UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

INGENIERÍA TÉCNICA DE TELECOMUNICACIÓN
SISTEMAS DE TELECOMUNICACIÓN

PROYECTO FINAL DE CARRERA

DEPARTAMENTO DE TEORÍA DE LA SEÑAL Y COMUNICACIONES

PREDICCIÓN DE EVENTOS DEPORTIVOS EMPLEANDO PROCESOS GAUSSIANOS: TENIS

AUTOR: JAVIER CÉSPEDES MARTÍN
TUTOR: FERNANDO PÉREZ-CRUZ

21 de Julio de 2011

Título: *PREDICCIÓN DE EVENTOS DEPORTIVOS EMPLEANDO PROCESOS GAUSSIANOS: TENIS.*

Autor: *JAVIER CÉSPEDES MARTÍN*

Tutor: *FERNANDO PÉREZ-CRUZ*

La defensa del presente Proyecto Fin de Carrera se realizó el día 21 de Julio de 2011; siendo calificada por el siguiente tribunal:

Presidente: *Harold Molina-Bulla*

Secretario *Jose Miguel Leiva Murillo*

Vocal *David Delgado Gómez*

Habiendo obtenido la siguiente calificación:

Calificación:

Presidente **Secretario** **Vocal**

Agradecimientos

En estos días en los que el final de esta etapa está tan cerca, solo cabe mirar hacia atrás y ver que hubo comienzos duros, que no todo salía como se esperaba, pero que casi como una constante en mi vida, el esfuerzo y la entrega hicieron que esto saliera adelante. Muy lejos queda aquel día en el que mi madre y yo cogimos aquel tren para venir, para enseñarme el trayecto que tenía que hacer, y después se volvió, ella sola. Uno de tantos gestos que ha hecho por mí, su apoyo y ánimos siempre me han acompañado a lo largo de esta y todas las etapa de mi vida.

MAMÁ TE QUIERO

Mis hermanos, porque yo sé que me quieren, y yo estaré para que ellos pasen esta etapa cuando llegue su momento de la mejor forma posible. Mi familia está conmigo y soy feliz, y no solo aquí Papá, Abuelo sé que estáis orgullosos, que se hace camino al andar y que mi esfuerzo ha servido la pena, y siempre vuestra sonrisa estará en mi mente. Abuela, porque tienes tus cosas, eres muy tuya pero también tienes unos grandes momentos.

Mis amigos, tanto los amigos del día a día, fueron unos grandes momentos en la residencia, allí comienza el desarrollo de la persona en un mundo nuevo, y las experiencias vividas duran por siempre, todos tenemos las nuestras y fue un placer estar con todos. A los que luego comenzaron la segunda parte, el piso. A los compañeros de clase, sobre todo a los amigos, esas reuniones informales, las risas del primer años, donde estábamos todos, por acabar las partidas de mus en los descansos, cuando aún quedaban diez minutos para ir a clase al grito de *¡ORDAGO!*. Por mis amigos en mi pueblo, porque ellos me tratan como uno más a pesar de no estar allí con ellos, son muy grandes y me apoyan mucho.

A todos y cada uno de los profesores que me han impartido asignaturas, porque aunque a veces sus decisiones no las comparto, al fin y al cabo han dejado parte en mí, y ello me ha llevado a este fin de etapa. No me puedo olvidar de mi Tutor, le estoy muy agradecido, porque el me dio la oportunidad de reunir dos de mis actividades preferidas de ocio y hacer este proyecto, sé que al principio le costó brindarme la oportunidad, que el trabajo que tiene es mucho, no sé si vio el entusiasmo que tenía o que es lo que vio, aunque por mail es difícil ver este tipo de cosas, pero iniciandonos con el mundo que vamos a tratar enseguida, realizó una *Apuesta a Caballo Ganador* y no para una sola carrera, sino a *largo plazo*, ahora me toca a mí cumplir con mi parte. Muchas Gracias.

A las personas experimentadas del mundo de las apuestas, porque cuando les he consultado algo referente a este mundo me han ayudado sin esperar nada a cambio, porque han sabido valorar el trabajo, porque han dado ánimos.

Para finalizar a todas y cada una de las personas que conozco, porque aunque lo que hayamos compartido sea poco o mucho, yo estoy formado con grandes y también con pequeños fragmentos de todos y cada uno de vosotros.

Debemos admitir con humildad que, mientras el número es puramente un producto de nuestra mente, el espacio tiene una realidad fuera de nuestra mente, de modo que no podemos describir completamente sus propiedades a priori.

Carl Friedrich Gauss

No creo que haya alguna emoción más intensa para un inventor que ver alguna de sus creaciones funcionando. Esa emoción hace que uno se olvide de comer, de dormir, de todo.

Nikola Tesla

Resumen

En la sociedad actual el deporte es una de las actividades de ocio más recurrentes, desde unos años atrás la combinación de deporte y apuestas producen emoción, pero esto no siempre es sinónimo de ganancias. Por lo que proponemos un sistema que basado en eventos deportivos ocurridos, obteniendo una serie de características significativas, seamos capaces de predecir el resultado en un evento deportivo futuro, en este caso para el Tenis, para una vez calculadas las probabilidades poder apostar. En la creación del modelo de estimación de probabilidades emplearemos Procesos Gaussianos para Clasificación.

Índice general

Lista de Figuras

Lista de Tablas

Capítulo 1

INTRODUCCIÓN

1.1. El Mundo de las Apuestas

Cuando hablamos de apuestas deportivas debemos tener en cuenta como en muchos otros aspectos de la vida las probabilidades, ya que ambas, apuestas deportivas y probabilidades están fuertemente relacionadas. Las apuestas deportivas están basadas en la probabilidad que da un *trader* a la ocurrencia de un evento, cuando la obtiene puede calcular la cuota y es que la cuota está relacionada con la probabilidad.

Existen 3 visualizaciones de cuotas distintas a elegir, que han sido establecidas por los corredores y/o casas de apuestas a lo largo de los años. El formato de cuota corriente en Europa (excepto en Inglaterra). En esta visualización de cuota (2.20): el importe apostado se multiplica por la cuota mostrada. Ejemplo: 100€ × 2.20 = 220€ ganancia de la apuesta. Descontando el importe apostado, la ganancia neta sería de 120€. Pero también existe el sistema de cuotas americano. En esta representación (+120) ó (-200), siempre se muestra la ganancia neta. Es decir, que se muestra o bien el importe a apostar para ganar 100€, o cuánto se ganaría si se apostase un importe de 100€. Un signo positivo antes del número en nuestro caso +120 indica, que en el caso de acertar la apuesta, se obtendrían 120€ de ganancia neta por un importe apostado de 100€. En el caso de un signo negativo -200, significa que se deben apostar 200€ para obtener 100€ de ganancia neta. Inglaterra tiene una representación propia. En este caso no se muestra la ganancia neta, sino la cuota neta (6/5). En nuestro ejemplo 6/5 corresponde a una cuota de Europa continental de 2,20. Porque, al dividir 6 entre 5, se obtiene la cuota neta de 1,20. Es decir, con un importe apostado de 100€, la ganancia neta sería de 120€.

Favorito	No Favorito
1.8	2.1
1,81	2,09
1,82	2,07
1,83	2,06
1,84	2,05
1,85	2,03
1,86	2,02
1,87	2,01
1,88	1,99
1,89	1,98
1,9	1,98
1,91	1,97
1,92	1,96
1,93	1,95
1,94	1,94

Tabla 1.1: *Evolución de Cuotas Pre-Partido*

La relación que en un sistema justo tienen probabilidad y cuota es el inverso. Pero esto nunca es así, entre la probabilidad que calcula el *trader* y la cuota que obtiene hay un margen de variación, está variación es la ganancia fija que obtiene la casa de apuestas para la que trabaja el *trader*.

La cuota no es algo que esté fijo en el tiempo y evoluciona de diversas formas. El principal movimiento que se produce en las cuotas que son fijadas por las casas de apuestas es para cubrir el riesgo, si bien el *trader* puede estimar una probabilidad, pongamos un ejemplo, en un partido de tenis el *trader* marca las siguientes cuotas (1.8 para el favorito) y (2.1 para el no favorito), la percepción para los apostantes es que es un partido muy igualado, con lo que se empezaran a apostar a la cuota del jugador no favorito, con lo que una posible evolución la veremos en la Tabla 1.1.

En realidad la mayoría de las veces las cuotas cambian por el dinero que entra en ellas, aunque si bien es cierto que las casas están pendientes, es posible que hagan algún movimiento en ellas para cubrirse más y tener más ganancias. Estos casos pueden ser lesiones, o rumores de malestar de un jugador, incluso últimamente los comentarios de los tenistas en sus redes sociales. Estos aspectos causan fuertes cambios en las cuotas y no lo que hemos observado en la Tabla 1.1.

1.2. Objetivo

Nuestro objetivo en adelante será calcular la probabilidad de ocurrencia de eventos mediante procesos gaussianos, para a partir de esta probabilidad que en nuestro caso y debido al estudio debe de ser mejor que las de las casas de apuestas, obtener la apuesta óptima basándonos en el Criterio de Kelly.

1.3. Estimación de Probabilidades

Una vez se ha comprendido como se obtienen las cuotas y su relación con las probabilidades lo realmente importante y donde radican las ganancias a largo plazo en el mundo de las apuestas deportivas es en obtener una adecuada estimación de probabilidades. Este es el principal problema al que nos vamos a enfrentar, además de proponer modelos para este largo plazo del que hablamos. Vamos a tratar con un deporte en particular, como es el Tenis, un deporte individual y por parejas en el que hay muchas estadísticas sobre los puntos, además es un deporte en el que no tenemos un tiempo determinado de duración de un partido, para que un partido acabe un jugador tiene que ganar una bola de partido, o también ganará si su oponente se retira o no se presenta. A pesar de que el tema de lesiones y reducción del rendimiento físico por parte de los jugadores es más acentuado que en otros deportes, en el tenis hay mucha estadística, simplemente la posición mundial que ocupan cada uno de los tenistas en el ranking ATP, es el primero de ellos.

Hay una literatura extensa, sobre la estimación de probabilidades en el tenis, en la sucesión de distintos eventos, debido sobre todo a que en el tenis hay muchos mercados en los que apostar. Debemos diferenciar entre eventos en directo y pre-evento, en los eventos en directo existe la posibilidad de apostar a cada punto, por lo que tendremos en cuenta el estado del jugador, marcador actual y otra serie de variables. En este caso las cuotas son actualizadas a cada momento y las probabilidades deben estimarse en un corto instante de tiempo.

1.4. Estado del Arte

En ocasiones para afrontar este gran problema de la estimación de probabilidades se utilizan simplificaciones, la más común en los métodos de predicción que se actualizan punto a punto es tomar cada uno de los puntos como independientes e idénticamente distribuidos (iid), de esta manera podemos desde un principio calcular las probabilidades del final del partido e ir actualizando conforme el partido va transcurriendo. Así es como en forma de árbol, muy similar a un algoritmo de Viterbi vamos obteniendo el resultado basándonos en las probabilidades de que el jugador que está sacando gane el punto, para ello los datos estadísticos principales son los puntos que consigue cada uno de los jugadores durante su saque y durante su resto y los combina para obtener la citada probabilidad [15] [27] [32] [23] [29] [28], este sistema utiliza la asunción anteriormente comentada. La cuestión es clara y debemos preguntarnos, ¿Realmente podemos asumir tal distribución en los puntos? Se ha estudiado en diversos puntos y el tratamiento de los puntos de esta forma no es exacto, ganar los primeros puntos da un plus sobre el punto que se disputa a continuación, la desviación con respecto a los puntos iid depende de la calidad del jugador, para los mejores jugadores la desviación es pequeña. Por ello aunque la asunción de que los puntos sean iid no es correcta es una buena aproximación y si se puede cambiar los parámetros que definen al jugador [22].

Aun así seguimos con nuestras preguntas ¿Es suficiente con obtener dos o tres datos estadísticos de ambos jugadores?, y además hacer una suposición en el comportamiento de los puntos. A parte de la estadística de puntuación y debido al nivel que estamos tratando, nivel profesional, en el que cualquier pequeño cambio es percibido, nos podemos encontrar que 0.5kg más de tensión de lo normal es percibido por el tenista y se encuentre incomodo en su juego y por supuesto no debemos olvidarnos de la superficie sobre la que se está jugando, en el circuito hay muy pocos torneos sobre césped, una superficie que puede beneficiar a ciertos jugadores con un buen saque, ya que la bola no bota mucho y si es muy rápida, es muy difícil restarla. Por ello la ausencia de este tipo de torneos beneficia las características de otros jugadores, además en superficies más duras es más frecuente encontrar lesiones entre los jugadores [16]. Los tenistas son personas como cualquiera de nosotros con lo que sus pensamientos, sentimientos, molestias y sensaciones tienen que estar presentes en todo momento mientras ellos están jugando, por ello también se ha analizado estadísticamente el lado psicológico de los jugadores durante un torneo. Uno de estos efectos es el comportamiento del jugador ante el marcador, el jugador que comienza sacando un

set y ambos jugadores mantienen su servicio, va en todo momento por delante en el marcador, con lo que tiene cierta ventaja psicológica [30] [24], aunque debido a las reglas del Tenis esto no debería de ser así, de hecho hay ciertos jugadores, que son grandes restadores que prefieren si ganan el sorteo comenzar restando, este hecho se debe a que salen muy fuerte en el partido y si consiguen un break, en estos momentos la ventaja que consiguen es mucho mayor, Rafael Nadal o David Nalbandian son dos claros ejemplos de este tipo de jugadores. Otro aspecto relevante en la psicología del tenista, son los momentos en los que se da un plus de entrega, son los momentos claves del encuentro, en realidad no están fijados, pero sí que se puede decir que las bolas de break y la mayoría de bolas a partir del 4-4 en cualquiera de los sets. Este plus de entrega lo pueden realizar los mejores jugadores y así en puntos antes de un *deuce* incrementar su esfuerzo y también reducir el esfuerzo en puntos menos importantes [34], acentuándose cuando estos jugadores han conseguido un break de ventaja, claros ejemplos de este comportamiento son Roger Federer y Pete Sampras. Nosotros estudiamos todo este mundo de la estadística, pero es que los propios jugadores de forma natural, en su cabeza y mientras que están jugando están pendientes de ello también [33], así es como por ejemplo Rafael Nadal cuando se enfrenta a Roger Federer le juega a este al revés y con bote alto, debido a que el revés que utiliza Federer, a una mano ante este tipo de golpe no es muy efectivo produciendo a Roger Federer muchos errores no forzados. Pero no solo se utiliza la estadística durante el partido, también se puede utilizar antes, para preparar un partido. Estudiar como juega tu rival puede ayudar a afrontar el partido de forma satisfactoria. Uno de los aspectos básicos en el tenis es el comienzo de un punto, el saque, hay varias estadísticas en torno a él. Estudiar como saca tu rival o donde resta mejor puede dar un plus para la victoria del partido [35].

1.5. Base de Datos

Todo el rato estamos hablando de estadística, pero ¿Qué es la estadística sin datos?, para nosotros este un apartado fundamental, para obtener los datos y después poder trabajar con ellos utilizamos 2 principales fuentes [2] [7], en el mundo de las apuestas es normal tener algunos datos, pero hemos querido obtener los más posibles. Para diferenciar entre unos datos y otros se ha preguntado a varios pronosticadores, que hacen públicas sus apuestas para que el resto de apostantes puedan seguir sus apuestas, estos pronosticadores tienen una estadística en cuanto a las ganancias que obtienen, con lo que son pronosticadores contrastados. Sus indicaciones nos

han llevado a confeccionar una base de datos con varios datos, incluso con cuotas para testear nuestro sistema.

Una de nuestras fuentes para captar la información es una hoja de cálculo, un gran avance para el posterior tratamiento, debido a la integración y la capacidad de trabajar con ella en Matlab, pero nos encontramos con el problema de que su información es insuficiente, por ello tenemos que recurrir a la segunda fuente, en esta fuente la información se encuentra en tablas de formato web, esta información es la misma que tiene la ATP [1], con el inconveniente que está nos fue más difícil de conseguir. Para la obtención de los datos tuvimos que hacer un pequeño código en java y obtener así los datos de cada uno de los partidos de nuestro interés y guardarlos, a la hora de guardarlos elegimos por compatibilizar con los datos que ya poseíamos, otra hoja de cálculo [9]. El problema que nos surge a continuación es que estamos en la posesión de dos hojas de cálculo, nuestro ideal es poseer únicamente una de ellas. Por lo que la solución más rápida sería cortar y pegar datos de una de ellas a la otra, pero nos encontramos con el problema que en ellas los partidos son los mismo pero el orden en el que se encuentran estos partidos son diferentes, entonces sería pasar uno a uno cada partido. Así pues en base a estas dos hojas de cálculo hicimos otra, mediante una combinación de las dos con otro código java, al igual que en el anterior utilizamos la herramienta un API específico para tratar con hojas de cálculo [9]. Esta hoja de cálculo es tratable por la que será nuestro software de trabajo Matlab. Para poder ir actualizando la base de datos, se puede hacer de 2 maneras: de manera manual, el número de partidos en un día pueden llegar a unos 30 como mucho en los *Grand Slam*, pero en el resto de torneos, que son mayoría pueden ser unos 8-10 partidos. Esta es la manera en la que tendremos que actualizar si queremos predecir día a día. En cambio sí hemos pasado un tiempo que no hemos realizado ninguna predicción tenemos la posibilidad de utilizar conjuntamente nuestros códigos y actualizar nuestra hoja de cálculo, ponerla al día y realizar las predicciones y posteriormente actualizar manualmente. Cuando dispusimos de la primera hoja de cálculo, fue para nosotros importante que los datos en ellas reflejadas fueran los necesarios, para ello nos pusimos en contacto de nuevo con los pronosticadores y quedaron satisfechos, son datos que no se pueden encontrar en ningún lugar de forma tan resumida y que es tan legible tanto para la lectura humana como para el procesamiento por parte de un computador.

Ahora vamos a ver los tipos de datos que vamos a tratar:

1. El Torneo, es interesante conocer el torneo en el que se disputa el partido, ya que la defensa

de puntos puede motivar a un jugador a hacer una buena actuación o por el contrario dejarse llevar y perder incluso en primera ronda ante un rival menor. A este hecho se le conoce en el mundo del tenis como "Coger el cheque", los grandes jugadores solicitan inscripción a torneos y son admitidos en 1ª Ronda, con lo que su premio es considerable.

2. Fecha de disputa del partido, es de utilidad para conocer y temporizar la sucesión de distintos eventos, por ejemplo la disputa del último partido.

3. El tipo de torneo que se disputa, existen 4 categorías distintas de torneo y uno que es único y especial. Los torneos son ATP 250, ATP 500, ATP 1000 y *Grand Slam*, el torneo único es la Copa de Maestros y es el último torneo de la temporada disputado entre los 8 mejores tenistas del ranking.

4. El tipo de pista, debido a las condiciones climatológicas es frecuente sobre todo en la época de invierno, ya que el único mes donde no se disputa ningún partido es en diciembre, que nos encontremos con pistas cubiertas y con las tradicionales pistas exteriores.

5. El tipo de superficie, en la actualidad existen 3 tipos principales de superficies, aunque en años anteriores eran 4, la superficie moqueta fue eliminada del ATP TOUR a partir de 2009. En la actualidad las superficies son: Superficie Dura, Hierba y Tierra Batida.

6. Ronda de la disputa del partido, al igual que el Torneo es interesante ver cómo se comporta cada jugador en una situación y en otra.

7. El número máximo de sets al que se disputa el partido, disponemos de 2 modalidades, los partidos al mejor de 3 ó 5 sets, los partidos al mejor de 5 sets están reservados para los *Grand Slam*, el resto de partidos son al mejor de 3 sets.

8. Ganador y Perdedor del partido, en conjunción con la fecha de los partidos nos indica rachas de los jugadores.

9. Rankings, indica la posición en el ranking ATP TOUR.

10. Los resultados de los partidos de forma completa, no solo nos indica el resultado en el número de sets, como puede ser 2-0, 2-1, 3-0, 3-1, 3-2, sino que además nos indica el resultado de los juegos en cada uno de los sets.

11. El tipo de final que tuvo el partido, en el tenis si el tenista se lesiona durante el partido o durante el torneo y ha disputado la 1ª ronda, el partido no se continua disputando o directamente no se disputa, siendo su finalización como retirado o como *Walk/Over* respectivamente. Si el tenista no disputa el partido de 1ª ronda, entonces otro jugador de la ronda previa que no se clasificó para la ronda final, *Lucky Loser*, ocupa su lugar.

12. El porcentaje de primeros servicios puestos en juego por el jugador.

13. Aces, el número de saques directos del jugador.

14. Las dobles faltas del jugador, el número de veces que el jugador no logra realizar el servicio en ninguna de las dos ocasiones que dispone para ello.

