en grandes acumulaciones de datos. Además, es posible que el algoritmo que revisa los datos con el objetivo de predecir el futuro mejore con el tiempo sin intervención humana. A este fenómeno es al que hace referencia el término automático. Dentro del aprendizaje automático, las técnicas existentes se dividen en aprendizaje supervisado y aprendizaje no supervisado, dependiendo de los procesos que sigan.

En otras palabras, el aprendizaje automático consiste en la creación de máquinas que se adaptan y cambian sus acciones o predicciones de tal manera que disminuya su error, o lo que es lo mismo, incrementando su precisión. [17]

El aprendizaje automático es uno de los mayores campos de investigación de la ciencia computacional actualmente. El objetivo de los investigadores es la construcción de algoritmos que puedan aprender para más tarde predecir comportamientos. Esto algoritmos funcionan creando modelos a partir de datos de ejemplo o entrenamiento. Mientras que la ciencia computacional clásica consiste en escribir código que resuelva problemas específicos, utilizando unas reglas de programación fijas, los algoritmos de aprendizaje automático no están hechos a la medida de un único problema, sino que se basa en los datos con los que se le entrene para crear un modelo que tiene unas reglas "autoenseñadas".

El aprendizaje automático es un campo muy complejo, es por esto que después de décadas de investigación, aún no hay ninguna máquina que tenga capacidades de aprendizaje similares a las humanas. Esto no quiere decir que no se haya avanzado: se han hecho grandes progresos para el procesamiento de grandes cantidades de datos y detección de patrones.

2.2.1 Aprendizaje supervisado

El objetivo de este tipo de aprendizaje es el de predecir los valores de las variables de salida de un sistema a partir de las variables de entrada. La forma de funcionamiento del aprendizaje supervisado consiste en entrenar al algoritmo previamente con una serie de datos de los que se conocen tanto las entradas como las salidas. Mediante este entrenamiento se consigue ajustar los parámetros del algoritmo para que proporcione una estimación de la salida buscada, con el mínimo error posible, a partir de las variables de entrada. El proceso de aprendizaje de este tipo de algoritmos se puede separar en tres fases diferenciadas: entrenamiento, validación y test.

Para realizar el entrenamiento del algoritmo, se necesitará un conjunto de datos llamado conjunto de entrenamiento o *training set*. El objetivo de esta fase será minimizar el error de estimación de las salidas en función de las entradas. Así, ajustará los parámetros del modelo dependiendo de los valores del *training set*. Si se utiliza un modelo de estimación muy complejo, en el que se intente explicar con mucha precisión el conjunto de entrenamiento, se puede caer en el sobreaprendizaje u *overfitting*. Este problema consiste en adaptar los parámetros del algoritmo a particularidades del *training set*, haciendo que el error real de la predicción aumente. Para evitar este problema se utilizarán modelos sin un elevado grado de complejidad, es decir, sencillos. [18]

La validación consiste en el ajuste de los metaparámetros mediante la comparación de diferentes modelos de predicción. Los metaparámetros son las propiedades de un mayor nivel a las características que se están estudiando. Algunos ejemplos de metaparámetro serían la

complejidad y la capacidad de aprendizaje del algoritmo o el número de clusters que se van a utilizar en un determinado método. Los metaparámetros no pueden ser aprendidos por el algoritmo directamente de los datos, por lo que deben ser predefinidos. Para elegir qué metaparámetros elegir, habrá que entrenar diferentes algoritmos y comparar sus resultados para escoger el que menor error produzca. [19] Es importante destacar que es en este punto del proceso donde se puede identificar el sobreaprendizaje.

Por último, el *test set* se utiliza para saber si la herramienta está bien calibrada. Este buen funcionamiento se mide a través de la eficacia que ofrezca el algoritmo y la medición de su error de generalización. Además, es importante que para evitar errores metodológicos de cada proceso no esté "contaminadas" por los otros procesos, siendo así independientes. En caso contrario, las conclusiones pueden diferir considerablemente de la realidad debido a una falta de robustez en el sistema.

Para la decisión de cuántos datos deberían ir a cada prueba no hay un único criterio. El más común consiste en dedicar aproximadamente la mitad de los datos al entrenamiento y un cuarto para validación y test respectivamente. Otra estrategia que es interesante por los mejores resultados que otorga aunque se necesite una mayor capacidad computacional es la llamada *cross-validation* o validación cruzada. Consiste en separar los conjuntos que pertenecen a validación y entrenamiento en subconjuntos mutuamente excluyentes, y repetir los procesos de entrenamiento y validación tantas veces como subconjuntos haya. De esta manera, cada vez se utilizará un subconjunto diferente para validación, siendo los demás dedicado al entrenamiento, y obteniendo un resultado final más robusto. [20]

Después de haber definido los parámetros del algoritmo y comprobado el correcto funcionamiento, éste podrá ser utilizado con nuevos conjuntos de datos formados solamente por entradas, estimando las salidas y los errores cometidos en el proceso. Además, existe la posibilidad de continuar el aprendizaje del algoritmo insertando el valor real de la salida cuando se tenga, de tal manera que lo compare con la estimación que había producido y así, dato a dato, minimizar los errores cometidos.

2.2.2 Aprendizaje no supervisado

El aprendizaje no supervisado es un tipo de aprendizaje automático cuya principal característica es que no existe un conjunto de datos que contenga los valores de la variable de salida para la fase de entrenamiento. Considerando los datos de entrada como variables aleatorias, en el aprendizaje no supervisado se buscan correlaciones y patrones para tratar de agrupar los datos con el objetivo de predecir su comportamiento futuro.

La base del aprendizaje de estos modelos son las probabilidades condicionadas y las diferencias cuantificadas (distancias) entre distintos elementos, y tienen diferentes aplicaciones, como estimar funciones de densidad de probabilidad o identificar subconjuntos (*clustering*). Esta segunda funcionalidad, el *clustering*, es la más común del aprendizaje no supervisado, y consiste en clasificar los datos de entrada en grupos de individuos similares.

2.3 CLUSTERING

2.3.1 Introducción al clustering

El clustering es un tipo de aprendizaje no supervisado en el que no existe un conocimiento a priori de como debe ser la salida. El término de análisis de clusters (utilizado por primera vez por Tryon en 1939), engloba un grupo de diferentes algoritmos y métodos para la agrupación de objetos similares en categorías. Un problema al que tienen que hacer frente continuamente los investigadores de diversas áreas es cómo organizar los datos observados en estructuras significativas.

En otras palabras, el análisis de clusters es una herramienta de análisis de datos que intenta clasificar diferentes objetos en grupos de tal manera que el grado de asociación entre dos objetos sea máximo si pertenecen al mismo grupo y mínimo si pertenecen a grupos diferentes. Es por esto por lo que el análisis de clusters puede ser utilizado para descubrir estructuras en los datos sin ofrecer una explicación o interpretación; es decir, el análisis de clusters sirve para encontrar estructuras sin explicar por qué existen.

Hay multitud de ejemplos de clustering en el día a día. Por ejemplo, un grupo de comensales compartiendo la misma mesa de un restaurante puede ser entendido como un cluster de gente, o en las tiendas de comida, los productos similares que se colocan en posiciones cercanas pueden ser también entendidos como clusters.

En este proyecto se han utilizado dos tipos diferentes de clustering, tal y como se explica más adelante en el documento: k-means y GMM.

2.3.2 Algoritmo k-means

El clustering k-means es un método de aprendizaje de cuantificación vectorial que surgió en primer lugar en el procesamiento de señales. El objetivo de este algoritmo es dividir n observaciones en k clusters, perteneciendo cada observación al cluster cuya media se encuentre a menor distancia. Esta clasificación resulta en la partición del espacio muestral en celdas o regiones de Voronoi.

Un diagrama de Voronoi es la división del plano en tantas regiones como puntos se tengan, de tal manera que cada punto tenga asignada la región que está más cerca de él que de cualquier otro punto. [21] En este caso, en lugar de haber tantas regiones como puntos, habrá tantas regiones como clusters (número que se fija antes del análisis), y el equivalente a cada punto del diagrama de Voronoi será la media de los puntos pertenecientes al cluster. Además, si el diagrama de Voronoi es una división del plano, el clustering utilizando el algoritmo k-means puede dividir espacios vectoriales con más dimensiones.

El término k-means fue utilizado por primera vez por James MacQueen en 1967 [22], aunque la idea se remonta al 1957 y fue desarrollada por Hugo Steinhaus [23]. El algoritmo estandarizado actualmente fue propuesto por Stuart Lloyd en 1957 como técnica de codificación modular de pulsos, aunque un fue publicado hasta 1982.

El algoritmo de k-means más común es una técnica iterativa. Dado un grupo de datos con k medias, el algoritmo está compuesto de dos pasos:

1. Asignación: se asigna cada observación al cluster con la media a la menor distancia euclídea, es decir, la media más cercana. Si cada una de las medias son m_1, m_2,…, m_k, este razonamiento queda resumido como:

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\| \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\} \tag{2.4}$$

   Donde cada x_p queda asignada exactamente a un $S^{(t)}$, incluso si pudiese ser asignado a más de uno por tener la misma distancia euclídea.

2. Actualización: Se calculan las nuevas medias para que sean los centroides de cada uno de los nuevos clusters:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \epsilon S_i^{(t)}} x_j \tag{2.5}$$

   Después de este segundo paso, se vuelve al primero para comprobar que cada punto sigue estando comprendido en el cluster con el centroide más cercano.

Se dice que el algoritmo ha convergido cuando las asignaciones de datos a clusters dejan de cambiar, y no hay garantía de que se haya encontrado la clasificación óptima, sino que puede tratarse de un óptimo local del sistema.

Por lo tanto, el punto débil más importante de este método es que depende en gran parte de las condiciones iniciales que se asignen al algoritmo.

2.3.3 Modelos de mezcla gaussiana o GMM (Gaussian Mixture Model)

En estadística, un modelo mixto es un modelo que se utiliza para representar subconjuntos de población dentro de una población general sin la necesidad de que se haya hecho una clasificación previa según el origen de los datos.

El modelo de mezcla gaussiana es más sofisticado que el algoritmo k-means, pero sin causar una gran diferencia computacional. Es más robusto en el sentido de que no depende tanto de las condiciones iniciales de la primera iteración.

Los modelos de mezcla gaussiana parten de la premisa de que todo grupo poblacional puede clasificarse mediante una serie de distribuciones gaussianas. Cada distribución quedará definida por su media y su matriz de covarianza, y cada dato del conjunto pertenecerá a un grupo diferente (a una gaussiana diferente) con una cierta probabilidad. De esta manera, cada dato será integrado en el cluster representado por una gaussiana en el que más probabilidades tenga de estar.

Al igual que el método k-means, es necesario fijar inicialmente el número de clusters en los que se van a dividir los datos. También se pueden añadir otro tipo de restricciones a las gaussianas, como fijar valores o parámetros en las matrices de covarianza.

Los diferentes parámetros que definen a las gaussianas se obtienen a partir del algoritmo de Esperanza-Maximización (EM). Este método iterativo se utiliza para encontrar la máxima verosimilitud o máximo a posteriori, estimando los parámetros de modelos estadísticos donde el modelo depende de variables latentes no observadas. Para que este algoritmo empiece a funcionar debe ser inicializado, inicialización que se puede hacer de forma aleatoria o escogida por alguna razón. Esta inicialización sobre todo afectará en el tiempo que tarde el algoritmo en converger y no tanto en su resultado final por tratarse de un algoritmo relativamente robusto.

Después de la inicialización, comienza el proceso iterativo. El primer paso es el cálculo de la probabilidad de cada dato de pertenecer a cada una de las gaussianas generadas por los parámetros iniciales. En este algoritmo la asignación que no es absoluta (el dato x_1 pertenece al cluster *A*), sino que es ponderada (el dato x_1 pertenece al cluster *A* con una probabilidad p_1 y al cluster *B* con una probabilidad p_2). De esta manera, la influencia que tiene cada dato en modificar los parámetros de cada cluster vendrá determinada por la probabilidad de

pertenencia a dicho cluster, siendo más importantes los elementos que mayor probabilidad de pertenencia tengan. Después del cálculo de las responsabilidades (las probabilidades de pertenencia), el siguiente paso es la maximización de las mismas. De esta manera, se calcularán los nuevos parámetros que maximicen la probabilidad de los datos de pertenecer a cada uno de los clusters según el criterio de máxima verosimilitud. Esta técnica se repetirá hasta la convergencia.

Como se expondrá más adelante en el documento, k-means y GMM son los dos tipos de clustering que se han utilizado en el proyecto.

2.4 PREDICCIÓN DEL MERCADO DE VALORES

Como se ha señalado en los apartados anteriores, el mercado de valores está influenciado por multitud de diferentes factores. De esta manera, encontrar los factores más explicativos de su comportamiento se convierte en una tarea dura y llena de desafíos. Al incluir demasiados factores en el modelo de predicción se pueden calcular y obtener correlaciones, pero no relaciones de causalidad. David Leinweber cubre este tema en su artículo *Stupid Data Miner Tricks, Overfitting the S&P 500,* donde muestra la perfecta correlación que existe entre la población de ovejas y la producción de mantequilla de Bangladesh con la evolución del S&P 500 durante diez años [24]. Además, sabiendo que los factores que cambian el comportamiento del mercado se encuentran en constante cambio, la figura del predictor del mercado parece algo imposible. El hecho de incluir muchos factores para la predicción hace que aparezcan correlaciones sin causalidad, mientras que incluir demasiado pocos factores hará que no se pueda predecir el comportamiento suficientemente bien. Con este horizonte de problemas, la predicción del mercado de valores es un auténtico reto.

2.4.1 Estado del arte

Existe un gran número de artículos y publicaciones sobre la predicción del mercado de valores utilizando Aprendizaje Automático. Investigadores japoneses consiguieron resultados considerablemente superiores a los de la competencia utilizando diferentes algoritmos de aprendizaje automático alimentados por datos semanales de la bolsa japonesa NIKKEI 225 [25]. El artículo presenta la variedad de algoritmos utilizados, algunos de los cuales están basados en redes neuronales y *decision forest*, pero descubrieron que uniendo SVM (Support Vector Machine) con otros métodos se conseguían los mejores resultados. El hecho de que esta combinación diese buenos resultados para la predicción da esperanza a este proyecto,

ya que se ha elegido abordar el problema en cuestión utilizando técnicas de clustering, más complejas y precisas que el SVM.

Una de las publicaciones con la investigación más exhaustiva detrás sobre la utilización de algoritmos de aprendizaje automático para la predicción del comportamiento del mercado de valores es el artículo *Surveying stock market forecasting techniques – Part II: Soft computing methods*. Escrito por George S Atsalakis y Kimon P VValavanis, muestra y compara las técnicas utilizadas por los autores de más de 100 publicaciones de investigación sobre este tema. La conclusión del documento es que las redes neuronales tienen el mejor rendimiento de entre los diferentes algoritmos de aprendizaje automático, al mismo tiempo que muestra la complejidad y los problemas que surgen en la selección del número correcto de capas y nodos de las redes. [26]

Por otro lado, los distintos métodos de agrupamiento relacionados con el aprendizaje automático también han sido investigados por un gran número de personas. De esta manera, se ha demostrado que las técnicas de agrupamiento (Ensemble Learning en inglés) han dado buenos resultados en la predicción de índices. [27]

Por lo tanto, en la investigación de la predicción del mercado de valores con aprendizaje automático no hay una sola teoría aceptada sobre qué tipo de técnica es más efectiva: algunos investigadores afirman que los SVM son el algoritmo óptimo para este tipo de predicción, mientras que otros aseguran que son las redes neuronales o que las técnicas de agrupamiento son con diferencia la mejor opción. Esta diversidad de opiniones es una clara muestra del gran número de diferentes algoritmos de aprendizaje automático existente y podría ser una prueba de que diferentes técnicas pueden ser más o menos efectivas dependiendo del mercado o del tipo de datos estudiados (precios, volúmenes, índices…).

Para la realización de este proyecto, y como se contará en detalle más adelante en este documento, se han decidido utilizar técnicas de agrupamiento o Ensemble Learning para estudiar su funcionamiento en la predicción del comportamiento del precio de acciones de diferentes mercados.

3 METODOLOGÍA

Este capítulo describe cómo se han obtenido los resultados. El apartado 3.1 muestra una breve descripción del entorno de prueba. En el capítulo 3.2 se explica qué datos se han utilizado en los experimentos, por qué, y cómo se han procesado antes de introducirlos en el algoritmo de clustering. La sección 3.3 describe la implementación de los algoritmos, la 3.4 señala cómo entender los resultados y para finalizar la 3.5 expone las limitaciones del proyecto.

3.1 ENTORNO DE PRUEBA

El objetivo de este proyecto es la detección de nichos de inversión con una alta rentabilidad y bajo nivel de variabilidad mediante técnicas de aprendizaje no supervisado, y concretamente de clustering. Para realizar este análisis se han construido ventanas de datos a clasificar en clusters y posteriormente se ha estudiado su evolución en el tiempo.

La metodología seguida para la creación de modelo y su evaluación es la estándar para este tipo de análisis. En primer lugar, se han dividido los datos obtenidos en tres ventanas de diferentes tamaños. La primera ventana se ha utilizado para el entrenamiento del algoritmo y la creación de clusters, y contiene aproximadamente el 50% de los datos. Con la segunda ventana, compuesta por un 30% de los datos aproximadamente, se ha realizado la fase de análisis, que consiste en la modificación de los metaparámetros para la obtención de resultados acordes a lo esperado, es decir, se han modificado el número de clusters y el tamaño de ventana hasta obtener clusters interesantes (con alta rentabilidad y baja variabilidad). Una vez logrados buenos resultados entrenando el algoritmo con ciertas combinaciones de parámetros y comprobándolo con los datos de análisis, se ha procedido a utilizar la última ventana que contiene el 20% restante de los datos, la de test, para hacer una segunda confirmar que los clusters que en las dos primeras etapas parecían buenos siguen siendo interesantes.

De esta manera, y como se explica más adelante en este documento, se han utilizado dos técnicas diferentes de clustering (k-means y GMM) y otras dos técnicas de reducción de dimensionalidad de los parámetros.

3.1.1 Notas de implementación

La implementación de algoritmos de aprendizaje automático puede consumir cantidades considerables de tiempo, mostrar una gran variedad de errores y es difícil de optimizar. De esta manera, se ha estimado oportuno utilizar el entorno de programación de Matlab, y concretamente el conjunto de funciones de machine learning y redes neuronales. Este entorno, junto a los paquetes de funciones nombrados, contiene los algoritmos que se

estimaron oportunos utilizar para el clustering y permite modificaciones en modos de cálculo, precisión o número de iteraciones, dando al programador una gran capacidad para escoger la metodología más conveniente en cada momento.

