



# Clasificación Automática de Galaxias en base a su Espectro - MEMORIA -

Arvin Daswani Daswani  
Pedro Martín Gómez  
Silvia Peraza Delgado  
Jose Luis Quintero García  
Vicente Tetuani Sánchez

## Índice

<b>Objeto</b>	<b>4</b>
<b>Definición del problema</b>	<b>5</b>
La investigación astrofísica	5
El Instituto de Astrofísica de Canarias	5
Qué es una galaxia	7
Espectroscopia de galaxias	8
Clasificación	9
La base de datos SDSS	12
Entendiendo el problema	13
Cómo actúan ahora los astrofísicos	14
<b>¿Quiénes somos?</b>	<b>15</b>
Galassify como solución al problema	15
El futuro de nuestra empresa	17
<b>Investigación y toma de datos</b>	<b>18</b>
Hipótesis	18
Identificación y validación	18
Priorización	24
Entrevistas	25
<b>Análisis y diagnóstico</b>	<b>26</b>
Análisis de la competencia	27
Análisis DAFO	28
<b>Plan estratégico</b>	<b>30</b>
Modelo de negocio	30
<b>Plan de acción</b>	<b>35</b>
Alcance	35
Misión, visión y objetivos	36
Métricas	37
Diferencia de costes entre selección manual y automatizada	37
Papers emitidos y de referencia	37
Subvenciones conseguidas	38
Contactos científicos realizados	39
Satisfacción de Clientes	39
Análisis de actividades y tareas	40

Mapa de procesos	40
Solución tecnológica	43
Análisis de los recursos	45
Talento humano	45
Recursos físicos	46
Gestión del tiempo	48
<b>Optimización de los resultados</b>	<b>49</b>
Claves	49
Cuenta de resultados	50
Ingresos	50
Gastos	52
Cuenta de pérdidas y ganancias	53
Flujo de caja	55
Balance	56
Indicadores económicos	57
<b>Bibliografía</b>	<b>58</b>
<b>Anexos</b>	<b>60</b>

## 1. Objeto

El objeto de este proyecto es el de resolver el problema que se les presenta a los astrofísicos al elegir objetos de estudio (galaxias en este caso) con unas características definidas.

El desarrollo tecnológico en los sensores instalados tanto en observatorios como en satélites, permiten obtener cada vez más cantidad de valiosísima información a estudiar. Cada uno de estos equipos no es más que un generador de datos de forma que, cada jornada, se obtienen varios terabytes de datos adicionales por cada uno de ellos.

Esta información se debe recopilar, almacenar, filtrar, transformar, analizar y visualizar antes de proceder al estudio del objeto en sí.

Sin embargo, tal cantidad de información es a la vez un problema, dado que resulta extraordinariamente complicado (por no decir imposible) seleccionar la mejor información de entre toda la disponible para cada estudio que se plantee.

En estos momentos, esta información se escoge a mano de entre todos los datasets disponibles, con lo que es improbable que dicha información sea la óptima.

Galassify viene a resolver una parte de este problema de forma eficaz y eficiente, empleando técnicas de Machine Learning y Deep Learning para clasificar de forma automática y ponderada las galaxias en función de sus espectros y a partir de unas especificaciones definidas con anterioridad. El objetivo último es la identificación de outliers, galaxias que por sus características no siguen las normas habituales respecto al resto de galaxias y cuya investigación aporta muchísimo más conocimiento.

Esto supondrá una optimización considerable del tiempo de los investigadores, así como una enorme mejora en los resultados de sus estudios, al elegirse las galaxias más adecuadas para su análisis en profundidad.

## 2. Definición del problema

Para entender el problema, primero vamos a contextualizar la situación de partida explicando qué es la astrofísica, a qué se dedica el IAC, qué es una galaxia, de qué está hecha, que es un espectro y cómo los diferentes componentes de una galaxia dejan su huella en el espectro.

A partir de ahí, explicaremos más en detalle el problema y cómo lo resuelven los investigadores a día de hoy.

### 2.1. La investigación astrofísica

La astrofísica es el desarrollo y estudio de la física aplicada a la astronomía. Es una parte moderna de la astronomía que estudia los astros como cuerpos de la física estudiando su composición, estructura y evolución.

La astrofísica emplea la física para explicar las propiedades y fenómenos de los cuerpos estelares a través de sus leyes, fórmulas y magnitudes. Si bien se usó originalmente para denominar la parte teórica de dicho estudio, la necesidad de dar explicación física a las observaciones astronómicas ha llevado a que los términos astronomía y astrofísica sean usados de forma equivalente.

Actualmente existen algo más de 10.000 astrofísicos que toman datos para sus investigaciones proporcionados por los más de 650 observatorios existentes en el planeta entre los que se encuentra el Instituto de Astrofísica de Canarias (IAC).

También hacen uso de centros de datos como el SDSS, creador de los mapas tridimensionales del universo con mayor detalle hechos hasta la fecha. Dichos mapas incluyen imágenes multicolor de un tercio del cielo y los espectros de más de tres millones de objetos astronómicos. Se trata, sin duda, de una fuente de conocimiento de alto valor para los investigadores.

### 2.2. El Instituto de Astrofísica de Canarias

El Instituto de Astrofísica de Canarias, también conocido por sus siglas IAC, es un centro de investigación español internacionalizado y seleccionado por el Gobierno español como "Centro de Excelencia Severo Ochoa". Está integrado por:

- ★ Su sede central en La Laguna (Tenerife).

- ★ El Centro de Astrofísica en La Palma (CALP).
- ★ El Observatorio del Teide (OT) en Izaña (Tenerife).
- ★ El Observatorio del Roque de los Muchachos (ORM), en Garafía (La Palma).

Los fines del IAC son la investigación astrofísica, el desarrollo de instrumentación científica ligada a la astronomía, la formación de personal investigador, la administración del Observatorio del Teide y del Observatorio del Roque de los Muchachos y la divulgación de la ciencia.

El IAC tiene su sede central en La Laguna (Tenerife), que es el lugar de trabajo habitual de la mayor parte de su personal científico, tecnológico y de soporte. También en dicha sede se concentran las instalaciones para desarrollar instrumentación científica. El IAC cuenta además con otra sede, el Centro de Astrofísica en La Palma (CALP).

La excepcional calidad del cielo de Canarias para la observación astronómica está protegida por ley. El IAC dispone de una Oficina Técnica para la Protección de la Calidad del Cielo (OTPC) que vigila la aplicación permanente de esta ley. También dispone de un grupo científico que se ocupa de hacer un seguimiento continuo de los parámetros que determinan la calidad astronómica de los observatorios del IAC (Grupo de Calidad del Cielo).

El Programa de Investigación del IAC comprende proyectos tanto de investigación astrofísica como de desarrollo tecnológico.

Entre las actividades del IAC se encuentra también la formación de investigadores, la enseñanza universitaria y la difusión cultural.

El IAC ha dedicado un gran esfuerzo al desarrollo tecnológico para el diseño y construcción de un gran telescopio de 10,4 metros de diámetro (Gran Telescopio CANARIAS, GTC), que está situado en el Observatorio del Roque de los Muchachos.

La Oficina de Transferencia de Resultados de Investigación (OTRI) del IAC, creada bajo el Plan Nacional de I+D, ha sido una de las pioneras dentro de la Red Nacional de oficinas OTRI.

## 2.3. Qué es una galaxia

Una galaxia es un sistema auto-gravitacional (es decir, cuyas propiedades son sólo el resultado de la interacción gravitacional) que consiste en estrellas (entre 10 millones y 100.000 millones), gas y polvo.



Galaxia de la Vía Láctea

En la foto podemos distinguir claramente estos tres elementos:

- ★ **Estrellas:** a simple vista, la Vía Láctea parece una nube blanquecina, pero en realidad está formada por una gran cantidad de estrellas, las cuales tienen masas y temperaturas variadas.
- ★ **Gas:** el gas contenido en la galaxia (esencialmente hidrógeno) se presenta en varios estados. La mayor parte del gas se encuentra en forma neutra, el gas HI. También hay grandes nubes frías de hidrógeno molecular H<sub>2</sub>, dentro de las cuales nacen las estrellas por colapso gravitacional. Después de su nacimiento, las estrellas de mayor masa ionizan el gas circundante. Forma entonces nubes de gas ionizado llamadas regiones “HII”, o nebulosas en emisión. En la fotografía, las nebulosas de emisión son claramente reconocibles por su tono rojo-rosado debido al color de la línea de emisión de hidrógeno H<sub>α</sub> a 656.3 nm.
- ★ **Polvo:** las galaxias también contienen polvo que las estrellas formaron durante su vida y fueron expulsadas al medio interestelar. Estas partículas de polvo tienen la propiedad de absorber la luz emitida por las estrellas, al igual que el polvo suspendido en el aire absorbe la luz del sol. Así, desde la Tierra solo podemos observar (en el campo de la luz visible) una pequeña parte de nuestra Galaxia. En la fotografía de la Vía Láctea de arriba, las regiones negras no son áreas vacías de estrellas y gases, sino áreas donde el polvo ha absorbido casi toda la luz del

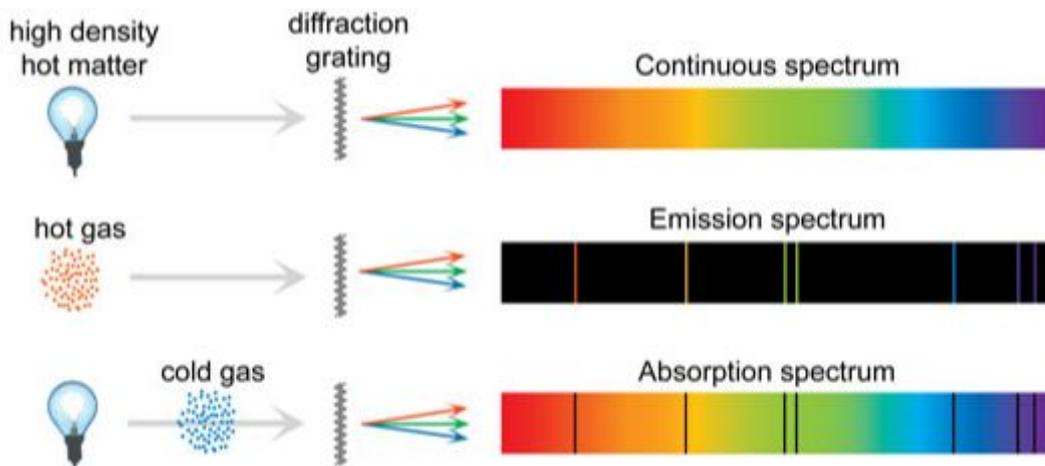
fondo. La absorción de polvo es un efecto cromático: el polvo absorbe mucha más luz azul que la luz roja.

Cada uno de estos componentes tiene su propia firma espectral. La herramienta fundamental de los astrónomos extragalácticos es el análisis de los espectros de galaxias a partir de los cuales los astrónomos pueden medir y estimar las siguientes propiedades:

- ★ La edad de las poblaciones de estrellas.
- ★ La composición química del gas.
- ★ La cantidad de polvo.
- ★ Cinemática interna (movimientos de estrellas y gas).
- ★ La distancia (desplazamiento rojo o redshift).

### 2.3.1. Espectroscopia de galaxias

Existen tres clases diferentes de espectros:



Tipos de espectro

- ★ **Espectro de emisión:** si consideramos una nube de hidrógeno caliente y tenue: los átomos chocan entre sí y la energía de las colisiones puede transferir un electrón a un nivel de energía más alto. Al desenergizarse, el electrón emite uno o más fotones de longitud de onda específicos para el elemento químico del cual se deriva, la mayor parte del tiempo hidrógeno. En el dominio visible, el resultado será un espectro llamado espectro de emisión, porque muestra líneas de emisión.
- ★ **Espectro térmico o cuerpo negro:** a diferencia de una débil nube, un cuerpo negro tiene una alta densidad. Antes de abandonar el cuerpo negro, los fotones

creados por los átomos sufren múltiples colisiones con ellos. Esto tiene el efecto de redistribuir los fotones ("termalización") en todas las longitudes de onda, dando así la distribución del cuerpo negro. Al analizar el espectro de un cuerpo negro obtenemos una banda continua de luz de rojo a violeta: es un espectro continuo. Todos los objetos en equilibrio térmico, es decir, caracterizados por una temperatura limpia (como las estrellas), tienen una emisión cercana a la de un cuerpo negro.

- ★ **Espectro de absorción:** un espectro de absorción se forma cuando un haz de luz continuo pasa a través de una nube delgada y fría. Solo los fotones correspondientes a las transiciones permitidas son absorbidos por los átomos de la nube y luego reemitidos en varias direcciones. Se obtiene un espectro continuo (cuerpo negro) con líneas negras correspondientes a las absorciones. El espectro de la luz solar es un ejemplo de espectro de absorción. Las líneas de absorción se forman cuando la luz del sol pasa por las capas delgadas de la atmósfera solar. De este modo podemos determinar la composición química de la atmósfera solar.

