VNiVERSiDAD
D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

FACVLTAD DE CIENCIAS

GRADO DE ESTADÍSTICA

TRABAJO DE FIN DE GRADO

ANÁLISIS ESPACIAL DEL TIRO EN BALONCESTO PROFESIONAL Y SU TRANSFORMACIÓN A LO LARGO DEL TIEMPO

Autor: Rodrigo Valladares Alonso

Tutor: José Luis Vicente Villardón

Salamanca, 2020

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

800 AÑOS VNiVERSiDAD D SALAMANCA 1218 - 2018

FACVLTAD DE CIENCIAS

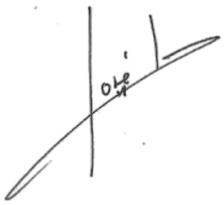
GRADO DE ESTADÍSTICA

TRABAJO DE FIN DE GRADO

ANÁLISIS ESPACIAL DEL TIRO EN BALONCESTO PROFESIONAL Y SU TRANSFORMACIÓN A LO LARGO DEL TIEMPO

Autor: Rodrigo Valladares Alonso

Tutor: José Luis Vicente Villardón



Salamanca, 2020

Resumen

La NBA ha sufrido grandes cambios a lo largo de toda su historia, mayoritariamente debidos a la implementación de nuevas reglas, a las mejoras en el físico de los jugadores, y a otros muchos factores. Sin embargo, el cambio posiblemente más significativo en la liga no ha tenido que ver con aspectos ligados directamente al juego, sino que ha venido de mano de la estadística. Los avances en tecnología y su implementación, junto con las técnicas de análisis de datos, han jugado un papel fundamental en la estrategia de los equipos en los últimos años. El objetivo principal de este trabajo es analizar la transformación que han sufrido los jugadores a lo largo del tiempo para dar una respuesta a por qué hoy se juega como se juega, así como encontrar los aspectos más relevantes ligados a esta transformación.

Así, se presenta una breve introducción a lo que es la NBA y cómo funciona su sistema de competición, además de hablar de aspectos importantes sobre baloncesto necesarios para el total entendimiento del análisis que se va a realizar. Después se lleva a cabo una recopilación de la información existente acerca de los principales cambios que se dan hoy en día en el juego y de la importancia de la estadística, para que el análisis que se realiza posteriormente esté lo mejor enfocado posible. Por último, se describen las técnicas utilizadas en los distintos análisis para mostrar a continuación los resultados más significativos de éstos de cara a las conclusiones.

Se comprueba cómo la transformación en el estilo de juego actual, debido a la implementación del análisis especializado, ha tenido un impacto importante en el cambio en el rol de las posiciones de los jugadores. Este cambio se ve especialmente acentuado en el apartado de tiro y mucho menos pronunciado en aspectos más profundos en los cuales los roles de cada posición han tenido que verse menos alterados. También se descubre la diferencia significativa que existe entre los tiros realizados cerca de la canasta con el resto de tiros, especialmente en los factores principales que determinan su éxito.

Abstract

The NBA has experienced major changes throughout its history, mostly due to the implementation of new rules, improvement in the physical abilities of the players, and many other factors. However, the most significant change in the league probably has not been directly related to the game but has come from the statistics. Advances in technology and its implementation, along with data analysis techniques, have played a big role in team strategy in recent years. The main objective of this paper is to analyze the transformation that players have suffered over time in order to provide an answer to why the game is played the way it is today, as well as to find the most relevant aspects linked to this transformation.

First it is shown a brief introduction to what the NBA is and how its competition system works, as well as talking about important aspects about basketball necessary for the total understanding of the analysis to be carried out. After that, a compilation of the existing information about the main changes that are taking place in the game today and the importance of statistics is analyzed, so that the analysis that is made afterwards is as focused as possible. Finally, the techniques used in the different analyses are described and then the most significant results are shown in order to draw conclusions.

It is verified how the transformation in the current style of play due to the implementation of the specialized analysis has had an important impact in the change over the role of the players' positions. This change is especially seen in the shooting section and much less pronounced in deeper aspects in which the roles of each position have had to be less altered. You will also discover the significant difference between shots taken close to the basket and the rest, especially in the main factors that determine their success.

ÍNDICE DE CONTENIDOS

1. INTRODUCCIÓN

1.1 El sistema de competición de la NBA

La NBA (*National Basketball Association*) es la liga profesional de baloncesto de Estados Unidos y Canadá, y sin lugar a duda, la más importante del mundo. En ella participan los mejores jugadores del planeta y mantiene uno de los sistemas de competición deportiva más exigentes que existe.

Actualmente la liga se divide en dos conferencias de quince equipos cada una: Conferencia Este y Conferencia Oeste. Y éstas a su vez se dividen en tres divisiones de cinco equipos:

- Conferencia Este (División Atlántico, División Central y División Sureste).
- Conferencia Oeste (División Noroeste, División Suroeste y División Pacífico).

Antiguamente estas divisiones eran más relevantes a la hora de la clasificación, ya que determinaban los equipos que alcanzaban la fase final en mejores posiciones. Pero este sistema no tenía en cuenta que un equipo en segunda posición de una división podría tener mejores resultados que el primer clasificado de otra y así ser perjudicado en la plaza conseguida para esa fase final. Por ello, hoy en día y desde 2014, las divisiones se utilizan arbitrariamente para contener un mismo número de equipos por zona geográfica y así adecuar el calendario lo mejor posible para cada equipo, y en última instancia, como criterio de desempate en caso de ser necesario.

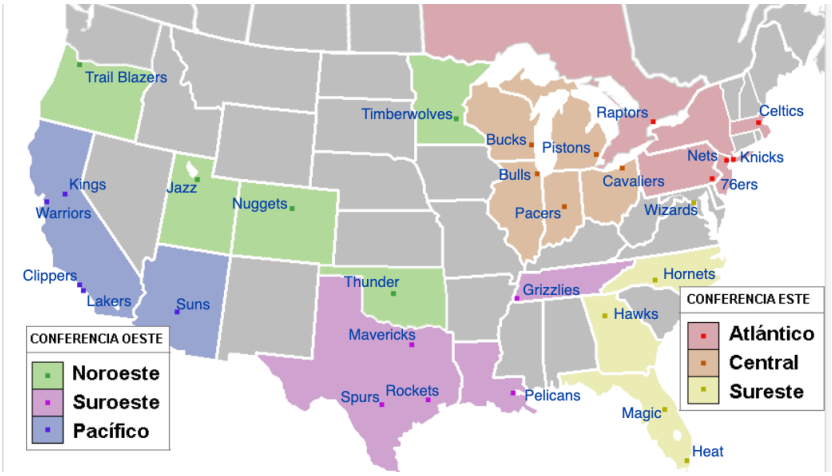


***Figura 1**: Distribución de los equipos de la NBA en sus respectivas conferencias y divisiones.*

El sistema de competición se divide en dos grandes bloques: la Temporada Regular y los *Playoffs*.

1.1.1 Temporada Regular

La Temporada Regular habitualmente da comienzo en la tercera o cuarta semana de octubre. En total consta de 1230 partidos, 82 por equipo, con el siguiente esquema para cada uno:

- 4 enfrentamientos contra los otros cuatro equipos de su división.
- 3 - 4 enfrentamientos contra el resto de los equipos de su conferencia.
- 2 enfrentamientos contra cada equipo de la otra conferencia.

La Temporada Regular finaliza entre la segunda y tercera semana de abril dejando a cada equipo con un balance de victorias y derrotas que se denomina *record*. Este *record* es el que determina la clasificación de los equipos, y por lo tanto los enfrentamientos de cara a la siguiente fase: los *Playoffs*.

Sin embargo, la clasificación se tiene en cuenta por conferencia. De los quince equipos de cada una de ellas se clasifican los ocho con un mejor *record*. De este modo un equipo de la Conferencia Oeste podría quedarse fuera de los *Playoffs* a pesar de tener mejor *record* que un equipo de la Conferencia Este que sí ha conseguido clasificarse, o viceversa.

1.1.2 Playoffs

Una vez se tiene los ocho equipos de cada conferencia que han conseguido clasificarse para los *Playoffs* da comienzo la lucha por el título. La forma de proceder en cada conferencia es la misma:

- El 1º se enfrenta al 8º
- El 4º se enfrenta al 5º
- El 3º se enfrenta al 6º
- El 2º se enfrenta al 7º

dando lugar así a unos cuartos de final, semifinales y final de conferencia. Los campeones de conferencia miden sus fuerzas en la gran final para ver quién se alza como campeón de la NBA.

Todas las rondas siguen el mismo esquema, una serie al mejor de siete partidos. Esto implica que es necesario ganar cuatro veces al mismo rival para poder avanzar. En cada serie el equipo con mejor *record* goza de la ventaja de campo. El modelo de alternancia es 2-2-1-1-1, así el equipo mejor clasificado juega los dos primeros partidos como local y los dos siguientes como visitante, y en caso de ser necesario quinto, sexto o incluso séptimo partido (lo cual es lo más frecuente ya que no demasiadas series se resuelven con un 4-0 para ninguno de los dos equipos implicados), al equipo con mejor *record* le correspondería jugar como local el quinto y el séptimo, y como visitante el sexto.

***Figura 2**: Esquema de los Playoffs de la temporada 2015/2016. Fuente: Sports Illustrated.*

1.2 Historia

El origen de la NBA se remonta a 1946, cuando los propietarios de los principales pabellones deportivos del noreste y medio-oeste de USA estaban en busca de otro espectáculo diferente al hockey sobre hielo y el boxeo que consiguiese llenar esos estadios. Estos magnates encontraron en el baloncesto una gran oportunidad y ello dio lugar a la BAA (*Basketball Association of America*), aunque ya existía una liga de baloncesto nacional, la NBL (*National Basketball League*). Estas dos se fusionaron en 1949 para dar lugar a lo que hoy se conoce como NBA.

Durante el transcurso de los años se han dado muchos cambios significativos: el número de equipos y los propios equipos, los jugadores son muy distintos a cómo eran antaño, se han ido modificando y adaptando las reglas, dimensiones de la cancha, zonas de tiro, interpretaciones, etc.

Cabe destacar que la NBA tiene un reglamento propio, diferente al estipulado por la FIBA (Federación Internacional de Baloncesto) y, por ende, distinto del resto de competiciones regidas por ésta, con diferencias notables, y que, de ser necesario para la completa comprensión de este trabajo, se explicarán más adelante.

Pese a todos estos cambios realizados a lo largo del tiempo, este trabajo se centra en los más importantes de cara a la perspectiva estadística y las posibles repercusiones que ha tenido en ésta. Pero antes y para un mejor entendimiento de estos cambios, se tiene que poner sobre el papel algunos conceptos importantes.

1.3 Juego y Conceptos Importantes

Hasta ahora se ha hablado de competición, historia, equipos… pero en ningún momento se ha hablado de baloncesto como tal.

El baloncesto es un deporte de pelota en el cual, en un partido de tiempo limitado, dos equipos se enfrentan para ver cuál de los dos es capaz de anotar más puntos que el otro. No es posible que un partido finalice en empate, en caso de que el marcador acabe en tablas al finalizar el tiempo reglamentario se irán añadiendo tiempos extra denominados prórrogas hasta que uno de los dos equipos terminé con ventaja.

Como se ha mencionado previamente, el funcionamiento de los partidos en la NBA tiene diferencias destacables al que, por ejemplo, se practica en Europa. Por ello, las explicaciones aquí recogidas se ceñirán estrictamente al baloncesto americano y su reglamento.

Un partido en la NBA está dividido en cuatro periodos de 12 minutos cada uno, y cada periodo extra, si fuese necesario, sería de 5 minutos. Aunque el juego se desarrolla en forma de cinco contra cinco, cada equipo puede convocar un máximo de 13 jugadores y hacer todos los cambios deseados durante el transcurso del partido.

Otro concepto importante a tener en cuenta es que, cada vez que un equipo empieza a tener el control del balón, dispone de 24 segundos para realizar un tiro que consiga al menos tocar el aro, de lo contrario el control de balón pasa a ser del otro equipo. Esta cuenta de 24 segundos se llama reloj de posesión. Pero si en una posesión, el reloj de posesión es inferior a 14 segundos para un equipo y éste captura un rebote ofensivo (que provenga de un tiro que haya tocado el aro) o provoca una falta del equipo rival, su reloj de posesión volverá a los 14 segundos.