3.2 DATOS

3.2.1 Selección de los datos

3.2.1.1 Qué y por qué

Tal y como se ha recogido en el capítulo 2, la elección de los datos para la construcción de un modelo de predicción del comportamiento del mercado de valores es una tarea difícil y prácticamente imposible. Hay multitud de publicaciones, tesis y libros escritos sobre la materia, y los expertos no parecen ponerse de acuerdo en cómo estaría formado el conjunto de datos óptimo para esta tarea.

De esta manera, para este proyecto se ha estimado oportuna la utilización de los precios de cierre de la acción, esto son, los precios que tienen las diferentes acciones con el cierre de mercado al final del día. Para la elección de estos datos se ha partido de la premisa de que el precio de la acción, como elemento último de representación del mercado, recoge información intrínseca de todos los diferentes factores que afectan al comportamiento del mercado de forma ponderada.

Lo que se ha buscado utilizando estos dos datos de cada acción es evitar el sobreajuste (overfitting), que provocaría un mal diseño del modelo. Se puede argumentar que estos datos no sean suficientes y que el modelo podría ser más preciso utilizando más indicadores, pero en todo momento se ha buscado obtener un modelo robusto y por lo tanto se ha penalizado más el sobreajuste que la falta de datos.

Con el objetivo de alimentar al algoritmo con la mayor cantidad de evoluciones de precios de acciones posible, se han elegido un total de 898 empresas, repartidas en los mercados que muestra la siguiente tabla:

Tabla 1. Mercados e índices utilizados para la realización del proyecto

Índice/ Mercado	País
IBEX 35	España
IBEX MEDIUM CAP	
IBEX SMALL CAP	
MERCADO CONTINUO	
FTSE 100	Reino Unido
MILAN	Italia
LISBOA	Lisboa
BRUSELAS	Bélgica
AMSTERDAM	Holanda
DAX 30	Alemania
VIENA	Austria
S&P 500	Estados Unidos
NASDAQ 100	
TORONTO TSX 60	Canadá

3.2.1.2 Granularidad

Los datos relativos a los diferentes mercados de valores y los datos financieros en general existen con una resolución desde milisegundos a décadas. Cuanto mayor es la resolución, igual que en cualquier otro tipo de sistema dinámico, mayor es la precisión y más detalles se pueden observar, pero a cambio de tener más ruido. Además, la información financiera en alta resolución no es accesible normalmente a no ser que se tenga algún tipo de cuenta premium que supondría un pago periódico. Así, se ha elegido una resolución diaria ya que, aunque puede haber fluctuaciones por eventos puntuales, al tener únicamente en cuenta el precio de cierre se ha entendido que estos cambios inusuales no tendrán un gran peso en la creación del modelo. Además, la utilización de datos diarios parece el equilibrio perfecto entre la gran cantidad de información (y ruido) que un periodo de horas y minutos podría tener con la robustez de los datos semanales, mensuales o anuales.

De esta manera, los datos diarios están libres del ruido intradía y son fáciles de encontrar en internet de forma gratuita. Además, los datos diarios parecen perfectos, aunque podrían surgir problemas con días singulares como los fines de semana, las vacaciones o las diferentes estaciones. Todos estos "problemas" se verán resueltos debido al tipo de aprendizaje automático utilizado, que como ya se ha explicado detecta estructuras internas en los datos y los agrupa según se parezcan más o menos.

Además, y como también se ha expuesto en el capítulo 2, existe ruido asociado a las noticias y rumores que en ocasiones pasa la franja intradía pero que son rápidamente ajustados por el propio mercado al día siguiente.

3.2.1.3 Periodo de tiempo

Una vez escogida la definición de tiempo a utilizar, es importante elegir el periodo de tiempo en el que se va a trabajar. La utilización de datos muy antiguos podría crear un modelo desfasado en el tiempo, que no tendría suficiente precisión. Por el contrario, cuanto más largo sea el periodo de tiempo, más datos se tendrán y por lo tanto mayor será la robustez del sistema, siendo de hecho imposible crear el modelo si no se llega a un mínimo de datos.

Para la realización de este proyecto se ha considerado interesante mirar hacia atrás en el tiempo en un periodo de 15 años en los casos en los que ha sido posible (de algunas empresas solo hay datos de los últimos 10 o 12). La siguiente imagen muestra la evolución del Standard and Poor's 500 durante los últimos 15 años:

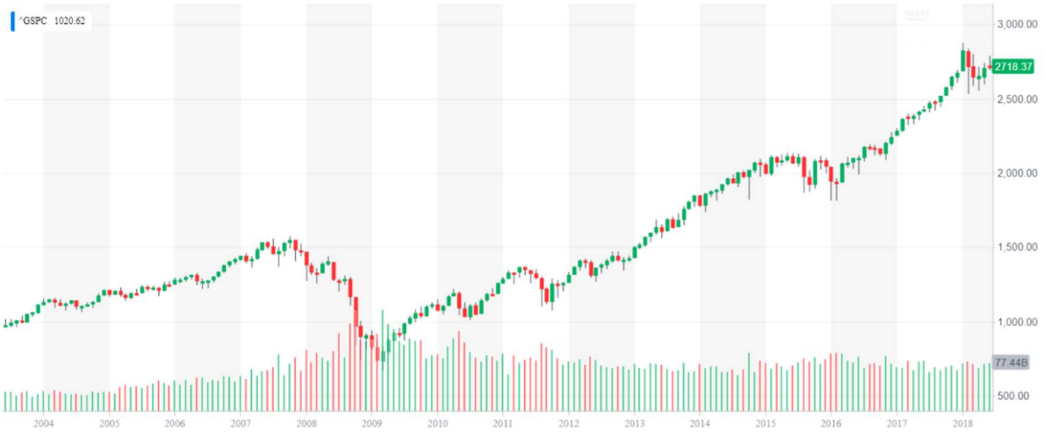


Ilustración 11. Evolución del índice Standard and Poor's 500 desde 2003

Se ha elegido este índice por ser uno de los más representativos de la economía global.

Para las distintas fases de entrenamiento, validación y test, se suelen dividir los datos en un 50%, 30% y 20% relativamente. En este caso, además se han querido recoger comportamientos similares, es decir, que cada conjunto de datos tenga una bajada y una recuperación. De esta manera, y para cumplir aproximadamente los porcentajes estándar se ha dividido el tiempo de la siguiente manera:

- Entrenamiento: desde junio de 2.003 hasta mayo de 2.011
- Validación: desde junio 2.011 hasta mayo de 2.015
- Test: desde junio de 2.015 hasta junio de 2.018

Con esta división de los datos, el entrenamiento tendrá la crisis financiera de 2.008, la validación el bajón de 2.011 y el test la caída de 2.016, quedando los porcentajes de datos en el tiempo aproximadamente como un 53% para el entrenamiento, 27% para la validación y

20% para el test. Estas proporciones son por lo tanto suficientemente parecidas a las indicativas 50-30-20 del procedimiento estándar.

3.2.1.4 Pruebas preliminares

Hay multitud de maneras y metodologías de entrenar algoritmos de aprendizaje automático. Una de las decisiones más importantes es la elección del periodo de tiempo como se acaba de explicar, en otras palabras, cuántos días de datos del pasado se utilizarán para predecir cuántos días del futuro. A modo de ejemplo, si eligiésemos utilizar tres días para predecir uno en el futuro, para predecir el movimiento de los precios del viernes necesitaríamos la información del martes, miércoles y jueves anteriores.

Una forma simple de encontrar un orden de magnitud para estos parámetros fue la realización de pruebas previas sencillas, utilizando los resultados para determinar los parámetros. Lo que se entiende por sencillo en este caso es la utilización de una menor cantidad de datos (en este caso solamente se utilizaron las acciones del IBEX 35), con el objetivo de reducir el tiempo de computación. Como se verá en la parte de experimentos, la elección del periodo de tiempo es una de las variables que se cambia de valor para obtener diferentes resultados, aunque sí permanece constante la relación utilizada, que es de uno a uno. Lo que se intenta expresar diciendo que la relación es de uno a uno es que, si se quieren predecir treinta días en el futuro, se están utilizando treinta días hasta el momento, y que si se quieren predecir cincuenta, se necesitará el histórico de los últimos cincuenta días.

Tras realizar estas pruebas simples, y teniendo en mente el objetivo inicial del proyecto que es la creación de una herramienta para el medio-largo plazo, se determinó que la predicción de menos de treinta días no tendría sentido y que la relación de días previstos e histórico sería de uno a uno, aunque en el apartado de experimentos se verá algún caso en el que también se ha probado a cambiar esta relación.

3.2.2 Preparación de los datos

3.2.2.1 Generación de ventanas

Una vez obtenidos los datos, lo primero que se ha hecho ha sido generar las ventanas de datos, es decir, escoger qué datos se van a utilizar para el proyecto. Tal y como se ha explicado en el apartado anterior, se ha elegido un período de tiempo de quince años. De esta manera, todos los datos anteriores a un periodo de quince años desde el día 8 de junio de 2018 se han descartado para el estudio ya que se ha considerado que el sistema ha cambiado de forma suficientemente significativa como para que significasen una distorsión en el entrenamiento

del modelo más que una aportación de información. La siguiente figura representa la elección de la generación de la ventana de datos con los quince años, teniendo en cuenta que habrá datos más antiguos de unas acciones y no tan antiguos de otras:

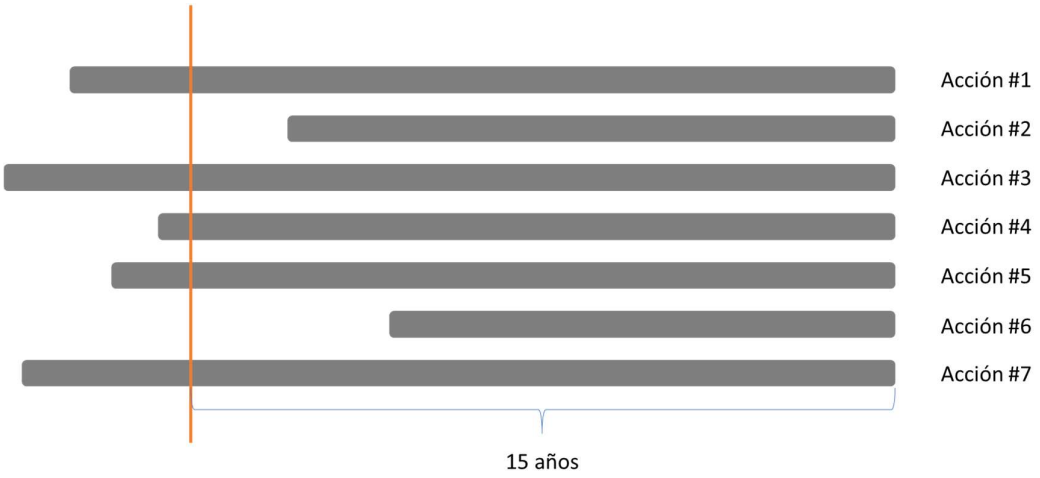


Ilustración 12. Formación de ventana de 15 años de antigüedad

En los casos en los que la acción no haya tenido valores suficientes para los quince años que se han requerido, la matriz se ha rellenado con NaN (Not a Number) en Matlab y después se ha procedido a su "limpieza" tal y como se detalla en el siguiente punto.

3.2.2.2 Limpieza de ventanas

Un paso posterior a la creación de las ventanas ha sido su limpieza. Por limpieza se entiende la eliminación de valores atípicos que significarían ruido para cualquiera de las fases del modelado (entrenamiento, validación y test) y de los valores no numéricos (NaN) que provocarían errores de cálculo. De esta manera, se han descartado todos los datos que de un día para otro hayan sufrido un cambio de más del 50% en su valor. También otras situaciones que podrían haber creado errores en la ejecución de los algoritmos como podría ser un hueco en los datos (que no se tengan por ejemplo los datos de tres días seguidos) se han evitado en esta "limpieza de ventanas".

En este análisis el volumen de datos utilizados para cada una de las fases (entrenamiento, validación y test) es tan alto que la eliminación de los "elementos problemáticos" no variará los resultados, mientras que la mejora en el funcionamiento del algoritmo será significativa ya que se evitarán errores propios de los datos.

3.2.2.3 Creación de las matrices de entrenamiento, validación y test

Como ya se ha comentado en este documento, es muy importante la división de la matriz con todos los datos en tres matrices de las proporciones aproximadas 50-30-20% como esquematiza la siguiente figura:

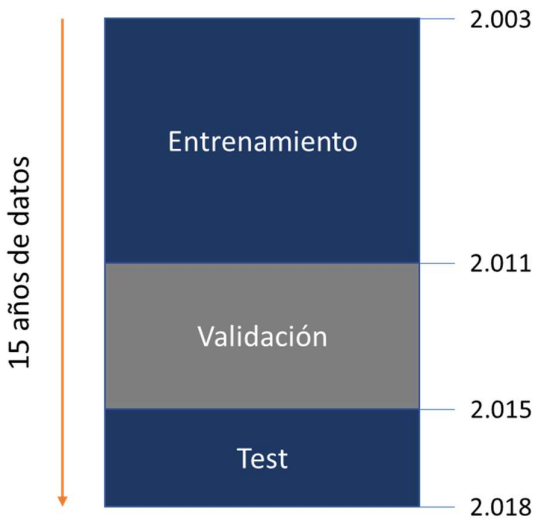


Ilustración 13. Esquema de la división de los datos en las matrices de entrenamiento, validación y test de resultados

De esta manera, y debido a la estructura de los datos, la división de ventanas se ha hecho de datos ordenados en el tiempo. Esto significa que se ha creado la matriz de entrenamiento con los primeros datos que se tienen, es decir, con los años de los primeros ocho años aproximadamente, la matriz de validación con los datos de los siguientes cuatro años, y la de test con el restante de los datos.

Se ha considerado que el mercado actual se comportará de manera más similar a los datos más recientes, y por lo tanto si el modelo pasa el filtro de la validación con este conjunto de datos, será un modelo válido.

3.2.2.4 Creación de las ventanas de entrada y salida

La forma en la que se ha diseñado el sistema es relativamente simple: se divide una ventana de un determinado número de días que se puede cambiar como parámetro en dos. Estas dos ventanas se ha escogido que tengan una relación de tamaño uno a uno tal y como se ha explicado en el capítulo previo.

Una vez se ha dividido la ventana en dos, se ha llamado a la primera ventana “ventana de entrada” y a la segunda “ventana de salida”. De esta manera, se clasifica con el modelo de clustering cada ventana según su ventana de entrada y se estudia el comportamiento de la

ventana de salida. Este comportamiento quedará resumido en unos parámetros que se exponen más adelante en este documento y que servirán para clasificar el cluster según su rentabilidad esperada y sus variabilidades interna y externa. El siguiente esquema muestra de forma gráfica la estructura de estas dos ventanas de entrada y salida:

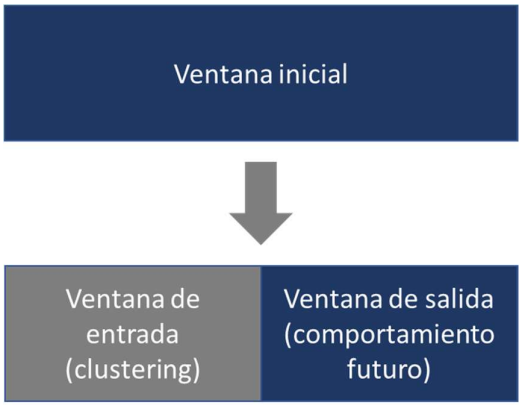


Ilustración 14. División de la ventana inicial en las ventanas de entrada y salida

3.2.2.5 Normalización

El precio de una acción dependerá básicamente del número de acciones que tenga la empresa y de su valor de mercado, siendo el precio de la acción cociente entre el valor y el número de acciones:

$$Precio\ de\ la\ acción = \frac{Valor\ de\ mercado\ de\ la\ empresa}{Número\ de\ acciones\ de\ la\ empresa} \quad (3.1)$$

Con el objetivo de homogeneizar los datos para poder realizar un análisis de conjunto en el que se comparen rentabilidades o evolución, en otras palabras, para poder encontrar patrones y estructuras internas en la evolución de los datos con independencia del precio de la acción, resulta importante la normalización de las ventanas. Dicha normalización se ha realizado dividiendo todos los valores de la ventana por el primer valor de la ventana y se ha hecho por separado para la de entrada y la de salida (se han dividido los valores de la ventana de entrada por el primer valor de la ventana de entrada y los de la ventana de salida por el primer valor de la ventana de salida).

3.2.2.6 Creación de matriz de parámetros

Por último, para la clasificación de las ventanas de entrada no se ha trabajado con los precios finales brutos, sino que se han sacado parámetros de cada ventana como pueden ser el número de máximos y mínimos o el valor del máximo y el mínimo absoluto de la ventana. En

un principio se hizo la clasificación de los precios sin utilizar parámetros, es decir, con la evolución de los precios normalizados, sin obtener resultados interesantes que por otro lado sí se han obtenido trabajando con parámetros.

