UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

ANÁLISIS DE RESULTADOS DE ENCUENTROS DE FÚTBOL MEDIANTE TÉCNICAS DE MACHINE LEARNING

Eduardo Radío Gallego
Tutor: Manuel Sánchez-Montañés Isla

JUNIO 2017

ANÁLISIS DE RESULTADOS DE PARTIDOS DE FÚTBOL MEDIANTE TÉCNICAS DE MACHINE LEARNING

AUTOR: Eduardo Radío Gallego
TUTOR: Manuel Sánchez-Montañés Isla

Dpto. Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid

junio de 2017

Resumen (castellano)

El Trabajo de Fin de Grado que se presenta a continuación tiene como principal objetivo el estudio del resultado de aplicar algunas de las técnicas de aprendizaje automático más populares actualmente al campo de la predicción de partidos de fútbol en la Primera División de la Liga Española.

Este estudio ha dado lugar a la creación de un Sistema de Predicción de Resultados de Fútbol, lo suficientemente robusto como para ser capaz de generar beneficios sin necesidad del uso de ninguna estrategia adicional, como podrían ser la selección de apuestas según su relación beneficio/riesgo o las conocidas como apuestas de sistema.

Se han analizado tres temporadas distintas pertenecientes a los años 2012-13, 2014-15 y 2015-16, pero el diseño utilizado permite adaptar el sistema rápidamente a otras temporadas, competiciones e incluso deportes. Esto es posible gracias a que gran parte del proceso de recogida de datos y generación de atributos está automatizado y es fácilmente configurable.

Posteriormente, se procedió a implementar y observar los resultados que iban obteniendo los algoritmos k-NN, regresión logística, *random forest* y máquinas de soporte virtual, con unos u otros atributos, hasta elegir la combinación más adecuada.

Finalmente, elegido el modelo, se rescató el histórico de cuotas de entonces y se simuló que todos los algoritmos comenzaban con una banca e iban apostando a todo lo que predecían. Los resultados se pueden encontrar en el apartado correspondiente.

Palabras clave

Aprendizaje automático, apuestas deportivas, fútbol.

Abstract (English)

The Bachelor Thesis that is presented has as its main objective the study of the results of applying some of the most popular machine learning techniques nowadays to predict matches from the First Division of the Spanish League.

This study has led to the creation of a football results predictor system, robust enough to generate, in many cases, benefits without the need of using any additional strategy, such as selecting bets based on their risk/profit ratio, or the "system pick" bets.

Three seasons have been analysed: the ones from 2012-13, 2014-15 and 2015-16, although the design allows the adaptation to other seasons, championships and even sports. This is made possible because the data gathering process and attribute generation is highly automated and easily customisable.

Later, the results obtained by the k-NN, logistic regression, support vector machines and random forest algorithms, with a different set of attributes were compared, in order to pick the best combination.

Once the model chosen, the bet share history was put together and all the algorithms were set to use it, starting with a fixed amount and betting on all the results the saw fit. Results can be found on the appropriate chapter.

Keywords

Machine Learning, sport bets, soccer.

INDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

Cada vez son más las personas que se aventuran en el mundo de las apuestas, sobre todo desde que estalló el *boom* de los juegos de azar online. Sin embargo, muy pocas se detienen a evaluar fríamente cómo están invirtiendo su dinero. A menudo, influenciadas por las grandes ganancias potenciales que podrían recibir o por el sentimiento de afecto hacia un determinado club de fútbol, se toman decisiones precipitadas. El resultado, como norma general, es el de un balance negativo a largo plazo para este tipo de usuarios.

Por otra parte, dentro del campo en auge del Machine Learning, hay multitud de herramientas que ayudan a realizar modelos predictivos que permiten realizar predicciones a futuro en situaciones de incertidumbre. En este TFG se aborda el uso de este tipo de herramientas para elaborar modelos predictivos para resultados de encuentros de fútbol, y el estudio de su rentabilidad teórica si se hubiesen aplicado en la realidad. Las aplicaciones de este tipo de herramientas son, aparte de ayudar al usuario en sus apuestas, la detección de fraude en eventos deportivos (por ejemplo, investigando si hay equipos cuyos resultados deportivos tienen una probabilidad anormalmente baja de acuerdo con el modelo), o la mejora de la estimación inicial de las cuotas en las casas de apuestas.

1.2 Objetivos

Cabe destacar tres objetivos principales:

1. Generar una base de datos adecuada que incluya información relevante sobre los encuentros deportivos. Para ello, se recopilarán bases de datos de resultados, se condensarán en una sola, y se generarán atributos sintéticos que contengan información relevante.

2. Obtener modelos predictivos cuya tasa de aciertos supere de manera significativa al priori (es decir, predecir como resultado el resultado mayoritario hasta el momento, independientemente de los detalles del encuentro). Esto serviría para confirmar que la generación de la base de datos y la elección de atributos realizada en el anterior objetivo son adecuados.

3. Construir un sistema básico de recomendación de apuestas capaz de generar beneficios a largo plazo. Para ello se estudiará la rentabilidad de varios modelos predictivos y se escogerá el que mayores beneficios retorne.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

Capítulo 1. Introducción.

Capítulo 2. Estado del arte.

Capítulo 3. Diseño.

Capítulo 4. Desarrollo.

Capítulo 5. Pruebas y resultados.

Capítulo 6. Conclusiones y trabajo futuro.

2 Estado del arte

2.1 Introducción a las apuestas deportivas

2.1.1 Situación actual en España

Las casas de apuestas forman parte de uno de esos sectores en España a los que no les afectan los períodos de inestabilidad económica. Su crecimiento es constante en el paso del tiempo y las webs dedicadas a los juegos de azar están en pleno auge, como se puede ver en la figura 1. En concreto, las apuestas deportivas online son la modalidad que más crece en España dentro de los juegos de esta industria, con más de 2 millones de personas registradas a día de hoy [1].

Existen numerosas webs dedicadas a este negocio, como Bet365 [2], Bwin [3], William Hill [4] o Betfair [5], entre otras.

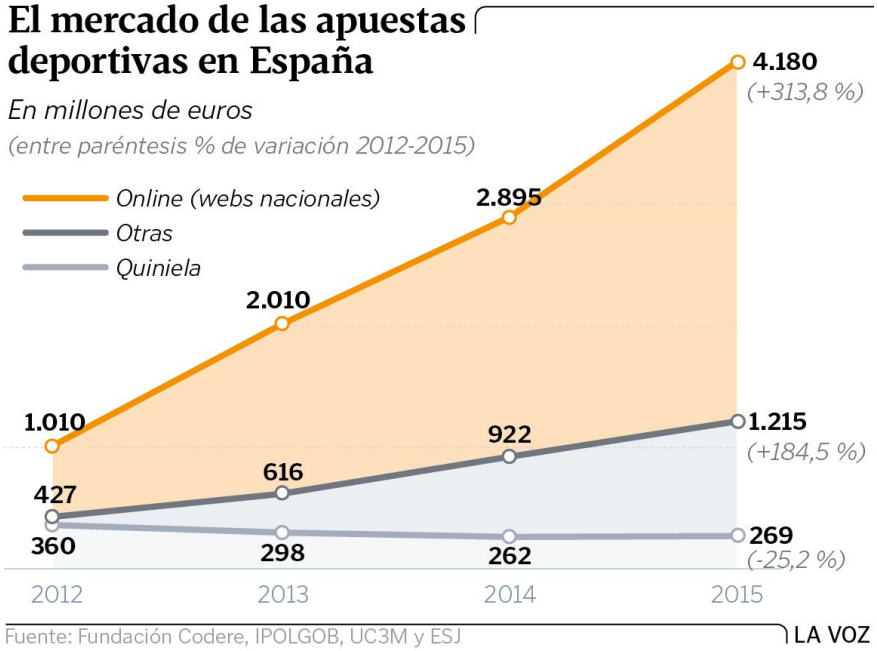

Figura 1. Evolución del volumen de negocio entre los años 2012 y 2015.

