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 — 13/01/2017
Tutores: Álvar Arnaiz González
Cristóbal José Carmona del Jesus

UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Grado en Ingeniería en Informática

D. Álvar Arnaiz González y D. Cristóbal José Carmona del Jesus profesores del departamento de Ingeniería Civil, Área de Lenguajes y Sistemas 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.

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, 12 de enero de 2017

Vº. Bº. del Tutor:

D. Álvar Arnaiz González

Vº. Bº. del co-tutor:

D. Cristóbal José Camona del Jesus

Resumen

La gran cantidad de información disponible a la que podemos acceder a través de Internet y la creciente capacidad de las técnicas de tratamiento automático de datos, está produciendo que disciplinas como la minería de datos estén ganando atención. Toda esta cantidad de información no siempre se utiliza, quedando «olvidada» en servidores aguardando ser explotada. Esta es la idea con la que surge este proyecto de fin de grado, tratar de buscar en la web información de partidos de fútbol, utilizarla para entrenar una técnica de clasificación y poder ser capaz de predecir resultados de próximos partidos. Obviamente, la tarea resulta ambiciosa e interesante.

Más concretamente, este proyecto busca diseñar e implementar una herramienta capaz de generar predicciones en partidos de fútbol, con el objetivo de proporcionar al usuario la información necesaria para que pueda orientarse a la hora de llevar a cabo apuestas deportivas. Todo ello se realizará en una interfaz web cliente-servidor pensada en la facilidad de uso para usuarios de cualquier edad.

Descriptores

Scraping, bases de datos, Drupal, PHP, apuestas, predicción, minería de datos, redes neuronales

Abstract

The big amount of information available nowadays through Internet and the increasing capacity of automatic data-processing techniques are leading to disciplines such as data mining being gaining attention. All this amount of information is not always used, being «forgotten» in servers waiting to be exploited. Based on this idea, this project try to search in the net information about football matches and use it to train a technique of classification which can be able to predict results of upcoming matches. Obviously, the task is ambitious and interesting.

More specifically, this project try to find how to design and implement a tool capable of generating predictions in football matches, with the aim of providing the user the necessary information in order to he can orientate himself when making sports bets. All of this will be done in a client-server web interface designed for ease of use for users of any age.

Keywords

Scraping, Bases de datos, Drupal, PHP, Backpropagation, data mining, neural networks

Índice general

Índice de figuras

Introducción

El Aprendizaje automático (Machine Learning) va siendo cada vez más habitual en nuestro día a día, necesitamos que las máquinas sean capaces de aprender de su propia experiencia en vez de tener que pre-programarlas por nuestra cuenta.

A su vez, vemos como las casas de apuestas son cada vez más numerosas, esto se debe a la gran popularidad que poseen y a que el negocio se encuentra en alza. En concreto vamos a tratar con las apuestas deportivas, que nos ofrecen una infinidad de posibilidades en las que apostar. Nos vamos a centrar en los resultados finales de los partidos de primera división.

Para el desarrollo de estas técnicas es necesario tener una cantidad de datos suficientemente grande con la que trabajar, estos datos los obtendremos mediante técnicas de web scraping.

La idea de este proyecto es unir una técnica de aprendizaje automático y las apuestas deportivas, está técnica utiliza las redes neuronales y partiendo de la información previa, debe ser capaz de realizar un pronóstico.

Objetivos del proyecto

Actualmente podemos encontrar una gran cantidad de información sobre cualquier cosa que podamos imaginar, esta información muchas veces aparece reflejada en estadísticas. El fútbol no es una excepción, en los análisis de los partidos se pueden detectar que algunos equipos tienden a perder cuando, por ejemplo tienen una menor posesión. 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.

2.1. Predecir resultados de próximos partidos

La red neuronal debe ser capaz de pronosticar el resultado de un partido de fútbol basándose en la experiencia adquirida de los pasados partidos y de las rachas calculadas, con lo que se va a llevar a cabo el entrenamiento de la red neuronal. Finalizado el entrenamiento se realizará el pronóstico, donde el valor predicho se considera clase que es el pronóstico a mostrar al usuario, 1X2.

2.2. Diseño web centrado en el usuario

Esta información la hacemos llegar al usuario a través de un entorno sencillo donde el usuario no solo 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 los 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.

2.3. Automatización del algoritmo

La intervención humana deber ser mínima y por este motivo es necesario automatizar el proyecto. La necesidad de automatización viene de que no puede haber una persona pendiente de cuándo se deben ejecutar cada uno de los algoritmos. Es por ello que será necesaria la creación de un servicio de Linux que sepa cuando debe ejecutar cada uno de los algoritmos.