Durante la etapa de obtención de resultados, y tal como se explica en el capítulo 4, se han hecho pruebas con diferentes combinaciones de parámetros a partir de una primera lista de treinta, de los cuales se han mantenido veinticuatro debido a que los otros seis presentaban una alta correlación con alguno de los que se han utilizado finalmente, añadiendo ruido al sistema. Los veinticuatro parámetros utilizados por lo tanto son:

- Máximo de la ventana
- Mínimo de la ventana
- Media móvil
- Valor final de la ventana
- Mínimo entre el primer valor de la ventana y el valor máximo
- Mínimo entre el valor máximo y el último valor de la ventana
- Máximo entre el primer valor de la ventana y el valor mínimo
- Máximo entre el valor mínimo y el último valor de la ventana
- Número de muestras desde la última hasta que se produce el mínimo
- Número de muestras desde la última hasta que se produce el máximo
- Número de muestras desde la última hasta que se produce el mínimo anterior al máximo
- Número de muestras desde la última hasta que se produce el mínimo posterior al máximo
- Número de muestras desde la última hasta que se produce el máximo anterior al mínimo
- Número de muestras desde la última hasta que se produce el máximo posterior al mínimo
- Desviación estándar de los datos de la ventana
- Varianza de los datos
- Pendiente de la línea de regresión lineal
- Valor inicial
- Diferencia entre valor final y valor inicial
- Desviación superior máxima de los datos de la ventana con la recta que une el valor inicial con el valor final

- Desviación inferior máxima de los datos de la ventana con la recta que une el valor inicial con el valor final
- Índice de fuerza relativa o RSI
- Banda superior de Bollinger
- Banda inferior de Bollinger

3.2.2.7 Estructura de los datos

Una vez explicada la forma de las ventanas y la importancia de la normalización y parametrización de los datos, queda hacer un último apunte sobre la formación de las ventanas y la estructura de los datos. En este caso, se ha escogido la formación de ventanas deslizantes para poder utilizar parámetros deslizantes tal y como se ha indicado en el apartado anterior. La siguiente figura muestra esquemáticamente cómo se producen ventanas de la misma acción con el avance del tiempo:

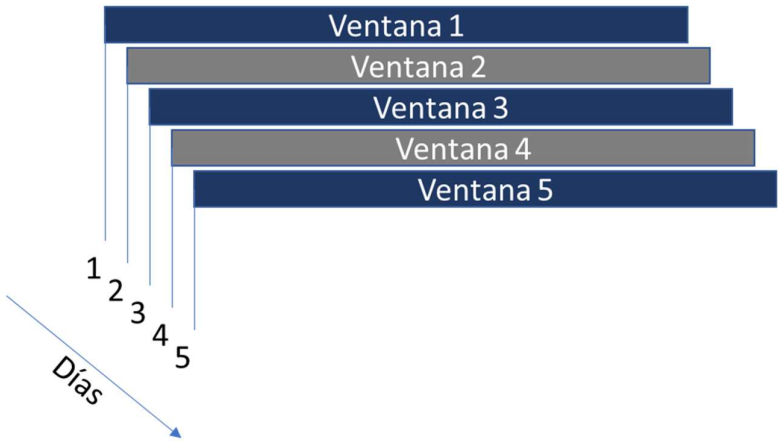


Ilustración 15. Ventanas en el tiempo

Imagine el lector que las ventanas de la figura tienen una duración de treinta días comenzando por el día 1. La ventana 2 tendrá todos los datos que tiene la ventana 1, excepto el primero, y el que era el segundo dato de la ventana 1 será ahora el primero de la ventana 2. Además, en la ventana 2 habrá un dato que no había en la ventana 1, que es el último. De esta manera, se van formando las diferentes ventanas que contienen toda la información de la ventana anterior excepto el primer dato y desplazan los datos una posición hacia la izquierda en el vector, dejando hueco para el nuevo dato que se incorporará al final del mismo.

3.2.2.8 Tendencias del mercado

Al utilizar datos históricos financieros, es importante recalcar que la mayoría de los periodos de tiempo tienen una tendencia general creciente ya que la mayoría de los mercados han

crecido en el tiempo. Por lo tanto, el modelo tendrá más datos de tendencias alcistas y será normal que unos clusters tengan más miembros que otros, aunque no se debe despreciar ningún tipo de información pues cada cambio en el desarrollo del mercado se puede aprovechar gracias a la gran variedad de productos financieros que hay en el mercado.

Tal y como se expone en la sección en la que se explica la elección del periodo de tiempo estudiado, se incluyó a propósito la crisis financiera de 2008 para así tener más probabilidades de predecir posibles futuras caídas similares, aunque se hayan tenido que añadir una gran cantidad de cambios atípicos al modelo a cambio.

3.3 LOS ALGORITMOS

En el proyecto se han utilizado algoritmos relacionados con el aprendizaje automático para dos tareas en particular: la reducción de dimensionalidad de la matriz de parámetros y el agrupamiento de las ventanas de entrada en clusters.

3.3.1 Reducción de dimensionalidad

A partir de cada ventana de precios se han sacado treinta parámetros diferentes. Era necesario reducir la dimensión de treinta a dos o tres por dos razones principales:

1. Treinta parámetros son demasiados para el algoritmo de clustering, ya que tardaría mucho tiempo en ejecutar y se necesitaría una potencia computacional mayor a la que se tiene para la ejecución de este proyecto
2. Entre los parámetros hay correlaciones altas, y tal y como se expone en el estado del arte, uno de los mayores problemas de los algoritmos de aprendizaje no supervisado es la correlación sin causalidad. Con la reducción de dimensionalidad lo que se consigue entre otras cosas es eliminar las altas correlaciones lineales que existen entre los diferentes parámetros, construyendo de esta manera un modelo más robusto.

Así, se han probado dos tipos diferentes de reducción de dimensionalidad. El primero y el que se ha terminado utilizando es el análisis de componentes principales o PCA (Principal Component Analysis). PCA es la principal técnica de reducción de dimensionalidad linear, y consiste en el mapeo de los datos en un espacio de menores dimensiones de tal manera que la varianza de los datos en las dimensiones reducidas sea máxima. Para lograr esta varianza máxima, se sacan los autovalores y los autovectores. De esta manera, los autovectores asociados a los autovalores más altos serán las componentes principales del nuevo espacio de dimensionalidad reducida, y la matriz de datos se multiplicará por la matriz de paso,

obteniéndose de esta forma las componentes de cada uno de los datos en el nuevo espacio, en este caso, los vectores con tres componentes que mejor resumen o explican la información que tenían los vectores de treinta parámetros.

El otro método que se ha utilizado en el proyecto pero que se ha descartado por la carga computacional que supone es t-SNE (t-Distributed Stochastic Neighbor Embedding), que se vale de la utilización de redes neuronales para respetar de una forma mejor que el PCA las distancias euclídeas que hay entre los puntos del espacio que se quiere reducir. Esta técnica resulta muy interesante porque organiza los datos de una manera mucho más agrupada y separando unos grupos del otro, práctica que agiliza considerablemente la labor del algoritmo de clustering. Esta mayor separación entre grupos se traduciría en que si se representan en dos o tres dimensiones los nuevos puntos en el espacio de menor dimensionalidad, a simple vista sería fácil reconocer las agrupaciones y por lo tanto los clusters que formarán los algoritmos de clustering. Para implementarlo reduciendo la carga computacional, también se ha probado a reducir la dimensionalidad a cinco dimensiones con la metodología PCA para después aplicar t-SNE y reducir de cinco a dos, pero la carga computacional seguía siendo demasiado elevado.

3.3.2 Clustering

Tal y como se ha expuesto en el apartado 2.3 dedicado al clustering, en este proyecto se han utilizado dos estrategias diferentes de clustering: k-means y GMM.

K-means es la metodología que se comenzó utilizando por ser computacionalmente más ligera, aunque GMM es el que ha dado mejores resultados y ha terminado siendo el preferido, ya que con una diferencia computacional baja, el modelo es considerablemente más robusto con respecto a los valores iniciales que se den a los clusters, tal y como queda expuesto en 2.3.

3.4 ESQUEMA DE METODOLOGÍA

Una vez expuesta la metodología seguida en el proyecto, se ha considerado visual y oportuno realizar el siguiente esquema para una mejor comprensión por parte del lector:

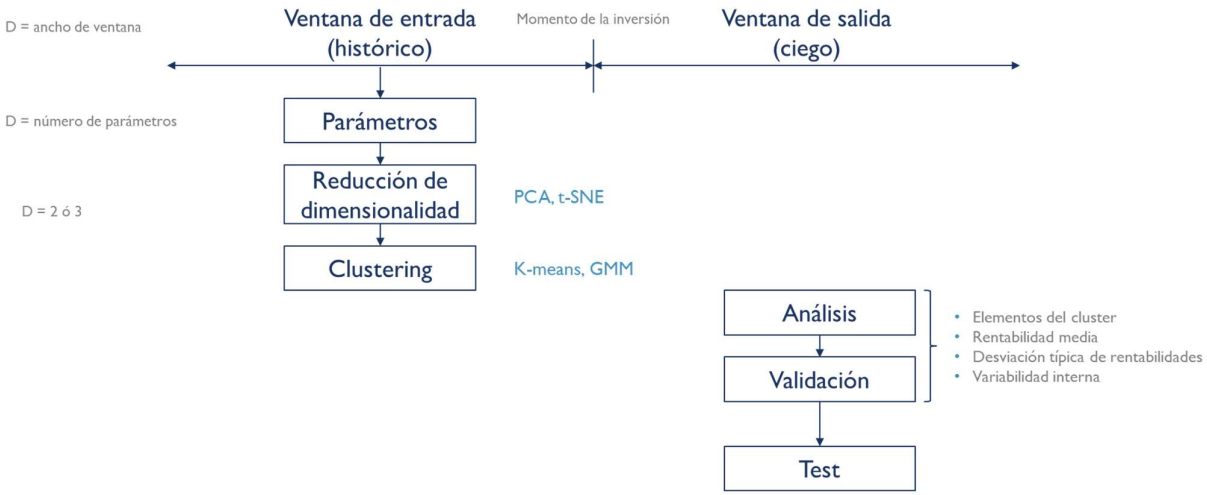


Ilustración 16. Resumen esquemático de la metodología seguida

3.5 ENTENDIENDO LOS RESULTADOS

Para que los diferentes experimentos sean rápidos de comparar, se ha elegido un formato de tabla con cinco columnas:

1. Cluster: muestra el número de cluster del que se están mostrando las características en la fila
2. Members: representa el porcentaje de observaciones que pertenecen a ese cluster. Un cluster será tanto más importante cuanto más porcentaje de observaciones represente.

$$Members\ n\ (\%) = \frac{número\ de\ elementos\ en\ el\ clúster\ n}{número\ de\ elementos\ clasificados} \times 100 \qquad (3.2)$$

   Siendo n el número indicativo del cluster.

3. Expected growth: rentabilidad media anual de las ventanas de salida cuyas ventanas de entrada se corresponden con el cluster de la fila. Los datos están ordenados según el crecimiento esperado.

$$Expected\ growth\ n\ (\%) = \sum_{i=1}^{N} r_i \times \frac{252}{N \times ldv} \times 100 \qquad (3.3)$$

   Siendo r_i la rentabilidad del elemento i del cluster n, N el número de elementos del clúster n y ldv la longitud de la ventana escogida como metaparámetro.

4. External variation: variación de las rentabilidades de las ventanas de salida de cada uno de los clusters. Es una medida de la similaridad entre las rentabilidades de las ventanas de salida de cada cluster. Un cluster será tanto mejor cuanto menor sea esta variación, pues eso significará que las ventanas de salida se comportan de forma similar en el periodo de tiempo estudiado (tamaño de la ventana de salida).

$$Ev = \sqrt{\frac{\sum_{i=1}^{N}(r_i - \bar{r})^2}{N}} \qquad (3.4)$$

Donde N es el número de elementos del clúster n, r_i la rentabilidad de la muestra i y $\bar{r}$ la media de rentabilidades.

Cabe añadir que todos los clusters con un valor en el campo External variation mayor de uno han sido descartados, ya que una desviación típica del 100% se ha considerado el máximo para considerar a un cluster nicho de mercado objetivo. Esta decisión se ha tomado en base a la idea de que una desviación típica superior al 100% implicaría la posibilidad de perder dinero.

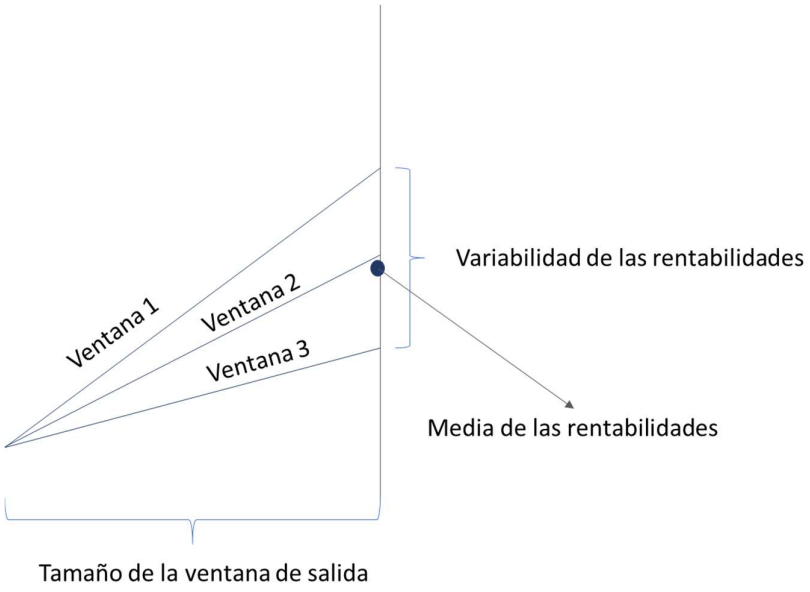


Ilustración 17. Representación del significado de Std

5. Internal variation: es una medida de la variación interna de cada acción. En otras palabras, es la variabilidad con respecto a la línea que une el primer valor de la ventana con el último de la curva que describen los diferentes puntos. Igual que Std, el cluster será tanto mejor cuanto menor sea la variabilidad interna.

$$Iv(n) = \frac{\sum_{j=1}^{NC} \sqrt{\frac{\sum_{i=1}^{N} (x_{i,k} - \overline{x_{\iota,k}})^2}{N}} \times \frac{1}{r_k}}{NC} \quad (3.5)$$

Siendo N el número de datos de muestras i de la ventana k, $x_{i,k}$ cada una de las muestras de la ventana k, $\overline{x_{\iota,k}}$ cada uno de los puntos de la recta de regresión correspondientes a cada $x_{i,k}$, r_k la rentabilidad final de la ventana k y NC el número de miembros del cluster n.

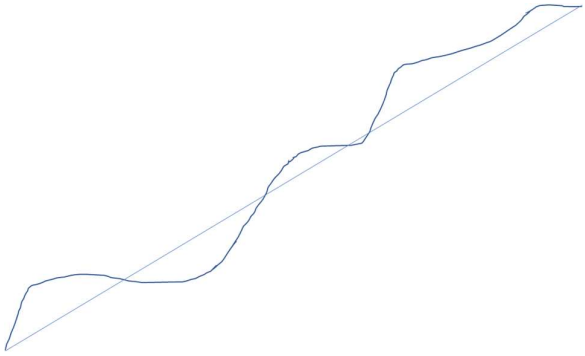


Ilustración 18. Representación de la variabilidad interna de la ventana Std 2

4 RESULTADOS

Con el objetivo de encontrar clusters objetivo, es decir, que tengan un alto nivel de rentabilidad y baja variación tanto interna como externa, se han probado diferentes combinaciones de metaparámetros.

Cabe destacar que no hay una combinación de metaparámetros óptima, ya que dependiendo de la naturaleza de las acciones y de los comportamientos que puede tener el precio de una acción, puede haber diferentes periodos de tendencia por ejemplo. Esto implica que, por ejemplo, si el tamaño de la ventana es muy grande no se podrán detectar las tendencias con un periodo de tiempo menor, y si el tamaño de la ventana es pequeño se perderán las tendencias interesantes a mayor plazo.

Además, tal y como se ha expuesto en los capítulos anteriores, el clustering es un método de análisis de datos que agrupa valores que se han comportado de manera similar. Así, la hipótesis que se quiere valorar en este proyecto es: *Si un valor se ha comportado de manera similar al resto del cluster en un determinado periodo de tiempo (durante la ventana de entrada), ¿es posible conocer su comportamiento futuro a partir de los comportamientos futuros de los elementos del mismo cluster?* De esta manera, si se cumple la hipótesis y además los datos previstos para un cluster en particular incluyen una rentabilidad interesante, positiva o negativa, y baja variabilidad, se habrá encontrado un nicho de inversión objetivo.

Es también importante señalar que la variabilidad entre elementos del cluster, es decir, la desviación típica con respecto a la rentabilidad media del cluster, es un parámetro más importante que la variabilidad interna, o en otras palabras, la variación de la evolución de un precio con respecto a la línea recta de evolución.

Tras varias pruebas, se decidió que los parámetros a utilizar para el clustering serían los nombrados en el apartado 3.2.2.6. Este conjunto de parámetros minimiza las correlaciones lineales entre elementos a la hora de hacer la reducción de dimensionalidad y por lo tanto añade menos ruido al modelo. Esta es la razón por la que este metaparámetro (el conjunto de parámetros a partir de los precios de cierre) no es uno de los que se ha cambiado para obtener los diferentes resultados, sino que se ha mantenido constante.

Con el objetivo de abordar la obtención de resultados de forma sistemática, se plantearon los siguientes casos teniendo en cuenta la capacidad computacional que se tiene para el proyecto, y que buscan explorar el mayor número de opciones diferentes posibles para dar con ejemplos de clusters interesantes de tal manera que se demuestre su existencia, aunque quede para futuros desarrollos del proyecto la optimización en la búsqueda de los mismos.

4.1 ¿CUÁNDO ES UN CLUSTER BUENO?

Tal y como se ha explicado a lo largo del documento, un cluster será tanto mejor cuanto:

- Mayor porcentaje de los datos contenga
- Mayor sea, en valor absoluto, su rentabilidad
- Menor sea la desviación de rentabilidades de los diferentes elementos del cluster
- Menor sea la variabilidad interna de cada elemento del cluster

Estas medidas indicativas, no obstante, se tienen que hacer cuantificables para poder valorar si un cluster es interesante o no. Es por esto que, para que un cluster sea considerado nicho de inversión, debe cumplir que:

- Represente al menos un 1% de los datos totales en las fases de entrenamiento y validación. Esto se hace porque si un cluster representa menos del 1%, podría ser producto del sobreajuste (overfitting), y por lo tanto no representar un nicho interesante de inversión, sino el agrupamiento de datos atípicos en el pasado.
- Su rentabilidad, en valor absoluto, sea superior al 20% anual. En primer lugar, se busca valor absoluto y no solamente rentabilidad negativa, porque si se tiene la certeza de que un determinado valor va a bajar su precio, el inversor se puede posicionar en corto y obtener beneficios del cambio de precio. Además, se ha considerado que un 20% de rentabilidad anual es un rendimiento financiero más que aceptable por estar considerablemente por encima del crecimiento medio de la bolsa española en los últimos veinticinco años, que se encuentra alrededor de un 5,1% anual. [28]
- Su desviación típica entre rentabilidades futuras de elementos del cluster es menor de uno. Con una desviación típica media de uno, en el peor escenario a una distancia de una desviación típica de la media, el inversor no ganaría nada con la inversión, pero tampoco perdería dinero (sin tener en cuenta el coste de oportunidad). De esta manera, todo cluster que ha mostrado una variabilidad externa superior a uno ha sido automáticamente descartado en la búsqueda de nichos de inversión.
- La media de variabilidades internas de los elementos del cluster sea inferior a uno. Aunque esta variabilidad no sea tan importante como la recién nombrada, también se valora en un activo que su precio no oscile y que se comporte de la forma más parecida a la tendencia que puede marcar por ejemplo una media móvil. En este caso, y tal y como se ha explicado en el capítulo de metodología, se ha comparado la evolución de cada ventana de salida con una línea recta.