El objetivo de cada equipo es claro: conseguir más puntos que su rival, pero ¿de dónde pueden proceder estos puntos? La cancha (Figura 3) está estructurada de forma en la que las canastas obtenidas por tiros realizados dentro de la línea de tres puntos situada a 7,24 metros del aro (6,71 metros en las esquinas) suman 2 puntos, mientras que las obtenidas por tiros realizados más allá de esta línea suman 3 puntos. También es posible obtener canastas de 1 punto si éstas proceden de la línea de tiros libres tras haberse producido una falta que obligue a un jugador a ir a esta línea.

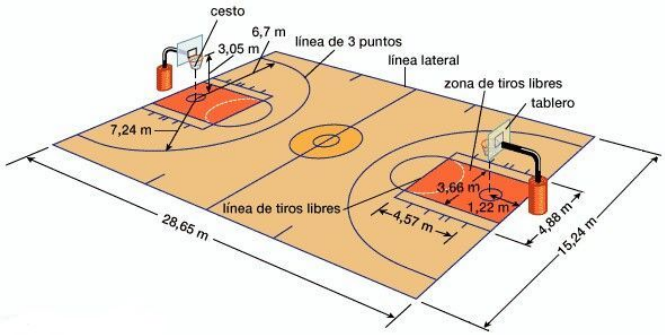


***Figura 3**: Dimensiones de una cancha de baloncesto en la NBA.*

Durante el tiempo que dura un partido ocurren cientos de acciones de interés, tanto defensivas como ofensivas. Muchas quedan recogidas en las estadísticas oficiales del partido, que dan lugar al denominado *boxscore*: puntos, asistencias, rebotes, porcentajes de acierto, etc. Otras tantas no son cuantificables objetivamente y por lo tanto no se recogen en ningún lugar, a pesar de ser importantes en el juego y convenientes de analizar: posicionamiento correcto, toma de decisiones, *timing* de las acciones, etc.

Quedan aquí presentadas las acciones cuantificables que forman parte de la estadística oficial. Estos datos se recogen de forma individual para cada jugador y el agregado de éstos conforma el global del equipo.

- Minutos (**MIN**): Minutos disputados por un jugador.
- Puntos (**PTS**): Puntos obtenidos por un jugador.
- Asistencias (**AST**): Pases realizados por un jugador que inmediatamente después se han convertido en canasta.
- Rebotes ofensivos (**OREB**): Captura del balón conseguida por un jugador después de que el tiro de un compañero de su equipo o un tiro propio haya sido errado.
- Rebotes defensivos (**DREB**): Captura del balón conseguida por un jugador después de que el tiro de un rival haya sido errado.
- Robos (**STL**): Balón recuperado, que estaba en posesión del rival, por un jugador.
- Tapones (**BLK**): Tiro del rival desviado legalmente por un jugador y que por lo tanto no se ha convertido en canasta.
- Tiros de campo intentados (**FGA**): Suma de tiros de 2 y 3 puntos intentados por un jugador.
- Tiros de campo convertidos (**FGM**): Suma de tiros de 2 y 3 puntos conseguidos por un jugador.
- Porcentaje de acierto de los tiros (**FG%**): Cociente entre los tiros convertidos y los intentados por un jugador.
- Tiros de 3 intentados (**3PA**): Tiros exclusivamente de 3 puntos intentados por un jugador.
- Tiros de 3 convertidos (**3PM**): Tiros exclusivamente de 3 puntos convertidos por un jugador.
- Porcentaje de acierto de los tiros de 3 (**3P%**): Cociente entre los tiros de 3 puntos convertidos y los intentados por un jugador.
- Tiros libres intentados (**FTA**): Tiros libres intentados por un jugador.
- Tiros libres convertidos (**FTM**): Tiros libres convertidos por un jugador.
- Porcentaje de acierto de los tiros libres (**FT%**): Cociente entre los tiros libres convertidos y los intentados por un jugador.
- Pérdidas (**TOV**): pérdidas de balón realizadas por un jugador que otorgan la posesión al otro equipo.
- Faltas personales (**PF**): Faltas personales realizadas por un jugador (Un jugador puede cometer un máximo de 6 faltas personales tras las cuales es expulsado del partido y no puede volver a participar en él).
- Más menos (+/-): Diferencia entre los puntos obtenidos y encajados por el equipo durante el tiempo que un jugador ha estado en pista.

Boston Celtics																			
PLAYER	MIN	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	PTS	+/-
Jayson Tatum F	35:54	11	22	50.0	3	9	33.3	5	6	83.3	1	5	6	2	0	4	0	1	30	5
Gordon Hayward F	36:04	10	19	52.6	3	7	42.9	4	4	100	2	8	10	5	2	2	1	4	27	-4
Daniel Theis C	33:34	8	12	66.7	2	2	100	2	2	100	2	4	6	3	0	0	2	5	20	2
Marcus Smart G	35:44	6	14	42.9	2	6	33.3	2	2	100	0	5	5	5	1	0	0	1	16	12
Kemba Walker G	30:19	3	12	25.0	2	8	25.0	3	4	75.0	0	3	3	2	2	0	0	0	11	-1
Romeo Langford	9:46	0	4	0.0	0	1	0.0	0	0	0.0	1	0	1	0	0	0	0	1	0	3
Semi Ojeleye	21:10	1	5	20.0	1	4	25.0	0	0	0.0	1	5	6	0	1	0	0	1	3	1
Brad Wanamaker	20:23	1	2	50.0	0	0	0.0	3	3	100	1	1	2	1	1	1	0	1	5	3
Enes Kanter	14:26	1	3	33.3	0	0	0.0	0	1	0.0	3	4	7	1	1	1	1	0	2	1
Grant Williams	2:40	0	0	0.0	0	0	0.0	0	0	0.0	0	0	0	0	0	1	0	1	0	-7
Carsen Edwards	DNP - COACH'S DECISION																			
Javonte Green	DNP - COACH'S DECISION																			
Robert Williams III	DNP - COACH'S DECISION																			
Totals:		41	93	44.1	13	37	35.1	19	22	86.4	11	35	46	19	8	9	4	15	114	3

***Tabla 1**: Ejemplo del boxscore de los Boston Celtics en su partido contra los Indiana Pacers celebrado el 10 de marzo de 2020. Fuente: NBA.*

Todas estas variables son las más básicas que se pueden encontrar en las estadísticas tradicionales de un partido. ¿Son útiles? Sí. ¿Sirven para un análisis significativo? No lo suficiente. Ésta es una de las razones por las que las herramientas de análisis estadístico aplicadas a la NBA han evolucionado enormemente en los últimos años.

1.4 Importancia y evolución de las Estadísticas en la NBA

Entonces, ¿por qué se necesita profundizar en el análisis de todas estas variables e ir más allá? La respuesta es sencilla y se encuentra en la propia naturaleza de la estadística: cuanta más información se tenga de un suceso y mejor sea el análisis de esa información, mejores resultados se podrán obtener.

La estadística deportiva es una rama que está en auge y no ha parado de crecer en los últimos años. En parte gracias a la especialización, y también por los grandes avances en tecnología para poder obtener datos y procesarlos.

Las estadísticas tradicionales no son nada complicadas de conseguir. A poco que alguien entienda de baloncesto, esté atento a un partido apuntando todo lo que pasa en él y pueda dejarlo reflejado en el papel, se haría fácilmente con el *boxscore* de ese partido. Hoy en día existen una gran cantidad de aplicaciones que permiten hacerlo de forma intuitiva. Y así, podría afirmar, sin tener gran idea de estadística, quién ha sido el jugador que ha anotado más puntos, qué equipo ha conseguido capturar más rebotes ofensivos o incluso atreverse a afirmar quién ha sido el mejor defensor basándose en los robos que éste ha tenido. Pero estaría pasando por alto uno de los pilares más importantes de la estadística: **correlación no implica causalidad**. Cualquiera podría afirmar todo esto sin tener en cuenta que el perfil de los equipos que se han enfrentado ya indicaba a priori cuál iba a ser el equipo con más rebotes ofensivos; o que ese jugador que él ha puesto como un gran defensor ha conseguido tantos robos porque se ha encargado de defender al peor jugador del equipo rival.

En las estadísticas tradicionales sólo se ven acciones que toman muy poca parte del tiempo total de un partido, todo lo que hay detrás queda de la mano de una estadística más avanzada que no demasiada gente es capaz de manejar o interpretar, pero que es, sin duda alguna, enormemente ventajosa de cara a obtener resultados.

Como es lógico, un entrenador profesional no necesita saber nada de estadística para hacer un análisis intuitivo y suficientemente exhaustivo del juego, pero puede estar pasando por alto hechos importantes que no pueden ser descubiertos sin echar mano a esa estadística que desconoce. Lo que antes era suficiente no lo es nunca más. Ya no basta con saber de dónde proceden las victorias para potenciarlas y las derrotas para corregirlas. Ahora se deben tener en cuenta todos los factores posibles que intervienen en estos procesos y ser capaces de optimizarlos al máximo.

Hoy en día no hay una sola franquicia en la NBA que no cuente con su propio equipo de analistas de datos. Obviamente esto no ha sido siempre así, y tampoco han sido cambios introducidos de un día para otro. Pero ¿dónde se sitúa este *boom* de interés por el análisis de datos en la NBA? Si se tiene que fijar un punto de inflexión sería entre las temporadas 2009/2010 y 2013/2014. Esto no significa que antes no hubiera equipos interesados en ver que hay detrás de los meros números proporcionados por un papel al acabar un partido, de hecho algunos ya lo hacían. Entonces, ¿por qué esas fechas en concreto? Es bien sabido que la mayoría de las veces el problema a la hora de trabajar con datos no reside en lo que es puramente el análisis de éstos, sino en su obtención.

1.5 Data tracking con SportVU y Second Spectrum

En la temporada 2009/2010 la NBA llegó a un acuerdo con *SportVU*, una compañía de obtención y procesamiento de datos, y empezó a implementar en los principales pabellones de la liga un sistema de vídeo basado en seis cámaras capaz de captar el movimiento de cada jugador, así como el del balón y las acciones que sucedían en la pista incluyendo sus coordenadas, a 25 imágenes por segundo. Al final

de la temporada 2013/2014 todos los pabellones de la competición contaban con ese sistema de obtención de datos.

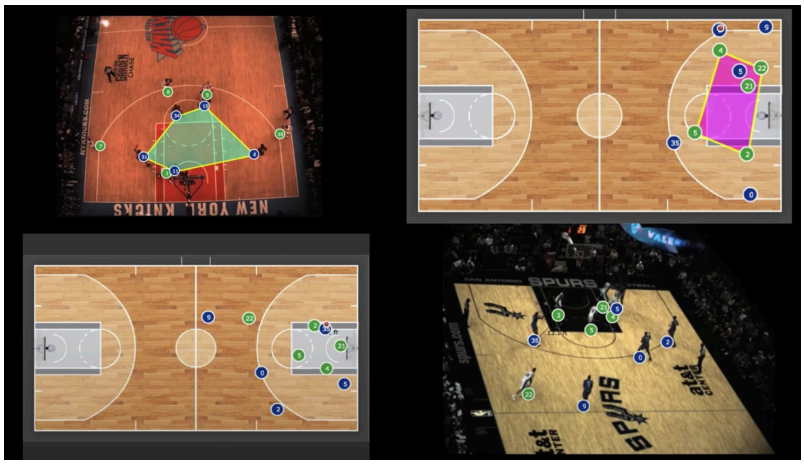


***Figura 4**: Ejemplos del funcionamiento del sistema de data tracking empleado por la NBA con la tecnología de SportVU. Fuente: STATS.*

Teniendo en cuenta que un partido sin prórrogas dura 48 minutos - lo equivalente a 2880 segundos - y que por cada segundo se tiene 25 veces toda la información de lo que está sucediendo en la pista, esto se traduce en un total de 72000 imágenes en las que se tiene las coordenadas de cada jugador en la pista (*x, y; t*) y las del balón (*x, y, z; t*). Todo esto de un solo partido y, como se ha explicado anteriormente, cada equipo tan sólo en la Temporada Regular juega un total de 82 partidos. Se obtiene de esta forma una cantidad enorme de datos que requiere de un análisis muy refinado y potente para poder sacar información útil de ellos.

Después de la temporada 2016/2017 el proveedor oficial paso de ser *SpotVU* a *Second Spectrum*, aunque sus sistemas no guardan grandes diferencias.