Conceptos teóricos

Es fundamental pensar en el usuario final de una página web en el desarrollo de la misma, ya que una interfaz intuitiva y sencilla permitirá una buena interacción entre el usuario y esta. Es por ello que se ha realizado un Diseño Centrado en el Usuario.

3.1. Diseño web centrado en el usuario

El diseño esta basado en que el usuario pueda conseguir sus objetivos, como encontrar información comunicarse y aprender.

Para tratar de conseguir los objetivos del usuario, es necesario conocer estos desde el principio del desarrollo. Por ello es necesario conocer el usuario, el uso que se va a dar de la aplicación o que es lo que necesita.

Centrar el diseño en los usuario lleva a investigar la reacción ante el diseño, conocer la experiencia de uso e innovar, siempre teniendo en cuanta mejorar la experiencia del usuario [4].

El Diseño Web Centrado en el Usuario se divide en varias fases. Como podemos observar en la Figura 3.1 algunas de las fases son iterativas.

Planificación

En necesaria una correcta planificación donde se identifiquen los objetivos, necesidades y requerimientos. En esta fase va a dividirse en etapas más pequeñas:

Modelado del usuario

Lo primero es tener en cuenta aquellas clases o perfiles de usuarios, que vamos a necesitar. Tendremos en cuenta la información a la que necesitan

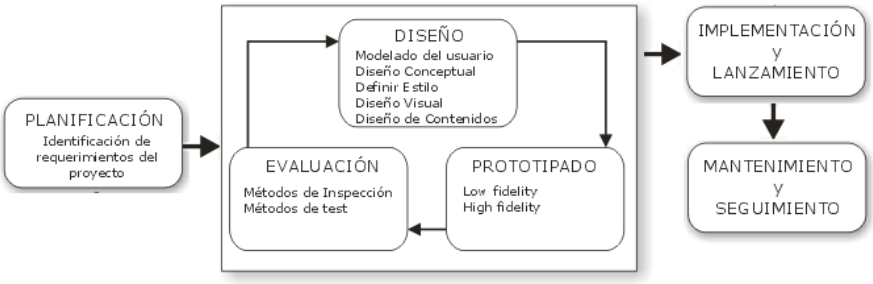

Figura 3.1: Fases del Diseño Centrado en el Usuario. Imagen extraida de [14]

acceder, las condiciones de acceso y su experiencia y conocimientos. Así se tener en cuenta para quien diseñar y lo que el usuario puede esperar encontrar.

Diseño conceptual

Se va a elaborar la estructura del sitio web, con estructura nos referimos a las conexiones y relaciones entre páginas y la posible navegación entre estas mismas. La arquitectura puede ser descendente donde nos desplazamos de lo general a las partes y ascendente, donde el desempeño global se logra con bloques mínimos de información.

Diseño visual

La facilidad con la que encontrar información, tiene que ver con la distribución que se hace de esta. La zona superior, por ejemplo, tiene una mayor jerarquía visual que las inferiores y es por ello que encontramos información más valiosa. Para variar la jerarquía de la información se utiliza diferente tamaño de letra o diferente fuente.

Lo importante es mantener una coherencia de diseño en todas las páginas. Un elemento que nos puede ayudar a lograr esto, son las páginas maestras.

Diseño de contenidos

El texto que nos encontremos debe tener un diseño y es que no puede ser caótico, de manera que encontrar la información resulte una quimera. Algunas características que debe tener son: tono cercano y familiar, conciso y preciso y que todo párrafo resuma una idea. Todo ello facilita enormemente que el usuario pueda localizar más fácilmente la información que desea.

Prototipado

Es necesario hacer prototipos de las interfaz del sitio que no tienen por que ser iguales al modelo final, pero nos servirán para evaluar la usabilidad

del sitio web sin tener que implementarla.

En un primer momento se puede realizar un prototipo en papel, donde veamos de una manera sencilla la distribución de la web. También existe software capaz de ayudarnos a realizar prototipos con este software incluso podemos interactuar con las diferentes páginas.

Evaluación de la usabilidad

Aunque hay diversas maneras de evaluar la usabilidad de un sitio web nos vamos a centrar en el test con usuarios ya que es el que se ha llevado a cabo.

Se va a realizar un análisis con un grupo de usuarios utilizando la web y se va apuntando aquellos problemas con los que se van encontrando. Los mismos usuarios muchas veces proporcionan algunas ideas que nos pueden permitir solventar algunos problemas. Es importante que los usuarios con los que se lleva a cabo la evaluación de usabilidad, tengan el mismo perfil o similar al del usuario final. Por ejemplo si elaboramos un sitio web para un tipo de usuario más experto, es normal que un usuario básico se encuentre con una infinidad de problemas, que el usuario experto no.

