GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Predicción del éxito de la inversión de un Venture Capital en una Startup utilizando algoritmos de Machine Learning.

Autor: Armando Sala López

Director: Miguel Ángel Sanz Bobi

Madrid

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

Predicción del éxito de la inversión de un Venture Capital en una Startup utilizando algoritmos de Machine Learning.

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2022/2023 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.: Armando Sala López Fecha: 04/07/2022

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Firmado por SANZ BOBI MIGUEL ANGEL - ***6599** el día 05/07/2023 con un certificado emitido por AC FNMT Usuarios

Fdo.: Miguel Ángel Sanz Bobi Fecha: ……/ ……/ ……

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Predicción del éxito de la inversión de un Venture Capital en una Startup utilizando algoritmos de Machine Learning.

Autor: Armando Sala López

Director: Miguel Ángel Sanz Bobi

Madrid

PREDICCIÓN DEL ÉXITO DE LA INVERSIÓN DE UN VENTURE CAPITAL EN UNA STARTUP UTILIZANDO ALGORITMOS DE MACHINE LEARNING.

Autor: Sala López, Armando.
Director: Sanz Bobi, Miguel Ángel.

RESUMEN

Este proyecto se centra en el desarrollo de un modelo de predicción para el éxito de las inversiones en startups. Se utiliza un enfoque basado en clústeres y se analiza cada cluster por separado. Los resultados muestran un modelo efectivo, confiable y útil para la toma de decisiones informadas en el ámbito del capital riesgo y emprendimiento.

PALABRAS CLAVE: Venture Capital (Capital Riesgo), Machine Learning, Clustering, Startup.

1. INTRODUCCIÓN

El capital riesgo, también conocido como *venture capital*, consiste en la inversión a largo plazo por parte de compañías especializadas en la gestión de recursos financieros, de forma minoritaria y temporal, en pequeñas y medianas empresas, con grandes perspectivas de crecimiento y/o rentabilidad.

La industria del *venture capital* experimentó un fuerte crecimiento durante la primera década del siglo XXI. No obstante, la crisis financiera de 2008 provocó fuertes correcciones en la economía mundial y, en particular, en la industria del *venture capital*. Sin embargo, la recuperación económica global y la intensificación del desarrollo tecnológico a mediados de la segunda década del siglo provocó un fuerte movimiento en el ecosistema *startup* que desencadenó otra muy buena década para el capital riesgo. Con la conclusión de la década, la pandemia del coronavirus generó mucha inseguridad en la economía global, afectando al *venture capital* y al ecosistema *startup* de igual forma que al resto de sectores. En la actualidad, tras un 2021 donde se batieron récords de inversión

en la industria del *venture capital*, ésta parece que continua con una tendencia positiva. Dicha tendencia viene reflejada por fuertes inversiones, aunque muy por debajo de las cifras de 2021. No obstante, en el mercado más potente del capital riesgo del mundo, Estados Unidos, si la captación de fondos hubiese seguido la tendencia de los 3 primeros trimestres del año, el 2022 se podría haber convertido en el año que más capital se ha levantado en la historia. Esto muestra el atractivo que sigue teniendo el sector del venture capital como método de inversión alternativa.

2. ESTADO DE LA CUESTIÓN

La inversión en *startups* es el gran reto al que los fondos de *venture capital* se enfrentan. Existen distintas etapas en las que se puede encontrar una *startup* que busca financiación de un fondo de *Venture Capital*. Cuanto más joven sea la empresa, menos métricas existen para poder calcular un potencial retorno de la *startup*. Por lo tanto, en estos casos, gran parte del análisis recae sobre el estudio, con una gran componente subjetiva, del equipo que forma la compañía, el *background*[1] de los miembros de este y la química entre ellos [1]. Existen otros puntos de análisis igual de importantes, como el mercado al que pertenece la compañía o el producto, pero la dificultad que supone elaborar métricas fiables en estas etapas da lugar a un mayor componente subjetivo y, por lo tanto, históricamente, no ha existido una gran necesidad del uso de herramientas como el *machine learning*.

No obstante, el uso de herramientas de inteligencia artificial ha experimentado una fuerte evolución durante los últimos años [2]. En la actualidad, algunos de los principales fondos del mundo utilizan algoritmos desarrollados *in-house*[2] o elaborados por compañías que se dedican exclusivamente a este sector [3]. Existen distintos modelos que se aplican en

[1] El background es el conjunto de conocimientos y experiencias que constituyen el bagaje de una persona.
[2] In-house se refiere a una actividad u operación que se realiza dentro de una empresa, en lugar de depender de la subcontratación.

la industria, desde algoritmos dedicados a la *due diligence*[3] y el análisis de compañías hasta algoritmos que descubren potenciales *startups* para invertir [4].

3. OBJETIVOS

El principal objetivo del trabajo es desarrollar un modelo basado en algoritmos de *machine learning* que pueda ser útil para los fondos de capital riesgo a la hora de analizar oportunidades de inversión. La herramienta permitirá a los fondos obtener mayor cantidad de información relevante para estimar el potencial éxito de la inversión en una *startup*. Además, los *founders*[4] podrían aprovecharse de un modelo como éste para poder tomar decisiones que les permitan desarrollar su empresa de la manera óptima.

Es cierto que un análisis de este tipo permite identificar tendencias en el mercado y su evolución, por lo que otro de los objetivos del trabajo será identificar tendencias en el mercado que, en este caso, dada la base de datos, englobaría movimientos en el ecosistema *startup* hasta 2013. En particular, centraremos el análisis en el efecto que tuvo la crisis financiera de 2007-2008 en el desarrollo de las startups y sus valoraciones. Además, resulta interesante el estudio del éxito de una inversión en función del tipo de *exit*[5] (IPO vs Adquisición en el mercado privado).

Finalmente, resultará interesante conocer el impacto que tienen los estudios que tienen los fundadores de las *startups* y la relevancia que tiene ese nivel de cualificación sobre el éxito de las empresas. En particular se concentra el estudio en el impacto que tiene el prestigio de la institución académica, donde los fundadores cursan los estudios, en el éxito de la startup.

---

[3] Due diligence, o diligencia debida en español, es la operación previa de revisión de una persona o empresa de cara a evaluar posibles riesgos al establecer una relación económica con ella

[4] Se entiende por *founders* aquellas personas físicas o jurídicas que constituyeron la *startup.*

[5] El exit se refiere al momento en el que los inversores o fundadores de la empresa obtienen un retorno de su inversión inicial al vender total o parcialmente sus participaciones en la compañía.

4. METODOLOGÍA

4.1 PREPROCESADO

El preprocesado de datos ha tenido un gran peso en la consecución de los objetivos planteados previamente. La fuente de información utilizada en este TFG , compuesta a su vez por 14 bases de datos distintas, contenía un alto nivel de datos nulos o no validos que han tenido que ser modificados o eliminados según sus características. Tras haber procesado todas las tablas individualmente, se obtienen un total de seis tablas que forman parte del análisis exploratorio a las que se referirán como 'Conjunto de Bases de Datos Procesadas' a lo largo del análisis exploratorio.

La base de datos final es una combinación de las tablas del 'Conjunto de Bases de Datos Procesadas' que asegura el mejor balance entre cantidad de casos (766 filas) y variables (25 columnas). Contiene información referente a la startup, como el sector y el método de exit. Además, también incluye información con respecto a las rondas de financiación, como el año de la primera y última inversión, el número de rondas y la suma total del capital levantado, entre otras variables. Asimismo, contiene datos sobre las oficinas y sobre la formación académica de los fundadores. Finalmente, la base de datos final incluye información sobre las relaciones que tienen las startups con otros miembros del ecosistema emprendedor y los participantes en los procesos de captación de fondos.

	category_code	status	invested_companies	first_funding_at	last_funding_at	funding_rounds	funding_total_usd	relationships	region	country_code	...	funds	people	companies
0	web	acquired	0	2008	2008	1	5000000.0	14	sf bay	USA	...	1	2	0
1	games_video	acquired	0	2007	2007	1	3000000.0	6	sf bay	USA	...	2	4	0
2	photo_video	acquired	0	2007	2013	3	15286415.0	2	nashville	USA	...	2	0	0
3	hardware	acquired	0	2011	2011	1	28000000.0	7	sf bay	USA	...	2	8	0
4	enterprise	acquired	0	2009	2012	5	142000000.0	38	sf bay	USA	...	21	5	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
761	biotech	ipo	0	2010	2013	4	128614965.0	6	bedminster	USA	...	5	0	0
762	biotech	ipo	0	2011	2012	2	35400000.0	17	sacramento	USA	...	8	0	0
763	manufacturing	ipo	0	2012	2012	1	20000000.0	6	santa barbara	USA	...	1	0	0
764	messaging	ipo	0	2007	2012	5	44000000.0	32	sf bay	USA	...	13	0	3
765	security	ipo	0	2006	2012	5	328600000.0	35	sf bay	USA	...	12	0	0

Figura 1. Fragmento de la Base de Datos Final.

4.2 ANÁLISIS EXPLORATORIO

El análisis exploratorio se divide en dos etapas. En primer lugar, se analizan las bases de datos pertenecientes al 'Conjunto de Bases de Datos Procesadas', permitiendo una mayor comprensión sobre el ecosistema startup. Esta primer análisis de la base de datos resulta clave, pues a medida que se va añadiendo contenido a la base de datos el número de casos completos va reduciendo y por lo tanto se pierde la imagen general del ecosistema.

Una vez se ha analizado el ecosistema global, obteniendo así, un mayor conocimiento del sector y de las variables que componen el banco de datos, se profundiza en la base de datos final, que se utilizará en los modelos. El análisis de la base de datos final permite identificar tendencias y características de la base de datos, dando lugar a una mejor toma de decisiones y un mejor diseño de los modelos. Además, este análisis está diseñado para confirmar que la base de datos final, que no deja de ser una muestra de los datos que se ven en las seis bases de datos procesadas, sigue las tendencias del ecosistema, identificadas en el estudio sector. En la Tabla 1 se muestran los resultados obtenidos para ambas bases de datos, mostrando las similitudes en la variable que representa la región donde se encuentra la oficina.

Tabla 1. Comparativa de las Regiones Más Comunes en el Ecosistema y en la Base de Datos Final

Región	Ranking en Ecosistema	Ranking en Base de Datos Final
SF Bay	1	1
New York	2	3
London	3	7
Los Angeles	4	4
Boston	5	2
Seattle	8	5

4.3 CLUSTERING Y MODELOS DESCARTADOS

La elaboración de los modelos comienza con un método de clustering diseñado para conocer en profundidad los diferentes grupos de datos y añadir una variable que aumente la precisión del modelo. Para ello, se crea la base final de datos procesada que se prepara aplicando normalización a las variables numéricas y OneHotEncoding a las variables categóricas. Tras separar la base de datos final procesada en el conjunto de entrenamiento y el de prueba, se diseña un algoritmo de clustering que selecciona un número óptimo de 6 clústeres.

Las etiquetas de los clústeres se añaden como variable a la base de datos final procesada y se prueban los siguientes modelos:

- Regresión Lineal
- Random Forest
- AdaBoostRegressor
- SVR
- Árboles de Decisión
- GradientBoostingRegressor
- XGBRegressor
- Redes Neuronales

No obstante, todos estos modelos son descartados en favor del modelo final.

4.4 MODELO FINAL

El primer paso en el diseño del modelo final, al igual que se ha realizado para los modelos descartados, es eliminar los valores extremos utilizando el método del valor z. Este método elimina una sola entrada que su variable objetivo tenía un valor de 9949110360000$. Posteriormente, se aplica el preprocesado, donde se normalizan las variables numéricas y se codifican las categóricas. Finalmente, se separan los datos en el conjunto de entrenamiento y el conjunto de prueba.

En segundo lugar, se realiza un análisis para identificar el número de clústeres óptimos. El método escogido para conocer el número de clústeres es KMeans. En este caso, el número de casos se reduce, dando lugar a un número de idóneo de seis clústeres.

Tras haber analizado los nuevos grupos formados a partir del KMeans se procede a aplicar todas los modelos distintos mencionados previamente a cada clúster. Sin embargo, las redes neuronales no se han aplicado a los clústeres. Esto se debe a que, dado el bajo número de casos en cada grupo y su bajo desempeño con la base de datos completa, no tiene sentido implementar este modelo en este caso.

Para cada grupo, se escogió el modelo que mejor se adapta y mejores resultados genera, aplicándole a los modelos la técnica de hyperparameter tuning para maximizar el rendimiento del modelo final. Los modelos obtenidos para cada grupo son los siguientes:

- Grupo 0: Random forest.
- Grupo 1: AdaBoost.
- Grupo 2: Gradient Boosting.
- Grupo 3: SVR.
- Grupo 4: Random Forest.
- Grupo 5: AdaBoost.

Tras obtener los modelos óptimos para cada clúster, se combina el algoritmo de clustering junto a los modelos utilizando métodos de ensemble para obtener el modelo final completo.

5. ANÁLISIS DE LOS RESULTADOS

El modelo final supone una ligera mejoría en cuanto a las métricas utilizadas si se compara con el resto de los modelos elaborados, incluyendo las redes neuronales y los algoritmos de boosting, entre otros. Además, a diferencia de algunos de los modelos descartados, las métricas de error utilizadas para analizar el modelo no varían tras cada ejecución. Esto genera un modelo mucho más confiable y útil de cara a predecir el éxito de la inversión en startups de manera precisa. El hecho de que los resultados obtenidos a partir del modelo final sean los más potentes muestran la ventaja que supone subdividir el conjunto de datos en clústeres, donde los patrones y relaciones entre los datos son mucho más claras y dan lugar a mejores predicciones. Además, a pesar de que la subdivisión en clústeres da lugar a conjuntos de datos más pequeños, donde extraer relaciones y patrones fiables es mucho más preciso, este modelo sigue siendo capaz de dar un rendimiento superior al resto de modelos.

```
RMSE: 1.2935939478679512
Average Error ($): 2.6458660907546063
MAPE: 5.391202549824616
Average MAPE (%): 218.46714710822704
R^2: 0.3595023411703795
```

Figura 2. Resultados del Modelo Final.

El modelo final tiene un error medio de 2.65 dólares. Este resultado es realmente sorprendente pues la variable objetivo contiene valores entre 30,000 y 104,000,000,000. Por lo tanto, se puede asegurar que el modelo es capaz de predecir con un nivel de precisión aceptable el éxito de la inversión.

6. CONCLUSIONES

En primer lugar, tras la elaboración y posterior análisis de los modelos y sus resultados, se puede concluir que el modelo final es efectivo y cumple con el principal objetivo de generar un modelo que permita a los fondos de capital riesgo y emprendedores tomar decisiones más informadas. El haber obtenido un modelo final que tan solo tiene un error promedio de 2.5 dólares muestra un alto grado de fiabilidad y desde luego sirve como una herramienta útil para analizar oportunidades de inversión para los fondos. Además, la propia naturaleza de las inversiones en startups tiene un alto componente de varianza, donde casos extremos y grandes éxitos son comunes. Por lo tanto, se puede asumir una métrica MAPE ligeramente más alta pues estos casos son relativamente escasos a lo largo del tiempo. Además, todo el análisis exploratorio previo y los resultados obtenidos de otros modelos también aportan información realmente útil al emprendedor, que puede usarla como una fuente de información adicional.

Finalmente, todas los objetivos relacionados con identificar la importancia relativa han sido cumplidos utilizando el análisis de los clústeres. Combinando la información sobre los clústeres obtenidos y sus valores medios de la variable objetivo por clúster, se pueden extraer conclusiones relevantes sobre las variables a estudiar.

```
Cluster 3: Median = 352000000.0, Mean = 2526095394.2720337
Cluster 4: Median = 200000000.0, Mean = 1112472978.7234042
Cluster 2: Median = 102500000.0, Mean = 214875666.4057971
Cluster 0: Median = 100000000.0, Mean = 485432790.6976744
Cluster 1: Median = 89000000.0, Mean = 273467411.7816092
Cluster 5: Median = 78400000.0, Mean = 408730647.9166667
```

Figura 3. Valores Medios y Medianos del Precio de Exit por Clúster.

En primer lugar, con respecto al método de exit, el único clúster donde todas las operaciones son IPOs es el clúster 3, que como se ha mencionado previamente presenta los valores más altos en la variable objetivo. Además, es el clúster 4, que es el segundo con las valoraciones más altas, el siguiente grupo con mejor ratio de IPOs/Adquisiciones. Por otro lado, se observa que el clúster 1, que es el que tiene os valores medios de venta/IPO más bajos, es el único grupo donde todas las operaciones son adquisiciones. También se ha de mencionar como las pocas operaciones de IPO en el grupo 5 son las que hacen que el valor medio sea muy superior al mediano. Estos hechos demuestran que existe una clara relación entre el tipo de salida y el valor de la compañía, donde los inversores que participan en compañías cuyo exit sea una IPO tienden a tener participaciones en una empresa con mayor valoración.

En segundo lugar, se observa cierta relación entre el tiempo en el que han ido avanzando las startups, primero a través de sus rondas de financiación y, posteriormente mediante la operación de venta o salida a bolsa, y el éxito de la inversión. La crisis financiera que ocurre entre el 2007 y el 2008 afecta significativamente a la economía mundial y con ello al sector del venture capital. Si se analizan las fechas de las operaciones de las startups se observan ciertos patrones que muestran el efecto del mercado y la situación económica sobre el ecosistema y el éxito de las startups. Por ejemplo, los dos clústeres con peores cifras de venta/IPO, el 1 y el 5, son aquellos que levantaron sus rondas de financiación entre 2006-2008. Esto se puede deber a que durante la crisis las valoraciones sufrieron y dio lugar a rondas de financiación más ajustadas y menos capital para operar. Por el contrario, las compañías que más tarde han cerrado sus procesos de levantamiento de capital han sido los que mejores valoraciones han tenido. Este patrón se ve acentuado al analizar los valores para el clúster con los exits más altos. Las empresas del clúster 3, de

media, realizan su última ronda de financiación en 2011 donde la economía global estaba experimentando una fuerte recuperación tras la crisis. Por lo tanto, estas tendencias demuestran una clara relación entre el ciclo económico y el éxito que pueda tener una startup y sus inversores en su proceso de exit. Además, muestran el impacto significativo que tuvo la crisis financiera en el ecosistema startup.

Finalmente, en la figura 7 se muestra una relación menos evidente entre la universidad y el éxito de la compañía, aunque no tan fuerte como la relación con el tipo de exit. El clúster 5, que tiene menos fundadores/CEOs provenientes de universidades prestigiosas, se encuentran entre las categorías con las peores valoraciones. Por otro lado, las categorías 4 y 3, presentan las valoraciones más altas y se destacan por tener un alto porcentaje de fundadores/CEOs provenientes de las mejores instituciones (Top 100 del mundo).
Sin embargo, es importante mencionar que el clúster 1, el cual tiene la segunda valoración más baja en términos de mediana, contiene una gran cantidad de fundadores en las mejores universidades del mundo. Además, el clúster 2, que se sitúa en la tercera posición en cuanto al valor mediano de la variable objetivo, no tiene ningún alumno que haya asistido a la universidad. Lo mismo ocurre con el clúster 0, que se encuentra en la cuarta posición. Esto sugiere que a pesar del impacto claro que tiene la formación académica de los puestos altos de una startup sobre el éxito de esta, no es una condición imprescindible para que la compañía sea exitosa.

7. REFERENCIAS

[1] Chae, T. (Mayo 2019). How Do VCs Evaluate Early Stage Startups Versus Later Stage Ones? https://www.forbes.com/sites/quora/2019/05/01/how-do-vcs-evaluate-early-stage-startups-versus-later-stage-ones/?sh=62a9b1793b3e

[2] Foy, P. (Abril 2020). Applications of AI and Machine Learning in Venture Capital. https://www.mlq.ai/ai-machine-learning-venture-capital/

[3] Corea, F. (Mayo 2019). Data-driven VCs. https://francesco-ai.medium.com/data-driven-vcs-839f2454d22

[4] Nunes, J. (Septiembre 2022). Venture Capital 2.0 — the revolution of Machine Learning & Data-Driven VC. https://medium.com/included-vc/venture-capital-2-0-the-revolution-of-machine-learning-data-driven-vc-5ecd62b76fb

PREDICTING THE SUCCESS OF VENTURE CAPITAL INVESTMENT IN A STARTUP USING MACHINE LEARNING ALGORITHMS

Author: Sala López, Armando.
Supervisor: Sanz Bobi, Miguel Ángel.

ABSTRACT

This project focuses on developing a prediction model for the success of investments in startups. It utilizes a cluster-based approach and analyses each cluster individually. The results demonstrate an effective, reliable, and useful model for making informed decisions in the venture capital and entrepreneurship field.

KEYWORDS: Venture Capital, Machine Learning, Clustering, Startup.

1. INTRODUCTION

Venture capital involves long-term investment by specialized companies in the management of financial resources, in minority and temporary stakes, in small and medium-sized enterprises with high growth and/or profitability prospects.

The venture capital industry experienced significant growth during the first decade of the 21st century. However, the 2008 financial crisis resulted in major corrections in the global economy and, particularly, in the venture capital industry. Nevertheless, the global economic recovery and intensified technological development in the mid-2010s sparked a strong movement in the startup ecosystem, leading to another promising decade for venture capital. With the conclusion of the decade, the coronavirus pandemic created significant uncertainty in the global economy, impacting both venture capital and the

startup ecosystem, much like other sectors. Currently, following a record-breaking year of investment in the venture capital industry in 2021, it appears to continue with a positive trend. This trend is reflected by strong investments, albeit below the figures of 2021. However, in the world's most robust venture capital market, the United States, if fund raising had followed the trend of the first three quarters of the year, 2022 could have become the year with the highest capital raised in history. This shows the strong market and the potential of the industry.

2. STATE OF THE ART

Investing in startups is the great challenge that venture capital funds face. There are different stages at which a startup seeking funding from a venture capital fund can be found. The younger the company, the fewer metrics exist to calculate the potential return of the startup. Therefore, in these cases, much of the analysis relies on the subjective study of the team that forms the company, their professional background, and the chemistry between them [1]. There are other equally important points of analysis, such as the market to which the company belongs or the product, but the difficulty of developing reliable metrics at these stages leads to a higher subjective component, and historically, there has not been a great need for the use of tools like machine learning.

However, the use of artificial intelligence tools has experienced significant evolution in recent years [2]. Currently, some of the world's leading funds use in-house developed algorithms or algorithms created by companies dedicated exclusively to this sector [3]. There are different models applied in the industry, from algorithms dedicated to due diligence and company analysis to algorithms that discover potential startups for investment [4].

3. OBJECTIVES

The main objective of this thesis is to develop a machine learning model that can be useful for venture capital funds in analysing investment opportunities. The tool will enable funds to obtain more relevant information to estimate the potential success of investing in a

startup. Additionally, founders could benefit from such a model to make decisions that optimize their company's development.

Alternatively, an analysis of this kind allows identifying market trends and their evolution. Therefore, another objective of the work is to identify trends in the market, which, in this case, would encompass movements in the startup ecosystem up until 2013. Specifically, the analysis will focus on the impact of the 2007-2008 financial crisis on the development and valuations of startups. Additionally, studying the success of an investment based on the type of exit (IPO vs. private market acquisition) is of interest.

Lastly, it will be valuable to understand the impact of the educational background of startup founders and the relevance of their qualifications on the success of the companies. In particular, the study concentrates on the influence of the reputation of the academic institution where founders pursued their studies on the startup's success.

4. METHODOLOGY

4.1 DATA PREPROCESSING

The data preprocessing played a significant role in achieving the previously defined objectives. The information source used in this thesis consisted of 14 different databases, which contained a high level of missing or invalid data that had to be modified or eliminated based on their characteristics. After individually processing all the tables, a total of six tables were obtained, which form part of the exploratory analysis referred to as the "Processed Database Set" throughout the exploratory analysis.

The final database is a combination of the tables from the "Processed Database Set" that ensures the best balance between the number of cases (766 rows) and variables (25 columns). It contains information related to the startup, such as the sector and exit method. Additionally, it includes information about the funding rounds, such as the year of the first and last investment, the number of rounds, and the total raised capital, among other variables. Furthermore, it contains data about the offices and the educational background of the founders. Finally, the final database includes information about the relationships

that startups have with other members of the entrepreneurial ecosystem and participants in the fundraising processes.

	category_code	status	invested_companies	first_funding_at	last_funding_at	funding_rounds	funding_total_usd	relationships	region	country_code	...	funds	people	companies
0	web	acquired	0	2008	2008	1	5000000.0	14	sf bay	USA	...	1	2	0
1	games_video	acquired	0	2007	2007	1	3000000.0	6	sf bay	USA	...	2	4	0
2	photo_video	acquired	0	2007	2013	3	15286415.0	2	nashville	USA	...	2	0	0
3	hardware	acquired	0	2011	2011	1	28000000.0	7	sf bay	USA	...	2	8	0
4	enterprise	acquired	0	2009	2012	5	142000000.0	38	sf bay	USA	...	21	5	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
761	biotech	ipo	0	2010	2013	4	128614965.0	6	bedminster	USA	...	5	0	0
762	biotech	ipo	0	2011	2012	2	35400000.0	17	sacramento	USA	...	8	0	0
763	manufacturing	ipo	0	2012	2012	1	20000000.0	6	santa barbara	USA	...	1	0	0
764	messaging	ipo	0	2007	2012	5	44000000.0	32	sf bay	USA	...	13	0	3
765	security	ipo	0	2006	2012	5	328600000.0	35	sf bay	USA	...	12	0	0

Figure 1. Fragment of the Final Database.

4.2 EXPLORATORY DATA ANALYSIS

The exploratory analysis is divided into two stages. Firstly, the databases belonging to the "Processed Database Set" are analyzed, allowing for a better understanding of the startup ecosystem. This initial analysis of the database is crucial because as more content is added to the database, the number of complete cases decreases, and therefore, the overall picture of the ecosystem is lost.

Once the overall ecosystem has been analyzed, gaining a greater knowledge of the sector and the variables comprising the database, the focus shifts to the final database, which will be used in the models. The analysis of the final database helps identify trends and characteristics, leading to better decision-making and model design. Moreover, this analysis is designed to confirm that the final database, which is essentially a sample of the data seen in the six processed databases, follows the trends of the ecosystem identified in the sector study. Table 1 presents the results obtained for both databases, showcasing the similarities in the variable representing the region where the office is located.

Table 1. Comparison of the Most Common Regions in the Ecosystem and the Final Database.

Región	Ranking en Ecosistema	Ranking en Base de Datos Final
SF Bay	1	1
New York	2	3
London	3	7
Los Angeles	4	4
Boston	5	2
Seattle	8	5

4.3 CLUSTERING AND DISCARDED MODELS

Regarding clustering and discarded models, the model development starts with a clustering method designed to gain in-depth knowledge about the different data groups and add a variable that enhances the model's accuracy. For this purpose, the processed final database is created by applying normalization to the numerical variables and OneHotEncoding to the categorical variables. After splitting the processed final database into training and testing sets, a clustering algorithm is designed to determine the optimal number of 6 clusters.

The cluster labels are added as a variable to the processed final database, and the following models are tested:

- Linear Regression
- Decision Trees
- Random Forest
- GradientBoostingRegressor
- AdaBoostRegressor
- XGBRegressor
- SVR
- Neural Networks

However, all these models are discarded in favour of the final model.

4.4 FINAL MODEL

The first step in designing the final model, as done for the discarded models, is to remove outliers using the z-score method. This method removes a single entry where the target variable had a value of $994,911,036,000. Subsequently, preprocessing is applied, where numeric variables are normalized, and categorical variables are encoded. Finally, the data is split into the training set and the test set.

Secondly, an analysis is performed to identify the optimal number of clusters. The chosen method to determine the number of clusters is KMeans. In this case, the ideal number of clusters is six.

After analysing the new groups formed by KMeans, all the previously mentioned different models are applied to each cluster. However, neural networks are not applied to the clusters. This is because, given the low number of cases in each group and their poor performance with the complete database, it does not make sense to implement this model in this case.

For each group, the model that best fits and generates the best results is selected, applying hyperparameter tuning to maximize the performance of the final model. The obtained models for each group are as follows:

- Cluster 0: Random forest.
- Cluster 1: AdaBoost.
- Cluster 2: Gradient Boosting.
- Cluster 3: SVR.
- Cluster 4: Random Forest.
- Cluster 5: AdaBoost.

After obtaining the optimal models for each cluster, the clustering algorithm is combined with the models using ensemble methods to obtain the complete final model.

5. RESULTS ANALYSIS

The final model represents a slight improvement in terms of the metrics used compared to the other developed models, including neural networks, and boosting algorithms, among others. Furthermore, unlike some of the discarded models, the error metrics used to analyse the model do not vary after each execution. This generates a much more reliable and useful model for accurately predicting the success of investment in startups. The fact that the results obtained from the final model are the most powerful demonstrates the advantage of subdividing the dataset into clusters, where patterns and relationships between the data are much clearer and lead to better predictions. In fact, although the

subdivision into clusters results in smaller datasets, where extracting reliable relationships and patterns is more precise, this model still performs better than the other models.

```
RMSE: 1.2935939478679512
Average Error ($): 2.6458660907546063
MAPE: 5.391202549824616
Average MAPE (%): 218.46714710822704
R^2: 0.3595023411703795
```

Figure 2. Results of the Final Model.