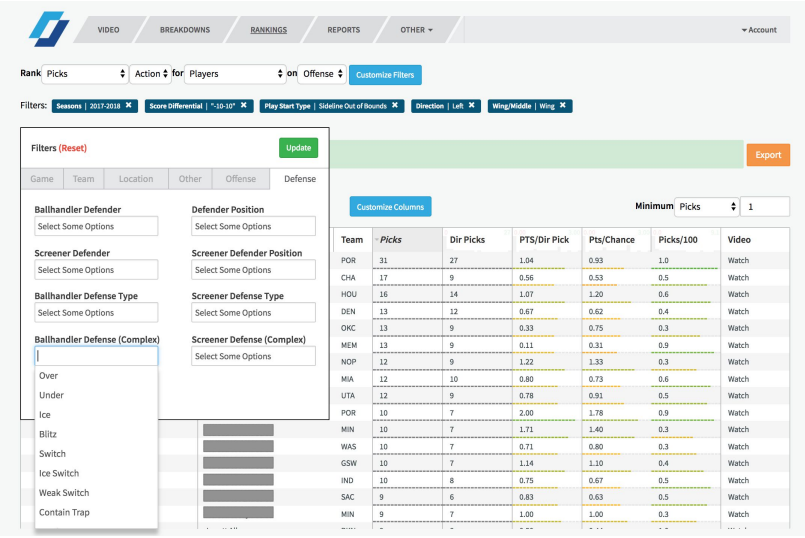


***Figura 5**: Ejemplo de base de datos proporcionada por Second Spectrum. Fuente: Second Spectrum.*

Se encuentran una infinidad de datos: tiros desde una posición en concreto, procedencia de la mayor parte de las asistencias de un jugador, puntos por el número de veces que toca un jugador el balón y una lista interminable de diferentes posibilidades y combinaciones. Y todo esto, cuando lo que realmente le importa a cualquier equipo, es conseguir más puntos que su rival y así ganar el partido.

1.6 ¿Cómo ha cambiado la estadística el juego?

Todos estos avances han propiciado una serie de cambios significativos en el juego de cara a una mejora en la obtención de los resultados y en la eficiencia. Pero el mayor impacto de este nuevo uso de la tecnología se ha visto de forma más clara en dos aspectos directamente correlacionados y que más adelante se explicará su porqué: la línea de tres puntos y los jugadores.

1.6.1 La línea de tres puntos.

Sin duda la mayor revolución en la NBA se debe a la nueva perspectiva que se tiene sobre la línea de tres puntos. No siempre ha existido, pero desde que se introdujo en la temporada 1979/1980 su impacto sobre el juego no ha dejado de crecer. Se presenta un sencillo gráfico del artículo *"The 3-Point Revolution (Shea, 2018)"* en el que puede observarse claramente este cambio tan contundente.

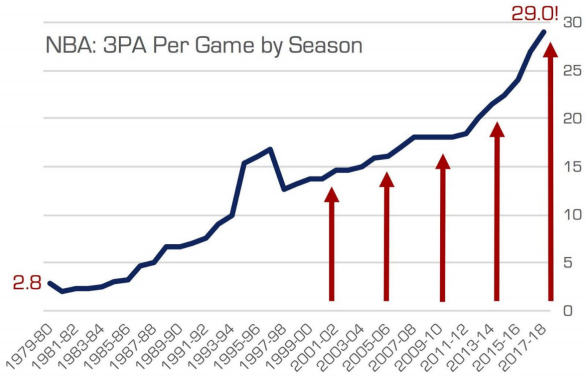


***Gráfico 1**: Gráfico del número medio de tiros de 3 puntos intentados en un partido por un equipo a lo largo del tiempo. Fuente: ShotTracker.*

La diferencia es abismal. En la primera temporada en la que se encuentra la línea de tres puntos cada equipo lanzaba una media de 2,8 triples por partido. Hoy en día esa cifra asciende hasta los 29 intentos por encuentro. Y no, no es que se hagan más tiros que en 1980 como se ilustra en *"A whole new ball game: Quantifying changes in NBA basketball over the past 30 years (Thinking Machines Data Science, 2018)"*.

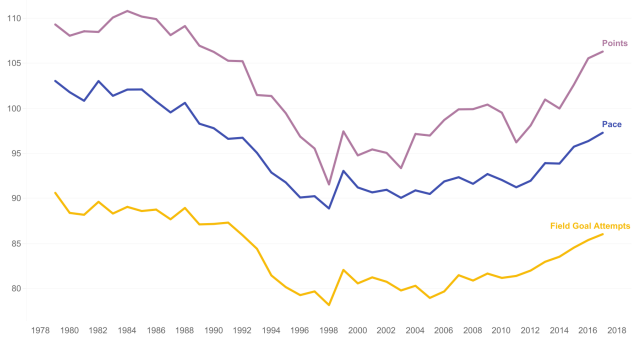


***Gráfico 2**: Gráfico del número medio de tiros intentados (línea amarilla), puntos (línea rosa) y ritmo (línea azul) por partido y por un equipo a lo largo del tiempo. Fuente: Thinking Machines.*

De hecho, **sucede justamente lo contrario**: se hacen menos tiros, se meten menos puntos y se tiene un ritmo (variable que mide la cantidad de posesiones, tanto ofensivas como defensivas, que promedia un equipo por partido) más bajo.

Se podría pensar que se debe a una mejora significativa en los porcentajes de acierto de estos tiros ya que, en los primeros años, no era algo a lo que los jugadores estuvieran acostumbrados. Y sí, una mejora ha habido, de eso no hay duda. Mientras que la media de acierto en los tiros de 2 puntos sólo se ha visto incrementada en un 2%, los porcentajes de acierto en los tiros por detrás de la línea de tres puntos han mejorado en un 10%. A priori puede parecer un cambio poco relevante, pero es uno de los factores más importantes en esta revolución.

Entonces, ¿qué más aparte de la mejora en el acierto de estos tiros? **Las matemáticas**.

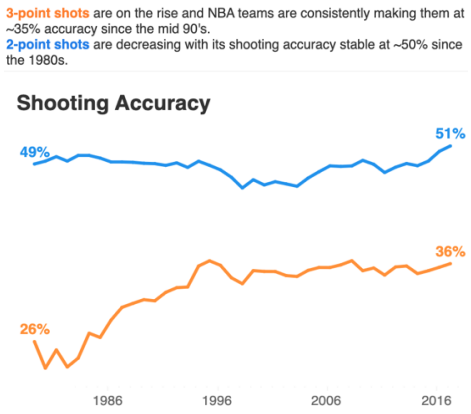


***Gráfico 3**: Gráfico con los porcentajes de acierto a lo largo del tiempo para los tiros de 2 y 3 puntos. Fuente: Thinking Machines*

El análisis estadístico profundo ha jugado un papel fundamental en este cambio. Para entender todo esto hay que remontarse al hombre que cambió la filosofía de un equipo para poner mucho más eco en los tiros de 3 puntos: *Daryl Morey*. *Morey* es el *General Manager* de los *Houston Rockets*, y lo que es más importante, un firme creyente en la estadística avanzada aplicada al deporte. *Morey*, a pesar de todas las críticas recibidas por el estilo de juego de su equipo, se mantiene firme. Los *Rockets* basan sus posesiones en hacer tiros muy cercanos al aro o acabar con un tiro de 3 puntos. Puede que para muchos no sea lo más entretenido de ver, pero de lo que no hay duda es que da sus resultados.

Para comprender lo que *Morey* entendió mucho antes que la mayoría de los equipos se tiene que hacer uso del análisis espacial de las diferentes zonas desde donde se realizan los tiros y los porcentajes de acierto asociados a estas posiciones. En el artículo *"How Mapping Shots In The NBA Changed It Forever (Goldsberry, 2019)"* se encuentran las ilustraciones necesarias para ello.

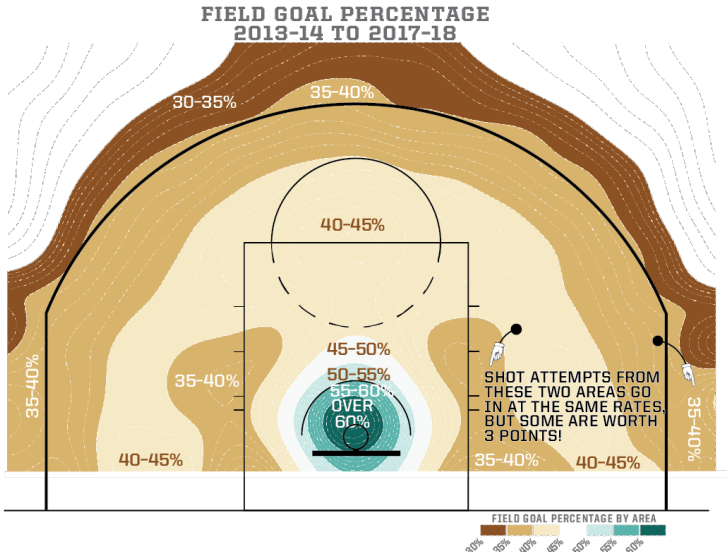


***Gráfico 4**: Shot Chart de los porcentajes de acierto por zonas del campo de todos los tiros realizados entre las temporadas 2013/2014 y 2017/2018. Fuente: Kirk Goldsberry.*

Como es lógico, los tiros de un partido se ven afectados por muchas variables y es imposible a priori sobre el papel fijar exactamente los tiros que va a realizar un equipo. Pero la estrategia juega un papel fundamental en las decisiones que toman los jugadores. Y para optimizar cualquier estrategia no hay nada más ventajoso que usar datos basados en la ciencia.

Con la información sobre la probabilidad en promedio en la cual se tiene éxito en los tiros desde las diferentes localizaciones posibles, simplemente basta con calcular el resultado que se espera obtener

desde cada una. O en términos estadísticos, la esperanza matemática de puntos que se tiene para cada zona del campo.

Si se supone entonces que desde una zona más alejada de la línea de tres puntos se tiene una probabilidad de acierto del 33%, como cada tiro desde esa zona en caso de ser convertido hace sumar 3 puntos es fácil ver que si de cada 3 tiros que se hacen se acierta 1 y éste vale 3 puntos, se han necesitado 3 tiros para conseguir 3 puntos. La esperanza matemática en este caso es de 1 punto por cada tiro realizado. Aplicando esto a todas las zonas del campo se obtiene el Gráfico 5.

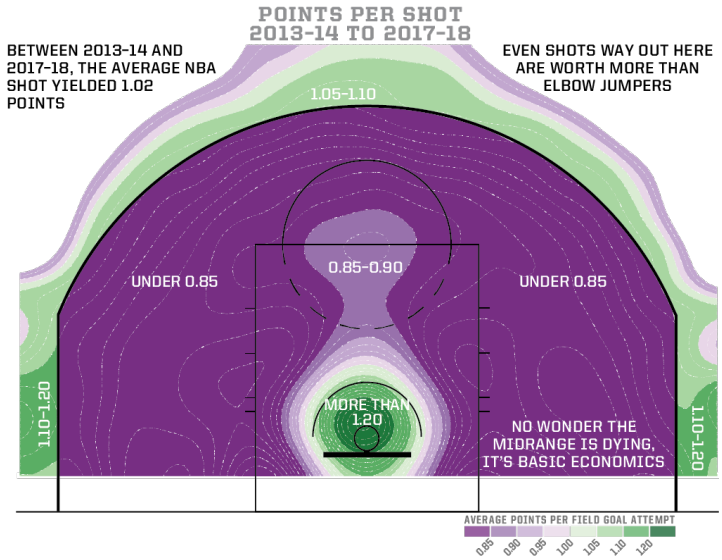


***Gráfico 5**: Shot chart de los puntos esperados por cada zona de tiro en base a sus porcentajes de acierto. Fuente: Kirk Goldsberry.*

La evidencia es clara, los únicos tiros que tienen más valor a priori que los triples son los que se hacen en la zona más cercana a la canasta (a excepción de los tiros libres que han sido excluidos de los gráficos debido a que las condiciones en las que se dan estos tiros son completamente diferentes al resto).

Como se ha mencionado, *Morey* creía fervientemente en esta teoría. Para observar mejor este fenómeno se presenta una comparación entre los tiros que han realizado los *Rockets* en las últimas 3 temporadas y el equipo con los considerados dos mejores triplistas de la liga: los *Warriors*.

