Mondragon Unibertsitatea

Facultad de Empresariales

Grado en

Business Data Analytics

Informe final

Reto: 07 - Grupo Spri

Curso: 2º 2022-2023

Equipo: Morado

URL del repositorio en GitHub:

Ruta de la carpeta de Google Drive que contiene el proyecto:

https://drive.google.com/drive/u/1/folders/17sqrU7sZF3QV-4g8omCTIz0wX61HG7Ah

Autores:

- Paula Arnaiz Cuenca
- Izaskun Benegas Moreno
- Iker Benito de la Hera
- Aitor Fernandez de Retana Zuñiga
- Libe Xiang Galdos Alberdi
- Unai Martinez Leal

22-23_2_R7_Morado

Índice

Índice de figuras

Índice de tablas

1. Introducción

A lo largo de este informe se dará respuesta al reto planteado por el grupo Spri, que consiste en recibir información sobre la **valoración de una *startup* futura y también las posibilidades de ser adquirida** (total o parcialmente).

El trabajo a lo largo del reto se ha realizado con diferentes **herramientas** para poder **tratar** los **datos**, entre las cuales destacan el lenguaje de programación Python, con el objetivo de cumplir con los entregables y aportar una clara visualización del resultado.

En primer lugar se ha realizado una pequeña **introducción** del proyecto, seguido del desarrollo de la problemática, una descripción del cliente y los principales objetivos que se deben completar.

A continuación, se desarrolla todo el **preprocesamiento** de los **datos**. En este se mencionan las fuentes de datos utilizadas y seguidamente se realiza una extensa explicación sobre el *data discovering* y *data cleaning*.

Por último se presenta la **bibliografía** y varios **apéndices**. El primero de ellos está relacionado con la interpretación **financiera** y el segundo está conectado con la interpretación **matemática**.

2. 2. Identificar la problemática

a. 2.1 Sobre el cliente

El grupo "*spri*" es la entidad del Departamento de Desarrollo Económico, Sostenibilidad y Medio Ambiente del **Gobierno Vasco**. Esta entidad trabaja junto a empresas para **facilitar el acceso a la digitalización, ciberseguridad e iniciarlas**. Buscan una forma de expandir el negocio de sus clientes a otros países o a buscar espacios físicos, pabellones u oficinas, en los que instalarles.

La empresa indica en su página web que en el plan de gestión de 2021 tiene como marco general el Programa para la **Reactivación Económica y el Empleo de Euskadi** (2020-2024), BERPIZTU.

El programa se estructura en dos ejes verticales: un primer eje de **reactivación Económica** y un segundo eje de **dinamización de empleo**, y sus ejes, a su vez, se despliegan en 12 políticas de actuación que constituyen el marco de referencia de diferentes medidas e instrumentos en materia de recuperación y estimulación de la economía y el empleo a desarrollar hasta el 2024.

Su misión es apoyar, impulsar y contribuir a la **mejora competitiva de las empresas vascas**, colaborando con ello a la generación de riqueza en Euskadi y a la mejora del

bienestar de su ciudadanía mediante un desarrollo humano sostenible, en el ámbito de la política de Promoción Económica del Gobierno Vasco.

En cuanto a la **visión** tiene como objetivo ser el **referente** en las actividades que contribuyen a la **promoción económica** y mejora de la competitividad de las empresas vascas.

Sus **valores** son.

1. Cercanía: Ser accesible y aplicar un **trato personalizado** a cada empresa demandante a sus servicios.
2. Integridad: Aplicar el **código ético** de SPRI a todas sus actividades.
3. Transparencia: La obligación de dar cuenta de manera fiel de todas sus actuaciones tanto a nivel externo como interno.
4. Innovación: La **mejora continua** para adaptarse a las nuevas necesidades.
5. Compromiso social: El compromiso de realizar sus actividades teniendo presente **el mayor beneficio social** alcanzable, con el mejor uso de los recursos.

i.

b. 2.2 Sobre la problemática

c. 2.2 Objetivos

Calcular la valoración de una empresa en función de sus datos financieros y características cualitativas.

Predecir mediante **modelos de clasificación** si una empresa será: adquirida totalmente (100%), adquirida parcialmente o no adquirida (seguirá su curso sin accionistas).

3. 3. Recoger y almacenar los datos

a. 3.1 Fuentes de datos utilizadas

Los datos empleados para proporcionar información sobre el **desarrollo y la evolución que tendrán las *startups* de Euskadi en un futuro próximo**, han sido proporcionados por Grupo Spri. Concretamente, buscan conocer la valoración futura de una empresa, y también las posibilidades que tendrá de ser adquirida (total o parcialmente).

DESCRIPCIÓN DE VARIABLES

Variable	Detalles
ASSET	Activo
ASSET_TYPE	Tipo de activo
BRAND	Marca del activo
SEGMENT	Segmento de negocio
CHECKIN_TIME	Fecha y hora del checkin
CHECKOUT_TIME	Fecha y hora del checkout
BOOKING_TIME	Momento en el que se hizo la reserva
ADULT_COUNT	Cantidad de adultos en la reserva
CHILD_COUNT	Cantidad de niños en la reserva
Codigo_NIF	Número de identificación fiscal
name_dealroom	Nombre de la empresa, en Dealroom
profile_url	Enlace al perfil de la empresa dentro de Dealroom
website	Página web de la empresa
tagline	Valores/cualidades más importantes (breve descripción de la empresa)

total_funding	Financiación total recibida, en millones de euros
first_funding_date	Fecha de la primera financiación recibida
last_funding_date	Fecha de la última financiación recibida
last_funding	Importe recibido en la última ronda, en millones de euros
last_round	Naturaleza de la última ronda de financiación (de qué tipo ha sido)
total_rounds	Nº total de rondas de financiación realizadas
n_empleados_dealroom	Número de empleados en 2021
ownerships	Lista con la naturaleza de las entidades propietarias de la empresas
b2b_b2c	Forma(s) de hacer relaciones y transacciones comerciales
revenue_models	Fuentes de ingresos (*saas*, *manufacturing*, *marketplace & ecommerce*)
growth_stage	Etapa de crecimiento en la que se encuentra la empresa
company_status	Estatus de la compañía (*acquired* / *low-activity* / *operational*)
valuation_2022	Valoración de la empresa en 2022, en millones de EUR

a. Tabla 01: Descripción de las variables

En cuanto a los distintos conjuntos de datos proporcionados por Grupo Spri, este ha proporcionado un total de **4 datasets** que contenían todo tipo de i**nformación sobre las empresas a analizar.**

Por una parte, se ha dispuesto de datos asociados a las **rondas de financiación** y a diversas **características cualitativas** de 412 *startups*, las cuales están incluidas en este dataset llamado “df_dealroom_modif.xlsx”.

Por otra parte, también se ha dispuesto de información sobre **balances** y cuentas de los resultados obtenidos a través de la **herramienta SABI** (Sistema de Análisis de Balances Ibéricos). Estos resultados se han separado en dos datasets distintos, “df_sabi_modif_1.xlsx” el cual contiene 11 columnas que corresponden a **información general** de las 412 empresas a su **estado actual.**

El otro dataset, “df_sabi_modif_2.xlsx” muestra datos de 412 empresas en formato largo. Este dataset contiene **2 filas por empresa (una por año: 2020 y 2021)**. En esta se visualizan 39 variables de las cuales 38 son variables **financieras** y la otra es el **número de empleados**.

Por último, a estos 3 datasets se le sumó un cuarto dataset, llamado “df_sabi_modif_3.xlsx” el cual mostraba más **información financiera sobre las empresas.**

b. 3.2 Procesamiento de los datos

En la *Figura 01* se pueden ver los **pasos seguidos en el preprocesamiento de los datos originales**, los cuales se seguirán en orden para la explicación del proceso llevado a cabo en la limpieza de los datos.

i. Figura 01: Pasos del preprocesamiento

1. Visualización datos

El primer paso llevado a cabo en el proyecto ha sido la lectura de los ficheros de datos de los que se disponía, que en este caso eran cuatro, de los cuales **tres eran datos de la base de datos sabi**, y el **cuarto fichero de dealroom**. Antes de cargar todos los csvs, se ha decidido seleccionar todas las variables de cada fichero, excepto del tercero de sabi, del que tan solo se han cogido **2 variables financieras que se consideraban relevantes** para completar el resto de los datos. Los ficheros de sabi_2 y sabi_3 tienen dos filas por cada empresa que había, ya que una de ellas corresponde al 2020 y la otra al 2021, mientras que el sabi_1 y dealroom solo tenían una fila por empresa.

A continuación, se han mirado las **distintas variables** que había en cada *dataframe*, así como la cantidad de **missings** (datos ausentes) en cada una de ellas. Además, se han graficado matrices de los missings que hay por filas, y se ha encontrado que hay **8 empresas que no disponen de datos para el 2020**, y que hay otras que tienen cerca de **15 datos ausentes en un mismo año**. También se han realizado boxplots para la visualización de algunas variables, y ver así la distribución de las mismas, así como los valores **outliers** (extremos).

2. Correcciones y conversión variables

También se han hecho algunas correcciones en los datos originales, como la eliminación de las **tildes**, la sustitución de los **"n.s."** por valores Nan (ausentes) para que se puedan tratar como tales. Adicionalmente, se han creado **variables dummies** (binarias) a partir de algunas variables, así como la forma jurídica y si las empresas son **b2b o b2c**, es decir si sus **clientes son otras empresas o personas**. Otra corrección realizada ha sido pasar las **fechas a su respectivo formato** para hacer cálculos con ellas.

3. Unión de los dataframes

Después de hacer estas primeras transformaciones, **se han juntado los 4 dataframes** con los que se ha empezado a trabajar. Primero se han juntado los dataframes de sabi, y posteriormente el de dealroom, los de sabi **por medio del código NIF y el año**, y con el de dealroom solamente por el NIF. Como resultado, se ha obtenido un dataframe de 800 filas, con algo más de 100 variables.

4. Creación de variables

Respecto a la creación de variables, en primera instancia se ha centrado en calcular **información** un poco **general** de las empresas, como los **años que han transcurrido desde su creación o desde su última financiación**, o si la empresa se considera una **startup** (menos de 7 años en el mercado) o no.

5. Reemplazo de valores ausentes

Inicialmente se han **reemplazado** tan solo algunas variables, como por ejemplo el **número de empleados de sabi con la columna de empleados de dealroom**, ya que es más fiable esta primera fuente, pero en caso de no tener datos, se cogen los de dealroom. Después, se han aplicado **fórmulas financieras para rellenar los valores ausentes** de las variables de financiación, como por ejemplo el activo, patrimonio neto o pasivo. Se ha dado prioridad a este método ya que **desde el punto de vista financiero es lo que más se puede aproximar a la realidad**, y lo que más sentido tiene.

6. Eliminación de variables

Tras hacer la primera imputación, se han borrado aquellas **variables que no se consideraban importantes**, entre ellas las **descriptivas** sobre las empresas, como por ejemplo la página web, el nombre completo o la forma jurídica. Además, también se han eliminado aquellas variables financieras que tenían **muchos missings o que no fueran relevantes para el modelado**, así como el inmovilizado, la rotación de las existencias y el free capital. Adicionalmente, se han pasado los activos fijos y líquidos que tenían valores negativos a cero, puesto que esta **variable no puede ser menor a cero.**

7. Creación *dataframe* con missings

Como hay algunos **modelos que funcionan con missings**, se crea un **dataframe con los missings** para utilizarlo posteriormente en el modelado, y **ver si se mejora el score.** Esto podría ser útil porque no haría falta preocuparse en el futuro de los missings que pudiera tener alguna empresa, y el proceso de preprocesado sería más sencillo. Además, al no tener que imputar datos, no se estaría introduciendo ningún dato inventado, y **se estaría trabajando con los datos reales de los que se dispone**.

8. Más imputaciones de missings

Como se ha explicado en el paso 5, las imputaciones de los missings se ha empezado aplicando las **fórmulas financieras correspondientes**, y después se han probado 4 **métodos distintos** para imputar los missings restantes:

4. **Valor** año **anterior/posterior**: se ha imputado el valor del año anterior o posterior en caso de que el valor de un año estuviera vacío y el del otro año tuviera un valor. Por ejemplo, en el caso de la variable pasivo total una empresa no tiene valor para el 2020 pero sí para el 2021, por lo que se le asigna el valor del 2021.