4.2 PRUEBAS CON DIFERENTES METAPARÁMETROS

Tal y como se ha explicado en capítulos anteriores de este documento, los dos metaparámetros más importantes por su importancia en los resultados finales son el tamaño de ventana y el número de clusters. Por un lado, el tamaño de ventana es indicativo del periodo de tiempo en el que se están buscando las tendencias, ya que diferentes acciones dependiendo de su naturaleza y de las circunstancias se comportarán de maneras que pueden resultar interesantes a un inversor dependiendo del periodo de tiempo en el que se estudian. Por otra parte, el número de clusters es un indicativo directo de la cantidad de agrupaciones que se decide hacer con los datos que se tienen. Cuanto mayor sea el número de clusters, mayor riesgo de sobreajuste (overfitting), pero un número demasiado pequeño hará prácticamente imposible a su vez que haya poca variabilidad entre elementos del mismo cluster porque agrupará elementos más diferentes en el mismo cluster.

Además, hay que tener en cuenta que una de las limitaciones de este proyecto es la capacidad computacional, y que cuanto mayor sean el tamaño de la venta y el número de clusters, mayor serán las necesidades de computación. De esta manera, teniendo en cuenta estas limitaciones y queriendo valorar los datos en diferentes órdenes de magnitud de tiempo (tamaño de ventana), se han estudiado en primer lugar los siguientes casos:

	Tamaño de la ventana				
Número de clusters		30	50	100	150
	30	Caso I	Caso II	Caso III	Caso IV
	50	Caso V	Caso VI	Caso VII	Caso VIII
	100	Caso IX	Caso X	Caso XI	Caso XII
	150	Caso XIII	Caso XIV	Caso XV	Caso XVI

Tabla 2. Resumen del enfoque sistemático para la prueba de metaparámetros

En cada caso se ha entrenado el algoritmo con la combinación de metaparámetros que aparece en la tabla y se han estudiado las características de los diferentes clusters.

Para analizar cada caso representado en la tabla, se ha actuado de manera similar:

1. En primer lugar, se ha corrido el algoritmo de entrenamiento y formación del modelo. De esta manera, se han encontrado clusters que podrían resultar interesantes debido a su alta rentabilidad y baja variabilidad.
2. Una vez detectados estos clusters, se ha procedido a ajustar los datos reservados a la validación a los clusters creados y se han estudiado en detalle los clusters que habían

resultado interesantes durante el entrenamiento para ver si la repetitividad en los resultados es alta, es decir, si se siguen manteniendo unas características de rentabilidad y variabilidades similares. Tras este análisis, se han descartado aquellos clusters que habían sido elegidos en primer lugar y que han aumentado considerablemente su variabilidad o disminuido su rentabilidad.

3. Con los clusters que han pasado los filtros de los dos puntos anteriores, se ha procedido a realizar el test. De nuevo se han distribuido los datos, esta vez del conjunto de test, a los clusters existentes. Así, se han vuelto a estudiar las características de los datos agrupados en los clusters señalados tras el segundo paso para comprobar que siguen comportándose de forma similar con los nuevos datos. De esta manera, y tal y como se había hecho en el apartado de validación, se han descartado los clusters que en la fase de test han proporcionado cambios considerables a peor en rentabilidad o variabilidad.

Una vez explicado el procedimiento seguido, a continuación se muestran los resultados de cada uno de los casos estudiados:

4.2.1 Caso I: tamaño de ventana 30 y agrupación en 30 clusters

La siguiente tabla muestra los resultados de la fase de entrenamiento del Caso I:

Tabla 3. Resultados entrenamiento Caso I

Cluster	Members	Expected_growth	External_variation	Internal_variation
27	3,97%	-11,65%	5,22	0,03
23	1,03%	-9,20%	6,94	0,03
1	12,91%	-9,15%	7,06	0,03
29	0,33%	-7,67%	8,70	0,03
21	2,94%	-6,46%	11,98	0,03
15	6,78%	-5,83%	12,72	0,03
3	6,92%	-3,30%	20,35	0,03
6	1,81%	-2,49%	30,44	0,03
30	0,06%	0,32%	250,52	0,03
7	5,11%	2,18%	31,85	0,03
10	0,50%	6,10%	14,57	0,04
5	1,03%	8,57%	11,96	0,04
17	0,22%	9,73%	9,78	0,04
12	3,38%	11,34%	10,18	0,04
19	3,09%	16,64%	5,57	0,04
8	0,07%	20,12%	5,48	0,04
22	1,55%	22,79%	4,82	0,04
11	0,75%	25,68%	7,11	0,07
2	3,11%	25,83%	5,92	0,05

18	0,31%	27,46%	4,44	0,05
14	7,47%	29,06%	3,90	0,04
4	1,40%	43,37%	2,79	0,05
20	6,60%	56,97%	3,09	0,06
28	0,22%	67,35%	2,30	0,06
13	0,01%	72,54%	2,75	0,07
24	0,06%	75,62%	3,64	0,09
26	17,73%	103,61%	1,89	0,07
16	0,16%	131,34%	1,67	0,07
9	7,12%	146,69%	1,65	0,08
25	3,38%	163,41%	2,35	0,10

En este primer caso, ninguno de los clusters ha pasado el corte de la fase de entrenamiento debido a la alta variabilidad (el criterio es que sea menor que 1 o ligeramente superior) entre las rentabilidades de los elementos del mismo cluster. La causa de esta alta variabilidad es que comparado con el volumen de datos del entrenamiento (ocho años, aproximadamente novecientas empresas), el tamaño de ventana y el número de clusters son demasiado pequeños y por lo tanto hay una alta variedad entre elementos agrupados en el mismo cluster. Se podría decir por lo tanto, que al ser unos metaparámetros tan relajados se ha producido una sobregeneralización en el sistema.

4.2.2 Caso II: tamaño de ventana 50 y agrupación en 30 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso II:

Tabla 4.Resultados entrenamiento Caso II

Cluster	Members	Expected_growth	External_variation	Internal_variation
23	33,95%	-8,95%	5,23	0,04
28	5,32%	-8,73%	5,54	0,04
9	0,85%	-7,00%	8,00	0,04
10	2,23%	-6,77%	7,54	0,04
1	1,56%	-4,99%	9,85	0,04
2	0,50%	-4,51%	12,37	0,04
25	0,35%	-3,67%	14,63	0,04
18	5,44%	-3,03%	19,65	0,04
17	3,34%	1,12%	67,01	0,05
27	4,80%	3,50%	19,62	0,05
26	0,02%	4,30%	13,80	0,04
14	1,15%	6,20%	11,72	0,05
4	1,02%	10,11%	7,52	0,05
8	3,64%	10,51%	7,29	0,05
12	0,08%	12,46%	6,72	0,05
20	2,13%	15,09%	5,29	0,05

13	2,34%	18,58%	4,86	0,06
16	2,48%	21,71%	4,28	0,05
24	0,21%	25,03%	4,29	0,07
30	4,08%	29,48%	3,65	0,06
5	0,04%	31,21%	4,06	0,07
3	0,20%	34,63%	3,42	0,06
7	5,30%	48,14%	2,24	0,07
19	0,43%	52,85%	2,26	0,08
22	6,07%	66,87%	3,24	0,10
6	3,35%	70,94%	2,63	0,09
11	2,04%	72,30%	1,88	0,10
15	6,38%	108,86%	1,83	0,09
21	0,05%	110,72%	2,38	0,11
29	0,65%	128,90%	1,55	0,11

En este caso, igual que en el anterior, no se ha obtenido en la fase de entrenamiento ningún cluster que cumpliese el requisito mínimo de variabilidad interna.

4.2.3 Caso III: tamaño de ventana 100 y agrupación en 30 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 5. Resultados entrenamiento Caso III

Cluster	Members	Expected_growth	External_variation	Internal_variation
1	7,99%	-7,34%	4,44	0,05
18	0,59%	-6,49%	5,30	0,05
16	0,16%	-5,51%	6,15	0,05
17	1,13%	-5,02%	6,78	0,05
12	0,21%	-4,83%	7,77	0,06
29	0,40%	-1,58%	25,54	0,06
30	1,64%	-0,87%	49,30	0,07
5	9,13%	1,49%	30,31	0,06
9	2,42%	3,89%	14,63	0,08
8	3,94%	4,02%	10,78	0,06
10	1,17%	4,40%	11,61	0,07
22	8,16%	5,26%	9,13	0,07
24	3,02%	5,91%	11,04	0,08
21	0,20%	7,19%	7,30	0,07
25	0,56%	8,28%	5,85	0,07
13	17,66%	10,11%	5,58	0,07
15	15,00%	15,37%	4,47	0,09
7	9,01%	16,08%	3,84	0,08
11	1,13%	21,05%	2,81	0,07
27	0,64%	25,32%	2,50	0,08
20	1,44%	38,86%	1,93	0,09

19	1,83%	40,18%	2,92	0,12
4	0,03%	42,63%	1,98	0,09
2	3,10%	53,08%	1,95	0,09
6	2,78%	60,73%	2,83	0,14
14	0,02%	99,93%	1,06	0,12
23	1,61%	121,03%	1,02	0,10
28	0,12%	127,19%	1,45	0,13
3	3,74%	146,95%	0,84	0,09
26	1,16%	247,65%	0,58	0,10

En este caso, hay cuatro clusters que parecen interesantes debido a su baja variabilidad externa, es decir, que los elementos del cluster tienen una rentabilidad a futuro en el tamaño de ventana escogido similar. El siguiente filtro es el de rentabilidad, que se ha marcado en un mínimo del 20% de crecimiento anual y que los cuatro clusters cumplen. Sin embargo, al comprobar la representatividad del cluster, el cluster 14 (marcado con letras verdes pero color de fila normal) no es suficientemente representativo, pues un 0,02% es un porcentaje muy bajo de los datos (el límite se ha situado cerca del 1% para considerar que un cluster es significativo).

Estos clusters interesantes aparecen señalados en color verde. Así, se ha pasado a la fase de validación, probando los clusters interesantes y obteniendo los siguientes resultados:

Tabla 6. Resultados validación Caso III

Cluster	Members	Expected_growth	External_variation	Internal_variation
23	0,00%	NaN	NaN	NaN
3	0,00%	NaN	NaN	NaN
26	0,00%	NaN	NaN	NaN

Como se puede observar, no hay datos en el conjunto de validación que encajen en estos clusters, y por eso aparecen los valores NaN. De esta manera, en el Caso III tampoco se han apreciado clusters interesantes o nichos de inversión objetivo.

4.2.4 Caso IV: tamaño de ventana 150 y agrupación en 30 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 7. Resultados entrenamiento Caso IV

Cluster	Members	Expected_growth	External_variation	Internal_variation
23	8,53%	-10,88%	1,85	0,06
22	3,59%	-9,14%	2,54	0,06
12	17,38%	-7,26%	3,64	0,06

3	0,20%	-5,82%	4,37	0,06
1	4,75%	-5,13%	5,66	0,07
8	0,14%	-2,79%	11,11	0,07
18	0,01%	0,07%	438,19	0,07
24	3,67%	2,45%	16,98	0,09
14	4,27%	2,50%	14,95	0,08
25	0,61%	3,61%	10,89	0,08
5	1,76%	4,62%	7,87	0,08
10	10,04%	5,99%	10,63	0,12
9	0,41%	10,04%	5,39	0,10
2	7,29%	10,22%	4,73	0,09
30	0,30%	14,05%	3,94	0,10
27	0,02%	14,82%	3,89	0,10
11	0,01%	18,51%	5,04	0,13
26	8,99%	20,95%	3,37	0,11
4	1,12%	49,30%	1,75	0,14
19	0,03%	51,66%	1,89	0,13
29	0,03%	54,19%	1,87	0,11
28	9,08%	58,61%	1,34	0,11
20	3,37%	68,36%	1,15	0,09
7	2,38%	77,36%	0,13	0,06
15	6,19%	91,92%	1,91	0,13
6	1,16%	116,42%	1,11	0,13
21	1,17%	132,02%	0,72	0,12
17	0,27%	143,27%	0,42	0,07
16	0,17%	150,44%	0,39	0,10
13	3,05%	169,05%	1,27	0,12

Se han marcado en verde los clusters cuya representatividad, rentabilidad media y variabilidades son mejores que los filtros explicados en el apartado 4.1. De esta manera, se han probado los clusters que pueden ser objetivo con los datos de validación, obteniendo los siguientes resultados:

Tabla 8. Resultados validación Caso IV

Cluster	Members	Expected_growth	External_variation	Internal_variation
20	0,86%	49,23%	2,03	0,09
7	0,57%	59,69%	2,37	0,32
6	1,71%	103,79%	1,47	0,32
21	1,29%	154,30%	0,87	1,59

Es muy importante la repetitividad en los resultados de los diferentes conjuntos de datos. De esta manera, se puede observar que dos clusters (marcados en rojo) han sido descartados por

el incremento en la variación externa, que además han venido acompañados de una baja representatividad.

Por último, se ha probado el conjunto de datos de test en los dos clusters que continúan pareciendo interesantes tras la fase de validación, obteniendo la siguiente tabla:

Tabla 9. Resultados test Caso IV

Cluster	Members	Expected_growth	External_variation	Internal_variation
6	2,41%	290,02%	0,98	0,33
21	0,01%	68,43%	1,71	0,58

El cluster 21 se ha descartado por su aumento en la variación externa y su disminución en la representatividad. De esta manera, del Caso IV se ha obtenido un cluster objetivo, el 6, demostrando la existencia de este tipo de resultados y respondiendo a una de las principales cuestiones que se plantean al principio de la memoria y que se expone en más profundidad en el capítulo de conclusiones.

4.2.5 Caso V: tamaño de ventana 30 y agrupación en 50 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 10. Resultados entrenamiento Caso V

Cluster	Members	Expected_growth	External_variation	Internal_variation
30	1,58%	-12,31%	4,91	0,03
33	18,29%	-10,94%	5,69	0,03
16	0,13%	-9,77%	6,58	0,03
27	0,50%	-8,60%	8,01	0,03
2	0,21%	-8,46%	7,48	0,03
31	0,03%	-6,73%	10,08	0,03
42	0,21%	-6,45%	10,75	0,03
8	1,10%	-6,25%	11,84	0,03
47	0,46%	-6,14%	10,27	0,03
13	0,71%	-5,25%	15,09	0,03
28	4,43%	-4,66%	17,04	0,03
11	0,96%	-3,69%	20,60	0,03
19	3,38%	-1,24%	60,31	0,03
45	0,03%	1,32%	57,25	0,03
29	0,54%	1,92%	49,10	0,04
44	5,93%	2,31%	43,51	0,04
21	0,11%	2,97%	27,25	0,03
43	0,17%	3,78%	23,86	0,04
41	3,19%	4,91%	17,47	0,03
36	0,08%	8,82%	8,47	0,03

23	2,90%	9,87%	10,14	0,04
1	0,10%	11,07%	8,52	0,04
15	2,26%	11,34%	9,91	0,04
46	1,11%	14,66%	6,62	0,04
37	0,50%	16,97%	5,94	0,04
10	0,40%	19,90%	5,38	0,04
40	2,90%	23,31%	4,12	0,04
25	2,52%	23,89%	6,60	0,06
24	1,49%	24,61%	4,70	0,05
32	5,92%	26,06%	6,15	0,06
38	1,48%	26,49%	4,09	0,04
49	0,42%	26,66%	5,87	0,07
50	4,36%	27,32%	5,07	0,05
26	0,02%	29,58%	3,89	0,05
12	0,03%	32,01%	3,78	0,05
4	4,67%	32,53%	4,00	0,05
5	0,55%	33,08%	5,37	0,06
22	1,19%	36,41%	5,70	0,08
9	0,12%	45,97%	2,97	0,05
7	1,99%	59,47%	2,30	0,06
17	1,93%	59,96%	4,64	0,10
48	5,32%	60,68%	4,29	0,10
39	0,41%	63,84%	2,51	0,05
20	0,89%	69,81%	3,54	0,08
18	1,61%	90,39%	2,05	0,07
3	2,86%	96,65%	2,22	0,07
35	8,84%	118,35%	2,68	0,08
34	0,05%	139,70%	2,46	0,10
14	0,31%	146,03%	1,37	0,07
6	0,84%	197,87%	1,66	0,09

Igual que en los dos primeros casos, todos los clusters han sido descartados en la fase de entrenamiento por ser la variación externa superior a uno en todos los casos. Este resultado también se puede explicar por la sobregeneralización que supone un tamaño de ventana de

4.2.6 Caso VI: tamaño de ventana 50 y agrupación en 50 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 11. Resultados entrenamiento Caso VI

Cluster	Members	Expected_growth	External_variation	Internal_variation
12	3,82%	-8,86%	5,30	0,04
26	1,12%	-8,29%	5,67	0,04
30	29,08%	-8,26%	5,85	0,04
1	4,15%	-7,90%	6,62	0,04

3	0,22%	-6,77%	7,03	0,04
42	0,56%	-6,06%	9,32	0,04
13	1,46%	-5,96%	9,02	0,04
29	0,07%	-4,72%	11,02	0,04
18	0,11%	-3,14%	17,96	0,04
21	2,63%	-2,84%	19,50	0,04
11	1,48%	-2,76%	22,84	0,05
20	3,28%	-1,48%	37,17	0,04
7	2,26%	-0,90%	65,24	0,04
32	0,39%	0,50%	141,54	0,05
15	6,77%	2,94%	18,87	0,04
48	0,02%	3,87%	15,43	0,04
2	0,17%	4,83%	13,63	0,04
24	4,25%	4,95%	15,90	0,05
33	0,90%	5,45%	12,59	0,05
19	3,23%	9,40%	8,60	0,05
38	2,12%	9,46%	7,15	0,04
4	1,15%	10,77%	7,79	0,05
31	0,04%	11,35%	6,06	0,05
49	1,33%	11,47%	7,74	0,06
6	1,61%	13,51%	7,33	0,06
25	3,77%	14,69%	5,17	0,05
22	0,02%	15,37%	5,65	0,06
45	0,08%	16,57%	5,14	0,05
10	2,51%	18,51%	4,86	0,05
39	2,04%	22,89%	4,24	0,06
35	0,98%	23,19%	3,81	0,05
41	0,70%	23,84%	4,00	0,06
47	1,90%	24,15%	4,49	0,06
50	0,39%	26,92%	4,52	0,07
40	0,93%	35,10%	3,63	0,07
34	0,09%	41,75%	2,41	0,07
36	0,02%	43,03%	2,72	0,09
46	0,60%	50,84%	3,59	0,09
14	1,05%	53,19%	2,56	0,07
5	0,30%	57,37%	1,97	0,08
17	0,54%	60,25%	2,84	0,09
28	5,85%	62,96%	3,46	0,12
9	0,07%	64,26%	2,44	0,08
8	0,05%	69,21%	3,02	0,10
27	2,11%	72,19%	2,38	0,09
37	0,15%	80,58%	1,68	0,11
43	1,13%	115,48%	2,04	0,11
44	1,50%	132,74%	1,61	0,09
16	0,52%	210,93%	1,20	0,11
23	0,47%	232,93%	1,52	0,10

Para este caso no se ha encontrado ningún cluster que cumpla los requisitos mínimos de representación, rentabilidad y variabilidades, por lo tanto no se ha pasado a la fase de validación.