Como consecuencia, la ludopatía está más presente que nunca y cada vez son más jóvenes los afectados. Algunas asociaciones que luchan contra esta patología exigen que cambie la regulación de la publicidad en este sector de forma análoga a como ya se hizo con el tabaco o el alcohol, y que puede llegar a copar el 46% del espacio publicitario durante la retransmisión de este tipo de eventos [6].

2.1.2 Historia de las apuestas deportivas

Las apuestas deportivas son el juego de azar más antiguo del que se tiene constancia [7]. Hace más de dos mil años, con la aparición de las Olimpiadas, los griegos fueron los primeros en apostar a resultados de eventos deportivos. Más tarde, los romanos adoptaron

este juego como una modalidad de negocio, con los gladiadores de los circos romanos como protagonistas.

Sin embargo, no fue hasta los siglos XVIII y XIX, cuando se popularizaron las apuestas de carreras de galgos y caballos en Reino Unido. Medio siglo más tarde llegó a los Estados Unidos, comenzando definitivamente su expansión global.

A comienzos del siglo XXI el juego online ha supuesto una revolución en el sector. Desde su aparición, se ha producido un crecimiento exponencial del número de compañías, así como de su facturación y del número de usuarios [8].

2.1.3 El Negocio de las apuestas deportivas

Desde su creación como un modelo de negocio, las empresas dedicadas a este sector han ido mejorando sus estrategias para anticiparse a los apostantes y asegurarse así los beneficios. Y eso no es todo, algunas ya están integrando en sus plataformas web servicios de *exchange* o intercambio de apuestas, donde se apuesta contra otros apostantes y la casa se lleva una comisión por cada operación.

Las cuotas

Para entender un poco mejor cómo funciona este negocio, es necesario entender qué son y cómo funcionan las cuotas. Si se pregunta a cualquier persona que alguna vez haya realizado algún tipo de apuesta deportiva "¿qué es la cuota?" lo más probable es que responda algo así como "la cantidad de euros que recibes por cada euro apostado". Es decir, que si se acierta sobre un suceso en el que se invierte X€ y cuya cuota en el momento de la apuesta tenía un valor igual a Y, se reciben (X·Y) €.

Desde el punto de vista de probabilidad la cuota, en situaciones de equilibrio, debe coincidir con el período que tendría el suceso en condiciones idénticas a las actuales. En la siguiente figura se observa que la cuota para la apuesta "gana el Club Deportivo Tenerife" en su encuentro con el Cádiz Club de Fútbol es, según la casa de apuestas, 2.0, por lo cual el Tenerife debería ganar una de cada dos veces al Cádiz si se repitiera el encuentro en condiciones idénticas.

Figura 2. Ejemplo de cuotas asignadas al mercado 1-X-2 de un partido en Betfair.

Partiendo de esta última definición, se deduce la **probabilidad implícita** mediante la siguiente fórmula:

$$\mathbf{P(suceso) = 1 / cuota_{suceso}}$$

Si sumáramos las probabilidades implícitas de los tres sucesos (gana local, empatan y gana visitante), la probabilidad total superaría el 100%. Esto se debe a que las cuotas son ligeramente bajadas por la propia casa de apuestas para poder obtener un margen de beneficio conocido como *overround* [9].

Cuando una casa de apuestas decide ofrecer un determinado mercado relativo a un evento deportivo las cuotas asignadas a los distintos posibles sucesos son previamente estudiadas. Cuanto mayor es la repercusión del evento, mayores son los esfuerzos por acercarse a la probabilidad real [10].

Sin embargo, ninguna empresa es perfecta. Es común encontrar variaciones entre las cuotas de las diferentes casas de apuestas debido a que es imposible conocer exactamente la probabilidad de este tipo de resultados y cada una utiliza sus estrategias particulares. Por ello el apostante considera como buenas aquellas apuestas para las que estima una probabilidad mayor a la probabilidad implícita en la cuota.

2.1.4 Los Tipsters deportivos

Los *Tipsters Deportivos* son un perfil que ha resurgido en los últimos años, y cuya labor consiste en proporcionar información adicional sobre un determinado evento deportivo para incrementar la garantía de éxito a la hora de realizar una apuesta. Los *Tipsters* pueden darse en forma de aplicación, web, foro, o canal de chat. Estos "consejeros" suelen especializarse en un determinado deporte y, aunque no suelen revelar cómo realizan el estudio previo de los eventos, muchos afirman basarse primordialmente en su experiencia y conocimiento personales.

2.1.5 Fraudes deportivos

A lo largo de la historia se ha hecho evidente que la corrupción está muy presente en el deporte de alto rendimiento. De hecho, en la actualidad hasta 50 clubes son sospechosos de haber realizado algún tipo de fraude en la Liga 2016-17. Además, muchas veces son los propios jugadores quienes apuestan a los partidos que más tarde van a jugar [11].

Pese a que se suceden año tras año y son hasta cierto punto predecibles, este tipo de fraudes es muy difícil de demostrar. Sin embargo, las casas de apuestas no son ajenas a este problema, y utilizan todos los mecanismos a su disposición para prevenir e incluso aprovecharse de esta situación.

2.1.6 Sistemas de recomendación de apuestas / evaluación de riesgos

Existen multitud de proyectos presentes y pasados cuyo objetivo se basa en tratar de predecir y recomendar a sus clientes las mejores opciones de inversión en eventos deportivos. Entre los métodos que podríamos considerar como más rigurosos desde el punto de vista estadístico se han encontrado tres tipos bastante comunes a la hora de pronosticar partidos de fútbol:

Métodos basados en el estudio de la probabilidad implícita otorgada por la casa de apuestas. Estos sistemas son posiblemente los más comunes.

- Ventajas: se aprovecha el estudio realizado por la casa de apuestas.
- Inconvenientes: es muy complicado obtener beneficios a largo plazo mediante este método, ya que está diseñado por la propia casa para hacer perder dinero a los jugadores. Por tanto, se hace imprescindible dotar de más información al sistema con otro tipo de parámetros [12].

Métodos basados en usar algoritmos de clasificación como redes bayesianas, que se nutren de atributos subjetivos y muy específicos determinados por un experto.

- Ventajas: tiene buenos resultados en general a largo plazo [13].
- Inconvenientes: necesitan contar con un experto en fútbol que conozca al detalle el estado del equipo que se quiere pronosticar, así como los jugadores disponibles en cada partido u otros parámetros relevantes. Además, debe actualizarse el modelo con cada cambio de jugadores importantes y solamente es válido un modelo por equipo [13].

Métodos basados en la utilización de técnicas de Aprendizaje Automático (Machine Learning) sin necesidad de que un experto aporte información.

- Ventajas: utilizando los programas o librerías adecuados, es sencillo de implementar. En combinación con una buena estrategia de apuestas, que minimice riesgos y maximice beneficios, puede dar buenos resultados [14].
- Inconvenientes: sin una estrategia, es necesaria una base de datos muy informativa, con atributos bien escogidos. La cuota acaba ocupando un papel demasiado predominante en muchos, bien por falta de conocimiento de otros factores importantes, o bien por un exceso de confianza en ésta.

Este proyecto tiene, como principal diferencia respecto a los métodos anteriores, que no utiliza el histórico de cuotas en la labor de predicción. Esta decisión se tomó con el objetivo de poder explotar al máximo el sistema y encontrar aquellas situaciones en las que se maximice la diferencia entre la probabilidad otorgada por el modelo desarrollado en este TFG y la probabilidad implícita en la cuota.

2.2 Aprendizaje Automático

2.2.1 Introducción

El aprendizaje automático (del inglés, *Machine Learning*) es una rama de la inteligencia artificial que tiene como objetivo desarrollar técnicas que permitan a las computadoras crear modelos capaces de clasificar o diferenciar elementos según sus características. Se trata, por tanto, de un proceso de inducción del conocimiento [15]. En la actualidad sigue siendo un campo en continuo desarrollo, dado su potencial para predecir todo tipo de sucesos y las ventajas que ofrece respecto a los modelos clásicos o tradicionales que se basan en el estudio estadístico.

En ML hay diferentes familias de algoritmos dependiendo del tipo de problema: no supervisado (donde uno de los problemas principales es el de *clustering*), supervisado (donde los problemas principales son clasificación y regresión), semisupervisado, y técnicas de aprendizaje por refuerzo.

