TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Machine Learning aplicado a las cuotas en las apuestas deportivas

Autor: Pablo Bustillo Rodríguez

Director: Josep Anton Sànchez Espigares

Departamento: Estadística e Investigación Operativa

Convocatoria: Junio 2017-2018

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resumen

En este trabajo, se implementarán diferentes técnicas de Aprendizaje automático (Machine Learning) y se aplicarán en las apuestas deportivas. En concreto, el caso que se va a estudiar, consiste en predecir el resultado final de diferentes partidos de fútbol de la liga Italiana. Para ello, se contará con las cuotas de 7 casas de apuestas diferentes para las 5 últimas temporadas.

Entre los métodos usados, se encuentra el de KNN, SVM, QDA y Árboles de decisión entre otros. El objetivo principal, consistirá, en determinar que combinación de clasificador y casa de apuestas, produce un beneficio más grande. Para ello, se realizará un estudio de validación cruzada y se ajustarán los parámetros de cada clasificador, para obtener el mejor resultado.

Palabras claves: Aprendizaje automático, apuestas, validación cruzada, predicción, clasificación, SVM, Arboles de decisión y precisión.

Abstract

In this work, different techniques of Machine Learning will be implemented and applied in sports betting. In particular, the case that will be studied, is to predict the final result of different football matches in the Italian league. To do this, we will have the odds of 7 different betting houses for the last 5 seasons.

Among the methods used, is that of KNN, SVM, QDA and Decision Trees among others. The main objective, will consist, in determining which combination of classifier and betting house, produces a bigger benefit. For this, a cross-validation study will be carried out and the parameters of each classifier will be adjusted to obtain the best result.

Keywords : Machine Learning, bets, crossvalidation, prediction, clasification, SVM, Decision Trees and precision

Clasification: 62-04 Explicit machine computation and programs (not the theory of computation or programming); 62M20 Prediction [See also 60G25]; filtering [See also 60G35, 93E10, 93E11]; 62P05 Applications to actuarial sciences and financial mathematics

ÍNDICE

Tabla de Ilustraciones

Tabla de datos

Glosario:

1.- Cuotas: Cantidades que las casas de apuestas pagan por los diferentes pronósticos de los eventos que ofrecen en su programa de deportes.

2.- Overround : Margen de beneficio de la casa de apuestas

3.- Tipster**:** Pronosticador que realiza análisis de distintos eventos deportivos

4.- Value: Error en la estimación de las Casas de apuestas respecto a la probabilidad de que algo determinado ocurra en un evento deportivo, como la victoria de un equipo de fútbol.

5.- Surebets: Apuestas con ganancias seguras

6.- Bankroll: Dinero que se ha decidido destinar para apostar

8.-Stake: Nivel de confianza de que una apuesta sea ganadora. Esta confianza definirá la cantidad de dinero a apostar

1.- Presentación y Objetivos

Este trabajo, se centrará en la implementación de 5 técnicas de Aprendizaje Automático (Machine Learning) en el estudio de las apuestas deportivas.

El caso que se va a estudiar, consiste en predecir el resultado final de diferentes partidos de fútbol de la liga Italiana. Es decir, si ganará el equipo local, habrá empate o ganará el equipo visitante. Para ello, se contará con las cuotas de 7 casas de apuestas diferentes para las 5 últimas temporadas.

La motivación principal de este trabajo es poder aprender más sobre las técnicas de Machine Learning, como es el caso de KNN, SVM, QDA, CART y Naive Bayes. Además, se quiere apreciar, si realmente estas técnicas pueden llegar a ser efectivas en el mundo de las apuestas deportivas. Se ha de tener en cuenta, que el sector de las casas de apuestas, controla diferentes factores que pueden ocurrir, para maximizar su beneficio y no dejar posibilidad de entrar en perdidas.

La predicción del resultado final, es difícil, debido a la gran cantidad de factores aleatorios y no aleatorios que ocurren antes y durante el partido, y que influyen en el resultado final. Es por ello, que a priori, no se esperan predicciones de acierto altas, pero si se busca que el retorno final(€) sea positivo.

Una de las principales curiosidades que se conocen antes de realizar el trabajo, es que una persona con conocimientos de fútbol pero no considerada experta es capaz de tener un acierto cercano al 40%, y si se predice siempre el resultado como victoria del local, se tiene a la larga un acierto superior al 45%. Por lo tanto, uno de los objetivos en este trabajo será el de encontrar un algoritmo o técnica de aprendizaje adecuado, que pueda alcanzar tasas de acierto superiores al 50% como mínimo.

El objetivo principal, será poder observar, que combinación de clasificador y casa de apuestas, produce un beneficio más grande. Para ello, se realizaran, las diferentes técnicas nombradas, cuidando el procedimiento de cada una. Se proporcionará como resultados: las matrices de confusión y precisión, el retorno neto(€) según el clasificador utilizado para cada una de las casas de apuestas, diferentes gráficas de evolución del beneficio/pérdida y técnicas de mejora para maximizar los beneficios entre otros.

2.- Conceptos sobre las apuestas deportivas

2.1.- Historia de las apuestas deportivas

La historia de las apuestas deportivas no es tan reciente como pensamos, y para nada es una novedad en nuestra sociedad. Su origen data de hace muchos siglos, para ser más precisos, de la antigua Grecia, donde sus ciudadanos comenzaron a realizar apuestas al ganador de diferentes deportes mientras los animaban.

Ilustración 1 Carrera de Caballos

De la antigua Grecia las apuestas dieron el salto y aterrizaron en el Imperio Romano, donde se hicieron muy famosas en los Coliseos con los combates de Gladiadores y las carreras de carros.

Con el paso de los años su afición comenzó a ser mayor y en la Edad Media, las apuestas consiguieron hacer famosos a deportes que estaban apareciendo en aquella época, como eran los torneos de caballeros y el tiro al arco.

Las apuestas eran un método de pasatiempo y disfrute a la vez de los deportes favoritos en distintas épocas, pero todo cambio a partir del Siglo XVII, principalmente en el 1780, donde Inglaterra fue el país donde se estableció la gran revolución de las apuestas deportivas. Empezaron a abrirse lugares destinados a hacer apuestas y sobre todo había un deporte rey en estas apuestas, las carreras de caballos.

En el Siglo XIX llegaron definitivamente a América, donde fueron acogidas con gran expectación y luego más tarde, sobre el 1930, comenzó la gran difusión en los periódicos de páginas especializadas en apuestas deportivas.

En esa época las apuestas adquirieron una gran importancia y se extendieron rápidamente por todos los países de Europa y Norteamérica, convirtiéndose en un acontecimiento que atraía a las masas.

La última gran revolución es la que ha protagonizado Internet en la historia de las apuestas deportivas. Todo comenzó en Canadá y los Estados Unidos, donde se crearon empresas cuya especialidad eran las apuestas deportivas. Consideraron que gracias a la revolución tecnológica que había ya en nuestra sociedad, lo mejor para todas las personas acostumbradas a apostar, sería el facilitarles todo y que pudieran así apostar en línea.

2.2.- Situación actual

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. En concreto, las apuestas deportivas online son la modalidad que más crece dentro de los juegos de esta industria, con más de 2 millones de personas registradas a día de hoy.

Existen numerosas webs dedicadas a este negocio, como Bet365 , Bwin , William Hill o Betfair, entre otras.

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 ocupar el 46% del espacio publicitario durante la retransmisión de este tipo de eventos.

El siguiente gráfico muestra como las apuestas deportivas online se han cuadruplicado desde 2012. En Junio de 2011, se movieron 100 millones de euros y en Diciembre de 2016 unos 400 millones de euros.

Ilustración 2 Cantidades jugadas por mes

2.3.- ¿Qué son las cuotas?

Cuando se habla de cuota en las apuestas online se hace referencia a las cantidades que las casas de apuestas pagan por los diferentes pronósticos de los eventos que ofrecen en su programa de deportes**.**

Estas cuotas son ofrecidas a partir de las probabilidades estadísticas que existen en cada evento, siendo más pequeñas cuanto mayor posibilidad de éxito.

Cuando una persona acude por primera vez a una casa de apuestas a realizar una apuesta debe tener en claro los siguientes puntos:

La casa de apuestas ofrecerá un abanico de deportes, divididos estos en diferentes eventos correspondientes a cada deporte.

Cada evento dispondrá de una serie de pronósticos que pueden suceder a lo largo del evento ofertado. Cada pronóstico tiene una cuota.

La cuota mínima que se ofrece para un pronóstico es de 1.01, la más baja posible, esta cuota indica que se gana un céntimo por euro apostado. Del mismo modo, una cuota de 2.5, quiere decir que se gana (beneficio neto) un euro y medio por cada euro apostado.

Ilustración 3 Posibilidad y riesgo de ganar según la cuota

A continuación se muestra un sencillo ejemplo:

En el siguiente encuentro, Francia vs Alemania, se muestra la cuota que las casas de apuestas pagan en un momento determinado.

Francia (1.80) - Empate (4.20) - Alemania (3.50)

La victoria de Francia tiene una cuota de 1.80. Si decidimos apostar 2€, multiplicaremos la cuota por la cantidad de dinero, y así averiguamos el retorno o importe de beneficio bruto en el caso de acertar: 1.80 x 2 = 3.60 €. Para determinar el beneficio neto, se ha de restar esta cantidad menos el dinero apostado inicialmente. En este caso : 3.6-2=1.6€

Por otra parte, en el caso de fallar, la pérdida corresponde a la apuesta inicial de 2€.

2.4.- Diferentes Casas de apuestas

En este grupo están las más comunes y tradicionales de Internet, en las que los usuarios apuestan contra la casa y éstas son las que fijan las cuotas. Algunos ejemplos son William Hill, Betfair, Bwin, Bet365, etc…

Ilustración 4 Casas de apuestas

Ofrecen la posibilidad de apostar a eventos de múltiples deportes, ofertando apuestas a diferentes sucesos en cada evento, como quien marca el primer gol, resultado exacto, resultados con hándicap o simplemente, qué equipo acabará ganando el partido. En la mayoría de eventos se puede apostar días e incluso semanas antes de que den comienzo, lo que puede suponer una gran ventaja de cara al apostante si posee suficiente información para obtener una buena cuota. Como gran desventaja tiene que a veces limitan la cantidad invertida en algunos eventos.

Como se ha comentado en la página anterior, el funcionamiento de estas cuotas es de la siguiente manera:

Beneficios: (Cuota Ganadora X Cantidad Apostada) – Cantidad Apostada

Pérdidas: Cantidad Apostada

En la siguiente tabla se muestra un ejemplo de las cuotas de dos casas de apuestas de un encuentro de la Primera División Española antes del inicio del partido:

Tabla 1 Cuotas según la casa de apuestas

Jornada 29	Bet365	William Hill
Zaragoza	7,50	7,50
Empate	4,50	4,20
Real Madrid	1,40	1,44

El resultado final del encuentro fue 1-1. Si se juega un euro a cada apuesta de cada casa, los resultados hubiesen sido los siguientes:

Beneficios (Bet365): (4,50 X 1) – 1 (Apostado a favor del empate) = 3,50 €

Pérdidas (Bet365): 1 (Apostado a favor del Zaragoza) + 1 (Apostado a favor del Real Madrid) = 2 €

Beneficios (William Hill): (4,20 X 1) – 1 (Apostado a favor del empate) = 3,20 €

Pérdidas (William Hill): 1 (Apostado a favor del Zaragoza) + 1 (Apostado al Madrid) = 2 €

2.5.- Overround

A la hora de calcular la probabilidad de resultado final de un encuentro, la suma de probabilidades de victoria local, empate y victoria visitante debería ser del 100%, sin embargo las casas de apuestas, al tener unos amplios márgenes de beneficios, la suma de estas, está por encima del 100%, ya que las cuotas son inferiores a lo pronosticado. La diferencia se denomina Overround y es el margen de beneficio de la casa de apuestas.

En el siguiente ejemplo se muestra las probabilidades estimadas por la casa de apuestas William Hill antes de un encuentro:

Tabla 2 Cuotas para el partido Osasuna vs Levante

Osasuna	Empate	Levante
2.03	3.32	2.81

Si no se normalizan las cuotas, las probabilidades de ganar que estima la casa de apuestas son:

Probabilidad del Local: **PL** = 1 / Cuota del Local *100= 49.26%
Probabilidad de Empate: **PE** = 1 / Cuota de Empate *100= 30.12 %
Probabilidad del Visitante: **PV** = 1 / Cuota del Visitante *100= 26.25%
Probabilidad Total: **PT** = PL +PE + PV =105.63%

Overround= 105.63 – 100 = 5.63% (Margen de beneficios)

Esto también puede ser representado mediante el Payout, que es la inversa del Overround y se define como la media de dinero que van a recibir todos los usuarios por cantidad invertida.

Para saber las probabilidades reales de las casas de apuestas se tiene que hacer una normalización de éstas. Se define la siguiente matriz.

$$A = \begin{pmatrix} Cuota\ Local & 0 & 0 & -1 & 0 \\ 0 & Cuota\ Empate & 0 & -1 & 0 \\ 0 & 0 & Cuota\ Visitante & -1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Se tiene un problema inverso, por lo que se calcula la matriz reducida de la matriz de cuotas, mediante el comando rref en Matlab.

$$B = rref(A) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0.4664 \\ 0 & 1 & 0 & 0 & 0.2852 \\ 0 & 0 & 1 & 0 & 0.2485 \\ 0 & 0 & 0 & 1 & 0.9467 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & ProbNorm\,L \\ 0 & 1 & 0 & 0 & ProbNorm\,E \\ 0 & 0 & 1 & 0 & ProbNorm\,V \\ 0 & 0 & 0 & 1 & Q \end{pmatrix}$$

- **Probabilidad Local :** **PL**= Q / Cuota Local * 100 = 46.64%
- **Probabilidad Empate :** **PE**= Q / Cuota Empate * 100 = 28.52%
- **Probabilidad Visitante:** **PV**=Q / Cuota Visitante * 100 = 24.85
- **Probabilidad Total:** **PT** = PL +PE + PV =100%

Si el número de apuestas se repartiera de manera equitativa, es decir, que el 46,64% del dinero invertidos fuera a favor del equipo local, el 28,52% a favor del empate y el 24,85% a favor del visitante, la casa de apuestas obtendría un beneficio del 5,63%, por lo que, si la cantidad total jugada al evento fuera de 10000 €, la casa de apuestas se llevaría 563€.

En este caso, las cuotas justas de apuestas serían:

$$Cuota\ justa\ Local = \frac{1}{46{,}64} * 100 = 2{,}14€ > 2{,}03€$$
$$Cuota\ justa\ Empate = \frac{1}{28{,}52} * 100 = 3{,}51€ > 3{,}32€$$
$$Cuota\ justa\ Visitante = \frac{1}{24{,}85} * 100 = 4{,}02€ > 3{,}81€$$

2.6.- Bonos de bienvenida

La competencia de las casas de apuestas por captar nuevos clientes es muy alta. Para ello, ofrecen bonos de bienvenida a los usuarios que hacen más atractiva su oferta. Estos bonos suelen consistir en una apuesta gratuita al usuario, que en el caso de fallar pueda recuperar su dinero ó regalar un porcentaje de la cantidad invertida como dinero gratuito para apostar.

Las casas de apuestas estudian muy bien estas ofertas ya que a largo plazo estas pérdidas iniciales se convierten en beneficios.

William Hill

William Hill ofrece al apostante el regalo de un bono del 100% de lo que ingrese, mediante un depósito mínimo de 10 € y bono gratuito de hasta 100 €. Las condiciones de retirada de dinero se permite cuando el usuario apuesta al menos 4 veces la cantidad del bono y del depósito inicial a cuotas superiores a 1.50 €, en el que sólo cuenta la primera apuesta que se realiza para el mercado de un evento. Con esto evitan que la gente apueste a cosas en las que haya un 50% de probabilidad, las cuotas estén aproximadamente a 1.90 € y hagan la apuesta a favor y en contra, perdiendo una mínima cantidad de dinero y aprovechándose de la rentabilidad del bono ofrecido.

Bwin

El bono de bienvenida de Bwin es muy similar al de William Hill. Se ofrece un bono de igual cantidad ingresada hasta 50 €, con la restricción de que el bono se tiene que gastar en cuotas superiores a 1.70€ para poder hacer la retirada del dinero. A pesar de que obligan a apostar a cuotas más altas, la cantidad a arriegar es mucho menor que con William Hill

Ilustración 5 Bonos de bienvenida

2.7.- Tipos de apuestas

Se encuentran las siguientes: Sencillas, combinadas, de sistema, apuestas en directo o en vivo (Live) y móviles.

Apuestas simples: Este tipo de apuestas consiste en elegir un único pronóstico de un evento (ejemplo: gana el Milan). Esa cuota se multiplicara por la cantidad que deseemos apostar, y el resultado de dicha multiplicación nos dará las ganancias finales. Para cobrar esas ganancias el pronóstico deberá ser acertado.

Apuestas combinadas: En este caso seleccionamos de los eventos ofertados más de un evento y de la misma forma elegimos un pronóstico de cada. Con cada cuota de los pronósticos seleccionados, las multiplicamos y obtendremos una cuota final, que habrá que multiplicar por la cantidad de dinero que deseemos apostar. En este tipo de apuestas tenemos que acertar todos los pronósticos por los que hemos apostado, porque en el caso de fallar uno, la apuesta será perdedora en su totalidad.

Apuestas de sistema: Este tipo de apuestas son muy parecidas a las combinadas, pero con la diferencia de que si uno de los pronósticos que hemos seleccionado se falla, disponemos de la posibilidad de recuperar parte de la apuesta. En estas apuestas se utiliza mucho la probabilidad y estadística, combinando entre si los pronósticos seleccionados para poder obtener ganancias o no.

Apuestas Live: Esta modalidad de apuestas son las consideradas apuestas en tiempo real, en directo o en vivo. Las apuestas Live nos ofrecen la posibilidad de apostar mientras se está celebrando un evento, donde irán variando los pronósticos dependiendo de los que vaya aconteciendo en el mismo.

Apuesta móvil: Las casas de apuestas han puesto al servicio de sus clientes, diferentes aplicaciones o Apps para poder acceder a su casa de apuestas, y desde esa plataforma móvil poder realizar sus apuestas sin tener que estar en un lugar físico.

2.8.- Los tipsters deportivos

Un tipster es un pronosticador que realiza análisis de distintos eventos deportivos y localiza cuotas con valor en ellos. Generalmente, suele ser un apostante ganador a largo plazo, cuyas estadísticas son su aval de cara al público.

Su "modelo de negocio" es doble, ya que por una parte gana dinero haciendo las apuestas con valor que descubre, y por otra, consigue ingresos extra compartiendo sus consejos con suscriptores, que normalmente han realizado un pago previo para poder tener acceso a sus consejos.

Función del tipster

Descubrir cuotas con valor en eventos deportivos, hacer las apuestas y compartir la información de inmediato con sus suscriptores o su público objetivo**.** La rentabilidad y la rapidez del tipster son dos características muy importantes a tener en cuenta. Seguir a un tipster ganador puede ser una gran estrategia para ganar dinero con las apuestas, siempre que se haga lo que dice y en el mismo momento en que publica un pronóstico.

Que deportes suelen pronosticar?

Los tipsters suelen centrarse en deportes como el fútbol, basket, tenis y béisbol entre otros. Son los deportes que más apuestas reciben y también los más conocidos por los apostantes, gracias, en parte, al gran seguimiento de que son objeto por parte de los distintos canales de televisión..

Donde encontrar pronósticos?

Los tipsters suelen hacer llegar sus pronósticos de pago o gratuitos por medio de distintas vías. Las webs y los blogs de apuestas, y las redes sociales (Facebook y Twitter) son los canales habituales de los tipsters gratuitos. Los de pago suelen utilizar vías de acceso restringido, como el e-mail, el whatsapp, Telegram o las páginas interiores de las webs

Tipsters reconocidos

Fútbol : Juan Gayá Salom

Básquet: Mallic Sergio

Tenis: TipsterApuesta

Ilustración 6 Juan Gayá, el tipster más seguido de Europa

2.9.- El Value

El value es un parámetro totalmente subjetivo, cada persona puede interpretar que una apuesta tiene o no tiene *value* en función de sus conocimientos y análisis previo. Incluso para algunos apostadores una apuesta concreta puede tener mucho o poco *valor*.

Se puede definir el *value* como un error en la estimación de las Casas de apuestas respecto a la probabilidad de que algo determinado ocurra en un evento deportivo, como la victoria de un equipo de fútbol.

Para entender este concepto tan importante, hay que dejar claro que las cuotas de las apuestas son la inversa de la probabilidad de que ocurra dicha apuesta. A continuación se explicará el modo de actuar de las casas de apuestas a la hora de poner las cuotas de las distintas competiciones deportivas.

Por ejemplo, en un partido de tenis entre Nadal y Federer, el oddsmaker (trabaja para la casa de apuestas) analiza el partido desde todos los puntos de vista posibles, estadísticas, informaciones, etc. y llega a la conclusión de que la probabilidad de que Nadal gane ese partido es del 50 %.

Ilustración 7 Cuotas de las casas vs cuotas personales

Por tanto, la cuota justa para la victoria de Nadal sería la inversa de dicha probabilidad 1/0.5 = 2.00. Y la cuota de Federer al ser su probabilidad también del 50 % sería la misma, 2.00. La pregunta es: ¿Por qué nunca en las casas de apuestas se ven cuotas 2.00 - 2.00? Esto es debido a que las casas de apuestas aplican un margen a cada cuota y es donde generan sus beneficios, de ahí que las cuotas en partidos al 50% sean del tipo 1.83 - 1.83.

Volviendo al tema del *value* y siguiendo con el ejemplo de ese partido de tenis. Un apostador analiza el partido desde su perspectiva y llega a la conclusión de que la probabilidad que él estima de que Nadal consiga la victoria es del 60 %. Así, la cuota justa para esa probabilidad sería 1/0.6 = 1.66. El apostador va a su casa de apuestas y comprueba como la cuota a favor de la victoria de Nadal es 1.83 . Para ese apostador esa cuota está sobrevalorada y, por tanto, para él esa apuesta tiene *value*, ya que la cuota correcta debería ser 1.66 y la que hay es 1.83. Cuando esto ocurre, se detecta el value en algún mercado, se debe apostar, pues la casa de apuestas está ofreciendo más de lo que debería ser.

2.10.- Surebets

Las apuestas seguras o las Surebet es una manera de apostar en la cual se elige un evento deportivo y se apuesta a todas las opciones que se ofrecen. De esta manera, si se juega por los resultados ofrecidos por un evento, se acertaran de manera segura y se obtendrán beneficios. Pero para que este procedimiento funcione, se tendrá que buscar el evento más interesante, siendo el tenis uno de los deportes con más repercusión en el Surebet.

Lo primero que se tendrá que hacer ante un partido de tenis, es comprobar que por ambos jugadores se paga una cuota superior a 2.00**,** para que apostando la misma cantidad a los dos jugadores, la victoria de cualquiera de ellos proporcione beneficios.

Lo más efectivo en este tipo de apuestas es elegir las cuotas de distintas casas de apuestas online, ya que cada una variará en función del evento, y puede que su valorización sea diferente al resto. En este caso hay que fijarse bien en las cuotas y ser rápido en la toma de decisión. Hay que tener en cuenta que las casas de apuestas detectan estas surebets y rectifican lo más rápido posible.

El problema que presentan las apuestas seguras, surebet, es que se dedica mucho tiempo y mucho estudio de eventos y cuotas para obtener un beneficio muy reducido. Y por lo general las apuestas seguras son muy complicadas de encontrar, por lo que es una modalidad de apuestas que no a todo llega a convencer, aunque da beneficios seguros, estos también son muy reducidos.

Ejemplo:

Tabla 3 Cuotas para el partido de Nadal vs Federer

Casas de apuestas/Ganador	Nadal	Federer
Bet365	2.12	1.75
William Hill	1.8	2.07

En este caso, si se apostara en Bet 365 un euro a que gana Nadal y un euro en William Hill a que gana Federer, se obtendrá un beneficio seguro gane quien gane. Para saber el beneficio neto, se restara la apuesta total, en este caso 2€.

- Si gana Nadal: 2.12 – 2 = 0,12€
- Si gana Federer: 2.07 - 2=0.07€

Claramente el beneficio es muy pequeño, pero si se apuestan cantidades muy grandes, el beneficio es significativo.

2.11.- Gestión del Bankroll

El término **"Bankroll"** es un término muy utilizado en el póker, pero también se puede usar para hablar de apuestas deportivas.

El "Bankroll" hace referencia al dinero que se ha decidido destinar para apostar.

Apostar gestionando el Bankroll se refiere a tener en cuenta en cada momento su estado, para hacer apuestas, ya que el objetivo es gestionar de forma óptima el bankroll para obtener beneficios y evitar una quiebra no deseada.

Hacer apuestas online con una buena gestión, se podría resumir en "apostar con cabeza". Al tener un control del dinero que se ingresó inicialmente y la evolución de este ingreso, se debe analizar cuánto dinero se puede apostar en una apuesta sin que el Bankroll sufra un gran golpe en caso de fallo. El factor riesgo de una apuesta es vital para decidir qué porcentaje del dinero se pondrá en juego.

Ilustración 8 Gestión del Bankroll según los resultados e inversión deportiva

Si el Bankroll está pasando por un mal momento, lo aconsejable es apostar poco a poco a apuestas con grandes probabilidades de acierto, es decir, ir apostando cautelosamente hasta que el Bankroll vaya recuperándose. Si el Bankroll está pasando por un buen momento, se puede arriesgar un poco más en las apuestas pero siempre dentro de la lógica.

Las apuestas online cada vez se están haciendo más populares y cada vez son más los que hacen apuestas deportivas, sin embargo, no es tan popular llevar un control del dinero destinado para apostar, y eso es un error.

2.12.- Que es el stake?

Uno de los términos más utilizados dentro del ámbito de las apuestas deportivas es el **Stake**. Este término se refiere al nivel de confianza de que una apuesta sea ganadora, esta confianza definirá la cantidad de dinero a apostar.

La confianza la podemos evaluar según una escala del 1 al 10, siendo el 10/10 una confianza absoluta en la apuesta, y si hablamos de una confianza de 1/10, estaremos hablando de una confianza mínima.

El grado de confianza, stake, y el porcentaje de bankroll a apostar estan íntimamente relacionados. Se recomienda apostar el 10% del Bankroll en eventos cuya confianza sea de 10/10, y el 5% en apuestas con un stake de 5/10. Para apuestas con un stake inferior a 5 en muchas ocasiones es mejor no apostar.

Antes de apostar, se deberá asignar un Stake a la apuesta. Se podrían diferenciar entre distintos niveles de Stake:

Tabla 4 Niveles de Stake

Stake	Rangos
1	Cuotas ≥ 7
2	4.5 < Cuotas < 7
3	3< Cuotas ≤ 4.5
4	2.5< Cuotas ≤ 3
5	2 ≤ Cuotas ≤ 2.5
6	1.8 < Cuotas < 2
7	1.5 < Cuotas ≤ 1.8
8	1.2 < Cuotas ≤ 1.5
9	1.1 < Cuotas ≤ 1.2
10	Cuotas ≤ 1.1

Stakes que van del 1 al 2: Son aquellas apuestas que vemos difíciles de ganar pero que a la vez tienen unas cuotas muy sobrevaloradas.

Stakes que van del 3 al 4: Son apuestas que nos generan algunas dudas por el momento en que se disputan.

Stakes que van del 5 al 6: Este tipo de Stake suelen ser el grado de confianza en el que la gente suele moverse cuando apuesta.

Stakes que van del 7 al 8: Son oportunidades que suelen aparecer de vez en cuando y pueden generar grandes cantidades de dinero. Suelen darse en partidos de fútbol donde uno de los equipos llega en un estado de forma muy bueno.

Stakes que van del 9 al 10: Son en las apuestas en eventos donde la victoria de un equipo sobre otro es clarísima debido a la superioridad de un equipo respecto del otro.

Por lo tanto, el Stake es una de las cosas a tener en cuenta a la hora de empezar a apostar para poder llevar una buena administración del Bankroll y apostar el porcentaje de dinero correcto en relación a la confianza de la apuesta

2.13.- Internet y las apuestas

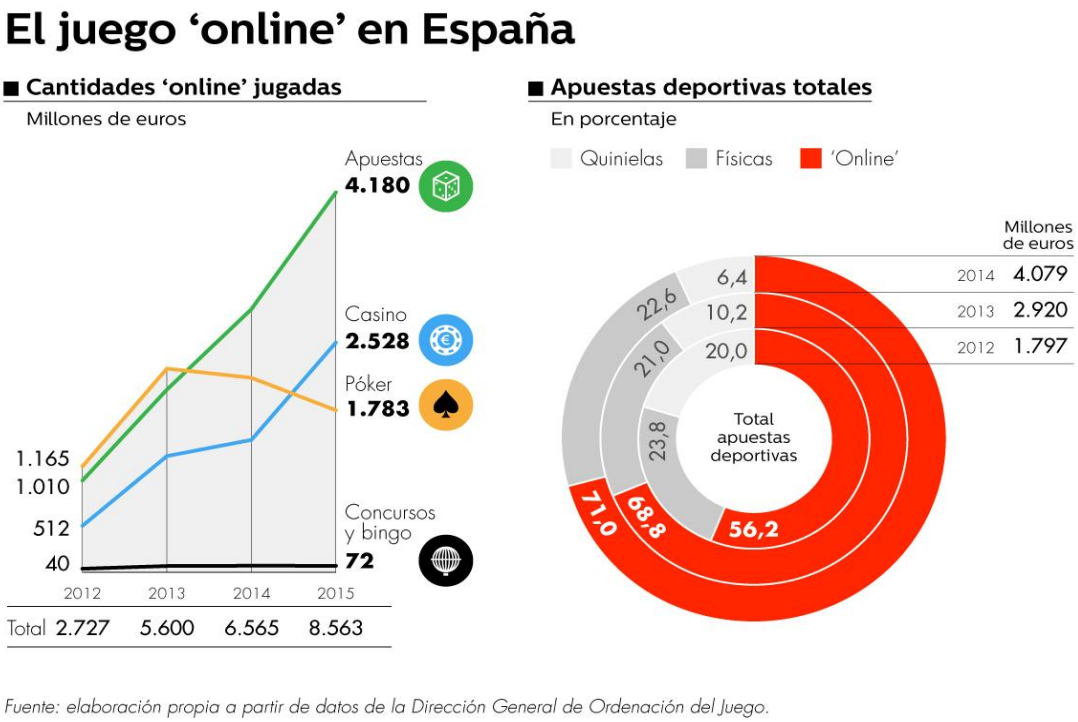
Ante tanta publicidad que encontramos en los medios de comunicación, como patrocinio en diferentes equipos deportivos y una masiva afición al deporte, las casas de apuestas deportivas están creciendo mucho, y no es para menos.

Internet se ha convertido en un medio totalmente revolucionario desde su aparición, ya que ha supuesto una de las revoluciones más notables en la historia de la comunicación y de la información, dado que con muy poco, se puede acceder a todo tipo de información y a una velocidad cada vez más alta.

De esto se están beneficiando las empresas dedicadas a las apuestas deportivas, ya que mediante la creación de plataformas muy avanzadas, ofrecen a todos los usuarios la posibilidad de realizar todo tipo de apuestas de una manera muy cómoda a través de Internet, en tiempo real y con diferentes sistemas de pago muy beneficiosos y prácticos para los usuarios.

Los mejores equipos del mundo están accediendo a que sus patrocinadores oficiales sean casas de apuestas. ¿A qué se debe?. Posiblemente sea porque los equipos de fútbol principalmente vean lo que el resto de la sociedad lleva percibiendo hace mucho tiempo. Las casas de apuestas por Internet se están haciendo un hueco muy importante en el mercado.

Equipos como el Manchester United tienen firmados acuerdos con casas de apuestas (en este caso Bwin) para los tres próximos años.



Fuente: elaboración propia a partir de datos de la Dirección General de Ordenación del Juego.

Ilustración 9 Evolución de las apuestas y cantidades jugadas en internet

2.14.- Estafas deportivas 

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.

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

Italia siempre ha sido objeto de los mayores escándalos de fútbol de todos los tiempos. A continuación se muestran 2 ejemplos

Totonero

El primer caso fue conocido como "**Totonero**" y apareció en la década de los '80. Este alboroto de la "lotería negra" en italiano, reveló que varios jugadores habían colocado *apuestas sobre los resultados de sus propios equipos*, una ola de sanciones golpeó a clubes populares como el AC Milan y la Lazio de Roma que descendieron de la Serie A a la Serie B (segunda división italiana) y el delantero estrella de la equipo nacional italiano, Paolo Rossi, fue suspendido por 2 años.


TOTOCALCIO

La Gazzetta dello Sport

CHE TRISTEZZA: MANETTE NEGLI STADI!

12 ARRESTI

In carcere il presidente del Milan Colombo e 11 calciatori coinvolti nelle scommesse

ECCO CHI SONO

Ilustración 10 Diario la Gazzeta

Calciopoli

Sin embargo, hasta ahora el mayor escándalo de partidos amañados en el fútbol italiano ocurrió en el 2006, fue conocido como "**Calciopoli**" y consistió en fraudes arbitrales en los partidos de la Serie A del fútbol italiano. Equipos como Juventus FC, AC Milan, Fiorentina y Lazio fueron acusados de *sobornar a los árbitros para controlar los resultados* de los partidos de este popular campeonato italiano. Los principales responsables de este fraude, Luciano Moggi y Antonio Giraudo, pertenecían a la Juventus FC, lo que trajo como consecuencia que el equipo de Turín descendiera a Serie B con 30 puntos de penalización y perdiera sus últimos 2 títulos o "scudetti".


Corriere dello Sport

CALCIOPOLI ECCO LA SENTENZA DI NAPOLI

Condannati!

A Moggi 5 anni e 4 mesi: «Ha creato l'associazione». A Bergamo 3 anni e 8 mesi. Puniti i Della Valle e Lotito

QUESTO CALCIO MALATO

Ilustración 11 Diario Corriere dello Sport

3.- Metodología

3.1.- Machine Learning

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. 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 el proceso KDD (Knowledge Discovery and Data Mining), nos encontramos con la tarea de minería de datos, nombre con el cual nos referimos comúnmente a todo el proceso KDD . La tarea de minería de datos trata el análisis automático, o semi-automático, de grandes conjuntos de datos con el objetivo de obtener a partir de ellos patrones desconocidos, grupos de datos, o dependencias entre ellos .

Esta extracción de conocimiento se consigue mediante técnicas de minería de datos. Con ellos obtenemos modelos de conocimiento a partir de los datos para poder realizar un análisis predictivo. 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.

Normalmente nos encontramos con dos tipos de problemas a resolver en el aprendizaje supervisado, la clasificación y la regresión.

Regresión: En los problemas de regresión se trata de observar cómo reacciona una variable de respuesta (variable dependiente) en función de una variable explicativa (variable independiente).

Clasificación: En los problemas de clasificación se trata de asignar una clase a un elemento entrante, el cual no tiene una categoría asignada.

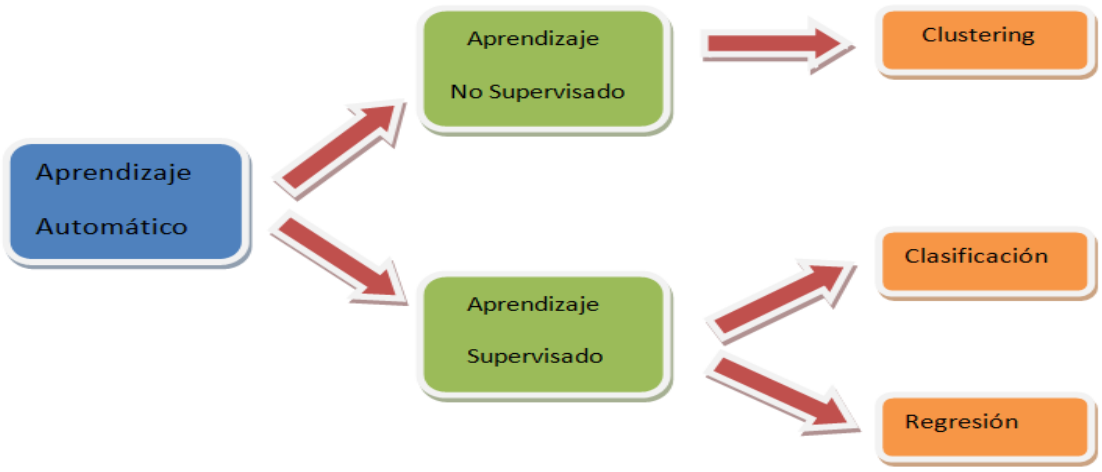


Ilustración 12 Esquema del aprendizaje automático

En este trabajo se utilizaran clasificadores. El término clasificador se utiliza en referencia al algoritmo utilizado para asignar a un elemento entrante no etiquetado, en una categoría concreta conocida. Dicho algoritmo permite ordenar o disponer por clases un conjunto de datos de entrada a partir de cierta información característica de estos.

Una forma de implementar un clasificador es seleccionar un conjunto de datos de entrenamiento y tratar de definir unas ciertas reglas que permitan asignar una clase a cualquier otro dato de entrada. En ocasiones, el término clasificador también es utilizado para referirse a la función matemática que implementa el algoritmo de clasificación.

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.

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 o la predicción de las pérdidas y ganancias de una empresa . A la aplicación de estos métodos en el ámbito empresarial se le conoce como Business Intelligence.

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. En ella se realizan competiciones por crear los mejores modelos predictivos y la comunidad es muy activa y colaboradora.

A continuación, se presentaran los 5 clasificadores que se realizaran en la parte práctica

- Naive Bayes (NB)
- K Nearest Neighbor (KNN)
- Decision Tree (CART)
- Quadratic Discriminant Analysis (QDA)
- Support Vector Machine (SVM)

3.2.- Naive-Bayes

Es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de *ingenuo*. Además, funciona correctamente cuando no se tienen demasiados datos.

En caso de que los atributos contengan valores missings, las entradas de la tabla correspondiente son omitidos por las predicciones.

Implementando en R Naive Bayes, incluirá los siguientes componentes: La distribución a priori de la clase por la variable dependiente, una lista de tablas, una por cada predictor de la variable.

Además, por cada variable categórica se dará una tabla dónde por cada nivel del atributo encontramos la probabilidad condicional dada por la clase objetivo. Por cada variable numérica se presenta la media y la desviación típica.

A continuación se muestra un sencillo ejemplo:

Se quiere clasificar una persona en hombre o mujer basándose en las características de sus medidas: peso, altura y número de pie.

Entrenamiento previo

Tabla 5 Datos de diferentes personas

sexo	altura (pies)	peso (lbs)	número de pie (inches)
hombre	6	180	12
hombre	5.92 (5'11")	190	11
hombre	5.58 (5'7")	170	12
hombre	5.92 (5'11")	165	10
mujer	5	100	6
mujer	5.5 (5'6")	150	8
mujer	5.42 (5'5")	130	7
mujer	5.75 (5'9")	150	9

Se realiza una distribución Gaussiana, se extraen los datos y se obtiene la media y la varianza de cada característica.

Tabla 6 Cálculos de la tabla 5

sexo	media (altura)	varianza (altura)	media (peso)	varianza (peso)	media (foot size)	varianza (foot size)
hombre	5.855	0.035033	176.25	122.92000	11.25	0.91667
mujer	5.4175	0.097225	132.5	558.33000	7.5	1.66670

En este caso se encuentra una distribución equiprobable, es decir que tienen la misma probabilidad. P(hombre)=0.5 y P(mujer)=0.5.

Testing

Ahora se reciben unos datos para ser clasificado como hombre o mujer

Tabla 7 Muestra a predecir

sex	altura (pies)	peso (lbs)	número de pie(inches)
muestra	6	130	8

A continuación se calcula la probabilidad a posteriori de los dos casos, según es hombre o mujer.

$$posteriori(hombre) = \frac{P(hombre)\, p(altura|hombre)\, p(peso|hombre)\, p(numerodepie|hombre)}{Evidencia}$$

$$posteriori(mujer) = \frac{P(mujer)\, p(altura|mujer)\, p(peso|mujer)\, p(numerodepie|mujer)}{Evidencia}$$

$$evidencia = P(hombre)\, p(altura|hombre)\, p(peso|hombre)\, p(numerodepie|hombre) + P(mujer)\, p(altura|mujer)\, p(peso|mujer)\, p(numerodepie|mujer)$$

Realizando los respectivos cálculos se obtiene que la posteriori más grande es el de la mujer, por eso se determina que los datos son de mujer.

3.3.- KNN

El método de los *k* vecinos más cercanos es un método de clasificación supervisada (Aprendizaje, estimación basada en un conjunto de entrenamiento y prototipos) que sirve para estimar la función de densidad $F(x/Cj)$ de las predictoras x por cada clase C_j.

Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase C_j a partir de la información proporcionada por el conjunto de prototipos. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.

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.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p}(x_{ri} - x_{rj})^2}$$ donde p= atributos y x = valores

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

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.

Elección del K:

La mejor elección de k depende fundamentalmente de los datos; generalmente, valores grandes de k reducen el efecto de *ruido* en la clasificación, pero crean límites entre clases parecidas. Un buen k puede ser seleccionado mediante una optimización de uso.

A menor K, más varianza.

A mayor K, más sesgo.

A continuación, se muestra una imagen con un ejemplo:

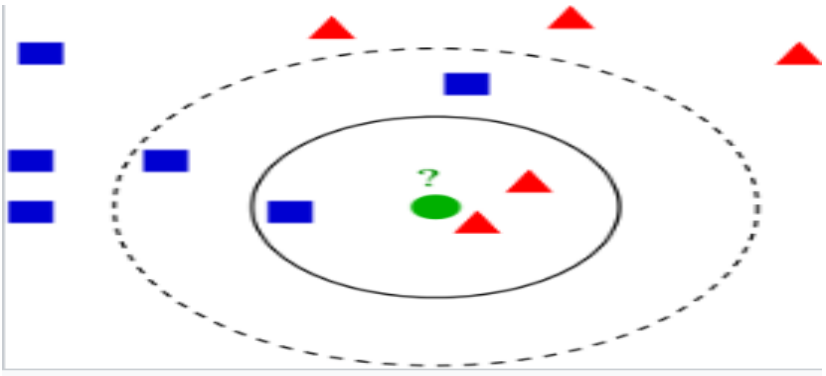


Ilustración 13 Ejemplo del algoritmo de KNN

En esta imagen, se observa un punto verde, que es el individuo que se quiere clasificar.

Para **k =3**, este es clasificado con la clase triángulo, ya que hay solo un cuadrado y 2 triángulos, dentro del círculo que los contiene.

Si **k =5**, este es clasificado con la clase cuadrado, ya que hay 2 triángulos y 3 cuadrados, dentro del círculo externo.

Inconvenientes:

Es necesario destacar que KNN es un aproximador universal, que requiere muestras de training (entrenamiento) grandes. Algunos de los inconvenientes son que hay problemas con las dimensiones elevadas, problemas de clases desequilibradas y efecto agujero negro (peores resultados que otros métodos)

3.4.- Árboles de Decisión (Decision Trees)

Un árbol de decisión es un mapa de los posibles resultados de una serie de elecciones relacionadas. Este clasificador, permite a un individuo u organización evaluar las posibles acciones entre sí en función de sus costos, probabilidades y beneficios. Se pueden usar para generar discusiones informales o para trazar un algoritmo que prediga matemáticamente la mejor opción.

Un árbol de decisión generalmente comienza con un solo nodo, que se bifurca en posibles resultados. Cada uno de esos resultados conduce a nodos adicionales, que se ramifican hacia otras posibilidades. Esto le da una forma arborescente.

Hay tres tipos diferentes de nodos: nodos aleatorios, nodos de decisión y nodos finales. Un nodo aleatorio, representado por un círculo, muestra las probabilidades de ciertos resultados. Un nodo de decisión, representado por un cuadrado, muestra una decisión que debe tomarse, y un nodo final muestra el resultado final de un camino de decisión

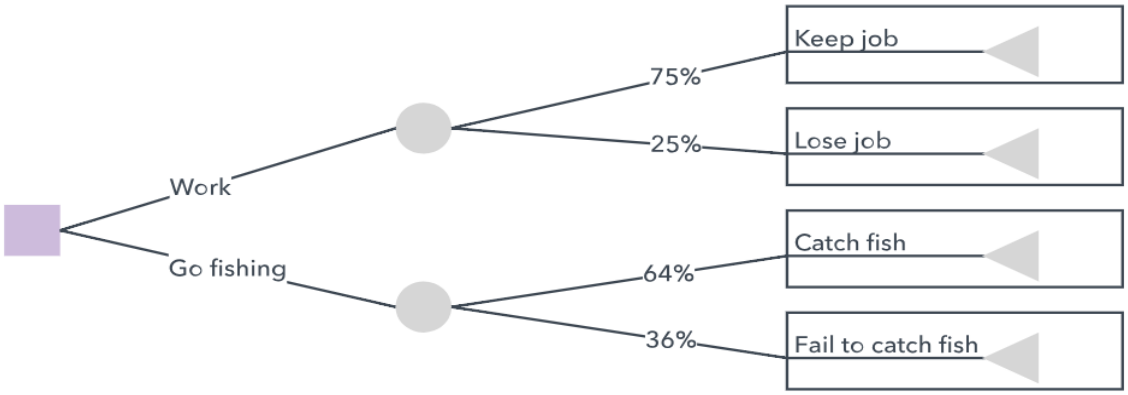


Ilustración 14 Ejemplo del árbol de decisión

Cuando se usa un árbol de decisión con un modelo de probabilidad acompañante, se puede utilizar para calcular la probabilidad condicional de un evento, o la probabilidad de que suceda, dado que ocurre otro evento. Para hacerlo, simplemente se comienza con el evento inicial, y se sigue el camino desde ese evento hasta el evento objetivo, multiplicando la probabilidad de cada uno de esos eventos juntos.

Ventajas y desventajas

Los árboles de decisión siguen siendo populares por razones como estas:

1.- Son fáciles de entender
2.- Cualquier información requiere una preparación mínima
3.- Se pueden agregar nuevas opciones a los árboles existentes
4.- Se puede escoger la mejor de varias opciones
5.- Se combina fácilmente con otras herramientas para la toma de decisiones

Sin embargo, los árboles de decisión pueden volverse excesivamente complejos. En tales casos, un diagrama de influencia más compacto puede ser una buena alternativa. Los diagramas de influencia reducen el enfoque a las decisiones críticas y los objetivos.