4.2.7 Caso VII: tamaño de ventana 100 y agrupación en 50 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 12. Resultados entrenamiento Caso VII

Cluster	Members	Expected_growth	External_variation	Internal_variation
43	4,04%	-8,46%	3,79	0,05
12	4,99%	-6,72%	4,86	0,05
47	1,16%	-6,62%	5,14	0,05
8	1,24%	-5,55%	5,97	0,05
44	0,01%	-5,35%	6,20	0,05
38	0,30%	-5,14%	7,62	0,06
20	1,99%	-4,38%	8,43	0,06
2	2,85%	-4,24%	8,85	0,06
31	5,26%	-3,86%	9,60	0,06
1	1,29%	-2,02%	19,42	0,06
21	0,54%	0,00%	9495,10	0,07
10	7,64%	0,61%	79,59	0,07
24	0,13%	2,58%	17,41	0,06
41	0,01%	2,94%	16,59	0,06
18	0,21%	3,04%	14,23	0,06
25	0,45%	3,23%	13,50	0,06
29	0,53%	4,07%	11,20	0,07
26	0,57%	5,80%	7,42	0,06
9	0,05%	6,58%	9,36	0,08
4	4,16%	9,05%	6,13	0,07
48	4,40%	9,34%	5,95	0,07
37	0,14%	9,59%	7,06	0,09
23	2,21%	10,80%	5,30	0,07
6	2,52%	12,74%	5,15	0,09
3	1,20%	16,33%	3,23	0,07
27	5,36%	16,76%	3,95	0,08
7	0,63%	17,36%	3,32	0,07
32	0,02%	22,07%	2,78	0,07
40	2,11%	23,55%	2,51	0,08
39	0,08%	23,63%	3,24	0,09
35	2,00%	28,00%	4,23	0,13
11	0,84%	33,62%	2,11	0,08
15	0,07%	37,06%	2,27	0,08
34	0,18%	40,53%	2,09	0,09

45	0,41%	41,38%	2,43	0,10
16	0,16%	46,44%	1,98	0,09
17	1,32%	48,29%	3,07	0,13
49	2,68%	48,43%	1,72	0,10
36	0,92%	58,97%	1,46	0,08
42	0,62%	68,33%	2,02	0,12
22	0,50%	90,60%	1,05	0,08
13	0,30%	99,06%	1,07	0,12
19	5,60%	111,97%	1,02	0,11
46	21,30%	112,66%	1,15	0,09
33	0,37%	125,63%	1,70	0,13
28	0,16%	134,72%	0,66	0,11
5	5,83%	145,54%	0,51	0,11
30	0,29%	162,68%	1,18	0,13
50	0,32%	175,81%	0,79	0,10
14	0,04%	244,12%	0,31	0,11

Tras analizar los clusters señalados en verde con los datos de validación, se han obtenido los siguientes resultados:

Tabla 13. Resultados validación Caso VII

Cluster	Members	Expected_growth	External_variation	Internal_variation
19	25,80%	346,71%	0,53	0,19
46	0,00%	NaN	NaN	NaN
5	0,00%	NaN	NaN	NaN

De esta manera, solamente un cluster (19) será probado con los datos de test para comprobar su validez:

Tabla 14. Resultados test Caso VII

Cluster	Members	Expected_growth	External_variation	Internal_variation
19	2,13%	7,59%	11,38	0,00

El cluster se ha descartado por su importante incremento en la variación externa, aunque tampoco cumpliría los requisitos de rentabilidad mínima y repetitividad.

4.2.8 Caso VIII: tamaño de ventana 150 y agrupación en 50 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 15. Resultados entrenamiento Caso VIII

Cluster	Members	Expected_growth	External_variation	Internal_variation

32	0,72%	-10,95%	1,82	0,06
44	1,71%	-9,27%	2,49	0,06
20	3,36%	-8,10%	3,15	0,06
36	0,03%	-7,72%	4,01	0,08
39	0,21%	-7,11%	3,76	0,07
27	0,15%	-6,65%	4,25	0,07
2	0,64%	-5,85%	4,85	0,07
22	0,64%	-5,79%	4,60	0,06
34	0,01%	-5,62%	4,95	0,07
28	3,79%	-5,05%	5,15	0,06
10	0,76%	-4,31%	7,08	0,07
29	1,92%	-4,27%	6,85	0,07
14	0,01%	-3,63%	8,39	0,07
41	3,01%	-0,18%	187,80	0,07
15	0,26%	1,05%	42,44	0,11
46	0,00%	1,84%	18,94	0,07
23	1,71%	2,68%	13,52	0,08
3	0,03%	4,57%	7,93	0,08
40	1,04%	5,87%	7,13	0,08
43	5,93%	6,59%	6,71	0,09
31	0,62%	7,47%	6,32	0,09
7	4,44%	7,85%	7,15	0,10
17	3,13%	8,71%	5,18	0,09
12	0,68%	9,49%	4,69	0,09
11	0,07%	10,61%	5,08	0,10
35	0,29%	12,22%	4,32	0,10
49	1,17%	14,53%	4,23	0,10
30	17,28%	16,13%	4,79	0,12
19	5,69%	16,47%	5,56	0,12
33	0,39%	17,58%	3,54	0,10
24	3,06%	19,11%	3,07	0,11
1	3,08%	20,12%	3,27	0,11
45	1,73%	26,70%	4,10	0,14
8	0,63%	36,23%	2,02	0,13
6	1,77%	53,48%	1,88	0,11
26	0,73%	59,13%	1,40	0,13
21	0,02%	61,83%	1,83	0,13
50	0,42%	73,39%	1,35	0,13
47	0,69%	76,04%	0,13	0,06
16	2,84%	80,51%	0,16	0,06
25	4,42%	83,31%	2,70	0,12
4	0,01%	84,06%	0,94	0,09
42	2,28%	87,36%	0,14	0,06
13	8,81%	88,28%	1,46	0,14
48	0,55%	106,00%	1,14	0,13
5	6,97%	113,38%	1,51	0,13

18	0,01%	122,51%	1,08	0,13
9	0,14%	149,22%	0,55	0,09
38	2,11%	153,86%	1,37	0,12
37	0,03%	159,49%	0,32	0,10

Tras analizar los clusters señalados en verde con los datos de validación, se han obtenido los siguientes resultados:

Tabla 16. Resultados validación Caso VIII

Cluster	Members	Expected_growth	External_variation	Internal_variation
16	7,77%	89,21%	1,49	0,01
42	1,87%	203,81%	1,13	0,01

El cluster 16 se ha seguido analizando ya que aunque ha aumentado su variación externa, sería un valor en el límite de decisión. De esta manera, solamente dos clusters (16 y 42) serán probados con los datos de test para comprobar su validez:

Tabla 17. Resultados test Caso VIII

Cluster	Members	Expected_growth	External_variation	Internal_variation
42	0,00%	NaN	NaN	NaN
47	0,00%	NaN	NaN	NaN

Los clusters se han descartado porque los datos de test no se corresponden con ninguno de los clusters, de ahí los NaN.

4.2.9 Caso IX: tamaño de ventana 30 y agrupación en 100 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 18. Resultados entrenamiento Caso IX

Cluster	Members	Expected_growth	External_variation	Internal_variation
39	1,26%	-279,30%	1,19	0,23
11	0,26%	-12,72%	4,84	0,03
84	2,27%	-11,52%	5,25	0,03
76	0,15%	-10,95%	5,64	0,03
3	0,02%	-9,70%	6,58	0,03
50	0,08%	-9,69%	6,58	0,03
51	1,29%	-9,57%	7,51	0,03
12	0,27%	-9,19%	7,09	0,03
53	1,30%	-8,42%	7,91	0,03

59	1,63%	-7,96%	7,78	0,03
79	3,26%	-7,86%	9,68	0,03
60	2,67%	-7,30%	9,51	0,03
20	0,16%	-6,83%	9,56	0,03
32	0,09%	-6,12%	11,27	0,03
38	0,04%	-6,01%	12,29	0,03
7	0,73%	-5,99%	13,40	0,03
94	1,55%	-5,79%	11,33	0,03
86	1,99%	-5,21%	15,12	0,03
21	0,21%	-4,94%	14,69	0,03
1	5,06%	-4,72%	15,14	0,03
42	2,70%	-4,57%	17,89	0,03
65	1,06%	-1,99%	45,98	0,04
80	0,69%	-0,81%	93,42	0,03
10	0,11%	0,09%	825,68	0,03
49	0,06%	1,20%	67,18	0,03
91	0,16%	1,74%	51,12	0,04
75	0,20%	1,92%	44,12	0,03
95	0,33%	2,02%	37,69	0,03
22	0,49%	2,34%	35,02	0,03
9	0,16%	4,10%	19,59	0,03
55	1,10%	4,60%	19,03	0,04
90	3,86%	5,05%	16,45	0,03
98	1,93%	5,16%	16,62	0,03
74	0,05%	5,36%	16,12	0,03
18	0,00%	5,38%	13,30	0,03
46	0,19%	5,51%	18,58	0,04
56	0,85%	6,40%	14,26	0,04
29	0,95%	6,97%	15,21	0,04
97	0,00%	7,12%	12,78	0,04
28	0,00%	7,63%	13,21	0,04
99	0,02%	7,88%	9,22	0,03
78	1,91%	8,35%	12,01	0,04
23	0,23%	9,52%	9,37	0,04
70	0,17%	9,82%	10,24	0,04
58	0,08%	11,47%	9,07	0,04
54	0,61%	11,61%	8,84	0,04
8	0,44%	13,07%	10,23	0,05
81	0,45%	13,65%	8,47	0,04
27	1,31%	15,02%	8,68	0,05
47	1,91%	15,22%	6,23	0,04
66	2,73%	17,95%	5,84	0,04
67	0,04%	18,07%	5,60	0,04
33	2,60%	18,23%	5,04	0,04
88	0,17%	18,74%	6,08	0,04
82	0,72%	18,88%	6,99	0,05

87	0,11%	19,25%	8,99	0,06
37	0,07%	20,54%	4,10	0,03
71	0,29%	20,89%	5,17	0,04
100	15,86%	21,61%	4,92	0,04
92	1,56%	22,65%	4,86	0,04
62	0,02%	23,19%	11,38	0,11
48	0,03%	23,88%	4,82	0,04
15	0,02%	23,97%	6,96	0,07
16	0,18%	24,43%	5,20	0,05
26	0,25%	24,99%	5,77	0,05
44	0,50%	25,97%	5,63	0,06
17	0,52%	26,61%	4,04	0,04
77	0,19%	27,56%	4,32	0,05
31	0,10%	27,71%	3,59	0,04
4	0,25%	30,18%	3,87	0,05
13	0,30%	31,09%	3,79	0,05
30	0,14%	32,17%	3,34	0,05
93	0,00%	33,93%	5,71	0,08
64	1,10%	35,98%	4,72	0,06
2	0,66%	36,17%	4,14	0,06
73	4,38%	37,99%	8,65	0,14
68	0,45%	38,64%	3,57	0,05
69	0,60%	41,06%	2,79	0,05
36	1,01%	41,68%	3,33	0,05
25	1,83%	43,44%	4,22	0,07
19	0,21%	45,88%	3,06	0,05
14	1,16%	46,36%	2,59	0,05
96	0,01%	47,74%	2,86	0,05
45	5,25%	53,09%	4,63	0,09
43	0,05%	57,38%	3,09	0,06
24	0,94%	63,06%	2,48	0,06
52	0,15%	69,55%	2,63	0,06
6	0,13%	72,93%	2,01	0,06
72	0,03%	75,54%	3,23	0,07
34	1,52%	78,33%	1,85	0,06
85	0,54%	85,40%	3,33	0,10
41	0,56%	116,18%	2,22	0,09
57	0,21%	118,16%	1,89	0,07
5	1,17%	142,43%	1,26	0,07
89	1,91%	147,04%	1,54	0,07
61	0,20%	171,20%	2,27	0,10
40	0,66%	181,64%	1,46	0,07
63	0,76%	184,16%	1,92	0,09
83	1,27%	184,22%	0,74	0,08
35	0,28%	399,17%	0,85	0,13

Tras analizar los clusters señalados en verde con los datos de validación, se han obtenido los siguientes resultados:

Tabla 19. Resultados entrenamiento Caso IX

Cluster	Members	Expected_growth	External_variation	Internal_variation
39	0,00%	-6,54%	12,40	0,00
5	0,00%	NaN	NaN	NaN
83	0,00%	NaN	NaN	NaN

Al haber sido descartados todos los clusters, no se ha procedido a la fase de test.

4.2.10 Caso X: tamaño de ventana 50 y agrupación en 100 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 20. Resultados entrenamiento Caso X

Cluster	Members	Expected_growth	External_variation	Internal_variation
31	2,96%	-9,37%	5,29	0,04
91	1,52%	-9,21%	5,06	0,04
49	0,11%	-9,21%	5,55	0,04
16	2,38%	-9,04%	5,06	0,04
95	0,97%	-8,28%	5,79	0,04
4	2,18%	-8,08%	6,10	0,04
99	1,59%	-7,80%	6,23	0,04
5	0,70%	-7,74%	6,50	0,04
47	0,02%	-7,67%	6,06	0,03
32	0,10%	-7,66%	7,03	0,04
73	0,06%	-6,93%	7,67	0,04
2	0,58%	-6,65%	8,73	0,04
61	0,01%	-6,55%	8,34	0,04
65	1,45%	-6,10%	8,59	0,04
86	0,07%	-6,00%	8,66	0,04
83	5,56%	-5,95%	8,31	0,04
70	0,55%	-5,86%	8,58	0,04
8	0,28%	-5,19%	10,47	0,04
60	0,25%	-4,85%	11,14	0,04
40	2,96%	-4,71%	13,08	0,04
97	0,01%	-4,23%	16,06	0,05
69	0,55%	-4,18%	13,62	0,04
66	0,11%	-4,06%	14,71	0,04
44	0,25%	-2,42%	23,30	0,04
84	0,46%	-1,90%	26,59	0,04
93	0,12%	-1,11%	62,15	0,05
18	0,51%	-1,07%	59,60	0,05
58	0,10%	-0,97%	52,71	0,04

85	1,79%	-0,85%	68,24	0,04
94	0,14%	-0,49%	143,93	0,05
17	1,38%	0,90%	63,35	0,04
75	0,50%	1,42%	39,74	0,04
53	0,46%	1,51%	43,71	0,04
92	0,29%	2,35%	28,97	0,05
1	0,03%	2,51%	25,45	0,04
41	0,03%	2,63%	26,79	0,05
79	0,70%	3,04%	19,10	0,04
29	1,37%	3,30%	17,43	0,04
56	0,95%	3,71%	24,28	0,06
45	0,83%	3,78%	20,99	0,05
37	0,76%	3,81%	20,57	0,05
20	1,00%	4,27%	13,04	0,04
62	0,71%	4,56%	16,86	0,05
88	1,83%	6,92%	9,53	0,04
7	0,25%	7,21%	9,71	0,05
27	0,07%	9,39%	8,59	0,05
39	18,84%	9,95%	6,72	0,04
6	0,17%	12,19%	5,63	0,04
82	1,50%	12,33%	6,69	0,05
52	0,38%	12,77%	5,90	0,05
71	0,03%	13,21%	7,85	0,06
43	1,57%	14,33%	5,87	0,05
72	1,13%	14,83%	6,27	0,06
34	0,00%	14,95%	5,99	0,06
42	0,10%	15,47%	4,96	0,05
14	0,20%	17,68%	4,92	0,05
55	0,26%	18,00%	5,13	0,06
74	0,79%	18,92%	5,06	0,06
24	0,00%	19,55%	4,96	0,06
22	1,73%	20,45%	4,06	0,05
48	1,31%	21,92%	5,24	0,06
12	0,31%	22,41%	4,26	0,06
57	0,03%	23,64%	3,79	0,06
38	0,19%	24,73%	4,49	0,06
80	0,86%	24,96%	3,73	0,05
100	0,72%	25,62%	4,17	0,06
25	0,01%	25,65%	3,53	0,05
68	0,13%	30,91%	3,44	0,07
33	1,15%	30,99%	3,94	0,07
26	3,24%	32,45%	3,30	0,06
87	0,13%	33,20%	3,32	0,06
50	0,34%	33,81%	3,75	0,07
98	0,44%	34,93%	3,89	0,07
78	1,68%	35,14%	3,11	0,06

64	2,90%	36,40%	2,78	0,06
51	0,08%	38,59%	2,38	0,09
59	0,07%	40,84%	2,19	0,13
96	0,07%	41,21%	3,92	0,08
23	0,60%	41,61%	3,78	0,07
19	0,27%	44,64%	2,36	0,07
11	0,02%	48,08%	2,44	0,07
28	0,58%	51,68%	2,17	0,07
13	1,41%	56,21%	1,81	0,12
21	4,03%	57,69%	1,94	0,10
90	0,92%	60,64%	3,27	0,09
76	1,20%	61,12%	2,43	0,07
10	0,33%	61,81%	1,83	0,08
46	1,20%	65,94%	1,98	0,07
3	0,03%	66,14%	2,50	0,08
63	0,10%	70,58%	2,13	0,08
36	3,25%	76,01%	1,52	0,10
54	1,30%	83,06%	1,96	0,08
30	0,60%	85,08%	2,49	0,11
67	0,29%	90,84%	1,50	0,10
77	1,74%	92,25%	2,56	0,09
89	0,13%	99,35%	1,93	0,11
81	0,33%	108,09%	2,73	0,13
9	0,33%	140,04%	1,68	0,10
15	1,29%	170,43%	1,72	0,10
35	0,22%	187,70%	1,46	0,12

Para este caso no se ha encontrado ningún cluster que cumpla los requisitos mínimos de representación, rentabilidad y variabilidades, por lo tanto no se ha pasado a la fase de validación.