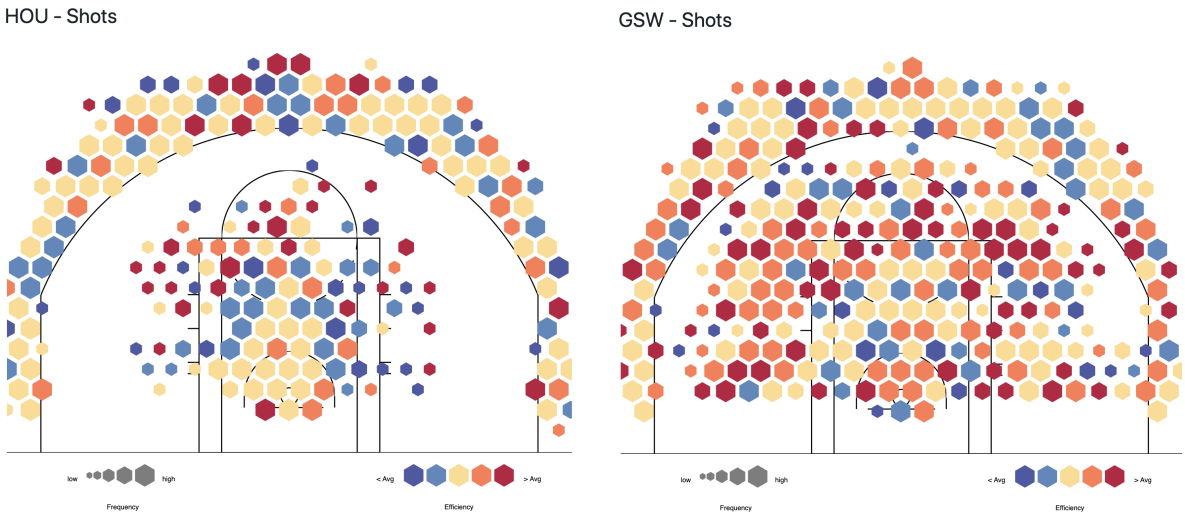


***Gráfico 6**: Shot Charts de todos los tiros de los Houston Rockets y los Golden State Warriors en las últimas 3 temporadas. El tamaño de los hexágonos hace referencia al volumen de tiros en esa posición y el color a su porcentaje de acierto. Fuente: PBP Stats.*

Esta transformación e impacto de los tiros de 3 puntos en la liga se ha traducido en una necesidad para los jugadores de adaptarse a este nuevo estilo de juego en el que tanto se premia la puntería desde la larga distancia.

1.6.2 Los jugadores.

Los perfiles tradicionales de los jugadores de baloncesto vienen definidos por dos aspectos esenciales: las características físicas y su rol dentro del equipo. Las **posiciones** en un equipo de baloncesto que pueden ocupar los jugadores son:

Posiciones exteriores:

- *Point Guard* (*PG*) o Base.
- *Shooting Guard* (*SG*) o Escolta.
- *Small Forward* (*SF*) o Alero.

Posiciones interiores:

- *Power Forward* (*PF*) o Ala Pívot.
- *Center* (*C*) o Pívot.

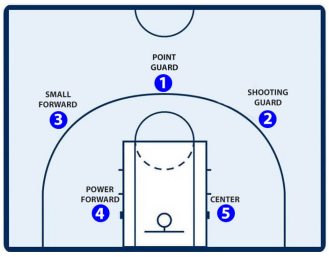


***Figura 6**: Disposición de los jugadores de un equipo en la pista por posiciones. Fuente: Basketball Phantom*

La teoría sobre el papel dice que los tres primeros (*PG*, *SG* y *SF*) forman la parte exterior del quinteto y los otros dos (*PF* y *C*) la interior. También, y en función de las características físicas de altura, envergadura, velocidad y potencia, el perfil tradicional dice que los jugadores exteriores son más pequeños y con mejores habilidades en el control del balón y el tiro, mientras que los interiores son más grandes, fuertes y los encargados de capturar los rebotes y meter canastas cerca del aro. Pero todo esto es, como se ha mencionado, teoría.

Es importante diferenciar los conceptos de posición y rol de un jugador, ya que será algo muy tratado a lo largo de este trabajo. La **posición** de un jugador viene definida, como se ha mencionado antes, por una serie de características tanto físicas como de habilidad, y es una forma de clasificar al jugador en un grupo más amplio en el que los jugadores de un mismo grupo tienen unas características similares. El **rol** es un concepto más difuso que hace referencia a la función del jugador dentro del equipo y es algo más particular de cada jugador independientemente de su posición. Sin embargo, también es importante conocer que cada una de las posiciones está asociada a una serie de roles precisamente por esas habilidades y cualidades físicas.

Esta teoría y generalidades en el pasado se ajustaba mucho más a la realidad de lo que lo hace en la actualidad. Para ahondar en esta idea se presenta una nueva variable: el porcentaje de Uso o *Usage rate*.

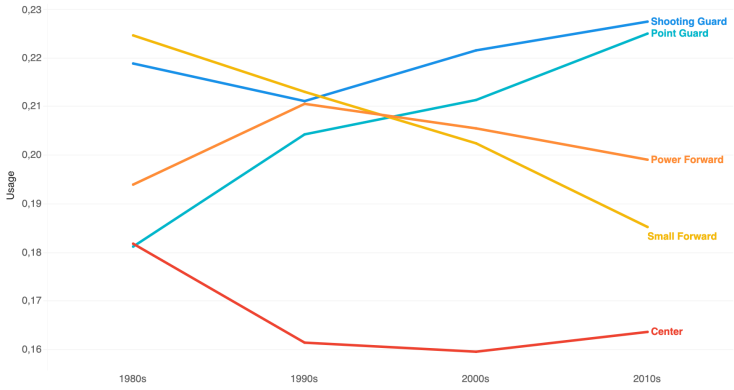


***Gráfico 7**: Gráfico con el Usage rate de las distintas posiciones que desempeñan los jugadores de un equipo a lo largo del tiempo. Fuente: Thinking Machines.*

¿Qué mide esta variable? Mide el porcentaje de posesiones ofensivas empleadas por un jugador que terminan en un tiro suyo, en conseguir tiros libres o en cometer una pérdida de balón. En otras palabras, la cantidad de posesiones ofensivas que acaban en las manos de ese jugador. El Gráfico 7 por lo tanto compara el porcentaje de posesiones que utilizan en promedio los jugadores dependiendo de su posición.

Se puede comprobar que los jugadores más perjudicados por el paso del tiempo son los *Centers* y los *Small Forwards*. Sobretodo estos últimos, que han pasado de ser los jugadores que más posesiones acaparaban a ser los segundos que menos lo hacen en la actualidad. ¿Se puede encontrar una respuesta científica a este fenómeno?

Se ha hablado de que el cambio más importante ha tenido lugar a raíz de la línea de tres puntos, así que presumiblemente la opción más lógica es que este cambio en los roles de las diferentes posiciones tenga una estrecha relación con su habilidad para lanzar triples con efectividad. En *"How the Three-Point Line Changed the NBA and the Game of Basketball (Meng, 2018)"* recogen este cambio a lo largo del tiempo para las diferentes posiciones (Gráfico 8).

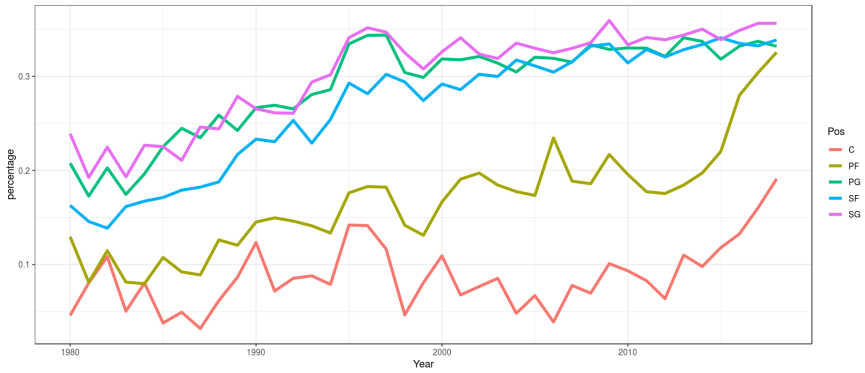


***Gráfico 8**: Gráfico con el porcentaje de acierto en los tiros de 3 puntos de las distintas posiciones que desempeñan los jugadores de un equipo a lo largo del tiempo. Fuente: NYC Data Science Academy.*

Cabe destacar que el incremento significativo que se produce a finales de los años 90 para todas las posiciones se debe a que en la temporada 1994/1995 y con el objetivo de frenar el descenso de puntos que se estaban consiguiendo en las temporadas previas, la NBA redujo la distancia de la línea de tres puntos de los 7,24 metros con 6,71 metros en las esquinas a una línea uniforme de 6,71 metros desde cualquier ángulo. Este cambio duró tres temporadas y se ve reflejado en ese incremento sustancial en el porcentaje de acierto para esas fechas.

Se ve que para las tres posiciones exteriores se experimenta una mejora mucho más lineal y constante que la de los interiores, los cuales precisamente mejoran más sus porcentajes con el *boom* del análisis estadístico favorecido por *SportVU*. Se trata de selección natural, los exteriores siempre habían sido los encargados de tirar de 3, pero en una liga que cada vez avanza a un mayor volumen de estos tiros necesita de todos los jugadores que puedan aportar a esta faceta. Adaptarse o morir lo llaman los expertos.

Esto daría una clara explicación al porqué los *Centers* son claramente los menos favorecidos a la hora del reparto de las posesiones. Pero si los porcentajes por posición de todas las demás posiciones son prácticamente idénticas, ¿Por qué los *Small Forwards* han sufrido una caída tan contundente? Para esta cuestión no se encuentra una respuesta respaldada meramente con números a pesar de que sí se deben tener en cuenta. Se trata mayormente una cuestión de cómo han evolucionado los roles de las demás posiciones. No es que los *SF* hayan perdido mucho, es que el resto han ido ganando más y más. Ha sido a costa principalmente de los *Point Guards* y su conversión a una figura mucho más anotadora. Antaño la figura del *PG* se asociaba con el jugador que más juego repartía, el encargado de hacer llegar el balón al resto de jugadores para que estos tuvieran la mejor opción para anotar. Hoy en día es completamente distinto, los bases han pasado a ser referentes, estrellas, máquinas de anotación. Y, al fin y al cabo, esto es un juego de suma constante, a mismo número de tiros por equipo, para que unos tiren más, otros tienen que tirar menos.

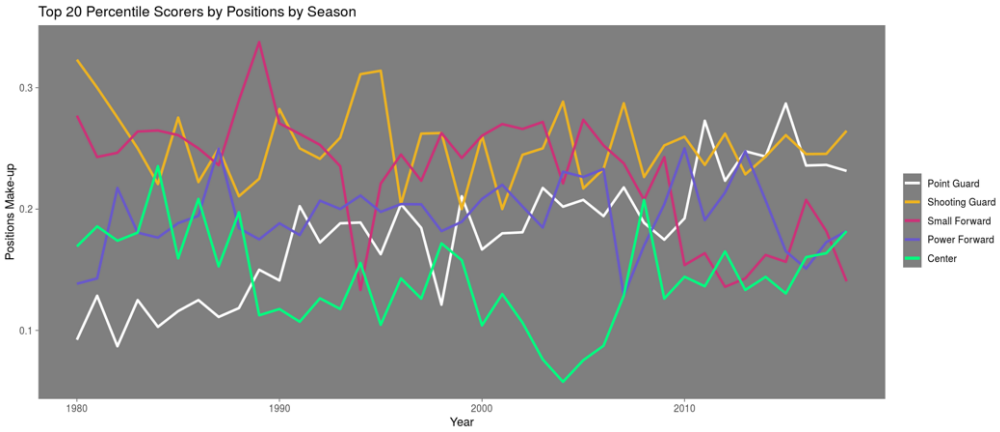


***Gráfico 9**: Porcentaje de jugadores por posición que forman parte del percentil 20 de mejores anotadores de la NBA a lo largo del tiempo. Fuente: NYC Data Science Academy.*

En la temporada 2018/2019 participaron un total de 530 jugadores por lo que, el 20% supone un total de 90. Basándose en el gráfico extraído de *NYC Data Science Academy* se puede decir que aproximadamente de estos 90 mejores anotadores 22 (24%) son *PG*, 23 (26%) son *SG*, 13 (14%) son *SF*, 16 (18%) son *PF* y 16 (18%) son *C* (véase Gráfico 9).