15. El porcentaje de puntos ganados por el jugador cuando realiza un primer servicio.

16. El porcentaje de puntos ganados por el jugador cuando realiza un segundo servicio.

17. El porcentaje de puntos ganados por el jugador cuando resta sobre el primer servicio.

18. El porcentaje de puntos ganados por el jugador cuando resta sobre el segundo servicio.

19. El porcentaje de puntos ganados con el servicio.

20. El porcentaje de puntos ganados al resto.

21. El porcentaje de puntos de break ganados.

22. Los puntos ganados en el partido por el jugador.

23. Las cuotas de cada uno de los partidos en diversas casas de apuestas.

Con los datos anteriormente comentados obtenemos un total de 49 variables que se introducirán en nuestro sistema, alguna de ellas son directamente datos, otras de ellas son la combinación y/o tratado de los mismos. Nuestro sistema predictor va a estar basado en las variables que obtendremos del tratamiento de la base de datos, así podremos obtener datos de la trayectoria de un jugador, su comportamiento en torneos similares, lo efectivo que es su saque, su resto, analizaremos más a fondo el comportamiento del jugador, no solo en los partidos anteriores, sino que tendremos una memoria residual de lo que pudiera haber hecho en el pasado. Para ello vamos a utilizar procesos gaussianos como herramienta para crear una máquina de aprendizaje

con reentrenamiento, por lo que deberemos de tener una buena base de datos y dividirla en datos de entrenamiento y datos de test, para que todo sea consistente en el tiempo, una vez que nos encontremos en el testado del sistema y cuando los partidos de test hayan sido predichos estos deben entrar a la máquina para entrenarla, esto es debido fundamentalmente a que el tenis se desarrolla como un torneo de K.O., así el jugador que gana pasa a la siguiente ronda y el jugador que pierde ya no disputará ningún partido más, por lo que es interesante para predecir un partido de 2ª ronda y posteriores conocer lo que el jugador ha hecho justo en la ronda previa. La salida de nuestro proceso gaussiano es la probabilidad de victoria de uno de los tenistas, el de mejor ranking, pero como las posibilidades son dos por teoría de probabilidad, tenemos la del adversario, ahora el siguiente paso consiste en lo que en el mundo de las apuestas se conoce como "buscar el value" para ello utilizaremos el Criterio de Kelly [21].

1.6. Herramientas

Ahora vamos a describir las herramientas que vamos a utilizar en todo nuestro predictor.

1.6.1. Java

Como anteriormente hemos visto para la creación de nuestra base de datos hemos utilizado el lenguaje java, el lenguaje java fue desarrollado por Sun Microsystems en los años 90, es un lenguaje de programación orientado a objetos y tiene una gran herencia del lenguaje C, aunque con una gran simplificación en los aspectos de bajo nivel ya que no es necesario el tratamiento de punteros, ni de memoria. Su gran versatilidad radica en que no se necesita ningún sistema operativo en concreto para que funcióne, utiliza una máquina virtual que puede ser instalada en cualquier sistema operativo y dispositivo, incluyendo dispositivos móviles. En el año 2006 se convierte en un lenguaje con licencia libre por lo que su utilización aumenta, así es como surgen aparte de las librerías (API) propias del código otras que son de máxima utilidad, este es nuestro caso hemos utilizado el JXL [9], es un API para el tratamiento de hojas de cálculo y nos permite crear, escribir, leer y la mayoría de operaciones que podamos imaginar sobre una hoja de cálculo. Los puntos fuertes de la filosofía de este lenguaje de programación son:

- Orientación a objetos
- Ejecución en diversos sistemas operativos

- Soporte en red
- Utilidad en sistemas remotos
- Fácil de usar y toma lo mejor de otros lenguajes, como puede ser C

Para nosotros el puntos fuerte es el tercero, por ello su elección para resolver nuestro problema, para obtener los datos estamos mandado una petición web y recibiendo las respuestas a ellas, mediante la red, creamos un flujo de datos por cada una de ellas, algo que para nosotros es de suma importancia en la actualización de la base de datos en un tiempo razonable [17] [20].

1.6.2. MATLAB

Otra de nuestras herramientas y nexo de unión entre todas ellas es MATLAB es un software matemático y de ingeniería que ofrece un desarrollo integrado y un lenguaje propio (Lenguaje M), está disponible en diferentes sistemas operativos, pero lo que a nosotros más nos interesa y por lo que definitivamente hemos optado por él, es su orientación a matrices, en su propio nombre ya nos indica esta característica básica para nosotros. "MATrix LABoratory", (laboratorio de matrices). De manera similar a como funciona java con sus API, MATLAB utiliza una serie de toolbox, estos son conjuntos de funciones, algunas al igual que pasa con el lenguaje se Sun Microsystems, son propias y se distribuyen de inicio con la compra de la licencia del software, pero otras son creadas por usuarios, centros de investigación o universidades. Este último es el caso que nosotros vamos a tratar, el toolbox de procesos gaussianos, ha sido desarrollado como un toolbox adicional por Carl Edward Rasmussen y Chris Williams [14].

1.6.3. Criterio de Kelly

John Larry Kelly Jr. Físico que trabajó en los laboratorios Bell, allí coincidió con Graham y Shannon. La herramienta que vamos a utilizar de él es el Criterio de Kelly, combina la teoría de la información con los juegos de apuestas y casino. A pesar que nunca utilizó su fórmula, está permite basándose en los conocimientos de una probabilidad de ocurrencia de un evento y una cuota en dicho evento calcular la cantidad óptima para maximizar las ganancias. En nuestro caso el desarrollo de este apartado lo uniremos a la gestión del *bank*, el *bank* es el dinero dedicado unicamente a las apuestas, debe ser una cantidad que no necesitemos en nuestra vida diaria, como si fuera una actividad de ocio más.

La simplificación de este criterio para eventos con 2 posibles desenlaces.

$$\langle Bank(\,\%)\rangle = \frac{(C \times Pr') - 1}{C - 1} \times 100 \tag{1.1}$$

- C= Cuota
- Pr'= Probabilidad de salida de nuestro predictor

Un ejemplo de aplicación de la formula (1.1), que podemos ver en la Tabla 1.2, cuando las cuotas son para el jugador favorito $C = 1{,}80$ y para el jugador no favorito $C = 2{,}1$, la probabilidad de victoria es la misma para los dos: $Pr' = 0{,}5$. Aplicando la fórmula para el jugador favorito tenemos:

$$\langle Bank(\,\%)\rangle = \frac{(1{,}8 \cdot 0{,}5) - 1}{1{,}8 - 1} \times 100 = -12{,}5\,\%$$

Como es menor a 0, NO se apuesta

Aplicando la fórmula para el jugador no favorito tenemos:

$$\langle Bank(\,\%)\rangle = \frac{(2{,}1 \cdot 0{,}5) - 1}{2{,}1 - 1} \times 100 = 4{,}545\,\%$$

El 4.545 % del total de nuestro *bank* es lo más recomendable apostar.

Aunque si bien, la Formula (1.1) es una simplificación, la posibilidad de utilización es muy potente, ya que podemos utilizarla en eventos con múltiples desenlaces.

- Primero:

$$\mathbf{t} = \mathbf{p} \times \mathbf{c} \tag{1.2}$$

donde:

  - **p** es el vector de probabilidades que arroja nuestro predictor.
  - **c** es el vector de cuotas que hay en el mercado.
  - Además obtendríamos un vector de como están ordenadas: **o**.

- Segundo: ordenaremos **t** de mayor a menor. De forma que si el mayor es menor a 1, habremos acabado, NO apostaremos.

- Tercero: En el caso que el mayor supere 1. Deberemos calcular por separado cada componente del vector **f** ya que en su cálculo depende de la posición que ocupe.

$$f_n = \frac{1 - \sum_{i=1}^{n} p_{o_i}}{1 - \sum_{i=1}^{n} \frac{1}{c_{o_i}}} \tag{1.3}$$

- Cuarto: Tomaremos como punto pivote el valor menor superior a 0, del vector **f** calculado en (1.3)

$$piv = \text{mín}\,\mathbf{f} \tag{1.4}$$

- Quinto: Para finalmente calcular el vector con la cantidad a apostar a cada uno de los mercados

$$\mathbf{a} = \left[\frac{p_1 - piv}{c_1}, \ldots, \frac{p_n - piv}{c_n}\right] \times 100 \tag{1.5}$$

Tomando únicamente los valores positivo.

Para mostrar lo anterior vamos a tomar el ejemplo del resultado exacto en un partido de tenis al mejor de 3 sets, los posibles resultados que se pueden dar es que el jugador que gane, lo haga con un marcador de 2-0 ó 2-1. En nuestro caso vamos a tomar que las probabilidades que arroja el predictor son:

- 30 % que gane el favorito 2-1
- 35 % que gane el favorito 2-0
- 10 % que gane el no favorito 2-1

- 25 % que gane el no favorito 2-0

$$\mathbf{p}=[0.3\ 0.35\ 0.1\ 0.25].$$

Las cuotas son:

- 3.5 que gane el favorito 2-1
- 2.5 que gane el favorito 2-0
- 8 que gane el no favorito 2-1
- 4.25 que gane el no favorito 2-0

$$\mathbf{c}=[3.5\ 2.5\ 8\ 4.25]$$

De manera que a partir (1.2) obtenemos $\mathbf{t}$=[1.0625 1.0500 0.8750 0.8000] y $\mathbf{o}$=[4 1 2 3]

Con (1.3) y (1.4) tenemos $\mathbf{f}$=[1.0000 0.9808 0.9395 1.2660] y piv=0.9395

Para concluir de (1.5) calculamos $\mathbf{a}$=[3.1579 -2.5789 -1.7434 2.8947].

Los mercados a apostar serán:

- El jugador favorito que gana 2-1 a cuota 3.5 un 3.1579 % del *bank*.
- El jugador no favorito gana 2-0 a cuota 4.25 un 2.89474 % del *bank*.

Así este criterio [21] nos ayudará en la cantidad optima a apostar y no solo eso, ya que este criterio también nos puede indicar lo que en el mundo de las apuestas y las finanzas se conoce como arbitraje, el arbitraje en las apuestas consiste en realizar diversas apuestas a un evento de manera que ocurra lo que ocurra se generen ganancias, esta es la premisa inicial, aunque también es cierto que se puede compensar y dejar todas las ganancias en un único desenlace o en diferentes posibilidades intermedias [25].

Kelly demostró que el criterio es válido cuando la apuesta se realiza varias veces sobre eventos con la misma probabilidad, esta aproximación nosotros la tenemos en cuenta, ya que la temporada de tenis en el ATP TOUR, tiene alrededor de 2600 partidos, con lo que ante una buena estimación de la probabilidad el número de eventos es suficiente, apoyandonos como hace Kelly en la ley de los grandes números.

1.6.4. Procesos Gaussianos

Los Procesos Gaussianos para máquinas de aprendizaje son herramientas Bayesianas no lineares de estimación y detección, que proporcionan un alto grado de confianza en sus estimaciones y predicciones. Así pues nos encontramos con los Procesos Gaussianos para Regresión (GPR) y los Procesos Gaussianos para Clasificación (GPC). GPR se caracteriza por una solución analítica teniendo en cuenta la matriz de covarianza, en base a ella se puede estimar la matriz de covarianza de los datos. Los GPR se propusieron en 1996, para posteriormente dar paso a los GPC. Está demostrada en la amplia literatura las múltiples utilidades, en el ámbito de las Telecomunicaciones [31] [18] y en otros aspectos como en los movimientos sismicos [26] o en otros campos como el que nos es de estudio.

Procesos Gaussianos para Regresión

Los GPR son una herramienta Bayesiana para que las máquinas de aprendizaje puedan predecir la probabilidad a posteriori para la salida (b_*) dada por la entrada (x_*) y el conjunto de entrenamiento ($\mathcal{D} = \{x_i,b_i\}_{i=1}^n$, con $x_i \in \mathbb{R}^d$ y $b_i \in \mathbb{R}$)

$$p(b_*|x_*, \mathcal{D}). \tag{1.6}$$

En GPR se asume una función con valores reales (función latente) de manera que como la función se genera mediante un proceso gaussiano, la función también es gaussiana, de media cero y cuya covarianza viene dada por la función $k(x,x')$. La función de covarianza, que también es llamada *kernel* relaciona cada punto de entrada y las características que describen el proceso gaussiano.

Para un conjunto finito de muestras entrada, el proceso gaussiano se convierte en una gaussiana multidimensional definida por su media y la matriz de covarianza. Así el proceso se convierte en:

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(0, \mathbf{K}), \tag{1.7}$$

donde $\mathbf{f}=[f(x_1),f(x_2),\ldots, f(x_n)]^\top$, $\mathbf{X} = [\mathrm{x}_1, x_2, \ldots, x_n]$ y $(\mathrm{K})_{ij} = (\mathrm{x}_i,x_j)$, $\forall\, \mathrm{x}_i,x_j \in \mathcal{D}$.

Una vez que tenemos etiquetados los puntos de test $\mathbf{b}=[b_1, b_2, \ldots, b_n]^\top$, con x_*, podemos calcular (1.6) utilizando las herramientas Bayesianas de la estadística: El Teorema de Bayes. Así pues aplicaremos Bayes para obtener la probabilidad de densidad a posteriori de la función latente:

$$p(\mathbf{f}, f(x_*)|\mathcal{D}, x_*) = \frac{p(\mathbf{b}|\mathbf{f}, \mathbf{X})p(\mathbf{f}, f(x_*)|\mathbf{X}, x_*)}{p(\mathbf{b}|\mathbf{X})}, \tag{1.8}$$

donde $\mathcal{D}$=$\{\mathbf{b},\mathbf{X}\}$, la probabilidad $p(\mathbf{f}, f(x_*)|\mathbf{X}, x_*)$ es el proceso gaussiano obtenido con (1.7) extendido a las entradas de test, $p(\mathbf{b}|\mathbf{f}, \mathbf{X})$ es la probabilidad de que la función latente pertenezca al conjunto de entrenamiento, $p(\mathbf{b}|\mathbf{X})$ es la evidencia del modelo y garantiza que la distribución de probabilidad pertenece a él. Un modelo utilizado para función de distribución:

$$p(\mathbf{b}|\mathbf{f}, \mathbf{X}) = \prod_{i=1}^{n} p(b_i|f(x_i), x_i) \tag{1.9}$$

Debido a que los datos etiquetados se han obtenido de manera iid, entonces asumimos que las observaciones de la función latente tienen ruido, $b_i = f(x_i) + \nu$ y que este ruido tiene una distribución gaussiana.

$$p(b_i|f(x_i), x_i) = \mathcal{N}(0, \sigma_\nu^2). \tag{1.10}$$

El cálculo de (1.8) y obtener Gaussianas Multidimensionales simplifica los cálculos para obtener (1.6)

Podemos obtener la densidad de probabilidad a posteriori para la salida en (1.6) para los datos de test condicionados a los datos de entrenamiento y x_* basándonos en la función latente:

$$p(b_*|x_*, \mathcal{D}) = \int p(b_*|f(x_*), x_*)p(f(x_*)|\mathcal{D}, x_*)df(x_*) \tag{1.11}$$

donde:

$$p(f(x_*)|\mathcal{D}, x_*) = \int p(f(x_*), \mathbf{f}|\mathcal{D}, x_*)d\mathbf{f}. \tag{1.12}$$

Podemos calcular (1.11) utilizando las propiedades de las gaussianas.

$$p(b_*|x_*, \mathcal{D}) = \mathcal{N}(\mu_{b_*}, \sigma_{b_*}^2), \tag{1.13}$$

donde

$$\mu_{b_*} = \mathbf{k}^\top \mathbf{C}^{-1}\mathbf{b}, \tag{1.14}$$

$$\sigma_{b_*}^2 = k(x_*, x_*) - \mathbf{k}^\top \mathbf{C}^{-1}\mathbf{k}, \tag{1.15}$$

y

$$\mathbf{k} = [k(x_1, x_*), k(x_2, x_*), \ldots, k(x_n, x_*)]^\top, \tag{1.16}$$

$$\mathbf{C} = \mathbf{K} + \sigma_\nu^2\mathbf{I}. \tag{1.17}$$

Procesos Gaussianos para Clasificación

Se pueden extender los GPR para resolver problemas de clasificación. En este entorno nos encontramos con muestras etiquetadas en un espacio finito, esta es máxima utilidad para problema que pensamos resolver, como estamos ante un problema binario, nos centraremos en que $b_i \in \{-1, 1\}$. Para GPC tenemos que cambiar la probabilidad del modelo en las observaciones, ya que ahora tenemos -1 o 1. Con lo que la función latente de x_i la podemos obtener utilizando la función $\Phi(\cdot)$:

$$p(b_i = 1|f(x_i), x_i) = \Phi(f(x_i)). \tag{1.18}$$

La función de respuesta, trata el valor de la función latente y lo lleva al intervalo (-1,1) que representa la probabilidad a posteriori para b_i.

Las integrales (1.11) y (1.12) ahora son intratables analíticamente. Con lo que deberemos resolver el problema con aproximaciones. Los dos métodos de aproximación son la aproximación de Laplace y *Expectation Propagation* (EP). La aproximación de Laplace se basa en obtener una serie de puntos, con ellos finalmente poder obtener la función de distribución, estos puntos son el punto máximo de la distribución y su curvatura en este punto. EP trata de obtener mayor información, para obtener la función de distribución en base al total de ella. Así pues EP tiene mayor carga computacional. Utilizando estas aproximaciones gaussianas en (1.8) nos permite obtener el resultado de (1.12) y finalmente poder resolver numéricamente la integral (1.11) que hasta el momento no era posible.

Funciones de Covarianza

En el planteamiento previo de GPC asumimos que $k(x, x')$ era conocida, sim embargo, para la mayoría de los problemas, la función de covarianza es desconocida, debemos obtenerla a partir de las muestras de entrenamiento. La función de covarianza describe la relación entre las entradas y las posibles soluciones que devolverá GPC. La función de covarianza debe recoger la información disponible sobre el problema en cuestión, por norma general se describe de forma paramétrica, como hiperparametros.

Si asumimos que los hiperparametros, θ, son desconocidos, la probabilidad a priori de los datos y la función latente son $p(\mathbf{b}|\mathbf{f}, \theta, \mathbf{X})$ y $p(\mathbf{f}|\mathbf{X}, \theta)$

Aplicando lo que ya hicimos anteriormente en el cálculo de la función latente nos encontramos que:

$$p(\mathbf{b}|\mathbf{X},\theta) = \int p(\mathbf{b}|\mathbf{f},\theta,\mathbf{X})p(\mathbf{f}|\mathbf{X},\theta)d\mathbf{f}. \tag{1.19}$$

Podemos utilizar una probabilidad a priori para el cálculo de los hiperparametros $p(\theta)$ de manera que nos sirva para calcular la probabilidad a posteriori. Se integraran para obtener las probabilidades a priori y posterior, aunque normalmente no serán analíticas lo que nos lleva a una integración por aproximación o por muestreo. Una alternativa sería maximizar la probabilidad obtenida en (1.19) que es utilizada para describir las propiedades de las muestras de test. Aunque este cambio, maximizar la probabilidad marginal, implica que no sea una solución puramente bayesiana. Pero incrementando las muestras de entrenamiento la probabilidad se convierte en una distribución.

Debido a la relación entre las variables las funciones de covarianza tienen que estar definidas de manera matricial, para que podamos utilizar las distribuciones gaussianas. Así pues una función de covarianza tendría esta forma:

$$k(x_i, x_j) = \alpha_1 \exp - \sum_{l=1}^{d} \gamma_l (x_{il} - x_{jl})^2 + \alpha_3 \delta_{ij} \tag{1.20}$$

donde $\theta = [\alpha_1, \gamma_1, \gamma_2, \ldots, \gamma_d, \alpha_3]$ son los hiperparametros. El primer término también denominado, termino de gauss o termino base, tiene una escala diferente según sea la dimensión del problema que estamos tratando [19].

1.7. Desarrollo

Para nosotros MATLAB es la herramienta fundamental, ya que nos permite cargar una hoja de cálculo, como es nuestra base de datos, para posteriormente trabajar con ella. Los trabajos que desempeña MATLAB, con el tratamiento y trabajo de los datos que poseemos, para transfórmalos y obtener definitivamente las 49 variables en las que se basa nuestro predictor se ocupará de obtener los datos de entrenamiento, entrenar el proceso, para posteriormente obtener los datos a predecir, sus variables y pasarlas por el proceso entrenado y que finalmente obtiene la estimación de la probabilidad, con esta podemos utilizar el Criterio de Kelly, por lo que MATLAB se encargará también de la gestión del *bank*, lo que como más adelante mostraremos nos permitirá aplicar diferentes estrategias de gestión del *bank*. Un esquema de nuestro desarrollo es el que podemos observar en la Figura 1.1 los módulos verdes son los que hemos desarrollado, los módulos blancos son entrada y salida de datos hacía y desde nuestros módulos.

