UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería en Informática

**TFG del Grado en Ingeniería Informática**

# Predicción de apuestas deportivas

Presentado por Nuño Basurto Hornillos
en Universidad de Burgos — 23/01/2017
Tutores: Alvar Arnaiz González
Cristobal José Carmona del Jesús

UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería en Informática

D. Alvar Arnaiz González, profesor del departamento de ingeniería civil, área de sistemas y lenguajes informáticos.

Expone:

Que el alumno D. Nuño Basurto Hornillos, con DNI 71295798-F, ha realizado el Trabajo final de Grado en Ingeniería Informática titulado Predicción de apuestas deportivas de TFG.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 27 de diciembre de 2016

| | |
|---|---|
| Vº. Bº. del Tutor: | Vº. Bº. del co-tutor: |
| D. nombre tutor | D. nombre co-tutor |

**Resumen**

Se obtienen datos de pasadas jornadas de La Liga y con un algoritmo genético de backpropagation obtenemos una predicción del resultado de los siguientes partidos.

**Descriptores**

Scrapping, Bases de datos, Drupal, PHP, Backpropagation

**Abstract**

Get data of the previous matchweeks of La Liga and with a genetic algorithm of backpropagation we get a prediction of the result of the next games.

**Keywords**

Scrapping, Bases de datos, Drupal, PHP, Backpropagation

# Índice general

# Índice de figuras

# Introducción

Este trabajo de fin de grado tiene como objetivo predecir resultados en partidos de fútbol, a partir de estadísticas de los equipos en anteriores jornadas. Para ello se utiliza machine learning, más en concreto un algoritmo genético de Backpropagation. La idea es obtener

El trabajo de fin de grado ha sido realizado sobre una máquina virtual montada sobre el programa Oracle VM VirualBox, el Sistema Operaivo elegido ha sido Ubuntu 14.04 LTS, se ha instalado un servidor apache en él se ha instalado PHP y MySQL la extracción de datos se ha llevado a cabo mediante técnicas de Scrapping, se obtienen tanto los datos de los partidos de cada jornada como la clasificación al final de la jornada. Estos datos los almacenamos en la Base de Datos que gestionamos con PHPMyAdmin. La gestión web ha sido llevada a cabo con Drupal 7 ya que es un gestor web open source y con una buena comunidad detrás, además de tener facilidad a la hora de trabajar con PHPMyAdmin.

# Objetivos del proyecto

Este apartado explica de forma precisa y concisa cuales son los objetivos que se persiguen con la realización del proyecto. Se puede distinguir entre los objetivos marcados por los requisitos del software a construir y los objetivos de carácter técnico que plantea a la hora de llevar a la práctica el proyecto.

Actualmente podemos encontrar una gran cantidad de información sobre cualquier cosa que podamos imaginar, esta información muchas veces es demostrada en estadísticas para refutarla. El fútbol no es una excepción, en los análisis de los partidos se puede detectar que algunos equipos pierden cuando tienen una menor posesión por ejemplo. Es este escenario donde se intenta plantear este proyecto dado que pudiendo observar las tendencias de los equipos en sus partidos y sus rachas, el algoritmo puede entender las tendencias de cada equipo.

El objetivo principal es calcular el resultado final de un partido de fútbol, el algoritmo irá aprendiendo a lo largo de la temporada según vaya disponiendo de una mayor cantidad de datos.

Esta información la hacemos llegar al usuario a través de un entorno sencillo donde el usuario no únicamente dispone de la información del resultado, sino que también le proporcionamos información de las diferentes cuotas de algunas casas de apuestas, para que de esta manera pueda elegir en cual apostar. La manera de mostrarle os datos al usuario es sencilla, "1"si se trata de la victoria del equipo local, "X"si el partido va a acabar empate y "2"si la victoria va a caer del lado visitante, se ha decidido mostrar de esta manera ya que en España estamos acostumbrados a esta simbología gracias a la Quiniela.

2.1

Se va a programas un algoritmo genético en PHP que tenga difere

# Conceptos teóricos

En aquellos proyectos que necesiten para su comprensión y desarrollo de unos conceptos teóricos de una determinada materia o de un determinado dominio de conocimiento, debe existir un apartado que sintetice dichos conceptos.