Arboles de decisión en aprendizaje automático y minería de datos

Un árbol de decisión también se puede usar para ayudar a crear modelos predictivos automáticos, que tienen aplicaciones en aprendizaje automático, extracción de datos y estadísticas. Conocido como el aprendizaje del árbol de decisión, este método tiene en cuenta las observaciones sobre un elemento para predecir el valor de ese elemento.

En estos árboles de decisión, los nodos representan datos en lugar de decisiones. Este tipo de árbol también se conoce como árbol de clasificación. Cada rama contiene un conjunto de atributos, o reglas de clasificación, que están asociados con una etiqueta de clase particular, que se encuentra al final de la rama..

Cada dato adicional ayuda al modelo a predecir con mayor precisión a cuál de un conjunto finito de valores pertenece el sujeto en cuestión. Esa información puede usarse como una entrada en un modelo de toma de decisiones más grande.

Random Forest

Un clasificador Random Forest consta de varios árboles diseñados para aumentar la tasa de clasificación.
Un árbol de decisión se considera óptimo cuando representa la mayoría de los datos con el menor número de niveles o preguntas. Los algoritmos diseñados para crear árboles de decisión optimizados incluyen CART, ASSISTANT, CLS e ID3 / 4/5. También se puede crear un árbol de decisiones creando reglas de asociación, colocando la variable de destino a la derecha.
Cada método debe determinar cuál es la mejor manera de dividir los datos en cada nivel. Los métodos comunes para hacerlo incluyen medir la impureza de Gini, la ganancia de información y la reducción de la varianza.

3.5.- QDA

El clasificador cuadrático o *Quadratic Discriminat Analysis (QDA)* se asemeja en gran medida al *LDA*, con la única diferencia de que el *QDA* considera que cada clase *k* tiene su propia matriz de covarianza $(\sum_k)$ y, como consecuencia, la función discriminante toma forma cuadrática.

$$log(P(Y=k|X=x)) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log(\pi_k)$$

Para poder calcular la probabilidad a posteriori , es necesario estimar, para cada clase, (∑k), μk y πk a partir de la muestra. Cada nueva observación se clasifica en aquella clase para la que el valor de la probabilidad *a* posteriori sea mayor. Hay que añadir que *QDA* genera límites de decisión curvos por lo que puede aplicarse a situaciones en las que la separación entre grupos no es lineal.

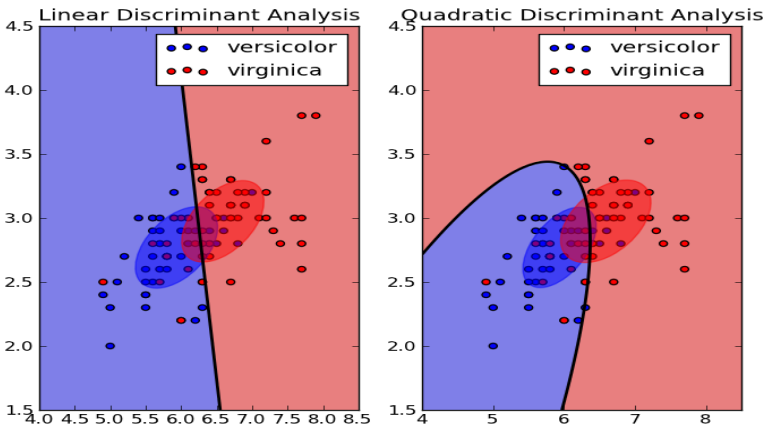


Ilustración 15 Ejemplo 1 del algoritmo LDA y QDA

El proceso de un análisis discriminante puede resumirse en 6 pasos:

1.- Disponer de un conjunto de datos de entrenamiento (*training data*) en el que se conoce a que grupo pertenece cada observación.

2.- Calcular las probabilidades previas (*prior probabilities*): la proporción esperada de observaciones que pertenecen a cada grupo.

3.- Determinar si la varianza o matriz de covarianzas es homogénea en todos los grupos. De esto dependerá que se emplee *LDA* o *QDA*.

4.- Estimar los parámetros necesarios para las funciones de probabilidad condicional, verificando que se cumplen las condiciones para hacerlo.

5.- Calcular el resultado de la función discriminante. El resultado de esta determina a qué grupo se asigna cada observación.

6.- Utilizar validación cruzada (*cross-validation*) para estimar las probabilidades de clasificaciones erróneas.

¿Cómo elegir entre LDA y QDA?

La diferencia es realmente una compensación de sesgo-varianza. Con los predictores p, la estimación de una matriz de varianza requiere estimar los parámetros p (p + 1) / 2. El QDA estima una matriz de varianza separada para cada clase, por lo que a medida que el número de predictores se vuelve alto, se experimenta un gasto computacional. Por el contrario, si se asume una matriz de varianza común, solo se tiene que hacer el cálculo una vez. LDA es un clasificador mucho menos flexible que QDA, por lo que tiene una varianza sustancialmente menor. Sin embargo, si la suposición de la varianza uniforme es muy baja, entonces LDA puede sufrir un alto sesgo. En general, LDA tiende a ser mejor que QDA si hay relativamente pocas observaciones de entrenamiento, por lo que la reducción de la varianza es crucial. Se recomienda QDA si el conjunto de entrenamiento es muy grande, de modo que la varianza del clasificador no es una preocupación importante.

Entre la regresión logística LDA y QDA, las cosas más importantes a tener en cuenta son el tipo de límite de decisión que se requiere. Si es altamente lineal, que LDA y Logistic puede resultar superior, si no es lineal, el borde puede asignarse a QDA.

Ejemplo:

Al tratar de clasificar las observaciones en tres clases (codificadas por colores):

LDA (diagrama izquierdo) proporciona límites de decisión lineales que se basan en la suposición de que las observaciones varían consistentemente en todas las clases. Sin embargo, al observar los datos, resulta evidente que la variabilidad de las observaciones dentro de cada clase difiere.

En consecuencia, QDA (gráfico derecho) es capaz de capturar las diferentes covarianzas y proporcionar límites de decisión de clasificación no lineales más precisos.

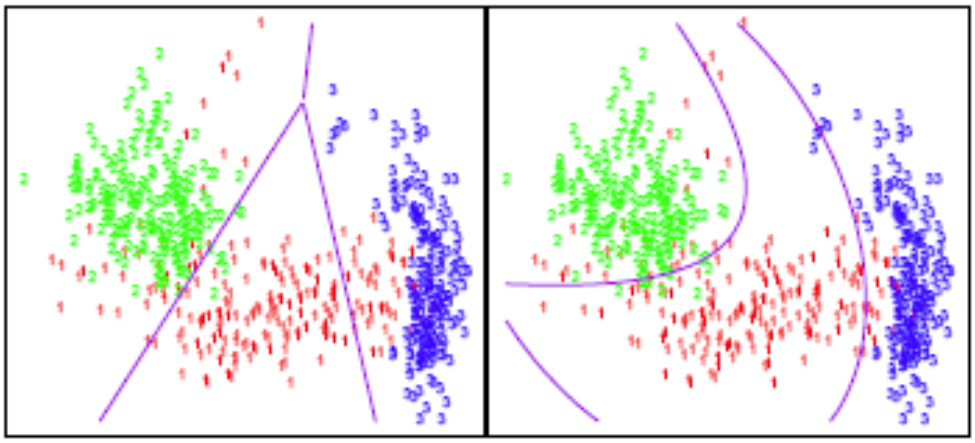


Ilustración 16 Ejemplo 2 del algoritmo LDA y QDA

3.6.- SVM

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.

Vectores de soporte

Son aquellos vectores que entrenan individuos que se encuentran en el hiperplano de separación óptimo y son los patrones más difíciles de clasificar. En cierto sentido, son los patrones más informativos

La complejidad de un clasificador SVM, depende de los puntos SV, no de la dimensión del espacio de características

SVM de margen rígido

El SVM de margen rígido es muy sensible al ruido en los datos, ya que un solo valor atípico puede determinar el hiperplano de separación

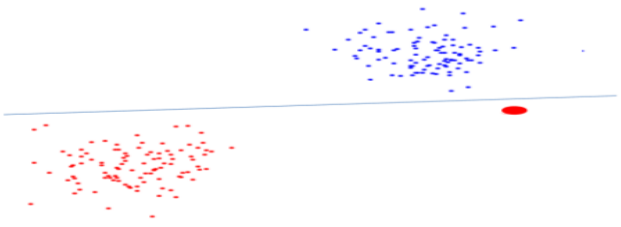


Ilustración 17 Ejemplo del algoritmo SVM de margen rígido

Además, las clases usualmente (siempre) se superponen. Por lo tanto, existe la necesidad de relajar la condición de separabilidad.

Hiperplano de margen suave

Si las clases se superponen, el problema ya no es separable

Luego, se permite cierta flexibilidad para algunos puntos (ξ i distancia del punto "mal clasificado" al margen correcto).

En este caso mal clasificado significa en el lado equivocado del margen

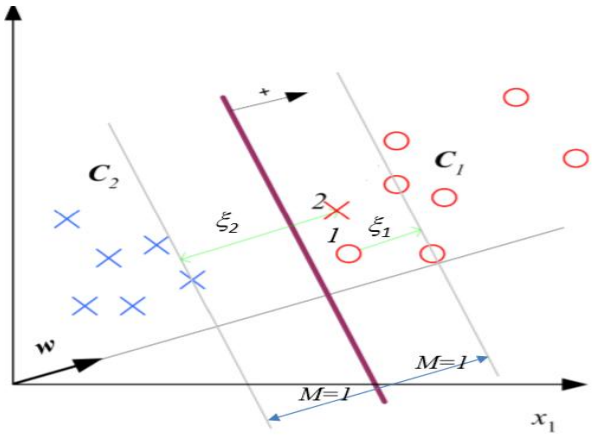


Ilustración 18 Ejemplo del hiperplano de margen suave

Rol de la C

C representa la tolerancia prevista con respecto a los puntos en el lado incorrecto del hiperplano de separación. Para C pequeña, más puntos incorrectos. Al establecer C muy alto, estamos en el caso del margen difícil. Disminuir C hace que el clasificador sacrifique la separabilidad lineal para ganar estabilidad, en un sentido en el que la influencia de cualquier punto de datos único ahora está limitada por C. De ahí una compensación entre el ancho del margen y el error de entrenamiento.

Alto C, bajo margen, solución muy nítida

Bajo C, alto margen, solución muy suave

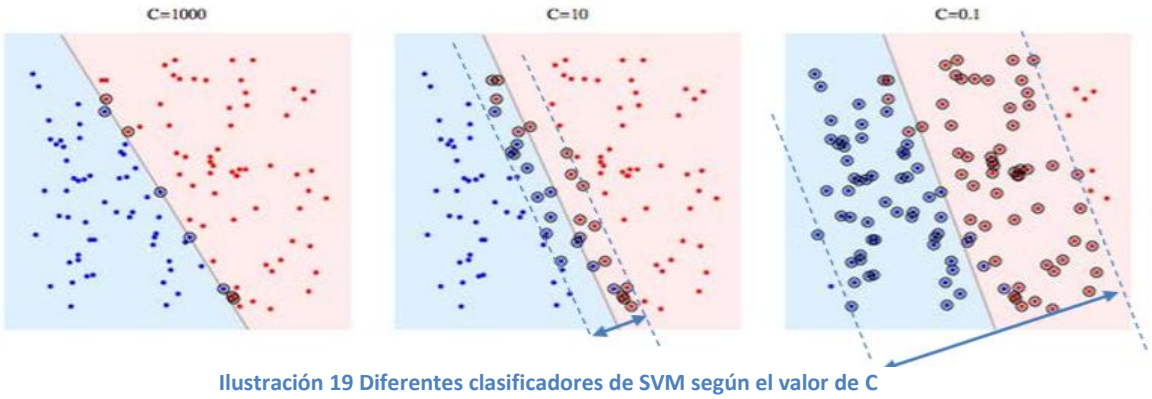


Ilustración 19 Diferentes clasificadores de SVM según el valor de C

3.7.- Validación cruzada (Cross-Validation)

La **validación cruzada** o **cross-validation** es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevará a cabo a la práctica

Contexto

La validación cruzada proviene de la mejora del método de retención o *holdout method*. Este consiste en dividir en dos conjuntos complementarios los datos de muestra, realizar el análisis de un subconjunto(denominado datos de entrenamiento o *training set*), y validar el análisis en el otro subconjunto (denominado datos de prueba o *test set*), de forma que la función de aproximación sólo se ajusta con el conjunto de datos de entrenamiento y a partir de aquí calcula los valores de salida para el conjunto de datos de prueba (valores que no ha analizado antes). La ventaja de este método es que es muy rápido a la hora de computar. Sin embargo, este método no es demasiado preciso debido a la variación de resultados obtenidos para diferentes datos de entrenamiento.

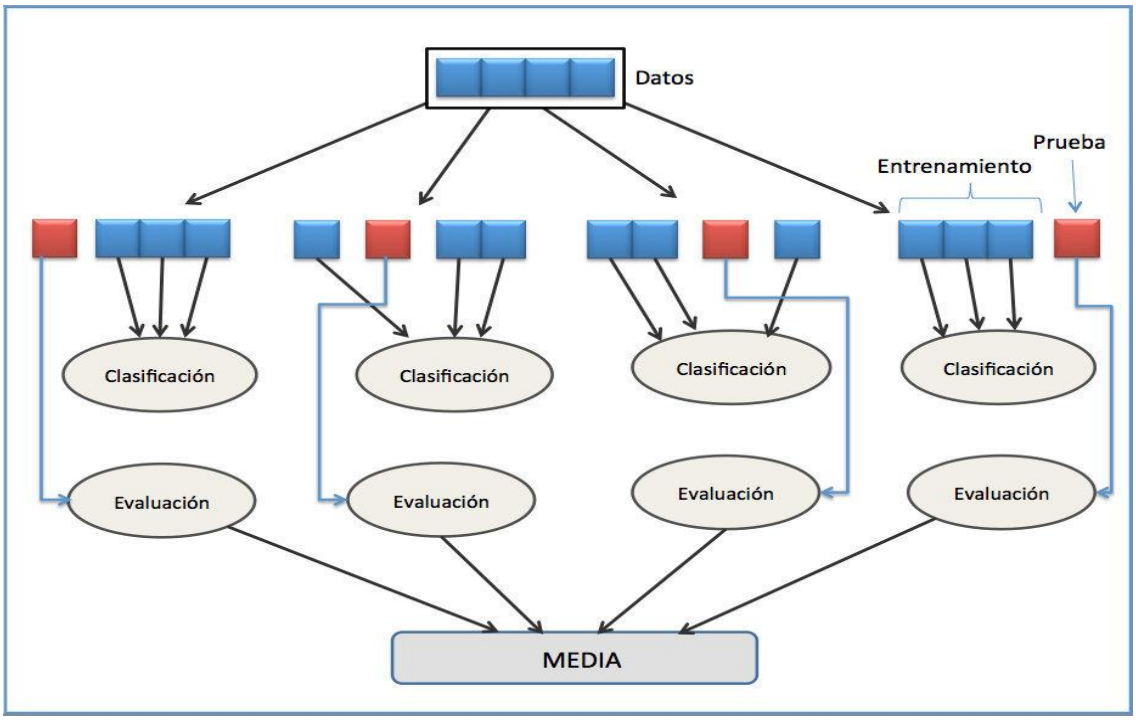


Ilustración 20 Esquema k-fold cross validation, con k=4 y un solo clasificador.

Objetivo de la validación cruzada

La validación cruzada es una manera de predecir el ajuste de un modelo a un hipotético conjunto de datos de prueba cuando no disponemos del conjunto explícito de datos de prueba.

Validación cruzada aleatoria

Este método consiste al dividir aleatoriamente el conjunto de datos de entrenamiento y el conjunto de datos de prueba. Para cada división la función de aproximación se ajusta a partir de los datos de entrenamiento y calcula los valores de salida para el conjunto de datos de prueba. El resultado final se corresponde a la media aritmética de los valores obtenidos para las diferentes divisiones

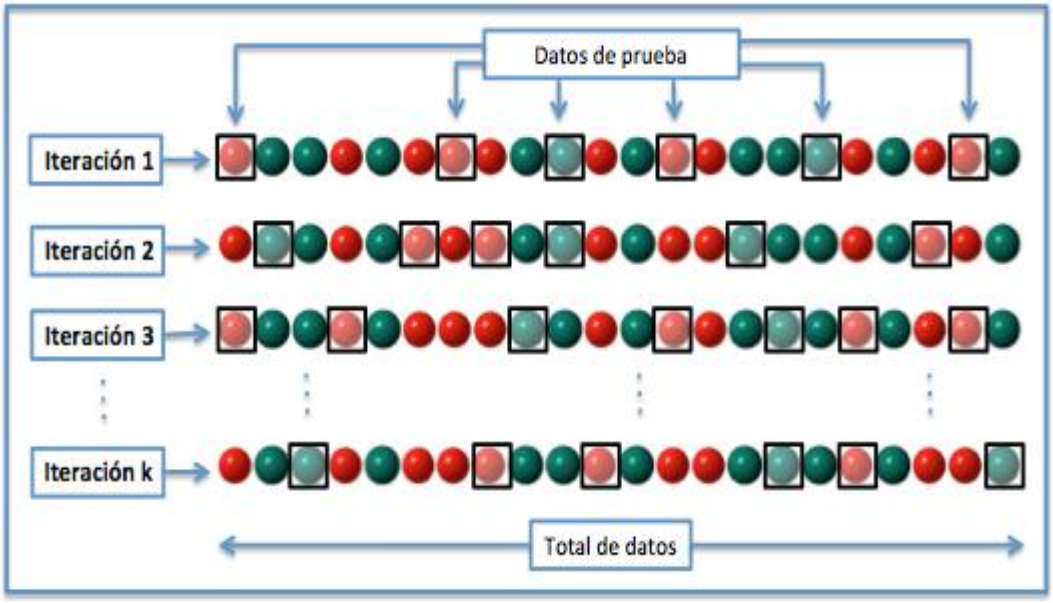


Ilustración 21 Elección aleatoria de los datos de entrenamiento y de prueba

Error de la validación cruzada aleatoria

En la validación cruzada aleatoria, se cogen muestras al azar durante k iteraciones y se realiza un cálculo de error para cada iteración. El resultado final también lo obtenemos a partir de realizar la media aritmética de los K valores de errores obtenidos, según la misma fórmula:

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

Es decir, se realiza el sumatorio de los K valores de error y se divide entre el valor de K.

4.- Resultados

4.1.- Base de datos

Este TFG se ha realizado a partir del estudio de una base de datos extraída de la página web Kaggle. Esta base, proporciona las cuotas de diferentes casas de apuestas para 5 temporadas de la liga italiana. Entre otras variables, se incluyen el equipo local, equipo visitante, el resultado final, quien ganaba en la media parte (como variable categórica), quien acabó ganando el partido (como variable categórica) y la fecha del encuentro.

Como variable respuesta se tendrá en cuenta la variable FTR, que hace referencia a que equipo acabó ganando el partido. Esta variable se compone de tres niveles:

- H (Home): Gana el equipo local

- D (Draw): Empate

- A (Away): Gana el equipo visitante

Por otro lado, como variables predictoras, se tendrán en cuenta 7 casas de apuestas, las cuales ofrecen cada una tres cuotas diferentes para cada partido, una por cada resultado(H,D,A). Por ejemplo, la casa de apuestas William&Hill (WH), proporciona una cuota para las personas que quieren apostar a que gana el equipo local (WHH), otra para las que quieren apostar a empate (WHD) y una tercera para el equipo visitante (WHA). Esto da lugar a un total de 7*3=21 variables predictoras.

Se ha tenido que adaptar la base de datos, excluyendo información no relevante para este trabajo, como por ejemplo las cuotas ofrecidas para aquellos jugadores que deseen apostar a que el equipo local o visitante gana por 2 o mas goles, cuotas para apuestas de córners…

Además, se observó en su momento, que había casas de apuestas que en ciertas jornadas de diferentes temporadas no presentaban las cuotas. Por lo tanto, debido a la falta de información se decidió excluirlas y seleccionar únicamente aquellas casas de apuestas que presentaban todas las cuotas para todos los partidos.

Una vez limpiada la base de datos, esta quedó formada por 1900 filas, que hacen referencia a los 1900 partidos que se jugaron en las últimas 5 temporadas de la liga italiana (380 partidos por temporada). A su vez , como se ha comentado, la base datos queda formada por 21 variables predictoras que son las cuotas para 7 casas de apuestas y la variable endógena “FTR”(Resultado final). Además, se incluyó el día y los nombres del equipo local y visitante, los cuáles no se utilizarán para aplicar los diferentes clasificadores (SVM,KNN…), sino únicamente como información adicional del evento.

A continuación se muestra una tabla con la descripción de los diferentes campos.

Tabla 8 Explicación de las variables

Date	Fecha del partido
HomeTeam	Equipo Local
AwayTeam	Equipo Visitante
FTR	Resultado final (H=gana local, D= Empate, A=gana visitante)
B365H	Bet365: Cuota para el equipo local
B365D	Bet365: Cuota para el empate
B365A	Bet365: Cuota para el equipo visitante
BWH	Bet&Win: Cuota para el equipo local
BWD	Bet&Win: Cuota para el empate
BWA	Bet&Win: Cuota para el equipo visitante
IWH	Interwetten: Cuota para el equipo local
IWD	Interwetten: Cuota para el empate
IWA	Interwetten: Cuota para el equipo visitante
LBH	Ladbrokes: Cuota para el equipo local
LBD	Ladbrokes: Cuota para el empate
LBA	Ladbrokes: Cuota para el equipo visitante
PSH	Pinnacle: Cuota para el equipo local
PSD	Pinnacle: Cuota para el equipo visitante
PSA	Pinnacle: Cuota para el equipo visitante
WHH	William Hill: Cuota para el equipo local
WHD	William Hill: Cuota para el empate
WHA	William Hill: Cuota para el equipo visitante
VCH	VC Bet: Cuota para el equipo local
VCD	VC Bet: Cuota para el empate
VCA	VC Bet: Cuota para el equipo visitante

En la siguiente página, se podrá observar un subconjunto de la base de datos, en la cual aparecen diferentes características acerca del partido y las cuotas para tres casas de apuestas.

Date	HomeTeam	AwayTeam	FTR	B365H	B365D	B365A	BWH	BWD	BWA	IWH	IWD	IWA
25/08/2012	Fiorentina	Udinese	H	2.05	3.2	3.75	2.05	3.3	3.9	2	3.2	3.4
25/08/2012	Juventus	Parma	H	1.3	5	10.5	1.28	5	11	1.35	4.5	7.3
26/08/2012	Atalanta	Lazio	A	2.4	3.2	3	2.4	3.2	2.95	2.6	3.1	2.6
26/08/2012	Chievo	Bologna	H	2.1	3.2	3.6	2.1	3.15	3.6	2.2	3	3.2
26/08/2012	Genoa	Cagliari	H	1.95	3.3	4	2	3.25	3.8	2	3.15	3.45
26/08/2012	Milan	Sampdoria	A	1.4	4.5	8	1.48	4.1	7.75	1.45	3.9	6.5
26/08/2012	Palermo	Napoli	A	3	3.25	2.38	3	3.4	2.4	2.7	3.2	2.4
26/08/2012	Pescara	Inter	A	4.33	3.6	1.8	5	3.55	1.7	4.3	3.4	1.75
26/08/2012	Roma	Catania	D	1.4	4.5	8	1.35	4.75	8.25	1.5	3.8	5.7
26/08/2012	Siena	Torino	D	2.25	3.1	3.4	2.25	3.15	3.25	2.2	3.1	3.1
01/09/2012	Bologna	Milan	A	3.6	3.2	2.1	3.8	3.25	2	3.2	3.2	2.1
01/09/2012	Torino	Pescara	H	1.83	3.4	4.5	1.83	3.5	4.6	1.9	3.3	3.7
02/09/2012	Cagliari	Atalanta	D	2.3	3	3.4	2.2	3	3.5	2.1	3.1	3.3
02/09/2012	Catania	Genoa	H	2.05	3.25	3.75	2	3.3	3.75	2	3.3	3.3
02/09/2012	Inter	Roma	A	1.95	3.6	3.75	2	3.5	3.8	1.85	3.3	3.9
02/09/2012	Lazio	Palermo	H	1.67	3.75	5	1.67	3.6	5.25	1.7	3.4	4.6
02/09/2012	Napoli	Fiorentina	H	1.73	3.5	5	1.75	3.5	4.75	1.7	3.5	4.4
02/09/2012	Parma	Chievo	H	2.05	3.2	3.8	2.05	3.2	3.7	2	3.2	3.4
02/09/2012	Sampdoria	Siena	H	1.83	3.4	4.5	1.83	3.3	4.5	1.85	3.3	3.9
02/09/2012	Udinese	Juventus	A	5	3.5	1.73	5.25	3.5	1.75	3.8	3.4	1.85
15/09/2012	Milan	Atalanta	A	1.5	4	7	1.42	4.1	8.25	1.45	3.9	6.5
15/09/2012	Palermo	Cagliari	D	2.2	3.2	3.4	2.25	3.3	3.3	2.2	3.2	
16/09/2012	Chievo	Lazio	A	2.8	3.1	2.6	2.85	3.2	2.45	2.7	3.05	2.5
16/09/2012	Fiorentina	Catania	H	1.91	3.3	4.2	1.83	3.45	4.25	1.85	3.3	3.9
16/09/2012	Genoa	Juventus	A	5.5	3.75	1.62	5.75	3.9	1.62	5	3.6	1.6
16/09/2012	Napoli	Parma	H	1.55	3.8	6.5	1.5	3.85	6.9	1.55	3.7	5.4
16/09/2012	Pescara	Sampdoria	A	2.8	3.1	2.6	2.85	3.1	2.5	2.65	3.05	2.55
16/09/2012	Roma	Bologna	A	1.4	4.75	7.5	1.37	4.75	7.7	1.37	4.2	7.2
16/09/2012	Siena	Udinese	D	2.5	3.2	2.88	2.45	3.15	2.9	2.5	3.05	2.7
16/09/2012	Torino	Inter	A	3.1	3.3	2.3	3.3	3.4	2.2	3.2	3.2	2.1

4.2.- Descriptiva

En primer lugar se va a realizar una descriptiva univariante de las diferentes variables.

Descriptiva de la variable Respuesta FTR

Tabla 9 Número de sucesos de cada nivel

	H	D	A
Frecuencia	869	481	550
%	45'74	25'32	28'95

La variable respuesta presenta 3 niveles:

- H (Home): Gana el equipo local
- D (Draw): Empate
- A (Away): Gana el equipo visitante

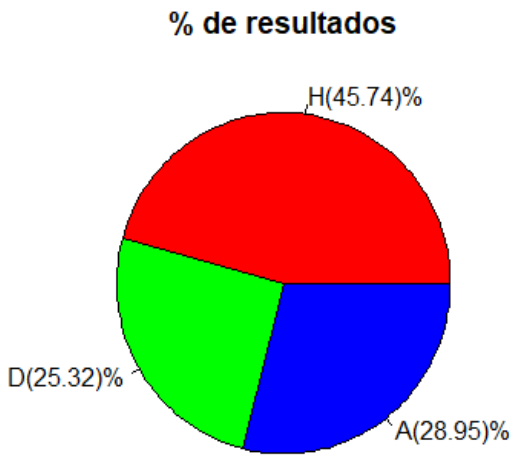


Ilustración 22 Gráfico circular de la variable respuesta

En la descriptiva se observa que:

- El equipo local gana el 45.74% de las veces (869 casos)
- Hay un 25.32% de empates (481 casos)
- El equipo visitante gana el 28.95% de las veces (550 casos)

Ahora se va a hacer el mismo estudio pero en vez de tener en cuenta los 5 años, se tendrá en cuenta año por año:

La tabla mostrará el porcentaje de veces que se ha dado cada resultado(H,D,A) respecto su año

Tabla 10 % de sucesos respecto el año

	H	D	A
2012-2013	46’58	25’26	28’16
2013-2014	47’63	23’68	28’68
2014-2015	40	31’58	28’42
2015-2016	46’05	25	28’95
2016-2017	48’42	21’05	30’53

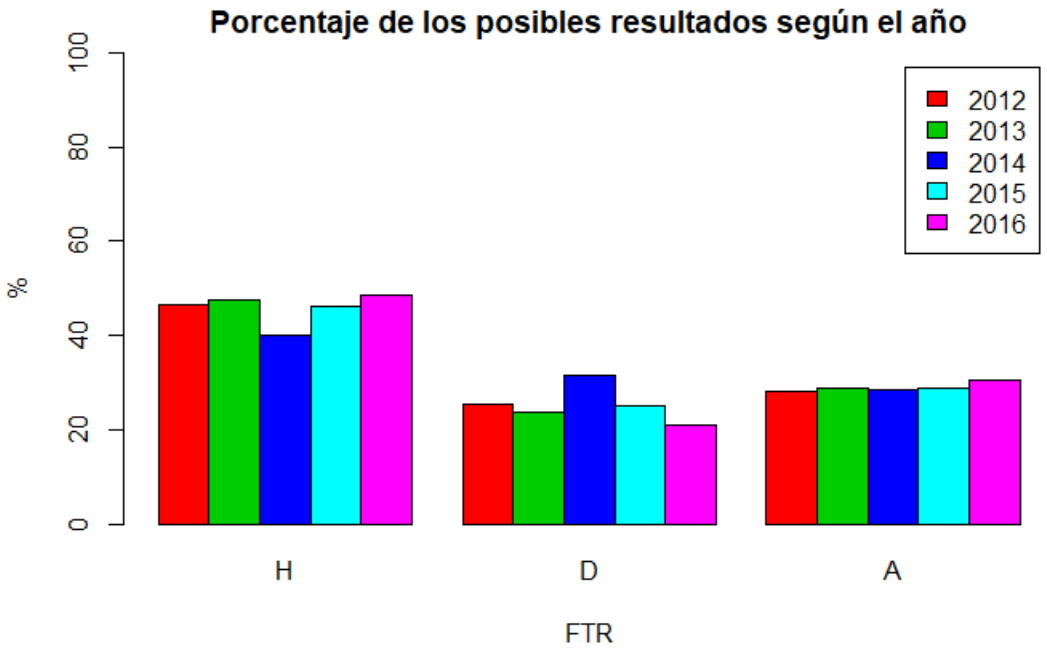


Ilustración 23 Gráfico de Barras según la variable respuesta y el año

En general, como era de esperar, la distribución de H,D,A es muy parecida los diferentes años.

Sin embargo, el año 2014 respecto el resto de años, presenta un porcentaje de victorias para el equipo local algo inferior y un porcentaje de empates más alto.

Descriptiva de las variables predictoras

A continuación, se muestra una tabla con la cuota mínima, la media y la cuota máxima para cada una de las casas de apuestas según si se apuesta a que gana el equipo local, empata o gana el visitante.

Por ejemplo, en el primer caso de todos, se observa que para la casa de apuestas Bet365, si se apuesta a favor del equipo local, la cuota más baja encontrada fue de 1'06€ y la más alta de 15€. Esta segunda cuota, corresponderá a un partido en el cual el equipo local es un equipo mucho más flojo que el visitante. Por lo tanto, como la probabilidad de que gane este equipo es muy baja, la cuota (o retorno bruto) será muy alta.

Casa de apuestas	Cuota Mínima	Media	Cuota Máxima
B365H	1.06	2.62	15.00
B365D	1.73	3.81	15.00
B365A	1.17	4.63	34.00
BWH	1.06	2.60	21.00
BWD	1.67	3.78	11.00
BWA	1.13	4.50	29.00
IWH	1.07	2.50	14.00
IWD	1.60	3.69	9.50
IWA	1.12	4.19	22.00
LBH	1.06	2.57	15.00
LBD	1.83	3.77	12.00
LBA	1.15	4.51	41.00
PSH	1.10	1.76	23.60
PSD	1.94	3.98	12.23
PSA	1.16	5.06	42.32
WHH	1.06	2.63	15.00
WHD	1.20	3.65	12.00
WHA	1.17	4.57	29.00
VCH	1.04	2.73	19.00
VCD	1.98	3.87	15.00
VCA	1.15	4.98	34.00

Ilustración 24 Cuota mínima, media y máxima según la casa de apuestas

Como todas las variables, hacen referencia a cuotas de 7 casas de apuestas para tres sucesos (A(Away): Gana el equipo visitante, D(Draw): Empate, H(Home): Gana el equipo local) se compararán los valores entre ellas.

En primer lugar, se ha observado que la cuota más alta corresponde a la casa de apuestas PS, hace referencia a la cuota que se pagaria a aquella persona que apostará a favor del equipo visitante y la cifra es de 42'320. Esto quiere decir, que por cada euro jugado el retorno es de 42'320. El beneficio neto si se apostaran 5€ a este suceso seria de 5*42'320-5=206'6€

Investigando más a fondo en la base de datos, se ha encontrado que este partido corresponde a Roma vs Genova, el cual se jugó el dia 5 de Mayo de 2017 y acabo con el resultado de victoria para el local. El motivo de la cuota tan alta, fue debido a que la probabilidad de que el Genova ganará a la Roma fuera de casa, era muy baja.

Realizando una comparativa de cuanto pagaban las otras casas de apuesta a este partido se observan diferencias muy significativas:

Tabla 11 Diferentes cuotas según la casa de apuestas

B365A	BWA	IWA	LBA	PSA	WHA	VCA
34	29	22	41	42.32	29	34

Por un lado, la casa de apuestas LB, es la que más se aproxima a los 42'32€ de la casa de apuestas que mejor paga, PS. Por el otro lado IW es la casa de apuestas que peor paga para este partido. En comparación con PS, la casa de apuesta IW, pagaría la victoria del Genova a 22€,casi la mitad.

En conclusión, el análisis de las diferentes cuotas, para las diferentes casas de apuestas, es muy importante, ya que la ganancia puede ser del doble. Esto supone que antes de apostar se tendrán que comparar las cuotas de todas las casas existentes y realizar la apuesta en aquella casa que mejor pague.

En segundo lugar, la cuota más baja encontrada, es de 1'04. Este partido corresponde también a Roma vs Genova el dia 5 de Mayo de 2017, y la casa de apuestas con esta cuota es VC. El valor de la cuota tan baja, es debido a que la probabilidad de que el Roma gane al Genova en su propia casa es muy alta. En concreto, tal como se explica en el apartado 2.3.- ¿Qué son las cuotas?, la probabilidad de que el Roma gane al Genova es de aproximadamente un 1/1’04*100=96’15%

Una cuota de 1'04, quiere decir que por cada euro jugado el retorno es de 1'04. El beneficio neto si se apostaran 5€ a este suceso seria de 5*1'04-5=0'2€.

Como se puede observar, la diferencia del premio en el caso de acertar un resultado u otro es muy grande. También es cierto, que acertar un partido con una cuota muy alta es muy difícil, ya que la probabilidad de dicho suceso no es muy probable. Por otro lado, apostar a una cuota muy baja, nos da la seguridad de que ese suceso ocurra, pero el retorno es muy pequeño.

Cada persona ha de pensar y decidir, si merece la pena realmente apostar 5€ a que gana el Roma al Genova, cuya probabilidad es de aproximadamente un 96%, sabiendo que el retorno es de 20 céntimos.

Estudio de la Variabilidad

Seguidamente, para ver la variabilidad que presentan las casas de apuestas, se muestra un boxplot para cada variable

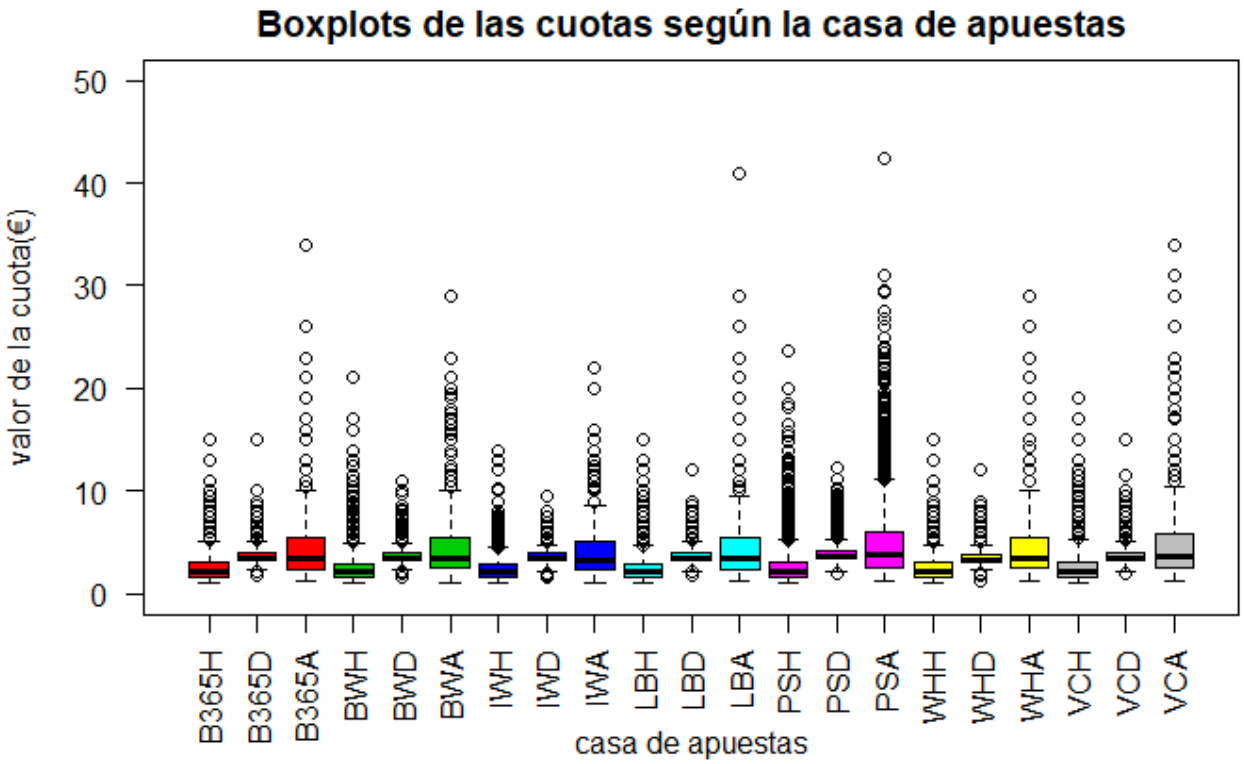


Ilustración 25 Boxplots de las cuotas de las 7 casas de apuestas, 21 variables predictoras

En el boxplot, se aprecia que hay casas de apuestas que presentan una variabilidad más grande que otras. Las que más variabilidad tienen son PS y LB.

Además, en todos los casos, la variabilidad es más grande en las apuestas de "A": Gana el equipo visitante, que corresponden a las variables B365A,BWA,IWA,LBA,PSA,WHA,VCA.

En el caso de las apuestas a victoria local (H) o empate (D), se aprecia que el Rango intercuartílico es significativamente más pequeño que en las apuestas a gana el equipo visitante (A).

Descriptiva Bivariante

A continuación se va a realizar un estudio acerca de los aciertos en los partidos para la conocida casa de apuestas Bet365.

Se va a mostrar cuales han sido las cuotas que se han pagado cuando:

- Gana el equipo local (H)

- Hay empate (D)

- Gana el equipo visitante (A)

Para ello, se van a seleccionar únicamente aquellas cuotas de la casa Bet365 que coincidan con el verdadero resultado (FTR)

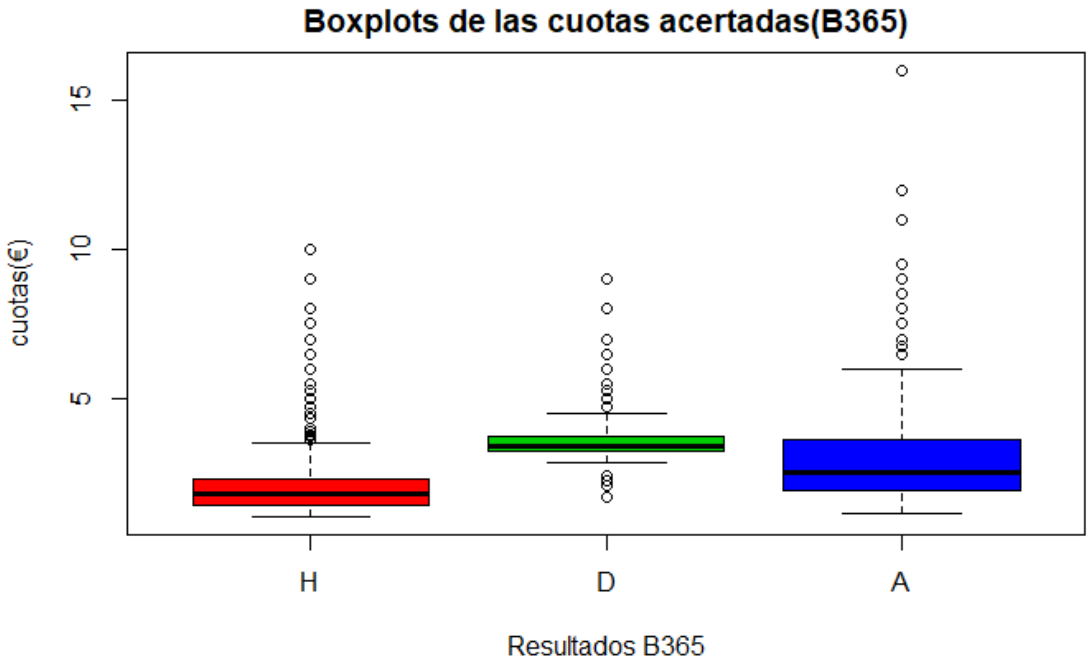


Ilustración 26 Boxplots de las cuotas acertadas, según la variable respuesta para B365

Tabla 12 Resumen de las cuotas acertadas

	Mínimo	1r Cuartil	Mediana	Media	3r Cuartil	Máximo
Gana Local (H)	1.06	1.44	1.83	2.055	2.3	10.00
Empate (D)	1.730	3.250	3.400	3.628	3.750	9.000
Gana Visitante(A)	1.170	1.910	2.500	3.017	3.600	16.000

Gana el equipo local:

Observando el boxplot de las cuotas acertadas para el equipo local(H), en la casa de apuestas B365, se aprecia que cuotas superiores a 3'59 se consideran outliers.

Para el cálculo del outlier se ha tenido en cuenta los siguientes cálculos:

C1: 1r cuartil

C3: 3R CUARTIL

IQR: Rango intercuartílico: C3 - C1 = 2'3 - 1'440 = 0'86

valores atípicos leves superiores: C3 + 1'5 * IQR = 2'3 + 1'5 * 0'86 = 3'59

Considerar como outlier una cuota de 3'59 puede resultar extraño ya que no es extremadamente alta. Esto es debido en primer lugar, a que las cuotas acertadas para equipos locales no suelen ser altas, inferiores a 1'8. Además, hay que comentar que la probabilidad de acertar esta cuota es baja: 1/3'59=0.279. Es decir, hay un 27'9% de probabilidades acertar este resultado.

Por otro lado, la característica a tener más en cuenta en este caso, es que ha habido un partido en el cual se ha acertado una cuota de 10. Este acierto es claramente un caso muy raro, ya que la probabilidad del suceso es de un 10%.

El retorno neto en el caso de jugar 5€ es de : 5*10-5=45€.

Añadir a este análisis específico que el encuentro corresponde a Crotone vs Inter el 9 de Abril de 2017, con el resultado de 2-1 para el Crotone. Se puede entender perfectamente, que el Crotone es un equipo futbolísticamente inferior al Inter, pero en ese encuentro hubo una sorpresa y se llevó el partido el equipo con probabilidades más bajas de ganar.

Observando el resumen inicial de esta variable:

Se puede ver claramente que el acierto normal suele ser de una cuota aproximadamente de entre 1'4 y 2'3.

Estas apuestas, suelen ser en algunos casos seguras, pero por el contrario, el beneficio neto no será excesivamente alto.

Empate:

Siguiendo con el estudio de la casa de apuestas B365, una de las más famosas dentro del sector, en el caso de los empates, el análisis del boxplot y el summary son los siguientes:

La cuota más alta acertada en una apuesta de empate, para esta casa de apuestas es de 9. Este encuentro corresponde al Napoli vs Palermo, jugado el 29 de enero de 2017:

En el caso de los empates, las cuotas normalmente suelen ser mejor pagadas que las apuestas a las victorias del local, debido a su dificultad de acierto. En este caso, suelen oscilar entre 3'25 y 3,75, dando como outliers valores más bajo de 2'5 Y superiores a 4.

C1: 1r cuartil

C3: 3R CUARTIL

IQR: Rango intercuartílico: C3-C1= 3'75 - 3'25=0'5

valores atípicos leves inferiores: C1 - 1'5 * IQR = 3'25 - 1'5 * 0'5 = 2'5

valores atípicos leves superiores: C3 + 1'5 * IQR = 3'75 + 1'5 * 0'5 = 4

Gana el equipo visitante:

En último lugar, se comentará el caso de las cuotas acertadas, en las apuestas a favor de la victoria del equipo visitante(A):

La cuota más alta acertada en una apuesta de victoria visitante, para esta casa de apuestas es de 16. Este encuentro corresponde al Juventus vs Sampdoria, jugado el 6 de enero de 2013, con victoria para el Sampdoria:

En el caso de las victorias visitantes, las cuotas normalmente suelen ser mejor pagadas que las apuestas a las victorias del local pero inferiores a las apuestas de empate. No obstante, se han encontrado aciertos con una cuota más extrema en apuestas al equipo visitante que al empate.

En este caso, las cuotas acertadas para equipos visitantes suelen oscilar entre 1'91 y 3'6, dando como outliers valores superiores a 6'135

5.- Resultados de los métodos de clasificación (Calibración)

5.1.-Introducción

En el apartado de calibración, a diferencia del de validación cruzada, en el cual se dividirán los datos en dos partes (datos de entrenamiento y datos de prueba), en este caso se va a realizar el estudio de todo el conjunto de datos y a continuación , se aplicará la predicción en el mismo conjunto de datos. Es decir, los datos de entrenamiento y de prueba son los mismos (1900 partidos).

Por lo tanto, la calibración puede entenderse de muchas maneras, pero generalmente se trata de dos cuestiones: cómo se distribuye el error y cómo se realiza la autoevaluación (estimación de la confianza o probabilidad).Estimar las probabilidades o los valores de confianza es crucial, en muchas aplicaciones reales. Por ejemplo, si tenemos probabilidades inexactas, las decisiones se pueden tomar con una buena evaluación de riesgos y costes, utilizando modelos de utilidad u otras técnicas de toma de decisiones.

