Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Departamento de Computación

TenisRank: Un nuevo ranking de jugadores de tenis basado en PageRank

Tesis presentada para optar al título de
Licenciado en Ciencias de la Computación

Alex Aronson

Director: Lic. Ernesto Mislej

Codirectora: Dra. Flavia Bonomo

Buenos Aires, 2015

TENISRANK: UN NUEVO RANKING DE JUGADORES DE TENIS BASADO EN PAGERANK

Ante la necesidad de lograr un ranking social, comprendido por todos aquellos seguidores del tenis, el ranking ATP se expone a quejas constantes de los jugadores y al mismo tiempo expone a nuevos tenistas a ser beneficiados con un buen torneo para poder comenzar a progresar en sus carreras. Al mismo tiempo, el ranking ATP no es lo suficientemente poderoso para lograr predecir con certezas quién será el vencedor de un partido en caso que nos basáramos únicamente en las posiciones.

Con el fin de combatir estos problemas surge la idea de la creación de un nuevo ranking que logre indicar cuáles son las chances reales de victoria de un jugador ante el comienzo de un nuevo torneo. Basándonos en el método de PageRank, generado por Larry Page y Sergey Brin, logramos generar un nuevo ranking que se destaca por la utilización de las carácteristicas del torneo para la generación del mismo.

Basándonos en un historial de 40000 partidos nos proponemos evaluar cómo se comporta el nuevo método creado en comparación de otros rankings existentes y así analizar si realmente logramos una mejora en el reflejo de la realidad.

Una vez obtenido el ranking, nos proponemos a, dado un partido, evaluar el ranking de los jugadores que lo disputan y las características del mismo y así poder indicar con que probabilidad saldrá victorioso el jugador con mejor posicionamiento.

Palabras claves: Sports Ranking, Tenis Ranking, PageRank, Ranking Methods

AGRADECIMIENTOS

A Ernesto por haberme guíado, por la paciencia, la buena predisposición y la buena onda en cada momento de todo el proceso que fuimos armando la tesis.

A Flavia por la confianza y por haberme ayudado a encontrar una idea que me permita acercar mis dos pasiones: El deporte y la computación.

A mis viejos por cada consejo brindado desde la primaria, la secundaria y la facultad hasta este momento. Por preocuparse siempre de mi evolución en la carrera e insistir para que no me de por vencido. Por apoyarme en cada una de mis decisiones tomadas. Me brindaron todo lo necesario para poder llegar aca de la mejor forma y estoy infinitamente agradecido.

A Melu por acompañarme durante toda la carrera incondicionalmente. Por comprender mis ataques de locura, mis nervios ante los examenes y ayudarme a superarlos. Por no haberme dejado caer nunca.

A mis hermanas por el apoyo de todos los días.

A mis abuelos por estar siempre pendientes y más contentos que yo de que haya llegado hasta acá.

A mis amigos por las mil salidas postergadas, por las palabras de aliento en todo momento.

A mis compañeros de facultad por haber ayudado a hacer de la facultad un lugar ameno.

A mis compañeros de trabajo por ayudarme a llegar a la meta sin poner trabas en el camino.

A Marcelo Albamonte por ayudarme a visualizar la cantidad de cosas que se pueden hacer en el deporte teniendo como base lo aprendido en la facultad.

A Jorge por ser mi soporte en la paradas bravas y acompañarme en los buenos momentos.

A mis profesores de la facultad, por todo lo aprendido durante estos años.

A mi zeide Ernesto y mi tío Hector.

A mi familia y a Melu.

Índice general

1. RANKINGS EN GENERAL

1.1. Introducción a los rankings

En nuestra sociedad meritocrática, el concepto de ranking es extremo. La sociedad suele consumir el mejor producto, los buscadores muestran el documento más relevante y la gente sigue las estadísticas de su equipo favorito.

El gran consumo de rankings en varios contextos hizo necesario el desarrollo de muchos algoritmos para poder lograr rankings más eficientes y confiables.

Tal como su definición lo indica, un ranking es una clasificación de mayor a menor, útil para establecer criterios de valoración. Es, generalmente, una lista que establecerá una relación entre un conjunto de elementos que se reúnen en la misma debido a una característica en común.

Es común el empleo de los rankings en varios ámbitos para establecer algunos niveles o determinarlos. Esto lo podemos ver por ejemplo en el ambiente musical tanto como en el mundo de las finanzas y de los negocios, donde es recurrente encontrarse con diferentes rankings ya sea de las empresas o compañías más exitosas como de los productos más vendidos en el lapso de un mes, de los hombres más ricos del mundo, etc.

En lo que respecta a la moda y la belleza, también tienen lo propio con los rankings que realizan las revistas especializadas como por ejemplo los mejores y los peores vestidos.

Por último, y en los que haremos foco, también en el mundo del deporte es factible encontrarse con infinidad de rankings tales como el ranking de los goleadores en el fútbol, de los mejores jugadores de tenis o el ranking FIFA que indica que selecciones de fútbol son las mejores, entre otros.

1.2. Rankings deportivos

Los rankings deportivos son diseñados a raíz de varios objetivos. Algunos de ellos son objetivos comerciales, por ejemplo garantizar la presencia de equipos o jugadores en algunos torneos.

Quizás con igual importancia tanto los medios periodísticos como los observadores interesados económicamente en el juego usan los rankings como una manera de asesorarse y evaluar quiénes son los mejores jugadores o equipos en la actualidad.

Más aún, los jugadores suelen mencionar como una de sus mayores motivaciones la posición en el ranking en que están calificados tanto ellos como su equipo.

Se puede discutir que ningún sistema de ranking refleja fehacientemente la per-

formance de cada jugador, sino que es un método simplista para determinar quién está jugando de la mejor manera en un momento dado de la temporada.

Los rankings deportivos tienen además y casi de forma principal una función social. Con el objetivo de lograr captar la mayor cantidad de personas que consuma los torneos profesionales se suelen generar rankings sencillos de entender, para que todos los espectadores puedan sentirse atraídos ante los diferentes partidos de los torneos.

Sin embargo, en épocas donde se permite contar con tanta información estadística, gracias a los medios digitales, que reflejan datos que antes eran imposibles de calcular fácilmente, se abre la oportunidad de generar rankings más exactos al momento de indicar quienes son realmente los mejores jugadores pero al mismo tiempo dificiles de comprender para la sociedad.

Otro uso importante que tienen los rankings es su carácter predictivo, donde se puede determinar quién va a ser el ganador de un partido mirando el posicionamiento de los jugadores en cuestión. Muchas casas de apuestas, toman estos factores como para poner el valor de pago para cada jugada.

1.3. Rankings predictivos vs Rankings clasificatorios

En general, la mayoría de los sistemas de rankings fallan en una de las dos categorías, o son predictivos o son clasificatorios.

El objetivo de los rankings clasificatorios es indicar a los equipos en base a su participación en una determinada competencia a lo largo de la temporada. De esta manera se puede elegir al campeón o utilizar también para indicar un conjunto de equipos que haya clasificado por su posición en el ranking para participar de un torneo particular. El objetivo de un ranking predictivo, sin embargo, es proveer el mejor pronóstico sobre el resultado de un partido en que se enfrentarán dos jugadores o equipos.

Un ranking clasificatorio objetivo debería tomar como factores determinantes quién fue el ganador, o la diferencia de puntuación entre los jugadores, o la combinación de ambas. El hecho de usar un buen criterio bien definido permite a los equipos o jugadores conocer con exactitud las consecuencias de ganar, empatar, o perder un partido. Esto se usa generalmente en las tablas de posiciones de los deportes donde cada equipo o jugador recibe una posición, siendo la más baja la mejor rankeada.

Por otro lado, para la creación de rankings predictivos con la mayor exactitud posible, está permitida la inclusión de cualquier información adicional que sea útil, como por ejemplo: goles a favor de un equipo, partidos ganados, goleadores en el caso del fútbol o características definidas del torneo en particular y su historial en esos torneos o versus esos rivales, entre otras cosas.

2. ATP RANKING

2.1. Un poco de historia

Es un objetivo universal de los jugadores: convertirse en el No. 1 del mundo. Los niños sueñan con lograrlo, el esfuerzo está puesto en ello. Sin embargo sigue siendo uno de los logros más esquivos en el deporte. En los 40 años que tienen de existencia los rankings de tenis, sólo 25 jugadores han llegado a la cumbre del (hoy) Emirates ATP Ranking, y sólo 16 terminaron la temporada como No. 1.

Tal como cuenta en [1], desde los inicios de la Era Abierta del tenis en 1968, las clasificaciones fueron en gran medida un cálculo subjetivo, generado por las asociaciones nacionales, diferentes circuitos y periodistas especializados que compilaron sus propias listas.

Algunos jugadores estaban en una lista de tenistas que podían ayudar a vender boletos para el evento, lo que los colocaba como prioridad por sobre los demás en la lista de aceptación en los torneos. Esto causó una gran preocupación en aquellos que no tenían un gran nombre y estaban al límite de entrar o no a los eventos.

En agosto de 1972 se hizo evidente que la recién creada Asociación de Tenistas Profesionales sería necesaria para establecer un sistema de clasificación único, sin opiniones personales ni prejuicios. Esta determinaría la forma de ingreso a los torneos y marcaría una muestra objetiva sobre la actuación de los jugadores. Doce meses después de la fundación de ATP, Ilie Nastase se convirtió en el primer No. 1.

Los torneos fueron inicialmente divididos en categorías –A,B,C, etc...–, lo que permitió a los organizadores del evento seleccionar a los jugadores de acuerdo a su clasificación ATP y determinar los cabezas de serie.

Los resultados de los torneos eran enviados a la ATP donde eran procesados para luego generarse el ranking en enormes hojas perforadas y una vez al mes publicar el nuevo ranking de jugadores.

En los siguientes años, después de 11 publicaciones en 1973, el ATP Ranking comenzó a publicarse con mayor frecuencia –1974 (11), ’75 (13), ’76 (23), ’77 (34), 78 (40)– hasta 1979, cuando se produjo una vez por semana, 43 en la temporada.

El ATP Ranking internacional de jugadores del 23 de agosto de 1973 era un sistema de promedio, acumulaba los puntos obtenidos durante un período de 52 semanas y se dividía por el total de torneos jugados (con un divisor mínimo de 12). Los torneos otorgaban puntos de acuerdo a los premios (mínimo 25 mil dólares), el tamaño del cuadro y su dificultad. El sistema basado en el mérito fue respaldado por los jugadores.

Luego de varios cambios en el sistema de conteo de puntos en el año 2000, para fomentar una mayor participación en Grand Slam y la serie de nueve torneos más im-

portantes del ATP (ahora conocido como ATP World Tour Masters 1000), el sistema de clasificación comenzó a contar 18 eventos para la mayoría de los jugadores. Los 13 resultados de Grand Slam y torneos ATP World Tour Masters 1000 contarían, al igual que las mejores cinco actuaciones de un jugador en los eventos de la Serie Internacional (ahora ATP World Tour 250 y 500).

2.2. Sistema de Ranking ATP

Los Rankings Emirates ATP son el método histórico objetivo basado en méritos para determinar la aceptación y siembra en todos los torneos para singles y dobles.

Como se explica en [2, 3], el período que toman los Rankings Emirates ATP son las últimas 52 semanas, sin embargo no se toman todos los torneos en los que participa un jugador durante ese período. El ranking ATP está basado en el total de puntos conseguidos por un jugador en los cuatro (4) Grand Slam, ocho (8) de los nueve (9) ATP Masters 1000, y sus mejores seis (6) resultados del resto de los torneos ATP en que participó. Para jugadores que no pertenecen al Top-30, y que no clasifican directamente a los torneos ATP Masters 1000 y Grand Slam, en caso de no jugarlos se toma en cuenta un séptimo torneo del resto de los torneos jugados durante el año. Al final de la temporada se juega el ATP World Tour Finals, con sede en Londres, donde los mejores siete (7) jugadores son clasificados automáticamente y el ganador de uno de los torneos Grand Slam (en caso de que ya esté clasificado irá el octavo clasificado en el Ranking ATP)

Level/Round	W	F	SF	QF	16	32	64	128
Grand Slams	2000	1200	750	360	180	90	45	10
Masters 1000	1000	600	360	180	90	45	10[25]	[10]
ATP 500	500	300	180	90	45	20	0	
ATP 250	250	150	90	45	20	[10]	0	

Sistema de puntuación por ronda alcanzada en cada tipo de torneo

2.3. Críticas al ranking ATP

Basándonos en lo que cuenta [4] existen muchas críticas de los jugadores al sistema de puntuación del ranking ATP, muchas de ellas hechas por los jugadores más importantes como Rafael Nadal.

En el actual ranking ATP las medidas están hechas sobre 52 semanas, lo que puede hacer al ranking algo inconsistente semana a semana. Por ejemplo, en el caso hipotético en el que un jugador gane un ATP Masters 1000, este acumulará esos 1000 puntos durante 51 semanas desde la final, pero si no lo repite el título, 53 semanas después de la

obtención del trofeo, no tendrá más esos 1000 puntos. La potencial oscilación entre dos posiciones en el ranking con fechas tan cercanas no refleja realmente el verdadero ranking de los jugadores.