Implementación y lanzamiento

Es muy recomendable elaborar un sitio web con los estándares, ya que esto nos asegura que en el futuro haya una compatibilidad del sitio además de permitir un futuro crecimiento del mismo. También es importante el uso de hojas de diseño y bases de datos, ya que permite que si en un futuro hay que readaptar las necesidades, resulte más sencillo.

Una vez lanzado el sitio web los primeros meses son muy importantes, los usuarios son primerizos y deben encontrar una facilidad de uso, tambien podemos hacer pequeños tutoriales que enseñen a utilizar el sitio web.

Para mantener una correcta promoción del sitio es necesario incluir la empresa en banners, en buscadores o publicitarse a través del correo electrónico.

Mantenimiento y seguimiento

Según va pasando el tiempo es importante realizar un seguimiento de los usuarios que se conectan periódicamente a la web, ya que es este el volumen importante de usuarios. Hay que analizar por qué aquellos usuarios que se conectan por primera vez no lo vuelven a hacer. Una forma de hacerlo es generar un apartado de opinión de usuarios, donde estos nos expongan su opinión del sitio web para analizarlo más adelante.

Otras preguntas que debemos realizarnos son: ¿quién lo utiliza?, ¿cuándo lo utiliza?, ¿desde dónde lo utiliza?, ¿qué utiliza? Es muy importante tener

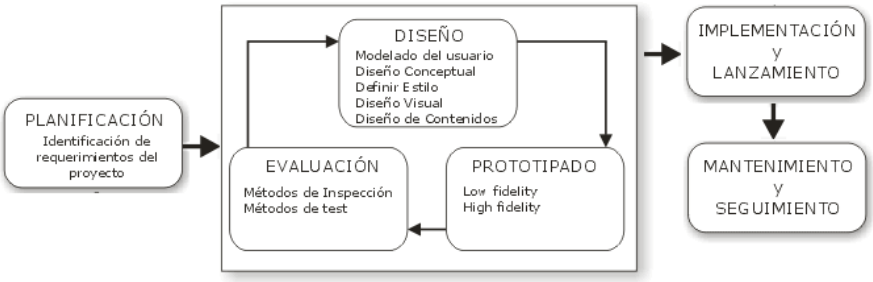

Figura 3.2: Esquema de funcionamiento de Aprendizaje Automático(Machine Learning) Imagen proporcionada por Álvar Arnaiz

una respuesta a estas preguntas, ya que nos puede permitir adaptarnos a los usuarios y de esta manera afianzarnos a ellos [4].

3.2. 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 cent no expandimos ral o en otros lugares de almacenamiento. Son diversas las técnicas que se pueden utilizar, en concreto se ha empleado el protocolo HTTP, mediante la observación de la estructura de la página vamos viendo como recorrerla y cuál es el sitio en el que encontramos los datos que queremos, una vez ahí los extraemos [18].

3.3. Machine Learning

En español denominado como aprendizaje automático, es 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 Learning se divide en dos áreas principales, el aprendizaje supervisado y el no supervisado, mientras que el primero requiere una etiqueta a predecir, en el segundo no se sabe. [19]

En esta figura 3.2 podemos ver un ejemplo sencillo del aprendizaje automático.

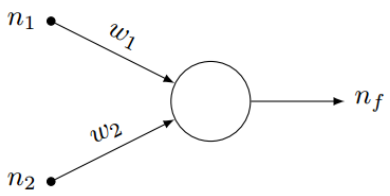

Figura 3.3: Neurona artificial, n_1 y n_2 son las entradas, mientras que n_f la salida. Imagen obtenida de [8].

3.4. Minería de datos

Es el proceso por el cual es posible detectar información tratable 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 [10].

3.5. Neuronas artificiales

Similares a la idea de neuronas biológicas, forman parte de redes neuronales artificiales. Reciben una serie de entradas y dan una salida similar a la figura 3.3. La salida se ve condicionada por tres funciones: función de propagación, función de activación y función de transferencia [8].

3.6. 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. Esto se logra modificando los pesos gracias al entrenamiento. Las redes se organizan por capas, habiendo capas de entrada, ocultas y de salida. Para que se comprenda la entrada se suele realizar un preprocesamiento [13].

