



**UNIVERSIDAD DE CASTILLA-LA MANCHA**

**ESCUELA SUPERIOR DE INGENIERÍA INFORMÁTICA**

**MÁSTER UNIVERSITARIO EN INGENIERÍA  
INFORMÁTICA**

**TRABAJO FIN DE MÁSTER**

**Servicio de análisis de cambios en repositorios de código para la  
identificación de fallos potenciales**

Diego Fermín Sanz Alonso

Mayo de 2019





**UNIVERSIDAD DE CASTILLA-LA MANCHA**

**ESCUELA SUPERIOR DE INGENIERÍA INFORMÁTICA**

**MÁSTER UNIVERSITARIO EN INGENIERÍA  
INFORMÁTICA**

**TRABAJO FIN DE MÁSTER**

**Servicio de análisis de cambios en repositorios de código para la  
identificación de fallos potenciales**

Autor: Diego Fermín Sanz Alonso

Directores: Pablo Bermejo López

Luis de la Ossa Jiménez

Mayo de 2019



## Resumen

Durante el desarrollo de una aplicación software en equipo, es muy común utilizar servicios de control de versiones como GitHub o BitBucket. Cuando un miembro del equipo aporta un incremento en alguna de las ramas del repositorio, esto lanza un proceso automático de Integración (CI) y Prueba Continua (CT). Cuando el proceso de build en CI o durante las pruebas en CT es posible que se produzca un error, entonces decimos que la pipeline se ha roto, y todo el equipo debe parar su trabajo y no enviar ningún incremento nuevo al repositorio hasta que el programador arregle el código que ha roto el proceso de CI o CT.

Este TFM asume que algunos de los commits de código que acaban rompiendo el pipeline pueden tener características que los diferencian de los cambios que no, como por ejemplo el número de archivos actualizados, o el Volumen de código añadido. Por ello, se pretende mostrar al usuario la información más relevante para la detección de fallos, a modo de advertencia, antes de enviar el nuevo incremento al repositorio compartido. Así, se podría evitar alertar o molestar a todo el equipo involucrado.

A partir de todo ello se desplegará un servicio al que el usuario pueda subir la última versión de su repositorio local y compararlo con el último push realizado al repositorio compartido online, de manera que sepa si dicho commit es probable que falle durante el build y las pruebas o no.

En este trabajo se explicarán todas las fases de creación del servicio, desde la recogida y el análisis de los datos hasta la implementación y el despliegue del servicio en la nube. Además, durante el desarrollo del proyecto se seguirá una metodología Scrum y se cumplirá con la Primera Vía de la filosofía DevOps, buscando entregar valor al cliente lo más rápido posible.



# Agradecimientos

Agradecimientos





# Índice de contenido

Resumen .....	v
Agradecimientos.....	vii
<b>CAPÍTULO 1 Introducción.....</b>	<b>1</b>
1.1 Motivación .....	1
1.2 Método y fases de trabajo .....	2
1.3 Estructura de la memoria .....	3
<b>CAPÍTULO 2 Estado del arte .....</b>	<b>5</b>
2.1 Sistemas de control de versiones.....	5
2.2 Integración Continua, Entrega Continua y Despliegue Continuo .....	7
2.3 Predicción de fallos mediante análisis de código.....	9
2.4 Conclusiones .....	11
<b>CAPÍTULO 3 Servicio de análisis .....</b>	<b>13</b>
3.1 Descripción del servicio .....	13
3.2 Selección de métricas a mostrar .....	16
3.3 Modelo de predicción.....	17
3.4 Conclusiones .....	18
<b>CAPÍTULO 4 Despliegue Continuo .....</b>	<b>19</b>
4.1 Desarrollo.....	20
4.2 Integración Continua.....	20
<b>CAPÍTULO 5 Análisis de datos .....</b>	<b>25</b>
5.1 Variable clase .....	25
5.2 Variables predictoras.....	27
<b>CAPÍTULO 6 Conclusiones y propuestas.....</b>	<b>29</b>
6.1 Conclusiones .....	29
6.2 Trabajo futuro.....	29
<b>Bibliografía.....</b>	<b>31</b>
<b>Anexos.....</b>	<b>35</b>
A.1. Anexo 1: Metodología de trabajo.....	35

A.2.	Dockerfiles para la construcción de los contenedores.....	42
A.3.	Contenido del Docker Compose.....	43
A.4.	Archivo de configuración de Travis .....	44

# Índice de figuras

IMAGEN 2.1: ESQUEMA GENERAL DE UN CONTROL DE VERSIONES DISTRIBUIDO.....	7
IMAGEN 2.2: DIFERENCIA ENTRE INTEGRACIÓN CONTINUA, ENTREGA CONTINUA Y DESPLIEGUE CONTINUO.....	8
IMAGEN 3.1: FLUJO DE TRABAJO COMÚN - SIN UTILIZAR NUESTRO SERVICIO.....	13
IMAGEN 3.2: FLUJO DE TRABAJO COMÚN - UTILIZANDO NUESTRO SERVICIO .....	14
IMAGEN 3.3: CONTEXTO DE USO DE NUESTRO SERVICIO .....	15
IMAGEN 3.4: PANTALLA INICIAL DEL SERVICIO .....	15
IMAGEN 3.5: PANTALLA DE RESULTADOS.....	16
IMAGEN 3.6: PANTALLA DE PREDICCIÓN DE FALLOS - VERSIÓN BETA .....	18
IMAGEN 4.1: PIPELINE DE DESARROLLO.....	19
IMAGEN 4.2: VISUALIZACIÓN DE INCIDENCIAS EN ZENHUB.....	21
IMAGEN 4.3: EJEMPLO BUILD EN TRAVIS CI.....	21
IMAGEN 4.4: DOCKER COMPOSE. ESQUEMA DE LOS CONTENDORES .....	23
IMAGEN 4.5: RESULTADO DEL CD EN TRAVIS .....	24
IMAGEN 5.1: DISTRIBUCIÓN DE LA VARIABLE "NÚMERO DE BUGS" .....	26
IMAGEN 5.2: DISTRIBUCIÓN DE LA VARIABLE "NÚMERO DE BUGS" EN ESCALA LOGARÍTMICA .....	26
IMAGEN 1.1: CONTROL FLOW DEL PROYECTO .....	37



# CAPÍTULO 1

## INTRODUCCIÓN

---

### 1.1 MOTIVACIÓN

Durante el desarrollo de una aplicación software en equipo, es muy común utilizar servicios de control de versiones como GitHub<sup>1</sup> o BitBucket<sup>2</sup>. A partir de estos repositorios, el equipo de desarrollo puede obtener una copia con el código en un momento determinado y trabajar sobre él, para posteriormente subir sus cambios al repositorio remoto, con el fin de que los demás tengan acceso a dichos cambios. Sin embargo, en función de la complejidad del proyecto, el tamaño del equipo y la agilidad que se quiera llevar en el proyecto, la integración del código mediante las técnicas tradicionales puede conllevar conflictos y generar problemas que supongan tiempo de desarrollo perdido.

Por ello, para evitar estos contratiempos y a la vez asegurar la calidad del código, se realiza lo que se conoce como Integración Continua [1], que consiste en que, cada vez que un desarrollador añada una funcionalidad nueva, esta se compile, se realice el build y se ejecuten un conjunto de pruebas automáticas con el fin de determinar si el funcionamiento de todos los módulos del proyecto es el esperado, de manera que siempre se esté trabajando sobre una versión estable del proyecto.

Si se sigue esta filosofía, cuando un miembro del equipo aporta un incremento en alguna de las ramas del repositorio, esto lanza un proceso automático de Integración (CI) y Prueba Continua (CT). Cuando se lanza el proceso de build en CI o durante las pruebas en CT es posible que se produzca un error, entonces decimos que la pipeline se ha roto, y todo el equipo debe parar su trabajo y no enviar ningún incremento nuevo al repositorio hasta que el programador arregle el código que ha roto el proceso de CI o CT.

Este TFM asume que algunos de los commits de código que acaban rompiendo el pipeline pueden tener características que los diferencian de los cambios que no, como por ejemplo el número de archivos actualizados, o el volumen de código añadido [2]. Por ello,

<sup>1</sup> <https://github.com/>

<sup>2</sup> <https://bitbucket.org/>

## CAPÍTULO 1: Introducción

se pretende mostrar un análisis descriptivo de las clases modificadas en un commit para que el usuario pueda decidir a raíz de estos datos si prefiere volver a revisar su código o si detecta, por ejemplo, que no cumple con los estándares de calidad del equipo. Asimismo, vamos a analizar si es posible crear un modelo que aprenda y pueda avisar al programador de estos 'commits smells', a modo de advertencia, antes de enviar el nuevo incremento al repositorio compartido. Así, se podría evitar alertar o molestar a todo el equipo involucrado.

Diversos artículos que apoyan esta teoría suelen basarse en entrevistas o encuestas para obtener feedback sobre los resultados de los commits [3]. En nuestro caso, utilizaremos la información de los propios commits, realizando un análisis estático del código modificado para poder extraer sus características y predecir si es probable que tenga fallos.

A partir de todo ello desplegará un servicio al que el usuario pueda subir la última versión de su repositorio local y compararlo con el último *push* realizado al repositorio compartido online, de manera que sepa si dicho commit es probable que falle durante el *build* y las pruebas o no. Los fallos que se intentarán detectar, por tanto, son exclusivamente los fallos de carácter funcional, ya que los fallos de compilación se detectarían automáticamente antes de que comience a hacerse el *build*.

### 1.2 MÉTODO Y FASES DE TRABAJO

Las fases de trabajo se pueden resumir en 4 fases principales, que son:

1. Adquisición de los datos.
2. Realizar un análisis estático de los cambios de código en un repositorio.
3. Tratar de crear un modelo predictivo a partir de la información extraída.
4. Creación de un servicio a partir las dos fases anteriores.

El desarrollo del proyecto seguirá la metodología de desarrollo Scrum [4], siguiendo recomendaciones para su adaptación en proyectos de Ciencia de Datos [5]. Dentro de esta metodología, los tutores ejercerán los papeles de Scrum Master, mientras que el alumno ejercerá las funciones del equipo de desarrollo y del Product Owner. El Product Owner es el encargado de crear y ordenar por prioridad las tareas del Product Backlog, así como de mantener el Burndown Chart, y por ende es la persona con una visión más global del proyecto, por lo que el alumno es el que puede hacer mejor este papel. Los tutores, por su parte, al tomar el papel de Scrum Masters serán los encargados de asegurarse de que se siguen las prácticas descritas en la metodología y de eliminar cualquier impedimento que pueda ir surgiendo debido a esta elección en la metodología:

- Se realizarán Sprints quincenales, juntando la Sprint Review y la Retrospectiva

del sprint que termina, con el Sprint Planning del nuevo sprint.

- Cada uno de los sprints quedará documentado, con la fecha de las reuniones y un resumen de los temas tratados en las mismas.
- Se mantendrá un dashboard online para dar transparencia al backlog del producto y del sprint.

Además, el marco de trabajo cumplirá con la Primera Vía de DevOps [1]. Esta parte de la filosofía DevOps busca agilizar lo máximo posible el flujo de trabajo para así entregar valor a los clientes cuanto antes. Para ello, se aplicarán los siguientes puntos:

- Control de versiones, no solo del código fuente y sus dependencias, sino también de las bases de datos y el modelo creado, siguiendo la actual corriente ‘Machine Learning is code’ [6] [7].
- CI enlazada con el repositorio de control de versiones, y CT.
- En el repositorio siempre se encuentra la última versión de los cuadernos Jupyter con los que realizamos los distintos análisis de datos y creación de modelos, tal y como se aconseja en [5].
- Posible despliegue automático (CD) una vez creado el modelo.



### 1.3 ESTRUCTURA DE LA MEMORIA

Este trabajo se dividirá en un total de 5 capítulos, incluyendo este capítulo introductorio.

El segundo capítulo hablará del estado del arte. En este capítulo entraremos más en detalle en el problema concreto que queremos abordar: avisar al programador de posibles fallos de código en un commit antes de realizar ningún tipo de compilación o prueba automática. Describiremos algunas de las vías de estudio que se han realizado para resolver este problema, así como la fuente de la cuál se han extraído los datos que usaremos de cara a la predicción de fallos.

En el siguiente capítulo se abordará la creación del servicio. En este capítulo explicaremos en qué consiste, en qué fase de trabajo de un equipo de desarrollo se puede utilizar y qué información se muestra al usuario. Además, explicaremos los diferentes

## CAPÍTULO 1: Introducción

modelos de predicción que hemos creado para tratar de analizar de manera automática si un cambio de código puede tener fallos o no.

El cuarto capítulo explica cómo se ha creado el servicio y qué métodos se han llevado a cabo para crear una pipeline de CI/CD activa cada vez que hay un incremento funcional de código.

En el quinto y último capítulo se extraerán diferentes conclusiones tras el diseño e implementación de este servicio. Adicionalmente, se presentarán diversas mejoras que pueden desarrollarse en trabajos futuros.

Por último, se incluirá un apéndice en el que se detalle la metodología que hemos utilizado, la cual se basa en Scrum pero ha tenido que ser ligeramente adaptada debido a las características propias de un trabajo de fin de máster y al contexto de la ciencia de datos, algo diferente a la construcción ágil de software por incrementos.