## 3.1. Scraping

Técnica utilizada para simular la navegación de un humano en Internet con el fin de extraer información del sitio web. Los datos pueden ser almacenados y analizados en una base de datos central o en otros lugares de almacenamiento. Son diversas las técnicas que se pueden utilizar, en concreto se ha empleado[2]

## 3.2. Machine Learning

Aquel proceso que le da a las computadoras la habilidad de aprender sin ser explícitamente programadas. Hay una segunda definición mucho más clara: "Se dice que un programa de computación aprende de la experiencia E con respecto a una tarea T y alguna medida de rendimiento P, si el rendimiento en T, medido por P, mejora con la experiencia E". El Machine Leraning se divide en dos áreas principales, el aprendizaje supervisado y el no supervisado, mientras que el primero requiere la intervención humana, el segundo no.[**?**]

## 3.3. Minería de datos

Es el proceso por el cual es posible detectar información procesable de conjuntos de datos, para después transformarla en una estructura comprensible y poder así utilizarla más tarde. Tiende a confundirse con el Machine Learning, pero su principal diferencia es que mientras este último se usa para reproducir

patrones conocidos y hacer predicciones basadas en patrones, la minería de datos descubre patrones desconocidos.

## 3.4. Neuronas artificiales

Similares a la idea de neuronas biológicas, forman parte de redes neuronales artificiales. Reciben una serie de entradas y dan una salida, la cual se ve condicionada por tres funciones: función de propagación, función de activación y función de transferencia.[1]

## 3.5. Redes neuronales artificiales

Imitan el funcionamiento de las redes neuronales biológicas, donde un conjunto de neuronas artificiales trabajan unidas, a fin de resolver problemas relacionados con el reconocimiento de formas o con la predicción. Se parte de un conjunto de datos de entrada significativo para conseguir que la red aprenda las propiedades deseadas.

## 3.6. Aprendizaje Supervisado

Se dispone de un conjunto de ejemplos de el cual se conoce la respuesta, por lo que el objetivo es marcar una regla o correspondencia de manera que seaposible aproximar la respuesta para todos los objetos que se presenten. La salida de la función puede ser un valor numérico o una clase.[3] El uso más extendido del aprendizaje supervisado consiste en hacer predicciones a futuro basadas en comportamientos o características ya vistas en los datos que se contienen.[**?**]

## 3.7. Backpropagation

También denominada propagación hacia atrás de errores, es un tipo de algoritmo con aprendizaje supervisado utilizado para el entrenamiento de las redes neuronales artificiales. Una vez aplicado un patrón, este se propaga a través del resto de capas hasta generar una salida, la cual se compara con la salida establecida como objetivo ya que se trata de aprendizaje supervisado. Se obtiene un error que se propaga hacia atrás cambiando los pesos, acercando así el algoritmo a un mejor resultado. A medida que entrena la red neuronal, las capas intermedias aprenden a organizarse para así reconocer diferentes características del espacio de entrada.[4]

## 3.8. Preprocesamiento

Los datos reales conducen en numerosas ocasiones a la extracción de patrones y reglas poco útiles, lo cual puede deberse al ruido de los datos o a su inconsistencia. La finalidad del preprocesamiento es la creación de información homogénea.

## 3.9. Demonios

También llamados servicios, son un tipo especial de procesos que se ejecutan en segundo plano.(Wkipeda) El sistema los inicia en el arranque, ejecutándose en este caso una tarea planificada y comprobando si la fecha del sistema coincide con las fechas de la jornada. Para la ejecución del demonio se emplea un gestor de servicios, el cron. El demonio también es posible ejecutarlo a mano ya que es posible que en un momento determinado no se encuentre el servidor iniciado y no haya sido posible la ejecución del demonio.

# Técnicas y herramientas

En esta sección se explica brevemente la metodología utilizada en el desarrollo del trabajo de fin de grado. También veremos las herramientas elegidas para el llevar a cabo el mismo y su elección frente a otras.