Por otro lado, todos los torneos del mismo nivel suman la misma cantidad de puntos, independientemente de los jugadores que participen del mismo. Al contar solamente 6 de los torneos no considerados "grandes", un jugador puede participar en tantos torneos de esos como pueda, con el objetivo de engrosar su puntaje y obtener así la posibilidad de ser preclasificado en los torneos "grandes" y obtener un cuadro de torneo más favorable.

Por último, el sistema de ranking actual penaliza a aquellos que obtienen un desafortunado draw en los torneos grandes. Si a un jugador siempre le toca jugar contra un Top-10 en primera ronda, sus chances de engrosar su ranking son pequeñas.

2.4. ATP como modelo predictivo

Si miráramos al Ranking ATP como un Ranking no solo clasificatorio sino como uno predictivo, podríamos tratar en base al posicionamiento de los jugadores, de intuir ante cada partido quién sería el ganador.

Considerando al jugador con número menor en el ranking como el jugador con mejor presente en ese momento dado, evaluamos en partidos desde el año 2005 a 2013 cómo funcionó este modelo predictivo.

De esta forma obtenemos:

AÑO	ATP
2005	66.430%
2006	66.494%
2007	65.455%
2008	66.990%
2009	68.072%
2010	66.641%
2011	67.758%
2012	67.908%
2013	65.892%
PROMEDIO TOTAL	66.849%

Eficacia del ranking ATP como sistema predictivo

De esta manera podemos decir que el ranking ATP acierta en aproximadamente 2/3 de los partidos tal como está establecido. También se puede interpretar que en aproximadamente 2/3 de los casos el jugador mejor posicionado es el ganador del encuentro.

La forma en que se está evaluando es contando la cantidad de partidos donde el jugador mejor posicionado ganó, es decir un HIT, sobre el total de partidos que se jugaron durante el período en el que se está evaluando.

3. PAGERANK

3.1. Introducción

El PageRank es un valor introducido por Google con el objetivo de valorar y diferenciar las páginas existentes de acuerdo a su importancia.

Mide la "autoridad" que tiene una página sobre varios temas concretos. Cuanto más autoridad, más posibilidades tendrá de salir en las primeras posiciones del ranking y figurar así en las búsquedas que traten sobre dicho tema.

Se mide conociendo la cantidad de enlaces que apuntan hacia esa página, junto con la autoridad de la página que la enlaza y la forma en que lo hace.

Cada enlace cuenta como un voto o una recomendación. Y es tan importante la recomendación como quién la recomienda y cómo lo hace. Esto es decir, que si el recomendador tiene mayor peso, su recomendación será de mayor importancia a la de otro sitio.

3.2. Historia del PageRank

Ante el crecimiento inmenso que tuvo internet y la gran cantidad de páginas heterogéneas existentes en la web, los motores de búsqueda se vieron obligados a perfeccionar sus rankings para poder brindar a usuarios inexpertos las mejores respuestas.

Inicialmente los motores de búsqueda funcionaban como crawlers de sitios web en la que se listaban términos encontrados en las páginas. Cuando se buscaba una palabra en el buscador, ésta era buscada en los listados de términos y se mostraban las páginas donde aparecían. El método de ordenamiento era basado en la cantidad de veces que una palabra era mencionada, lo que podía llevar a casos de spam donde cada uno en su site independientemente del tema repetía las palabras más buscadas para poder llevar tráfico a su web.

Analizando esto y considerando que si bien se había incrementado la cantidad de sitios, no dejaban de ser hipertextos que proveían información adicional, se creó un ranking llamado PageRank como método de computar cada página existente como un nodo de un gran grafo que formaba la web. Cada nodo tenía un grado de importancia que correspondía a si era una página importante. Esto se determinaba en base a cuántos links alcanzaban a cada nodo.

El algoritmo está patentado por los fundadores de Google, Larry Page y Sergey Brin, cuando eran alumnos de doctorado en la Universidad de Stanford en 1999 [6, 8]. Page fue alumno de doctorado de Terry Winograd, quién parece que le animó en la idea de trabajar en el PageRank, y Brin fue alumno de doctorado de Jeffrey D. Ullman, el

famoso autor del libro de compiladores con Aho, aunque pronto se unió a Page para trabajar en temas "más interesantes" que los que le ofrecía su director.

Ese algoritmo se utilizó para potenciar un nuevo motor de búsqueda llamado BackRub, que luego pasó a llamarse Google.

Internet es hoy lo que es gracias al gran esfuerzo que tuvieron por ordenar la información de forma relevante en un entorno en el que los grandes portales habían vendido los resultados de las búsquedas al mejor postor.

De los 24 millones de páginas web que su primera versión consiguió indexar, Internet ha crecido hoy hasta superar, según estimaciones, los 4.000 millones de direcciones distintas. Dado que el algoritmo ha de buscar no solo los vínculos de primer orden que una página recibe, sino también los de órdenes superiores, el problema real no se encuentra en la idea original de cómo medir la relevancia en Internet, sino en lograr indexar el mayor porcentaje de sitios existentes de Internet y en evaluar los vínculos que entran y salen de cada web.

3.3. Definición y Algoritmo

Basándose en el concepto de la centralidad de un grafo, en la que se determina un valor a un nodo en base a su ubicación, se puede determinar cuál es el nodo más influyente del grafo en su matriz de adyacencia. Un nodo con valor alto indica que está conectado a otros nodos también con valor considerable y mientras más alto el valor del nodo se va ubicando más al centro.

Utilizando estos conceptos surge la creación del PageRank que, justamente, es una función que asigna un valor a cada nodo correspondiente a una página web.

Considerando la web como un grafo dirigido conectado por los hiperlinks de las páginas que son apuntadas, es decir cada arco representa un vínculo entre el nodo saliente al nodo entrante. Se comienza por un sitio al azar y se empieza a hacer click en los distintos vínculos, navegando así de página en página.

El valor del PageRank de una página corresponde a la frecuencia con la que un navegador cualquiera visita esa página. Más tiempo pasa un usuario en una página, más importante pasa a ser su PageRank sobre esa página.

Debido al tamaño actual de la Web, el buscador de Google usa un valor iterativo aproximado de PageRank. Esto significa que a cada página se le asigna un valor inicial de PageRank, y después el PageRank de todas las páginas se calcula con cálculos cíclicos basados en la fórmula del algoritmo de PageRank.

Yendo al plano formal, como indica [18], el valor general del PageRank de cualquier página se puede representar como:

Sea u una página web, F_u el conjunto de páginas apuntadas por u y B_u el conjunto de páginas que apuntan a u. Sea $N_u = |F_u|$ el número de links de u y c el un factor usado para la normalización.

Comenzamos definiendo un ranking simple R como una versión simplificada de PageRank:

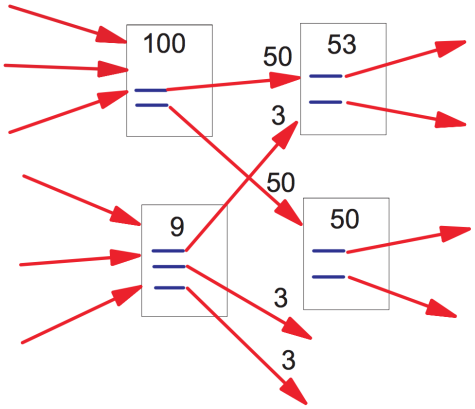
$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

Las páginas más populares son aquellas que más arcos reciben desde otras páginas o nodos del grafo. Una caracteristica importante del algoritmo de Pagerank es la de asignar pesos a las aristas. Estos son tomados en cuenta como herramienta principal para el cálculo final del ranking.

En el algoritmo de Pagerank cada página cuenta como un linkeo, sin embargo los pesos de las aristas se acumulan a medida que van conectandose los nodos del grafo.

El peso de la página es calculado sumando los pesos de los links que recibe, mientras que el peso de los arcos tiene como valor la división del peso de la página por cada link a los que hace referencia.

Esto provoca que las páginas más importantes tengan una mayor influencia en el valor del Pagerank que las páginas menos populares.



Cálculo de PageRank con peso en las aristas

Una vez calculado el peso que debe llevar cada arista, se genera una matriz de adyacencia para calcular las posiciones que cada página tomará en el ranking.

3.4. Usos del PageRank

El algoritmo de PageRank es muy útil como indicador de popularidad. Dado cualquier grafo dirigido, es capaz de indicar qué nodo es el más importante y así poder armar un ranking sobre cada uno de ellos.

Como se cuenta en [7], el algoritmo de Pagerank no es solo utilizado como uno de los principales factores para el posicionamiento de los sitios en Google. Considerando su capacidad de rankeo también lo utilizan grandes empresas para armar Rankings particulares, independientemente de los motores de búsqueda.

Por ejemplo:

- **Ranking de tweets en Twitter**: Se hace relacionando las menciones de un usuario a otro formando así un grafo donde los nodos son los usuarios y cada mención indica un arco dirigido de un usuario al otro
- **Sistemas de recomendación**: Se puede utilizar en sistemas de recomendación en base a los productos consumidos por un usuario. Por ejemplo Netflix puede recomendar películas basándose en casos de usuarios similares
- **Sugeridor de amigos en redes sociales**: Considerando que cada usuario alimenta y se alimenta de otros usuarios se puede generar un sistema de PageRank para sugerir nuevos amigos en base a los ya conectados.

3.5. PageRank como ranking de tenis

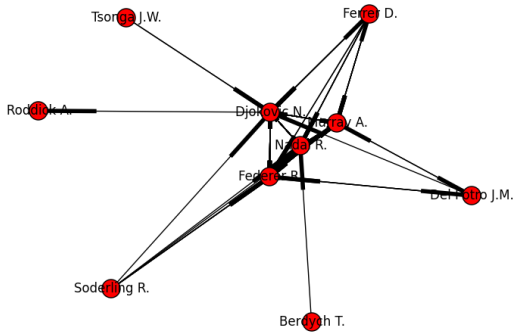
Basándonos en [9, 12, 13], existen diferentes alternativas al ranking ATP en las que se busca conseguir como principal objetivo un mejor reflejo de quiénes son los mejores jugadores en la actualidad, y en qué posición se ubica cada uno de ellos.

Uno de los algoritmos que se pueden tomar como modelo para generar un Ranking nuevo es el PageRank.

Basándonos en el paper *On the (Page)Ranking of Professional Tennis Players*, escrito por Dingle, Knottenbelt y Spanias [11], se genera un ranking de tenis nuevo, donde cada nodo es un jugador y se traza un arco orientado de conexión entre nodos ante el resultado de un partido. Cuando un jugador le gana a otro, se agrega un arco de peso 1 (uno) desde el nodo perdedor al nodo ganador.

Una vez evaluados todos los partidos se aplica el algoritmo de PageRank sobre ese grafo y así se obtiene un ranking basado exclusivamente en el historial de los resultados de los partidos, a diferencia del Ranking ATP donde se evalúa la instancia del torneo a la cual se llega.

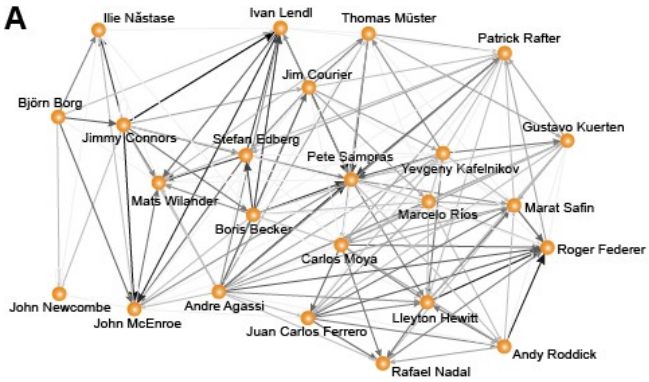
Generamos un grafo en base a los resultados de los partidos de un torneo.



A partir de este grafo el PageRank logra generar un ranking para los jugadores involcucrados

Posición	Jugador
1	Djokovic N.
2	Nadal R.
3	Federer R.
4	Murray A.
5	Ferrer D.
6	Del Potro J.M
7	Rodick A.
8	Tsonga J.W
9	Soderling R.
10	Berdych T.