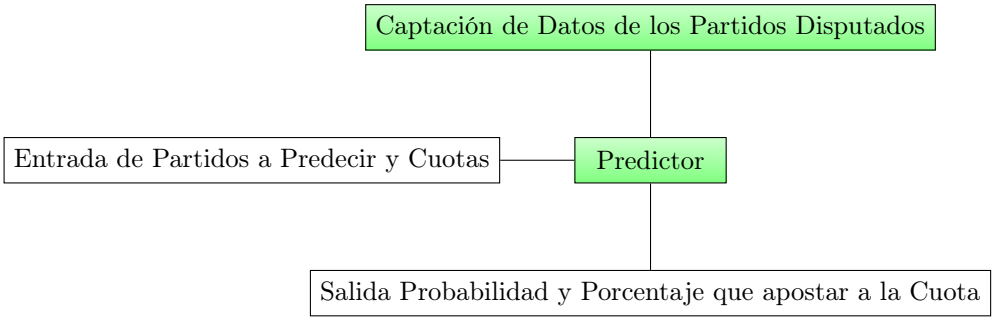


Figura 1.1: *Sistema Desarrollado*

En el ejemplo de la evolución de cuotas anteriores en ese partido tan igualado (50 %-50 %) Apostaríamos siempre que la cuota de uno de los jugadores fuera superior a 2. Podemos visualizar las cuotas que nuestro criterio se decantaría a apostar en la Tabla 1.2.

De esta manera apostaremos a las cuotas que están marcadas con el verde. La evolución en el tiempo no es un problema que trataremos, ya que habría que hacer un estudio, aunque lo habitual en nuestra experiencia en el mercado es cierta estabilización de cuotas durante las 12 últimas horas antes del partido en la mayoría de estos, claro está los casos de lesión y rumores se saldrían de esta media. Aun así las cuotas que vamos a utilizar para nuestras simulaciones serán las cuotas instantes antes del partido.

En la Tabla 1.2 tenemos marcadas en verde las cuotas a las que el Criterio de Kelly y como ya hemos comentado que las probabilidades son iguales para los dos tenistas, podemos visualizar en la columna de la derecha el porcentaje de nuestro *bank* que el criterio recomienda apostar.

Favorito	No Favorito	% del *Bank*
1,8	2,1	4.545 %
1,81	2,09	4.1284 %
1,82	2,07	3.271 %
1,83	2,06	2.8301 %
1,84	2,05	2.3809 %
1,85	2,03	1.4563 %
1,86	2,02	0.9803 %
1,87	2,01	0.495 %
1,88	1,99	
1,89	1,98	
1,9	1,98	
1,91	1,97	
1,92	1,96	
1,93	1,95	
1,94	1,94	

Tabla 1.2: *Cuotas Pre-Partido a Apostar*

Capítulo 2

VARIABLES Y ESTRATEGIAS

Ya hemos hablado un poco en el Capítulo 1 sobre la formación de la base de datos, cuales son los datos exactamente que posee, en este nuevo Capítulo vamos a tratar el tema de las variables, este es un tema que requiere ciertos conocimientos en el mundo del tenis, de su normativa, de los jugadores. Nosotros para solventarlo, aparte de nuestra pequeña experiencia en el mundo hemos optado por hablar con pronosticadores, estos pronosticadores son personas que llevan mucho tiempo viendo tenis, que conocen mucho el mundo, lo ven mucho y son capaces de prever un comportamiento entre 2 futuros jugadores por sus partidos previos. Ellos publican cada uno en su página web [12] [11] [6] [13], las apuestas que llevan, cada uno tiene un tipo de gestión del *bank* distinto.

Si crear una base de datos actualizada, completa y con información útil, es complicado. Pero no lo es menos que las variables que seleccionemos para nuestro predictor sean también acertadas, por ello la interacción con estos grandes pronosticadores se llevó a cabo en dos consultas independientes:

- En una primera ocasión se les preguntó de manera independiente y sin que conocieran las respuestas de los otros cuales serían los datos que para ellos son más interesantes en el mundo del tenis.

  Una vez que tuvimos está información fue cuando fuimos capaces de hacer la base de datos. Cuando la tuvimos completa, a cada uno de ellos se les pasó el resultado, sus comentarios fueron todos de gratitud, ya que hasta el momento no existe o al menos de forma gratuita, este tipo de base de datos, que sea tratable y fácilmente utilizable.

- Por lo que a continuación fuimos directamente a lo importante para nosotros, las variables, algunas de las variables que nos proponían eran directamente algunos datos de los capturados, cosa que nos facilitaría posteriormente las cosas, pero en otras ocasiones sus comentarios no eran datos directamente y para obtenerlas implementamos el tratado de datos, tanto a la hora de obtener las variables para los datos de entrenamiento, como para obtenerlas en el partido a pronosticar.

  Algo que es resaltable en todos ellos es que clasificaron las variables en tres tipos:

- Variables individuales del tenista
- Variables del partido y/o torneo
- Variables que relacionan a los dos tenistas

2.1. Variables individuales del tenista

Este tipo de variables son las que describen al tenista, entre todos los tipos de variables con las que vamos a trabajar las que clasificamos dentro de este tipo son las más numerosas, y es que el tenis al ser un deporte individual, se presta demasiado al estudio de la estadística, para poder caracterizar al tenista. Esto es lo que comentábamos anteriormente que son los datos que pueden mirar un tenista sobre su rival para enfrentarse a él, así en parte podrá estudiar como contrarrestar los puntos fuertes que tenga su rival.

Estas variables comprenden desde la variable 1 hasta la variable 19 para el primer jugador y para el segundo de ellos el numero será igual, para él estarán comprendidas entre la 20 y la 38. Este será el vector $\mathbf{x}$ de entrada.

Ahora vamos a pasar a analizar cada una de las variables.

Ranking: serán las variables 1 y 20 del vector, la variable 1 será para el jugador con mejor clasificación en él, y la variable 20 para el otro jugador. Ya comentamos anteriormente que este ranking indica la posición en el ATP TOUR, pero su cálculo es complejo y causa cierto desconcierto entre los jugadores. Su cálculo está basado en las distintas categorías de torneos, ya que dependiendo de su categoría se obtendrán una serie de puntos en función de la ronda en la que el tenista sea eliminado, pero la consecución de esta serie de puntos no significa sistemáticamente la suma de ellos a nuestro ranking, sino que antes de suma se descontará la cantidad de puntos obtenidos en el año anterior. Veamos esto con un ejemplo

En la última edición de Wimbledon 2011, antes de comenzar Rafa Nadal tenía 12.070 puntos, ostentando el primer puesto del ranking, mientras que Novak Djokovic, tenía 12.005 puntos, ostentando el segundo puesto. Ambos jugadores llegarón a la final. El campeón fue Novak Djokovic y el subcampeón Rafa Nadal. Al ser el ganador Novak Djokovic obtuvo 2.000 puntos correspondientes al campeón de un *Grand Slam*, y Rafa Nadal obtuvo 1.200, pero como ya hemos comentado no se sumaron directamente a los puntos que ya tenían, sino que descontarán los del año anterior. En el caso de Novak Djokovic el año anterior hizo semifinal, con lo que ganó 720 puntos, así debe sumar $2000-720 = 1280$ puntos, su ranking después de Wimbledon es de 13.285 puntos. Mientras que para Rafa Nadal que fue campeón el año anterior, con lo que obtuvo los 2.000 puntos, deberá sumar $1200 - 2000 = -800$ puntos, su ranking después de Wimbledon es de 11.270 puntos. Como podemos ver este sistema de puntuación es muy poco respetuoso con cambiar la asistencia a los torneos o con posibles lesiones, ya que para un tenista que un año ganase Roland Garros y Wimbledon, si la temporada siguiente sufriese una lesión de gravedad justo antes de Roland Garros y le tuviese apartado de las pistas, sufriría una gran pérdida de ranking.

Porcentaje promedio de puntos ganados con el servicio largo plazo: serán las variables 2 y 21 del vector, ordenados de manera idéntica a la primera variable y que será una constante en las 38 primeras variables. Esta variable es obtenida promediando el porcentaje de los puntos que obtiene el jugador de manera histórica. Así pues para el proceso de entrenamiento, en la creación de la matriz, está la obtendremos partido a partido, de manera que para entrenar el partido N, tomaremos este dato en cuestión en todos y cada uno de los partidos que esté involucrado este jugador hasta el partido $N-1$. En la predicción la medida se tomará igual. Esta medida la obtenemos directamente de la base de datos que hemos creado nosotros, pero no es una medida que se pueda encontrar, ya que la encontramos de manera más estructurada, y está basada en el porcentaje de primeros servicios puestos en juego (A), el porcentaje de puntos con el primer servicio (B) y el porcentaje de puntos con el segundo servicio (C). Finalmente calculamos esta medida $A \times B + (1 - A) \times C$. El servicio es un factor clave y son varios los jugadores que técnicamente no son grandes jugadores, pero tienen un servicio muy potente o certero y que aunque no ganan muchos partidos si que suponen que el ganarles sea un trabajo difícil. Vamos a observar en la Figura 2.1 y en la Figura 2.2 los valores para ambas variables.

Nota: *los valores de 0 y 1 son valores fuera de rango que se dan en caso de retida o que no se presente un jugador. Esta nota es valida para las Figuras 2.1 2.2 2.3 2.4 2.5 2.6 2.11 2.12*

2.13 2.14 2.15 2.16 2.17 2.18 2.19 2.20 2.21 2.22

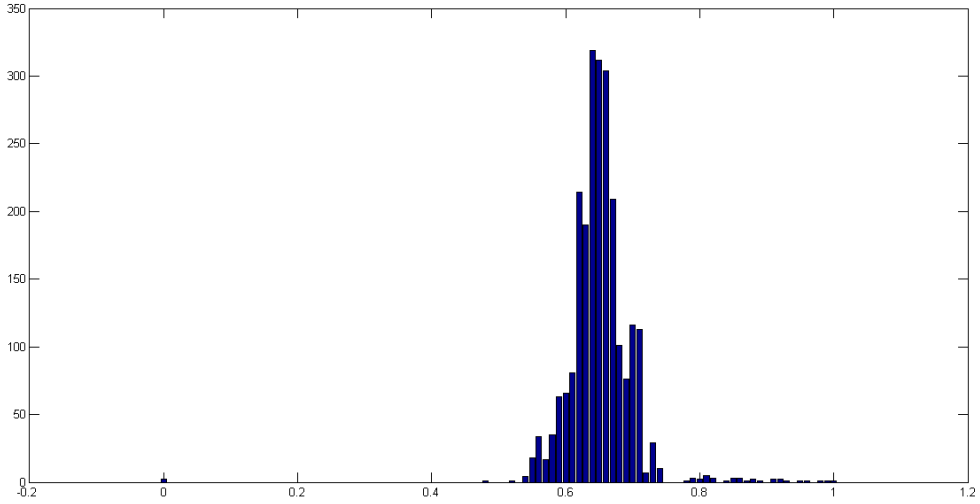


Figura 2.1: *Porcentaje promedio de puntos ganados con el servicio largo plazo para el jugador de mejor Ranking*

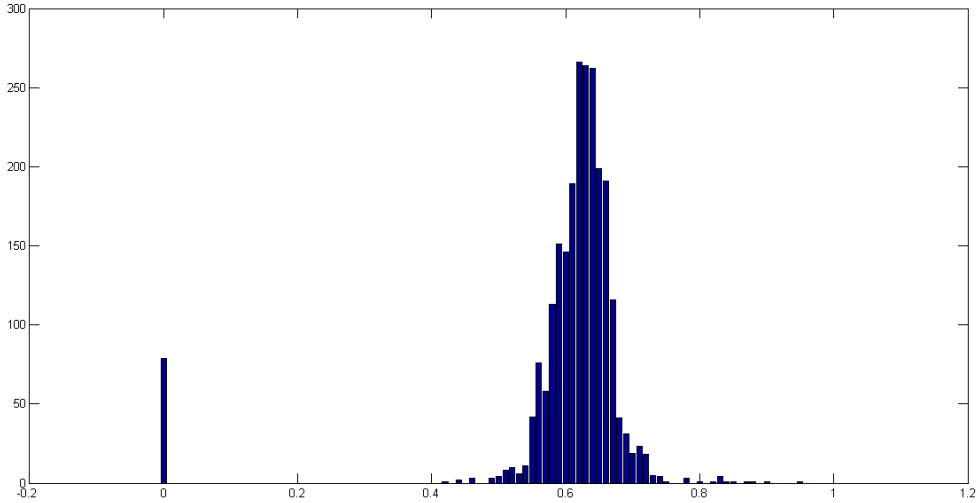


Figura 2.2: *Porcentaje promedio de puntos ganados con el servicio largo plazo para el jugador de peor Ranking*

Porcentaje promedio de puntos ganados con el resto largo plazo: serán las variables 3 y 22 del vector. Un buen restador se siente más tranquilo en el momento de recibir los saques, ya que sabe que puede presionar al rival, en el tenis es importante el resto y más en el tenis masculino que es muy dificil romper el saque. El objetivo de cualquier jugador es tratar de romper el saque al rival para ganar el partido. En este caso la obtención es directa de la base de datos y es un dato que tambien podemos obtener de manera más estructurada, sin embargo es más facil encontrar de manera directa que el anterior, y así lo hicimos. Al igual que antes en las figuras 2.3 y 2.4 vamos a observar este comportamiento.

Partidos acumulados: serán las variables 4 y 23 del vector. En este caso basándonos en el dato de la fecha de disputa del partido tratamos de buscar cuantos partidos ha disputado el jugador en un tiempo definido, para nuestro caso hemos utilizado 30 días, pero este tiempo se podría variar. Esta medida puede indicar dos cosas, que el jugador haya estado lesionado y que este valor sea pequeño, incluso cero, o que el jugador haya disputado muchos partidos y el valor sea alto, cosa que podría indicar un gran cansancio.

Días desde el último partido: serán las variables 5 y 24 del vector. Debido a que la variable anterior requiere poner un límite, en este caso únicamente buscaremos el tema de lesiones por lo que buscaremos el tiempo de inactividad en el ATP TOUR, también es una variable obtenida a partir de las fechas y que calculamos nosotros.

Si el jugador ganó el último torneo disputado: serán las variables 6 y 25 del vector. Este es un dato binario, de manera que nos indicará 1 que el jugador fue el campeón de un torneo que se disputó la semana anterior al partido que se vaya a jugar y -1 en el caso que no lo hubiese ganado. Es un dato que en principio no dice mucho debido a que el ATP TOUR son 65 torneos, y la mayoría de estos torneos solo los ganan los grandes jugadores. Pero la toma de este dato es debido a que en algunos casos el jugador que haya ganado un torneo podría tomarse un poco a risa, no tener la motivación suficiente para el torneo que va a disputar o incluso cansancio. Un caso de este tipo lo encontramos en la edición de Roland Garros 2011, el jugador español Nicolás Almagro (12 del Ranking en ese momento), disputó la semana antes el torneo de Niza, un torneo de categoría ATP 250, lo acabó ganando, pero al llegar a Roland Garros, y pese a que Nicolás Almagro es un especialista en la superficie del torneo parisino, cayó eliminado ante Lukas Kubot (122 del Ranking).

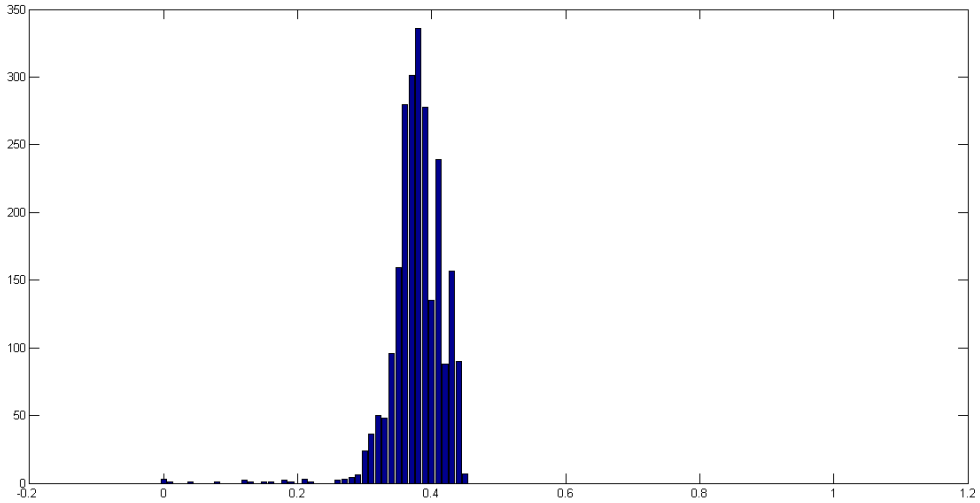


Figura 2.3: *Porcentaje promedio de puntos ganados con el resto largo plazo para el jugador de mejor Ranking*

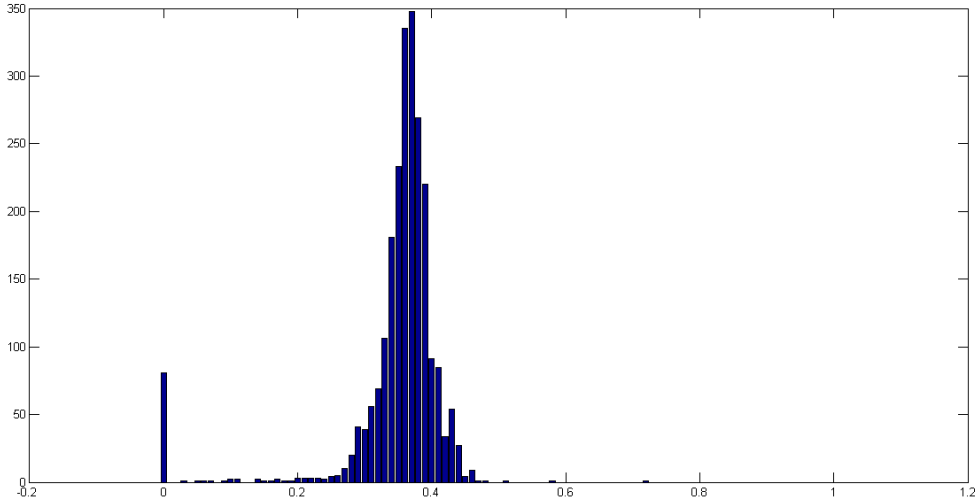


Figura 2.4: *Porcentaje promedio de puntos ganados con el resto largo plazo para el jugador de peor Ranking*

Número de puntos jugados el día anterior: serán las variables 7 y 26 del vector. Este dato viene a reflejar el cansancio reciente del jugador, aunque en los torneos grandes no se suele

dar mucho, que el tenista tenga que jugar muchos partidos en días consecutivos, esto si que ocurre en otros torneos de menor nivel, más si las condiciones meteorológicas no son buenas. Podríamos mirar el número de juegos, pero realmente ese número no es tan característico como el que proponemos debido a que esta variable mide cada vez que un jugador realiza un servicio. En cambio un juego se mide por puntos, pero no tienen todos por que tener el mismo número de puntos, ya que un juego como mínimo tiene 4 puntos, y no tiene número máximo. Así podemos en un set tener un resultado de 6-0 con 24 puntos disputados, o el mismo set 6-0 con 50 ó 60 puntos disputados. Y el cansancio es distinto.

Racha de partidos: serán las variables 8 y 27 del vector. Este dato viene reflejado con un valor positivo y contabiliza únicamente las victorias obtenidas por el jugador en un número de partidos definidos. Para nosotros este dato será 30 partidos, con este dato cogemos entre 3 y 4 meses de la mayoría de jugadores, aunque si el jugador es más mediocre cogeremos más tiempo, eso significará que no llega muy lejos en los torneos. Así pues este año, 2011, en algunos momentos de la temporada el tenista serbio Novak Djokovic ha sido el único que ha tenido este valor al máximo. Este cálculo se realiza a partir de los datos de ganadores y perdedores; y las fechas de disputa de partidos.

Racha de partidos en la superficie: serán las variables 9 y 28 del vector. Este dato es similar al anterior y contabiliza las victorias obtenidas por el jugador en la superficie que se dispute el partido que se va a entrenar o predecir. Para nosotros este dato será de 15 partidos, en esta ocasión no buscamos una trayectoria tan larga, ya que si pusiésemos un dato mayor por la topología del programa y el número de partidos en la superficie que normalmente se disputan, contabilizaríamos los del año anterior, cosa que para los primeros partidos que se disputan en la superficie está bien, pero no para cuando se han disputado un numero razonable de partidos. Unos quince partidos en la superficie para un jugador normal pueden ser 3 ó 4 torneos, dependiendo de las rondas que jueguen. Por la confección del calendario del ATP TOUR la variable anterior y esta tiene una alta colinealidad, debido a que lo normal es que los jugadores en la primera parte del calendario jueguen en pista dura, ya que el sexto torneo del año es el Open de Australia, a continuación se alternen pista dura y tierra batida hasta la disputa del torneo vigésimo en Miami (ATP 1000), luego comiencen a preparar exclusivamente Roland Garros el torneo trigésimo primero, a partir de aquí y en un mes jugar todos y cada uno de los partidos de hierba, en este caso y para la simulación de este tipo de superficie si que sería recomendable disminuir este valor, porque si no serían necesarias 2 temporadas para los mejores jugadores llegando incluso

hasta las 4 ó 5 para otros jugadores, hasta Wimbledon el torneo trigésimo sexto, para volver a la alternancia a los partidos de tierra y pista dura, estos últimos ya los más importantes ya en suelo de Norteamérica y Canadá para preparar el Open de Estados Unidos.