Dentro de los algoritmos de clasificación supervisada, algunos de los más populares actualmente, por tener buenas tasas de acierto en general, son *Random Forest* y Máquinas de Soporte Vectorial. Además de estas, se suelen estudiar otras técnicas más básicas para realizar comparativas de funcionamiento a nivel de tasas de acierto. Aquí consideraremos los siguientes algoritmos:

Clasificación mediante Vecinos más Próximos (k-NN)

Este algoritmo clasifica cada nuevo patrón buscando primero en la base de datos de entrenamiento aquellos k que se le parezcan más. Cuando los atributos de los patrones son números reales es habitual utilizar la distancia euclídea para determinar los k ejemplos más cercanos. A continuación, clasifica dicho patrón nuevo de acuerdo a la clase mayoritaria en esos k patrones, lo que equivale a realizar una votación entre ellos para decidir la clase que le corresponde al patrón nuevo [16].

Se trata de un aprendizaje "vago" ya que durante la fase de entrenamiento se encarga únicamente de almacenar datos sin construir explícitamente en esa fase ningún tipo de modelo o generalización. Esto le permite adaptarse con éxito a distintos dominios de un mismo problema [17].

Regresión Logística

Este método construye un modelo que intenta separar mediante fronteras lineales los elementos de las diferentes clases. Cuando se enfrenta a un problema con más de dos clases (multiclase) se aplica la estrategia conocida como o*ne-versus-all*, que consiste en dividir el problema en tantos problemas de dos clases como sea posible, y que cada uno de ellos realice un voto ponderado a la hora de etiquetar nuevos datos [18]. Debido a su sencillez el número de parámetros libres que tiene es pequeño, por lo que es un buen algoritmo sobre el que comparar resultados de otros más sofisticados.

Máquinas de Soporte Vectorial (SVMs)

La idea principal de esta técnica es construir un clasificador lineal en un espacio de atributos transformado, en el que el problema sea linealmente separable o esté cerca de serlo. Por ello, hay conceptualmente dos etapas diferentes: 1) Transformación del espacio de atributos a un espacio nuevo, normalmente de mucha más alta dimensionalidad. 2) Construcción en este espacio del separador lineal, al que además se le pide maximizar el margen de separación que deja entre una clase y la otra. Una de las claves de los SVMs es que abordan el paso 1) de manera implícita mediante el truco del kernel, de tal forma que no necesitan llevar explícitamente el problema al nuevo espacio transformado.

Desde su aparición en los años 90, esta técnica ha ido ganando popularidad y actualmente tiene una gran variedad de aplicaciones. Finalmente cabe mencionar que los

SVMs pueden funcionar con diferentes tipos de kernel, lo que permite que puedan adaptarse a diferentes tipos de problemas [19].

Clasificación mediante Random Forest

Este método de clasificación construye en el entrenamiento un conjunto de árboles de decisión. Las preguntas que realiza cada uno de esos árboles se seleccionan aleatoriamente, lo que asegura la diversidad en el conjunto de clasificadores. Por otra parte, los nodos finales no se construyen aleatoriamente sino que se etiquetan de acuerdo a la clase mayoritaria en ese nodo, por lo que la construcción de cada árbol no es totalmente aleatoria. La diversidad de los árboles construidos asegura que, al realizar la clasificación mediante votación, los aspectos relacionados con el sobreajuste tiendan a cancelarse. Este método actualmente es uno de los que mejores resultados proporciona, siendo muy utilizado [20].

Todos estos algoritmos comparten los mismos puntos débiles, queriendo destacar aquí dos muy relacionados:

- Por un lado, está la información no relevante: en todo problema de predicción hay atributos que no contienen información útil sobre la variable que se quiere predecir. Por otra parte, todos los atributos están influidos de alguna u otra forma por ruido, es decir, información aparentemente aleatoria y que no está correlacionada con la variable a predecir. Todo esto tiende a influir de manera negativa en la fase de aprendizaje del modelo.

- Por otro lado, está el sobreajuste. Este caso se da cuando un modelo predice perfecta o casi perfectamente sobre los datos de entrenamiento, pero falla estrepitosamente tratando de clasificar datos nuevos [21]. Esto se debe a que el modelo tiene demasiada flexibilidad (parámetros libres), tanta como para "aprender de memoria" los datos de entrenamiento. Debido a que el modelo ha tendido a "memorizar" los datos de entrenamiento, tiene una capacidad de generalización mala en patrones nuevos.

2.2.2 Aplicaciones del Machine Learning

En la actualidad hay una variedad enorme de aplicaciones actuales del ML. La utilización de estos algoritmos se ha convertido en fundamental en problemas tales como la predicción del comportamiento del clima, la demanda eléctrica, la evolución de los mercados financieros, la detección de malware y virus [22] o la predicción de las pérdidas y ganancias de una empresa [23]. A la aplicación de estos métodos en el ámbito empresarial se le conoce como Business Intelligence.

Otro campo de aplicación importante es la minería de datos en la web. Dicho campo surgió como respuesta a la necesidad de manejar e interpretar la gran cantidad de datos que se generan a diario en la red [24]:

En este campo son ampliamente utilizados los algoritmos de aprendizaje automático: los motores de búsqueda, el análisis de las redes sociales y los sistemas de recomendación utilizan este tipo de técnicas, por ejemplo. También el ML tiene un área de aplicación muy

relevante en el problema de la seguridad en internet, por ejemplo en algoritmos que aprenden a detectar virus y malware nuevo [25].

La visión artificial es otra área que se ha visto beneficiada por el desarrollo en los últimos años del Machine Learning. Numerosas librerías (si no todas) de visión artificial utilizan técnicas de ML para analizar las características de los datos, ajustar sus pesos, y en definitiva, aprender. Una de las más famosas, por tratarse de *open source*, es la de OpenCV. Por otra parte, la clasificación y el *clustering* en ML son el equivalente al reconocimiento y segmentación del espacio en la visión artificial (dónde se encuentran los distintos patrones) [26]

Dado el gran potencial que tiene el ML, existen diversas plataformas en Internet dedicadas a este campo. Una de las más conocidas y con una comunidad más grande es Kaggle [27]. En ella se realizan competiciones por crear los mejores modelos predictivos y la comunidad es muy activa y colaboradora.

2.3 Software para *Machine Learning*

Existe una enorme diversidad de herramientas disponibles para el ML. Aquí describo brevemente algunas de las más conocidas:

Librería Scikit-Learn. Se trata de una librería *open source* de Python para aprendizaje automático. El programa que se ha realizado en Python para este TFG utiliza los algoritmos facilitados por esta librería.

MATLAB. Es un entorno de desarrollo integrado con licencia propietaria. Se usa mucho en los centros de I+D y en las universidades. Ofrece una interfaz gráfica y una gran variedad de funcionalidades que lo hacen idóneo para este campo. Utiliza además su propio lenguaje de programación (M) [28].

Lenguaje de programación R. Este lenguaje apareció en el año 1993. Se distribuye bajo licencia de *software* libre. Es ampliamente utilizado en la minería de datos debido a que fue inicialmente diseñado para el análisis estadístico [29].

WEKA. Esta plataforma con licencia de *software* libre fue desarrollada por la Universidad de Waikato, Nueva Zelanda, de ahí su nombre (*Waikato Environment for Knowledge Analysis*). Dispone de una amplia gama de herramientas entre las que se incluyen: clasificación, regresión, *clustering* o visualización de los datos [30].

3 Diseño

3.1 Estructura del sistema

El sistema que se ha diseñado se compone de la siguiente estructura:

- Un programa **generador de bases de datos** de fútbol a partir de bases de datos existentes, cuya función es sacar a la luz información que subyace en los datos originales. De esta forma, se facilita la labor de los algoritmos que más tarde se utilizarán.

- El **módulo principal**, se nutre de los ficheros generados en el paso anterior para **instanciar y entrenar los** diferentes **clasificadores** que se pondrán a prueba. Este proceso de entrenamiento se realiza siguiendo siempre el orden cronológico de los datos. Dispone además de diferentes parámetros de configuración, para poder abordar los problemas desde diferentes perspectivas.