The final model has a mean error of $2.65. This result is truly remarkable considering that the target variable contains values ranging from $30,000 to $104,000,000,000. Therefore, it can be stated that the model is capable of predicting the success of investment with an acceptable level of accuracy.

6. CONCLUSIONS

Firstly, after the development and subsequent analysis of the models and their results, it can be concluded that the final model is effective and fulfils the main objective of providing a model that allows venture capital funds and entrepreneurs to make more informed decisions. Obtaining a final model with an average error of only $2.5 shows a high degree of reliability and serves as a useful tool for analysing investment opportunities for funds. Additionally, the nature of investments in startups has a high variance component, where extreme cases and large successes are common. Therefore, a slightly higher MAPE metric can be assumed as these cases are relatively rare over time. Furthermore, the previous exploratory analysis and the results obtained from other models also provide really useful information to entrepreneurs, which can be used as an additional source of information.

Finally, all the objectives related to identifying relative importance have been achieved using cluster analysis. By combining the information about the obtained clusters and their mean values of the target variable per cluster, relevant conclusions can be drawn about the variables under study.

```
Cluster 3: Median = 352000000.0, Mean = 2526095394.2720337
Cluster 4: Median = 200000000.0, Mean = 1112472978.7234042
Cluster 2: Median = 102500000.0, Mean = 214875666.4057971
Cluster 0: Median = 100000000.0, Mean = 485432790.6976744
Cluster 1: Median = 89000000.0, Mean = 273467411.7816092
Cluster 5: Median = 78400000.0, Mean = 408730647.9166667
```

Figure 3. Mean and Median Values of Exit Price by Cluster.

Firstly, regarding the exit method, the only cluster where all operations are IPOs is cluster 3, which, as previously mentioned, has the highest values in the target variable. Additionally, it is cluster 4, which is the second highest in terms of valuations, that is the next group with the best IPOs-to-Acquisitions ratio. On the other hand, it can be observed that cluster 1, which has the lowest average sale/IPO values, is the only group where all operations are acquisitions. It should also be mentioned how the few IPO operations in cluster 5 make the mean value much higher than the median. These facts demonstrate a clear relationship between the type of exit and the company's valuation, where investors participating in companies whose exit is an IPO tend to have stakes in a higher valued company.

Secondly, there is some relationship between the timeline of startup progress, first through their funding rounds, and later through the sale or IPO, and the investment success. The financial crisis that occurred between 2007 and 2008 significantly affected the global economy and, with it, the venture capital sector. If the dates of the startup operations are analyzed, certain patterns can be observed that show the effect of the market and the economic situation on the ecosystem and the success of startups. For example, the two clusters with the worst sale/IPO figures, 1 and 5, are those that raised their funding rounds between 2006-2008. This may be because during the crisis, valuations suffered, leading to tighter funding rounds and less capital to operate. Conversely, the companies that closed their capital raising processes later had the best valuations. This pattern is accentuated when analysing the values for the cluster with the highest exits. The companies in cluster 3, on average, completed their last funding round in 2011 when the global economy was experiencing a strong recovery after the crisis. Therefore, these trends demonstrate a clear relationship between the economic cycle and

the success a startup and its investors may have in their exit process. They also highlight the significant impact of the financial crisis on the startup ecosystem.

Finally, in Figure 7, a less evident relationship is shown between the university and the company's success, although not as strong as the relationship with the type of exit. Cluster 5, which has fewer founders/CEOs from prestigious universities, is among the categories with the lowest valuations. On the other hand, categories 4 and 3 have the highest valuations and are distinguished by having a high percentage of founders/CEOs from the top-ranked institutions (Top 100 in the world).

However, it is important to mention that cluster 1, which has the second-lowest valuation in terms of median, contains a large number of founders from the world's top universities. Additionally, cluster 2, which ranks third in terms of the median value of the target variable, does not have any founders who attended university. The same is true for cluster 0, which ranks fourth. This suggests that despite the clear impact of the educational background of key positions in a startup on its success, it is not a necessary condition for the company to be successful.

7. REFERENCES

[1] Chae, T. (Mayo 2019). How Do VCs Evaluate Early Stage Startups Versus Later Stage Ones? https://www.forbes.com/sites/quora/2019/05/01/how-do-vcs-evaluate-early-stage-startups-versus-later-stage-ones/?sh=62a9b1793b3e

[2] Foy, P. (Abril 2020). Applications of AI and Machine Learning in Venture Capital. https://www.mlq.ai/ai-machine-learning-venture-capital/

[3] Corea, F. (Mayo 2019). Data-driven VCs. https://francesco-ai.medium.com/data-driven-vcs-839f2454d22

[4] Nunes, J. (Septiembre 2022). Venture Capital 2.0 — the revolution of Machine Learning & Data-Driven VC. https://medium.com/included-vc/venture-capital-2-0-the-revolution-of-machine-learning-data-driven-vc-5ecd62b76fb

Índice de la Memoria

Índice de Figuras

Índice de Tablas

CAPÍTULO 1. INTRODUCCIÓN

El capital riesgo, también conocido como *venture capital*, consiste en la inversión a largo plazo por parte de compañías especializadas en la gestión de recursos financieros, de forma minoritaria y temporal, en pequeñas y medianas empresas, con grandes perspectivas de crecimiento y/o rentabilidad.

El origen del capital riesgo se remonta a antes de la segunda guerra mundial. Las familias más ricas del nuevo continente, entre las cuales se encontraban la familia Rockefeller o la familia Vanderbilt, utilizaban sus grandes fortunas para financiar proyectos empresariales privados. No obstante, el primer fondo moderno de capital de riesgo, también denominado *venture capital*, es fundado en 1946 por el presidente del MIT Karl Compton, el presidente del Banco de Boston (organismo perteneciente a la Reserva Federal Americana) Ralph Flanders, el presidente del fondo Massachusetts Inverstors Trust y el profesor de Harvard George F. Doriot, al que más adelante se apodaría como el padre del *venture capital*. El fondo, llamado ARDC, levantó[6] capital de distintas fuentes como universidades o compañías de seguros con el objetivo de invertir en compañías que desarrollaron tecnologías innovadoras durante la segunda guerra mundial.

El éxito del fondo atrajo gran atención generando, durante las siguientes décadas, un proceso de fuerte crecimiento del sector, destacando especialmente las correspondientes a los años setenta y ochenta, donde varios fondos realizaron operaciones de éxito muy

[6] En la terminología asociada a esta industria, *levantar capital* se entiende como la captación de fondos y recursos económicos de terceros para su aportación al proyecto

lucrativas. Además, durante aquellas décadas, se fundaron algunas de las compañías actualmente con mayor capitalización bursátil, como Apple, Microsoft o Fedex. Todas ellas recibieron apoyo de firmas de capital de riesgo que, más tarde, se verían recompensadas por su apuesta en lo que en su momento eran grandes ideas y que luego se convirtieron en líderes de sus mercados.

Sin embargo, es durante la burbuja de las Dot-Com[7] cuando la industria del capital de riesgo realmente explota, acabando la década de los noventa con más de 700 fondos activos y 143 billones de dólares bajo gestión. Tras la explosión de la burbuja aparecen nuevas firmas de *venture capital* con un enfoque más especializado, centrando su inversión en un número reducido de sectores. Además, acompañando a sus nuevos criterios de inversión, los fondos empiezan a implementar nuevas formas en las que ayudar a las *startups*. El rol más activo de los inversores ayuda a las *startups* en las tareas de captación de talento, desarrollo de negocio o marketing.

La industria del *venture capital* continuó su fuerte crecimiento durante la primera década del siglo XXI. No obstante, la crisis financiera de 2008 provocó fuertes correcciones en la economía mundial y, en particular, en la industria del *venture capital*. Sin embargo, la recuperación económica global y la intensificación del desarrollo tecnológico a mediados de la segunda década del siglo provocó un fuerte movimiento en el ecosistema *startup* que desencadenó otra muy buena década para el capital riesgo. Con la conclusión de la década, la pandemia del coronavirus generó mucha inseguridad en la economía global, afectando al *venture capital* y al ecosistema *startup* de igual forma que al resto de

[7] La burbuja de las punto-com (Dot-Com en inglés) se refiere al periodo comprendido entre 1997 y 2000. Durante dicho periodo se produjo un fuerte crecimiento de los valores económicos de las empresas relacionadas con Internet.

sectores. En la actualidad, tras un 2021 donde se batieron récords de inversión en la industria del *venture capital*, ésta parece que continua con una tendencia positiva. Dicha tendencia viene reflejada por fuertes inversiones, aunque muy por debajo de las cifras de 2021. No obstante, en el mercado más potente del capital riesgo del mundo, Estados Unidos, si la captación de fondos hubiese seguido la tendencia de los 3 primeros trimestres del año, el 2022 se podría haber convertido en el año que más capital se ha levantado en la historia.

1.1 ESTADO DE LA CUESTIÓN

La inversión en *startups* es el gran reto al que los fondos de *venture capital* se enfrentan. Existen distintas etapas en las que se puede encontrar una *startup* que busca financiación de un fondo de *Venture Capital*. Cuanto más joven sea la empresa, menos métricas existen para poder calcular un potencial retorno de la *startup*. Por lo tanto, en estos casos, gran parte del análisis recae sobre el estudio, con una gran componente subjetiva, del equipo que forma la compañía, el *background*[8] de los miembros de este y la química entre ellos [1]. Existen otros puntos de análisis igual de importantes, como el mercado al que pertenece la compañía o el producto, pero la dificultad que supone elaborar métricas fiables en estas etapas da lugar a un mayor componente subjetivo y, por lo tanto, históricamente, no ha existido una gran necesidad del uso de herramientas como el *machine learning*.

No obstante, el uso de herramientas de inteligencia artificial ha experimentado una fuerte evolución durante los últimos años [2]. En la actualidad, algunos de los principales fondos

[8] El background es el conjunto de conocimientos y experiencias que constituyen el bagaje de una persona.

del mundo utilizan algoritmos desarrollados *in-house*[9] o elaborados por compañías que se dedican exclusivamente a este sector [3]. Existen distintos modelos que se aplican en la industria, desde algoritmos dedicados a la *due diligence*[10] y el análisis de compañías hasta algoritmos que descubren potenciales *startups* para invertir [4].

La inteligencia artificial ha demostrado ser una herramienta imprescindible en la mayoría de los sectores. Esta realidad también se aplica al mundo del *venture capital*. Gracias a los modelos elaborados, los inversores pueden tomar decisiones más informadas y mejorar sus inversiones, que a su vez impactan positivamente en la rentabilidad recibida por el fondo y sus socios capitalistas [5]. La inteligencia artificial permite analizar los datos de manera objetiva, eliminando cualquier sesgo que un inversor pueda tener en base a su experiencia y que pueda nublar su objetividad [6].

Para entender de forma más elaborada como plantear el problema y el modelo, se utiliza el articulo redactado por Verónica Wu para McKinsey [7]. En el artículo se explican las fuentes de información y el proceso que se ha seguido para la construcción del algoritmo.

Por otro lado, para conocer en detalle la estructura mediante la que se implementan los modelos de *machine learning* en el sector de *venture capital*, el proyecto se basa en el artículo de Francisco Corea en la revista Forbes [8]. Al enumerar los pasos que siguen los

[9] In-house se refiere a una actividad u operación que se realiza dentro de una empresa, en lugar de depender de la subcontratación.
[10] Due diligence, o diligencia debida en español, es la operación previa de revisión de una persona o empresa de cara a evaluar posibles riesgos al establecer una relación económica con ella

profesionales de la industria, se puede utilizar el artículo como guía para seguir un procedimiento similar.

Finalmente, para comprender en todo su calado la profundidad que puede adquirir un proyecto de esta materia, se han analizado dos trabajos de investigación sobre el asunto. En el primero, de Thomas Rory, combina la analítica computacional con el *venture capital* para mejorar la toma de decisiones [9]. Al ser un trabajo extenso, se puede utilizar para ver la estructura general y el enfoque del estudio con el objeto de poder desarrollar un mejor trabajo de investigación. Por otro lado, el trabajo titulado *"Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments"*, [10] es una versión breve, similar al que se realizará en este trabajo de fin de grado. En este caso, se utilizan datos de la misma fuente, pero engloba distintos periodos temporales. Gracias a este trabajo se podrán comparar las metodologías que implementan y los análisis que se realizan. En consecuencia, se podrá entender el enfoque que expertos en el campo del análisis de datos le dan a un trabajo de estas características.

1.2 MOTIVACIÓN DEL PROYECTO

Como en la gran mayoría de sectores, el análisis de datos de forma masiva está empezando a ser una herramienta fundamental para la toma de decisiones. No obstante, en el mundo de la inversión aún existe cierto rechazo hacia este tipo de tecnologías. Esto se debe a que, en ocasiones, resulta complicado medir y procesar ciertas variables cualitativas que son fundamentales para el análisis de una potencial inversión. Por lo tanto, la motivación detrás de este Trabajo de Fin de Grado (TFG) es el análisis y la posibilidad, en su caso, de la elaboración de un modelo basado en técnicas de *machine learning* que puedan resultar útiles para los inversores, funcionando como fuente alternativa de información para el sector del *venture capital*.

1.3 OBJETIVOS

El principal objetivo del trabajo es llegar a plantear un modelo basado en algoritmos de *machine learning* que pueda ser útil para los fondos de capital riesgo a la hora de analizar oportunidades de inversión. La herramienta permitirá a los fondos obtener mayor cantidad de información relevante para estimar el potencial éxito de la inversión en una *startup*.

Por otro lado, igual que a las firmas de inversión les resulta interesante disponer de un modelo que proporcione información de calidad y que les asista en la búsqueda de potenciales empresas para la inversión, los *founders*[11] podrían aprovecharse de un modelo como éste para poder tomar decisiones que les permitan desarrollar su empresa de la manera óptima.

Es cierto que un análisis de este tipo permite identificar tendencias en el mercado y su evolución, por lo que otro de los objetivos del trabajo será identificar tendencias en el mercado que, en este caso, dada la base de datos, englobaría movimientos en el ecosistema *startup* hasta 2013. . En particular, centraremos el análisis en el efecto que tuvo la crisis financiera de 2007-2008 en el desarrollo de las startups y sus valoraciones. Además, resulta interesante el estudio del éxito de una inversión en función del tipo de exit[12] (IPO vs Adquisición en el mercado privado).

[11] Se entiende por *founders* aquellas personas físicas o jurídicas que constituyeron la *startup.*
[12] El exit se refiere al momento en el que los inversores o fundadores de la empresa obtienen un retorno de su inversión inicial al vender total o parcialmente sus participaciones en la compañía.

Finalmente, resultará interesante conocer el impacto que tienen los estudios que tienen los fundadores de las startups y la relevancia que tiene ese nivel de cualificación sobre el éxito de las empresas. En particular, el estudio se concentra en el impacto que tiene el prestigio de la institución académica, donde los fundadores cursan los estudios, en el éxito de la startup.

1.4 METODOLOGÍA

El desarrollo del TFG se divide en varias etapas.

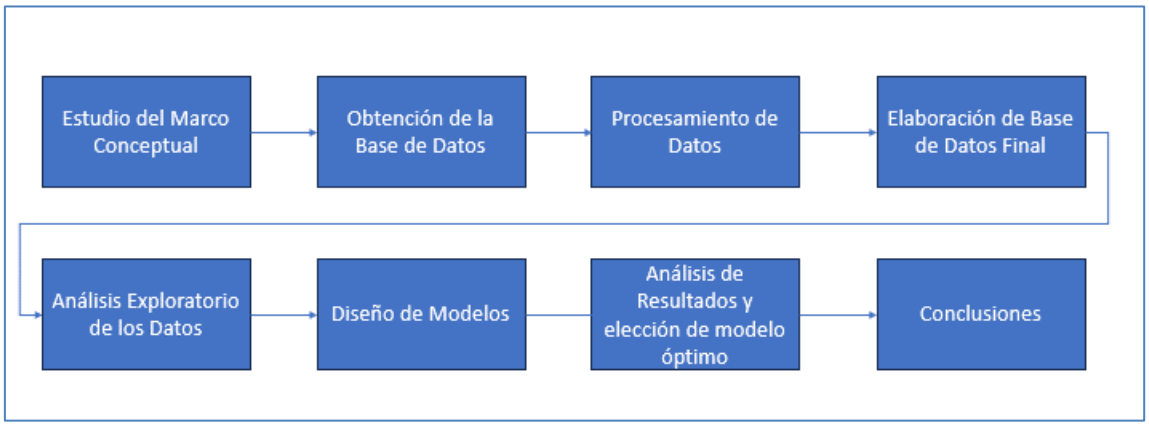


Figura 2. Diagrama de bloques mostrando relación entre las distintas etapas del TFG.

En primer lugar, tras el estudio del marco conceptual y la elaboración de los objetivos , se realizó un análisis exploratorio de los datos. Para ello primero se procesaron los datos para el análisis eliminando variables no relevantes y combinando las distintas bases de datos, entre otras cuestiones. El objetivo de esta limpieza de datos fue la simplificación de la base de datos, obteniendo sólo las variables que permitían un análisis preciso del problema presentado. Cuando se eliminaron los datos irrelevantes se combinaron las distintas bases de datos para juntando todas las variables que se estudiaron más adelante. Una vez se construyó la base de datos, se realizó un análisis exploratorio de los datos,

obteniendo las primeras observaciones de los datos a estudiar y dando una visión inicial sobre el problema a resolver.

En segundo lugar, se aplicaron los modelos de *machine learning* a los datos obtenidos. Para ello, se dividió la base de datos en dos, identificando la parte que dedicada al desarrollo del modelo (*train data*) y la elaborada para probar la fiabilidad del modelo con datos desconocidos y por lo tanto su aplicación para futuros casos (*test data*). El proceso comenzó utilizando los modelos más sencillos como Regresión Lineal para más tarde finalizar con modelos más complejos como Redes Neuronales. Durante este proceso se documentaron los resultados obtenidos. Por lo general, a medida que se aumentó la complejidad del modelo, se obtuvieron resultados de mayor fiabilidad eliminando sesgos y errores. Por lo tanto, seguir una estructura como la descrita permitió la explicación lineal y lógica del proceso. Para comparar los distintos modelos se utilizaron métricas de predicción como RMSE y MAPE. Para la comparativa de modelos se tuvieron en cuenta los resultados el *test*, pues al ser datos que los modelos no han procesado, son los proporcionan información relevante.

Más adelante, una vez se desarrollaron los distintos modelos, se escogió el mejor modelo y se realizó un examen más en detalle explicando las conclusiones que se pudieron extraer de él. Además, tras haber documentado el modelo óptimo, se escribieron las conclusiones obtenidas del estudio.

Finalmente, se realizó un repaso general del mismo, incluyendo el formato, la bibliografía y su contenido.

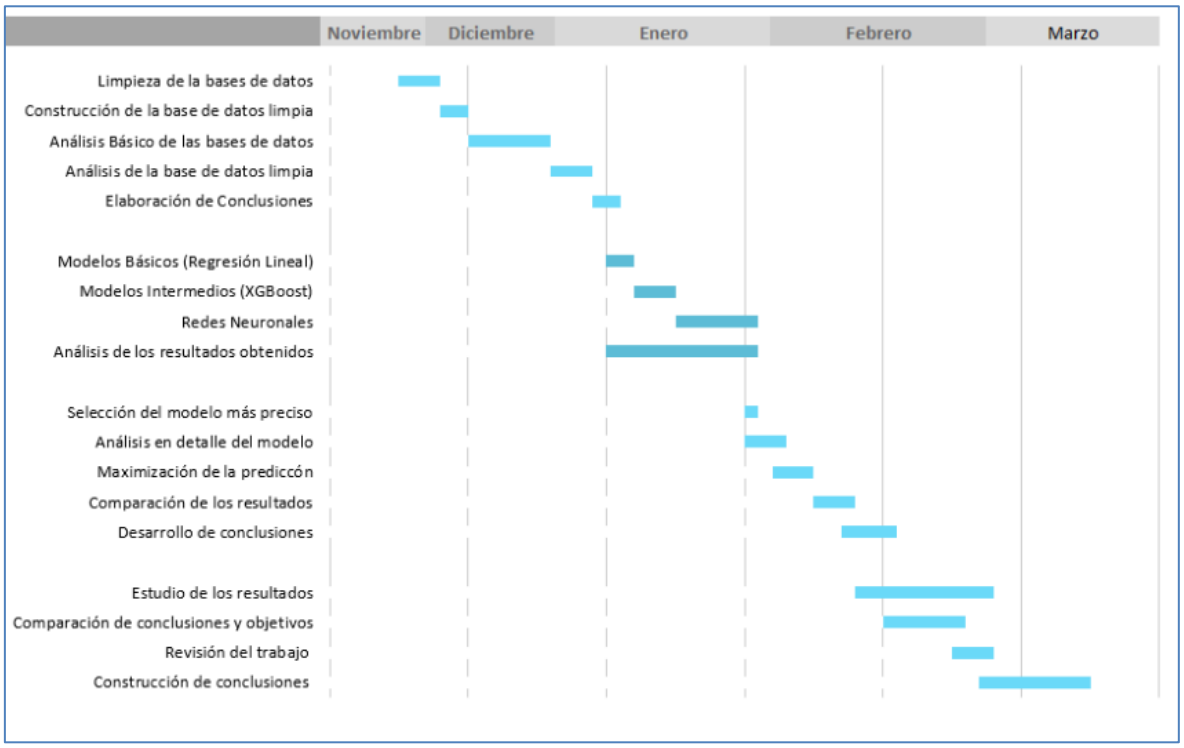


Figura 3. Cronograma del TFG utilizando GANTT

1.5 RECURSOS EMPLEADOS

Los recursos empleados consisten en una base de datos llamada Sartup Investments extraída de Kaggle [11]. Con respecto a la base de datos, es un extracto de la base de datos de *Crunchbase*, que comprende información relevante al mundo emprendedor hasta diciembre del año 2013. La herramienta Visual Studio (v. 1.79.1) se utilizó para programar, en lenguaje Python (v. 3.11.0), los modelos de *machine learning*. Para poder desarrollar todo el análisis y la implementación de los algoritmos se utilizan los siguientes paquetes:

La librería 'Pandas' proporciona estructuras de datos llamadas DataFrames que permiten manipular, limpiar y transformar datos de manera sencilla. También permite la importación de las bases de datos, que se encuentran en formato CSV.

En segundo lugar, la librería 'Numpy' contiene un conjunto de funciones y estructuras de datos para realizar operaciones numéricas.

En tercer lugar, para poder visualizar todo el contenido de una forma estructurada se utiliza el paquete 'matplotlib'. La librería proporciona una amplia variedad de herramientas para crear gráficos, diagramas y visualizaciones personalizables.

En cuanto al diseño de modelos se implementan dos librerías principales. Por un lado, 'sklearn' se utiliza para implementar las funciones de preprocesamiento de datos y aplicar las métricas que permiten la evaluación de modelos. Además, exceptuando las redes neuronales, todos los modelos de regresión y clustering se encuentran prediseñados en el paquete, lo que permite su fácil implementación.

Por otro lado, 'TensorFlow' y 'keras' son una biblioteca y un interfaz que simplifican la creación y entrenamiento de redes neuronales. Al importar tensorflow.keras, puedes aprovechar las funcionalidades de TensorFlow y utilizar Keras para diseñar y entrenar redes neuronales de manera eficiente.

Finalmente, destacar que también se utilizan otros paquetes de manera excepcional, como 'xgboost' para importar el modelo prediseñado de xgboost y 'skopt' para poder utilizar funciones relacionadas con el ajuste de los modelos.

La base de datos se compone de once documentos de Excel que contienen la siguiente información:

- Adquisiciones: Contiene información de las adquisiciones de *startups* por parte de otras empresas.

- Grados: Contiene información con respecto a la formación académica de los individuos involucrados en el mundo de las *startups* (inversores de *venture capital* incluidos y *business angels*[13]).
- Rondas de Financiación: Contiene información acerca de las rondas de financiación.
- Fondos: Contiene datos de los distintos fondos de *venture capital*.
- Inversiones: Contiene información con respecto a las distintas inversiones realizadas por las empresas de *capital riesgo*.
- IPOs: Contiene información acerca de las IPOs realizadas por algunas *startups*.
- Milestones: Contiene información acerca de distintos eventos en el ecosistema emprendedor.
- Objetos: Base de datos principal que contiene información base (incluye datos de personas, empresas y fondos involucrados en el sector).
- Oficinas: Contiene información sobre las oficinas de las *startups* y los fondos.
- Personas: Contiene información acerca de los individuos del mundo emprendedor.
- Relaciones: Contiene las relaciones entre las compañías y los individuos, incluyendo sus roles en la relación.

[13] Los ángeles inversores, o ángeles de negocios, son personas que invierten su dinero en la fase inicial de compañías emergentes a cambio de una participación en capital. Habitualmente, ejercen también un rol de mentor y ofrecen su consejo y experiencia a los emprendedores. [12]

CAPÍTULO 2. ANÁLISIS EXPLORATORIO DE LOS DATOS

El análisis exploratorio de los datos es una etapa fundamental en cualquier estudio de estas características. Permite descubrir relaciones y patrones en los datos, además de ayudar a identificar cualquier problema o anomalía en los mismos. Asimismo, facilita la elaboración de hipótesis y de objetivos. Finalmente, el análisis exploratorio permitirá la toma de decisiones informadas a la hora de diseñar el modelo final que sirva para predecir el éxito de la inversión de un fondo de *venture capital* en una *startup*.

2.1 PREPROCESADO Y LIMPIEZA DE DATOS

La limpieza de datos es crítica para su análisis . El preprocesado y la limpieza permite la identificación y posterior eliminación o modificación de datos redundantes y erróneos. Además, el preprocesado incluye la transformación de los datos en un formato adecuado y homogéneo, resultando en un análisis más preciso. Por lo tanto, el preprocesado y la limpieza de datos son fundamentales para garantizar la fiabilidad de los resultados obtenidos y para asegurarse de que los datos se utilicen de manera efectiva.

El preprocesado de datos ha tenido un gran peso en la consecución de los objetivos planteados previamente. La fuente de información utilizada en este TFG , compuesta a su vez por 14 bases de datos distintas, contenía un alto nivel de datos nulos o no validos que han tenido que ser modificados o eliminados según sus características. Para explicar de forma más clara como se elaboró la base de datos final se analizará el proceso realizado

a cada base de datos por separado. Posteriormente, se explicarán los contenidos de la base de datos final.

2.1.1 PREPROCESADO DE PERSONAS Y GRADOS

La base de datos de personas incluye dos columnas que representan un ID. La columna "id" solo hace referencia a la propia tabla, mientras que "objects_id" es una variable global que se utiliza en todas las bases de datos. Por otro lado, se tiene el lugar de nacimiento y la afiliación de la persona. Finalmente, el nombre y el apellido de la persona son los últimos datos incluidos.

Tabla 1. Base de datos de Personas Original

	id	object_id	first_name	last_name	birthplace	affiliation_name
0	1	p:2	Ben	Elowitz	NaN	Blue Nile
1	2	p:3	Kevin	Flaherty	NaN	Wetpaint
2	3	p:4	Raju	Vegesna	NaN	Zoho
3	4	p:5	Ian	Wenig	NaN	Zoho
4	5	p:6	Kevin	Rose	Redding, CA	i/o Ventures
...	...	...	...	...	...	...
226704	226705	p:268589	John	Pins	NaN	Unaffiliated
226705	226706	p:268590	David	Schulhof	NaN	Unaffiliated
226706	226707	p:268592	Matthew	D. Rosen	NaN	Unaffiliated
226707	226708	p:268593	Gordon	Hutchins	NaN	Unaffiliated
226708	226709	p:268597	Denise	Basow	NaN	Unaffiliated

226709 rows × 6 columns

La limpieza de la base de datos ha consistido en la eliminación de las columnas de "id", "birthplace" & "affiliation_name", para posteriormente eliminar las filas vacías. Además, como el nombre de la persona realmente no tendrá ningún valor de cara al modelo

predictivo, se crea una nueva columna llamada "name" donde se juntan las columnas de "first_name" y "last_name", que tampoco tendrá peso en el modelo final (se mantiene por si resultase útil en algún gráfico).

Tabla 2. Tabla Resultante Tras el Procesamiento de la base de datos de Personas

	object_id	name
0	p:2	Ben Elowitz
1	p:3	Kevin Flaherty
2	p:4	Raju Vegesna
3	p:5	Ian Wenig
4	p:6	Kevin Rose
...	...	...
226704	p:268589	John Pins
226705	p:268590	David Schulhof
226706	p:268592	Matthew D. Rosen
226707	p:268593	Gordon Hutchins
226708	p:268597	Denise Basow

226704 rows × 2 columns

Por otro lado, la base de datos de grados sí contiene información más interesante de cara al modelo. Esto implica que la preparación de la base de datos será más elaborada.