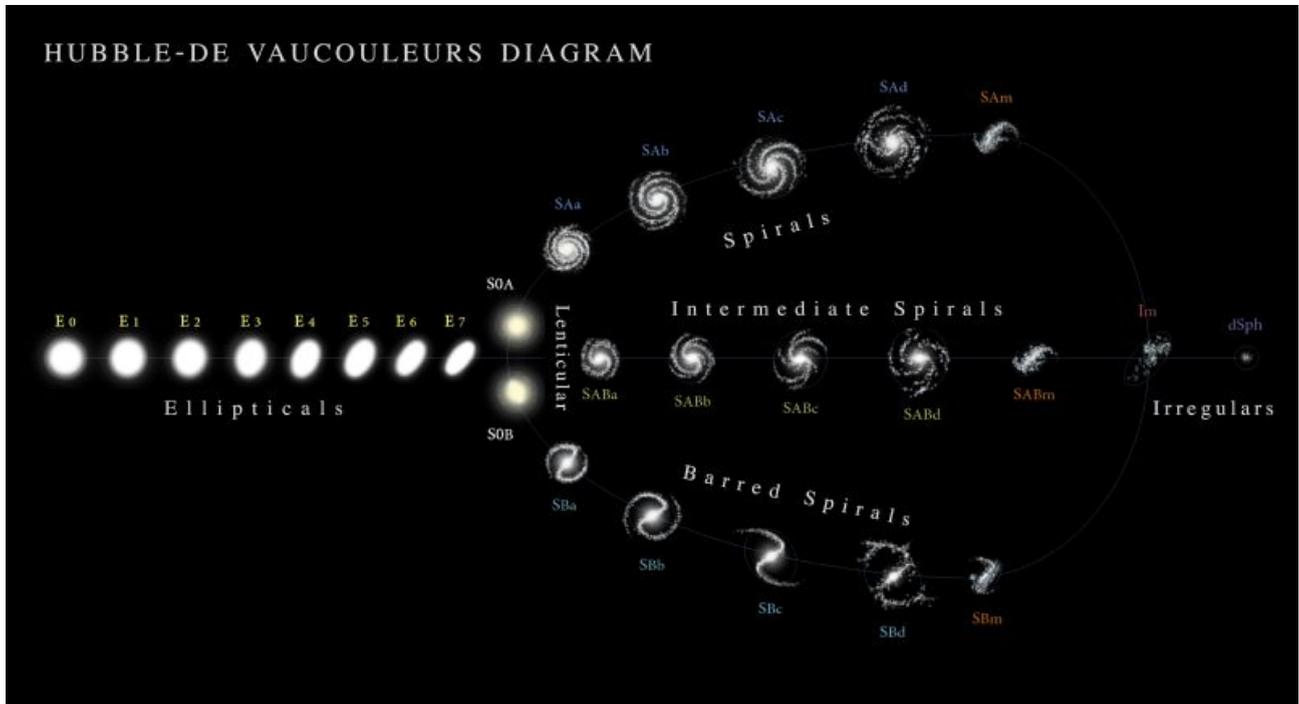
### 2.3.2. Clasificación

A día de hoy existen dos grandes sistemas de clasificación de galaxias:

#### ★ **Clasificación morfológica**

La forma de una galaxia nos da información sobre la misma, como por ejemplo su momento angular o la cantidad de materia.

Esta clasificación morfológica está estandarizada y es ampliamente utilizada (la inició Edwin Hubble en 1926).



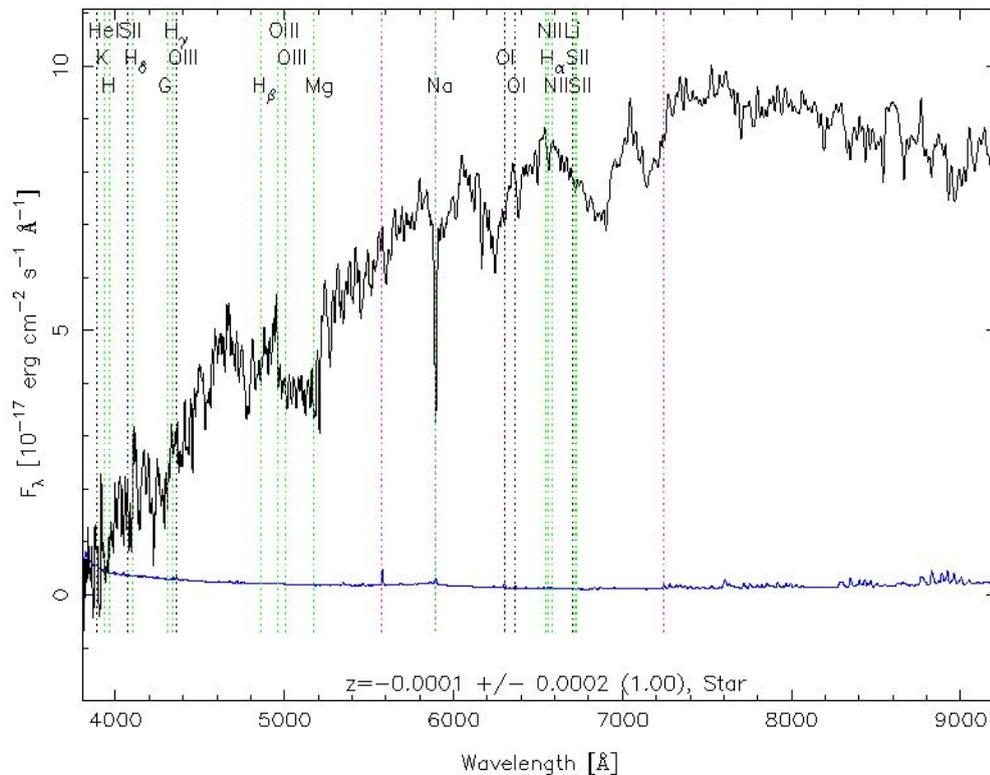
Clasificación morfológica de galaxias

### ★ Clasificación por espectros

Cada objeto celeste emite una radiación de energía que podemos recoger en un espectro. Este no es más que un histograma donde se recoge en un gráfico la cantidad de energía recibida para cada valor de longitud de onda en un intervalo concreto.

Para los efectos, vamos a tratar espectros de luz visible, esto es, espectros con longitud de onda comprendidos entre los 3800 y los 9200 Angstroms.

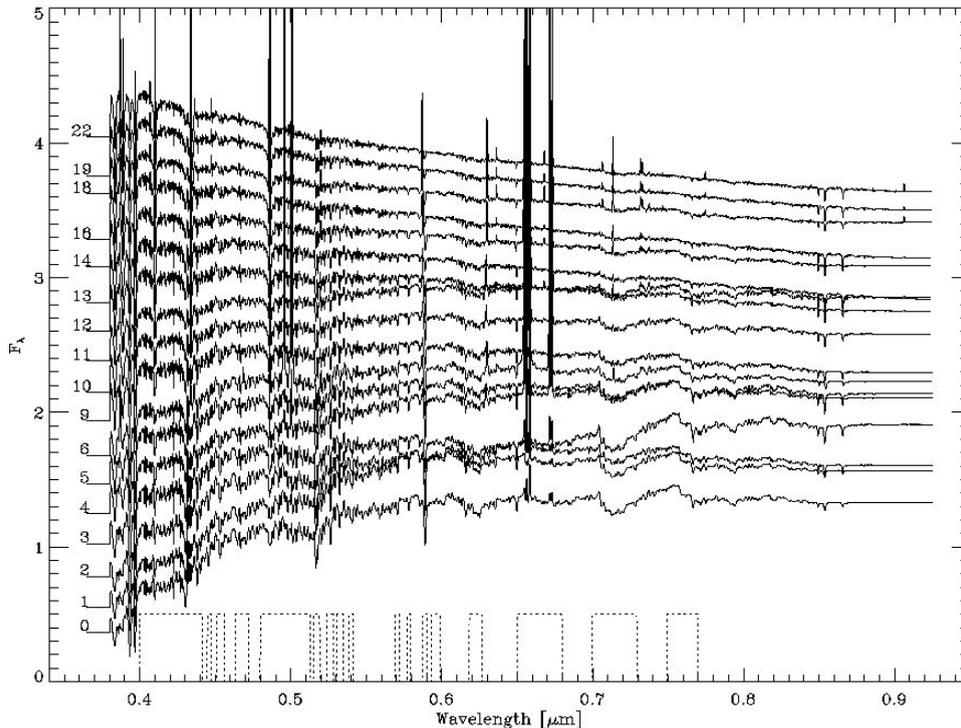
RA=163.02668, DEC= 0.90938, MJD=51909, Plate= 276, Fiber=402



Ejemplo de un espectro

A partir del espectro podemos conocer mucho sobre la clasificación del objeto emisor. Concretamente, nosotros nos vamos a centrar en galaxias (evidentemente, el espectro recibido es el del promedio de todos los objetos que haya dentro de esa galaxia; no conoceremos datos concretos de una estrella de esa galaxia, sino del conjunto completo). Y el espectro de una galaxia nos dará información valiosa sobre la composición promedio de la misma y, por tanto, también sobre su edad.

En estos momentos, no existe una clasificación normalizada de galaxias por su espectro.



Clasificación por espectros según ASK

Aquí vemos algunos de los espectro tipo clasificados según el sistema ASK empleado por el Dr. Sánchez Almeida.

## 2.4. La base de datos SDSS

La Sloan Digital Sky Survey (SDSS) es una base de datos pública para investigadores y estudiantes que recoge datos de objetos celestes del cielo norte (10.000 grados cuadrados) generados por telescopios terrestres.

La SDSS es un proyecto de inventario del universo y comenzó su labor en el año 2000, después de décadas de planificación y construcción. Ha tenido varias fases:

- ★ SDSS-I, entre 2000 y 2005
- ★ SDSS-II, entre 2005 y 2008
- ★ SDSS-II, entre 2008 y 2014
- ★ y SDSS-IV, desde 2014 a la actualidad

En los 5 primeros años recogió datos de 200 millones de objetos en 5 bandas ópticas. Cuando se complete el catálogo, la información en bruto serán unos 25 TB y la información procesada unos 13TB.

Los datos recogidos por los telescopios son procesados y cada imagen genera 400 atributos junto con una imagen recortada de 5 colores. Estos atributos pueden ser espectros calibrados, desplazamientos respecto al rojo, líneas de absorción y emisión, y muchos otros.

En el momento de la elaboración de este documento, la versión publicada más reciente es la DR15 (Data Release 15) que incluye 6 tipos de información:

- ★ Imágenes (Photo)
- ★ Espectros ópticos (Spectro)
- ★ Espectros de infrarrojos (Apogee)
- ★ Espectros IFU (MaNGA)
- ★ Librería de espectros estelares (MaStar)
- ★ Catálogo de otros datos como magnitudes o desplazamientos de rojo

En este proyecto utilizaremos solo las tablas relativas a espectro visible y la de desplazamiento al rojo, necesaria para corrección previa del espectro.

## 2.5. Entendiendo el problema

Como ya hemos comentado, la astrofísica es un campo científico con un uso intensivo de datos. En los últimos años se han ido desarrollando nuevos proyectos con los que se obtienen cada vez más información. Sin ir más lejos, en el SDSS hay catalogados más de 1.231 millones de objetos y el programa EUCLID, de la Agencia Espacial Europea, será incluso mayor.

Esto hace que cada vez sea más inviable el tratamiento manual de estos datasets para investigación de objetos concretos.

No obstante, aún no existe una estrategia definida para el tratamiento de estos datos, lo que abre un campo de acción interesantísimo para terceras empresas expertas en Big Data.

En el caso concreto que nos ocupa, nos centramos en el Instituto de Astrofísica de Canarias (IAC) y de investigaciones a realizar en base a datasets como el SDSS:

- ★ En el IAC no existe una estructura transversal definida para el tratamiento de los datos. Cada científico (o equipo creado al efecto de una determinada investigación) se ocupa de esta parte del trabajo.

- ★ Existe un germen de este hipotético departamento transversal, pero de momento sólo realiza labores de asesoramiento genérico y no hay planes concretos de desarrollo.
- ★ Los científicos compiten entre sí para obtener financiación para sus proyectos, de forma que existe una ocasional puesta en común de proyectos, incluyendo el tratamiento de los datos.
- ★ Por otro lado, las técnicas de Machine Learning avanzan a diario con nuevos desarrollos sobre los que muchos científicos no quieren / pueden estar permanentemente actualizándose. Prefieren dedicar su tiempo a labores en las que puedan aportar su valor, como analizar objetos interesantes que aparezcan tras el cribado de los datos, por lo que la aparición servicio de especialización en materia de big data puede ser una necesidad a corto plazo.

## 2.6. Cómo actúan ahora los astrofísicos

En estos momentos, los científicos parten para sus estudios de una selección manual. Para realizar esta selección se basan principalmente en su experiencia e intuición (ojo entrenado), utilizando como punto de partida estudios previos sobre clasificación de galaxias.

Sin embargo esta tarea, aunque importante, es bastante tediosa y los científicos prefieren dedicar su tiempo y experiencia en realizar los estudios que se deriven del análisis de las galaxias menos comunes. Estas galaxias de espectros diferentes (también llamadas outliers) son las que más información proporcionan, por lo que sería de gran ayuda para ellos un sistema que realizara esta diferenciación de manera rápida y fiable.

Si bien es cierto que la comunidad científica tiene conciencia de las ventajas de la aplicación de Machine Learning a sus estudios, sólo algunos científicos invierten parte del tiempo de sus investigaciones a la adquisición de este tipo de conocimientos y normalmente de forma muy específica para resolver únicamente el problema que les ocupa en cada investigación concreta.

### 3. ¿Quiénes somos?



**Galassify S.L.** es una empresa de consultoría en Big Data especializada en el ámbito científico, más concretamente en el campo de la astrofísica. Nos dedicamos principalmente al diseño e implementación de soluciones a medida para que nuestros clientes puedan obtener el máximo rendimiento de sus datos.

La empresa debe su nombre a nuestro primer proyecto, también denominado Galassify, que viene de las palabras Galaxy y Classify.

Galaxy + Classify = **Galassify**

#### 3.1. Galassify como solución al problema

La solución al problema propuesto en el apartado anterior podría consistir en utilizar un modelo de Machine Learning para la clasificación automática de galaxias a partir de su espectro.

Sería un modelo de aprendizaje no supervisado, dado que no tenemos un dataset ya clasificado. En todo caso, podría usarse un algoritmo semi-supervisado, realizando una clasificación manual de una parte pequeña del dataset.

El objetivo es aplicar el modelo sobre el dataset SDSS en su versión más reciente DR15, si bien tendremos en cuenta el volumen de los datos a tratar, de modo que la podamos trabajar en un primera entrega sobre un subconjunto de los datos.

Del modelo deben resultar una serie de clusters (en número a determinar) con los que clasificar las galaxias a través de su espectro.

Esta clasificación permitiría escoger a los científicos aquellas galaxias que presenten características más raras, bien porque pertenezcan a un cluster concreto o porque estén muy alejadas del centroide del cluster (aunque no se use K-means, éstas galaxias raras presentarán una clasificación en su cluster más difusa que la gran mayoría de las galaxias de ese cluster).

No se pretende obtener un sistema que valga para siempre ni siquiera para próximas releases de SDSS, sino tener un marco de trabajo actualizado y que permita una selección adecuada, quizá basada en aspectos concretos del espectro.

En suma, **Galassify** emitirá un listado ordenado de galaxias con unas características concretas definidas por el equipo de astrofísicos, reportando los siguientes beneficios:



**Ahorro de tiempo:** tiempo no dedicado a clasificación manual, sino aprovechado en tareas de mayor valor.

**Cribado óptimo** de forma que la relación de galaxias obtenidas serían las más adecuadas para el estudio en curso.

**Mejores resultados** al elegir para su estudio los ejemplos más significativos.

**Flexibilidad**, dado que el algoritmo podría modificarse con relativa sencillez para ajustarse a características de selección diferentes y, por tanto, para otro tipo de estudios potenciales.