5. **Código CNAE**: El segundo método consiste en agrupar las empresas por código CNAE y calcular la **mediana** de cada **variable** para cada código CNAE, y en caso de haber missings se imputa el valor de la mediana de esa variable para ese código CNAE. Por ejemplo, para el activo total una empresa del código CNAE 0111 no tiene valor para el 2020, por lo que se le asigna el valor de la mediana del activo total para el código CNAE 0111 en el 2020.

6. Más **fórmulas financieras**: Una vez realizadas estas imputaciones, se vuelven a aplicar las fórmulas financieras del paso 5 para calcular los valores de las variables que no se han imputado con los métodos anteriores.

7. **Regresión lineal**: Las pocas variables que todavía tenían missings se han imputado con una regresión lineal, para la cual se ha buscado una variable que tuviera una **correlación alta** con la variable que se quería imputar, y se ha usado como variable independiente para la regresión lineal. Por ejemplo, para la variable pasivo total se ha usado como variable independiente el activo total, ya que tienen una correlación alta.

9. Más correcciones

Tras hacer todas las imputaciones de los missings, se han realizado algunas **correcciones** en el df, tanto en las **variables originales** como en las **imputadas**. Entre otras correcciones, se ha asignado el **valor 1** a la variable de número de empleados en los casos en **los que eran negativo**s (ocasionado por la imputación por regresión) y se

han convertido los valores **infinitos positivos y negativos en valores muy grandes y muy pequeños,** respectivamente.

10. Creación de df finales

Antes de crear los df finales, se han **creado** algunas **variables nuevas**, como el **beneficio neto** restando los gastos de personal a los ingresos de explotación, el retorno sobre los activos (**ROA**) y el **margen de EBITDA**. Ahora sí, se han creado los df finales, que en este caso han salido 5, 1 con y 1 sin missings para cada uno de los 2 modelos, y un quinto para hacer gráficos. A pesar de esta distinción, el proceso de su creación es el mismo:

Se ha pasado la **columna de año a columnas**, es decir, se han creado 2 columnas por cada variable, una para el año 2020 y otra para el 2021, **quedando solamente una fila para cada empresa**, a diferencia de los df anteriores, en los que había dos por cada una.

Se ha realizado la **división de cada variable en el año 2021 entre la del año 2020**, para obtener el crecimiento de cada variable, y con eso se han **creado variables nuevas**, como el crecimiento de los ingresos de explotación, el crecimiento del beneficio neto, etc.

Se han **eliminado las variables del año 2020**, ya que no aportan información nueva.

Para el df de valoración, se han **quitado** las **empresas que no tienen valoración**, ya que no se pueden utilizar para el modelo.

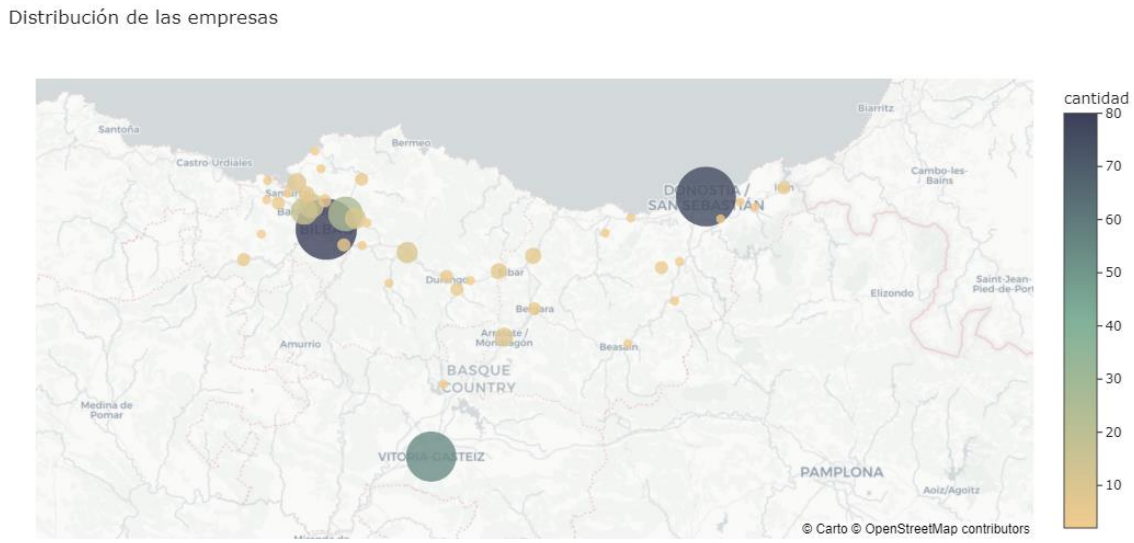
Por último, se han **creado variables nuevas** solo para el **df** de **valoración**, como el precio entre beneficio (**PER**) y el precio entre ventas (**P/Sales**).

a.

9. 4. Analizar y modelar los datos

a. 4.1 Análisis descriptivo

Para conocer mejor las **empresas con las que se trabaja** se han visualizado varios gráficos que aportan información relevante. Analizar esto, puede llegar a ser interesante para entender las **características básicas** que tienen las **startups** se encuentran en los dataset.

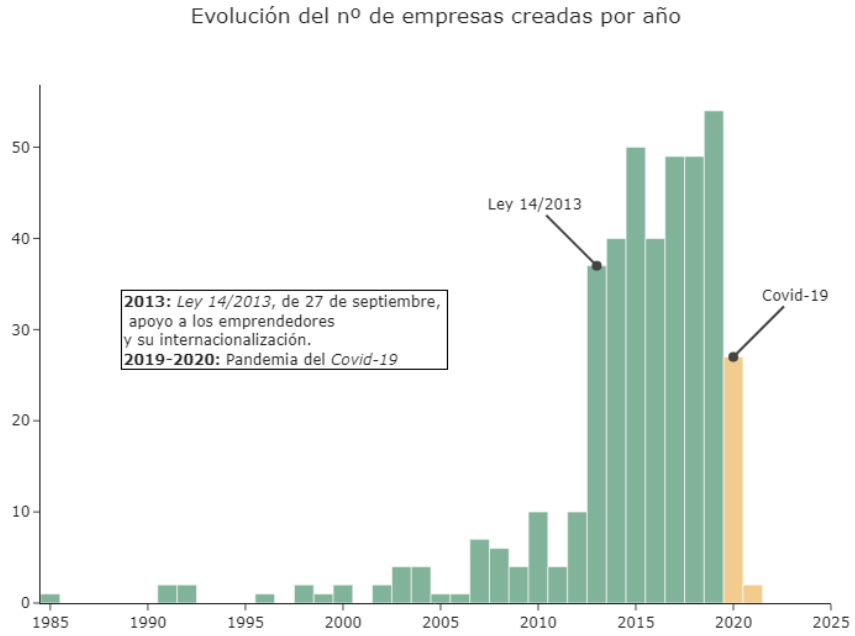


i. Figura 02: Mapa con la distribución de las empresas

Primero de todo, se ha decidido analizar la **ubicación** de las **startups** y, para ello, se ha realizado un gráfico de mapa visualizando el **número de empresas que hay en cada localidad**. Como era de esperar, son en las ciudades del País Vasco donde se concentran la gran mayoría de las empresas. Esto no ha resultado ser una gran sorpresa ya que, por normal general, son en las grandes ciudades donde se **encuentran más oportunidades para emprender**.

Asimismo, también es **mencionable** la gran cantidad de **empresas que se encuentran alrededor de Bilbao**. Con este dato también se puede intuir que de las tres ciudades, es Bilbao donde más empresas se encuentran. Así pues, se identifica una preferencia bastante clara en cuanto al lugar dónde se crea una empresa.

Por otro lado, con el fin de investigar en **qué año** hubo **más crecimiento** en las creaciones de las *startups* se ha realizado un gráfico **barras que muestra la evolución** que ha tenido el número de creación de empresas por cada año.

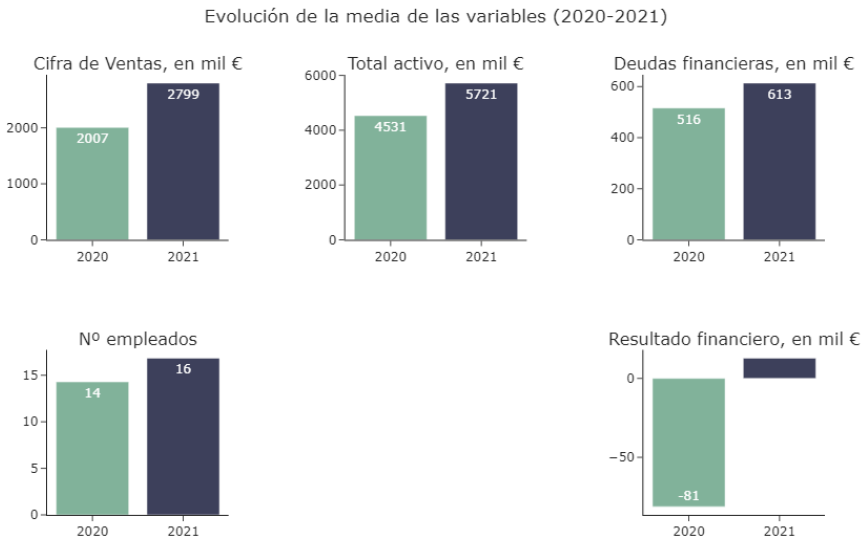


ii. Figura 03: Evolución del nº de empresas creadas por año

En la *Figura 03* son dos las cosas que destacan a primera vista. La primera de ellas es el gran **crecimiento** que sufrió la **cantidad** de **empresas creadas** en el año **2013**. En los años **previos al 2013** el número de empresas creadas apenas llegaba a las **10** empresas. No obstante, a **partir del 2013** se ve que este número crece hasta **40** empresas.

Lógicamente, los **factores** de este crecimiento repentino han podido haber sido varios, no obstante, habiendo investigado un poco se ha descubierto que el 27 de septiembre de 2013 se aprobó la **Ley 14/2013 en España**. Esta ley se aprobó con el fin de apoyar al emprendedor y la actividad empresarial, **favorecer su desarrollo**, **crecimiento e internacionalización** y fomentar la **cultura emprendedora** y un entorno favorable a la actividad económica. Por ello, entre otras muchas causas, podría encontrarse la aprobación de esta ley.

Por otro lado, puede llegar a sorprender también el **decrecimiento** tan grande que sufrió el número de creación de empresas en el 2020. En este caso, no cabe duda de la razón de este suceso que, efectivamente, sería la **aparición de la pandemia del *covid-19***.

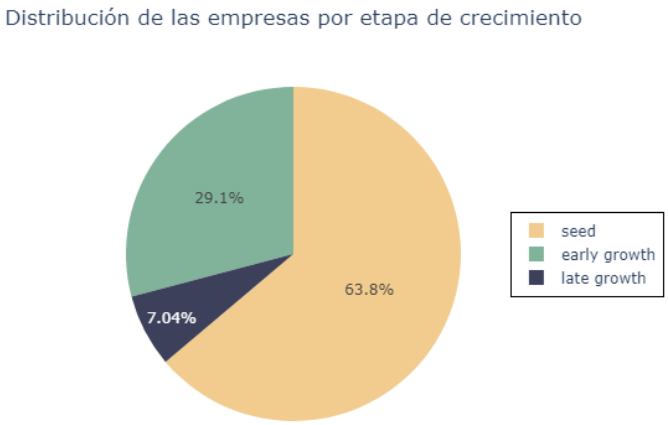


iii. Figura 04: Evolución de la media de las variables (2020-2021)

Por otro lado, resulta interesante examinar los **cambios** que suceden en diversas variables con el poder del **tiempo**. En este caso, se han decidido analizar las siguientes variables: *cifra de ventas, el activo total, las deudas financieras, nº empleados* y *el resultado financiero*. Como norma general, las **startups tienden a crecer en sus primeros años y con este gráfico de aquí se puede comprobar lo dicho.** Así pues, en general, las empresas han sufrido un aumento en sus ventas, activo, etc. y lógicamente, también lo han sufrido en sus **responsabilidades, es decir, en sus deudas.**