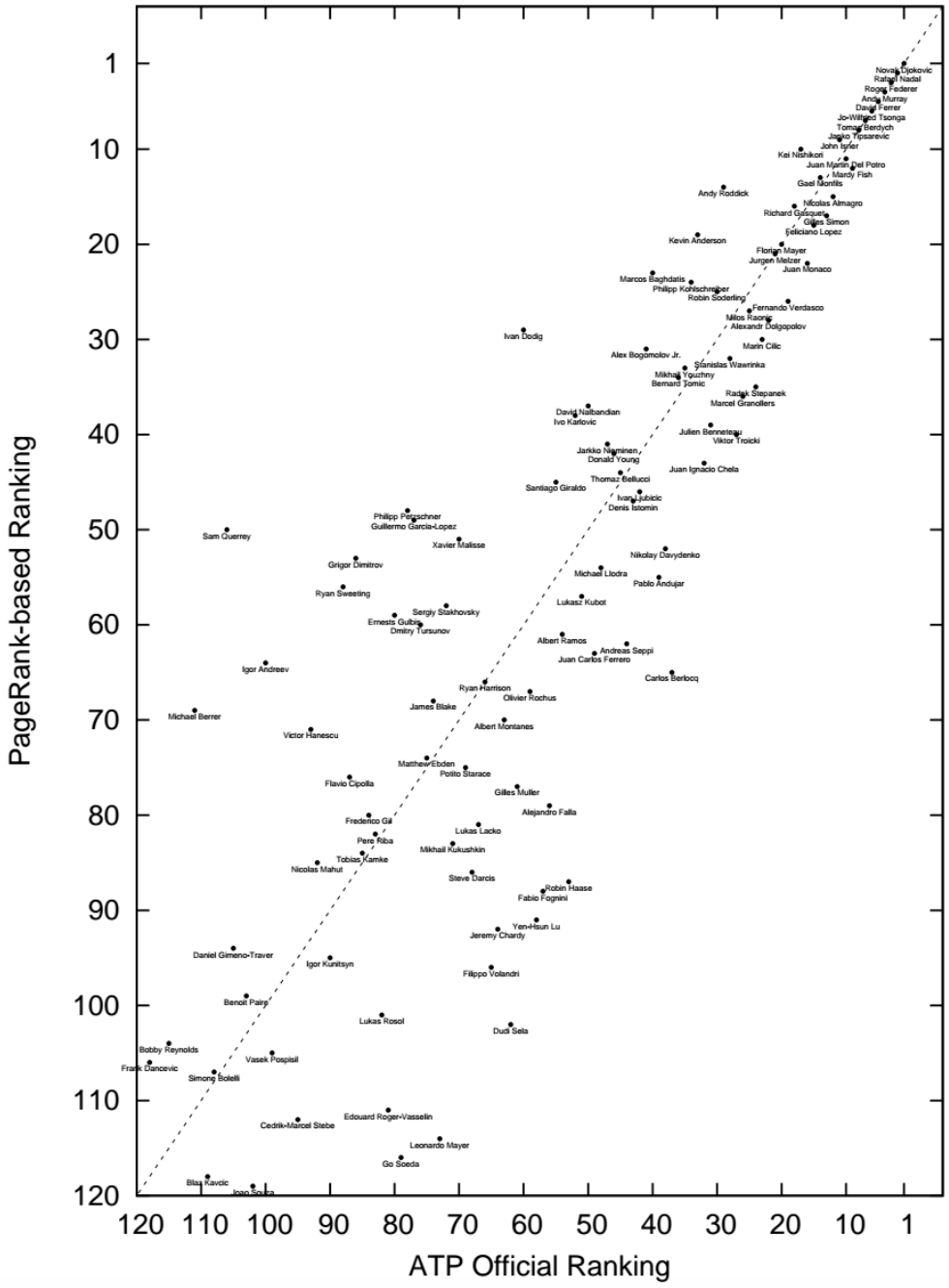
Como corolario, este método a diferencia del Ranking ATP, nos permite además armar un ranking con jugadores de épocas distintas.



Grafo generado por los resultados de los partidos ATP tomado de [10]

3.6. PageRank vs Ranking ATP

Comparando los rankings ATP vs el ranking generado a través de un algoritmo de PageRank podemos observar que los mismos pueden tener muchas variaciones entre el posicionamiento asignado a algunos jugadores por un sistema y por el otro.



Comparativa ranking ATP vs PAGERANK extraída del paper de Dingle, Knottenbelt y Spanias. [11]

POSICION	ATP	PAGERANK	DIFERENCIA
1	Djokovic	Djokovic	0
2	Nadal	Nadal	0
3	Murray	Federer	4
4	Ferrer	Murray	-1
5	Berdych	Ferrer	-1
6	Del Potro	Berdych	-1
7	Federer	Tsonga	1
8	Tsonga	Del Potro	-2
9	Gasquet	Isner	8
10	Wawrinka	Tipsarevic	11
11	Raonic	Almagro	4
12	Nishikori	Gasquet	-3
13	Haas	Wawrinka	-3
14	Janowicz	Raonic	-3
15	Almagro	Simon	1
16	Simon	Nishikori	-4
26	Lopez	Youzhny	-2
54	Hanescu	Nalbandian	50
55	Andujar	Berlocq	-7
56	Matosevic	Ramos	19
57	Delbonis	Giraldo	28
58	Stepanek	Hewitt	8
75	Ramos-Vinolas	Hanescu	-21
76	Carreno Busta	Matosevic	-20
77	Sela	Blake	23

Comparativa ranking ATP vs PAGERANK previa a US OPEN 2013

En el caso de la gráfica podemos ver que al momento de comenzar el US OPEN 2013 (Torneo Grand Slam) los mismos tienen variaciones salvo en los dos primeros lugares del ranking.

Para este torneo, una vez finalizado, el ranking ATP acertó los resultados en un $70,25\,\%$ mientras que el ranking armado por el PageRank tuvo una mayor eficiencia al acertar el $72,72\,\%$ de los resultados.

Finalmente evaluamos año a año la performance del PageRank en comparación con el ranking ATP.

AÑO	ATP	PAGERANK
2005	66.430%	66.785%
2006	66.494%	66.087%
2007	65.455%	66.939%
2008	66.990%	66.563%
2009	68.072%	69.516%
2010	66.641%	68.472%
2011	67.758%	67.524%
2012	67.908%	68.299%
2013	65.892%	68.010%
PROMEDIO TOTAL	**66.849%**	**67.577%**

Comparativa ranking ATP vs PAGERANK

Como se puede apreciar, este nuevo sistema de ranking logra una mayor precisión si lo evaluáramos como ranking clasificatorio.

Es decir, que tomar como referencia los resultados entre los jugadores y conectarlos para poder así generar un ranking, tiene una mayor eficacia al momento de predecir el resultado que evaluando únicamente la instancia del torneo a la que llegan como lo hace el ranking ATP.

4. PAGERANK PARAMÉTRICO

4.1. Motivación

Ante los buenos resultados que se obtuvieron utilizando el método de PageRank por sobre lo que brinda el ranking ATP, se nos ocurrió la idea de buscar una mejor alternativa que mejore a los ya existentes. Este nuevo método tenía que lograr contrarrestar las críticas de los jugadores al ranking utilizado hoy en día, al mismo tiempo ser más eficiente y lograr reflejar la verdadera posición que ocupan en un ranking los tenistas, comparado a los métodos ya evaluados.

Como hemos visto, la diferencia del ranking ATP comparado con el ranking PageRank es que el primero evalúa a los jugadores según la instancia a la que llegaron en los torneos jugados, mientras que el otro evalúa los resultados partido a partido que han tenido en el último tiempo.

Es por esto que nos preguntamos: *¿Es correcto evaluar solamente la instancia del torneo a la que llegan los jugadores?, ¿Es correcto evaluar solamente 52 semanas atrás para conocer el presente de un jugador?, ¿Todos los partidos ganados por un jugador ante otro tienen el mismo valor?, ¿No es importante el contexto en que se desarrolla cada victoria?.*

Es ante estas preguntas que surge la idea de combinar ambos modelos. Considerando el PageRank como un modelo más efectivo que el ranking ATP actual, lo utilizamos como esquema base para diseñar un nuevo ranking. En este nuevo sistema, no solo se tendrán en cuenta los resultados entre jugadores, también se considerarán otros atributos importantes como lo son la instancia del torneo en el que se desarrolló el encuentro (tal como lo hace ATP), la superficie en que se jugó cada partido, y la antigüedad de cada uno de los enfrentamientos.

Buscaremos el parámetro que mejor representa cada uno de estos atributos y se le asignará el peso correspondiente a cada arco del grafo. Luego se combinarán los distintos parámetros para conseguir un ranking generado por la evaluación de todos los factores mencionados.

4.2. Set de datos

Obtenidos del sitio `www.tennis-data.co.uk`, se cuenta con un set de datos de cada uno de los 923 torneos jugados entre los años 2000 y 2013.

Analizando cada uno de esos torneos, se pueden tomar resultados de casi 40000 partidos, en los que está detallado el tipo de superficie de cada partido (Hard, Clay, Grass), la fecha de cada partido, la importancia del torneo (ATP250, ATP500, MASTERS1000, Grand Slam, Master Cup).

También está indicado el resultado de partido set a set, el ranking ATP original de cada jugador al momento de desarrollarse el encuentro, y el monto de las apuestas que pagaba la victoria de cada jugador.

Todos estos datos fueron curados, se unificaron nombres de torneos, jugadores, tipo de torneo, y se volcaron en 3 tablas de una base de datos MySql para una mejor lectura. Las tablas están normalizadas por Jugadores, Torneos y Partidos. (Más info en el Apéndice)

Por otro lado, se hizo una captura de la información de la página web de la ATP (`www.atpworldtour.com`) en la que se procesó el top 100 del ranking ATP desde el año 1973.

4.3. Modelo

De manera similar a la que hace en [11], se generará por cada torneo el grafo dirigido correspondiente a los enfrentamientos del mismo.

Cada arco estará vinculado estrictamente al resultado del partido, y a medida que se recorre la lista de partidos en la que se mirarán una cantidad de años asignadas, previas al torneo que se evaluará, se agregan a un multigrafo dirigido, del que finalmente calcularemos el PageRank.

Nuestra propuesta consiste, a diferencia del PageRank mencionado, en incluir un valor como peso de cada arco, haciendo que cada partido tenga distinto valor en nuestro multigrafo.

Para ello tomaremos en cuenta 3 atributos que ayudarán a calcular el peso correspondiente para ese partido:

- Envejecimiento: Se evalúa cuán viejo es el partido a evaluar con respecto al torneo por el que se sacará el ranking
- Superficie: Se evalúa la diferencia de superficie en el partido a evaluar con respecto al torneo por el que se sacará el ranking
- Tipo de torneo e instancia alcanzada: Se evalúa el tipo de torneo y la instancia de ese torneo en que se jugó el partido al igual que lo hace ATP.

Como desarrollo de este modelo primero se analizará la antigüedad de años y torneos que miraremos hacia atrás para poder luego evaluar cada parámetro individualmente y encontrar el mejor valor para generar una combinación de ellos. De esta forma lograremos crear un PageRank paramétrico más eficaz que el ranking ATP y el PageRank ya existente.

Finalmente el peso de cada arco quedará indicado bajo la siguiente formula:

$$PESO_{Arco} = PESO_{Envejecimiento} * PESO_{Superficie} * PESO_{Instancia}$$

4.3.1. Antigüedad

¿Es correcto tomar únicamente 52 semanas como referencia para generar el ranking?

Como fue mencionado reiteradas veces, el ranking ATP se genera en base a la sumatoria de puntos conseguidos en los torneos que se desarrollaron en las últimas 52 semanas.

Para investigar si esto era correcto o se podía conseguir mejores resultados si miráramos mayor antigüedad de los partidos, evaluamos distintos parámetros de antigüedad para saber cuántos años hacia atrás es recomendable considerar para poder conseguir mejores resultados.

Para ello, cada cálculo de los distintos parámetros que serán tenidos en cuenta en el peso del arco, se evaluará teniendo en cuenta múltiples años de antigüedad y eligiendo el de mejor eficacia.

4.3.2. Envejecimiento

¿Es correcto evaluar todos los partidos del historial con el mismo peso?

A diferencia del ranking ATP o del PageRank generado en el capítulo anterior, nosotros estaremos evaluando muchos años para atrás para poder conseguir un ranking más eficaz que los anteriores.
Sin embargo, consideramos que no es correcto evaluar con el mismo peso a partidos que se jugaron hace mucho tiempo comparado a los enfrentamientos más recientes.
Es por esto que tomaremos el factor de envejecimiento como uno de nuestros parámetros. Para ello evaluaremos de qué manera se calculará un ranking que le de menor peso a los partidos más lejanos y que le de un mayor peso a aquellos que se jugaron últimamente.

Exponential Decay

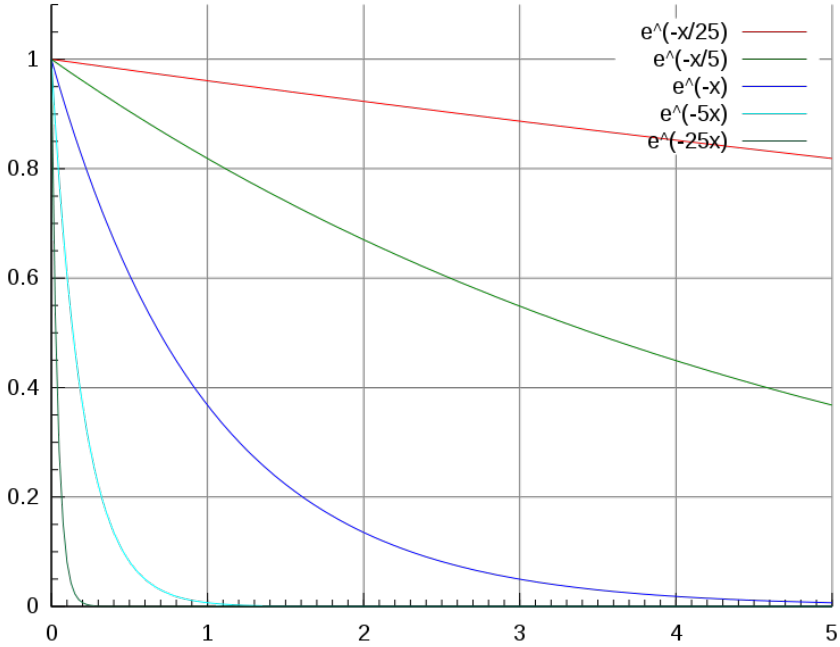
Para representar el factor de envejecimiento de la mejor manera, encontramos en el proceso de Exponential Decay una forma de expresar el peso, en el que los partidos más antiguos tenían menor peso que los nuevos.