**Escalabilidad**, ya que el algoritmo está preparado para trabajar con la DR15 del SDSS, pero podría adaptarse con relativa sencillez a próximas versiones.

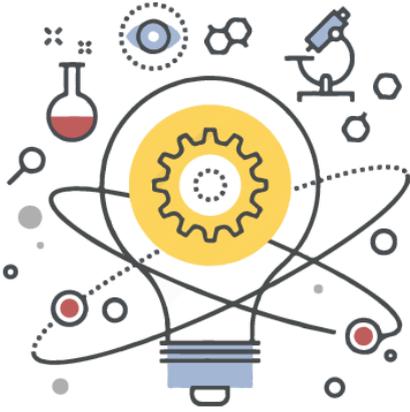
### 3.2. El futuro de nuestra empresa

Si bien Galassify se trata del primer proyecto dentro de nuestro portfolio, pretendemos que se convierta en la punta de lanza para la creación de una división especializada en astrofísica, dado el gran potencial económico que tienen los proyectos de Big Data en esta rama de la ciencia.

Gracias a Galassify podremos:

- ★ Evaluar el tamaño de este mercado con mayor precisión.
- ★ Conocer mucho mejor el funcionamiento del mismo, sus problemas y las palancas a desarrollar para poder actuar en él con eficacia.
- ★ Optimizar la forma en la que podríamos llegar a actuar en el futuro en este mercado (vías de selección de oportunidades, contacto con clientes potenciales, criterios de aprobación de proyectos, consideración de elementos de valor de los científicos y de los órganos de decisión, etc.).
- ★ Crear una imagen de marca en este mercado.

## 4. Investigación y toma de datos



Desde abril hasta julio de 2019 hemos llevado a cabo una serie de acciones con la finalidad de conocer en detalle el funcionamiento del IAC y comprender cuáles son sus principales necesidades en materia de Big Data y Business Intelligence.

La principal fuente información proviene de las entrevistas realizadas a distintos perfiles del IAC: dos astrofísicos dedicados a la investigación, un astrofísico experto en Machine Learning y el gerente de informática.

Por otro lado, nos hemos documentado leyendo también algunos papers publicados con aspectos de interés para nuestro estudio.

### 4.1. Hipótesis

A continuación planteamos las principales hipótesis del proyecto que determinarán la viabilidad de nuestra idea de negocio.

#### 4.1.1. Identificación y validación

##### **Hipótesis 1**

*Creemos que nuestros algoritmos podrán clasificar en función del espectro y sin ambigüedad (es decir, que no exista una separación clara entre los distintos clusters, con probabilidades de pertenencia en torno al  $50\% \pm 10\%$ ) al menos el 90% de las galaxias observadas.*

Durante las entrevistas, todos los investigadores han afirmado que el problema de la clasificación por espectros se ha podido resolver mediante algoritmos de ML. Como ejemplo nos han enseñado el estudio realizado por Dalya Baron usando un Random Forest, donde clasifica las galaxias y detecta aquellas más raras.

Así pues, queda patente que el uso de técnicas de ML para la clasificación y detección dan resultados positivos, por lo que estudiar nuevos algoritmos como DL pueden conseguir mejores resultados.

## **Hipótesis 2**

*Creemos que nuestro cliente tiene una base de datos demasiado grande como para extraer los datos más valiosos con eficacia de forma manual y no desea dedicar más de 40 horas por proyecto a esta selección. Además, el actual sistema deja en torno a un 75% de objetos valiosos sin evaluar.*

*Creemos que además, nuestro cliente aprecia los beneficios de poder ajustar el algoritmo de clasificación para detectar galaxias con otras características diferentes de las planteadas inicialmente, lo que podrá ser útil en otros proyectos.*

Esta necesidad podría resolverse con un sistema automático de clasificación que le entregue a nuestro cliente un listado ordenado de las galaxias más raras con unas especificaciones definidas.

En las entrevistas nº 2, 3 y 4, los 3 astrofísicos y el responsable de informática consultados nos han confirmado nuestra hipótesis, dado que los científicos necesitan de forma inevitable el uso de herramientas automáticas de selección de aquellos objetos que desean estudiar. Es más, el número de objetos que contiene el SDSS es enorme, pero las nuevas misiones espaciales, como EUCLID, implicarán bases de datos aún más grandes.

Lo sorprendente de este tipo de misiones es que conllevan un enorme desarrollo tecnológico, tanto en satélites como, sobre todo, en instrumentos de medida. Sin embargo, no llevan aparejado un programa de tratamiento de los datos que estos equipos generen.

Por otra parte, aunque hay una cierta incorporación del big data en los proyectos de investigación, no existe una organización definida al respecto. Por ejemplo, en el IAC se está creando un grupo de apoyo para asesorar y discutir tratamientos de los datos con big data, pero no hay una estructura oficial y la que hay no lo utiliza de forma sistemática.

En estos momentos, la selección de galaxias para su estudio es manual, basándose en la visualización de los espectros y deteniéndose en aquellos que, por la experiencia (ojo entrenado) de los científicos, resulta más llamativa. Ahora hay que confirmar si la galaxia en la que nos hemos detenido es o no interesante para un mayor estudio. Si no lo es, seguimos la búsqueda; si lo es, procedemos a su estudio. Esto deja indudablemente un enorme número de candidatos pendientes y la elección de ejemplares que quizá no sean los más interesantes de todos.

Por otro lado, el IAC favorece el uso de empresas externas locales (dentro de su programa de responsabilidad social corporativa), significando esto una mayor puntuación a la hora de valorar los proyectos presentados por los científicos y suponiendo, por tanto, una mayor probabilidad de que éstos se aprueben o solicitar subvenciones de mayor cuantía.

Todo esto supone que los científicos apreciarán como muy valiosa nuestra contribución, tanto por el valor del listado, como por el tiempo ahorrado y la relativa facilidad para financiar nuestro servicio. Los científicos estarán dispuestos a realizar ellos mismos un desarrollo automático, siempre que éste no les lleve demasiado tiempo (lo que podrían ser las 40 horas de nuestra hipótesis).

Por otra parte, lo que también comprobamos en las entrevistas es que estas empresas externas deberán contar con astrofísicos en su organización, dado que esto facilitaría mucho la comunicación con los científicos, sin que éstos deban explicar matices que a los legos en la materia se nos escaparían y sin el esfuerzo y tiempo que explicar todo esto supone.

Por tanto, si nuestra empresa quiere tener una división especializada en astrofísica para futuros proyectos, deberá contar con un astrofísico, bien en plantilla o bien como colaborador habitual subcontratado.

Para este caso concreto, en la entrevista 2 tenemos el deseo expreso de Ana Monreal y Jorge Sánchez de tener un listado ordenado de outliers tras una clasificación del SDSS en su DR15. Las futuras ampliaciones o modificaciones de este listado se verían al estudiar el resultado de este proyecto.

### **Hipótesis 3**

*Creemos que nuestro cliente usará Galassify porque prefiere dedicar su tiempo a la investigación en vez de en trabajos que percibe de menor valor, como lo es la selección manual de galaxias dentro de las posibilidades a su alcance. Por esto, preferirá pagar a un servicio externo hasta una cantidad de 51.000€ por este trabajo.*

En las entrevistas realizadas con varios investigadores del IAC hemos descubierto que existen dos tipos de astrofísicos bien diferenciados:

- ★ Los que viven por y para la investigación y el estudio que estén desarrollando en el momento y consideran que toda tarea de preparación, de selección o cualquier

otra índole que se distancie de sus objetivos es una pérdida de tiempo, por mucho que estas les ayuden a alcanzarlos.

- ★ Aquellos que son más inquietos y no les importa realizar desarrollos informáticos paralelos a sus investigaciones aunque esto les haga invertir más tiempo en entender, aprender y comprender los algoritmos.

En la actualidad hay muchos más astrofísicos en el primer grupo que en el segundo. A pesar de que cada vez son más los que prefieren desarrollar sus propias soluciones, siguen siendo una minoría. Ello se demuestra en que el grupo de Big Data que se ha creado dentro del IAC es aún muy pequeño, y no son más de 10 investigadores, de un total de casi 200 ([Personal del IAC](#)), los que acuden con frecuencia a las sesiones y de estos una amplia mayoría únicamente va por curiosidad pero no han llegado a aplicar nada en sus estudios. De esto podemos concluir que apenas el 5% de los investigadores están interesados en las técnicas de Machine Learning y Deep Learning que vamos a aplicar.

Por otra parte, más de la mitad de los entrevistados nos ha afirmado que creen que las técnicas de Big Data son muy capaces de ayudarles en sus investigaciones, pero no desean dedicar tiempo a aprender lo necesario para aplicarlas. De hecho, uno de los astrofísicos nos indicó que en una ocasión tuvo la iniciativa de realizar un desarrollo por sí mismo pero finalmente se dio cuenta de que necesitaba dedicarle mucho tiempo y esfuerzo para poder obtener resultados aprovechables.

El presupuesto que se dedica a la investigación es muy importante y las cantidades que se suelen ofrecer en las subvenciones son elevadas. Por ejemplo, en una única convocatoria de proyectos el Ministerio de Ciencia, Innovación y Universidades participa con 1.000.000€ y el Ministerio de Economía y Empresa con 5.000.000€. Los proyectos que optan a estas subvenciones realizan estudios económicos para determinar cuánto dinero necesitan y dado que se valora el desarrollo del tejido empresarial local no es de extrañar que en muchos proyectos se cuente con la contratación de terceros para realizar desarrollos informáticos que sirvan de apoyo al objetivo de la investigación.

## **Hipótesis 4**

*Creemos que conseguiremos a nuestros clientes a través de:*

- ★ *Listados de centros de investigación pública y privada en astrofísica base (no aplicada).*
- ★ *Networking y asistencia eventos de astrofísica.*
- ★ *Elaborando papers con los avances realizados y apareciendo en las investigaciones en las que se utilice nuestra herramienta para la generación de prestigio.*

Nuestros interlocutores nos han comentado que hoy por hoy no hay empresas que ofrezcan servicios de consultoría de ML especializada para astrofísica. Sin embargo nos cuentan que los investigadores demandan cada vez más este tipo de conocimientos para poder llevar a cabo sus investigaciones, lo que consiguen principalmente preguntando a otros investigadores (en reuniones informales) o buscando códigos en Github, ya que la primera comunidad de ML entre científicos aún están empezando a crearse.

Si logramos formar parte de esos grupos, publicamos nuestros estudios y compartimos parte de nuestros códigos, llegaremos fácilmente a los investigadores y podremos convertirnos en un referente.

## **Hipótesis 5**

*Creemos que nuestra principal fuente de ingresos será la venta del servicio de creación del listado ordenado de outliers en la clasificación de galaxias, bien sobre especificaciones definidas en el catálogo o bien mediante diseños a medida (de mayor valor).*

A través de las entrevistas realizadas, hemos comprobado que los científicos tienen un sistema de financiación que se basa en diversos programas que dotan con fondos locales, nacionales o internacionales. Los científicos preparan un dossier con su programa de estudio y que debe ser aprobado por el IAC y luego por el organismo competente en función de los fondos solicitados.

Se da la circunstancia de que si en este dossier se incorpora la contratación de empresas locales, esto implica una mayor probabilidad de aprobación del dossier.

Galassify es un programa especial dentro del portfolio de operaciones de nuestra empresa, una solución a medida para un programa específico de investigación del IAC.

Se tiene un equipo de científicos del IAC que desean investigar sobre la creación y evolución de las galaxias (se quiere determinar por qué en algunas galaxias se crean nuevas estrellas y otra serie de objetos y por qué otras desaparecen). Una de las formas de atacar este objetivo consiste en estudiar galaxias raras, aquellas con características no habituales y ver cómo estas se comportan y por qué.

Por tanto, el disponer de un listado ordenado de este tipo de galaxias resulta un adelanto muy importante en tiempo y calidad de estudio para ellos.

Por otro lado, el uso del SDSS para este tipo de estudio se ve muy beneficiado con el uso de técnicas de ML, pero aún siendo un volumen enorme de datos, puede ser relativamente manejable de forma manual en base a la experiencia de los investigadores. Sin embargo, los futuros programas espaciales serán absolutamente inmanejables sin técnicas de cribado automático. En estos programas el volumen de datos será mucho mayor y algoritmos semejantes a Galassify serán imprescindibles (es probable que el mismo algoritmo ya no sea válido en estos datasets con más variables).

De forma que se confirma que nuestro principal valor (y objeto de nuestra participación) será la entrega del listado considerado, siendo posible ampliar la cartera de proyectos futuros en líneas parecidas dentro del IAC o de otros Institutos de Investigación astrofísica internacionales. Nuestra participación en Galassify podría ser valiosísima en esos futuros proyectos.

## **Hipótesis 6**

*Creemos que nuestros clientes serán los equipos de astrofísicos de los distintos centros de investigación de astrofísica base del mundo.*

A raíz de las entrevistas que hemos mantenido con los científicos hemos determinado que los centros de investigación de astrofísica realizan investigación de base, no aplicada. El objetivo principal de los investigadores que se dedican al estudio de la astrofísica es la recopilación de información para construir una base de conocimiento que se va agregando a la información previa existente.