# CAPÍTULO 2

## ESTADO DEL ARTE

---

En este capítulo se explicarán los componentes imprescindibles en el desarrollo ágil de software durante los últimos años, como son los sistemas de control de versiones y las técnicas de Integración Continua, Entrega Continua y Despliegue Continuo. A continuación, se detallarán diferentes estudios que giran en torno al análisis de código y la predicción de fallos en función de los cambios en el mismo. Finalmente, se mostrará la fuente a partir de la cual hemos realizado nuestros análisis, explicando cómo se han conseguido.

### 2.1 SISTEMAS DE CONTROL DE VERSIONES

En términos generales, los proyectos de desarrollo software son realizados por más de una persona, en ocasiones alejados geográficamente. E incluso si estamos en el mismo centro de trabajo, es recomendable que cada persona pueda desarrollar el proyecto de manera independiente, y posteriormente tener la capacidad de combinar el trabajo del equipo de una manera ágil. Esto es controlado por el proceso conocido como control de versiones.

Los sistemas de control de versiones son herramientas software que ayudan a un equipo software a gestionar los cambios de código a través del tiempo [8]. Estos sistemas monitorizan los cambios de código que se realizan en cada modificación, de manera que el programador puede deshacer sus cambios si ha detectado un error, así como comparar versiones previas del código sin tener que interrumpir el trabajo de todos sus compañeros.

Estos sistemas, por tanto, tienen tres características fundamentales que los hacen casi imprescindible para el flujo de trabajo de casi cualquier proyecto software [8]. En primer lugar, mantienen un registro a largo plazo del histórico de cada archivo, lo que nos ayuda a tener una visión general de los cambios del proyecto. Esta característica es muy útil para volver a versiones anteriores con el fin de detectar errores en la aplicación, o para poder arreglar errores en versiones anteriores del proyecto.

Otra de las ventajas del control de versiones es que permiten diversificar el desarrollo, mediante un proceso conocido como ramificación, en el que un miembro del equipo puede, a partir de una versión específica, trabajar en una nueva funcionalidad o arreglar un fallo sin

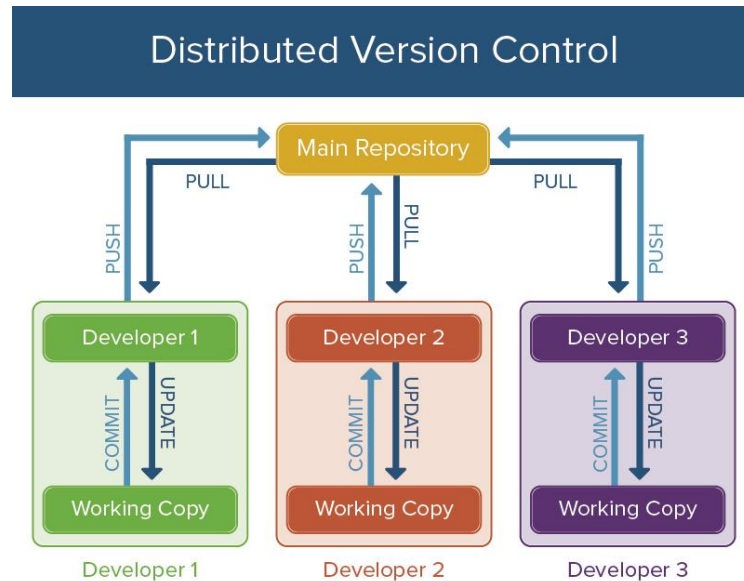
tener que preocuparse de que otro miembro del equipo esté realizando cambios de manera simultánea. Una vez se realizan los cambios en una rama, se debe hacer el proceso de fusión, que consiste en unir los cambios y verificar que los cambios de las dos ramas que se intentan unir no tienen conflictos. Un conflicto es una modificación de la misma parte del código por parte de dos ramas diferentes. Si esto ocurre, es necesario resolver el conflicto manualmente, decidiendo de qué manera se combinan ambas modificaciones.

La tercera ventaja más importante de estas herramientas es la trazabilidad. Al llevar un registro de cada cambio de software, es muy sencillo conectar esta información con herramientas de gestión de proyectos y de gestión de *bugs* o errores, con el fin de poder documentar el propósito de cada cambio, de manera que es posible tener una visión más global de los cambios realizados sin tener que leer el código fuente para entender las nuevas funcionalidades o los *fix* realizados.

Dentro de los sistemas de control de versiones, existen dos grupos bien diferenciados [9]. Por un lado, tenemos los sistemas de control de versiones centralizados. En este tipo de herramientas se mantiene una copia central del proyecto en un único repositorio, y los cambios que realizan los desarrolladores lo aplican sobre esa versión central. Uno de los sistemas más conocidos es Subversion (SVN) [10]. Sin embargo, estos sistemas tienen un problema cuando dos desarrolladores realizan cambios simultáneos, ya que uno puede sobrescribir el trabajo del otro antes de que el resto del equipo pueda visualizar esos cambios.

Por ello, son muchos más frecuentes los sistemas de control de versiones distribuidos. Como se puede ver en la Imagen 2.1 [9], la principal diferencia con los anteriores radica en que no hay un repositorio central para los cambios de información. En su lugar, cada desarrollador tiene su propio repositorio (que es una copia de alguna versión del repositorio principal) y realiza los cambios a partir de él. De esta manera no se sobrescribe el trabajo de otros, ya que cada uno tiene su propia copia local, y una vez se quiere llevar el trabajo al repositorio principal se puede comprobar si hay conflictos, código fuente cambiado por dos o más personas, para poder realizar los cambios adicionales pertinentes. Existen muchos

sistemas distribuidos de control de versiones, aunque el más conocido y usado con diferencia es Git [11].



*Imagen 2.1: Esquema general de un control de versiones distribuido*

## 2.2 INTEGRACIÓN CONTINUA, ENTREGA CONTINUA Y DESPLIEGUE CONTINUO

Los sistemas de control de versiones han ayudado mucho a la agilidad de proyectos software, ya que todos los miembros del equipo pueden realizar modificaciones de manera simultánea. Sin embargo, existe un problema derivado de esta concurrencia: no siempre sabemos qué cambios ha realizado la otra persona. Por lo tanto, al margen de los errores propios del desarrollo software que existen desde siempre, con este modelo podemos encontrar errores debido a cambios que realiza un miembro del equipo sobre una fracción del código de la que se nutren parte de los cambios añadido por otro miembro.

Al margen de este problema surge la necesidad de disponer de herramientas que garanticen la calidad del software de manera automática, ya que los proyectos tienden a crecer y hacer esto de manera manual sería cada vez más complicado. A raíz de esta problemática, en las prácticas modernas de desarrollo se han incorporado tres conceptos o técnicas: Integración Continua, Entrega Continua y Despliegue Continuo [12].

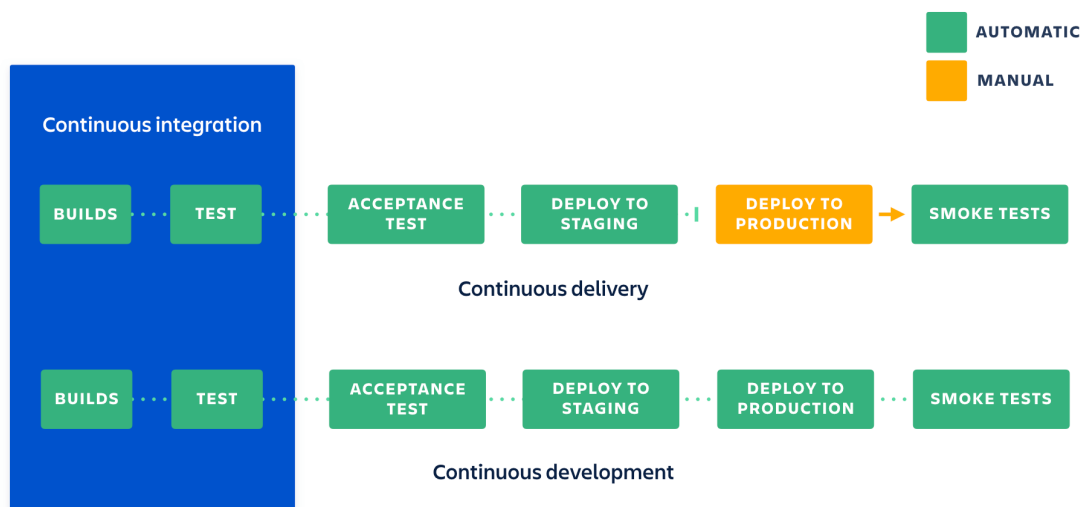
La Integración Continua o *Continuous Integration (CI)* es la solución más directa al problema planteado anteriormente. Esta técnica trata de hacer que los desarrolladores incorporen sus cambios en la rama principal con la mayor frecuencia posible. Cuando un

desarrollador hace los cambios, estos se validan creando una compilación y sometiéndola a pruebas automatizadas.

La Entrega Continua o *Continuous delivery* (CD) es una extensión de la Integración Continua, y amplía la automatización a la preparación de la aplicación del despliegue, de manera que siempre se pueda desplegar una versión estable de la aplicación.

El último concepto relacionado con estas prácticas es la que se conoce como Despliegue Continuo o *Continuous Deployment* (CD). Aquí se da un paso más en la automatización, ya que los proyectos que aplican Despliegue Continuo, cada vez que los cambios de un desarrollador pasan todas las pruebas del pipeline o flujo de trabajo, se despliegan de manera automática para que sean visibles de manera inmediata para el cliente final.

En la Imagen 2.2 [12] se aprecia la diferencia entre estos tres términos. La Integración Continua se refiere únicamente a validar los cambios de manera automática, mientras que los otros dos conceptos son más amplios e incluyen los distintos procesos de preparación del despliegue. La Entrega Continua cede el despliegue final a una decisión manual, mientras que el Despliegue Continuo automatiza incluso ese último paso.



*Imagen 2.2: Diferencia entre Integración Continua, Entrega Continua y Despliegue Continuo*

Aunque gracias a estas técnicas, que cada vez se implantan en más empresas, es posible mantener la calidad del software sin mermar notablemente la agilidad de los proyectos, surge un problema derivado de las pruebas automáticas. Si se sigue esta filosofía, cuando un miembro del equipo aporta un incremento en alguna de las ramas del repositorio, esto lanza

un proceso automático de Integración Continua. Cuando se lanza el proceso de build en CI es posible que se produzca un error. Es entonces cuando decimos que la pipeline se ha roto, y todo el equipo debe parar su trabajo y no enviar ningún incremento nuevo al repositorio hasta que el programador arregle el código que ha roto el proceso de CI. Además, en proyectos grandes el proceso de compilación y aceptación de las pruebas automáticas puede requerir un tiempo destacable, por lo que sería conveniente saber de antemano si es probable que los cambios contienen errores.

Debido a esto, una vía de investigación que puede beneficiar todavía más a la agilidad de los proyectos software modernos consiste en, una vez realizados los cambios por parte de un desarrollador, y antes de lanzar el proceso de CI, tener alguna manera de saber si es posible que existan fallos para revisarlos antes de lanzar estos procesos automáticos. Una de las alternativas para ello podría ser el análisis del código modificado, con el fin de extraer diferentes métricas y poder hacernos una idea de si los cambios pueden provocar interrupciones del pipeline.

### 2.3 PREDICCIÓN DE FALLOS MEDIANTE ANÁLISIS DE CÓDIGO

El análisis de los cambios de código en los equipos de desarrollo es un campo que se ha estudiado desde perspectivas muy diversas. Sin embargo, en términos generales podemos afirmar que este análisis es meramente exploratorio, llegando a distintas conclusiones a partir de encuestas o de la observación de distintos proyectos para ver sus puntos en común.

Existen estudios, por ejemplo, de la relación entre los patrones que se aprecian en los commits de los desarrolladores con la evolución del código de un proyecto. En el artículo [13] se determinan cuatro métricas para medir la actividad respecto a los commits y la evolución del código: cambios del commit, tiempo entre commits, autor del cambio y la evolución del código. A partir de estas cuatro métricas, extraen diferentes conclusiones, como qué cambios en clases que se utilizan en muchas partes del código provocan cambios en un número de archivos mayor.

Uno de los activos más importantes que podemos tener para tratar de resolver nuestro problema es la cantidad de datos que tenemos de manera abierta en repositorios públicos como GitHub. Proyectos como GHTorrent [14] se han encargado de recoger información relativa a los repositorios públicos de GitHub y la información de los usuarios durante más de un año. Esta base de datos presenta más de 900GB de información, y 10GB de metadatos, incluyendo casi 800.000 usuarios, más de 1,3 millones de repositorios y casi 30 millones de commits, y con el tiempo esta cifra aumenta, ya que se sigue actualizando con los eventos más recientes [15].

Sin embargo, necesitamos también información relativa a cuándo el estado del código en un commit concreto ha provocado un fallo o no. Existen otros artículos que analizan la categorización de los *issues* de GitHub a partir de las etiquetas asociadas en los distintos proyectos [16]. No obstante, en [16] se destaca que, aunque el uso de etiquetas favorece la resolución de issues en un repositorio, este sistema de etiquetado es muy poco utilizado, y la manera en que la usan los repositorios varía en cada caso, lo que hace complicado su análisis.