La fórmula de Exponential Decay está representada de la forma:

$$N(t) = N_0 e^{-\lambda t}.$$

Donde $N(t)$ es el peso que se le asignará al partido representado en el tiempo t, que indica la diferencia de tiempo entre el torneo para el que se está generando el ranking comparado con el torneo al que se le están mirando los resultados.

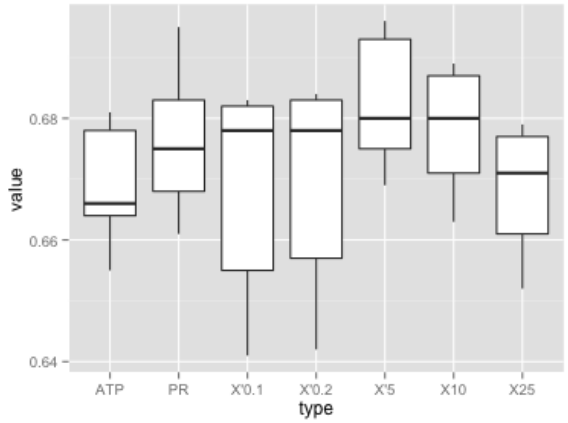
El modelo de Exponential Decay tiene como particularidad su rápido decrecimiento, logrando así valores muy bajos para aquellos partidos más antiguos.



Modificando el λ correspondiente se puede ajustar la velocidad de caída.

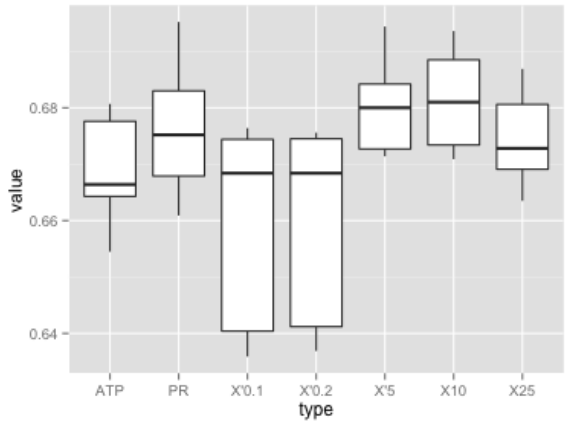
Considerando esto, evaluamos cuál es el mejor valor de λ y lo evaluamos con el nuevo sistema de ranking generado.

Para ello tomaremos como antigüedad de los torneos 3 y 5 años, es decir que el grafo estará compuesto por nodos y arcos correspondientes a partidos con esa antigüedad.



			ANTIGUEDAD DE 3 AÑOS				
AÑO	**ATP**	**PAGERANK**	0.1	0.2	5	10	25
PROMEDIO TOTAL	**66.849%**	**67.577%**	**66.820%**	**66.913%**	**68.237%**	**67.824%**	**66.836%**

Comparación del ranking ATP y el PageRank generado, comparado con el nuevo modelo tomando 3 años de antigüedad



			ANTIGUEDAD DE 5 AÑOS				
AÑO	**ATP**	**PAGERANK**	0.1	0.2	5	10	25
PROMEDIO TOTAL	**66.849%**	**67.577%**	**65.849%**	**65.899%**	**68.008%**	**68.150%**	**67.443%**

Comparación del ranking ATP y el PageRank generado, comparado con el nuevo modelo tomando 5 años de antigüedad

Como se puede observar, si fijáramos una antigüedad de 3 años con un valor de λ en el Exponential Decay de -5, conseguimos una mejora como ranking predictivo de 1,3 % con respecto al ranking ATP y de un 0,8 % con respecto al PageRank en la que todos sus arcos tienen pesos iguales.

4.3.3. Tipo de torneo e instancia alcanzada

¿Es lo mismo ganar partido de una primera fase que una final? ¿Y jugar un ATP 250 que un Grand Slam?

Como sabemos el tenis es un deporte de alta competitividad individual. Cada partido está cargado de presión y la mentalidad no es la misma cuanto más importante es el partido.
Es por esto que no consideramos correcto asignarle el mismo peso a partidos de distintos torneos y dentro de los mismos a distintas instancias.

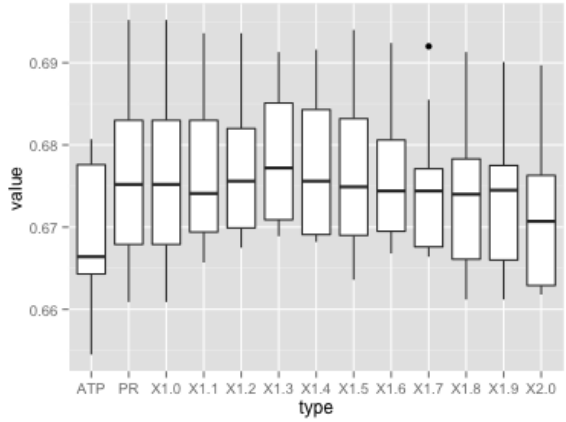
De la misma forma que lo hace la ATP, trataremos de generar el peso del arco del grafo en base a la instancia del torneo en la que se está disputando el partido evaluado para generar el PageRank.

Para ello utilizaremos una fórmula tal que se asigne una cantidad de puntos parecida a la que entrega la ATP. Esta formula es:

$$Peso_{instancia} = \frac{2000/\lambda^{Valor_{instancia}-1}}{2000}$$

Donde el valor de la instancia alcanzada en ese torneo es el valor numérico que representa el número de ronda alcanzada dependiendo del tipo de torneo.

Buscaremos entonces el λ entre 1 y 2 que mejor logre comportarse a modo de ranking predictivo y al mismo tiempo se acerque al ranking ATP.



AÑO	ATP	PAGERANK	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
PROMEDIO TOTAL	66.849%	67.577%	67.578%	67.639%	67.716%	67.763%	67.687%	67.573%	67.536%	67.482%	67.369%	67.341%	67.249%

Evaluación de λ correspondiente para puntuar la instancia del torneo alcanzada

Se puede observar como el λ en 1.3 logra estar apenas por encima de lo que logra predecir el PageRank existente aunque con una varianza mucho menor. Por otro lado si logramos una diferencia de casi 1 punto frente al ranking ATP.

4.3.4. Superficie

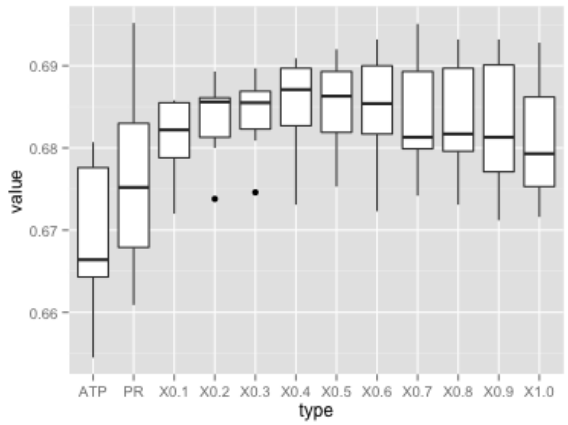
¿Cuanto influye en el historial los partidos jugando en las mismas superficies contra los de distintas superficies?

Existen tres tipos de superficies en el circuito de la ATP: *Dura, Polvo de ladrillo, Pasto*. Cada una de esas superficies tiene distintas características.

En polvo de ladrillo por ejemplo, la pelota tiene una velocidad más lenta y es más fácil deslizarse por la cancha. En cemento, la pelota cuenta con una velocidad más rápida y con piques más pronunciados, mientras que en pasto, además de tener una gran velocidad, la pelota no suele tomar altura.

Ante estas características, hay jugadores que saben desarrollar mejor su juego en algunas superficies, mientras que en otras no logran una buena performance. Es ante esta característica que surgió la idea de diferenciar el peso de cada arco que representa un partido del historial, en base a la superficie de ese partido con respecto a la superficie del torneo que buscamos predecir.

Generamos una matriz en la que a los arcos correspondientes a partidos jugados en la misma superficie que el torneo a predecir se les asigna como peso el valor más alto que es 1(uno), mientras a los arcos correspondientes a partidos jugados en otra superficie que la del torneo a predecir se les asigna como peso un valor que calcularemos.



AÑO	ATP	PAGERANK	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
PROMEDIO TOTAL	66.849%	67.577%	68.141%	68.394%	68.424%	68.520%	68.557%	68.508%	68.393%	68.347%	68.242%	68.105%

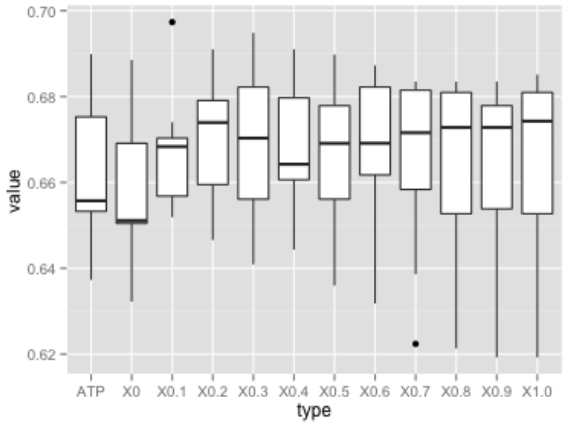
Evaluación de superficies bajo distintos valores de parámetros ante superficies distintas

Como se puede observar, evaluando el resto de las superficies con un parámetro de 0.5 se logra una eficacia de 1,7 % mayor con respecto al ranking ATP mientras que al PageRank estático logramos superarlo en casi 1 %.

Ya sabiendo estos resultados, evaluamos los resultados diferenciando en superficie. Cabe destacar que la mayor cantidad de partidos de la temporada se desarrollan en cancha dura, seguido luego de polvo de ladrillo, mientras que en pasto se juega el circuito

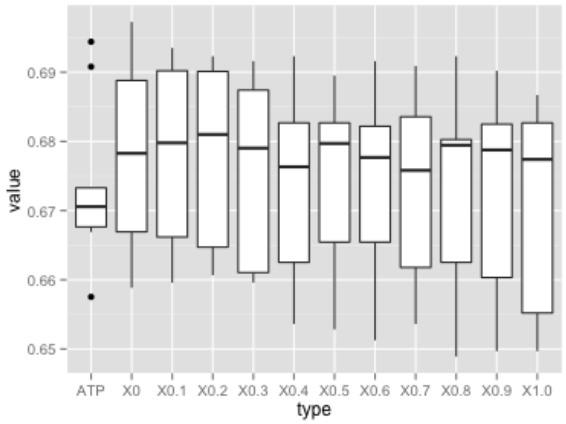
solo unas pocas semanas durante cada temporada.

Podemos observar que si bien cada superficie se comporta mejor ante un parámetro distinto, en cada una de ellas con la mejor elección nuestro modelo se comporta mejor que el ranking ATP en cualquiera de las superficies.



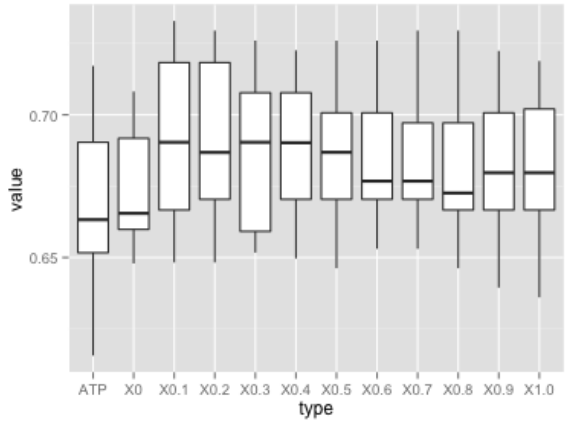
CLAY												
	ATP	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
PROMEDIO TOTAL	66.083%	65.698%	66.709%	66.925%	66.867%	66.711%	66.557%	66.650%	66.432%	66.265%	66.180%	66.273%

Evaluación de la superficie polvo de ladrillo bajo distintos valores de parámetros ante superficies distintas



HARD												
	ATP	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
PROMEDIO TOTAL	67.353%	67.861%	67.869%	67.814%	67.657%	67.430%	67.439%	67.298%	67.239%	67.246%	67.150%	67.053%

Evaluación de la superficie dura bajo distintos valores de parámetros ante superficies distintas



GRASS												
	ATP	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
PROMEDIO TOTAL	67.0181%	67.4533%	69.3553%	68.9427%	68.8134%	68.9709%	68.5859%	68.3960%	68.3967%	68.2814%	68.2428%	68.1278%

Evaluación de la superficie pasto bajo distintos valores de parámetros ante superficies distintas

4.3.5. Combinación de parámetros

Ya con la certeza de que nuestro modelo se comporta mejor que los anteriormente mencionados, buscamos la mejor combinación de parámetros para obtener un resultado más eficiente que cuando miramos los atributos particularmente.

Para ello combinaremos los parámetros y utilizaremos la multiplicación de los mismos como peso del arco dentro del multigrafo dirigido formado del que se calculará el PageRank.

Como primera intuición probamos la combinación de los mejores parámetros que obtuvimos particularmente. Es decir, tomamos una antigüedad de 5 años con Exponential Decay de -5 como λ, para el cálculo de la instancia del torneo alcanzada usaremos 1.3, y como superficie 0.5.