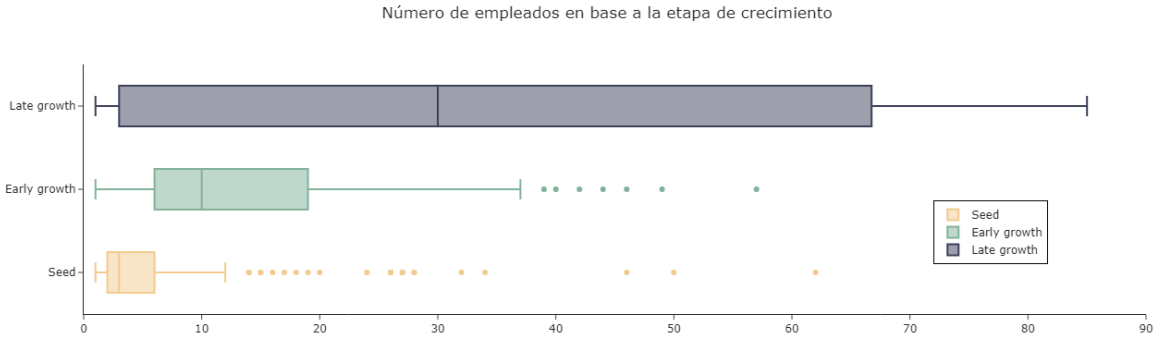
ii. 4.1.1 Análisis descriptivo: Dependiendo de la etapa de crecimiento

La situación en la que se encuentra una *startup*, es decir, la **etapa de crecimiento** en la que se encuentra una empresa puede ser **clave** para **determinar diversas características o datos sobre cada una de ellas**. Por ello, hacer un análisis más exhaustivo sobre la **etapa de crecimiento** de las empresas puede ser interesante para identificar si existen diferencias notorias entre las empresas que se encuentran en cada una de las fases.



i. Figura 05: Gráfico de tarta de la distribución de las empresas por etapa de crecimiento

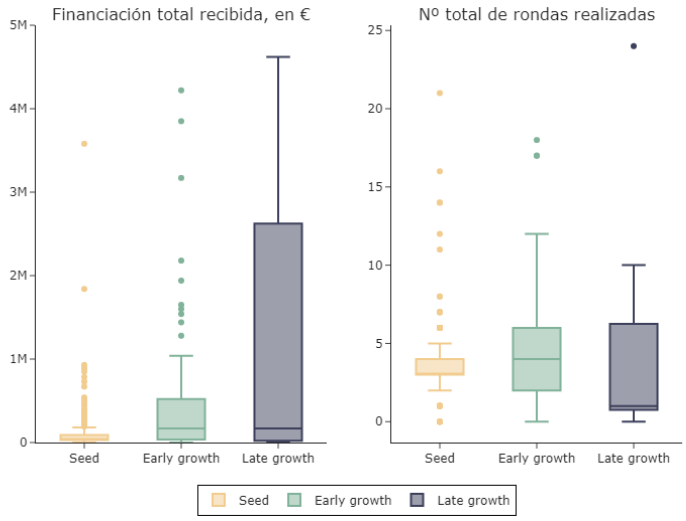
Como bien podemos observar en la *Figura 05* la mayoría de las empresas que se encuentran en el dataset se sitúan en la etapa de crecimiento *"***Seed***"* (ocupando el **63.75%** del total de empresas). Estas empresas, como su nombre indica, son empresas que **acaban de iniciar el proyecto**. En cambio, por otro lado, se encuentran las *startups* que se encuentran en la fase *"***Late growth***"* (ocupando el **7.06%** del total). Así pues, se puede ver que este pequeño número de empresas son los que ya **han conseguido algún que otro cliente fijo** y una **fuente de ingresos medianamente estable**. Como anteriormente se ha mencionado, a la hora de elegir la empresa al que se le realizará un plan financiero, tendrá que ser una empresa que esté en la fase *early growth* o *late growth*.



ii. Figura 06: Boxplots del número de empleados en base a la etapa de crecimiento

Como era de esperar, una de las características que mejor define a cada etapa de crecimiento de una *startup* es el **número de empleados** o el **tamaño del equipo** de trabajo que hay en cada una de las empresas. Si se analizan las empresas que dispone el dataset visualizamos que las *startups* que están en la fase ***seed*** disponen de **media**

una plantilla de tres trabajadores mientras que el número de empleados de las empresas que se encuentran en la fase ***late growth*** asciende mucho hasta los **30 empleados de media**.



iii. Figura 07: Boxplots de la financiación recibida y el nº total de rondas realizadas

Es importante también analizar aquellos datos que estén asociados a las **rondas de financiación** de las empresas. Por ello, se han visualizado tanto la **financiación total recibida** en el año **2021** como el **número total de rondas realizadas** en este mismo año. Algo curioso que se observa es que tanto en la financiación total como en las rondas totales realizadas la media de las empresas de la fase *late-growth* **está por debajo de las demás.**

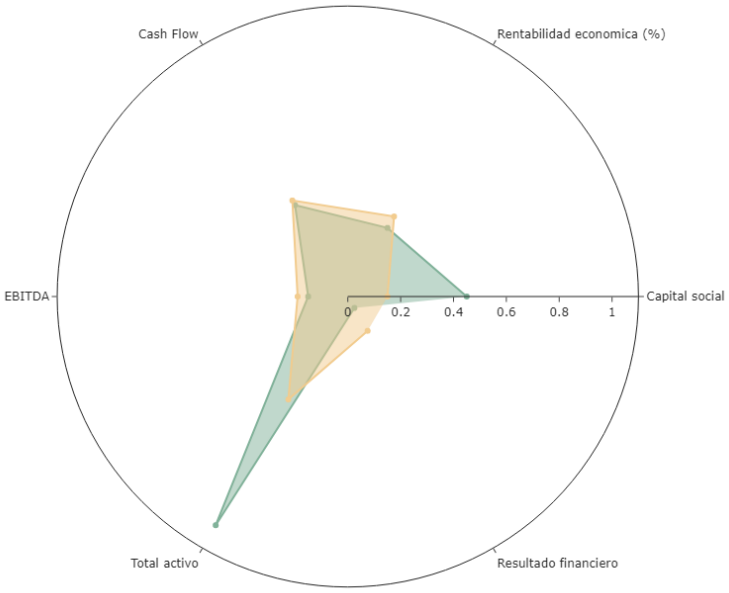
iii. 4.1.2 Análisis descriptivo: Wise Security Global

Uno de los objetivos que se han planteado ha sido el de realizar una **valoración para una de las empresas** que se encuentra en el dataset. Para ello, primero de todo, se ha seleccionado una *startup* que se encuentra por encima de la fase de crecimiento *"seed"*. Una vez barajadas diferentes opciones, finalmente, se ha seleccionado la empresa alavesa llamada ***Wise Security Global***.

Antes de empezar a valorar y a plantear un plan financiero para la empresa, con el fin de mantener una primera toma de contactos con dicha empresa, **conviene analizar rápidamente pequeños aspectos generales y alguna que otra variable financiera**. Así pues, primero de todo, se hará una comparación muy general de los **resultados o valores** que tiene la empresa alavesa con la media de los resultados del resto de empresas.

Wise Security Global vs Resto de empresas

Empresa	Cash flow mil EUR	EBITDA mil EUR	Resultado financiero mil EUR	Rentabilidad economica (%)	Capital social mil EUR	Total activo mil EUR
RESTO EMPRESAS	428.28	682.48	-250.32	-4.55	2817.97	25490.86
WISE GLOBAL SECURITY	497.47	590.24	-26.12	14.9	3.25	2701.66



i.

ii. Figura 08: Gráfico de radar. Wise Security Global vs. Resto de empresa

Primero de todo, se ha decidido hacer una **comparación** de los resultados de diferentes **indicadores financieros** mediante un gráfico de radar. Para la realización de dicho gráfico se han cogido, por una parte, los valores de los indicadores económicos de la **startup Wise Security Global** y, por otra parte, la media del **resto de empresas** que están en la misma fase que esta *startup*, la fase *late-growth*.

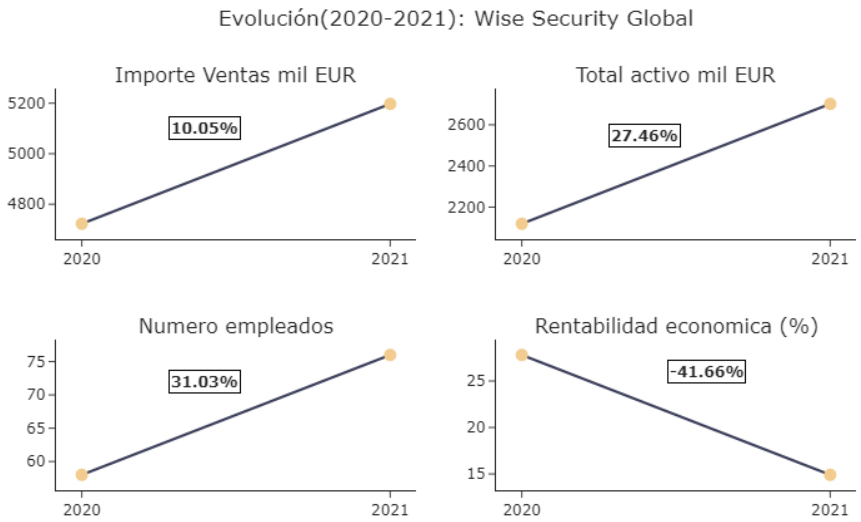
El objetivo de este gráfico es el de ver las principales **diferencias** que tiene la empresa a analizar con el resto de empresas que están en su misma fase. De esta forma, se conseguirá **tener una idea de la situación económica de la empresa** en comparación a las demás empresas.

Los **indicadores** que se han escogido para realizar la comparación han sido los siguientes: el cash flow, la rentabilidad económica, el capital social, el resultado financiero, el activo total y el EBITDA.

Debido a que son las variables que **más diferencias** muestran entre la **startup** y la **media del resto de empresas**, lo que más llama la atención, probablemente, sean los indicadores del **total activo** y la **capital social**. Como bien se puede observar en el gráfico el capital social de *Wise Global Security* estaría **por debajo de la media** y el total activo estaría **muy por debajo** de ella.

Por otro lado, estaría el indicador del EBITDA, donde el valor de la empresa a analizar se encontraría **ligeramente por debajo de la media**. No obstante, finalmente, son los siguientes tres indicadores de la empresa los que tendrían el valor por encima de la media del resto.

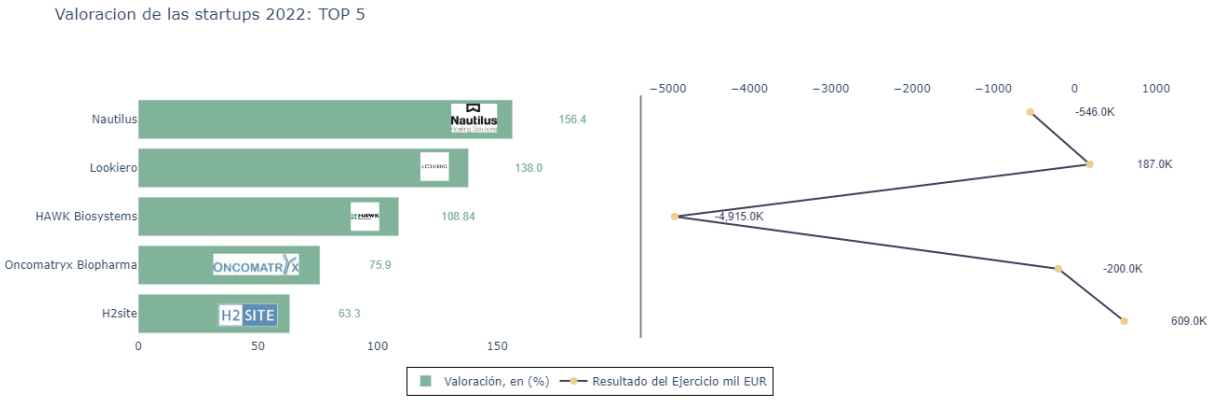
Como conclusión se puede decir que la empresa, en comparación al resto de las empresas que se encuentran en la misma fase de crecimiento, anda **escaso** de **posesiones**, **bienes y derechos**. No obstante, en los demás indicadores se observa que la empresa anda rondando por la media del resto por lo que, dichos valores negativos no son algo que deban de preocupar demasiado.



iii. Figura 09: Evolución de los indicadores de *Wise Security Global*

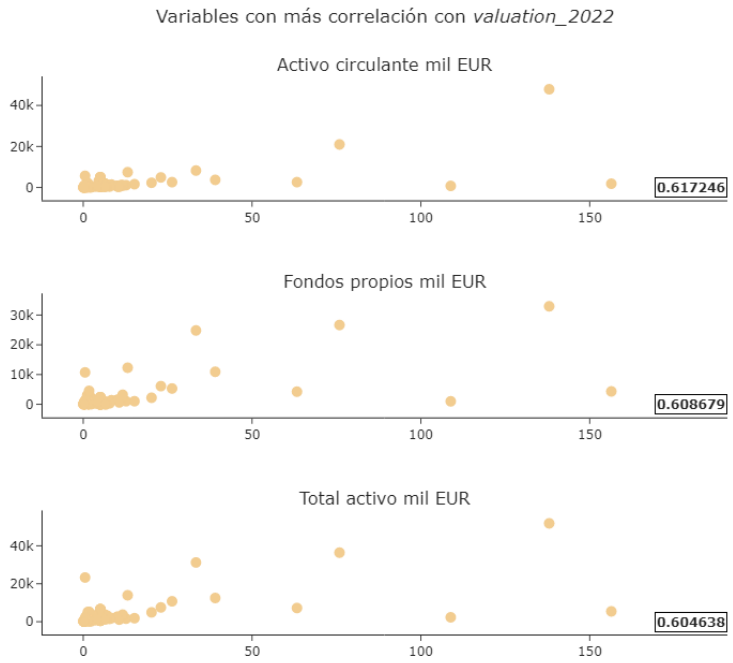
Asimismo, se ha realizado otro gráfico para analizar la **evolución** que han tenido algunos de los **indicadores** de *Wise Security Global*. Como solo se dispone de los datos de 2020 y 2021, el análisis de la evolución queda bastante pobre. No obstante, puede servir para ver que, efectivamente, la **tasa de crecimiento ha aumentado respecto al anterior año** tanto en el importe de ventas, en el activo total y en el número de empleados. Por otro lado, no pasa lo mismo con la **rentabilidad económica**, por lo que, a pesar de que esta siga siendo positiva, es algo que se puede tener en cuenta de cara al futuro.