Otra posible vía para tratar de encontrar los commits que han provocado un fallo de código son las técnicas de Integración Continua (CI). Cuando un desarrollador realiza un cambio de código y existe alguna herramienta de CI, automáticamente se lanzan una serie de pruebas tras la subida de ese nuevo código. Si las pruebas fallan, entonces decimos que la pipeline se ha roto, y todo el equipo debe parar su trabajo y no enviar ningún incremento nuevo al repositorio hasta que el programador arregle el código que ha roto el proceso de CI. Por ello, sería posible recoger información de bases de datos de los resultados de las pruebas de Integración Continua y cruzarlos con la información de los commits. Una base de datos válida para este tipo de estudios es TravisTorrent [17], que recopila el resultado de las pruebas realizadas por Travis-CI [18], un servicio de Integración Continua. En esta base de datos se puede observar el resultado de las pruebas y los commits que intervienen en cada subida de código, por lo que se podría cruzar con la información disponible en GHTorrent para estudiar qué clases han provocado fallos en cada subida de código. No obstante, esta vía de estudio tiene dos problemas principales: la enorme cantidad de información que se manejan en ambos casos; y los pocos proyectos que aparecen en ambas bases de datos para poder cruzar la información.

Existen otros estudios que hacen un análisis que buscan el mismo objetivo que el que nos hemos planteado. Un ejemplo lo encontramos en [19], donde se explica cómo se estudian varios proyectos de GitHub y, a partir de la información presente en los distintos commits de cada uno de ellos, se puede dilucidar qué clases han tenido bugs y en cuáles no se han detectado errores hasta ahora. A partir del estudio hecho en este artículo se ha generado una base de datos pública en la que se presentan las distintas clases de un conjunto de proyectos calificados en función de si presentan bugs o no [20].

Entrando un poco más en profundidad en [19] [20], el proceso que los autores han seguido para obtener la base de datos de *bugs* tiene varias partes. En primer lugar, se han seleccionado los proyectos, un total de 13 diferentes, todos ellos en lenguaje Java y en general con una gran cantidad de código, ya que son los más útiles para este tipo de análisis. Además, se han buscado proyectos con un número adecuados de commits que utilicen la etiqueta bug propia de GitHub, para poder diferenciar los cambios de código relacionado con errores y los que no, así como para permitir referenciar qué parte del código se ha modificado para solucionar un error. Dentro de los proyectos que cumplían estos requisitos, se han priorizado los que estaban en activo en el momento del análisis.

Una vez seleccionados los proyectos, mediante la API de GitHub se guardaron sus datos. Estos incluyen los usuarios que colaboran en el repositorio, los issues abiertos y cerrados y los datos de todos los commits. Entre todos los commits guardados se realizó un filtrado para seleccionar solo los relacionados con errores, mediante el uso de la etiqueta bug.

Por otra parte, se realizó una descarga de los repositorios en sus distintas versiones o releases. Para cada una de ellas, se realizó un análisis de código estático mediante SourceMeter [21]. Por último, se cruzó esta información con los commits extraídos en el paso anterior, para saber qué partes del código se modificaron para solucionar la incidencia.

Tras todo este análisis, los autores crearon en [20] una base de datos pública con versiones de un conjunto de proyectos en intervalos de 6 meses, y tras descartar aquellas versiones en las que no había errores suficientes. Para cada una de ellas tenemos las distintas clases que han sido modificadas entre versiones, y el número de *bugs* presentes en cada una de ellas. A partir de los datos que estos autores nos proporcionan, es posible intentar crear un modelo de predicción que aprenda a partir de estos datos para poder predecir en futuros commits de cualquier proyecto si un commit puede contener algún error.

## 2.4 CONCLUSIONES

Tras estudiar diversas fuentes de estudio de análisis estático de código, y de búsqueda de fallos potenciales a partir de cambios en el mismo, se ha encontrado una fuente de datos que incluye diferentes métricas de código estático, combinado con la presencia de errores o no, discriminando por cada clase del proyecto. A partir de estos datos, se realizará el análisis posterior y se seleccionará qué información se muestra al usuario en el servicio desarrollado.





# CAPÍTULO 3

## SERVICIO DE ANÁLISIS

En esta sección vamos a explicar el servicio ~~que se va a ofrecer~~ y en qué fase del ~~flujo de trabajo~~ se introduce para facilitar la tarea a los equipos de trabajo de desarrollo software. Además, se explicará qué procedimiento se ha escogido para seleccionar la información que se va a mostrar al usuario, para finalmente explicar los distintos modelos de predicción que hemos intentado crear para realizar esta búsqueda de fallos potenciales de manera automática.

### 3.1 DESCRIPCIÓN DEL SERVICIO

En la Imagen 3.1 se puede observar el flujo de trabajo típico para las empresas que siguen la filosofía *DevOps* y aplican técnicas automáticas de Integración Continua, en el círculo verde, y de despliegue continuo, en el círculo azul [22]. En general, los desarrolladores suben sus cambios al repositorio del equipo. A continuación, si se utilizan técnicas de Integración Continua o CI se compila el proyecto y se pasan una serie de pruebas automatizadas. Si todas las pruebas son correctas, acaba la fase de CI, y a partir de aquí podemos, o bien desplegar manualmente en el momento que deseemos, si seguimos las tendencias clásicas, o bien comenzar el proceso de Despliegue Continuo o CD, que comienza las operaciones necesarias para desplegar una nueva versión visible para el usuario final.

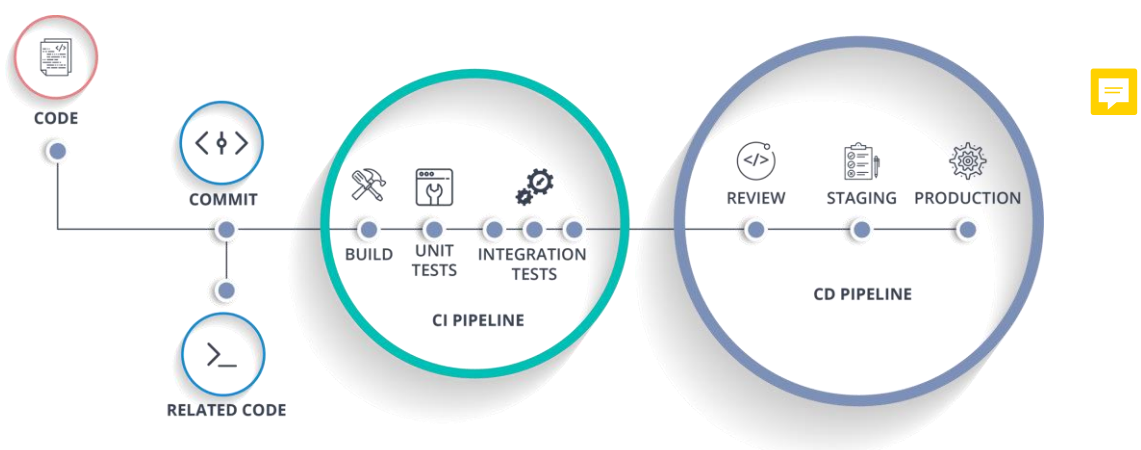
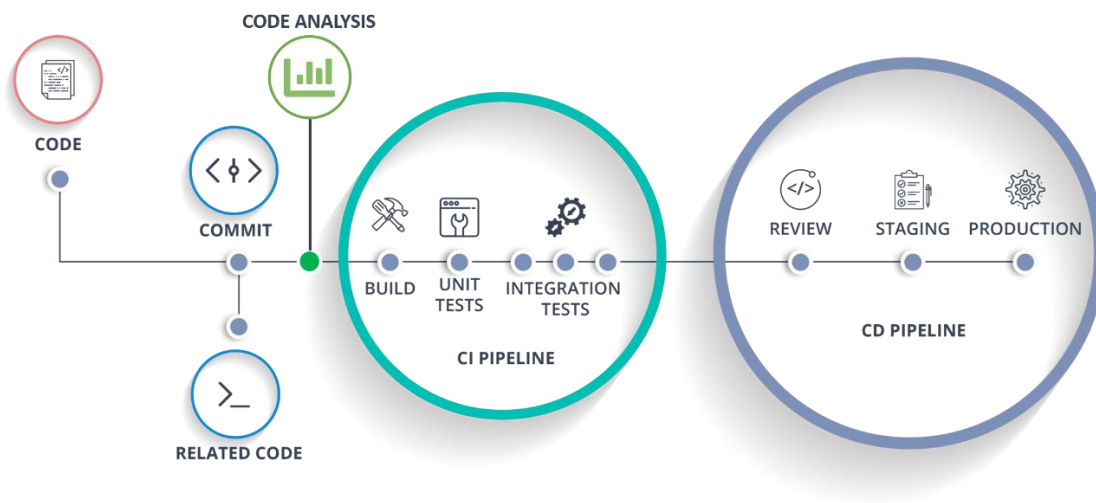


Imagen 3.1: Flujo de trabajo común - Sin utilizar nuestro servicio

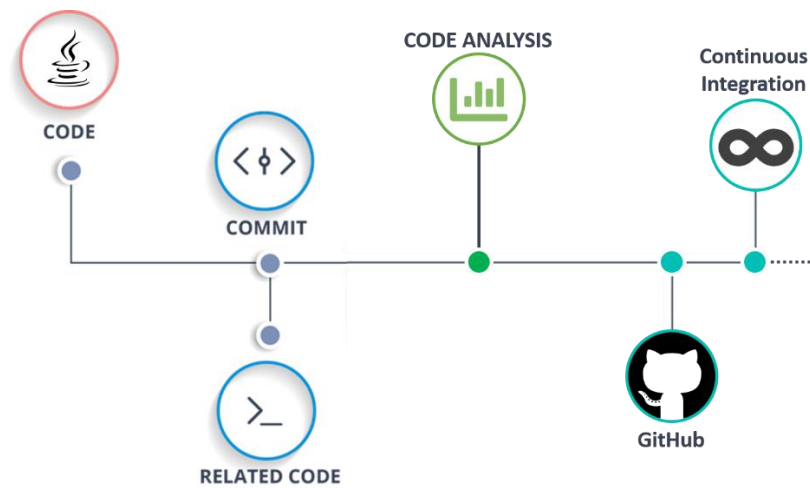
Sin embargo, el flujo anterior tiene un inconveniente, y es que generalmente no hay un solo desarrollador intentando subir sus cambios para añadir nuevas funcionalidades o arreglar fallos, sino que será un equipo entero el que trate de enviar sus modificaciones. Si uno de estos envíos falla, entonces se dice que la *pipeline* se ha roto, y todo el equipo debe parar su trabajo y no enviar ningún incremento nuevo al repositorio hasta que el programador arregle el código que ha roto el proceso de CI. Por ello, sería conveniente tener información adicional antes de lanzar este proceso, con el fin de saber antes de romper el flujo de trabajo de todo el equipo que las modificaciones incorporadas pueden provocar fallos.

Teniendo en cuenta este problema, nuestro servicio se incorporaría justo antes de lanzar el proceso de Integración Continua. En la Imagen 3.2 podemos ver el nuevo flujo de trabajo, con la inclusión de nuestro servicio en color verde. Los desarrolladores, una vez tengan sus cambios, podrán utilizar el servicio para recibir información acerca del código que han modificado. Nuestro servicio comprueba qué partes del proyecto se han modificado y ofrece distintas métricas calculadas a partir de un análisis estático del código que ha sufrido cambios. Tras ver estos datos, los desarrolladores pueden continuar el flujo de trabajo y subir sus cambios para comenzar el proceso de Integración Continua, o si observan que los datos son diferentes a los valores usuales o a los aceptados en los estándares de calidad de la empresa, revisar sus cambios sin interrumpir al resto del equipo.



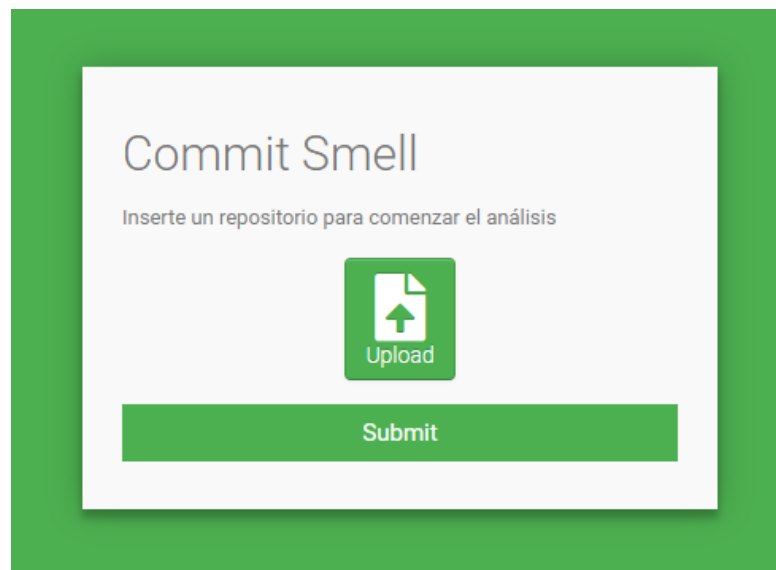
*Imagen 3.2: Flujo de trabajo común - Utilizando nuestro servicio*

En la Imagen 3.3 se puede apreciar más detalladamente en qué situación se puede utilizar nuestro servicio. El servicio presenta también algunas limitaciones. Por ejemplo, como se aprecia en el primer elemento del diagrama, el código analizado solo puede provenir del lenguaje de programación Java, ya que es el mejor estructurado y el que resulta más sencillo para analizar estáticamente. Además, otro requisito de nuestro servicio es que el proyecto debe estar alojado en GitHub, ya que el servicio estudia, a partir de su sistema de control de versiones, qué código se ha modificado.