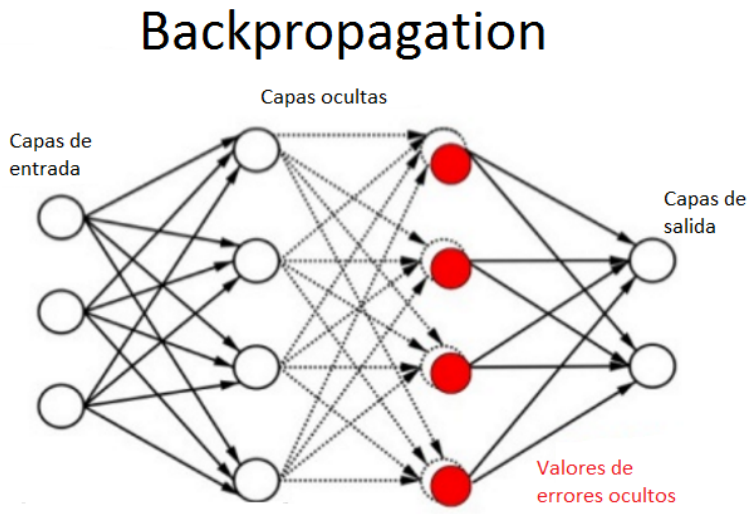

Figura 3.4: Esquema de una red Backpropagation. Imagen obtenida de [6]

3.7. 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 sea posible aproximar la respuesta para todos los objetos que se presenten. La salida de la función puede ser un valor numérico o una clase. 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].

3.8. 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, su esquema es similar al que podemos observar en la figura 3.4. 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 se entrena la red neuronal, las capas intermedias aprenden a organizarse para así reconocer diferentes características del espacio de entrada [11].

3.9. Preprocesamiento

El ruido y las anomalías de las bases de datos reales suelen conducir a la extracción de patrones y reglas poco útiles. La finalidad del preprocesamiento es la creación de información homogénea. En el proyecto se ha llevado a cabo la normalización [5].

3.10. Servicios

Aquellos procesos que se ejecutan en segundo plano, los denominamos servicios o demonios. 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, llamado 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 [15].

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 Sprint el equipo se reúne y decide cuales son los objetivos de cara al próximo Sprint [12].

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 [18]. Hay algunas herramientas que nos permiten la extracción de datos como *80legs* y *Scrapinghub*. Para el mejor aprendizaje de estas técnicas, se ha decidido llevarlo a cabo completamente a mano.

4.2. Herramientas

Trello

Trello [1] es una herramienta colaborativa que nos permite organizar nuestro proyecto en tablones. En el tablón ponemos diferentes tarjetas y en cada una de ellas una lista [1].

Hemos decidido usarla principalmente por el conocimiento de esta herramienta, ya que ha sido utilizada a lo largo de la carrera y sabemos de su sencillez. Otra razón que nos ha llevado a su uso es la aplicación para Android que nos permite estar conectados en cualquier momento y su integración en Drupal.

GitHub

Git Hub [2] 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.

El enlace a mi repositorio es el siguiente https://github.com/nunobasurto/apuestas_deportivas

Oracle VM Virtual Box

Virtual Box [3] 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 [4] 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 [16].

[1] https://trello.com
[2] https://github.com/
[3] https://www.virtualbox.org/
[4] http://httpd.apache.org/

Drupal

Drupal [5] es un gestor de contenido Open Source que posee una gran comunidad, es combinable con MySQL. Posee una gran cantidad 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 sistema gestor de contenido. 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.

Drupal no es el único gestor de contenido exixtente hay otros. Los más conocidos son *WordPress* y *Joomla*, *WordPress* es el que tiene una mayor facilidad de uso y extensibilidad en la plataforma, pero dado que somos usuarios más avanzados nos viene mejor Drupal, este tiene una mayor cantidad de funcionalidades aunque de mayor complejidad. El gestor *Joomla* ha realizado enormes avances en los últimos tiempos pero aún queda lejos de los dos gestores de referencia, es sin duda un gestor a tener en cuenta en los próximos años [2].

Sublime Text

Sublime Text [6] 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 [17].

Uno de los lenguajes con los que nos permite trabajar es PHP, que ha sido el lenguaje con el que hemos desarrollado gran parte del proyecto.

Zotero

Zotero [7] es una herramienta de gestión de información, que nos ayuda a gestionar las referencias bibliográficas. Obtenemos las referencias que deseamos utilizando la extensión para el navegador Chrome, que nos importa las referencias al programa *Zotero Standlone*. Desde aquí podemos exportar utilizando el formato BibTeX, permitiendo una correcta lectura en LaTeX.

[5] https://www.drupal.org/
[6] https://www.sublimetext.com/
[7] https://www.zotero.org/

Texmaker

TEXmaker [8] es una moderna plataforma que integra las diferentes herramientas que se necesitan para desarrollar documentos con LATEX. Inicialmente sin conocimiento de como utilizar la herramienta resulta costoso, con el tiempo ahorra mucho trabajo. Su uso viene motivado por tratarse de una de las mejores herramientas para LATEXy que tiene una licencia GNU GPL.