Porcentaje promedio de primeros servicios puestos en juego: serán las variables 10 y 29 del vector. En este dato lo que tratamos de observar es la eficiencia y seguridad del jugador con su saque, si a un valor alto en este punto añadimos una gran velocidad será difícil romper el saque. Esta medida es un dato que obtenemos directamente de nuestra base de datos, y lo único que hacemos es promediar para un número de partidos definidos, en nuestro caso será de 15 partidos, como hemos dicho anteriormente esto tomará la eficiencia y seguridad del jugador aproximadamente 3 torneos. Este dato unido con otros como puede ser la superficie pueden hacer que la victoria esté cerca, no es lo mismo meter un gran porcentaje de primeros servicios en hierba, donde la bola desliza y es más difícil restar; que hacerlo en tierra, donde la bola bota más y se frena un poco. Veamos la distribución en las figuras 2.5 y 2.6.

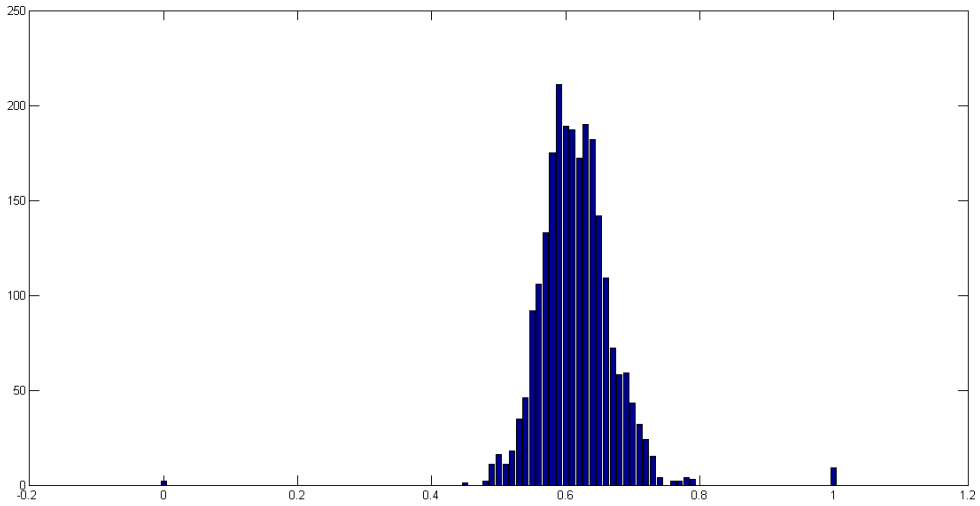


Figura 2.5: *Porcentaje promedio de primeros servicios puestos en juego por el jugador de mejor Ranking*

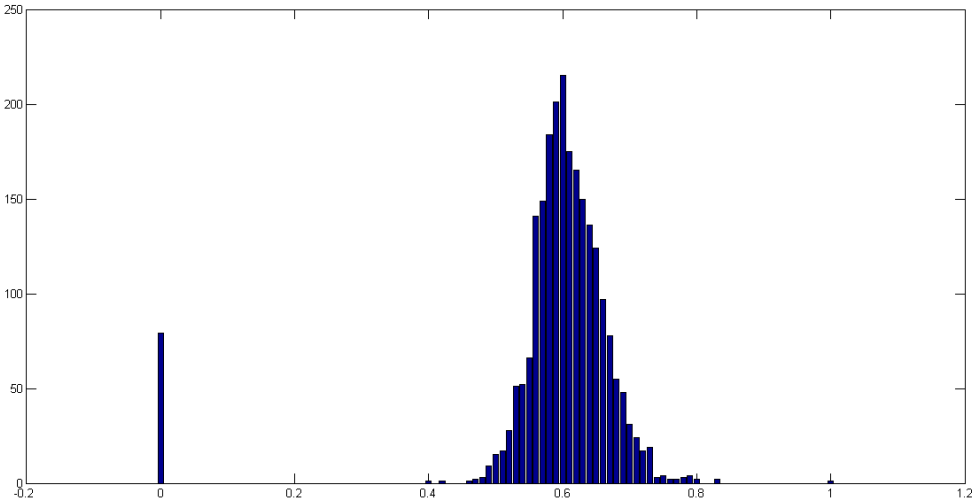


Figura 2.6: *Porcentaje promedio de primeros servicios puestos en juego por el jugador de peor Ranking*

Promedio de Aces: serán las variables 11 y 30 del vector. Los aces son los saques que el rival no logra golpear, por lo que tienen una mezcla de colocación y potencia. Este promedio se calcula sobre un total de 15 partidos. Aunque por las observaciones realizadas es un valor con una alta varianza. Y un jugador que en un partido puede hacer 16 aces en otro solamente hacer 4, ya que también influye el rival. Veamos la distribución en las figuras 2.7 y 2.8.

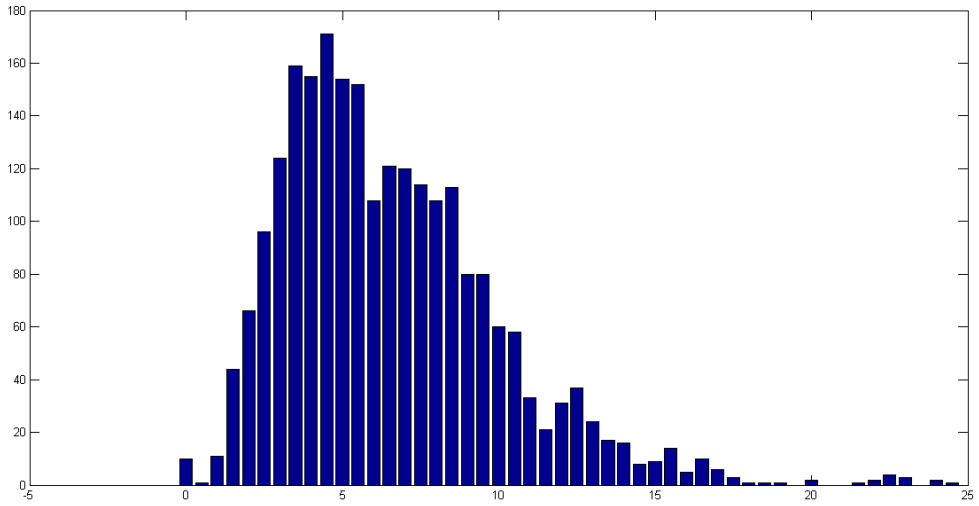


Figura 2.7: *Promedio de Aces del jugador de mejor Ranking*

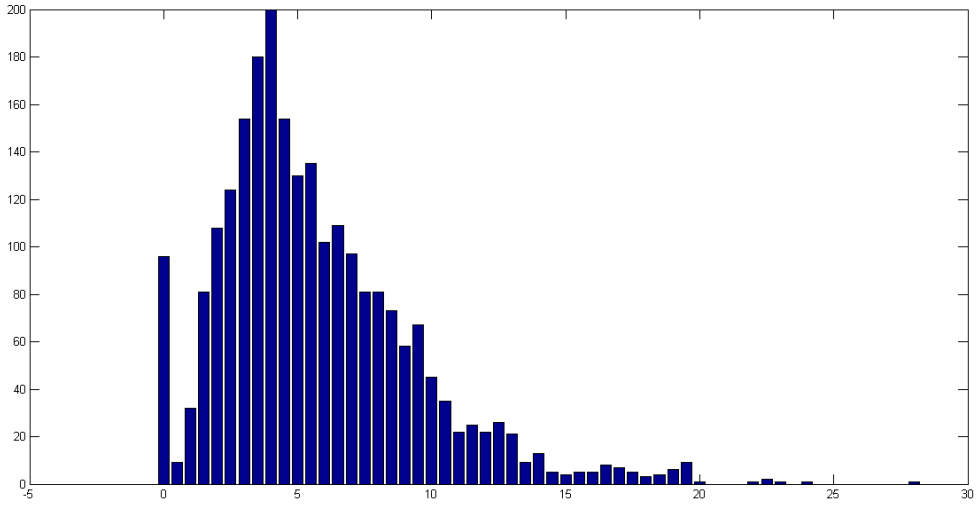


Figura 2.8: *Promedio de Aces del jugador de peor Ranking*

Promedio de Dobles Faltas: serán las variables 12 y 31 del vector. Las dobles faltas ocurren cuando el jugador no es capaz de poner en juego su saque en ninguna de las dos ocasiones de las que dispone. Promediaremos sobre los 15 partidos. Un valor alto implica puntos fáciles para el rival, y estar más cerca del break, aunque como ocurre con los Aces, también tiene una varianza alta. Veamos la distribución en las figuras 2.9 y 2.10.

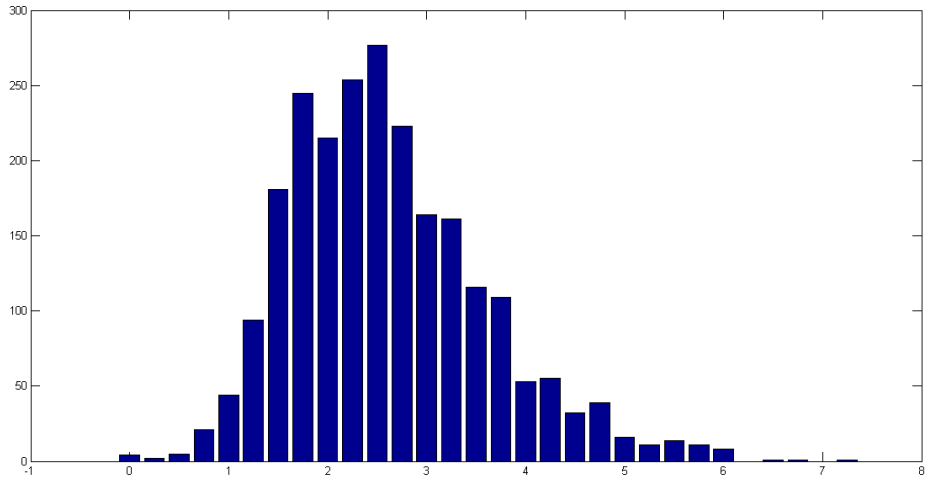


Figura 2.9: *Promedio de Dobles Faltas del jugador de mejor Ranking*

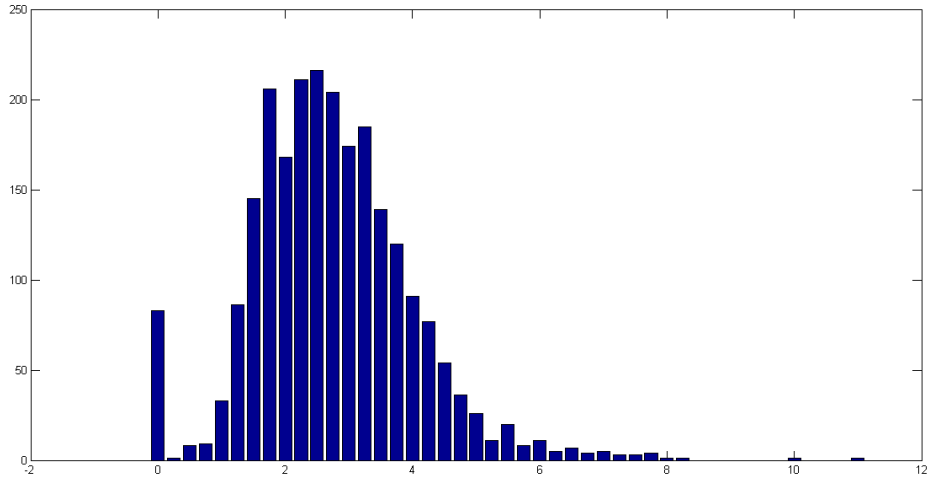


Figura 2.10: *Promedio de Dobles Faltas del jugador de peor Ranking*

Porcentaje promedio de puntos con el primer servicio: serán las variables 13 y 32 del vector. En este caso se indicará el porcentaje de los primeros saque que son punto para el jugador que saca. Nosotros como las anteriores la tomaremos en función de 15 partidos, para el ganador del partido este dato suele ser bastante alto. Veamos la distribución en las figuras 2.11 y 2.12.

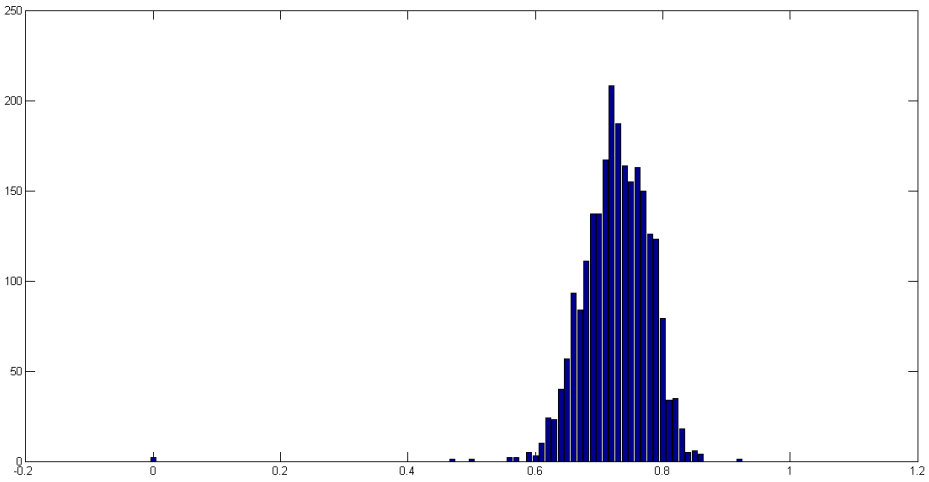


Figura 2.11: *Porcentaje promedio de puntos con el primer servicio del jugador de mejor Ranking*

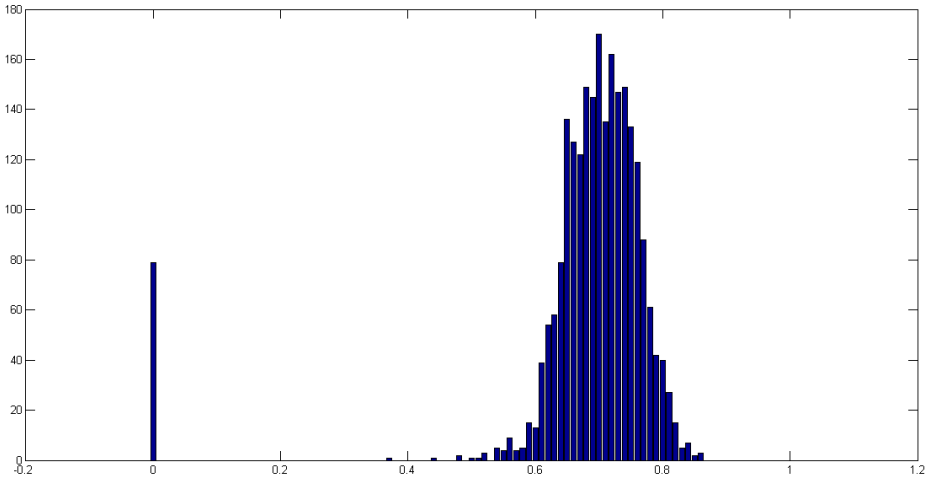


Figura 2.12: *Porcentaje promedio de puntos con el primer servicio del jugador de peor Ranking*

Porcentaje promedio de puntos con el segundo servicio: serán las variables 14 y 33 del vector. Es un caso muy similar que el anterior y procederemos con el mismo número de partidos, para el jugador este valor tiene que ser más importante ya que con este saque debido a su intención de no producir la doble falta reduce la fuerza y aumenta la precisión, pero esto da más ventaja al rival de golpear más fuerte. Veamos la distribución en las figuras 2.13 y 2.14.

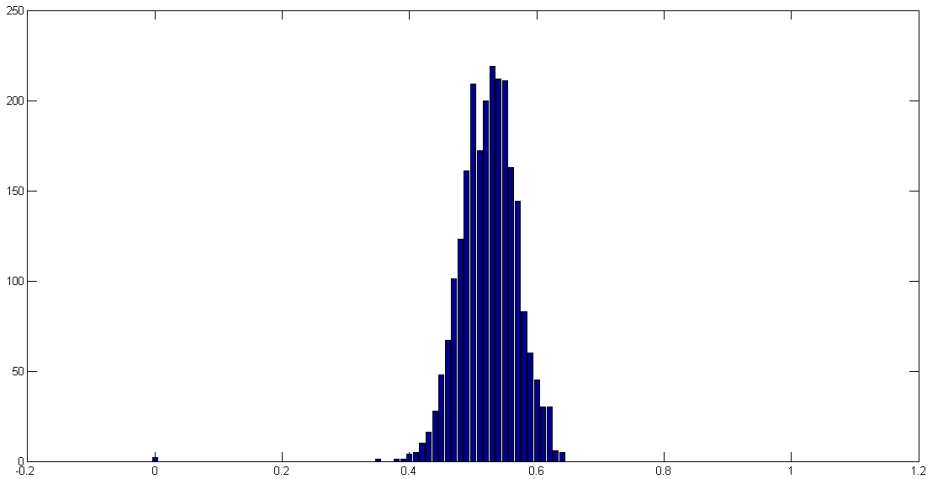


Figura 2.13: *Porcentaje promedio de puntos con el segundo servicio del jugador de mejor Ranking*

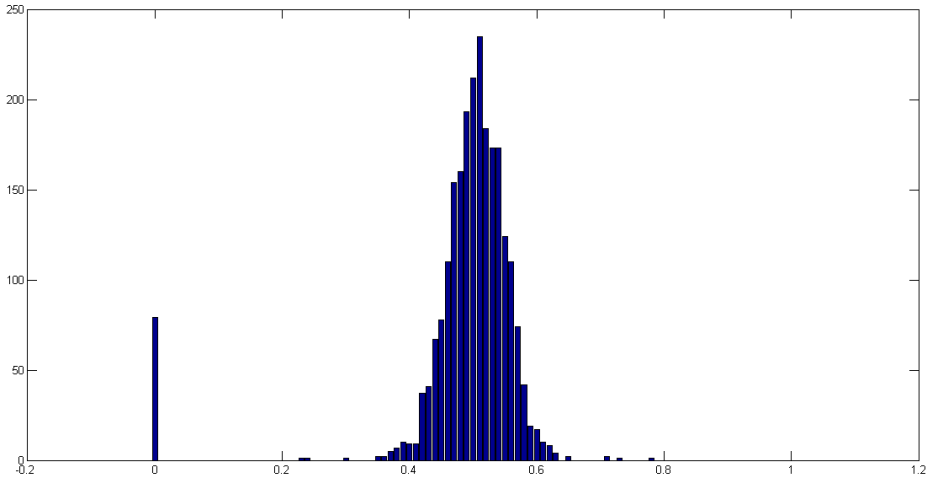


Figura 2.14: *Porcentaje promedio de puntos con el segundo servicio del jugador de peor Ranking*

Porcentaje promedio de puntos ganados al resto del primer servicio: serán las variables 15 y 34 del vector. Esta variable no suele ser alta, en comparación con las anteriores y si lo fuera sería indicativo de que el jugador es muy restador. También procederemos con el mismo número de partidos. Veamos la distribución en las figuras 2.15 y 2.16.