4.2.11 Caso XI: tamaño de ventana 100 y agrupación en 100 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 21. Resultados entrenamiento Caso XI

Cluster	Members	Expected_growth	External_variation	Internal_variation
20	18,04%	-8,09%	4,13	0,05
69	0,23%	-7,42%	4,44	0,05
18	0,01%	-7,39%	4,44	0,05
86	0,09%	-6,69%	5,00	0,05
62	2,03%	-6,41%	5,61	0,05
1	0,58%	-6,07%	5,25	0,05
14	0,15%	-6,00%	5,57	0,05
41	0,00%	-5,44%	6,59	0,05

92	0,19%	-5,20%	6,56	0,05
68	0,39%	-4,99%	7,73	0,06
78	0,06%	-4,37%	7,93	0,06
65	0,73%	-4,34%	8,42	0,06
49	0,31%	-4,32%	8,14	0,05
96	0,99%	-4,26%	7,88	0,05
44	0,98%	-4,19%	8,67	0,06
55	0,01%	-3,41%	11,84	0,06
51	1,33%	-3,24%	11,55	0,06
5	3,91%	-1,91%	21,11	0,06
26	0,63%	-1,59%	25,58	0,06
83	2,55%	-1,45%	28,06	0,06
87	0,03%	-1,03%	38,81	0,06
63	0,30%	-0,97%	37,77	0,06
57	0,28%	-0,58%	78,89	0,07
45	0,61%	-0,57%	72,71	0,06
22	0,08%	0,19%	237,77	0,07
15	3,25%	0,28%	146,68	0,06
56	0,02%	0,80%	53,17	0,06
93	0,01%	1,17%	37,87	0,06
60	1,86%	1,20%	36,22	0,06
17	1,67%	2,43%	20,64	0,07
43	0,36%	2,72%	14,93	0,06
75	0,45%	2,81%	16,80	0,07
10	0,93%	3,30%	15,01	0,07
38	0,65%	3,57%	14,46	0,07
70	0,46%	3,72%	16,29	0,08
34	0,02%	4,93%	9,75	0,06
95	0,03%	6,72%	8,33	0,08
6	0,75%	7,36%	7,13	0,08
79	0,24%	7,38%	7,64	0,07
12	0,03%	7,38%	6,26	0,06
42	3,71%	7,48%	5,98	0,07
73	0,24%	7,61%	7,40	0,07
90	1,67%	7,68%	6,64	0,07
94	4,30%	7,78%	5,46	0,06
84	1,69%	8,62%	7,81	0,09
59	0,04%	10,06%	4,88	0,07
50	0,75%	10,31%	4,68	0,07
53	0,13%	11,22%	4,60	0,07
29	0,55%	11,94%	3,99	0,07
67	0,45%	12,55%	4,90	0,08
74	1,31%	13,56%	4,23	0,08
61	0,25%	14,97%	6,29	0,12
85	0,50%	16,19%	3,64	0,07
32	0,27%	16,87%	3,19	0,07

88	1,10%	17,41%	4,16	0,08
66	0,37%	17,59%	3,48	0,08
47	1,10%	17,81%	4,68	0,10
19	0,41%	19,76%	5,04	0,11
2	1,07%	19,91%	3,09	0,08
48	1,70%	20,02%	3,60	0,09
30	0,07%	20,63%	2,86	0,07
24	2,28%	21,36%	2,64	0,07
58	0,82%	25,79%	3,00	0,09
31	0,08%	26,69%	4,36	0,13
99	1,12%	26,85%	2,71	0,09
54	1,23%	27,01%	3,24	0,09
35	0,77%	30,06%	3,78	0,12
33	1,58%	30,58%	2,27	0,08
77	4,38%	32,01%	1,83	0,08
23	0,97%	35,38%	2,68	0,11
39	0,14%	37,53%	2,19	0,09
7	0,00%	45,09%	2,44	0,11
100	0,41%	46,63%	1,21	0,09
4	0,36%	51,89%	1,99	0,10
25	0,13%	54,35%	1,16	0,09
13	0,13%	55,59%	1,84	0,10
76	0,15%	62,94%	1,69	0,10
21	0,15%	68,11%	1,24	0,11
52	0,28%	70,09%	1,48	0,09
72	0,00%	70,57%	0,25	0,10
81	0,01%	88,16%	1,13	0,09
91	0,02%	91,75%	0,63	0,10
80	2,42%	93,45%	0,64	0,15
82	1,44%	100,95%	1,33	0,15
40	0,28%	101,80%	1,10	0,12
97	2,87%	101,89%	0,99	0,12
9	1,11%	105,61%	1,05	0,08
89	0,17%	114,00%	0,88	0,10
46	0,04%	114,51%	0,89	0,13
71	0,62%	116,26%	1,73	0,13
64	0,02%	124,66%	1,44	0,12
36	4,76%	126,22%	0,48	0,10
11	0,73%	129,05%	1,74	0,13
16	0,78%	136,28%	1,16	0,09
27	1,47%	147,01%	0,84	0,10
3	1,23%	206,11%	0,38	0,10
37	0,05%	207,01%	0,91	0,14
28	0,01%	257,75%	0,21	0,11
98	0,07%	379,54%	0,53	0,08
8	0,02%	415,69%	0,37	0,11

Tras analizar los clusters señalados en verde con los datos de validación, se han obtenido los siguientes resultados:

Tabla 22. Resultados validación Caso XI

Cluster	Members	Expected_growth	External_variation	Internal_variation
80	0,00%	NaN	NaN	NaN
97	0,00%	NaN	NaN	NaN
9	0,00%	NaN	NaN	NaN
36	0,00%	NaN	NaN	NaN
27	0,00%	NaN	NaN	NaN
3	0,00%	NaN	NaN	NaN

Para este caso no se ha encontrado ningún cluster que cumpla los requisitos mínimos de representación, rentabilidad y variabilidades, por lo tanto no se ha pasado a la fase de validación.

4.2.12 Caso XII: tamaño de ventana 100 y agrupación en 100 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 23. Resultados entrenamiento Caso XII

Cluster	Members	Expected_growth	External_variation	Internal_variation
3	4,80%	-17,61%	2,92	0,13
97	0,21%	-11,24%	1,61	0,06
24	0,11%	-10,74%	1,89	0,06
43	0,53%	-8,47%	2,94	0,06
10	1,60%	-8,17%	3,20	0,07
1	0,61%	-7,89%	3,06	0,06
99	0,01%	-7,85%	4,03	0,08
72	0,01%	-7,23%	3,38	0,06
98	1,24%	-6,92%	3,90	0,06
68	0,24%	-6,89%	3,88	0,07
46	0,19%	-5,74%	4,96	0,07
25	0,41%	-5,52%	4,80	0,07
57	5,32%	-5,43%	4,60	0,06
49	0,03%	-5,09%	5,87	0,07
95	8,16%	-4,32%	6,60	0,07
4	0,39%	-3,89%	7,57	0,07
39	0,04%	-3,37%	10,05	0,08
15	1,63%	-2,84%	10,01	0,07
77	0,27%	-1,70%	19,15	0,07
23	0,32%	-1,69%	18,75	0,07

48	0,23%	-0,05%	669,48	0,07
81	0,00%	0,18%	183,58	0,08
54	0,26%	0,61%	61,02	0,07
13	2,23%	0,73%	47,76	0,07
6	1,42%	1,29%	31,32	0,08
83	0,07%	1,56%	24,26	0,07
76	0,12%	1,59%	21,11	0,07
93	0,05%	1,67%	19,46	0,07
91	0,17%	2,46%	16,32	0,09
86	1,11%	2,58%	13,93	0,08
29	0,02%	3,45%	11,93	0,09
36	0,19%	3,56%	11,00	0,09
92	0,27%	4,49%	8,95	0,08
16	0,19%	4,58%	8,20	0,08
20	0,06%	5,51%	10,72	0,10
34	0,35%	5,56%	7,81	0,09
79	0,10%	5,77%	7,35	0,08
18	0,43%	6,03%	6,17	0,08
66	0,89%	7,28%	5,27	0,08
55	0,28%	7,73%	5,90	0,10
17	0,63%	8,23%	8,86	0,12
9	1,22%	8,59%	5,66	0,09
88	6,73%	8,90%	5,96	0,10
42	0,99%	8,95%	5,77	0,09
27	0,04%	9,25%	7,75	0,13
84	1,16%	10,18%	4,55	0,10
70	0,14%	10,24%	5,06	0,10
100	0,82%	11,17%	4,38	0,08
5	5,06%	11,25%	4,26	0,09
71	0,01%	11,68%	4,91	0,10
47	0,05%	11,71%	3,34	0,11
19	0,01%	11,99%	4,49	0,09
41	0,01%	12,73%	4,71	0,10
33	1,46%	13,11%	4,31	0,11
80	0,41%	13,30%	3,52	0,09
74	0,00%	13,45%	3,39	0,09
40	8,07%	14,54%	3,67	0,12
44	0,10%	14,76%	6,52	0,13
2	0,12%	15,41%	4,09	0,11
85	0,00%	15,74%	4,70	0,10
12	0,02%	16,07%	3,22	0,10
32	0,18%	17,13%	4,77	0,14
63	0,22%	18,16%	3,72	0,11
30	0,43%	21,49%	3,24	0,11
89	0,01%	25,15%	3,28	0,12
64	1,93%	26,99%	2,68	0,11

38	0,12%	31,87%	2,46	0,12
96	2,95%	37,71%	1,23	0,10
75	0,01%	37,73%	2,19	0,15
28	1,35%	39,43%	2,49	0,12
87	0,35%	42,06%	1,78	0,13
37	2,93%	45,63%	2,04	0,12
94	0,02%	60,12%	1,57	0,14
90	0,45%	60,50%	1,95	0,14
58	0,08%	60,79%	1,29	0,12
35	4,07%	61,06%	1,36	0,14
67	1,57%	62,50%	1,31	0,11
60	0,04%	73,41%	0,15	0,06
56	0,54%	76,20%	0,15	0,06
11	0,84%	77,43%	1,92	0,13
22	0,23%	80,84%	0,07	0,05
7	0,09%	81,63%	0,14	0,06
61	0,02%	82,19%	1,04	0,13
78	0,89%	89,11%	0,81	0,12
31	0,15%	92,29%	1,46	0,14
82	1,61%	101,01%	1,63	0,14
26	0,28%	101,72%	1,24	0,12
73	1,29%	115,26%	1,01	0,13
14	0,66%	117,34%	0,91	0,09
21	0,34%	129,92%	1,59	0,13
50	0,39%	129,92%	0,46	0,06
51	2,88%	130,32%	1,03	0,13
53	0,58%	131,88%	0,86	0,09
8	0,14%	138,38%	0,50	0,09
69	1,27%	146,00%	0,37	0,08
45	0,01%	146,19%	0,85	0,14
62	3,23%	151,72%	1,27	0,11
52	4,94%	154,30%	0,28	0,11
59	0,77%	170,91%	1,47	0,13
65	0,56%	185,58%	0,26	0,09

Tras analizar los clusters señalados en verde con los datos de validación, se han obtenido los siguientes resultados:

Tabla 24. Resultados validación Caso XII

Cluster	Members	Expected_growth	External_variation	Internal_variation
96	5,71%	37,03%	2,60	0,19
73	0,00%	NaN	NaN	NaN
51	0,00%	NaN	NaN	NaN
69	0,00%	NaN	NaN	NaN
52	0,00%	NaN	NaN	NaN

Al haber sido descartados todos los clusters, no se ha procedido a la fase de test.

4.2.13 Caso XIII: tamaño de ventana 30 y agrupación en 150 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 25. Resultados entrenamiento Caso XIII

Cluster	Members	Expected_growth	External_variation	Internal_variation
87	1,87%	-33,01%	3,60	0,06
136	0,03%	-20,39%	17,92	0,15
93	0,27%	-12,39%	4,93	0,03
119	0,22%	-11,59%	5,23	0,03
39	0,81%	-11,58%	5,36	0,03
140	0,05%	-10,14%	6,40	0,03
61	0,83%	-10,06%	6,44	0,03
92	0,08%	-9,76%	6,48	0,03
62	0,03%	-9,68%	6,64	0,03
150	0,01%	-9,59%	6,66	0,03
43	0,23%	-9,59%	6,87	0,03
1	0,09%	-9,52%	6,63	0,03
7	0,50%	-9,36%	7,94	0,03
41	1,00%	-9,27%	6,88	0,03
34	8,20%	-8,67%	8,03	0,03
28	0,87%	-8,54%	7,86	0,03
27	0,16%	-8,45%	9,02	0,03
71	0,19%	-8,35%	7,91	0,03
69	0,20%	-8,29%	8,05	0,03
38	0,27%	-7,92%	9,74	0,03
105	0,09%	-7,48%	9,62	0,03
85	1,42%	-7,08%	8,90	0,03
35	0,04%	-6,32%	11,23	0,03
147	0,60%	-6,05%	11,36	0,03
137	0,07%	-5,90%	14,48	0,04
115	1,32%	-5,84%	12,51	0,03
65	1,08%	-5,76%	14,11	0,03
58	0,96%	-5,02%	15,54	0,03
40	0,01%	-4,43%	15,36	0,03
80	0,07%	-4,41%	16,53	0,03
15	0,62%	-3,95%	16,02	0,03
110	0,12%	-3,84%	18,91	0,03
94	0,00%	-3,79%	20,99	0,03
103	1,62%	-3,70%	22,42	0,04
22	1,55%	-3,51%	20,73	0,03
55	0,91%	-3,02%	25,95	0,03
67	0,04%	-1,81%	41,51	0,03

134	0,68%	-1,74%	54,88	0,04
122	2,08%	-1,27%	53,61	0,03
88	1,58%	-0,34%	233,76	0,03
97	0,60%	0,06%	1304,40	0,03
26	0,02%	0,99%	79,00	0,03
36	1,98%	1,52%	54,43	0,03
68	0,15%	1,83%	37,69	0,03
141	0,13%	2,68%	29,08	0,03
104	0,71%	3,01%	27,25	0,03
149	0,63%	3,13%	22,34	0,03
24	0,02%	3,41%	28,98	0,04
13	0,02%	3,61%	24,07	0,04
31	0,05%	4,03%	20,83	0,03
64	0,05%	4,22%	21,80	0,04
144	0,11%	4,64%	20,83	0,04
14	0,15%	4,65%	17,93	0,03
98	0,74%	4,81%	15,57	0,03
127	0,65%	5,70%	15,62	0,04
16	1,01%	5,73%	12,50	0,03
4	0,36%	6,04%	15,79	0,04
72	0,40%	7,07%	11,11	0,03
100	0,59%	7,10%	13,02	0,04
74	0,23%	7,57%	11,62	0,04
120	0,57%	8,16%	12,55	0,04
47	3,92%	8,38%	11,97	0,04
130	0,06%	9,11%	11,89	0,04
102	0,40%	9,64%	9,84	0,04
121	0,61%	10,73%	11,53	0,05
46	0,44%	10,78%	8,78	0,04
96	1,14%	11,92%	8,43	0,04
11	0,77%	12,06%	8,13	0,04
125	1,01%	12,79%	8,10	0,04
83	0,26%	12,85%	6,13	0,03
131	1,48%	13,25%	9,99	0,05
109	2,03%	13,36%	6,57	0,04
106	0,18%	14,02%	6,14	0,03
112	0,47%	15,03%	7,45	0,04
126	0,33%	15,04%	7,86	0,04
56	0,21%	15,63%	5,01	0,03
66	0,01%	15,82%	6,44	0,04
139	0,03%	16,46%	6,24	0,04
108	0,35%	16,51%	5,99	0,04
5	1,30%	18,58%	4,77	0,04
70	0,09%	18,70%	7,55	0,05
19	0,27%	19,54%	5,91	0,04
79	0,66%	21,06%	5,34	0,04

51	0,01%	21,80%	8,42	0,07
128	3,33%	21,93%	4,96	0,04
123	0,14%	24,03%	4,40	0,04
20	0,00%	24,09%	4,50	0,04
23	0,69%	24,42%	6,59	0,05
60	0,04%	24,58%	5,81	0,06
8	0,01%	24,87%	9,46	0,09
59	0,31%	25,18%	4,32	0,04
133	1,48%	26,32%	3,69	0,04
82	3,08%	26,51%	5,58	0,06
57	1,02%	27,06%	4,18	0,04
54	0,01%	27,26%	3,60	0,04
129	0,64%	28,54%	3,95	0,04
148	0,29%	28,55%	3,96	0,05
86	0,73%	28,73%	4,16	0,05
18	0,11%	28,89%	4,18	0,05
44	0,39%	30,22%	3,76	0,05
3	0,01%	31,22%	4,29	0,05
75	0,48%	31,27%	4,05	0,05
91	0,54%	31,67%	5,61	0,06
76	0,86%	32,33%	4,65	0,06
73	1,16%	33,44%	3,96	0,05
99	0,99%	34,10%	6,02	0,08
142	0,04%	36,03%	3,99	0,06
53	0,10%	36,63%	3,36	0,05
21	0,54%	37,09%	3,12	0,05
113	0,81%	38,52%	3,48	0,05
32	0,05%	40,75%	2,71	0,05
25	0,17%	40,92%	2,78	0,05
63	0,35%	41,40%	5,82	0,09
52	0,02%	41,78%	4,23	0,07
45	1,34%	43,42%	3,24	0,05
81	0,05%	44,66%	3,30	0,05
17	0,03%	45,08%	2,92	0,05
2	0,10%	52,15%	2,47	0,07
30	12,88%	53,41%	2,56	0,06
12	0,15%	53,54%	3,23	0,06
143	0,45%	53,84%	5,22	0,08
37	0,71%	53,88%	4,38	0,08
135	0,41%	54,39%	4,13	0,08
116	0,01%	60,45%	2,50	0,05
9	0,66%	66,41%	2,44	0,05
118	0,31%	67,17%	2,22	0,05
6	0,33%	68,85%	2,79	0,07
107	0,46%	70,32%	2,56	0,06
89	0,17%	76,66%	1,96	0,06

48	0,32%	76,67%	1,98	0,05
50	0,29%	84,29%	1,79	0,06
84	0,00%	87,07%	1,81	0,05
42	0,62%	92,48%	2,32	0,07
138	0,25%	97,00%	3,51	0,11
146	0,08%	97,07%	2,99	0,09
10	0,01%	97,18%	1,97	0,07
78	0,39%	98,40%	2,03	0,06
101	0,00%	118,21%	1,21	0,07
111	0,28%	142,01%	1,65	0,07
124	0,82%	154,49%	1,15	0,08
117	0,68%	159,22%	1,93	0,08
90	0,12%	161,05%	1,80	0,09
114	0,04%	169,81%	1,37	0,07
77	0,31%	177,63%	1,26	0,07
132	0,00%	185,00%	1,39	0,07
49	0,05%	195,27%	1,39	0,08
33	2,07%	201,17%	0,70	0,08
95	0,25%	206,46%	2,01	0,10
29	0,29%	271,61%	1,74	0,11
145	0,74%	395,32%	1,02	0,14

Tras analizar los clusters señalados en verde con los datos de validación, se han obtenido los siguientes resultados:

Tabla 26. Resultados entrenamiento Caso XIII

Cluster	Members	Expected_growth	External_variation	Internal_variation
101	0,00%	NaN	NaN	NaN
124	0,00%	NaN	NaN	NaN
33	0,00%	NaN	NaN	NaN
145	0,00%	NaN	NaN	NaN

Al haber sido descartados todos los clusters, no se ha procedido a la fase de test.