- Por último, el **simulador** de apuestas, que monitoriza la tasa de aciertos y el rendimiento económico previsto por cada uno de los clasificadores desde la primera jornada que todavía no se ha usado como entrenamiento, hasta la última jornada de la temporada. Éste último no aparece representado en la figura 3 dado que se encuentra incorporado al módulo principal.

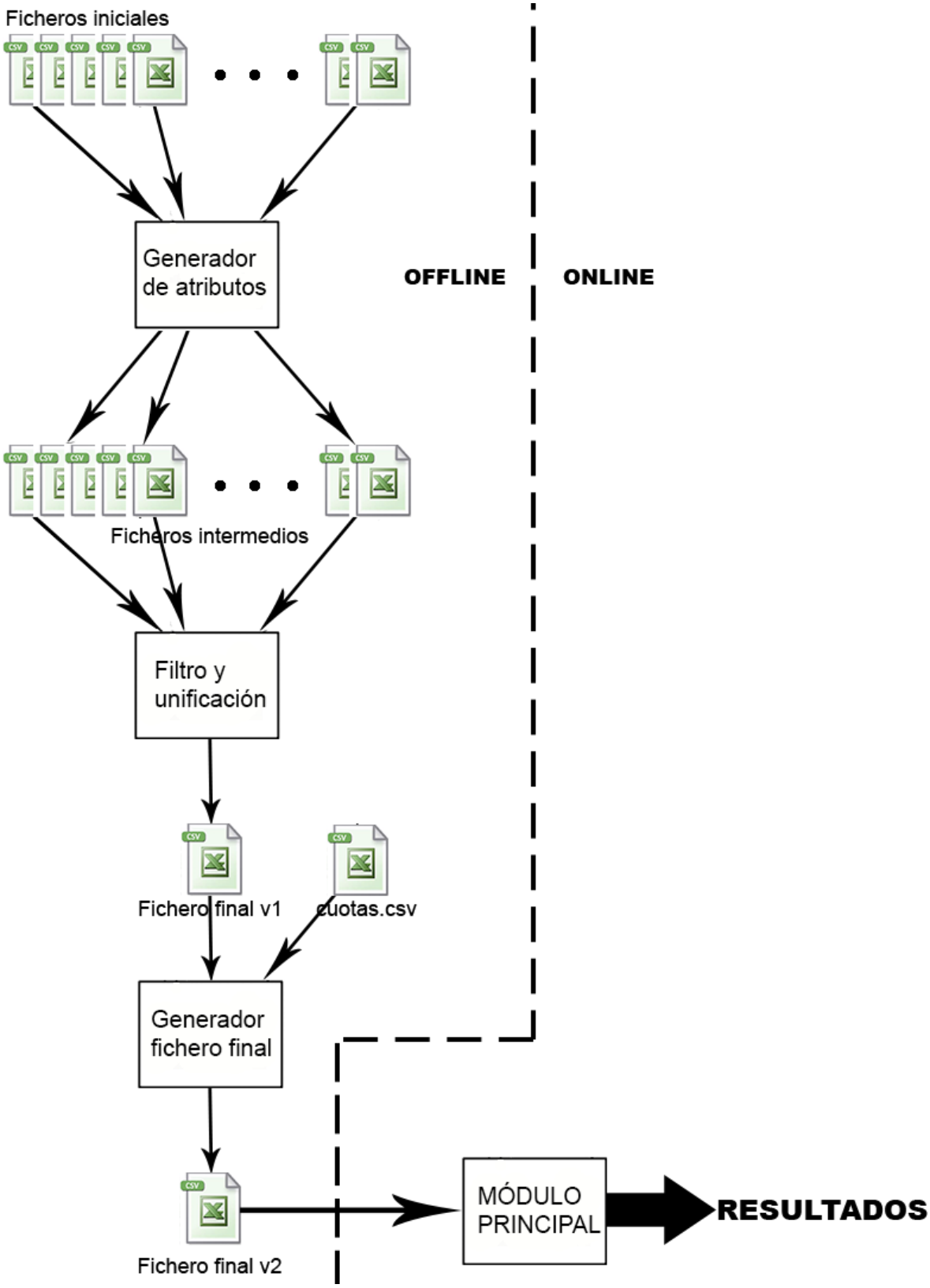

Figura 3. Diagrama de la estructura del sistema.

3.2 Base de Datos

Los datos utilizados han sido obtenidos de Internet, (concretamente de www.bdfutbol.com) y están disponibles de forma pública. Esta base de datos inicial incluye todos los partidos de todos los equipos de la liga española durante una temporada concreta.

Los datos se encuentran divididos en un fichero por cada equipo, y la información que caracteriza inicialmente a cada partido es la siguiente:

- Nombre de la competición
- Jornada/ronda
- Fecha
- Nombres de los equipos local y visitante
- Goles marcados en dicho partido por cada equipo.

Copa del Rey	9	29/1/2015	Athletic Club	1	0	Malaga
Liga	21	1/2/2015	Levante	0	2	Athletic Club
Liga	22	8/2/2015	Athletic Club	2	5	Barcelona
Copa del Rey	10	11/2/2015	Athletic Club	1	1	Espanyol
Liga	23	14/2/2015	Granada	0	0	Athletic Club
Europa League	Dieciseisavos de Final(1)	19/2/2015	Torino	2	2	Athletic Club
Liga	24	22/2/2015	Athletic Club	1	0	Rayo Vallecano
Europa League	Dieciseisavos de Final(2)	26/2/2015	Athletic Club	2	3	Torino
Liga	25	1/3/2015	Eibar	0	1	Athletic Club
Copa del Rey	11	4/3/2015	Espanyol	0	2	Athletic Club
Liga	26	7/3/2015	Athletic Club	1	0	Real Madrid

Tabla 1. Ejemplo del fichero inicial perteneciente al Athletic Club de Bilbao

3.2.1 Generación de atributos sintéticos

El objetivo de este proceso es generar una serie de atributos sintéticos cuya función es aportar información extra con la que trabajarán los algoritmos de aprendizaje automático y que se encuentra implícita en la base de datos inicial. Esto se realiza con cada equipo de manera individual, almacenando sus resultados en ficheros separados.

3.2.2 Filtro y unificación en un único fichero

En este segundo proceso se llevan a cabo una serie de tareas para preparar la base de datos para su correcta ejecución:

- Eliminación de partidos que no son de Liga. Hasta ahora estos partidos habían sido importantes para generar algunos atributos sintéticos, como la importancia del siguiente partido, cuando éste no es de Liga.

- Eliminación de atributos que ya no son necesarios (sirvieron para generar los atributos sintéticos).

- Unificación en un solo fichero que contiene todos los partidos de Liga, ordenados por fecha. En este paso, se “entrelazan” los partidos de los distintos ficheros, para

que ahora contengan la información adicional de los dos equipos involucrados en cada partido.

- Gestión de los partidos aplazados. En ocasiones, debido a problemas de calendario, los clubes se ven forzados a aplazar ciertos partidos de Liga. Esto supone un problema para nuestros modelos, ya que la fecha y la jornada de estos partidos no se corresponde.

  Para solventarlo, se modifican estas jornadas asignándolas como valor de jornada la media entre la jornada anterior y la jornada siguiente entre las que se haya el encuentro en realidad (por ejemplo, si el partido aplazado se jugó finalmente entre la jornada 19 y la 20, se le asigna un valor de jornada de 19.5).

3.2.3 Generación de más atributos y formato del fichero final

A partir del fichero unificado con todos los partidos, se simula una competición de Liga para comprobar que los datos son correctos y para obtener algunos atributos nuevos, como la posición del equipo en la tabla de clasificación en ese momento.

Finalmente, se genera un fichero en formato CSV [31] listo para ser procesado por el sistema.

3.3 Atributos Escogidos

La enorme variedad de parámetros que podrían haberse escogido deriva de la gran complejidad del problema al que nos enfrentamos. Como consecuencia, es común que a medida que se generan unos, se vayan deduciendo otros, convirtiendo el proceso de creación y desarrollo en un bucle. En la sección de Trabajos Futuros se recogen algunas observaciones y mejoras que podrían realizarse.