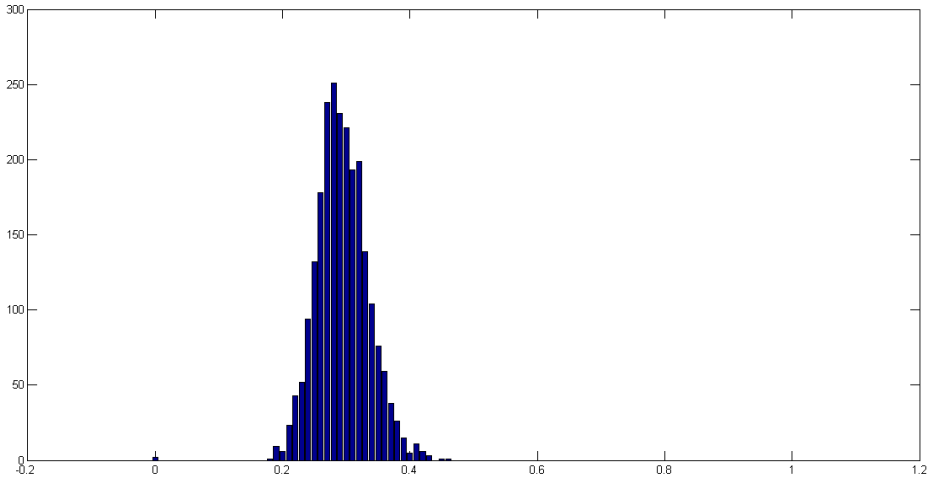


Figura 2.15: *Porcentaje promedio de puntos ganados al resto del primer servicio del jugador de mejor Ranking*

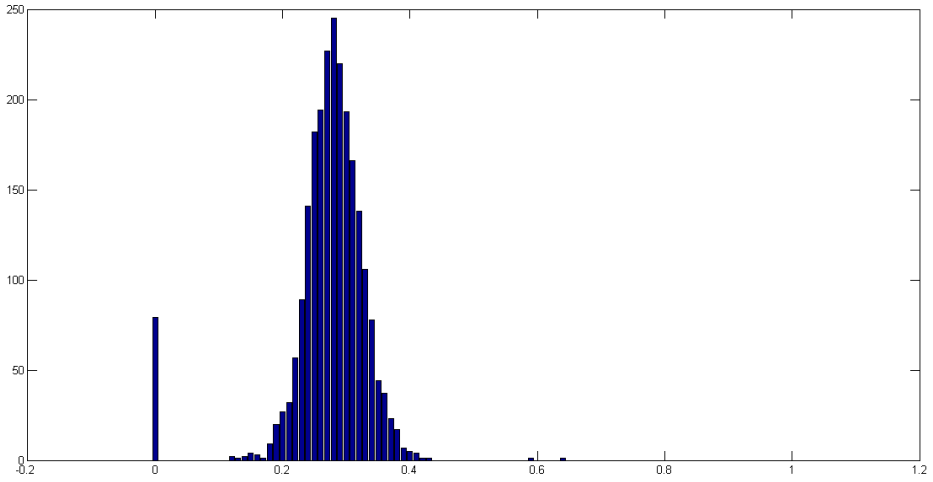


Figura 2.16: *Porcentaje promedio de puntos ganados al resto del primer servicio del jugador de peor Ranking*

Porcentaje promedio de puntos ganados al resto del segundo servicio: serán las variables 16 y 35 del vector. Esta normalmente suele ser más alta que la anterior por lo que comentábamos en el saque con el segundo servicio. También tomaremos el promedio sobre 15 partidos. Veamos la distribución en las figuras 2.17 y 2.18.

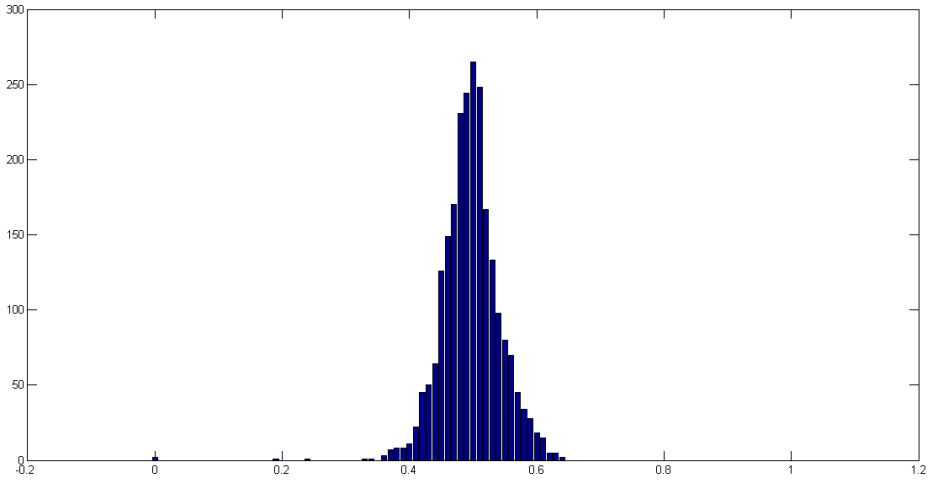


Figura 2.17: *Porcentaje promedio de puntos ganados al resto del segundo servicio del jugador de mejor Ranking*

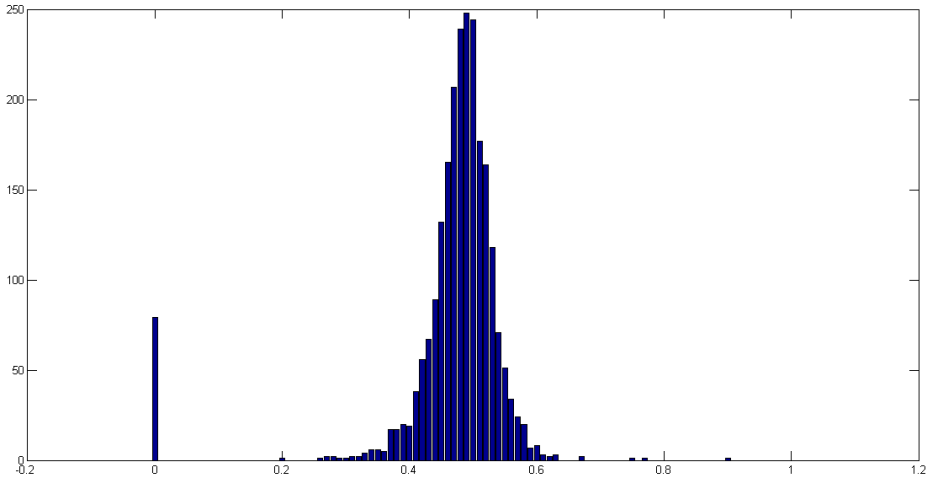


Figura 2.18: *Porcentaje promedio de puntos ganados al resto del segundo servicio del jugador de peor Ranking*

Porcentaje promedio de puntos ganados con el servicio corto plazo: serán las variables 17 y 36 del vector. Es el mismo dato que utilizábamos para obtener las variables 2 y 21, con la diferencia que antes buscábamos un comportamiento a largo plazo del jugador, por así decirlo de temporadas, y en este caso buscamos un rendimiento de los últimos torneos, ya que al igual que en las anteriores variables utilizaremos un valor de 15 partidos. Veamos la distribución en las figuras 2.19 y 2.20.

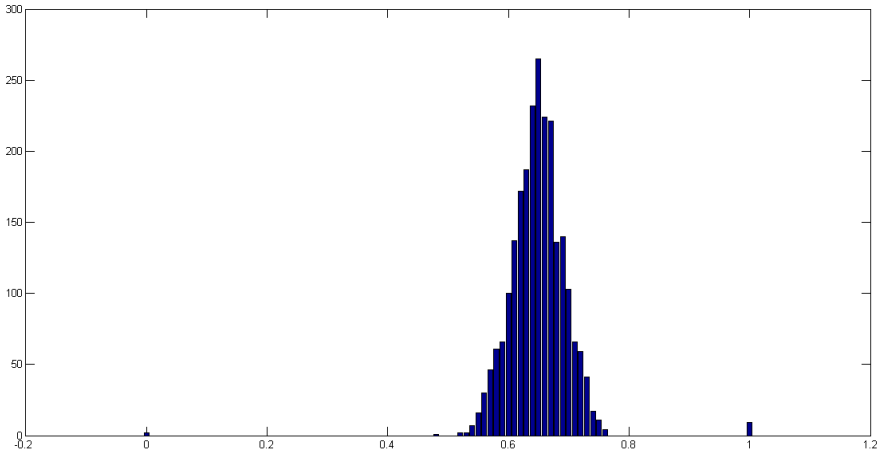


Figura 2.19: *Porcentaje promedio puntos ganados con el servicio corto plazo del jugador de mejor Ranking*

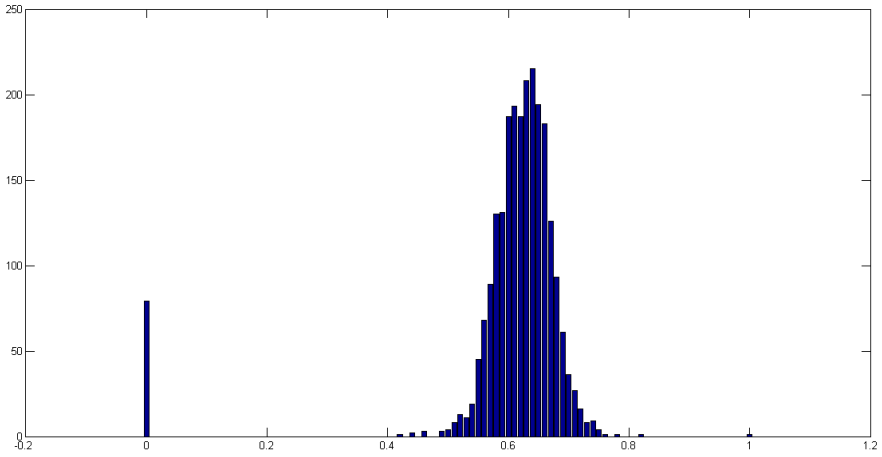


Figura 2.20: *Porcentaje promedio puntos ganados con el servicio corto plazo del jugador de peor Ranking*

Porcentaje promedio de puntos ganados con el resto corto plazo: serán las variables 18 y 37 del vector. Al igual que las anteriores 2 variables, utilizamos los mismos datos que en las variables 3 y 22, pero sobre un tiempo de 15 partidos. Veamos la distribución en las figuras 2.21 y 2.22.

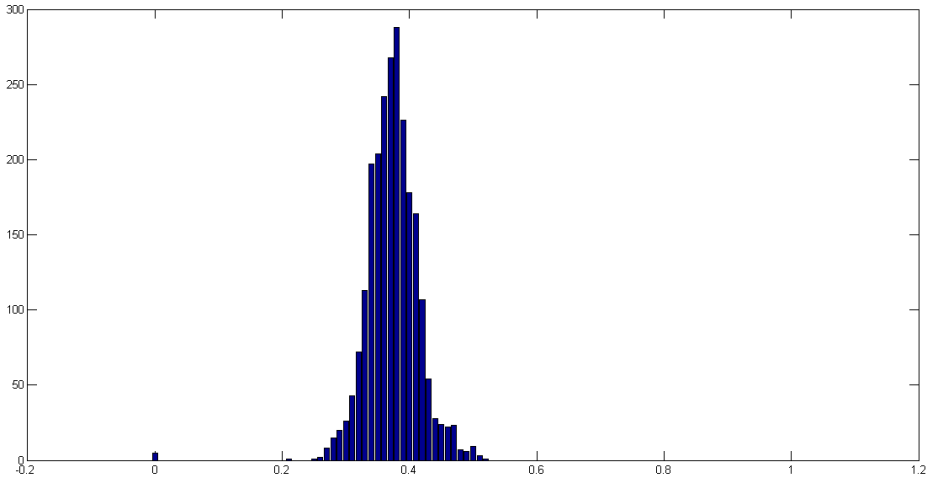


Figura 2.21: *Porcentaje promedio puntos ganados con el resto corto plazo del jugador de mejor Ranking*

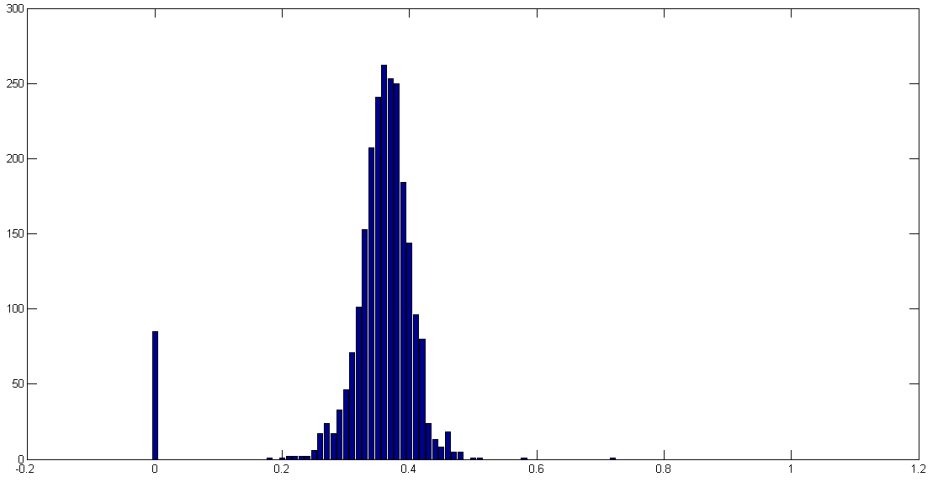


Figura 2.22: *Porcentaje promedio puntos ganados con el resto corto plazo del jugador de peor Ranking*

Porcentaje promedio de roturas de servicio convertidas: serán las variables 19 y 38 del vector. En este caso también utilizaremos el promedio de los últimos 15 partidos. Pero al igual que sucedería con los Aces y las Dobles faltas esta medida tiene una excesiva varianza. Veamos la distribución en las figuras 2.23 y 2.24.

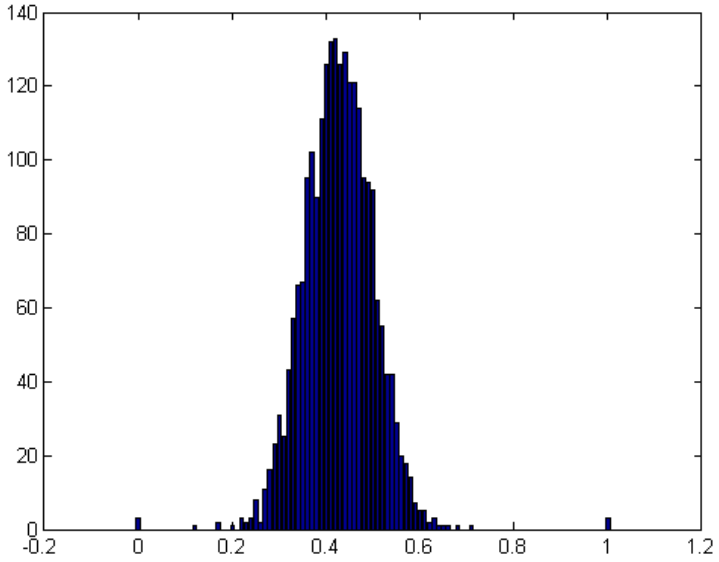


Figura 2.23: *Porcentaje promedio de roturas de servicio convertidas del jugador de mejor Ranking*

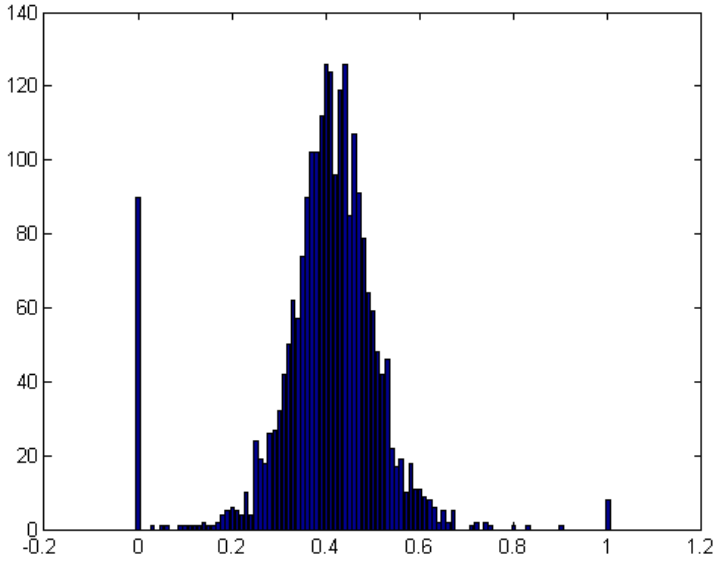


Figura 2.24: *Porcentaje promedio de roturas de servicio convertidas del jugador de peor Ranking*

2.2. Variables del torneo y/o partido

En este apartado trataremos la varibles que definen el partido que se disputa entre los dos tenistas, y que son totalmente ajenos a ellos.

Tipo de pista: será la variable 39 del vector. En el tipo de variable será de valor binario, 1 pista descubierta y -1 pista cubierta. La mayoría de los partidos son del primer tipo, pero en ciertas ocasiones debido a las condiciones meteorológicas, por ejemplo el frio, ya que recordemos que el último torneo de la temporada es en Noviembre, la Copa de Maestros en Londres, y es cubierta. También nos encontramos con la lluvia y es que antiguamente se podía suspender los partidos el tiempo que fuera necesario, pero actualmente existen muchos intereses económicos detrás, que el retraso del partido no beneficia. La pista descubierta, en el juego hace que el viento no afecte, cosa que puede beneficiar a un jugador y perjudicar al otro.

Superficie: será la variable 40 del vector. Este tipo de variable será de valor ternario, 1 para superficie dura, 2 para tierra batida y 3 para hierba. En la actualidad la mayoría de los partidos se desarrollan en las dos primeras superficies y es donde los jugadores hacen la mayor parte de sus puntos. Pero en el circuito hay auténticos especialistas de hierba que en esta superficie pueden ganar a un top. Además la superficie condiciona la mayoría de los parámetros de los jugadores que hemos tratado en el apartado 2.1.

Tipo de Torneo: será la variable 41 del vector. Este tipo de variable tiene 5 posibilidades y su codificación será los puntos que otorgan al ganador, ordenados de mayor a menor importancia tenemos:

- 41 torneos ATP 250
- 11 torneos ATP 500
- 9 torneos ATP 1000
- La Copa de Maestros
- 4 *Grand Slam*

Este tipo de variable es útil para la moral de jugador, hay torneos que los jugadores top no quieren disputar y simplemente van como entrenamiento o como se dice en el mundo van a "recoger el cheque".

Ronda: será la variable 42 del vector. Este tipo de variable tiene 7 posibilidades, aunque no en todos los torneos puede tener todos los valores, esto es así porque los cuadros tienen distintos tamaños, los *Grand Slam* tienen las 7 rondas completas pero el resto no, el más pequeño tiene 5 rondas. Para nosotros todos los torneos tienen 1ª y 2ª ronda, codificados como 1 y 2; además de Cuartos de Final, Semifinal y Final, codificados como 5, 6 y 7 respectivamente. Esta variable tiene sentido junto con el torneo y la defensa de puntos con la motivación de los jugadores.

Número máximo de sets: será la variable 43 del vector. Este tipo tiene dos posibilidades, que el partido sea de 3 sets o que sea de 5 sets, aunque la mayoría de los partidos son de 3 sets, los *Grand Slam* y los partidos de Copa Davis aún siguen siendo a 5 sets. Este dato es importante, ya que para un jugador no favorito dar la sorpresa ante un jugador top es más difícil si tiene que ganarle 3 sets al favorito.

La defensa de puntos de los jugadores: serán las variables 44 y 45 del vector. Estas variables tienen sentido para la motivación y entrega del jugador, si un jugador tiene que defender puntos muy posiblemente se entregue más en el partido por la muy probable pérdida de puntos en el ranking. Esta variable es una variable que obtenemos en base a los datos fecha, torneo, y la ronda en la que fue eliminado el tenista el año anterior.

2.3. Variables que relacionan a los dos tenistas

Este tipo de variables vamos a tratar sobre todo los enfrentamiento que han disputado con anterioridad ambos jugadores, ya que es posible que un jugador se adapte perfectamente a como juega el rival pero este no sea capaz de cambiar su juego para que su rival no lo tenga tan fácil.

Partidos ganados por el jugador: serán las variables 46 y 47 del vector. Tendrá un valor a partir de 0 y será un sumatorio de todas las victorias que haya conseguido éste jugador frente a su rival. Un ejemplo sería, tras la disputa de Wimbledon 2011, de los 28 partidos disputados entre Rafa Nadal (nº2) y Novak Djokovic(nº1), el vector $\mathbf{x}$ de un partido que se disputase con estos jugadores, tendría por variable 46 el valor 12 y por variable 47 el valor 16.

Partidos ganados por el jugador sobre la superficie del partido: serán las variables 48 y 49. Tendrán los mismos rangos de valores que las dos anteriores. Es posible que en esta variable haya colinealidad con las anteriores, pero también es posible que no, ya que uno de los jugadores puede ser muy bueno en pista dura y tierra batida y el otro muy bueno en hierba.

2.4. Estrategias

Como ya comentamos en el Capítulo 1 nuestra estrategia va a estar basada en torno a las predicciones que obtengamos y al Criterio de Kelly.

Pero llegado el momento realizar una apuesta en cada partido con el total del porcentaje que obtenemos del Criterio de Kelly puede llegar a ser demasiado arriesgado. Aparte de ello que debido a la simultaneidad de partidos en la mayoría de los torneos no podríamos apostar las futuribles ganancias que produciría la primera apuesta, incluso darle una importancia excesiva a los primeros encuentros.