4.2.14 Caso XIV: tamaño de ventana 50 y agrupación en 150 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 27. Resultados entrenamiento Caso XIV

Cluster	Members	Expected_growth	External_variation	Internal_variation
29	0,54%	-10,02%	4,63	0,04
145	0,12%	-9,60%	5,46	0,04
44	0,28%	-9,29%	4,97	0,04
143	0,28%	-9,26%	4,87	0,04

82	0,08%	-9,03%	5,49	0,04
137	1,14%	-8,85%	5,36	0,04
6	0,04%	-8,77%	5,55	0,04
138	0,54%	-8,66%	5,36	0,04
56	0,43%	-8,55%	5,34	0,03
31	1,15%	-8,35%	5,85	0,04
127	0,65%	-7,96%	6,02	0,04
25	1,12%	-7,93%	6,86	0,04
140	0,58%	-7,22%	7,43	0,04
22	0,18%	-7,22%	6,68	0,04
134	0,10%	-7,20%	7,22	0,04
109	0,11%	-7,06%	7,05	0,04
119	0,05%	-6,70%	7,52	0,04
34	1,08%	-6,54%	7,52	0,04
48	0,09%	-6,02%	9,31	0,04
89	0,04%	-5,93%	8,12	0,03
47	0,80%	-5,87%	9,25	0,04
26	0,53%	-5,33%	9,96	0,04
69	0,00%	-5,31%	10,80	0,04
62	0,36%	-5,23%	9,64	0,04
86	1,49%	-5,05%	10,25	0,04
18	0,52%	-5,05%	11,89	0,04
120	0,17%	-4,91%	11,15	0,04
97	0,22%	-4,79%	12,62	0,04
1	1,19%	-4,49%	13,83	0,05
77	0,33%	-3,98%	15,04	0,04
90	1,03%	-3,00%	22,87	0,05
79	0,41%	-2,84%	19,52	0,04
41	0,04%	-2,79%	19,65	0,04
93	1,23%	-2,19%	24,26	0,04
39	0,41%	-2,16%	26,87	0,04
85	0,11%	-0,92%	57,57	0,04
13	0,13%	-0,87%	71,29	0,05
66	0,14%	-0,76%	69,56	0,04
64	1,00%	-0,58%	119,10	0,05
149	0,37%	-0,44%	115,21	0,04
30	0,79%	-0,19%	343,03	0,05
59	0,14%	0,14%	458,42	0,04
103	0,71%	1,54%	35,16	0,04
150	2,77%	1,92%	32,37	0,04
104	1,15%	2,03%	27,65	0,04
88	0,00%	2,25%	25,69	0,04
8	1,30%	2,41%	31,68	0,05
60	1,02%	2,54%	25,08	0,04
49	0,39%	2,75%	25,63	0,05
11	0,01%	3,19%	18,85	0,04

101	0,05%	3,33%	16,49	0,04
102	0,46%	3,55%	24,75	0,05
100	0,03%	3,58%	19,17	0,05
72	0,28%	3,61%	17,50	0,04
131	0,21%	3,88%	15,70	0,04
9	6,80%	4,36%	17,70	0,05
40	0,02%	4,60%	17,19	0,05
4	0,23%	5,29%	11,75	0,04
141	0,89%	5,58%	14,58	0,05
113	1,35%	5,67%	12,58	0,05
91	0,50%	6,01%	10,65	0,04
55	0,74%	6,16%	13,83	0,06
128	0,39%	6,85%	11,89	0,05
45	0,63%	7,64%	7,47	0,04
63	0,00%	8,12%	9,07	0,05
118	0,76%	8,15%	9,21	0,05
68	0,27%	8,24%	11,29	0,06
106	0,17%	8,30%	8,22	0,05
130	0,47%	8,44%	9,96	0,06
144	0,22%	10,20%	6,20	0,04
21	0,40%	10,59%	5,74	0,04
142	1,83%	10,69%	5,84	0,12
117	0,03%	11,23%	5,64	0,04
81	0,14%	12,04%	5,96	0,05
94	0,16%	12,51%	6,58	0,05
110	0,03%	12,61%	6,66	0,05
43	0,55%	12,93%	6,32	0,05
74	0,06%	14,45%	5,40	0,06
10	0,61%	14,56%	5,22	0,05
35	0,00%	14,91%	4,80	0,05
135	1,57%	15,46%	5,06	0,05
133	1,74%	15,97%	5,83	0,06
136	0,01%	16,16%	4,78	0,05
105	0,08%	16,24%	4,98	0,05
12	1,31%	16,79%	4,97	0,05
36	0,70%	16,80%	5,24	0,06
15	0,11%	17,11%	6,26	0,07
52	0,72%	17,93%	4,87	0,06
38	12,18%	18,36%	4,90	0,06
27	0,46%	19,14%	5,00	0,06
42	0,60%	19,57%	5,18	0,06
54	0,00%	19,90%	4,24	0,05
58	3,19%	20,10%	5,02	0,06
115	0,27%	20,20%	4,92	0,06
114	0,13%	20,37%	5,60	0,07
71	0,02%	20,44%	4,95	0,06

126	0,54%	20,61%	4,99	0,06
46	0,01%	20,90%	6,16	0,09
132	0,01%	21,01%	4,34	0,05
37	0,40%	21,34%	5,31	0,06
67	0,37%	21,84%	4,85	0,06
95	0,11%	22,73%	4,61	0,06
125	0,21%	23,47%	4,88	0,07
32	1,27%	24,10%	3,92	0,05
61	0,66%	27,69%	3,40	0,06
124	0,56%	30,83%	3,57	0,06
129	0,10%	30,83%	3,54	0,07
70	0,08%	31,74%	3,01	0,06
3	0,90%	31,86%	3,14	0,06
24	0,37%	36,04%	3,81	0,07
2	0,02%	36,46%	3,32	0,08
80	0,09%	36,86%	1,99	0,12
123	0,98%	37,38%	3,28	0,07
28	0,11%	37,57%	3,49	0,08
75	0,25%	39,44%	2,49	0,06
107	0,01%	39,76%	2,41	0,07
53	0,38%	41,11%	2,31	0,09
76	0,50%	41,53%	2,52	0,07
73	0,78%	44,98%	1,90	0,07
112	1,89%	45,05%	3,77	0,08
146	0,00%	45,30%	3,13	0,08
16	0,00%	45,54%	2,85	0,07
14	0,20%	48,56%	3,49	0,08
51	0,41%	49,94%	2,21	0,08
96	0,31%	50,24%	2,19	0,10
148	0,89%	54,59%	2,06	0,07
122	0,75%	54,73%	1,95	0,09
5	0,14%	55,18%	2,21	0,08
78	0,14%	58,29%	2,10	0,08
17	0,06%	60,72%	1,61	0,08
108	0,48%	63,43%	2,11	0,07
111	0,44%	68,73%	2,14	0,09
7	0,27%	74,63%	2,32	0,07
33	0,55%	75,65%	3,28	0,13
99	0,67%	76,39%	1,13	0,10
57	0,29%	79,49%	2,00	0,10
20	1,77%	92,32%	2,54	0,10
87	1,35%	92,85%	2,38	0,11
147	0,01%	96,82%	1,34	0,09
116	1,41%	98,83%	2,34	0,14
84	0,21%	112,16%	2,19	0,09
19	0,00%	117,27%	1,75	0,09

98	5,27%	117,94%	1,25	0,11
121	0,47%	127,93%	1,46	0,11
23	0,89%	145,69%	0,97	0,11
92	0,33%	168,20%	0,77	0,11
139	0,02%	212,99%	1,08	0,14
83	0,03%	242,58%	0,63	0,08
50	2,80%	299,26%	1,48	0,11
65	0,44%	495,65%	0,89	0,11

Tras analizar el cluster señalado en verde con los datos de validación, se han obtenido los siguientes resultados:

Tabla 28. Resultados validación Caso XIV

Cluster	Members	Expected_growth	External_variation	Internal_variation
98	0,00%	NaN	NaN	NaN

Al haber sido descartados todos los clusters, no se ha procedido a la fase de test.

4.2.15 Caso XV: tamaño de ventana 100 y agrupación en 150 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 29. Resultados entrenamiento Caso XV

Cluster	Members	Expected_growth	External_variation	Internal_variation
134	0,61%	-8,75%	3,62	0,05
5	0,05%	-7,56%	4,40	0,05
63	1,85%	-7,32%	4,89	0,06
78	0,36%	-7,24%	4,73	0,05
29	3,48%	-7,17%	4,72	0,05
35	0,01%	-6,95%	4,69	0,05
82	0,24%	-6,03%	5,36	0,05
26	0,03%	-5,96%	5,77	0,05
64	0,75%	-5,68%	6,32	0,06
86	0,02%	-5,63%	6,01	0,05
60	0,47%	-5,44%	6,34	0,05
127	0,54%	-5,33%	6,25	0,05
24	0,68%	-5,24%	6,92	0,05
67	0,00%	-4,91%	7,75	0,06
34	0,12%	-4,77%	6,63	0,05
53	0,78%	-4,69%	7,81	0,06
112	0,81%	-4,60%	7,00	0,05
21	0,04%	-4,25%	8,92	0,06
39	0,37%	-3,87%	9,07	0,06
84	0,91%	-3,74%	10,25	0,06

87	1,70%	-3,73%	10,07	0,06
123	0,30%	-3,59%	10,22	0,06
102	0,27%	-3,35%	11,12	0,06
65	0,58%	-3,25%	12,20	0,06
110	0,00%	-3,20%	13,35	0,07
141	0,58%	-2,62%	14,37	0,06
43	0,50%	-2,53%	15,26	0,06
137	0,33%	-2,30%	14,61	0,05
136	2,33%	-2,24%	17,16	0,06
120	0,01%	-2,22%	18,52	0,06
16	0,16%	-2,15%	19,40	0,06
72	1,23%	-2,11%	18,29	0,06
117	0,11%	-1,97%	22,00	0,06
1	0,55%	-1,93%	20,09	0,06
103	3,11%	-1,51%	32,44	0,07
9	0,01%	-1,08%	34,86	0,06
7	0,29%	-0,34%	131,02	0,06
132	0,76%	0,04%	1153,80	0,07
4	0,76%	0,20%	232,32	0,07
104	2,11%	0,72%	61,09	0,06
17	0,01%	0,83%	49,59	0,06
32	0,08%	0,95%	54,12	0,07
88	0,07%	1,45%	29,64	0,06
3	0,15%	1,60%	26,50	0,06
148	0,73%	1,63%	23,24	0,06
90	0,10%	3,00%	16,51	0,07
147	0,02%	3,21%	15,52	0,07
107	0,04%	3,24%	14,04	0,06
75	0,23%	3,53%	13,14	0,06
57	0,26%	3,64%	14,88	0,08
20	0,07%	3,70%	12,11	0,06
100	0,00%	4,89%	8,07	0,06
128	0,42%	5,58%	9,02	0,07
81	0,08%	5,87%	8,80	0,07
38	0,70%	6,73%	6,48	0,06
40	0,13%	7,11%	8,72	0,08
76	1,15%	7,64%	5,33	0,06
95	0,01%	8,08%	5,64	0,07
108	0,04%	8,15%	6,30	0,07
61	1,73%	8,30%	6,47	0,08
50	0,64%	8,57%	6,82	0,07
111	0,70%	8,91%	5,65	0,07
89	2,00%	8,95%	5,17	0,07
144	2,02%	9,24%	6,97	0,09
27	1,14%	9,49%	5,35	0,06
130	0,17%	9,95%	5,54	0,07

13	1,52%	10,06%	4,82	0,07
121	0,53%	10,21%	7,70	0,10
62	0,12%	10,42%	4,81	0,07
19	0,05%	10,69%	5,21	0,07
135	0,00%	12,07%	4,57	0,08
97	1,06%	12,21%	5,02	0,07
129	0,39%	12,41%	3,82	0,07
56	0,00%	12,78%	4,85	0,08
142	0,98%	14,66%	4,11	0,08
68	0,89%	14,97%	4,02	0,08
28	0,01%	15,11%	4,22	0,08
31	0,66%	15,69%	4,73	0,09
146	0,09%	15,74%	4,53	0,09
11	0,00%	15,87%	3,27	0,07
45	0,23%	16,27%	3,27	0,07
66	15,66%	16,54%	3,67	0,08
126	0,19%	17,15%	3,18	0,08
124	0,89%	17,75%	3,72	0,08
55	0,47%	17,92%	5,09	0,11
2	0,98%	19,23%	3,51	0,08
73	4,50%	20,58%	2,82	0,07
15	1,03%	20,69%	3,36	0,07
92	0,32%	20,83%	4,70	0,11
18	0,65%	21,41%	4,19	0,09
83	0,03%	22,37%	2,44	0,07
143	0,37%	22,60%	2,82	0,08
37	0,08%	23,65%	2,61	0,08
119	0,02%	23,76%	4,15	0,11
113	0,10%	26,16%	2,41	0,08
122	0,01%	26,93%	2,58	0,09
138	0,12%	27,74%	2,71	0,08
118	0,00%	29,05%	2,90	0,08
85	0,19%	29,33%	3,39	0,11
41	0,45%	31,00%	3,25	0,09
22	0,08%	32,44%	2,18	0,08
149	2,00%	32,48%	1,87	0,08
12	0,71%	34,19%	2,15	0,08
150	0,09%	34,99%	2,46	0,09
23	0,00%	35,30%	2,73	0,10
131	0,02%	37,82%	2,36	0,09
114	0,97%	42,23%	2,02	0,10
58	0,55%	45,10%	2,31	0,10
101	0,02%	46,12%	1,87	0,09
93	0,23%	46,52%	2,25	0,10
99	0,23%	47,18%	2,23	0,10
139	0,19%	49,84%	2,13	0,09

91	0,62%	52,66%	1,27	0,09
46	0,11%	52,86%	1,24	0,09
70	0,04%	53,60%	2,10	0,12
125	0,00%	57,59%	3,20	0,11
140	0,60%	58,99%	1,99	0,11
79	0,02%	62,12%	1,54	0,08
44	0,13%	67,96%	1,28	0,08
52	0,26%	69,18%	0,20	0,10
96	1,00%	72,93%	1,26	0,10
54	0,10%	76,12%	1,46	0,11
49	0,43%	78,71%	2,25	0,14
51	0,29%	81,67%	1,37	0,11
106	0,01%	91,68%	0,65	0,10
77	0,20%	93,98%	0,63	0,08
48	8,14%	96,49%	0,87	0,12
71	0,18%	97,53%	0,07	0,08
94	1,16%	103,49%	0,93	0,11
42	0,30%	108,14%	1,28	0,09
133	0,09%	109,43%	0,33	0,10
33	0,44%	110,48%	0,99	0,09
8	0,01%	117,92%	1,58	0,13
6	2,64%	121,50%	1,14	0,09
109	0,26%	132,64%	0,77	0,15
69	0,92%	137,33%	1,34	0,12
115	0,34%	145,56%	0,65	0,11
14	0,12%	151,13%	0,53	0,10
47	0,06%	152,49%	0,48	0,09
30	0,09%	159,52%	1,02	0,11
59	0,49%	172,53%	1,61	0,15
80	1,09%	205,09%	0,14	0,12
145	0,50%	241,70%	0,15	0,11
10	0,27%	256,22%	0,63	0,13
116	0,00%	276,11%	0,21	0,13
74	0,11%	288,31%	0,14	0,11
36	0,02%	368,31%	0,56	0,08

Para este caso no se ha encontrado ningún cluster que cumpla los requisitos mínimos de representación, rentabilidad y variabilidades, por lo tanto no se ha pasado a la fase de validación.