Como resultado de esta combinación obtenemos:

AÑO	ATP	PAGERANK	PAGERANK PARAMÉTRICO
2005	66.430%	66.785%	68.632%
2006	66.494%	66.087%	67.530%
2007	65.455%	66.939%	68.163%
2008	66.990%	66.563%	68.191%
2009	68.072%	69.516%	69.204%
2010	66.641%	68.472%	68.784%
2011	67.758%	67.524%	69.162%
2012	67.908%	68.299%	68.926%
2013	65.892%	68.010%	68.419%
PROMEDIO TOTAL	**66.849%**	**67.577%**	**68.557%**

Evaluación con el mejor parámetro calculado individualmente

A pesar de obtener un mejor resultado, optamos por buscar parámetro a parámetro cuál es la mejor combinación posible para obtener todavía un resultado mejor.

Para ello armamos un algoritmo que hará una búsqueda greedy en la que fijará dentro de un vector tres de los cuatro valores que estamos iterando y el otro lo iteraremos hasta encontrar el mejor resultado, luego iremos cambiando el vector fijando distintos valores. Esto lo haremos hasta que luego de varias vueltas, consigamos obtener el mejor resultado, es decir que luego de reintentar con todos los parámetros nuevamente no se modificó ninguno, por lo que nuestro vector es un óptimo local para el problema.

Generando un set de testeo, compuesto por los partidos de 90 torneos a razón de 10 por año, evaluamos los mejores parámetros, obteniendo como resultado que la combinación:

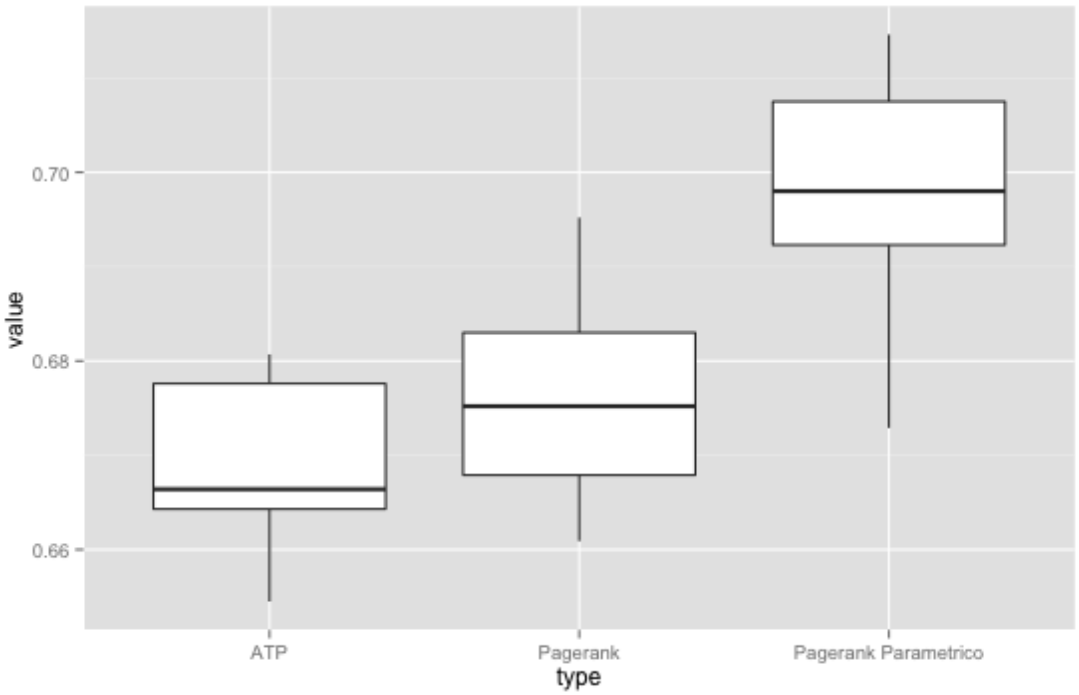
- Antigüedad: 4 años
- Envejecimiento: -5 como valor de Exponential Decay
- Superficie: 0,3 como valor para superficies distintas
- Torneo e instancia: 1,7 como valor de λ para darle un puntaje a la instancia del torneo jugado

Combinando estos valores obtuvimos:

AÑO	ATP	PAGERANK	PAGERANK PARAMÉTRICO
2005	66.430%	66.785%	70.755%
2006	66.494%	66.087%	69.143%
2007	65.455%	66.939%	69.767%
2008	66.990%	66.563%	67.290%
2009	68.072%	69.516%	70.190%
2010	66.641%	68.472%	71.464%
2011	67.758%	67.524%	69.231%
2012	67.908%	68.299%	69.799%
2013	65.892%	68.010%	70.915%
PROMEDIO TOTAL	**66.849%**	**67.577%**	**69.839%**

Evaluación con el mejor parámetro calculado luego de obtener los mejores resultados evaluados con el algoritmo

Como se puede ver, se consigue una aproximación al 70 % de aciertos en los torneos evaluados, logrando 3 puntos de mejora con respecto a lo que la ATP suele acertar en su ranking semanal y 2,2 con respecto al modelo existente de PageRank.



Boxplot comparativo de rankings predictivos, ATP, Pagerank y Pagerank paramétrico

En el gráfico de tipo *boxplot* podemos notar que lo que refiere a la mediana del Pagerank Paramétrico está por encima de lo que es el resto de los rankings estudiados. Se puede notar que el mínimo resultado obtenido con el modelo sugerido, alcanza la media del pagerank ya existente y supera al ranking ATP.

También se puede observar que en el modelo sugerido tal como muestra la tabla ronda el 70,0 % de aciertos, contando incluso con resultados por sobre esta medición.

Utilizando el test estadístico ANOVA (Analisis de la varianza) podemos confirmar que no se establece una diferencia considerable en la eficacia mostrada entre el ATP comparado al Pagerank existente (p = 0.14). Mientras tanto vemos una gran diferencia entre los resultados otorgados por el ranking ATP y el Pagerank Paramétrico (p = 0,000024) y también tiene una diferencia importante con respecto al Pagerank ya existente (p = 0.00079)

4.3.6. Resultados de parámetros desglosados

Ante el hallazgo de que el modelo planteado lograba un carácter predictivo mejor que el resto de los modelos existentes, comenzamos a buscar si este se comportaba de manera similar en distintas situaciones planteadas existentes en el marco de los torneos y partidos de tenis.

Es por esto que hicimos un desglose de los partidos en base a superficie, ranking entre jugadores y tipo de torneo para encontrar en qué marcos nuestro modelo predecía mejor y en cuales no lo hacía de la mejor manera.

- Superficie: En estos casos entrenamos con todas las superficies pero solo testeamos sobre una sola superficie en particular

HARD			
AÑO	**ATP**	**Pagerank**	**Pagerank Paramétrico**
2005	66.764%	66.178%	68.452%
2006	67.057%	65.886%	67.716%
2007	66.768%	67.525%	69.114%
2008	65.752%	65.752%	66.850%
2009	69.441%	69.930%	70.350%
2010	67.296%	68.577%	68.847%
2011	67.329%	66.918%	69.247%
2012	69.080%	69.012%	69.621%
2013	66.691%	68.769%	68.398%
Promedio general	67.353%	67.616%	68.733%

Comparación por cancha dura

CLAY			
AÑO	**ATP**	**Pagerank**	**Pagerank Paramétrico**
2005	64.749%	66.736%	67.678%
2006	65.568%	64.659%	65.682%
2007	64.253%	67.647%	67.647%
2008	68.990%	67.548%	68.510%
2009	65.806%	69.250%	68.635%
2010	65.574%	68.096%	70.366%
2011	67.528%	68.020%	68.881%
2012	67.970%	67.970%	68.222%
2013	65.328%	67.153%	67.762%
Promedio general	66.196%	67.453%	68.153%

Comparación por polvo de ladrillo

GRASS			
AÑO	ATP	Pagerank	Pagerank Paramétrico
2005	71.717%	71.044%	72.391%
2006	65.169%	70.037%	69.288%
2007	66.327%	64.626%	65.986%
2008	69.039%	70.819%	71.886%
2009	68.858%	70.242%	70.588%
2010	66.207%	68.966%	66.897%
2011	70.548%	69.178%	69.521%
2012	61.566%	65.480%	67.616%
2013	63.732%	66.901%	67.254%
Promedio general	67.018%	68.588%	69.047%

Comparación por pasto

Como se puede observar, nuestro modelo logra obtener mejores resultados en todas las superficies, logrando una mayor posibilidad de predicción en aquellos partidos jugados en grass donde se acerca al promedio general del modelo y logrando 2 % más que el ranking ATP.

Se puede observar también que tanto en cancha dura como en polvo de ladrillo, nuestro modelo es el de mayor cantidad de aciertos aunque la diferencia con los otros modelos es menor.

- Ranking: En estos casos evaluaremos los partidos entre jugadores de ranking similar y que solo pertenecen a ese grupo de ranking delimitado.

1 a 10			
AÑO	ATP	Pagerank	Pagerank Parametrico
2005	70.270%	78.049%	76.60%
2006	68.085%	71.698%	71.11%
2007	66.667%	69.565%	68.42%
2008	60.345%	64.815%	64.91%
2009	63.529%	66.250%	59.15%
2010	63.462%	62.963%	68.25%
2011	67.568%	65.217%	69.62%
2012	76.471%	73.913%	78.16%
2013	58.974%	71.642%	67.11%
Promedio general	66.152%	69.346%	69.260%

Comparación partidos entre top ten

11 a 50			
AÑO	**ATP**	**Pagerank**	**Pagerank Parametrico**
2005	60.172%	55.655%	58.69%
2006	61.165%	53.041%	58.36%
2007	57.547%	59.281%	60.57%
2008	55.963%	50.485%	56.13%
2009	62.162%	64.151%	60.47%
2010	55.799%	57.049%	57.05%
2011	55.848%	57.317%	56.16%
2012	58.333%	55.128%	58.28%
2013	53.333%	55.455%	58.07%
Promedio general	57.814%	56.396%	58.200%

Comparación partidos del ranking 11 al 50

50 en Adelante			
AÑO	**ATP**	**Pagerank**	**Pagerank Parametrico**
2005	58.100%	59.538%	59.848%
2006	59.091%	58.142%	59.783%
2007	56.364%	60.358%	58.978%
2008	58.534%	59.463%	58.518%
2009	56.377%	56.364%	58.177%
2010	56.855%	60.156%	60.614%
2011	60.931%	59.814%	61.485%
2012	56.831%	58.413%	57.517%
2013	56.825%	60.699%	58.742%
Promedio general	57.767%	59.216%	59.296%

Comparación partidos del ranking 50 en adelante

En esta comparativa podemos observar que en los tres casos nuestro modelo supera al ranking ATP mientras que tiene una cantidad de aciertos similar a lo que respecta al modelo del Pagerank original.

En el caso de los partidos entre jugadores top ten logramos una diferencia mayor a 3 puntos con respecto a lo que es el ranking ATP. Lo que indica que para partidos entre los grandes jugadores donde la diferencia es muy menor ya que los jugadores involucrados son del top ten mundial, nuestro modelo logra, al igual que el Pagerank original, comportarse de una manera bastante mejor que lo que hace el ranking ATP.

Para el resto de lo casos el modelo creado logra una mejora para los dos modelos existentes, aunque la diferencia no logra ser abultada, nuestro ranking creado logra acertar con mayor certeza quien será el ganador de un partido o de un torneo si juegan jugadores entre esas posiciones.

- Tipo de torneo: En este caso evaluaremos los resultados según la importancia del torneo que se encuentra en disputa. Cabe destacar que los torneos más importantes cuentan con mayor cantidad de participantes y por ende una mayor cantidad de partidos para analizar.

ATP 250			
AÑO	ATP	Pagerank	Pagerank Parametrico
2005	67.180%	67.180%	67.488%
2006	65.468%	65.217%	66.304%
2007	63.365%	66.509%	66.116%
2008	67.475%	66.639%	67.726%
2009	64.677%	65.576%	66.067%
2010	64.446%	66.696%	67.215%
2011	66.725%	66.115%	67.509%
2012	62.839%	63.562%	64.467%
2013	62.285%	64.642%	64.823%
Promedio general	64.940%	65.793%	66.413%

Comparación torneos ATP 250

ATP 500			
AÑO	ATP	Pagerank	Pagerank Parametrico
2005	64.747%	65.060%	65.899%
2006	62.264%	66.667%	64.623%
2007	66.667%	63.904%	71.094%
2008	63.037%	64.419%	62.751%
2009	70.029%	72.245%	72.622%
2010	65.775%	67.286%	67.914%
2011	68.519%	64.880%	71.693%
2012	70.516%	69.944%	73.464%
2013	63.057%	69.131%	68.153%
Promedio general	66.068%	67.060%	68.690%

Comparación torneos ATP 500

MASTERS 1000			
AÑO	ATP	Pagerank	Pagerank Parametrico
2005	62.478%	64.977%	66.954%
2006	67.185%	61.792%	68.566%
2007	62.063%	67.448%	66.298%
2008	63.670%	61.318%	64.607%
2009	68.980%	70.317%	71.429%
2010	63.941%	66.578%	68.401%
2011	63.216%	69.841%	65.250%
2012	69.388%	72.482%	70.315%
2013	66.543%	65.605%	68.207%
Promedio general	65.274%	66.706%	67.781%

Comparación torneos Masters 1000

GRAND SLAM			
AÑO	ATP	Pagerank	Pagerank Parametrico
2005	71.047%	69.815%	73.511%
2006	71.837%	70.816%	70.612%
2007	72.765%	71.518%	73.597%
2008	72.279%	72.279%	74.127%
2009	74.743%	76.591%	74.949%
2010	74.948%	74.741%	74.741%
2011	74.948%	72.464%	75.362%
2012	74.845%	73.196%	73.402%
2013	75.260%	75.676%	75.052%
Promedio general	73.630%	73.011%	73.928%

Comparación torneos Grand Slam

MASTERS CUP			
AÑO	**ATP**	**Pagerank**	**Pagerank Parametrico**
2005	53.333%	53.333%	73.333%
2006	66.667%	80.000%	73.333%
2007	100.000%	53.333%	53.333%
2008	66.667%	73.333%	73.333%
2009	53.333%	53.333%	60.000%
2010	86.667%	93.333%	86.667%
2011	60.000%	53.333%	73.333%
2012	93.333%	86.667%	86.667%
2013	66.667%	80.000%	73.333%
Promedio general	71.852%	69.630%	72.593%

Comparación torneos Masters Cup

Como se puede observar, es importante destacar que en los cinco tipos de torneos analizados nuestro modelo logra una mejor predicción que el resto de los modelos analizados.