Atributos 1 y 2: Días hasta el partido siguiente y días desde el partido anterior

Los equipos profesionales de fútbol se encuentran a menudo inmersos en una serie de partidos consecutivos concentrados en un intervalo corto de tiempo. En estos casos los entrenadores se ven obligados a hacer cambios en la plantilla titular para dar descanso a jugadores “clave” y esto puede afectar al rendimiento del conjunto.

Con el objetivo de reducir el número de valores posibles y de facilitar la clasificación de instancias de los algoritmos, se ha realizado la siguiente discretización sobre ambos atributos:

Diferencia de días	Valor discreto
< 3	0
4, 5	1
> 5	2

Tabla 2. Correspondencias entre diferencia de días y su valor discreto

Atributo 3: Importancia del partido siguiente

La relevancia de este atributo reside en el mismo principio que los dos primeros: la necesidad de los técnicos de reservar a jugadores importantes para que jueguen a pleno rendimiento en los momentos decisivos de la temporada, como puedan ser las competiciones europeas.

De hecho, es tal la correlación que existe entre éste y los atributos 1 y 2, que se deja como estudio futuro una posible fórmula que los sintetice en una única variable.

Por el momento, se ha decidido utilizar la siguiente ponderación para representar esta cualidad:

Valor	Liga	Champions League o Europa League	Copa del Rey
0	Partido normal		Partido normal
1		Fase de grupos	Octavos y cuartos de final
2	Derby	Octavos y cuartos de final	Semifinal
3		Semifinal y final	Final

Tabla 3. Correspondencias entre la importancia de un partido y su valor discreto.

Atributos 4 y 5: Media de goles a favor del equipo local y del equipo visitante

Esta medida resulta muy efectiva a la hora de predecir resultados de partidos de fútbol. Esto es debido a que, si se observan un poco los datos históricos acumulados [25], es fácil ver que los equipos que más encuentros ganan son los más goleadores.

El valor de estos atributos es continuo.

Atributos 6 y 7: Media de goles en contra del equipo local y Media de goles en contra del equipo visitante

En este caso se cumple, de forma similar al caso anterior, que los equipos que más goles reciben suelen ser también los peor clasificados.

El valor de estos atributos es continuo.

Atributos 8 y 9: Aspiraciones del equipo local y visitante

Conocer qué aspiraciones reales tiene cada equipo en cada partido puede ayudar a realizar un pronóstico acertado. No es lo mismo jugar con un objetivo claro que sin él.

A continuación, se detalla cómo se ha codificado este atributo. El número de puntos de diferencia ha sido escogido en base al número de puntos que obtiene el equipo cuando gana un partido, en el caso del fútbol, tres.

Puntos de diferencia	Líder	Puestos de Champions League o Europa League	Salvación
> 6	0	0	0
< 7	1	1	1
< 4	2	2	2

Tabla 4. Correspondencias entre las aspiraciones de un equipo y su valor discreto.

4 Desarrollo

4.1 Fases de Desarrollo

Fase 1. Generador de bases de datos

En primer lugar, dado que uno de los objetivos es crear un modelo predictivo, será necesario extender nuestra base de datos inicial con nuevos atributos que permitan un análisis más exhaustivo de la base de datos de resultados deportivos. Todo esto se automatizó desarrollando para ello un programa en Python.

Fase 2: Implementación de modelos predictivos

A continuación, se procedió a implementar los algoritmos mediante otro programa en Python y utilizando las funciones facilitadas por la librería *Scikit-Learn.*

Fase 3: Simulación de apuestas

Finalmente, se han asociado las cuotas correspondientes a los partidos de los que se dispone y se ha simulado que los modelos apuestan por sí mismos a todo aquello que predicen. Esta funcionalidad sólo está disponible para el modelo anual, debido a que fue el que mejores resultados obtuvo inicialmente.

4.2 Modelos Implementados

Se han probado los 4 clasificadores descritos anteriormente (regresión logística, k-NN, *random forest* y *support vector machine)* sobre dos modelos distintos: anual y bianual.

4.2.1 Opciones de configuración.

En todos los modelos, los algoritmos comienzan a **aprender a partir una determinada jornada y a predecir a partir de otra**. Esto es necesario ya que al comienzo de la temporada hay ciertos parámetros que todavía no contienen información fiable y hay que dejar un cierto margen para que se estabilicen sus valores.

Es posible elegir entre 3 tipos de problema distintos a los que el sistema se enfrentará:

1. Gana local, empatan o gana visitante.
2. Gana local o no.
3. Gana visitante o no.

4.2.2 Modelo anual

Los algoritmos van pronosticando resultados semana tras semana, desde la decimotercera hasta la última jornada de la temporada. Las cinco primeras son descartadas por tener demasiado ruido y atributos “inmaduros”, como la media de goles.

4.2.3 Modelo bianual

Aquí se realiza el mismo procedimiento que en el modelo anual, pero en lugar de terminar al final de la temporada, guarda todo lo aprendido hasta entonces y lo aplica a la siguiente temporada, desde la quinta jornada. Las cinco primeras son de nuevo descartadas, debido en este caso a que hay tres equipos que han ascendido y por lo tanto carecemos de cierta información acerca de ellos todavía.

5 Pruebas y resultados

Se muestran a continuación los resultados del modelo anual para cada temporada y tipo de problema implementado.

En cada imagen se muestran, primero, la tasa de aciertos del clasificador a priori y de los 4 algoritmos seleccionados y, después, la evolución de estos 4 algoritmos suponiendo que cada uno empieza con 100 fichas en la jornada 13 y apuestan 1 ficha a todos los partidos restantes.

1. Gana local, empatan o gana visitante (1-X-2).

Se trata del problema que más beneficios retorna en general, posiblemente debido a que las cuotas resultantes de un problema de 3 clases son mayores que las del problema de 2 clases.

```
--------- TASAS DE ACIERTO EN LA TEMPORADA: 2012-13 ---------
PRIORI:                 49.1452991453%   -->    115/234
KNN:                    47.4358974359%   -->    111/234
LOG REG:                52.5641025641%   -->    123/234
RAND FOREST:            52.9914529915%   -->    124/234
SVC:                    52.1367521368%   -->    122/234
```

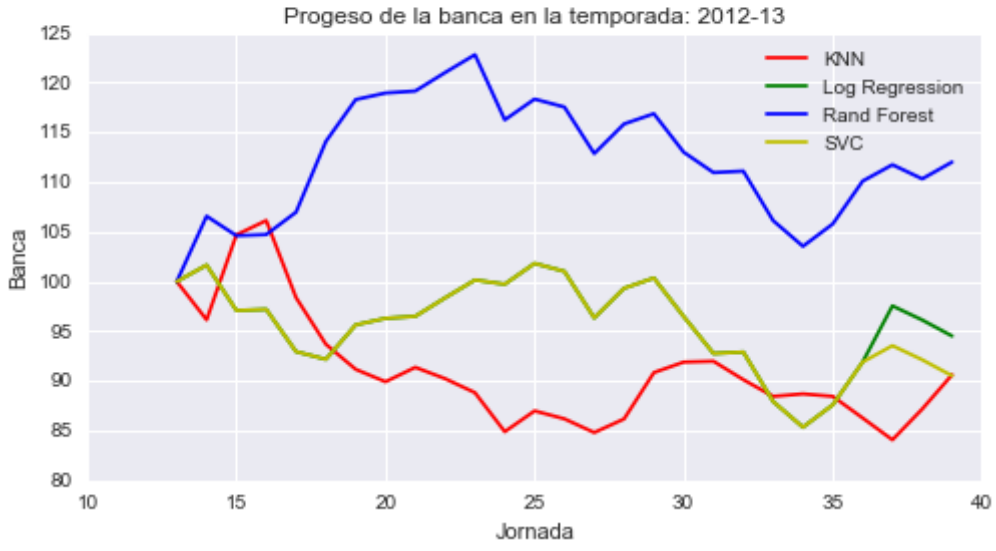

Figura 4. Tasas de acierto y evolución de las bancas de cada algoritmo correspondientes a la temporada 2012-13. Mercado 1-X-2.

```
--------- TASAS DE ACIERTO EN LA TEMPORADA: 2014-15 ---------
PRIORI:                 46.4%   -->     116/250
KNN:                    55.6%   -->     139/250
LOG REG:                53.6%   -->     134/250
RAND FOREST:            54.4%   -->     136/250
SVC:                    53.6%   -->     134/250
```