Primeramente, se eliminan las columnas "id", "created_at" y "updated_at" que simplemente tienen información relevante al ID del caso (dentro de la tabla) y la creación y la última actualización correspondiente a la fila, pues la base de datos de Crunchbase

está en constante actualización. Además, se elimina la variable "graduation_date" por su alto porcentaje de nulos. Esta columna contenía la fecha de graduación para cada caso.

Tabla 3. Base de datos de Grados sin modificar

id	object_id	degree_type	subject	institution	graduated_at	created_at	updated_at
1	p:6117	MBA	MBA	Santa Clara University	NaN	19/02/2008 03:17	19/02/2008 03:17
2	p:6136	BA	English, French	Washington University, St. Louis	01/01/1990	19/02/2008 17:58	25/02/2008 00:23
3	p:6136	MS	Mass Communication	Boston University	01/01/1992	19/02/2008 17:58	25/02/2008 00:23
4	p:6005	MS	Internet Technology	University of Greenwich	01/01/2006	19/02/2008 23:40	25/02/2008 00:23
5	p:5832	BCS	Computer Science, Psychology	Rice University	NaN	20/02/2008 05:28	20/02/2008 05:28
...	...	...	...	...	...	...	...
109606	p:268589	CPA	NaN	American Institute of Certtified Public Accoun...	NaN	12/12/2013 14:28	12/12/2013 14:28
109607	p:268527	MS & BS	Engineering	South China University of Technology	NaN	12/12/2013 14:31	12/12/2013 14:31
109608	p:268527	PhD	Engineering	Clarkson University	NaN	12/12/2013 14:31	12/12/2013 14:31
109609	p:268528	B.S.	Electrical Engineering	Colorado State University	NaN	12/12/2013 14:38	12/12/2013 14:38

En segundo lugar, se estudia la cantidad de filas nulas en las columnas restantes. La columna referente a la institución de estudio del grado solo contiene 16 valores nulos. Sin embargo, el tipo de grado (PhD, Master, Grado, Bachiller) y el enfoque (medicina, ADE, derecho) del grado tienen 11.218 y 28.309 valores nulos respectivamente. Se puede observar cómo es significativamente superior el valor de nulos en el enfoque del grado. Además, se identifica que dentro de la columna de tipo de grado se tienen valores como BBA o MBA, que en si ya explican el enfoque del grado en el propio nombre. Por lo tanto, lo que se hace es que, si la columna de "subject" está vacía se llena con el valor de la columna tipo de grado reduciendo el número de nulos a 6.146.

Por otro lado, al ser una base de datos que se actualiza a lo largo del tiempo, es posible que ciertos criterios de redacción varíen. Por ello, se normalizan las variables tipo *string* para asegurarnos la mayor homogeneidad posible. Esta estandarización consiste en la eliminación de signos de puntuación, mayúsculas (que se pasan a minúsculas) y se elimina cualquier contenido entre paréntesis.

Finalmente, en ocasiones, al anotar la institución, se anota la facultad en lugar de la universidad. Por lo tanto, para eliminar las escuelas como *Wharton School of Business* (en lugar de *University of Pennsylvania*) o *McCombs School of Business* (en caso de la Universidad de Texas en Austin) se parten aquellos registros que contengan el guion y se mantiene la primera entrada, que por lo general es el nombre de la universidad.

Finalmente, se combinan ambas tablas, para así poder asociar la base de datos al nombre de la persona. Como se observa, la base de datos resultante no mantiene las personas que no tiene registrado ningún grado a su ID. Esto se debe a que, de esta forma, la base de datos final tiene menos valores nulos y se podrán realizar mejores visualizaciones sobre ella. Además, de cara a la construcción de la base de datos final, aquellas personas que no tenga grado de ningún tipo incluirán los valores nulos en esas columnas.

Tabla 4. Tabla resultante de las bases de datos Personas y Grados procesadas

	object_id	name	degree_type	subject	institution
0	p:2	Ben Elowitz	bs	electrical engineering/computer science	university of california berkeley
1	p:2	Ben Elowitz	bs	applied mathematics	university of california berkeley
2	p:3	Kevin Flaherty	bba	bba	washington university in st louis
3	p:3	Kevin Flaherty	mba	mba	indiana university
4	p:5	Ian Wenig	degree	advanced business professional course	the aji network
...	...	...	...	...	...
109051	p:268528	Edward J. Treska	b.s.	electrical engineering	colorado state university
109052	p:268528	Edward J. Treska	j.d.	j.d.	university of san diego school of law
109053	p:268560	Drew Langloh	mba	mba	samford university
109054	p:268589	John Pins	b.s.	accounting	iowa state university
109055	p:268589	John Pins	cpa	cpa	american institute of certtified public accoun...

109056 rows × 5 columns

2.1.2 PREPROCESADO DE OFICINAS

La tabla de oficinas ha experimentado un proceso de limpieza contundente.

En primer lugar, como en todas las bases de datos, se elimina el ID de la tabla (que no el global) y las columnas que hacen referencia a la creación y actualización del dato. Asimismo, se elimina la descripción, pues contiene demasiados nulos.

En segundo lugar, se elimina el detalle que se tiene acerca de la localización de la oficina. Esto se debe a que, en el fondo, la dirección exacta, incluyendo el código ZIP y la calle, no tiene ninguna utilidad para un modelo predictivo. No obstante, si se mantiene la ciudad, la región y el código del país, además de la longitud y la latitud de cara a poder dibujar algún gráfico que requiera esos datos. Finalmente, se deshecha la variable de "state_code", pues, aunque gran parte de las empresas tienen la sede en USA, no es el caso de todas (45% Nulls).

De cara al modelo, las variables región y país son las más interesantes. Esto se debe a que, aunque tener en cuenta la ciudad sea más preciso que región, la variable "city" tiene 5.168 valores nulos y la columna "region" 0. Además, aunque no parezca lógico tener en cuenta 2 variables que representan la localización como "country_code" y "region", la realidad es que la variable país hace referencia al país de origen de la empresa y la de región representa la zona donde se encuentran las oficinas de las startups.

Posteriormente, una vez normalizadas las ciudades y las regiones, se procede a rellenar de forma óptima la base de datos. En este caso, tras analizar la base de datos, se identifican varias entradas con ciudades, pero la región vacía. Además, muchas de estas ciudades se encuentran en otras entradas en la base de datos (otras líneas). Por lo tanto, para reducir

el número de nulos al mínimo se crea un diccionario con las ciudades y su respectiva región (tomando el primer valor de la agrupación, pues se asume que es el mismo valor en todas). Después, se utiliza la función "map" para rellenar las regiones con valores vacíos que si se encuentran completas en nuestro diccionario.

Tras haber obtenido la base de datos limpia para oficinas, se crea una nueva tabla de datos de cara a la elaboración de gráficos en el EDA. En primer lugar, se elabora una base de datos donde se representan las ciudades en un mapa incluyendo burbujas o un *heatmap* para mostrar la cantidad de oficinas dentro del mundo *startup* que se encuentran en cada una de las ciudades.

Ciudades (cities):

1. Se crea una tabla con las líneas con las que tienen latitud y longitud completas (vale sólo con latitud porque siempre que una es 0 la otra también).
2. Para la nueva tabla se agrupan por ciudad calculando el valor medio de latitud y longitud para cada ciudad.
3. Se combinan junto a la tabla cities_loc (oficinas final sin nulos en ciudad) y se eliminan ciudades innecesarias.

Al trabajar con oficinas, se ha llegado a la conclusión de que existen varias oficinas asociadas a los distintos valores de la columna "object_id" (empresas). Este dato puede ser interesante pues es una buena representación del alcance y prestigio del fondo o empresa y que debería tener un impacto sobre el resultado de la inversión. Por lo tanto, valorando dicha suma como una potencial variable a utilizar en el modelo predictivo, se crea una tabla con esa cuenta para posteriormente añadírsela a la base de datos final.

2.1.3 PREPROCESADO DE FONDOS

La limpieza de la base de datos de fondos ha sido la más compleja. Para poder normalizar todas las cifras económicas (pasarlas a USD), se ha importado la base de datos de la OECD con los tipos de cambio de las principales monedas del mundo vs el dólar (medias mensuales) entre 1999-2013. La limpieza ha seguido el siguiente proceso:

1. Se importan ambas bases de datos
2. Utilizando todos los valores mensuales desde 1993 hasta 2013, se crea un valor medio para cada tipo de cambio (que se utilizará para aquellos fondos sin fecha de fundación y que por lo tanto no se le puede aplicar un valor de la base de datos).
3. Se crean dos columnas (mes y año) en cada tabla para poder combinar las bases de datos.
4. Se eliminan columnas innecesarias de la tabla de fondos y se combinan ambas bases de datos.

Tabla 5. Extracto de la Tabla con información del Tipo de Cambio (OECD)

	raised_currency_code	Value	Year	Month
0	AUD	1.574645	1999	01
1	AUD	1.564095	1999	02
2	AUD	1.586022	1999	03
3	AUD	1.559240	1999	04
4	AUD	1.513076	1999	05
...	...	...	...	...
5785	CRC	455.307385	Average	Average
5786	COP	2153.284182	Average	Average
5787	ARS	3.366973	Average	Average
5788	SAR	3.750000	Average	Average
5789	USD	1.000000	Average	Average

La tabla resultante es muy sencilla, conteniendo el nombre y el ID del fondo (sin valor analítico), la fecha de fundación (con 115 valores nulos) y la cantidad levantada (171 0s). Finalmente, se combinan el fondo con la base de datos de oficinas para obtener más detalle sobre cada fondo.

Tabla 6. Tabla Final para la base de datos de Fondos

	object_id	name	funded_at	raised_amount	region	city	country_code	Nº of Offices
0	f:371	Second Fund	2008-12-16	300000000.0	shanghai	shanghai	CHN	2.0
1	f:371	Second Fund	2008-12-16	300000000.0	sf bay	menlo park	USA	2.0
2	f:17	Sequoia Israel Fourth Fund	2008-12-17	200750000.0	sf bay	menlo park	USA	8.0
3	f:17	Sequoia Israel Fourth Fund	2008-12-17	200750000.0	shanghai	shanghai	CHN	8.0
4	f:17	Sequoia Israel Fourth Fund	2008-12-17	200750000.0	beijing	beijing	CHN	8.0
...	...	...	...	...	...	...	...	...
2900	f:7331	JANVEST Technologies LP I	2013-03-19	5.0	chicago	chicago	USA	3.0
2901	f:7331	JANVEST Technologies LP I	2013-03-19	5.0	sf bay	san francisco	USA	3.0
2902	f:7791	Fund II	2013-12-11	5000000.0	portland	portland	USA	1.0
2903	f:15091	Fund I	2013-12-11	1100000.0	dallas	dallas	USA	1.0
2904	f:5920	Rochester Angel Fund	2013-12-12	2300000.0	west henrietta	west henrietta	USA	1.0

2.1.4 PREPROCESADO DE INVERSIONES Y RONDAS DE FINANCIACIÓN

La base de datos de inversiones nos dibuja las relaciones que existen entre los inversores y las empresas invertidas. El primer paso en el preprocesado, como para cada base de datos, ha sido la eliminación de las columnas que no tienen relevancia. La base de datos se utilizará más adelante para construir la base de datos final.

La base de datos de rondas de financiación es de las más interesantes. Una ronda de inversión o financiación es un proceso mediante el cual una *startup* recibe una cantidad de capital determinada que necesita para el desarrollo de su negocio, gracias a la participación de diferentes inversores. El propósito de los inversores es que la empresa crezca para así recuperar el capital invertido. Por ello, los inversores se convierten en

socios de la empresa y adquieren una parte de su capital social y, por tanto, una parte del control de esta.

La base de datos contiene información sobre las distintas rondas de financiación que han levantado las distintas *startups*, incluyendo su tipo, la fecha de la ronda, la cantidad levantada y las valoraciones *pre-money*[14] y *post-money*[15]. Además, identifica el número de participantes y si es o no la primera o última ronda. Es cierto que esta base de datos contiene muchos valores nulos sobre todo en las columnas de valoración. Esto se debe a que muchas rondas, sobre todo iniciales son privadas y no se suele revelar información acerca de las cantidades. Además, con el aumento de la popularidad y uso de deuda convertible, en rondas iniciales, no es necesaria la valoración de las empresas pues esta deuda se convierte en base a valoraciones futuras.

[14] El valor *pre-money* de una empresa es el valor que tiene antes de que entren a aportar capital los inversores externos.
[15] El valor *post-money*, que no es más que una derivación del anterior, es el valor que tiene tras haberse realizado esa inversión de capital. [13]

Tabla 7. Extracto de las Columnas relacionadas con las Valoraciones pre-ronda y post-ronda (Mayormente NAs)

pre_money_valuation_usd	pre_money_valuation	pre_money_currency_code	post_money_valuation_usd	post_money_valuation
0.0	0.0	NaN	0.0	0.0
0.0	0.0	USD	0.0	0.0
115000000.0	115000000.0	USD	0.0	0.0
525000000.0	525000000.0	USD	0.0	0.0
0.0	0.0	NaN	0.0	0.0
...	...	...	...	...
0.0	0.0	USD	0.0	0.0
0.0	0.0	USD	0.0	0.0
0.0	0.0	USD	0.0	0.0

2.1.5 PREPROCESADO DE IPOS Y ADQUISICIONES Y CREACIÓN DE OPERACIONES

La base de datos de IPOS contiene información acerca de aquellas *startups* que hayan salido a mercados de valores (mayormente NASDAQ). Sorprendentemente, a pesar de que estas operaciones están altamente analizadas y requieren un alto nivel de transparencia, esta base de datos contiene un alto nivel de nulos.

Por otro lado, se tiene la base de datos de adquisiciones, que por lo general suelen tener un carácter privado y es más común que las cifras de las operaciones se mantengan secretas.

Ambas bases de datos son clave pues la variable de venta y la variable que representa la cuantía por la que salieron a bolsa servirán de variable objetivo para el diseño del algoritmo. Esto se debe a que, en cierta medida, supone la mejor forma de ejemplificar la

salida de un inversor de VC. Es cierto que estas operaciones no siempre representan una salida de capital, pero es muy común que sí sea el caso y, por lo tanto, será la variable utilizada para medir el éxito de la inversión.

Tanto para IPO como para Adquisiciones, el objetivo es homogeneizarlas para así poderlas combinar en un único *dataset*. Este *dataset* contiene la variable objetivo ("price_amount") que se utiliza en el modelo predictivo final. Además, contiene otras variables interesantes como la fecha cuando se vendió/salió a bolsa o el número de operaciones en las que participa la empresa (que se calcula más adelante). Al combinar las dos bases, creando la base de datos de Operaciones, se ha de considerar al comprador en las IPOs como el público general, pues la base de datos de Adquisiciones contiene una variable donde se identifica el comprador y se necesita una estructura común para poder combinar bases de datos.

Tabla 8. Extracto de Operaciones(Combinación de Adquisiciones e IPOs)

acquiring_object_id	object_id	price_amount	currency_code	acquired_at // public_at	Year
c:11	c:10	2.000000e+07	USD	2007-05-30	2007
c:59	c:72	6.000000e+07	USD	2007-07-01	2007
c:24	c:132	2.800000e+08	USD	2007-05-01	2007
c:59	c:155	1.000000e+08	USD	2007-06-01	2007
c:212	c:215	2.500000e+07	USD	2007-07-01	2007
...	...	...	...	...	...
General Public	c:10704	1.100000e+09	USD	2013-11-14	NaN
General Public	c:71350	2.000000e+07	USD	2011-05-02	NaN
General Public	c:12	1.810000e+10	USD	2013-11-07	NaN
General Public	c:1105	7.500000e+08	USD	2013-11-06	NaN
General Public	c:185523	3.860000e+04	USD	1999-12-09	NaN

2.1.6 PREPROCESADO DE OBJETOS

La base de datos Objetos contiene la gran mayoría de la información que se encuentra en las otras bases de datos. En ella, se combinan las personas, los fondos y las *startups* (además de otras compañías que tienen presencia en el ecosistema como compradores o inversores). Por ello, existen muchos valores nulos en la mayoría de las columnas. De cara a la simplificación, se divide la base de datos de objetos en 3 bases de datos principales:

1. Compañías
2. Personas
3. Fondos

No obstante, antes de poder hacer eso, se eliminan todas las columnas que no aportan valor.

Una vez dividida la base de datos de objetos, partiendo de compañías, se puede elaborar la base de datos final.

2.1.7 CREACIÓN DE LA BASE DE DATOS FINAL

Tras haber procesado todas las tablas individualmente, se obtienen un total de seis tablas que forman parte del análisis exploratorio a las que se referirán como 'Conjunto de Bases de Datos Procesadas' a lo largo del análisis exploratorio. Sin embargo, para poder implementar los algoritmos es necesario combinar las bases de datos para así poder crear la base de datos final. Además, la base de datos final también se utilizará para identificar patrones en el análisis exploratorio de los datos. La creación de la base de datos final se explica a continuación.

Primeramente, se combinan las compañías con las bases de datos de oficinas y operaciones. Esto permite añadir mucha información que puede ser relevante para la creación del modelo. Además, con la base de datos de operaciones, se fija la variable objetivo para cada caso. Esto genera una base de datos con un total de 9188 casos.

En segundo lugar, se eliminarán aquellas filas que tengan un valor nulo en la variable objetivo pues no podrán ser utilizadas por el modelo, reduciendo el número de casos a 2515. Además, al observar que la variable que describe la categoría de las empresas tiene gran cantidad de nulos, se intenta rellenar con la información que describe la actividad principal de la empresa (columna "description"). No obstante, en los casos donde no se consigue rellenar esa categoría, se utiliza la propia web de Crunchbase para completar todos los valores de la columna. Asimismo, para aumentar la homogeneidad y permitir al modelo encontrar patrones, se eliminan los días y los meses de las variables de tipo fecha, para sólo utilizar el año.

Tras haber procesado la nueva base de datos, se añade la información referente a las rondas de financiación y a los distintos inversores, de nuevo procesando la nueva información para permitir un diseño del modelo óptimo.

Finalmente, se añaden, a cada empresa, las relaciones con los distintos individuos, seleccionando sólo la más importante (CEO[16] // Founder). Esto permite añadir nuevas columnas como el ranking de la universidad donde se ha formado el individuo asociado a la compañía o el número de títulos obtenidos, completando asi la base de datos que se utilizará en el modelo. Sin embargo, la implementación de la información de las rondas de financiacion y de los estudios de los fundadores generan muchas filas con valores nulos. Por lo tanto, para permitir el funcionamiento de los modelos se eliminan aquellas filas con valores nulos, reduciendo el número total de casos a 766. Sin embargo, cada uno

[16] CEO (Chief Executive Officer) se traduce al español como director ejecutivo. Es el cargo ejecutivo de mayor rango en una empresa u organización.

de los casos cuenta con 25 columnas con información relevante, proporcionando más información y características sobre los datos. Esto puede permitir que el modelo aprenda patrones más complejos y tome decisiones más precisas.

2.2 ANÁLISIS EXPLORATORIO

2.2.1 ANÁLISIS EXPLORATORIO DEL CONJUNTO DE BASES DE DATOS PROCESADAS

En primer lugar, se analizarán los datos de las seis base de datos obtenidas tras procesar los datos originales, esto permitirá conocer en mayor detalle el ecosistema *startup* y algunas tendencias que más adelante se verán reflejadas en el modelo.

Esta primer análisis de la base de datos resulta clave, pues a medida que se va añadiendo contenido a la base de datos el número de casos completos va reduciendo y por lo tanto se pierde la imagen general del ecosistema. Por lo tanto, es imprescindible hacer un primer análisis global para darle un mejor contexto al trabajo, además de poder diseñar un modelo más apropiado.

2.2.1.1 Universidades Más Populares

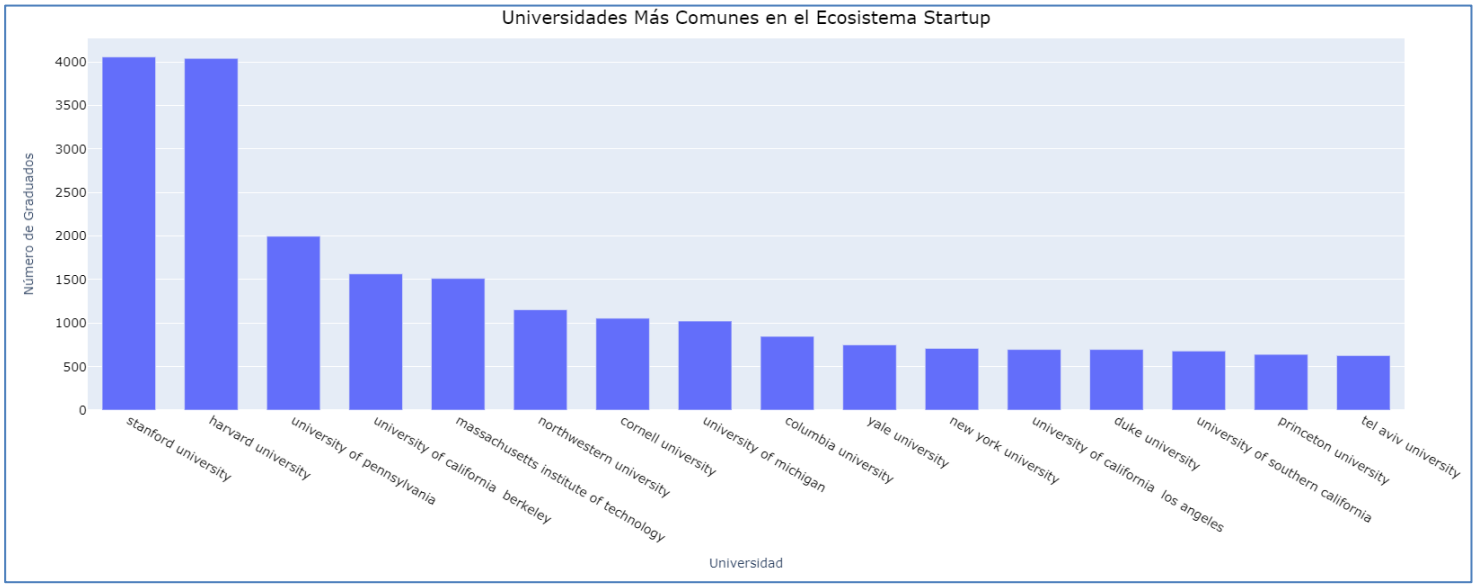
La formación académica tiene un gran peso sobre el desarrollo laboral de una persona en la mayoría de las carreras profesionales. Por lo tanto, es interesante estudiar la formación académica de los individuos relacionados con el sector.

Se pueden extraer varias conclusiones de la *Figura 3.* En primer lugar, se puede observar claramente que la educación universitaria en Estados Unidos era diferente en cuanto a

materia de emprendimiento hasta el año 2013. Tan solo una de las 20 universidades no se encontraba en Estados Unidos y es la Universidad de Tel Aviv. La institución israelita es y era de las mejores del país y se encontraba entre las mejores del mundo según el ranking QS. Además, el ecosistema emprendedor israelita era el segundo más importante del mundo, después de Estados Unidos, siendo Tel Aviv el mejor ecosistema emprendedor tras Silicon Valley [14], lo que potencia la alta recurrencia que tenía esta universidad como centro de formación de individuos del ecosistema emprendedor.

Por otro lado, no es casualidad que el resto de las universidades se encontrasen todas entre las más prestigiosas del mundo. Esto se debe a que son estas universidades las que más soporte daban a los alumnos, además de tener una gran cultura y mentes brillantes para poner en práctica proyectos de emprendimiento [15]. Asimismo, eran estas universidades las que más financiación reciben para proyectos de investigación [16], dando muchas más posibilidades a sus alumnos para desarrollar ideas innovadoras.

Finalmente, hay que destacar la existencia de dos instituciones muy potentes en el ecosistema. En la costa este de EE. UU., Harvard es el centro más escogido por las personas del mundo del emprendimiento. Alternativamente, Stanford no solo lidera la



costa oeste, sino que es la universidad que cuenta con mayor representación en el ecosistema.

Figura 4. Estudio del Ecosistema Startup - Universidades más Comunes

2.2.1.2 Estudio de las IPOs

El término IPO se refiere a la oferta pública inicial, que significa que una empresa privada decide vender sus acciones al público por primera vez, lo que permite a los inversores participar en el desarrollo de la empresa [17].

Las IPOs son una de las dos formas de Exit que se estudian en este trabajo. Sin embargo, a diferencia de las adquisiciones privadas, las ofertas públicas iniciales son operaciones con una relación más directa con la economía. Por lo tanto, es importante entender la relación que existe entre las IPOs y el ciclo económico.

En el gráfico se identifican dos momentos clave. En primer lugar, podemos observar el efecto que tuvo la burbuja Dot-Com sobre los mercados bursátiles. El número de IPOs aumenta significativamente en un espacio de dos años, para luego caer a los mismos niveles pre-burbuja. Esto se debe a que durante el periodo desde el año 1990 hasta el año 2000, la creciente popularidad de Internet generó la creación de muchas empresas tecnológicas. Sin embargo, las valoraciones exageradas de estas empresas no estaban respaldadas por un modelo de negocio sostenible dando lugar a una burbuja en el mercado de valores de tecnología de la información. Cuando la burbuja estalló a principios de la década de 2000, muchas empresas tecnológicas quebraron y la confianza de los inversores bajo, dando lugar a valoraciones más mesuradas y menos IPOs [18].

Por otro lado, se puede ver esa tendencia creciente en el número de salidas a bolsa anuales, destacando el año 2013 como un punto de inflexión. Esto se debe a que 2013 fue un año bueno, económicamente hablando, donde la economía ya presentaba una estado saludable sin rastro de remanentes de la crisis financiera de 2007-2008. Esta año no solo fue bueno en el ecosistema emprendedor, de hecho 2013 fue el mejor año para IPOs desde el año 2000 [19].

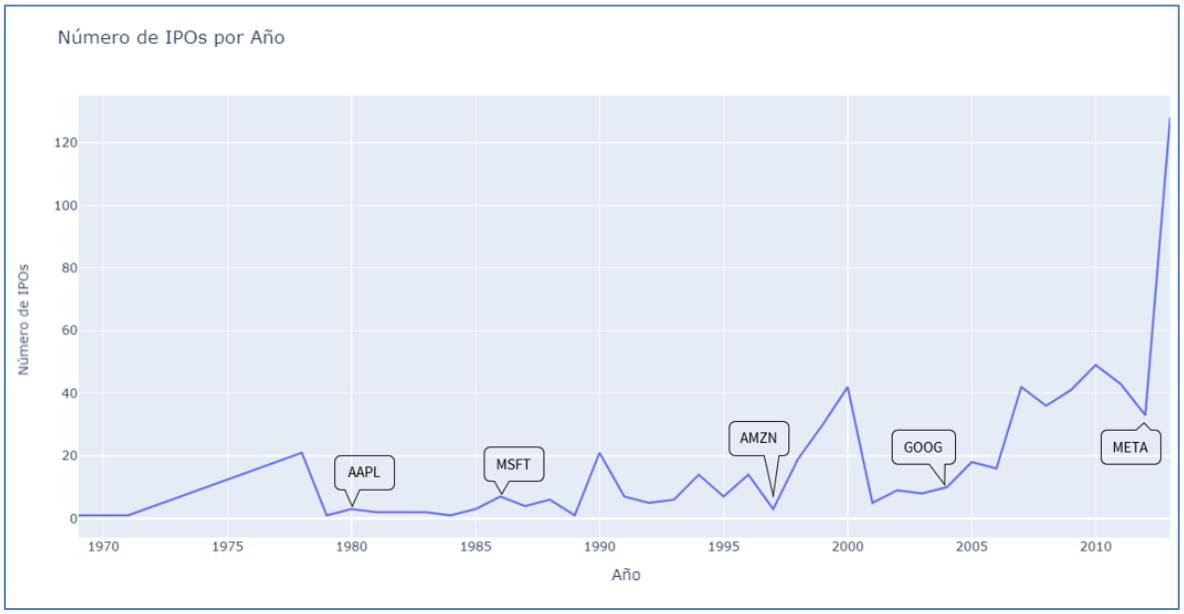


Figura 5. Estudio del Ecosistema Startup - Evolución de las IPOs

2.2.1.3 Regiones con Mayor Actividad

Es una realidad que en el aspecto económico existen países y regiones más productivas y eficientes. También es cierto que hay zonas geográficas con mayor conocimiento en ciertos sectores/industrias como puede ser Nueva York en servicios financieros o La Rioja en producción de vino. Estas ventajas suponen un atractivo añadido para las startups a la hora de elegir sus sedes. Además, en la propia industria del Venture Capital y el Ecosistema emprendedor existen regiones como Silicon Valley, que destacan por su alta concentración de startups y conocimiento. Por lo tanto, es importante conocer cuáles son

esos puntos neurálgicos alrededor del mundo y si tienen un efecto sobre el éxito de la startup.