[8] http://www.xm1math.net/texmaker

Aspectos relevantes del desarrollo del proyecto

En este capítulo vamos a destacar los puntos claves del proyecto y como se han llevado a cabo.

5.1. Lenguaje para el desarrollo del proyecto

Desde un primer momento se decidió 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 entrenamiento, 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:

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

En un primer momento se pensó introducir todos los partidos anteriores como instancias, independientemente el equipo hubiese disputado el partido como local o como visitante, para diferenciarlo se establecía un campo en la tabla *clasificacion_jornada*, llamado *local_visitante*, este campo se encuentra a 0 si el equipo juega como local o 1 si juega como visitante. Dado que los resultados obtenidos no eran óptimos se planteó otra posibilidad: si el equipo iba a disputar el partido como local se tomaban únicamente los datos de los partidos como local para ese equipo y para el rival lo mismo, se iban a tomar los datos como visitante.

Jornada 12	Resultado	Pronóstico 1	Pronóstico 2
Alavés - Espanyol	0-1	1	2
Athletic - Villareal	1-0	X	X
Barcelona - Málaga	0-0	1	1
Deportivo - Sevilla	2-3	1	1
Valencia - Granada	1-1	X	X
Betis - Las Palmas	2-0	X	X
Atlético - Real Madrid	0-3	1	2
Sporting - Real Sociedad	1-3	1	2
Eibar - Celta	1-0	1	1
Leganés - Osasuna	2-0	X	X

Cuadro 5.1: Pruebas jornada 12.

Jornada 13	Resultado	Pronóstico 1	Pronóstico 2
Celta - Granada	3-1	1	1
Espanyol - Leganés	3-0	1	X
Málaga - Deportivo	4-3	2	1
Osasuna - Atlético	0-3	2	2
Real Madrid - Sporting	2-1	X	1
Real Sociedad - Barcelona	1-1	2	2
Villareal - Alavés	0-2	X	2
Sevilla - Valencia	2-1	X	1
Eibar - Betis	3-1	1	X
Las Palmas - Athletic	3-1	X	1

Cuadro 5.2: Pruebas jornada 13.

Se realizó un estudio durante tres jornadas, en concreto las jornadas 12, 13 y 14, para elegir una de las opciones planteadas previamente. En las tablas 5.1, 5.2 y 5.3 podemos observar los resultados obtenidos por los algoritmos. *Pronóstico 1* es el primer método donde se utilizaban todos los partidos, mientras que el segundo método donde diferenciabamos local y visitante es *Pronóstico 2*.

Una vez analizados los datos vemos como es el segundo método el que consigue mejores resultados.

Jornada 14	Resultado	Pronóstico 1	Pronóstico 2
Alavés - Las Palmas	1-1	X	X
Athletic - Eibar	3-1	1	X
Barcelona - Real Madrid	1-1	X	2
Deportivo - Real Sociedad	5-1	X	2
Valencia - Málaga	2-2	2	2
Betis - Celta	3-3	X	X
Atlético - Espanyol	0-0	2	1
Sporting - Osasuna	3-1	X	1
Granada - Sevilla	2-1	X	2
Leganés - Villareal	0-0	1	X

Cuadro 5.3: Pruebas jornada 14.

5.3. Optimización del algoritmo de Backpropagation

El problemas de las redes neuronales, es que estas necesitan ajustarse para su correcto funcionamiento, ya que no es lo mismo pasar instancias con un número reducido de columnas que instancias muy grandes.

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 la salida deseada. 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 iteraciones que van a ejecutarse, 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 iteraciones para el aprendizaje de la red neuronal.

Se utilizó la jornada 14 para ver como realizar el ajuste de la red neuronal, extraíamos el error que nos encontrábamos realizando el training, en la tabla 5.4 podemos observar el error obtenido.

Para realizar las pruebas se tuvo que aumentar el tiempo de espera del servidor, ya que no podía ejecutar a la vez todos los partidos. Como se ha podido observar en la tabla se han encontrado mejores resultados en las ejecuciones con 15 y con 20 neuronas cuyos valores son similares. Dado que es costoso para el servidor una ejecución con muchas neuronas, se optó por es-

Iteraciones	Neuronas					
	5	**10**	**15**	**20**	**25**	**30**
1000	0.9865	0.5632	0.2122	0.2902	0.3113	0.8873
2000	0.9865	0.4312	0.0145	0.0193	0.0105	0.7846
3000	0.9865	0.3214	0.0035	0.0055	0.0096	0.6575
4000	0.9865	0.2214	0.0035	0.0036	0.0096	0.5664
5000	0.9865	0.2132	0.0034	0.0034	0.0095	0.5664

Cuadro 5.4: Pruebas realizadas de la red neuronal.

coger la mejor solución con menos numero de iteraciones y neuronas, es decir, la configuración con 15 neuronas y 3 000 iteraciones.