Por otro lado, el dataset contiene una variable muy interesante llamada "**valuation_2022**" con las valoraciones de las empresas. Es importante tener en cuenta dicho variable ya que más adelante se tratará de **predecir el valor de este para aquellas startups que no tengan valor en esta columna.**



iv. Figura 10: TOP 5 Startups con mayor valoración

Primero de todo, para conocer cuales son las empresas que **mejor valoración** tienen en el dataset, se ha decidido hacer un top 5 de las empresas que mejor valoración tienen: *Nautilus, Lookiero, HAWK Biosystems, Oncomatryx Biopharma* y *H2site*. Por otro lado, al costado se visualiza la variable "Resultado del Ejercicio mil EUR", que muestra el valor que han obtenido las **5 empresas en este indicador financiero**. En este caso, se observa que no hay una **relación directa** entre cuán alta sea la **valoración** de una *startup* y su **resultado financiero**.



v. Figura 11: Variables con más correlación con *valuation_2022*

Finalmente, se han visualizados las 3 variables que **más correlacionan** con "valuation_2022" mediante gráficos de dispersión. Las variables serían las siguientes: "Activo circulante mil EUR" (0.6172 correlación), "Fondos propios mil EUR" (0.6086 correlación) y "Total activo mil EUR" (0.6046 correlación). Tener conocimiento de esto puede **resultar de gran ayuda**, sobre todo, a la hora de **seleccionar las variables** para el **modelo de predicción** de "**valuation_2022**".

b. 4.2 Modelar los datos y Visualizar los resultados obtenidos

Una **predicción** es la acción de anunciar un hecho futuro a través del empleo de los datos existentes. Las predicciones tratan de **anticiparse al futuro**, sin embargo, al existir todo tipo de dificultades, estas nunca pueden ser totalmente fiables. Por tanto, el objetivo es **acercarse lo máximo posible a las creencias y a los modelos estadísticos.**

Para poder obtener la mejor predicción posible, se han llevado a cabo muchas modalidades de modelos. Desde modelos de **regresión** para predecir la **valoración** de las empresas, hasta modelos de **clasificación** para saber si una empresa será **adquirida** totalmente, adquirida parcialmente, o no adquirida.

Con el fin de llevar a cabo estos modelos, se ha empleado todo tipo de **modelos sencillos**, y a estos se les han sumado modelos de **Ensemble** *learning* y otros tipos de modelos de alto rendimiento.

- **Datos utilizados**

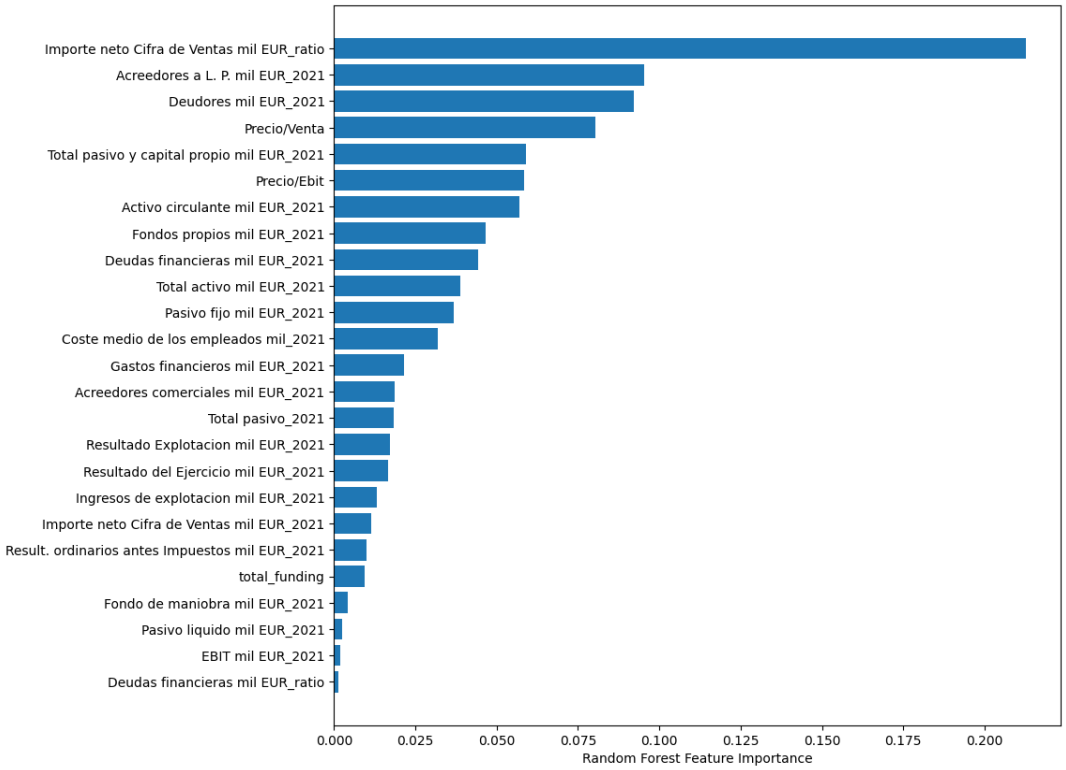
Para crear el los dataframes de valoración y de adquisición, se ha tomado de base el dataframe final creado una vez realizado el preprocesamiento de datos. A pesar de contener **4 datasets** al comienzo, se ha logrado juntar toda la información relevante en **uno único**.

Una vez se ha conseguido tener toda la información en un único dataframe, este se ha dividido en dos. Uno de ellos tendrá la columna "**valuation_2022**" como **variable a predecir** y se le eliminará la columna "Porcentaje_adquisición_cat", y en cambio, "df_adquisión" tendrá la columna "**Porcentaje_adquisicion_cat**" como **variable a predecir** y se le eliminará la columna "valuation_2022".

Tras contener estos dos dataframes, se debe realizar una **selección de variables** debido a la gran cantidad de variables en comparación con el número de empresas. Por ello, se ha empleado el método "**feature_importances**" para mostrar la **importancia** de

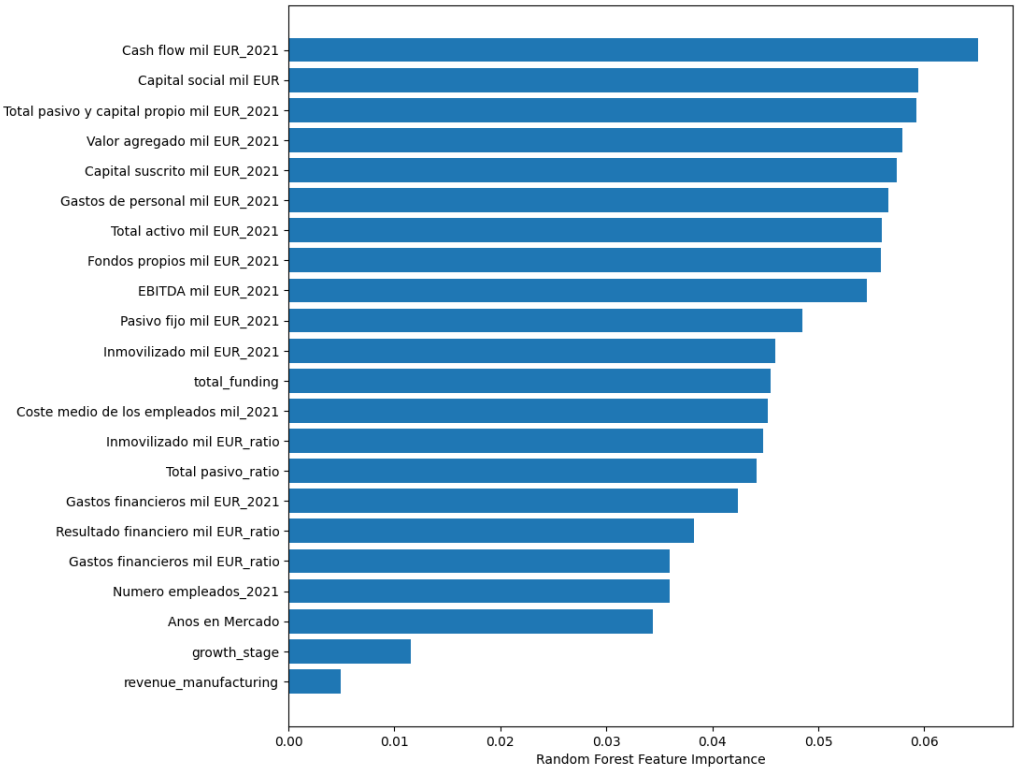
cada variable en la predicción del modelo. Con esas importancias, se han creado varios gráficos para poder visualizar la importancia de cada variable sobre la variable a predecir en cada caso. Estas importancias se han ordenado de mayor a menor para poder identificar fácilmente aquellas que tienen importancia en las variables objetivo. (Se han omitido los que tenía una importancia menor a 0.00025)

Dataframe valoración:



i. FIgura 12: Importancia variables valoración

Dataframe adquisición



ii. FIgura 13: Importancia variables clasificación

Además de tener en cuenta la importancia de las variables sobre la variable a predecir, también se ha tenido en cuenta el **número de valores** que se han tenido que **imputar** en esas variables y la **correlación** con la variable a predecir.

Esto es, al comienzo del preprocesamiento de los datos, muchos valores de las variables fueron imputados debido a la **ausencia de información**. Por tanto, muchas de las variables contienen valores imputados, los cuales pueden provocar una **falta de fiabilidad** en estas variables. A consecuencia de ello, se ha **comparado el número de instancias imputadas por variable con el ratio de importancia** sobre la variable a predecir, y con ello, se ha realizado la selección de las mismas. Además, como se ha mencionado anteriormente, también se ha tenido en cuenta la correlación con la variable predictora para hacer una primera criba, la cual ha sido analizada mediante una matriz de correlación.

Una vez realizada la selección de variables, los dos dataframes, contienen un total de **21 variables**, siendo el dataframe de valoración el más pequeño, con tan solo 60 instancias, y el *dataframe* adquisición el que contiene todas las empresas (412).

- **Modelado**

Como se ha mencionado anteriormente, la fase de modelado consta de **dos secciones** principales: en primer lugar, la predicción de la **valoración** de un conjunto de empresas y, en segundo lugar, la predicción de la **adquisición** de las empresas, es decir, si una empresa será adquirida en su totalidad, parcialmente o no será adquirida. Para una

mejor comprensión de ambos modelos, se ha dividido el proceso de modelado en modelos de regresión y modelos de clasificación.

- Valoración de empresa:

El dataframe de valoración de empresas contiene un total de 60 empresas (filas) y 20 variables. Al ser un **modelo supervisado**, se ha **separado** la variable predictora de la variable a predecir, y se han separado los datos en train y test para **entrenar** los modelos y **posteriormente testear**.

Cabe destacar que al ser un dataframe con un número de filas muy bajo en proporción al número de variables, en todos los modelos se ha hecho uso de la **validación cruzada** para evaluar la capacidad del modelo. El objetivo de la validación cruzada es **conocer la capacidad** que tiene el modelo para entrenar sobre los datos existentes, usando cada vez una parte del conjunto de datos como test. En este caso, se ha utilizado la **validación cruzada** sobre los datos de train, y al final se han cogido los datos de **test para probar los mejores modelos** con datos que el modelo no ha visto previamente.