Además, su integración con otros modelos (por ejemplo, multiclasificadores) o con conocimientos previos se vuelve más sólida.En la clasificación, las probabilidades se pueden entender como grados de confianza. En este contexto, y en lugar de rediseñar cualquier método existente para obtener directamente buenas probabilidades o una mejor distribución de errores, algunas técnicas de calibración se han desarrollado hasta la fecha.

Una técnica de calibración, es cualquier técnica de posprocesamiento que apunta a mejorar la estimación de probabilidad o mejorar la distribución de errores de un modelo predictivo. Dado una técnica de calibración general, podemos usarla para mejorar cualquier método existente de aprendizaje automático: árboles de decisión, redes neuronales, métodos kernel, métodos basados en instancias, métodos bayesianos, etc., pero también puede aplicarse a modelos hechos a mano, sistemas expertos o modelos combinados. Dependiendo de la tarea, se pueden aplicar diferentes técnicas de calibración.

Tabla 13 Técnicas de calibración

Tipo	Tarea	Problema	Global/Local	¿Qué calibra?
CD	Clasificación	La distribución de la clase esperada, es diferente de la real	Global/Local	Predicción
CP	Clasificación	La probabilidad estimada de los aciertos es diferente de la verdadera proporción	Local	Probabilidad/confidencial
RD	Regresión	La respuesta esperada es diferente de la respuesta media real	Global/Local	Predicción
RP	Regresión	El error esperado de los intervalos de confianza son demasiado estrechos o anchos	Local	Probabilidad/confidencia

5.2- Naive-Bayes

Información a priori

Como se ha comentado anteriormente, el clasificador probabilístico fundamentado en el teorema de Bayes proporcionará :

1.- La distribución a priori de cada clase de la variable dependiente (variable respuesta). En este caso la variable dependiente es FTR, con niveles H (gana el equipo local), D (hay empate) y A(gana el equipo visitante)

2.- Y una lista de tablas, una por cada predictor de la variable. Por cada variable numérica se presenta la media y la desviación típica. Debido a la gran cantidad de variables predictoras, se mostraran las 3 primeras:

Tabla 14 Resultados de Naive Bayes

```
Naive Bayes Classifier for Discrete Predictors
Call: naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

      H           D           A
0.4573684  0.2531579   0.2894737

Conditional probabilities:

 B365H

      [,1]      [,2]
H 2.055017  0.9622076
D 2.566570  1.3606968
A 3.573109  2.2969715

 B365D

      [,1]      [,2]
H 4.038032  1.1628611
D 3.627692  0.7202484
A 3.616309  0.6413496

 B365A

      [,1]      [,2]
H 5.900840  4.090135
D 4.190873  2.725484
A 3.017182  1.725448
```

Precisión

Tabla 15 Precisión (NB)

Aciertos	%
867/1900	45’63

Se puede observar, que Naive Bayes tiene una capacidad de predicción no muy buena. No obstante, un porcentaje de acierto inferior a 50, no quiere decir que siempre vayan a haber perdidas aseguradas. Es decir, en un primer momento, se puede pensar que a mayor número de aciertos , mayor sera el beneficio, pero esto realmente no es del todo cierto. Es possible que un clasificador no sea el que acierte más pero tenga el retorno neto más grande, debido a que es capaz de predecir correctamente partidos donde las cuotas están muy bien pagadas, como es el caso de los empates.

Un ejemplo sencillo seria el siguiente:

Se realizan 5 apuestas a diferentes partidos, 4 de ellas a que gana el local y la última a empate. Imaginemos que se apuesta un euro por partido.

Supongamos que las 4 cuotas donde hemos apostado que gana el local se pagan a 1’5€ y la apuesta del empate tiene una cuota de 6’25€. En el caso que únicamente acertemos uno de los cinco partidos y este sea el del empate, la precisión que hemos tenido ha sido de 1/5=20%, pero el retorno neto ha sido de -4 +6’25-1=1’25€. Resultado positivo.

-4€ -> perdidos en las apuestas de gana el local

6’25-1 -> ganancias por acertar la cuota del empate (se resta el euro jugado del beneficio total)

Matriz de confusión

Las columnas, hacen referencia al verdadero resultado y las filas a la predicción. Es decir, la suma de las columnas, es realmente el verdadero número de veces que ha ocurrido cada suceso (H,D,A). Es por eso, que todas todas sumaran 869,481 y 550.

Por otro lado, los valores que se encuentran en color verde, en la diagonal, hacen referencia, a los valores bien clasificados.

Se han predicho 296 victorias del local, 237 empates y 334 victorias del visitante correctamente. Por otro lado, se han predicho 411 empates, que realmente eran victorias del local.

	H	D	A
H	296	82	38
D	411	237	178
A	162	162	334

Retorno neto

El retorno neto obtenido, una vez se ha apostado un euro a a cada partido, es el siguiente:

Tabla 16 Retorno según la casa de apuestas (NB)

Casa de apuestas	retorno
B365	-20.94
BW	-27.20
IW	-37.38
LB	-29.39
PS	+27.99
WH	-29.92
VC	-2.05

Todas las casas de apuestas excepto PS, proporcionan perdidas. El hecho que PS devuelva un beneficio, es debido a que las cuotas que ha proporcionado en algunos partidos han sido bastante más altas que el resto de casas. De hecho, PS, ofreció para varios partidos, cuotas excesivamente altas para sucesos muy poco probables, que acabaron sucediendo. Ese es el principal motivo por el que marca la diferencia.

10 primeros encuentros

En la tabla que se muestra a continuación, se refleja el dinero en forma de beneficio o perdida, que se va acumulando, partido tras partido. En este caso, solo se muestran las 10 primeras variaciones, para los 10 primeros partidos según la casa de apuestas que se elija.

Tabla 17 Variación del beneficio o pérdida por jornada (NB)

Jornada	B365	BW	IW	LB	PS	WH	VC
1	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
2	-0.70	-0.72	-0.65	-0.7	-0.67	-0.67	-0.67
3	-1.70	-1.72	-1.65	-1.7	-1.67	-1.67	-1.67
4	-2.70	-2.72	-2.65	-2.7	-2.67	-2.67	-2.67
5	-3.70	-3.72	-3.65	-3.7	-3.67	-3.67	-3.67
6	-4.70	-4.72	-4.65	-4.7	-4.67	-4.67	-4.67
7	-3.32	-3.32	-3.25	-3.4	-3.17	-3.17	-3.17
8	-2.52	-2.62	-2.50	-2.7	-2.33	-2.37	-2.34
9	-3.52	-3.62	-3.50	-3.7	-3.33	-3.37	-3.34
10	-1.42	-1.47	-1.40	-1.6	-1.21	-1.17	-1.22

Se observa, como en la primera jornada se falló la apuesta, por eso marca -1€ en todas las casas(1ra fila). En cambio, en el segundo partido se acierta. Por lo tanto, se reduce la perdida inicial. En B365 marca -0.70, ya que la cuota que se ha acertado es de 1.30. Entonces, 1’30-1(apuesta inicial) = 0.30€ de beneficio en el segundo partido. Este beneficio restando el euro perdido del primer partido, nos sitúa tras la segunda jornada en -1+0’3= -0.7€.

Gráfico

La tabla de 10 valores que se ha mostrado anteriormente, es realmente una tabla de 1900 valores para cada casa de apuesta. Gracias a esta tabla construida, se puede representar un gráfico, para poder apreciar con más detalle los movimientos que se han realizado y como ha influido cada casa de apuestas en el retorno.

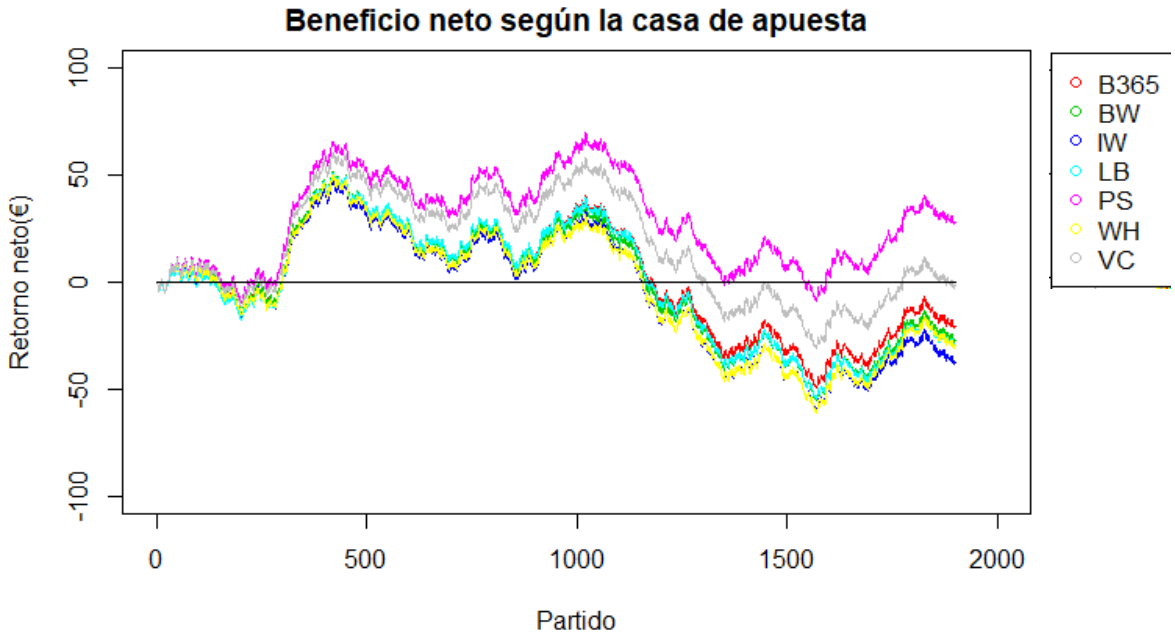


Ilustración 27 Variación del Beneficio neto según la casa de apuestas y jornada para (NB)

En un primer momento, se mantiene con valores positivos y después negativos. No es hasta el partido 280, donde el beneficio aumenta significativamente. No obstante, después de los 1000 partidos, el beneficio disminuye, produciendo que cuando se llega al partido 1900, tan solo una casa de apuestas (PS=Pinnacle), se mantiene en positivo.

Mejora del beneficio

Apreciando anteriormente la tabla de retorno neto según la casa de apuestas, se ha visto que la mejor casa para apostar utilizando el método de la Naive Bayes es PS, con un beneficio de 27'99€.

Este resultado se puede mejorar significativamente si en vez de apostar siempre a la misma casa (PS), se comparan las cuotas con diferentes casas y se elige aquella que claramente aporta un beneficio mayor. Si se tiene en cuenta esta estrategia, se maximizará el beneficio, hasta encontrar el óptimo. En este caso, después de aplicar un código de recorrido en R Studio, el retorno neto aumenta hasta los 50'17€. Este suceso es muy relevante, ya que casi se duplican los beneficios, únicamente teniendo en cuenta el factor comentado.

Tabla 18 Mejora del beneficio (NB)

Antes de la mejora	Despúes de la mejora
27'99€	50.17€

Además, a continuación, se va a mostrar para los partidos acertados, cuantas veces se apostaría a cada casa, si se tiene en cuenta el hecho de maximizar el beneficio eligiendo la casa con la cuota más alta.

Tabla 19 Número de veces que se apuesta a cada casa (NB)

B365	BW	IW	LB	PS	WH	VC
75 veces	108 veces	203 veces	105 veces	392 veces	268 veces	162 veces

La casa en la cual más veces se apostaría es PS seguida de WH. Una posible pregunta, es porqué si sumamos las veces que se ha de apostar en cada casa de apuestas, esta no es igual a 1900? Es decir, 75 + 108 + 203 + 105 + 392 +268 + 162= 1316

Este suceso, es debido a dos cosas:

En primer lugar, la tabla muestra las veces que se apuesta en cada casa, en el caso que se haya acertado el encuentro, por lo tanto no se ha tenido en cuenta las veces que se ha perdido.

En segundo lugar, hay encuentros(partidos) en los que las casas presentan las mismas cuotas. Por lo tanto, si hay tres casas que ofrecen una cuota que es la misma para las tres y a la vez la mas alta de las 7, en la tabla aparecerá tres veces, una para cada casa. Hay que tener en cuenta que en caso de que haya casas que presenten cuotas iguales, no importa donde se apueste, a no ser que exista alguna promoción o bono de bienvenida en alguna casa e interese aprovechar la oportunidad.

En último lugar, se mostrará el siguiente gráfico:

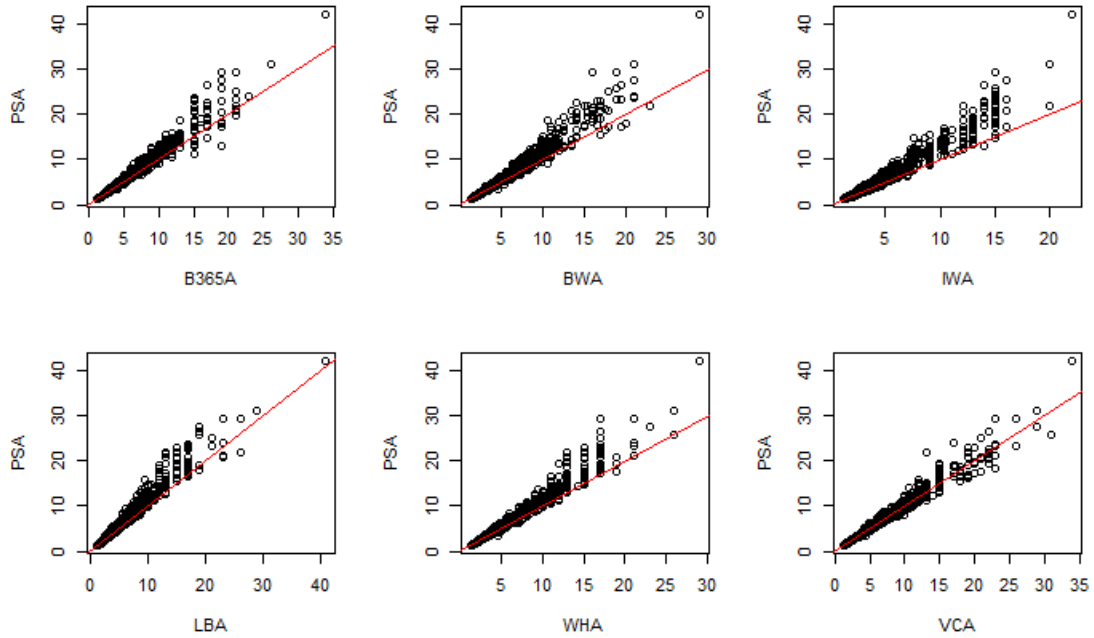


Ilustración 28 Comparación de las cuotas de PS respecto el resto de casas (Para partidos de victoria visitante)

Este gráfico compara la única casa de apuestas que ha dado beneficio (PS) con el resto de casas. Se proyectaran las cuotas de las diferentes casas de apuestas para las variables predictoras de gana el equipo visitante, que son 7 (B365A,BWA,IWA,LBA,,PSA,WHA,VCA). Se puede observar claramente en el gráfico de arriba a la derecha, que casi todas las cuotas que se pagan en PS, son superiores a la de la casa de apuesta IW. Por lo tanto, tiene sentido, que cuando se acierte un partido el retorno sea superior en PS que en IW

5.3.- KNN

Información a priori

Para determinar el parámetro k, se han probado diferentes (k=1,5,7..), y se ha elegido aquel que proporciona un retorno neto más grande. Es decir, en vez de fijarse en el porcentaje de aciertos, el estudio se ha centrado en el beneficio. Claramente, nos interesa más un clasificador que proporcione un beneficio más alto, que uno que tiene un porcentaje de acierto mayor pero el beneficio no es tan alto. La explicación de este curioso suceso viene a continuación.

En la tabla que se muestra a continuación aparecen el número de aciertos, su respectivo porcentaje teniendo en cuenta que hay 1900 partidos y el beneficio de retorno en € si se apuesta a una casa en concreto.

Tabla 20 Número de aciertos, % i retorno para los diferentes K

K	aciertos	%	B365	BW	IW	LB	PS	WH	VC
1	851/1900	44'8	-9.52	-18.96	-40.50	-22.32	54.47	-11.40	25.80
5	932/1900	49'0	-2.78	-11.26	-22.58	-10.59	48.65	5.06	26.43
7	961/1900	50'6	14.08	9.77	-1.87	8.56	65.40	24.37	43.59
10	1004/1900	52'8	68.52	65.91	59.21	62.48	119.25	85.44	100.09
20	1012/1900	53'3	12.35	13.22	9.26	8.82	58.18	36.52	41.72
50	1029/1900	54'2	21.78	25.13	23.48	20.51	66.89	51.91	52.04
100	1026/1900	54'0	-44.82	-37.37	-39.07	-46.55	-1.09	-9.12	-13.48

Se puede apreciar que el parametro k con el que se aciertan más partidos corresponde a k =50, en cambio se observa que el retorno no es tan alto como k=10, el cual acierta, con un porcentaje menor. Esto es debido a que k=10 acierta más empates que k=50 (87 vs 37 respectivamente. Se pueden observar estos valores en la matrices de confusión de la página siguiente). Hay que recordar que los empates están mejor pagados, por lo tanto acertar un empate, produce como consecuencia un beneficio más alto. Por otro lado; k =50 acierta más partidos de equipos visitantes, y aunque también se suelen pagar bien estas apuestas, seguramente, se acertaron partidos, donde la cuota del visitante era muy baja debido a que era el favorito. Esto ocasiona que aunque se haya acertado el resultado, el retorno del cuota no sea muy alto.

Observando las matrices de confusión de otros k=1,k=5 y k=7, se observa que el porcentaje de acierto es menor que k =10, pero se aciertan más empates. De hecho cuando más pequeño es k, se llegan a acertar más empates. De acuerdo con la explicación anterior de la comparación entre k =10 y k =50, el hecho de acertar más empates con k más pequeño puede llevar a la pregunta de si realmente tendría que dar un beneficio mayor estos casos.

La explicación en este caso seria la siguiente:

Para k=1, aunque es el que más empates acierta, el porcentaje de aciertos total (44'8%) respecto el resto es significativamente inferior.

Para k =5 y K=7, predicen correctamente bastantes empates(103 y 99 respectivamente) pero seguramente las cuotas de los empates que acierta k =10 están mejor pagados.

Por lo tanto, un k muy pequeño o muy grande, da retornos negativos. Por otro lado, la precisión de acierto ha sido mejor para k grandes que para k pequeños.

K=1

	H	D	A
H	484	217	173
D	205	133	143
A	180	131	234

K=5

	H	D	A
H	575	250	195
D	144	103	101
A	150	128	254

K=7

	H	D	A
H	609	254	198
D	118	99	99
A	142	128	253

K=10

	H	D	A
H	644	251	191
D	101	87	86
A	124	143	273

K=20

	H	D	A
H	679	290	214
D	62	53	56
A	128	138	280

K=50

	H	D	A
H	693	308	223
D	36	37	28
A	140	136	299

K=100

	H	D	A
H	704	319	226
D	15	7	9
A	150	155	315

El estudio se realizará con k=10

Precisión

Tabla 21 Precisión (KNN)

Aciertos	%
1004/1900	52’8

Matriz de confusión

	H	D	A
H	644	251	191
D	101	87	86
A	124	143	273

Se observa, como se han predicho correctamente 644 apuestas al equipo local,87 empates y 273 apuestas al equipo visitante.

Retorno neto

Tabla 22 Retorno según la casa de apuestas (KNN)

Casa de apuestas	retorno
B365	68.52
BW	65.91
IW	59.21
LB	62.48
PS	119.25
WH	85.44
VC	100.09

En cuanto al retorno neto, se aprecia como la casa de apuestas PSA es la que proporciona un beneficio mayor. Además si comparamos su beneficio de 119’25€ con la casa de apuestas IW que tiene un beneficio de 59’21€, se puede ver que el beneficio es el doble. Por lo tanto el factor de la casa de apuestas es una de las características que debe tener más en cuenta un jugador.

10 primeros encuentros

Del total de 1900 partidos/encuentros se muestra a continuación, como varía el beneficio o pérdida en cada partido, si se apuesta 1€ por encuentro.

En este caso se observa que en el encuentro número 10, el beneficio que lleva el jugador que ha empleado el método de knn(k=10) para la casa de apuestas B365 es de 6'08€.

Tabla 23 Variación del beneficio o pérdida por jornada (KNN)

Jornada	B365	BW	IW	LB	PS	WH	VC
1	1.05	1.05	1.00	1.15	1.10	1.10	1.10
2	1.35	1.33	1.35	1.45	1.43	1.43	1.43
3	0.35	0.33	0.35	0.45	0.43	0.43	0.43
4	1.45	1.43	1.55	1.55	1.66	1.58	1.63
5	2.40	2.43	2.55	2.55	2.69	2.68	2.68
6	1.40	1.43	1.55	1.55	1.69	1.68	1.68
7	2.78	2.83	2.95	2.85	3.19	3.18	3.18
8	3.58	3.53	3.70	3.55	4.03	3.98	4.01
9	7.08	7.28	6.50	6.75	7.97	7.48	7.81
10	6.08	6.28	5.50	5.75	6.97	6.48	6.81

Gráfico

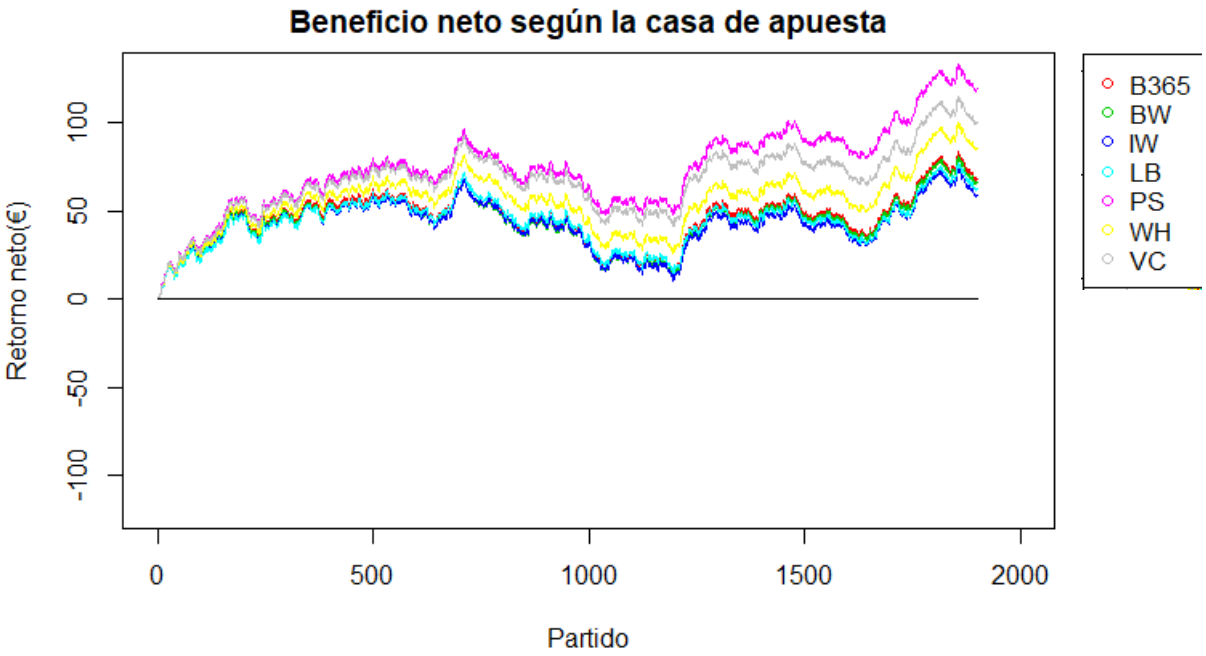


Ilustración 29 Variación de Beneficio neto según la casa de apuestas y jornada (KNN)

El gráfico deja claro que es un método bastante bueno, ya que se ha mantenido en un beneficio constante y positivo para todas las casas.

Mejora del beneficio

Además, se puede mejorar el beneficio de 119’25 € obtenidos en PSA, utilizando el método explicado en el apartado anterior de Naive Bayes, en el cual si se elige la casa que mejor paga en cada partido, el beneficio aumenta en este caso hasta los 148’31€.

Tabla 24 Mejora del beneficio (KNN)

Antes de la mejora	Despúes de la mejora
119.25	148.31

En los casos, donde se ha acertado el resultado, se ha contabilizado el numero de veces que se jugaría en cada casa. Teniendo en cuenta que la casa escogida seria la de la cuota mejor pagada.

Tabla 25 Número de veces que se apuesta a cada casa (KNN)

B365	BW	IW	LB	PS	WH	VC
77	123	281	120	378	355	224

Es decir, 77 veces se ha apostado y acertado el partido en B365, 123 en BW, 281 en IW…

5.4.- Árboles de Decisión(CART)

Los árboles de decisión son un método usado en distintas disciplinas como modelos de predicción. Estos son similares a diagramas de flujo, en los que llegamos a puntos en los que se toman decisiones de acuerdo a una regla.

De manera general, lo que hace este algoritmo es encontrar la variable independiente que mejor separa los datos en grupos, que corresponden con las categorías de la variable objetivo. Esta mejor separación es expresada con una regla. A cada regla corresponde un nodo.

Primero de todo, se aplica este método en R Studio y se obtiene un árbol completo con un gran número de nodos y ramificaciones .

Con la finalidad de reducir la varianza del modelo y así disminuir el *test error*, se somete al árbol a un proceso de *pruning(poda)*. Este proceso, intenta encontrar el árbol más sencillo (menor tamaño) que consigue explicar las observaciones.

Para ello se utilizará la tabla siguiente:

Tabla 26 Tabla utilizada para la elección del CP

Root node error: 1031/1900 = 0.54263

n= 1900

	CP	nsplit	rel error	xerror	xstd
1	0.1357905	0	1.00000	1.00000	0.021062
2	0.0179437	1	0.86421	0.87488	0.021112
3	0.0080828	3	0.82832	0.85257	0.021080
4	0.0058196	6	0.80407	0.85742	0.021088
5	0.0043647	7	0.79825	0.86130	0.021094
6	0.0041222	9	0.78952	0.86615	0.021101
7	0.0038797	16	0.75752	0.87973	0.021118
8	0.0033948	19	0.74588	0.87391	0.021111
9	0.0029098	27	0.71775	0.88070	0.021119
10	0.0026673	31	0.70417	0.89428	0.021130
11	0.0024248	35	0.69350	0.89525	0.021131
12	0.0019399	37	0.68865	0.90107	0.021134
13	0.0016166	51	0.66052	0.90592	0.021136
14	0.0015357	55	0.65373	0.90689	0.021137
15	0.0014549	84	0.59457	0.91465	0.021139
16	0.0012932	87	0.58972	0.95538	0.021125
17	0.0010000	98	0.57129	0.95829	0.021122

Utilizaremos los parámetros de complejidad (CP), como una penalización para controlar el tamaño del árbol. En resumen, cuanto mayor es el parámetro de complejidad, menos decisiones contiene el árbol (nsplit). El valor rel error representa la desviación media del árbol al que se refiera dividida entre la desviación media del árbol nulo (nsplit = 0). El valor xerror es el valor medio estimado mediante un procedimiento de *cross validation* y xstd es el error estándar del error relativo.

El error que tendremos que observar para la decisión de la poda es el xerror. Se escogerá aquel que tenga un xerror más pequeño. En este caso xerror=0.85257, que corresponde a un cp de 0.0080828 y un número de divisiones (nsplit) igual a 3.

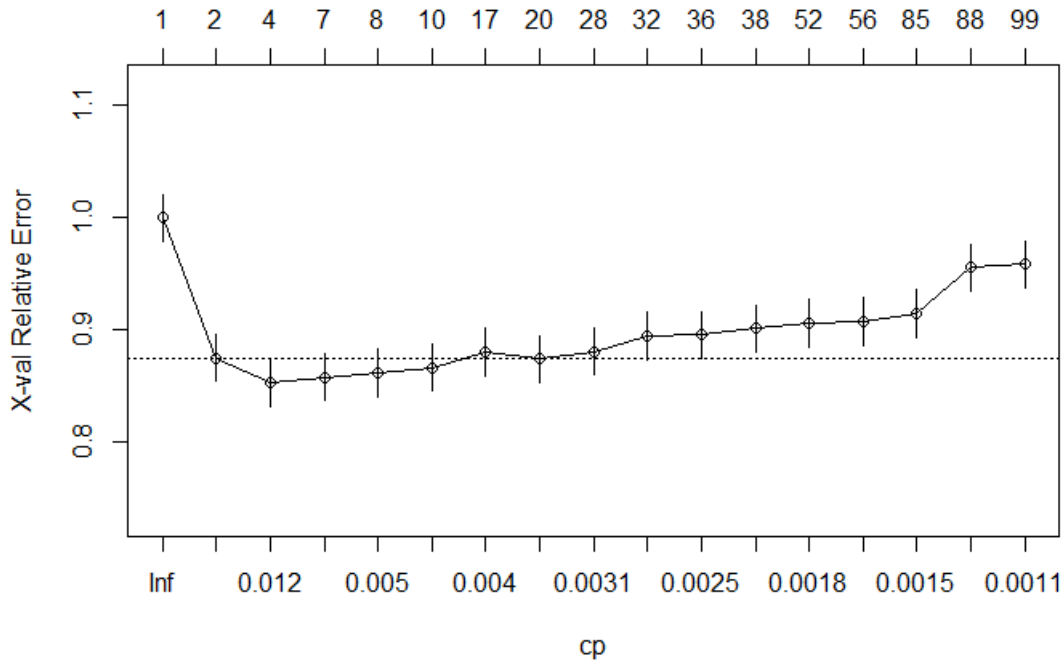


Ilustración 30 X-val error relativo según el CP

En el gráfico que se acaba de mostrar, se muestra en el eje X, los diferentes valores de CP y en el eje Y, el xerror. Se aprecia, como el xerror disminuye al principio y con valores CP más altos crece.

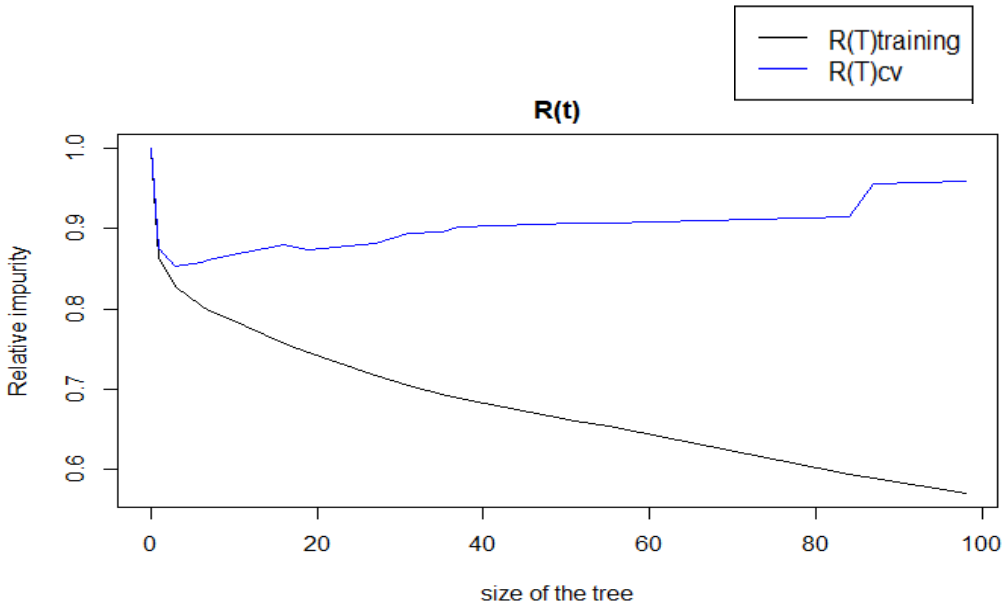


Ilustración 31 impureza según el tamaño del árbol

Este segundo gráfico, muestra, que cuando el tamaño del árbol es más grande, la impureza(error) para los datos de entrenamiento va disminuyendo progresivamente. Por otro lado, no ocurre los mismo en la validación cruzada(línea negra)

A continuación, se muestra el esquema del árbol de clasificación podado. Cada inciso, nos indica un nodo y la regla de clasificación que le corresponde. Siguiendo estos nodos, podemos llegar a las hojas del árbol, que corresponde a la clasificación de los datos.

```
n= 1900

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 1900 1031 H (0.4573684 0.2531579 0.2894737)
   2) PSA>=3.865 904  325 H (0.6404867 0.2267699 0.1327434) *
   3) PSA< 3.865 996  566 A (0.2911647 0.2771084 0.4317269)
     6) PSA>=2.175 680  438 H (0.3558824 0.3000000 0.3441176)
      12) PSA>=3.025 340  210 H (0.3823529 0.3441176 0.2735294) *
      13) PSA< 3.025 340  199 A (0.3294118 0.2558824 0.4147059) *
     7) PSA< 2.175 316  120 A (0.1518987 0.2278481 0.6202532) *
```

Ilustración 32 Esquema del árbol de clasificación podado

Todo lo anterior resulta mucho más claro si se visualiza, así que se creará una gráfica usando el modelo

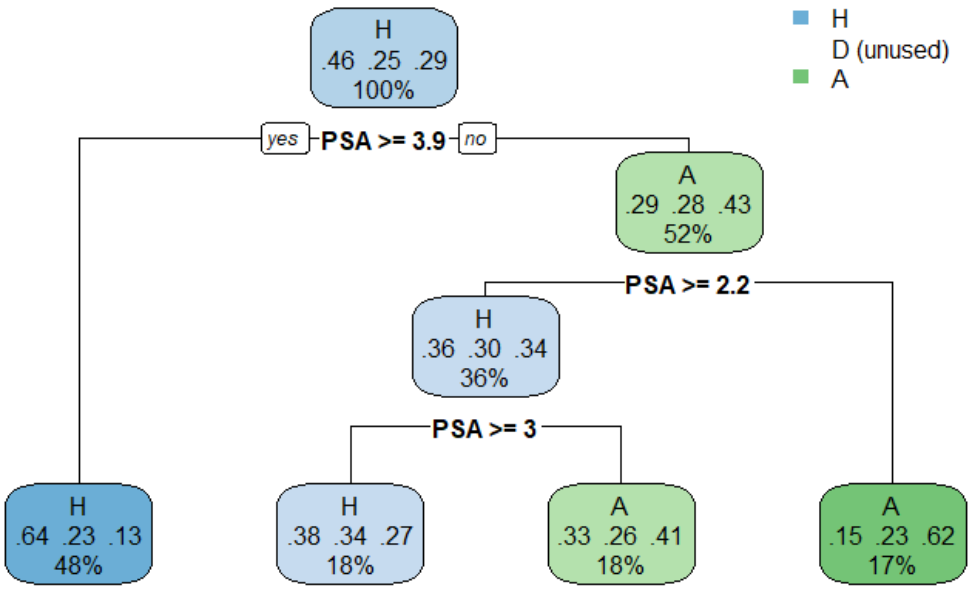


Ilustración 33 Árbol de decisión óptimo

En estos gráficos, cada uno de los rectángulos representa un nodo de nuestro árbol, con su regla de clasificación. Cada nodo está coloreado de acuerdo a la categoría mayoritaria entre los datos que agrupa. Esta es la categoría que ha predicho el modelo para ese grupo. Arriba a la derecha, aparece D (unused), esto quiere decir, que no se van a predecir empates, únicamente victorias del local o visitante.

Primero de todo, se parte del nodo inicial (el de arriba de todo), este contiene 3 valores 0'46, 0'25 y 0'29, que hacen referencia a proporción de casos pertenecen a cada categoría. Es decir, en un primer momento sin hacer ninguna división aún, se tiene unl 46% de casos en la predicción de gana el local(H), 25% en empates(D) y 29% en victorias del visitante(A). Además, aparece un 100%, que indica la proporción del total de datos que han sido agrupados allí. Es del 100% ya que no se ha divido de momento.

A continuación, aparece PSA >3'9.La variable predictora PSA, corresponde a las cuotas de la casa de apuestas Pinnacle, que ofrece las cuotas que se pagan si se apuesta a que gana el visitante.

Esta es la variable que tiene una mayor importancia, ya que es la que divide en un primer momento los datos. Para cada uno de los partidos, se observara el valor de esta variable, y en el caso que sea mayor o igual a 3`9, la rama del gráfico irá por la izquierda, en otro caso por la derecha.

Los datos son separados en grupos a partir de la regla obtenida. Después, para cada uno de los grupos resultantes, se repite el mismo proceso. Se realizará de manera recursiva hasta que es imposible obtener una mejor separación. Cuando esto ocurre, el algoritmo se detiene. Cuando un grupo no puede ser partido mejor, se le llama nodo terminal u hoja.

Una característica muy importante en este algoritmo es que una vez que alguna variable ha sido elegida para separar los datos, ya no es usada de nuevo en los grupos que ha creado. Se buscan variables distintas que mejoren la separación de los datos.

Las **principales ventajas** de este método son su interpretabilidad, pues nos da un conjunto de reglas a partir de las cuales se pueden tomar decisiones. Este es un algoritmo que no es demandante en poder de cómputo comparado con procedimientos más sofisticados y, a pesar de ello, que tiende a dar buenos resultados de predicción para muchos tipos de datos.

Sus **principales desventajas** son que este en tipo de clasificación "débil", pues sus resultados pueden variar mucho dependiendo de la muestra de datos usados para entrenar un modelo. Además es fácil sobre ajustar los modelos, esto es, hacerlos excelentes para clasificar datos que conocemos, pero deficientes para datos conocidos.

En último lugar, se añadirá, que el rectángulo en el extremo inferior izquierdo de la gráfica tiene 64% de casos en la predicción de gana el local(H), 23% en empates(D) y 13% en victorias del visitante(A). Estas proporciones nos dan una idea de la precisión de nuestro modelo al hacer predicciones. En este nodo, se elige H, ya que contiene una proporción de H, más grande respecto del resto. De este modo, las reglas que conducen al rectángulo que acabamos de mencionar nos dan un 64% de clasificaciones correctas

Precisión

Tabla 27 Precisión (CART)

Aciertos	%
1046/1900	55’05

Matriz de confusión

	H	D	A
H	709	322	213
D	0	0	0
A	160	159	337

Hay que comentar como curiosidad, que el método de árboles de decisión no ha predicho ningún empate. No obstante, teniendo en cuenta la dificultad que supone acertar correctamente el resultado del partido en las apuestas deportiva, tiene el porcentaje de acierto más alto respecto el resto e clasificadores con un 55’05%.

Retorno neto

Tabla 28 Retorno según la casa de apuestas (CART)

Casa de apuestas	retorno
B365	4.11
BW	11.93
IW	8.58
LB	2.87
PS	48.85
WH	41.23
VC	37.67

Todas las casas de apuestas, presentan un retorno neto positivo,empleando el método de arboles de decisión. De nuevo PS(Pinnacle), se mantiene como cabeza de serie en la oferta de cuotas como casa de apuesta, ya que su retorno neto es el mayor.

10 primeros encuentros

Tabla 29 Variación del beneficio o pérdida por jornada (CART)

Jornada	B365	BW	IW	LB	PS	WH	VC
1	1.05	1.05	1.00	1.15	1.10	1.10	1.10
2	1.35	1.33	1.35	1.45	1.43	1.43	1.43
3	0.35	0.33	0.35	0.45	0.43	0.43	0.43
4	1.45	1.43	1.55	1.55	1.66	1.58	1.63
5	2.40	2.43	2.55	2.55	2.69	2.68	2.68
6	1.40	1.43	1.55	1.55	1.69	1.68	1.68
7	2.78	2.83	2.95	2.85	3.19	3.18	3.18
8	3.58	3.53	3.70	3.55	4.03	3.98	4.01
9	2.58	2.53	2.70	2.55	3.03	2.98	3.01
10	1.58	1.53	1.70	1.55	2.03	1.98	2.01

A diferencia del método anterior KNN, en la jornada 10 el benefició no es tan alto. No obstante, el gráfico ayudará a visualizar el beneficio o perdida que se tiene en cada momento, en cada uno de los 1900 partidos.

Gráfico

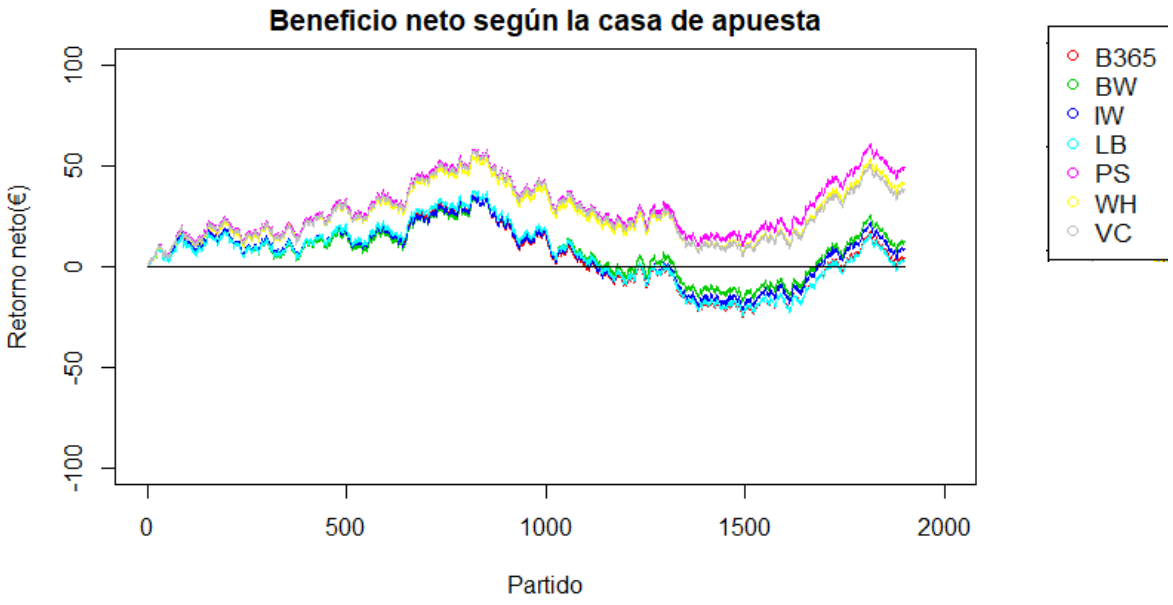


Ilustración 34 Variación de Beneficio neto según la casa de apuestas y jornada (CART)

Se aprecia como el movimiento de las 7 casas de apuestas en los 1900 encuentros es de una manera similiar a KNN.Es cierto que en este caso, hay 4 casas de apuestas que pasan por un periodo de pérdidas pero todas ellas acaban con retorno positivo.

Mejora del beneficio

Tabla 30 Mejora del beneficio (CART)

Antes de la mejora	Despúes de la mejora
48.85	81.78

Hasta 81’78€ se pueden alcanzar utilizando el método de mejora.

Por otro lado WH, es la casa de apuestas en la que más veces se apostaría y se acertaría el resultado. Esta casa corresponde a la segunda mejor en retorno positivo final después de PSA.

Tabla 31 Número de veces que se apuesta a cada casa (NB)

B365	BW	IW	LB	PS	WH	VC
69	144	314	126	363	416	245

5.5- QDA

Información a priori

La siguiente técnica de predicción es la del análisis de discriminación. Se trata de una generalización de discriminación de Fisher que pretende caracterizar o bien separar las clases. Es decir, utilizando esta técnica se quieren encontrar combinaciones de las variables que mejor expliquen la base de datos.

A la hora de elegir entre análisis de discriminación lineal versus análisis de discriminación cuadrático, se ha tenido en cuenta la siguiente condición:

"Cada predictor que forma parte del modelo se distribuye de forma normal en cada una de las clases de la variable respuesta. En el caso de múltiples predictores, las observaciones siguen una distribución normal multivariante en todas las clases."

Por lo tanto, como se ha detectado que hay variables que no siguen la distribución normal, el conjunto que formará con otras variables, no seguirá una distribución multivariante. Por este motivo, no se puede aplicar el análisis de discriminación lineal y se opta por su versión cuadrática.

La tabla que se presenta a continuación, muestra la probabilidad a priori de los diferentes niveles de la variable respuesta y la media de la variables predictoras.

Tabla 32 Resultados a priori de QDA

```
Call:
qda(FTR ~ ., data = datos[, -c(1, 2, 3, 4)], prior = apriori)

Prior probabilities of groups:
        H         D         A
0.4573684 0.2531579 0.2894737

Group means:

   B365H   B365D   B365A     BWH     BWD     BWA     IWH     IWD     IWA     LBH     LBD
H 2.055017 4.038032 5.900840 2.050173 4.006778 5.685086 2.023406 3.881853 5.219321 2.042785 3.977204
D 2.566570 3.627692 4.190873 2.537360 3.599834 4.087630 2.451040 3.541601 3.823119 2.525655 3.597796
A 3.573109 3.616309 3.017182 3.537545 3.590055 2.979800 3.307291 3.520909 2.879636 3.456291 3.586545
      LBA     PSH     PSD     PSA     WHH     WHD     WHA     VCH     VCD     VCA
H 5.718400 2.123049 4.244419 6.543211 2.085190 3.852451 5.789367 2.114154 4.121392 6.406617
D 4.072765 2.676008 3.769792 4.512412 2.577672 3.484615 4.115572 2.649459 3.665405 4.466507
A 2.968418 3.846964 3.753836 3.193545 3.549673 3.480127 3.026400 3.775545 3.642982 3.168818
```

Precisión

Tabla 33 Precisión (QDA)

Aciertos	%
882/1900	46’42

Matriz de confusión

	H	D	A
H	367	109	56
D	253	135	114
A	249	237	380

El método de QDA, aunque tiene un porcentaje de acierto bastante bajo (46’42%), es uno de los métodos que más empates llega a predecir correctamente. Por otro lado, en comparación con el resto de clasificadores,predice pocas victorias del local y del visitante correctamente.

Retorno neto

Tabla 34 Retorno según la casa de apuestas (QDA)

Casa de apuestas	retorno
B365	-84.11
BW	-89.26
IW	-98.98
LB	-91.77
PS	-33.30
WH	-79.11
VC	-55.60

Todas las casas de apuestas presentan una perdida considerable. Por lo tanto, antes de realizar el caso real de Validación, en el cual se separan los datos en test y train, se puede decir que a priori no parece un buen método.