Se pueden observar excelentes resultados de aciertos en los torneos Grand Slam y en Masters Cup, alcanzando casi el 74 % en los primeros y logrando un $72,6\,\%$ en los segundos. Sin embargo, no es en estos torneos donde se logra una mayor diferencia en cuanto a los rankings existentes.

En los torneos ATP 500 y Masters 1000 se logra una diferencia con respecto al ATP de aproximadamente 2,5 %, lo que hace nuestro modelo mucho más confiable.

5. ESTIMACIÓN DE LA PROBABILIDAD DE VICTORIA

5.1. ¿Qué es la probabilidad de victoria?

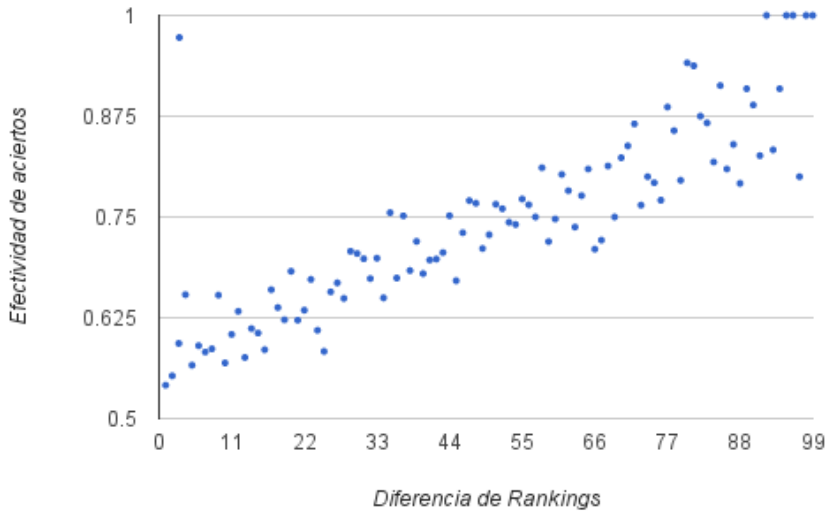
Como hemos visto, logramos crear un modelo capaz de predecir los resultados de un torneo dado con una efectividad mejor que lo que lo hace hoy en día el ranking ATP, y el modelo de Pagerank existente.

Ante este panorama surge la idea de no solo poder predecir el resultado de un partido basándonos en el ranking, sino además poder determinar con qué probabilidad el resultado del jugador de mejor ranking será victorioso frente a uno de peor ranking.

Es aquí donde surge la probabilidad de victoria, que determina, dados r_1 y r_2 (los rankings de los jugadores que se van a enfrentar) y considerando su diferencia, cuán firmes son las posibilidades de que el mejor posicionado triunfe.

5.2. ¿Cómo la calculamos?

Para poder encontrar esta probabilidad de victoria, primero determinaremos la eficacia que hemos tenido en en nuestro ranking si agrupáramos por la diferencia de ranking. Es decir, por cada partido, analizamos no sólo el resultado sino, considerando la diferencia de ranking entregado por nuestro modelo, cómo se comportó para ese delta.



Evaluación de eficacia de victorias en base a diferencia de ranking provisto por nuestro modelo

Analizando el gráfico obtenido, queremos encontrar la función que logre acercarse lo mejor posible a cada punto del gráfico, para ello utilizaremos un modelo de regresión que nos permitirá alcanzar el objetivo de la mejor manera.

Basandonos en el paper [15] en la que se trabaja con un esquema similar para selecciones de fútbol y considerando que nuestro gráfico se asemeja a una curva exponencial, buscaremos la exponencial que logra cubrir todos los puntos indicados.

Definiremos primero una función modelo que consideramos más apropiada en base a lo obtenido. Para ello usaremos una función logística que se caracteriza por representar el crecimiento de organismos desde un pequeño estado inicial, durante el cual el crecimiento es proporcional al tamaño, hasta la última etapa cuando el tamaño se aproxima a una asíntota.

Considerando que nuestro gráfico comienza en el punto inicial donde la diferencia de ranking entre los jugadores es mínima y luego va creciendo hasta el punto donde a medida que la diferencia del ranking va creciendo la cantidad de aciertos ya se considera permanente podemos indicar que tiene un comportamiento similar a una asíntota. Es por eso que podemos utilizar la funćion logística sugerida en nuestro caso.

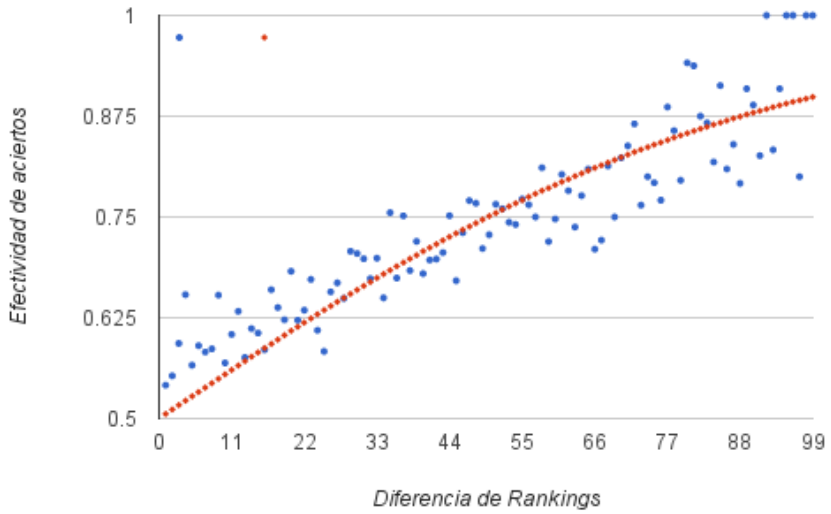
Ésta función es del tipo:

$$P_{Victory} = \frac{1}{1 + e^{\frac{r1-r2}{a}}}$$

Luego debemos cumplir los siguientes pasos:

1. Estimar los parámetros del modelo de regresión. Este proceso es llamado ajuste del modelo a los datos.
2. Chequear qué tan bueno es el modelo ajustado. El resultado de este chequeo puede indicar si el modelo es razonable o si el ajuste original debe ser modificado.

Cumpliendo estos pasos, podemos determinar a través del modelo de regresión que $a = 45.321$. En base a este parámetro obtenemos



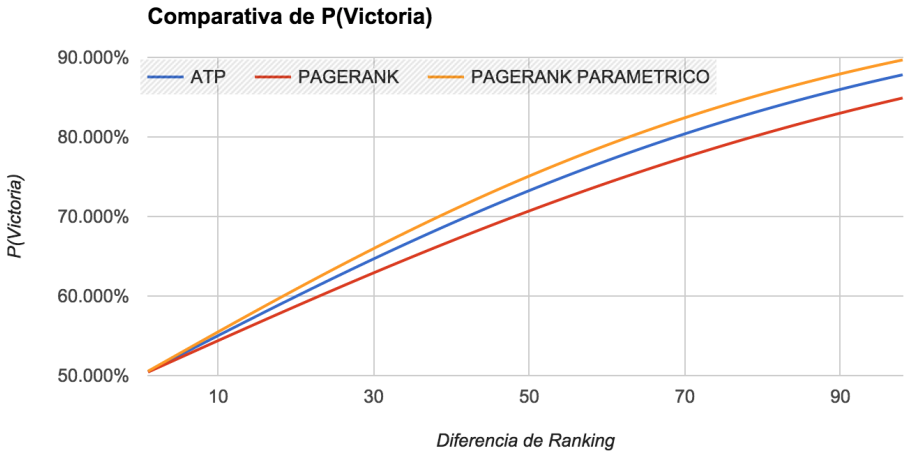
Evaluación de eficacia de victorias en base a diferencia de ranking y curva de regresión

Podemos indicar que siendo esta una curva razonable podemos pasar a evaluar como se comporta la misma en base a nuestro modelo.

5.3. Comparativa de la P(Victoria)

Ya conociendo la función que representa nuestra probabilidad de victoria, y reafirmando que la curva obtenida logra una forma coherente frente a los puntos marcados, nos enfocamos en poder comparar cuán optimista es nuestra probabilidad de victoria frente a la probabilidad de victoria que se puede obtener de la curva basada en el ranking ATP o en el modelo de Pagerank existente.

Para ello calculamos para el top 100 de los jugadores de cada modelo cuál es la probabilidad de victoria de cada diferencia.



Comparación de probabilidad de victoria de los modelos existentes

Se puede observar que el modelo sugerido logra en el marco de la probabilidad de victoria un carácter más optimista que el PageRank existente y mucho más aún que el ranking ATP.

Tal como muestra el gráfico, a partir de los 20 puestos de diferencia de ranking ya logra en nuestro modelo sugerido empezar a desprenderse del ranking ATP, logrando una probabilidad de victoria mayor al 60 % para aquel que tiene un mejor ranking.

5.3.1. Especificación en base a parámetros especificados

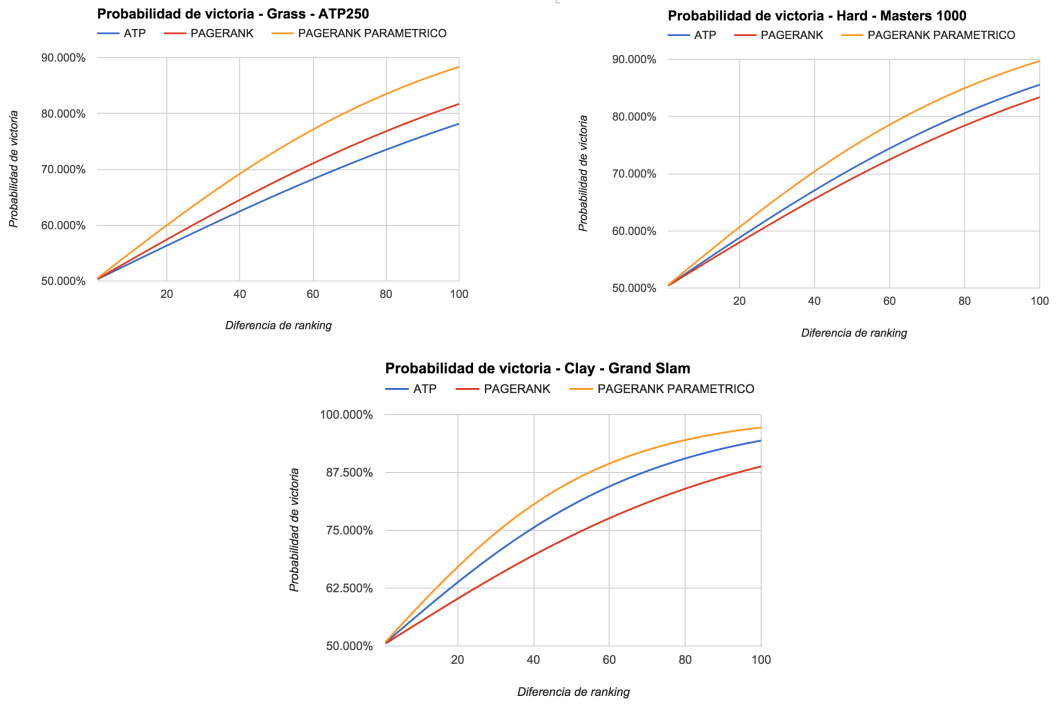
Como hemos visto, en el marco general nuestro ranking generado logra una probabilidad de victoria más optimista que el resto de los rankings. Podemos entonces buscar la probabilidad de victoria teniendo en cuenta los atributos que fueron evaluados al momento de generar el ranking.

Se puede en base a esto armar un árbol de decisión cuyas aristas pueden responder a las preguntas ¿En qué superficie se jugó?, ¿Qué tipo de torneo es el que se está evaluando?.

Utilizando la misma comparativa usada en la probabilidad generalizada, llegamos a las hojas de la combinación *Superficie - Tipo de torneo*, donde nuestro ranking sugerido logra ser más optimista que el resto de los rankings.