*Imagen 3.3: Contexto de uso de nuestro servicio*

Este servicio se proporciona vía web, de manera que cualquier usuario pueda enviar el estado de un repositorio e iniciar un análisis. En la Imagen 3.4 se puede ver la pantalla inicial del servicio, en la que se pide el repositorio de GitHub en el que se está trabajando.



*Imagen 3.4: Pantalla inicial del servicio*

Tras introducir el repositorio comprimido, comenzará el procesamiento de los cambios realizados y se mostrarán una serie de métricas que nos pueden orientar en una posible detección de fallos. En la Imagen 3.5 se puede ver una pantalla con unos resultados de ejemplo. El análisis resultante consiste en una serie de 10 métricas que suelen estar relacionadas con la presencia de *bugs*, de manera que el usuario puede ver si se presenta alguna anomalía.

Class	CBO	NLE	RFC	CXMR	WMC	DMR	CPMR	TNLA	WI	SMR
Hammer	0	1	5	0	6	13	0	2	13	0
User	0	1	9	0	15	19	0	5	19	0
AppTest	0	2	1	0	3	5	0	0	5	0

Información sobre las métricas:

- **CBO:** *Coupling Between Object classes*, número de usos de otras clases. Un valor muy alto de este valor indica que es muy dependiente de otros módulos, y por tanto más difícil de testear y utilizar, además de muy sensible a cambios. Si tiene un valor alto quizás debería revisar sus cambios.
- **NLE:** *Nesting Level Else-If*, grado de anidamiento máximo de cada clase (bloques de tipo if-else-if cuentan como 1 nivel)
- **RFC:** *Response set For Class*: combinación de número de métodos locales y métodos llamados de otras clases.
- **CXMR:** *Complexity Metric Rules*, violaciones en las buenas prácticas relativas a métricas de complejidad. Si es distinto de 0, quizás deba revisar sus cambios.
- **WMC:** *Weighted Methods per Class*, número de caminos independientes de una clase. Se calcula como la suma de la complejidad ciclométrica de los métodos locales y bloques de inicialización.
- **DMR:** *Documentation Metric Rules*, violaciones de buenas prácticas relativas a la cantidad de comentarios y documentación.
- **CPMR:** *Coupling Metric Rules*, violaciones en las buenas prácticas relativas al acoplamiento de las clases. Si es distinto de 0, quizás deba revisar sus cambios.
- **TNLA:** *Total Number of Local Attributes*, número de atributos locales de cada clase.
- **WI:** *Warning Info*, advertencias de tipo *WarningInfo* en cada clase
- **SMR:** *Size Metric Rules*, violaciones en las buenas prácticas relativas al tamaño de las clases. Si es distinto de 0, quizás deba revisar sus cambios.

Imagen 3.5: Pantalla de resultados

### 3.2 SELECCIÓN DE MÉTRICAS A MOSTRAR

Uno de los puntos fuertes del servicio desarrollado es la selección de las métricas a mostrar al usuario. Ofrecer una cantidad excesiva de métricas podría saturar al usuario y dificultar el uso del servicio. Por ello, se ha optado por estudiar qué características son las más útiles de cara a una posible detección de fallos.

La selección de variables se ha realizado a partir de [20], la que disponíamos de una clasificación de las clases de un conjunto de proyectos, indicándonos para cada una de ellas distintas métricas de código estático junto con el número de bugs que presentaban. En total tenemos un total de 106 métricas diferentes, por lo que es conveniente seleccionar un subconjunto de estas.

Se ha realizado un análisis exploratorio de cada una de estas métricas, para ver la distribución de cada una de ellas a lo largo de las clases recogidas en [20]. Hemos descartado variables que tuvieran una correlación muy fuerte con otras, ya que reflejaban prácticamente la misma información, así como aquellas que no tuvieran una variabilidad notable en sus valores. Siguiendo esos criterios se han logrado eliminar hasta 35 métricas del conjunto de 106 métricas que disponíamos.

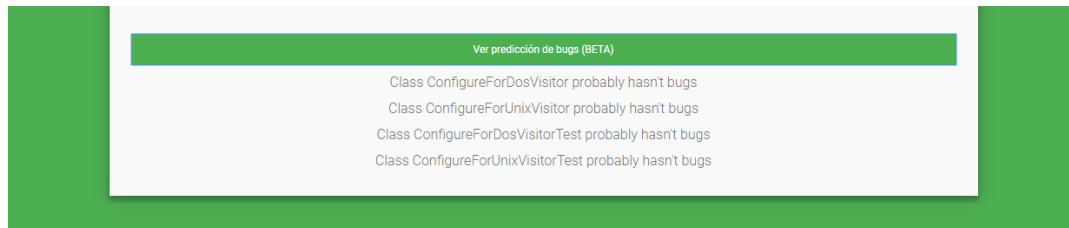
Una vez realizado esta primera valoración de las métricas, y puesto que el objetivo de este servicio es prevenir la presencia de *bugs*, hemos ordenado las métricas restantes en función de la correlación de cada una con respecto al número de bugs de cada clase. Finalmente, se han descartado algunas más por representar información similar a otras métricas de las que disponíamos, para mostrar un total de 10 métricas que pueden ser indicativas de la presencia de errores en los cambios de código de los desarrolladores.

### 3.3 MODELO DE PREDICCIÓN

Al tener en [20] una base de datos en la que se indica la presencia de errores en cada una de las clases de una serie de proyectos, se ha valorado la posibilidad de elaborar un modelo de predicción que, a partir de estos datos, fuera capaz de predecir si los cambios de código que se envían a nuestro servicio tienen una probabilidad alta de presentar fallos o no.

En primer lugar, se ha intentado crear un modelo cuya salida pueda ser interpretable, con el objetivo de poder explicar al usuario las razones por las que el modelo ha predicho que los cambios pueden contener fallos. Por ello, se ha intentado crear un árbol de decisión, ya que con este modelo podemos saber qué variables ha tomado en cuenta el modelo y a partir de qué valores considera que la probabilidad de presentar fallos es alta; y un modelo de regresión lineal con regularización Lasso, ya que en este modelo podemos saber la importancia de cada variable para predecir el valor final de la variable clase. Sin embargo, los resultados obtenidos no han sido positivos, ya que los modelos tenían un alto porcentaje de acierto sobre el conjunto de entrenamiento, es decir, sobre los datos que utilizamos para aprender, pero no sobre el conjunto de test, aquel que apartamos del entrenamiento del modelo para estudiar el rendimiento real del mismo. Se ha producido, por tanto, un fenómeno conocido como sobreajuste, el cual no nos permite concluir que el modelo es capaz de predecir cambios de código en nuevos repositorios.

Tras los resultados de los modelos anteriores, se optó por crear modelos de predicción más potentes pero que careciesen de interpretabilidad. En concreto, se utilizó un modelo conocido como Random Forest, que consiste en utilizar un conjunto de modelos de árbol de decisión y combinar la salida de todos ellos, y un modelo denominado XGBoost [23], un algoritmo que aplica la técnica del gradiente descendiente de manera optimizada para permitir utilizar grandes conjuntos de datos. No obstante, los resultados son similares a los modelos anteriores. En consecuencia, los modelos generados no son tan buenos como los que se comentan en [20], por lo que se ha incluido en el servicio como una funcionalidad en fase beta, pero centrando los resultados en la selección de métricas mencionada en la sección anterior. En la Imagen 3.6 tenemos una imagen de la parte del servicio relativa al uso de este modelo de predicción. Esta funcionalidad se encuentra en la parte final de la pantalla y solo se mostrará si el usuario está interesado.



*Imagen 3.6: Pantalla de predicción de fallos - versión Beta*

### 3.4 CONCLUSIONES

Tras realizar distintos estudios sobre la creación de modelos de predicción que indiquen de manera automática si hay una alta probabilidad de tener fallos potenciales, hemos visto que los resultados no son los esperados. En consecuencia, este servicio se centrará en mostrar al usuario métricas claves para la identificación de estos fallos, de manera que pueda observar de manera muy rápida si hay algún valor fuera de lo común o de los estándares de calidad del proyecto.

Se recomienda el uso de este servicio cuando se utilizan técnicas de CI, justo antes de subir los cambios al repositorio online, con el fin de poder detectar fallos antes de realizar la compilación y las pruebas automáticas y, por lo tanto, de romper el flujo de trabajo, aunque se puede utilizar en proyectos que no apliquen Integración Continua.

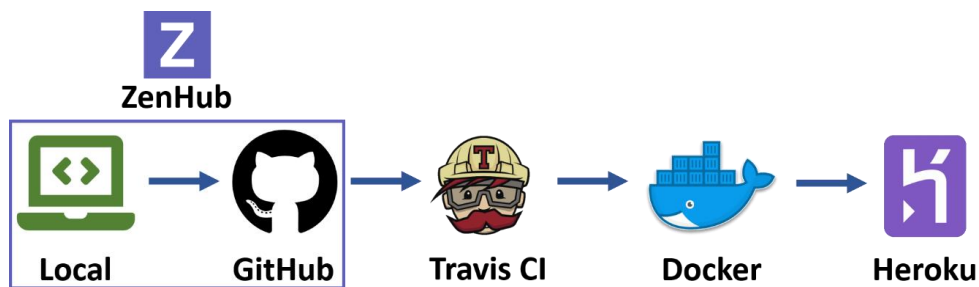
# CAPÍTULO 4

## DESPLIEGUE CONTINUO



En este capítulo se entrará más en detalle sobre cómo se ha implementado el servicio y la estructura utilizada para su despliegue. Se enumerarán y explicarán cada una de las tecnologías utilizadas durante la evolución de este servicio.

En la Imagen 4.1 se muestra el *pipeline* o flujo de trabajo adoptado durante este trabajo. En primer lugar, se realizan los cambios convenientes en el entorno de trabajo del usuario. A continuación, cuando se tiene un cambio que se quiere llevar al despliegue final, se suben los cambios al repositorio remoto, en nuestro caso GitHub. También se ha utilizado ZenHub<sup>3</sup>, una herramienta que complementa a GitHub para la gestión ágil del proyecto.



*Imagen 4.1: Pipeline de desarrollo*

Tras subirse a GitHub, automáticamente Travis CI [18], que es la herramienta de Integración Continua que hemos utilizado, ejecutará el plan de pruebas que hayamos creado para comprobar que los desarrollos realizados no han provocado variaciones en el comportamiento usual del servicio. Si las pruebas no se pasan, se avisaría al usuario y se procedería a arreglar los fallos. Si las pruebas son correctas, se despliega automáticamente. Para ello, se crean una serie de contenedores Docker<sup>4</sup> con funciones diferenciadas, y estos contenedores se despliegan con Heroku<sup>5</sup>, dejando el servicio completamente funcional.

<sup>3</sup> <https://www.zenhub.com/>

<sup>4</sup> <https://www.docker.com/>

<sup>5</sup> <https://www.heroku.com/>

### 4.1 DESARROLLO

El proyecto se ha desarrollado con el lenguaje de programación Python, utilizando un framework denominado Django<sup>6</sup>. Django es un framework de alto nivel de Python para desarrollo web que trata de conseguir un desarrollo rápido y un diseño limpio. Este framework se basa en una arquitectura software conocida como modelo-vista-controlador [24], o simplemente MVC, que trata de separar el desarrollo en tres capas diferenciadas: el modelo, que define estructura de los datos; el controlador, que gestiona las peticiones del usuario; y la vista, que maneja la presentación del contenido.

La parte del controlador es la encargada de gestionar los distintos tipos de llamada que realiza el usuario. En este caso, solo se ha proporcionado una posible URL, a la cual el usuario accede e introduce el repositorio que desea analizar.

En la parte de la vista se crean las distintas plantillas que se utilizan para mostrar la información. Aquí se han tenido que crear tres diferentes: una para permitir al usuario introducir su repositorio; otro para indicarle que la petición se está procesando y debe esperar; y una última para mostrar los resultados del análisis una vez acabado.

Por último, en cuanto al modelo de datos, no se ha creado una estructura específica para este servicio. En esta capa se suelen definir las entidades que forman la base de datos si se necesitan almacenar los datos a largo plazo. Sin embargo, ya que en nuestro caso únicamente queremos ofrecer los resultados cuando se procesan, pero no guardarlos durante más tiempo, tan solo habrá que guardar la información temporalmente y mostrar al usuario los resultados cuando estén calculados.

Todo el código ha sido gestionado utilizando GitHub para el control de versiones. Además, para facilitar la gestión ágil del proyecto se ha utilizado ZenHub<sup>7</sup>, lo que nos permite manejar y visualizar qué incidencias estamos tratando en cada momento, como se puede ver en el ejemplo de la Imagen 4.2.