10 primeros encuentros

Tabla 35 Variación del beneficio o pérdida por jornada (QDA)

Jornada	B365	BW	IW	LB	PS	WH	VC
1	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
2	-0.70	-0.72	-0.65	-0.70	-0.67	-0.67	-0.67
3	1.30	1.23	0.95	0.80	1.67	1.21	1.58
4	0.30	0.23	-0.05	-0.20	0.67	0.21	0.58
5	-0.70	-0.77	-1.05	-1.20	-0.33	-0.79	-0.42
10	-1.52	-1.67	-1.90	-2.20	-1.00	-1.49	-1.09

Se puede apreciar como las perdidas están presentes desde las primeras jornadas

Gráfico

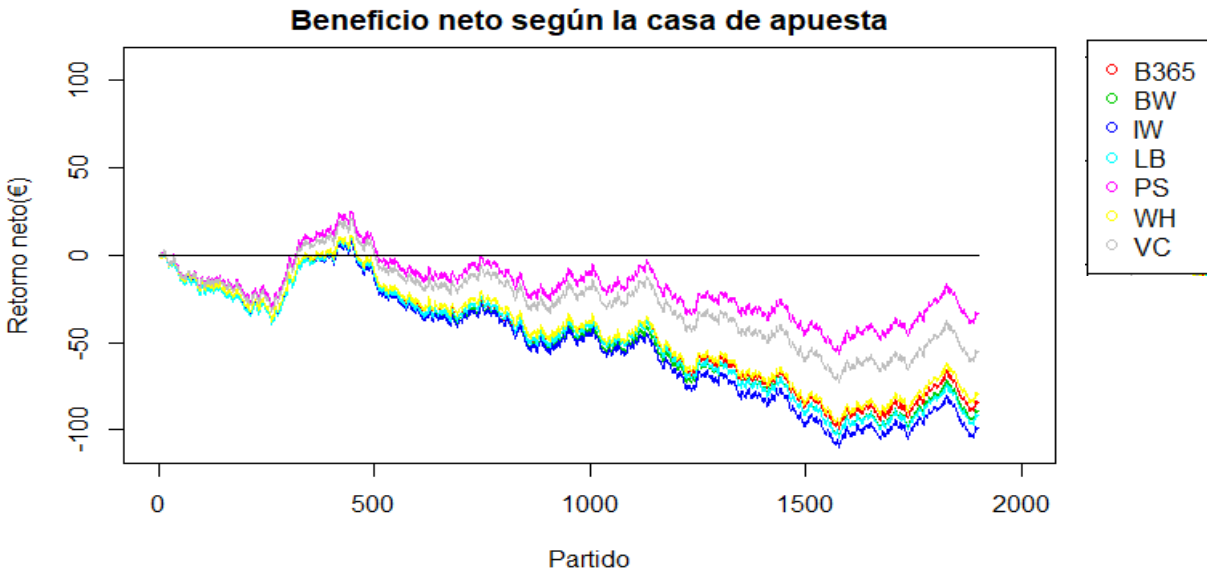


Ilustración 35 Variación de Beneficio neto según la casa de apuestas y jornada (QDA)

Aunque hay un instante en que todas las casas llegan a situarse en un retorno positivo, lo cierto es que se mantienen casi en todo momento en pérdidas.

Mejora del beneficio

Tabla 36 Mejora del beneficio (QDA)

Antes de la mejora	Despúes de la mejora
-33’30	-10.24

Teniendo en cuenta el método de la mejora del beneficio, en el cual se apuesta a las casas con cuotas más altas, la perdida más baja que se ha obtenido apostando en PSA(-33’30), se puede reducir hasta los -10.24€.

El número de veces que se apostará en cada casa es:

Tabla 37 Número de veces que se apuesta a cada casa (QDA)

B365	BW	IW	LB	PS	WH	VC
70	106	225	107	374	279	192

5.6- SVM

Información a priori

En último lugar, se mostrará la información a priori para el método de maquinas de soporte vectorial.

El mejor kernel utilizado es el radial, con un coste de 1 y un parámetro gamma de *0.04761905.*

Además el número de vectores soporte es de 1565

Tabla 38 Información a priori (SVM)

```
Call:
svm.default(x = x, y = y)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.04761905

Number of Support Vectors:  1565

 ( 620 464 481 )


Number of Classes:  3

Levels:
 H D A
```

Precisión

Tabla 39 Precisión (SVM)

Aciertos	%
1042/1900	54’84

Matriz de confusión

	H	D	A
H	750	348	260
D	0	2	0
A	119	131	290

SVM(Support Vector Machines), predice correctamente un gran número de partidos donde gana el local, pero no es nada preciso en los empates.

En cuanto a las apuestas de la victoria para el equipo visitante, se encuentra por debajo de la mayoría de métodos estudiados.

Retorno neto

Tabla 40 Retorno según la casa de apuestas (SVM)

Casa de apuestas	retorno
B365	-15.03
BW	-5.95
IW	-10.61
LB	-16.95
PS	30.26
WH	22.34
VC	18.36

Se observa también en SVM, como PS(Pinnacle) es la mejor casa de apuestas con diferencia para todos los métodos.

No obstante, utilizando SVM (Máquinas de vector Soporte), hay 4 casas que producen perdidas

10 primeros encuentros

Tabla 41 Variación del beneficio o pérdida por jornada (SVM)

Jornada	B365	BW	IW	LB	PS	WH	VC
1	1.05	1.05	1.00	1.15	1.10	1.10	1.10
2	1.35	1.33	1.35	1.45	1.43	1.43	1.43
3	0.35	0.33	0.35	0.45	0.43	0.43	0.43
4	1.45	1.43	1.55	1.55	1.66	1.58	1.63
5	2.40	2.43	2.55	2.55	2.69	2.68	2.68
6	1.40	1.43	1.55	1.55	1.69	1.68	1.68
7	2.78	2.83	2.95	2.85	3.19	3.18	3.18
8	3.58	3.53	3.70	3.55	4.03	3.98	4.01
9	2.58	2.53	2.70	2.55	3.03	2.98	3.01
10	1.58	1.53	1.70	1.55	2.03	1.98	2.01

En la tabla se aprecia como en la jornada 10, las casas de apuestas PS,WH y VC, ya van marcando diferencias respecto el resto

Gráfico

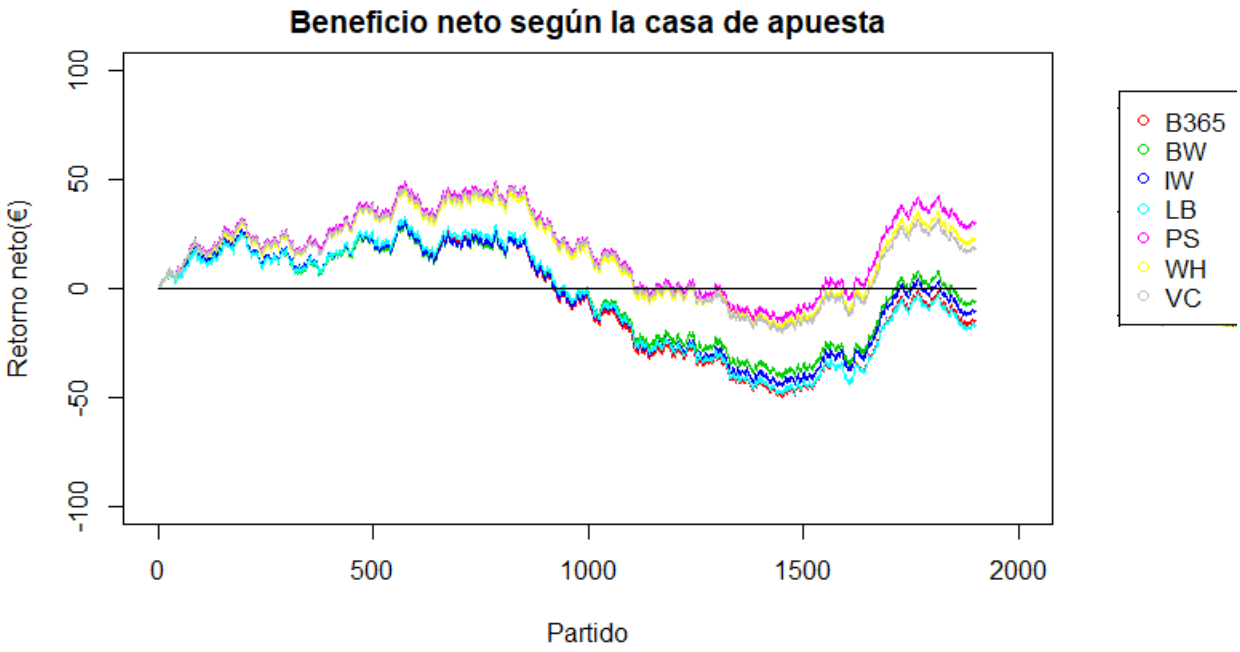


Ilustración 36 Variación de Beneficio neto según la casa de apuestas y jornada (SVM)

Claramente queda representado, que hasta el partido 900, las 7 casas de apuestas se mantenían con un retorno positivo para el jugador. En cambio, una vez llegado a ese momento, las 7 casas de apuestas pasan por momentos donde el beneficio es negativo. Una vez se llega a la jornada 1400, donde todas las casas de apuestas se encontraban en retornos negativos para el jugador, cuando se llega al final del estudio, en la jornada 1900, tres de ellas(PS,WH Y VC) acaban con beneficio y el resto con pérdidas.

Mejora del beneficio

El beneficio pasa de los 30’26 de la casa Pinnacle hasta los 62’12 apostando a diferentes casas de apuestas. Es decir, se dobla la ganancia utilizando este método.

Tabla 42 Mejora del beneficio (SVM)

Antes de la mejora	Despúes de la mejora
30.26	62.12

El número de veces que se apostará en cada casa es:

Tabla 43 Número de veces que se apuesta a cada casa (SVM)

B365	BW	IW	LB	PS	WH	VC
68	148	313	125	365	416	241

5.7.- Otros métodos : Cuota mínima (MIN)

En último lugar se va explicar el método de la cuota mínima. A diferencia del resto de métodos, este no aparece como un clasificador más dentro de la minería de datos, sino que únicamente es una opción a tener en cuenta en el ámbito de las apuestas. El método consiste en el siguiente: Se escoge de cada partido, la cuota mínima, y se apuesta a ese evento. Por ejemplo, si las cuotas en el partido Milán vs Parma fueran las siguientes:

Tabla 44 Cuotas para el partido Milán vs Parma

Milán	Empate	Parma
1.33	5.75	6.91

En este caso se apostaría a la victoria del Milán, ya que la cuota es la más baja. Como se ha comentado anteriormente, una cuota más baja, significa que la probabilidad de acertar es más alta. Esto es debido a que la inversa de la cuota es aproximadamente la probabilidad de acertar el evento, teniendo en cuenta el margen de beneficio de la casa de apuestas. En este caso, aplicando el método de la cuota mínima lo que se va a hacer, es apostar a favor del suceso más probable.

Este método, resulta uno de los más jugados por miles de personas, debido a que es un método que tiene bastante sentido a priori. Apostar por el resultado más probable siempre da tranquilidad al apostador. El problema principal de este método es que el retorno en caso de acertar no es muy alto.

En el ejemplo anterior el retorno neto por euro jugado es de 1.33-1=0.33€.

A continuación se va mostrar el resultado de aplicar este método en la calibración:

En primer lugar se va a presentar el número de aciertos, su porcentaje y la matriz de confusión

Precisión

Tabla 45 Precisión (MIN)

Aciertos	%
1043/1900	54’89

Dentro de los métodos estudiados, es de lo que más acierta, después de CART.

Matriz de confusión

	H	D	A
H	735	335	245
D	1	4	1
A	133	142	304

En la matriz de confusión se puede observar que se han acertado correctamente 735 apuestas a que gana el local, 4 empates y 304 apuestas a que gana el visitante.

Si observamos las 3 columnas, que hacen referencia al verdadero valor, se observa que la suma de 735 + 1 +133=869, este valor hace referencia al verdadero numero de veces que ha ganado el local de los 1900. Por otro lado 335 + 4 + 142 =481 es el número de empates y 245 + 1 +304=550 es el numero de victorias del visitante.

Por lo tanto, el valor 133 de la primera columna quiere decir que 133 veces de las 1900 se ha predicho que ganaría el visitante, cuando realmente lo ha hecho el local, por lo tanto es un error en nuestra predicción. De la misma manera el valor 335 de la segunda columna, quiere decir que 335 veces se ha predicho que gana el local, cuando realmente el resultado final ha acabado en empate. Por último, comentar que sumando la primera fila 735 + 335 + 245=1315, se aprecia que de los 1900 encuentros, se ha apostado 1315 veces a que gana el local, ya que 1315 veces, la cuota más baja correspondía con la victoria del equipo que jugaba en casa

Retorno neto

Tabla 46 Retorno según la casa de apuestas (MIN)

Casa de apuestas	Retorno
B365	-14.58
BW	-5.40
IW	-8.71
LB	-15.48
PS	31.14
WH	23.28
VC	19.44

En este último caso, se aprecia como 3 casas de apuestas (PS,WH y VC) devuelven un retorno neto positivo. Por otro lado, las otras 4 casas devuelven un retorno neto negativo, siendo LB la más desfavorecida.

10 primeros encuentros

Tabla 47 Variación del beneficio o pérdida por jornada (MIN)

Jornada	B365	BW	IW	LB	PS	WH	VC
1	1.05	1.05	1.00	1.15	1.10	1.10	1.10
2	1.35	1.33	1.35	1.45	1.43	1.43	1.43
3	0.35	0.33	0.35	0.45	0.43	0.43	0.43
4	1.45	1.43	1.55	1.55	1.66	1.58	1.63
5	2.40	2.43	2.55	2.55	2.69	2.68	2.68
6	1.40	1.43	1.55	1.55	1.69	1.68	1.68
7	2.78	2.83	2.95	2.85	3.19	3.18	3.18
8	3.58	3.53	3.70	3.55	4.03	3.98	4.01
9	2.58	2.53	2.70	2.55	3.03	2.98	3.01
10	1.58	1.53	1.70	1.55	2.03	1.98	2.01

Gráfico

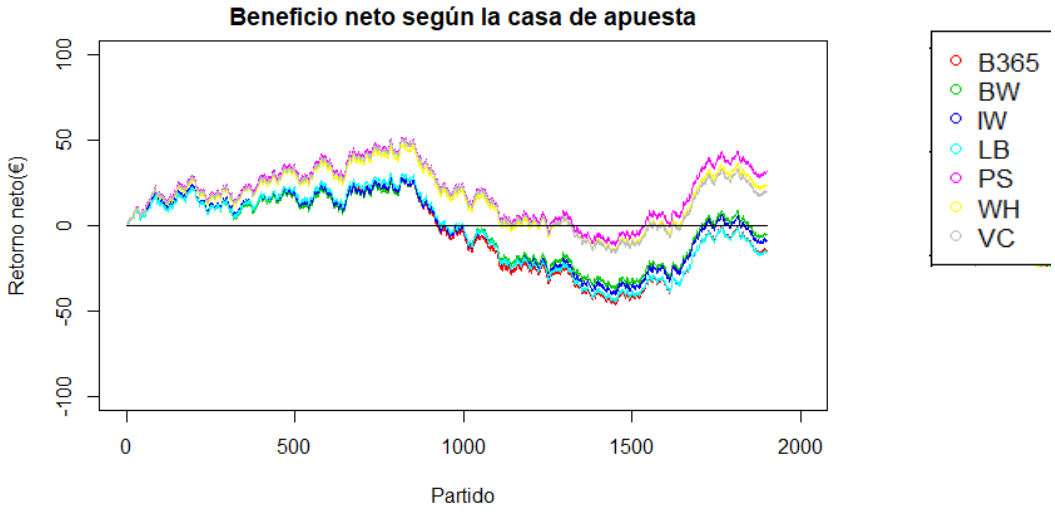


Ilustración 37 Variación de Beneficio neto según la casa de apuestas y jornada (MIN)

La principal característica de este gráfico, es que se distribuye de una manera muy parecida al clasificador SVM. Por lo tanto, se puede entender que en este trabajo, SVM, a la hora de clasificar, utiliza la técnica de la elección de la cuota mínima.

Mejora del beneficio

Apreciando anteriormente la tabla de retorno neto según la casa de apuestas, se ha visto que la mejor casa para apostar utilizando el método de la cuota mínima es PS, con un beneficio de 31.14€.

Tabla 48 Mejora del beneficio (MIN)

Antes de la mejora	Despúes de la mejora
31.14€.	62.7€

El número de veces que se apostará en cada casa es:

Tabla 49 Número de veces que se apuesta a cada casa (MIN)

B365	BW	IW	LB	PS	WH	VC
65	145	314	127	366	415	242

Sorprendentemente, la casa en la cual más veces se apostaría es WH seguida de PS. Una posible pregunta es por que PS devuelve el retorno positivo más alto, si realmente se apuesta más veces a WH. Esto es debido, básicamente a que la casa de apuestas PS para unos ciertos encuentros, presenta unas cuotas muy elevadas en comparación con el resto de casas, hasta el punto de llegar a convertirse en la casa mejor pagada.

5.8.- Resumen de las ganancias según la casa de apuestas y método de clasificación usado

Tabla 50 Porcentaje de acierto y retorno neto según el método y casa de apuesta

Metodo	% Acierto	B365	BW	IW	LB	PS	WH	VC
NV	45’63	-20.94	-27.20	-37.38	-29.39	27.99	-29.92	-2.05
KNN	52’8	68.52	65.91	59.21	62.48	119.25	85.44	100.09
CART	55’05	4.11	11.93	8.58	2.87	48.85	41.23	37.67
QDA	46’42	-84.11	-89.26	-98.98	-91.77	-33.30	-79.11	-55.60
SVM	54’84	-15.03	-5.95	-10.61	-16.95	30.26	22.34	18.36
MIN	54’89	-14.58	-5.40	-8.71	-15.48	31.14	23.28	19.44

En primer lugar se aprecia que en 21 de las 42 posibles combinaciones se tiene un beneficio y en las otras 21 una pérdida. Además, hay que añadir que los clasificadores KNN y CART dan resultados positivos en el beneficio, independientemente de la casa de apuestas elegida. No obstante, las diferencias entre casas son relevantes, ya que se obtiene un beneficio superior en PS (Pinnacle) que en el resto de casas posibles.

La mejor combinación posible, es utilizar el clasificador KNN con un parametro de k=10 y escoger como casa de apuestas PS. El beneficio neto seria de 119’25€

Por otro lado, la peor combinación, seria utilizar el clasificador QDA para la casa de apuestas IW. La pérdida seria de 98’98€

5.9.- Gráficos según la casa de apuesta

A continuación se va a mostrar una gráfica para casa de apuestas, teniendo en cuenta el clasificador. Es el mismo estudio que el anterior pero ahora se comparan clasificadores dentro de cada casa.

El resumen básicamente es el siguiente:

-El método KNN aparece el primero en todos lo casos y QDA el último.

- Para todas la casa de apuestas, el método SVM y MIN siguen la misma distribución de retorno neto, debido a que SVM clasifica de la misma manera que el método e la cuota mínima.

- Únicamente PS, tiene 5 de los 6 métodos con retorno positivo

-Las gráficas de B365,BW,IW y LB son prácticamente las mismas, con pequeñas diferencia si se observa la parte final de la gráfica.

B365

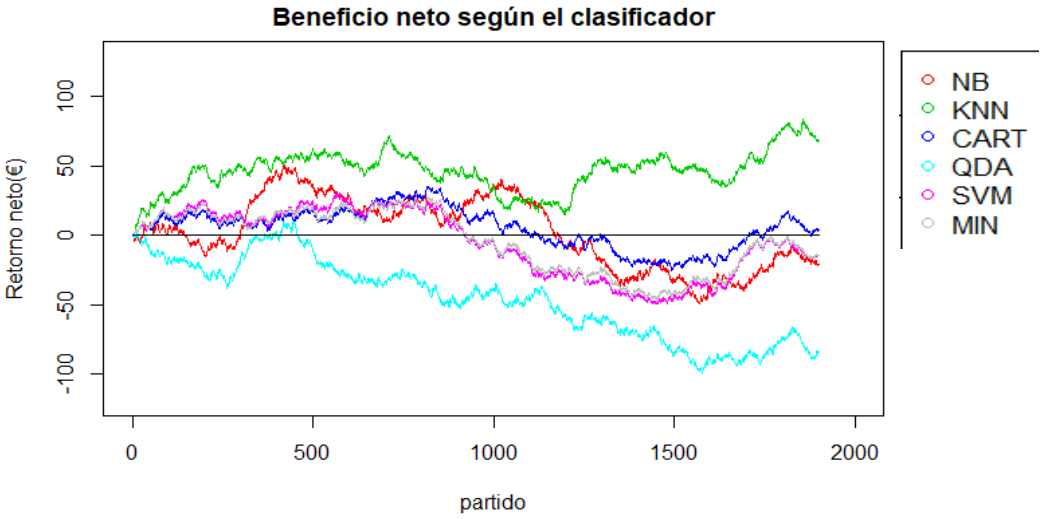


Ilustración 38 Retorno para B365 según el clasificador

BW

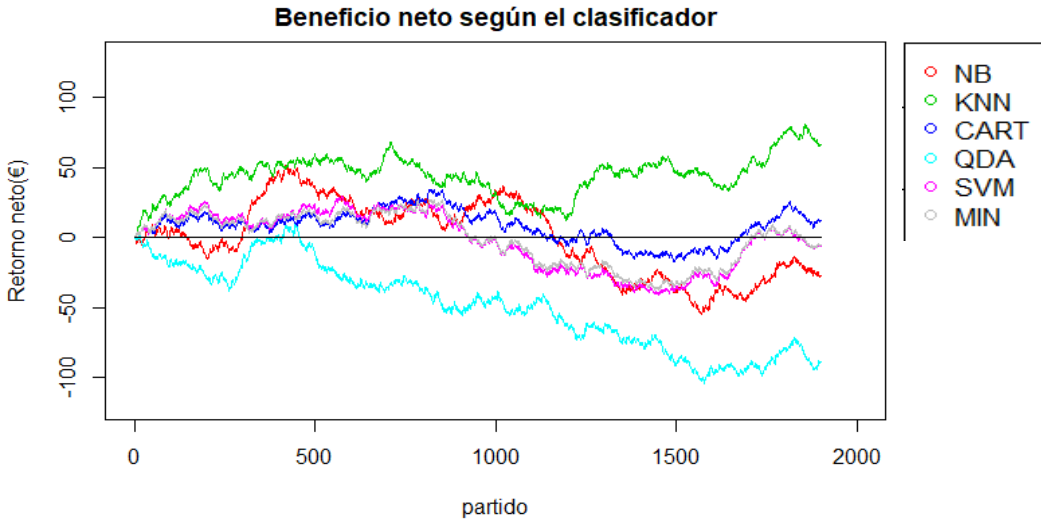


Ilustración 39 Retorno para BW según el clasificador

IW

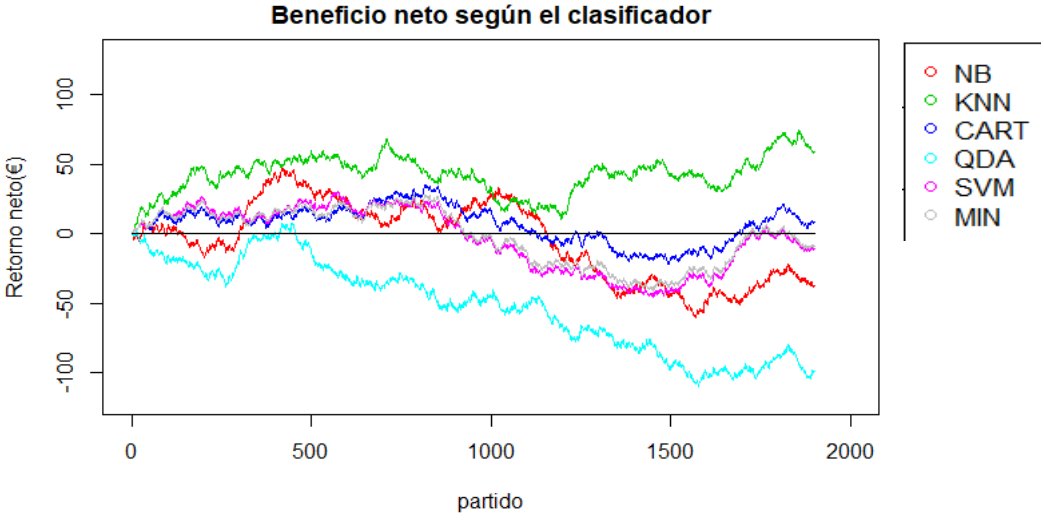


Ilustración 40 Retorno para IW según el clasificador

LB

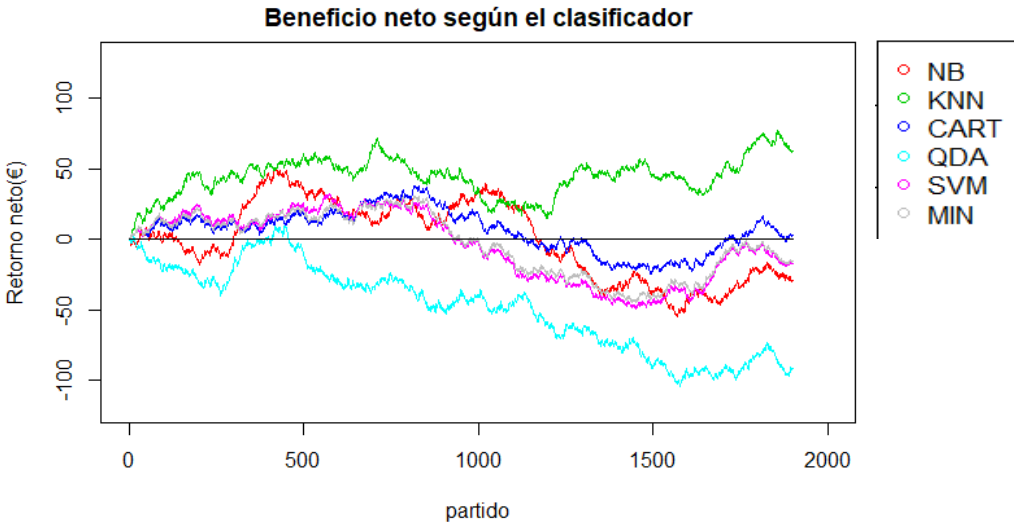


Ilustración 41 Retorno para LB según el clasificador

PS

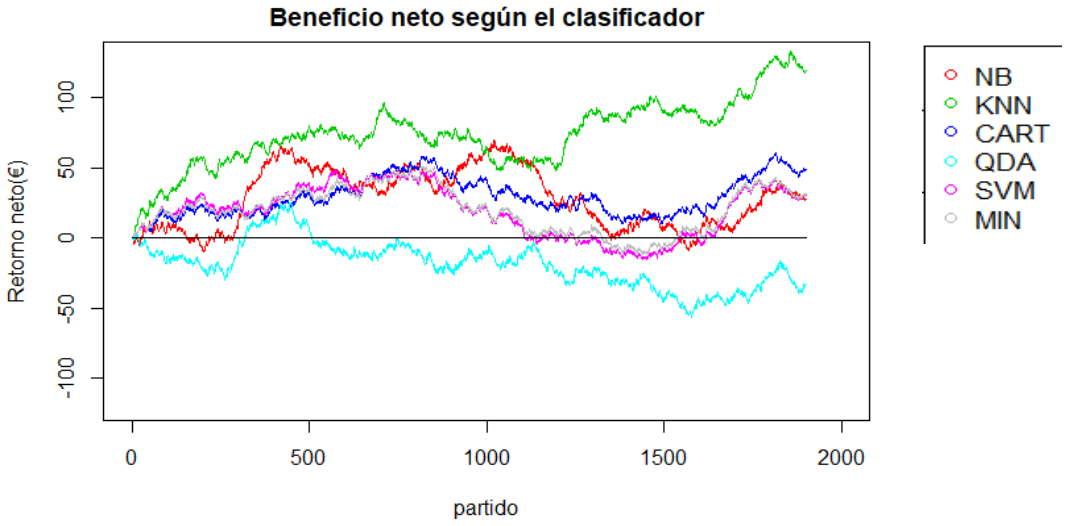


Ilustración 42 Retorno para PS según el clasificador

WH

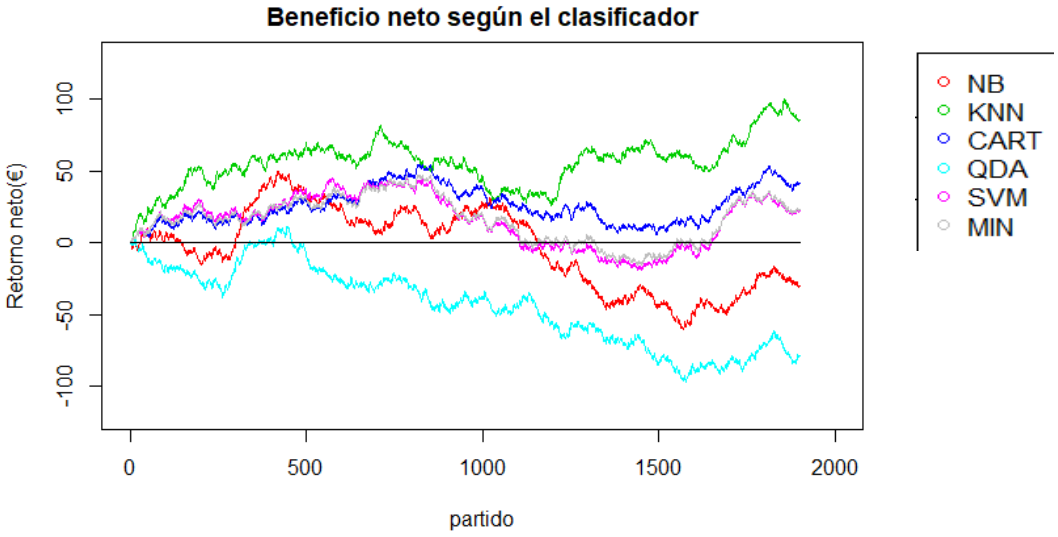


Ilustración 43 Retorno para WH según el clasificador

VC

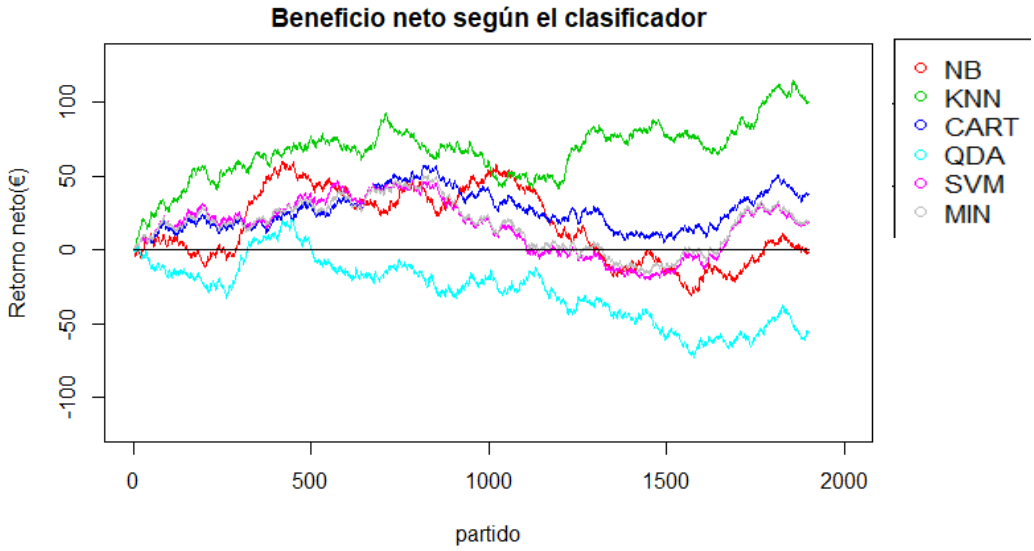


Ilustración 44 Retorno para VC según el clasificador

6.- Validación cruzada(Crossvalidation)

6.1.- Introducción

La **validación cruzada** o **cross-validation** es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevará a cabo a la práctica

Este proceso consiste en dividir en dos conjuntos complementarios los datos de muestra, realizar el análisis de un subconjunto (denominado datos de entrenamiento o *training set*), y validar el análisis en el otro subconjunto (denominado datos de prueba o *test set*), de forma que la función de aproximación sólo se ajusta con el conjunto de datos de entrenamiento y a partir de aquí calcula los valores de salida para el conjunto de datos de prueba (valores que no ha analizado antes).

El objetivo de la validación cruzada, consiste en estimar el nivel de ajuste de un modelo a un cierto conjunto de datos de prueba independientes de las utilizadas para entrenar el modelo.

6.2.- Aplicación

A continuación se va a mostrar para 5 métodos de Minería de datos (NV,KNN,CART.SVM Y QDA) los diferentes valores que se toman (% de acierto), si se tiene en cuenta 9 partes o pliegues (folds) como datos de entrenamiento y una parte (fold) como datos de test (Validación). En el caso a estudiar, como hay 1900 partidos, el 90% formaran parte de los datos de entrenamiento y el 10% de los datos de prueba. Los partidos que formaran parte de los datos de entrenamiento y de test en cada una de las 10 iteraciones serán escogidos al azar.

En este resultado, se va proporcionar la validación cruzada dentro de la muestra(in-sample) y fuera de la muestra (out-sample). El segundo caso es el que realmente interesa.

Para ello, se calculará la media para las dos series de valores.

cv-in : proceso interno de validación cruzada

cv-out : proceso externo de validación cruzada

Los valores que se presentaran a continuación, están en tanto por uno, por lo tanto el Fold 1 de NV en cv-out (crossvalidation out sample)= 0.5368, quiere decir que si se dividen los 1900 datos en 10 partes (190 datos por parte), y la parte 1 son los datos test y las partes 2,3,4,5,6,7,8,9 y 10 son los datos de entrenamiento, el porcentaje de acierto en la parte de test será del 53'68%. Por otro lado el valor 0.4491, quiere decir que se ha aplicado la predicción en la parte de entrenamiento (parte 2,3,4,5,6,7,8,9 y 10) en vez de la parte test (parte1) y su acierto es del 44'91%

4.5.1- Naive-Bayes

Tabla 51 Precisiones (%) de los diferentes Folds (NB)

	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
cv-in	0.4491	0.4491	0.4295	0.4561	0.4614	0.4588	0.4590	0.4672	0.4556	0.4672
cv-out	0.5368	0.4947	0.4346	0.4053	0.4158	0.4180	0.4263	0.4158	0.4947	0.5105

Tabla 52 Media de todos los Folds (NB)

	Media
cv-in	0.4553
cv-out	0.4552

4.5.2.- KNN

Tabla 53 Precisiones (%) de los diferentes Folds (KNN)

	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
cv-in	0.5579	0.5550	0.5535	0.5573	0.5637	0.5476	0.5468	0.5591	0.5550	0.5614
cv-out	0.5421	0.5474	0.5497	0.5316	0.5211	0.5503	0.5526	0.5211	0.5474	0.5211

Tabla 54 Media de todos los Folds (KNN)

	Media
cv-in	0.5557
cv-out	0.5384

4.5.3.- Decision Trees

Tabla 55 Precisiones (%) de los diferentes Folds (CART)

	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
cv-in	0.5292	0.5637	0.5477	0.5520	0.5363	0.5523	0.5322	0.5550	0.5730	0.5544
cv-out	0.5473	0.5526	0.5654	0.5052	0.4842	0.5344	0.5210	0.5316	0.5474	0.4947

Tabla 56 Media de todos los Folds (CART)

	Media
cv-in	0.5496
cv-out	0.5284

4.5.4- QDA

Tabla 57 Probabilidad a priori de la variable respuesta

	H	D	A
A priori	0.4574	0.2532	0.2895

Tabla 58 Precisiones (%) de los diferentes Folds (QDA)

	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
cv-in	0.4781	0.4649	0.4696	0.4743	0.4766	0.4702	0.4561	0.4760	0.4626	0.4775
cv-out	0.4579	0.4737	0.5131	0.4737	0.4526	0.4603	0.4842	0.4421	0.4368	0.4421

Tabla 59 Media de todos los Folds (QDA)

	Media
cv-in	0.4706
cv-out	0.4637

4.5.5- SVM

Tabla 60 Precisiones (%) de los diferentes Folds (SVM)

	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
cv-in	0.5526	0.5450	0.5400	0.5421	0.5509	0.5523	0.5497	0.5497	0.5491	0.5474
cv-out	0.5000	0.5737	0.6178	0.5947	0.5316	0.5132	0.5263	0.5263	0.5368	0.5316

Tabla 61 Media de todos los Folds (SVM)

	Media
cv-in	0.5479
cv-out	0.5452

6.3.- Resumen

In_sample

El clasificador con un acierto más alto en la media del estudio de la validación cruzada in sample, corresponde a KNN.

Tabla 62 Medias de los Folds para cada clasificador

	Media CV-IN
NB	0.4553
KNN	0.5557
CART	0.5496
QDA	0.4706
SVM	0.5479

Observando el gráfico de boxplots (diagramas de caja) se observa como la variabilidad es muy parecida en todos los clasificadores, pero el porcentaje de acierto es más alto en KNN,CART Y SVM respecto NB y QDA.

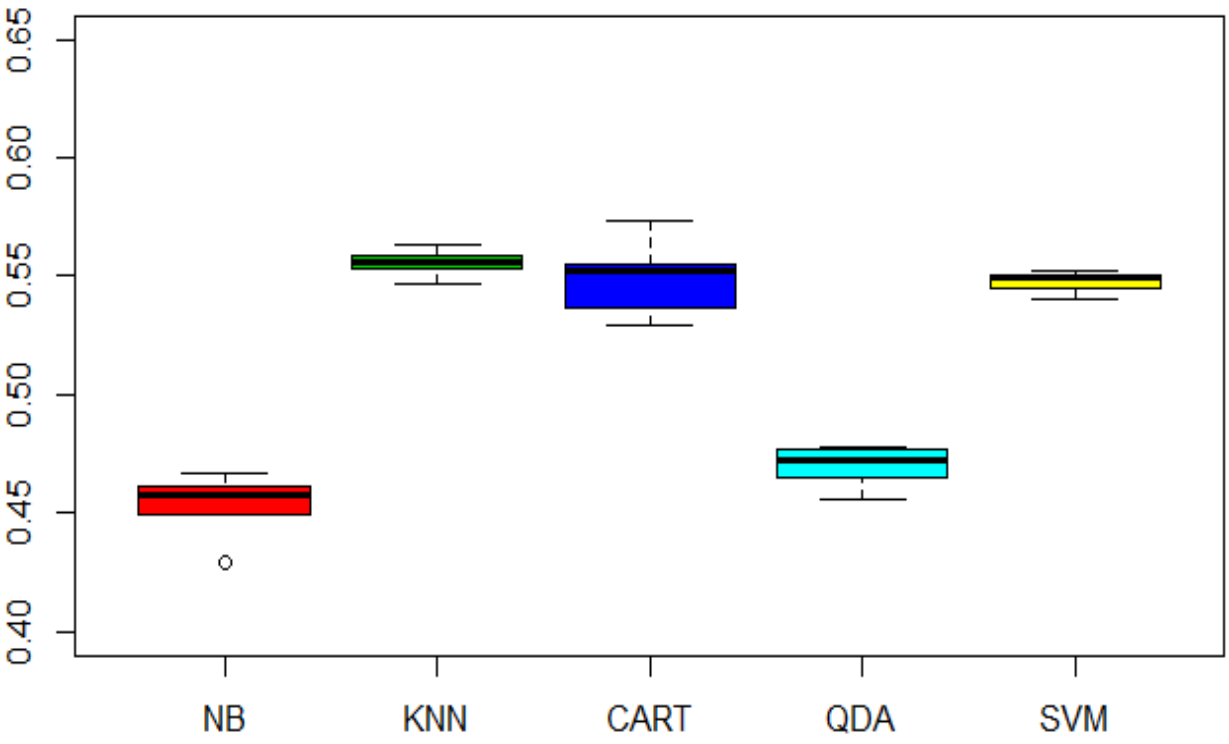


Ilustración 45 Boxplots para cada uno de los clasificadores (In-sample)

Out_sample

Por otro lado, en el estudio de la validación cruzada out-sample, el método con una media de acierto más alta corresponde a SVM,seguido de KNN.

Tabla 63 Medias de los Folds para cada clasificador

	Media CV-OUT
NB	0.4552
KNN	0.5384
CART	0.5284
QDA	0.4637
SVM	0.5452

En el gráfico de la validación cruzada out simple, se observa como la variabilidad en este caso es más grande. Los clasificadores NB y SVM presentan una variabilidad bastante elevada respecto el resto.

Np obstante SVM, llega a alcanzar precisiones muy altas de más del 60%, pero también algunas no tan altas cerca del 50%. Por otro lado, KNN, con la variabilidad más baja de todos los métodos, mantiene en sus predicciones un porcentaje de acierto muy parecido en toda la validación cruzada, alrededor de 54-55%.

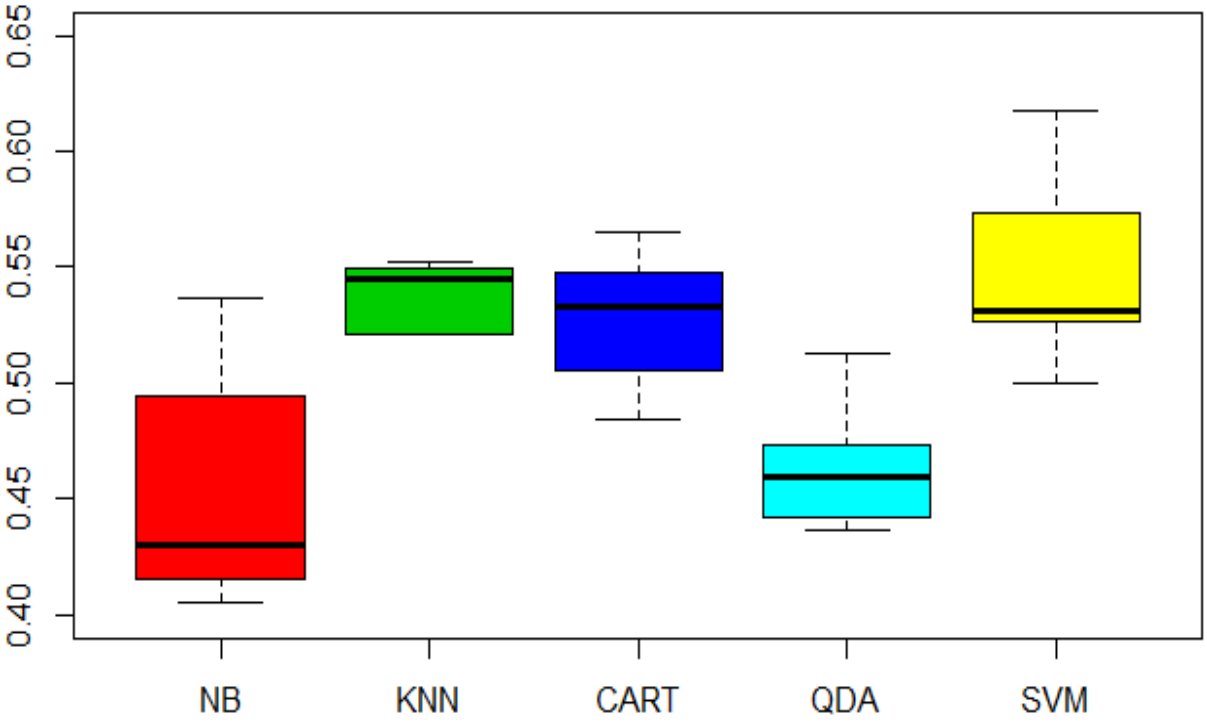


Ilustración 46 Boxplots para cada uno de los clasificadores (Out-sample)

7.- Validación Real (Out of Sample Forecast)

7.1.- Introducción

En último lugar, se va aplicar el caso de Validación Real, que tiene las siguientes diferencias respecto la calibración y validación cruzada.

Por un lado, en la calibración, todos los datos (1900 partidos) forman parte, tanto de los datos de entrenamiento como de test. Se utiliza para la autoevaluación de los datos en un primer lugar.

Por otro lado, en la validación cruzada (validación cruzada), se han realizado diez predicciones diferentes, utilizando de manera aleatoria el 90% de los datos (1710 partidos) como datos de entrenamiento) y el 10% restante (190 partidos) como datos de prueba. Se utiliza para mirar el ajuste del problema.

En este último lugar, se va a realizar un caso real. Es decir, como los partidos están ordenados por fecha, el primero, es el más antiguo y el último el más actual, se va a partir de la información recogida, de los primeros 1710 partidos (datos de entrenamiento) y se predecirán los últimos 190 partidos(datos de prueba).

Ilustración 47 Separación se los datos de entrenamiento y de prueba en orden

A continuación, se van a exponer los resultados de los 6 métodos, pero en vez de ir observando los resultados método por método, se realizará una comparación de los 7 a la vez, para apreciar sus similitudes y diferencias.

En primer lugar, se mostrará la información a priori de cada uno de los métodos.

Naive Bayes

En la tabla (código extraído de R Studio), se muestra en primer lugar las probabilidades a priori de H (gana el equipo local), D (hay empate) y A(gana el equipo visitante).

Además se mostrará, para cada variable predictora, la media y la desviación típica.

Debido a la gran cantidad de variables predictoras, se mostraran las 3 primeras:

Tabla 64 Probabilidad a priori i probabilidades condicionadas

Naive Bayes Classifier for Discrete Predictors

Call: naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

H	D	A
0.4590643	0.2567251	0.2842105

Conditional probabilities:

B365H

Y	[,1]	[,2]
H	2.051146	0.8937942
D	2.543257	1.3147296
A	3.442778	2.1147555

B365D

Y	[,1]	[,2]
H	3.956803	1.0303611
D	3.598018	0.6654335
A	3.558416	0.5710368

B365A

Y	[,1]	[,2]
H	5.716535	3.776832
D	4.171686	2.594030
A	3.031852	1.690868

Los valores a priori de la validación real, comparados con el apartado de la calibración, son muy parecidos. El mismo hecho, ocurre con las medias y desviaciones típicas.

KNN

Para determinar el parámetro k, igual que en el primer proceso de la calibración, se han probado diferentes, y se ha elegido aquel que proporciona un retorno neto más grande. Es decir, en vez de escoger el k que proporciona un porcentaje de aciertos más alto, el estudio se ha centrado en el que proporciona un beneficio más alto.