Poniendo como ejemplo algunos de los gráficos de las hojas obtenidas, se puede observar que a diferencia del ranking generalizado, el optimismo y la probabilidad de victoria de los distintos rankings logran hacerse más pronunciadas o menos dependiendo de los parámetros con los que se calculan.

Entonces, dado un partido en el que conocemos el posicionamiento de cada jugador en cada uno de los rankings evaluados, podemos determinar conociendo además el tipo de torneo en el que se juega el partido y su superficie, la probabilidad que tiene cada jugador de salir victorioso en ese partido.



A partir de esto, teniendo el cuenta los números vistos, si se toma como referencia el ranking del modelo sugerido, podemos conseguir ante menor diferencia de ranking, una apuesta más importante en la que se determina quién saldrá ganador del encuentro, ya que muestra ser de un carácter más optmista que los otros dos rankings vistos.

5.4. Evaluación

Ante la obtención de las probabilidades de victoria y ante la muestra de que nuestro modelo es más optimista que los modelos existentes, nos enfocamos en evaluar la perfomance en cada uno de ellos. Para ello utilizamos el método de AUROC (Area under ROC).

Como se indica en [16, 17], el análisis de curvas ROC constituye un método estadístico para determinar la exactitud diagnóstica de los tests realizados, siendo utilizadas con tres propósitos específicos:

- Determinar el punto de corte de una escala continua en el que se alcanza la sensibilidad y especificidad más alta.
- Evaluar la capacidad discriminativa del test diagnóstico, es decir, su capacidad de diferenciar con qué probabilidad ganará un partido un jugador con una diferencia de ranking dada.
- Comparar la capacidad discriminativa de los tests diagnósticos que expresan sus resultados como escalas continuas.

El área bajo la curva ROC es un excelente indicador global del desempeño de una prueba diagnóstica ya que hace factible expresarlo en un número simple.

Evaluando la $P(victoria)$ para los 3 modelos obtenemos:

ATP	PAGERANK	PAGERANK PARAMETRICO
0.7040	0.6776	0.7137

Comparación de AUROC

Como se puede observar, nuestro modelo es el de mejor performance, sin embargo, a diferencia de la eficacia mostrada en el capítulo anterior, la diferencia con el ranking ATP es menor, mientras que la diferencia con el Pagerank es considerable.

Si calculamos el AUROC haciendo que cada partido vaya por su rama correspondiente en el árbol de decisión, encontramos una mejora, aunque la misma no es muy grande.

ATP	PAGERANK	PAGERANK PARAMETRICO
0.7070	0.6812	0.7168

Comparación de AUROC

La utilización de la $P(victoria)$ nos permite ampliar el abanico de uso de nuestro ranking. De esta manera es posible fijar un umbral de satisfacción acorde a nuestras expectativas de uso. En el “Trabajo Futuro” ampliaremos este aspecto.

6. TRABAJO A FUTURO

Como hemos visto, los rankings nos rodean en cada una de nuestras actividades del día a día. A partir de este trabajo hemos logrado introducir un nuevo sistema de posicionamiento en un ámbito deportivo como lo es el tenis, sin embargo el hecho de haber creado un nuevo sistema y animarnos además a intentar predecir resultados de tenis nos abre las puertas a muchos trabajos que se pueden hacer a futuro.

- Aplicación de otros rankings conocidos al ambiente del tenis.

  Como se puede leer en el libro Who's the #1 [5], se han desarrollado a lo largo de la historia muchos modelos de generación de rankings[14], utilizados hoy en día para clasificar equipos en varios deportes. Estos métodos, como el Massey, el Colley, el Keener, el Elo, entre otros, son utilizados mayoritariamente en la generación de estadísticas para la NFL o el Ajedrez, pero no para el tenis. Si bien a través de nuestro modelo se fueron utilizando diferentes técnicas que utilizan sus métodos, ya que el PageRank en si las incluye, creemos que se puede trabajar aplicando puramente estos métodos para lograr distintos rankings de tenis y así poder descubrir cuál es el ranking que mejor logra clasificar a los jugadores en la actualidad.

- Aplicación de nuestro método de PageRank paramétrico en distintos deportes.

  El método generado demostró comportarse mejor que el ranking ATP con el que se clasifica a los jugadores de tenis hoy en día. Ante estos resultados favorables se abre la puerta para poder aplicar este sistema de clasificación en otros deportes donde los resultados son de Ganar/Perder y nos permitan la generación de un PageRank en el que se puedan dirigir arcos con distintos pesos evaluando las características del partido jugado.

  Es así entonces que en cualquier deporte de paleta se puede aplicar el ranking, como en ligas particulares como los son la NFL o la NBA.

  En cuanto a deportes como el fútbol, se puede aplicar siempre y cuando se evalúe para una liga nacional, pero no a nivel internacional. Este ranking, de la forma en que está creado, no permite saber en estos deportes quién sería el mejor equipo del mundo, ya que los partidos entre equipos de interligas son muy pocos para contemplar ese resultado.

- Aplicación de nuestro método de PageRank paramétrico en distintos ambientes.

  Saliendo del marco deportivo, se puede además buscar la forma de aplicación de rankings en otros ambientes como la moda o la cocina o en todos aquellos rubros donde a lo largo del año se hagan concursos o torneos donde se pueda clasificar a los participantes.

  Por ejemplo en los concursos de cocina que se realizan se genera un grafo tal como si fuese un torneo de tenis y luego a lo largo del año se van acumulando esos torneos para poder lograr indicar quien es el mejor chef, aplicando en los arcos el peso del

tópico del concurso que se realiza. Misma aplicación se puede hacer con diseñadores o incluso con jugadores de videojuegos.

- Herramienta para ayuda en apuestas.

  Ante los resultados obtenidos surge una gran pregunta que es ¿Se puede utilizar la clasificación generada para apostar?

  Si bien a simple vista parece que los números obtenidos son buenos e incluso, con la utilización de la P(Victoria), podemos ante un partido indicar cuál es la chance concreta de que gane un jugador ante otro, la tesis realizada abre la puerta a un análisis más concreto sobre el tema.

  Creemos que se puede aplicar la creación de un asistente de apuestas en el que se indique qué apuestas son las más confiables para realizar y cuáles entregarán mayor dinero. Teniendo la probabilidad de victoria ante cada partido, podemos indicar un valor mínimo de probabilidad por el que es recomendable hacer una apuesta.

  También se puede intentar indicar a largo plazo los ganadores de los torneos para poder así obtener mayor ganancia ante una apuesta en caso de salir victorioso.

  Otra de las funciones de esta herramienta puede ser la simulación de cuánto dinero obtendríamos poniendo un monto inicial y apostando siempre al ganador que indica nuestro ranking en base a la clasificación. Esto se puede hacer utilizando nuestra base de datos ya que contamos con los valores de las apuestas en el historial del partido.

- Planificación de temporada de los jugadores.

  A partir de la probabilidad de victoria y del modelo generado, se puede trabajar en la planificación de la temporada de un jugador, encontrando qué torneos son los recomendables para presentarse y llegar mejor preparado, priorizando los que potencialmente le darán una mejora radical en el ranking.

  Se puede hacer esto mismo por partido, incluso yendo a jugar sabiendo cuáles son las chances concretas (teóricas) de victoria.

  Aprovechando además que nuestro ranking se puede diferenciar fácilmente por superficie, tipo de torneo e instancia, se pueden conocer las fortalezas y debilidades de los jugadores y ver como se comportan los mismos ante un partido con estos atributos, lo que permite a los entrenadores volcar los trabajos de entrenamiento para crecer en esos aspectos y lograr mejoras sustanciales en la performance de sus dirigidos.

- Rachas

  Un factor que se le puede agregar al modelo con el que se genera nuestro rankings es la contemplación de rachas en los jugadores. Ocurre en todas las temporadas que hay momentos en el año donde hay jugadores que inician una serie de victorias o derrotas seguidas que los hacen crecer o caer mucho en el ranking. Esto es algo que no suele comportarse bien a futuro ya que muchas veces son momentos aislados en las carreras de los deportistas y no logran reflejar realmente su nivel

  Por lo tanto, se podría hacer que evaluando las rachas al momento de generar el ranking, los partidos en que estos jugadores salieron victoriosos o derrotados debido a una racha tengan un peso distinto que permita que nuestro modelo no se confunda y logre entregarnos un ranking que realmente refleje a nivel clasificatorio quién es el mejor jugador para ese momento de manera que los resultados de los partidos reflejen la clasificación otorgada.

7. CONCLUSIONES

Ante la idea de poder lograr un ranking social, comprendido por todos aquellos seguidores del tenis, el ranking ATP se expone a quejas constantes de los jugadores y al mismo tiempo expone a nuevos tenistas a ser beneficiados con un buen torneo para poder comenzar a progresar en sus carreras.

Al mismo tiempo, analizando los resultados obtenidos, se puede observar que el ranking ATP no es lo suficientemente poderoso para lograr predecir con certezas quién será el vencedor de un partido en caso en que nos basáramos únicamente en las posiciones.

Con el fin de combatir estos problemas surge la idea de la creación de un nuevo ranking que logre indicar cuáles son las chances reales de victoria de un jugador ante el comienzo de un nuevo torneo. Intentamos generar un nuevo ranking que se destaca por la utilización de las características del torneo para la generación del mismo.

En primera instancia implementamos un ranking que utiliza el algoritmo de ordenamiento de Google, llamado PageRank, para poder lograr un ordenamiento de posiciones de los jugadores en la que se le da más importancia a los partidos jugados por los jugadores entre sí que a la instancia del torneo que fue alcanzada por un jugador.

Este tipo de ranking además, por la característica del algoritmo utilizado, permite darle un mayor rédito a aquellos jugadores que lograron ganarle a los mejores jugadores del torneo y del circuito.

Sin embargo, al momento de comparar su capacidad de predicción, lograba ser apenas mejor que el actual, pero no lo suficientemente como para afirmar que es un mejor predictor.

Fue entonces que se nos ocurrió implementar un ranking similar a este último pero donde no solo se miraría el resultado de los partidos sino que se podría además evaluar otros atributos que estos mostraban. Siendo uno de ellos el que tomaba en cuenta el ranking ATP. Creamos así el modelo del PageRank Paramétrico donde hacemos una combinación de los rankings ya existentes y además miramos otros factores para evaluarlo.

Este nuevo modelo se comporta como el algoritmo del PageRank, pero se le agrega a cada arco un peso distinto, basándonos en la superficie, instancia del torneo, tipo de torneo, y antigüedad de los historiales de los partidos evaluados.

Como se pudo observar, hemos obtenido grandes mejoras con respecto a los modelos existentes. Nuestro modelo logra predecir los resultados de los partidos y torneos con respecto al ATP con un porcentaje de mejora que ronda el 3 %.

Nuestro modelo además logró una mejor capacidad de predecir ante cualquier desglose que hemos realizado, ya sea por partidos entre jugadores de ranking similar, tipo de torneo, o tipo de superficie.

Una vez lograda una mejora en el algoritmo empleado, nos dimos lugar a entregar más información con respecto a la posibilidad de predecir un resultado. Es por eso que integramos a nuestro análisis la probabilidad de victoria. Esta probabilidad nos indica, dado un partido y evaluando la diferencia del ranking entre los jugadores que intervienen en el mismo, con qué porcentaje nuestro modelo puede predecir que ganará el mejor posicionado.

Para ello, basándonos en un modelo de regresión logística, hemos comparado cómo se comportaba nuestro modelo frente a los otros modelos existentes. Esta comparación nuevamente dió como más optimista a nuestro modelo por sobre el resto. Hacía falta entonces un cálculo de certezas para ver que nuestro optimismo no indique algo que no era certero. Utilizando la medida AUROC pudimos observar que nuestro modelo, además de ser más optimista, tiene un alto margen de certeza en lo que indica.

Dado todos estos resultados, podemos afirmar que, a lo largo de este trabajo, hemos creado un nuevo ranking de tenis robusto, capaz de predecir con un alto porcentaje el resultado de los partidos, y ante mayor detalle del mismo, como la diferencia del ranking, poder decir con qué certeza vamos a lograr intuir quién será el ganador del encuentro.

Bibliografía

[1] James B. *El ranking que cambió el tenis.* Obtenido de `http://www.atpworldtour.com/news`

[2] ATP World Tour *Emirates ATP Rankings FAQ.* Obtenido de `http://www.atpworldtour.com/en/rankings/rankings-faq`

[3] ATP World Tour *About the ATP Challenger* Obtenido de `http://www.atpworldtour.com/en/corporate/history`

[4] Bryant H. *How to fix tennis' big problems* 2013. Obtenido de `http://espn.go.com/tennis`

[5] LLangville, Amy N., and Carl D. Meyer. *Who is #1?: the science of rating and ranking.* Princeton University Press, 2012.

[6] Brin, Sergey, and Lawrence Page. *Reprint of: The anatomy of a large-scale hypertextual web search engine.* Computer networks 56.18 (2012): 3825-3833.

[7] Ashish G. *Applications of PageRank to Recommendation Systems.* Obtenido de `http://web.stanford.edu/class/msande233/handouts/lecture8.pdf`

[8] Page L., Brin S., Motwani R. and Winograd T *The PageRank Citation Ranking: Bringing Order to the Web* 7th International World Wide Web Conference, Brisbane, Australia, 1998.

[9] Irons, David J., Stephen Buckley, and Tim Paulden. *Developing an improved tennis ranking system.* Journal of Quantitative Analysis in Sports 10.2 (2014): 109-118.