### 4.2 INTEGRACIÓN CONTINUA

Cada vez que se hace un nuevo incremento de nuestro servicio, se suben los cambios a GitHub. Al subirlo, y gracias a la integración que Travis CI tiene con GitHub, el proyecto

<sup>6</sup> <https://www.djangoproject.com>

<sup>7</sup> <https://www.zenhub.com>



pasa una serie de pruebas automáticas para asegurar que el comportamiento sigue siendo el esperado en todos los componentes de nuestro servicio. Si las pruebas que hemos definido fallan, se avisa al equipo y el despliegue se anula, con el objetivo de que arreglen los defectos encontrados. Si son correctas, como se ve en el ejemplo de la Imagen 4.3, comienza el proceso de Despliegue Continuo.

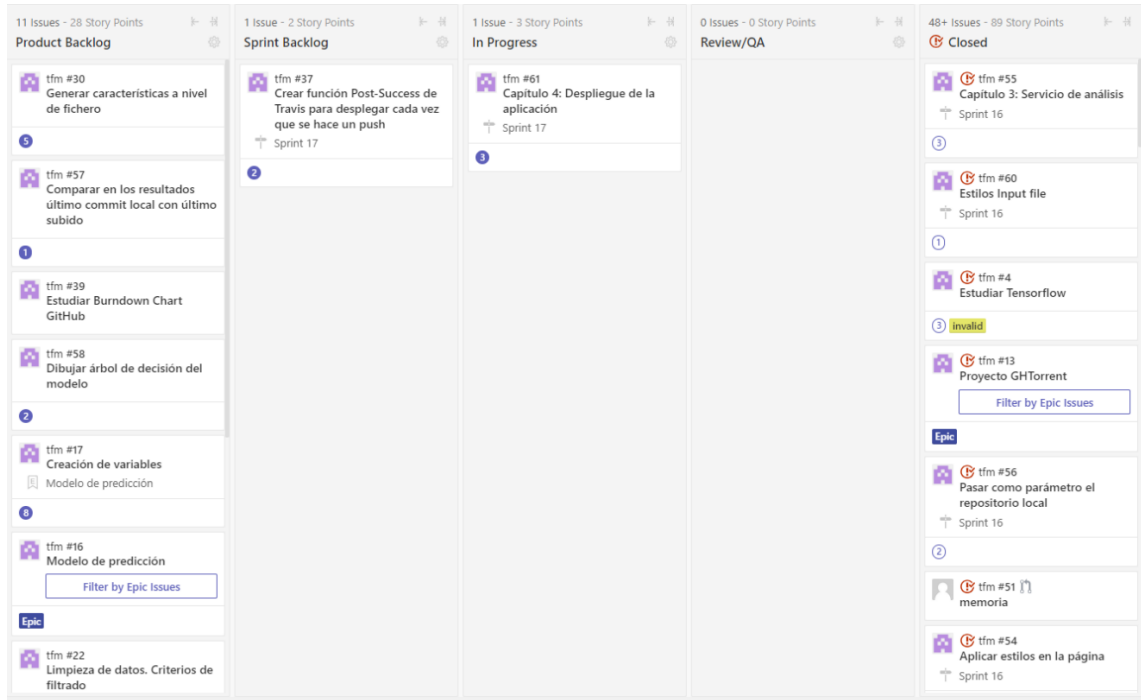


Imagen 4.2: Visualización de incidencias en ZenHub

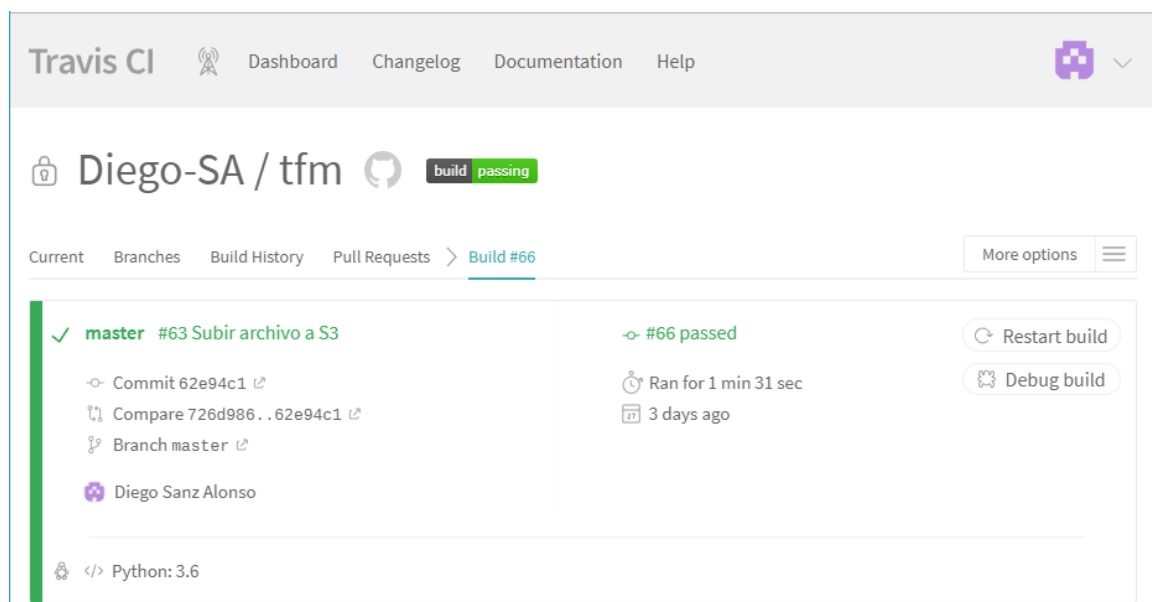


Imagen 4.3: Ejemplo Build en Travis CI

Para realizar el proceso de Despliegue Continuo se ha definido en la configuración de Travis que, tras pasar correctamente las pruebas, comience la creación de los contenedores y el posterior despliegue. La aplicación está desplegada en la nube con Heroku<sup>8</sup>. Esta plataforma permite ejecutar aplicaciones en múltiples lenguajes de programación y con una escalabilidad muy rápida. En nuestro caso, desplegaremos en Heroku utilizando Docker Compose, una herramienta para orquestar un conjunto de contenedores Docker de manera simultánea [25]. En total, la aplicación está formada por 3 contenedores Docker, que se organizan como muestra la Imagen 4.4.

El primero de ellos, al que se ha denominado *Web*, es el encargado de las peticiones web, de manera que gestionará todas las llamadas que los usuarios realicen. Este contenedor será el responsable de recibir el repositorio que se va a analizar y guardarlo en Amazon S3<sup>9</sup>, el servicio de almacenamiento de Amazon, además de mostrar los resultados finales.

El segundo contenedor, denominado *Worker*, hará las funciones más pesadas, de forma que se ejecuten en segundo plano y el usuario no tenga grandes tiempos de espera. La función principal de este contenedor es la de procesar los repositorios que los usuarios proporcionan para realizar el análisis explicado en el capítulo anterior. Este contenedor descarga los repositorios almacenados en S3 por el contenedor anterior.

El tercer y último contenedor es el encargado de comunicar los dos anteriores, de manera que cuando un usuario realiza una petición para analizar su repositorio, se almacene mediante el primer contenedor, y este mande un mensaje al segundo para que realice el análisis. Para este contenedor se ha utilizado una librería de Python denominada Redis Queue (RQ) [26].

En la Imagen 4.4 se puede ver un esquema de los contenedores que conforman nuestro servicio y las interacciones entre ellos. Como se aprecia, los contenedores Web y Worker se basan en nuestra aplicación de desarrollada en Django, y la Web se encarga de subir información a Amazon S3 mientras que el Worker solo la descarga. Por su parte, el tercer contenedor parte de una imagen de Redis, siendo la encargada de recibir las peticiones de los otros dos contenedores para informar cuándo debe empezar un proceso o cuándo ha terminado.

Hay dos razones por la que se utiliza un servicio externo como Amazon S3 de forma temporal para el envío de información entre Web y Worker. En primer lugar, el sistema de archivos de Heroku es efímero<sup>10</sup> [27], por lo que si se intenta procesar un repositorio y antes

<sup>8</sup> <https://www.heroku.com/>

<sup>9</sup> <https://aws.amazon.com/es/s3/>

<sup>10</sup> Cada vez que se reinicia el servicio, los datos almacenados se eliminan

de acabar se para o se reinicia alguna de las instancias, el proceso fallaría puesto que el repositorio ya no estaría almacenado. Además, en el despliegue con Heroku no se permite el uso de volúmenes de datos compartidos entre contenedores, por lo que necesitamos una herramienta que actúe de intermediaria entre ellos.

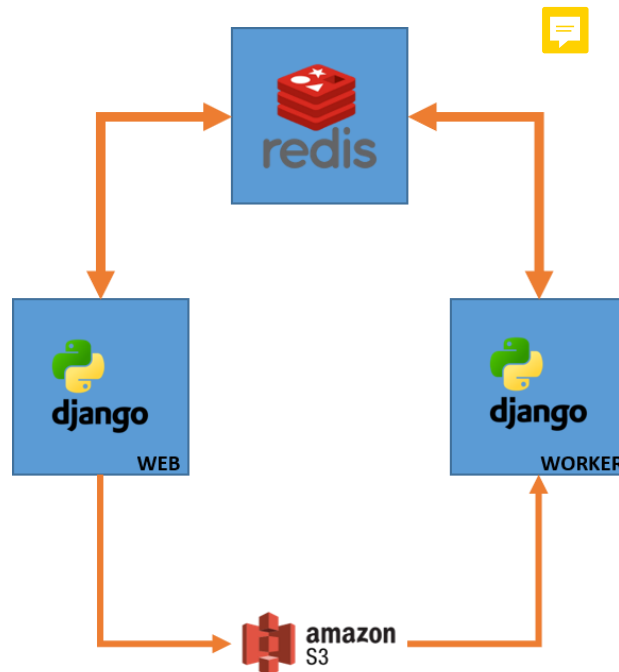
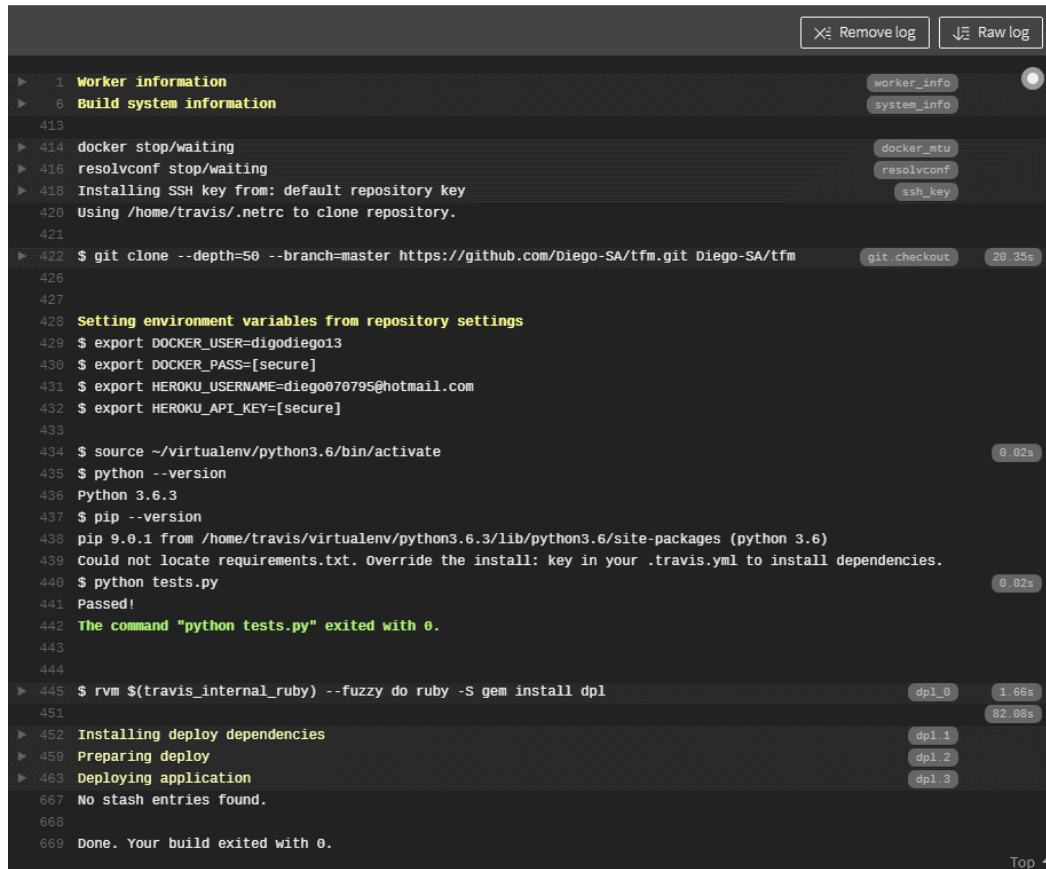


Imagen 4.4: Docker Compose. Esquema de los contenedores

Para desplegar el servicio debemos, en primer lugar, crear la configuración de los contenedores Docker de la web y el worker. Dicha configuración puede consultarse en el Anexo A.2. A continuación, es necesario definir la configuración de Docker Compose, indicando en cada uno de los tres contenedores su origen, que para los dos primeros partirá de las imágenes creadas en los archivos Docker mencionados antes, y para el contenedor de Redis se usará una imagen predeterminada, como se aprecia en la imagen anterior. Esta configuración puede verse en el Anexo A.3. Una vez hechos estos archivos de configuración, Travis CI es capaz de desplegarlo automáticamente en Heroku, de manera que el servicio se quede completamente operativo con cada incremento de código realizado. La configuración establecida en Travis se encuentra en el Anexo A.4.