En un primer momento, se puede pensar que ha mayor número de aciertos , mayor sera el beneficio, pero esto realmente no es del todo cierto. Es posible que un clasificador no sea el que acierte más pero tenga el retorno neto más grande, debido a que es capaz de predecir correctamente partidos donde las cuotas están muy bien pagadas, como es el caso de los empates.

Por lo tanto, nos interesa el clasificador que proporcione un beneficio más alto.

En la tabla que se muestra a continuación, aparece para cada k, el número de aciertos, su respectivo porcentaje de acierto y el beneficio de retorno en € si se apuesta a una casa en concreto.

Tabla 65 Aciertos, porcentaje y retorno según la casa de apuestas y el parámetro K

K	aciertos	%	B365	BW	IW	LB	PS	WH	VC
1	94/190	49’47	2.76	1.13	-4.04	-0.15	7.20	2.54	4.86
3	104/190	54’73	12.73	12.24	10.51	11.79	18.11	15.23	14.30
5	102/190	53’68	3.14	1.93	0.03	1.28	6.93	4.26	4.43
7	111/190	58’42	19.46	18.20	15.53	17.32	23.56	20.54	20.37
10	110/190	57’89	4.99	4.27	2.33	3.36	8.88	7.11	5.92
20	110/190	57’89	-0.30	-0.49	-2.07	-1.62	3.10	1.31	0.57

Observando la tabla, se observa que en este caso, el método que más acierta, es a la vez, el que más beneicio produce que corresponde a k=7.

Por lo tanto, se escogerá este valor para tenerlo en cuenta en la predicción.

CART

Primero de todo, se aplica este método en R Studio y se obtiene un árbol completo con un gran número de nodos y ramificaciones .

Con la finalidad de reducir la varianza del modelo y así disminuir el *test error*, se somete al árbol a un proceso de *pruning(poda)*. Este proceso, intenta encontrar el árbol más sencillo (menor tamaño) que consigue explicar las observaciones.

Para ello se utilizará la tabla siguiente:

Tabla 66 Posibles divisiones del árbol para determinar el parámetro CP

```
Root node error: 925/1710 = 0.54094

n= 1710

          CP nsplit rel error  xerror     xstd
1  0.1232432      0   1.00000 1.00000 0.022277
2  0.0421622      1   0.87676 0.88324 0.022330
3  0.0080000      2   0.83459 0.86162 0.022301
4  0.0064865      7   0.79459 0.87243 0.022317
5  0.0054054      8   0.78811 0.87459 0.022320
6  0.0043243     10   0.77730 0.88108 0.022328
7  0.0037838     12   0.76865 0.88432 0.022332
8  0.0032432     16   0.75135 0.90162 0.022346
9  0.0027027     17   0.74811 0.92865 0.022352
10 0.0025225     23   0.73081 0.93405 0.022351
11 0.0021622     30   0.70595 0.94486 0.022347
12 0.0018018     49   0.65730 0.95135 0.022343
13 0.0016216     53   0.64757 0.97081 0.022324
14 0.0014414     61   0.63459 0.98703 0.022301
15 0.0010811     72   0.61081 0.98703 0.022301
16 0.0010000     90   0.58919 1.01946 0.022234
```

Como se ha comentado en el apartado de la calibración, se utilizaran los siguientes parámetros:

CP : Parámetro de complejidad (penalización para controlar el tamaño del árbol)

Nsplit : es el número de divisiones del árbol

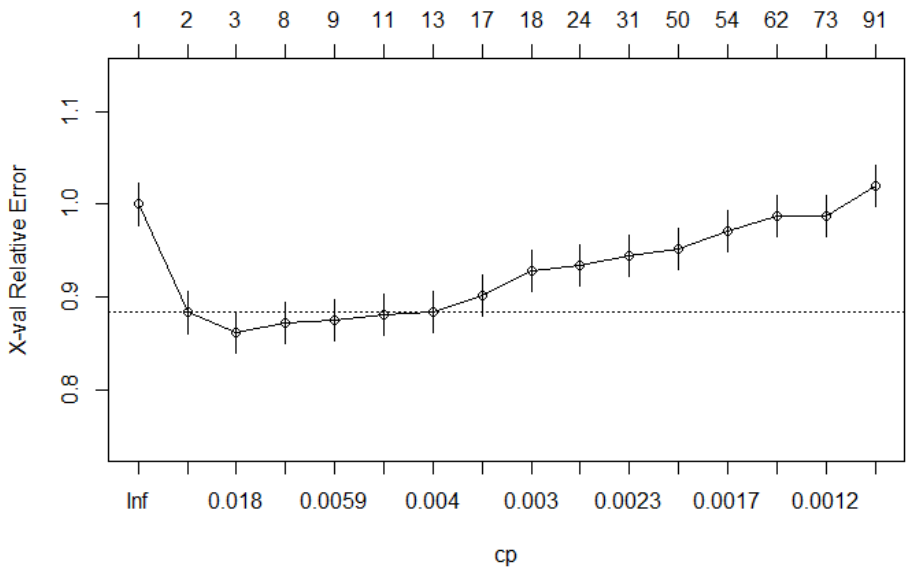
rel error : Desviación media del árbol al que se refiera dividida entre la desviación media del árbol nulo (nsplit = 0).

xerror : valor medio estimado mediante un procedimiento de validación cruzada

xstd : error estándar del error relativo.

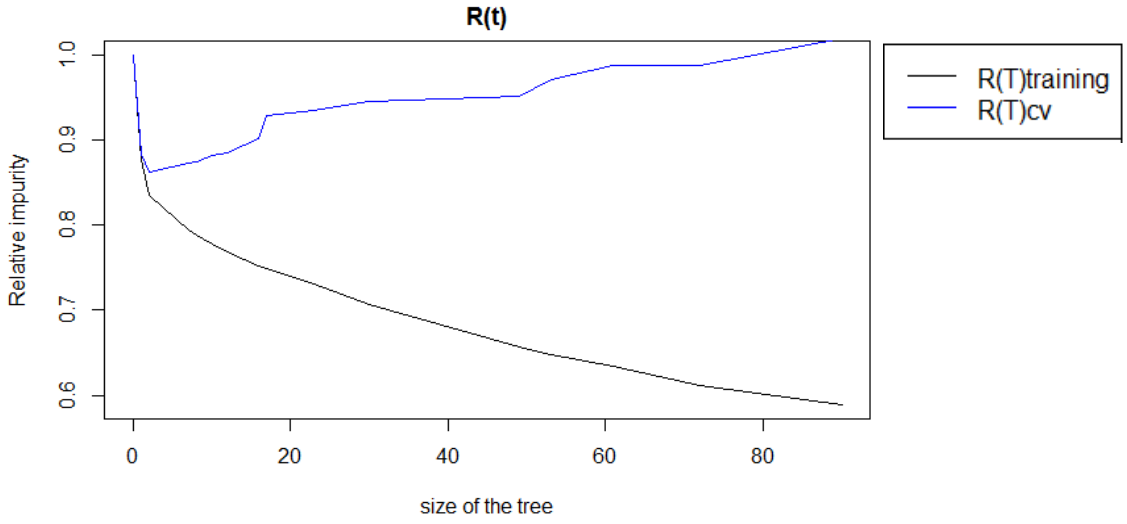
El error que tendremos que observar para la decisión de la poda es el xerror. Se escogerá aquel que tenga un xerror más pequeño. En este caso xerror=0.86162, que corresponde a un cp de 0.0080 y un número de divisiones (nsplit) igual a 2.

Tabla 67 X-error según el CP



En el gráfico que se acaba de mostrar, se muestra en el eje X, los diferentes valores de cp y en el eje Y, el xerror. Se aprecia, como el xerror disminuye al principio y con valores CP más altos crece.

Tabla 68 Variación de la impureza según el tamaño del árbol



Este segundo gráfico, muestra, que cuando el tamaño del árbol es más grande, la impureza (error) para los datos de entrenamiento va disminuyendo progresivamente. Por otro lado, no ocurre los mismo en la crosvalidación (línea negra)

A continuación, se muestra el esquema del árbol de clasificación podado. Cada inciso, nos indica un nodo y la regla de clasificación que le corresponde. Siguiendo estos nodos, podemos llegar a las hojas del árbol, que corresponde a la clasificación de los datos.

Tabla 69 Esquema del árbol de clasificación podado

```
n= 1710

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 1710 925 H (0.4590643 0.2567251 0.2842105)
  2) PSA>=3.865 813 293 H (0.6396064 0.2287823 0.1316113) *
  3) PSA< 3.865 897 518 A (0.2954292 0.2820513 0.4225195)
    6) BWH< 2.925 476 294 H (0.3823529 0.3172269 0.3004202) *
    7) BWH>=2.925 421 185 A (0.1971496 0.2422803 0.5605701) *
```

Todo lo anterior resulta mucho más claro si se visualiza, así que se creará una gráfica usando el modelo.

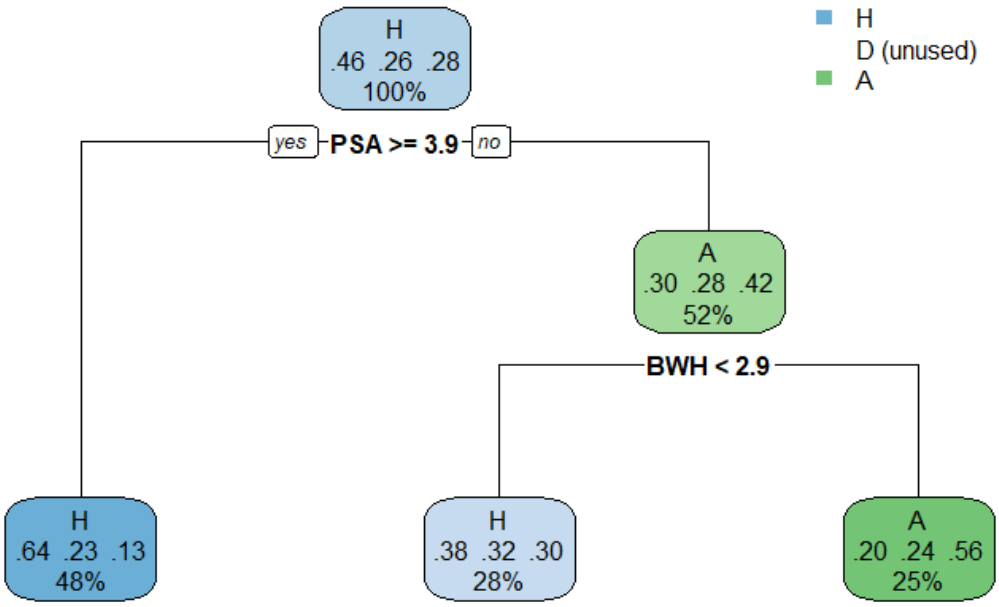


Ilustración 48 Árbol de decisión óptimo

A diferencia de la grafica del apartado de calibración, este gráfico presenta 2 divisiones (nsplits) en vez de 3. Otra diferencia, es que en Calibración solo se separaban las clases a partir del predictor PSA, en cambio, en este caso, influye el predictor BWH.

Por otro lado, en los dos casos, el cp es prácticamente el mismo y no clasifican empates

QDA

La siguiente técnica de predicción es la del análisis de discriminación. Se trata de una generalización de discriminación de Fisher que pretende caracterizar o bien separar las clases. Es decir, utilizando esta técnica se quieren encontrar combinaciones de las variables que mejor expliquen la base de datos.

A la hora de elegir entre análisis de discriminación lineal versus análisis de discriminación cuadrático, se ha tenido en cuenta la siguiente condición:

"Cada predictor que forma parte del modelo se distribuye de forma normal en cada una de las clases de la variable respuesta. En el caso de múltiples predictores, las observaciones siguen una distribución normal multivariante en todas las clases."

Por lo tanto, como se ha detectado que hay variables que no siguen la distribución normal, el conjunto que formará con otras variables, no seguirá una distribución multivariante. Por este motivo, no se puede aplicar el análisis de discriminación lineal y se opta por su versión cuadrática.

La tabla que se presenta a continuación, muestra la probabilidad a priori de los diferentes niveles de la variable respuesta y la media de la variables predictoras.

Tabla 70 Probabilidades a priori i medias de las variables predictoras

Call:
qda(FTR ~ ., data = datos_train, prior = apriori)

Prior probabilities of groups:

H	D	A
0.4590643	0.2567251	0.2842105

Group means:

	B365H	B365D	B365A	BWH	BWD	BWA	IWH	IWD	IWA	LBH	LBD
H	2.051146	3.956803	5.716535	2.049261	3.928038	5. 526318	2.026127	3.823223	5.088777	2.044306	3.905465
D	2.543257	3.598018	4.171686	2.518497	3.565877	4.084214	2.431572	3.524396	3.814852	2.508975	3.568975
A	3.442778	3.558416	3.031852	3.415021	3.530658	2.993025	3.193333	3.478601	2.895062	3.341626	3.529198

	LBA	PSH	PSD	PSA	WHH	WHD	WHA	VCH	VCD	VCA
H	5.533439	2.120191	4.157834	6.370038	2.085185	3.768204	5.651389	2.111236	4.038803	6.288930
D	4.059590	2.654009	3.737107	4.524601	2.560979	3.443599	4.115399	2.633052	3.634009	4.475604
A	2.983621	3.707243	3.684198	3.217058	3.425391	3.409486	3.048251	3.651893	3.579630	3.196975

SVM

En último lugar, se mostrará la información a priori para el método de maquinas de soporte vectorial.

El mejor kernel utilizado es el radial, con un coste de 1 y un parámetro gamma de *0.04761905.*

Además el número de vectores soporte es de 1418

Tabla 71 Tipo de SVM,Kernel, coste y gamma utilizados

```
Call:
svm(formula = FTR ~ ., data = datos_train, type = "C-classification")

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.04761905

Number of Support Vectors:  1418
```

7.2.- Precisión

Tabla 72 Número de aciertos y precisión según el clasificador

Clasificador	Número de aciertos	Precisión %
NV	104/190	54’74
KNN(K=7)	111/190	58’42
CART	110/190	57’89
QDA	103/190	54’21
SVM	111/190	58’42
MIN	111/190	58’42

En el estudio de la precisión, se observa como hay un triple empate en el número de aciertos máximo. Tanto KNN, SVM como MIN aciertan el 58’42 % de los partidos.

No obstante, como ya se ha comentado anteriormente, el objetivo de este trabajo, no es determinar que clasificador predice con más acierto, si no que clasificador devuelve un retorno neto positivo más grande. Como se observará más adelante, hay diferencias muy significativas en el retorno en aquellos clasificadores que tienen de hecho la misma la precisión.

A continuación, se van a mostrar las matrices de confusión conjuntamente, para todos los métodos.

7.3.- Matrices de confusión

Antes que nada, hay que comentar, que de los 190 partidos de los datos de prueba, el número de victorias del equipo local, empates y victorias del equipo visitante, son los siguientes:

H = 84 D = 42 A = 64

NB

	H	D	A
H	44	9	6
D	19	14	12
A	21	19	46

KNN

	H	D	A
H	66	25	16
D	4	6	9
A	14	11	39

CART

	H	D	A
H	59	19	13
D	0	0	0
A	25	23	51

QDA

	H	D	A
H	56	18	11
D	10	3	9
A	18	21	44

SVM

	H	D	A
H	72	28	25
D	0	0	10
A	12	14	39

MIN

	H	D	A
H	71	28	24
D	0	0	0
A	13	14	40

Antes de pasar a explicar las conclusiones de las matrices de confusión, hay que comentar que las columnas, hacen referencia al verdadero resultado y las filas a la predicción.

Es decir, la suma de las columnas, es realmente, el verdadero número de veces que ha ocurrido cada suceso (H,D,A). Es por eso, que todas todas suman 84,42 y 64.

Por otro lado, los valores que se encuentran en color verde, en la diagonal, hacen referencia, a los valores bien clasificados.

En las matrices de confusión que se acaban de presentar, se puede observar lo siguiente:

En primer lugar, el método que más aciertos tiene sobre la victoria del equipo local es SVM, seguido de MIN. Estos métodos aciertan 72 y 71 veces correctamente que gana el local de las 84 posibles.

En cuanto a empates, el método que más acierta es Naive Bayes(NB) con un total de 14. Aunque NB, en el global ha sido un método que no tiene un nivel de precision de los más altos, el hecho de haber acertado empates, puede producir que el beneficio sea mayor, incluso que un método que tiene una mayor precisión pero no ha acertado empates, como es el caso de CART, SVM y MIN.

En victorias del equipo visitante, el mejor método en este caso ha sido CART con un total de 51 de los 64 posibles.

Por último, comentar que los clasificadores de SVM y MIN, siguen una distribución muy parecida a la hora de clasificar resultados. En cuanto al retorno, se tendrá que observar a continuación si también son parecidos.

7.4.- Retorno neto

A continuación, se muestra para cada método, el beneficio o pérdida que supone apostar a los 190 encuentros para las diferentes casas de apuestas.

Tabla 73 Retorno neto según el método y casa de apuesta

	B365	BW	IW	LB	PS	WH	VC
NV	7.58	7.39	4.18	6.88	11.80	8.07	8.38
KNN	19.46	18.20	15.53	17.32	23.56	20.54	20.37
CART	2.79	2.16	1.13	1.60	5.82	3.62	3.44
QDA	-12.22	-11.83	-14.97	-13.23	-8.56	-10.55	-11.05
SVM	-6.03	-5.84	-7.27	-6.88	-2.31	-3.87	-4.91
MIN	-6.23	-5.99	-7.27	-6.88	-2.48	-3.75	-5.16

En primer lugar se aprecia que en 21 de las 42 posibles combinaciones se tiene un beneficio y en las otras 21 una pérdida. Además, hay que añadir que los clasificadores NB, KNN Y CART dan resultados positivos en el beneficio, independientemente de la casa de apuestas elegida. No obstante, las diferencias entre casas son relevantes, ya que se obtiene un beneficio superior en PS (Pinnacle) que en el resto de casas posibles.

La mejor combinación posible, es utilizar el clasificador KNN con un parametro de k=7 y escoger como casa de apuestas PS.

Por otro lado, la peor combinación, seria utilizar el clasificador QDA para la casa de apuestas IW.

El beneficio neto en el primer caso seria de 23’56€. Es decir, si se apostará un euro a cada encuentro, de los 190€ apostados inicialmente, se acabaría con 190+23.56=213.56€.

En el Segundo caso, seria al revés, de los 190€ apostados inicialmente, se acabaría con 190-14.97=175’03€.

A continuación, se va mostrar una tabla , que ordena el retorno de mayor a menor, teniendo en cuenta la combinación de clasificador y la casa de apuesta elegida.

Tabla 74 Clasificación del método y casa de apuestas según el retorno

POSICIÓN	CLASIFICADOR	CASA DE APUESTAS	RETORNO €
1	KNN	PS	23.56
2	KNN	WH	20.54
3	KNN	VC	20.37
4	KNN	B365	19.46
5	KNN	BW	18.20
6	KNN	LB	17.32
7	KNN	IW	15.53
8	NV	PS	11.80
9	NV	VC	8.38
10	NV	WH	8.07
11	NV	B365	7.58
12	NV	BW	7.39
13	NV	LB	6.88
14	CART	PS	5.82
15	NV	IW	4.18
16	CART	WH	3.62
17	CART	VC	3.44
18	CART	B365	2.79
19	CART	BW	2.16
20	CART	LB	1.60
21	CART	IW	1.13
22	SVM	PS	-2.31
23	MIN	PS	-2.48
24	MIN	WH	-3.75
25	SVM	WH	-3.87
26	SVM	VC	-4.91
27	MIN	VC	-5.16
28	SVM	BW	-5.84
29	MIN	BW	-5.99
30	SVM	B365	-6.03
31	MIN	B365	-6.23
32	SVM	LB	-6.88
33	MIN	LB	-6.88
34	SVM	IW	-7.27
35	MIN	IW	-7.27
36	QDA	PS	-8.56
37	QDA	WH	-10.55
38	QDA	VC	-11.05
39	QDA	BW	-11.83
40	QDA	B365	-12.22
41	QDA	LB	-13.23
42	QDA	IW	-14.97

7.5.- Evolución del beneficio o pérdida

Seguidamente, se va mostrar para el método de clasificación NB, el progreso que ha seguido la cartera del jugador partido tras partido, en el caso que se apostará 1€ por encuentro al resultado que nos da el predictor.

Tabla 75 Variación del beneficio o pérdida por jornada (Ejemplo NB)

Partido	B365	BW	IW	LB	PS	WH	VC
1	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
2	-2.00	-2.00	-2.00	-2.00	-2.00	-2.00	-2.00
3	-1.86	-1.85	-1.88	-1.82	-1.82	-1.83	-1.83
4	0.64	0.65	0.52	0.68	0.91	0.47	0.77
5	-0.36	-0.35	-0.48	-0.32	-0.09	-0.53	-0.23
6	0.55	0.56	0.37	0.53	0.81	0.42	0.64
7	-0.45	-0.44	-0.63	-0.47	-0.19	-0.58	-0.36
8	0.80	0.81	0.57	0.73	1.11	0.67	0.89
9	1.80	1.86	1.87	1.78	2.21	1.77	1.94
10	0.80	0.86	0.87	0.78	1.21	0.77	0.94
11	-0.20	-0.14	-0.13	-0.22	0.21	-0.23	-0.06
12	-1.20	-1.14	-1.13	-1.22	-0.79	-1.23	-1.06
13	-2.20	-2.14	-2.13	-2.22	-1.79	-2.23	-2.06
14	-1.70	-1.64	-1.63	-1.72	-1.28	-1.70	-1.56
15	-1.15	-1.11	-1.08	-1.17	-0.72	-1.13	-1.01
…							
80	13.72	13.76	11.79	13.15	15.96	14.16	14.37
81	14.12	14.18	12.29	13.55	16.41	14.60	14.82
82	15.92	15.98	13.94	15.30	18.27	16.40	16.62
83	14.92	14.98	12.94	14.30	17.27	15.40	15.62
84	15.36	15.43	13.44	14.70	17.70	15.84	16.06
85	14.36	14.43	12.44	13.70	16.70	14.84	15.06
86	14.50	14.58	12.56	13.84	16.84	15.01	15.19
87	14.90	14.98	12.96	14.24	17.26	15.41	15.55
88	16.30	16.38	14.26	15.54	18.66	16.81	16.93
89	15.30	15.38	13.26	14.54	17.66	15.81	15.93
90	15.87	15.98	13.81	15.09	18.23	16.38	16.53
…							
180	9.39	9.19	5.69	8.72	13.20	9.84	10.24
181	9.68	9.45	5.96	9.01	13.48	10.13	10.49
182	10.41	10.25	6.96	9.74	14.26	10.88	11.24
183	9.41	9.25	5.96	8.74	13.26	9.88	10.24
184	8.41	8.25	4.96	7.74	12.26	8.88	9.24
185	7.41	7.25	3.96	6.74	11.26	7.88	8.24
186	7.74	7.58	4.36	7.07	11.63	8.24	8.57
187	6.74	6.58	3.36	6.07	10.63	7.24	7.57
188	6.80	6.64	3.43	6.13	10.98	7.30	7.61
189	6.98	6.82	3.63	6.28	11.15	7.47	7.76
190	7.58	7.39	4.18	6.88	11.80	8.07	8.38

En la jornada 190, se muestra de hecho, el retorno final neto de este método. En el caso de PS, nos encontramos con una pérdida en la jornada 15 de -0.72€, una ganancia en la jornada 90 de 18.23€ y en la jornada 190 un beneficio neto total de 11.80€. Este ultimo valor (11’80€), en la tabla de dos páginas antes, se encuentra como la octava mejor opción posible de las 42 existentes en este trabajo.

Estas tablas, realizadas con R Studio, se han aplicado para los 6 métodos y han dado lugar a la gráficas que se presentan a continuación. De este modo, se facilita la visualización del proceso.

7.6.- Gráficos

Los 6 gráficos que se muestran a continuación, reflejan el retorno neto que se tiene desde el primer partido hasta el ultimo.

NB

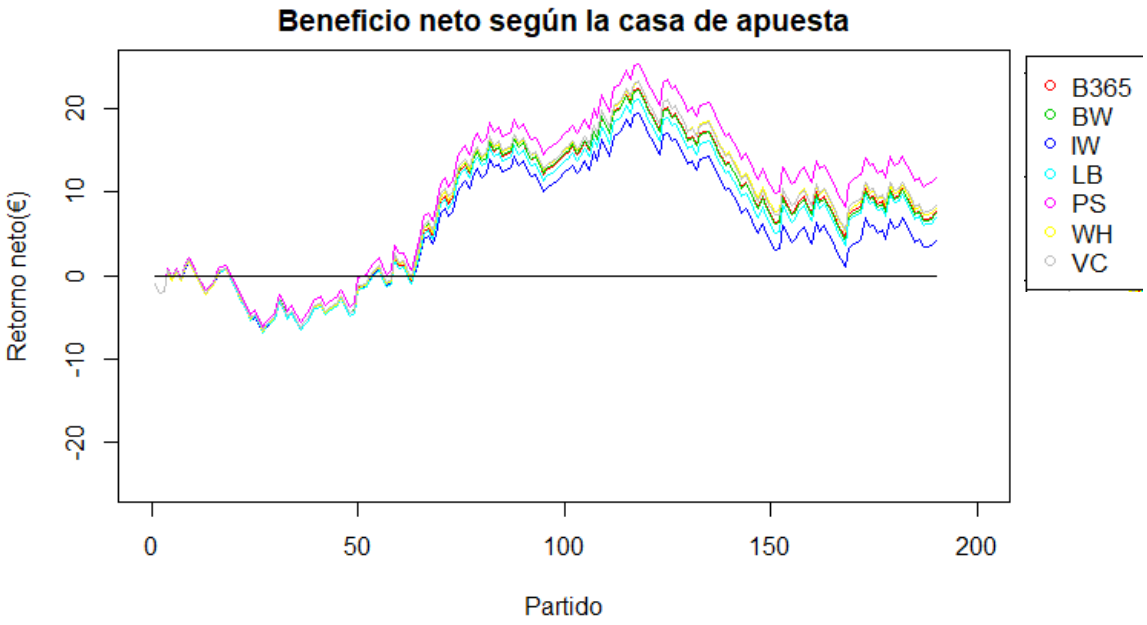


Ilustración 49 Variación del retorno según la casa de apuestas y jornada (NB)

El gráfico de NB, muestra como hasta el encuentro 60 se mantiene en pérdidas, y despúes el retorno neto es positivo hasta el final. La casa de apuestas PS, lidera en todos los métodos como la mejor casa para apostar.

En comparación con el primer apartado de los resultados, el punto 4.3 de calibración, en el cual solo una casa (PS) presentaba un retorno positivo, en este caso las 7 casas se mantienen en beneficio. Hay que apreciar que este es el caso real, es decir, el que realmente hay que tener en cuenta. El gráfico de NB en calibración es únicamente informativo, pero no aplicable, ya que realiza apuestas sobre partidos que ya se habrían jugado. En cambio en el caso presente de Validación Real, se ha realizado el estudio en únicamente la muestra test, que la componen los últimos 190 partidos de los 1900 totales.

KNN(K=7)

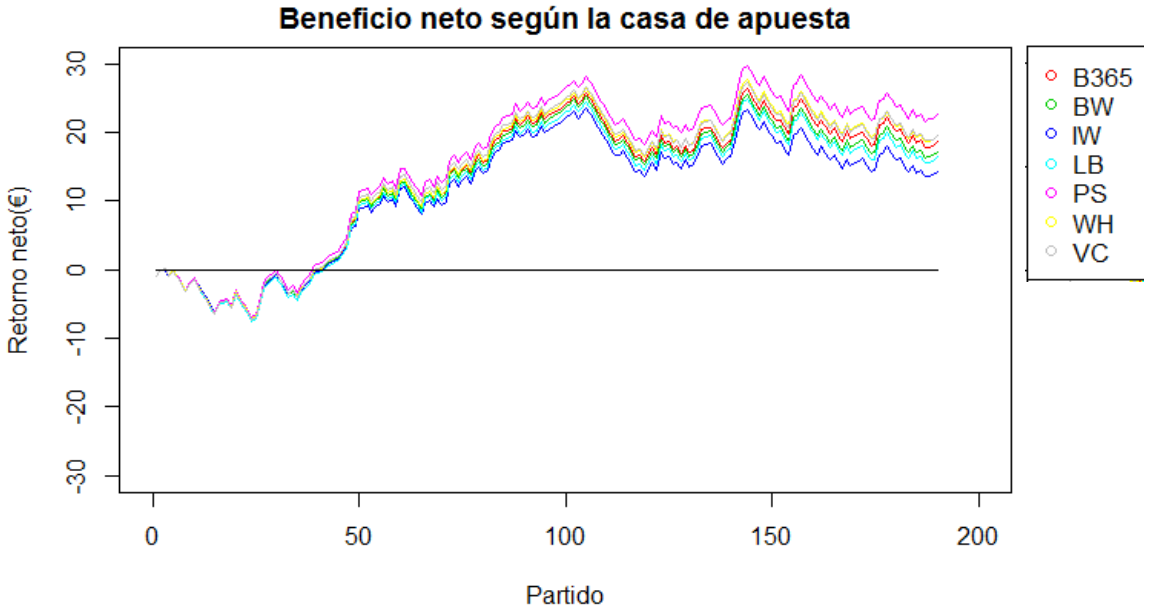


Ilustración 50 Variación del retorno según la casa de apuestas y jornada (KNN)

El gráfico de KNN, refleja como claramente es el método con el retorno positivo más grande.

Además, igual que en NB, todas las casas de apuestas acaban con un retorno positivo

CART

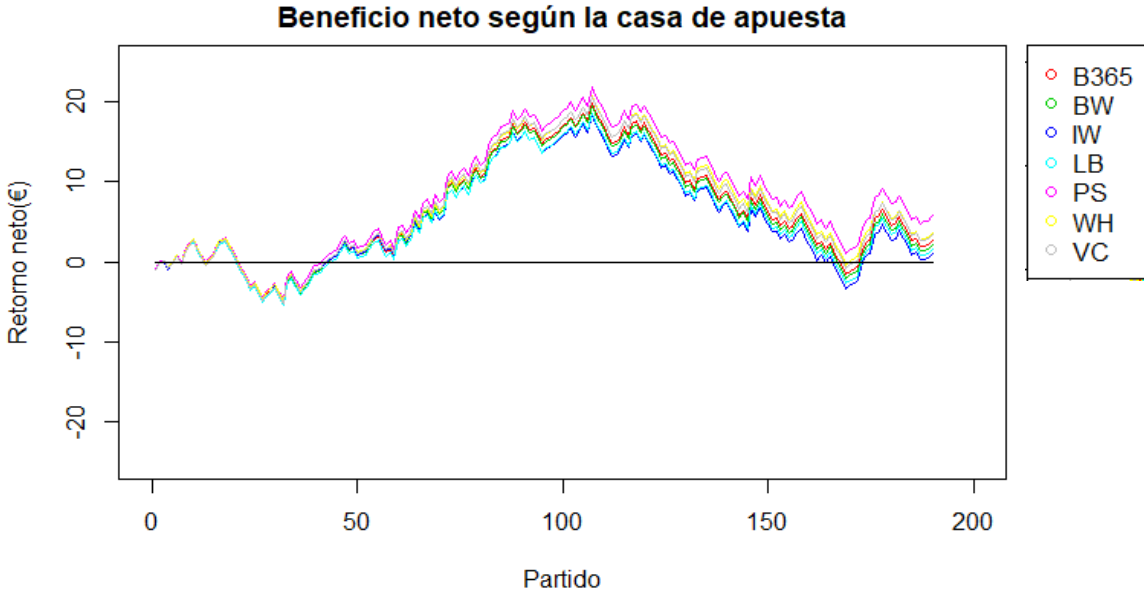


Ilustración 51 Variación del retorno según la casa de apuestas y jornada (CART)

Cart, presenta un gráfico en forma de pirámide. A partir del partido 50 empieza a crecer y una vez alcanza el beneficio máximo en el partido 105, empiezan a descender los beneficios.

No obstante, termina el estudio en positivo, para todas las casas.

QDA

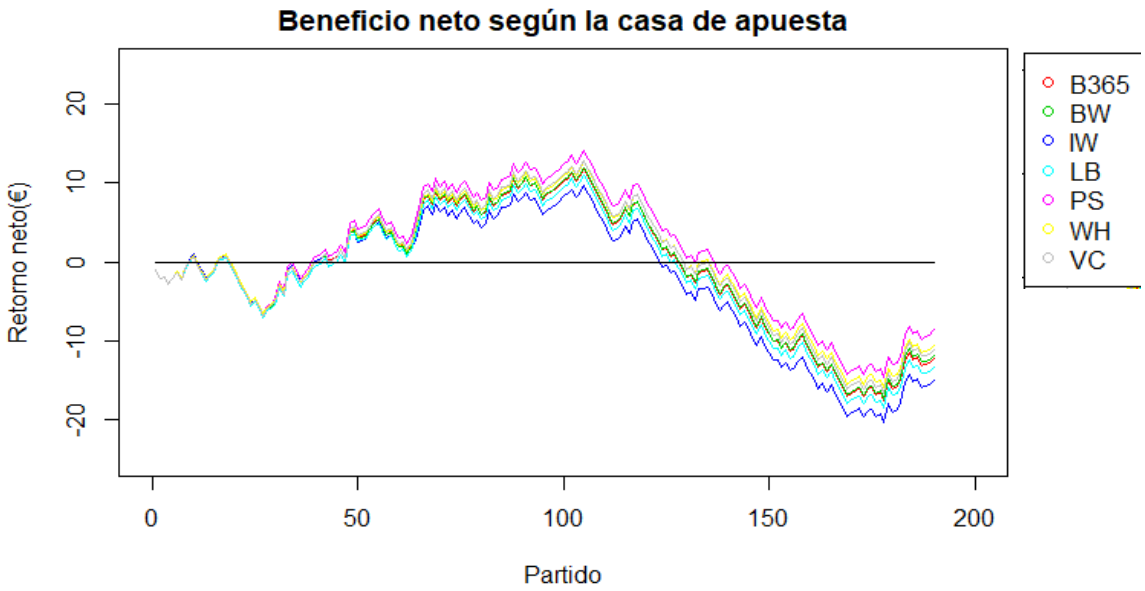


Ilustración 52 Variación del retorno según la casa de apuestas y jornada (QDA)

En QDA, se aprecia como del partido 43 hasta el 100 funciona bien, pero a partir de este encuentro la predicción junto con el beneficio caen en picado. Aunque, en el tramo final se observa una recuperación, esta no es suficiente como para poder alcanzar valores positivos.

SVM

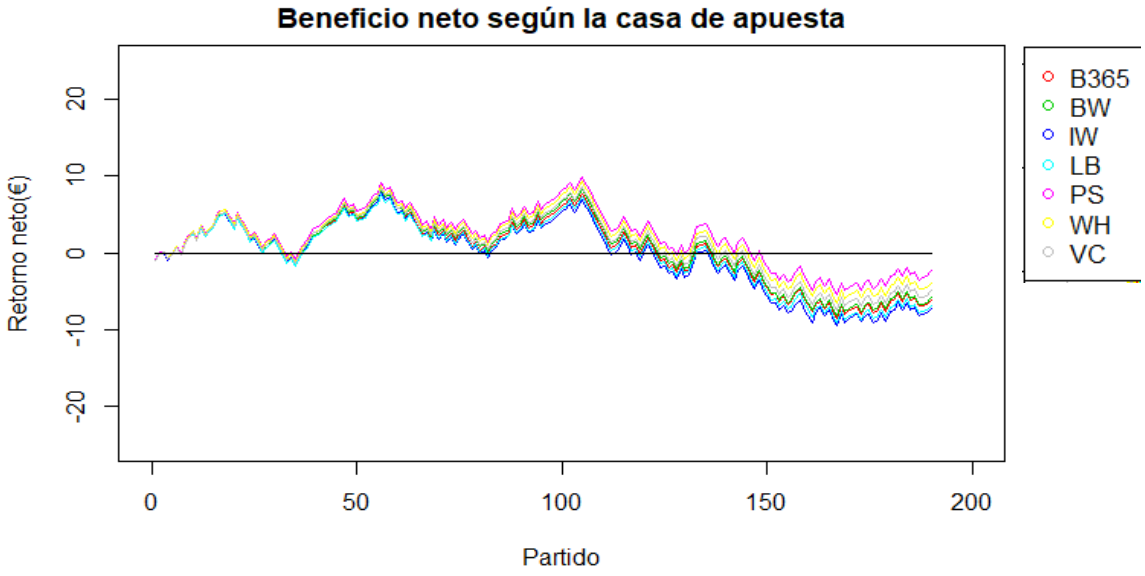


Ilustración 53 Variación del retorno según la casa de apuestas y jornada (SVM)

Tanto SVM como MIN, presentan gráficas prácticamente idénticas, esto quiere decir, que el método de clasificación SVM, realiza la clasificación del resultado del partido, escogiendo la cuota con el valor más bajo y apostando a ese resultado.

Este tipo de métodos, el cuál es utilizado por muchos jugadores, no da un buen resultado y es que si realmente este método funcionará las casas de apuestas tendrían un grave problema.

MIN

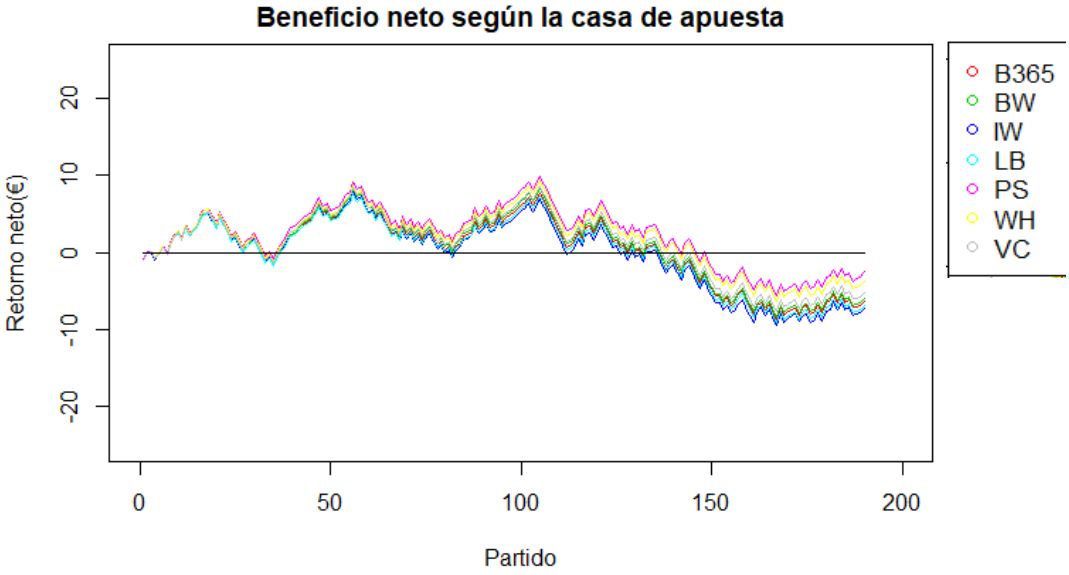


Ilustración 54 Variación del retorno según la casa de apuestas y jornada (MIN)

7.7.- Mejora del beneficio o reducción de la pérdida

Existe un método, para poder mejorar las ganancias. Este consiste en escoger la casa de apuestas que ofrezca una cuota más alta para cada partido independiente. Es decir, si en un partido se decide apostar a la victoria del local, se comparan las cuotas de las 7 casas de apuestas, y se realiza la apuesta en aquella que sea más alta. No tendría sentido apostar en un partido a una casa que por el mismo resultado vaya a pagar menos.

Apreciando anteriormente la tabla de retorno neto según las casas de apuestas, se ha observado, que la mejor casa para apostar utilizando cualquiera de los 6 métodos, corresponde a PSA. Por lo tanto, la tabla presentada a continuación, mostrará el beneficio de PSA para cada método y el beneficio del método de la mejora, ahora explicado

Tabla 76 Mejora del beneficio según el clasificador

Clasificador	Antes de la mejora	Después de la mejora
NV	11.80	14.92
KNN(K=7)	23.56	25.44
CART	5.82	9.18
QDA	-8.56	-6.13
SVM	-2.31	0.38
MIN	-2.48	0.23

Además, a continuación, se va a mostrar para los partidos acertados, cuantas veces se apostaría a cada casa, si se tiene en cuenta el hecho de maximizar el beneficio eligiendo la casa con la cuota más alta.

Tabla 77 Número de veces que se apuesta a cada casa según el clasificador

Clasificador	B365	BW	IW	LB	PS	WH	VC
NV	8	17	30	13	48	37	9
KNN(K=7)	9	13	29	10	54	36	14
CART	8	14	32	11	51	40	11
QDA	7	17	30	11	48	35	12
SVM	6	14	33	10	55	38	11
MIN	6	14	33	10	54	39	11

8.- Conclusiones

La conclusiones finales, harán referencia al caso de validación real, es decir, partimos de la información de las cuotas de 1710 partidos (datos de entrenamiento) y se predicen los 190 partidos últimos (datos de prueba).

1.- En primer lugar, la precisión más alta de acierto ha sido de 111 partidos de un total de 190. Esto corresponde a un 58'42 % de acierto y los dos métodos que comparten este resultado son SVM y KNN con un parámetro de k =7.

2.- En las matrices de confusión se ha podido observar lo siguiente:

2.1.- El método que más aciertos tiene sobre la victoria del equipo local es SVM.Este método acierta 72 veces correctamente que gana el local de las 84 posibles.

2.2.- En cuanto a empates, el método que más acierta es Naive Bayes(NB) con un total de 14 sobre 42. Claramente, el acierto de un empate ha quedado demostrado que es de lo más complicado.

2.3.- En victorias del equipo visitante, el mejor método en este caso ha sido CART(Arbol de decisión) con un total de 51 de los 64 posibles.

3.- Una de las observaciones mas curiosas, es que obtener una precisión de acierto más alta que el resto, no garantiza que los beneficios sean también más altos. Esto es debido, a que puede existir un clasificador que no tenga una precisión muy alta, pero en cambio acierte empates o resultados poco probables, lo que conlleva a un beneficio más alto, debido a que la cuota será también más alta

Este hecho, ha ocurrido en el trabajo, con el clasificador Naive Bayes, el cual es el segundo peor de los 5 en predecir pero en cambio, una vez acabado el partido 190, se obtiene un retorno neto para todas las casas de apuestas.

4.- El principal objetivo del trabajo, era determinar la combinación de que clasificador y casa de apuestas, se debe elegir para obtener un beneficio más alto. Este suceso, corresponde al método de clasificación KNN con parámetro k=7 , el cual se debe aplicar en la casa de apuestas PS(Pinnacle). El retorno neto positivo obtenido ha sido de 23'56€, en el caso que se apostara 1 € a cada encuentro.

Este suceso, puede dar ganancias más significativas si se aumenta la apuesta inicial. Es decir, si en vez de apostar un euro por partido, se decidiera jugar 5€, el retorno en este caso seria de 23'56*5=117'8€.

5.- La casa de apuestas PS(Pinnacle), ha sido de las 7 casas estudiadas, la mejor en todos los clasificadores. El motivo principal, es debido, a que en la mayoría de partidos, se han ofrecido cuotas más altas respecto el resto.

6.- Una de las conclusiones más relevantes, es que, si en vez de jugar siempre a la misma casa de apuestas, se apuesta para cada partido en aquella que ofrece la cuota más alta, el beneficio aumentará al largo del tiempo.

Este suceso, ha quedado demostrado, para todos los casos. En conclusión, si se utiliza como método de clasificación KNN con k=7, pero, en vez de escoger la casa de apuestas Pinnacle, se apuesta a aquella que en cada partido diferente ofrezca una cuota mayor, el beneficio pasará de los 23'56€ a los 25'44€. Por lo tanto, este hecho, a la larga, produce ganancias significativas.