4.2.16 Caso XVI: tamaño de ventana 150 y agrupación en 150 clusters

La siguiente tabla muestra los resultados del entrenamiento del Caso III:

Tabla 30. Resultados entrenamiento Caso XVI

Cluster	Members	Expected_growth	External_variation	Internal_variation
75	0,51%	-14,06%	3,66	0,13
89	2,80%	-11,18%	1,57	0,05
52	1,55%	-10,81%	1,90	0,06
4	3,55%	-9,01%	2,73	0,06
143	0,00%	-8,64%	2,60	0,06
100	0,03%	-7,90%	3,27	0,06
95	0,00%	-7,81%	3,36	0,07
98	0,01%	-7,75%	3,21	0,06
28	0,06%	-7,71%	4,07	0,08
71	0,47%	-7,17%	3,90	0,06
62	4,38%	-7,00%	3,83	0,07
132	13,85%	-6,98%	3,39	0,07
2	0,00%	-6,89%	3,24	0,06
90	0,10%	-6,76%	4,01	0,07
47	0,73%	-6,54%	4,17	0,07
25	0,65%	-6,24%	4,14	0,06
67	0,50%	-6,16%	4,70	0,07
38	0,38%	-5,40%	5,41	0,07
58	0,46%	-5,35%	5,37	0,07
138	0,09%	-5,15%	5,48	0,07
12	0,19%	-4,78%	5,50	0,06
91	0,00%	-4,60%	6,32	0,07
96	0,01%	-4,52%	7,27	0,08
15	0,42%	-4,40%	6,47	0,07
20	2,03%	-3,42%	18,09	0,11
14	0,03%	-2,84%	15,11	0,09
111	0,46%	-2,23%	14,36	0,07
72	0,73%	-2,09%	15,80	0,07
40	0,56%	-2,03%	14,92	0,07
105	0,04%	-1,71%	18,41	0,07
94	0,92%	-1,67%	16,09	0,07
141	0,00%	-1,47%	20,57	0,07
125	0,44%	-0,54%	61,83	0,08
133	2,33%	-0,38%	84,08	0,07
97	0,06%	-0,02%	1527,60	0,07
116	0,04%	0,48%	105,20	0,13
11	0,43%	0,92%	37,69	0,07
53	0,50%	1,09%	33,75	0,08
10	0,91%	1,18%	29,02	0,07
122	0,38%	1,48%	49,83	0,15
3	0,34%	1,64%	22,82	0,08
31	0,00%	2,09%	16,47	0,08
34	0,89%	2,10%	18,11	0,07
18	0,14%	2,35%	14,38	0,08

147	0,16%	2,49%	13,44	0,07
57	0,04%	2,98%	14,13	0,09
73	0,22%	3,45%	9,70	0,07
60	1,53%	3,48%	12,60	0,09
50	0,05%	3,57%	10,57	0,08
114	5,10%	4,04%	10,48	0,09
33	0,09%	4,24%	8,48	0,08
137	2,06%	4,31%	9,88	0,09
142	1,42%	4,47%	11,01	0,09
92	0,01%	4,89%	9,03	0,11
115	0,04%	5,19%	7,58	0,08
109	0,30%	5,23%	7,05	0,07
130	0,25%	5,60%	7,31	0,09
102	0,75%	5,95%	6,41	0,09
70	0,72%	6,69%	6,86	0,10
59	0,10%	7,02%	6,04	0,09
88	0,02%	7,91%	5,40	0,13
44	0,39%	8,15%	6,53	0,10
66	0,07%	8,39%	6,01	0,10
145	0,01%	8,51%	8,81	0,12
135	0,01%	8,63%	5,94	0,10
27	1,38%	9,39%	4,80	0,09
108	0,60%	9,80%	4,24	0,09
26	0,00%	9,92%	10,43	0,13
19	0,21%	10,01%	3,88	0,08
149	0,50%	10,25%	4,63	0,10
48	6,25%	10,35%	4,74	0,09
61	1,02%	10,64%	8,97	0,18
39	0,42%	10,83%	5,16	0,10
80	0,19%	10,89%	3,55	0,08
24	0,13%	11,35%	5,13	0,10
84	0,07%	11,52%	3,57	0,09
51	0,01%	11,81%	5,18	0,10
148	0,01%	12,18%	3,58	0,09
43	0,07%	12,52%	3,97	0,09
126	0,08%	13,14%	3,55	0,09
85	0,02%	15,30%	6,40	0,13
144	0,00%	15,63%	3,27	0,10
110	0,00%	16,43%	4,02	0,10
123	0,47%	16,77%	3,86	0,11
1	0,69%	17,07%	3,99	0,11
120	0,07%	17,14%	3,01	0,09
16	0,01%	18,18%	3,78	0,10
129	0,07%	18,91%	3,18	0,11
29	2,39%	20,34%	3,16	0,11
79	2,48%	21,12%	3,27	0,10

81	0,40%	21,62%	2,82	0,10
134	0,13%	25,43%	3,14	0,13
127	0,04%	27,75%	2,04	0,11
46	0,06%	28,17%	2,61	0,12
45	0,17%	29,87%	2,60	0,12
55	0,53%	30,70%	3,59	0,12
103	0,02%	32,00%	2,59	0,14
56	2,20%	32,25%	1,96	0,10
17	0,14%	38,31%	1,84	0,13
131	1,27%	42,15%	2,48	0,12
78	0,06%	46,26%	1,17	0,09
104	0,71%	47,03%	1,64	0,14
37	0,02%	51,43%	2,17	0,14
101	0,06%	53,44%	1,44	0,13
63	2,07%	55,54%	1,33	0,12
6	0,01%	56,30%	1,83	0,12
99	0,01%	59,33%	1,30	0,10
21	1,32%	64,76%	1,89	0,13
106	0,14%	65,73%	1,18	0,07
9	0,28%	67,48%	3,05	0,12
35	0,44%	67,78%	1,21	0,12
124	0,00%	69,03%	1,84	0,15
22	0,02%	69,13%	0,13	0,06
140	0,55%	70,12%	1,44	0,14
77	0,22%	71,74%	1,27	0,08
41	0,08%	73,86%	2,33	0,14
7	0,00%	76,20%	0,15	0,06
42	0,00%	79,41%	0,10	0,06
54	0,00%	81,63%	0,14	0,06
30	0,17%	83,13%	0,83	0,13
87	0,00%	88,72%	1,11	0,13
113	0,12%	89,88%	1,38	0,12
86	1,64%	94,22%	0,92	0,14
76	0,01%	98,94%	1,49	0,13
36	0,43%	100,09%	1,00	0,10
112	0,62%	107,30%	0,10	0,08
118	0,11%	112,72%	0,63	0,09
93	0,00%	113,33%	1,22	0,13
107	0,04%	123,72%	0,64	0,12
8	0,29%	129,92%	0,46	0,06
150	0,05%	134,46%	0,54	0,11
139	1,19%	141,55%	0,91	0,10
64	1,41%	141,60%	0,21	0,12
49	0,17%	145,29%	0,91	0,14
32	0,67%	152,60%	0,28	0,08
117	0,00%	153,48%	0,39	0,10

65	0,70%	156,88%	0,85	0,11
74	1,67%	157,04%	1,20	0,11
121	0,00%	160,11%	0,49	0,09
5	0,02%	172,23%	0,34	0,08
69	0,70%	177,54%	1,28	0,13
23	0,31%	181,63%	0,27	0,09
136	4,72%	208,78%	0,28	0,12
146	0,40%	227,06%	0,11	0,08
13	0,27%	NaN	NaN	NaN
68	0,00%	NaN	NaN	NaN
82	0,63%	NaN	NaN	NaN
83	0,24%	NaN	NaN	NaN
119	0,08%	NaN	NaN	NaN
128	0,00%	NaN	NaN	NaN

Tras analizar los clusters señalados en verde con los datos de validación, se han obtenido los siguientes resultados:

Tabla 31. Resultados validación Caso XVI

Cluster	Members	Expected_growth	External_variation	Internal_variation
35	0,41%	111,37%	1,07	0,00
86	0,00%	NaN	NaN	NaN
36	0,00%	NaN	NaN	NaN
112	0,00%	NaN	NaN	NaN
139	0,00%	NaN	NaN	NaN
64	0,00%	NaN	NaN	NaN
32	0,00%	NaN	NaN	NaN
65	0,00%	NaN	NaN	NaN
74	0,00%	NaN	NaN	NaN
136	0,00%	NaN	NaN	NaN
146	0,00%	NaN	NaN	NaN

Por último, se ha probado el conjunto de datos de test en el cluster que continúa pareciendo interesante tras la fase de validación, obteniendo el siguiente resultado:

Tabla 32. Resultados test Caso XVI

Cluster	Members	Expected_growth	External_variation	Internal_variation
35	0,00%	20,61%	4,59	0,00

El cluster que parecía interesante tras la fase de validación no ha pasado el test, así que será descartado.

5 CONCLUSIONES

5.1 CONCLUSIONES SOBRE LOS OBJETIVOS DEL PROYECTO

En primer lugar, cabe destacar la importancia que supone la capacidad computacional para este proyecto. La ejecución de algoritmos complejos para el análisis de cantidades tan altas de datos (precio de cierre diario de más de 850 empresas durante quince años) permitirá análisis más exhaustivos cuanto mayor sea esta capacidad. Además, y como se ha señalado en el capítulo de metodología, hay otros algoritmos como el t-SNE para la reducción de dimensionalidad, que directamente se han descartado por la carga que suponían tras unas pruebas iniciales.

Además, una de las conclusiones más importantes de este proyecto sería intentar dar respuesta, una vez expuestos el modo de trabajo y los resultados, a la pregunta que se planteaba en la introducción del proyecto:

¿Es posible generar beneficios en el mercado de valores utilizando técnicas de análisis de datos, y concretamente, de aprendizaje no supervisado?

Tal y como se expuso en el apartado de predictibilidad, predecir los precios de las acciones del mercado de valores y sus evoluciones en el tiempo es una tarea muy complicada. Además, se han mostrado un gran número de fuentes que aseguran que el mercado de valores está gobernado por el paseo aleatorio, añadiendo que es imposible realizar predicciones significativas y que la mejor manera de invertir es comprando acciones aleatorias de los diferentes índices que se deseen seguir en un volumen suficientemente grande como para replicar el comportamiento del mercado como conjunto. Así, responder a esta pregunta de forma definitiva es algo muy complejo, prácticamente imposible, debido a la existencia de una cantidad ingente de factores que modelan el comportamiento.

Por un lado, hay resultados de este proyecto que han mostrado la existencia de nichos de inversión muy atractivos debido a su alta rentabilidad y su baja variabilidad, y cómo pueden ser detectados por los algoritmos de clustering utilizados. Por otra parte, también ha quedado patente lo difícil que es dar con estas combinaciones de metaparámetros que dan forma al algoritmo, ya que hay un equilibrio muy fino entre hacer un análisis demasiado general y sobreajustar el modelo de tal manera que los resultados no sean explicativos para conjuntos de datos diferentes. Además de estas causas por las que pueden no encontrarse resultados satisfactiorios, hay veces en las que simplemente no hay estructuras internas entre los datos que sirvan para agruparlos de tal manera que queden expuestos los nichos de inversión atractivos.

Otras veces, el cambio en el mercado con el tiempo puede ser tan significativo que un conjunto de metaparámetros que había dado buenos resultados en las etapas de entrenamiento y validación no se comporte de la misma manera con los datos reservados para el test, algo que también se ha visto en los resultados.

También habría que destacar al intentar responder esta pregunta que la enorme cantidad de factores que afectan al comportamiento del mercado de valores hace que las técnicas de aprendizaje no supervisado sean seguramente la mejor aproximación al problema que supone la búsqueda de nichos de inversión con unas determinadas características. Además, hay nichos de inversión que en este trabajo se han descartado por tener demasiada variabilidad o

por no ofrecer suficiente rentabilidad, que podrían ser interesantes a los ojos de un inversor de un perfil diferente. Por ejemplo, un inversor más afín al riesgo pondría los límites de variabilidades más altos, mientras que alguien interesado en operar con derivados financieros podría estar interesado en valores que se mantengan en el mismo precio o que tengan un precio máximo detectado porque haya algún producto en el mercado que cubra este tipo de casos.

Así, el mundo del mercado de valores es tan complejo que incluso el mismo resultado puede ser interpretado de formas completamente diferentes.

De esta manera, la mejor respuesta a la pregunta sería que, tras la ejecución de este proyecto, se mantiene la opinión de que el comportamiento del mercado de valores es predecible parcialmente con las complicaciones que ello conlleva, y que por lo tanto sí es posible generar ventaja con respecto a otros inversores mediante la utilización de técnicas de aprendizaje automático.

5.2 FUTUROS DESARROLLOS

- Utilización de otros algoritmos de aprendizaje no supervisado computacionalmente más pesados como por ejemplo ksom, que utiliza redes neuronales.
- Utilización de algoritmos de reducción de dimensionalidad más complejos como podría ser el ya propuesto anteriormente en este documento t-SNE.
- Búsqueda de parámetros a partir de datos de diferente naturaleza como puedan ser el volumen o indicadores que no utilicen únicamente el precio de cierre.
- Creación de una herramienta de inversión a partir de los resultados obtenidos.
- Utilización de más datos de mercados que no hayan sido analizados en este proyecto.

6 BIBLIOGRAFÍA

[1] «Diccionario de la Real Academia Española de la lengua,» [En línea]. Available: http://dle.rae.es/?id=0KZwLbE. [Último acceso: 2018 06 27].

[2] J. M. P. y. L. H. S. David M Cutler, «What moves stock prices?,» 1989. [En línea]. Available: https://www.nber.org/papers/w2538. [Último acceso: 28 06 2018].

[3] «Random Walk Theory,» [En línea]. Available: https://www.investopedia.com/terms/r/randomwalktheory.asp. [Último acceso: 28 06 2018].

[4] A. W. L. y. A. C. MacKinlay, «Stock Market Prices Do Not Follow Random Walks: Evidence From a Simple Specification Test,» 02 1987. [En línea]. Available: https://www.nber.org/papers/w2168. [Último acceso: 28 06 2018].

[5] E. F. Fama, «The Behavior of Stock-Market Prices,» 01 1965. [En línea]. Available: https://www.jstor.org/stable/2350752?seq=1#page_scan_tab_contents. [Último acceso: 29 06 2018].

[6] R. S. y. M. A. Zaman, «Market reaction to Business Week 'Inside Wall Street' column: A self-fulfilling prophecy,» 05 1996. [En línea]. Available: https://www.sciencedirect.com/science/article/pii/0378426695000259. [Último acceso: 29 06 2018].

[7] F. Black, «Noise,» 07 1986. [En línea]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1986.tb04513.x. [Último acceso: 29 06 2018].

[8] A. Puerro, «La psicología de la masa en los mercados: homenaje a José Penso de la Vega,» [En línea]. Available: https://labolsacomoestadistica.wordpress.com/2015/02/26/la-psicologia-de-la-masa-en-los-mercados-homenaje-a-jose-penso-de-la-vega/. [Último acceso: 12 07 2018].

[9] N. Fuller, «The Most Successful Price Action Trader in History: Munehisa Homma,» 10 2017. [En línea]. Available: http://www.learntotradethemarket.com/forex-articles/most-successful-price-action-trader-in-history-munehisa-homma. [Último acceso: 12 07 2018].

[10] «Technical Analysis of Stocks and Trends,» [En línea]. Available: https://www.investopedia.com/terms/t/technical-analysis-of-stocks-and-trends.asp. [Último acceso: 30 06 2018].

[11] J. Kuepper, «Technical Analysis: Chart Types,» [En línea]. Available: https://www.investopedia.com/university/technical/techanalysis7.asp. [Último acceso: 01 07 2018].

[12] J. Kuepper, «Technical Analysis: Chart Patterns,» [En línea]. Available: https://www.investopedia.com/university/technical/techanalysis8.asp. [Último acceso: 01 07 2018].

[13] J. J. Murphy, Análisis técnico de los mercados financieros, Nueva York: Gestión 2000, 1999.

[14] C. Murphy, «Moving Averages: What Are They?,» Investopedia, [En línea]. Available: https://www.investopedia.com/university/movingaverage/movingaverages1.asp. [Último acceso: 07 06 2018].

[15] «Bollinger Band,» Investopedia, [En línea]. Available: https://www.investopedia.com/terms/b/bollingerbands.asp. [Último acceso: 07 07 2018].

[16] J. Oliveira, «La tecnología que enseña a los robots a ´pensar´como humanos,» 3 julio 2017. [En línea]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.2964&rep=rep1&type=pdf. [Último acceso: 28 noviembre 2017].

[17] J. G. C. y. T. M. M. Ryszard S michalski, Machine learning: An artificial intelligence approach, 2013.

[18] G. H. A. K. I. S. R. S. Nitish Srivasta, «Dropout: A Simple Way to Prevent Neural Networks from Overfitting,» junio 2014. [En línea]. Available:

https://www.cs.toronto.edu/~hinton/absps/JMLRdropout.pdf. [Último acceso: 30 noviembre 2017].

[19] X. Amatriain, «What are hyperparameters in machine learning?,» 1 agosto 2016. [En línea]. Available: https://www.quora.com/What-are-hyperparameters-in-machine-learning. [Último acceso: 30 noviembre 2017].

[20] R. Kohavi, «A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,» 1995. [En línea]. Available: http://robotics.stanford.edu/users/ronnyk.link/accEst.pdf. [Último acceso: 30 noviembre 2017].

[21] C. Grima, «El diagrama de Voronoi, la forma matemática de dividir el mundo,» 08 05 2017. [En línea]. Available: https://www.abc.es/ciencia/abci-diagrama-voronoi-forma-matematica-dividir-mundo-201704241101_noticia.html. [Último acceso: 14 07 2018].

[22] J. MacQueen, «Some methods for classification and analysis of multivariate observations,» Berkeley, 1967. [En línea]. Available: https://projecteuclid.org/euclid.bsmsp/1200512992. [Último acceso: 14 07 2017].

[23] H. Steinhaus, «Sur la division des corps matériels en parties,» 1956. [En línea]. Available: http://www.laurent-duval.eu/Documents/Steinhaus_H_1956_j-bull-acad-polon-sci_division_cmp-k-means.pdf. [Último acceso: 14 07 2017].

[24] D. Leinweber, «Stupid Data Miner Tricks, overfitting the S&P 500,» [En línea]. Available: https://www.researchgate.net/publication/247907373_Stupid_Data_Miner_Tricks_Overfitting_the_SP_500. [Último acceso: 11 07 2018].

[25] Y. N. y. S.-Y. W. Wei Huang, «Forecasting stock market movement direction with support vector machine,» 01 2005. [En línea]. Available: https://www.researchgate.net/publication/220472531_Forecasting_stock_market_movement_direction_with_support_vector_machine. [Último acceso: 11 07 2018].

[26] G. S. A. y. K. P. Valavanis, «Surveying stock market forecasting techniques - Part II: Soft computing methods,» 01 04 2009. [En línea]. Available: https://scholar.google.com/citations?user=bbMxXyMAAAAJ&hl=en#d=gs_md_cita-d&p=&u=%2Fcitations%3Fview_op%3Dview_citation%26hl%3Den%26user%3DbbMx

XyMAAAAJ%26citation_for_view%3DbbMxXyMAAAAJ%3Au5HHmVD_uO8C%26tzom%3D-120. [Último acceso: 12 07 2018].

[27] B. Y. y. A. A. Yuehui Chen, «Flexible neural trees ensemble for stock index modeling,» 01 2007. [En línea]. Available: https://www.sciencedirect.com/science/article/pii/S0925231206002736. [Último acceso: 12 07 2018].

[28] J. C. Hidalgo, «El Ibex cumple 25 años con una ganancia anual del 5,1% más el 4% en dividendos,» Expansion, 14 01 2017. [En línea]. Available: http://www.expansion.com/mercados/2017/01/14/5878b52ce5fdeaeb768b4678.html. [Último acceso: 18 07 2018].

[29] T. M. Mitchell, Machine Learning, McGraw-Hill, 1997.