Otra estrategia vendría por parte de la casa de apuestas a elegir, apostar en casas de apuestas que tengan el margen de beneficio lo más bajo posible para ellas, en este apartado tenemos a casas como:

- PinnacleSports
- Betfair

Estas casas reparten en torno al 98 % de lo que reciben por apuestas.

$$\text{Porcentaje a repartir} = \frac{1}{\frac{1}{c1} + \frac{1}{c2}} \times 100. \tag{2.1}$$

donde c1 y c2 son las cuotas de los jugadores

Para calcular esto utilizaremos la ecuación (2.1),un ejemplo sería con c1=1.575 y c2=2.6

$$\frac{1}{\frac{1}{1{,}575} + \frac{1}{2{,}6}} \times 100 = 98{,}083\,\%$$

Debemos evitar a toda cosa las casas que su porcentaje sea menor, un ejemplo sería Bwin ó Betclic

- Para Bwin
$$\frac{1}{\frac{1}{1{,}4} + \frac{1}{2{,}75}} \times 100 = 92{,}771\,\%$$
- Para Betclic
$$\frac{1}{\frac{1}{1{,}5} + \frac{1}{2{,}40}} \times 100 = 92{,}308\,\%$$

Cuotas tomadas el día previo al partido a disputar el día 14 de Julio de 2011 entre Youznhy y Ferrero [10] [4] [5]

Una de las cosas que hemos podido aprender de los grandes pronosticadores, es que tener multitud de casas puede ayudar, así pues, en este caso hemos encontrado 2 casas distintas con cada una cuotas superiores a la otra de manera que en estos momentos el porcentaje que podría llegar a obtener sería menor, de (2.1) combinando la cuota de PinnacleSports y Bwin tendremos

$$\frac{1}{\frac{1}{1{,}575}+\frac{1}{2{,}75}} \times 100 = 100{,}14\,\%$$

Estamos ante una situación de arbitraje, podríamos ganar apostando en ambas casas, el único inconveniente que tendríamos sería que las reglas son distintas, debido a que una casa de apuestas en caso de retirada da como ganador al jugador que pasada de ronda, solo si, se ha jugado un set completo y otra desde el momento en que se ha jugado una bola de partido.

Para terminar vamos a contar la estrategia que hemos utilizado para las simulaciones, en nuestro caso, vamos a utilizar la división que realizan los torneos por rondas, y en cada una de ellas utilizaremos nuestro *bank*. Si en primera ronda se disputan 32 partidos por que tengamos un cuadro de 64 participantes, dividiremos nuestro *bank* actual entre 32. Lo normal es que antes de disputar los partidos de segunda ronda hayan podido acabar todos los de la primera, con lo que volveremos a dividir. En este método prioriza las rondas finales, algo que si el sistema no funcionase bien podría ser un problema, pero también es cierto que de los partidos de favoritos es de los que más datos hay y los que se pueden predecir mejor. Además nosotros hemos puesto un límite para que la apuesta máxima no sea superior a una cantidad.

Basándonos en nuestras herramientas básicas, como son la estimación de probabilidades y el Criterio de Kelly nos podemos encontrar con una gran variedad de estrategias y aunque nosotros nos hemos centrado en esta. Dividir el *bank* entre los partidos del día, dividir el *bank* entre cada uno de los partidos del torneo, apostar únicamente a los partidos en los que las herramientas nos digan que lo hagamos al favorito. Un estudio detallado de la gran diversidad de ellas sería, un gran trabajo de futuro.

Capítulo 3

RESULTADOS Y SIMULACIONES

Para nuestras simulaciones hemos utilizado la estrategía que comentabamos al final del Capitulo 2. Para nosotros tanto la cantidad con la que iniciamos la estrategía como el límite por apuesta es de 100€. Uno de los inconvenientes que nos podemos encontrar con esta estrategía es la posibilidad que el porcentaje a apostar sea tan pequeño que no superemos el límite. En nuestro caso utilizaremos la casa de apuestas Pinnacle, a no ser que se diga lo contrario; y su apuesta mínima es de 1€, aunque para nuestras simulaciones no tendremos en cuenta esta límitación. Si bien es cierto que una manera sencilla de suprimir este inconveniente sería aumentar el *bank* inicial. Además en nuestras simulaciones el entrenamiento se realizará con los partidos desde Roland Garros 2010 hasta el ultimo partido disputado que se va a predecir, en nuestro caso los Torneos que vamos a predecir pertenecen al nucleo central de la temporada de tierra:

1. Estoril
    - ATP 250
    - Del 25/4/2011 al 1/5/2011
    - 27 partidos simulados de 27 partidos
2. Munich
    - ATP 250
    - Del 24/4/2011 al 1/5/2011
    - 30 partidos simulados de 31 partidos

3. Belgrado
    - ATP 250
    - Del 25/4/2011 al 1/5/2011
    - 26 partidos simulados de 27 partidos
4. Madrid
    - ATP 1000
    - Del 1/5/2011 al 8/5/2011
    - 54 partidos simulados de 55 partidos
5. Roma
    - ATP 1000
    - Del 8/5/2011 al 15/5/2011
    - 53 partidos simulados de 55 partidos
6. Niza
    - ATP 250
    - Del 15/5/2011 al 21/5/2011
    - 27 partidos simulados de 27 partidos
7. Rolan Garros
    - *Grand Slam*
    - Del 22/5/2011 al 5/6/2011
    - 124 partidos simulados de 127 partidos

En total disponemos de la simulación de 341 partidos.

Una vez que disponemos de la estrategia llega el momento de la decisión de que variables incorporar al predictor, en este apartado lo primero es decir que al utilizar todas las variables, las decisiones eran iguales fuese quien fuesen jugadores implicados, encontrandonos con una situación de subajuste, así que debíamos decidir cuales iban a ser las variables. A pesar de que no realizamos todas y cada una de las combinaciones que se debería de hacer en una selección

de variables, debido a que es un proceso costoso. Si que realizamos una oposición de la mayor parte de las variables individuales de los jugadores.

Podemos ver la oposición de rankings en la Figura 3.1

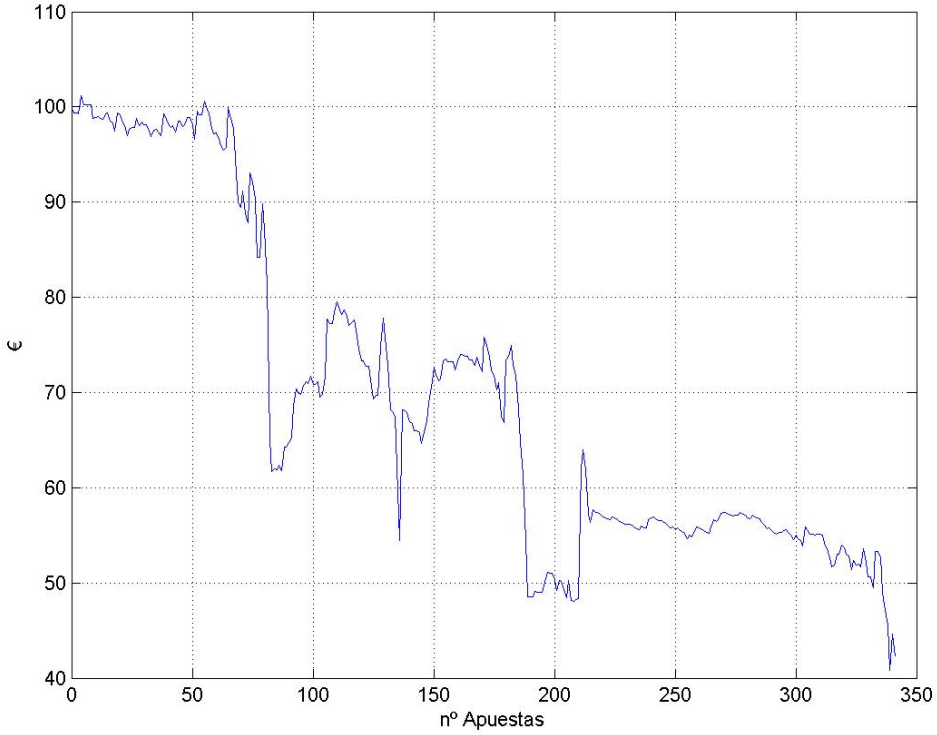


Figura 3.1: *Simulación oposición Ranking.*

Si analizamos las figuras 3.1, 3.2 y 3.3 podemos observar que en las primeras rondas de cada torneo obtenemos ciertas ganancias, y que cuando llegan las rondas finales, que es cuando según nuestra estrategia realizamos apuestas de mayor importe, entramos en perdidas, sería interesante mirar este aspecto en otro tipo de estrategias.

La oposición de porcentaje promedio puntos ganados con servicio largo plazo 3.2

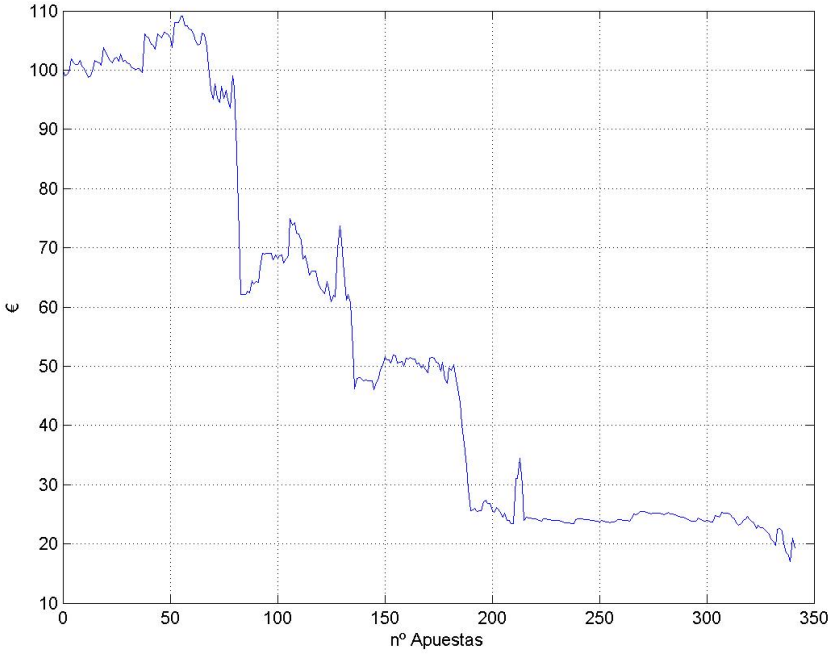


Figura 3.2: *Simulación oposición porcentaje promedio puntos ganados con servicio largo plazo.*

La oposición de porcentaje promedio puntos ganados con el resto largo plazo 3.3

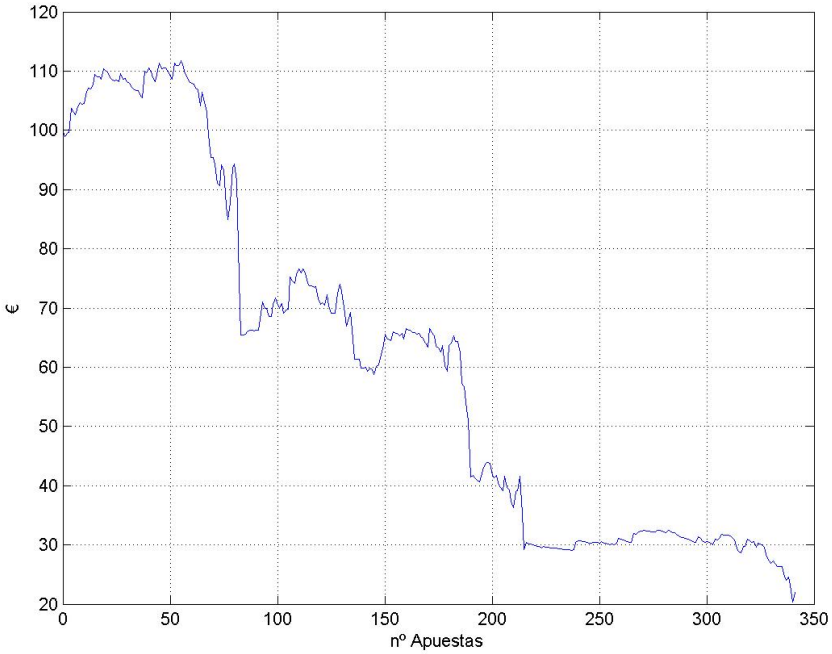


Figura 3.3: *Simulación oposición porcentaje promedio puntos ganados con el resto largo plazo.*

Las variables 4, 5, 6 y 7 y sus respectivas homologas del rival creemos que son variables para afinar los resultados y que en ningún caso por ellas mismas pueden dar un resultado coherente por lo que no se realizó su oposición.

La oposición de la racha de victorias 3.4

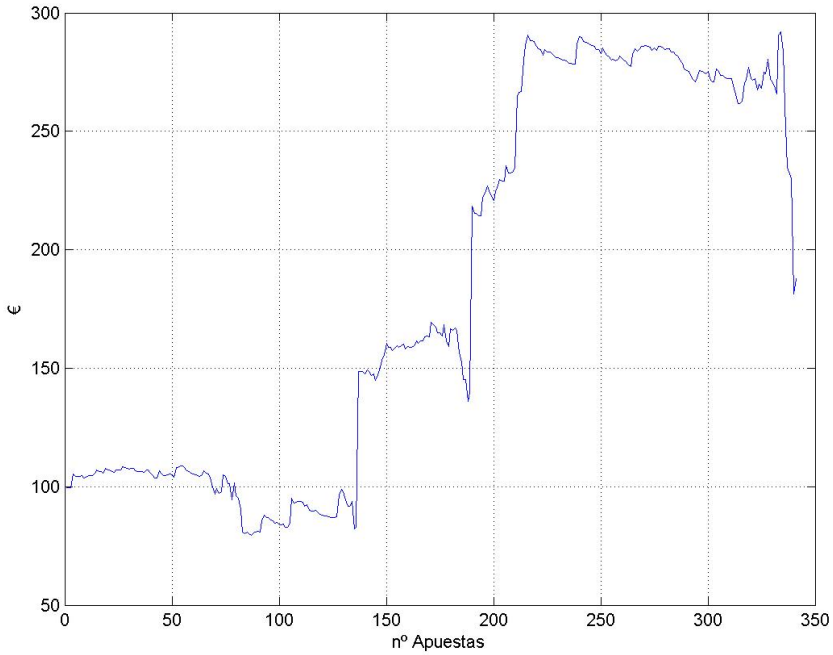


Figura 3.4: *Simulación oposición racha de victorias.*

Si analizamos la Figura 3.4 podemos observar como se producen ciertas ganancias en las primeras rondas, pero también se producen ganancias en rondas avanzadas.

La oposición de la racha de victorias en superficie 3.5

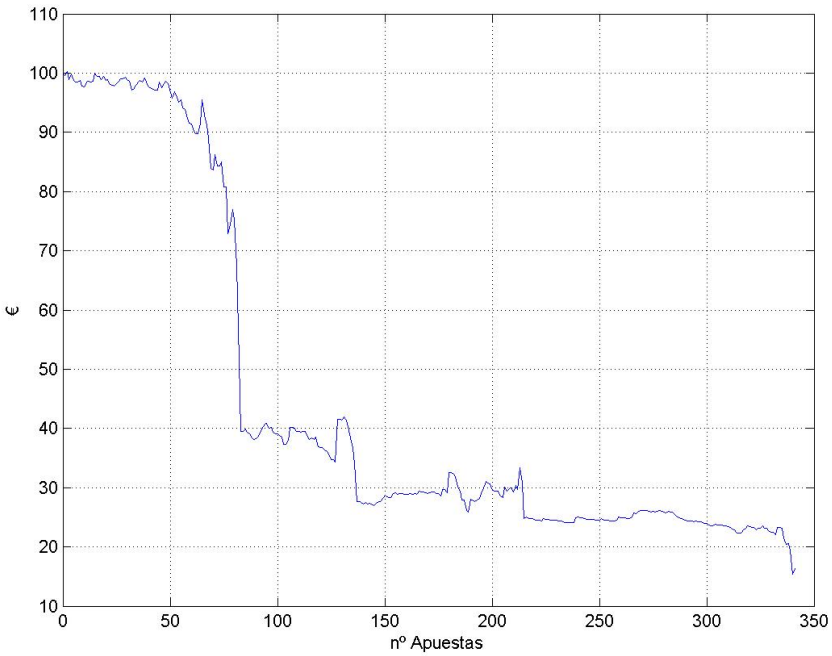


Figura 3.5: *Simulación oposición racha de victorias en superficie.*

Al analizar la Figura 3.5 y en contraposición a lo que podiamos pensar al analizar 3.4 vemos que su comportamiento es totalmente distinto, como ya comentamos en el Capitulo 2 al respecto de esta variable, se debería de realizar un estudio en la ventana de partidos a seleccionar.