En la temporada 1979/1980 participaron un total de 287 jugadores. Aplicando el mismo razonamiento que se ha usado anteriormente se obtiene que aproximadamente de los 57 mejores anotadores 5 (9%) eran *PG*, 18 (32%) son *SG*, 16 (28%) eran *SF*, 8 (14%) eran *PF* y 10 (17%) eran *C* (véase Gráfico 9).

Las posiciones no son lo que eran.

1.7 Estado del Arte.

Pese a ser algo relativamente novedoso existen una gran cantidad de artículos, trabajos e investigaciones dedicadas al análisis de los aspectos técnicos y tácticos del juego. De hecho, desde hace ya varios años en USA y concretamente en *Boston*, tiene cada año una convención dedicada exclusivamente a este apartado: la *MIT Sloan Sports Analytics Conference*. Fue fundada en 2006 por *Daryl Morey*, del que ya se ha hablado previamente, y *Jessica Gelman*. En dicha convención se presentan trabajos sobre diversos deportes, pero con los que más se trabaja son: fútbol, baloncesto, béisbol y fútbol americano. Durante dos días se exponen los mejores trabajos recibidos por la organización y estos quedan publicados en su página web. Es cierto que existen multitud de otras fuentes en las que encontrar información sobre el análisis deportivo, y especialmente de la NBA por ser una liga tan famosa a nivel mundial, pero sí es cierto que, en toda la literatura accesible sobre el tema, las investigaciones más interesantes y especializadas se recopilan en este evento. Por lo tanto, este apartado va a tratar concretamente las publicaciones aportadas por esta organización a lo largo de los múltiples años que ha estado presente e impulsando el análisis deportivo.

Los 3 bloques más grandes en los que se ha centrado el análisis de datos aplicado a la NBA podrían resumirse en:

- Tiro.
- Toma de decisiones.
- Predicción de resultados y variables más importantes para ellos.

En el apartado de tiro se destacan tres trabajos especialmente relevantes. El primero *"How To Get An Open Shot (Lucey, Bialkowski, Carr, Yue & Matthews, 2014)"* da una explicación sobre las variables más relevantes a la hora de conseguir un tiro abierto (sin oposición) porque estos tiros tienen porcentajes de acierto significativamente más altos que los que sí están presionados. El segundo *"CourtVision: New*

Visual and Spatial Analytics for the NBA (Goldsberry, 2012)" formula dos nuevas variables que tienen en cuenta no sólo cómo de bien tira un jugador sino su versatilidad a la hora de tirar desde distintas localizaciones con el fin de encontrar los mejores tiradores basándose no sólo en números brutos. Y el tercero *"The Hot Hand: A New Approach to an Old Fallacy (Bocskocsky, Ezekowitz & Stein (2014)"* da evidencias ciéntificas para desmitificar el fenómeno conocido como "mano caliente", el cual asume que los distintos tiros que un mismo jugador hace son dependientes entre ellos, y que a una mayor racha de canastas conseguidas, más posibilidad existe de que el siguiente tiro también sea un éxito. Lo que todos estos trabajos tienen en común es que a través de la ciencia le dan un enfoque distinto a lo que toda la vida se ha dado por hecho sin evidencias que lo respaldaran.

Luego, la toma de decisiones es posiblemente la faceta en la que más se haya podido investigar más exhaustivamente, y todo ello gracias al *data tracking* y la capacidad de monitorizar el movimiento de los jugadores y los resultados que obtienen en las diferentes acciones que realizan. Hay un trabajo especialmente destacable en este campo *"Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data (Cervone, D'Amour, Bornn & Goldsberry (2014)"* se define una variable denominada "*Expected Possession Value* (*EPV*)" la cual mide los puntos esperados de esa posesión basándose en qué jugador tiene el balón, su localización en el campo y sus porcentajes de acierto en dicha localización; de esta manera es posible cuantificar las variaciones en esta variable dependiendo de las decisiones que tome el jugador en posesión del balón y poder determinar si la decisión ha sido correcta (El *EPV* aumenta) o incorrecta (El *EPV* baja).

Y en el tercer apartado, la predicción de resultados y variables más importantes para ellos también es posible encontrar un espectro muy amplio de investigaciones, aunque destaca una especialmente interesante: "*Are the ‹‹Four Factors›› Indicators of One Factor? (Baghal, 2012)"* que a su vez hace alusión al libro "*Basketball on Paper* (*Oliver, 2002*)" en el cual se habla de que los cuatro factores más importantes a la hora de la victoria de un equipo son: ***Effective Field Goald Percentage*** (variable que intenta ajustar mejor el porcentaje de acierto de un jugador o equipo ponderando que los tiros de 3 puntos valen más que los de 2), ***Free Throws Rate*** (una variable que mide el porcentaje del de puntos que provienen de los tiros libres respecto al total de puntos de un equipo), ***Turnover Per Possession*** (variable que mide el número de pérdidas por posesión de un equipo) y ***Offensive Rebounding Percentage*** (variable que mide el porcentaje de rebotes ofensivos conseguidos por el equipo de interés respecto a la suma de rebotes ofensivos conseguidos por el equipo de interés y los rebotes defensivos conseguidos por el otro equipo). *Baghal* en su estudio, y a través de modelos de regresión, incluye una perspectiva diferente y trata esos cuatro factores no para evaluar su implicación en la victoria del equipo, sino que hace una distinción entre esos factores en ataque y en defensa para poder medir su calidad por separado, además de añadir como factor el salario del equipo para comprobar que su impacto en la calidad del ataque y la defensa, así como en el porcentaje de victorias, es significativo, lo cual resulta ser así para la calidad ofensiva pero no la defensiva, aunque la que mayor peso tiene en la victoria también se descubre que es la ofensiva.

2. OBJETIVOS

Este trabajo tiene dos objetivos principales:

1. El primero consiste en analizar la evolución que han sufrido los jugadores de la NBA a lo largo del tiempo y, especialmente, cómo se ha visto reflejada esta evolución en sus posiciones y en los roles asociados a estas posiciones.

2. El segundo se basará en un análisis de todas las variables que pueden ser de interés en el momento de la realización de un tiro, con el fin de descubrir cuáles son las más significativas y que puedan servir como indicadores de su éxito. Todo ello teniendo en cuenta que no todos los tiros son iguales y, por lo tanto, no deben estar afectados por las mismas variables.

3. MATERIAL Y MÉTODOS

Para la realización de este trabajo se ha recurrido a diferentes técnicas estadísticas para el correcto análisis de los eventos de interés. De la misma forma, se han extraído diversas bases de datos sobre jugadores y tiros, sus estadísticas, variables y resultados.

Debido a la variedad en estas bases de datos y todo el tratamiento que se les ha aplicado, para una mejor comprensión cada una se detallará previamente a su respectivo análisis en la sección de Resultados.

Las técnicas estadísticas que se han empleado han sido técnicas de ***Clustering***, ***Clasificación*** y de ***Regresión Logística***.

3.1. Clustering

Los métodos de clustering son un conjunto de técnicas estadísticas que sirven para, a partir de un conjunto de datos, crear grupos que contengan individuos similares entre ellos, pero diferentes con los otros grupos, en base a una serie de variables de interés. Forma parte de las técnicas de aprendizaje no supervisado, ya que se desconoce la clase de la que forma parte cada individuo en el supuesto de que ésta existiera. A la hora de crear los grupos existen muchas formas de medir las distancias entre los individuos, pero las más conocidas y empleadas son la distancia euclídea y la distancia Manhattan.

La **distancia euclídea** es la suma cuadrática de las distancias entre cada observación respecto al centroide.

$$d_{euc}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

La **distancia Manhattan** funciona de una forma muy parecida, aunque se considera más robusta ya que no eleva las distancias al cuadrado y así se ve menos afectada por los posibles *outliers*.

$$d_{man}(p,q) = \sum_{i=1}^{n}|(p_i - q_i)|$$

Dentro de todas las técnicas de clustering que existen, en este trabajo se van a emplear en concreto 3 de las más utilizadas: ***K-Means Clustering***, ***Partitioning Around Medoids (PAM)*** y ***Hierarchical Clustering***.

3.1.1. K-Means

El método *K-Means* forma parte de un conjunto más amplio denominado *Partitioning Clustering,* en el cual es el usuario el que tiene que indicar el número de *clusters* a formar por el algoritmo a priori. La forma de actuar entonces una vez definido el número *K* de *clusters* que han de ser formados asegura que todas las observaciones van a ser asignadas a un solo *cluster*. De esta manera no puede quedar ninguna observación sin ser asignada y tampoco ser asignada a varios a la vez. El algoritmo encuentra los *K* mejores *clusters*, entendiendo como mejores clusters aquellos cuya varianza interna sea lo más pequeña posible. Para ello a cada *cluster* se le asigna un ***centroide***, el punto dentro del *cluster* que hace que las distancias de todas las observaciones de ese *cluster* a ese punto sean lo menor posible.

Considérese $C_1, C_2 \dots, C_K$ el conjunto de *K clusters* formados por el algoritmo. Tomando $x_1, x_2 \dots, x_i$ como el total de i observaciones que forman parte del C_k y sea μ_k el *centroide* del C_k, entonces para medir la varianza interna de cada *cluster* C_k las dos medidas más comunes son:

- La suma de las distancias euclídeas al cuadrado entre cada observación respecto al *centroide*.

$$W(C_k) = \sum_{x_i \in C_k}^{n} (x_i - \mu_k)^2$$

- La suma de las distancias euclídeas al cuadrado entre todos los pares de observaciones que forman el *cluster*, dividida entre el número de observaciones del *cluster*.

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Encontrar la varianza mínima de todos los *clusters* $\sum_{k=1}^{K} W(C_k)$ es un proceso complejo y aunque el algoritmo pueda no encontrar la división óptima, al menos sí será muy aproximada a ésta (máximo local).

Los pasos que sigue el algoritmo son los siguientes:

1. Asignación aleatoria de cada observación a uno de los *K clusters*.
2. Asignar a cada *cluster* formado un *centroide*.
3. Asignar cada observación al *cluster* cuyo *centroide* esté más cercano.
4. Repetir pasos 2 y 3 hasta que las observaciones no cambien de *cluster* o se alcance el número de iteraciones definidas por el usuario.

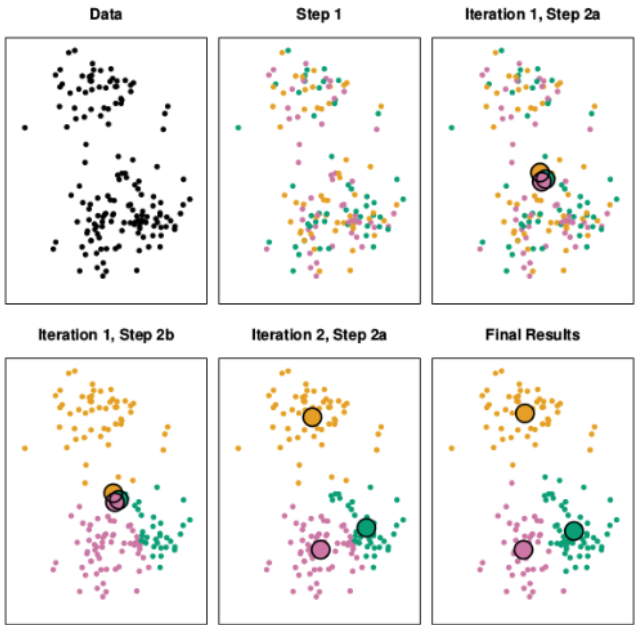


***Figura** 7: Funcionamiento del algoritmo K-Means. Fuente: RPubs.*

3.1.2. Partitioning Around Medoids (PAM)

El algoritmo *PAM* es el más empleado dentro de las técnicas de *K-Medoids*. Al igual que *K-Means* forma parte del conjunto de *Partitioning Clustering*. De hecho, funcionan de una forma muy parecida, solo que *K-Medoids* no emplea un *centroide* como *K-Means*, siendo éste un punto que optimiza las distancias a todas las observaciones, sino que utiliza lo que se denomina como *medoid*, que es lo mismo, pero en vez de un punto espacial, toma una observación en concreto.

Los pasos que sigue este algoritmo son:

1. Selecciona *K* observaciones aleatorias como *medoids* iniciales (aunque pueden ser especificados).
2. Calcula la matriz de distancias de todas las observaciones en caso de no haber sido calculada previamente.
3. Asigna cada observación al *medoid* más cercano.
4. Para cada uno de los *clusters* creados, comprobar si alguna otra observación tomada como *medoid* consigue minimizar la distancia promedio del *cluster*, si esto ocurre, seleccionar la observación que consigue una mayor reducción como nuevo *medoid*.
5. Si en el paso 4 para algún *cluster* su *medoid* ha cambiado volver al paso 3, de lo contrario finalizar el algoritmo.