La gráfica muestra los puntos con mayor actividad del mundo en el ámbito del emprendimiento. Para mostrar la actividad, las burbujas representan el total de oficinas que se encuentran dentro de esas regiones. Como se observa en la gráfica, África y Sudamérica no tienen ningún centro neurálgico. Además, al tener datos hasta 2013, el fuerte crecimiento que han experimentado ecosistemas del sureste asiático como Beijing o Seúl no se ven representados [20].

Por otro lado, se observan tres regiones clave, separadas geográficamente, en la industria. En primer lugar, en Europa, la capital británica destaca como punto de mayor actividad del sector, hasta el año 2013, seguido, muy de lejos, por Berlín y Paris. El posicionamiento de Londres como el líder de la región se debe a varios motivos.

En primer lugar, el estatus de Londres como centro financiero de Europa y el puente con el continente americano daba lugar a mayores facilidades en la captación de fondos y capital. Por otro lado, el gobierno británico facilitaba el levantamiento de capital para las startups, ofreciendo rebajas fiscales a aquellos que aportasen capital a compañías autorizadas por el gobierno [21].

Las otras dos regiones más potentes se encuentran en los lados opuestos del continente americano. Por un lado, en la costa oeste, San Francisco se situaba como la principal región (conocida como Silicon Valley) del mundo para el ecosistema emprendedor. Existen diversos motivos por los que Silicon Valley se posicionaba al frente de la lista.

Las startups de Silicon Valley levantaban entre 2 y 3 veces más, en etapas tempranas. Además, en San Francisco había un 46% más de startups que contaban con la ayuda de mentores. Por otro lado, Nueva York contaba con emprendedores mejor formados. De hecho, los fundadores en Nueva York eran el doble de probables de tener un PhD[17] que aquellos de San Francisco. Finalmente, las startups en etapas más avanzadas levantaban de media un 27% más si se encontraban en la ciudad de la gran manzana [22].

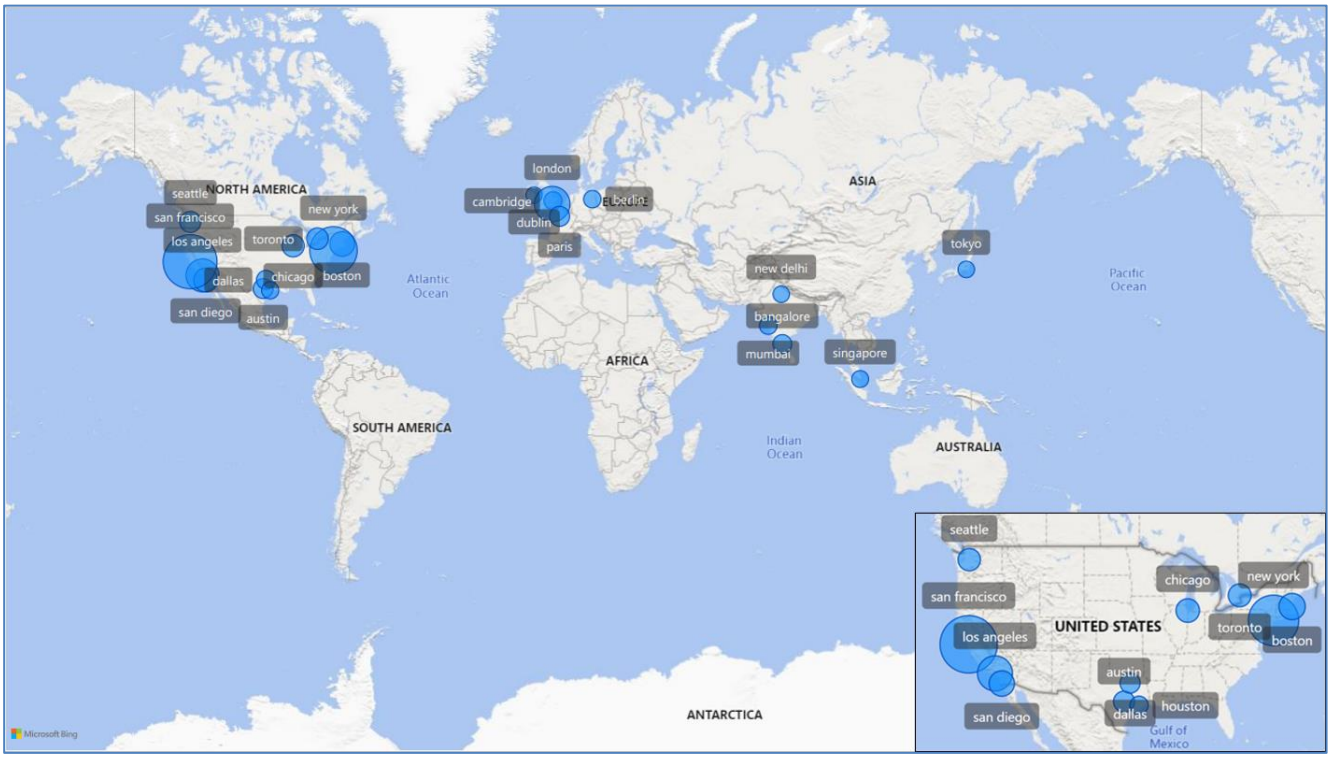


Figura 6. Estudio del Ecosistema Startup - Regiones con Mayor Actividad

[17] The term PhD or Doctor of Philosophy is an abbreviation of the Latin phrase 'philosophiae doctor'. A PhD degree typically involves students independently conducting original and significant research in a specific field or subject, before producing a publication-worthy thesis. [23]

2.2.1.4 Capital Levantado por Tipo de Ronda

A lo largo de sus vidas, las startups experimentan distintas rondas de financiación. Estas rondas tienen un impacto significativo sobre el desarrollo y la planificación de las startups. En consecuencia, en busca de un mayor entendimiento sobre la industria, se analizan los tamaños de las distintos tipos de ronda de financiación que experimentaron las startups entre 1999 y 2013.

Existen distintos tipos de ronda que una startup puede levantar. Cada ronda se define por la etapa en la que se encuentra la empresa y las rondas previas que ha recibido.

Con el objetivo de entender en mayor medida el grafico, se explican a continuación los distintos tipos de rondas utilizando el glosario de Crunchbase.

I. **Venture** – Esta categoría hace referencia a una mezcla de varias rondas que no han sido catalogadas como ninguna de las anteriores. Una venture round hace referencia a cualquier tipo de ronda de financiación a partir de la Serie A utilizada por una startup para levantar capital.
II. **Angel** – Este tipo de ronda está diseñada para startups en etapas muy iniciales donde la compañía necesita capital para empezar a operar. Los inversores en este tipo de ronda pueden ser familiares, amigos o inversores profesionales llamados Angel Investors[18].
III. **Series A & Series B** – Este tipos de rondas de financiación son para empresas en etapas tempranas y suelen tener un tamaño de entre $1M - $30M. Para que ocurra una Series B, tiene que haber una Series A primero.

[18] Los ángeles inversores, o ángeles de negocios, son personas que invierten su dinero en la fase inicial de compañías emergentes a cambio de una participación en capital. Habitualmente, ejercen también un rol de mentor y ofrecen su consejo y experiencia a los emprendedores.

IV. Series C+ – Las rondas de la serie C y posteriores son para empresas más avanzadas y establecidas. Estas rondas suelen ser de $10 millones o más y a menudo son mucho más grandes.

V. Other – Hace referencia a otros tipos de rondas.

VI. Private Equity – Una ronda de inversión de capital privado es liderada por una firma de capital privado o un fondo de cobertura y es una ronda en etapa avanzada. Es una inversión menos arriesgada porque la empresa está más establecida y las rondas suelen ser de más de $50 millones.

VII. Crowdfunding – Hace referencia a dos tipos de rondas similares. El Equity Crowdfunding permite a los usuarios individuales invertir en empresas a cambio de acciones. Por lo general, los inversores invierten pequeñas cantidades de dinero, aunque se forman sindicatos para permitir que un individuo lidere la evaluación de una inversión y combine fondos de un grupo de inversores individuales. Por otro lado, el Product Crowdfunding es una ronda de financiación donde una empresa proporcionará su producto, que a menudo aún está en desarrollo, a cambio de capital. Este tipo de ronda también se completa típicamente a través de una plataforma de financiamiento.

VIII. Post IPO – Engloba 3 tipos de rondas que pueden ocurrir tras la salida a bolsa de la compañía. Una ronda de Capital Post-IPO tiene lugar cuando las empresas invierten en una compañía después de que ésta ya se haya hecho pública. En segundo lugar, una ronda de Deuda Post-IPO tiene lugar cuando las empresas prestan dinero a una compañía después de que ésta ya se haya hecho pública. Similar al financiamiento de deuda, una compañía promete reembolsar el principal y los intereses adicionales sobre la deuda. Finalmente, una ronda secundaria post-IPO tiene lugar cuando un inversor compra acciones de una compañía de otros accionistas existentes en lugar de hacerlo directamente de la compañía, y ocurre después de que la compañía ya se haya hecho pública.

Tras haber explicado los distintos tipos, se puede observar claramente como a mayor tamaño de la compañía, mayores son sus capacidades para levantar capital y mayores son

las rondas de financiación. Esto se ve claramente al analizar los tamaños de ronda medios desde el tipo de ronda "Angel" hasta las "Series C+".

Por otro lado, son las rondas posteriores a las IPOs las más grandes del ecosistema. Esto se debe a que la salida a bolsa de las compañías implica una valoración previa a la IPO deben tener una capitalización superior a los $550M$ durante los últimos 12 meses [24] y como se ha visto previamente, a mayor valoración, mayor es la capacidad de levantar capital y mayores son las rondas.

Finalmente, se observa que las rondas de crowdfunding son las más bajas, probablemente causado por la inflexibilidad de las rondas una vez son lanzadas y la falta de inversores corporativos que suelen aportar los grandes tickets.

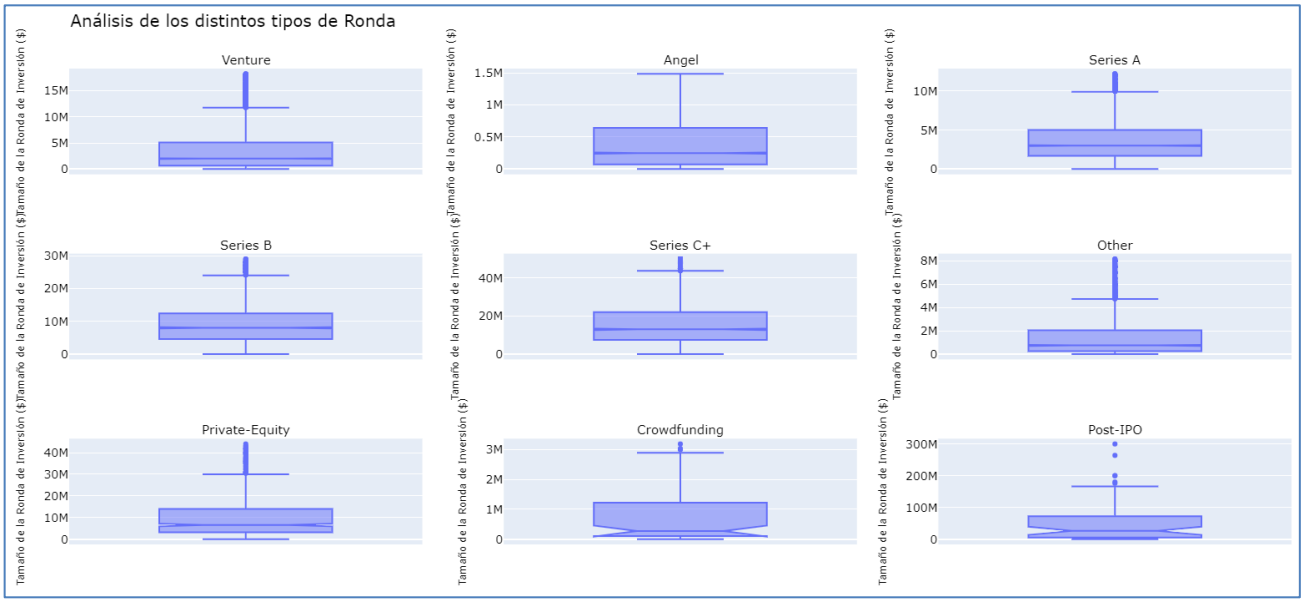


Figura 7. Estudio del Ecosistema Startup - Capital Levantado por Tipo de Ronda

2.2.2 ANÁLISIS EXPLORATORIO DE LA BASE DE DATOS FINAL

Una vez se han analizado el conjunto de bases de datos procesadas, obteniendo así, un mayor conocimiento del sector y de las variables que componen el banco de datos, se profundiza en la base de datos final, que es la que se utilizará en los modelos.

El análisis de la base de datos final permite identificar tendencias y características de la base de datos, dando lugar a una mejor toma de decisiones y un mejor diseño de los modelos. Además, este análisis está diseñado para confirmar que la base de datos final, que no deja de ser una muestra de los datos que se ven en las seis bases de datos procesadas, sigue las tendencias del ecosistema, identificadas en el estudio sector.

2.2.2.1 Capital Levantado por Ronda

Tras haber analizado las distintas rondas de financiación del ecosistema, se analiza si la base final sigue las tendencias del mercado.

A partir de la figura 7, se puede confirmar que los patrones vistos en el ecosistema startup se siguen en la base de datos final. Sin embargo, es cierto que, en el caso de la base de datos final, las rondas catalogadas como Venture tienen un tamaño ligeramente superior. Como este tipo de ronda hace referencia a aquellas que no han sido catalogadas como ninguna de las anteriores, incluye distintos tipos de ronda, que para esta muestra son rondas más avanzadas que en el ecosistema en general.

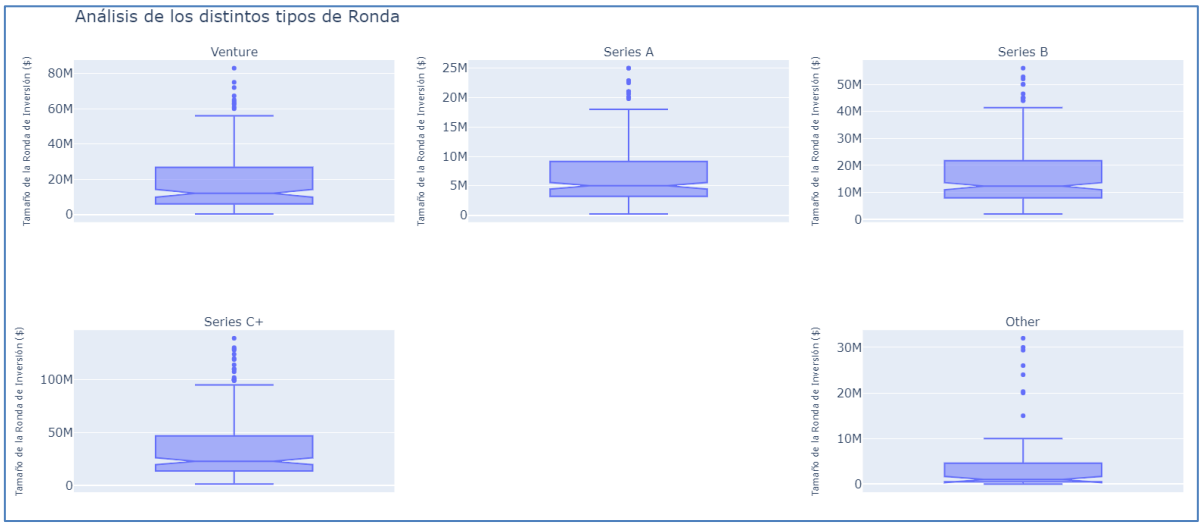


Figura 8. Estudio de la Base de Datos Final - Capital Levantado por Ronda

2.2.2.2 Instituciones de preferencia entre Fundadores y Directivos

Al incluir los estudios de los fundadores y directivos de las startups, resulta interesante realizar un análisis del tipo de institución más popular entre los altos cargos en las startups de la base de datos final.

Los datos son realmente sorprendentes, pues entre las startups restantes, son aquellas personas que han dejado la universidad o que directamente no la han empezado las más comunes. Algunos de los mayores emprendedores de la historia como Bill Gates o Mark Zuckerberg dejaron la universidad para centrarse en sus proyectos. Sin embargo, este gráfico muestra que esta tendencia es mucho más común de lo que parece.

Por otro lado, de entre aquellas personas que montan startups, la mayoría proviene de alguna de las universidades en el Top 50 del mundo, siguiendo esa tesis inicial vista en el estudio del ecosistema.

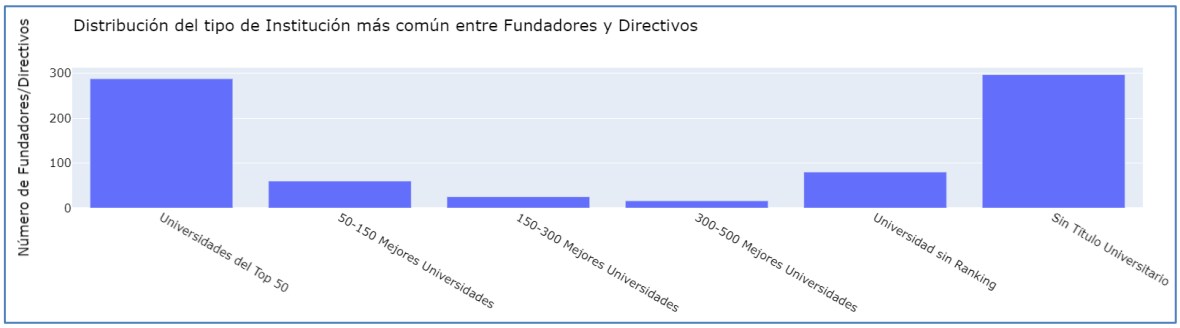


Figura 9. Estudio de la Base de Datos Final - Distribución de Instituciones de preferencia entre Fundadores y Directivos

2.2.2.3 Regiones más Comunes

Finalmente, como se ha visto en el estudio general, la región en la que se encuentran las startups es una característica importante y, por lo tanto, es importante analizar si existe concordancia entre las tendencias globales y la base de datos final.

Como se observa en la Tabla 1, las regiones más comunes siguen un patrón similar al visto en el análisis del ecosistema, encontrando en la lista algunas de las regiones más activas del mundo. No obstante, el hecho de que la compañía Crunchbase sea americana puede hacer que la recolección de información de empresas situadas dentro de los Estados Unidos sea más sencilla, generando un sesgo en la base de datos final (sobrerrepresentación de compañías americanas porque compañías en otras regiones se descartan por la falta de información).

Tabla 9. Regiones Más Comunes en la Base de Datos Final

Región	Número de Startups
San Francisco	274
Boston	64
Nueva York	51
Los Ángeles	32
Seattle	32
San Diego	18
Londrés	17
Washington DC	14
Austin	13
Tel Aviv	11

CAPÍTULO 3. MODELOS DESCARTADOS

En este apartado del proyecto se explicará los modelos diseñados inicialmente, antes de haber logrado el algoritmo que mejor se adapta a los datos y que por lo tanto proporciona mejores resultados. Además, se presentarán y compararán los resultados obtenidos con cada modelo. El proceso desarrollado previo a la elaboración de los modelos se presenta a continuación.

3.1 CLUSTERING

3.1.1 PROCESO DE MODELADO

En primer lugar, antes de realizar los modelos se realizó un estudio de los datos previo, utilizando técnicas de clustering [25]. El clustering es una técnica de aprendizaje no supervisado que se usa para agrupar datos similares en conjuntos llamados "clústeres". Los clústeres se generan a partir de la identificación de patrones y relaciones entre los datos, agrupando aquellos casos que son parecidos entre sí. En este trabajo, el objetivo que tiene la realización de dicho análisis es el de la segmentación de los datos. Al identificar los distintos clústeres, se pueden conocer los distintos grupos dentro de una base de datos. Esta segmentación puede proporcionar una comprensión más profunda de la estructura subyacente de los datos y ayudar a identificar subgrupos o segmentos específicos que pueden requerir un enfoque o tratamiento diferenciado. En particular, ha el clustering permitió conocer si existían relaciones suficientes entre los datos como para poder seguir con la elaboración de los modelos predictivos. Además, los resultados del clustering tendrán otros usos durante la elaboración de los modelos [26].

Por otro lado, el clustering puede ayudar a identificar patrones ocultos o estructuras en los datos sin la necesidad de etiquetas o categorías previas. Esto puede ser especialmente útil cuando se tiene un conjunto de datos no etiquetados o cuando no se conocen las

características clave que definen las clases en los datos. Además, los clústeres pueden proporcionar una forma de inicializar o preparar los modelos de predicción. Esto puede ayudar a mejorar la precisión y el rendimiento de los modelos de predicción al adaptarse a las características y patrones únicos de cada clúster.

En el proceso de desarrollo de un algoritmo de clustering, el primer paso es preparar los datos para el modelado. Esto implica abordar tanto las variables categóricas como las numéricas, con el objetivo de permitir que los modelos comprendan de manera óptima los patrones en los datos. Durante esta etapa, se utiliza el método de OneHotEncoder para codificar las variables categóricas. Este método crea características binarias para cada categoría de la variable categórica, sin repetir los valores. Cada nueva columna representa si el caso pertenece a esa categoría, siendo 1 si el caso tiene esa categoría como atributo y 0 en caso contrario [27]. La codificación de variables categóricas es esencial, ya que la mayoría de los modelos utilizados en este tipo de proyectos requieren datos numéricos. El codificador permite ajustar varios parámetros para adaptar el algoritmo a las necesidades del problema. En este caso, se ha implementado una configuración que permite al algoritmo manejar datos desconocidos de forma sencilla. Cuando el algoritmo encuentra una categoría desconocida durante la transformación, establece todas las características binarias correspondientes en 0. Esto evita errores en caso de que aparezcan nuevas categorías en los datos de prueba que no estaban presentes en los datos de entrenamiento. Además, se ha configurado el algoritmo para descartar la primera columna de cada variable, lo cual ayuda a evitar la multicolinealidad.

En cuanto a las variables numéricas, se realiza un procesamiento que implica la normalización de dichas variables. La normalización consiste en escalar los valores de las variables para que estén en una escala común. Este proceso permite eliminar sesgos que puedan existir en los datos debido a las diferentes escalas de las variables. Además,

asegura que todas las variables contribuyan de manera equitativa al modelo. Si una variable tiene una escala mucho mayor que las demás, puede dominar el modelo y tener un impacto desproporcionado en los resultados [28]. Durante la selección de las variables numéricas, se ha excluido la variable objetivo 'price_amount'. Esto se debe a que el objetivo del clustering es agrupar las observaciones en función de sus características similares, y la variable objetivo podría introducir un sesgo en la formación de los grupos.

En segundo lugar, se identifica el número óptimo de clústeres. Aunque existen varios métodos para determinar este valor, en este proyecto se ha utilizado el método del codo (Elbow Curve). Este método implica comparar la suma de los cuadrados de las distancias entre los clústeres (SSE, por sus siglas en inglés) con el número de clústeres. El punto donde la disminución de SSE se reduce significativamente se considera el número ideal de grupos [29]. Este método recibe su nombre debido a que, al representar gráficamente la comparativa, la curva tiene una forma similar a un codo. En la figura 9 se muestra que el número óptimo de clústeres, según los resultados del modelo, es un valor entre 6 y 8. Como el número de grupos tiene que ser un valor entero, ambos valores se probarán más adelante, escogiendo el que ofrezca mejores resultados.

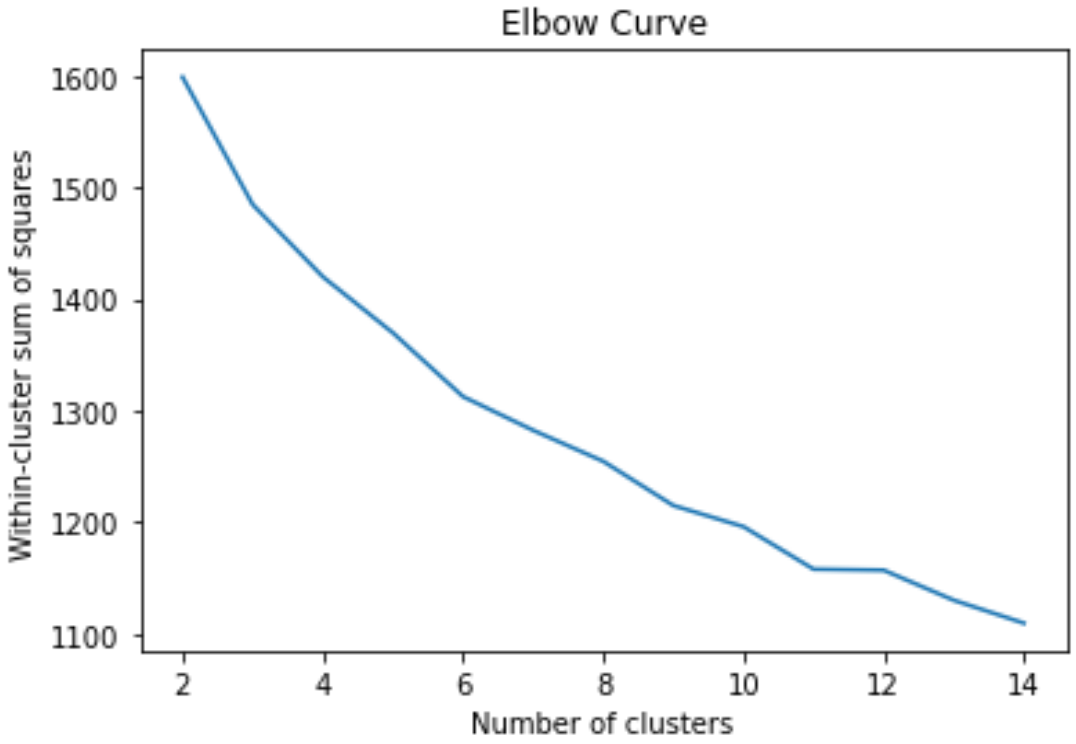


Figura 10. Estudio del Número Óptimo de Clústeres para KMeans

Una vez identificado el número de grupos ideal, se procede a la implementación del algoritmo de clustering. En este caso, se utilizan dos algoritmos de clustering, de los cuales se escogerá el que mejor resultados de, aumentando así la robustez del modelo. En primer lugar, se implementa el método KMeans con el número de clústeres ideal extraído de la Figura 9. En este caso, se probarán valores del 5-8 pues el punto de inflexión no se distingue claramente en la gráfica. El otro algoritmo utilizado es DBSCAN. DBSCAN es un algoritmo de clustering que identifica áreas densas de puntos en el espacio de características y forma grupos en función de la densidad de los puntos. Este algoritmo escoge de forma automática el número ideal de clústeres. No obstante, se han de escoger los valores de los hiperparámetros. Para ello se sigue el procedimiento explicado en el trabajo realizado por Tara Mullin "DBSCAN Parameter Estimation Using Python" [30].

Tras haber probado ambos métodos de clustering con diversos parámetros, se procede a comparar los resultados. Para ello se hace uso de las siguientes métricas:

1. Davies-Bouldin Index:

El índice Davies-Bouldin se calcula como la similitud promedio de cada clúster con su clúster más similar. Este índice Davies-Bouldin representa la "similitud" promedio de los clústeres, donde la similitud es una medida que relaciona la distancia entre clústeres con el tamaño del clúster. Un modelo con un índice Davies-Bouldin más bajo tiene una mejor separación entre los clústeres [31].

2. Calinski-Harabasz Index:

El índice Calinski-Harabasz (CH) también es una medida de evaluación interna para la calidad del clustering. El objetivo es maximizar el índice CH, lo que implica una mayor cohesión dentro de los clústeres y una mayor separación entre los clústeres [31].

Los resultados obtenidos son los siguientes:

Algoritmo	Calinski-Harabasz	Davies-Bouldin
KMeans con 5 Clústeres	97.7	1.74
KMeans con 6 Clústeres	95	1.28
KMeans con 7 Clústeres	93.3	1.45
KMeans con 8 Clústeres	92	1.41
DBSCAN	55.4	2.83

Figura 11. Comparativa entre Modelos de Clustering.

Según los resultados de la Figura 10 el algoritmo con mejores resultados para ambos casos es el KMeans con 6 clústeres, pues tiene el segundo valor más alto en la métrica CH y el valor más bajo en el índice DB. Por lo tanto, se selecciona dicho algoritmo para el proceso de clustering.