9.- Bibliografía

ALUJA, Tomàs, et al. Aprender de los datos: el análisis de componentes principales: una aproximación desde el Data Mining. Barcelona: EUB, 1999

HAND, D. J. Construction and assessment of classification rules. Chichester [etc.]: Wiley, 1997

ALUJA, Tomàs , Construction of classification rules

ALUJA, Tomàs, Slides Svm

http://foroapuestas.forobet.com/la-charla-de-apuestas/58641-historial-de-cuotas.html

https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

https://msdn.microsoft.com/es-es/library/bb895174(v=sql.120).aspx

https://es.wikipedia.org/wiki/Clasificador_bayesiano_ingenuo

http://www.apuestasonline.es/guia-apuestas/historia-apuestas/

http://rstudio-pubs-static.s3.amazonaws.com/233932_1ea94e1d67694a58bb71419b3ad49c5a.html

https://www.kaggle.com/datasets

https://es.wikipedia.org/wiki/M%C3%A1quinas_de_vectores_de_soporte

https://www.igi-global.com/chapter/calibration-machine-learning-models/36983

https://rpubs.com/anish20/decisionTreesinR

10.- Código de R

````
---
title: "Machine Learning, aplicado a las cuotas, en las apuestas deportivas
author: "Pablo Bustillo""
date: "7 de marzo de 2018"
output: html_document
---



```{r}
library(readxl)
#datos <- read_excel("TODO.xlsx")
#datos <-read_excel("C:/Users/pabustil/Desktop/ORACLE/pablo/TFG/TODO.xlsx")
#datos <- read_excel("C:/Users/pablo.bustillo/Downloads/TODO.xlsx")

datos <- read_excel("TODO.xlsx")
datos <- as.data.frame(datos)
class(datos)
str(datos)
names(datos)
head(datos)
nrow(datos)
summary(datos)

```

```{r}
datos$HomeTeam <-as.factor(datos$HomeTeam)
datos$AwayTeam <-as.factor(datos$AwayTeam)
datos$FTR <-as.factor(datos$FTR)
datos$FTR=factor(datos$FTR,level=c("H","D","A"))

datos$B365H <-as.numeric(datos$B365H)
datos$B365D <-as.numeric(datos$B365D)
datos$B365A <-as.numeric(datos$B365A)

datos$BWH <-as.numeric(datos$BWH)
datos$BWD <-as.numeric(datos$BWD)
datos$BWA <-as.numeric(datos$BWA)

datos$IWH <-as.numeric(datos$IWH)
datos$IWD <-as.numeric(datos$IWD)
datos$IWA <-as.numeric(datos$IWA)

datos$LBH <-as.numeric(datos$LBH)
datos$LBD <-as.numeric(datos$LBD)
datos$LBA <-as.numeric(datos$LBA)

datos$PSH <-as.numeric(datos$PSH)
datos$PSD <-as.numeric(datos$PSD)
datos$PSA <-as.numeric(datos$PSA)

datos$WHH <-as.numeric(datos$WHH)
datos$WHD <-as.numeric(datos$WHD)
datos$WHA <-as.numeric(datos$WHA)

datos$VCH <-as.numeric(datos$VCH)
datos$VCD <-as.numeric(datos$VCD)
datos$VCA <-as.numeric(datos$VCA)


#ja està corregit a la base de dades
#which(datos[,20]>41)
#datos[datos[,18]>1000,18]=datos[datos[,18]>1000,18]/1000
#datos[datos[,20]>100,20]=datos[datos[,20]>1000,20]/1000
#datos[datos[,24]>1000,24]=datos[datos[,24]>1000,24]/1000
#summary(datos$PSA)
````
````

```
```

```{r}
sum(is.na(datos))
which(is.na(datos))
```

#Descriptiva de la variable Respuesta FTR

```{r}
class(datos$FTR)
table(datos$FTR)
freq <- table(datos$FTR)/1900
freq
pie(table(datos$FTR),col=c(3,4,2))
barplot(table(datos$FTR),col=c(2,4,3))

#
# Pie Chart with Percentages
slices <- c(869,481,550)
lbls <- c("H","D","A")
pct <- round(slices/sum(slices)*100,2)
lbls <- paste(lbls,"(", pct,")","%",sep="") # add percents to labels

pie(slices,labels = lbls, col=rainbow(length(lbls)),
    main="% de resultados")
```

La variable respuesta presenta 3 niveles:
A(Away): Gana el equipo visitante
D(Draw): Empate
H(Home): Gana el equipo local

En la descriptiva se observa que:

- El equipo local gana el 45.74% de las veces(869 casos)
- El equipo visitante gana el 28.95% de las veces(550 casos)
- Hay un 25.32% de empates (481 casos)

Ahora se va a hacer el mismo estudio pero en vez de tener en cuenta los 5 años, se tendrá en cuenta año por año:

```{r}
#2012-2013
t1 <- table(datos$FTR[1:380])
t1
freq1 <- t1/380


#2013-2014
t2 <- table(datos$FTR[381:760])
t2
freq2 <- t2/380


#2014-2015
t3 <- table(datos$FTR[761:1140])
t3
freq3 <- t3/380


#2015-2016
t4 <- table(datos$FTR[1141:1520])
```

```
t4
freq4 <- t4/380


#2016-2017
t5 <- table(datos$FTR[1521:1900])
t5
freq5 <- t5/380




freq1
freq2
freq3
freq4
freq5
```

2012-2013: H=0.4657895
        D=0.2526316
        A=0.2815789

2013-2014: H=0.4763158
        D=0.2368421
        A=0.2868421

2014-2015: H=0.4000000
        D=0.3157895
        A=0.2842105

2015-2016: H=0.4605263
        D=0.2500000
        A=0.2894737

2016-2017: H=0.4842105
        D=0.2105263
        A=0.3052632

En general, como era de esperar, la distribución de H,D,A es muy parecida los diferentes años.
Sin embargo, el año 2014-2015 respecto el resto de años, presenta un porcentaje de victorias para el equipo local bastante inferior y el porcentaje de empates más alto.

#Gráfico
```{r}

graf_años <-
data.frame("H"=c(46.58,47.63,40,46.05,48.42),"D"=c(25.26,23.68,31.58,25,21.05),"A"=c(28.16,28.68,28.42,28.95,30.53))
row.names(graf_años)=c("2012","2013","2014","2015","2016")


graf_años <- as.matrix(graf_años)



barplot(graf_años,
      legend=TRUE,
      beside=TRUE,
      main="Porcentaje de los posibles resultados según el año",
      ylab="%",
      xlab="FTR",
      col=c(2:6),
      ylim=c(0,100),
      xpd=F)

```


#Descriptiva de las variables predictoras

#las=2, gira los nombres del eje x 45 grados
```{r}
summary(datos[,-c(1,2,3,4,5)])
boxplot(datos[,-c(1,2,3,4,5)],main="Boxplots de las cuotas según la casa de apuestas",ylim=c(0,50),ylab="valor de la cuota(???)",xlab="casa de apuestas",las=2,col=c(rep(2:8,each=3)))
par(mfrow=c(3,7))
for(i in 6:26){
#hist(datos[,i])
}



```
En el boxplot, se aprecia que hay casas de apuestas que presentan una variabilidad más grande que otras. Las que más variabilidad tienen son PS y LB.
Además,en todos los casos, la variabilidad es más grande en las apuestas de "A":Gana el equipo visitante, que corresponden a las variables B365A,BWA,IWA,LBA,PSA,WHA,VCA. A continuación, se pueden apreciar las comparaciones

```{r}
apply(datos[,6:26],2,max)
ap <-apply(datos[,6:26],2,function(x) c(min(x),max(x),mean(x)))
rownames(ap) <- c("min","max","mean")
ap
```

Como todas las variables, hacen referencia a cuotas de 7 casas de apuestas para tres sucesos (A(Away): Gana el equipo visitante, D(Draw): Empate, H(Home): Gana el equipo local) se compararán los valores entre ellas.

```{r}
max(datos[,-c(1,2,3,4,5)])
```

-En primer lugar, se ha observado que la cuota más alta corresponde a la casa de apuestas PS, hace referencia a la cuota que se pagaria a aquel jugador que apostará a favor del equipo visitante y la cifra es de 42'320.
Esto quiere decir, que por cada euro jugado el retorno es de 42'320. El beneficio neto si se apostaran 5??? a este suceso seria de 5*42'320-5=206'6???
```{r}
datos[datos$PSA>42,]
datos[datos$PSA>42,c(8,11,14,17,20,23,26)]
```
Investigando más a fondo en la base de datos, se ha encontrado que este partido corresponde a Roma vs Genova, el cual se jugó el dia 5 de Mayo de 2017 y acabo con el resultado de victoria para el local.
El motivo de la cuota tan alta, fue debido a que la probabilidad de que el Genova ganará a la Roma fuera de casa, era muy baja.

Realizando una comparativa de cuanto pagaban las otras casas de apuesta a este partido se observan diferencias muy significativas:

```
 B365A BWA IWA LBA   PSA WHA VCA
   34  29  22  41 42.32  29  34
```

Por un lado, la casa de apuestas LB, es la que más se aproxima a los 42'32??? de la casa de apuestas que mejor paga, PS.
Por el otro lado IW es la casa de apuestas que peor paga para este partido. En comparación con PS, la casa de apuesta IW, pagaria la victoria del Genova a 22???,casi la mitad.
En conclusión, el análisis de las diferentes cuotas, para las diferentes casas de apuestas, es muy importante, ya que la ganancia puede ser del doble.

Esto supone que antes de apostar se tendran que comparar las cuotas de todas las casas existentes y realizar la apuesta en aquella casa que mejor pague.

```{r}
min(datos[,-c(1,2,3,4,5)])
datos[datos$VCH==1.04,]

```

-En segundo lugar, la cuota más baja encontrada, es de 1'04. Este partido corresponde también a Roma vs Genova el dia 5 de Mayo de 2017, y la casa de apuestas con esta cuota es VC. El valor de la cuota tan baja, es debido a que la probabilidad de que el Roma gane al Genova en su propia casa es muy alta.

Una cuota de 1'04, quiere decir que por cada euro jugado el retorno es de 1'04. El beneficio neto si se apostaran 5??? a este suceso seria de 5*1'04-5=0'2???.

Como se puede observar, la diferencia del premio en el caso de acertar un resultado u otro es muy grande. También es cierto, que acertar un partido con una cuota muy alta es muy difícil, ya que la probabilidad de dicho suceso no es muy probable. Por otro lado, apostar a una cuota muy baja, nos da la seguridad de que ese suceso ocurra, pero el retorno es muy pequeño.
Cada jugador ha de pensar y decidir, si merece la pena realmente apostar 5??? a que gana el Genova al Roma, sabiendo que el retorno es de 20 céntimos.

#Descriptiva Bivariante

A continuación se va a realizar un estudio acerca de los aciertos en los partidos para las diferentes casas de apuestas.
Se va a mostrar cuales han sido las cuotas ganadoras en el caso que se haya apostado a empate(D), o que gana el equipo local(H) o visitante(A).

```{r}
#Grafico para Bet365
head(datos)

resultat=data.frame(t(apply(datos,1,function(el) {
 cuo=el[8]
 if(el[5]=="H") {
  cuo=el[6]} else{
   if (el[5]=="D"){
    cuo=el[7]
   }
  }
  cbind(el[5],cuo)})))
colnames(resultat)=c("FTR","CuotaB365")

resultat$CuotaB365=as.numeric(as.character(resultat$CuotaB365))
resultat$FTR=factor(resultat$FTR,levels=c("H","D","A"))

head(resultat)

boxplot(CuotaB365~FTR,resultat,main="Boxplots de las cuotas acertadas(B365)",col=2:4,ylab="cuotas(???)", xlab="Resultados B365")

by(resultat$CuotaB365,resultat$FTR,summary)


```

```{r}
#Índices de las variables que corresponden a H,D o A.
#H: c(6,9,12,15,18,21,24)
#D: c(7,10,13,16,19,22,25)
#A: c(8,11,14,17,20,23,26)

#H
for(i in c(6,9,12,15,18,21,24)){
plot(datos[datos$FTR=="H",5],datos[datos$FTR=="H",i])
summary(datos[datos$FTR=="H",i])
}

#plot(datos[datos$FTR=="H",5],datos[datos$FTR=="H",6],)
summary(datos[datos$FTR=="H",6])


#D
for(i in c(7,10,13,16,19,22,25)){
plot(datos[datos$FTR=="D",5],datos[datos$FTR=="D",i])
summary(datos[datos$FTR=="D",i])
}

#plot(datos[datos$FTR=="D",5],datos[datos$FTR=="D",7],)
```

```
summary(datos[datos$FTR=="D",7])

#A
for(i in c(8,11,14,17,20,23,26)){
plot(datos[datos$FTR=="A",5],datos[datos$FTR=="A",i])
}
```

```{r}
plot(datos[datos$FTR=="H",5],datos[datos$FTR=="H",6],)
summary(datos[datos$FTR=="H",6])
```

Observando el boxplot de las cuotas acertadas para el equipo local(H), en la casa de apuestas B365, se aprecia que cuotas superiores a 3'59 se consideran outliers.

Para el cálculo del outlier se ha tenido en cuenta los siguientes cálculos:

C1: 1r cuartil
C3: 3R CUARTIL
IQR: Rango intercuartílico: C3 - C1 = 2'3 - 1'440 = 0'86

valores atípicos leves superiores: C3 + 1'5 * IQR = 2'3 + 1'5 * 0'86 = 3'59

Considerar como outlier una cuota de 3'59 puede resultar extraño ya que no es extremadamente alta. Esto es debido en primer lugar, a que las cuotas acertadas para equipos locales no suelen ser altas, inferiores a 1'8. Además, hay que comentar que la probabilidad de acertar esta cuota es baja: 1/3'59=0.279. Es decir, hay un 27'9% de acertar este resultado.

Por otro lado, la característica a tener más en cuenta en este caso, es que ha habido un partido en el cual se ha acertado una cuota de 10. Este acierto es claramente un caso muy raro, ya que la probabilidad del suceso es de un 10%.
El retorno neto en el caso de jugar 5??? es de : 5*10-5=45???.

Añadir a este análisis específico que el encuentro corresponde a Crotone vs Inter el 9 de Abril de 2017, con el resultado de 2-1 para el Crotone. Se puede entender perfectamente, que el Crotone es un equipo futbolísticamente inferior al Inter, pero en ese encuentro hubo una sorpresa y se llevó el partido el equipo con probabilidades más bajas de llevárselo.
```{r}
datos[datos$FTR=="H" & datos$B365H==10,c(2,3,4)]
```

Observando el summary de esta variable:
```{r}
summary(datos[datos$FTR=="H",6])
```
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.060   1.440   1.830   2.055   2.300  10.000
```

Se puede ver claramente que el acierto normal suele ser de una cuota aproximadamente de entre 1'4 y 2'3.
Estas apuestas, suelen ser seguras, pero por el contrario el beneficio neto no será excesivamente alto.

Siguendo con el estudio de la casa de apuestas B365, una de las más famosas dentro del sector, en el caso de los empates, el boxplot y el summary son los siguientes:
```{r}
summary(datos[datos$FTR=="D",7])
plot(datos[datos$FTR=="D",5],datos[datos$FTR=="D",7],)
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.730   3.250   3.400   3.628   3.750   9.000
```

La cuota más alta acertada en una apuesta de empate, para esta casa de apuestas es de 9. Este encuentro corresponde al Napoli vs Palermo, jugado el 29 de enero de 2017:
```{r}


```

```
datos[datos$FTR=="D" & datos$B365D>8,c(2,3,4,7)]
```
En el caso de los empates, las cuotas normalmente suelen ser mejor pagadas que las apuestas a las victorias del local, debido a su dificultad de acierto. En este caso, suelen oscilar entr 3'25 y 3,75, dando como outliers valores más bajo de 2'5 Y superiores a 4.

C1: 1r cuartil
C3: 3R CUARTIL
IQR: Rango intercuartílico: C3-C1= 3'75 - 3'25=0'5

valores atípicos leves inferiores: C1 - 1'5 * IQR = 3'25 - 1'5 * 0'5 = 2'5
valores atípicos leves superiores: C3 + 1'5 * IQR = 3'75 + 1'5 * 0'5 = 4

En último lugar, se comentará el caso de las cuotas acertadas, en las apuestas a favor de la victoria del equipo visitante(A):

```{r}
summary(datos[datos$FTR=="A",8])
plot(datos[datos$FTR=="A",5],datos[datos$FTR=="A",8],)
```

La cuota más alta acertada en una apuesta de victoria visitante, para esta casa de apuestas es de 16. Este encuentro corresponde al Juventus vs Sampdoria, jugado el 6 de enero de 2013, con victoria para el Sampdoria:
```{r}
datos[datos$FTR=="A" & datos$B365A>15,c(2,3,4,8)]
```

En el caso de las victorias visitantes, las cuotas normalmente suelen ser mejor pagadas que las apuestas a las victorias del local pero inferiores a las apuestas de empate. No obstante, se han encontrado aciertos con una cuota más extrema en apuestas al equipo visitante que al empate.
En este caso, las cuotas acertadas para equipos visitantes suelen oscilar entre 1'91 y 3'6, dando como outliers valores superiores a 6'135

C1: 1r cuartil
C3: 3R CUARTIL
IQR: Rango intercuartílico: C3 - C1 = 3'6 - 1'91 = 1'69

valores atípicos leves inferiores: C1 - 1'5 * IQR = 1'91 - 1'5 * 1'69 = -0'625(no se tendrá en cuenta)
valores atípicos leves superiores: C3 + 1'5 * IQR = 3'6 + 1'5 * 1'69 = 6'135

#MACHINE LEARNING

#Naive Bayes

```{r}
# Install these packages first if working on your computer

# Includes pairs.panels
# install.packages("psych", dependencies = TRUE)
library(psych)

# Includes naiveBayes
# install.packages("e1071", dependencies = TRUE)
library(e1071)

# Includes confusionMatrix
# install.packages("caret", dependencies = TRUE)
library(caret)

```

#Naive Bayes
```{r}
#Naive Bayes
```

```
datos
head(datos)
m <- naiveBayes (FTR ~ ., data = datos[,c(-1,-2,-3,-4)])  #
#mean(dada$ADMIp[dada$ind_class==1])-> 1r valor de admip de la clase 1

#m <- naiveBayes (FTR ~ ., data = datos[,5:8])  #

(tab.nb <- table(predict(m, datos[,c(-1,-2,-3,-4)]), datos[,5]))

tab.nb

sum(diag(tab.nb))/sum(tab.nb)#53'05% de acierto/45.6%
```

```{r}
#########################################################
#Jornades acertades
#########################################################


ofi <- datos$FTR #ofi:resultado oficial
pro <- predict(m, datos[,c(-1,-2,-3,-4)])#predicción(pronóstico)

acertados <- function(ofi,pro){
 s <- 0
 for(i in 1:length(ofi)){
   if(ofi[i]==pro[i]){
     s <- s +1
   }
 }
 return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/1900
```



```{r}


#########################################################
#Càlcul de guanys
#########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- predict(m, datos[,c(-1,-2,-3,-4)])#predicción(pronóstico)


guanys <- function(ofi,pro,datos){
 dineroB365 <- 0
 dineroBW   <- 0
 dineroIW   <- 0
 dineroLB   <- 0
 dineroPS   <- 0
 dineroWH   <- 0
 dineroVC   <- 0

 for(i in 1:length(ofi)){
   if(ofi[i]==pro[i]){

     if(ofi[i]== "H"){
       dineroB365 <- dineroB365 + (datos[i,6]-1)
       dineroBW   <- dineroBW   + (datos[i,9]-1)
       dineroIW   <- dineroIW   + (datos[i,12]-1)
       dineroLB   <- dineroLB   + (datos[i,15]-1)
       dineroPS   <- dineroPS   + (datos[i,18]-1)
       dineroWH   <- dineroWH   + (datos[i,21]-1)
       dineroVC   <- dineroVC   + (datos[i,24]-1)

     }else if(ofi[i]=="D"){
```

```
        dineroB365 <- dineroB365 + (datos[i,7]-1)
        dineroBW   <- dineroBW   + (datos[i,10]-1)
        dineroIW   <- dineroIW   + (datos[i,13]-1)
        dineroLB   <- dineroLB   + (datos[i,16]-1)
        dineroPS   <- dineroPS   + (datos[i,19]-1)
        dineroWH   <- dineroWH   + (datos[i,22]-1)
        dineroVC   <- dineroVC   + (datos[i,25]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos[i,8]-1)
        dineroBW   <- dineroBW   + (datos[i,11]-1)
        dineroIW   <- dineroIW   + (datos[i,14]-1)
        dineroLB   <- dineroLB   + (datos[i,17]-1)
        dineroPS   <- dineroPS   + (datos[i,20]-1)
        dineroWH   <- dineroWH   + (datos[i,23]-1)
        dineroVC   <- dineroVC   + (datos[i,26]-1)

      }
   }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
   }


  }
   g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))
   a <- c(1,2,3)
  return(g)
   return(a)
}

guanys(ofi,pro,datos)
```

```{r}

############################################################
#Gráfico
############################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- predict(m, datos[,c(-1,-2,-3,-4)])#predicción(pronóstico)

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){
```

```
      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos[i,6]-1)
        dineroBW   <- dineroBW   + (datos[i,9]-1)
        dineroIW   <- dineroIW   + (datos[i,12]-1)
        dineroLB   <- dineroLB   + (datos[i,15]-1)
        dineroPS   <- dineroPS   + (datos[i,18]-1)
        dineroWH   <- dineroWH   + (datos[i,21]-1)
        dineroVC   <- dineroVC   + (datos[i,24]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos[i,7]-1)
        dineroBW   <- dineroBW   + (datos[i,10]-1)
        dineroIW   <- dineroIW   + (datos[i,13]-1)
        dineroLB   <- dineroLB   + (datos[i,16]-1)
        dineroPS   <- dineroPS   + (datos[i,19]-1)
        dineroWH   <- dineroWH   + (datos[i,22]-1)
        dineroVC   <- dineroVC   + (datos[i,25]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos[i,8]-1)
        dineroBW   <- dineroBW   + (datos[i,11]-1)
        dineroIW   <- dineroIW   + (datos[i,14]-1)
        dineroLB   <- dineroLB   + (datos[i,17]-1)
        dineroPS   <- dineroPS   + (datos[i,20]-1)
        dineroWH   <- dineroWH   + (datos[i,23]-1)
        dineroVC   <- dineroVC   + (datos[i,26]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }

    B365[i] <- dineroB365
    BW[i] <- dineroBW
    IW[i] <- dineroIW
    LB[i] <- dineroLB
    PS[i] <- dineroPS
    WH[i] <- dineroWH
    VC[i] <- dineroVC

  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

    graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos)
graf

```

```{r}
 plot(1:1900, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
 ylim=c(-100,100),xlim=c(0,2000))

lines(1:1900,graf$BW,type="l",pch=2,col=3)
lines(1:1900,graf$IW,type="l",pch=2,col=4)
lines(1:1900,graf$LB,type="l",pch=2,col=5)
lines(1:1900,graf$PS,type="l",pch=2,col=6)
lines(1:1900,graf$WH,type="l",pch=2,col=7)
```

```
lines(1:1900,graf$VC,type="l",pch=2,col=8)
lines(1:1900,rep(0,1900),type="l",pch=2,col=1)

 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=1:7,xjust=1)


```

La gráfica muestra, que si se realizan todas las apuestas en la casa PS, siguiendo como predictor/clasificador Naive Bayes, se obtendrá un beneficio neto positivo.
Faltaria añadir la linea en la gráfica que mostrará el beneficio neto si se apuesta en cada partido a la casa con la cuota más alta. Claramente seria el beneficio neto más alto.

También falta hacer el grafico teniendo en cuenta todos los metodos de predicción para una casa en concreto.(en el markdown graf)

#Gráfico apostando a la mejor cuota

```{r}
```
```
##########################################################
#Càlcul de guanys
##########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- predict(m, datos[,c(-1,-2,-3,-4)])#predicción(pronóstico)

B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0


guanys <- function(ofi,pro,datos){
 dinero <- 0


 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
     dinero <- dinero + max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,6]-1){
      B365 <- B365 + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,9]-1){
      BW <- BW + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,12]-1){
      IW <- IW + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,15]-1){
      LB <- LB + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,18]-1){
      PS <- PS + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,21]-1){
      WH <- WH + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,24]-1){
      VC <- VC + 1
```

```
                        }


    }else if(ofi[i]=="D"){

     dinero <- dinero + max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))
     if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,7]-1){
      B365 <- B365 + 1
     }
     if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,10]-1){
      BW <- BW + 1
     }
      if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,13]-1){
      IW <- IW + 1
     }
     if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,16]-1){
      LB <- LB + 1
     }
     if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,19]-1){
      PS <- PS + 1
     }
     if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,22]-1){
      WH <- WH + 1
     }
     if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,25]-1){
      VC <- VC + 1
                        }

    }else if(ofi[i]=="A"){

     dinero <- dinero + max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))

    if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,8]-1){
      B365 <- B365 + 1          }
     if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,11]-1){
      BW <- BW + 1
     }
      if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,14]-1){
      IW <- IW + 1
     }
     if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,17]-1){
      LB <- LB + 1
     }
     if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,20]-1){
      PS <- PS + 1
     }
     if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,23]-1){
      WH <- WH + 1
     }
     if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,26]-1){
      VC <- VC + 1
                        }

   }
  }else{dinero <- dinero - 1
```

```
   }


 }
   g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
 return(g)
}

guanys(ofi,pro,datos)


```
#Programar una funcion que devuelve cuantas veces se ha apostado en cada casa


```{r}
par(mfrow=c(2,3))
for (i in c(8,11,14,17,23,26)){
plot(datos[,c(i,20)])#no tendria que ser plot(datos[,c(8,26)])
abline(c(0,1),col=2)
}
par(mfrow=c(1,1))

```

```{r}
#comparació entre cases
sum(datos$PSA-datos$B365A)
sum(datos$PSA-datos$BWA)
sum(datos$PSA-datos$IWA)
sum(datos$PSA-datos$LBA)
sum(datos$PSA-datos$WHA)
sum(datos$PSA-datos$VCA)
```

```{r}
#si apostem al equip amb la quota més baixa:
predi=function(cuota){c("H","D","A")[which.min(cuota)]}
pr=apply(datos[,6:8],1,predi)#per cada fila diu quina és la quota mínima
(tab.nb <- table(pr, datos[,5]))
sum(diag(tab.nb))/sum(tab.nb)#0.5489474



```

```{r}
##########################################################
#Càlcul de guanys
##########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- pr


guanys <- function(ofi,pro,datos){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
```

```
 dineroVC   <- 0

 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
      dineroB365 <- dineroB365 + (datos[i,6]-1)
      dineroBW   <- dineroBW   + (datos[i,9]-1)
      dineroIW   <- dineroIW   + (datos[i,12]-1)
      dineroLB   <- dineroLB   + (datos[i,15]-1)
      dineroPS   <- dineroPS   + (datos[i,18]-1)
      dineroWH   <- dineroWH   + (datos[i,21]-1)
      dineroVC   <- dineroVC   + (datos[i,24]-1)

    }else if(ofi[i]=="D"){
      dineroB365 <- dineroB365 + (datos[i,7]-1)
      dineroBW   <- dineroBW   + (datos[i,10]-1)
      dineroIW   <- dineroIW   + (datos[i,13]-1)
      dineroLB   <- dineroLB   + (datos[i,16]-1)
      dineroPS   <- dineroPS   + (datos[i,19]-1)
      dineroWH   <- dineroWH   + (datos[i,22]-1)
      dineroVC   <- dineroVC   + (datos[i,25]-1)

    }else if(ofi[i]=="A"){
      dineroB365 <- dineroB365 + (datos[i,8]-1)
      dineroBW   <- dineroBW   + (datos[i,11]-1)
      dineroIW   <- dineroIW   + (datos[i,14]-1)
      dineroLB   <- dineroLB   + (datos[i,17]-1)
      dineroPS   <- dineroPS   + (datos[i,20]-1)
      dineroWH   <- dineroWH   + (datos[i,23]-1)
      dineroVC   <- dineroVC   + (datos[i,26]-1)

    }
  }else{dineroB365 <- dineroB365 - 1
      dineroBW   <- dineroBW   - 1
      dineroIW   <- dineroIW   - 1
      dineroLB   <- dineroLB   - 1
      dineroPS   <- dineroPS   - 1
      dineroWH   <- dineroWH   - 1
      dineroVC   <- dineroVC   - 1
  }


 }
   g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))
 return(g)
}

guanys(ofi,pro,datos)
```



```{r}


##########################################################
#Gráfico
##########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- pr#predicción(pronóstico)

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()
```

```
guanys <- function(ofi,pro,datos){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

       if(ofi[i]== "H"){
         dineroB365 <- dineroB365 + (datos[i,6]-1)
         dineroBW   <- dineroBW   + (datos[i,9]-1)
         dineroIW   <- dineroIW   + (datos[i,12]-1)
         dineroLB   <- dineroLB   + (datos[i,15]-1)
         dineroPS   <- dineroPS   + (datos[i,18]-1)
         dineroWH   <- dineroWH   + (datos[i,21]-1)
         dineroVC   <- dineroVC   + (datos[i,24]-1)

       }else if(ofi[i]=="D"){
         dineroB365 <- dineroB365 + (datos[i,7]-1)
         dineroBW   <- dineroBW   + (datos[i,10]-1)
         dineroIW   <- dineroIW   + (datos[i,13]-1)
         dineroLB   <- dineroLB   + (datos[i,16]-1)
         dineroPS   <- dineroPS   + (datos[i,19]-1)
         dineroWH   <- dineroWH   + (datos[i,22]-1)
         dineroVC   <- dineroVC   + (datos[i,25]-1)

       }else if(ofi[i]=="A"){
         dineroB365 <- dineroB365 + (datos[i,8]-1)
         dineroBW   <- dineroBW   + (datos[i,11]-1)
         dineroIW   <- dineroIW   + (datos[i,14]-1)
         dineroLB   <- dineroLB   + (datos[i,17]-1)
         dineroPS   <- dineroPS   + (datos[i,20]-1)
         dineroWH   <- dineroWH   + (datos[i,23]-1)
         dineroVC   <- dineroVC   + (datos[i,26]-1)

       }
    }else{dineroB365 <- dineroB365 - 1
         dineroBW   <- dineroBW   - 1
         dineroIW   <- dineroIW   - 1
         dineroLB   <- dineroLB   - 1
         dineroPS   <- dineroPS   - 1
         dineroWH   <- dineroWH   - 1
         dineroVC   <- dineroVC   - 1
    }

    B365[i] <- dineroB365
    BW[i] <- dineroBW
    IW[i] <- dineroIW
    LB[i] <- dineroLB
    PS[i] <- dineroPS
    WH[i] <- dineroWH
    VC[i] <- dineroVC

  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

    graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos)
graf

```
```
```

```{r}
plot(1:1900, graf$B365,type="l",
pch=100, col=2, xlab="Partido",
ylab="Retorno neto(???)",
main="Beneficio neto según la casa de apuesta",
ylim=c(-100,100),xlim=c(0,2000))

lines(1:1900,graf$BW,type="l",pch=2,col=3)
lines(1:1900,graf$IW,type="l",pch=2,col=4)
lines(1:1900,graf$LB,type="l",pch=2,col=5)
lines(1:1900,graf$PS,type="l",pch=2,col=6)
lines(1:1900,graf$WH,type="l",pch=2,col=7)
lines(1:1900,graf$VC,type="l",pch=2,col=8)
lines(1:1900,rep(0,1900),type="l",pch=2,col=1)

#legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
#pch=c(1,1),col=2:8,xjust=1)


```


#Gráfico apostando a la mejor cuota

```{r}
##########################################################
#Càlcul de guanys
##########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- pr#predicción(pronóstico)

B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0


guanys <- function(ofi,pro,datos){
 dinero <- 0


 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
     dinero <- dinero + max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,6]-1){
      B365 <- B365 + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,9]-1){
      BW <- BW + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,12]-1){
      IW <- IW + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,15]-1){
      LB <- LB + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,18]-1){
      PS <- PS + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
```

```
                        datos[i,24]-1))==datos[i,21]-1){
  WH <- WH + 1
 }
 if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                        datos[i,24]-1))==datos[i,24]-1){
  VC <- VC + 1
                        }


}else if(ofi[i]=="D"){

 dinero <- dinero + max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,7]-1){
  B365 <- B365 + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,10]-1){
  BW <- BW + 1
 }
  if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,13]-1){
  IW <- IW + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,16]-1){
  LB <- LB + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,19]-1){
  PS <- PS + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,22]-1){
  WH <- WH + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,25]-1){
  VC <- VC + 1
                        }

}else if(ofi[i]=="A"){

 dinero <- dinero + max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))

 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,8]-1){
  B365 <- B365 + 1          }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,11]-1){
  BW <- BW + 1
 }
  if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,14]-1){
  IW <- IW + 1
 }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,17]-1){
  LB <- LB + 1
 }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,20]-1){
  PS <- PS + 1
 }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,23]-1){
  WH <- WH + 1
 }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
```

```
                            datos[i,26]-1))==datos[i,26]-1){
        VC <- VC + 1
                            }

     }
   }else{dinero <- dinero - 1

   }


 }
   g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
 return(g)
}

guanys(ofi,pro,datos)


```

#KNN
```{r}

##########################################

# k mejor?

library(class)
s <- data.frame(k=c(),prob=c())

for(i in c(1:10,20,30,40,50,60,70,80,90,100)){

myknn.cv <- knn.cv(datos[,-c(1,2,3,4,5)], datos[,5], k =i, prob = F)
(tab.knn <- table(myknn.cv, datos[,5]))
tab.knn
p <-sum(diag(tab.knn))/sum(tab.knn)#0.52%

s[i,1] <-i
s[i,2] <-p

}
s



##########################################
#knn k=5

library(class)

myknn.cv <- knn.cv(datos[,-c(1,2,3,4,5)], datos[,5], k = 5, prob = F)
tab.knn <- table(myknn.cv, datos[,5])
tab.knn
sum(diag(tab.knn))/sum(tab.knn)#0.52%


##########################################

#knn k=7

library(class)

myknn.cv <- knn.cv(datos[,-c(1,2,3,4,5)], datos[,5], k = 7, prob = F)
tab.knn <- table(myknn.cv, datos[,5])
tab.knn
sum(diag(tab.knn))/sum(tab.knn) #0.546%

##########################################
##########################################

#knn k=10
```

```
library(class)

myknn.cv <- knn.cv(datos[,-c(1,2,3,4,5)], datos[,5], k = 10, prob = F)
tab.knn <- table(myknn.cv, datos[,5])
tab.knn
sum(diag(tab.knn))/sum(tab.knn)#0.52%


##########################################

#knn k=50

library(class)

myknn.cv <- knn.cv(datos[,-c(1,2,3,4,5)], datos[,5], k = 50, prob = F)
tab.knn <- table(myknn.cv, datos[,5])
tab.knn
sum(diag(tab.knn))/sum(tab.knn) #0.546%

##########################################

#knn k=100

library(class)

myknn.cv <- knn.cv(datos[,-c(1,2,3,4,5)], datos[,5], k = 20, prob = F)
tab.knn <- table(myknn.cv, datos[,5])
tab.knn
sum(diag(tab.knn))/sum(tab.knn) #0.537%
```

```{r}


##########################################
#percentatge d'encerts
##########################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- myknn.cv#predicción(pronóstico)

acertados <- function(ofi,pro){
  s <- 0
  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){
      s <- s +1
    }
  }
  return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/1900


#########################################################
#Càlcul de guanys
#########################################################




ofi <- datos$FTR #ofi:resultado oficial
pro <-myknn.cv  #predicción(pronóstico)


guanys <- function(ofi,pro,datos){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
```

```
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0



  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos[i,6]-1)
        dineroBW   <- dineroBW   + (datos[i,9]-1)
        dineroIW   <- dineroIW   + (datos[i,12]-1)
        dineroLB   <- dineroLB   + (datos[i,15]-1)
        dineroPS   <- dineroPS   + (datos[i,18]-1)
        dineroWH   <- dineroWH   + (datos[i,21]-1)
        dineroVC   <- dineroVC   + (datos[i,24]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos[i,7]-1)
        dineroBW   <- dineroBW   + (datos[i,10]-1)
        dineroIW   <- dineroIW   + (datos[i,13]-1)
        dineroLB   <- dineroLB   + (datos[i,16]-1)
        dineroPS   <- dineroPS   + (datos[i,19]-1)
        dineroWH   <- dineroWH   + (datos[i,22]-1)
        dineroVC   <- dineroVC   + (datos[i,25]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos[i,8]-1)
        dineroBW   <- dineroBW   + (datos[i,11]-1)
        dineroIW   <- dineroIW   + (datos[i,14]-1)
        dineroLB   <- dineroLB   + (datos[i,17]-1)
        dineroPS   <- dineroPS   + (datos[i,20]-1)
        dineroWH   <- dineroWH   + (datos[i,23]-1)
        dineroVC   <- dineroVC   + (datos[i,26]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }


  }
    g <- data.frame(casa =c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,
                                                  dineroPS,dineroWH,dineroVC))

    return(g)
  }


  guanys(ofi,pro,datos)

```
```
Conclusions importants:

Tot i que amb k=10, tenim un percentatge d'encert menor que amb k=50 i k=100, s'observa que els beneficis són més alts.
Això és degut a que s'encerten partits on la quota és mes gran.

```{r}


##########################################################
#Gráfico
##########################################################
ofi <- datos$FTR #ofi:resultado oficial
```
```

```
pro <- myknn.cv#predicción(pronóstico)

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

       if(ofi[i]== "H"){
         dineroB365 <- dineroB365 + (datos[i,6]-1)
         dineroBW   <- dineroBW   + (datos[i,9]-1)
         dineroIW   <- dineroIW   + (datos[i,12]-1)
         dineroLB   <- dineroLB   + (datos[i,15]-1)
         dineroPS   <- dineroPS   + (datos[i,18]-1)
         dineroWH   <- dineroWH   + (datos[i,21]-1)
         dineroVC   <- dineroVC   + (datos[i,24]-1)

       }else if(ofi[i]=="D"){
         dineroB365 <- dineroB365 + (datos[i,7]-1)
         dineroBW   <- dineroBW   + (datos[i,10]-1)
         dineroIW   <- dineroIW   + (datos[i,13]-1)
         dineroLB   <- dineroLB   + (datos[i,16]-1)
         dineroPS   <- dineroPS   + (datos[i,19]-1)
         dineroWH   <- dineroWH   + (datos[i,22]-1)
         dineroVC   <- dineroVC   + (datos[i,25]-1)

       }else if(ofi[i]=="A"){
         dineroB365 <- dineroB365 + (datos[i,8]-1)
         dineroBW   <- dineroBW   + (datos[i,11]-1)
         dineroIW   <- dineroIW   + (datos[i,14]-1)
         dineroLB   <- dineroLB   + (datos[i,17]-1)
         dineroPS   <- dineroPS   + (datos[i,20]-1)
         dineroWH   <- dineroWH   + (datos[i,23]-1)
         dineroVC   <- dineroVC   + (datos[i,26]-1)

       }
    }else{dineroB365 <- dineroB365 - 1
         dineroBW   <- dineroBW   - 1
         dineroIW   <- dineroIW   - 1
         dineroLB   <- dineroLB   - 1
         dineroPS   <- dineroPS   - 1
         dineroWH   <- dineroWH   - 1
         dineroVC   <- dineroVC   - 1
    }

    B365[i] <- dineroB365
    BW[i] <- dineroBW
    IW[i] <- dineroIW
    LB[i] <- dineroLB
    PS[i] <- dineroPS
    WH[i] <- dineroWH
    VC[i] <- dineroVC

  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))
```

```
   graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos)
graf

```

```{r}
 plot(1:1900, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
 ylim=c(-120,130),xlim=c(0,2000))

lines(1:1900,graf$BW,type="l",pch=2,col=3)
lines(1:1900,graf$IW,type="l",pch=2,col=4)
lines(1:1900,graf$LB,type="l",pch=2,col=5)
lines(1:1900,graf$PS,type="l",pch=2,col=6)
lines(1:1900,graf$WH,type="l",pch=2,col=7)
lines(1:1900,graf$VC,type="l",pch=2,col=8)
lines(1:1900,rep(0,1900),type="l",pch=2,col=1)

 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=1:7,xjust=1)


```


#Gráfico apostando a la mejor cuota

```{r}
############################################################
#Càlcul de guanys
############################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- myknn.cv#predicción(pronóstico)

B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0


guanys <- function(ofi,pro,datos){
  dinero <- 0


  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dinero <- dinero + max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                                 datos[i,24]-1))
        if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                                 datos[i,24]-1))==datos[i,6]-1){
          B365 <- B365 + 1
        }
        if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                                 datos[i,24]-1))==datos[i,9]-1){
          BW <- BW + 1
        }
        if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                                 datos[i,24]-1))==datos[i,12]-1){
```

```
  IW <- IW + 1
 }
 if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                        datos[i,24]-1))==datos[i,15]-1){
  LB <- LB + 1
 }
 if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                        datos[i,24]-1))==datos[i,18]-1){
  PS <- PS + 1
 }
 if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                        datos[i,24]-1))==datos[i,21]-1){
  WH <- WH + 1
 }
 if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                        datos[i,24]-1))==datos[i,24]-1){
  VC <- VC + 1
                        }


}else if(ofi[i]=="D"){

 dinero <- dinero + max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,7]-1){
  B365 <- B365 + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,10]-1){
  BW <- BW + 1
 }
  if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,13]-1){
  IW <- IW + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,16]-1){
  LB <- LB + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,19]-1){
  PS <- PS + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,22]-1){
  WH <- WH + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,25]-1){
  VC <- VC + 1
                        }

}else if(ofi[i]=="A"){

 dinero <- dinero + max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))

 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,8]-1){
  B365 <- B365 + 1          }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,11]-1){
  BW <- BW + 1
 }
  if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,14]-1){
  IW <- IW + 1
 }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,17]-1){
```

```
         LB <- LB + 1
        }
        if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,20]-1){
         PS <- PS + 1
        }
        if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,23]-1){
         WH <- WH + 1
        }
        if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,26]-1){
         VC <- VC + 1
                          }

      }
    }else{dinero <- dinero - 1

    }


  }
    g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(g)
}

guanys(ofi,pro,datos)


```` 


#EXTRA(ABAJO)


```{r}
# Install these packages first if working on your computer

# Includes pairs.panels
# install.packages("psych", dependencies = TRUE)
library("psych")

# Includes naiveBayes
# install.packages("e1071", dependencies = TRUE)
library("e1071")

# Includes confusionMatrix
# install.packages("caret", dependencies = TRUE)
library(caret)



# Let us look again at the data
wb <- datos[,-c(1,2,3,4)]
head(wb)
pairs.panels(wb)



# Split data into training and validation parts
# set random seed to some value so that results are consistent

set.seed(15092011)

wb.size <- length(datos$FTR)
wb.train.size <- round(wb.size * 0.7) # 70% for training
wb.validation.size <- wb.size - wb.train.size # The rest for testing
```

```
wb.train.idx <- sample(seq(1:wb.size), wb.train.size) # Indeces for training
wb.train.sample <- wb[wb.train.idx,]
wb.validation.sample <- wb[-wb.train.idx,]

##### Validate NB classifiers, check their performance and refine them

# Let's see the performance of all variables, excluding ind_class
classf <- naiveBayes(
  subset(wb.train.sample, select = -FTR),
  wb.train.sample$FTR, laplace=1)
classf
preds <- predict(classf,
              subset(wb.validation.sample, select = -FTR))
table(preds, wb.validation.sample$FTR)
round(sum(preds == wb.validation.sample$FTR, na.rm=TRUE) /
       length(wb.validation.sample$FTR), digits = 2)#~0.56%

# We could better report the performance using "caret" package
#   which has lots of other very useful functions in there
confusionMatrix(table(preds, wb.validation.sample$FTR))



##### And then compare NB with k-NN

# Create a simple k-NN classifier
# Install this package if working on your computer
# install.packages("class", dependencies = TRUE)
library(class)

preds <- knn(
  subset(wb.train.sample, select = -FTR),
  subset(wb.validation.sample, select = -FTR),
  factor(wb.train.sample$FTR),
  k = 3, prob=TRUE, use.all = TRUE)
confusionMatrix(table(preds, wb.validation.sample$FTR))#0.4579

preds <- knn(
  subset(wb.train.sample, select = -FTR),
  subset(wb.validation.sample, select = -FTR),
  factor(wb.train.sample$FTR),
  k = 10, prob=TRUE, use.all = TRUE)
confusionMatrix(table(preds, wb.validation.sample$FTR))#0.5281

preds <- knn(
  subset(wb.train.sample, select = -FTR),
  subset(wb.validation.sample, select = -FTR),
  factor(wb.train.sample$FTR),
  k = 20, prob=TRUE, use.all = TRUE)
confusionMatrix(table(preds, wb.validation.sample$FTR))#0.5474

preds <- knn(
  subset(wb.train.sample, select = -FTR),
  subset(wb.validation.sample, select = -FTR),
  factor(wb.train.sample$FTR),
  k = 50, prob=TRUE, use.all = TRUE)
confusionMatrix(table(preds, wb.validation.sample$FTR))#0.5561

preds <- knn(
  subset(wb.train.sample, select = -FTR),
  subset(wb.validation.sample, select = -FTR),
  factor(wb.train.sample$FTR),
  k = 100, prob=TRUE, use.all = TRUE)
confusionMatrix(table(preds, wb.validation.sample$FTR))#0.5544

```
#EXTRA(ARRIBA)


```
```

#Decision Trees

```{r}
######
####ARBOLES DE DECISION
######
set.seed(15092011)
library(rpart)

pl<-rpart(FTR~., datos[,-c(1,2,3,4)])#,cp=0.001
pl
plot(pl)
text(pl, use.n=T)
#predicciones para el data de entrenamiento
pll=predict(pl,data=datos[,-c(1,2,3,4,5)],type = "class")
(mat<-table(datos$FTR,pll))
error_rate<-(sum(mat)-sum(diag(mat)))/sum(mat)
sum(diag(mat))/sum(mat)#0.548%
```

#Árbol Óptimo

```{r}
#predicciones para el data de test
pll=predict(pl,newdata=dada_ler[,-18],type = "class")
(mat<-table(pll))


######
#PROFILING
######
dades$ind_class<-dada$ind_class
catdes(data.frame(dada$ind_class,dades[,3:19]),num.var=1,proba = 0.001)

######
#ARBOL OPTIMO(caso bcnses)
######
set.seed(15092011)
p2 = rpart(ind_class~.,data= dada[,-18], control=rpart.control(cp=0.001, xval=10))
p2
plot(p2)
text(p2,pretty = 0)
text(p2, use.n=T)
printcp(p2)

plot(p2$cptable[,2],p2$cptable[,3],type="l",xlab="size of the tree",ylab="Relative impurity",main="R(t)")
lines(p2$cptable[,2],p2$cptable[,4],col="blue")
legend("topright",c("R(T)training","R(T)cv"),col=c("black","blue"),lty=1)

#observando el grafico vemos que el cp que tiene 6 splits es de 0.010582
plotcp(p2)
printcp(p2)
prune(p2,cp = 0.010582)

#sin embargo recurimos a la tÃ©cnica dada en clase
p2$cptable = as.data.frame(p2$cptable)
ind = which.min(p2$cptable$xerror)

xerr <- p2$cptable$xerror[ind]
xstd <- p2$cptable$xstd[ind]

i = 1
while (p2$cptable$xerror[i] > xerr+xstd) i = i+1

alfa = p2$cptable$CP[i] #sacamos el parametro optimo

# sacamos el arbol resultante
p1 <- prune(p2,cp=alfa)
summary(p1)
plot(p1)
```

```
text(p1,use.n = T)
rpart.plot(p1)

#predicciones para los datos de entreno, del ano 1988
p11=predict(p1,data=dada[,-c(17,18)],type = "class")
(mat<-table(dada$ind_class,p11))
error_rate<-(sum(mat)-sum(diag(mat)))/sum(mat) #error_rate <- [1] 0.1169355
#predicciones para los datos de test, del aÃ±o 1996
pll=predict(p1,newdata=dada_ler[,-18],type = "class")
(mat<-table(pll))

# IMPORTANCE OF VARIABLES IN THE TREE DEFINITION
bp=barplot(p1$variable.importance, ylim = c(0,100),col=c("green","red"),main="Importancia por variable")
text(x=bp, y=p1$variable.importance, labels=round(p1$variable.importance,1), pos=3,xpd=NA)

###PERO SI LO DEJAMOS TAL CUAL OBTENEMOS UN ERROR RATE MÃ□S PEQUEÃ'O###









######
#ARBOL OPTIMO(caso futbol)
######
set.seed(15092011)
p2 = rpart(FTR~.,data= datos[,-c(1,2,3,4)], control=rpart.control(cp=0.001, xval=10))

plot(p2)
text(p2,pretty = 0)
text(p2, use.n=T)
printcp(p2)

plot(p2$cptable[,2],p2$cptable[,3],type="l",xlab="size of the tree",ylab="Relative impurity",main="R(t)")
lines(p2$cptable[,2],p2$cptable[,4],col="blue")
legend("topright",c("R(T)training","R(T)cv"),col=c("black","blue"),lty=1)

#observando el grafico vemos que el cp que tiene 6 splits es de 0.010582
plotcp(p2)
printcp(p2)
prune(p2,cp = 0.01)
prune(p2,cp = 0.0080828)