3.1.3. Hierarchical Clustering

Las técnicas de *Hierarchical Clustering* forman parte de un conjunto de la misma magnitud que las técnicas de *Partitioning Clustering.* La diferencia consiste en que no es necesario establecer el número de *clusters* a priori, si no que, en base a las observaciones y sus variables, el propio algoritmo encuentra el número óptimo de *clusters*. Dentro de las técnicas de Hierarchical Clustering se distinguen dos aproximaciones operativas: las técnicas de *Agglomerative Clustering (bottom-up)* y las de *Divisive Clustering (top-down).*

Agglomerative Clustering toma primeramente cada observación como un *cluster* y las va agrupando hasta converger a una "rama central", mientras que el ***Divisive Clustering*** funciona al contrario, de un único *cluster* que comprende a todas las observaciones se va estructurando hasta llegar al punto en que cada observación supone un *cluster* diferente.

Los pasos que sigue el algoritmo del ***Agglomerative Clustering*** son:

1. Cada observación se toma como un *cluster* individual (hojas).
2. Se inicia un proceso iterativo hasta que todas las observaciones pasan a formar parte de un único *cluster*:
   2.1. Se calcula la distancia entre cada par posible de los *n clusters*. El usuario define la distancia y el *linkage* para medir la similitud entre observaciones y grupos.
   2.2. Los dos *clusters* más similares se fusionan dando lugar a *n-1 clusters*.
3. Determinar dónde cortar la estructura de árbol generada para establecer los *clusters*.

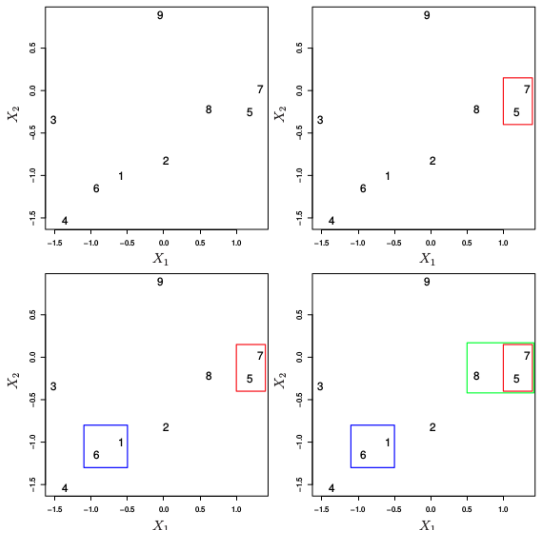


Figura 8*: Funcionamiento del algoritmo Agglomerative de Hierarchical Clustering. Fuente: Libro "An Introduction to Statistical Learning: With Applications in R".*

Los pasos que sigue el algoritmo del ***Divisive Clustering*** son:

1. Todas las *n* observaciones forman el mismo *cluster*.
2. Se inicia un proceso iterativo en el que hasta conseguir *n* clusters:
   2.1. Se calcula para cada *cluster* la mayor de las distancias entre pares de observaciones (diámetro del *cluster*).
   2.2. Se selecciona el *cluster* con mayor diámetro y para éste:
      2.2.1. Se calcula la distancia media de cada observación con las demás.
      2.2.2. La observación más alejada pasa a ser un nuevo *cluster*.
      2.2.3. Las anteriores observaciones se reasignan al nuevo *cluster* o al viejo dependiendo de cuál esté más próximo.

3.2. Clasificación

Los métodos de clasificación son técnicas estadísticas de aprendizaje supervisado que tienen el objetivo de determinar a qué grupo de la variable categórica de interés pertenece un individuo basándose en la información contenida en otros parámetros relacionados con esa variable. Existen multitud de técnicas de clasificación, pero en este trabajo se han utilizado las tres siguientes: ***Árboles de Clasificación***, ***Árbol C.5*** y ***K-Vecinos Más Cercanos.***

3.2.1. Árboles de Decisión

Los *árboles de decisión o clasificación* funcionan como algoritmos de particiones sucesivas. Es una técnica explicativa y descomposicional que utiliza un proceso de división secuencial, iterativo y descendente que, partiendo de una variable dependiente, forma grupos homogéneos definidos específicamente mediante combinaciones de variables independientes en las que se incluyen la totalidad de los casos recogidos en la muestra.

Se comienza con un nodo inicial dividiendo la variable dependiente a partir de una primera variable independiente escogida para dar lugar a dos conjuntos homogéneos de datos. Por ejemplo, de la variable X_1 se escoge un punto c de tal manera que las observaciones para las que el valor de su $X_1 \leq c$ compondrán un nuevo nodo y las observaciones para las que el valor de su $X_1 > c$ compondrán el otro nuevo nodo. En cada uno de estos nodos se repite el proceso de seleccionar una variable y un punto de corte para dividir la muestra y el proceso termina cuando todas las observaciones hayan sido clasificadas correctamente.

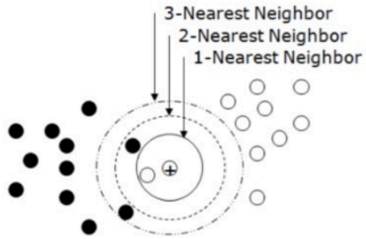
3.2.2. Árbol C5.0

El algoritmo *C5.0* es uno de los más usados en el ámbito de árboles de decisión. Su mecanismo se basa en la construcción de un árbol de forma descendente. Para ello toma como primer atributo el que mejor clasifica individualmente los datos de entrenamiento y pasa a ser la raíz del árbol. Entonces una rama y su nodo se crean para cada valor posible del atributo en cuestión. Los ejemplos de entrenamiento son repartidos en los nodos descendentes de acuerdo con el valor que tengan para el atributo de la raíz. Para los nodos generados se repite el mismo proceso en busca de los siguientes atributos más relevantes. Generalmente el algoritmo se detiene cuando los ejemplos de entrenamiento comparten el mismo valor para el atributo que está siendo probado. Sin embargo, es posible utilizar otros criterios para finalizar la búsqueda.

3.2.3. K-Vecinos Más Cercanos

El algoritmo de los *k-vecinos más cercanos* funciona de una manera sencilla: cada observación es clasificada en un grupo basándose en las *k* observaciones más cercanas a él. Es decir, calcula la distancia

de la nueva observación a cada una de las ya existentes, y ordena dichas distancias de menor a mayor para ir seleccionando el grupo al que pertenecer. Este grupo será, por tanto, el de mayor frecuencia con menores distancias. En la Figura 9 se muestra de forma visual el funcionamiento progresivo del algoritmo.



Web Data Mining, de Bing Liu

***Figura 9**: Funcionamiento del algoritmo de los k-vecinos más cercanos. Fuente: Web Data Mining.*

Para este ejemplo entonces:

- k = 1 el círculo con el signo "+" será clasificado como blanco.
- k = 2 el círculo con el signo "+" no tendría un criterio más fuerte que el otro para ser clasificado.
- k = 3 el círculo con el signo "+" será clasificado como negro.
- …

3.3. Regresión Logística

3.3.1. Regresión Logística Simple

El método de regresión logística simple es un método de regresión que sirve para estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. La clasificación es una de las principales aplicaciones que tiene esta técnica, con la que las observaciones se clasifican en un grupo u otro en base a una probabilidad obtenida a partir del valor de la variable independiente (también llamado predictor).

A pesar de poder ajustar un modelo de regresión lineal por mínimos cuadrados $\beta_0 + \beta_1 x$ codificando la variable cualitativa como 1 y 0, esta aproximación presenta el problema de que, al tratarse de una recta, para los valores extremos del predictor se obtienen valores de la variable dependiente menores que 0 o mayores que 1 y no pueden ser tomados como probabilidades. Para evitar estos problemas la regresión logística transforma este valor generado por la regresión lineal empleando una función que siempre produce valores comprendidos entre 0 y 1. Una de las funciones más empleadas es la función logística o sigmoide:

$$\text{función sigmoide} = \sigma(x) = \frac{1}{1+e^{-x}}$$

Entonces sustituyendo la *x* de esta función por la función lineal ($\beta_0 + \beta_1 x$) se obtiene que:

$$P(Y = k|X = x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}} = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

donde $P(Y = k|X = x)$ se interpreta como la probabilidad de que la variable cualitativa *Y* tome el valor *k* dado que el predictor *X* tome el valor *x*.

Empleando su versión logarítmica se puede ajustar esta función de forma sencilla con métodos de regresión lineal, obteniendo así lo que se conoce como *LOG of ODDs*:

$$\ln\left(\frac{P(Y=k|X=x)}{1-P(Y=k|X=x)}\right)=\beta_0+\beta_1 x$$

En regresión lineal simple el valor de la variable dependiente *Y* adquiere su valor en función al de la variable independiente *X*. Sin embargo, en regresión logística lo que se obtiene es la probabilidad de que la variable respuesta *Y* pertenezca al nivel de referencia 1 en función del valor que adquiera el predictor *X* y mediante el uso de *LOG of ODDs*.

Supóngase que la probabilidad de que un evento sea verdadero es de 0.666, por lo que la probabilidad de evento falso es de 1 - 0.666 = 0.333. Los *ODDs* o razón de probabilidad de verdadero se definen como el ratio entre la probabilidad de evento verdadero y la probabilidad de evento falso p / q. En este caso los *ODDs* de verdadero son 0.666 / 0.333 = 2, lo que equivale a decir que se esperan 2 eventos verdaderos por cada evento falso.

La trasformación de probabilidades a *ODDs* es monotónica, si la probabilidad aumenta también lo hacen los *ODDs*, y viceversa. El rango de valores que pueden tomar los *ODDs* es de [0,∞]. Dado que el valor de una probabilidad está acotado entre [0,1] se recurre a una trasformación logit (existen otras) que consiste en el logaritmo natural de los *ODDs*. Esto permite convertir el rango de probabilidad previamente limitado a [0,1] a [−∞,+∞].

Los *ODDs* y el logaritmo de *ODDs* cumplen que:

- Si p(verdadero) = p(falso), entonces odds(verdadero) = 1
- Si p(verdadero) < p(falso), entonces odds(verdadero) < 1
- Si p(verdadero) > p(falso), entonces odds(verdadero) > 1

A diferencia de la probabilidad que no puede exceder el 1, los *ODDs* no tienen límite superior.

- Si odds(verdadero) = 1, entonces logit(p) = 0
- Si odds(verdadero) < 1, entonces logit(p) < 0
- Si odds(verdadero) > 1, entonces logit(p) > 0
- La transformación logit no existe para p = 0

Una vez obtenida la relación lineal entre el logaritmo de los *ODDs* y la variable predictora *X*, se tienen que estimar los parámetros β_0 y β_1. La combinación óptima de valores será aquella que tenga la máxima verosimilitud (*maximum likelihood*), es decir el valor de los parámetros β_0 y β_1 con los que se maximiza la probabilidad de obtener los datos observados.

Existen diferentes técnicas estadísticas para calcular la significación de un modelo logístico en su conjunto (*p-value* del modelo). Todos ellos consideran que el modelo es útil si es capaz de mostrar una mejora respecto a lo que se conoce como modelo nulo, el modelo sin predictores, sólo con β_0. Dos de los más empleados son:

- ***Wald chi-square***: La hipótesis nula que plantea este método es que la variable dependiente tiene un valor distinto de 0, por lo que rechazar esta hipótesis nula sugiere que las variables independientes empleadas pueden ser excluidas sin que el modelo se vea alterado significativamente.

- ***Likelihood ratio***: Usa la diferencia entre la probabilidad de obtener los valores observados con el modelo logístico creado y las probabilidades de hacerlo con un modelo sin relación entre las variables. Para ello, calcula la significación de la diferencia de residuos entre el modelo con predictores y el modelo nulo. El estadístico tiene una distribución Chi-cuadrado con grados de libertad equivalentes a la diferencia de grados de libertad de los dos modelos comparados. Si se compara respecto al modelo nulo, los grados de libertad equivalen al número de predictores del modelo generado.

Para determinar la significancia individual de cada uno de los predictores introducidos en un modelo de regresión logística se emplea el **estadístico *Z*** y el ***Wald chi-test***.