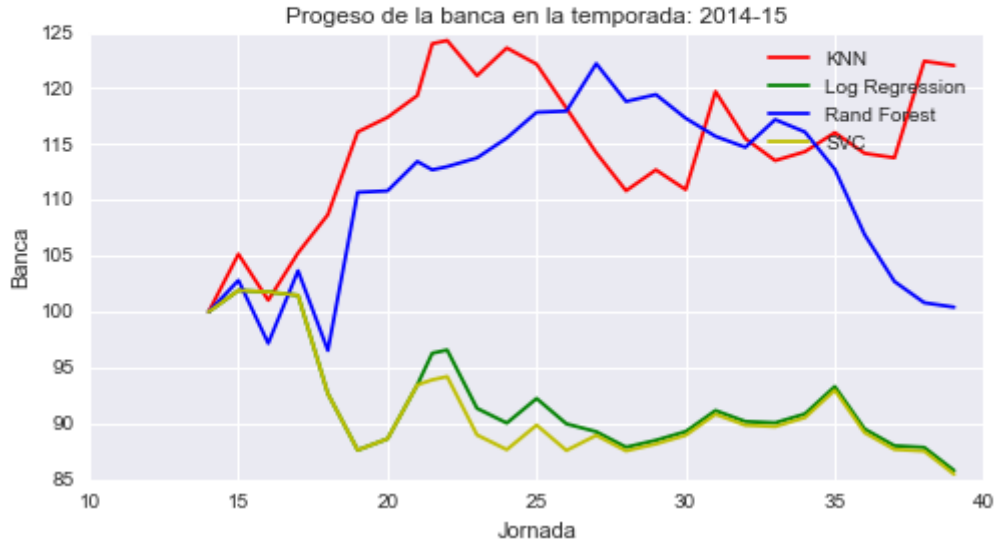

Figura 5. Tasas de acierto y evolución de las bancas de cada algoritmo correspondientes a la temporada 2014-15. Mercado 1-X-2.

```
--------- TASAS DE ACIERTO EN LA TEMPORADA: 2015-16 ---------
PRIORI:                 49.6%   -->     124/250
KNN:                    58.0%   -->     145/250
LOG REG:                55.2%   -->     138/250
RAND FOREST:            56.8%   -->     142/250
SVC:                    54.0%   -->     135/250
```

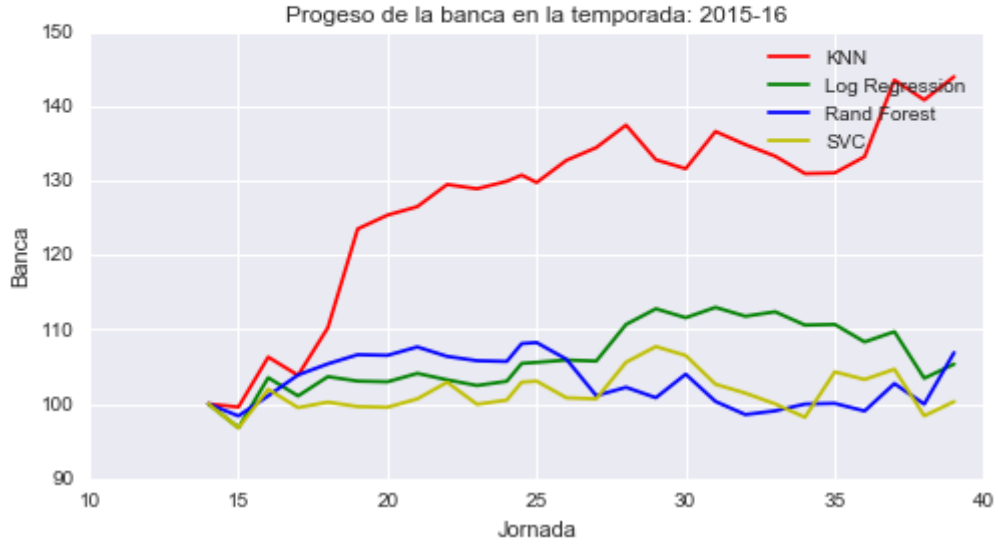

Figura 6. Tasas de acierto y evolución de las bancas de cada algoritmo correspondientes a la temporada 2015-16. Mercado 1-X-2.

2. Gana local o no (doble oportunidad 1-X2).

Puede observarse que, en este mercado, pese a obtener una tasa de aciertos mayor, las cuotas resultantes de combinar las clases correspondientes al empate y a la victoria visitante son muy bajas y el balance de los tres años de la banca es negativo, salvo para k-NN.

```
--------- TASAS DE ACIERTO EN LA TEMPORADA: 2012-13 ---------
PRIORI:                 47.8632478632%   -->    112/234
KNN:                    58.547008547%    -->    137/234
LOG REG:                61.1111111111%   -->    143/234
RAND FOREST:            58.547008547%    -->    137/234
SVC:                    59.4017094017%   -->    139/234
```

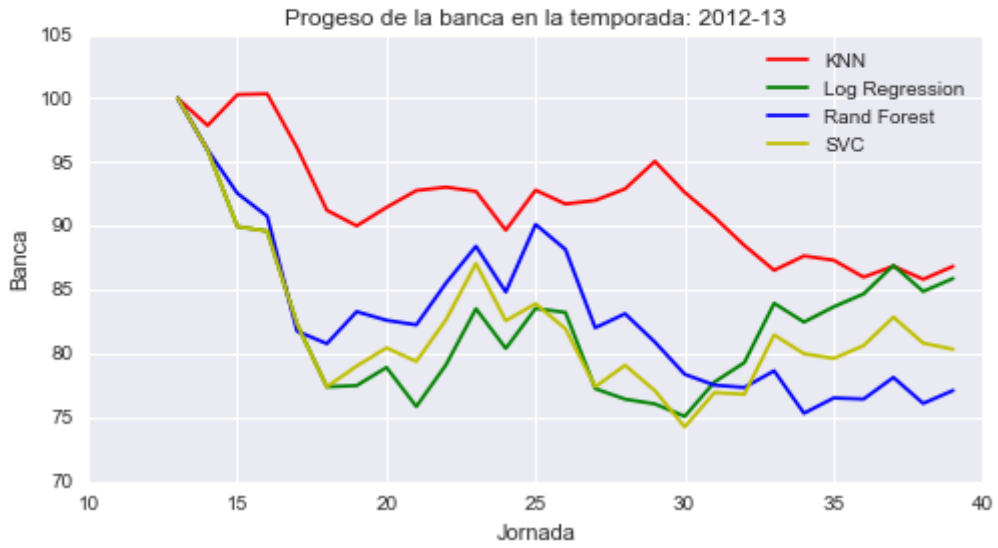

Figura 7. Tasas de acierto y evolución de las bancas de cada algoritmo correspondientes a la temporada 2012-13. Mercado 1-X2.

```
--------- TASAS DE ACIERTO EN LA TEMPORADA: 2014-15 ---------
PRIORI:                 53.6%   -->    134/250
KNN:                    71.2%   -->    178/250
LOG REG:                66.4%   -->    166/250
RAND FOREST:            66.4%   -->    166/250
SVC:                    66.8%   -->    167/250
```