La columna "***valuation_2022***" al tratarse de una columna con valores numéricos muy altos y en la que hay una **gran diferencia** de valoración entre algunas empresas, tiene una varianza muy elevada, lo cual **afecta al rendimiento** de ciertos algoritmos de aprendizaje automático. Por tanto, con el fin de **reducir esa varianza** y que no afecte al rendimiento del modelo, se ha escalado la variable objetivo en la parte de entrenamiento con el uso del **logaritmo neperiano**. Este escalado será posteriormente revertido para poder realizar una comparación correcta sobre los datos de testeo.

Ya con todas las divisiones y escalados correspondientes realizados, se pasa al apartado de la explicación de los modelos empleados. En este apartado se van a tratar todo tipo de modelos, tanto simples como modelos de ***Ensemble learning*** y modelos de aprendizaje automático más complejos.

Para conocer qué variables son las que **mejor** explican la variable a predecir, y comprender los resultados de la regresión lineal realizada, se ha realizado una matriz de **correlación** para **visualizar** estas relaciones. Mediante esta, se ha podido conocer que hay variables que muestran una alta correlación sobre "*valuation_2022*", como por ejemplo "Activo circulante", "Fondos propios", "Total activo" y "Total pasivo y capital propio", las cuales rondan un **0,60** de correlación .

Esta matriz de correlación nos sirve para poder identificar patrones previos y correlaciones entre las variables y comprender mejor el rendimiento de las distintas variables en las regresiones lineales.

En cuanto a los modelos de ***Ensemble learning***, se ha realizado un ***Stacking***, el cual **combina** las **predicciones** de varios modelos de aprendizaje automático e individuales para **mejorar** la precisión de la predicción final. En este caso, el modelo coge como

estimadores base los modelos de *KNeighborsRegressor*, *DecisionTreeRegressor*, *SVR* y *LinearRegression*, y como estimador final el *TweedieRegressor*.

También se ha empleado *BaggingRegressor*, el cual combina múltiples modelos de regresión para **mejorar** el rendimiento del modelo. Se define el modelo *BaggingRegressor* con *RandomForestRegressor* como estimador base y se realiza una validación cruzada, previamente explicada.

Además del *Stacking* y el *Bagging*, también se emplean *AdaBoostRegressor* y *GradientBoostingRegressor*, que son algoritmos que se basan en el método de ***boosting***. *AdaBoostRegressor* utiliza un modelo de *boosting* que se enfoca en los **errores cometidos en las predicciones anteriores** y ajusta el peso de los casos para reducir el error en la siguiente iteración. *GradientBoostingRegressor*, por otro lado, **ajusta los residuos** de los modelos anteriores en cada iteración, lo que permite una mejor **aproximación** de la función objetivo.

Por último, se emplean **dos** modelos *Boosting* adicionales, como son ***XGBoost*** y ***CatBoost***. *XGBoost* es una librería optimizada para ***gradient boosting*** que se enfoca en **mejorar** la **velocidad** y **escalabilidad** del algoritmo. *CatBoost*, por otro lado, es una librería de ***gradient boosting*** que se enfoca en manejar de manera **eficiente** datos categóricos. A diferencia de otras librerías, *CatBoost* no requiere una transformación previa de los datos categóricos en variables numéricas.

Todos los resultados de los modelos de regresión mencionados anteriormente han sido medidos utilizando **dos** métricas comunes: el **coeficiente de determinación** (*R2 score*) y el **error absoluto medio** (*MAE*) a partir de los resultados de la validación cruzada.

Tabla de resultados de valoración

Model	R2 score	MAE
Random Forest Regressor	0.72	0.47
Bagging Regressor	0.69	0.60
AdaBoost Regressor	0.68	0.54
CatBoost	0.67	0.44
XGBoost	0.62	0.46
GradientBoost Regressor	0.62	0.44
Decision Tree Regressor	0.55	0.47
Stacking Regressor	0.25	0.97

Linear Regression	-130.30	2.67

b. Tabla 02: Resultados modelos valoración

A través de esta tabla de **resultados**, se puede ver que los modelos *Random Forest Regresso*r, *Bagging* y *AdaBoost* presentan los **mejores** resultados tanto en términos de *R2 score* como de *MAE*. Estos, al ser modelos de *Ensemble learning* de árboles, tienen capacidad de ajustar datos no lineales. Estos modelos tienen la **habilidad** de **capturar relaciones no lineales** y **complejas** entre las variables **predictoras** y la variable objetivo, lo cual hace que tengan un **buen** rendimiento.

Además, estos modelos utilizan **técnicas de regularización** y reducción de la varianza, lo que permite **evitar** el sobreajuste y **mejorar** la generalización del modelo. Por lo tanto, estos modelos son más **adecuados** para conjuntos de datos complejos y no lineales, lo que resulta en una mejor capacidad predictiva y una mejor interpretación de los resultados, el cual puede ser el motivo de esta disposición de resultados.

- Clasificación de empresa:

Como se ha mencionado anteriormente, el *dataframe* de clasificación de empresas contiene un total de 412 empresas (filas) y 21 variables. Al ser un **modelo supervisado**, se han separado las variables **predictoras** de la variable a **predecir**, y se han separado los datos en train y test; para poner entrenar y posteriormente testear. A pesar de haber separado los datos en train y test, se vuelve a dividir el apartado de train, en training y validación. Esta división adicional permite **ajustar** los hiper parámetros del modelo en el conjunto de validación, mientras se utiliza el conjunto de entrenamiento para ajustar los parámetros del modelo. De esta forma, se puede **evitar** el **sobreajuste** del modelo al conjunto de entrenamiento y se puede evaluar su capacidad en datos nuevos y desconocidos.

Primeramente, se prueba a realizar un modelo mediante *AutoML* de *TPOT*, el cual **evalúa automáticamente diferentes modelos** de aprendizaje automático. Este, devuelve un *accuracy* de 0.64 presentada en esta matriz de confusión.
En segundo lugar, se entrenan **modelos simples** como, Regresión logística, *Random forest*, *SVC*, *KNeighbors*, *Decision Tree* y *GaussianNB*. Todos estos modelos se prueban con los datos de ***cross validation***, y estos devuelven como mejor resultado el Random Forest, con un accuracy de **0.59**.

Posteriormente, se comienza a emplear modelos de *Ensemble learning*. En el primer modelo de *Ensemble learning*, el **Stacking**, se emplean una lista de modelos simples como estimadores, un modelo **SVC como estimador** final y se le aplica el *gridsearch* para encontrar los mejores hiper parámetros. Una vez realizada la predicción sobre los datos de cross validation, este devuelve un accuracy de 0.67.

Como segundo modelo de Ensemble learning, se prueba el Bagging. Para encontrar los mejores hiper parámetros para el modelo **Bagging**, se emplea un **GridSearch sobre**

un conjunto de hiperparametros para saber cuales son los que mejor se comportan. En este caso, se emplea el Random forest como modelo estimador y los mejores hiperparametros como parámetros en el modelo. Posteriormente, se predice sobre los datos de cross validation y devuelve un accuracy de **0.61**, con esta matriz de confusión.

Posteriormente, a este mismo modelo se le aplica y una **optimización bayesiana** sobre los hiper parámetros para mejorar su rendimiento. Se vuelve a emplear como estimador el random forest, y al predecir sobre los datos de cross validation, devuelve un accuracy de **0.61**, el cual se ve reflejado en esta matriz de confusión.

Por último, se emplean algoritmos como *AdaBoost*, *GradientBoosting*, *XGBoost* y *Catboost*, que son modelos que se basan en el método de *Boosting*. En este caso, **CatBoost** es el que mejor *accuracy* muestra, **0.57**. Estos datos se pueden ver en la tabla 03.

Model	Accuracy
TPOT	0.64
Optimización Bagging GridSearch	0.61
Bagging GridSearch	0.61
Random Forest	0.59
CatBoost	0.57
Stacking	0.57
AdaBoost	0.56
GradientBoost	0.55
XGBoost	0.55

c. Tabla 03: Resultados modelos clasificación

Para poder conocer **cómo se comportan estos modelos con los datos de test**, se han seleccionado **3 modelos** que mejor comportamiento mantienen en los datos de prueba una vez visualizados sus *accuracy* en *cross validation*.

En primer lugar, el modelo de *Stacking* ha mostrado un *accuracy* de 0.40 sobre los datos de prueba **bajando 0.17** sobre *cross validation*. Esta es su matriz de confusión.

En segundo lugar, el modelo de *GridSearch Bagging* ha mostrado un *accuracy* de 0.57 sobre los datos de prueba **bajando 0.04** sobre *cross validation*. Este modelo baja en muy pequeña proporción su *accuracy* sobre los datos de prueba, lo cual nos indica el correcto funcionamiento del modelo. Esta es su matriz de confusión.

Por último, el modelo de *Optimización GridSearch Bagging* ha mostrado un *accuracy* de 0.60 sobre los datos de prueba **bajando 0.01** sobre *cross validation*. Este es el **mejor modelo respecto a los datos de *cross validation***; esta es su matriz de confusión.

Discusión crítica modelos y resultados

ii. En este apartado se **debatirán** los **resultados obtenidos** por los modelos, así como su **veracidad** y cómo se podrían **mejorar**. En primer lugar, cabe mencionar que hoy en día existen muchos métodos de valoración de empresas, y que hay personas e incluso compañías que se dedican a esto de continuo, y aun así no siempre sacan valoraciones **acertadas**. Es decir, que es un **campo muy complicado**, subjetivo y difícil de predecir sobre todo, por lo que hacer un modelo fiable es **muy complicado**.

Además, con los **pocos datos** de los que se disponían, la **complejidad** del modelo es todavía **mayor**. Tan **solo** había **60 empresas** en todo el conjunto de datos que tuvieran una valoración, y por ende, solo se han podido utilizar las mismas para el entrenamiento de los algoritmos. Así, es **muy probable** que los modelos **sufran** de ***overfitting***, y que no sean capaces de generalizar correctamente a nuevos datos. De hecho, tras haber entrenado los modelos y haberlos evaluado con el conjunto de test, se ha comprobado que los resultados obtenidos son **mucho peores** que los obtenidos con el conjunto de train, lo que confirma la sospecha de que los modelos están sufriendo de *overfitting*.

Una posible **solución** a este problema sería **aumentar** el conjunto de datos con el que se entrena el modelo, pero en este caso no es posible, ya que no se dispone de más datos. En caso de que los hubiera, el modelo sería más capaz de aprender los patrones entre las variables y el target, y por ende, sería capaz de predecir mejor los datos nuevos que se le introducen.

Otra solución para mejorar los resultados sería la **creación** de nuevas variables, en este caso **más complejas** y que sean más descriptivas de la empresa. Para esto estaría bien disponer de más información sobre las empresas, y además, **conocimiento técnico** sobre la materia. Por otro lado, se podría probar con otros algoritmos de *machine learning*, como **redes neuronales**, que son más complejos y capaces de aprender patrones más complejos, o incluso hacer una combinación de métodos más personales y tradicionales, con algoritmos de *machine learning*.

Respecto al modelo de **adquisición**, se puede decir que los resultados obtenidos **han sido mejores que los obtenidos con el modelo de valoración**. Esto se debe a que el modelo de clasificación tiene un **rango** de valores **más reducido**, había más datos disponibles para entrenar el modelo (412 frente a 60), y además, el modelo de adquisición se enfoca en una **pregunta más concreta y específica**, lo que hace que sea más fácil de predecir con precisión.

Sin embargo, por lo general, los modelos utilizados **tienden** a clasificar las empresas en la categoría de **“no adquirida” y "parcialmente adquirida"**, lo que se debe a que hay **pocos registros** que pertenezcan a la categoría de “**completamente adquirida**”. Esto se puede solucionar de varias formas, entre ellas aumentando el número de registros de la categoría minoritaria mediante el uso de balanceo de datos, o bien, reduciendo el número de registros de las otras categorías. En este caso, no se ha hecho ninguna de estas dos cosas, ya que **no había datos suficientes para quitar instancias, y además, no se querían añadir empresas artificiales.**

c. 4.3 Conclusiones

En conclusión, la valoración de empresas es una tarea muy complicada de hacer, y por ello hay cientos de personas que se dedican diariamente a hacerlo. En consecuencia, existen muchos métodos y diferentes entre sí para hacer las valoraciones. Encima, la valoración de startups es un campo muy subjetivo, que suele estar basada en la experiencia, y en vivencias de las personas que se dedican a ello.