A diferencia de la regresión lineal, en la que β_1 se corresponde con el cambio promedio en la variable dependiente *Y* debido al incremento en una unidad del predictor *X*, en regresión logística, β_1 indica el cambio en el logaritmo de *ODDs* debido al incremento de una unidad de *X*, o lo que es lo mismo, multiplica los *ODDs* por e^{β_1}.

Una vez estimados los coeficientes del modelo logístico, es posible conocer la probabilidad de que la variable dependiente pertenezca al nivel de referencia, dado un determinado valor del predictor. Para ello se emplea la ecuación del modelo:

$$\hat{p}(Y = 1|X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

Para conseguir clasificar las observaciones entonces es necesario establecer un *threshold* de probabilidad a partir de la cual se considera que la variable pertenece a uno de los niveles. Por ejemplo, eligiendo un *threshold* de 0.5 se puede asignar una observación al grupo 1 si $\hat{p}(Y = 1|X) > 0.5$ y de lo contrario al grupo 0.

A diferencia de los modelos de regresión lineal, en los modelos logísticos no existe un equivalente a R^2 que determine exactamente la varianza explicada por el modelo. Se han desarrollado diferentes métodos conocidos como $pseudoR^2$ que intentan aproximarse al concepto de R^2 pero que, aunque su rango oscila entre 0 y 1, no se pueden considerar equivalentes.

- ***McFadden's***: $R^2_{McF} = 1 - \frac{ln\hat{L}(modelo)}{ln\hat{L}(modelo\ nulo)}$, siendo $\hat{L}$ el valor de likelihood de cada modelo. La idea de esta fórmula es que, $ln(\hat{L})$, tiene un significado análogo a la suma de cuadrados de la regresión lineal. De ahí que se le denomine $pseudoR^2$.

- Otra opción bastante extendida es el test de ***Hosmer-Lemeshow***. Este test examina mediante un *test chi-cuadrado de Pearson* si las proporciones de eventos observados son similares a las probabilidades predichas por el modelo, haciendo subgrupos.

3.3.2. Regresión Logística Múltiple

La regresión logística múltiple es una extensión de la regresión logística simple. Sigue los mismos principios que ésta, pero ampliando el número de predictores, que pueden ser tanto continuos como categóricos.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$$

$$logit(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i$$

Y por lo tanto el valor de la probabilidad de que *Y* sea igual a 1 puede calcularse con la inversa del logaritmo natural.

$$P(Y = 1) = \frac{e^{\beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i}}{1 + e^{\beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i}}$$

4. RESULTADOS

4.1. Análisis Del Cambio En El Rol De Los Jugadores Por Su Posición

Primeramente, se va a analizar la evolución con el paso de los años de los roles por posición de los jugadores. Para este propósito se utilizarán las técnicas de *clustering*. Como se ha explicado previamente en la sección 1.6.2, cada jugador tiene asociada una posición de las 5 posibles dependiendo de sus características, capacidades y su rol dentro del equipo. Por lo tanto, con las técnicas de *clustering* **lo que se quiere comprobar es si existen, o han existido, patrones que diferencien a los jugadores por sus posiciones tomando como referencia las estadísticas de sus tiros**. En pocas palabras, si los jugadores de cada posición comparten un patrón de tiros similar.

En "*Scoring and Shooting Abilities of NBA Players (Piette, Anand & Zhang, 2010)*" se definen diferentes variables para describir el rendimiento de los jugadores en base a sus habilidades de tiro y anotación. Lo que se va a plantear en este análisis es, con las variables ya existentes, ver la evolución temporal de los jugadores considerando las variables existentes.

La base de datos utilizada para este apartado ha sido extraída de la página web de *basketball-reference*, uno de los directorios más fiables y populares sobre estadísticas del baloncesto estadounidense. Se han seleccionado las temporadas 1979/1980, 1994/1995, 2009/2010 y 2018/2019, y para cada una de ellas se han recopilado las siguientes variables de todos sus jugadores (en totales de la temporada):

- ***Player****:* Nombre del jugador.
- ***Position*** (***Pos****):* Posición del jugador.
- ***Minutes Played*** (***MP****):* Minutos disputados por el jugador.
- ***3 Points Made*** (***3P****):* Tiros de 3 puntos convertidos por el jugador.
- ***3 Points Attempted*** (***3PA****):* Tiros de 3 puntos intentados por el jugador.
- ***2 Points Made*** (***2P****):* Tiros de 2 puntos convertidos por el jugador.
- ***2 Points Attempted*** (***2PA****):* Tiros de 2 puntos intentados por el jugador.
- ***Free Throws Made*** (***FT****):* Tiros libres convertidos por el jugador.
- ***Free Throw Attempted*** (***FTA****):* Tiros libres intentados por el jugador.

Y se han definido las siguientes variables para tener en cuenta la diferencia de minutos que ha disputado cada jugador:

- ***3 Points Made Per Minute*** (***3P/MIN****):* Tiros de 3 puntos convertidos por minuto.
- **3** ***Points Attempted Per Minute*** (***3PA/MIN****):* Tiros de 3 puntos intentados por minuto.
- ***2 Points Made Per Minute*** (***2P/MIN****):* Tiros de 2 puntos convertidos por minuto.
- ***2 Points Attempted Per Minute*** (***2PA/MIN***): Tiros de 2 puntos intentados por minuto.

Esta forma de homogeneizar los datos es simple, pero adecuada porque es evidente que hay muchos sesgos que pueden venir determinados por el tiempo de juego.

También se han llevado a cabo una serie de modificaciones con el objetivo de formar un *set* de datos más robusto con el que poder trabajar. En primer lugar, se han eliminado todos los jugadores que hayan

disputado menos de 1000 minutos en una misma temporada, lo que equivaldría generalmente a los jugadores menos importantes en el equipo. Para los jugadores que han formado parte de varios equipos durante la misma temporada, se ha tenido en cuenta las estadísticas totales de la temporada. Por último, pese a que los jugadores denominados con dos posiciones constituyen un porcentaje muy reducido, sólo se ha tenido en cuenta la posición en la que han disputado más minutos.

Una vez realizado todo ello se ha acabado con un total de 197 jugadores para la temporada 1979/1980, 227 jugadores para la temporada 1994/1995, 252 jugadores para la temporada 2009/2010 y 276 jugadores para la temporada 2018/2019.

Además, para cada temporada y a través de ***R Studio*** (software que se utilizará para el desarrollo de todas las técnicas estadísticas de este trabajo) se han establecido 4 *subsets* de datos diferentes combinando las variables disponibles de cada temporada considerada:

1. *Subset 1:* ***Todos los tiros*** ("*Player*", "*Pos*", "*3P*", "*2P*", "*FT*", "*3PA*", "*2PA*", "*FTA*") para ver las diferencias usando el volumen de todos los tiros posibles, intentados y convertidos, de cada jugador.

2. *Subset 2:* ***Tiros por minuto*** ("*Player*", "*Pos*", "*3P/MIN*", "*2P/MIN*", "*3PA/MIN*", "*2PA/MIN*") para ver las diferencias usando sólo los tiros, intentados y convertidos, ajustados a los minutos disputados por cada jugador.

3. *Subset 3****: Tiros intentados*** ("*Player*", "*Pos*", "*3PA*", "*2PA*", "*FTA*") para ver las diferencias usando sólo los tiros intentados de cada jugador.

4. *Subset 4:* ***Triples*** ("*Player*", "*Pos*", "*3PA*", "*3PA/MIN*") para ver las diferencias usando sólo el volumen de todos los tiros de tres puntos, intentados e intentados por minuto, de cada jugador.

En total para los 16 *subsets* que se han construido se han llevado a cabo todas las técnicas de *clustering* citadas anteriormente. Sin embargo, en el cuerpo principal de este trabajo sólo aparecerán las figuras que hayan llevado a resultados más significativos o sean de verdadero interés para la validación de las hipótesis. El resto de las figuras, al igual que todo el código empleado en su obtención, vendrán incluidas de forma digital en el anexo.

El primer paso que se ha realizado para todos los *subsets* ha sido comprobar el número óptimo de *clusters* necesarios utilizando como criterio la distancia euclídea. Para el *subset* de triples en 1980 el gráfico que se genera es el siguiente (Gráfico 10).

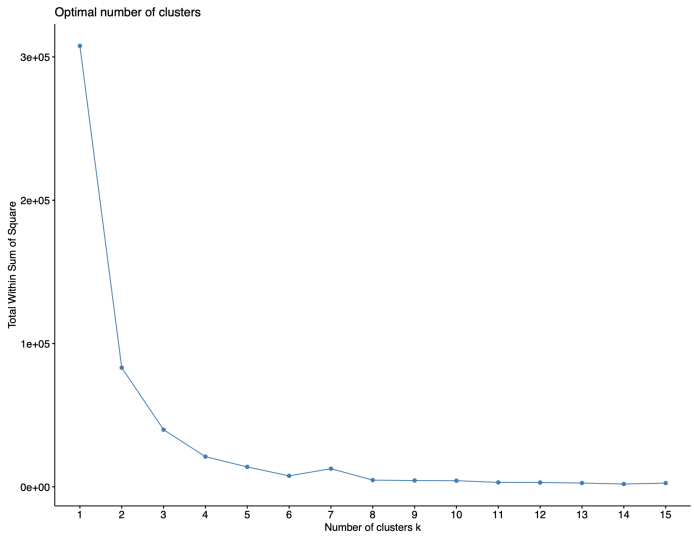


Gráfico 10*: Gráfico de la suma de cuadrados de los errores y los posibles clusters a ser formados para el subset de triples de la temporada 1979/1980. Generado con R Studio.*

Se puede comprobar que para este ejemplo bastaría con generar 3 *clusters*. Sin embargo, y aunque la suma de cuadrados de los errores que se reduce es poco significativa, es más interesante de cara al análisis que se está realizando crear 5 *clusters* por cuestiones metodológicas.

Pasando a analizar el primer evento de interés, y que da continuación a lo explicado en la introducción: ¿Cómo ha afectado a los jugadores el tiro de 3 puntos? Para este análisis se utilizará de cada temporada el *subset* de triples y como técnica de *clustering* un *k-means* con $k = 5$.

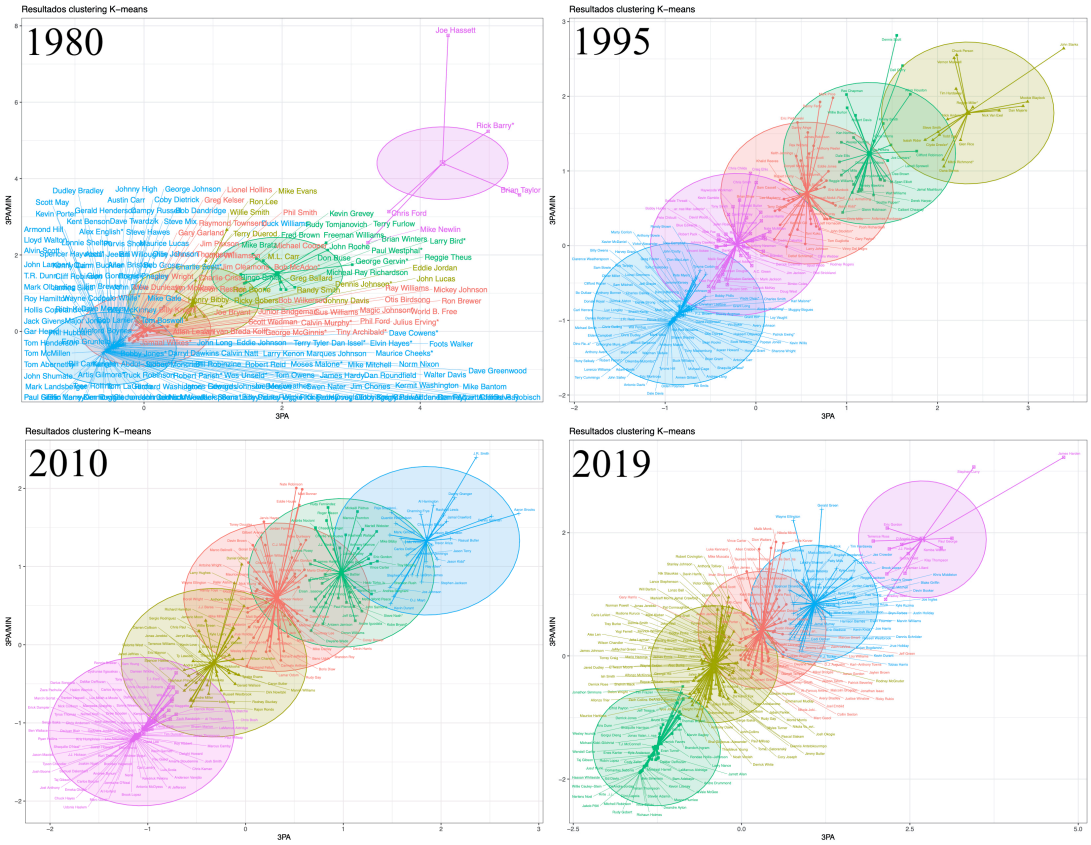


***Gráfico 11**: Representación del método k-means con k = 5 para las cuatro temporadas de interés con los subsets de triples. Generado con R Studio.*

Lo primero que se puede ver en este gráfico (Gráfico 11) es la diferencia sustancial que existe entre la temporada 1979/1980, la cual engloba a casi todos sus individuos en el *cluster* con menos tiros, y todas las demás. La explicación más lógica a este fenómeno se encuentra en que, como se ha explicado en la sección 1.6.1, aquella fue la primera temporada en la que se incorporó el tiro de tres puntos, por lo que es de esperar que fuera un aspecto del juego menos desarrollado. El mismo razonamiento da una explicación al hecho de que exista un *cluster* con sólo 5 individuos, que muestra el carácter incipiente de la importancia de reforzar este tipo de tiro.