[10] Radicchi, Filippo, and Matjaz Perc. *Who is the best player ever? A complex network analysis of the history of professional tennis.* PloS one 6.2 (2011): e17249.

[11] Dingle, Nicholas, William Knottenbelt, and Demetris Spanias. *On the (page) ranking of professional tennis players.* Computer Performance Engineering. Springer Berlin Heidelberg, 2013. 237-247.

[12] Spanias, A. Demetris, and B. William Knottenbelt. *Tennis Player Ranking using Quantitative Models.* Manuscrito

[13] Blackburn, McKinley L. *Ranking the performance of tennis players: an application to womens professional tennis.* Journal of Quantitative Analysis in Sports 9.4 (2013): 367-378.

[14] Barrow, Daniel, et al. *Ranking rankings: an empirical comparison of the predictive power of sports ranking methods.* Journal of Quantitative Analysis in Sports 9.2 (2013): 187-202.

[15] Dormagen, David. *Development of a Simulator for the FIFA World Cup 2014.* Bachelorarbeit FU Berlin 13 (2014). Manuscrito

[16] Anagnostopoulos C., Hand D.J., Adams N.M *Measuring classification performance: the hmeasure package* Department of Mathematics, South Kensington Campus, Imperial College London 2012

[17] Fawcett, Tom. *ROC graphs: Notes and practical considerations for researchers.* Machine learning 31 (2004): 1-38.

[18] Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets.* Cambridge University Press, 2014.

7. APÉNDICE

7.1. Base de datos

Como hemos mencionado en la sección en que hablamos del set de datos, contamos con archivos que contenían información de aproximadamente 40000 partidos.

Una vez obtenidos estos archivos, no fue fácil la tarea de interpretar lo que los archivos *.csv* entregaban. Los archivos tenían formatos distintos entre si, de algunos años contábamos con más información y de algunos otros no contábamos con toda la información requerida.

Como primera instancia teníamos que sanitizar los nombres de los jugadores, ya que había por ejemplo partidos en los que jugaba Juan Martin Del Potro y otro donde jugaba J.M Del Potro.

Para ello hicimos un script que se encargaba de evaluar aquellos jugadores con un solo partido y evaluar si había un jugador con nombre similar ante ese nombre o simplemente eran jugadores con un solo partido ATP en esos 10 años evaluados.

Misma condición ocurrió con los tipos de torneos, y sus nombres, ya que a lo largo de los años los tipos de torneos se fueron renombrando. Por ejemplo los ATP 250 y ATP 500 antes tenían nombres de International Series. Toda esta información se necesitó incorporar para luego tener un set de datos confiable.

Otras de las complicaciones encontradas es que el orden de la información brindada por cada año era distinta. En algunos archivos teníamos el resultado en una columna mientras que en un año diferente esta información estaba en otra, es por esto que se reordenó cada archivo para poder, a través de un script, interpretar de manera similar cada uno de los archivos.

Una vez lograda una sanitización completa de los datos obtenidos nos dispusimos a distribuir la información en 3 tablas.

Para ello, recorriendo en cada archivo obtenido un partido por línea, generamos tres tablas que contenían la información necesaria.

- Jugadores: Jugador_ID - Nombre del jugador
- Torneos: Torneo_ID - Nombre - Año - Semana - Superficie - Cantidad de sets - Tipo de torneo - Lugar - Techo
- Partidos: Jugador_ID - Ganador_ID - Perdedor_ID - Ranking ATP Ganador - Ranking ATP Perdedor - WSet1 - LSet1 - WSet2 - LSet2 - WSet3 - LSet3 - WSet4 - LSet4 - WSet5 - LSet5 - Sets Ganador - Sets Perdedor - Partido completo - Pago apuesta ganador - Pago apuesta perdedor

Era importante ante cada jugador evaluar que no exista ya en nuestra base y poder asociar cada jugador de los torneos a nuestra base. Al mismo tiempo en la data obtenida contábamos con la fecha exacta del partido por lo que el script que insertaba los torneos calculaba, dada una fecha, a qué semana y año correspondía.

Una complicación extra que tuvimos que sortear fue que no todos los partidos contaban con la información del ranking ATP, por lo que generamos un script que crawleaba el site oficial de la ATP (www.atpworldtour.com) para obtener el ranking correspondiente a esos jugadores en esos partidos.

7.2. Implementación del nuevo modelo de PageRank

7.2.1. Pagerank Paramétrico

Para poder realizar el cálculo del PageRank paramétrico, hemos desarrollado en Python los scripts correspondientes que nos permiten calcular el Pagerank dado un set de torneos dado.

El proceso de cálculo fue el siguiente:

- Generación de grafo por torneo

  Para generar un ranking para un torneo en particular, nuestro script mira todos los torneos con una antigüedad máxima de una cantidad de años determinada. Por cada uno de estos torneos, miramos cada partido que se jugó y trazamos un arco dirigido desde el perdedor al ganador. Para cada arco trazado previamente se le calcula el peso que se le dará a ese partido. Para ello, tal como fue explicado, se calcula cuánto suma ese partido en cuanto a antigüedad, en cuanto a superficie, tipo, e instancia de torneo, comparado con el torneo original para el que se quiere obtener el ranking.

  Una vez evaluado el torneo, cada grafo se une a un multigrafo dirigido.

- Obtención de un PageRank por torneo y generación de un ranking

  Una vez obtenido el multigrafo con todos los partidos que se evalúan para el torneo, a través de la librería, se utiliza el método *pagerank_scipy* que nos otorga un puntaje por cada vértice del multigráfo. Ese puntaje es el que, según sus resultados, recibe un jugador en la clasificación otorgada.

  Logrados estos puntajes se asocia cada jugador a un nombre y se ordena el arreglo de manera descendiente en cuanto al puntaje. Ante este ordenamiento se obtiene un arreglo donde el primer elemento es el jugador número uno del ranking y así obtenemos un ranking general.

- Evaluación de resultados en base al ranking obtenido

  Ya sabiendo como se realiza la generación del PageRank, evaluamos por cada año como se comporta éste cálculo en cada torneo. Es por eso que recorriendo torneo a torneo, se genera el ranking correspondiente al torneo y se evalúa partido a partido si el ranking logró indicar correctamente que aquel que fue el ganador del partido tenía un mejor posicionamiento en el ranking que el perdedor. En caso que esto sea cierto, lo considermaos un *Hit*; en caso de que no sea así, lo consideramos un *Miss*. Cabe destacar que únicamente evaluamos los partidos que resultaron ser completos. Una vez obtenidos los hit y los miss de todo el año, evaluamos cómo se comportó para ese año nuestro modelo de PageRank paramétrico.

La implementación para el cálculo del PageRank permite poder, a través de parámetros, indicar qué atributos se quieren evaluar para una corrida. Eso nos permitió la diferenciación por superficie, antigüedad, tipo de torneo, instancia, entre otras corridas realizadas para ir mejorando nuestro modelo. Para aquellos jugadores nuevos en el ranking, se asigna en nuestro PageRank manualmente un ranking muy alto, ya que no tiene historial que se vea reflejado en el multigrafo que se utiliza para calcular el PageRank.

7.2.2. Cálculo de mejores parametros

Para encontrar la mejor combinación de parámetros que utlizaríamos para calcular el valor de cada arista utilizamos el siguiente algoritmo

Algorithm 1 Algoritmo buscador de mejores parámetros.

$yearDone \leftarrow true$
$surfaceDone \leftarrow true$
$tournamentDone \leftarrow true$
$exponentialDone \leftarrow true$

$exponential \leftarrow [5, 10]$
$surfaces \leftarrow$ [00,1.,1]
$tournaments \leftarrow$ [11,1.,2]
$years \leftarrow$ [1.,6]

$maxResult \leftarrow 0$
$bestExponentialDecay \leftarrow 0$
$bestTournamentWeight \leftarrow 0$
$bestYearBefore \leftarrow 0$
$bestSurfaceWeight \leftarrow 0$

while $yearDone$ and $surfaceDone$ and $tournamentDone$ and $exponentialDone$ **do**
  for y in $years$ **do**
    $(resultado, year) \leftarrow max(processYear(y))$
  end for

  if $resultado == maxResult$ **then**
    $yearDone \leftarrow false$
  end if

  if $resultado > maxResult$ **then**
    $bestYearBefore \leftarrow year$
    $maxResult \leftarrow resultado$
  end if

  for e in $exponential$ **do**
    $(resultado, exponential) \leftarrow max(processExponential(e))$
  end for

  if $resultado == maxResult$ **then**
    $yearDone \leftarrow false$
  end if

  if $resultado > maxResult$ **then**
    $bestExponentialDecay \leftarrow exponential$
    $maxResult \leftarrow resultado$
  end if

  for s in $surface$ **do**
    $(resultado, surface) \leftarrow max(processSurface(s))$
  end for

  if $resultado == maxResult$ **then**
    $surfaceDone \leftarrow false$
  end if

  if $resultado > maxResult$ **then**
    $bestSurfaceWeight \leftarrow surface$
    $maxResult \leftarrow resultado$
  end if

  for t in $tournaments$ **do**
    $(resultado, tournament) \leftarrow max(processTournament(t))$
  end for

  if $resultado == maxResult$ **then**
    $tournamentDone \leftarrow false$
  end if

  if $resultado > maxResult$ **then**
    $bestTournamentWeight \leftarrow tournament$
    $maxResult \leftarrow resultado$
  end if
end while

Algorithm 2 Algoritmo buscador del mejor año.

procedure PROCESSYEAR(y)
$bestYearBefore \leftarrow y$

$resultado \leftarrow evaluarTorneoConMejoresParametros()$

return $resultado$

end procedure

Algorithm 3 Algoritmo buscador del mejor *exponential*.

procedure PROCESSEXPONENTIAL(e)
$bestExponentialDecay \leftarrow e$

$resultado \leftarrow evaluarTorneoConMejoresParametros()$

return $resultado$

end procedure

Algorithm 4 Algoritmo buscador del mejor parámetro para superficie.

procedure PROCESSSURFACE(s)
$bestSurfaceWeight \leftarrow s$

$resultado \leftarrow evaluarTorneoConMejoresParametros()$

return $resultado$

end procedure

Algorithm 5 Algoritmo buscador del mejor λ para torneo.

procedure PROCESSTOURNAMENT(t)
$bestTournamentWeight \leftarrow t$

$resultado \leftarrow evaluarTorneoConMejoresParametros()$

return $resultado$

end procedure

Donde *evaluarTorneoConMejoresParámetro* es la función que calcula la eficacia del Pagerank con los parámetros que estan en los flags como los mejores parámetros ante esa corrida.

7.2.3. Probabilidad de victoria y AUROC

Para el cálculo de probabilidad de victoria y su posterior análisis de AUROC, hemos generado una tabla en la base de datos que contiene como información: ranking del ganador, ranking del perdedor y número del partido. Para poder completar esta tabla, hemos evaluado cada partido almacenado, y guardado el ranking que nos entregó nuestro modelo como resultado para cada jugador.

- Cálculo de probabilidad de victoria

  Evaluado cada registro de la tabla generada, calculamos la diferencia de ranking de cada partido y además cruzamos con el resto de las tablas la información para saber la superficie y el tipo de torneo que se estaba disputando. Luego agrupamos por diferencia de ranking y calculamos la eficacia que tuvo de predicción nuestro ranking para esa diferencia.

  Eso nos da lugar a generar dos vectores. Por un lado el vector *xdata*, donde tenemos todas las diferencias de ranking, y el vector *ydata*, que tendrá la eficacia de nuestro ranking para esas diferencias.

  Es aquí donde tratamos de minimizar el error y generar la curva que logre acercar cada uno de esos puntos como lo indicamos en la sección de "Probabilidad de victoria" previamente. Para ello calculamos la curva que acerca a la función utilizando el método *curve_fit* de la librería *sklearn*. Una vez obtenido el mejor parámetro para nuestra función del modelo de regresión utilizado, logramos poder evaluar dada una diferencia de ranking, cuan optimista es nuestro modelo al momento de predecir el resultado.

  Dado que también se podía parametrizar la búsqueda, ya que contábamos con la información de los partidos como la superficie y el tipo de torneo, fuimos generando esto mismo por cada combinación posible.

  Esto nos dio lugar a la generación de un árbol de decisión donde dado, además de la diferencia de ranking, las características del partido, podíamos dar una mayor certeza de como podía llegar a ser el resultado del partido.

- Cálculo de AUROC

  Para el cálculo del área bajo la curva ROC recorrimos nuevamente cada partido, y armamos las tuplas donde el primer elemento representaba si el partido fue un *hit* o un *miss* y el segundo elemento fue la $P(Victoria)$ de ese partido.

  De la misma forma se replicó el resultado de forma opuesta; esto nos permitió la generación de un gráfico tipo S necesario para poder calcular de manera correcta la curva ROC y por ende así obtener su área bajo la curva.

  Para el cálculo de la curva ROC generamos dos arrays: *yTrue*, que tenía los primeros elementos de cada tupla, y el *yScore*, que tenia el segundo elemento. Luego, con la función *roc_curve*, logramos generar la curva ROC que nos permitía indicar el *False Positive Rate* y el *True Negative Rate*.

  Para poder obtener el AUROC utilizamos el método *roc_auc_score*, que dados los arrays generados, ya nos entregaba el resultado del área.