## 4.1. Técnicas

### Scrum

Es aquel proceso que se aplica un conjunto de prácticas para trabajar en equipo y de esta manera obtener el mejor resultado posible en un proyecto.

Se realizan entregas parciales hasta realizar la entrega del producto final. Estas entregas parciales las denominamos Sprints, al final de cada Srpint el equipo se reúne y decide cuales son los objetivos de cara al próximo Sprint.

Scrum es indicado para aquellos proyectos en los que tenemos requisitos poco definidos, que están abiertos cambios, ya que lo indicado entonces es realizar objetivos a corto plazo y así agilizar el desarrollo del mismo.

### Web scraping

Técnica utilizada para simular la navegación de un humano en Internet con el fin de extraer información del sitio web. Los datos pueden ser almacenados y analizados en una base de datos central o en otros lugares de almacenamiento. Son diversas las técnicas que se pueden utilizar, en concreto hemos realizado a través de peticiones al protocolo HTTP.

## 4.2. Herramientas

### Trello

Trello (https://trello.com/) es una herramienta colaborativa que nos permite organizar nuestro proyecto en tablones. En el tablero ponemos diferentes tarjetas y en cada una de ellas una lista. Hemos decidido usarla principalmente por su sencillez y facilidad de uso. Otra razón que nos ha llevado a su uso es la aplicación para Android que nos permite estar conectados en cualquier momento.

### GitHub

Git Hub (https://github.com/) es una plataforma que nos permite el desarrollo colaborativo del software, alojando el repositorio de código en ella. Su elección se debe al conocimiento de su funcionamiento, además de su buena adaptación a Ubuntu, sistema operativo en el que se ha trabajado y que mendiante sencillos comandos nos permite utilizarla fácilmente.

### Oracle VM Virtual Box

Virtual Box (https://www.virtualbox.org/) es un software libre con el cual hemos trabajado a lo largo de toda la carrera, cumple bien nuestras necesidades además de ser sencillo en su uso. Hay algunas alternativas como por ejemplo Hyper X que se pueden instalar en Windows, pero es necesario tener el sistema operativo en una versión Pro a la cual no tenemos acceso.

### Servidor web Apache

Apache (http://httpd.apache.org/) es un servidor web Open Source y multiplataforma, utilizado para realizar servicio a paginas web estáticas o dinámicas. Su elección de debe a que es el servidor web más conocido, además de tratarse de un software Open Source. Se adapta perfectamente a nuestras necesidades en el trabajo de fin de grado.

### Drupal

Drupal (https://www.drupal.org/) es un gestor de contenido Open Source que posee una gran comunidad, es combinable con MySQL. Posee una gran comunidad de módulos sobre los que apoyarse. Su sencillez de combinación con MySQL y el hecho de que se trata de un software Open Source, nos han hecho decantarnos por este CMS. La gran cantidad de módulos también han contribuido a su uso. La versión elegida ha sido Drupal 7 dado que esta tiene una madurez mucho mayor que Dupal 8, sobre todo para la elaboración de una pequeña web como la nuestra, la falta de tiempo para aventurarse y la comunidad detrás han sido claves.

### Sublime Text

Sublime Text (https://www.sublimetext.com/) nos permiten editar código fuente de un programa, ayudando en la simplificación de la escritura y resaltando la sintaxis haciendo mas sencillo escribir el código es un editor de texto gratuito, que no libre que nos permite trabajar con una gran variedad de idiomas. En concreto nos permite trabajar con PHP, que ha sido el lenguaje en el que hemos desarrollado gran parte del proyecto.

### Zotero

Zotero (https://www.zotero.org/) es una herramienta de gestión de información que nos ayuda a gestionar las referencias bibliográficas. Obetenemos las referencias que deseamos utilizando la extensión para el navegador Chrome y lo exportamos en formato BibTex a Latex.

### TexMaker

TexMaker (http://www.xm1math.net/texmaker) es una moderna plataforma que integra as diferentes herramientas que se ncesitan para desarrollar documentos con LaTeX. Su uso viene motivada por tratarse de una de las mejores herramientes para LaTeX y que tiene una licencia GPL.