#sin embargo recurimos a la tÃ©cnica dada en clase
p2$cptable = as.data.frame(p2$cptable)
ind = which.min(p2$cptable$xerror)

xerr <- p2$cptable$xerror[ind]
xstd <- p2$cptable$xstd[ind]

i = 1
while (p2$cptable$xerror[i] > xerr+xstd) i = i+1

alfa = p2$cptable$CP[i] #sacamos el parametro optimo(0.008082768)

# sacamos el arbol resultante
p1 <- prune(p2,cp=alfa)
summary(p1)
plot(p1)
text(p1,use.n = T)
```

```
library(rpart.plot)
rpart.plot(p1)

#predicciones para los datos de entreno, del ano 1988
p11=predict(p1,data=datos[,-c(1,2,3,4,5)],type = "class")
(mat<-table(datos$FTR,p11))
error_rate<-(sum(mat)-sum(diag(mat)))/sum(mat) #error_rate <- [1] 0.1169355
1-error_rate
#predicciones para los datos de test, del aÃ±o 1996
pll=predict(p1,newdata=dada_ler[,-18],type = "class")
(mat<-table(pll))

# IMPORTANCE OF VARIABLES IN THE TREE DEFINITION
bp=barplot(p1$variable.importance, ylim = c(0,200),col=c("green","red"),main="Importancia por variable")
text(x=bp, y=p1$variable.importance, labels=round(p1$variable.importance,1), pos=3,xpd=NA)

###PERO SI LO DEJAMOS TAL CUAL OBTENEMOS UN ERROR RATE MÃ□S PEQUEÃ'O###
```



```{r}
##########################################
#percentatge d'encerts
##########################################
ofi <- datos$FTR #ofi:resultado oficial
#pro <- predict(pl,data=datos[,-c(1,2,3,4,5)],type = "class")#predicción(pronóstico)
pro <- p11

acertados <- function(ofi,pro){
 s <- 0
 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){
   s <- s +1
  }
 }
 return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/1900



#########################################################
#Càlcul de guanys
#########################################################
ofi <- datos$FTR #ofi:resultado oficial
#pro <- predict(pl,data=datos[,-c(1,2,3,4,5)],type = "class")#predicción(pronóstico)
pro <- p11


guanys <- function(ofi,pro,datos){
 dineroB365 <- 0
 dineroBW  <- 0
 dineroIW  <- 0
 dineroLB  <- 0
 dineroPS  <- 0
 dineroWH  <- 0
 dineroVC  <- 0

 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
     dineroB365 <- dineroB365 + (datos[i,6]-1)
     dineroBW  <- dineroBW  + (datos[i,9]-1)
     dineroIW  <- dineroIW  + (datos[i,12]-1)
     dineroLB  <- dineroLB  + (datos[i,15]-1)
     dineroPS  <- dineroPS  + (datos[i,18]-1)
     dineroWH  <- dineroWH  + (datos[i,21]-1)
```

```
        dineroVC   <- dineroVC   + (datos[i,24]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos[i,7]-1)
        dineroBW   <- dineroBW   + (datos[i,10]-1)
        dineroIW   <- dineroIW   + (datos[i,13]-1)
        dineroLB   <- dineroLB   + (datos[i,16]-1)
        dineroPS   <- dineroPS   + (datos[i,19]-1)
        dineroWH   <- dineroWH   + (datos[i,22]-1)
        dineroVC   <- dineroVC   + (datos[i,25]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos[i,8]-1)
        dineroBW   <- dineroBW   + (datos[i,11]-1)
        dineroIW   <- dineroIW   + (datos[i,14]-1)
        dineroLB   <- dineroLB   + (datos[i,17]-1)
        dineroPS   <- dineroPS   + (datos[i,20]-1)
        dineroWH   <- dineroWH   + (datos[i,23]-1)
        dineroVC   <- dineroVC   + (datos[i,26]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }


  }
    g <- data.frame(casa =c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,
                                                  dineroPS,dineroWH,dineroVC))
  return(g)
}

guanys(ofi,pro,datos)
```
```{r}


#########################################################
#Gráfico
#########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- p11#predicción(pronóstico)

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos[i,6]-1)
```

```
        dineroBW   <- dineroBW   + (datos[i,9]-1)
        dineroIW   <- dineroIW   + (datos[i,12]-1)
        dineroLB   <- dineroLB   + (datos[i,15]-1)
        dineroPS   <- dineroPS   + (datos[i,18]-1)
        dineroWH   <- dineroWH   + (datos[i,21]-1)
        dineroVC   <- dineroVC   + (datos[i,24]-1)

       }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos[i,7]-1)
        dineroBW   <- dineroBW   + (datos[i,10]-1)
        dineroIW   <- dineroIW   + (datos[i,13]-1)
        dineroLB   <- dineroLB   + (datos[i,16]-1)
        dineroPS   <- dineroPS   + (datos[i,19]-1)
        dineroWH   <- dineroWH   + (datos[i,22]-1)
        dineroVC   <- dineroVC   + (datos[i,25]-1)

       }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos[i,8]-1)
        dineroBW   <- dineroBW   + (datos[i,11]-1)
        dineroIW   <- dineroIW   + (datos[i,14]-1)
        dineroLB   <- dineroLB   + (datos[i,17]-1)
        dineroPS   <- dineroPS   + (datos[i,20]-1)
        dineroWH   <- dineroWH   + (datos[i,23]-1)
        dineroVC   <- dineroVC   + (datos[i,26]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }

    B365[i] <- dineroB365
    BW[i] <- dineroBW
    IW[i] <- dineroIW
    LB[i] <- dineroLB
    PS[i] <- dineroPS
    WH[i] <- dineroWH
    VC[i] <- dineroVC

  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

    graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos)
graf

```

```{r}
 plot(1:1900, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
 ylim=c(-100,100),xlim=c(0,2000))

lines(1:1900,graf$BW,type="l",pch=2,col=3)
lines(1:1900,graf$IW,type="l",pch=2,col=4)
lines(1:1900,graf$LB,type="l",pch=2,col=5)
lines(1:1900,graf$PS,type="l",pch=2,col=6)
lines(1:1900,graf$WH,type="l",pch=2,col=7)
lines(1:1900,graf$VC,type="l",pch=2,col=8)
lines(1:1900,rep(0,1900),type="l",pch=2,col=1)
```

```
 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=1:7,xjust=1)


```


#Gráfico apostando a la mejor cuota

```{r}
############################################################
#Càlcul de guanys
############################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- p11#predicción(pronóstico)

B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0


guanys <- function(ofi,pro,datos){
 dinero <- 0


 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
     dinero <- dinero + max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,6]-1){
      B365 <- B365 + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,9]-1){
      BW <- BW + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,12]-1){
      IW <- IW + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,15]-1){
      LB <- LB + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,18]-1){
      PS <- PS + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,21]-1){
      WH <- WH + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,24]-1){
      VC <- VC + 1
                              }


    }else if(ofi[i]=="D"){

     dinero <- dinero + max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                              datos[i,25]-1))
     if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                              datos[i,25]-1))==datos[i,7]-1){
```

```
        B365 <- B365 + 1
       }
       if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                          datos[i,25]-1))==datos[i,10]-1){
        BW <- BW + 1
       }
        if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                          datos[i,25]-1))==datos[i,13]-1){
        IW <- IW + 1
       }
       if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                          datos[i,25]-1))==datos[i,16]-1){
        LB <- LB + 1
       }
       if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                          datos[i,25]-1))==datos[i,19]-1){
        PS <- PS + 1
       }
       if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                          datos[i,25]-1))==datos[i,22]-1){
        WH <- WH + 1
       }
       if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                          datos[i,25]-1))==datos[i,25]-1){
        VC <- VC + 1
                          }

      }else if(ofi[i]=="A"){

       dinero <- dinero + max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))

      if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,8]-1){
        B365 <- B365 + 1          }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,11]-1){
        BW <- BW + 1
       }
        if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,14]-1){
        IW <- IW + 1
       }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,17]-1){
        LB <- LB + 1
       }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,20]-1){
        PS <- PS + 1
       }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,23]-1){
        WH <- WH + 1
       }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,26]-1){
        VC <- VC + 1
                          }

      }
    }else{dinero <- dinero - 1

    }


  }
   g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(g)
}
```

guanys(ofi,pro,datos)

```

La casa de apuestas WH, ha sido la casa en la cual más veces se ha escogido, Debido a que ofrecia cuotas más altas en las predicciones que nos ha dado el clasificador. Esto no quiere decir que sea la casa de apuestas que más beneficio da, ya que en todos los casos es PSA, debido a que en unos ciertos partidos, la diferencia de cuotas de PS respecto WC era más significativa, en cambio WH se ha escogido antes que PSA en partidos donde WH ofrecia una cuota unos centimos superior.

#otro
```{r}
p1 = rpart(FTR~., data=datos[,-c(1,2,3,4)])
p1
plot(p1)
text(p1, use.n=T)
p1l=predict(p1,data=dd[learn,])
p1lp=NULL
p1lp[p1l<0.5]=0
p1lp[p1l>=0.5]=1
table(dict[learn],p1lp)
p1t = predict(p1,newdata=dd[-learn,])
p1tp=NULL
p1tp[p1t<0.5]=0
p1tp[p1t>=0.5]=1
table(dict[-learn],p1tp)
```

#QDA

```{r}
plot(datos[, 6:26], col = datos$FTR, pch = 19)#no és lineal

```

```{r}
#no va bien
#################
#miramos si los predictores que usamos en nuestro modelo se distribuyen segun
#la distribucion normal multivariante
mlibrary(nortest)
levels(dada$ind_class)<-1:6
a<-dada[,-18]
for(j in 1:3){
for(i in 1:17){
  print(ad.test(datos[datos$FTR==j, i]))
}
}
#hay p-valores inferiores a 0.05
#observamos que hay variables que no siguen la distribucion normal, por tanto,
#hemos de recurrir al analisis de discriminacion cuadratica

```

```{r}
apriori<-as.vector(table(datos$FTR)/length(datos$FTR))
###
##discriminante
###
library(MASS)
#d1 <- qda(FTR~., datos[,-c(1,2,3,4)], prior = c(1,1,1)/3)
d1 <- qda(FTR~., datos[,-c(1,2,3,4)], prior = apriori)
#no va porque no se alcanza invertir la matriz


############
d1_88<-predict(d1,datos[,-c(1,2,3,4)],type="class")
(mat<-table(datos$FTR,d1_88$class))

```
error_rate<-(sum(mat)-sum(diag(mat)))/sum(mat)
sum(diag(mat))/sum(mat)#0.55%
```

```{r}
########################################
#percentatge d'encerts
########################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- d1_88$class
class(pro)

acertados <- function(ofi,pro){
 s <- 0
 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){
   s <- s +1
  }
 }
 return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/1900
```

```{r}
#
#########################################################
#Càlcul de guanys
#########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- d1_88$class


guanys <- function(ofi,pro,datos){
 dineroB365 <- 0
 dineroBW  <- 0
 dineroIW  <- 0
 dineroLB  <- 0
 dineroPS  <- 0
 dineroWH  <- 0
 dineroVC  <- 0

 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
     dineroB365 <- dineroB365 + (datos[i,6]-1)
     dineroBW  <- dineroBW  + (datos[i,9]-1)
     dineroIW  <- dineroIW  + (datos[i,12]-1)
     dineroLB  <- dineroLB  + (datos[i,15]-1)
     dineroPS  <- dineroPS  + (datos[i,18]-1)
     dineroWH  <- dineroWH  + (datos[i,21]-1)
     dineroVC  <- dineroVC  + (datos[i,24]-1)

    }else if(ofi[i]=="D"){
     dineroB365 <- dineroB365 + (datos[i,7]-1)
     dineroBW  <- dineroBW  + (datos[i,10]-1)
     dineroIW  <- dineroIW  + (datos[i,13]-1)
     dineroLB  <- dineroLB  + (datos[i,16]-1)
     dineroPS  <- dineroPS  + (datos[i,19]-1)
     dineroWH  <- dineroWH  + (datos[i,22]-1)
     dineroVC  <- dineroVC  + (datos[i,25]-1)

    }else if(ofi[i]=="A"){
     dineroB365 <- dineroB365 + (datos[i,8]-1)
     dineroBW  <- dineroBW  + (datos[i,11]-1)
     dineroIW  <- dineroIW  + (datos[i,14]-1)
     dineroLB  <- dineroLB  + (datos[i,17]-1)
     dineroPS  <- dineroPS  + (datos[i,20]-1)
```

```
      dineroWH   <- dineroWH   + (datos[i,23]-1)
      dineroVC   <- dineroVC   + (datos[i,26]-1)

    }
  }else{dineroB365 <- dineroB365 - 1
      dineroBW   <- dineroBW   - 1
      dineroIW   <- dineroIW   - 1
      dineroLB   <- dineroLB   - 1
      dineroPS   <- dineroPS   - 1
      dineroWH   <- dineroWH   - 1
      dineroVC   <- dineroVC   - 1
  }


 }
  g <- data.frame(casa =c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,
                                                dineroPS,dineroWH,dineroVC))
 return(g)
}

guanys(ofi,pro,datos)




```

```{r}


#########################################################
#Gráfico
#########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- d1_88$class#predicción(pronóstico)

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos){
 dineroB365 <- 0
 dineroBW   <- 0
 dineroIW   <- 0
 dineroLB   <- 0
 dineroPS   <- 0
 dineroWH   <- 0
 dineroVC   <- 0

 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
      dineroB365 <- dineroB365 + (datos[i,6]-1)
      dineroBW   <- dineroBW   + (datos[i,9]-1)
      dineroIW   <- dineroIW   + (datos[i,12]-1)
      dineroLB   <- dineroLB   + (datos[i,15]-1)
      dineroPS   <- dineroPS   + (datos[i,18]-1)
      dineroWH   <- dineroWH   + (datos[i,21]-1)
      dineroVC   <- dineroVC   + (datos[i,24]-1)

    }else if(ofi[i]=="D"){
      dineroB365 <- dineroB365 + (datos[i,7]-1)
      dineroBW   <- dineroBW   + (datos[i,10]-1)
      dineroIW   <- dineroIW   + (datos[i,13]-1)
```

```
        dineroLB   <- dineroLB   + (datos[i,16]-1)
        dineroPS   <- dineroPS   + (datos[i,19]-1)
        dineroWH   <- dineroWH   + (datos[i,22]-1)
        dineroVC   <- dineroVC   + (datos[i,25]-1)

       }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos[i,8]-1)
        dineroBW   <- dineroBW   + (datos[i,11]-1)
        dineroIW   <- dineroIW   + (datos[i,14]-1)
        dineroLB   <- dineroLB   + (datos[i,17]-1)
        dineroPS   <- dineroPS   + (datos[i,20]-1)
        dineroWH   <- dineroWH   + (datos[i,23]-1)
        dineroVC   <- dineroVC   + (datos[i,26]-1)

       }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }

    B365[i] <- dineroB365
    BW[i] <- dineroBW
    IW[i] <- dineroIW
    LB[i] <- dineroLB
    PS[i] <- dineroPS
    WH[i] <- dineroWH
    VC[i] <- dineroVC

  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

    graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos)
graf

```

```{r}
 plot(1:1900, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
 ylim=c(-110,110),xlim=c(0,2000))

lines(1:1900,graf$BW,type="l",pch=2,col=3)
lines(1:1900,graf$IW,type="l",pch=2,col=4)
lines(1:1900,graf$LB,type="l",pch=2,col=5)
lines(1:1900,graf$PS,type="l",pch=2,col=6)
lines(1:1900,graf$WH,type="l",pch=2,col=7)
lines(1:1900,graf$VC,type="l",pch=2,col=8)
lines(1:1900,rep(0,1900),type="l",pch=2,col=1)

 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=1:7,xjust=1)


```


#Gráfico apostando a la mejor cuota

```{r}
##########################################################
```

```
#Càlcul de guanys
########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- d1_88$class#predicción(pronóstico)

B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0


guanys <- function(ofi,pro,datos){
 dinero <- 0


 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
     dinero <- dinero + max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,6]-1){
      B365 <- B365 + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,9]-1){
      BW <- BW + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,12]-1){
      IW <- IW + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,15]-1){
      LB <- LB + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,18]-1){
      PS <- PS + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,21]-1){
      WH <- WH + 1
     }
     if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                              datos[i,24]-1))==datos[i,24]-1){
      VC <- VC + 1
                              }


    }else if(ofi[i]=="D"){

     dinero <- dinero + max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                              datos[i,25]-1))
     if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                              datos[i,25]-1))==datos[i,7]-1){
      B365 <- B365 + 1
     }
     if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                              datos[i,25]-1))==datos[i,10]-1){
      BW <- BW + 1
     }
      if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                              datos[i,25]-1))==datos[i,13]-1){
      IW <- IW + 1
     }
     if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
```

```
                        datos[i,25]-1))==datos[i,16]-1){
        LB <- LB + 1
       }
       if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,19]-1){
        PS <- PS + 1
       }
       if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,22]-1){
        WH <- WH + 1
       }
       if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,25]-1){
        VC <- VC + 1
                        }

      }else if(ofi[i]=="A"){

       dinero <- dinero + max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))

      if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,8]-1){
        B365 <- B365 + 1          }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,11]-1){
        BW <- BW + 1
       }
        if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,14]-1){
        IW <- IW + 1
       }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,17]-1){
        LB <- LB + 1
       }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,20]-1){
        PS <- PS + 1
       }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,23]-1){
        WH <- WH + 1
       }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,26]-1){
        VC <- VC + 1
                        }

      }
   }else{dinero <- dinero - 1

   }


 }
   g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
 return(g)
}

guanys(ofi,pro,datos)


```


```{r}
#SVM
library (e1071)
```

```
## classification mode
# default with factor response:
#model <- svm(datos$FTR ~ ., data = datos[,-c(1,2,3,4)])

# alternatively the traditional interface:
#x <- subset(iris, select = -Species)
x <- datos[,-c(1,2,3,4,5)]
y <- datos[,5]
model <- svm(x, y)

#model <- svm(datos[,-c(1,2,3,4,5)],datos[,5])

summary(model)

# test with train data
pred <- predict(model, x)
# (same as:) pred <- fitted(model)

# Check accuracy:
tt <-table(pred, y)
sum(diag(tt))/sum(tt)#0.5484211
```


```{r}
#otro svm
svm.model <- svm(FTR ~ ., data=datos[,-c(1,2,3,4)], type="C-classification",
 kernel="radial", cost=1,gamma=0.04761904762)
 predict <- fitted(svm.model)
 cm <- table(predict, datos$FTR)
 cm
 accuracy <- sum(diag(cm))/sum(cm)
accuracy
```



```{r}



##########################################
#percentatge d'encerts
##########################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- pred#predicción(pronóstico)

acertados <- function(ofi,pro){
  s <- 0
  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){
      s <- s +1
    }
  }
  return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/1900
#########################################################
#Càlcul de guanys
#########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- pred#predicción(pronóstico)


guanys <- function(ofi,pro,datos){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
```

```
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
   if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos[i,6]-1)
        dineroBW   <- dineroBW   + (datos[i,9]-1)
        dineroIW   <- dineroIW   + (datos[i,12]-1)
        dineroLB   <- dineroLB   + (datos[i,15]-1)
        dineroPS   <- dineroPS   + (datos[i,18]-1)
        dineroWH   <- dineroWH   + (datos[i,21]-1)
        dineroVC   <- dineroVC   + (datos[i,24]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos[i,7]-1)
        dineroBW   <- dineroBW   + (datos[i,10]-1)
        dineroIW   <- dineroIW   + (datos[i,13]-1)
        dineroLB   <- dineroLB   + (datos[i,16]-1)
        dineroPS   <- dineroPS   + (datos[i,19]-1)
        dineroWH   <- dineroWH   + (datos[i,22]-1)
        dineroVC   <- dineroVC   + (datos[i,25]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos[i,8]-1)
        dineroBW   <- dineroBW   + (datos[i,11]-1)
        dineroIW   <- dineroIW   + (datos[i,14]-1)
        dineroLB   <- dineroLB   + (datos[i,17]-1)
        dineroPS   <- dineroPS   + (datos[i,20]-1)
        dineroWH   <- dineroWH   + (datos[i,23]-1)
        dineroVC   <- dineroVC   + (datos[i,26]-1)

      }
   }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
   }


  }
    g <- data.frame(casa =c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,
                                                  dineroPS,dineroWH,dineroVC))
  return(g)
}

guanys(ofi,pro,datos)


```


```{r}


##########################################################
#Gráfico
##########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- pred#predicción(pronóstico)

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
```

```
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

       if(ofi[i]== "H"){
         dineroB365 <- dineroB365 + (datos[i,6]-1)
         dineroBW   <- dineroBW   + (datos[i,9]-1)
         dineroIW   <- dineroIW   + (datos[i,12]-1)
         dineroLB   <- dineroLB   + (datos[i,15]-1)
         dineroPS   <- dineroPS   + (datos[i,18]-1)
         dineroWH   <- dineroWH   + (datos[i,21]-1)
         dineroVC   <- dineroVC   + (datos[i,24]-1)

       }else if(ofi[i]=="D"){
         dineroB365 <- dineroB365 + (datos[i,7]-1)
         dineroBW   <- dineroBW   + (datos[i,10]-1)
         dineroIW   <- dineroIW   + (datos[i,13]-1)
         dineroLB   <- dineroLB   + (datos[i,16]-1)
         dineroPS   <- dineroPS   + (datos[i,19]-1)
         dineroWH   <- dineroWH   + (datos[i,22]-1)
         dineroVC   <- dineroVC   + (datos[i,25]-1)

       }else if(ofi[i]=="A"){
         dineroB365 <- dineroB365 + (datos[i,8]-1)
         dineroBW   <- dineroBW   + (datos[i,11]-1)
         dineroIW   <- dineroIW   + (datos[i,14]-1)
         dineroLB   <- dineroLB   + (datos[i,17]-1)
         dineroPS   <- dineroPS   + (datos[i,20]-1)
         dineroWH   <- dineroWH   + (datos[i,23]-1)
         dineroVC   <- dineroVC   + (datos[i,26]-1)

       }
    }else{dineroB365 <- dineroB365 - 1
         dineroBW   <- dineroBW   - 1
         dineroIW   <- dineroIW   - 1
         dineroLB   <- dineroLB   - 1
         dineroPS   <- dineroPS   - 1
         dineroWH   <- dineroWH   - 1
         dineroVC   <- dineroVC   - 1
    }

    B365[i] <- dineroB365
    BW[i] <- dineroBW
    IW[i] <- dineroIW
    LB[i] <- dineroLB
    PS[i] <- dineroPS
    WH[i] <- dineroWH
    VC[i] <- dineroVC

  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

    graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos)
```

```
graf

```

```{r}
 plot(1:1900, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
 ylim=c(-100,100),xlim=c(0,2000))

lines(1:1900,graf$BW,type="l",pch=2,col=3)
lines(1:1900,graf$IW,type="l",pch=2,col=4)
lines(1:1900,graf$LB,type="l",pch=2,col=5)
lines(1:1900,graf$PS,type="l",pch=2,col=6)
lines(1:1900,graf$WH,type="l",pch=2,col=7)
lines(1:1900,graf$VC,type="l",pch=2,col=8)
lines(1:1900,rep(0,1900),type="l",pch=2,col=1)

 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=2:8,xjust=1)


```


#Gráfico apostando a la mejor cuota

```{r}
#########################################################
#Càlcul de guanys
#########################################################
ofi <- datos$FTR #ofi:resultado oficial
pro <- pred#predicción(pronóstico)

B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0


guanys <- function(ofi,pro,datos){
  dinero <- 0


  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dinero <- dinero + max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                                 datos[i,24]-1))
        if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                                 datos[i,24]-1))==datos[i,6]-1){
          B365 <- B365 + 1
        }
        if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                                 datos[i,24]-1))==datos[i,9]-1){
          BW <- BW + 1
        }
        if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                                 datos[i,24]-1))==datos[i,12]-1){
          IW <- IW + 1
        }
        if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                                 datos[i,24]-1))==datos[i,15]-1){
          LB <- LB + 1
        }
        if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
```

```
                        datos[i,24]-1))==datos[i,18]-1){
  PS <- PS + 1
 }
 if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                        datos[i,24]-1))==datos[i,21]-1){
  WH <- WH + 1
 }
 if(max(c(datos[i,6]-1,datos[i,9]-1,datos[i,12]-1,datos[i,15]-1,datos[i,18]-1,datos[i,21]-1,
                        datos[i,24]-1))==datos[i,24]-1){
  VC <- VC + 1
                        }


}else if(ofi[i]=="D"){

 dinero <- dinero + max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,7]-1){
  B365 <- B365 + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,10]-1){
  BW <- BW + 1
 }
  if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,13]-1){
  IW <- IW + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,16]-1){
  LB <- LB + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,19]-1){
  PS <- PS + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,22]-1){
  WH <- WH + 1
 }
 if(max(c(datos[i,7]-1,datos[i,10]-1,datos[i,13]-1,datos[i,16]-1,datos[i,19]-1,datos[i,22]-1,
                        datos[i,25]-1))==datos[i,25]-1){
  VC <- VC + 1
                        }

}else if(ofi[i]=="A"){

 dinero <- dinero + max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))

 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,8]-1){
  B365 <- B365 + 1          }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,11]-1){
  BW <- BW + 1
 }
  if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,14]-1){
  IW <- IW + 1
 }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,17]-1){
  LB <- LB + 1
 }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                        datos[i,26]-1))==datos[i,20]-1){
  PS <- PS + 1
 }
 if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
```

```
                          datos[i,26]-1))==datos[i,23]-1){
        WH <- WH + 1
       }
       if(max(c(datos[i,8]-1,datos[i,11]-1,datos[i,14]-1,datos[i,17]-1,datos[i,20]-1,datos[i,23]-1,
                          datos[i,26]-1))==datos[i,26]-1){
        VC <- VC + 1
                          }

     }
  }else{dinero <- dinero - 1

  }


 }
  g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
 return(g)
}

guanys(ofi,pro,datos)


```


#svm(tune)
```{r}
#SVM con Kernel Lineal
#Se procede a dejar una semilla para hacer una elección del mejor valor de parámetro cost

set.seed (1)
#Se hace cross-validation de k-fold, con k=10. Esto mediante la función tune()


tune.out=tune(svm,FTR~.,datos[,-c(1,2,3,4)],kernel="linear",type='C-classification',scale=FALSE,ranges=list(cost=c(0.001,0.01)))
tune.out
summary(tune.out)

#Se elige el mejor modelo con el mejor valor para el parámetro cost

bestmod=tune.out$best.model
summary(bestmod)
# Gráfica del mejor modelo con kernel lineal

plot(bestmod,Datos1)

#Predicción
Pred=predict(bestmod,Datos1)

table(predicción=Pred,Valores_reales=Datos1$Clus)



#Kernel radial
set.seed (1)
tune.out=tune(svm,factor(Clus)~.,data=Datos1,kernel="radial",type='C-classification',scale=FALSE,ranges=list(cost=c(0.001,0.01))
tune.out
summary(tune.out)
bestmod_r=tune.out$best.model
summary(bestmod_r)
plot(bestmod_r,Datos1)
Pred=predict(bestmod_r,Datos1)
table(predicción=Pred,Valores_reales=Datos1$Clus)
```

```
data(iris)
 ## tune `svm' for classification with RBF-kernel (default in svm),
 ## using one split for training/validation set

 obj <- tune(svm, Species~., data = iris,
          ranges = list(gamma = 2^(-1:1), cost = 2^(2:4)),
          tunecontrol = tune.control(sampling = "fix")
         )

 ## alternatively:
 ## obj <- tune.svm(Species~., data = iris, gamma = 2^(-1:1), cost = 2^(2:4))

 summary(obj)
 plot(obj)




 datos
 ## tune `svm' for classification with RBF-kernel (default in svm),
 ## using one split for training/validation set

 obj <- tune(svm, FTR ~ ., data = datos[,-c(1,2,3,4)],
          ranges = list(gamma = 2^(-1:1), cost = 2^(2:4)),
          tunecontrol = tune.control(sampling = "fix")
         )

 ## alternatively:
 ## obj <- tune.svm(Species~., data = iris, gamma = 2^(-1:1), cost = 2^(2:4))

 summary(obj)
 plot(obj)


```
```


#Redes neuronales
```{r}
#Redes Neuronales
red=nnet(type~.,data=spam.train,size=2)
Pred=predict(red,spam.test,type="classable(Predicción=Pred,Valores_Reales=spam.test$type)
```
```


```{r}
#ANN
library (neuralnet)


## classification mode
# default with factor response:
#model <- svm(datos$FTR ~ ., data = datos[,-c(1,2,3,4)])

# alternatively the traditional interface:
#x <- subset(iris, select = -Species)
x <- datos[,-c(1,2,3,4,5)]
mi=min(x[,1:21])*0.90
ma=max(x[,1:21])*1.1
x[,1:21]=(x[,1:21]-mi)/(ma-mi)
y <- datos[,5]
dades=data.frame(x,FTR=y,contrasts(y)[y,])

expl=paste0(names(x), collapse = "+")
f <- as.formula(paste0("D+A~",expl))
```

```
net <- neuralnet(f, dades, lifesign="full",lifesign.step =5000,hidden = c(6,6),rep=1, err.fct="ce",linear.output=F,threshold = 0.3)
plot(net)



# test with train data
pred <- compute(net,dades[,1:21])$net.result
# (same as:) pred <- fitted(model)

# Check accuracy:
tt <-table(pred, y)
sum(diag(tt))/sum(tt)#0.5484211
```

```
######################
###CROSSVALIDATION###
######################


```{r}
#Cross validation
library(caret)
set.seed(15092011)
fold2<-createFolds(datos$FTR, k=10)#divide los datos en 10 partes
train_set<-datos

```



```{r}
###NaiveBayes###
cv_in1<- sapply(fold2, function(x){
  train_fold <- train_set[-x,]
  test_fold <- train_set[x,]
  classifier<-naiveBayes (FTR ~ ., data = train_fold[,-c(1,2,3,4)])
  y_pred<- predict(classifier, newdata=train_fold, type="class")
```
```

```
  cm<- table(train_fold$FTR,y_pred)
  accuracy<-sum(diag(cm))/sum(cm)
  return(accuracy)
})

#####################################
cv_out1<- sapply(fold2, function(x){
  train_fold <- train_set[-x,]
  test_fold <- train_set[x,]
  classifier<-naiveBayes (FTR ~ ., data = train_fold[,-c(1,2,3,4)])
  y_pred<- predict(classifier, newdata=test_fold, type="class")
  cm<- table(test_fold$FTR,y_pred)
  accuracy<-sum(diag(cm))/sum(cm)
  return(accuracy)
})
mean(as.numeric(cv_in1))
mean(as.numeric(cv_out1))


```



```{r}

###KNN###DUDA CON EL cv_in2

cv_in2<- sapply(fold2, function(x){
  train_fold <- train_set[-x,]
  test_fold <- train_set[x,]
  classifier<-knn(train_fold[,-c(1,2,3,4,5)],train_fold[,-c(1,2,3,4,5)],train_fold[,5], k = 50)
  cm<- table(train_fold$FTR,classifier)
  accuracy<-sum(diag(cm))/sum(cm)
  return(accuracy)
})
mean(as.numeric(cv_in2))

#####################################
cv_out2<- sapply(fold2, function(x){
  train_fold <- train_set[-x,]
  test_fold <- train_set[x,]
  classifier<-knn(train_fold[,-c(1,2,3,4,5)],test_fold[,-c(1,2,3,4,5)],train_fold[,5], k = 50)
  cm<- table(test_fold$FTR,classifier)
  accuracy<-sum(diag(cm))/sum(cm)
  return(accuracy)
})
mean(as.numeric(cv_out2))



```



```{r}
###ARBOLES DE DECISION###
library(rpart)
cv_in3<- sapply(fold2, function(x){
  train_fold <- train_set[-x,]
  test_fold <- train_set[x,]
  p2 = rpart(FTR~.,data= train_fold[,-c(1,2,3,4)], control=rpart.control(cp=0.001, xval=10))

p2$cptable = as.data.frame(p2$cptable)
ind = which.min(p2$cptable$xerror)

xerr <- p2$cptable$xerror[ind]
xstd <- p2$cptable$xstd[ind]
i = 1
while (p2$cptable$xerror[i] > xerr+xstd) i = i+1
```

```
alfa = p2$cptable$CP[i] #sacamos el parametro optimo

# sacamos el arbol resultante
p1 <- prune(p2,cp=alfa)

y_pred=predict(p1,newdata=train_fold[,-c(1,2,3,4)],type = "class")
cm<- table(train_fold$FTR,y_pred)
accuracy<-sum(diag(cm))/sum(cm)
return(accuracy)
})
mean(as.numeric(cv_in3))
##########################################
cv_out3<- sapply(fold2, function(x){
 train_fold <- train_set[-x,]
 test_fold <- train_set[x,]
 p2 = rpart(FTR~.,data= train_fold[,-c(1,2,3,4)], control=rpart.control(cp=0.001, xval=10))

p2$cptable = as.data.frame(p2$cptable)
ind = which.min(p2$cptable$xerror)

xerr <- p2$cptable$xerror[ind]
xstd <- p2$cptable$xstd[ind]
i = 1
while (p2$cptable$xerror[i] > xerr+xstd) i = i+1

alfa = p2$cptable$CP[i] #sacamos el parametro optimo

# sacamos el arbol resultante
p1 <- prune(p2,cp=alfa)

y_pred=predict(p1,newdata=test_fold[,-c(1,2,3,4)],type = "class")
cm<- table(test_fold$FTR,y_pred)
accuracy<-sum(diag(cm))/sum(cm)
return(accuracy)
})
mean(as.numeric(cv_out3))
```

```{r}
###QDA###
apriori<-as.vector(table(datos$FTR)/length(datos$FTR))
fold2<-createFolds(datos$FTR, k=10)
cv_in4<- sapply(fold2, function(x){
 train_fold <- train_set[-x,]
 test_fold <- train_set[x,]
 classifier<-qda(FTR~., train_fold[,-c(1,2,3,4)], prior = apriori)
 y_pred<- predict(classifier, newdata=train_fold, type="class")
 cm<- table(train_fold$FTR,y_pred$class)
 accuracy<-sum(diag(cm))/sum(cm)
 return(accuracy)
})
mean(as.numeric(cv_in4))
#####################################################
fold2<-createFolds(datos$FTR, k=10)
cv_out4<- sapply(fold2, function(x){
 train_fold <- train_set[-x,]
 test_fold <- train_set[x,]
 classifier<-qda(FTR~., train_fold[,-c(1,2,3,4)], prior = apriori)
 y_pred<- predict(classifier, newdata=test_fold, type="class")
 cm<- table(test_fold$FTR,y_pred$class)
 accuracy<-sum(diag(cm))/sum(cm)
 return(accuracy)
})
mean(as.numeric(cv_out4))



```

```{r}
###SVM###



# we can do the same for the SVM
# We again use d, where we had turned all predictors to numeric
d <- datos
d$fold = cut(1:nrow(d), breaks=10, labels=F)

svm.accuracies = c()

for (i in 1:10) {
  m.svmi = svm(d[d$fold != i,-c(1,2,3,4,5)],
    d[d$fold != i,]$FTR)

  predictions = predict(m.svmi, d[d$fold == i, -c(1,2,3,4,5)])

  numcorrect = sum(predictions == d[d$fold ==
    i,]$FTR)

  svm.accuracies = append(numcorrect / nrow(d[d$fold == i,]), svm.accuracies)
}

svm.accuracies
mean(svm.accuracies)
```


```{r}
###SVM###
cv_in5<- sapply(fold2, function(x){
  train_fold <- train_set[-x,]
  test_fold <- train_set[x,]
  classifier<-svm (FTR ~ ., data = train_fold[,-c(1,2,3,4)], type="C-classification",kernel="radial", cost=1,gamma=0.04761904762)
  y_pred<- predict(classifier, newdata=train_fold, type="class")
  cm<- table(train_fold$FTR,y_pred)
  accuracy<-sum(diag(cm))/sum(cm)
  return(accuracy)
})
#####################################
cv_out5<- sapply(fold2, function(x){
  train_fold <- train_set[-x,]
  test_fold <- train_set[x,]
  classifier<-svm(FTR ~ ., data = train_fold[,-c(1,2,3,4)], type="C-classification",kernel="radial",cost=1,gamma=0.04761904762)
  y_pred<- predict(classifier, newdata=test_fold, type="class")
  cm<- table(test_fold$FTR,y_pred)
  accuracy<-sum(diag(cm))/sum(cm)
  return(accuracy)
})
mean(as.numeric(cv_in5))
mean(as.numeric(cv_out5))




```


```{r}
#otro cv de svm
svm.model <- svm(FTR ~ ., data=datos[,-c(1,2,3,4)], type="C-classification",
 kernel="radial", cost=1,gamma=0.04761904762)
 predict <- fitted(svm.model)
 cm <- table(predict, datos$FTR)
 cm
 accuracy <- sum(diag(cm))/sum(cm)
accuracy
```

```

```{r}

resul_in=data.frame(NB=cv_in1,KNN = cv_in2,CART=cv_in3,QDA=cv_in4,SVM=cv_in5)
boxplot(resul_in,ylim=c(0.4,0.65),col=c(2:5,7))

resul_out=data.frame(NB=cv_out1,KNN=cv_out2,CART=cv_out3,QDA=cv_out4,SVM=cv_out5)
boxplot(resul_out,ylim=c(0.4,0.65),col=c(2:5,7))



resul_mean_in <-
t(data.frame(NB=mean(as.numeric(cv_in1)),KNN=mean(as.numeric(cv_in2)),CART=mean(as.numeric(cv_in3)),QDA=mean(as.nume
ric(cv_in4)),SVM=mean(as.numeric(cv_in5))))
resul_mean_in

resul_mean_out <-t(data.frame(NB=mean(as.numeric(cv_out1)),KNN =
mean(as.numeric(cv_out2)),CART=mean(as.numeric(cv_out3)),QDA=mean(as.numeric(cv_out4)),SVM=mean(as.numeric(cv_out5))
))
resul_mean_out
```


```{r}
plot(datos[, 6:8], col = datos$FTR, pch = 19)

```
```

######################
###VALIDACION###
######################

#Caso Real (Validación)

Ahora se pondran como datos train y test:

#Naive Bayes
```{r}
#NV
#IN
datos_train <- datos[1:1710,-c(1,2,3,4)]
datos_test <- datos[1711:1900,-c(1,2,3,4)]

 classifier<-naiveBayes(FTR ~ ., datos_train)
 y_pred<- predict(classifier, datos_train, type="class")
 cm<- table(datos_train$FTR,y_pred)
 accuracy<-sum(diag(cm))/sum(cm)
  accuracy

#OUT
  library(psych)
library(e1071)
library(caret)
datos_train <- datos[1:1710,-c(1,2,3,4)]
datos_test <- datos[1711:1900,-c(1,2,3,4)]

 classifier<-naiveBayes(FTR ~ ., datos_train)
 #classifier
 y_pred1<- predict(classifier, datos_test, type="class")
 cm<- table(datos_test$FTR,y_pred1)
 cm
 accuracy<-sum(diag(cm))/sum(cm)
  accuracy
```

```{r}
############################################################
#Jornades acertades
############################################################


ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred1#predicción(pronóstico)

acertados <- function(ofi,pro){
 s <- 0
 for(i in 1:length(ofi)){
   if(ofi[i]==pro[i]){
     s <- s +1
   }
 }
 return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/190
```

```{r}


############################################################
#Càlcul de guanys
############################################################


ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred1#predicción(pronóstico
```

```
guanys <- function(ofi,pro,datos_test){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos_test[i,2]-1)
        dineroBW   <- dineroBW   + (datos_test[i,5]-1)
        dineroIW   <- dineroIW   + (datos_test[i,8]-1)
        dineroLB   <- dineroLB   + (datos_test[i,11]-1)
        dineroPS   <- dineroPS   + (datos_test[i,14]-1)
        dineroWH   <- dineroWH   + (datos_test[i,17]-1)
        dineroVC   <- dineroVC   + (datos_test[i,20]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos_test[i,3]-1)
        dineroBW   <- dineroBW   + (datos_test[i,6]-1)
        dineroIW   <- dineroIW   + (datos_test[i,9]-1)
        dineroLB   <- dineroLB   + (datos_test[i,12]-1)
        dineroPS   <- dineroPS   + (datos_test[i,15]-1)
        dineroWH   <- dineroWH   + (datos_test[i,18]-1)
        dineroVC   <- dineroVC   + (datos_test[i,21]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos_test[i,4]-1)
        dineroBW   <- dineroBW   + (datos_test[i,7]-1)
        dineroIW   <- dineroIW   + (datos_test[i,10]-1)
        dineroLB   <- dineroLB   + (datos_test[i,13]-1)
        dineroPS   <- dineroPS   + (datos_test[i,16]-1)
        dineroWH   <- dineroWH   + (datos_test[i,19]-1)
        dineroVC   <- dineroVC   + (datos_test[i,22]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }


  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

  return(g)

}

guanys(ofi,pro,datos_test)
```

```{r}

#########################################################
#Gráfico
#########################################################
```

```
ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred1#predicción(pronóstico

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos_test){
 dineroB365 <- 0
 dineroBW   <- 0
 dineroIW   <- 0
 dineroLB   <- 0
 dineroPS   <- 0
 dineroWH   <- 0
 dineroVC   <- 0

 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

     if(ofi[i]== "H"){
      dineroB365 <- dineroB365 + (datos_test[i,2]-1)
      dineroBW   <- dineroBW   + (datos_test[i,5]-1)
      dineroIW   <- dineroIW   + (datos_test[i,8]-1)
      dineroLB   <- dineroLB   + (datos_test[i,11]-1)
      dineroPS   <- dineroPS   + (datos_test[i,14]-1)
      dineroWH   <- dineroWH   + (datos_test[i,17]-1)
      dineroVC   <- dineroVC   + (datos_test[i,20]-1)

     }else if(ofi[i]=="D"){
      dineroB365 <- dineroB365 + (datos_test[i,3]-1)
      dineroBW   <- dineroBW   + (datos_test[i,6]-1)
      dineroIW   <- dineroIW   + (datos_test[i,9]-1)
      dineroLB   <- dineroLB   + (datos_test[i,12]-1)
      dineroPS   <- dineroPS   + (datos_test[i,15]-1)
      dineroWH   <- dineroWH   + (datos_test[i,18]-1)
      dineroVC   <- dineroVC   + (datos_test[i,21]-1)

     }else if(ofi[i]=="A"){
      dineroB365 <- dineroB365 + (datos_test[i,4]-1)
      dineroBW   <- dineroBW   + (datos_test[i,7]-1)
      dineroIW   <- dineroIW   + (datos_test[i,10]-1)
      dineroLB   <- dineroLB   + (datos_test[i,13]-1)
      dineroPS   <- dineroPS   + (datos_test[i,16]-1)
      dineroWH   <- dineroWH   + (datos_test[i,19]-1)
      dineroVC   <- dineroVC   + (datos_test[i,22]-1)

     }
  }else{dineroB365 <- dineroB365 - 1
      dineroBW   <- dineroBW   - 1
      dineroIW   <- dineroIW   - 1
      dineroLB   <- dineroLB   - 1
      dineroPS   <- dineroPS   - 1
      dineroWH   <- dineroWH   - 1
      dineroVC   <- dineroVC   - 1
  }

  B365[i] <- dineroB365
  BW[i] <- dineroBW
  IW[i] <- dineroIW
  LB[i] <- dineroLB
  PS[i] <- dineroPS
  WH[i] <- dineroWH
  VC[i] <- dineroVC

 }
  g <- data.frame(casa
```

```
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

   graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos_test)
graf

```

```{r}
 plot(1:190, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
 ylim=c(-25,25),xlim=c(0,200))

lines(1:190,graf$BW,type="l",pch=2,col=3)
lines(1:190,graf$IW,type="l",pch=2,col=4)
lines(1:190,graf$LB,type="l",pch=2,col=5)
lines(1:190,graf$PS,type="l",pch=2,col=6)
lines(1:190,graf$WH,type="l",pch=2,col=7)
lines(1:190,graf$VC,type="l",pch=2,col=8)
lines(1:190,rep(0,190),type="l",pch=2,col=1)

 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=1:7,xjust=1)


```
La gráfica muestra, que si se realizan todas las apuestas en la casa PS, siguiendo como predictor/clasificador Naive Bayes, se obtendrá un beneficio neto positivo.
Faltaria añadir la linea en la gráfica que mostrará el beneficio neto si se apuesta en cada partido a la casa con la cuota más alta. Claramente seria el beneficio neto más alto.

También falta hacer el grafico teniendo en cuenta todos los metodos de predicción para una casa en concreto.(en el markdown graf)

#Gráfico apostando a la mejor cuota

```{r}
#########################################################
#Càlcul de guanys
#########################################################
ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred1#predicción(pronóstico


B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0


guanys <- function(ofi,pro,datos_test){
  dinero <- 0


  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dinero <- dinero + max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-
1,datos_test[i,17]-1,
                                 datos_test[i,20]-1))
        if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                                 datos_test[i,20]-1))==datos_test[i,2]-1){
```

```
        B365 <- B365 + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,5]-1){
        BW <- BW + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,8]-1){
        IW <- IW + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,11]-1){
        LB <- LB + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,14]-1){
        PS <- PS + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,17]-1){
        WH <- WH + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,20]-1){
        VC <- VC + 1
                          }


      }else if(ofi[i]=="D"){

       dinero <- dinero + max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-
1,datos_test[i,18]-1,
                          datos_test[i,21]-1))
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,3]-1){
        B365 <- B365 + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,6]-1){
        BW <- BW + 1
       }
        if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,9]-1){
        IW <- IW + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,12]-1){
        LB <- LB + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,15]-1){
        PS <- PS + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,18]-1){
        WH <- WH + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,21]-1){
        VC <- VC + 1
                          }

      }else if(ofi[i]=="A"){

       dinero <- dinero + max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-
1,datos_test[i,19]-1,
                          datos_test[i,22]-1))

      if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,4]-1){
        B365 <- B365 + 1         }
```

```
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,7]-1){
         BW <- BW + 1
        }
         if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,10]-1){
         IW <- IW + 1
        }
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,13]-1){
         LB <- LB + 1
        }
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,16]-1){
         PS <- PS + 1
        }
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,19]-1){
         WH <- WH + 1
        }
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,22]-1){
         VC <- VC + 1
                          }