La oposición porcentaje promedio saques puestos en juego 3.6

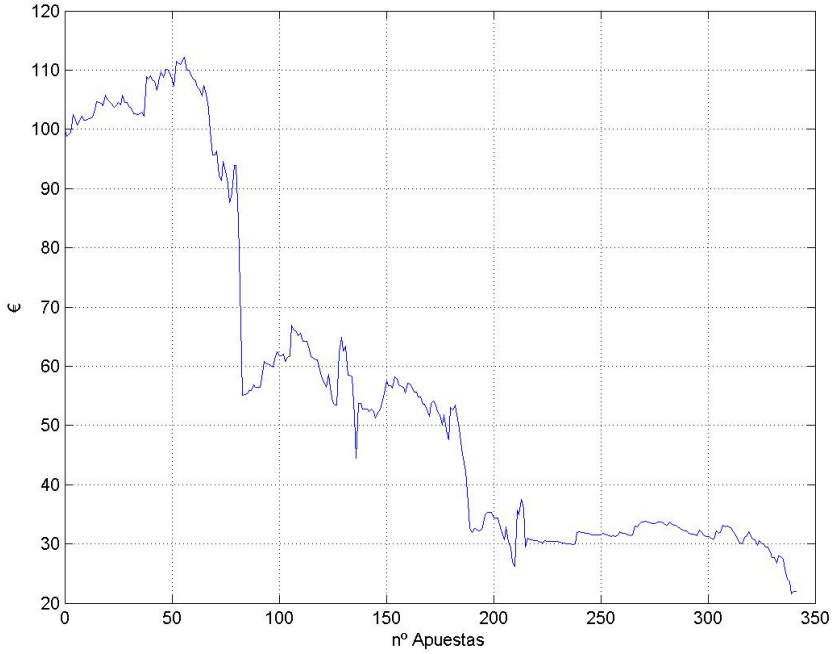


Figura 3.6: *Simulación oposición porcentaje promedio saques puestos en juego.*

La oposición promedio de Aces 3.7

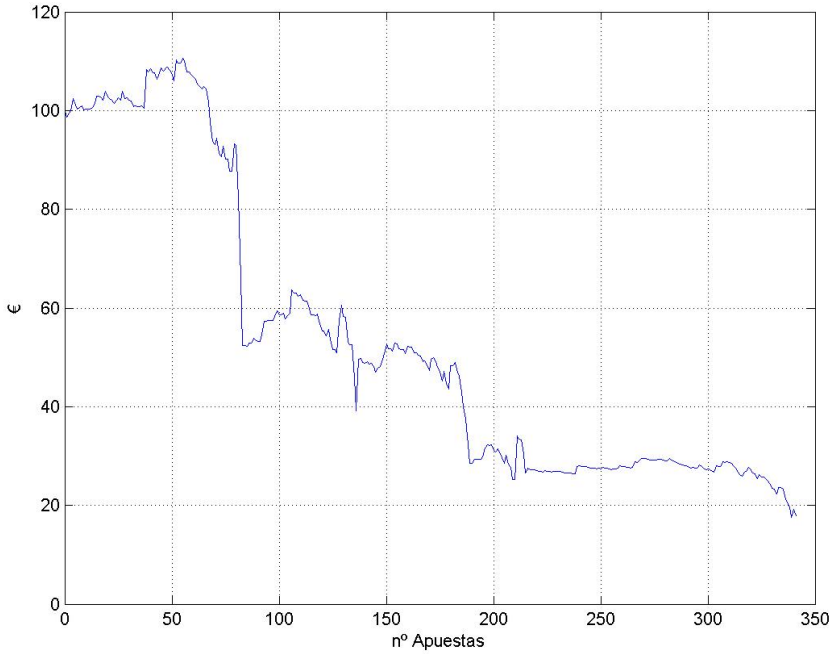


Figura 3.7: *Simulación oposición promedio de Aces.*

La oposición promedio de Dobles Faltas 3.8

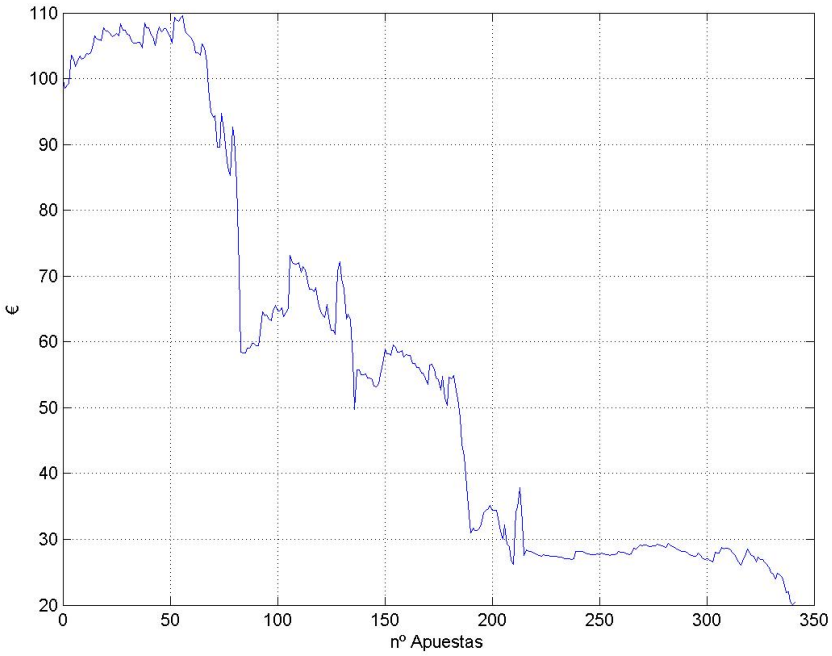


Figura 3.8: *Simulación oposición promedio de Dobles Faltas.*

La oposición porcentaje promedio de puntos con el primer servicio 3.9

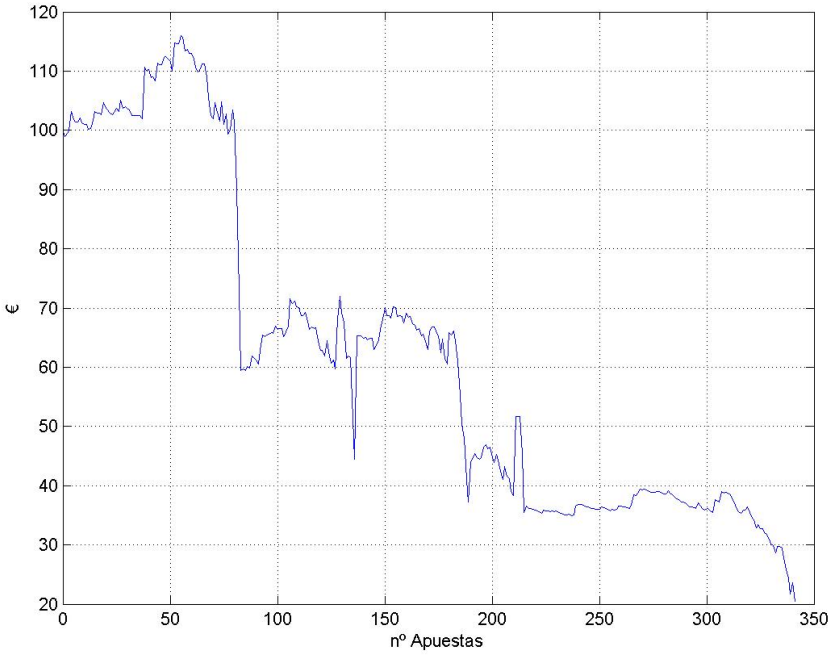


Figura 3.9: *Simulación oposición porcentaje promedio de puntos con el primer servicio.*

La oposición porcentaje promedio de puntos con el segundo servicio 3.10

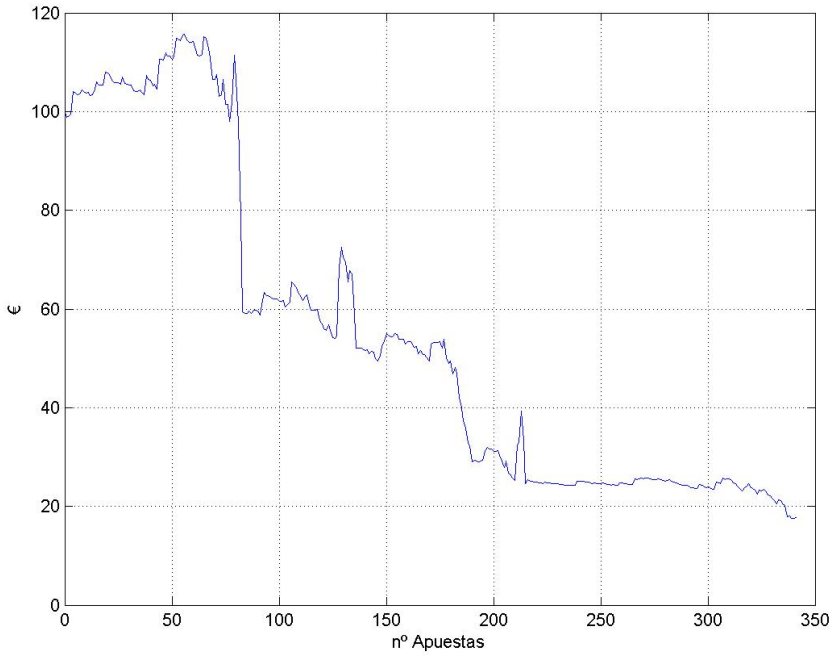


Figura 3.10: *Simulación oposición porcentaje promedio de puntos con el segundo servicio.*

La oposición porcentaje promedio de puntos ganados al resto del primer servicio 3.11

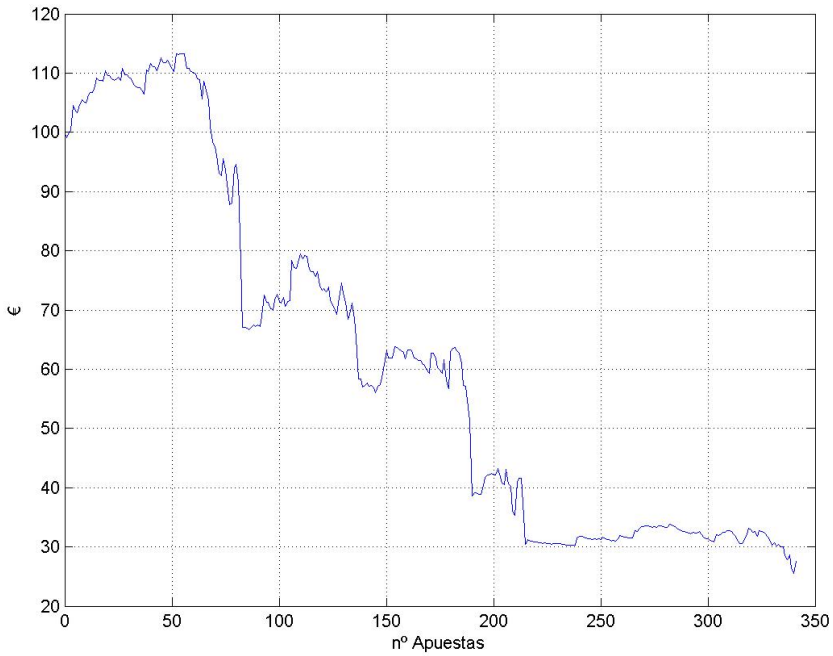


Figura 3.11: *Simulación oposición porcentaje promedio de puntos ganados al resto del primer servicio.*

La oposición porcentaje promedio de puntos ganados al resto del segundo servicio 3.12

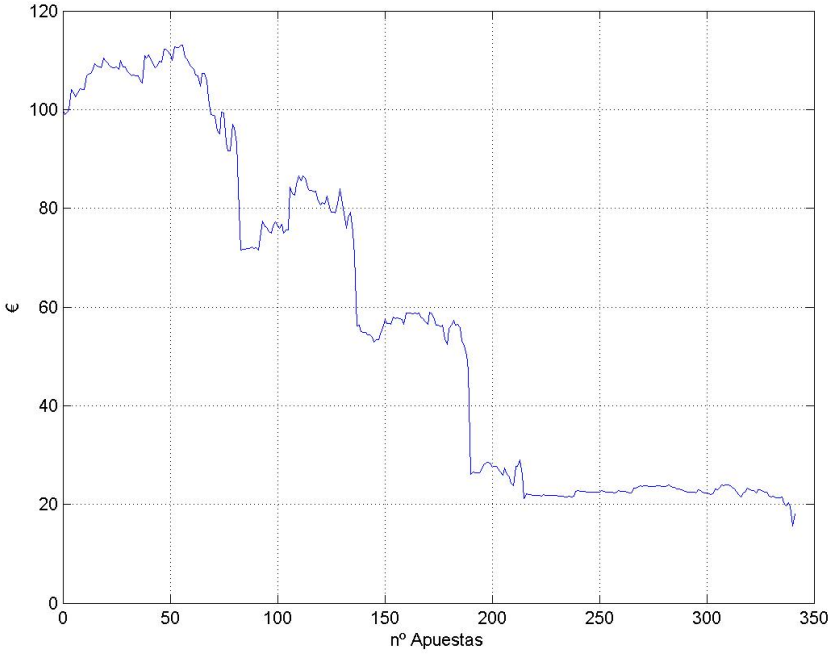


Figura 3.12: *Simulación oposición porcentaje promedio de puntos ganados al resto del segundo servicio.*

La oposición de porcentaje promedio puntos ganados con servicio corto plazo 3.13

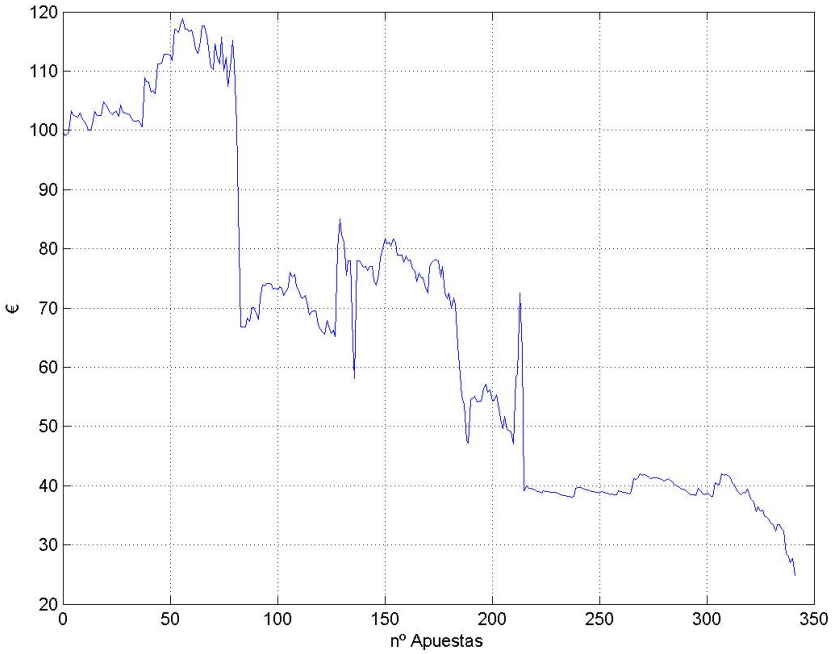


Figura 3.13: *Simulación oposición porcentaje promedio puntos ganados con servicio corto plazo.*

La oposición de porcentaje promedio puntos ganados con el resto corto plazo 3.14

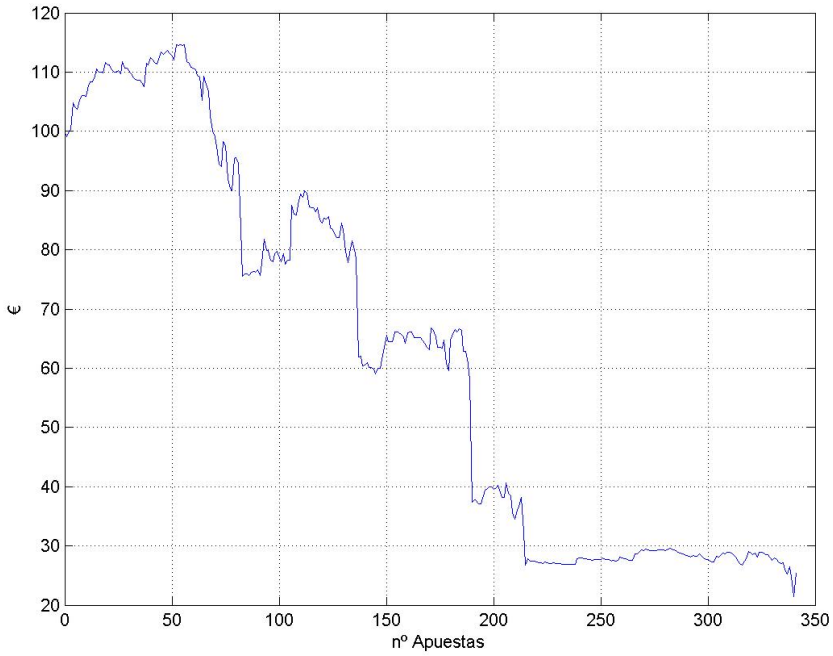


Figura 3.14: *Simulación oposición porcentaje promedio puntos ganados con el resto corto plazo.*

La oposición de porcentaje promedio puntos de rotura convertidos 3.15

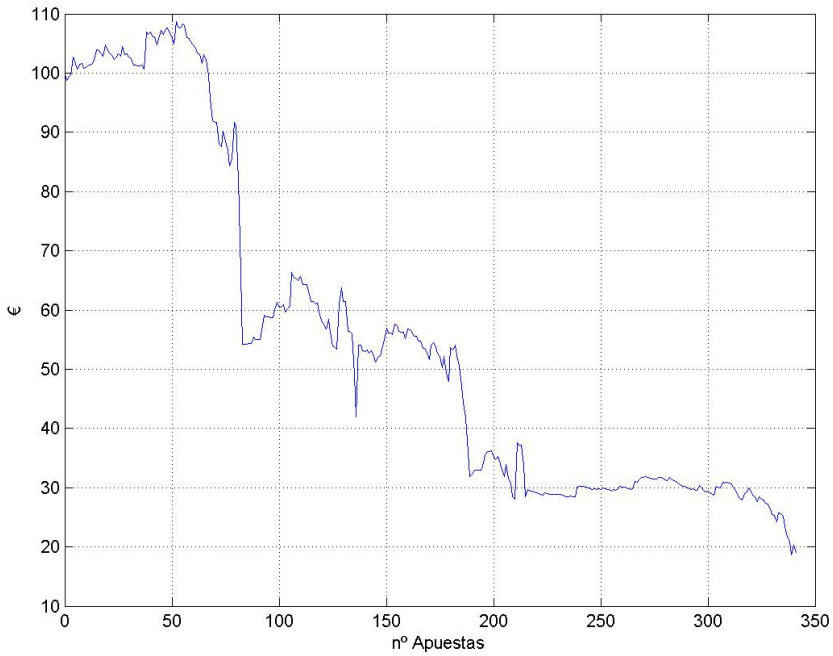


Figura 3.15: *Simulación oposición porcentaje promedio puntos de rotura convertidos.*

El analisis de las Figuras desde la 3.6 hasta la 3.15 es el mismo al que realizamos para las Figuras 3.1, 3.2 y 3.3

En estos momentos ya tenemos una variable interesante, por lo que la utilizaremos como base. Para tratar de afinar vamos a utilizar más variables.

La oposición de las variables 8, 15, 27, 34 3.16

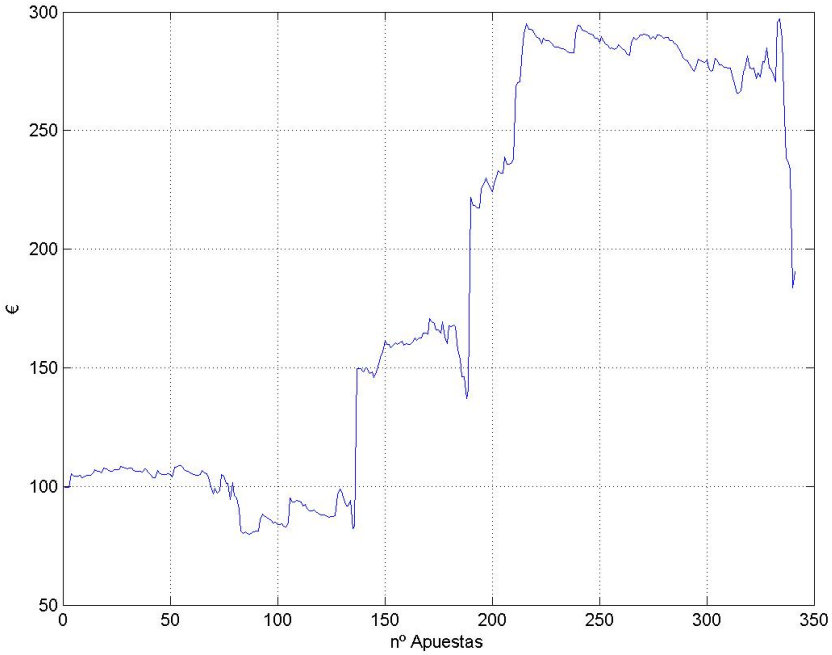


Figura 3.16: *Simulación oposición de las variables 8, 15, 27, 34.*

La oposición de las variables 1, 8, 20, 27 3.17

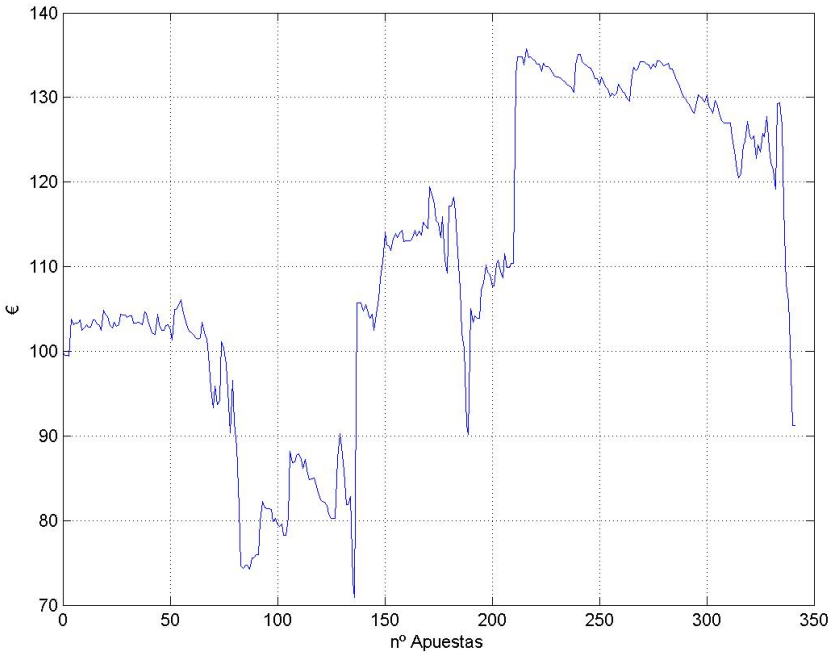


Figura 3.17: *Simulación oposición de las variables 1, 8, 20, 27.*

Podemos observar como 3.16 es parecida a 3.4 y mejor que 3.17

La oposición de las variables 8, 15, 17, 18, 27, 34, 36, 37

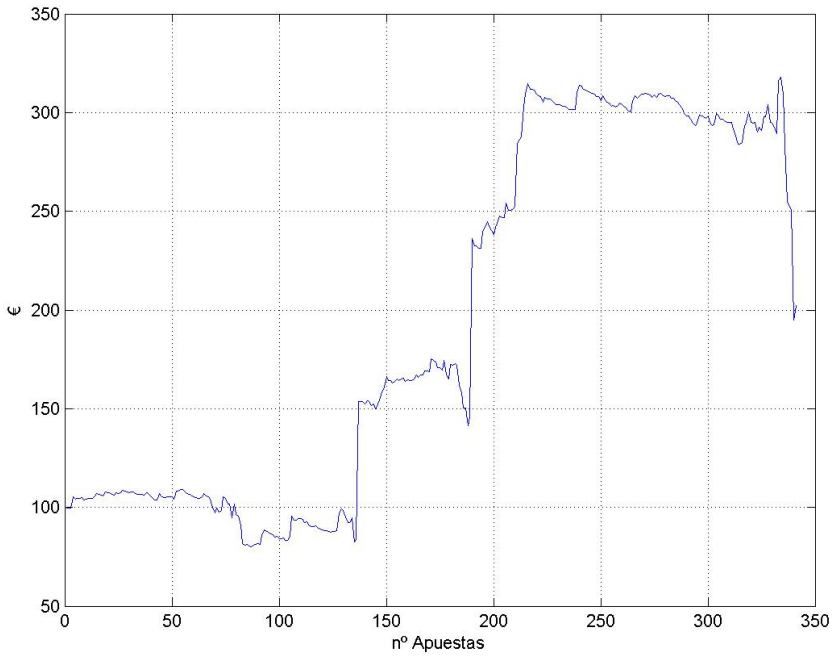


Figura 3.18: *Simulación oposición de las variables 8, 15, 17, 18, 27, 34, 36, 37.*