# Aspectos relevantes del desarrollo del proyecto

Este apartado pretende recoger los aspectos más interesantes del desarrollo del proyecto, comentados por los autores del mismo. Debe incluir desde la exposición del ciclo de vida utilizado, hasta los detalles de mayor relevancia de las fases de análisis, diseño e implementación. Se busca que no sea una mera operación de copiar y pegar diagramas y extractos del código fuente, sino que realmente se justifiquen los caminos de solución que se han tomado, especialmente aquellos que no sean triviales. Puede ser el lugar más adecuado para documentar los aspectos más interesantes del diseño y de la implementación, con un mayor hincapié en aspectos tales como el tipo de arquitectura elegido, los índices de las tablas de la base de datos, normalización y desnormalización, distribución en ficheros3, reglas de negocio dentro de las bases de datos (EDVHV GH GDWRV DFWLYDV), aspectos de desarrollo relacionados con el WWW... Este apartado, debe convertirse en el resumen de la experiencia práctica del proyecto, y por sí mismo justifica que la memoria se convierta en un documento útil, fuente de referencia para los autores, los tutores y futuros alumnos.

## 5.1. Lenguaje para el desarrollo del proyecto

Desde un primer momento se determinó el empleo de PHP como lenguaje central en la implementación del proyecto dado que permite el desarrollo web de contenido dinámico. El proyecto se ha llevado a cabo con Sublime Text, lo cual ha permitido una implementación del código más sencilla ya que iba mostrando posibles predicciones de texto. PHP ha hecho posible aplicar alguna técnica de programación orientada a objetos, que se puede observar en el código del algoritmo de Backpropagation donde existen dos clases. Se barajó la opción de implementar el código en Python y dejarlo embebido sobre PHP

pero la idea inicial se descartó debido a que algunas funciones como las de acceso a la base de datos son únicas desde Drupal.

### PHP en Drupal

La base de datos de MySQL estaba vinculada con Drupal, por lo que existe acceso desde Drupal a la base de datos. Para el acceso, el código PHP debía tener algunas sentencias exclusivas. La función db_query tiene una sintaxis similar a las que se conocen ya que el contenido es una consulta MySQL sencilla. En cambio, otras funciones como db_update, db_select o db_insert tienen una estructura completamente diferentes a las conocidas aunque son sencillas de utilizar.

Para poder desarrollar todo el código sobre Sublime Text y que no hubiese necesidad de copiarlo continuamente en los nodos de Drupal, fue necesario añadir un inclue_once tanto en el nodo de Drupal como en el script.

## 5.2. Datos de entrada a la red neuronal

Es preciso darle un sentido a los datos que se han ido recopilando mediante los algoritmos de scraping, ya que el objetivo es darle a la red neuronal unos datos a partir de los cuales pueda ir aprendiendo y encontrar patrones. Son dos los algoritmos de scraping que fueron recopilando los datos que más tarde han sido empleados como entrada a la red Neuronal.

El primer algoritmo extrae todas las estadísticas de cada uno de los partidos de una jornada y los almacena en la base de datos. El segundo extrae la posición y las estadísticas de cada equipo en cada jornada, para almacenarlos posteriormente en la base de datos. A partir de estos datos recopilados se calculan las rachas de los equipos, donde se tiene en cuenta los puntos de cada equipo en las últimas jornadas acontecidas, los goles a favor y los goles en contra. De esta forma se favorece al algoritmo para conocer la tendencia del equipo en las últimas jornadas. En el training, por cada una de las instancias se le pasan al algoritmo un total de 78 columnas, donde las 30 primeras contienen los estadísticas del partido, de la 30 a la 54 las rachas y las estadísticas del equipo local en la clasificación y de la 54 a la 78 lo correspondiente al equipo visitante. Como función objetivo se establece 0 si ganó el equipo local, 1 si ganó el equipo visitante y 0.5 si el resultado final fue empate. Antes de la entrada de los datos al algoritmo se realiza una normalización de los mismos respecto a la misma columna del resto de instancias, siendo la normalización:

$$v' = \frac{v - min_a}{max_a - min_a}$$

## 5.3. Optimización del algoritmo de Backpropagation

Cuando los datos se encuentran en la red neuronal es preciso optimizar el número de neuronas que se van a utilizar. Con un número de neuronas pequeño, la red neuronal puede ser incapaz de aprender todos los patrones existentes, teniendo problemas para devolver el ouput deseado. En cambio, si el número de neuronas es demasiado grande, la red neuronal se queda sin margen de adaptación a cambios. Otro factor a tener en cuenta son las epochs que va a ejecutarse el training, ya que ante un pequeño número el algoritmo puede no aprender suficiente, y si el número es muy alto termina memorizando cada una. Esto supone que no se logren los resultados deseados en el test, por lo que se ha llevado a cabo un estudio para determinar cual es la cantidad más adecuada de neuronas y de epochs para el aprendizaje de la red neuronal. De esta manera, se considera la mejor configuración para la red neuronal aquella que menos error ha dado.

## 5.4. Interfaz sencilla para el usuario

Con todos los datos necesarios es importante conocer cómo mostrarlos al usuario que accede al sitio web. Dado que el objetivo de esta página es su uso con apuestas deportivas, se ha hecho un algoritmo de scraping que recopila las cuotas de las casas de apuestas para cada partido. A partir de los resultados que se han predicho, se muestran las cuotas de las cuatro casas de apuestas principales en España, ofreciendo al usuario la posibilidad de apostar en ellas. En una de la páginas del sitio web se puede observar el balance general a lo largo de la temporada, donde es posible acceder a cada una de las jornadas y ver más detalladamente lo ocurrido a lo largo de la temporada.

## 5.5. Automatización del funcionamiento

Dado que la obtención de datos debe hacerse periódicamente y la ejecución manual por parte del usuario es costosa, se ha automatizado la ejecución de todos los algoritmos en base al momento en el que deban ejecutarse. Por ejemplo, solo es posible obtener los resultados de una jornada una vez finalizada esta, o es posible fijarse en las cuotas de las casas de apuestas, las cuales pueden ir variando a lo largo de la semana debido a que no son valores fijos. Para cada jornada se establece una fecha_antes y una fecha_despues, los algoritmos de scraping de resultados se ejecutan una vez finalizada la jornada, es decir, si son posteriores a fecha_despues, algo similar ocurre con el scraping de casas de apuestas solo que en este caso lo mejor es obtener los datos justo antes de la jornada para tener los datos más recientes. Para el algoritmo de backpropagation no es necesario el día exacto en el que ejecutarse, pero debe

ejecutarse siempre una vez finalizada la jornada anterior, ya que son necesarias las rachas y los datos de la pasada jornada.

Todo esto lo logramos gracias al uso de demonios, también llamados servicios, que comparan la fecha actual a la fecha de la jornada, ya sea la anterior o posterior, y en caso de coincidir ejecutan el algoritmo deseado. El demonio se ejecuta todos los dias a la misma hora esto se debe a que no todas las jornadas comienzan el Viernes y acaban el Lunes si no que algunas jornadas se disputan entre semana.

# Trabajos relacionados

Este apartado sería parecido a un estado del arte de una tesis o tesina. En un trabajo final grado no parece obligada su presencia, aunque se puede dejar a juicio del tutor el incluir un pequeño resumen comentado de los trabajos y proyectos ya realizados en el campo del proyecto en curso.

# Conclusiones y Líneas de trabajo futuras

Todo proyecto debe incluir las conclusiones que se derivan de su desarrollo. Éstas pueden ser de diferente índole, dependiendo de la tipología del proyecto, pero normalmente van a estar presentes un conjunto de conclusiones relacionadas con los resultados del proyecto y un conjunto de conclusiones técnicas. Además, resulta muy útil realizar un informe crítico indicando cómo se puede mejorar el proyecto, o cómo se puede continuar trabajando en la línea del proyecto realizado.

# Bibliografía

[1] Servidor HTTP Apache, June 2016. Page Version ID: 91988556.

[2] Web scraping, November 2016. Page Version ID: 94838651.

[3] phpMyAdmin contributors. phpMyAdmin.

[4] Mazur. A Step by Step Backpropagation Example, March 2015.