No obstante, para asegurar un rendimiento óptimo del modelo, se realiza una última comprobación. Al aplicar el algoritmo, se obtiene una etiqueta con el grupo al que

pertenece cada caso. Utilizando la etiqueta, se diseña un árbol de decisión para analizar qué tan buena es la elección de la cantidad de clústeres y, en este caso, qué número de clústeres produce un mejor modelo (6 - 8). Para elaborar el árbol de decisión, se separan los datos en conjuntos de entrenamiento y prueba, y se fija como variable objetivo la etiqueta del clúster al que pertenece cada caso (fila en la base de datos). Tras generar el árbol de decisión, se utiliza el mismo para hacer una predicción, utilizando el conjunto de datos de prueba, y se calcula la precisión del modelo (accuracy en inglés) para evaluar su rendimiento. En este caso, el modelo elegido es aquel que cuenta con seis clústeres, donde la precisión alcanza un valor del 99%, siendo un valor superior a los obtenidos con el resto de algoritmos.

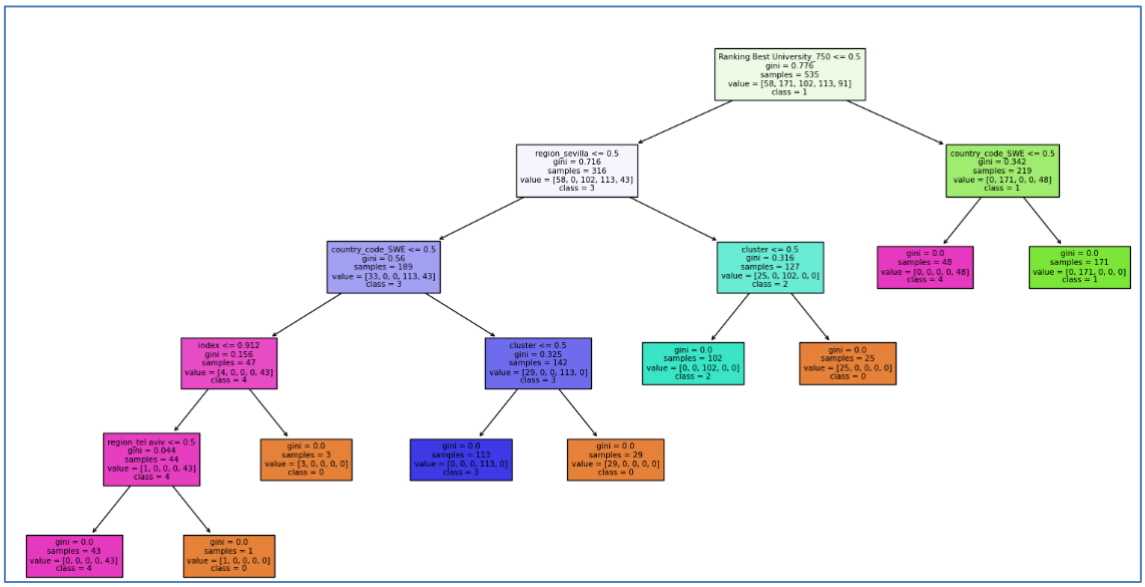


Figura 12. Representación de una sección del árbol de decisión completo.

La información extraída del aprendizaje no supervisado puede aportar gran valor al análisis. Sin embargo, para realmente conocer la lógica que ha seguido el algoritmo es necesario realizar un análisis profundo sobre los distintos grupos, permitiendo la identificación de relaciones y patrones que dan lugar a este reparto en clústeres.

3.1.2 ANÁLISIS DE LOS CLÚSTERES

El análisis de los clústeres es una herramienta imprescindible en la consecución de cualquier modelo que utilice estos datos. Mediante este algoritmo se podrán identificar y comprender de manera más precisa las particularidades y similitudes entre los clústeres, brindando una base sólida para el posterior análisis e interpretación de los resultados obtenidos.

De cara al análisis de los clústeres, se utilizarán boxplots (gráficos de caja) para comparar las variables numéricas y gráficos de barras para mostrar la distribución de las variables categóricas. La elección de boxplots se debe a que permiten visualizar de manera clara y concisa las diferencias en la distribución de los valores entre los clústeres, facilitando la identificación de patrones y tendencias.

En cuanto a los gráficos de barras, se utilizan debido a que permiten representar la frecuencia o proporción de cada categoría en cada clúster, lo que brinda una visión general de cómo se distribuyen las categorías en relación con los diferentes clústeres.

Es importante mencionar que, se analizaran aquellas variables que sus gráficos muestren signos claros de diferenciación entre los clústeres. Esto implica que se analizarán con detalle aquellos casos en los que los boxplots o gráficos de barras muestren indicios claros de que existen diferencias significativas entre los clústeres, lo que puede ser indicativo de características distintivas de cada grupo. Además, para permitir una mejor visualización de los gráficos, se han eliminado los datos extremos de las gráficas.

En primer lugar, la primera variable donde hay una diferencia significativa entre los clústeres es la variable 'first_funding_at'. Esta variable representa el año en el que la startup recibe su primera inversión. En la Figura 12 se observa como los clústeres 0 y 3 destacan por recibir su primera inversión después de 2007, mientras que los clústeres 1,4 y 5 recibieron la suya en 2006. Esta tendencia se ve continuada si se analiza la variable 'last_funding_at', que representa el año de la última inversión. En la Figura 13 se observa como el clúster 3 es el que más tarde recibe la inversión. Por otro lado, con una diferencia de dos años con el grupo 3, los clústeres 0, 2 y 4 reciben su última inversión en 2009. Finalmente, las startups en los clústeres 1 y 5 son las que, de media, antes cierran sus procesos de financiación. Estas son las primeras variables que permiten diferenciar los clústeres de manera más clara. Finalmente, es importante destacar que el momento en el que ocurren las rondas de financiación, también afecta al momento de la venta o salida a bolsa de la compañía. Los grupos 0 y 3, de media, realizan su operación de exit en 2011, siendo los grupos más tardíos en realizar la operación. Por otro lado, los fundadores de los clústeres 1 y 2 venden sus startups o salen a bolsa, de media, en 2010, siendo los que más pronto realizan la operación de exit.

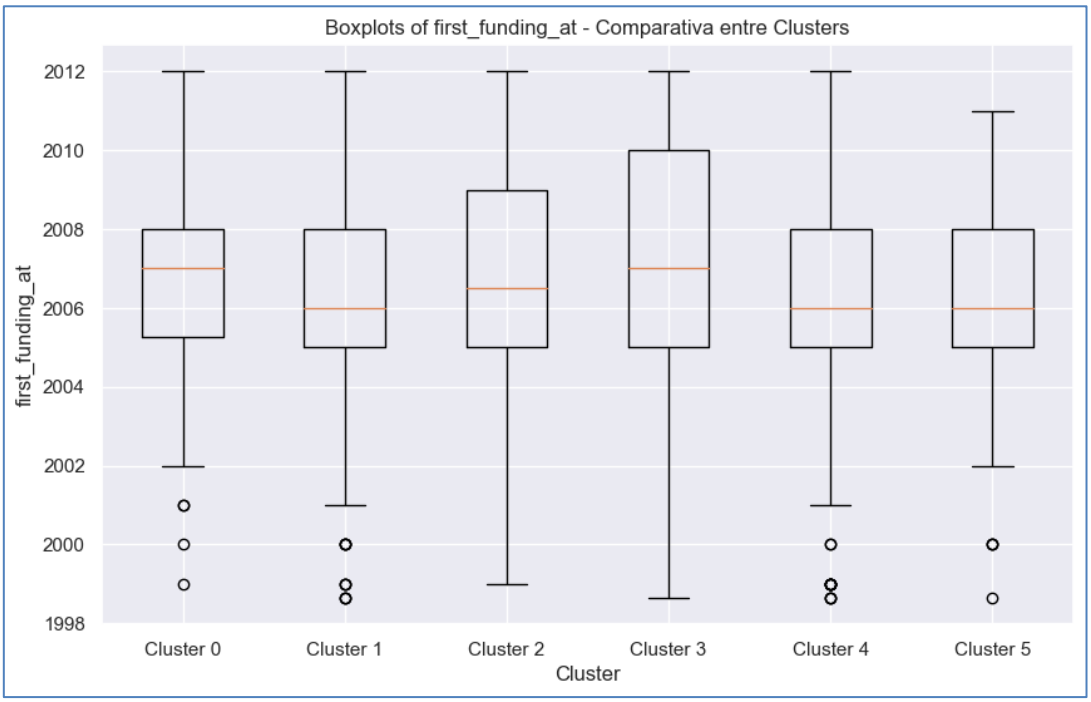


Figura 13. Gráficos de Caja con el año de la primera inversión recibida por la Startup.

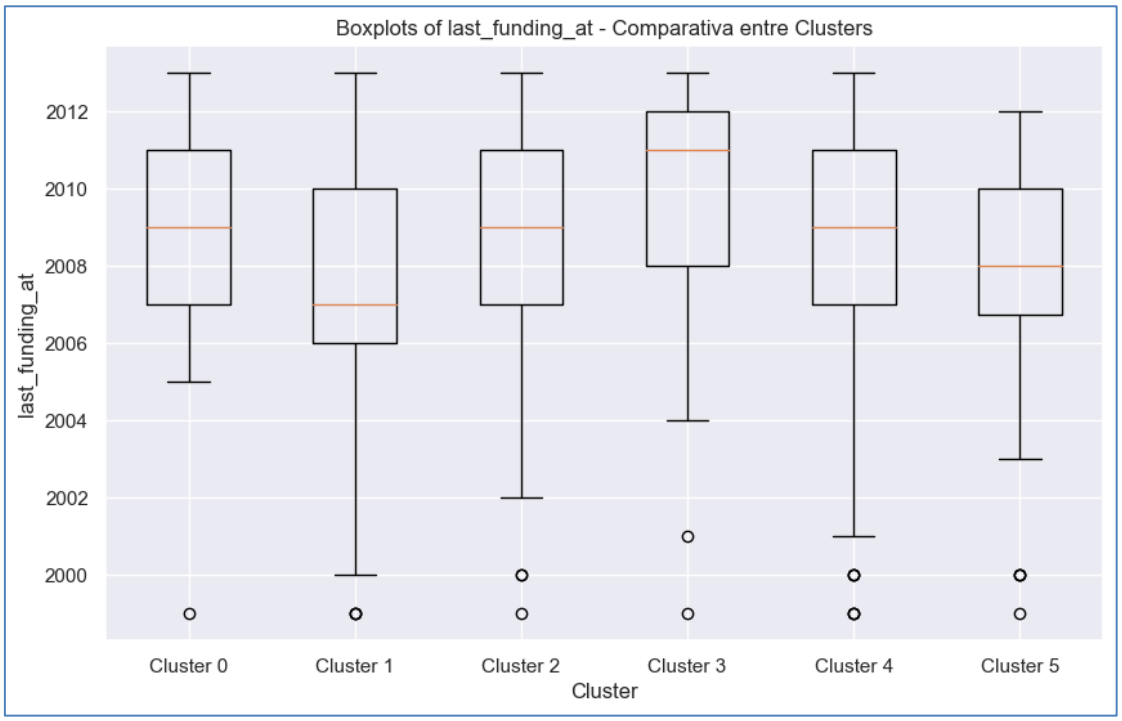


Figura 14. Gráficos de Caja con el año de la última inversión recibida por la Startup.

En segundo lugar, otra de las características clave para diferenciar los grupos es la cantidad recibida en forma de financiación. En la Figura 14 se observa como el grupo 3 recibe de media una cantidad muy superior al resto de clústeres. Alternativamente, el clúster 5 es el que menos capital levanta de media. La financiación en una startup es clave en su desempeño y su capacidad de crecimiento y por lo tanto es una característica muy importante y que se ha de tener en cuenta. Hay que destacar también la alta cantidad de valores extremos que aparecen en las gráficas incluso tras haber eliminado aquellos valores más altos. Esto demuestra la variabilidad de los atributos y las dificultades que existen al predecir y analizar en el sector del venture capital.

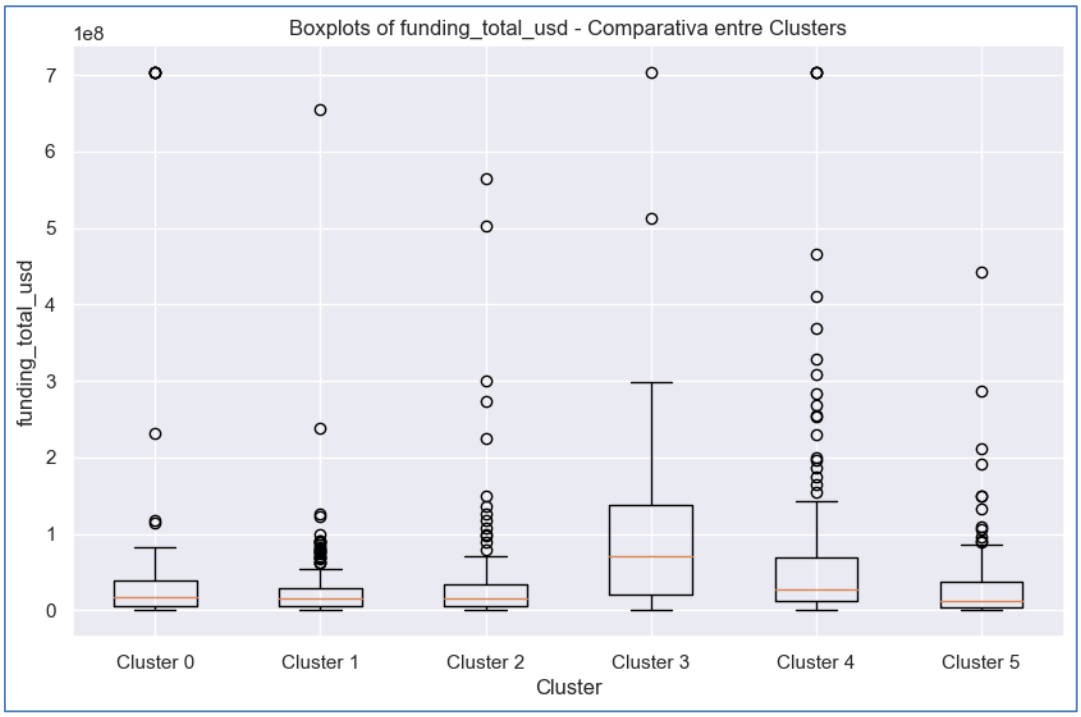


Figura 15. Gráficos de Caja con el total del capital levantado por las Startups.

Por otro lado, el número de medio de relaciones de las startups si varía bastante entre los grupos generados. Gracias a la Figura 15, se pueden identificar los clústeres que contienen a las startups más conectadas y con mayores relaciones dentro de la industria. En este caso, son las startups del grupo 3 las que cuentan con más conexiones dentro de la industria. Además, la diferencia es muy significativa con los clústeres 1,2 y 5 que son los grupos con las startups menos conectadas. En el mundo del venture capital las relaciones son claves para varios aspectos como la captación de fondos o el acceso a información y conocimientos por lo que un alto número de relaciones puede tener un impacto positivo en el desempeño de la startup.

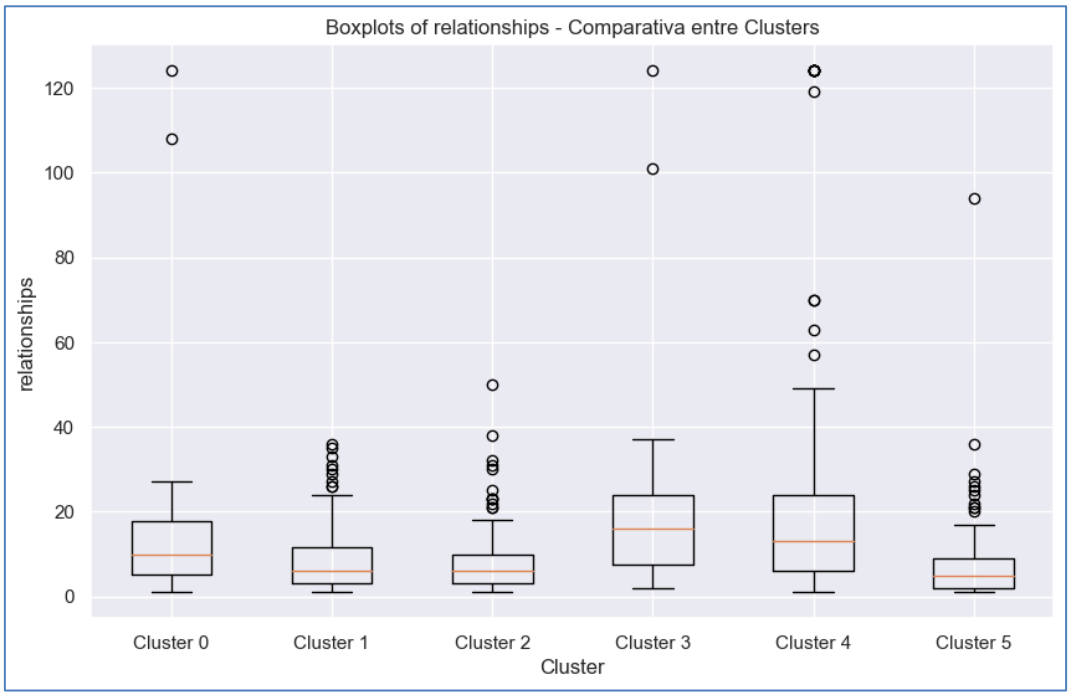


Figura 16. Gráficos de Caja con el número de relaciones.

Si se profundiza en el número de relaciones que están vinculados a la inversión de la startup se puede analizar el número de participantes en las rondas de financiación. Esta variable representa la suma de las empresas, fondos e individuos que han participado en el levantamiento del capital por parte de la startup. En este caso, el grupo con el valor más alto es el clúster 4, que previamente ocupada la segunda posición en el número de relaciones. Siguiendo de cerca al grupo 4, se encuentran el grupo 3 y el grupo 0. Finalmente, el clúster 1 se compone por startups que tienen de media 4 participantes y los clústeres 2 y 5 tienen una media de 3 participantes. Es cierto que el número de participantes no es directamente proporcional al capital levantado, pero si existe cierta relación que se puede apreciar si se analizan la Figura 16 y la Figura 14.

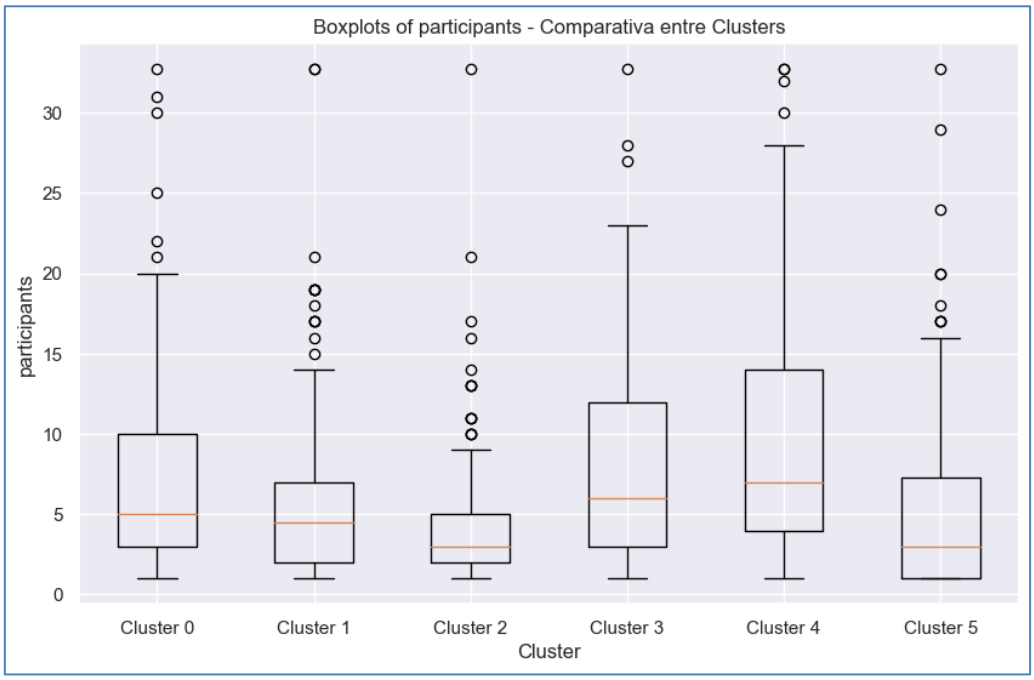


Figura 17. Gráficos de Caja con el total de participantes en las rondas de financiación.

En el caso de las variables de Series A y Series B el recorte de valores extremos del 95% no es suficiente para permitir obtener una gráfica clara sobre las distribuciones de cada clúster. Sin embargo, si se pueden apreciar varios datos relevantes sobre las startups. Por ejemplo, mientras que en las Series A las startups del grupo 0 reciben cantidades mayores de financiacion, de media, en el caso de las Series B, son las startups del grupo 3 las que más capital reciben.

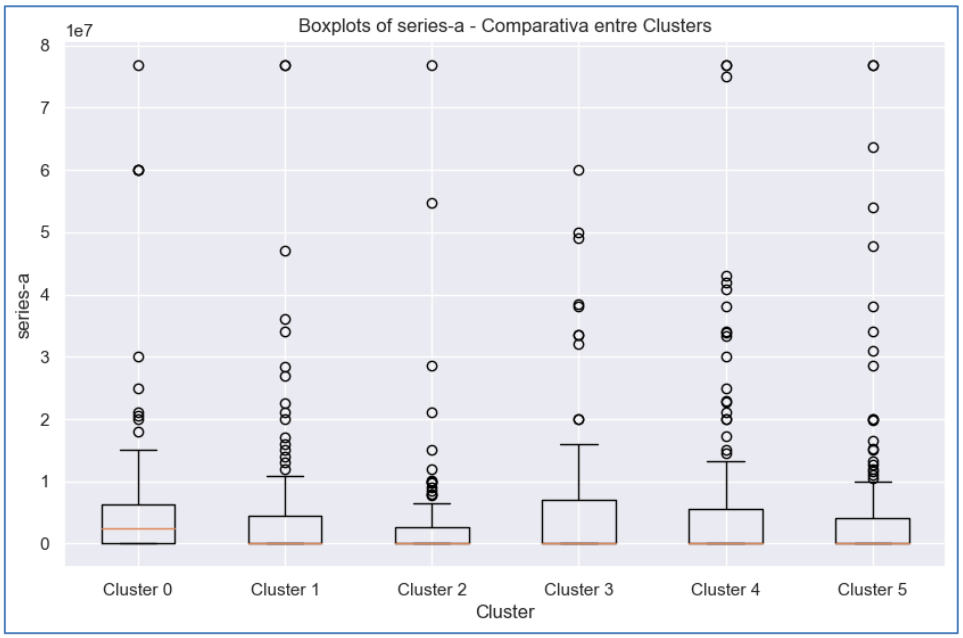


Figura 18. Gráficos de Caja con el tamaño de las Series A de las Startups.

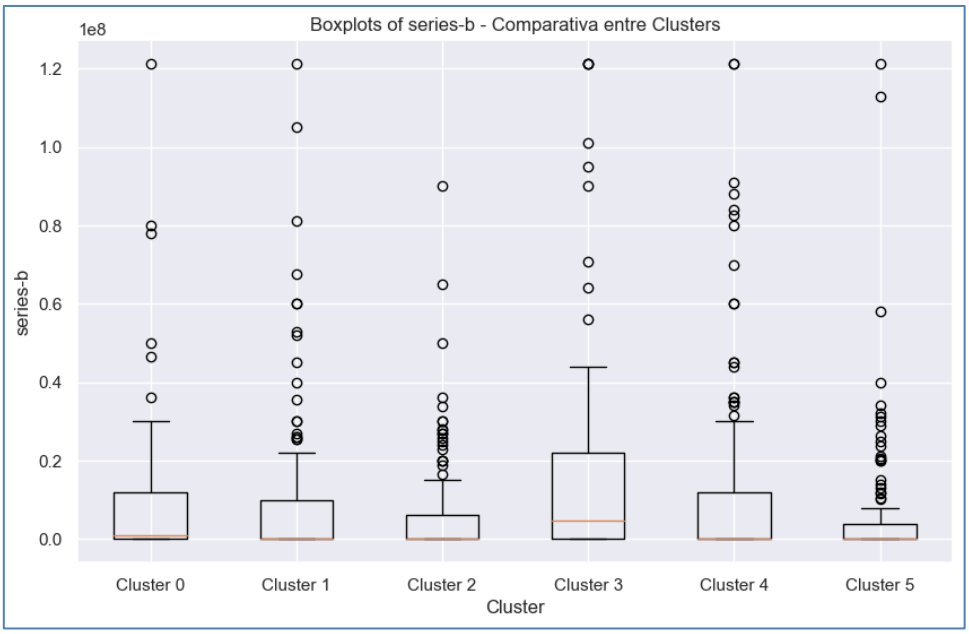


Figura 19. Gráficos de Caja con el tamaño de las Series B de las Startups.

En cuanto al tipo de operación de operación de salida de las startups existen varias diferencias claras entre los clústeres. En la Figura 19 se observa como el único grupo

donde la operación única y donde domina la IPO es en el cluster 3. Además, entre aquellos grupos que cuentan con los dos tipos de operaciones, el que mejor ratio tiene entre ambos es el grupo 4 donde se observa que aproximadamente 1 de cada 4 operaciones es una IPO. Por el contrario, el grupo 1 solo contiene operaciones de adquisición y en los grupos 0, 2 y 4 el número de startups cuya operación de exit haya sido una IPO es muy reducido.

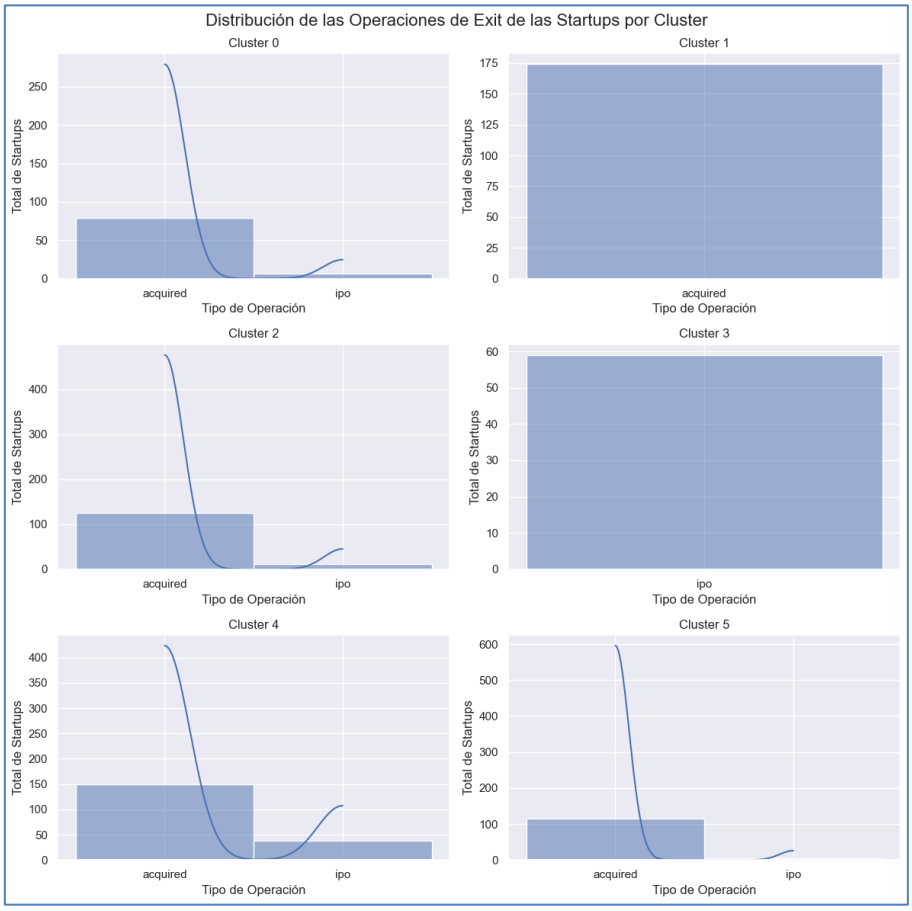


Figura 20. Distribuciones del tipo de operación de salida de la Startup.

En el caso de la regiones donde se encuentran las startups también se perciben diferencias muy significativas. Como se ha visto a lo largo de todo el proyecto, el área de San Francisco tiene una presencia muy potente en el ecosistema. En este caso, los grupos 0 y 4 agrupan a todas las startups con oficinas en la región de la bahía de San Francisco, esto incluye las zonas de Palo Alto, Sillicon Valley o Menlo Park. Por otro lado, destaca el

cluster 5 que de forma orgánica ha agrupado a la mayoría de los startups que se encuentran en los centros más potentes fuera de los Estados Unidos. Por ello, se observa una alta recurrencia de startups con sede en Londres, Tel Aviv, París o Toronto. Finalmente, el resto de los clústeres se reparten las startups que se encuentran en los centros de negocios más potentes de Estados Unidos, excluyendo San Francisco. En particular, destaca la fuerte presencia de Boston, Nueva York, Seattle y Los Ángeles.