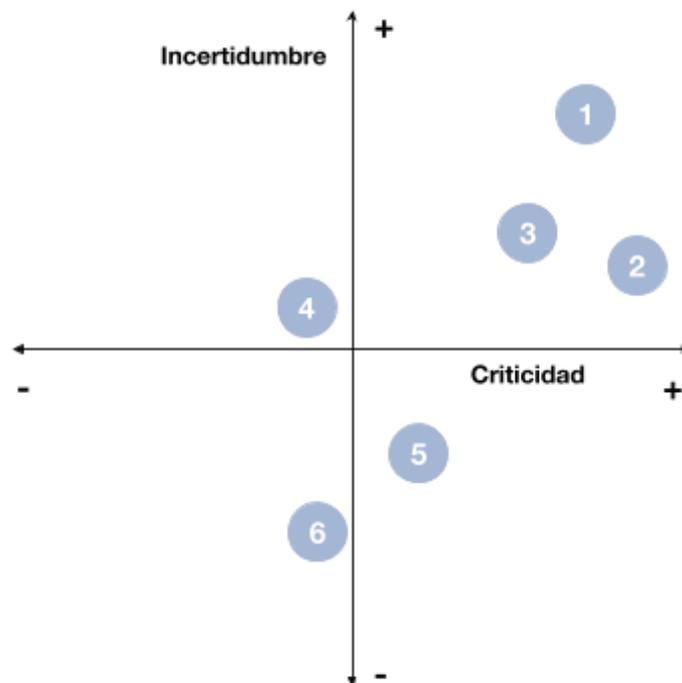
Hemos confirmado que son los propios astrofísicos los que determinan sus programas de investigación. Ellos están interesados en estudiar una materia en concreto y hacia ella dirigen sus trabajos, por tanto, serán ellos directa o indirectamente los principales clientes de nuestra actividad. Como ya hemos comentado, Galassify es una solución a medida para un programa específico dentro del IAC, pero cualquier investigador que requiera de soluciones de Machine Learning adaptadas a sus estudios puede ser objeto

de un desarrollo por nuestra parte. Ya ha quedado también demostrado que estas técnicas son capaces de colaborar en numerosos tipos de investigación dentro del mundo de la astrofísica.

Por otra parte, dado que los diferentes programas y subvenciones públicos de I+D+I son la principal fuente de financiación para la investigación, que en la mayor parte de ellos se valora el contar con colaboradores externos con el fin de impulsar la actividad de las empresas locales, y que son los propios investigadores los responsables de determinar sus necesidades para los estudios, es correcto afirmar que van a ser los equipos astrofísicos de los centros de investigación nuestros principales clientes.

### 4.1.2. Priorización

Una vez identificadas la hipótesis, hemos clasificado cada una de ellas en base a su criticidad y su incertidumbre. En el siguiente gráfico podemos ver el resultado:



Priorización de hipótesis

## 4.2. Entrevistas

A lo largo de los puntos anteriores hemos ido viendo los principales aspectos comentados en las entrevistas realizadas. En general, las entrevistas nos han servido para matizar y concretar los aspectos contemplados en nuestras hipótesis iniciales, reforzando aún más la idea del gran valor que aporta un listado de outliers para los investigadores.

Igual de interesante es el hecho de que este tipo de soluciones serán prácticamente obligatorias en el futuro. Se abre una vía de gran interés a explorar, como es la participación en el diseño de futuras misiones donde se recoja, no sólo la forma de extraer datos, sino que se planifique ya su tratamiento posterior.

También puede ser una vía de futuro la elaboración de software especializado en el tratamiento de datos astrofísicos, que se ajuste con facilidad a las necesidades de los diferentes proyectos de investigación y con distintos datasets.

Se adjunta en el **Anexo de gestión** los guiones preparados y las actas de las entrevistas realizadas.

## 5. Análisis y diagnóstico

A partir de las entrevistas realizadas con personal del IAC, detectamos las causas que resultan en el problema que queremos resolver. Hemos englobado las causas en 4 áreas, tal y como se puede ver en el siguiente diagrama causa-efecto:



Análisis Fish Bone

### El IAC como institución

Presentan dos aspectos que influyen el problema a tratar:

- ★ Carencia de personal especializado o con conocimientos sobre Business Intelligence y Big Data en general.
- ★ Competencia entre investigadores. Los investigadores compiten por sus proyectos y financiación, aunque también comparten el resultado de sus investigaciones e inquietudes.

## **Instrumentos de observación**

- ★ El perfeccionamiento y desarrollo de los equipos de medida supone la posibilidad de conseguir cada vez mayor cantidad de información sobre los cuerpos celestes.
- ★ Además, cada vez se consiguen datos de cuerpos más lejanos.

## **Los datos**

En este aspecto, como consecuencia de lo visto antes, existen dos factores que influyen:

- ★ El volumen de datos a tratar es de un tamaño considerable. Su manipulación usando algoritmos / técnicas “tradicionales” se presenta como algo inviable.
- ★ La fuente de datos que usan actualmente, general diferentes releases, cada una de ellas distinta, lo que dificulta el tratamiento de dichos datos.

## **Los investigadores**

En relación a los investigadores, existen varios aspectos a tener en cuenta:

- ★ Falta de conocimiento sobre técnicas automáticas de manipulación y análisis de datos (o interés por adquirirlo).
- ★ Falta de tiempo.
- ★ Por tanto, la selección de los datos a investigar puede no ser la más idónea, empleando métodos intuitivos y manuales.

Como conclusión, disponer de un volumen de datos cada vez mayor, dificulta la selección de aquellos considerados como los más adecuados para un estudio concreto. Es el caso de nuestro proyecto, donde hacemos referencia al estudio de las galaxias.

## **5.1. Análisis de la competencia**

Entre los distintos competidores que hemos encontrado sondeando el mercado hemos encontrado uno que destaca por encima del resto: DataRobot.

DataRobot es una empresa americana especializada en Machine Learning que ha trabajado con la NASA y cuentan con presencia en España, ya que disponen de una oficina en nuestro país. Por otra parte, dispone de agente comercial (Felipe Ortín), que ya ha tratado de colaborar con el IAC.

Datarobot es un software que tiene embebidos varios algoritmos y va mostrando el resultado de la ejecución de los mismos. Básicamente se trata de una solución en la que el usuario introduce su dataset, especifica el objetivo del análisis y el software se encarga de la creación y ejecución de diferentes modelos basados en Machine Learning. Luego el usuario es el que debe determinar qué modelo y con qué parámetros le interesa más cuál o cuáles quiere ejecutar.

En cualquier caso se trata de un producto genérico que no permite soluciones a medida ni está especializado en el área de astrofísica.

## 5.2. Análisis DAFO

El análisis DAFO es imprescindible para minimizar los problemas y sacar el mayor beneficio posible de las oportunidades, ya que nos va a ayudar en el desarrollo e implementación de una estrategia empresarial adecuada.



Análisis DAFO

Nuestro modelo de análisis DAFO está dividido en los siguientes factores internos y externos:

## Factores internos

### ★ Debilidades

- Dificultad a la hora de evaluar los resultados de nuestros algoritmos.
- Posibilidad de elegir incorrectamente el algoritmo más adecuado o que éste no defina adecuadamente los clusters requeridos.
- Falta de experiencia en el tratamiento de algoritmos complejos.
- Necesidad de un astrofísico en el equipo en próximos proyectos para facilitar la comunicación con los científicos y ganarnos su confianza.
- Costes de desarrollo muy elevados.

### ★ Fortalezas

- Estamos especializados en astrofísica.
- Somos un equipo multidisciplinar, con experiencia en el desarrollo de proyectos complejos.
- Alta capacidad de aprendizaje y adaptación del equipo.

## Factores externos

### ★ Amenazas

- Desconfianza por parte de los investigadores.
- Los investigadores prefieren no gastar su presupuesto en consultoras externas.

### ★ Oportunidades

- Somos los primeros. Se trata de un nicho de mercado donde no existe otra solución igual.
- Falta de conocimientos de Big Data por parte de los investigadores.
- Los presupuestos de las investigaciones pueden permitirse la consultoría.
- Tenemos contactos con el IAC y Galassify nos procurará experiencia y prestigio.
- La política de responsabilidad social del IAC favorece la contratación de empresas locales externas.
- No existen proyectos concretos de tratamiento de datos en los proyectos de inventarios celestes.
- Los científicos no disponen del tiempo necesario para realizar ellos mismos las labores de cribado de datos.

## 6. Plan estratégico

Como parte de nuestro plan estratégico hemos definido el modelo de negocio que se detalla a continuación, junto con los elementos clave cuya definición van a guiar el diseño de nuestra estrategia corporativa: misión, visión y objetivos.

### 6.1. Modelo de negocio

A continuación definiremos nuestro modelo de negocio siguiendo el Business Canvas Model.

<p><b>Asociados clave</b> </p> <ul style="list-style-type: none"> <li>• Científicos del IAC.</li> <li>• Proveedor de la arquitectura BI.</li> </ul>	<p><b>Actividades clave</b> </p> <ul style="list-style-type: none"> <li>• Definición del alcance.</li> <li>• Seguimiento y control.</li> <li>• Comprensión de los datos y tratamiento de los mismos.</li> <li>• Desarrollo de modelos.</li> <li>• Presentación de resultados.</li> </ul>	<p><b>Propuesta de valor</b> </p> <ul style="list-style-type: none"> <li>• Solvencia y eficacia en el estudio de las galaxias.</li> <li>• Elección de galaxias de mayor valor.</li> <li>• Detección de outliers.</li> <li>• Dedicación de esfuerzo en otras tareas de mayor valor.</li> <li>• Criterios de clasificación objetivos.</li> <li>• Modelo con capacidad de evolucionar.</li> <li>• Compartición de datos.</li> </ul>	<p><b>Relación con clientes</b> </p> <ul style="list-style-type: none"> <li>• Relación directa con el cliente.</li> <li>• El tipo de vínculo es transaccional, únicamente para este proyecto.</li> <li>• La relación es personal.</li> </ul>	<p><b>Clientes</b> </p> <ul style="list-style-type: none"> <li>• Astrofísicos</li> <li>• Investigadores</li> <li>• Científicos</li> </ul>	
<p><b>Recursos clave</b> </p> <ul style="list-style-type: none"> <li>• Base datos SDSS.</li> <li>• Arquitectura BI.</li> </ul>		<p><b>Canales</b> </p> <ul style="list-style-type: none"> <li>• Entrevistas personales</li> <li>• Correo electrónico</li> <li>• Foros y grupos de trabajo</li> </ul>			
<p><b>Estructura de costes</b> </p> <ul style="list-style-type: none"> <li>• Personal</li> <li>• Oficina</li> <li>• Arquitectura BI</li> </ul>			<p><b>Fuentes de ingresos</b> </p> <ul style="list-style-type: none"> <li>• Venta de servicios de consultoría</li> <li>• Mantenimiento/soporte de soluciones de clientes</li> </ul>		



### Segmentos de clientes

Para la realización del presente proyecto, nuestros clientes serán los investigadores del IAC que nos han encargado el desarrollo de una solución para un problema muy concreto.

Pese a tener un único cliente, podría ser viable una adaptación de la solución a desarrollar para otros clientes enfocados en la investigación y desarrollo pero para otras áreas en las que sea necesario una solución basada en técnicas de Machine Learning.



### **Propuesta de valor**

Con el sistema de clasificación automática, los científicos podrán abordar con mucha más solvencia y eficacia el estudio de las galaxias, dada la ingente cantidad de cuerpos que resultan del SDSS.

Además, podrán elegirse aquellas galaxias que tengan características de mayor valor para su estudio, como es el caso de nuestra solución, que facilita la detección de outliers o galaxias con características especiales o poco comunes.

Una vez desarrollado el modelo, éste podría evolucionar conforme surjan nuevas Data Releases del SDSS.

Además, los datos podrán compartirse con facilidad a toda la comunidad científica a través de nuestra web.

En definitiva lograremos que los investigadores, al no tener que realizar esta clasificación, puedan centrarse en tareas de mayor valor, además de contar con criterios de clasificación más objetivos.



### **Relación con clientes**

El tipo de relación establecido con nuestro cliente, el IAC, es directa. Es necesaria la relación bidireccional entre el equipo de trabajo y el cliente para comprender las necesidades del mismo y para diseñar una solución a medida.

Esta relación se realizará principalmente mediante entrevistas personales periódicas y correo electrónico directo entre el equipo de trabajo y el cliente.

El tipo de vínculo es transaccional, establecido únicamente para este proyecto. Una vez entregado el desarrollo y validado por el cliente, finaliza la relación. En todo caso, podría existir una relación posterior en proyectos futuros.



## Canales de distribución

Los canales establecidos con el equipo del IAC para abordar el proyecto son los siguientes:

- 1. Reuniones presenciales:** Las reuniones con el equipo del IAC serán presenciales y periódicas, de forma que aseguren un correcto control del avance del proyecto.
- 2. Correo electrónico:** La resolución de dudas puntuales y el envío de información entre ambas partes se llevará a cabo mediante correo electrónico.



## Actividades clave

Las actividades clave de nuestro proyecto son:

- 1. Definición del Alcance:** Es fundamental definir al comienzo del proyecto los límites del análisis: sobre qué datos se trabajará, qué se realizará con los mismos, los resultados que se esperan obtener de los mismos y cómo se presentarán y/o comunicarán.
- 2. Comprensión de los datos y tratamiento:** Se realizará un análisis exhaustivo de los datos con el objetivo de comprender cómo están organizados y lo que significan. Para ello, se utilizará la información disponible en la fuente de los mismos además de la que proporcionen los interlocutores del cliente.
- 3. Desarrollo de algoritmos de clasificación:** En esta actividad se prepara el entorno de trabajo que deberá soportar el tratamiento de toda la información. Se realiza la carga de datos en el mismo y las transformaciones necesarias para homogeneizar y normalizar la información. Se aplicarán las técnicas de clasificación que se consideren oportunas para el tipo de datos.
- 4. Presentación de resultados:** Los resultados del análisis se presentarán mediante el medio que se seleccione en la definición del alcance (web, informe, artículo, ...). Si se considera oportuno se puede realizar una presentación conjunta a los miembros del IAC que se consideren.

- 5. Comités de Control y Seguimiento:** La gestión y coordinación del proyecto es clave por lo que se debe crear un comité de seguimiento.
- Establecer los miembros del comité.
  - Definir la periodicidad de las reuniones y lo que se espera obtener de las mismas.



### Recursos clave

Para la puesta en marcha de los módulos de software a realizar así como su despliegue y publicación final, será muy importante apoyarse en una plataforma de cloud computing. En este caso, el proveedor elegido será Microsoft Azure.

Esta plataforma ofrece multitud de servicios especializados en tareas relacionados con el Big Data, por ejemplo:

- ★ Machine Learning
- ★ IoT
- ★ App services (web y restfull api)
- ★ Datalakes
- ★ Base de datos sql y nosql
- ★ Repositorios de código fuente basados en Git
- ★ Integración con aplicaciones de visualización y analítica como PowerBI
- ★ La posterior puesta a disposición del modelo a la comunidad científica sería a través de una página web



### Asociados clave

Nuestros asociados clave más importantes de este proyecto serán los científicos del IAC patrocinadores del proyecto: Ana Monreal Ibero y Jorge Sánchez Almeida. Ellos guiarán el desarrollo del proyecto con relación a la interpretación de los datos y a la selección de los resultados más adecuados.