Además de las dificultades ya mencionadas, existe el impedimento de que se disponían de pocos datos, y así los modelos quedan demasiado sesgados, aparte de que son poco precisos. En definitiva, el machine learning no es la mejor herramienta para hacer predicciones de valoraciones de startups.

11. 5. Transformar los negocios

La empresa elegida para la realización de la valoración ha sido *Wise Security Global*, que es una empresa de ciberseguridad cuyo **objetivo** principal consiste en **proteger** la actividad de sus clientes generando entornos **ciberseguros** y **ciber confiables**, que les permitan mantener y mejorar la confianza de sus *stakeholders*. Se ha escogido esta empresa porque no tenía datos ausentes, tenía una cuenta de resultados bastante estable, y pertenecía a un sector del que se pueden conseguir empresas similares a las que poder compararse.

Es por ésto que, ***Mandiant*** al ser una empresa de **ciberseguridad,** ha sido escogida para comparar con *Wise Security Global* ya que las dos compañías pertenecen al **mismo sector** y es la que más se puede acercar respecto a las actividades realizadas en los dos negocios. Fue fundada en 2004 por Kevin Mandia y ha trabajado con numerosas empresas y organizaciones gubernamentales para ayudar en la **detección** y **prevención** de ataques cibernéticos. Además, hace un par de años fue adquirida por ***Google***, y cotiza en la bolsa de valores **estadounidense**.

A continuación se van a detallar los 3 métodos de valoración realizados para hallar el valor de Wise Security Global, que son el flujo de caja libre, el de múltiplos, y el subjetivo.

Método 1: FLUJO DE CAJA LIBRE

El primer método es el del **flujo de caja libre**. En la tabla 04 se representan los valores de los Free Cash Flows (FCF) de los años 2021 al 2024, así como las 5 variables que se utilizan para su cálculo. Los únicos datos de los que se disponían eran los del 2021, y se han hecho una **serie de estimaciones para los años futuro**s. En primer lugar, se ha analizado la evolución del EBIT en los años anteriores al 2021 y se ha visto que en ese año tuvo una caída en el mismo respecto al 2020, probablemente por el repunte que hubo en el 20 en la tecnología y en la ciberseguridad, y una posterior caída. Sin embargo, se ha pensado que el EBIT podría **aumentar teniendo en cuenta la trayectoria que los beneficios de la empresa tenían antes del COVID**, y se ha incrementado la variable en un 30% (al igual que los impuestos, que van de la mano). Todos los valores de la tabla 04 están representados en miles de euros.

Año	EBIT	Impuestos	Amortización	Inversión en Capex	Inversión Activo Circulante	FCF
2021	428,859	66,63897	161,381	220,5	361,5	-58,3985
2022	556,4	86,58	161,381	220,5	361,5	49,201
2023	656,552	102,1644	161,381	220,5	361,5	133,7689
2024	774,731	120,5539	161,381	220,5	361,5	233,55846

a. Tabla 04: Predicción de FCF a 3 años

Luego, para los demás años, el crecimiento del EBIT se ha proyectado teniendo en cuenta que el crecimiento de las empresas **se suele frenar ligeramente**, pero sigue siendo una empresa en desarrollo, y que tiene un gran potencial, por lo que se ha considerado que crecerá un 18% al año. Las amortizaciones, inversiones en capex, e inversiones en activo circulante se han mantenido **constantes** a lo largo de los años puesto que **no se tiene información suficiente sobre su trayectoria pasada.**

En función de cómo han evolucionado los FCF en los primeros años, se ve que al principio su crecimiento es mayor, por ejemplo pasa de 49 a 133, es decir, que hay un salto cerca de 160%, **pero más adelante el crecimiento es menor** (75%) ya que pasa de 133 a 233. Por tanto, se ha asumido que en los próximos años el *Free Cash Flow* va a crecer más paulatinamente que hasta el 2024, y se han reducido los incrementos. Por ejemplo, en el año 2025 se ha asumido un **aumento** del 37% y en el 2026 del 30%. Después de hacer los cálculos de los FCF, se han actualizado los valores obtenidos en función de los años de diferencia respecto al 2021 y un WACC de un 9%, que se ha escogido en base al dato de la empresa Mandiant, que es también una empresa de ciberseguridad que cotiza en bolsa, y que se tomará como referencia posteriormente para hacer otras valoraciones. El resultado de los mismos se pueden ver en la tabla 04 y se muestran en miles de euros. La suma total de los FCF es de 835.251€.

Año	**FCF**
2021	-58,3985
2022	$49,201 / (1 + 0,09)^1 = 45,14$
2023	$133,7689 / (1 + 0,09)^2 = 112,59$
2024	$233,55846 / (1 + 0,09)^3 = 180,35$
2025	$319,975 / (1 + 0,09)^4 = 226,68$
2026	$415,967 / (1 + 0,09)^5 = 270,35$

b. Tabla 05: Predicción de FCF a 5 años

Además de los FCF de los 5 siguientes años, se ha calculado el valor residual de la compañía teniendo en cuenta que el valor del FCF del año 2026 actúa como una renta perpetua, y que el crecimiento del mismo a largo plazo es del 5%. El valor residual sería el valor que tiene la empresa en el año 5 en base a lo que generará en el futuro. Al estar en otro momento en el tiempo, hay que actualizar este valor mediante el WACC una vez más. En la tabla 05 se ven los datos utilizados para su cálculo, y en el apéndice la fórmula empleada.

Tasa de	**WACC**	**Crecimiento a perpetuidad**	**FCF año 5**	**Valor residual**

crecimiento				
15%	9%	5%	415.967 €	7.772.560 €

c. Tabla 06: Cálculo de valor residual FCF

Por tanto, hasta ahora se ha conseguido el valor actual de los FCF de los 5 siguientes años (835.251€), así como el valor residual en valor actual (7.772.560 €) también. Así, se pueden sumar ambos valores y se obtendría una **valoración estimada para la empresa, que en este caso sería de 8.607.067€.**

Método 2: MÉTODO POR MÚLTIPLOS

EV / EBITDA

Tras hablar sobre el método del flujo de caja libre para la valoración de la empresa, se procederá a mencionar el segundo método llevado a cabo: los múltiplos. Los múltiplos relacionan el valor de una empresa con el indicador al que se hace referencia. El valor de la sociedad se calcula multiplicando el indicador de referencia por un número. Lo normal es que el número usado sea similar al de otras empresas similares y en momentos similares de la economía. Es por esto que a la hora de hacer las comparaciones los datos se han obtenido teniendo en cuenta el eje temporal, 2020-2021.

Los múltiplos ofrecen la gran ventaja de que son muy intuitivos y fáciles de calcular. No obstante, sacrifican una parte importante de información que se debe tener en cuenta. Los más habituales son: EBITDA, PER, Múltiplo de Ventas, PCF, PVC, etc.

Hay múltiplos que hacen referencia al valor del negocio como el EBITDA. Otros como el PER hacen referencia directamente al valor de las acciones. Para pasar del valor del negocio al de las acciones hay que quitar la deuda financiera neta.

El primer método de múltiplos utilizado es el EV/EBITDA. En este caso, el valor de la empresa es el resultado de multiplicar un múltiplo por el **EBITDA** (*Earnings Before Interest, Taxes, Depreciation and Amortization*). En cuanto a Wise Security, el EBITDA en el 2021 era de 590 mil euros.

Es un ratio muy usado ya que calcula el valor de la compañía teniendo en cuenta los recursos que genera una empresa con independencia de políticas de financiación, políticas fiscales etc (por eso es previo a amortizaciones, gastos financieros, impuestos), ya que se basa en los recursos que genera la empresa a través de sus operaciones.

En este caso se han cogido como referencia países vecinos (Italia y Francia) para determinar el ratio de EV/ EBITDA, el cual ronda un 9% en el sector tecnológico. Normalmente en los sectores tradicionales, el múltiplo que se suele utilizar es entre 5 o 6 pero en este caso al tratarse de compañías tecnológicas e innovadoras y por tanto, se

ha utilizado un 8,8. Sin embargo, este múltiplo es una estimación teniendo en cuenta que cada sector es distinto y a su vez, las empresas que los forman también.

No obstante, no deja de ser una simplificación que limita bastante. Por ejemplo hay empresas intensivas en capital donde las inversiones anuales son importantes o empresas en que por la tipología de sector el comportamiento de las necesidades de circulante es diferente a otras.

El valor obtenido tras la multiplicación del EBITDA (590.241€) por el ratio de EV/ EBITDA (8,8) es 5.194.012 €.

Beneficio por acción (PER)

El *Price Earnings Ratio* (PER) mide cuántas veces una compañía cotiza su cifra de beneficio. El ***PER*** se puede obtener, bien tomando la totalidad de la compañía, o bien, los datos por acción. Sirve para calcular el valor de las acciones normalmente, pero también sirve para cuando las empresas no cotizan en bolsa.

Dicho esto, este múltiplo puede ser una herramienta muy útil para comparar el valor de una empresa con otra del mismo sector. En este caso, tras un cálculo teniendo en cuenta que el PER de Mandiant es de 31,85 y siguiendo la fórmula correspondiente para saber cúal sería el valor equivalente de *Wise Security Global*, aunque la compañía no cotice en bolsa, se ha llegado al valor de que ésta se encuentra rondando los 6 **millones** de euros. Cabe destacar que, comparándola con la compañía llamada **Mandiant** es podría ser que el múltiplo fuera menor, ya que, al tratarse de una empresa en estado de **crecimiento** no tiene la misma robustez que la otra compañía, pero ésto quiere decir que al tener un PER menor puede indicar que la empresa está infravalorada, y podría ser una oportunidad para que los inversores depositen capital en ella, lo que beneficiaría mucho a *Wise Security Global*.

El valor obtenido tras hacer la multiplicación de los beneficios de *Wise Security* (336.096) por el PER de Mandiant (31,85) es de 6.319.078 €.

EV/Ventas

El Múltiplo de EV/Ventas divide el valor de la empresa entre la cifra de ventas. Presenta como gran **ventaja** con respecto al ratio Precio/Ventas que tiene en cuenta la **estructura financiera** de la empresa. Para el cálculo de la valoración de *Wise*, se ha cogido un ratio estándar de EV/Ventas, que es **9,5**, y se ha multiplicado por las ventas de la compañía (5.250.000) y se ha obtenido una valoración de **49.875.000 €**.

En este caso, para ser una *startup* en fase de **crecimiento** la empresa está siendo valorada a un precio **excesivamente alto** de 49.875.000€. Esto significa que la empresa tiene un valor de casi 50 millones lo que es muy improbable ya que puede tener muchas ventas pero a su vez muchos costes y no generar beneficio alguno, cosa que no se tiene en cuenta en este método.

Así pues, este método **sólo** tiene en cuenta las **ventas** de la empresa y no considera otros factores importantes como su rentabilidad, la calidad de sus activos, su capacidad para **generar** flujo de efectivo y su **posición** en el mercado. Por lo tanto, también sería importante el conjunto de las métricas mencionadas anteriormente y los otros métodos de valoración vistos anteriormente que tienen mayor **fiabilidad** y uso.

Método 3: VALORACIÓN SUBJETIVA

La valoración de las empresas **no** es un estudio fácil y esto resulta ser más irrebatible si de *startups* se habla. De hecho, valorar una *startup* es una de las tareas más complejas que puede encontrar un analista financiero, sobre todo, si se encuentran en la etapa de *early-stage*. En este caso, se parte con un poco de ventaja ya que la *startup* que se ha seleccionado para hacer la valoración ha sido una que se encuentra en la fase *late growth*.

Hay que tener en cuenta que las *startups* suelen tener un *cash flow* negativo y un histórico de datos financieros inexistente o escaso. Por ello, debido a la falta de indicadores de rendimiento financiero que tienen las *startups*, no es posible utilizar las mismas técnicas que se utilizan para las empresas más establecidas. Al fin y al cabo, en la mayoría de los métodos tradicionales de valoración de empresas, se usan indicadores financieros que las *startups* no suelen tener, por ello, se considerarán otros métodos de valoración que se usan especialmente para analizar este tipo de casos, más en concreto métodos **cualitativos**.

Tanto el método del *Step-Up Valuation* como el del *Risk-Factor Summation* son dos métodos que se suelen utilizar para *startups* en fases iniciales (*early-stage*). Como bien se ha comentado, la *startup* a analizar llamada *Wise Security Global*, se encuentra ya en una fase más **avanzada**, no obstante, aplicar cualquiera de estos métodos puede llegar a ser interesante ya que permite valorar la empresa de forma **sencilla**.