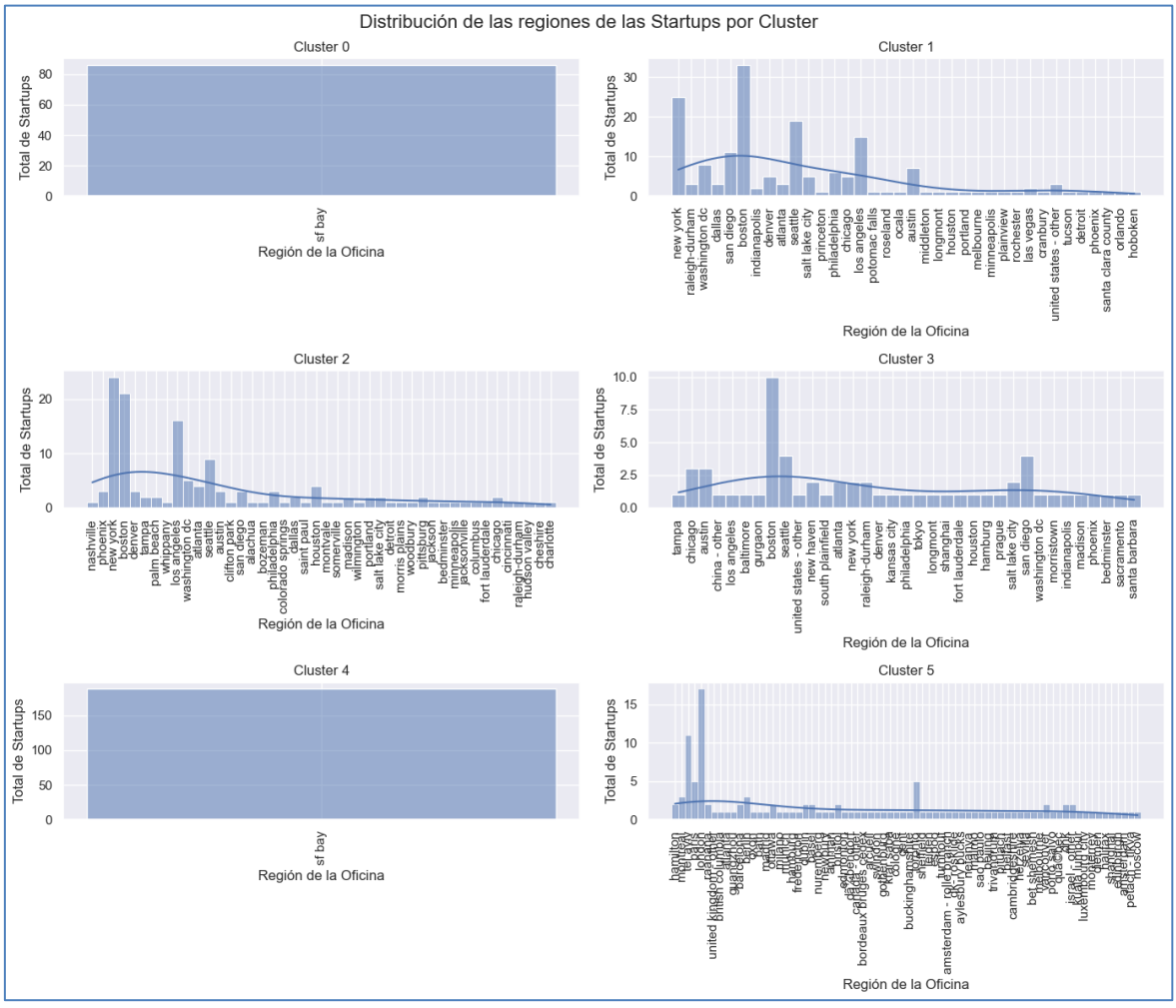


Figura 21. Distribuciones de la región de las oficinas de las Startups.

Finalmente, la universidad de procedencia de los fundadores de las startups también permite identificar diferencias claras entre los clústeres formados. Se puede observar en

la Figura 21 que hay mucha polaridad en las clases y existe una diferenciación significativa entre clústeres. Por un lado, los grupos 0, 2 y 5 cuentan con un alto número de fundadores sin grado universitario. Alternativamente, los fundadores de los clústeres 1,3 y 4 destacan por el prestigio de las universidades donde han realizado sus estudios. En particular, la mayoría de los fundadores de las startups del clúster 4 se han educado en las mejores universidades del mundo.

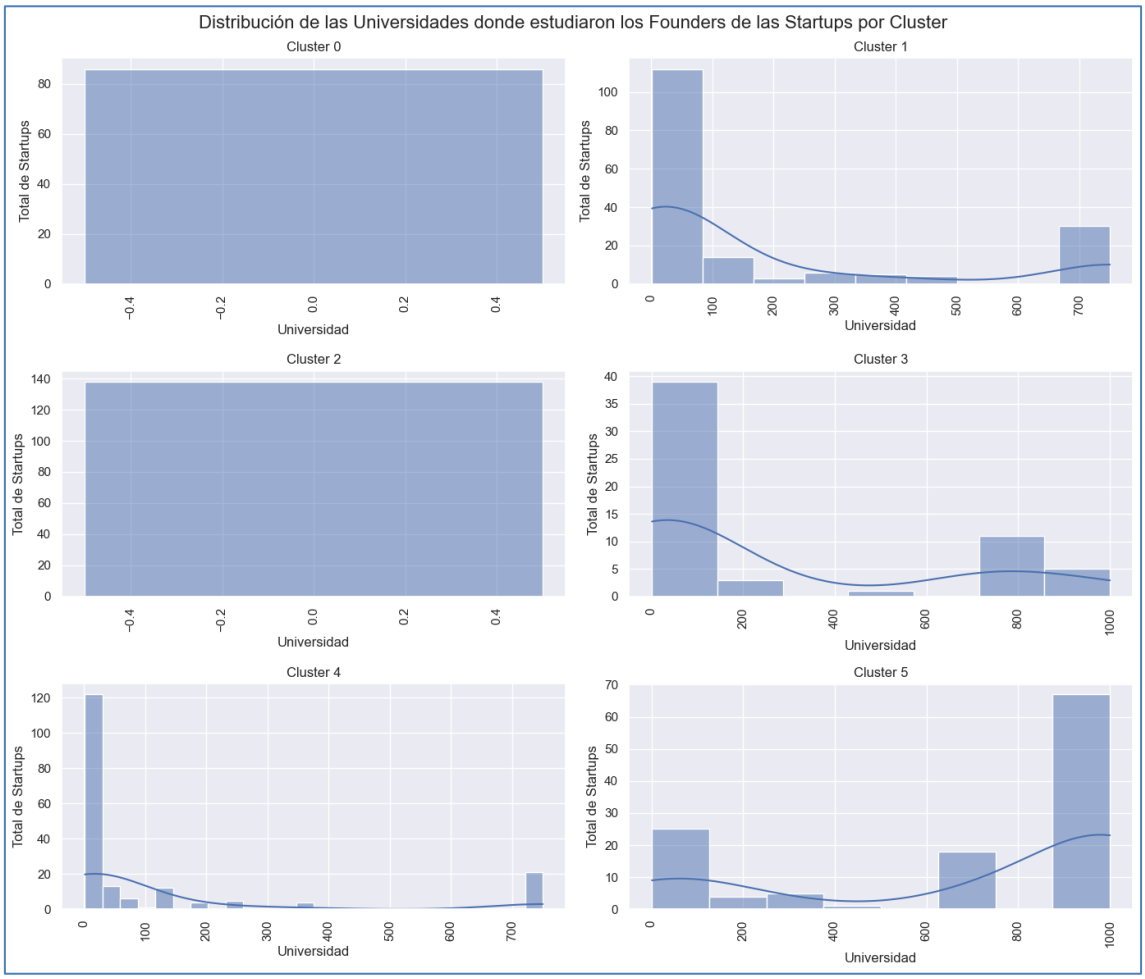


Figura 22. Distribuciones de las universidades donde estudiaron los fundadores de las Starups.

En conclusión, existen características que definen y permiten identificar a cada clúster de manera clara y por lo tanto su implementación en los modelos facilitará la labor predictiva de los algoritmos.

3.2 MODELOS DESCARTADOS

El primer paso, utiliza la base de datos con las variables categóricas y numéricas ya procesadas, desarrollada durante la implementación de los algoritmos de clustering. Para

asegurar un rendimiento óptimo de los modelos, se realizan ajustes enfocados a modificar la variable objetivo y mejorar las predicciones. Al tratar con una base de datos donde la variable objetivo varía significativamente es necesario procesar la variable para maximizar la capacidad predictiva de los modelos. Por lo tanto, tras estudiar múltiples opciones, se implementó un método para la eliminación de valores extremos [32]. El procedimiento utilizado utiliza el puntaje z (z-score) para eliminar las filas con valores extremos en la variable objetivo. El puntaje Z, también conocido como puntuación estándar, es una medida estadística que indica cuántas desviaciones estándar un valor determinado está por encima o por debajo de la media de una distribución. El z-score se utiliza para definir un umbral de corte, que en este caso se ha fijado en 3 desviaciones standard, considerando cualquier valor por encima como extremo. Eliminando valores atípicos se consigue obtener una distribución de datos más representativa y eliminar valores atípicos que puedan perjudicar a los algoritmos predictivos.

En segundo lugar, para terminar de ajustar la variable objetivo y permitir que los modelos puedan dar mejores resultados, se utilizan los logaritmos para procesar la variable. La transformación logarítmica permite lidiar con datos donde la varianza aumenta a medida que aumenta el valor de la variable objetivo, que es precisamente el problema al que se enfrenta este proyecto. La implementación de este tipo de modificación permite reducir el sesgo y facilitar la identificación de patrones por parte de los algoritmos de predicción [33]. Tras haber aplicado ambas técnicas de preprocesamiento a la variable objetivo, se obtiene la base de datos que se utilizará en el diseño de modelos, a la que se referirá como ‘data_modelos’ de aquí en adelante.

Tras analizar los clústeres, se procede a añadir la variable que contiene las etiquetas a la base de datos ‘data_modelos’. Esta variable puede aportar gran valor a los modelos de

predicción y mejorar los resultados. Además, en este apartado se comparan los resultados de los modelos.

El primer paso consta de la descripción de las medidas utilizadas para el estudio y comparación de los modelos.

3.2.1 MÉTRICAS DE ERROR

Las métricas utilizadas para comparar los modelos son RMSE y MAPE. Estas métricas permiten el acceso a información sobre le calidad de los modelos y sus capacidades predictivas.

El RMSE (error cuadrático medio) mide la desviación estándar de los errores en las predicciones. En otras palabras, la métrica calcula como de dispersos se encuentran los residuales, que son valores que miden la distancia entre los datos y la superficie de regresión. La fórmula para el RMSE entre dos puntos es la siguiente:

$$RMSE = \sqrt{(valores\ predichos - valores\ reales)^2}$$

El motivo detrás de elevar al cuadrado y luego aplicar la raíz cuadrada es para tener en cuenta aquellos casos donde las diferencias son negativas, pues al calcular el error cuadrático medio para todo un conjunto de datos el sumatorio entre aquellas diferencias positivas con las negativas, el error real se vería reducido. La ecuación que calcula el error cuadrático medio de un modelo con N casos es la siguiente [34]:

$$RMSE = \sqrt{\sum_{i=1}^{N} (valor\ predicho_i - valor\ real_i)^2 \Big/ N}$$

Al haber transformado los datos utilizando logaritmos es necesario trabajar con métricas de error que permitan lidiar con este tipo de ajustes. En el caso del RMSE el proceso para conseguir la métrica utilizando la escala original al aplicar la operación logarítmica

inversa al resultado obtenido. Esto proporcionará una evaluación más precisa de la calidad del modelo en relación con la transformación realizada en los datos.

En segundo lugar, se utiliza la métrica de MAPE (error porcentual absoluto medio). Esta medida calcula la precisión del sistema de predicción. El MAPE mide la diferencia porcentual promedio entre los valores predichos y los datos originales. De nuevo, para calcular el valor del MAPE en la escala original es necesario aplicar la transformación inversa. La fórmula del error porcentual absoluto medio es la siguiente:

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Donde, n representa el número de casos.

A_t representa el valor Real.

F_t representa el valor Predicho.

Tras presentar las métricas que se utilizaron se analizan los modelos descartados explicando y comparando su funcionamiento y sus resultados.

3.2.2 REGRESIÓN LINEAL

Tras haber preparado la variable objetivo para su implementación en los algoritmos predictivos, comienza la creación de los modelos. El primer modelo que se utiliza es una regresión lineal. La regresión lineal es uno de los algoritmos más utilizados en el ámbito del aprendizaje automático (machine learning). A pesar de ser un modelo sencillo, las regresiones lineales tienen una amplia gama de aplicaciones en problemas reales y pueden proporcionar una alta capacidad predictiva. El objetivo principal de este algoritmo es encontrar la mejor línea recta que se ajuste a todos los datos, minimizando la diferencia entre los valores reales y los valores predichos [35].

Los resultados de la regresión lineal tras la transformación inversa, o, en otras palabras, ajustado a la escala original, dan valores infinitos. Esto implica que el modelo no es capaz de predecir de forma precisa los valores de IPO/adquisición de la empresa, y que por lo tanto no es capaz de predecir el éxito de la inversión en startups.

```
Linear Regression RMSE: 21986237128.957985
Average Error ($): inf
Linear Regression MAPE: 535029783728858.2
Average MAPE (%): inf
```

Figura 23. Resultados de la Regresión Lineal con el Conjunto Test.

A pesar de los pésimos resultados de la regresión lineal, conviene analizar los coeficientes de las variables más significativas del modelo. Esto dará una mejor imagen sobre cómo afectan algunas de las variables al modelo y además permitirá conocer cuáles son las variables más relevantes.

Variable	Coefficient	P-Value
region_sheffield	-6.299688	0.000032
status_ipo	0.947940	0.000305
region_indianapolis	3.027884	0.006850
Nº de Operaciones	2.139468	0.007969
category_code_web	-0.861336	0.015033
series-a	2.307135	0.016567
people	-2.637670	0.025706
first_funding_at	-2.171999	0.028845
relationships	5.526787	0.039207
funding_rounds	-1.977028	0.044353

Figura 24. Variables Más Significativas.

En la figura 23 se observa el impacto de las variables más importantes del modelo. En primer lugar, se observa el impacto negativo que tiene el tener la oficina de tu startup en Sheffield. Esto muestra la importancia que tiene una buena localización en el éxito de la startup. El rodearse de talento y de personas afines tiene un impacto significativo sobre el éxito de la startup por lo que situarse fuera de estos centros neurálgicos del ecosistema

puede tener un impacto negativo como el visto en el caso de la variable 'region_sheffield'. En segundo lugar, se observa la alta importancia que tiene la variable IPO y su impacto positivo sobre el éxito de las compañías. Esto se debe a que por lo general las IPOs dan lugar a mayores valoraciones que las adquisiciones por parte de empresas privadas o fondos [36]. Finalmente, otra variable destacable es la que hace referencia a la series A. El coeficiente sugiere que, a mayores valores de la variable, mayores son las probabilidades de éxito. En particular, resulta interesante este caso, pues las Series A se consideran rondas iniciales, lejos aún del momento de salida. Sin embargo, el modelo identifica esta variable como una de las más importantes, mostrando la importancia que tiene un buen comienzo y el apoyo financiero al inicio sobre el éxito de la startup.

Por otro lado, a raíz de los resultados, resulta conveniente realizar un análisis de colinealidad. El análisis de colinealidad se realiza para identificar la presencia de alta correlación entre las variables independientes en un modelo de regresión. La colinealidad ocurre cuando hay una fuerte relación lineal entre dos o más variables predictoras, lo que puede causar problemas en la interpretación del modelo y en la estimación de los coeficientes.

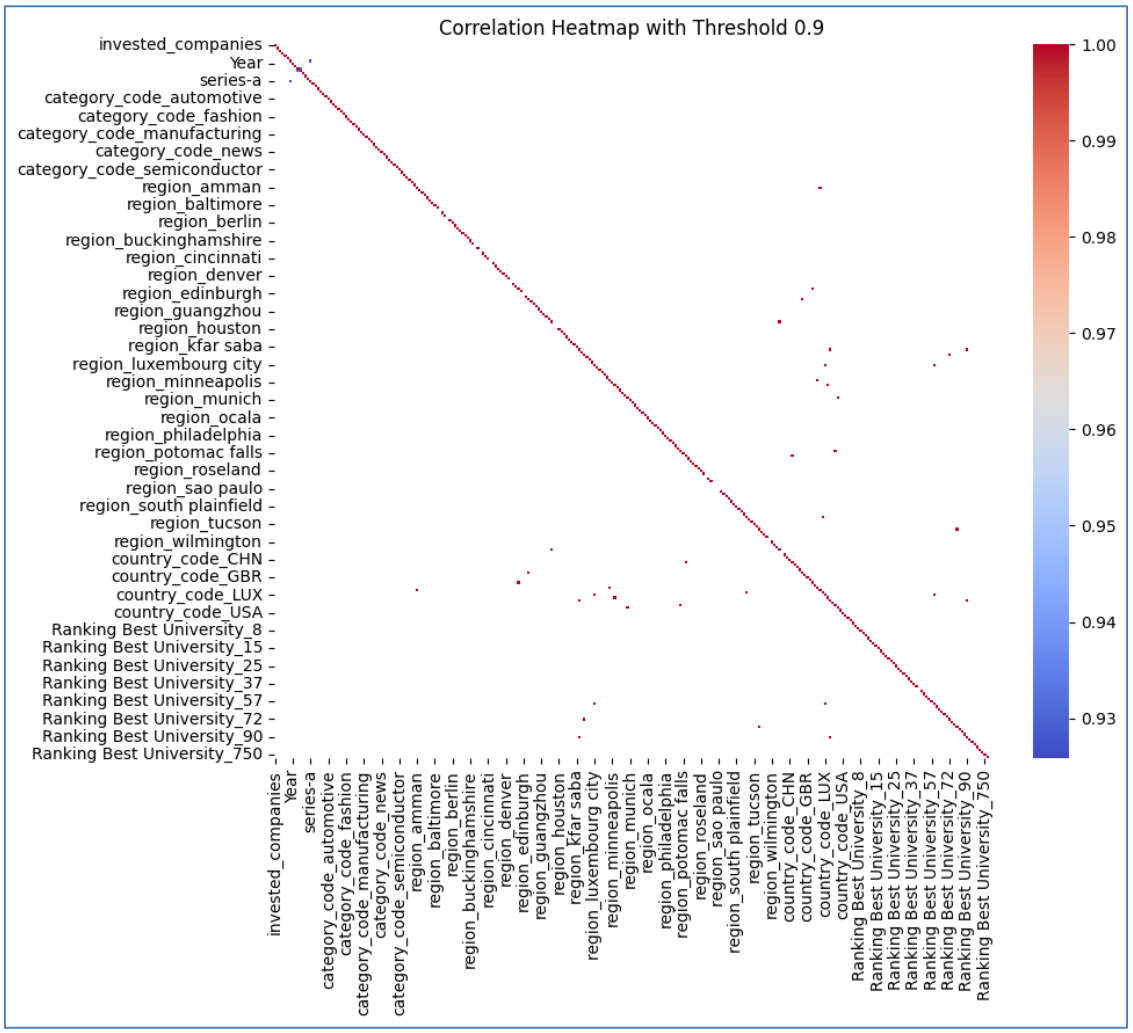


Figura 25. Gráfico mostrando aquellas variables correlacionadas con un coeficiente superior al 90%.

En la Figura 24 se observan aquellas variables que están altamente correlacionadas, con un coeficiente de correlación superior al 90%. Utilizando este gráfico se procede a realizar un modelo eliminando todas las variables en el gráfico. El objetivo de realizar un modelo sin las variables altamente correlacionadas es que se pueden lograr los siguientes objetivos:

1. Reducción de la redundancia: Si dos variables están altamente correlacionadas, es probable que proporcionen información similar al modelo. En este caso, se puede

considerar eliminar una de las variables para reducir la redundancia y simplificar el modelo.

2. Estabilidad de los coeficientes: La presencia de colinealidad puede hacer que los coeficientes de regresión sean inestables y difíciles de interpretar [37].
3. Interpretación adecuada: Si las variables están altamente correlacionadas, es difícil atribuir cambios en la variable de respuesta a una variable específica. Esto puede dificultar la interpretación adecuada de los efectos individuales de las variables predictoras en el modelo[38].

Sin embargo, tras eliminar al menos una de las variables que tenían correlación de cada pareja y realizar un modelo con la nueva base de datos, los resultados obtenidos no presentan ninguna mejora con respecto a los del modelo inicial de regresión lineal.

```
Resultados Test
Decision Tree Regression RMSE: 1.9220821482878058e+22
Average Error ($): inf
Decision Tree Regression MAPE: 4781118570098622.0
Average MAPE (%): inf
```

Figura 26. Resultados de la regresión lineal tras eliminar variables correlacionadas.

3.2.3 ÁRBOLES DE DECISIÓN

El segundo modelo implementado es un árbol de decisión. Los árboles de decisión utilizan una estructura de árbol para tomar decisiones basadas en las características de los datos. Este algoritmo separa los datos en subconjuntos que van disminuyendo en tamaño a medida que el árbol va implementando nuevas ramas. Estas ramas consisten en reglas de decisión que constan de dos opciones.

Los árboles de decisión tienen la capacidad de trabajar con variables categóricas y numéricas. Estos algoritmos dividen el conjunto de datos en diferentes nodos o ramas, utilizando criterios como la ganancia de información o la reducción de la varianza para determinar la mejor forma de dividir los datos en cada nodo. Cada división en el árbol

representa una regla de decisión basada en una característica o atributo particular. En el caso de las variables numéricas se suelen usar los operadores de mayor dé y menor que, generando dos subconjuntos de datos para dividir por rama.

Después de evaluar el modelo de regresión lineal y compararlo con los modelos basados en árboles de decisión, se observó una mejora significativa en la calidad de las predicciones. Los árboles de decisión han demostrado un rendimiento notablemente superior en términos de precisión y capacidad de generalización. Esta mejora significativa en la calidad de las predicciones representa un avance importante en el desarrollo del proyecto, ya que indica que los modelos basados en árboles son más capaces de capturar las relaciones y patrones subyacentes en los datos.

```
Resultados Train:
Decision Tree Regression RMSE: 0.0
Average Error ($): 0.0
Decision Tree Regression MAPE: 0.0
Average MAPE (%): 0.0
Resultados Test
Decision Tree Regression RMSE: 1.7133014766253982
Average Error ($): 4.547245380035833
Decision Tree Regression MAPE: 7.169376401554898
Average MAPE (%): 1298.0342744168654
```

Figura 27. Resultados de los Árboles de Decisión

Como se observa en la Figura 26, el árbol de decisión es capaz de reducir a 0 el error del modelo. Esto sugiere un rendimiento bastante bueno. Además, con el conjunto de prueba el error medio del modelo es de 4.55 dólares. Resulta sorprendente ver como a pesar de que el modelo tiene que tratar con cifras de adquisiciones e IPOs que rondan entre los 10 y los 1000 millones de dólares el algoritmo es capaz de predecir con cierta precisión el valor de la variable objetivo. Con respecto al MAPE, el error test obtenido por esta métrica sugiere un rendimiento muy inferior al representado por el RMSE. Esto se puede

deber a la naturaleza sesgada de los datos. Mientras el RMSE se ve afectado por los errores absolutos de las predicciones el MAPE mide los errores relativos en forma de porcentaje. Entonces, al tener unos datos extremadamente altos en la variable objetivo, incluso una pequeña discrepancia en las predicciones puede resultar en un MAPE alto debido a la naturaleza del cálculo porcentual [39].

Sin embargo, los increíbles resultados presentados por el conjunto train muestran claramente uno de los problemas de los árboles de decisión, que es el sobreajuste a los datos de entrenamiento, si no se controla su crecimiento. Este problema, conocido como sobreajuste (Overfitting), puede llevar a un rendimiento excelente en el conjunto de entrenamiento, pero un rendimiento deficiente en datos desconocidos, como el conjunto de prueba. Para evitar este problema, se aplican técnicas de poda (pruning) para mejorar la precisión y la generalización del modelo. El pruning consiste en eliminar ramas o nodos del árbol que no aportan un beneficio significativo a la precisión, reduciendo así la complejidad del árbol. Esto ayuda a evitar el sobreajuste y mejorar la capacidad de generalización del modelo [40]. En la figura 27 se muestran los resultados con el conjunto de prueba.

```
Resultados Train:
Decision Tree Regression RMSE: 0.81283409926455
Average Error ($): 1.2542878173146512
Decision Tree Regression MAPE: 3.526015499359382
Average MAPE (%): 32.988271164758046
Resultados Test
Decision Tree Regression RMSE: 1.4155228033289509
Average Error ($): 3.1186391414651773
Decision Tree Regression MAPE: 5.847277300505051
Average MAPE (%): 345.290251479394
```

Figura 28. Resultados del Algoritmo de Árboles de Decisión tras implementar poda.

Como se observa en la Figura, a pesar de haber aumentado el error de los datos con el conjunto de entrenamiento, los resultados del conjunto de prueba se han reducido

significativamente, demostrando una mayor capacidad de generalización del modelo y un mejor desempeño frente a datos desconocidos, que es lo que se busca de un modelo de inteligencia artificial.

3.2.4 RANDOM FOREST

El algoritmo de random forest se construye a partir de la combinación de múltiples árboles de decisión individuales, obteniendo así un modelo más robusto y preciso. Cuando se realiza una predicción con un Random Forest, cada árbol individual del conjunto emite su propia predicción y luego se promedian los resultados (en el caso de regresión) para obtener una predicción final. Esta agregación de múltiples predicciones ayuda a reducir el sesgo y la varianza, mejorando así la precisión general del modelo. Los random forest permiten conocer cuáles son las variables más significativas de un modelo.

El random forest ha sido el modelo descartado que segundo mejor rendimiento ha dado. El error promedio obtenido muestra una ligera pero contundente mejoría sobre los modelos anteriores. Uno de los motivos detrás de este alto rendimiento son la alta robustez de este algoritmo frente a valores atípicos, que son muy comunes en esta base de datos [41]. Además, los random forest pueden manejar de forma eficiente conjuntos de datos con variables categóricas y numéricas. No obstante, el problema que ocurre con este modelo es que, al tener un componente aleatorio, hay un cierto grado de variabilidad y cada vez que se entrena el algoritmo los resultados varían significativamente.

```
RMSE: 1.3074985450705552
Average Error ($): 2.696914471323427
MAPE: 6.490693980872678
Average MAPE (%): 657.9805244457303
```

Figura 29. Resultados del modelo de Random Forest.

Finalmente, hay que añadir que a este modelo se le han aplicado técnicas de ajuste de hiperparámetros, reduciendo el error medio hasta 2.60$ y convirtiéndolo así en el tercer mejor modelo.

Cuando se dice que una variable es importante según un Random Forest, significa que esa variable tiene un impacto significativo en la predicción realizada por el modelo. En otras palabras, esa variable tiene una fuerte relación con la variable objetivo que se está tratando de predecir.

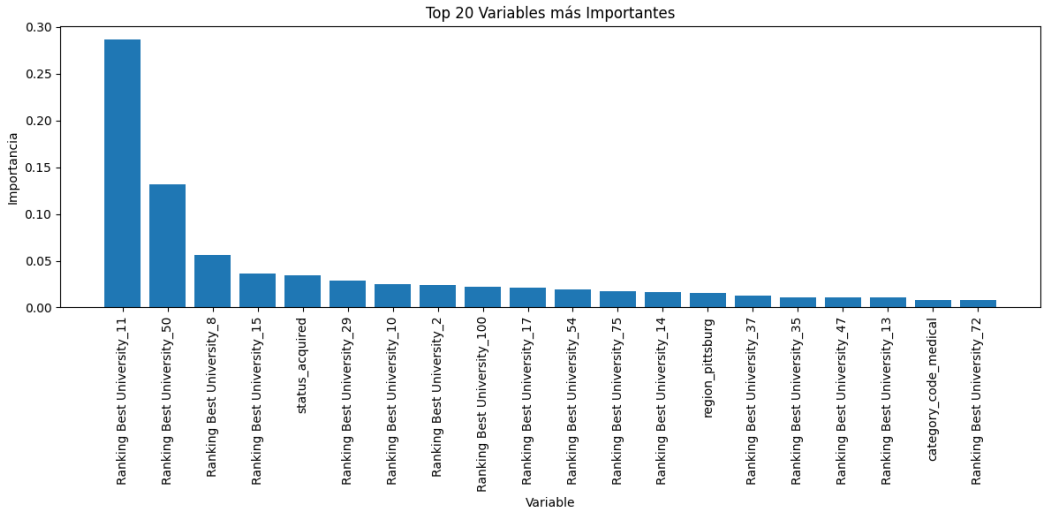


Figura 30. Variables más significativas.