Por otro lado, también tendremos como asociada la empresa de servicios Cloud utilizada para el almacenamiento del proyecto, de la computación y del interfase de salida de datos.



### Fuentes de ingresos

Los ingresos proceden de:

- ★ La venta de nuestros servicios de consultoría de Big Data, que estimamos serán 4 proyectos al año.
- ★ El coste de mantenimiento/soporte de nuestras soluciones año a año.



### Estructura de costes

En el apartado económico se desglosarán los costes en mayor detalle, pero el principal gasto será el personal de la empresa, y gastos asociados al mantenimiento de la oficina.

El desarrollo de la solución estará apoyada sobre la plataforma Microsoft Azure. El coste total, por mes, dependerá de los diferentes planes de precio aplicados a los servicios consumidos en la plataforma (servidores sql/nosql, machine learning, app services, etc.).

En esta plataforma cloud, el coste se calcula en función del uso y los recursos que se asignen a cada uno de los servicios mencionados anteriormente, por lo que hay una correlación directa entre este coste y la facturación de la empresa. A continuación, se muestra una estimación para un mes:

Producto / servicio	\$/mes
Servidor SQL	5,53\$
App service web	38,69\$
App service api	38,69\$
Machine Learning Studio	Gratis
Servicio Azure Machine Learning (entrenamiento de modelos)	49,64\$
PowerBI Embedded (30 horas):	30,24\$
Otros servicios	10,00\$
<b>TOTAL</b>	<b>124,45\$</b>

## 7. Plan de acción

Dentro de nuestro plan describimos las claves esenciales para convertir la estrategia en acción y monitorizar su implantación efectiva.

### 7.1. Alcance

En el presente proyecto abordaremos únicamente una **primera fase**, en la que se definirá un algoritmo general de clasificación de galaxias en función de su espectro con el fin de detectar outliers.

En este algoritmo general, trataremos una **sección concreta de los espectros**, sección que variará en otras fases para obtener clasificaciones más ajustadas a los requisitos deseados.

El motivo de plantear esta generalización es la de poder ajustar un alcance inicial que servirá de muestra de las capacidades del algoritmo.

En cuanto a los **datos**, por motivos de practicidad y capacidad de computación, el trabajo técnico se ha centrado en dos bases de datos:

- ★ **MyBestDR7**: se trata de una base de datos pública y reducida de la SDSS DR7 con fines experimentales. Una vez filtrados los datos que nos interesan, estaríamos hablando de unas 500 galaxias que se encuentran disponibles en la Base de Datos de nuestra arquitectura BI.
- ★ **SDSS DR15**: es la última versión disponible de la SDSS. Su volumen sobrepasa las capacidades de la arquitectura BI por lo que se ha decidido descargar los datos en local y trabajar con una muestra aleatoria de los mismos.

### 7.1.1. Misión, visión y objetivos

Hemos definido quiénes somos y en quién queremos convertirnos, estableciendo así la dirección a tomar para conseguir los objetivos de nuestra compañía.



MISIÓN

*“Ofrecer soluciones de Big Data a nuestros clientes para que puedan centrarse en lo verdaderamente importante”*



VISIÓN

*“Convertirnos en un referente mundial, trabajando codo con codo con nuestros clientes, haciendo de sus problemas los nuestros”*

Teniendo en cuenta los análisis previos se han desarrollado los siguientes objetivos y estrategias:

**Objetivo 1:** Darnos a conocer en la comunidad científica.

- ★ **Estrategia 1:** Asistencia a eventos.
- ★ **Estrategia 2:** Participar en grupos de trabajo.
- ★ **Estrategia 3:** Publicar nuestros propios estudios.

**Objetivo 2:** Ofrecer soluciones que cumplan con las expectativas del cliente.

- ★ **Estrategia 1:** Conocer el grado de satisfacción de nuestros clientes.
- ★ **Estrategia 2:** Utilizar estándares.
- ★ **Estrategia 3:** Desarrollar planes de formación continua.

**Objetivo 3:** Escalar la solución a otros programas de investigación astrofísica.

- ★ **Estrategia 1:** Ofrecer nuestros servicios a clientes con problemas parecidos.
- ★ **Estrategia 2:** Crear un porfolio de investigadores, donde se registre qué está estudiando y qué problemas tiene.

## 7.1.2. Métricas

Con el objeto de medir los conceptos validados anteriormente, se definen una serie de indicadores clave de negocio (KPIs) que nos darán objetividad y precisión evaluar si estamos cumpliendo los objetivos.

### 7.1.2.1. Diferencia de costes entre selección manual y automatizada

En estos momentos, la clasificación previa al estudio conlleva una serie de horas que los científicos consideran de poco valor: se trata únicamente de seleccionar los objetos que serán incorporados al estudio a realizar.

Estas horas de trabajo previo tienen un valor económico que llamaremos **Importe Manual (Im)**.

Por contra, la selección automática reduciría a cero esas horas, eliminando del pipeline de la investigación esas horas de bajo valor percibido. Por supuesto, la clasificación automática tiene un coste para el estudio (el coste de nuestros servicios), que llamaremos **Importe Automático (Ia)**.

Para cada proyecto, por tanto, se puede definir un KPI llamado **Diferencia de Costes (DC)** que resulta:

$$DC = \frac{(Im - Ia) \times 100}{Im}$$

Este KPI se puede obtener al inicio de **Galassify**, así como de otros proyectos posteriores similares, en base a preguntas a los Clientes sobre el valor Im.

### 7.1.2.2. Papers emitidos y de referencia

El uso de **Galassify** implica que la clasificación automática generará galaxias de estudio óptimas, frente a la selección manual cuyos resultados son limitados debido a que no se están considerando el 100% de los datos, como ya se indicó anteriormente.

Esta mejora de las galaxias de estudio implica, sin duda, resultados científicos de mayor calidad. La calidad de estos resultados resulta difícil de medir, pero podemos tener una aproximación a través del número de Artículos Científicos (Papers) emitidos por los científicos que hayan empleado **Galassify** (y otros proyectos posteriores) y, sobre todo, del número de referencias a estos Papers contenidos en los trabajos de otros investigadores.

Llamaremos **Papers emitidos (Pe)** al número de papers generados donde haya sido usado Galassify para el estudio. Por otro lado, llamaremos **Papers referenciados (Pr)** al número de referencias a estos Papers desde otros.

Evidentemente, estos dos KPIs no pueden medirse de forma instantánea, sino que podrán medirse con el paso del tiempo. Definimos un periodo de revisión trimestral durante un dos años, con los siguientes objetivos:

Pe = 5

Pr = 10

Este KPI tiene por objeto hacer un seguimiento de resultados que indique el éxito del proyecto y que pueda utilizarse comercialmente para futuros proyectos.

#### 7.1.2.3. Subvenciones conseguidas

Con el mismo espíritu que el KPI anterior, la consecución de subvenciones para la realización de estudios científicos donde se emplee **Galassify**, será una medida del éxito del proyecto.

Mediremos, también trimestralmente, el importe de subvenciones conseguidas por proyectos donde se haya empleado **Galassify** (podrá detectarse mediante los papers generados a la finalización del estudio), denominando a este KPI **Importe de Subvenciones Logradas (ISL)** y **Número de Subvenciones Logradas (NSL)**.

El objetivo a 2 años es de:

ISL = 2.000.000€

NSL = 10

Estos cuatro últimos KPIs definidos (Pe, Pr, ISL y NSL) no impactarán económicamente de forma directa en **Galassify**, pero si nos darán una medida de prestigio en la comunidad científica, que podrá usarse sobre todo en fases posteriores del proyecto.

#### 7.1.2.4. Contactos científicos realizados

**Galassify** es una empresa de asesoría e implantación de proyectos de big data especializada en Astrofísica, de forma que nuestros Clientes trabajan fundamentalmente

en este sector. Para la consecución de nuevos proyectos, acordes a los objetivos económicos fijados en nuestro Plan Económico, es necesario realizar contactos con científicos reputados que puedan contar con nuestros servicios.

Utilizando los canales adecuados, realizaremos un mínimo de 10 contactos al mes (2 por miembro del equipo), siendo necesario que al menos 2 de ellos sean calificados como “contactos con potencial”, es decir, aquellos con los que haya altas probabilidades de contratación.

Para el seguimiento de actividades comerciales, en un principio usaremos la aplicación **Zoho CRM**. Se realizarán reuniones mensuales de seguimiento comercial para, además, actualización de estos objetivos, acordes a los objetivos de contratación anual.

#### 7.1.2.5. Satisfacción de Clientes

Al finalizar el proyecto, así como después de 6 meses de la entrega, se realizará una encuesta a nuestros Clientes con el fin de determinar su grado de satisfacción. Entre las preguntas del cuestionario, vendrá incluida la pregunta “¿Recomendaría los servicios de Galassify a algún colega cercano? (1-10)”. El valor de esta pregunta será el NPS (Net Promoter Score), con el que se hará una media. El objetivo es que dicho KPI supere el valor de 8.

En caso de que el NPS no supere el valor de 8 en cualquiera de los Clientes, será necesaria una Reunión de Cliente específica del equipo donde se analice lo ocurrido y se fijen acciones concretas para conocer más del problema y mejorar los resultados futuros.

Hay que tener en cuenta que, una vez realizados los esfuerzos comerciales iniciales, este mercado muy probablemente se mueva fundamentalmente con el “boca a boca”, por lo que **el NPS se convierte en nuestro principal KPI**.

## 7.2. Análisis de actividades y tareas

En los siguientes apartados detallamos el mapa de procesos y la solución tecnológica.

### 7.2.1. Mapa de procesos

En el siguiente gráfico se refleja en mapa de procesos que forman parte de nuestra empresa. Gracias al correcto desempeño de cada uno de ellos, somos capaces de aportar valor a nuestros clientes, logrando como objetivo clientes satisfechos.



Mapa de procesos

### Procesos estratégicos

★ **Dirección estratégica:** A través de la dirección estratégica definiremos el rumbo a seguir por la empresa y estableceremos el cómo vamos a hacerlo, de ahí la importancia de este proceso. Entre las actividades que comprende destacamos:

- Formar una visión de hacia dónde se dirige la organización.
- Marcar unos objetivos específicos.
- Elaborar una estrategia para lograr dichos objetivos.
- Poner en práctica dicha estrategia.
- Evaluar el resultado y realizar las acciones correctivas oportunas.

- ★ **Gestión de las comunicaciones:** Será indispensable para la continuidad del negocio llevar una gestión de las comunicaciones exquisita, tanto desde el punto de vista de nuestros clientes, como de los clientes potenciales. Debemos tener presente que la cercanía con la comunidad científica nos permitirá conocer sus necesidades en detalle, aspecto indispensable para poder proponer soluciones a medida.

También deberá tenerse en cuenta que, en sectores tan acotados como éstos, la mejor carta de presentación a futuros clientes, es la satisfacción de nuestros propios clientes, que hablarán a nuestro favor a otros miembros de la comunidad.

- ★ **Mejora continua:** Los instrumentos científicos son cada vez más potentes, recogen más información y son capaces de detectar objetos celestes más lejanos. Si a este hecho le sumamos que la ciencia es una materia donde cada día se descubren nuevos avances, debemos estar siempre al corriente de tales acontecimientos y evaluar si nuestros servicios se ajustan a dichas necesidades o debemos mejorar en algún aspecto, ya sea de tipo infraestructura tecnológica o algorítmica. Al mismo tiempo será vital estar al día en los últimos avances en algoritmia en Machine Learning y Deep Learning.

## Procesos de apoyo

- ★ **Gestión financiera:** Consiste en administrar los recursos económicos de la empresa de modo que sean suficientes para cubrir los gastos, y así la empresa pueda seguir en funcionamiento.
- ★ **Gestión de personal:** Abarca todas las tareas rutinarias administrativas del departamento de recursos humanos. Entre sus actividades centrales se encuentran:
  - Elaboración de contratos laborales.
  - Encargarse de los pagos de salarios, seguridad social, etc.
  - Gestión de los expedientes de los empleados.
  - Registro de las vacaciones, permisos, bajas, etc.
  - Capacitación del personal.
  - Formación.
  - Control horario.

Inicialmente la plantilla estará formada por los 5 miembros del equipo de trabajo de Galassify.

El talento de nuestros empleados representa el pilar fundamental para lograr la satisfacción de nuestro clientes, es por ello que debemos llevar políticas de

fidelización, minimizando la rotación de nuestros trabajadores y fomentando las actividades formativas.

- ★ **Mantenimiento de la infraestructura:** Al tratarse de una solución basada en un infraestructura tecnológica, es indispensable llevar un mantenimiento de la misma. Con tal fin se desarrollarán y ejecutarán planes mantenimiento de equipos, copias de seguridad y demás elementos que formen parte del ecosistema tecnológico de la empresa.

### Procesos clave

- ★ **Análisis:** En primer lugar estudiaremos las necesidades del cliente, para poder identificar tanto el problema como los resultados que esperan obtener de nuestra solución. Acto seguido estableceremos un alcance asumible dentro del tiempo establecido para el proyecto. Una vez definido el alcance, realizaremos un análisis exhaustivo de los datos con el objetivo de comprender su significado y cómo se relacionan entre sí. Para ello, se utilizará la información disponible en la fuente de los mismos además de la que proporcionen los interlocutores del cliente.
- ★ **Diseño:** En esta etapa diseñaremos los distintos elementos que forman parte de nuestra solución, incluyendo desde la arquitectura necesaria para soportar el tratamiento de los datos, como los distintos modelos de Machine Learning que vayamos a utilizar (previsiblemente durante la etapa de desarrollo, se verá qué algoritmos arrojan mejores resultados y se podrá optar por probar algoritmos no incluidos en el diseño inicial).
- ★ **Desarrollo:** En esta actividad se prepara el entorno de trabajo que deberá soportar el tratamiento de toda la información. Se realiza la carga de datos en el mismo y las transformaciones necesarias para homogeneizar y normalizar la información. Se aplicarán las técnicas de clasificación que se consideren oportunas para el tipo de datos.

Entiéndase como “entorno de trabajo”, la infraestructura creada para dar soporte a la solución técnica aportada para resolver el problema planteado por el cliente. En la sección **Infraestructura BD/BI** del documento **Anexo Técnico**, se describe en detalle las diferentes herramientas, sistemas y procesos que forman parte de la misma.