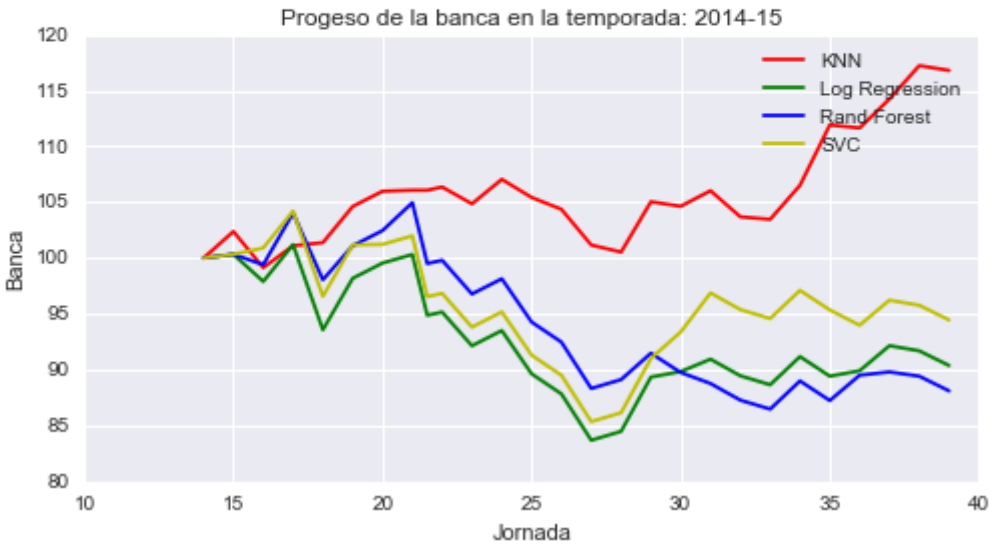

Figura 8. Tasas de acierto y evolución de las bancas de cada algoritmo correspondientes a la temporada 2014-15. Mercado 1-X2.

```
--------- TASAS DE ACIERTO EN LA TEMPORADA: 2015-16 ---------
PRIORI:                 50.4%   -->    126/250
KNN:                    62.8%   -->    157/250
LOG REG:                62.0%   -->    155/250
RAND FOREST:            65.2%   -->    163/250
SVC:                    61.2%   -->    153/250
```

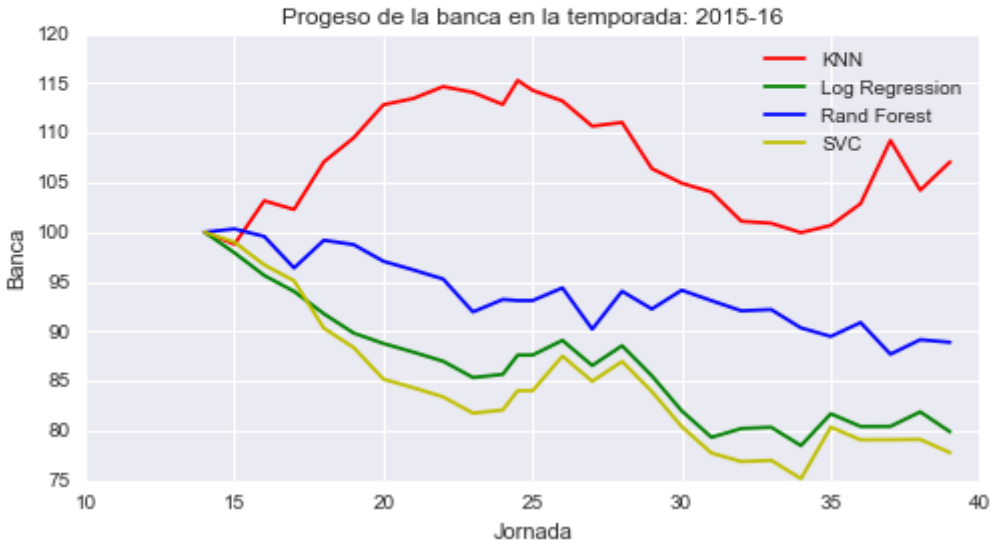

Figura 9. Tasas de acierto y evolución de las bancas de cada algoritmo correspondientes a la temporada 2015-16. Mercado 1-X2.

3. Gana visitante o no (doble oportunidad 1X-2).

Aquí ocurre algo similar al caso anterior. Las tasas de aciertos son las más altas de los tres modelos, pero es evidente que no se trata de una buena estrategia a largo plazo.

La única temporada que resulta rentable es la 2015-16, en la que todos los algoritmos salen ganando. Sin embargo, lo más probable es que se trate de una excepción por la alta predictibilidad de dicha temporada.

```
--------- TASAS DE ACIERTO EN LA TEMPORADA: 2012-13 ---------
PRIORI:                 73.0769230769%   -->    171/234
KNN:                    72.6495726496%   -->    170/234
LOG REG:                73.0769230769%   -->    171/234
RAND FOREST:            73.0769230769%   -->    171/234
SVC:                    73.0769230769%   -->    171/234
```

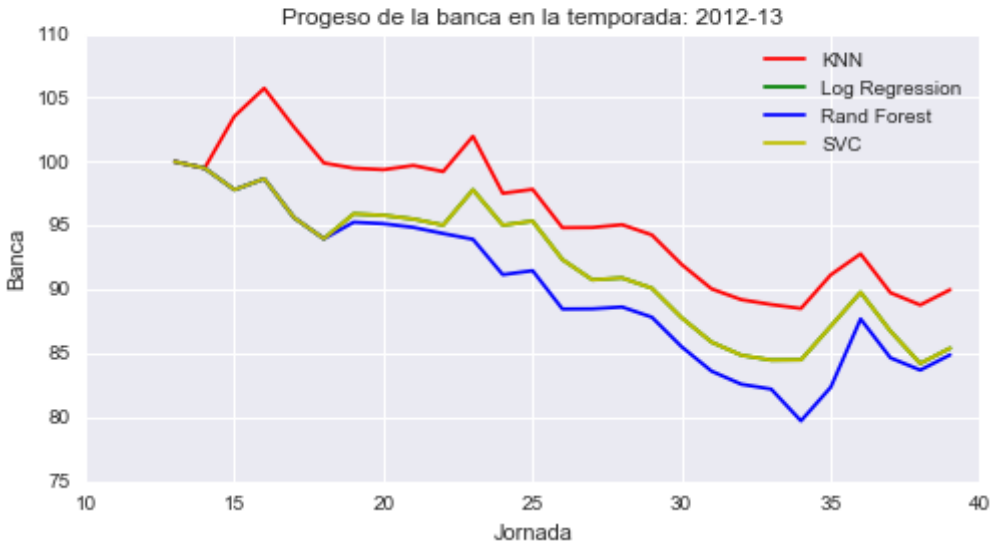

Figura 10. Tasas de acierto y evolución de las bancas de cada algoritmo correspondientes a la temporada 2012-13. Mercado 1X-2.

```
--------- TASAS DE ACIERTO EN LA TEMPORADA: 2014-15 ---------
PRIORI:                 69.6%   -->    174/250
KNN:                    72.0%   -->    180/250
LOG REG:                68.4%   -->    171/250
RAND FOREST:            72.0%   -->    180/250
SVC:                    68.0%   -->    170/250
```

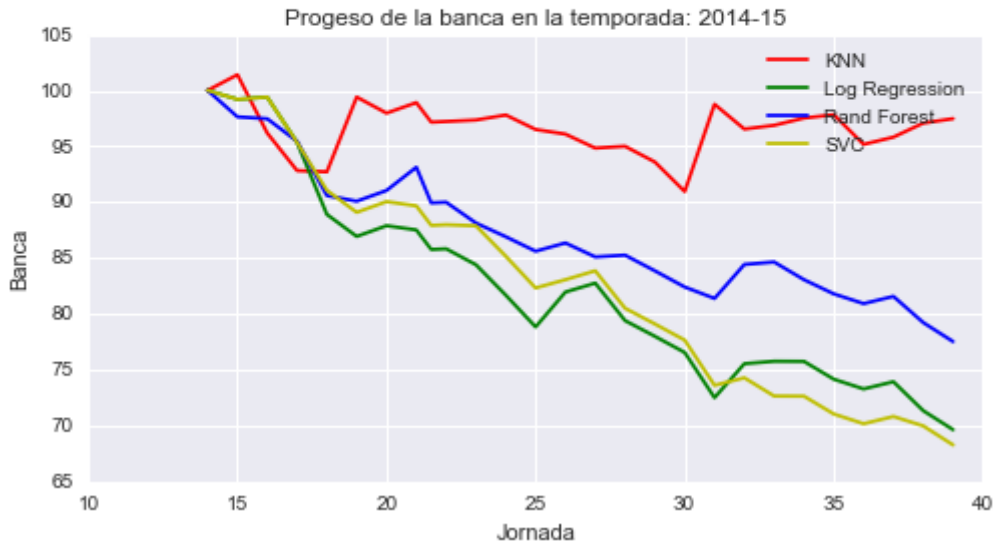

Figura 11. Tasas de acierto y evolución de las bancas de cada algoritmo correspondientes a la temporada 2014-15. Mercado 1X-2.

```
--------- TASAS DE ACIERTO EN LA TEMPORADA: 2015-16 ---------
PRIORI:                 74.0%   -->    185/250
KNN:                    77.2%   -->    193/250
LOG REG:                78.0%   -->    195/250
RAND FOREST:            76.4%   -->    191/250
SVC:                    78.0%   -->    195/250
```