Como se observa en la Figura 29, todas las variables entre las 20 más importantes son variables categóricas. Además, gran parte de esas variables están relacionadas con la universidad donde se ha formado el fundador o CEO la startup. Este detalle es importante, pues uno de los objetivos del proyecto es conocer la influencia que tienen los estudios de los fundadores sobre el éxito de las startups. Por otro lado, el hecho de que una startup sea adquirida también es muy importante a la hora de predecir el valor de la compra. Finalmente, la región de Pittsburg y las categoría de medical completan la lista de variables más significativas del modelo. El hecho de que la educación sea un componente clave y además en uno de los mejores modelos obtenidos muestra la importancia y el impacto que puede llegar a tener sobre una startup.

3.2.5 ALGORITMOS DE BOOSTING

El boosting es una técnica de ensamblado de modelos que combina múltiples modelos más débiles para crear un modelo más fuerte y preciso. En lugar de entrenar un solo modelo grande, el boosting entrena iterativamente una secuencia de modelos más simples, denominados 'weak learners', donde cada uno se enfoca en corregir los errores del modelo anterior [42]. El objetivo de los modelos de boosting es reducir el sesgo (bias). En este caso, se han implementado y comparado tres algoritmos.

El primer algoritmo de boosting implementado es AdaBoost. Adaboost es un algoritmo de aprendizaje secuencial. Esto implica que distintos modelos débiles se generan secuencialmente y los errores de los modelos previos son aprendidos por los sucesores, permitiendo que el modelo les dé más importancia a los errores más difíciles de corregir [43].

El segundo método implementado es el gradient boosting. A diferencia de AdaBoost, Gradient Boosting ajusta iterativamente los modelos débiles para minimizar la función de

pérdida mediante la optimización de los gradientes de esta función. En cada iteración, el nuevo modelo se ajusta a los residuos (diferencia entre las predicciones actuales y los valores reales) del modelo anterior. Una vez se obtienen los resultados de los modelos débiles se combinan de forma ponderada basándose en el rendimiento de cada modelo [44].

Finalmente, se ha implementado XGBoost, que es una implementación optimizada y altamente eficiente de Gradient Boosting. XGBoost es un algoritmo de ensamblado basado en árboles de decisión. Las optimizaciones del modelo que hacen de XGBoost un modelo tan potente son las siguientes [45]:

- Paralelización: XGBoost aborda el proceso de construcción secuencial de árboles utilizando una implementación paralelizada. Al poder construir de forma paralela, el tiempo de ejecución se reduce. Además, se intercambia el orden de los bucles mediante la inicialización a través de un escaneo global de todas las instancias y la clasificación mediante hilos paralelos, mejorando el rendimiento del algoritmo al compensar cualquier sobrecarga de paralelización en la computación.
- Poda de árboles: El algoritmo de XGBoost utiliza el parámetro 'max_depth' para la poda de los árboles. Este enfoque mejora significativamente el rendimiento computacional.
- Optimización de hardware: Este algoritmo ha sido diseñado para hacer un uso eficiente de los recursos de hardware.
- Mejoras algorítmicas: XGBoost también implementa otras mejoras algorítmicas como la regularización para evitar el sobreajuste o la validación cruzada para mejorar la capacidad de generalización del modelo.

Los resultados obtenidos de los modelos de Boosting suponen una mejora sobre los árboles de decisión, pero no consiguen mejorar los resultados de los random forest. El boosting destaca por su capacidad para mejorar el rendimiento predictivo y reducir el sesgo. Este efecto se ve reflejado en los resultados obtenidos, donde el RMSE muy bajos. De nuevo surge el problema relacionado con la alta varianza y el efecto de los valores

atípicos sobre el MAPE, pero este efecto es ligeramente inferior al visto en modelos anteriores.

```
GradientBoostingRegressor RMSE score: 1.3099123621964048
GradientBoostingRegressor Average Error ($): 2.7058489255183225
GradientBoostingRegressor MAPE score: 5.494225690179402
GradientBoostingRegressor Average MAPE (%): 242.28307673546337


XGBRegressor RMSE score: 1.4092411743538382
XGBRegressor Average Error ($): 3.092848467025873
XGBRegressor MAPE score: 5.92648957952351
XGBRegressor Average MAPE (%): 373.8363682932482


AdaBoostRegressor RMSE score: 1.4647109266368288
AdaBoostRegressor Average Error ($): 3.326292443616743
AdaBoostRegressor MAPE score: 6.125485724102018
AdaBoostRegressor Average MAPE (%): 456.3668134095827
```

Figura 31. Resultados para Modelos de Boosting.

Por otro lado, resulta curioso observar unos resultados mejores para Gradient Boosting que para XGBoost, pues a priori, XGBoost es una mejora del Gradient Boosting. No obstante, existen algunas explicaciones posibles para que ocurra esta situación. Por ejemplo, es posible que el tamaño de la base de datos, que en este caso es pequeño, reduzca la efectividad del modelo, pues los efectos de la paralelización no se explotan de forma óptima. Otra posible razón es que el modelo de Gradient Boosting, al ser más simple, se ajuste mejor a los datos [46].

Por otro lado, en el caso de los algoritmos de boosting, a diferencia de lo experimentado con los random forest, si se observa una clara mejoría tras la aplicación de algoritmos de ajuste de hiperparámetros.

En primer lugar, el AdaBoost es capaz de reducir el error medio hasta 3.04 $ por predicción, que supone una mejora del 10%. En cuanto al algoritmo de GradientBoosting, el ajuste de los hiperparámetros reduce el error medio a 2.61$, que en este caso es una reducción mucho menor que en el caso del AdaBoost. Esto se puede deber a que

inicialmente el modelo de GradientBoosting se encuentra más cerca de maximizar la capacidad predictiva del modelo. Finalmente, el caso de XGBoost es el más sorprendente de los 3, pues el ajuste de hiperparámetros es capaz de reducir el error medio hasta 2.54$ por predicción, que supone una mejora de 45 céntimos de media por predicción. Además, convierte al modelo de XGBoost en la mejor alternativa al modelo final.

3.2.6 SUPPORT VECTOR REGRESSION (SVR)

El modelo SVR es un algoritmo basado en las máquinas de vectores de soporte (support vector machines) utilizado en casos de regresión. El objetivo del modelo es encontrar una función que se ajuste mejor a los datos de entrenamiento, reduciendo al mínimo el margen de error [47].

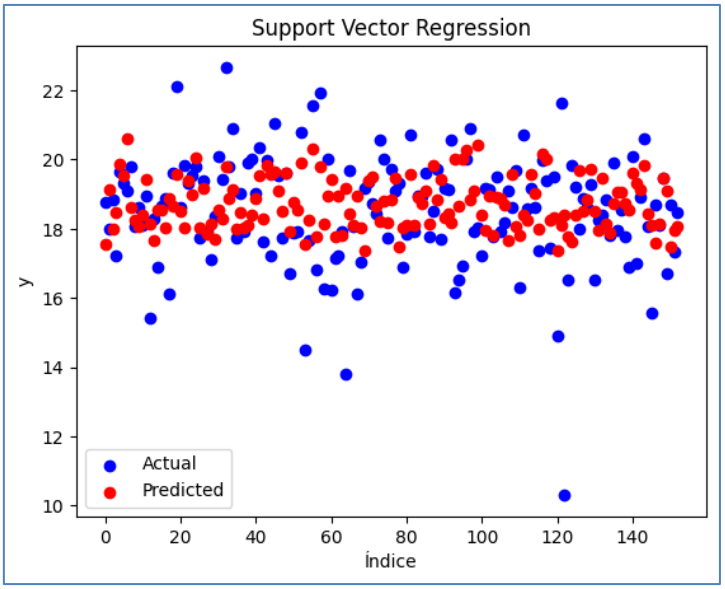


Figura 32. Resultados del Modelo SVR

A priori los resultados obtenidos sugieren un peor desempeño del modelo SVR frente al visto en la regresión lineal. Sin embargo, se presta mayor atención, se observa que el eje

Y no tiene ningún exponente y que la diferencia entre los valores actuales y los predichos es en realidad mucho menor a la del modelo de regresión lineal. Este modelo presenta resultados ligeramente inferiores a los obtenidos por los modelos de boosting.

```
Resultados Train:
SVR Regression RMSE: 1.3521762008937361
Average Error ($): 2.865829204308865
SVR MAPE: 5.225139948200164
Average MAPE (%): 184.88718328242632
Resultados Test
SVR RMSE: 1.5607624567068386
Average Error ($): 3.762451023911596
SVR MAPE: 6.499132385751725
Average MAPE (%): 663.5647969582001
```

Figura 33. Resultados para SVR.

Tras la aplicación de algoritmos de ajuste los hiperparámetros, el error medio por predicción se reduce hasta 3.64$. Esta variación no es muy significativa, por ello, si graficamos la comparativa entre los valores actuales y los predichos por el mejor modelo, no se puede apreciar la diferencia con la figura 32.

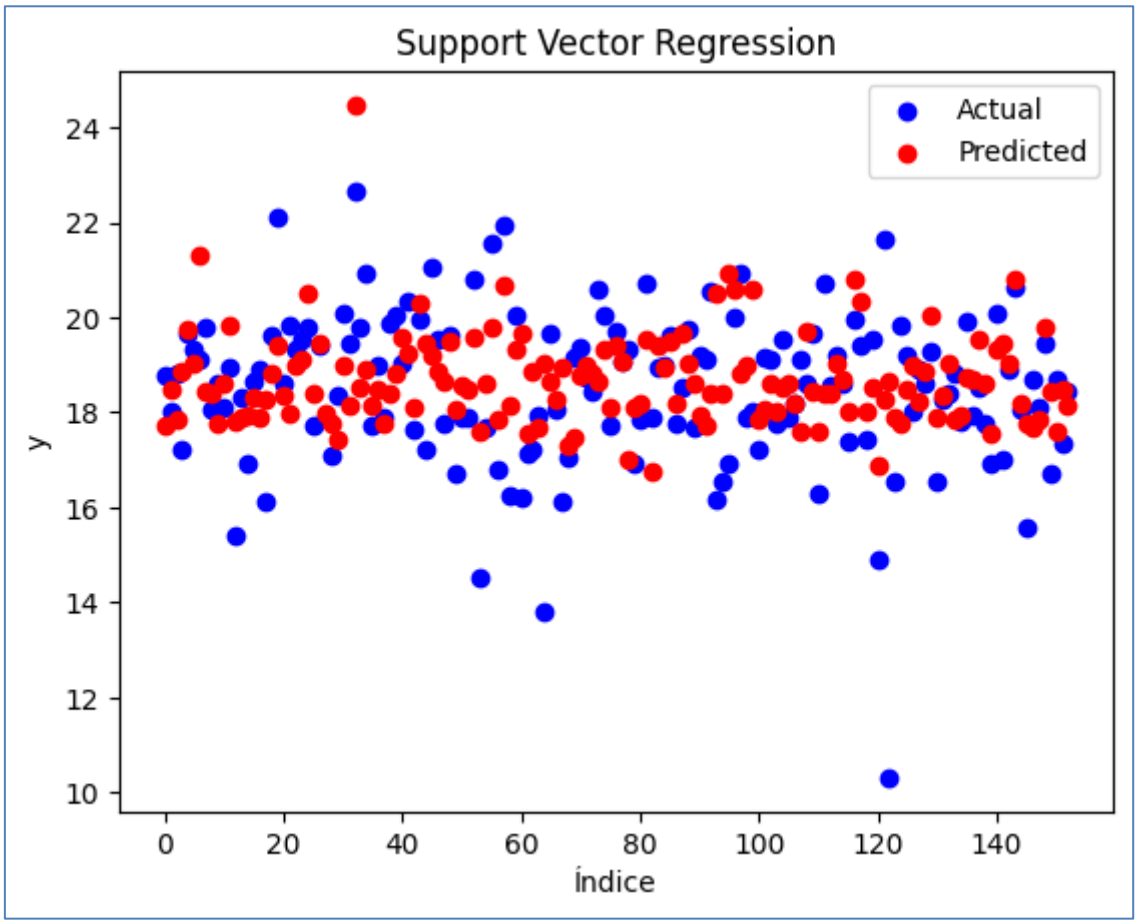


Figura 34. Comparativa Valores Actuales vs Valores Predichos por el Modelo con parámetros tuneados.

3.2.7 REDES NEURONALES

El último modelo que se descartó fue uno basado en redes neuronales. Las redes neuronales consisten en neuronas (nodos) interconectadas entre sí. Los nodos procesan las entradas que reciben de las capas anteriores y luego transmiten su salida a los nodos de la siguiente capa. Las capas de estas neuronas se apilan juntas para formar una red [48]. La estructura típica de una red neuronal comienza con una capa de entrada que recibe los datos en bruto y termina con la capa de salida encargada de realizar la predicción final. Las capas ocultas analizan los patrones y aprenden de ellos para futuras estimaciones.

Cada conexión entre los nodos tiene un peso asignado, que se aprende durante la fase de entrenamiento. Además de los pesos, los nodos tienen un término de sesgo para ajustar sus salidas y funciones de activación que determinan su salida dada una entrada [49].

La capa de entrada para el modelo final utiliza una capa Densa con 128 unidades. En general, surge un desafío significativo al seleccionar el número de neuronas en una capa. Este desafío consiste en el equilibrio entre la velocidad de predicción (o clasificación en este caso) y el error de predicción [50]. Un mayor número de neuronas podrá identificar patrones más complejos, pero a costa de aumentar el tiempo de entrenamiento del algoritmo. Esta capa utiliza la función de activación ReLU[19] para implementar la no linealidad en las redes neuronales. La función de activación lineal rectificada (ReLU) supera el problema del desvanecimiento del gradiente[20], permitiendo que los modelos aprendan más rápido y tengan un mejor rendimiento. Tras la capa inicial, se implementa otra capa con las mismas características con el objetivo de detectar los patrones más complejos y las relaciones profundas de los datos. Finalmente, tras las dos capas densas se añade una capa de salida con un solo nodo, el cual aplicará una función de activación lineal para predecir cada caso. En la Figura 34 se observa la estructura de la red neuronal diseñada.

[19] La función de activación lineal rectificada o ReLU en resumen es una función lineal por partes que devolverá la entrada directamente si es positiva; de lo contrario, devolverá cero [51].
[20] El problema del desvanecimiento del gradiente surge cuando se agregan más capas a las redes neuronales que utilizan ciertas funciones de activación. Los gradientes de la función de pérdida se acercan a cero, lo que dificulta el entrenamiento de la red [52].

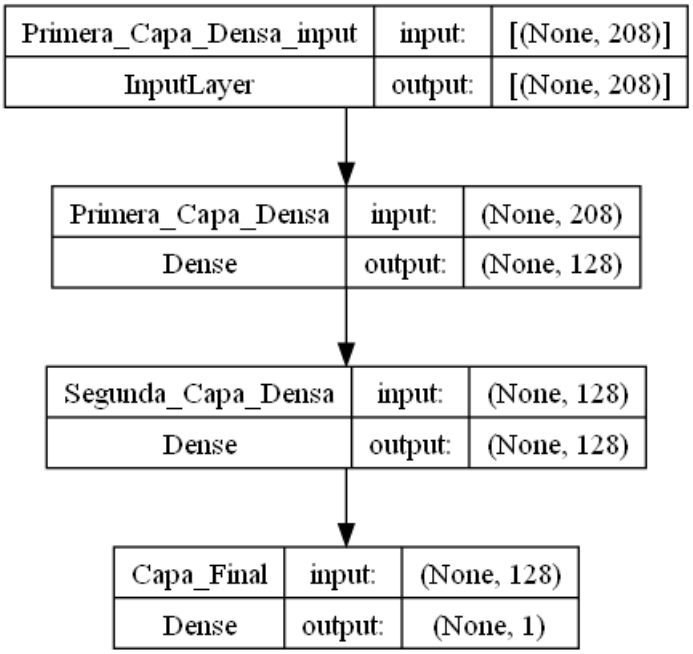


Figura 35. Estructura de la Red Neuronal

Por otro lado, las redes neuronales han mostrado un empeoramiento en la capacidad predictiva comparado con varios modelos previos. Una posible razón es que las redes neuronales requieren grandes conjuntos de datos para generalizar adecuadamente y evitar el sobreajuste. Si el conjunto de datos es pequeño, las redes neuronales pueden tener dificultades para aprender patrones significativos y pueden dar lugar a sobreajustes. Además, las redes neuronales son modelos complejos con muchos parámetros ajustables, lo que puede dificultar su optimización y configuración adecuada.

```
Resultados Train
Custom Neural Network RMSE score: 1.3521762008937361
Custom Neural Network Average Error ($): 2.865829204308865
Custom Neural Network MAPE score: 5.225139948200164
Custom Neural Network Average MAPE ($): 184.88718328242632

---

5/5 [==============================] - 0s 3ms/step
Resultados Test
Custom Neural Network RMSE score: 1.9863809197719264
Custom Neural Network Average Error ($): 6.289106109455754
Custom Neural Network MAPE score: 9.93627436623387
Custom Neural Network Average MAPE ($): 20665.604478171772

---
```

Figura 36. Resultados de las Redes Neuronales.

.

CAPÍTULO 4. MODELO FINAL

En esta capítulo se presenta un análisis en detalle del modelo final y de los resultados obtenidos.

4.1 EXPLICACIÓN DEL MODELO FINAL

El primer paso en el diseño del modelo final, al igual que se ha realizado para los modelos descartados, es eliminar los valores extremos utilizando el método del valor z. Posteriormente, se aplica el preprocesado descrito previamente, donde se normalizan las variables numéricas y se codifican las categóricas. Finalmente, se separan los datos en el conjunto de entrenamiento y el conjunto de prueba.

En segundo lugar, se replica el análisis del número de clústeres que se realizó inicialmente para entender el conjunto de datos. El método escogido para conocer el número de clústeres es KMeans. En este caso, tras haber aplicado en dos ocasiones la separación de los datos en los conjuntos, el número de idóneo de clústeres es seis. No obstante, algunos de los grupos generados son ligeramente distintos a los obtenidos en el primer algoritmo de clustering.

Al contar con una menor cantidad de casos debido a la división previa de los datos en conjuntos de entrenamiento y prueba, algunos de los clústeres generados en este modelo son ligeramente diferentes a los obtenidos en el primer modelo de clustering. Sin embargo, en su mayoría, los clústeres generados en este modelo comparten características muy similares a los obtenidos anteriormente. Es importante destacar que los grupos 5 presentan una distribución prácticamente idéntica en cuanto a las variables categóricas y

los rangos de valores en los atributos numéricos. Lo mismo ocurre con el grupo 4 de este modelo y el grupo 1 del primer algoritmo de clustering. Además, se observa una notable similitud entre los grupos 0 y 3, así como entre los grupos 2 y 4. En cuanto a los clústeres 1 y 3 de este modelo, es importante destacar las diferencias significativas que presentan en comparación con sus respectivos clústeres del primer algoritmo de clustering. Estas diferencias serán detalladas en su sección correspondiente, donde se analizará en profundidad cada uno de los clústeres y se explicarán las características distintivas que los separan de sus contrapartes en el primer modelo.

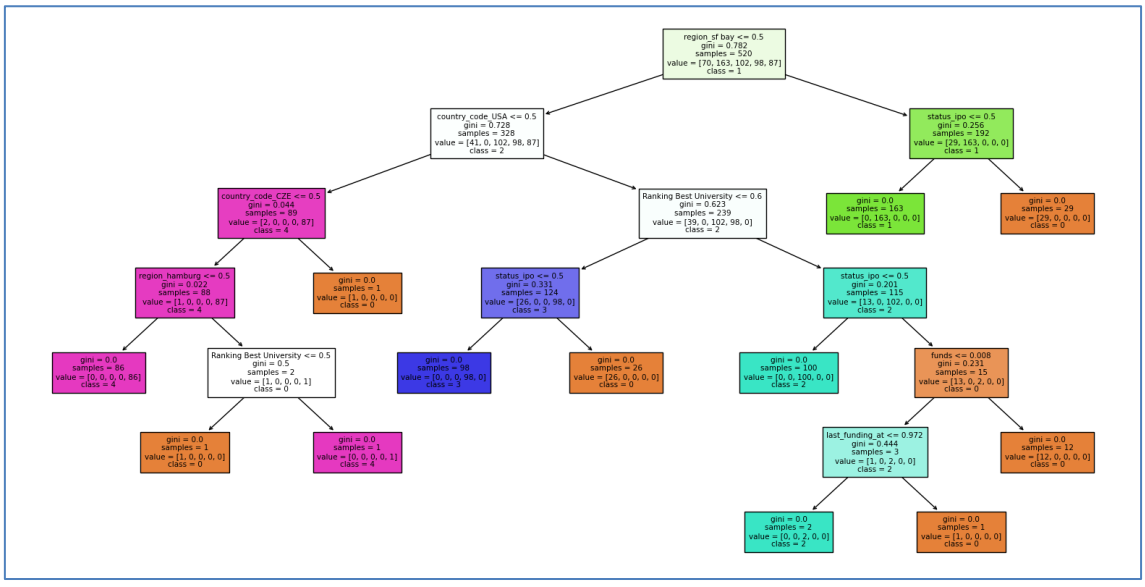


Figura 37. Árbol de Decisión Obtenido para la Diferenciación de los Clústeres.

Tras haber analizado los nuevos grupos formados a partir del KMeans se procede a aplicar todas los modelos distintos analizados previamente a cada clúster. Esto implica que a cada grupo se le aplicarán los siguientes modelos:

- Regresión Lineal
- Árboles de Decisión
- Random Forest

- GradientBoostingRegressor
- AdaBoostRegressor
- XGBRegressor
- SVR

Como se observa, las redes neuronales no se han aplicado en estos modelos. Esto se debe a que, dado el bajo número de casos en cada grupo y su bajo desempeño del modelo descartado, no tiene sentido implementar este modelo en este caso.

Para cada grupo, se escogió el modelo que mejor se adapta y mejores resultados genera, aplicándole a los modelos la técnica de hyperparameter tuning para maximizar el rendimiento del modelo final. A continuación, se presenta el análisis realizado a los nuevos clústeres y se explican los distintos modelos obtenidos para cada grupo.

4.1.1 GRUPO 0

El clúster 0, con una población de 78 startups, es uno de los clústeres que se identifican con mayor facilidad. Para este grupo, el mejor algoritmo es el Random Forest. Los parámetros ajustados para obtener el mejor modelo son los siguientes [53]:

1. 'n_estimators': Este parámetro se utiliza para decidir el número de árboles de decisión en el bosque. Cuanto más alto sea el número de estimadores, mejor rendimiento tendrá el modelo. No obstante, el tiempo de entrenamiento y la complejidad del modelo aumentan junto al rendimiento.
2. ‘max_depth’: La profundidad máxima de un árbol de decisión. Este valor determina la cantidad de divisiones antes de llegar a los nodos terminales. Si el valor es demasiado bajo el modelo no se ajustará correctamente a los datos y la capacidad predictiva sufrirá. No obstante, un valor demasiado alto da lugar a overfitting [54].
3. 'min_samples_split': El ‘min_samples_split' representa el número mínimo de muestras requeridas para dividir un nodo interno del árbol. Si el número de

muestras aumenta, cada árbol podrá considerar más muestras en cada nodo. Un valor más alto reduce la complejidad del modelo y disminuye el riesgo de sobreajuste [55].

4. 'min_samples_leaf': Este parámetro define el número mínimo de muestras requeridas en una hoja del árbol de decisión. A mayor valor de este parámetro, mayores son las probabilidades de sufrir underfitting.

En este caso, el método seleccionado para encontrar el valor óptimo de los parámetros es la optimización bayesiana. El enfoque de búsqueda bayesiana se basa en construir un modelo probabilístico de los hiperparámetros y su efecto en la métrica de evaluación del modelo, como la precisión o el error [56]. La ventaja de la búsqueda bayesiana es que permite trabajar con rangos de valores más grandes. Además, el algoritmo implementa validación cruzada que permite diseñar un modelo más robusto y mejorar la capacidad de generalización del algoritmo.

Los valores de los hiperparámetros del modelo óptimo son:

max_depth: 11

min_samples_leaf: 11

min_samples_split: 20

n_estimators: 46

4.1.2 GRUPO 1

El grupo 1 encuentra más similitudes con el clúster 0 del primer modelo. Sin embargo, como se ha mencionado previamente, las similitudes vistas en este clúster son más débiles que las encontradas en otros clústeres. Por lo tanto, las diferencias encontradas entre ambos grupos se explicarán en detalle.

En primer lugar, el número de relaciones de este clúster es ligeramente inferior a su contraparte del primer modelo. Además, ocurre una situación parecida con la variable participantes, donde el grupo obtenido en este modelo cuenta con menos participantes en sus rondas de financiación. Finalmente, con respecto a las variables numéricas, las utlimas diferencias se encuentran en las variables que analizan el tamaño de las series a y series b de las startups, donde el tamaño medio de las rondas es ligeramente inferior en el caso de las startups que forman este clúster.

En cuanto a las variables categóricas, las principales diferencias se encuentran en el atributo de región. Con respecto a la región, el grupo 1 cuenta con startups divididas entre los principales centros de startups de los Estados Unidos, destacando una frecuencia muy superior de San Francisco a la de otras ciudades. En el caso del grupo 0 del primer modelo de clustering, las oficinas se encuentran todas en la región de San Francisco.

En conclusión, a pesar de que es cierto que en la mayoría de las variables los resultados son similares y las diferencias son muy poco significativas es importante analizarlos pues es posible que den a clasificaciones distintas.

El método AdaBoost fue el seleccionado para trazar los patrones entre las muestras del grupo 1. El grupo 1 es el clúster más poblado con 154 muestras y los parámetros modificados para ajustar el modelo fueron los siguientes:

1. 'learning_rate': El learning rate determina el impacto de cada árbol en el modelo final. Un valor más bajo implica una contribución más pequeña de cada árbol, lo que hace que el modelo sea más conservador y reduzca el riesgo de sobreajuste. Sin embargo, a cambio de un modelo más robusto, el coste computacional aumenta.
2. 'n_estimators': Explicado en el apartado 3.4.1.

3. 'base_estimator': Este parámetro se utiliza para definir el estimador base que utilizará el modelo.
4. 'loss': El parámetro 'loss' especifica la función de pérdida utilizada para optimizar el modelo durante el entrenamiento.

Los valores de los hiperparámetros del modelo óptimo son:

base_estimator: DecisionTreeRegressor(max_depth=3)

learning_rate: 0.5

loss: 'exponential'

n_estimators: 100

En este caso, el algoritmo utilizado para realizar el ajuste de los hiperparámetros es GridSearchCV. GrideSearchCV, a diferencia de BayesSearchCV, comprueba todas las posibles combinaciones de parámetros especificados mediante la cuadrícula (grid) de parámetros. Esta técnica es muy compleja computacionalmente cuando hay una gran cantidad de combinaciones, por lo que se ha de tener cuidado al implementar esta técnica. En este caso, al contar con un espacio de parámetros limitado, resulta interesante implementar un método como el de GridSearchCV para poder comprobar todas las potenciales combinaciones.

4.1.3 GRUPO 2

En el caso del grupo 2, que cuenta con 117 startups, el método con mejor rendimiento fue el de Gradient Boosting. Como se observa más adelante, algunos hiperparámetros de Gradient Boosting son los mismos a los que utilizan los random forest. Esto se debe a que ambos utilizan árboles de decisión como componentes principales. Los hiperparámetros modificados fueron los siguientes [57]:

5. 'learning_rate': Explicado en el apartado 3.4.2.
6. 'n_estimators': Explicado en el apartado 3.4.1.
7. 'subsample': El subsample especifica la proporción de muestras a utilizar en cada árbol de decisión. Un valor menor que 1.0 hace que cada árbol se entrene con una fracción de las muestras totales, lo que puede ayudar a reducir la varianza y el sobreajuste.
8. 'min_samples_split': Explicado en el apartado 3.4.1.
9. 'min_samples_leaf': Explicado en el apartado 3.4.1.
10. 'max_depth': Explicado en el apartado 3.4.1.
11. 'max_features': El número máximo de características a considerar al buscar la mejor división en cada nodo. 'sqrt' significa que se considerarán la raíz cuadrada de todas las características, mientras que 'log2' significa que se considerarán el logaritmo en base 2 de todas las características.
12. 'max_leaf_nodes': Este parámetro establece el número máximo de nodos permitidos en el árbol. Limitar este número puede ayudar a controlar la complejidad del modelo y reducir el riesgo de sobreajuste.
13. 'validation_fraction': Esta fracción de los datos de entrenamiento se reservará como conjunto de validación durante el entrenamiento del modelo. Es utilizado para evaluar el rendimiento del modelo en datos no vistos y ajustar los hiperparámetros en función de esta evaluación.

Para este grupo, el algoritmo de ajuste utilizado ha sido GridSearchCV. Los valores de los hyperparámetros del modelo óptimo son:

max_features: 'log2'

learning_rate: 0.1

max_leaf_nodes: 15

max_depth: 4

min_samples_leaf: 4

min_samples_split: 3

n_estimators: 100

validation_fraction: 0.2

subsample: 0.75

4.1.4 GRUPO 3

El grupo 3 de este modelo es más parecido al clúster 2 del modelo de clustering inicial.