- ★ **Entrega:** La entrega podrá llevarse a cabo en distintos formatos según las necesidades del cliente. Para el IAC en concreto, se elaborará un documento en formato PDF en el que poder consultar los análisis y tareas así como los

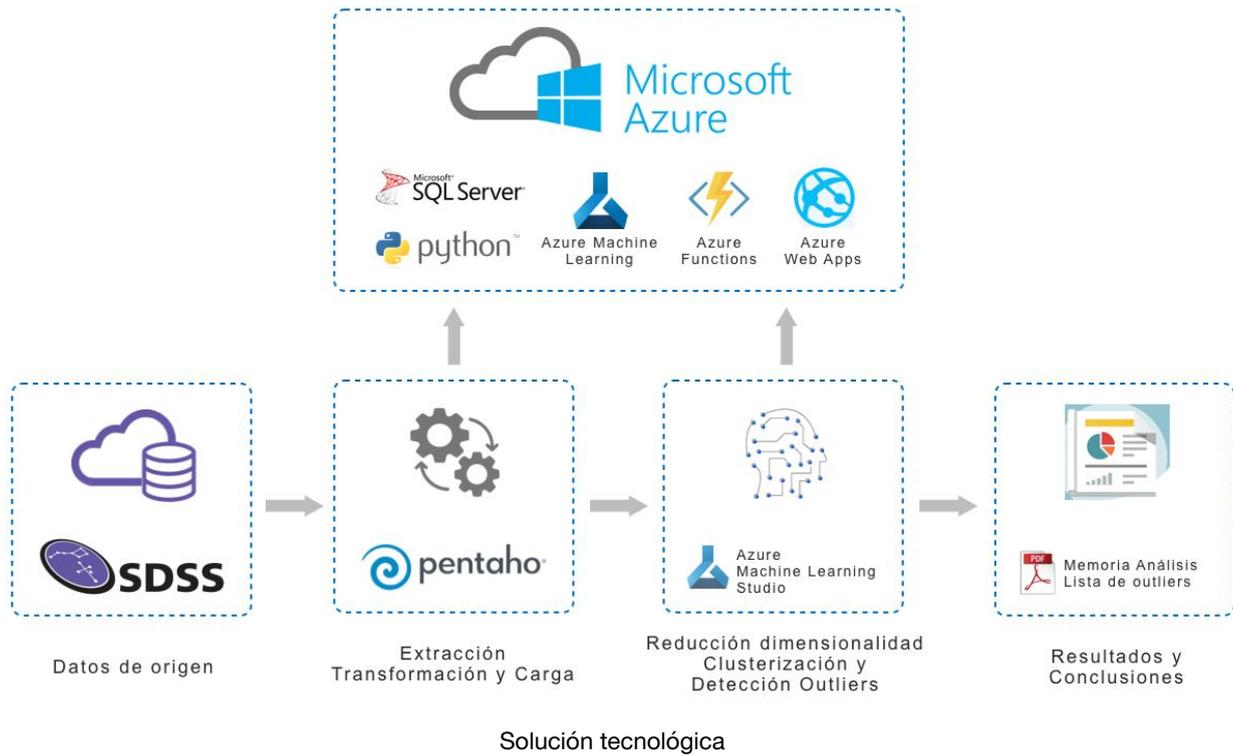
resultados obtenidos (lista de outliers) en diferentes ejecuciones de las herramientas desarrolladas.

- ★ **Seguimiento y control:** La gestión y coordinación del proyecto es clave por lo que se debe crear un comité de seguimiento, establecer los miembros y definir la periodicidad de las reuniones y lo que se espera obtener de las mismas. Es fundamental trabajar de forma iterativa presentando al cliente resultados parciales de los progresos, con la finalidad de detectar desviaciones en los resultados lo antes posible, ahorrando tiempo y aumentando el grado de satisfacción del cliente.

### 7.2.2. Solución tecnológica

En este apartado se describen las herramientas que formarán parte de la solución tecnológica. La siguiente figura ilustra cómo se relacionan diferentes elementos de software como:

- ★ **Fuente de datos de partida:** SSDS.
- ★ **Bases de datos:** SQL Server.
- ★ **Lenguajes de programación:** Python.
- ★ **Herramientas ETL:** Pentaho.
- ★ **Herramientas para Machine Learning / Deep learning:**
  - AZ ML Service.
  - AZ ML Studio.
- ★ **Infraestructura Cloud y arquitectura basada en microservicios:**
  - Azure Functions.
  - Azure Web Apps.



La figura anterior pretende describir, además, un workflow en el que se puede adivinar las distintas fases que es necesario ejecutar para dar respuesta al problema planteado. A grandes rasgos, dichas fases sería:

1. Adquisición de los datos de origen con la información de las galaxias a clasificar.
2. Realizar una primera batería de tareas ETL sobre datos originales darán como resultado un conjunto de datasets para ser tratados en la siguiente fase. Se trata de tareas enfocadas en normalizar y preparar la información obtenida de la fuente original para facilitar su posterior uso en la fase de aplicación de los diferentes algoritmos.
3. En esta fase, se realizan nuevas transformaciones de la información y se llevarán a cabo las tareas de cálculos de nuevas columnas en los diferentes dataset, se realizará la reducción de dimensionalidad, la clusterización y la obtención de outliers.
4. Una vez aplicados los algoritmos, se culmina el proceso con la presentación de los resultados acompañado de una descripción del estudio realizado, fases realizadas y resultados obtenido para resolver el problema.

En el anexo técnico, se describen con más detalle y con un enfoque más técnico, la solución tecnológica propuesta.

### 7.3. Análisis de los recursos

En este apartado se detallan los recursos necesarios para la ejecución del proyecto.

#### 7.3.1. Talento humano

El equipo humano estará formado por los 5 miembros del grupo, personal altamente técnico que se dividirá en función de las necesidades en los siguientes roles:

	Role	Responsabilidades	Habilidades
JP	Jefe de proyecto	Lidera y coordina los distintos proyectos. Planifica, establece objetivos, supervisa la evolución de las tareas, interlocuta e informa al cliente y decide sobre la incorporación de cambios a lo largo de la vida del proyecto	Capacidad de organización Liderazgo Comunicación eficaz Capacidad de negociación Atención al detalle Identificación de problemas Conocimientos de astrofísica Big Data y Business Intelligence Visualización de resultados
ANA	Analista BI	Recogida y análisis de los requisitos de los clientes	Comunicación eficaz Identificación de problemas Conocimientos de astrofísica Big Data y Business Intelligence Visualización de resultados
ARQ	Arquitecto BI	Diseña, monta y mantiene la infraestructura de BI necesaria para dar soporte a las soluciones de nuestros clientes	Microsoft Azure Microsoft SQL Server Pentaho Python
DE	Data Engineer	Recoge, almacena, suministra y trata previamente los datos	Sistemas de Bases de Datos Herramientas ETL APIs de Datos Soluciones de almacenamiento Modelo de datos
DS	Data Scientist	Interpretación y modelado de los datos para resolver los problemas de los clientes	Computación distribuida Estadística y matemática Machine Learning y Deep Learning

La metodología de proyectos utilizada será SCRUM. Los roles podrán ser asignados de forma dinámica en función de las necesidades de los proyectos.

### 7.3.2. Recursos físicos

#### ★ Estación de trabajo

Con independencia de la existencia de una oficina como sede de la empresa, se contempla y se fomenta la modalidad de teletrabajo, de forma que cada miembro del equipo de trabajo debe disponer de su propio ordenador con conexión a internet.



Haciendo uso de la cuenta de Google, compartimos la información del proyecto mediante el uso del servicio **Drive**.



Para el seguimiento y control de tareas hemos optado por usar la herramienta gratuita **Trello**.



Como herramienta de trabajo colaborativa hemos implantado **Slack**.

#### ★ Instalaciones

Se alquilará una oficina para mantener la operativa básica de la empresa, si bien como se comentó anteriormente, la mayor parte del trabajo se realizará de forma remota.

Las reuniones con los clientes podrán llevarse a cabo tanto en las propia oficina de la empresa, como en las dependencias del cliente.

En cuanto a las reuniones internas, se realizarán preferiblemente mediante videoconferencia. Durante esta primera fase, para las reuniones físicas hemos optado por 2 vías:



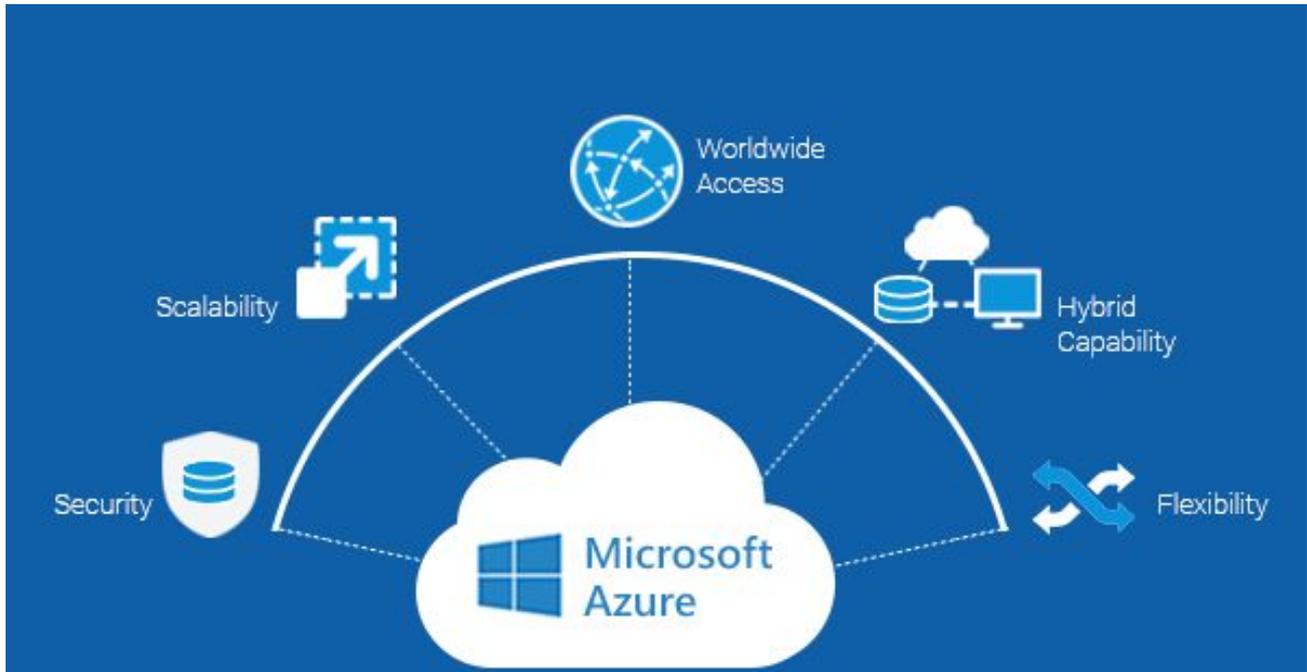
TF CoworkIN, instalaciones del Parque Científico y Tecnológico de Tenerife (PCTT) en la dársena pesquera.



Carrels de la Facultad de La Laguna

## ★ Arquitectura Big Data

Para el despliegue de la solución tecnológica se usará Microsoft Azure como base ya que permite partir de una arquitectura básica que será detallada en el **Anexo técnico**.



La base del problema a resolver es la aplicación de diferentes etapas de transformación para preparar los datos que luego serán usados por los algoritmos de machine learning y deep learning.

Para poder lograr unos resultados lo más preciso posible, es necesario gestionar un importante volumen de datos. Tanto la adquisición de este volumen de datos de su fuente original (SDSS), como las transformaciones necesarias así como el entrenamiento de modelos de ML/DL, deben ser realizados en un entorno con gran capacidad de cómputo. En este sentido, un proveedor como Microsoft Azure es el aliado perfecto permitiendo realizar dichos trabajos en un tiempo y coste económicos más que aceptable.

Microsoft Azure, no solo permite tareas propias relacionadas con el BigData, además permite preparar la solución para ser escalada, desplegable a través de Internet siguiendo arquitecturas clásicas de web o microservicios y de permitir la interconectividad con terceras herramientas si fuera necesario a través del uso, por ejemplo, apis restfull.

## 7.4. Gestión del tiempo

A partir de los procesos clave definidos previamente, hemos construido el cronograma del proyecto. En primer se han desglosado las tareas, se han asignado recursos y por último hemos establecido los plazos para cada una de ellas.

Nuestro proyecto arranca la segunda semana de abril de 2019, más concretamente el día 8 cuando se celebra la reunión de kick off con el cliente. La entrega final del primer sprint para el cliente se realizará el 19 de septiembre de 2019.

TAREAS	RECURSOS	SEMANAS					
		ABRIL	MAYO	JUNIO	JULIO	AGOSTO	SEPTIEMBRE
<b>Análisis</b>		■	■	■			
Estudiar requisitos cliente	ANA	■					
Determinar el alcance	ANA, JP	■					
Analizar los datos	ANA		■	■			
<b>Diseño</b>				■	■		
Diseñar infraestructura BI	ARQ			■	■		
Diseño modelo de datos	DE			■	■		
Seleccionar modelos analíticos	DS				■	■	
Diseñar formato de entrega	DE					■	■
<b>Desarrollo</b>					■	■	■
Montar infraestructura BI	ARQ				■	■	■
Desarrollar procesos ETL	DE				■	■	■
Ejecutar modelos analíticos	DS					■	■
Memoria con los resultados	DE						■
<b>Entrega</b>							■
Presentación solución	JP						■
<b>Seguimiento</b>		■	■	■	■	■	■
Reuniones internas	JP, ANA, ARQ, DE, DS	■	■	■	■	■	■
Reuniones cliente	JP, ANA	■	■	■	■	■	■

Cronograma del proyecto

## 8. Optimización de los resultados

En el presente apartado evaluamos la viabilidad de Galassify como empresa desde el punto de vista financiero. Para ello hemos realizado un estudio económico y financiero, donde trabajamos con una proyección de la empresa a 5 años.

Como ya hemos indicado anteriormente, el primer proyecto que abordaremos es Galassify y, si bien es cierto que aspiramos a lograr varios proyectos al año de tratamiento automático de datos en el campo de la Astrofísica, en principio no descartamos la realización de otro tipo de proyectos que puedan contribuir a los objetivos estratégicos de la empresa.