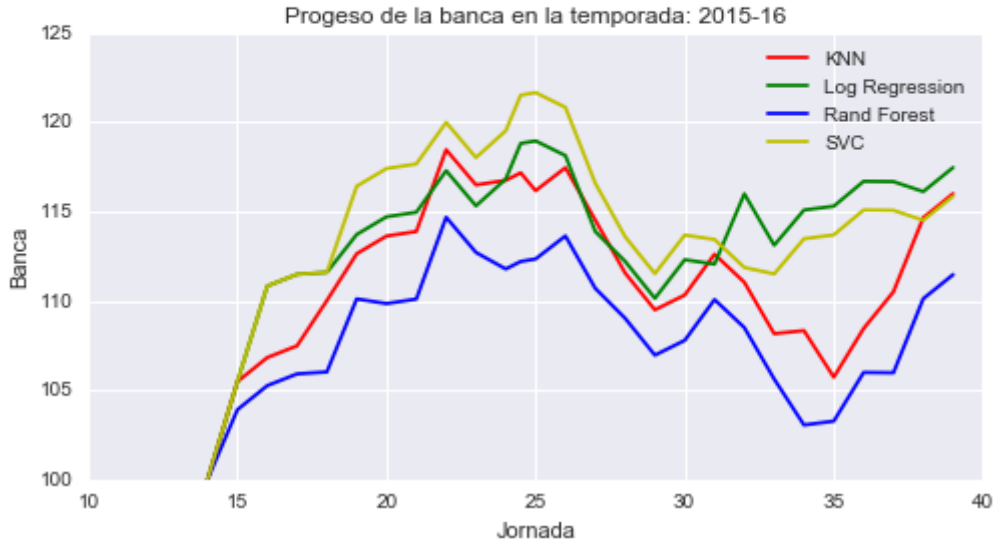

Figura 12. Tasas de acierto y evolución de las bancas de cada algoritmo correspondientes a la temporada 2015-16. Mercado 1X-2.

6 Conclusiones y trabajo futuro

6.1 Conclusiones

- La elección de los atributos iniciales es la parte más importante del proceso.
- Al tratarse de atributos inventados y/o subjetivos, generarlos es la parte que más tiempo requiere.
- Es posible ganar a las casas de apuestas a largo plazo, incluso contando con que a veces se dan temporadas muy impredecibles.
- El algoritmo que mejor ha funcionado, debido posiblemente a su fácil adaptación a las distintas temporadas, ha resultado ser k-NN, pese a tratarse de uno de los más sencillos.
- Como puede observarse en las pruebas realizadas, el problema de predecir resultados de fútbol tiene la particularidad de que los algoritmos que más tasa de acierto tienen no son necesariamente los que más beneficio retornan. Es el caso de regresión logística, que en muy pocas ocasiones logró mantenerse por encima del saldo inicial y, sin embargo, siempre tiene una tasa de aciertos bastante aceptable y similar al de los demás.

6.2 Trabajo futuro

Siempre es conveniente seguir con la labor de transformación de los datos para optimizar el proceso de aprendizaje. A continuación, recojo algunos puntos a tener en cuenta sobre este y otros aspectos mejorables:

- Síntesis: el número de días transcurridos entre el partido actual y el siguiente, podrían fusionarse por separado con la importancia del partido siguiente, dando como resultado dos nuevos atributos.
- Síntesis: las medias de goles recibidos y marcados se podrían expresar como la diferencia de goles media, obteniendo así un nuevo atributo.
- Distinción entre la media de goles marcados o encajados según se juegue como local o como visitante.
- Explotación de nuevos mercados: podrían adaptarse los modelos para tratar de predecir otro tipo de mercados que quizás resulten ser más lucrativos, como, por ejemplo, los que especulan con los saques de esquina.
- Adaptación a otras competiciones y deportes.

Referencias

[1] *Las apuestas deportivas dominan el sector del juego en España*, Expansión, 12/06/2017

[2] Bet365, www.bet365.es

[3] Bwin, www.bwin.es

[4] William Hill, www.williamhill.es

[5] Betfair, www.betfair.es

[6] Laura G. del Valle, *Cerco a las apuestas deportivas*, La Voz de Galicia, 14/06/2017.

[7] Jesús Miguel Gómez-Roso Jareño, TFG sobre "Origen de las apuestas deportivas", facultad de derecho y ciencias de la Universidad de Castilla-La Mancha, 14/07/2014.

[8] Stephen R Clarke. *Adjusting true odds to allow for vigorish*. Swinburne University of Technology.

[9] Daniel Martín Domínguez, *Análisis de resultados deportivos y estimación implícita de probabilidades: fútbol*, TFG Universidad Carlos III de Madrid.

[10] Íñigo Domínguez, *50 partidos bajo sospecha de amaño en Segunda B y Tercera*, El País, 25/05/2017. http://deportes.elpais.com/deportes/2017/05/24/actualidad/1495646908_304724.html

[11] Daniel Martín Domínguez, *Análisis de resultados deportivos y estimación implícita de probabilidades: fútbol*, Conclusiones, TFG Universidad Carlos III de Madrid.

[12] A Joseph, NE Fenton, M Neil. *Predicting football results using Bayesian nets and other machine learning techniques*, Knowledge-Based Systems, 2006.

[13] Fernando Valera Guardiola, *Sistema de predicción de resultados en eventos deportivos y su aplicación en las apuestas*, Conclusiones, PFC Universidad Carlos III de Madrid, 2013.

[14] Wikipedia, Aprendizaje automático, https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico

[15] T. Cover, P. Heart, *Nearest neighbour pattern classification*, IEEE Transactions on Information Theory, 13 (1967), pp. 21-27

[16] LE Peterson, *K-nearest neighbor*, Scholarpedia.org, 2009.

[17] Breiman, *L. Mach Learn*, 1996, 24: 123.

[18] Jhon D. Kelleher, Brian Mac Namee, Aoife D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics*, 2015.

[19] Bernhard Scholkopf, Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond,* 2001.

[20] Richard O. Duda, David G. Stork, Peter E. Hart, *Pattern Classification.*

[21] H Chen, RHL Chiang, VC Storey, *Bussiness intelligence and analytics: From big data to big impact*, MIS quarterly, 2012.

[22] Wikipedia, Kaggle, https://en.wikipedia.org/wiki/Kaggle

[23] Ivan Firdausi, Charles Lim, Alva Erwin. *Analysis of Machine Learning Techniques Used in Behavior-Based Malware Detection*. Advantages in computing, control and telecommunication technologies, Second International Conference. Diciembre de 2010.

[24] Liu, B. *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer-Verlag (2009) .

[25] Estadísticas históricas de la Liga Española, http://www.laliga.es/estadisticas-historicas

[26] Gary Bradsky, Adrian Kaehler, *Learning OpenCV: Computer Vision with OpenCV Library*, 2008.

[27] Kaggle, www.kaggle.com

[28] MATLAB, https://es.mathworks.com/products/matlab.html

[29] Wikipedia, *R (lenguaje de programación)*, https://es.wikipedia.org/wiki/R_(lenguaje_de_programaci%C3%B3n)

[30] WEKA, http://www.cs.waikato.ac.nz/ml/weka/

[31] *Common Format and MIME Type for Comma-Separated Values (CSV) Files*, RFC 4180, 2005.