La primera diferencia que se observa en los clústeres es que, en las variables temporales, que incluyen 'first_funding_at', ' last_funding_at' y 'year', pues en el grupo 3 todas las operaciones ocurren más tarde. Por ejemplo, en el caso del año de la última inversión, la media de las startups de este clúster reciben esa financiación en el año 2011 y en el grupo 2 del modelo de clustering inicial esa última ronda ocurre en 2009 de media.

Con respecto a las variables categóricas, existen diferencias significativas en las variables 'status', 'region' y 'Best University Ranking'. En el caso del método de exit, a diferencia del clúster obtenido por el primer modelo donde la diferencia entre las adquisiciones y las IPOs es muy significativa, en el grupo 3, esta diferencia se ve reducida con menos de 40 adquisiciones y 9 IPOs. En cuanto a la región donde se encuentra la oficina, mientras las startups de este grupo están todas situadas en el área de Boston, las startups del clúster 2 del modelo inicial se reparten en todas las ciudades americanas importantes del ecosistema, exceptuando San Francisco. Finalmente, mientras el cluster 2 del primer modelo cuante con cero graduados entre los fundadores de las startups, las startups de

este grupo tienen mayor cantidad de fundadores entra las 100 mejores universidades que fundadores sin título universitario.

En el caso de este grupo, que contiene un total de 48 casos, el mejor modelo es el SVR. Por lo tanto, se le aplicará el ajuste de los hiperparámetros. En este caso, los siguientes hiperparámetros han sido estudiados [58]:

1. 'C': Es un parámetro utilizado para la regularización del modelo. Define el parámetro de penalización para los errores de clasificación. La potencia de la regularización es inversamente proporcional al valor de C.
2. 'epsilon': Define la distancia permitida entre la variable objetivo real y la predicción del modelo antes de que se considere un error. Cuanto más bajo sea épsilon, más sensible será el modelo a errores, ajustando más las predicciones. Esta variable se utiliza en el modelo épsilon-SVR.
3. 'kernel': Este hiperparámetro determina el tipo de kernel utilizado en el SVM. Un kernel es una función que mapea los datos de entrada a un espacio de mayor dimensión [59].
4. 'gamma': El coeficiente gamma se utiliza como el coeficiente del kernel seleccionado. Gamma permite encontrar los subespacios que diferencien con mayor facilidad los puntos del espacio. Un valor bajo de gamma da lugar a predicciones más conservadoras, pues resulta en distancias mayores entre las observaciones que se encargan de separar los subespacios de SVM [60].
5. 'shrinking': Cuando está establecido en 'True', se aplica la reducción de margen, lo que puede acelerar el tiempo de entrenamiento. Cuando está establecido en 'False', no se utiliza esta técnica y puede llevar a un tiempo de entrenamiento más largo, pero puede producir un modelo más preciso.
6. 'tol': Este hiperparámetro define la tolerancia para el criterio de parada. Indica la precisión requerida para considerar que el modelo ha convergido. Los valores bajos de este parámetro dan lugar a mayores tiempos de entrenamiento.

El método seleccionado para realizar el ajuste de los parámetros es el de optimización bayesiana. El algoritmo de ajuste genera la siguiente combinación de parámetros como la óptima para los datos es la siguiente:

C: 61.74622791438141

epsilon: 0.7776107350396038

gamma: 'auto'

kernel: 'rbf'

shrinking: True

tol: 0.0005986738347359701

4.1.5 GRUPO 4

En el clúster 4 se utiliza un random forest como modelo base, pues es el que mejor traza los patrones y relaciones existentes en la base de datos. Tras implementar el ajuste de los hiperparámetros utilizando BayesSearchCV, los parámetros obtenidos son los siguientes:

max_depth: 11

min_samples_leaf: 11

min_samples_split: 20

n_estimators: 46

4.1.6 GRUPO 5

Finalmente, el grupo 5 hace uso de AdaBoost para conseguir los mejores resultados. Al aplicar el método de GridSearchCV sobre los mismos parámetros que en el clúster 1, los valores obtenidos son los siguientes:

base_estimator: DecisionTreeRegressor(max_depth=3)

learning_rate: 0.4

loss: 'exponential'

n_estimators: 100

Tras obtener los modelos óptimos para cada clúster, se combina el algoritmo de clustering junto a los modelos utilizando métodos de ensemble para obtener el modelo final completo.

4.2 RESULTADOS DEL MODELO FINAL

El modelo final supone una ligera mejoría en cuanto a las métricas utilizadas si se compara con el resto de los modelos vistos previamente. Además, a diferencia del modelo de random forest, las métricas de error utilizadas para analizar el modelo no varían tras cada ejecución. Esto genera un modelo mucho más confiable y útil de cara a predecir el éxito de la inversión en startups de manera precisa. El hecho de que los resultados obtenidos a partir del modelo final sean los más potentes muestran la ventaja que supone subdividir el conjunto de datos en clústeres, donde los patrones y relaciones entre los datos son mucho más claras y dan lugar a mejores predicciones. Además, a pesar de que la subdivisión en clústeres da lugar a conjuntos de datos más pequeños, donde extraer relaciones y patrones fiables es mucho más preciso, este modelo sigue siendo capaz de dar un rendimiento superior al resto de modelos.

```
RMSE: 1.2479884369984369
Average Error ($): 2.483328969610181
MAPE: 5.117113672888357
Average MAPE (%): 165.8530813626918
R^2: 0.4038676095873679
```

Figura 38. Resultados del Modelo Final.

El modelo final tiene un error medio de 2.48 dólares. Este resultado es realmente sorprendente pues la variable objetivo contiene valores entre 30,000 y 104,000,000,000. Por lo tanto, se puede asegurar que el modelo es capaz de predecir con un nivel de precisión aceptable el éxito de la inversión.

CAPÍTULO 5. CONCLUSIONES

Dados los objetivos establecidos al principio de esta tesis, se pueden extraer varias conclusiones relevantes.

En primer lugar, tras la elaboración y posterior análisis de los modelos y sus resultados, se puede concluir que el modelo final es efectivo y cumple con el principal objetivo de generar un modelo que permita a los fondos de capital riesgo tomar decisiones más informadas. El haber obtenido un modelo final que tan solo tiene un error promedio de 2.5 dólares muestra un alto grado de fiabilidad y desde luego sirve como una herramienta útil para analizar oportunidades de inversión para los fondos. Además, la propia naturaleza de las inversiones en startups tiene un alto componente de varianza, donde casos extremos y grandes éxitos son comunes. Por lo tanto, se puede asumir una métrica MAPE ligeramente más alta pues estos casos son relativamente escasos a lo largo del tiempo.

El segundo objetivo marcado era el de elaborar un algoritmo útil para los emprendedores, el cual les permitiese tomar decisiones más informadas, incrementando las posibilidades de éxito. Este modelo, al predecir con tanta precisión el valor futuro de la startup permitirá a los founders conocer más detalles que facilitan la toma de decisiones. Además, todo el análisis exploratorio previo y los resultados obtenidos de otros modelos también aportan información realmente útil al emprendedor, que puede usarla como una fuente de información adicional.

Finalmente, todas los objetivos relacionados con identificar la importancia relativa han sido cumplidos utilizando el análisis de los clústeres. Si se analizan los valores medios y los diagramas de caja, por clúster, para la variable objetivo, se puede conocer que grupo recoge a las empresas más exitosas. Si se combina esa información con los gráficos que

representan la información por clúster sobre el método de exit, el año del exit y la formación universitaria se pueden extraer conclusiones sobre la importancia de las variables.

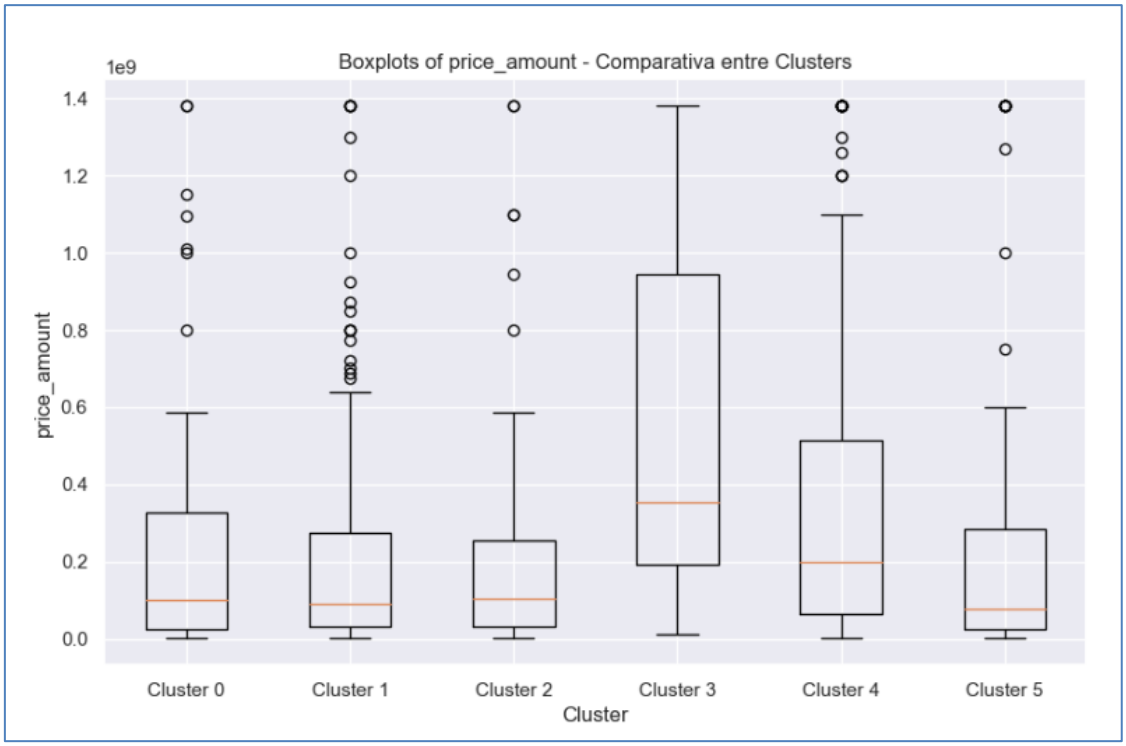


Figura 39. Rango de Precios por Clúster (tras eliminar valores extremos).

En la figura se observa como el clúster 3 es el más exitoso, pues es el que presenta valores de exit más altos. Además, cabe destacar que se han eliminado valores dentro por encima del percentil 95 de euros de valoración, pues permite ver de manera más clara la distribución de valores para todos los clústeres. Esta eliminación da lugar a la pérdida de 5 valores del clúster 0, algunos llegando a ser superiores a los 100B$. El clúster 4 es el siguiente clúster con valoraciones más altas y los grupos 5 y 1 tienen las valoraciones más bajas. Sin embargo, en la figura 38 se observa como el clúster 5 tiene un valor medio muy superior al del clúster 1, superando incluso al clúster 2. Esto se debe a que hay compañías

en el clúster 5 que han recibido valoraciones muy altas y están aumentando el valor medio de la transacción.

```
Cluster 3: Median = 352000000.0, Mean = 2526095394.2720337
Cluster 4: Median = 200000000.0, Mean = 1112472978.7234042
Cluster 2: Median = 102500000.0, Mean = 214875666.4057971
Cluster 0: Median = 100000000.0, Mean = 485432790.6976744
Cluster 1: Median = 89000000.0, Mean = 273467411.7816092
Cluster 5: Median = 78400000.0, Mean = 408730647.9166667
```

Figura 40. Valores Medios y Medianos del Precio de Exit por Clúster (Ordenados descendientemente en función del valor mediano).

En primer lugar, con respecto al método de exit, si se analiza la Figura 19, el único clúster donde todas las operaciones son IPOs es el clúster 3, que como se ha mencionado previamente presenta los valores más altos en la variable objetivo. Además, es el clúster 4, que es el segundo con las valoraciones más altas, el siguiente grupo con mejor ratio de IPOs/Adquisiciones. Por otro lado, se observa que el clúster 1, que es el que tiene os valores medios de venta/IPO más bajos, es el único grupo donde todas las operaciones son adquisiciones. También se ha de mencionar como las pocas operaciones de IPO en el grupo 5 son las que hacen que el valor medio sea muy superior al mediano. Estos hechos demuestran que existe una clara relación entre el tipo de salida y el valor de la compañía, donde los inversores que participan en compañías cuyo exit sea una IPO tienden a tener participaciones en una empresa con mayor valoración.

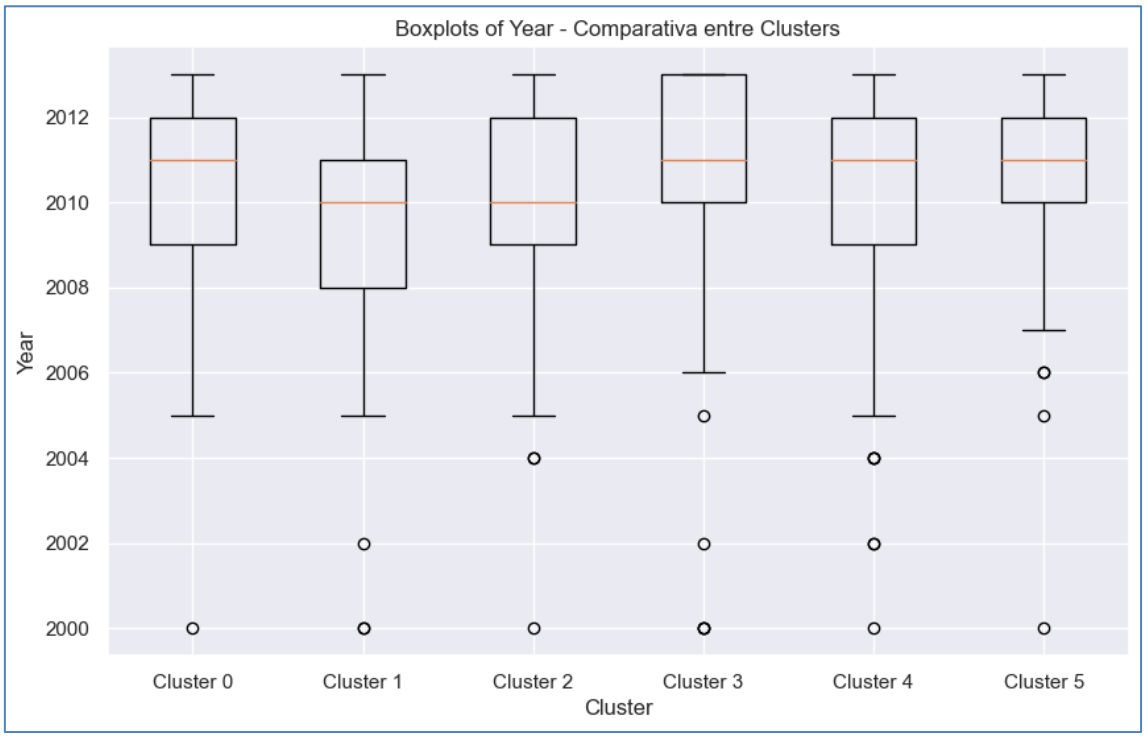


Figura 41. Distribuciones del Año de la operación de Exit por clúster.

En segundo lugar, si se analizan las Figuras 12, 13 y 40 se observa cierta relación entre el tiempo en el que han ido avanzando las startups, primero a través de sus rondas de financiación y, posteriormente mediante la operación de venta o salida a bolsa, y el éxito de la inversión. La crisis financiera que ocurre entre el 2007 y el 2008 afecta significativamente a la economía mundial y con ello al sector del venture capital. Si se analizan las fechas de las operaciones de las startups se observan ciertos patrones que muestran el efecto del mercado y la situación económica sobre el ecosistema y el éxito de las startups. Por ejemplo, los dos clústeres con peores cifras de venta/IPO, el 1 y el 5, son aquellos que levantaron sus rondas de financiación entre 2006-2008. Esto se puede deber a que durante la crisis las valoraciones sufrieron y dio lugar a rondas de financiación más ajustadas y menos capital para operar. Por el contrario, las compañías que más tarde han cerrado sus procesos de levantamiento de capital han sido los que mejores valoraciones han tenido. Este patrón se ve acentuado al analizar los valores para el clúster con los exits

más altos. Las empresas del clúster 3, de media, realizan su última ronda de financiación en 2011 donde la economía global estaba experimentando una fuerte recuperación tras la crisis. Por lo tanto, estas tendencias demuestran una clara relación entre el ciclo económico y el éxito que pueda tener una startup y sus inversores en su proceso de exit. Además, muestran el impacto significativo que tuvo la crisis financiera en el ecosistema startup.

Finalmente, en la figura 29 se puede observar una relación significativa entre el éxito de una compañía y la universidad donde el fundador/CEO ha cursado sus estudios. De las 20 variables más relevantes, 17 de ellas están relacionadas con la institución educativa de los fundadores/CEOs.

Por otro lado, en la figura 21 se muestra una relación menos evidente entre la universidad y el éxito de la compañía, aunque no tan fuerte como la relación con el tipo de exit. El clúster 5, que tiene menos fundadores/CEOs provenientes de universidades prestigiosas, se encuentran entre las categorías con las peores valoraciones. Por otro lado, las categorías 4 y 3, presentan las valoraciones más altas y se destacan por tener un alto porcentaje de fundadores/CEOs provenientes de las mejores instituciones (Top 100 del mundo).

Sin embargo, es importante mencionar que el clúster 1, el cual tiene la segunda valoración más baja en términos de mediana, contiene una gran cantidad de fundadores en las mejores universidades del mundo. Además, el clúster 2, que se sitúa en la tercera posición en cuanto al valor mediano de la variable objetivo, no tiene ningún alumno que haya asistido a la universidad. Lo mismo ocurre con el clúster 0, que se encuentra en la cuarta posición. Esto sugiere que a pesar del impacto claro que tiene la formación académica de los puestos altos de una startup sobre el éxito de esta, no es una condición imprescindible para que la compañía sea exitosa.

REFERENCIAS

[1] Chae, T. (Mayo 2019). How Do VCs Evaluate Early-Stage Startups Versus Later Stage Ones? https://www.forbes.com/sites/quora/2019/05/01/how-do-vcs-evaluate-early-stage-startups-versus-later-stage-ones/?sh=62a9b1793b3e

[2] Foy, P. (Abril 2020). Applications of AI and Machine Learning in Venture Capital. https://www.mlq.ai/ai-machine-learning-venture-capital/

[3] Corea, F. (Mayo 2019). Data-driven VCs. https://francesco-ai.medium.com/data-driven-vcs-839f2454d22

[4] Nunes, J. (Septiembre 2022). Venture Capital 2.0 — the revolution of Machine Learning & Data-Driven VC. https://medium.com/included-vc/venture-capital-2-0-the-revolution-of-machine-learning-data-driven-vc-5ecd62b76fb

[5] Chen, X. J. (Junio 2021). How AI Is Transforming Venture Capital. https://www.brinknews.com/how-ai-is-transforming-venture-capital/

[6] Shahat, M. (Marzo 2023). Revolutionizing Startup Funding: How AI is Changing the Investor Selection Process. https://www.linkedin.com/pulse/revolutionizing-startup-funding-how-ai-changing-investor-shahat/

[7] Wu, V. (Junio 2017). A machine-learning approach to venture capital. McKinsey Quarterly. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/a-machine-learning-approach-to-venture-capital

[8] Corea, F. (Febrero 2019). Using Machine Learning in Venture Capital. https://www.forbes.com/sites/cognitiveworld/2019/09/12/using-machine-learning-in-venturecapital/?sh=205217be239b

[9] Stone, T. R. (Octubre 2014). Computational Analytics for *Venture Capital*.

[10] Arroyo, J., Corea, F., Jiménez-Diaz, G., & Recio-García, J. A. (Agosto 2019). Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments.

[11] Cirtautas, J. (2019). Startup Investments. https://www.kaggle.com/datasets/justinas/startup-investments

[12] (Febrero 2018). ¿Qué es un 'business angel'? https://www.bbva.com/es/que-es-un-business-angel/

[13] " Qué es la valoración pre-money y la valoración post-money." (Noviembre 2021). https://www.plantillaspyme.com/blog-pymes/finanzas-y-contabilidad/que-es-la-valoracion-pre-money-y-la-valoracion-post-money

[14] Rouco, M. (Diciembre 2012). Tel Aviv, mejor ecosistema emprendedor del mundo tras Silicon Valley. https://www.elmundo.es/elmundo/2012/12/07/economia/1354896497.html

[15] (October 2013). These are the top universities graduating the most VC-backed entrepreneurs. https://www.geekwire.com/2013/top-universities-producing-vcbacked-entrepreneurs/

[16] Britt, R. (November 2013). Higher Education R&D Expenditures Remain Flat in FY 2012. https://www.nsf.gov/statistics/infbrief/nsf14303/

[17] (Marzo 2022). ¿Qué es una IPO y cómo funciona? https://aptki.com/publicaciones/que-es-una-ipo-y-como-funciona/

[18] (Marzo 2023). Dotcom Bubble. https://corporatefinanceinstitute.com/resources/capital-markets/dotcom-bubble/

[19] Risen, T. (Diciembre 2013). 2013: Best Year for Stock IPOs Since 2000. https://www.usnews.com/news/articles/2013/12/30/2013-best-year-for-stock-ipos-since-2000

[20] Startup Genome. Global Startup Ecosystem Ranking 2022 (Top 30 + Runners-Up). https://startupgenome.com/article/global-startup-ecosystem-ranking-2022-top-30-plus-runners-up

[21] Farr, C. (Julio 2013). Why London has all the ingredients for a successful startup ecosystem. https://venturebeat.com/business/london-has-all-the-ingredients-for-a-successful-startup-ecosystem/

[22] Shontell, A. (Abril 2012). Why More Startups Succeed In Silicon Valley: 22 Fascinating Research Findings. https://www.businessinsider.com/silicon-valley-vs-new-york-startup-genome-findings-2012-4

[23] Higginbotham, D. (June 2022). What is a PhD? https://www.prospects.ac.uk/postgraduate-study/phd-study/what-is-a-phd

[24] Stapleton, C. (Enero 2022). What Are the Listing Requirements for the NASDAQ? https://www.investopedia.com/ask/answers/nasdaq-listing-requirements/

[25] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., … Lin, C.-T. (2017). A review of clustering techniques and developments. Neurocomputing, 267, 664–681. doi:10.1016/j.neucom.2017.06.053

[26] Harichan. (Junio 2022). https://saxon.ai/blogs/identify-patterns-and-trends-from-large-data-sets-with-clúster-analysis/

[27] Yadav, D. (Diciembre 2019). Categorical encoding using Label-Encoding and One-Hot-Encoder. https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd

[28] De Dieu Nyandwi, J. (Agosto 12, 2021). The Ultimate and Practical Guide on Feature Scaling. https://jeande.medium.com/the-ultimate-and-practical-guide-on-feature-scaling-d03fbe2cb25e

[29] Yildirim, S. (Marzo 2020). K-Means Clustering – Explained. https://towardsdatascience.com/k-means-clustering-explained-4528df86a120

[30] Mullin, T. (Julio 10, 2020). DBSCAN Parameter Estimation Using Python. https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd

[31] Das, A. (Enero 9, 2022). Deciding number of Clusters using Gap Statistics, Davies-Bouldin Index, & Calinski-Harabasz Index for K-Means and Hierarchical Clustering using Python. https://medium.com/mlearning-ai/deciding-number-of-clusters-using-gap-statistics-davies-bouldin-index-calinski-harabasz-index-2ce9acfb6118

[32] Mahmood, S. (Mayo 2022). Outlier Detection (Part 1). https://towardsdatascience.com/outlier-detection-part1-821d714524c#:~:text=Usually%20z%2Dscore%20%3D3%20is,similar%20to%20standard%20deviation%20method.

[33] Benoit, K. (Marzo 2011). Linear Regression Models with Logarithmic Transformations. https://kenbenoit.net/assets/courses/ME104/logmodels2.pdf

[34] Glen, S. RMSE: Root Mean Square Error. From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

[35] Mali, K. (Octubre 2021). Everything you need to Know about Linear Regression. https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/

[36] Likos, P. (Noviembre 20, 2022). IPO vs Acquisition: Advantages and Disadvantages. https://www.sofi.com/learn/content/ipo-vs-acquisition/

[37] Wu, Songhao. (Mayo 19, 2020). Multicollinearity in Regression. https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea

[38] Bhandari, A. (Marzo 20, 2020). Multicollinearity | Causes, Effects and Detection Using VIF. https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

[39] Glen, S. Mean Absolute Percentage Error (MAPE). From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/mean-absolute-percentage-error-mape/

[40] Krueger, E. (Junio 2021). Build Better Decision Trees with Pruning. https://towardsdatascience.com/build-better-decision-trees-with-pruning-8f467e73b107

[41] Wang, Y. (Febrero 2023). What Are The Advantages And Disadvantages Of Random Forest? https://www.rebellionresearch.com/what-are-the-advantages-and-disadvantages-of-random-forest

[42] Thron, J. (Junio 2020). What is Boosting in Machine Learning? https://towardsdatascience.com/what-is-boosting-in-machine-learning-2244aa196682

[43] Kurama, V. (2020). A Guide to AdaBoost: Boosting To Save The Day. https://blog.paperspace.com/adaboost-optimizer/

[44] Saini, A. (Septiembre 2021). Gradient Boosting Algorithm: A Complete Guide for Beginners. https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

[45] Morde, V. (Abril 2019). XGBoost Algorithm: Long May She Reign! https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

[46] Jain, A. (Febrero 2016). Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python.

[47] Sethi, A. (Marzo 2020). Support Vector Regression Tutorial for Machine Learning. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

[48] Hardesty, L. (April 14, 2017). Explained: Neural networks. MIT News Office. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

[49] Picton, P. (1994). What is a Neural Network? In *Introduction to Neural Networks*, Palgrave, *London*,(pp. 1–12). Doi:10.1007/978-1-349-13530-1_1

[50] Krishnan S. (Septiembre 2021). How do determine the number of layers and neurons in the hidden layer? https://medium.com/geekculture/introduction-to-neural-network-2f8b8221fbd3

[51] Brownlee, J. (Enero 2019). A Gentle Introduction to the Rectified Linear Unit (ReLU). https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/

[52] Wang, C. (Enero 2019). The Vanishing Gradient Problem. https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484

[53] Srivastava, T. (Junio 2015). Tuning the parameters of your Random Forest model. https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/

[54] Ram, S. (Octubre 2020). Mastering Random Forests: A Comprehensive Guide. https://towardsdatascience.com/mastering-random-forests-a-comprehensive-guide-51307c129cb1

[55] Fraj, M. B. (Diciembre 2017). In Depth: Parameter Tuning for Random Forest. https://medium.com/all-things-ai/in-depth-parameter-tuning-for-random-forest-d67bb7e920d

[56] Michael, R. (Septiembre 2020). Optimizing Hyperparameters the right way. https://towardsdatascience.com/optimizing-hyperparameters-the-right-way-3c9cafc279cc

[57] Jain, A. (Febrero 2016). Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python. https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/

[58] Bhattacharyya, I. (Junio 2018). Support Vector Regression Or SVR. https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff

[59] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011).

Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[60] Vaquerizo, R. (Octubre 2016). El parámetro gamma, el coste, la complejidad de un SVM. https://analisisydecision.es/el-parametro-gamma-el-coste-la-complejidad-de-un-svm/

ANEXO I: ALINEACIÓN DEL PROYECTO CON LOS ODS

Los ODS tienen como propósito erradicar la pobreza, proteger el planeta y asegurar la prosperidad para todos. Existen distintos objetivos para asegurar el cumplimiento de estos tres pilares.

Por lo que respecta a este trabajo de fin de grado, podría alinearse con la mayoría de los ODS dado que una gran cantidad de *startups* se han creado bajo una filosofía de desarrollo sostenible y este trabajo tiene como objetivo desarrollar una herramienta que impacte positivamente en el ecosistema *startup*.

No obstante, si hubiese que destacar un número reducido de ODS con los que el TFG se alinea destacaría los ODS Nº8 (Trabajo Decente y Crecimiento Económico) y Nº9 (Industria, Innovación e Infraestructura). Por un lado, las *startups* emplean a un porcentaje significativo de la población. Además, son uno de los motores de la economía, generando un valor casi equivalente a una economía del G7. Por lo tanto, este trabajo que pretende mejorar la toma de decisiones en el mercado *startup* podría impactar positivamente en la consecución del ODS Nº8.

Por otro lado, el objetivo de desarrollo sostenible Nº9 hace referencia a la innovación que está directamente relacionada con el trabajo pues se está presentando un modelo que tiene el potencial de ayudar a mejorar las inversiones de las empresas de capital de riesgo. Dichas inversiones son las que financian a las *startups* que a su vez son las que llevan a cabo gran parte de la innovación en la actualidad.