A continuación, se describirán los números generales de la empresa, de forma que el presente proyecto se incluye como uno más de los proyectos que compondrán el portfolio de la compañía.

### 8.1. Claves

- ★ El funcionamiento de Galassify como empresa se basa en la contratación, ejecución y facturación de 4 proyectos anuales por valor medio de 75.000€.
- ★ Como ingresos se consideran tanto los proyectos principales como el soporte de mantenimiento de los mismos (reentrenamiento de modelos, corrección de bugs, pequeñas incorporaciones de features, etc.).
- ★ El gasto principal son las nóminas de los 5 miembros del equipo.
- ★ No se prevé contratación de personal en los 5 primeros años de funcionamiento.
- ★ Para oficina podría optarse por un espacio de coworking para reuniones periódicas y teletrabajo el resto del tiempo. En cualquier caso y como previsión de máximos, se considera en gastos toda la estructura de una pequeña oficina.
- ★ Durante los primeros años, será importante un plan de marketing con el que conocer a la mayor cantidad posible de astrofísicos en centros distintos. Después, el crecimiento esperado vendrá vía recomendación.

- ★ Nuestra empresa tiene una fuerte componente de responsabilidad social corporativa (RSC). Queremos participar en la medida de nuestras posibilidades en acciones de carácter social y de mejora del medio ambiente en la sociedad local, lo que se incluye en nuestros presupuestos.
- ★ La financiación inicial de la empresa se logra mediante la aportaciones de un capital social de 40.000 € (5 socios a 8.000 € por socio), la facturación de los servicios a Clientes y quedarían por cubrir 16.450 € de flujo de caja negativo el primer año y 12.742 € el segundo año. Estos flujos negativos deberán cubrirse mediante inversores externos.
- ★ La única inversión en activos a realizar es la compra de equipos informáticos por valor de 9.000 €, a amortizar en 3 años.
- ★ Se considera un IPC anual de un 2%.
- ★ Para el cálculo del VAN se ha considerado un tipo de interés para el inversor de un 10%.

## 8.2. Cuenta de resultados

### 8.2.1. Ingresos

Los ingresos proceden principalmente de la venta de proyectos de consultoría de Big Data, que estimamos serán 4 proyectos al año, con un importe medio aproximado de 75.000 € por proyecto. A lo que sumamos el coste de mantenimiento/soporte de nuestras soluciones año a año. Hemos valorado este coste de soporte en 15.000 €/año por proyecto.

INGRESOS					
Concepto	Año 1	Año 2	Año 3	Año 4	Año 5
Nº nuevos proyectos	4	4	4	4	4
Precio anual producto	75.000,00 €	75.000,00 €	75.000,00 €	75.000,00 €	75.000,00 €
<b>Total ingresos Ventas</b>	<b>300.000,00 €</b>				
Soporte anual producto	15.000,00 €	15.000,00 €	15.000,00 €	15.000,00 €	15.000,00 €
<b>Total ingresos Soporte</b>	<b>- €</b>	<b>60.000,00 €</b>	<b>120.000,00 €</b>	<b>180.000,00 €</b>	<b>240.000,00 €</b>
<b>Total ingresos</b>	<b>300.000,00 €</b>	<b>360.000,00 €</b>	<b>420.000,00 €</b>	<b>480.000,00 €</b>	<b>540.000,00 €</b>

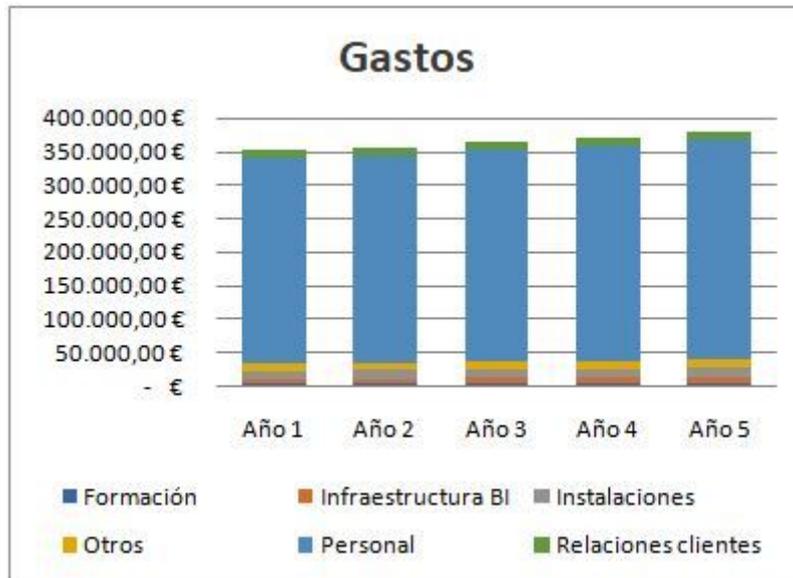


Ingresos estimados a 5 años

## 8.2.2. Gastos

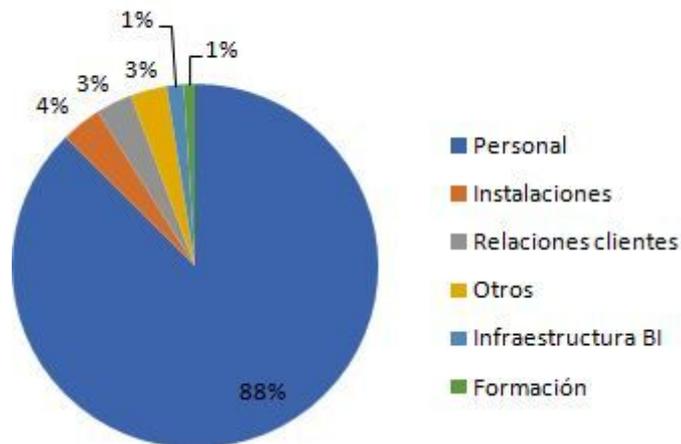
El principal gasto es el de personal, más concretamente las partidas destinadas al pago de las nóminas de los 5 trabajadores y la seguridad social correspondiente. Todo ello calculado en base a un sueldo medio de 45.000 €/año por trabajador.

GASTOS					
Concepto	Año 1	Año 2	Año 3	Año 4	Año 5
Alquiler de oficinas	7.200,00 €	7.344,00 €	7.490,88 €	7.640,70 €	7.793,51 €
Conservación oficinas	1.200,00 €	1.224,00 €	1.248,48 €	1.273,45 €	1.298,92 €
Suministro eléctrico	1.800,00 €	1.836,00 €	1.872,72 €	1.910,17 €	1.948,38 €
Agua	840,00 €	856,80 €	873,94 €	891,41 €	909,24 €
Limpieza	1.200,00 €	1.224,00 €	1.248,48 €	1.273,45 €	1.298,92 €
IBI	350,00 €	357,00 €	364,14 €	371,42 €	378,85 €
Suscripciones software	1.250,00 €	1.275,00 €	1.300,50 €	1.326,51 €	1.353,04 €
Gastos Cloud proyectos	4.000,00 €	4.080,00 €	4.161,60 €	4.244,83 €	4.329,73 €
Servicios Cloud	- €	816,00 €	1.664,64 €	2.546,90 €	3.463,78 €
Asesoría	3.000,00 €	3.060,00 €	3.121,20 €	3.183,62 €	3.247,30 €
Publicidad	3.000,00 €	3.060,00 €	3.121,20 €	3.183,62 €	3.247,30 €
Atenciones con Clientes	500,00 €	510,00 €	520,20 €	530,60 €	541,22 €
Desplazamiento	5.500,00 €	5.610,00 €	5.722,20 €	5.836,64 €	5.953,38 €
Dietas	2.310,00 €	2.356,20 €	2.403,32 €	2.451,39 €	2.500,42 €
Material de oficina	500,00 €	510,00 €	520,20 €	530,60 €	541,22 €
Comunicaciones	3.000,00 €	3.060,00 €	3.121,20 €	3.183,62 €	3.247,30 €
Mensajerías	300,00 €	306,00 €	312,12 €	318,36 €	324,73 €
Cuotas asociaciones	500,00 €	510,00 €	520,20 €	530,60 €	541,22 €
Sueldos y salarios	192.857,14 €	192.857,14 €	196.714,28 €	200.648,57 €	204.661,54 €
Pagas extras	32.142,86 €	32.785,72 €	33.441,43 €	34.110,26 €	34.792,47 €
Seguridad Social	82.500,00 €	84.150,00 €	85.833,00 €	87.549,66 €	89.300,65 €
Formación	3.500,00 €	3.570,00 €	3.641,40 €	3.714,23 €	3.788,51 €
Gastos sociales diversos	1.750,00 €	1.785,00 €	1.820,70 €	1.857,11 €	1.894,26 €
Mejora medio ambiente	1.750,00 €	1.785,00 €	1.820,70 €	1.857,11 €	1.894,26 €
Otros gastos	500,00 €	510,00 €	520,20 €	530,60 €	541,22 €
<b>Total</b>	<b>351.450,00 €</b>	<b>355.437,86 €</b>	<b>363.378,93 €</b>	<b>371.495,48 €</b>	<b>379.791,33 €</b>



Gastos estimados a 5 años

### Distribución del gasto



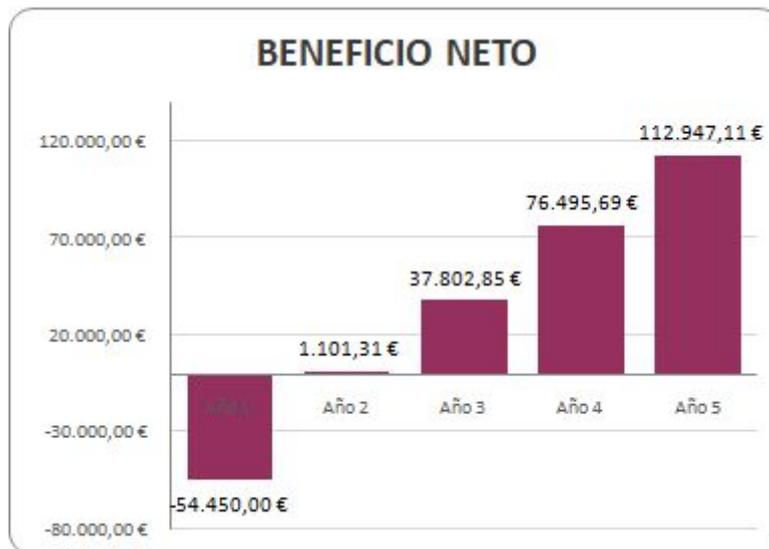
Distribución anual del gasto

AMORTIZACIÓN			
Concepto	Inversión	Amortiz. anual	Amort. anual €
Equipos informáticos	9.000,00 €	33,3%	3.000,00 €

### 8.2.3. Cuenta de pérdidas y ganancias

A partir de los ingresos y los gastos estimados previamente, calculamos la cuenta de resultados:

CUENTA DE RESULTADOS					
Concepto	Año 1	Año 2	Año 3	Año 4	Año 5
Total ingresos	300.000,00 €	360.000,00 €	420.000,00 €	480.000,00 €	540.000,00 €
Gastos por Estructura	351.450,00 €	355.437,86 €	363.378,93 €	371.495,48 €	379.791,33 €
Margen Operativo (EBITDA)	- 51.450,00 €	4.562,14 €	56.621,07 €	108.504,52 €	160.208,67 €
Amortización	3.000,00 €	3.000,00 €	3.000,00 €	- €	- €
EBIT	- 54.450,00 €	1.562,14 €	53.621,07 €	108.504,52 €	160.208,67 €
Intereses	- €	- €	- €	- €	- €
EBT	- 54.450,00 €	1.562,14 €	53.621,07 €	108.504,52 €	160.208,67 €
Impuestos	- €	460,83 €	15.818,21 €	32.008,83 €	47.261,56 €
<b>BENEFICIO NETO</b>	<b>- 54.450,00 €</b>	<b>1.101,31 €</b>	<b>37.802,85 €</b>	<b>76.495,69 €</b>	<b>112.947,11 €</b>
<b>MARGEN BENEFICIO</b>	<b>-18,15 %</b>	<b>0,31 %</b>	<b>9 %</b>	<b>15,94 %</b>	<b>20,92 %</b>



Beneficio neto

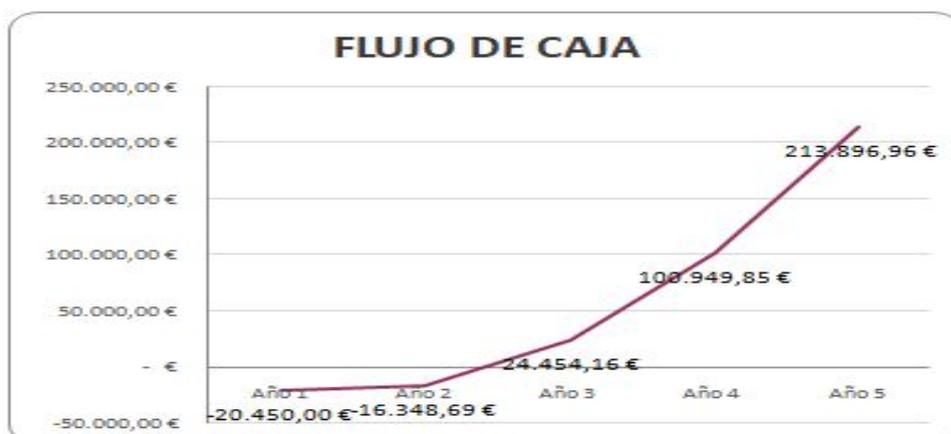
### 8.3. Flujo de caja

El estado de flujos de caja muestra los movimientos reales de efectivo que se han producido en nuestra empresa durante un periodo determinado de tiempo. Con ello lograremos identificar los cambios que se producen en el saldo de caja y tesorería de la empresa, al relacionar el saldo del período anterior con el saldo que tendremos al final del período analizado.

A continuación, hemos proyectado el flujo de caja en los próximos 5 años.