5.4. Diseño web centrado en 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.

Dado que se ha realizado un diseño web centrado en el usuario, se elaboró un prototipo en papel (véase imagen 5.5) de como se iban a elaborar las pantallas de *Informe Jornada* e *Informe General*. En la parte superior de la imagen vemos un diseño utilizado para implementar *Informe Jornada* (véase imagen 5.6), donde si se acierta aparecen las cuotas y si se comete un error aparece un *-1* y abajo finalmente se realiza el sumatorio. En la parte inferior está el diseño utilizado para el *Informe General* (véase imagen 5.7), donde tenemos cada jornada con el sumatorio y en la parte inferior el sumatorio global.

5.5. Automatización del funcionamiento

Dado que la obtención de datos debe hacerse periódicamente y la ejecución manual por parte del usuario es tediosa, 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 finali-

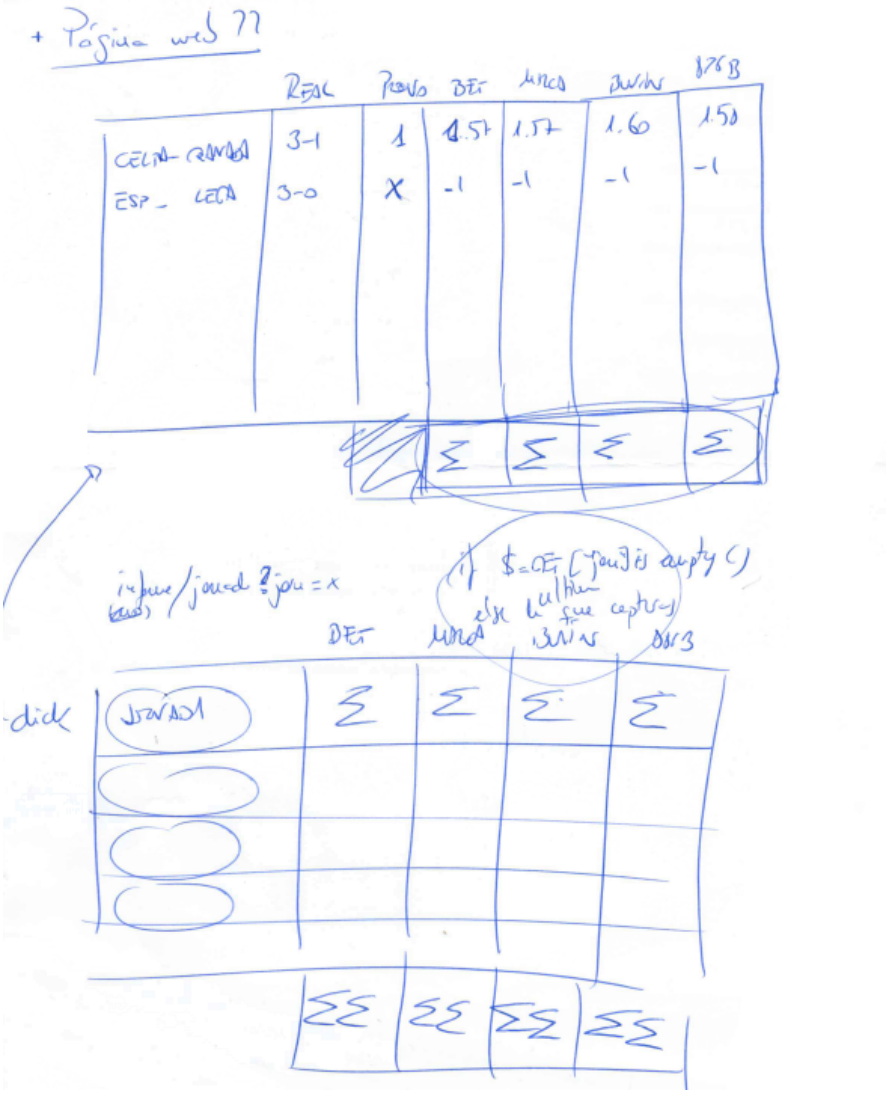

Figura 5.5: Prototipos diseño web.