Como bien se ha dicho, el uso del método ***Step-Up Valuation*** es muy sencillo. Únicamente se deben de listar 10 factores o logros ya preestablecidos y puntuar cada uno de ellos con una puntuación de de 0 o de 1: si se cumple se dará la puntuación de un 1, sino de un 0. Las declaraciones establecidas ayudarán a valorar a la *startup* correctamente ya que medirán tanto la escalabilidad, el mercado competitivo y otros muchos factores.

Una vez las puntuaciones son establecidas, se procede a hacer un cálculo donde se suele sumar 250.000€ por cada vez que se haya puntuado un factor con un 1. No obstante, se ha llegado a la conclusión de que con dicho valor la *startup* en cuestión no puede llegar a una valoración mayor a 2.500.000€. En este caso, como la empresa está en la **fase 3** se ha considerado que la valoración también debe de ser **mayor**, se ha decidido hacer una pequeña **modificación** al método y establecer el valor de 750.000€ por cada 1 establecido. De esta forma, se podrá obtener una valoración más *"adecuada"* de la empresa Wise Security Global. En la primera columna de la tabla 07 se pueden

ver los factores que se han considerado, y en la segunda columna su puntuación (1 o 0).

FACTORES DE STEP UP	1 = SÍ, 0 = NO
1. El tamaño del mercado supera los 5.000.000.000€	0
2. Negocio escalable	1
3. Los fundadores han tenido éxito previamente	0
4. Los fundadores están comprometidos a tiempo completo	1
5. MVP desarrollado, desarrollo de clientes en marcha	1
6. Modelo de negocio validado por clientes de pago	1
7. Firma con asociaciones industriales importantes	1
8. Hoja de ruta de ejecución elaborada y en curso	1
9. IP emitida o tecnología protegida	1
10. Entorno competitivo favorable	1
Total de factores de Step Up	**8**

Estimación del Pre-Money Valuation: 8 x $750.000 = $6.000.000

d. Tabla 07. Cálculo de la valoración de la empresa mediante el método del Step-Up Valuation

Para empezar, el tamaño del sector de la empresa en el mercado no llega de momento a los 5.000.000.000€, por lo tanto, no es posible asignarle el valor 1 a este logro. No obstante, como dato, es de mencionar que el tamaño del mercado actual ronda muy **cerca** de dicho valor. Por otro lado, la **escalabilidad** del negocio es algo que resulta indudable debido al tipo de negocio. Al fin y al cabo, se trata de una empresa **especializada en ciberseguridad**, cosa que a día de hoy busca tener más personas y empresas.

Otro de los logros que no se cumple es el 3º ya que no hay historial de que ninguno/a de los fundadores o co-fundadores de la empresa haya tenido éxito previamente. Sin embargo, esto no parece un inconveniente ya que a pesar de que no se cumpla este factor el resto de los que quedan sí que se cumplirían. Al fin y al cabo, se analiza una ***startup*** que a día de hoy ya está **muy avanzada** con productos mínimos viables desarrollados, hoja de ruta de ejecución ya elaborada, el modelo validado por clientes de pago y la propiedad intelectual ya establecida. Asimismo, resulta **indudable** tanto el entorno competitivo en el que se encuentra la empresa y la firma con asociaciones importantes (*Microsoft Partners, Blueliv, proofpoint,* etc.) como el compromiso a tiempo completo que tienen los fundadores de este.

Finalmente, después de haber obtenido un 1 en 8 de los factores y haber hecho los cálculos respectivos, se ha **estimado** que el Pre-Money Valuation de la empresa Wise Security Global es de **6.000.000,00 EUR**.

En resumen, tras ver los 3 métodos de valoración distintos, se han obtenido un total de 5 valoraciones para la empresa Wise Security Global, que se pueden ver en la tabla 08. Teniendo en cuenta que la mayoría de ellas excepto la del múltiplo de EV/ Ventas obtienen unos resultados similares, se asume que la valoración puede ser acertada. Por tanto, haciendo una especie de media ponderada entre las 4 valoraciones, se obtiene una valoración final de 6.530.000€. Sin embargo, tiene más sentido definir un rango en el que podría estar, que sería más bien entre 6 y 7 millones de euros. Aun así, como se ha dicho anteriormente, esto es solamente una estimación, y hay muchos factores que pueden afectar a que este valor sea diferente.

Método	Valoración obtenida
Free Cash Flow	8.607.067 €
Múltiplo: EV / EBITDA	5.194.012 €
Múltiplo: PER	6.319.078 €
Múltiplo: EV / Ventas	49.875.000 €
Subjetivo: *Step-Up Method*	6.000.000 €

e.

f. Tabla 08: Valoraciones obtenidas según el método utilizado

b.

c. 5.1 Implicaciones legales y éticas

A la hora de hacer una **valoración de una *startup*** se pueden llegar a plantear varias implicaciones legales y éticas. Lo primero que hay que tener en cuenta es que la valoración de una empresa es **subjetiva** y puede llegar a crear **conflictos de intereses** entre las diferentes personas involucradas en el proceso de valoración.

En el caso de los **inversores**, estos podrían estar interesados en inflar la valoración de la *startup* para que **aumente el beneficio** que van a obtener. Por otro lado, a la ***startup*** le puede interesar más tener una valoración baja para **llamar la atención de más inversores**. Estos procesos pueden llegar a ser éticamente cuestionables, ya que con ellos se puede engañar a los inversores o al público en general.

Por otro lado, hay factores que no son relevantes para determinar si una *startup* va a tener éxito a largo plazo y que afectan a la hora de realizar una valoración. Por ejemplo, la **popularidad** actual en el mercado o la **presencia de inversionistas** variables ayudan

a que el valor de una *startup* se infle. Cuando esto ocurre, la valoración es posible que no refleje la verdadera calidad o potencial de la *startup*, por lo que se está engañando a los inversores.

Otro problema ético que puede relacionarse con la valoración de una *startup* es la **falta de transparencia** que hay en el proceso. Las *startups* es probable que no revelen información esencial para los inversores, por lo que la dificultad de valorar la *startup* aumenta en gran medida. Por lo que es fundamental que una *startup* sea transparente y que aporte datos concisos y completos a los inversores para asegurarles que la valoración es justa y precisa.

En pocas palabras, la valoración de *startups* puede sugerir varias implicaciones éticas que se deben considerar cuidadosamente. A la hora de realizar la valoración, todas las personas implicadas en el proceso deben ser honradas y transparentes para garantizar que la valoración pueda reflejar la calidad y potencial de la *startup*.

Otra implicación ética importante a la hora de valorar una *startup* es la importancia de la **participación activa de los responsables** de las *startups* en el proceso en el que se genera el modelo de valoración. Los responsables de las *startups* son los que poseen un profundo conocimiento acerca de su negocio y pueden aportar información crítica para realizar una evaluación más precisa de su valor. Si esa información no se incluye en el proceso, el modelo no va a reflejar con precisión el verdadero potencial de la empresa.

Además, si la forma de **valoración es incorrecta** y se le da un **valor** excesivamente **alto**, se va a asignar más dinero del necesario para la empresa. Esto puede llevar a la *startup* y a sus inversores a tomar decisiones imprudentes, como la expansión innecesaria o la inversión en proyectos que no son rentables. Por lo tanto, la **estabilidad financiera** de la empresa se puede poner **en peligro**, lo que podría afectar de una manera negativa a sus empleados, inversores y otros interesados.

Por otro lado, si a una *startup* se le da un **valor inferior**, puede hacer que su **capacidad** para **atraer inversores y obtener financiamiento** sea **limitada**, por lo que su crecimiento y desarrollo también se va a ver afectado. Esto puede crear problemas especialmente a las *startups* que ofrecen soluciones innovadoras y valiosas para problemas sociales y económicos importantes.

En conclusión, la valoración de ***startups*** plantea importantes implicaciones éticas que deben ser cuidadosamente consideradas. Los gerentes de empresas emergentes deben participar en el proceso de valoración para garantizar la **precisión** y la **equidad** en la evaluación del valor de la empresa. Además, es importante que los inversores y otras partes interesadas involucradas en la valoración sean **transparentes** y **honestos** en su evaluación de las nuevas empresas para garantizar que los recursos se asignen de manera justa y eficiente para el éxito de la empresa y en beneficio de todas las partes interesadas.

12. 6. Bibliografía

Manifiesto grupo spri. Recuperado el 13 de febrero de 2023.

https://www.spri.eus/es/quienes-somos/

Datos Financieros de Mandiant. Recuperado el 10 de Marzo de 2023

https://es.investing.com/equities/fireeye-inc-income-statement

Crecimiento de Wise Security Global. Recuperado el 12 de Marzo de 2023

https://app.dealroom.co/companies/wise_security_global

Métodos de valoración empresariales. Recuperado el 12 de Marzo de 2023

https://mudle.mondragon.edu/eteo/pluginfile.php/282156/mod_resource/content/1/SEMINARIO%20DE%20VALORACION%20PARA%20GRADO%20DE%20DATOS.pdf

Datos financieros extra Wise Security Global. Recuperado el 12 de Marzo de 2023

https://infocif.economia3.com/ficha-empresa/wise-security-global-sl

13. Apéndice I Finanzas

Para la valoración de la empresa se han tenido en cuenta diferentes métodos. En primer lugar, se ha calculado mediante el método del ***Free Cash Flow***. El *Free Cash Flow* representa el flujo de fondos que genera la empresa sin importar cómo se financia. Es decir, es el flujo de caja obtenido de las actividades de explotación de la empresa una vez han sido deducidas las inversiones para mantener el negocio. Se calcula de la siguiente manera:

FCF = EBIT - Impuestos + Amortizaciones – Inversión en Capex - Inversión en activo circulante.

Por otro lado también se ha dado el uso de método del **DCF** (Descuento de Flujos de caja) para hallar el valor terminal: Flujo de Caja Libre con su múltiplo y el Flujo de Caja Libre para cada año proyectado.

Para hallar el **FCL** para cada año proyectado se ha conseguido mediante la siguiente forma:

EBIT - Impuestos sobre EBIT + Amortización - Inversión (+-) Fondo de maniobra (↑↓) * $1/(1+WACC)^n$ = FCL Descontado

Siguiendo con el cálculo del valor del negocio se ha procedido a seguir con el **valor terminal**:

Último FCL proyectado * Tasa de crecimiento / (WACC-crecimiento a la perpetuidad) * $1/ (1+WACC)^{n-1}$ = Valor terminal descontado.

Para finalizar con el primer método tras todo lo mencionado anteriormente se ha procedido a ver el **valor final de la empresa**:

$\sum$ FCL descontados + Valor terminal descontado = Valor del negocio

En segundo lugar, se ha llevado a cabo la valoración de la empresa por distintos múltiplos: EV/ EBITDA, EBITDA, EV/ VENTAS. Éstos, se han utilizado con el fin de tener otros aspectos en cuenta a la hora de valorar la compañía elegida. Sus fórmulas matemáticas son las siguientes:

EV / EBITDA (*Earnings Before Interest, Taxes, Depreciation and Amortization*):

Es un ratio muy usado ya que calcula el valor de la compañía teniendo en cuenta los recursos que genera una empresa con independencia de políticas de financiación, políticas fiscales etc (por eso es previo a amortizaciones, gastos financieros, impuestos).

Valor de la empresa / EBITDA = Ratio usado

PRICE EARNINGS RATIO (PER):

El Price Earnings Ratio mide cuántas veces una compañía cotiza su cifra de beneficio. El PER se puede obtener, bien tomando la totalidad de la compañía, o bien, los datos por acción.

Precio de Mercado por acción / Beneficio neto por acción = Ratio usado.

EV / Ventas:

Este ratio divide el valor de la empresa entre la cifra de ventas

Valor de la empresa / Ventas de la misma = Ratio usado

15. Apéndice II Modelado

·Criterios de evaluación:

- **Accuracy**: El *accuracy*, o precisión, es una métrica utilizada en problemas de clasificación para medir la proporción de **predicciones correctas realizadas** por un modelo **sobre** el **total de predicciones realizadas**.

  Se calcula dividiendo el número de predicciones correctas entre el número total de predicciones:

$$Accuracy = TP + TN/TP + TN + FN + FP$$

- **Recall**: El *recall*, o *sensitivity*, es un criterio de evaluación utilizado en clasificación para medir la **capacidad** de un modelo para **identificar todos los casos positivos**. En otras palabras, el recall mide la proporción de casos positivos que han sido correctamente identificados por el modelo respecto a los casos positivos actuales.

$$Sensitivity = recall = TP/TP + FN$$

- **Specificity**: El criterio de evaluación de especificidad es una medida que evalúa la capacidad de un modelo de **clasificación** para identificar correctamente las **muestras negativas**.