      }
    }else{dinero <- dinero - 1

    }


  }
    g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(g)
}

guanys(ofi,pro,datos_test)


```

```
```{r}
#KNN

#IN
library(class)

#OUT
datos_train <- datos[1:1710,-c(1,2,3,4,5)]
datos_test <- datos[1711:1900,-c(1,2,3,4,5)]

y_pred2<-knn(datos_train,datos_test,datos[1:1710,5], k = 10)

  cm<- table(datos[1711:1900,5],y_pred2)
  cm
  accuracy<-sum(diag(cm))/sum(cm)
  accuracy
```



```{r}
#########################################################
#Jornades acertades
```
```

```
#########################################################


ofi <- datos[1711:1900,5] #ofi:resultado oficial
pro <- y_pred2#predicción(pronóstico)

acertados <- function(ofi,pro){
 s <- 0
 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){
   s <- s +1
  }
 }
 return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/190
```


```{r}


#########################################################
#Càlcul de guanys
#########################################################


ofi <- datos[1711:1900,5] #ofi:resultado oficial
pro <- y_pred2#predicción(pronóstico


guanys <- function(ofi,pro,datos_test){
 dineroB365 <- 0
 dineroBW  <- 0
 dineroIW  <- 0
 dineroLB  <- 0
 dineroPS  <- 0
 dineroWH  <- 0
 dineroVC  <- 0

 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
     dineroB365 <- dineroB365 + (datos_test[i,1]-1)
     dineroBW  <- dineroBW  + (datos_test[i,4]-1)
     dineroIW  <- dineroIW  + (datos_test[i,7]-1)
     dineroLB  <- dineroLB  + (datos_test[i,10]-1)
     dineroPS  <- dineroPS  + (datos_test[i,13]-1)
     dineroWH  <- dineroWH  + (datos_test[i,16]-1)
     dineroVC  <- dineroVC  + (datos_test[i,19]-1)

    }else if(ofi[i]=="D"){
     dineroB365 <- dineroB365 + (datos_test[i,2]-1)
     dineroBW  <- dineroBW  + (datos_test[i,5]-1)
     dineroIW  <- dineroIW  + (datos_test[i,8]-1)
     dineroLB  <- dineroLB  + (datos_test[i,11]-1)
     dineroPS  <- dineroPS  + (datos_test[i,14]-1)
     dineroWH  <- dineroWH  + (datos_test[i,17]-1)
     dineroVC  <- dineroVC  + (datos_test[i,20]-1)

    }else if(ofi[i]=="A"){
     dineroB365 <- dineroB365 + (datos_test[i,3]-1)
     dineroBW  <- dineroBW  + (datos_test[i,6]-1)
     dineroIW  <- dineroIW  + (datos_test[i,9]-1)
     dineroLB  <- dineroLB  + (datos_test[i,12]-1)
     dineroPS  <- dineroPS  + (datos_test[i,15]-1)
     dineroWH  <- dineroWH  + (datos_test[i,18]-1)
     dineroVC  <- dineroVC  + (datos_test[i,21]-1)
```

```
      }
   }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
   }


 }
   g <- data.frame(casa =c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,
                                                dineroPS,dineroWH,dineroVC))

 return(g)

}

guanys(ofi,pro,datos_test)

```



```{r}


##########################################################
#Gráfico
##########################################################
ofi <- datos[1711:1900,5] #ofi:resultado oficial
pro <- y_pred2#predicción(pronóstico

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos_test){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos_test[i,1]-1)
        dineroBW   <- dineroBW   + (datos_test[i,4]-1)
        dineroIW   <- dineroIW   + (datos_test[i,7]-1)
        dineroLB   <- dineroLB   + (datos_test[i,10]-1)
        dineroPS   <- dineroPS   + (datos_test[i,13]-1)
        dineroWH   <- dineroWH   + (datos_test[i,16]-1)
        dineroVC   <- dineroVC   + (datos_test[i,19]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos_test[i,2]-1)
        dineroBW   <- dineroBW   + (datos_test[i,5]-1)
        dineroIW   <- dineroIW   + (datos_test[i,8]-1)
        dineroLB   <- dineroLB   + (datos_test[i,11]-1)
```

```
        dineroPS   <- dineroPS   + (datos_test[i,14]-1)
        dineroWH   <- dineroWH   + (datos_test[i,17]-1)
        dineroVC   <- dineroVC   + (datos_test[i,20]-1)

       }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos_test[i,3]-1)
        dineroBW   <- dineroBW   + (datos_test[i,6]-1)
        dineroIW   <- dineroIW   + (datos_test[i,9]-1)
        dineroLB   <- dineroLB   + (datos_test[i,12]-1)
        dineroPS   <- dineroPS   + (datos_test[i,15]-1)
        dineroWH   <- dineroWH   + (datos_test[i,18]-1)
        dineroVC   <- dineroVC   + (datos_test[i,21]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }

    B365[i] <- dineroB365
    BW[i] <- dineroBW
    IW[i] <- dineroIW
    LB[i] <- dineroLB
    PS[i] <- dineroPS
    WH[i] <- dineroWH
    VC[i] <- dineroVC

  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

    graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos_test)
graf

```

```{r}
 plot(1:190, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
 ylim=c(-25,25),xlim=c(0,200))

lines(1:190,graf$BW,type="l",pch=2,col=3)
lines(1:190,graf$IW,type="l",pch=2,col=4)
lines(1:190,graf$LB,type="l",pch=2,col=5)
lines(1:190,graf$PS,type="l",pch=2,col=6)
lines(1:190,graf$WH,type="l",pch=2,col=7)
lines(1:190,graf$VC,type="l",pch=2,col=8)
lines(1:190,rep(0,190),type="l",pch=2,col=1)

 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=1:7,xjust=1)


```
La gráfica muestra, que si se realizan todas las apuestas en la casa PS, siguiendo como predictor/clasificador Naive Bayes, se obtendrá un beneficio neto positivo.
Faltaria añadir la linea en la gráfica que mostrará el beneficio neto si se apuesta en cada partido a la casa con la cuota más alta. Claramente seria el beneficio neto más alto.

También falta hacer el grafico teniendo en cuenta todos los metodos de predicción para una casa en concreto.(en el markdown graf)

#Gráfico apostando a la mejor cuota

```{r}
#########################################################
#Càlcul de guanys
#########################################################
ofi <- datos[1711:1900,5] #ofi:resultado oficial
pro <- y_pred2#predicción(pronóstico


B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0


guanys <- function(ofi,pro,datos_test){
 dinero <- 0


 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
      dinero <- dinero + max(c(datos_test[i,1]-1,datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-
1,datos_test[i,16]-1,
                               datos_test[i,19]-1))
      if(max(c(datos_test[i,1]-1,datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,
                               datos_test[i,19]-1))==datos_test[i,1]-1){
        B365 <- B365 + 1
      }
      if(max(c(datos_test[i,1]-1,datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,
                               datos_test[i,19]-1))==datos_test[i,4]-1){
        BW <- BW + 1
      }
      if(max(c(datos_test[i,1]-1,datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,
                               datos_test[i,19]-1))==datos_test[i,7]-1){
        IW <- IW + 1
      }
      if(max(c(datos_test[i,1]-1,datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,
                               datos_test[i,19]-1))==datos_test[i,10]-1){
        LB <- LB + 1
      }
      if(max(c(datos_test[i,1]-1,datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,
                               datos_test[i,19]-1))==datos_test[i,13]-1){
        PS <- PS + 1
      }
      if(max(c(datos_test[i,1]-1,datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,
                               datos_test[i,19]-1))==datos_test[i,16]-1){
        WH <- WH + 1
      }
      if(max(c(datos_test[i,1]-1,datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,
                               datos_test[i,19]-1))==datos_test[i,19]-1){
        VC <- VC + 1
                               }


    }else if(ofi[i]=="D"){

      dinero <- dinero + max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-
1,datos_test[i,17]-1,
                               datos_test[i,20]-1))
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                               datos_test[i,20]-1))==datos_test[i,2]-1){
        B365 <- B365 + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
```

```
                        datos_test[i,20]-1))==datos_test[i,5]-1){
        BW <- BW + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,8]-1){
        IW <- IW + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,11]-1){
        LB <- LB + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,14]-1){
        PS <- PS + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,17]-1){
        WH <- WH + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,20]-1){
        VC <- VC + 1
                        }

      }else if(ofi[i]=="A"){

       dinero <- dinero + max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-
1,datos_test[i,18]-1,
                        datos_test[i,21]-1))

      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,3]-1){
        B365 <- B365 + 1          }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,6]-1){
        BW <- BW + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,9]-1){
        IW <- IW + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,12]-1){
        LB <- LB + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,15]-1){
        PS <- PS + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,18]-1){
        WH <- WH + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,21]-1){
        VC <- VC + 1
                        }

      }
   }else{dinero <- dinero - 1

   }


  }
   g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(g)
}

guanys(ofi,pro,datos_test)
```

```
```

```{r}
#CART
library(rpart)
#IN
datos_train <- datos[1:1710,-c(1,2,3,4)]
datos_test <- datos[1711:1900,-c(1,2,3,4)]

  p2 = rpart(FTR~.,datos_train, control=rpart.control(cp=0.001, xval=10))

p2$cptable = as.data.frame(p2$cptable)
ind = which.min(p2$cptable$xerror)

xerr <- p2$cptable$xerror[ind]
xstd <- p2$cptable$xstd[ind]
i = 1
while (p2$cptable$xerror[i] > xerr+xstd) i = i+1

alfa = p2$cptable$CP[i] #sacamos el parametro optimo

# sacamos el arbol resultante
p1 <- prune(p2,cp=alfa)

y_pred3=predict(p1,datos_train,type = "class")
cm<- table(datos_train$FTR,y_pred3)
accuracy<-sum(diag(cm))/sum(cm)
accuracy



#OUT
library(rpart)
datos_train <- datos[1:1710,-c(1,2,3,4)]
datos_test <- datos[1711:1900,-c(1,2,3,4)]

  p2 = rpart(FTR~.,datos_train, control=rpart.control(cp=0.001, xval=10))

p2$cptable = as.data.frame(p2$cptable)
p2$cptable
ind = which.min(p2$cptable$xerror)


xerr <- p2$cptable$xerror[ind]
xstd <- p2$cptable$xstd[ind]
i = 1
while (p2$cptable$xerror[i] > xerr+xstd) i = i+1

alfa = p2$cptable$CP[i] #sacamos el parametro optimo
alfa
# sacamos el arbol resultante
p1 <- prune(p2,cp=alfa)
library(rpart.plot)
rpart.plot(p1)

y_pred3=predict(p1,datos_test,type = "class")
cm<- table(datos_test$FTR,y_pred3)
cm
accuracy<-sum(diag(cm))/sum(cm)
accuracy

```


```{r}
######
```

```
#ARBOL OPTIMO(caso futbol)
######
set.seed(15092011)

p2 = rpart(FTR~.,datos_train, control=rpart.control(cp=0.001, xval=10))
p2
plot(p2)
text(p2,pretty = 0)
text(p2, use.n=T)
printcp(p2)

plot(p2$cptable[,2],p2$cptable[,3],type="l",xlab="size of the tree",ylab="Relative impurity",main="R(t)")
lines(p2$cptable[,2],p2$cptable[,4],col="blue")
legend("topright",c("R(T)training","R(T)cv"),col=c("black","blue"),lty=1)

#observando el grafico vemos que el cp que tiene 6 splits es de 0.010582
plotcp(p2)
printcp(p2)
prune(p2,cp = 0.001)

#sin embargo recurimos a la tÃ©cnica dada en clase
p2$cptable = as.data.frame(p2$cptable)
ind = which.min(p2$cptable$xerror)

xerr <- p2$cptable$xerror[ind]
xstd <- p2$cptable$xstd[ind]

i = 1
while (p2$cptable$xerror[i] > xerr+xstd) i = i+1

alfa = p2$cptable$CP[i] #sacamos el parametro optimo(0.008082768)

# sacamos el arbol resultante
p1 <- prune(p2,cp=alfa)
summary(p1)
plot(p1)
text(p1,use.n = T)
library(rpart.plot)
rpart.plot(p1)

#predicciones para los datos de entreno, del ano 1988
p11=predict(p1,data=datos_test,type = "class")
(mat<-table(datos_test$FTR,p11))
error_rate<-(sum(mat)-sum(diag(mat)))/sum(mat) #error_rate <- [1] 0.1169355
1-error_rate
```

```{r}
############################################################
#Jornades acertades
############################################################


ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred3#predicción(pronóstico)

acertados <- function(ofi,pro){
 s <- 0
 for(i in 1:length(ofi)){
   if(ofi[i]==pro[i]){
     s <- s +1
   }
 }
 return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/190
```

```
```

```{r}


#########################################################
#Càlcul de guanys
#########################################################


ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred3#predicción(pronóstico


guanys <- function(ofi,pro,datos_test){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos_test[i,2]-1)
        dineroBW   <- dineroBW   + (datos_test[i,5]-1)
        dineroIW   <- dineroIW   + (datos_test[i,8]-1)
        dineroLB   <- dineroLB   + (datos_test[i,11]-1)
        dineroPS   <- dineroPS   + (datos_test[i,14]-1)
        dineroWH   <- dineroWH   + (datos_test[i,17]-1)
        dineroVC   <- dineroVC   + (datos_test[i,20]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos_test[i,3]-1)
        dineroBW   <- dineroBW   + (datos_test[i,6]-1)
        dineroIW   <- dineroIW   + (datos_test[i,9]-1)
        dineroLB   <- dineroLB   + (datos_test[i,12]-1)
        dineroPS   <- dineroPS   + (datos_test[i,15]-1)
        dineroWH   <- dineroWH   + (datos_test[i,18]-1)
        dineroVC   <- dineroVC   + (datos_test[i,21]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos_test[i,4]-1)
        dineroBW   <- dineroBW   + (datos_test[i,7]-1)
        dineroIW   <- dineroIW   + (datos_test[i,10]-1)
        dineroLB   <- dineroLB   + (datos_test[i,13]-1)
        dineroPS   <- dineroPS   + (datos_test[i,16]-1)
        dineroWH   <- dineroWH   + (datos_test[i,19]-1)
        dineroVC   <- dineroVC   + (datos_test[i,22]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }


  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))
    a <- c(1,2,3)
  return(g)
```

```
   return(a)
}

guanys(ofi,pro,datos_test)

```


```{r}


#########################################################
#Gráfico
#########################################################
ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred3#predicción(pronóstico

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos_test){
 dineroB365 <- 0
 dineroBW   <- 0
 dineroIW   <- 0
 dineroLB   <- 0
 dineroPS   <- 0
 dineroWH   <- 0
 dineroVC   <- 0

 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

     if(ofi[i]== "H"){
       dineroB365 <- dineroB365 + (datos_test[i,2]-1)
       dineroBW   <- dineroBW   + (datos_test[i,5]-1)
       dineroIW   <- dineroIW   + (datos_test[i,8]-1)
       dineroLB   <- dineroLB   + (datos_test[i,11]-1)
       dineroPS   <- dineroPS   + (datos_test[i,14]-1)
       dineroWH   <- dineroWH   + (datos_test[i,17]-1)
       dineroVC   <- dineroVC   + (datos_test[i,20]-1)

     }else if(ofi[i]=="D"){
       dineroB365 <- dineroB365 + (datos_test[i,3]-1)
       dineroBW   <- dineroBW   + (datos_test[i,6]-1)
       dineroIW   <- dineroIW   + (datos_test[i,9]-1)
       dineroLB   <- dineroLB   + (datos_test[i,12]-1)
       dineroPS   <- dineroPS   + (datos_test[i,15]-1)
       dineroWH   <- dineroWH   + (datos_test[i,18]-1)
       dineroVC   <- dineroVC   + (datos_test[i,21]-1)

     }else if(ofi[i]=="A"){
       dineroB365 <- dineroB365 + (datos_test[i,4]-1)
       dineroBW   <- dineroBW   + (datos_test[i,7]-1)
       dineroIW   <- dineroIW   + (datos_test[i,10]-1)
       dineroLB   <- dineroLB   + (datos_test[i,13]-1)
       dineroPS   <- dineroPS   + (datos_test[i,16]-1)
       dineroWH   <- dineroWH   + (datos_test[i,19]-1)
       dineroVC   <- dineroVC   + (datos_test[i,22]-1)

     }
  }else{dineroB365 <- dineroB365 - 1
       dineroBW   <- dineroBW   - 1
       dineroIW   <- dineroIW   - 1
       dineroLB   <- dineroLB   - 1
```

```
      dineroPS   <- dineroPS   - 1
      dineroWH   <- dineroWH   - 1
      dineroVC   <- dineroVC   - 1
   }

   B365[i] <- dineroB365
   BW[i] <- dineroBW
   IW[i] <- dineroIW
   LB[i] <- dineroLB
   PS[i] <- dineroPS
   WH[i] <- dineroWH
   VC[i] <- dineroVC

 }
   g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

   graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
 return(graf)
}

graf <-guanys(ofi,pro,datos_test)
graf

```

```{r}
 plot(1:190, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
 ylim=c(-25,25),xlim=c(0,200))

lines(1:190,graf$BW,type="l",pch=2,col=3)
lines(1:190,graf$IW,type="l",pch=2,col=4)
lines(1:190,graf$LB,type="l",pch=2,col=5)
lines(1:190,graf$PS,type="l",pch=2,col=6)
lines(1:190,graf$WH,type="l",pch=2,col=7)
lines(1:190,graf$VC,type="l",pch=2,col=8)
lines(1:190,rep(0,190),type="l",pch=2,col=1)

 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=1:7,xjust=1)


```
La gráfica muestra, que si se realizan todas las apuestas en la casa PS, siguiendo como predictor/clasificador Naive Bayes, se obtendrá un beneficio neto positivo.
Faltaria añadir la linea en la gráfica que mostrará el beneficio neto si se apuesta en cada partido a la casa con la cuota más alta. Claramente seria el beneficio neto más alto.

También falta hacer el grafico teniendo en cuenta todos los metodos de predicción para una casa en concreto.(en el markdown graf)

#Gráfico apostando a la mejor cuota

```{r}
##########################################################
#Càlcul de guanys
##########################################################
ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred3#predicción(pronóstico


B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0
```

```
guanys <- function(ofi,pro,datos_test){
 dinero <- 0


 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

     if(ofi[i]== "H"){
      dinero <- dinero + max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-
1,datos_test[i,17]-1,
                          datos_test[i,20]-1))
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,2]-1){
       B365 <- B365 + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,5]-1){
       BW <- BW + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,8]-1){
       IW <- IW + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,11]-1){
       LB <- LB + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,14]-1){
       PS <- PS + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,17]-1){
       WH <- WH + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,20]-1){
       VC <- VC + 1
                          }


     }else if(ofi[i]=="D"){

      dinero <- dinero + max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-
1,datos_test[i,18]-1,
                          datos_test[i,21]-1))
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,3]-1){
       B365 <- B365 + 1
      }
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,6]-1){
       BW <- BW + 1
      }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,9]-1){
       IW <- IW + 1
      }
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,12]-1){
       LB <- LB + 1
      }
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,15]-1){
       PS <- PS + 1
      }
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,18]-1){
       WH <- WH + 1
```

```
        }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,21]-1){
         VC <- VC + 1
                          }

      }else if(ofi[i]=="A"){

       dinero <- dinero + max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-
1,datos_test[i,19]-1,
                          datos_test[i,22]-1))

      if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,4]-1){
         B365 <- B365 + 1          }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,7]-1){
         BW <- BW + 1
       }
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,10]-1){
         IW <- IW + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,13]-1){
         LB <- LB + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,16]-1){
         PS <- PS + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,19]-1){
         WH <- WH + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,22]-1){
         VC <- VC + 1
                          }

     }
   }else{dinero <- dinero - 1

   }


 }
   g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
 return(g)
}

guanys(ofi,pro,datos_test)


```
```
```{r}
#QDA

library(MASS)
#IN

datos_train <- datos[1:1710,-c(1,2,3,4)]
datos_test <- datos[1711:1900,-c(1,2,3,4)]

classifier<-qda(FTR~., datos_train, prior = apriori)
 y_pred4<- predict(classifier, datos_train, type="class")
```
```

```
  cm<- table(datos_train$FTR,y_pred4$class)
  accuracy<-sum(diag(cm))/sum(cm)
  accuracy


#OUT
datos_train <- datos[1:1710,-c(1,2,3,4)]
datos_test <- datos[1711:1900,-c(1,2,3,4)]

classifier<-qda(FTR~., datos_train, prior = apriori)
  y_pred4<- predict(classifier, datos_test, type="class")
  cm<- table(datos_test$FTR,y_pred4$class)
  cm
  accuracy<-sum(diag(cm))/sum(cm)
  accuracy
```

```{r}
############################################################
#Jornades acertades
############################################################


ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred4$class#predicción(pronóstico)

acertados <- function(ofi,pro){
  s <- 0
  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){
      s <- s +1
    }
  }
  return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/190
```

```{r}


############################################################
#Càlcul de guanys
############################################################


ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred4$class#predicción(pronóstico


guanys <- function(ofi,pro,datos_test){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos_test[i,2]-1)
        dineroBW   <- dineroBW   + (datos_test[i,5]-1)
```

```
        dineroIW   <- dineroIW   + (datos_test[i,8]-1)
        dineroLB   <- dineroLB   + (datos_test[i,11]-1)
        dineroPS   <- dineroPS   + (datos_test[i,14]-1)
        dineroWH   <- dineroWH   + (datos_test[i,17]-1)
        dineroVC   <- dineroVC   + (datos_test[i,20]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos_test[i,3]-1)
        dineroBW   <- dineroBW   + (datos_test[i,6]-1)
        dineroIW   <- dineroIW   + (datos_test[i,9]-1)
        dineroLB   <- dineroLB   + (datos_test[i,12]-1)
        dineroPS   <- dineroPS   + (datos_test[i,15]-1)
        dineroWH   <- dineroWH   + (datos_test[i,18]-1)
        dineroVC   <- dineroVC   + (datos_test[i,21]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos_test[i,4]-1)
        dineroBW   <- dineroBW   + (datos_test[i,7]-1)
        dineroIW   <- dineroIW   + (datos_test[i,10]-1)
        dineroLB   <- dineroLB   + (datos_test[i,13]-1)
        dineroPS   <- dineroPS   + (datos_test[i,16]-1)
        dineroWH   <- dineroWH   + (datos_test[i,19]-1)
        dineroVC   <- dineroVC   + (datos_test[i,22]-1)

      }
   }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
   }


 }
   g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))
   a <- c(1,2,3)
 return(g)
   return(a)
}

guanys(ofi,pro,datos_test)

```



```{r}


############################################################
#Gráfico
############################################################
ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred4$class#predicción(pronóstico

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos_test){
 dineroB365 <- 0
 dineroBW   <- 0
 dineroIW   <- 0
```

```
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
   if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos_test[i,2]-1)
        dineroBW   <- dineroBW   + (datos_test[i,5]-1)
        dineroIW   <- dineroIW   + (datos_test[i,8]-1)
        dineroLB   <- dineroLB   + (datos_test[i,11]-1)
        dineroPS   <- dineroPS   + (datos_test[i,14]-1)
        dineroWH   <- dineroWH   + (datos_test[i,17]-1)
        dineroVC   <- dineroVC   + (datos_test[i,20]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos_test[i,3]-1)
        dineroBW   <- dineroBW   + (datos_test[i,6]-1)
        dineroIW   <- dineroIW   + (datos_test[i,9]-1)
        dineroLB   <- dineroLB   + (datos_test[i,12]-1)
        dineroPS   <- dineroPS   + (datos_test[i,15]-1)
        dineroWH   <- dineroWH   + (datos_test[i,18]-1)
        dineroVC   <- dineroVC   + (datos_test[i,21]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos_test[i,4]-1)
        dineroBW   <- dineroBW   + (datos_test[i,7]-1)
        dineroIW   <- dineroIW   + (datos_test[i,10]-1)
        dineroLB   <- dineroLB   + (datos_test[i,13]-1)
        dineroPS   <- dineroPS   + (datos_test[i,16]-1)
        dineroWH   <- dineroWH   + (datos_test[i,19]-1)
        dineroVC   <- dineroVC   + (datos_test[i,22]-1)

      }
   }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
   }

   B365[i] <- dineroB365
   BW[i] <- dineroBW
   IW[i] <- dineroIW
   LB[i] <- dineroLB
   PS[i] <- dineroPS
   WH[i] <- dineroWH
   VC[i] <- dineroVC

  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

    graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos_test)
graf

```

```{r}
 plot(1:190, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
```

```
 ylim=c(-25,25),xlim=c(0,200))

lines(1:190,graf$BW,type="l",pch=2,col=3)
lines(1:190,graf$IW,type="l",pch=2,col=4)
lines(1:190,graf$LB,type="l",pch=2,col=5)
lines(1:190,graf$PS,type="l",pch=2,col=6)
lines(1:190,graf$WH,type="l",pch=2,col=7)
lines(1:190,graf$VC,type="l",pch=2,col=8)
lines(1:190,rep(0,190),type="l",pch=2,col=1)

 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=1:7,xjust=1)


```
La gráfica muestra, que si se realizan todas las apuestas en la casa PS, siguiendo como predictor/clasificador Naive Bayes, se obtendrá un beneficio neto positivo.
Faltaria añadir la linea en la gráfica que mostrará el beneficio neto si se apuesta en cada partido a la casa con la cuota más alta. Claramente seria el beneficio neto más alto.

También falta hacer el grafico teniendo en cuenta todos los metodos de predicción para una casa en concreto.(en el markdown graf)

#Gráfico apostando a la mejor cuota

```{r}
```
###
#Càlcul de guanys
###
ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred4$class#predicción(pronóstico

B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0

guanys <- function(ofi,pro,datos_test){
  dinero <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dinero <- dinero + max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-
1,datos_test[i,17]-1,
                                 datos_test[i,20]-1))
        if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                                 datos_test[i,20]-1))==datos_test[i,2]-1){
          B365 <- B365 + 1
        }
        if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                                 datos_test[i,20]-1))==datos_test[i,5]-1){
          BW <- BW + 1
        }
        if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                                 datos_test[i,20]-1))==datos_test[i,8]-1){
          IW <- IW + 1
        }
        if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                                 datos_test[i,20]-1))==datos_test[i,11]-1){
          LB <- LB + 1
        }
        if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                                 datos_test[i,20]-1))==datos_test[i,14]-1){
```
```

```
        PS <- PS + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,17]-1){
        WH <- WH + 1
       }
       if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                          datos_test[i,20]-1))==datos_test[i,20]-1){
        VC <- VC + 1
                          }


      }else if(ofi[i]=="D"){

       dinero <- dinero + max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-
1,datos_test[i,18]-1,
                          datos_test[i,21]-1))
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,3]-1){
        B365 <- B365 + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,6]-1){
        BW <- BW + 1
       }
        if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,9]-1){
        IW <- IW + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,12]-1){
        LB <- LB + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,15]-1){
        PS <- PS + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,18]-1){
        WH <- WH + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                          datos_test[i,21]-1))==datos_test[i,21]-1){
        VC <- VC + 1
                          }

      }else if(ofi[i]=="A"){

       dinero <- dinero + max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-
1,datos_test[i,19]-1,
                          datos_test[i,22]-1))

      if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,4]-1){
        B365 <- B365 + 1         }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,7]-1){
        BW <- BW + 1
       }
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,10]-1){
        IW <- IW + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,13]-1){
        LB <- LB + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,16]-1){
        PS <- PS + 1
       }
```

```
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                           datos_test[i,22]-1))==datos_test[i,19]-1){
          WH <- WH + 1
        }
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                           datos_test[i,22]-1))==datos_test[i,22]-1){
          VC <- VC + 1
                           }

      }
    }else{dinero <- dinero - 1

    }


  }
    g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(g)
}

guanys(ofi,pro,datos_test)


```
```

Sabiendo que SVM es el mejor, ahora se pondran como datos train y test:

```{r}
#SVM

#IN
datos_train <- datos[1:1710,-c(1,2,3,4)]
datos_test <- datos[1711:1900,-c(1,2,3,4)]


 classifier<-svm(FTR ~ ., data = datos_train, type="C-classification",kernel="radial",cost=1,gamma=0.04761904762)
 y_pred5<- predict(classifier, newdata=datos_train, type="class")
 cm<- table(datos_train$FTR,y_pred5)
 accuracy<-sum(diag(cm))/sum(cm)
 accuracy #0.5842105263

#OUT
datos_train <- datos[1:1710,-c(1,2,3,4)]
datos_test <- datos[1711:1900,-c(1,2,3,4)]


 classifier<-svm(FTR ~ ., data = datos_train, type="C-classification",kernel="radial",cost=1,gamma=0.04761904762)
 y_pred5<- predict(classifier, newdata=datos_test, type="class")
 cm<- table(datos_test$FTR,y_pred5)
 cm
 accuracy<-sum(diag(cm))/sum(cm)
 accuracy #0.5842105263

```
  El valor de 0.58, es posible ya que si observamos los folds de cv_out5 hay casos que llegan hasta el 0.59

   Fold01     Fold02     Fold03     Fold04     Fold05     Fold06     Fold07     Fold08     Fold09
0.5947368421 0.4894736842 0.5315789474 0.5526315789 0.5263157895 0.5526315789 0.5578947368 0.5343915344
0.5340314136
    Fold10
0.5736842105
```

```{r}
#########################################################
#Jornades acertades
#########################################################


ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred5#predicción(pronóstico)

acertados <- function(ofi,pro){
 s <- 0
 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){
   s <- s +1
  }
 }
 return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/190
```



```{r}


#########################################################
#Càlcul de guanys
#########################################################


ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred5#predicción(pronóstico


guanys <- function(ofi,pro,datos_test){
 dineroB365 <- 0
 dineroBW  <- 0
 dineroIW  <- 0
 dineroLB  <- 0
 dineroPS  <- 0
 dineroWH  <- 0
 dineroVC  <- 0

 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
     dineroB365 <- dineroB365 + (datos_test[i,2]-1)
     dineroBW  <- dineroBW  + (datos_test[i,5]-1)
     dineroIW  <- dineroIW  + (datos_test[i,8]-1)
     dineroLB  <- dineroLB  + (datos_test[i,11]-1)
     dineroPS  <- dineroPS  + (datos_test[i,14]-1)
     dineroWH  <- dineroWH  + (datos_test[i,17]-1)
     dineroVC  <- dineroVC  + (datos_test[i,20]-1)

    }else if(ofi[i]=="D"){
     dineroB365 <- dineroB365 + (datos_test[i,3]-1)
     dineroBW  <- dineroBW  + (datos_test[i,6]-1)
     dineroIW  <- dineroIW  + (datos_test[i,9]-1)
     dineroLB  <- dineroLB  + (datos_test[i,12]-1)
     dineroPS  <- dineroPS  + (datos_test[i,15]-1)
     dineroWH  <- dineroWH  + (datos_test[i,18]-1)
     dineroVC  <- dineroVC  + (datos_test[i,21]-1)

    }else if(ofi[i]=="A"){
     dineroB365 <- dineroB365 + (datos_test[i,4]-1)
     dineroBW  <- dineroBW  + (datos_test[i,7]-1)
     dineroIW  <- dineroIW  + (datos_test[i,10]-1)
```

```
        dineroLB   <- dineroLB   + (datos_test[i,13]-1)
        dineroPS   <- dineroPS   + (datos_test[i,16]-1)
        dineroWH   <- dineroWH   + (datos_test[i,19]-1)
        dineroVC   <- dineroVC   + (datos_test[i,22]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }


  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))
    a <- c(1,2,3)
  return(g)
    return(a)
}

guanys(ofi,pro,datos_test)

```



```{r}


#########################################################
#Gráfico
#########################################################
ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred5#predicción(pronóstico

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos_test){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos_test[i,2]-1)
        dineroBW   <- dineroBW   + (datos_test[i,5]-1)
        dineroIW   <- dineroIW   + (datos_test[i,8]-1)
        dineroLB   <- dineroLB   + (datos_test[i,11]-1)
        dineroPS   <- dineroPS   + (datos_test[i,14]-1)
        dineroWH   <- dineroWH   + (datos_test[i,17]-1)
        dineroVC   <- dineroVC   + (datos_test[i,20]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos_test[i,3]-1)
```

```
        dineroBW   <- dineroBW   + (datos_test[i,6]-1)
        dineroIW   <- dineroIW   + (datos_test[i,9]-1)
        dineroLB   <- dineroLB   + (datos_test[i,12]-1)
        dineroPS   <- dineroPS   + (datos_test[i,15]-1)
        dineroWH   <- dineroWH   + (datos_test[i,18]-1)
        dineroVC   <- dineroVC   + (datos_test[i,21]-1)

       }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos_test[i,4]-1)
        dineroBW   <- dineroBW   + (datos_test[i,7]-1)
        dineroIW   <- dineroIW   + (datos_test[i,10]-1)
        dineroLB   <- dineroLB   + (datos_test[i,13]-1)
        dineroPS   <- dineroPS   + (datos_test[i,16]-1)
        dineroWH   <- dineroWH   + (datos_test[i,19]-1)
        dineroVC   <- dineroVC   + (datos_test[i,22]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }

    B365[i] <- dineroB365
    BW[i] <- dineroBW
    IW[i] <- dineroIW
    LB[i] <- dineroLB
    PS[i] <- dineroPS
    WH[i] <- dineroWH
    VC[i] <- dineroVC

  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

    graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos_test)
graf

```

```{r}
 plot(1:190, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
 ylim=c(-25,25),xlim=c(0,200))

lines(1:190,graf$BW,type="l",pch=2,col=3)
lines(1:190,graf$IW,type="l",pch=2,col=4)
lines(1:190,graf$LB,type="l",pch=2,col=5)
lines(1:190,graf$PS,type="l",pch=2,col=6)
lines(1:190,graf$WH,type="l",pch=2,col=7)
lines(1:190,graf$VC,type="l",pch=2,col=8)
lines(1:190,rep(0,190),type="l",pch=2,col=1)

 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=1:7,xjust=1)


```
La gráfica muestra, que si se realizan todas las apuestas en la casa PS, siguiendo como predictor/clasificador Naive Bayes, se obtendrá un beneficio neto positivo.
Faltaria añadir la linea en la gráfica que mostrará el beneficio neto si se apuesta en cada partido a la casa con la cuota más alta. Claramente seria el beneficio neto más alto.

También falta hacer el grafico teniendo en cuenta todos los metodos de predicción para una casa en concreto.(en el markdown graf)

#Gráfico apostando a la mejor cuota

```{r}
#########################################################
#Càlcul de guanys
#########################################################
ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred5#predicción(pronóstico

B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0

guanys <- function(ofi,pro,datos_test){
 dinero <- 0

 for(i in 1:length(ofi)){
  if(ofi[i]==pro[i]){

    if(ofi[i]== "H"){
     dinero <- dinero + max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                              datos_test[i,20]-1))
     if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                              datos_test[i,20]-1))==datos_test[i,2]-1){
      B365 <- B365 + 1
     }
     if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                              datos_test[i,20]-1))==datos_test[i,5]-1){
      BW <- BW + 1
     }
     if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                              datos_test[i,20]-1))==datos_test[i,8]-1){
      IW <- IW + 1
     }
     if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                              datos_test[i,20]-1))==datos_test[i,11]-1){
      LB <- LB + 1
     }
     if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                              datos_test[i,20]-1))==datos_test[i,14]-1){
      PS <- PS + 1
     }
     if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                              datos_test[i,20]-1))==datos_test[i,17]-1){
      WH <- WH + 1
     }
     if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
```

```
                        datos_test[i,20]-1))==datos_test[i,20]-1){
        VC <- VC + 1
                        }


      }else if(ofi[i]=="D"){

       dinero <- dinero + max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-
1,datos_test[i,18]-1,
                        datos_test[i,21]-1))
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,3]-1){
        B365 <- B365 + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,6]-1){
        BW <- BW + 1
       }
        if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,9]-1){
        IW <- IW + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,12]-1){
        LB <- LB + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,15]-1){
        PS <- PS + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,18]-1){
        WH <- WH + 1
       }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,21]-1){
        VC <- VC + 1
                        }

      }else if(ofi[i]=="A"){

       dinero <- dinero + max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-
1,datos_test[i,19]-1,
                        datos_test[i,22]-1))

      if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                        datos_test[i,22]-1))==datos_test[i,4]-1){
        B365 <- B365 + 1          }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                        datos_test[i,22]-1))==datos_test[i,7]-1){
        BW <- BW + 1
       }
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                        datos_test[i,22]-1))==datos_test[i,10]-1){
        IW <- IW + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                        datos_test[i,22]-1))==datos_test[i,13]-1){
        LB <- LB + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                        datos_test[i,22]-1))==datos_test[i,16]-1){
        PS <- PS + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                        datos_test[i,22]-1))==datos_test[i,19]-1){
        WH <- WH + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                        datos_test[i,22]-1))==datos_test[i,22]-1){
        VC <- VC + 1
```

```
                              }

      }
   }else{dinero <- dinero - 1

   }


 }
   g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
 return(g)
}

guanys(ofi,pro,datos_test)


```


Cuota mínimina

```{r}

predi=function(cuota){c("H","D","A")[which.min(cuota)]}
pr=apply(datos[,6:8],1,predi)#per cada fila diu quina és la quota mínima
y_pred6 <- pr[1711:1900]
(tab.nb <- table(y_pred6, datos[1711:1900,5]))
#0.5489474
```



```{r}
############################################################
#Jornades acertades
############################################################


ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred6#predicción(pronóstico)

acertados <- function(ofi,pro){
 s <- 0
 for(i in 1:length(ofi)){
   if(ofi[i]==pro[i]){
     s <- s +1
   }
 }
 return(s)
}

acertados(ofi,pro)
acertados(ofi,pro)/190
```



```{r}


############################################################
#Càlcul de guanys
############################################################


ofi <- datos_test$FTR #ofi:resultado oficial
```

```
pro <- y_pred6#predicción(pronóstico


guanys <- function(ofi,pro,datos_test){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos_test[i,2]-1)
        dineroBW   <- dineroBW   + (datos_test[i,5]-1)
        dineroIW   <- dineroIW   + (datos_test[i,8]-1)
        dineroLB   <- dineroLB   + (datos_test[i,11]-1)
        dineroPS   <- dineroPS   + (datos_test[i,14]-1)
        dineroWH   <- dineroWH   + (datos_test[i,17]-1)
        dineroVC   <- dineroVC   + (datos_test[i,20]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos_test[i,3]-1)
        dineroBW   <- dineroBW   + (datos_test[i,6]-1)
        dineroIW   <- dineroIW   + (datos_test[i,9]-1)
        dineroLB   <- dineroLB   + (datos_test[i,12]-1)
        dineroPS   <- dineroPS   + (datos_test[i,15]-1)
        dineroWH   <- dineroWH   + (datos_test[i,18]-1)
        dineroVC   <- dineroVC   + (datos_test[i,21]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos_test[i,4]-1)
        dineroBW   <- dineroBW   + (datos_test[i,7]-1)
        dineroIW   <- dineroIW   + (datos_test[i,10]-1)
        dineroLB   <- dineroLB   + (datos_test[i,13]-1)
        dineroPS   <- dineroPS   + (datos_test[i,16]-1)
        dineroWH   <- dineroWH   + (datos_test[i,19]-1)
        dineroVC   <- dineroVC   + (datos_test[i,22]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }


  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))
    a <- c(1,2,3)
  return(g)
    return(a)
}

guanys(ofi,pro,datos_test)

```



```{r}


##########################################################
```

```
#Gráfico
#########################################################
ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred6#predicción(pronóstico

B365 <- c()
BW <- c()
IW <- c()
LB <- c()
PS <- c()
WH <- c()
VC <- c()


guanys <- function(ofi,pro,datos_test){
  dineroB365 <- 0
  dineroBW   <- 0
  dineroIW   <- 0
  dineroLB   <- 0
  dineroPS   <- 0
  dineroWH   <- 0
  dineroVC   <- 0

  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dineroB365 <- dineroB365 + (datos_test[i,2]-1)
        dineroBW   <- dineroBW   + (datos_test[i,5]-1)
        dineroIW   <- dineroIW   + (datos_test[i,8]-1)
        dineroLB   <- dineroLB   + (datos_test[i,11]-1)
        dineroPS   <- dineroPS   + (datos_test[i,14]-1)
        dineroWH   <- dineroWH   + (datos_test[i,17]-1)
        dineroVC   <- dineroVC   + (datos_test[i,20]-1)

      }else if(ofi[i]=="D"){
        dineroB365 <- dineroB365 + (datos_test[i,3]-1)
        dineroBW   <- dineroBW   + (datos_test[i,6]-1)
        dineroIW   <- dineroIW   + (datos_test[i,9]-1)
        dineroLB   <- dineroLB   + (datos_test[i,12]-1)
        dineroPS   <- dineroPS   + (datos_test[i,15]-1)
        dineroWH   <- dineroWH   + (datos_test[i,18]-1)
        dineroVC   <- dineroVC   + (datos_test[i,21]-1)

      }else if(ofi[i]=="A"){
        dineroB365 <- dineroB365 + (datos_test[i,4]-1)
        dineroBW   <- dineroBW   + (datos_test[i,7]-1)
        dineroIW   <- dineroIW   + (datos_test[i,10]-1)
        dineroLB   <- dineroLB   + (datos_test[i,13]-1)
        dineroPS   <- dineroPS   + (datos_test[i,16]-1)
        dineroWH   <- dineroWH   + (datos_test[i,19]-1)
        dineroVC   <- dineroVC   + (datos_test[i,22]-1)

      }
    }else{dineroB365 <- dineroB365 - 1
        dineroBW   <- dineroBW   - 1
        dineroIW   <- dineroIW   - 1
        dineroLB   <- dineroLB   - 1
        dineroPS   <- dineroPS   - 1
        dineroWH   <- dineroWH   - 1
        dineroVC   <- dineroVC   - 1
    }

    B365[i] <- dineroB365
    BW[i] <- dineroBW
    IW[i] <- dineroIW
    LB[i] <- dineroLB
    PS[i] <- dineroPS
    WH[i] <- dineroWH
    VC[i] <- dineroVC
```

```
  }
    g <- data.frame(casa
=c("B365","BW","IW","LB","PS","WH","VC"),retorno=c(dineroB365,dineroBW,dineroIW,dineroLB,dineroPS,dineroWH,dineroVC))

    graf <- data.frame(B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
  return(graf)
}

graf <-guanys(ofi,pro,datos_test)
graf

```

```{r}
 plot(1:190, graf$B365,type="l",
 pch=100, col=2, xlab="Partido",
 ylab="Retorno neto(???)",
 main="Beneficio neto según la casa de apuesta",
 ylim=c(-25,25),xlim=c(0,200))

lines(1:190,graf$BW,type="l",pch=2,col=3)
lines(1:190,graf$IW,type="l",pch=2,col=4)
lines(1:190,graf$LB,type="l",pch=2,col=5)
lines(1:190,graf$PS,type="l",pch=2,col=6)
lines(1:190,graf$WH,type="l",pch=2,col=7)
lines(1:190,graf$VC,type="l",pch=2,col=8)
lines(1:190,rep(0,190),type="l",pch=2,col=1)

 #legend("topleft",legend=c("B365","BW","IW","LB","PS","WH","VC"),
 #pch=c(1,1),col=1:7,xjust=1)


```
La gráfica muestra, que si se realizan todas las apuestas en la casa PS, siguiendo como predictor/clasificador Naive Bayes, se obtendrá un beneficio neto positivo.
Faltaria añadir la linea en la gráfica que mostrará el beneficio neto si se apuesta en cada partido a la casa con la cuota más alta. Claramente seria el beneficio neto más alto.

También falta hacer el grafico teniendo en cuenta todos los metodos de predicción para una casa en concreto.(en el markdown graf)

#Gráfico apostando a la mejor cuota


```{r}
############################################################
#Càlcul de guanys
############################################################
ofi <- datos_test$FTR #ofi:resultado oficial
pro <- y_pred6#predicción(pronóstico


B365 <- 0
BW <- 0
IW <- 0
LB <- 0
PS <- 0
WH <- 0
VC <- 0


guanys <- function(ofi,pro,datos_test){
  dinero <- 0


  for(i in 1:length(ofi)){
    if(ofi[i]==pro[i]){

      if(ofi[i]== "H"){
        dinero <- dinero + max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-
1,datos_test[i,17]-1,
```

```
                        datos_test[i,20]-1))
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,2]-1){
       B365 <- B365 + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,5]-1){
       BW <- BW + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,8]-1){
       IW <- IW + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,11]-1){
       LB <- LB + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,14]-1){
       PS <- PS + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,17]-1){
       WH <- WH + 1
      }
      if(max(c(datos_test[i,2]-1,datos_test[i,5]-1,datos_test[i,8]-1,datos_test[i,11]-1,datos_test[i,14]-1,datos_test[i,17]-1,
                        datos_test[i,20]-1))==datos_test[i,20]-1){
       VC <- VC + 1
                        }


     }else if(ofi[i]=="D"){

      dinero <- dinero + max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-
1,datos_test[i,18]-1,
                        datos_test[i,21]-1))
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,3]-1){
       B365 <- B365 + 1
      }
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,6]-1){
       BW <- BW + 1
      }
       if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,9]-1){
       IW <- IW + 1
      }
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,12]-1){
       LB <- LB + 1
      }
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,15]-1){
       PS <- PS + 1
      }
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,18]-1){
       WH <- WH + 1
      }
      if(max(c(datos_test[i,3]-1,datos_test[i,6]-1,datos_test[i,9]-1,datos_test[i,12]-1,datos_test[i,15]-1,datos_test[i,18]-1,
                        datos_test[i,21]-1))==datos_test[i,21]-1){
       VC <- VC + 1
                        }

     }else if(ofi[i]=="A"){

      dinero <- dinero + max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-
1,datos_test[i,19]-1,
                        datos_test[i,22]-1))
```

```
      if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,4]-1){
        B365 <- B365 + 1          }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,7]-1){
        BW <- BW + 1
       }
        if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,10]-1){
        IW <- IW + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,13]-1){
        LB <- LB + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,16]-1){
        PS <- PS + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,19]-1){
        WH <- WH + 1
       }
       if(max(c(datos_test[i,4]-1,datos_test[i,7]-1,datos_test[i,10]-1,datos_test[i,13]-1,datos_test[i,16]-1,datos_test[i,19]-1,
                          datos_test[i,22]-1))==datos_test[i,22]-1){
        VC <- VC + 1
                          }

     }
   }else{dinero <- dinero - 1

   }


 }
   g <- data.frame(casa = "Mejor casa",retorno=dinero,B365=B365,BW=BW,IW=IW,LB=LB,PS=PS,WH=WH,VC=VC)
 return(g)
}

guanys(ofi,pro,datos_test)


```