Informe de la Jornada

Jornada 17

Partido	Resultado	Pronóstico	bet365	marcapuestas	bwin	888sport
Athletic-Alavés	0-0	1	-1.00	-1.00	-1.00	-1.00
Celta-Málaga	3-1	**1**	1.80	1.85	1.87	1.85
Espanyol-Deportivo	1-1	**X**	3.25	3.15	3.25	3.10
Osasuna-Valencia	3-3	**X**	3.50	3.60	3.60	3.60
Real Madrid-Granada	5-0	**1**	1.05	1.01	1.10	1.10
R. Sociedad-Sevilla	0-4	**2**	3.00	3.00	3.10	3.10
Villarreal-Barcelona	1-1	2	-1.00	-1.00	-1.00	-1.00
Real Betis-Leganés	2-0	X	-1.00	-1.00	-1.00	-1.00
Eibar-Atlético	0-2	X	-1.00	-1.00	-1.00	-1.00
Las Palmas-Sporting	1-0	**1**	1.53	1.55	1.57	1.55
Balance			**4.13**	**4.16**	**4.49**	**4.30**

Figura 5.6: Ventana Informe Jornada.

Informe General

Jornada

Jornada	bet365	marcapuestas	bwin	888sport
1	-1.20	-1.24	-1.13	-1.09
2	4.80	5.16	4.71	4.75
3	-0.15	-0.05	-0.21	-0.08
4	0.85	1.04	0.82	0.89
5	0.30	0.27	0.29	0.22
6	-3.75	-3.81	-3.74	-3.78
7	-5.10	-5.07	-4.92	-4.87
8	0.75	0.89	0.60	0.94
9	-3.39	-3.29	-3.14	-3.27
10	-1.40	-1.36	-1.31	-1.44
11	2.02	2.90	2.56	2.64
12	6.18	5.35	5.55	5.65
13	7.44	7.41	7.48	7.44
14	1.45	1.45	1.80	1.50
15	-2.60	-2.67	-2.76	-2.66
16	-2.03	-2.03	-2.15	-2.00
17	4.13	4.16	4.49	4.30
Balance final:	**8.30**	**9.11**	**8.94**	**9.14**

Figura 5.7: Ventana Informe General.

zada 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

Se va a hablar otros trabajos que utilizan técnicas de Machine Learning en la predicción de apuestas deportivas, como vamos a ver no solo se llevan a cabo en fútbol.

6.1. Neural Network Prediction of NFL Football Games

Este trabajo de Joshua Kahn trata de la predicción de futuros partidos en la NFL. El utiliza el lenguaje Perl y Matlab para la red neuronal. Los datos que utiliza lo extrae de la página oficial de la NFL con hojas de cálculo, en mi proyecto en cambio se están utilizando técnicas de web scraping y el almacenamiento en bases de datos.

J. Kahn utiliza técnicas de aprendizaje supervisado, más en concreto back-propagation, lo mismo que yo [9].

Aunque tiene un enfoque diferente a la hora de insertar datos y el deportes es diferente, el objetivo del proyecto y las técnicas utilizadas hacen que sea similar al mío.

6.2. Sistema de predicción de resultados en eventos deportivos

Trabajo de fin de grado llevado a cabo por Fernando Valera Guardiola, en la universidad Carlos III de Madrid. Ha utilizado WEKA para la creación de modelos de predicción, también ha utilizado un algoritmo genético para que la combinación de apuestas, maximice el beneficio minimizando el riesgo.

La recopilación de datos no fue capaz de realizarla de manera automática, por lo que tuvo que hacerla manualmente, algo muy tedioso ya que se invierte

una gran cantidad de tiempo. Por contra el proyecto que he desarrollado recopila los datos mediante web scraping, además de ser automático gracias al demonio de Linux.

Sin duda es un enorme trabajo el que llevó a cabo que no solo predice los partidos de la liga, si no también de otras grandes ligas europeas y de otros deportes como el baloncesto donde tiene en cuenta la NBA. También destacar que el tiempo de realización del trabajo han sido 8 meses [7].

Conclusiones y Líneas de trabajo futuras

7.1. Conclusiones

El proyecto que se ha llevado a cabo me parece muy ambicioso, el número de técnicas que abarca es bastante grande, gracias a ello he podido aprender bastante. Me hubiese gustado disponer de más tiempo para realizar estudios con una mayor exactitud ello hubiese permitido mejores resultados en la red neuronal. Pese a ello me encuentro muy satisfecho del trabajo llevado a cabo, ya que teniendo en cuenta el tiempo disponible me parece que ha sido un buen proyecto.