En la Imagen 4.5 se puede apreciar el resultado de nuestro pipeline de CI/CD. En primer lugar, y tras cargar una serie de variables de entorno definidas en Travis, se ejecutan las pruebas. Si estas son correctas, como en este caso, comienza el despliegue. Para ello, en primer lugar, se instalan las dependencias necesarias para nuestro proyecto. A continuación, se prepara el despliegue mediante la creación del Docker Compose explicado antes. Finalmente, con Heroku, se despliega el resultado final para que el usuario final pueda acceder a nuestro servicio vía web como se muestra en la Imagen 3.4.



The image shows a Travis CI build log for a project named 'Diego-SA/tfm'. The log is displayed in a dark-themed interface with a 'Remove log' button at the top right. The build process includes several steps: 'Worker information', 'Build system information', 'docker stop/waiting', 'resolvconf stop/waiting', 'Installing SSH key from: default repository key', and 'Using /home/travis/.netrc to clone repository'. The main build step is a git clone command, which takes 20.35s. This is followed by setting environment variables from repository settings, including DOCKER\_USER, DOCKER\_PASS, HEROKU\_USERNAME, and HEROKU\_API\_KEY. The build then sources the virtualenv, checks the python and pip versions, and runs 'python tests.py', which passes. The final step is 'rvm \$(travis\_internal\_ruby) --fuzzy do ruby -S gem install dpl', which takes 82.08s. The build concludes with 'Installing deploy dependencies', 'Preparing deploy', and 'Deploying application', all of which are successful. The final message is 'Done. Your build exited with 0.'.

```
1 Worker information
6 Build system information
413
414 docker stop/waiting
416 resolvconf stop/waiting
418 Installing SSH key from: default repository key
420 Using /home/travis/.netrc to clone repository.
421
422 $ git clone --depth=50 --branch=master https://github.com/Diego-SA/tfm.git Diego-SA/tfm
426
427
428 Setting environment variables from repository settings
429 $ export DOCKER_USER=digodiego13
430 $ export DOCKER_PASS=[secure]
431 $ export HEROKU_USERNAME=diego070795@hotmail.com
432 $ export HEROKU_API_KEY=[secure]
433
434 $ source ~/virtualenv/python3.6/bin/activate
435 $ python --version
436 Python 3.6.3
437 $ pip --version
438 pip 9.0.1 from /home/travis/virtualenv/python3.6.3/lib/python3.6/site-packages (python 3.6)
439 Could not locate requirements.txt. Override the install: key in your .travis.yml to install dependencies.
440 $ python tests.py
441 Passed!
442 The command "python tests.py" exited with 0.
443
444
445 $ rvm $(travis_internal_ruby) --fuzzy do ruby -S gem install dpl
451
452 Installing deploy dependencies
459 Preparing deploy
463 Deploying application
667 No stash entries found.
668
669 Done. Your build exited with 0.
```

Imagen 4.5: Resultado del CD en Travis



# CAPÍTULO 5

## ANÁLISIS DE DATOS

---

En este capítulo se hará un resumen del estudio realizado con los datos de la base de datos de bugs disponible gracias a [20]. Se explicará el análisis exploratorio realizado para estudiar tanto la variable que queremos predecir, también llamada variable clase, que en este caso es el número de bugs; como las variables predictoras, que son las distintas métricas que se extraen de cada clase del proyecto. Además, se enumerarán las variables escogidas a mostrar al usuario, ya que como se comentó en el CAPÍTULO 3 el servicio finalmente sigue un método descriptivo.

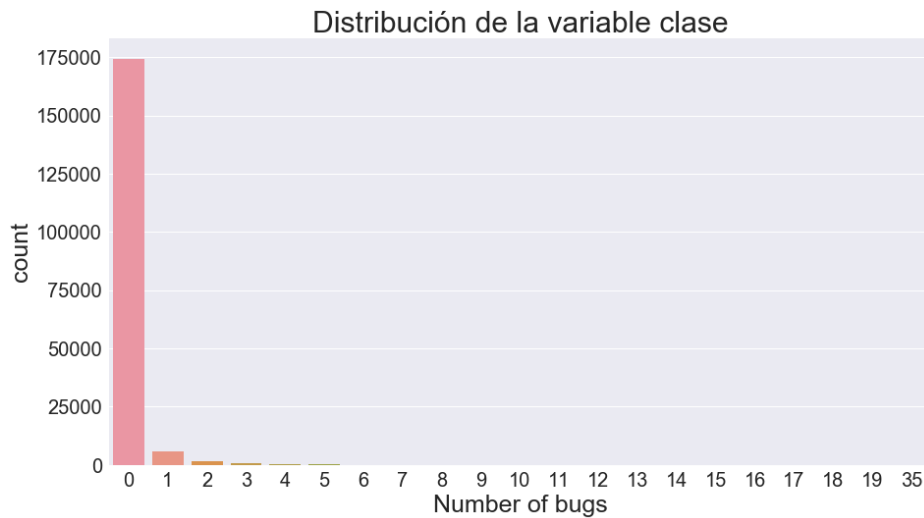


En la base de datos de [20] disponemos de un total de 182.671 registros. Cada uno de ellos hace referencia a una clase de un proyecto Java, y en todos ellos vienen recogidas distintas métricas de código estático (un total de 105) y el número de bugs que presentan. El análisis de datos completo se puede encontrar en el repositorio del proyecto, aunque aquí se comentarán los puntos principales del estudio.

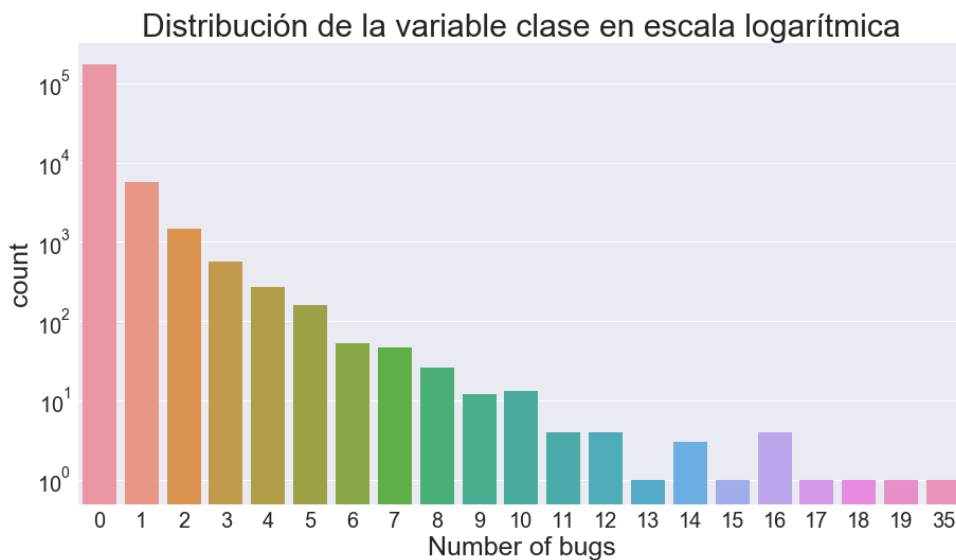
### 5.1 VARIABLE CLASE

La primera variable que se debe analizar es la más importante. En este caso es el número de errores que existen en una clase. El objetivo del servicio consiste en ayudar a los desarrolladores a encontrar fallos en sus cambios de código, por lo que todo el análisis debe centrarse en este punto.

En la Imagen 5.1 se muestra una gráfica con la distribución de los números de bugs presentes en cada una de las clases disponibles en la base de datos. Se puede apreciar como casi todos los valores son igual a 0. Esto provoca que no se vea con precisión el valor en el resto de los valores. Para subsanar esto, en la Imagen 5.2 se ha cambiado el eje y a una escala logarítmica. En esta escala se puede observar que en las clases que tienen bugs, predomina un número reducido de errores.



*Imagen 5.1: Distribución de la variable "Número de bugs"*



*Imagen 5.2: Distribución de la variable "Número de bugs" en escala logarítmica*

Sin embargo, no necesitamos un valor exacto del número de errores en cada clase. El objetivo del servicio es ayudar al desarrollador a que detecte que sus cambios presentan fallos o no, pero determinar el número es una tarea más compleja. Además, de cara a la creación de un posible modelo de predicción, es conveniente hacer una partición binaria de esta variable, discriminando entre clases que no contienen bugs, y clases que sí lo tienen.

Si realizamos esta partición binaria, nos encontramos con un total de 8.275 clases con bugs, frente a las 174.396 clases que no tienen errores conocidos. Esto supone tan solo el 4,53% de los datos, por lo que nos encontramos ante un conjunto de datos claramente

desbalanceado. Si se quiere crear un modelo de predicción es conveniente tener este factor en cuenta.

## 5.2 VARIABLES PREDICTORAS

Tras estudiar la variable clase y binarizarla, se puede comenzar el análisis del resto de características. Hay un total de 106 variables, representando cada una de ellas una métrica de código estático. Debido a este número tan elevado de características, se procederá a intentar eliminar variables que no muestren información relevante.

Esta ausencia de información puede deberse a dos factores principales. En primer lugar, puede existir una varianza muy baja entre los valores de una de las características, por lo que no aportará valor al estudio. Aquí hay que tener en cuenta que, al tener una base de datos desbalanceada, una varianza no muy alta no tiene por qué suponer que la métrica que se intenta estudiar no sea útil. Por ello, únicamente se eliminarán del estudio aquellas variables que tengan prácticamente todos sus valores iguales, ya que solo aportarían ruido en nuestro análisis.

El segundo punto a tener en cuenta es la correlación. Existen métricas que pueden aportar una información muy similar a otra. Por poner un ejemplo más específico, una de las variables disponibles es el número de líneas de código de una clase. Esta variable podría ser muy similar al número de líneas lógicas de la clase, que solo variaría si existiesen dos instrucciones en la misma línea física, algo no muy común. Sin embargo, es necesario calcular la correlación para asegurar la similaridad de información de estas clases. En este caso, la correlación es del 98.4%, por lo que se puede eliminar una de las dos sin apenas perder información.

Siguiendo el primer criterio se han eliminado 14 variables, y teniendo en cuenta el segundo se han eliminado 21. Esto provoca que el conjunto final de variables descienda a 76, reduciendo en un 33% el número de características de nuestra base de datos.

Esta limpieza de los datos permitiría una creación de un modelo de predicción con mucho menos ruido que el conjunto de datos original. Sin embargo, ya que en este proyecto nos hemos centrado en el carácter descriptivo, el número de métricas sigue siendo excesivo, por lo que hay que hacer una selección entre las variables restantes.

Para realizar esta selección se ha calculado la correlación entre cada una de las variables predictoras con la variable clase, y se han ordenado en función de este valor. Posteriormente, se han analizado las variables en orden descendente en función de esta correlación, eliminando variable que, aunque no tuvieran una correlación muy alta entre



ellas, mostraban una información similar. Este tipo de variables no ayudarán al usuario, ya que si ponemos variables muy parecidas el valor de la información será escasa.

Por ejemplo, una de las variables con una de las correlaciones más alta con la variable clase es el grado de anidamiento. Existen dos métricas que muestran esta información, aunque con un matiz diferencial. Una de ellas cuenta sentencias consecutivas de if-else-if como una sola instrucción, mientras que la otra las cuenta como dos. Estas variables no están tan correlacionadas como podría parecer a priori (aunque tiene un valor alto, un 81,59%), por lo que se han mantenido ambas en el estudio. A pesar de ello, de cara al usuario final es indiferente mostrar una u otra, por lo que la otra se excluirá de los resultados.

Con todo, se han seleccionado un total de 10 métricas que se mostrarán al usuario. Estas métricas son las siguientes:

- *Coupling between Object classes*, número de usos de otras clases. Clases con este valor muy alto serán muy dependientes de otros módulos, y por tanto más difíciles de testear y utilizar, además de muy sensibles a cambios.
- *Nesting Level Else-If*, grado de anidamiento máximo de una clase, contando bloques de tipo if-else-if como un solo nivel.
- *Response set for Class*, combinación de número de métodos locales y métodos llamados de otras clases.
- *Complexity Metric Rules*, violaciones en las buenas prácticas relativas a métricas de complejidad.
- *Weighted Methods per Class*, número de caminos independientes de una clase. Se calcula como la suma de la complejidad ciclomática de los métodos locales y bloques de inicialización.
- *Documentation Metric Rules*, violaciones de buenas prácticas relativas a la cantidad de comentarios y documentación.
- *Coupling Metric Rules*, violaciones en las buenas prácticas relativas al acoplamiento de las clases.
- *Total Number of Local Attributes*, número de atributos locales de cada clase.
- *Warning Info*, advertencias de tipo *WarningInfo* en cada clase.
- *Size Metric Rules*, violaciones en las buenas prácticas relativas al tamaño de las clases.