  Específicamente, la especificidad se define como la proporción de **muestras negativas actuales** que son **correctamente clasificadas** como negativas.

$$Specificity = TN/TN + FP$$

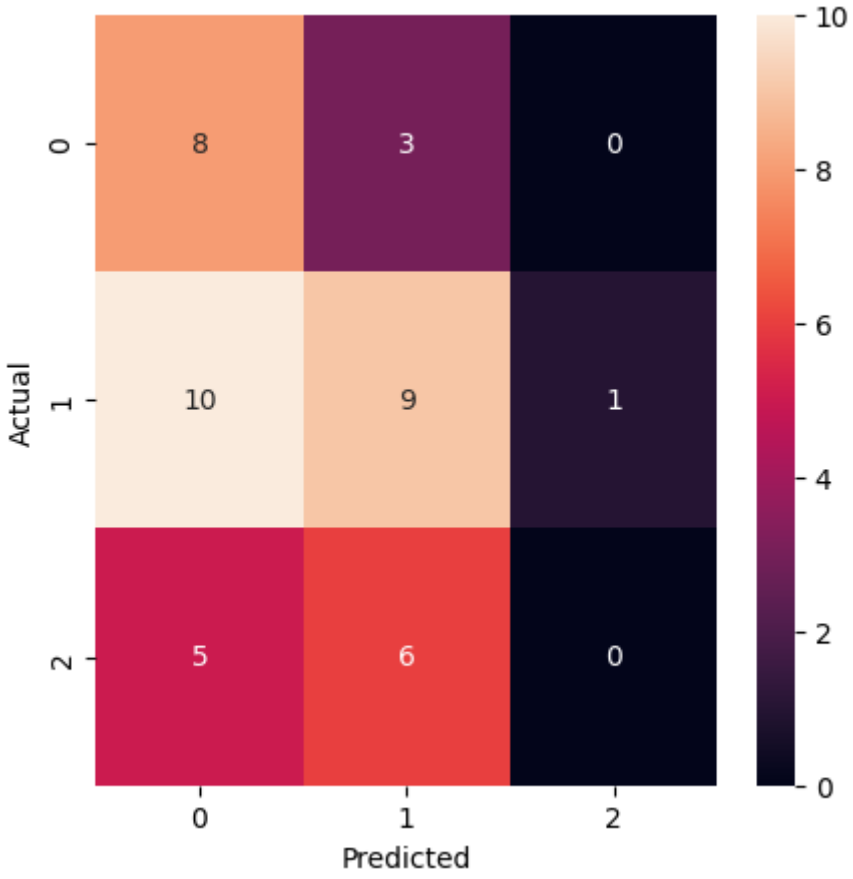
- **Precision**: La precisión (*precision* en inglés) es una medida de la exactitud de un modelo de clasificación. La precisión mide la **proporción de las predicciones clasificadas como positivas que también lo son realmente**.

  La precisión se calcula dividiendo el número de verdaderos positivos (VP) entre la suma de verdaderos positivos y falsos positivos (FP)

$$Precision = TP/TP + FP$$

·Matrices de confusión:

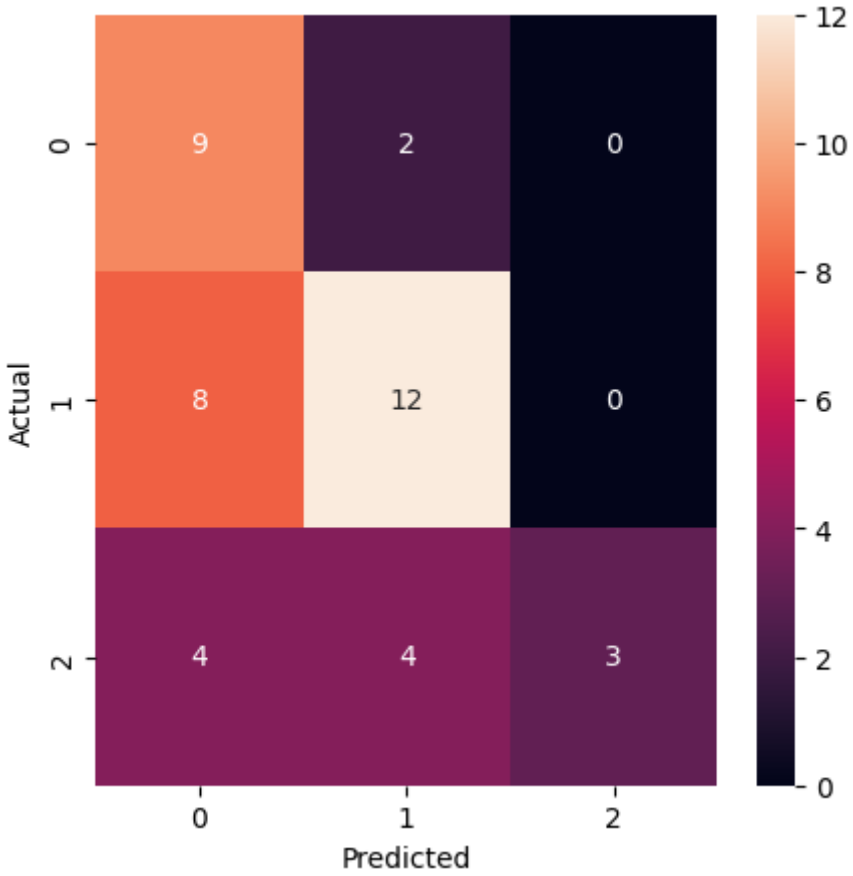
Stacking confusion matrix



i. Figura 14: Matriz de confusión de modelo Stacking

En esta matriz de confusión, la cual tiene un *accuracy* de **0.4** sobre los datos de *cross validation*, se puede ver como el modelo de *Stacking* muestra resultados no muy buenos, donde apenas realiza clasificaciones sobre la clase 2, lo cual **puede estar provocando fallos** en las predicciones de las otras dos clases. Por ello, se puede ver que a pesar de que gran parte de veces hace predicciones correctas, muchas otras falla en sus clasificaciones.

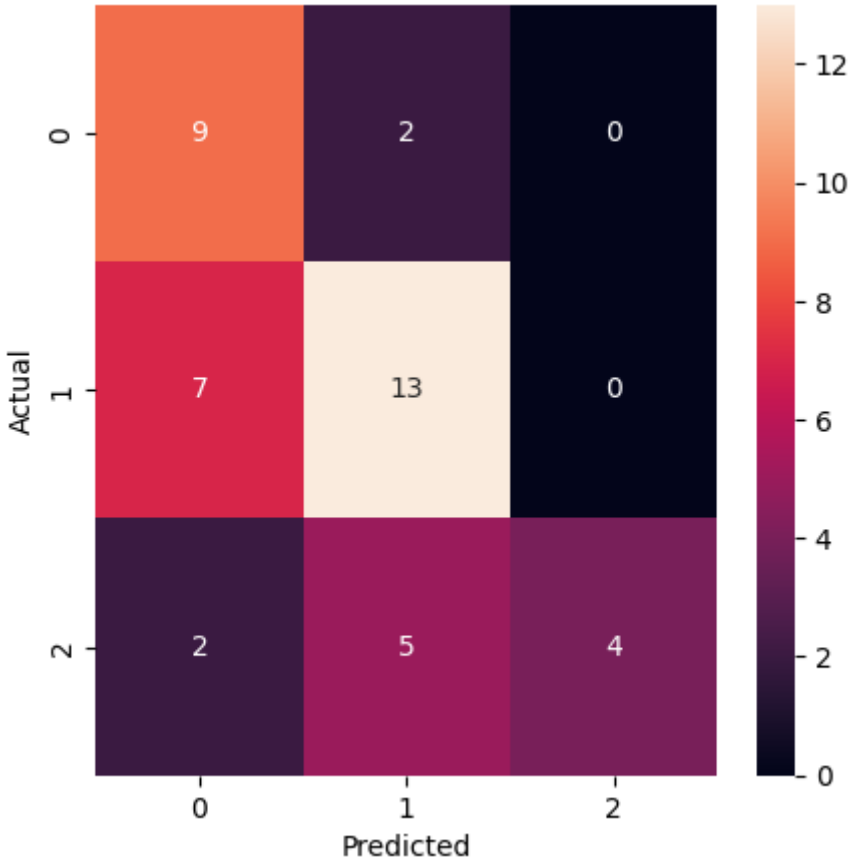
GridSearch Bagging confusion matrix



ii. Figura 15: Matriz de confusión de modelo Bagging con GridSearch

En esta matriz de confusión, la cual tiene un *accuracy* de **0.57** sobre los datos de *cross validation*, se puede ver como el modelo de *Bagging con GridSearch* muestra resultados algo mejores que el anterior modelo. Por ejemplo, de la clase 2, a pesar de solo clasificar 3 empresas, **acierta las tres**, mostrando una ligera contradicción en la clase 1. En cambio, en la clase 2, los resultados son bastante buenos, **donde acierta un gran porcentaje** de muchas predicciones sobre esa clase.

Optimización GridSearch Bagging Confusion matrix



iii. Figura 16: Matriz de confusión de modelo *Bagging* con *GridSearch* y optimización

En esta matriz de confusión, la cual tiene un *accuracy* de 0.6 sobre los datos de *cross validation*, se puede ver como el modelo de *Bagging con GridSearch y optimización* de hiperparámetros, muestra resultados aún mejores que el anterior modelo. Por ejemplo, de la clase 2, a pesar de solo clasificar 4 empresas, acierta las **cuatro**, mostrando una ligera contradicción en la clase 1. En cambio, en la clase 2, los resultados son bastante buenos, donde acierta un gran porcentaje de muchas predicciones sobre esa clase. Este modelo es **muy similar al modelo anterior pero optimizado**, el cual muestra una matriz de confusión casi idéntica, pero con **mejores resultados**.

·Explicación de modelos:

DESARROLLO MATEMÁTICO DEL ALGORITMO RANDOM FOREST:

El algoritmo *Random Forest* es un algoritmo de aprendizaje supervisado que combina las salidas de varios árboles de decisión con el objetivo de conseguir un resultado final. Dicho de otra manera, en la fase de entrenamiento, el algoritmo construye varios árboles de decisión individuales. Más tarde, se recogen las predicciones obtenidas por cada uno

de los árboles para así, combinando dichas predicciones, realizar una predicción final. Debido a que este algoritmo entrena varios modelos y combina sus resultados, forma parte de las técnicas del ***ensemble learning*** o aprendizaje conjunto.

Puesto que el *Random Forest* está compuesto por múltiples árboles de decisión, conviene analizar brevemente el funcionamiento de estos.

Los **árboles de decisión** son algoritmos de aprendizaje supervisado que sirven tanto para problemas de clasificación como para problemas de regresión. Este algoritmo tiene una estructura de árbol jerárquico que consiste de un nodo raíz, de ramas, de nodos internos y de nodo hojas.

Los árboles de decisión se basan en la división en sub-conjuntos de los datos, no obstante, es importante mencionar que no siempre se usa la misma técnica para dividir las ramas. En este caso, en los modelos de *Random Forest* que se han construido, se han utilizado tanto el criterio de *gini* como el criterio de la *entropía*. A continuación se observan las ecuaciones matemáticas de cada uno de los criterios.

a) Impureza de gini

$$GINI(V) = 1 - \sum_{i=1}^{nc} P(ci)^2$$

b) Entropía de la información

$$H(X) = -\sum P(X = x) \cdot log_2 \cdot (P(X = x))$$

Por un lado, la **impureza de gini** mide el grado de impureza de una variable e indica la probabilidad que tiene una variable de no ser clasificada correctamente si fuese elegida aleatoriamente. Por otro lado, la **entropía de la información**, mide la incertidumbre de las variables y es la suma de la probabilidad de que un valor ocurra multiplicado por el logaritmo de la probabilidad de que ese mismo valor ocurra en base dos. Asimismo, la impureza de gini será usada en algoritmos *CART* mientras que la entropía de la información se utilizará en algoritmos como el *C4.5*.

Finalmente, es interesante hablar sobre el método ***bagging*** ya que el *Random Forest* es una modificación o combina aspectos de este método. El *Bagging* es una técnica del *ensemble learning* y trabaja entrenando algoritmos simples paralelamente. Esta técnica trabaja especialmente bien reduciendo la varianza y evitando el sobreajuste.