La oposición de las variables 8, 15, 27, 34, 39, 40, 41, 42, 44, 45

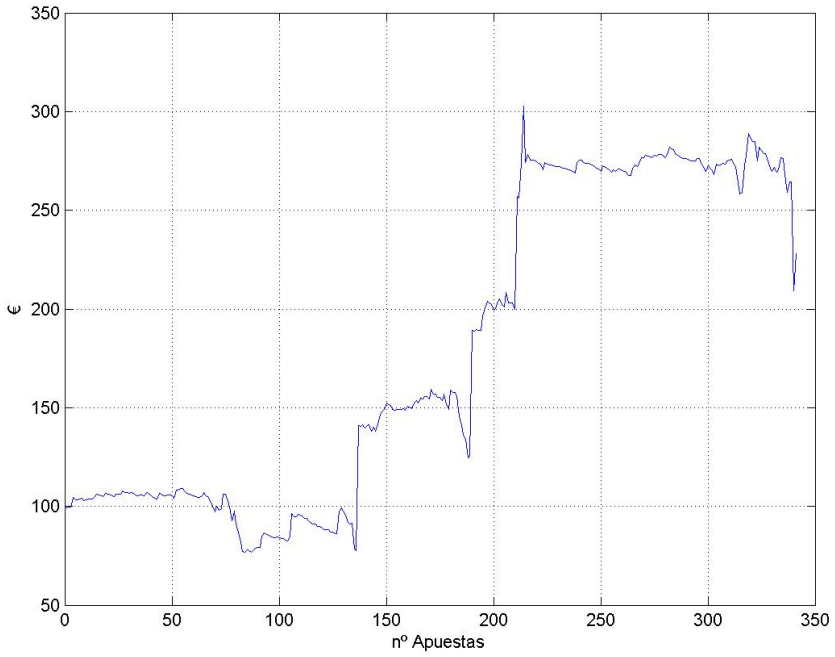
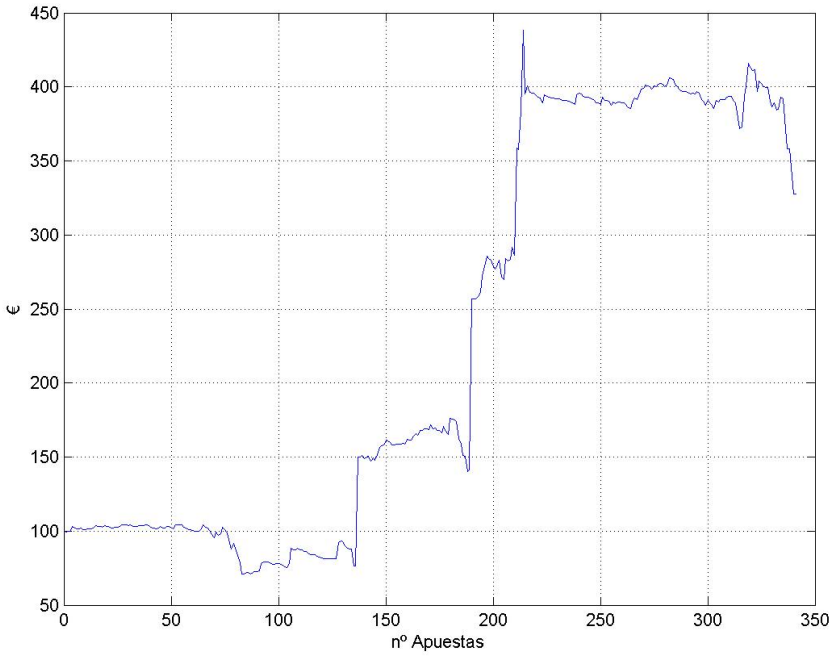


Figura 3.19: *Simulación oposición de las variables 8, 15, 27, 34, 39, 40, 41, 42, 44, 45.*

Con la prueba de las distintas variables vamos observando como nuestras simulaciones van siendo mejores, pero aún tenemos que mejorar.

La oposición de las variables 2, 3, 4, 8, 9, 10, 21, 22, 23, 27, 28, 29, 39, 40, 41, 42, 43, 44, 45, 46,



47, 48, 49

Figura 3.20: *Simulación oposición de las variables 2, 3, 4, 8, 9, 10, 21, 22, 23, 27, 28, 29, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49.*

Antes de llegar aquí hemos probado con varias variables, pero creemos que con lo mostrado es representativo. En la Figura 3.20 observamos que nuestro *bank* final es más de 3 veces superior a nuestro *bank* inicial.

Una de las aproximaciones que hemos tomado en todas las gráficas desde la 3.1 hasta 3.20 para que computacionalmente fuera viable es que el proceso de entrenamiento utilizar la aproximación de Laplace, y con los hiperparametros calculados en la predicción utilizar la aproximación EP.

Figura 3.20 utilizando tanto para entrenamiento como para predicción la función Laplace

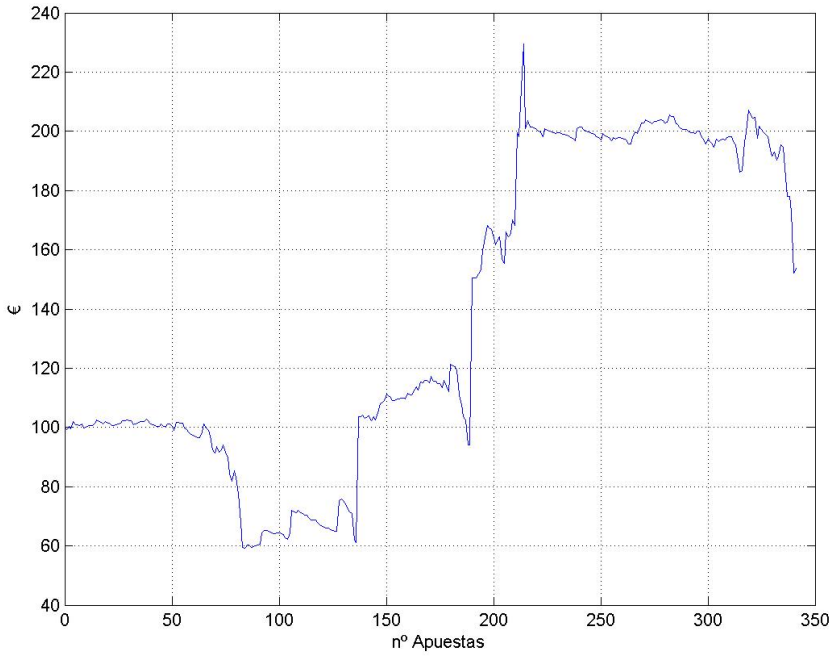


Figura 3.21: *Simulación 3.20 utilizando aproximación de Laplace.*

Comparando las Figuras 3.20 y 3.21 observamos que es mejor la aproximación inicial que habíamos tomado. Aunque la mayor parte del cómputo, que es el entrenamiento se realiza con una aproximación y que la parte de la predicción se realice con la otra, realizamos una buena predicción.

Hemos hablado anteriormente de la importancia de elegir una casa de apuestas, en la figura 3.22 podemos observar la comparativa entre la casas Pinnacle y Bet365 [3].

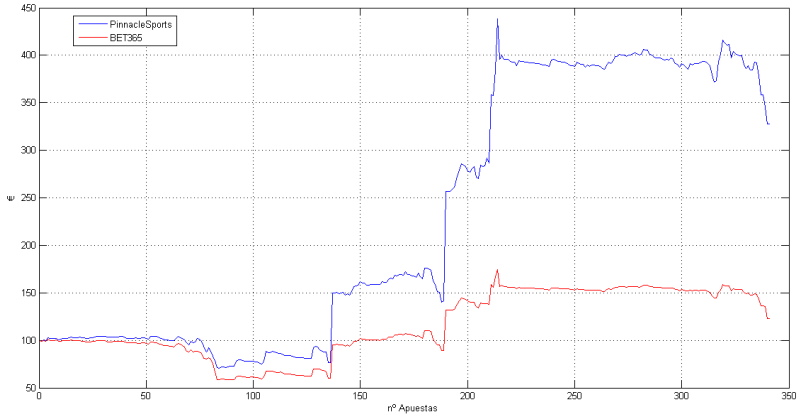


Figura 3.22: *Simulación comparativa Pinnacle vs Bet365.*

Se realizó una prueba piloto para tratar de utilizar el predictor y predecir los partidos de la edición de Roland Garros 2011 en tiempo real es decir día a día, el tiempo necesario para poner el sistema en funcionamiento no era mucho, y era más bien sencillo los pasos:

1. Actualizar la base de datos, para la 1ª ronda es recomendable la forma semiautomática que utilizamos al principio, para el resto de rondas, el tiempo varía en función de los partidos disputados, pero no superior a 15-20 minutos.

2. Búsqueda de los partidos de día, se puede realizar en la web del torneo o en la página de la ATP, aproximadamente 5 minutos

3. Búsqueda de cuotas para estos partidos en la casa o casas de nuestra elección, si unimos este paso con el anterior nos ahorraremos los 5 minutos anteriores, en este paso 15-20 minutos, depende del número de casas que manejemos y lo familiarizados que estemos con el entorno.

4. Introducción de los datos recolectados al predictor. 5 minutos

5. Espera de los resultados del predictor. Aproximadamente 30 min.

6. realización de las apuesta recomendadas por el predictor, este paso también varía en función del número de partidos, pero tardaremos lo mismo que en el paso 3.

En total aproximadamente 1 hora y media.

Después de realizar la prueba observamos un fallo clave en el tratamiento de datos, con lo que los resultados de esta prueba no son válidos y no se pueden mostrar, pero el haberlo hecho, si que nos ha ayudado para obtener una estimación real de lo que costaría en tiempo utilizar el sistema. Y poder a valorar la incursión de más módulos para la automatización del proceso.

Capítulo 4

CONCLUSIONES Y LÍNEAS FUTURAS

4.1. Conclusiones

Nuestro objetivo es conseguir un mejor sistema de estimación de probabilidades que el de las casas de apuestas, para los partidos de tenis, a través del cálculo de probabilidades no es posible establecer que nuestro sistema es mejor. Hemos utilizado una serie de herramientas adicionales que apoyadas en la estimación de probabilidades crean un modelo de estrategia económica de manera tal, que si el modelo consigue ganancias, implicará que nuestra estimación de probabilidades propuesta es mejor que la de las casas de apuestas. Para relacionar la mejora en la estimación de probabilidades y ganancias utilizamos el Criterio de Kelly, que en base a las probabilidades calculadas lo más reales posibles y la ley de los grandes números produce unas ganancias crecientes de forma exponencial. Una vez establecidas las herramientas, el siguiente paso es el de definir una estrategia, es posible que la nuestra no sea la mejor, pero hemos podido comprobar que nos lleva al objetivo. Para poder realizar las simulaciones necesitamos la elección de las variables o características significativas para nuestro predictor y aquí viene uno de nuestros grandes problemas, para simularlo necesitamos la estrategia económica, en el caso que el predictor fuera un producto de mercado, deberíamos preguntarle al cliente cuál sería la manera que pretende gestionar su *bank*, su estrategia económica. Y en base a esto realizar el módulo del Criterio de Kelly, debido a que en nuestro caso es el que gestiona la estrategia económica.

Hemos podido observar como la aproximación EP es más exacta que la aproximación de Laplace, también tiene mayor coste computacional, pero obtiene una mejor estimación de las probabilidades. Además a medida que vamos aportando más variables al proceso gaussiano este

va realizando mejores predicciones.

4.2. Líneas Futuras

A lo largo de todo nuestro trabajo hemos estado hablando de realizar apuesta a la victoria de partidos en la categoría masculina del circuito de tenis. Pero es solo una pequeña muestra de lo extensible que puede ser nuestro trabajo. Ya que el trabajo está preparado para siguiendo en el circuito masculino de tenis realizar apuestas en otros mercados, como pueden ser el nº de sets, la ocurrencia de tiebreak, etc. Los cambios a realizar para ello no serían sustanciales ya que las variables no cambiarían. Sin embargo en este trabajo no se han incluido debido a que no hemos podido acceder de forma automática, a una base de datos de cuotas que reflejen la ocurrencia de estos eventos en los partidos del circuito. Para en un futuro poder trabajar con ello se podrían obtener de las páginas web de las casas de apuestas con algún script, crear una cierta base de datos, en poco tiempo tendríamos bastentes con los que empezar porque partidos de tenis hay durante todo el año, excepto el mes de diciembre.

Otra de las líneas de trabajo futuras es en lo que hemos venido hablando durante este trabajo, las ventanas de partidos en las variables y las estrategias económicas.

Además todo esto se podría implementar para el circuito femenino, el mismo tipo de variables, todo el código valdría para el circuito femenino, con la excepción que en esta categoría el número de set de todos y cada uno de los partidos, incluidos los partidos de los Grand Slam son a 3 set, por lo que tendría incluso mayor homogeneidad, el resto de mercados también serían factibles.

Pero no nos quedemos solo ahí, ya que el trabajo se puede extender y no solo cambiar ciertas cosas, para abarcar más. Podemos avanzar hacia atrás, en el proceso de captación de datos y hacía delante en el proceso de apuesta, se podría desarrollar basado en nuestro predictor un sistema automático de apuesta para que el usuario, si tiene total confianza en él, no tuviera que realizar ninguna acción y estaría desatendido lo que nos ahorraría todo el tiempo que se tarda en poner en funcionamiento el sistema y realizar las apuestas. En la Figura 4.1 vemos los módulos en verde que tenemos desarrollados y en rojo los que proponemos a desarrollar.

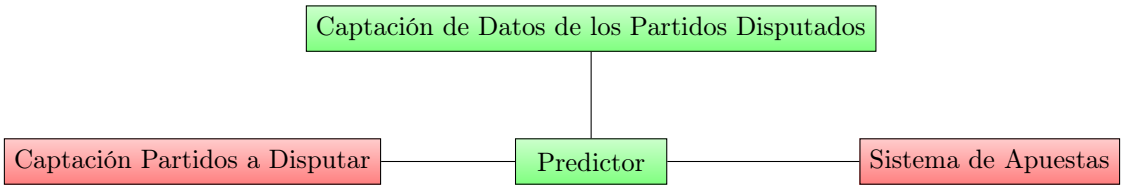


Figura 4.1: *Propuesta de Sistema Futuro*

APÉNDICES

APÉNDICE A

PRESUPUESTO DEL PROYECTO

En este apéndice se presentan justificados los costes globales de la realización de este Proyecto Fin de Carrera. Nos vamos a centrar en los costes imputables a gastos de personal, no obtante comentaremos también los de material, aunque estos últimos en los proyectos de este tipo no suelen ser excesivos, debido a que el ordenador, las licencias de software, luz... serían compartidos con otros proyectos. Se pueden deducir de las Tablas A.1 y A.2.

En la Tabla A.1 se muestran las fases del proyecto y el tiempo aproximado para cada una de ellas. Así pues, se desprende que el tiempo total dedicado por el Ingeniero ha sido de 280 horas. Teniendo en cuenta que el sueldo medio de un Ingeniero de Telecomunicaciones con una experiencia entre 0 y 3 años [8] establece un sueldo anual bruto de 25.000€, tenemos que el sueldo semanal es de 446,42€ así y como 280h son 7 semanas de trabajo el coste de personal serán 3125€.

En la Tabla A.2 se recogen los costes de material desglosados en equipo informático, (este valor es el correspondiente a un ordenador de gama media comprado anualmente amortizado de manera semanal) y gastos varios no atribuibles (material fungible, llamadas telefónicas, despla-

Tabla A.1: *Fases del Proyecto*

Fase 1	*Documentación*	60 *horas*
Fase 2	*Desarrollo del software*	40 *horas*
Fase 3	*Análisis de la base de datos*	140 *horas*
Fase 4	*Redacción de la memoria del proyecto*	40 *horas*

Tabla A.2: *Costes de material*

Ordenador de gama media	200€
Gastos varios	100€

Tabla A.3: *Presupuesto*

Concepto	**Importe**
Coste Personal	3.125€
Coste Material	300€
Base Imponible	3.425€
I.V.A (18 %)	616,5€
TOTAL	4.041,5€

zamientos...). Ascienden, pues, a un total de 3.640 €.

A partir de estos datos, el presupuesto total es el mostrado en la Tabla A.3.

Bibliografía

[1] Atp world tour. Website. http://es.atpworldtour.com/.

[2] Base de datos principal. Website. http://tennis.matchstat.com.

[3] Bet365. Website. http://www.bet365.com.

[4] Betclic. Website. https://es.betclic.com/.

[5] Bwin. Website. https://www.bwin.com/es/.

[6] Daniel mateos-sportyy. Website. http://www.sportyy.com/es.

[7] Excell con partidos disputados. Website. http://www.tennis-data.co.uk.

[8] Infojobs. Website. http://www.infojobs.net/.

[9] Jxl api excell. Website. http://jexcelapi.sourceforge.net.

[10] Pnnaclesports. Website. http://www.pinnaclesports.com/.

[11] Sergi-estoestenis. Website. http://www.estoestenis.com/.

[12] Speculators-miramiapuesta. Website. http://miramiapuesta.com/.

[13] T.hansen. Website. http://betintennis.blogabet.com/.

[14] Toolbox gp. Website. http://www.gaussianprocess.org/gpml/code/matlab/doc/.

[15] T. Barnett and S. R. Clarke. Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, 16:113–120, 2005.

[16] T. Barnett and G. Pollard. How the tennis surface affects player performance and injuries. *Society of Tennis Medicine and Science*, 12:34–37, 2007.

[17] J. Byous. *Java technology: The early year.* Sun Developer Network, 1998.

[18] J. J. M.-F. Fernando Pérez-Cruz and S. Caro. Nonlinear channel equalization with gaussian processes for regression. *IEEE Transactions on Signal Processing*, 56:5283–5286, 2008.

[19] J. J. M.-F. Fernando Pérez-Cruz and P. M. Olmos. Joint nonlinear channel equalization and soft ldpc decoding with gaussian processes. *IEEE Transactions on Signal Processing*, 58:1183–1192, 2010.

[20] G. S. James Gosling, Bill Joy and G. Bracha. *The Java language specification.* Addison-Wesley, 2005.

[21] J. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35:917–926, 1956.

[22] F. J. Klaassen and J. R. Magnus. Are points in tennis independent and identically distributed? evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, 96:500–509, 2001.

[23] F. J. Klaassen and J. R. Magnus. Forecasting the winner of a tennis match. *European Journal of Operational Research.*, pages 257–267, 2003.

[24] J. R. Magnus and F. J. Klaassen. On the advantage of serving first in a tennis set: four years at wimbledon. *The Statistician*, 48:247–256, 1999.

[25] B. R. Marshall. Arbitrage opportunities in sports betting markets.

[26] R. H. Mauricio Alvarez and E. Duque. Clasificación de eventos sísmicos empleando procesos gaussianos. Technical report, Universidad Técnologica de Pereira, Agosto 2007.

[27] P. K. Newton and K. Aslam. Monte carlo tennis. *Society for Industrial and Applied Mathematics*, 46:722–742, 2006.

[28] P. K. Newton and J. B. Keller. Probability of winning at tennis i. theory and data. *Studies in applied Mathematics*, 114:241–269, 2005.

[29] A. J. O'Malley. Probability formulas and statistical analysis in tennis. *Journal of Quantitative Analysis in Sports*, 4, 2008.

[30] G. Pollard and T. Barnett. Fairer service exchange mechanisms for tennis when some psychological factors exist. *Journal of Sports Science and Medicine*, 5:548–555, 2006.

[31] F. Pérez-Cruz and J. J. Murillo-Fuentes. Digital communication receivers using gaussian processes for machine learning. *Journal on Advances in Signal Processing*, 2008.

[32] M. A.-K. M. N. H. Takahashi Hiroo, Wada Tomohito. An analysis of the time duration of ground strokes in grand slam men's singles using the computerised scorebook for tennis. *International Journal of Performance Analysis in Sport*, 8:96–103, 2008.

[33] P. Triana. Probability distributions in tennis. *Business Strategy Review*, 18:89–91, 2007.

[34] A. B. Tristan Barnett and S. R. Clarke. Optimal use of tennis resources. *Proceedings of the 7th Australian Conference on Mathematics and Computers in Sport*, pages 57–65, 2004.

[35] D. M. Tristan Barnett and G. Pollard. Applying match statistics to increase serving perfomance. *Med Sci Tennis*, 13:24–27, 2008.