# CAPÍTULO 6

## CONCLUSIONES Y PROPUESTAS

---

### 6.1 CONCLUSIONES

El objetivo principal del trabajo consistía en poder advertir al programador de posibles fallos en su código antes de subirlo a su repositorio compartido online, con el fin de reducir el riesgo de que se interrumpa el flujo de trabajo de todo el equipo. Tras el desarrollo del proyecto, se ha conseguido crear un servicio descriptivo que muestra al usuario un subconjunto de las métricas más importantes desde el punto de vista de la detección de fallos.

No solo se ha implementado el servicio, sino que además se ha conseguido desplegar en la nube, de manera que cualquier usuario puede utilizar nuestro servicio vía web. Además, se ha creado un flujo de trabajo que incluye técnicas de Integración Continua y Despliegue Continuo, por lo que cada vez que se desarrolle un incremento nuevo de código se pasará un conjunto de pruebas automáticamente y, si no hay ningún fallo, el servicio se despliega automáticamente con los nuevos cambios, reduciendo así el tiempo entre el desarrollo de una nueva funcionalidad o un arreglo de algún fallo y la llegada de ese cambio al usuario final.

Además del servicio descriptivo, se ha intentado crear un modelo de predicción que nos diga automáticamente si las clases modificadas por un desarrollador tenían una alta probabilidad de fallo o no. Sin embargo, los resultados en este aspecto no son tan positivos como se esperaba, por lo que se ha mantenido como una versión en fase Beta que el usuario puede consultar, pero lo que impera será el carácter descriptivo del servicio.

### 6.2 TRABAJO FUTURO

Ya que el servicio se ha logrado desplegar y se ha creado el flujo de CI/CD, incluir nuevas funcionalidades debería ser una tarea sencilla. Si no se modifica la arquitectura, solo habría que realizar los cambios de código pertinentes, sin tener que preocuparnos de la parte operacional o de sistemas.

Uno de los puntos en los que se puede avanzar en este trabajo es la reducción de restricciones en los proyectos válidos para el estudio. Actualmente solo se tiene la capacidad de analizar proyectos con lenguaje de programación Java que estén alojados en GitHub. Se podrían incluir nuevos lenguajes de programación, teniendo en cuenta que las métricas pueden ser diferentes y habría que hacer un análisis exhaustivo en cada caso. Además, se podría permitir el estudio de repositorios en otras plataformas, como Bitbucket o GitLab<sup>11</sup>. En este caso habría que estudiar cómo proporcionar el repositorio al servicio, así como identificar los archivos que se han modificado localmente, pero la parte frontal del servicio no debería sufrir modificaciones en este aspecto.

Otro de los aspectos que se podrían tratar consiste en crear un modelo de predicción con unos buenos resultados. Para esto sería necesario buscar otra base de datos, ya que con la que se ha obtenido en [20] no se ha podido crear un modelo con un gran acierto. Una de las vías en las que se podría investigar consiste en buscar variables distintas a métricas estáticas de la clase. Por ejemplo, se podrían buscar características a nivel de fichero, si el proyecto contiene varias clases en un solo fichero, ya que es de esperar que, si eso ocurre, estas clases estén relacionadas entre sí y fallos en una de ellas puedan repercutir en las demás.

Otro tipo de características que podrían dar mucho valor a la hora de la creación de un modelo predictivo son aquellas relativas a información histórica de la clase. La mayoría de las plataformas de gestión de repositorio tienen control de versiones, por lo que a través del mismo podríamos sacar distinta información, como el número de veces que se ha cambiado una clase o el número de desarrolladores diferentes que han hecho algún tipo de cambio sobre la misma.

<sup>11</sup> <https://about.gitlab.com/>

# BIBLIOGRAFÍA

- [1] G. Kim, J. Humble y P. Debois, *The DevOps Handbook*, IT Revolutions, 2016.
- [2] Verifysoft, «Halstead metrics,» 10 Agosto 2017. [En línea]. Available: [https://www.verifysoft.com/en\\_halstead\\_metrics.html](https://www.verifysoft.com/en_halstead_metrics.html).
- [3] C. Treude, L. Leite y M. Aniche, «Unusual events in GitHub repositories,» *The Journal of Systems & Software*, nº 142, pp. 237-247, 2018.
- [4] «Guía de Scrum,» Noviembre 2017. [En línea]. Available: [www.scrum.org](http://www.scrum.org).
- [5] R. Jurney, *Agile Data Science 2.0*, O'Reilly Media, 2017.
- [6] D. Petrov, «Data Version Control: iterative machine learning,» FullStackML, Mayo 2017. [En línea]. Available: <https://www.kdnuggets.com/2017/05/data-version-control-iterative-machine-learning.html>.
- [7] T. Volk, «Seven steps to move a DevOps team into the ML and AI world,» DevOpsAgenda, Abril 2018. [En línea]. Available: <https://devopsagenda.techtarget.com/opinion/Seven-steps-to-move-a-DevOps-team-into-the-ML-and-AI-world>.
- [8] Atlassian, «What is version control,» [En línea]. Available: <https://www.atlassian.com/git/tutorials/what-is-version-control>. [Último acceso: 07 Abril 2019].
- [9] Smartsheet, «Software Version Control,» [En línea]. Available: <https://www.smartsheet.com/software-version-control>. [Último acceso: 07 Abril 2019].
- [10] Apache Software Foundation, «Apache Subversion,» [En línea]. Available: <https://subversion.apache.org/>. [Último acceso: 2019 Abril 04].

- [11] Git, «Git,» [En línea]. Available: <https://git-scm.com/>. [Último acceso: 2019 Abril 07].
- [12] Atlassian, «Continuous Integration vs Delivery vs Deployment,» Atlassian, [En línea]. Available: <https://es.atlassian.com/continuous-delivery/principles/continuous-integration-vs-delivery-vs-deployment>. [Último acceso: 07 Abril 2019].
- [13] Y. Weicheng, B. Shen y B. Xu, «MiningGitHub: Why Commit Stops,» *Asia-Pacific Software Engineering Conference*, n° 20, pp. 165-169, 2013.
- [14] G. Gousios, «The GHTorrent Dataset and Tool Suite,» de *Proceedings of the 10th Working Conference on Mining Software Repositories*, San Francisco, CA, USA, 2013, pp. 233-236.
- [15] G. Georgios, «The GHTorrent Project,» 2013. [En línea]. Available: <http://ghtorrent.org/>. [Último acceso: 13 Marzo 2019].
- [16] J. Cabot, J. L. Cánovas Izquierdo, V. Cosentino y B. Rolandi, «Exploring the Use of Labels to Categorize Issues in Open-Source Software projects,» *International Conference on Software Analysis, Evolution and Reengineering*, n° 22, 2015.
- [17] M. Beller, G. Gousios y A. Zaidman, «TravisTorrent: Synthesizing Travis CI and GitHub for Full-Stack Research on Continuous Integration,» *Proceedings of the 14th working conference on mining software repositories*, 2017.
- [18] Travis CI, «Travis CI,» 2011. [En línea]. Available: <https://travis-ci.org/>. [Último acceso: 13 Marzo 2019].
- [19] P. Gyimesi, G. Gyimesi y Z. Tóth, «Characterization of Source Code Defects by Data Mining Conducted on GitHub,» *Computational Science and Its Applications - ICCSA*, vol. 9159, pp. 47-62, 2015.
- [20] Z. Tóth, P. Gyimesi y R. Ferenc, «A Public Bug Database of GitHub Projects and its Application in Bug Prediction,» *Computational Science and Its Applications - ICCSA*, vol. 9789, pp. 625-638, 2016.

- [21] FrontendArt, «SouceMeter - Free-to-use, Advanced Source Code Analysis Suite,» 2016. [En línea]. Available: <https://www.sourcemeter.com/resources/java/>. [Último acceso: 03 Abril 2019].
- [22] S. Tuli, «Learn How to Set UP a CI/CD Pipeline From Scratch,» 10 Agosto 2018. [En línea]. Available: <https://dzone.com/articles/learn-how-to-setup-a-cicd-pipeline-from-scratch>. [Último acceso: 18 Abril 2019].
- [23] XGBoost, «XGBoost Documentation,» [En línea]. Available: <https://xgboost.readthedocs.io/en/latest/>. [Último acceso: 19 Abril 2019].
- [24] Django, «FAQ: General | Documentación Django,» [En línea]. Available: <https://docs.djangoproject.com/es/2.2/faq/general/>. [Último acceso: 17 Mayo 2019].
- [25] Heroku Dev Center, «Local Development with Docker Compose,» 05 Febrero 2019. [En línea]. Available: <https://devcenter.heroku.com/articles/local-development-with-docker-compose>. [Último acceso: 02 Mayo 2019].
- [26] V. Driessen. [En línea]. Available: <https://python-rq.org/>. [Último acceso: 2019 Mayo 02].
- [27] Heroku Dev Center, «Dynos and the Dyno Manager,» 2019 Enero 22. [En línea]. Available: <https://devcenter.heroku.com/articles/dynos#ephemeral-filesystem>. [Último acceso: 17 Mayo 2019].



# Anexos



## A.1. Anexo 1: Metodología de trabajo

Como se comenta en el capítulo introductorio, el proyecto se ha desarrollado siguiendo una metodología Scrum. En total se han desarrollado un total de (POR DETERMINAR) tareas, las cuales se listan a continuación (algunos códigos de referencia están ausentes porque pertenecen a tareas creadas para hacer pruebas del funcionamiento de ZenHub o de tareas que se rechazaron):

- #1: Estudiar conceptos básicos DevOps
- #2: Estudiar Travis-CI
- #3: Estudiar ZenHub
- #5: Desarrollar aplicación web básica
- #10: Mirar proyecto TravisTorrent
- #11: Estudiar página TravisTorrent
- #12: Descargar Base de Datos de Travis Torrent
- #13: ProyectoGHTorrent
- #14: Estudiar y descargar Bases de Datos
- #18: Creación modelo de predicción
- #19: Leer artículo “Unusual events in GutHub repositories”
- #20: Buscar base de datos reducida de GitHubTorrent
- #23: Anteproyecto TFM
- #24: Extraer información de 5 proyectos
- #25: Estudiar las variables de la BDD
- #26: Leer papers
- #27: Leer papers de predicción
- #28: Generar modelo de predicción
- #29 Conseguir extraer las características de un proyecto
- #31: Preparar código para crear el servicio
- #32: Estudiar DeployHub
- #33: Estudiar Heroku
- #34: Ejecutar SourceMeter en Heroku
- #35: Eliminar archivos tras análisis de un repositorio
- #36: Ejecutar análisis de SourceMeter en un worker
- #37: Crear función Post-Success de Travis para desplegar cada vez que se hace un push
- #42: Estudiar estructura de la memoria

## Anexo A.1. Imágenes de las células de estudio

- #43: Análisis exploratorio de los datos
- #44: Modelos de predicción interpretables
- #45: Estudiar grafo que proporciona SourceMeter
- #46: Memoria TFM. Capítulo 1 - Introducción
- #47: Modificar elementos del conjunto de entrenamiento
- #48: Support Vector Machine
- #49: Estructura Capítulo 2 - Estado del arte
- #50: Aislar proyectos del conjunto de datos
- #52: Mostrar variables más correlacionadas con la variable clase
- #53: Corregir y ampliar Capítulo 2 - Estado del arte
- #54: Aplicar estilos en la página
- #55: Capítulo 3 - Servicio de análisis
- #56: Pasar como parámetro el repositorio local
- #60: Estilos input file
- #61: Capítulo 4 - Despliegue de la aplicación
- #63: Subir el archivo a S3 para que pueda usarlo el Worker
- #64: Escribir Anexo - Metodología de trabajo

(En ZenHub no veo la opción de Burndown Chart de todo el proyecto, solo hay Burndown Report de un sprint específico. Además, miré un proyecto de GitHub que decía que te lo generaba, pero no llevamos la misma estructura que en los ejemplos por lo que no lo genera bien).