FLUJO DE CAJA					
Concepto	Año 1	Año 2	Año 3	Año 4	Año 5
Aporte de los socios	40.000,00 €	- €	- €	- €	- €
Saldo Inicial	40.000,00 €	- 20.450,00 €	- 16.348,69€	24.454,16 €	100.949,85 €
Total ingresos por Ventas	300.000,00 €	360.000,00 €	420.000,00 €	480.000,00 €	540.000,00 €
Total Ingresos	300.000,00 €	360.000,00 €	420.000,00 €	480.000,00 €	540.000,00 €
Hardware	- 9.000,00 €	- €	- €	- €	- €
Total Inversiones	- 9.000,00 €	- €	- €	- €	- €
Gastos por Estructura	- 351.450,00 €	- 355.437,86 €	- 363.378,93 €	- 371.495,48 €	- 379.791,33 €
Impuestos	- €	- 460,83 €	-15.818,21€	- 32.008,83 €	- 47.261,56 €
Total Gastos	- 351.450,00 €	- 355.898,69 €	- 379.197,15 €	- 403.504,31 €	- 427.052,89 €
<b>SALDO FINAL</b>	<b>- 20.450,00 €</b>	<b>- 16.348,69 €</b>	<b>24.454,16€</b>	<b>100.949,85 €</b>	<b>213.896,96 €</b>

Tal y como se observa, la evolución de los flujos de caja es creciente a lo largo de los 5 años, si bien es cierto que el aumento del flujo de caja los primeros años no es tan significativo como al final, siendo negativo durante los 2 primeros años. En cualquier caso, se trata de un proyecto viable.



Flujo de caja a 5 años

## 8.4. Balance

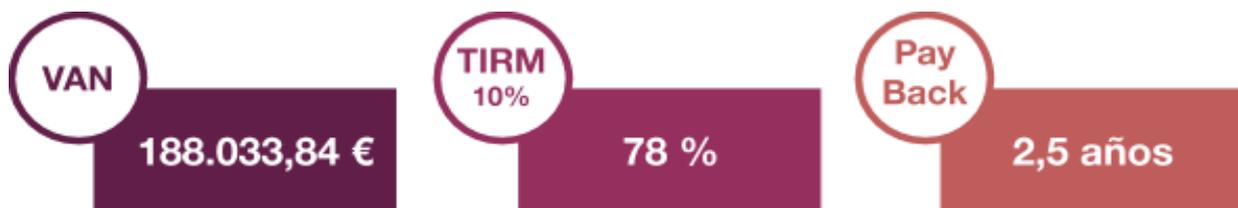
Balance previsto durante los primeros 5 años:

<b>BALANCE</b>					
<b>ACTIVO</b>					
<b>Concepto</b>	<b>Año 1</b>	<b>Año 2</b>	<b>Año 3</b>	<b>Año 4</b>	<b>Año 5</b>
Inversión	9.000,00 €	9.000,00 €	9.000,00 €	9.000,00 €	9.000,00 €
Amortización	- 3.000,00 €	- 6.000,00 €	- 9.000,00 €	- 9.000,00 €	- 9.000,00 €
Activo No Corriente	6.000,00 €	3.000,00 €	- €	- €	- €
Efectivo	-20.450,00 €	-16.348,69 €	24.454,16 €	100.949,85 €	213.896,96 €
Activo Corriente	-20.450,00 €	-16.348,69 €	24.454,16 €	100.949,85 €	213.896,96 €
<b>TOTAL ACTIVO</b>	<b>-14.450,00 €</b>	<b>-13.348,69 €</b>	<b>24.454,16 €</b>	<b>100.949,85 €</b>	<b>213.896,96 €</b>
<b>PASIVO</b>					
Capital	40.000,00 €	40.000,00 €	40.000,00 €	40.000,00 €	40.000,00 €
Reservas	- €	-54.450,00 €	-53.348,69 €	-15.545,84 €	60.949,85 €
Resultado del Ejercicio	-54.450,00 €	1.101,31 €	37.802,85 €	76.495,69 €	112.947,11 €
Patrimonio Neto	-14.450,00 €	-13.348,69 €	24.454,16 €	100.949,85 €	213.896,96 €
<b>TOTAL PASIVO</b>	<b>-14.450,00 €</b>	<b>-13.348,69 €</b>	<b>24.454,16 €</b>	<b>100.949,85 €</b>	<b>213.896,96 €</b>

## 8.5. Indicadores económicos

A continuación aportamos los siguientes indicadores para analizar más en detalle la rentabilidad del proyecto empresarial:

- ★ **Valor Actual Neto (VAN):** nos da información sobre el valor, a tiempo presente, de los flujos de caja futuros, teniendo en cuenta una tasa de descuento y habiendo descontado previamente la inversión inicial. La tasa de descuento representa el coste de oportunidad, la rentabilidad mínima esperada. Si un proyecto de inversión presenta un VAN positivo, dicho proyecto se considera rentable. En nuestro caso hemos elegido una del tasa de descuento del 10% y el VAN obtenido es un valor positivo, por lo que el proyecto es rentable.
- ★ **Tasa Interna de Retorno (TIR):** indica la rentabilidad del proyecto. Más concretamente nos da información sobre la tasa de descuento para la que el VAN se hace cero. En nuestro caso, la TIR obtenida es superior a la rentabilidad mínima exigida, por lo que efectivamente estamos ante un proyecto rentable.
- ★ **Tasa Interna de Retorno Modificada (TIRM):** es un método de valoración de inversiones que mide la rentabilidad de una inversión en términos relativos (en porcentaje), cuya principal cualidad es que elimina el problema de la inconsistencia que puede surgir al aplicar la TIR.
- ★ **Plazo de Recuperación (Pay-Back):** el periodo de tiempo requerido para recuperar el capital inicial de una inversión. En nuestro caso el payback obtenido indica que la inversión se recuperará entre el segundo y tercer año.



## 9. Bibliografía

- ★ [Wikipedia] Astrofísica. <https://es.wikipedia.org/wiki/Astrof%C3%ADsica>
- ★ [Wikipedia] Clasificación morfológica de las galaxias.  
[https://es.wikipedia.org/wiki/Clasificaci%C3%B3n\\_morfol%C3%B3gica\\_de\\_las\\_galaxias](https://es.wikipedia.org/wiki/Clasificaci%C3%B3n_morfol%C3%B3gica_de_las_galaxias)
- ★ Sloan Digital Sky Survey. <https://www.sdss.org/>
- ★ SDSS Data Release 15. <http://skyserver.sdss.org/dr15/en/home.aspx>
- ★ EUCLID. <http://sci.esa.int/euclid/>
- ★ DataRobot. <https://www.datarobot.com/>
- ★ Automatic Unsupervised Classification of All SDSS/DR7 Galaxy Spectra, J. Sánchez Almeida, et al. <https://arxiv.org/abs/1003.3186>
- ★ The weirdest SDSS galaxies: results from an outlier detection algorithm, D. Baron et al. <https://arxiv.org/abs/1611.07526>
- ★ Qualitative Interpretation of Galaxy Spectra, J. Sánchez Almeida et al., <https://arxiv.org/abs/1207.3928>
- ★ Machine Learning in Astronomy: A Practical Overview, D. Baron, <https://arxiv.org/abs/1904.07248>
- ★ A Machine Learning Approach to Mass Spectra Classification with Unsupervised Feature Selection, M. Ceccarelli et al., <http://www.isa.cnr.it/dacierno/Paperspdf/cibb08.pdf>
- ★ Automatic spectral classification of stellar spectra with low signal-to-noise ratio using artificial neural networks, S.G. Navarro et al., [https://www.researchgate.net/publication/258561512\\_Automatic\\_spectral\\_classification\\_of\\_stellar\\_spectra\\_with\\_low\\_signal-to-noise\\_ratio\\_using\\_artificial\\_neural\\_networks](https://www.researchgate.net/publication/258561512_Automatic_spectral_classification_of_stellar_spectra_with_low_signal-to-noise_ratio_using_artificial_neural_networks)
- ★ Classification of spectra of emission line stars using machine learning techniques, P. Bromová et al., [https://www.researchgate.net/publication/271917808\\_Classification\\_of\\_Spectra\\_of\\_Emission\\_Line\\_Stars\\_Using\\_Machine\\_Learning\\_Techniques](https://www.researchgate.net/publication/271917808_Classification_of_Spectra_of_Emission_Line_Stars_Using_Machine_Learning_Techniques)
- ★ Clustering with deep learning: Taxonomy and new methods, E. Aljalbout et al., <https://arxiv.org/abs/1801.07648>
- ★ Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization (DEPICT), K. Ghasedi Dizaji et al., <https://arxiv.org/abs/1704.06327>
- ★ Deep clustering with convolutional autoencoders (DCEC), X. Guo et al., [https://www.researchgate.net/publication/320658590\\_Deep\\_Clustering\\_with\\_Convolutional\\_Autoencoders](https://www.researchgate.net/publication/320658590_Deep_Clustering_with_Convolutional_Autoencoders)
- ★ Unsupervised Deep Embedding for clustering analysis, (DEC), J. Xie et al., <https://arxiv.org/abs/1511.06335>
- ★ Variational Deep Embedding: An unsupervised and generative approach to clustering (VaDE), Z. Jiang et al., <https://arxiv.org/abs/1611.05148>
- ★ Machine learning approaches and pattern recognition for spectral data, T. Villmann et al., [https://www.researchgate.net/publication/221165361\\_Machine\\_learning\\_approches\\_and\\_pattern\\_recognition\\_for\\_spectral\\_data](https://www.researchgate.net/publication/221165361_Machine_learning_approches_and_pattern_recognition_for_spectral_data)
- ★ CNN features are also great at unsupervised classification, J. Guërin et al., <https://arxiv.org/abs/1707.01700>
- ★ Convolutional clustering for unsupervised learning, A. Dundar et al., <https://arxiv.org/abs/1511.06241>

- ★ Explanatory analysis of spectroscopic data using machine learning od simple, interpretable rules, R. Goodacre,  
<https://www.sciencedirect.com/science/article/abs/pii/S0924203103000456?via%3Dihub>
- ★ Deep Learning with Keras, A. Gulli,  
<https://www.amazon.es/Deep-Learning-Keras-Antonio-Gulli/dp/1787128423>
- ★ Advanced Deep Learning with Keras, R. Atienza,  
<https://www.amazon.com/Advanced-Deep-Learning-Keras-reinforcement-ebook/dp/B078N8RDCP>
- ★ TP Analyse de spectres avec SpexStage de MeudonCorrigéM. Puech – M. Rodrigues  
<https://owncloud.iac.es/index.php/s/qLg1rqvwLGZbJgJ>
- ★ QUALITATIVE INTERPRETATION OF GALAXY SPECTRA J. Sánchez Almeida, R. Terlevich E. Terlevich, R. Cid Fernandes, and A. B. Morales-Luis  
<https://owncloud.iac.es/index.php/s/fFAdXJyuKSPg2qN>
- ★ Statics, Data Mining, and Machine Learning in Astronomy. A Practical Python Guide for the Analisis of Survey Data. Željko Ivezić, Andrew J. Connolly, Jacob T. VanderPlas, and Alexander Gray  
<https://owncloud.iac.es/index.php/s/4uCVI8I7hhFpOcq>
- ★ MACHINE LEARNING IN ASTRONOMY: A PRACTICAL OVERVIEW. Dalya Baron  
[http://research.iac.es/winterschool/2018/media/summaries/ml\\_summary\\_dbaron.pdf](http://research.iac.es/winterschool/2018/media/summaries/ml_summary_dbaron.pdf)
- ★ FITS File handling (astropy.io.fits)  
<http://docs.astropy.org/en/stable/io/fits/>
- ★ Astropy Tutorials. Lia R. Corrales  
<http://learn.astropy.org/FITS-images.html>
- ★ Análisis y visualización de datos con Pandas & Matplotlib. Stephanie Frias  
<https://code.likeagirl.io/an%C3%A1lisis-y-visualizaci%C3%B3n-de-datos-con-pandas-matplotlib-85ee4d7b4cad>
- ★ Pandas DataFrame Reference  
<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html>
- ★ Pandas en Python, con ejemplos. Ricardo Moya.  
<https://jarroba.com/pandas-python-ejemplos-parte-i-introduccion/>
- ★ Set up a Windows Data Science Virtual Machine on Azure  
<https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/provision-vm>
- ★ Variational autoencoders. Jeremy Jordan  
<https://www.jeremyjordan.me/variational-autoencoders/>
- ★ Variational Autoencoders Explained.  
<http://anotherdatum.com/vae.html>
- ★ Python: 3 Manuscripts in 1 book: - Python Programming For Beginners - Python Programming For Intermediates - Python Programming for Advanced  
<https://www.amazon.es/Python-Manuscripts-Programming-Beginners-Intermediates-ebook/dp/B07CQPHC1N/?tag=hrefdi-21&ie=UTF8>
- ★ Diving into Gaussian Mixture Modeling  
<https://medium.com/@pmdev/diving-into-gaussian-mixture-modeling-b87976081097>
- ★ Unsupervised Learning and Data Clustering  
<https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a>
- ★ Unsupervised Learning of Gaussian Mixture Models on a SELU auto-encoder (Not another MNIST)

- <https://towardsdatascience.com/unsupervised-learning-of-gaussian-mixture-models-on-a-selu-auto-encoder-not-another-mnist-11fceccc227e>
- ★ Gaussian Mixture Model clustering: how to select the number of components (clusters)  
<https://towardsdatascience.com/gaussian-mixture-model-clusterization-how-to-select-the-number-of-components-clusters-553bef45f6e4>
  - ★ Gaussian Mixture Models Clustering Algorithm Explained  
<https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e>
  - ★ 10 Tips for Choosing the Optimal Number of Clusters  
<https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92>
  - ★ Gaussian Mixture Models Explained  
<https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95>
  - ★ Gaussian Mixture Modelling (GMM)  
<https://towardsdatascience.com/gaussian-mixture-modelling-gmm-833c88587c7f>
  - ★ 101 Machine Learning Algorithms for Data Science with Cheat Sheets  
<https://www.r-bloggers-com.cdn.ampproject.org/c/s/www.r-bloggers.com/101-machine-learning-algorithms-for-data-science-with-cheat-sheets/amp/>
  - ★ AI and Analytics in Production. Ted Dunning & Ellen Friedman  
[https://get.oreilly.com/ind\\_ai-and-analytics-in-production.html?utm\\_medium=email&utm\\_source=topic+optin&utm\\_campaign=aieu19&utm\\_content=nem3a+ai+and+analytics+in+production](https://get.oreilly.com/ind_ai-and-analytics-in-production.html?utm_medium=email&utm_source=topic+optin&utm_campaign=aieu19&utm_content=nem3a+ai+and+analytics+in+production)

## 10. Anexos

- ★ Anexo de Gestión
- ★ Anexo Técnico