Los conocimientos adquiridos durante la realización de la carrera universitaria, me han ayudado de gran manera a llevar a cabo el proyecto. Las principales asignaturas sobre las que me he apoyado son:

- **Computación neuronal y evolutiva:** Sin duda la asignatura con la que guarda más relación ya que todos los conceptos de redes neuronales, así como el algoritmo de backpropagation han sido extraídos de aquí.
- **Sistemas operativos:** La navegación por Ubuntu mediante líneas de comandos, así como la creación de demonios son conocimientos vistos en esta asignatura.
- **Bases de datos:** El funcionamiento de la base de datos y las diferentes consultas llevadas a cabo, no hubiesen sido posible realizarlas sin los conocimientos tratados en esta asignatura. Aunque vimos PostgreSQL y aquí se ha trabajado con MySQL las similitudes son grandes.
- **Gestión de proyectos:** Sin duda otra de las asignaturas fundamentales en la planificación y gestión del proyecto, las metodologías ágiles han sido necesarias.

En cuanto a los conocimientos de la carrera que se han echado un poco de menos, una asignatura en la que aprender programación web sobre la cual apoyarnos para el desarrollo de los scripts en PHP o el web scraping.

7.2. Líneas de trabajo futuras

La escalabilidad que posee este proyecto es muy grande, sin duda puede continuarse en un entorno de apuestas deportivas incluyendo otros deportes, o es más en el propio fútbol cabe la posibilidad de pronosticar una mayor cantidad de tipos de apuestas como por ejemplo los goles a favor, los corners etc.

Incremento de la información

En lo primero que trabajaría sería en ofrecer al algoritmo una mayor cantidad de información con la que aprender, como los jugadores disponibles que tiene o los partidos de descanso que ha tenido el equipo si se ha disputado una competición entre semana.

Sin duda hay más factores a valorar que no han podido tenerse en cuenta, ya sea por falta de tiempo o por falta de recursos para recopilar esos datos.

Ampliación del entorno de actuación

El trabajo es ampliable a otros deportes como el baloncesto o el balonmano, sin llevar a cabo una gran modificación del mismo. Sin duda hubiese sido atractivo abarcar un mayor número de ligas europeas así como las competiciones europeas.

Fuera del fútbol, hubiese sido atractivo tratar con la NBA, sin duda la liga de baloncesto con mayor número de seguidores en el mundo, donde la gran cantidad de partidos que se disputan nos darían muchos datos con los que trabajar.

Pronóstico en tiempo real

Sin duda se trata de una de las tareas más complejas y a la vez más atractivas sobre las que trabajar en este proyecto. Que el usuario pueda acceder en tiempo real durante la disputa de partido a los pronósticos, supone un cambio en el proyecto pero seguiría los objetivos marcados por el mismo.

Bibliografía

[1] Acerca de Trello.

[2] WordPress, Drupal o Joomla. ¿Cuál es mejor?, January 2016.

[3] Alejandrocassis. Aprendizaje Supervisado, October 2015.

[4] Universidad de Granada. Diseño Web Centrado en el Usuario: Usabilidad y Arquitectura de la Información.

[5] Salvador García, Julián Luengo, and Francisco Herrera. Data Preparation Basic Models. In *Data Preprocessing in Data Mining*, number 72 in Intelligent Systems Reference Library, pages 39–57. Springer International Publishing, 2015. DOI: 10.1007/978-3-319-10247-4_3.

[6] Akash Goel. Backpropagation in neural networks, 2016.

[7] Fernando Valera Guardiola. Sistema de predicción de resultados en eventos deportivos y su aplicación en las apuestas, 2013.

[8] Guillermo Julian. Las redes neuronales: qué son y por qué están volviendo, 2014.

[9] Joshua Kahn. Neural network prediction of nfl football games, 2003.

[10] Cesar Krall. Minería de datos. ¿qué es? ¿para qué sirve?, 2010.

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

[12] Dave McKenna. The agile principles. In *The Art of Scrum*, pages 3–25. Springer, 2016.

[13] Monografias. Redes neuronales artificiales - fundamentos, modelos y aplicaciones, 2012.

[14] Diego Sanz. Diseño centrado en el usuario, 2014.

[15] Jesús Torres. Servicios y demonios en Linux, May 2013.

[16] Wikipedia. Servidor http apache — wikipedia, la enciclopedia libre, 2016. [Internet; descargado 11-enero-2017].

[17] Wikipedia. Sublime text — wikipedia, la enciclopedia libre, 2016. [Internet; descargado 11-enero-2017].

[18] Wikipedia. Web scraping — wikipedia, la enciclopedia libre, 2016. [Internet; descargado 11-enero-2017].

[19] Wikipedia. Machine learning — wikipedia, the free encyclopedia, 2017. [Online; accessed 5-January-2017].