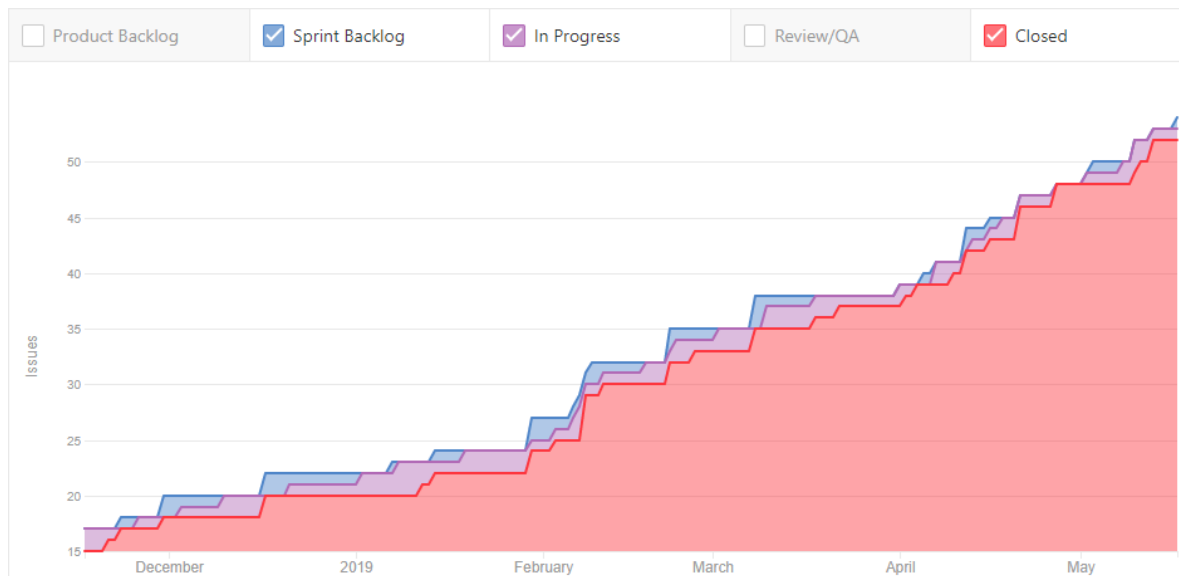


(Además, ZenHub solo deja mostrar gráficos a partir de los últimos 6 meses, no deja un período anterior)

En la imagen Imagen 6.1 se puede ver el ritmo del proyecto a lo largo del tiempo. En la gráfica aparecen en color rojo el número de tareas cerradas; en morado, las tareas que están en proceso en cada momento; y en azul, las tareas que estaban en el Sprint Backlog.

Estas tareas se han realizado en un total de 20 sprints (QUIZÁS VARÍE), y para cada uno de ellos se ha registrado la fecha de inicio y de fin, un resumen de la manifestación de ese sprint y un resumen de la review del mismo. A continuación, se muestran las fechas y los temas tratados en cada uno de estos sprints.





*Imagen 6.1: Control Flow del proyecto*

- Sprint 1



- Fecha inicio: 06/07/2018
- Fecha fin: 20/07/2018
- Sprint planning: Al ser la primera toma de contacto, se decidió comenzar un estudio de la filosofía DevOps y herramientas básicas durante el desarrollo del proyecto: GitHub, TravisCI y ZenHub.
- Tareas desarrolladas: #1, #2
- Sprint review: no dio tiempo para ver todas las capacidades de ZenHub, por lo que se dejó esa tarea para el siguiente sprint.

- Sprint 2

- Fecha inicio: 20/07/2018
- Fecha fin: 07/09/2018
- Sprint planning: además de continuar con el estudio de ZenHub, se decidió estudiar también Django y TensorFlow.
- Tareas desarrolladas: #3, #5
- Sprint review: Se consiguió tener una visión básica de los tres componentes.

- Sprint 3

- Fecha inicio: 07/09/2018
- Fecha fin: 25/09/2018
- Sprint planning: se definió el problema que íbamos a abordar (indicar la probabilidad de fallo en un nuevo incremento de un desarrollador en un repositorio compartido). A raíz del problema, se decidió consultar estado del arte sobre el tema, además de investigar sobre TravisTorrent.
- Tareas desarrolladas: #12, #19, #11, #10, #14

## Anexo A.1. Imágenes de las células de estudio

- Sprint review: Las bases de datos proporcionadas por TravisTorrent eran excesivamente grandes para hacer un tratamiento de los datos.
- Sprint 4
  - Fecha inicio: 25/09/2018
  - Fecha fin: 09/10/2018
  - Sprint planning: se decide buscar bases de datos más pequeñas, además de ver cómo tratarlas, limpiarlas y hacer unión de tablas.
  - Tareas desarrolladas: #20
  - Sprint review: el estudio de los datos dependía de unos permisos otorgados por TravisTorrent que tardaron en llegar por lo que no se avanzó mucho en este sprint.
- Sprint 5
  - Fecha inicio: 09/10/2018
  - Fecha fin: 31/10/2018
  - Sprint planning: para poder analizar los datos más en profundidad, se decidió extraer los datos de un conjunto de 5 proyectos e intentar cruzar la información de TravisTorrent con GitHubTorrent. Además de hacer el anteproyecto.
  - Tareas desarrolladas: #23, #24
  - Sprint review: los datos de los 5 proyectos se sacaron al final del sprint, por lo que su análisis se dejó para el Sprint 6.
- Sprint 6
  - Fecha inicio: 31/10/2018
  - Fecha fin: 16/11/2018
  - Sprint planning: en este sprint se analizan los datos de los 5 proyectos además de leer documentación relacionada con nuestro caso de estudio.
  - Tareas desarrolladas: #26, #25
  - Sprint review: se encontraron unos artículos en los que se proporciona una base de datos de bugs pública, por lo que se decide continuar investigando por esta vía en futuros sprints.
- Sprint 7
  - Fecha inicio: 16/11/2018
  - Fecha fin: 30/11/2018
  - Sprint planning: en este sprint se analizan los papers relativos a la base de datos de bugs, tanto el análisis de los artículos como la descarga de los datos.
  - Tareas desarrolladas: #27
  - Sprint review: los datos se consiguen descargar satisfactoriamente.
- Sprint 8
  - Fecha inicio: 30/11/2018
  - Fecha fin: 14/12/2018

- Sprint planning: en este sprint se decide elaborar un modelo de predicción con los datos descargados, así como ver cómo extraer estos datos de un commit cualquiera.
- Tareas desarrolladas: #29, #28
- Sprint review: en un entorno local se consiguen extraer los datos satisfactoriamente, por lo que debemos pasar a realizar este proceso en un servicio en la nube.
- Sprint 9
  - Fecha inicio: 14/12/2018
  - Fecha fin: 14/01/2019
  - Sprint planning: se decide estudiar dos entornos posibles para realizar el despliegue del servicio: DeployHub y Heroku.
  - Tareas desarrolladas: #32, #33
  - Sprint review: DeployHub tenía una documentación más desactualizada y tenía una complejidad notable, por lo que se opta por elegir Heroku como plataforma de despliegue.
- Sprint 10
  - Fecha inicio: 14/01/2019
  - Fecha fin: 30/01/2019
  - Sprint planning: en este sprint se intenta dejar el servicio desplegado en Heroku.
  - Tareas desarrolladas: #34, #31
  - Sprint review: durante el sprint surgieron problemas derivados de la ejecución de SourceMeter en Heroku, por lo que no se consiguió tener una versión estable desplegada.
- Sprint 11
  - Fecha inicio: 30/01/2019
  - Fecha fin: 08/02/2019
  - Sprint planning: en este sprint tratamos de arreglar los problemas del sprint anterior, así como gestionar el borrado de ficheros tras procesarlos, y el uso de workers que ejecuten la tarea en segundo plano.
  - Tareas desarrolladas: #35, #36
  - Sprint review: los errores se consiguieron subsanar, por lo que en los siguientes sprints se decidió pasar a mejorar el modelo de predicción.
- Sprint 12
  - Fecha inicio: 08/02/2019
  - Fecha fin: 22/02/2019
  - Sprint planning: en este sprint se realiza un análisis de los datos, así como un estudio de la estructura de la memoria. Tras el análisis, se intenta crear un modelo de predicción con buenos resultados.
  - Tareas desarrolladas: #43, #42, #18

## Anexo A.1. Imágenes de las células de estudio

- Sprint review: los modelos creados no consiguen dar buenos resultados, por lo que esa tarea se pasa al sprint siguiente.
- Sprint 13
  - Fecha inicio: 22/02/2019
  - Fecha fin: 08/03/2019
  - Sprint planning: en este sprint nos centraremos en la creación de un modelo de predicción. Concretamente, intentaremos crear modelos interpretables, para poder explicar al usuario el origen de nuestra decisión. Además, comenzamos con la memoria, escribiendo el capítulo introductorio.
  - Tareas desarrolladas: #45, #46, #44
  - Sprint review: los modelos creados siguen sin proporcionar resultados aceptables.
- Sprint 14
  - Fecha inicio: 08/03/2019
  - Fecha fin: 29/03/2019
  - Sprint planning: se intenta una nueva alternativa para crear un modelo con buenos resultados: probar Support Vector Machine, realizar una modificación de los datos de entrenamiento, etc. Además, en este sprint se define la estructura del capítulo relativo al estado del arte.
  - Tareas desarrolladas: #48, #47, #50, #49
  - Sprint review: tras un tercer sprint estudiando modelos de predicción y seguir sin obtener buenos resultados, se decide parar esta vía de estudio.
- Sprint 15
  - Fecha inicio: 29/03/2019
  - Fecha fin: 05/04/2019
  - Sprint planning: se opta por un método descriptivo, por lo que el siguiente paso consiste en ver las 10 variables más importantes en la predicción de fallos. Además, continuar con el capítulo del estado del arte.
  - Tareas desarrolladas: #53, #52
  - Sprint review: El servicio consigue hacerse descriptivo.
- Sprint 16
  - Fecha inicio: 05/04/2019
  - Fecha fin: 29/04/2019
  - Sprint planning: se decide modificar el servicio para que el usuario proporcione directamente su repositorio local, de manera que podamos ver los cambios y mostrar las métricas de los cambios locales.
  - Tareas desarrolladas: #56, #60, #55, #54
  - Sprint review: el servicio queda acabado, así que lo que queda de proyecto se centrará en la documentación.
- Sprint 17
  - Fecha inicio: 29/04/2019

- Fecha fin: 17/05/2019
- Sprint planning: continuar con la memoria, explicando la implementación del servicio. En cuanto al servicio, se decide que el repositorio suba a Amazon S3 para poder procesarlo posteriormente por el worker. Además, acabar de crear el flujo de trabajo de CI/CD.
- Tareas desarrolladas: #37, #63
- Sprint review: el pipeline de CI/CD queda finalizado, por lo que solo nos queda acabar la memoria en los sprints restantes.

(A partir de aquí no hay reuniones como tales, no sé si poner que no ha habido planning ni review y poner simplemente resumen de cada uno. Creo 3 sprints ya que queda mes y medio para entregarlo.)

- Sprint 18
  - Fecha inicio: 17/05/2019
  - Fecha fin: 31/05/2019
  - Sprint planning: acabar el capítulo de la implementación y comenzar a explicar la metodología de trabajo
  - Sprint review: ¿?
- Sprint 19
  - Fecha inicio: 31/05/2019
  - Fecha fin: 14/06/2019
  - Sprint planning: ¿?
  - Sprint review: ¿?
- Sprint 20
  - Fecha inicio: 14/06/2019
  - Fecha fin: 28/06/2019
  - Sprint planning: ¿?
  - Sprint review: ¿?



## A.2. Dockerfiles para la construcción de los contenedores

Dockerfile Web Container	
1	FROM paberlo/alpine-scikit-django-jdk8
2	
3	# variables para dockerizacion
4	ENV HOME=/home
5	ENV PROYECTO=tfm_diego
6	
7	#copiar codigo
8	COPY webapp/ \$HOME/\$PROYECTO
9	
10	# cambiar punto de entrada al directorio del proyecto
11	WORKDIR \$HOME/\$PROYECTO
12	
13	CMD python3 manage.py runserver 0.0.0.0:\$PORT
14	FROM paberlo/alpine-scikit-django-jdk8

Dockerfile Worker Container	
1	FROM paberlo/alpine-scikit-django-jdk8
2	
3	# variables para dockerizacion
4	ENV HOME=/home
5	ENV PROYECTO=tfm_diego
6	
7	#copiar codigo
8	COPY webapp/ \$HOME/\$PROYECTO
9	
10	# cambiar punto de entrada al directorio del proyecto
11	WORKDIR \$HOME/\$PROYECTO
12	
13	CMD python3 worker.py
14	FROM paberlo/alpine-scikit-django-jdk8

## A.3. Contenido del Docker Compose

Docker Compose	
1	<b>version:</b> '3.6'
2	<b>services:</b>
3	<b>web:</b>
4	<b>image:</b> tfm-diego-web:latest
5	<b>command:</b> python3 -u manage.py runserver 0.0.0.0:8000
6	<b>depends_on:</b>
7	- redis
8	<b>environment:</b>
9	- REDIS_HOST=redis
10	- PYTHONUNBUFFERED=0
11	- LOCAL=true
12	<b>ports:</b>
13	- "8000:8000"
14	<b>worker:</b>
15	<b>image:</b> tfm-diego-worker
16	<b>command:</b> python3 -u worker.py
17	<b>depends_on:</b>
18	- redis
19	<b>environment:</b>
20	- PYTHONUNBUFFERED=0
21	- LOCAL=true
22	<b>redis:</b>
23	<b>image:</b> redis:4.0.14-alpine

## A.4. Archivo de configuración de Travis

Travis CI Configuration	
1	<b>language:</b> python
2	<b>python:</b>
3	- '3.6'
4	<b>script:</b>
5	- python tests.py
6	<b>deploy:</b>
7	<b>provider:</b> script
8	# <a href="https://docs.travis-ci.com/user/deployment/script">https://docs.travis-ci.com/user/deployment/script</a>
9	<b>script:</b> bash \$TRAVIS_BUILD_DIR/travis_deploy.sh
10	<b>on:</b>
11	<b>branch:</b> master