También se puede comprobar que, con el paso del tiempo y hasta 2010, los grupos se van homogeneizando poco a poco. Por último, en 2019 se puede ver el efecto contrario, ahora se aprecia un grupo que se desmarca del resto por su alta especialización en este apartado, liderado además por *James Harden*, jugador de los *Houston Rockets*, el equipo pionero en la revolución del triple. Además, si se observa bien, ahora el grupo que contiene más individuos ya no es el más bajo de la gráfica que corresponde con los jugadores que menos triples asumen: ahora la gran mayoría de jugadores saben tirar bien desde la línea de tres puntos (o por lo menos en un volumen mucho mayor).

Con esto se puede ver que, como se ha explicado en la introducción, el cambio en la perspectiva del tiro de 3 puntos ha obligado a los jugadores a adaptarse a este estilo de juego, pero ¿se puede también encontrar similitudes en las variables de tiro de jugadores de una misma posición?

Ya que en 1980 todavía era muy pronto para tener en cuenta el triple como un factor determinante y en 2019 los jugadores de todas las posiciones se han adaptado lo suficiente como para que no sea tampoco algo diferencial, el análisis de esta cuestión sólo se va a realizar para las temporadas 1994/1995 y 2009/2010. En concreto se van a emplear los *subsets* de todos los tiros.

Para ello se ha realizado un *PAM* a cada *subset* de datos. Se ha comprobado qué jugadores se toman como *medoids* y qué posición ocupan, y cómo de buena es la clasificación del resto de jugadores en los *clusters* formados. A continuación, se incluyen los gráficos de ambos años para la técnica *PAM* (Gráfico 12).

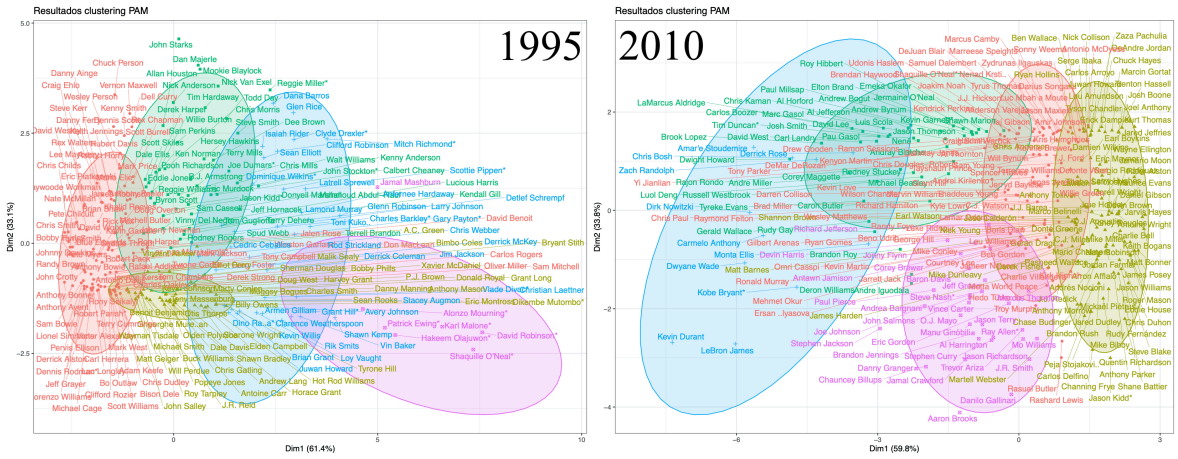


***Gráfico 12**: Gráfico del método PAM para los subsets de todos los tiros de las temporadas 1994/1995 y 2009/2010. Generado con R Studio.*

En la temporada 1994/1995 se encuentran como *medoids* a: *Winston Garland* (*PG*), *Jim Jackson* (*SG*), *Chris Mills* (*SF*), *Karl Malone* (*PF*) y *Tony Massenburg* (*C*). Cada *medoid* corresponde a una posición diferente, pero aún se tiene que comprobar cómo se ajusta cada *cluster* a esas posiciones. Los *clusters* que se han formado están demasiado descompensados, así que sólo se deberían considerar los 3 centrales, ya que el primero y que más individuos contiene es una mezcla de frecuencias muy similares de todas las posiciones, y el quinto sólo contiene 7 individuos por lo que no es representativo.

	grupo real				
cluster	C	PF	PG	SF	SG
1	14	18	21	17	18
2	18	15	5	4	6
3	1	2	17	10	15
4	3	12	5	13	6
5	5	1	0	1	0

***Tabla 2**: Tabla de asignación de clusters por el método PAM a cada posición para los jugadores de la temporada 1994/1995 y el subset de todos los tiros. Generada con R Studio.*

Observando el resto sí se aprecian diferencias refuerzan la hipótesis de la que se ha partido. El *cluster* número 2 contiene mayoritariamente jugadores que ocupan posiciones de jugadores interiores (*PF* y *C*). El *cluster* número 3, al contrario, contiene mayoritariamente a los jugadores exteriores (*PG*, *SG* y *SF*). Y el *cluster* número 4, aunque no se ajusta tan bien a la idea previa, sí que marca una diferencia al contener mayoritariamente a jugadores intermedios (*SF* y *PF*) lo que podría deberse a ser jugadores versátiles que podrían ocupar varias posiciones distintas. Para visualizar este análisis se presenta la tabla en la que se muestra la asignación de los *clusters* a los jugadores por su posición (Tabla 2).

Algo muy similar ocurre con la temporada 2009/2010. Se encuentran como *medoids* a *Jason Terry* (*SG*), *Mike Miller* (*SG*), *Carmelo Anthony* (*SF*), *Jonas Jerebko* (*PF*), y *Pau Gasol* (*C*). Esta vez se tiene un

medoid para cada posición excepto la de *Point Guard*, aunque la compensación se encuentra justo en la siguiente posición, en la que hay 2 *Shooting Guards.* Sin embargo, los *clusters* que se crean siguen estando desproporcionados e igualmente sólo hay 3 que nos muestren diferencias significativas, el número 3, que correspondería con los jugadores interiores y los números 5 y 2 que corresponderían mayoritariamente con los jugadores exteriores. De la misma forma que para la temporada 1994/1995 la información queda recogida en la siguiente tabla (Tabla 3).

	grupo real				
cluster	C	PF	PG	SF	SG
1	14	24	18	14	14
2	14	7	24	21	18
3	16	12	5	6	2
4	1	3	1	3	3
5	0	4	9	5	14

Tabla 3*: Tabla de asignación de clusters por el método PAM a cada posición para los jugadores de la temporada 2009/2010 con las variables de todos los tiros. Generada con R Studio.*

Con el fin de encontrar una mejor adaptación de los *clusters* a las posiciones se comprobó que el *subset* de tiros por minuto para la temporada 1994/1995 cumplía de una manera mucho más precisa con este objetivo. El resultado del *PAM* para este conjunto de datos fue el siguiente (Gráfico 13).

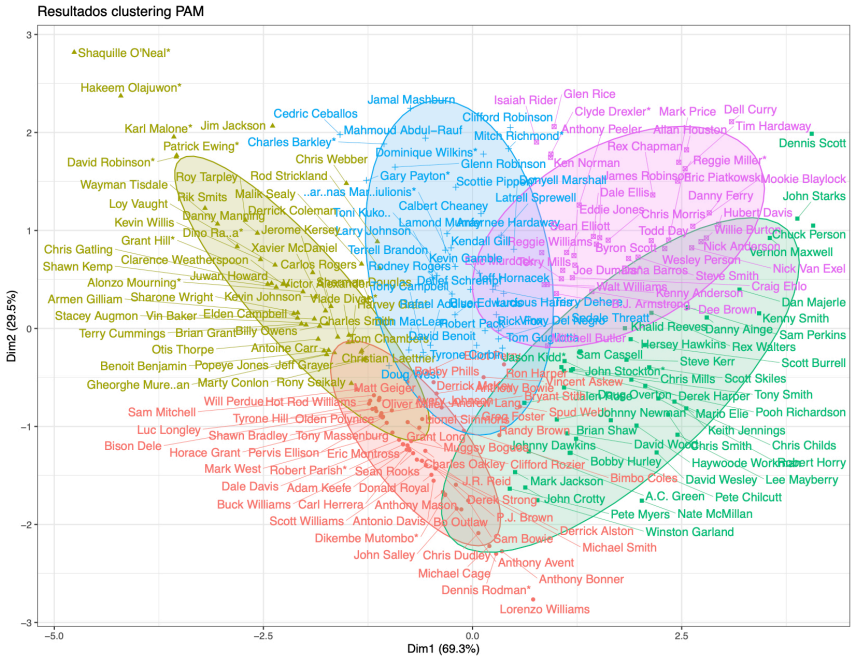


Gráfico 13*: Gráfico del método PAM para el subset de tiros por minuto de la temporada 1994/1995. Generado con R Studio.*

Los medoids resultantes para este subset fueron *Derek Harper* (*PG*), *Šarūnas Marčiulionis* (*SG*), *Chris Morris* (*SF*), *J.*R. Reid (*PF*) y *Vin Baker* (*PF*). Al igual que con el anterior análisis no se obtiene un *medoid* por cada posición, pero se obtienen 2 *Power Forwards* y ningún *Center*, otra vez la diferencia es entre posiciones adyacentes. Se puede comprobar que los *clusters* formados son mucho más homogéneos en cuanto al número de individuos que forman parte de cada uno. Además, si se analizan puede comprobarse que, aunque no clasifican correctamente cada posición concreta, se aprecia muy bien la diferencia entre jugadores interiores y exteriores. Los dos primeros *clusters* están compuestos mayoritariamente por jugadores interiores (*PF* y *C*). El tercero es el más heterogéneo y podría contener a los jugadores más versátiles que pueden desempeñar distintas posiciones. Y por fin, los dos últimos claramente engloban a los jugadores exteriores (*PG*, *SG* y *SF*).

A continuación, se presenta la tabla de asignación de los *clusters* (Tabla 4).

	grupo real				
cluster	C	PF	PG	SF	SG
1	13	13	3	3	4
2	15	11	2	3	2
3	13	15	8	10	3
4	2	4	18	22	19
5	2	7	26	11	23

***Tabla 4**: Tabla de asignación de clusters por el método PAM a cada posición para los jugadores de la temporada 1994/1995 con las variables de tiros por minuto. Generada con R Studio.*

Por último y para seguir en esta línea se presenta un análisis mediante un *Hierarchical Clustering*. Se examina cómo se van formando las ramas del árbol dependiendo del número de clusters generados y si las ramas reproducen los roles de interior y exterior. Para este propósito, en vez de sólo poner el foco en 5 grupos distintos, parece adecuado seleccionar las diferencias con solamente 2 clusters. Para esta técnica se utilizará el último *subset* de datos para las cuatro temporadas, el de tiros intentados. Se requerirá incluir una variable binaria nueva para este subset a la que se denominará como "EXT/INT" y que definirá si un jugador es exterior (*PG, SG, SF*) o si es interior (*PF, C*).

Analizando los resultados de las diferentes temporadas, ninguna de ellas cumple con esta idea de la separación de patrones basándose en los tiros intentados entre interiores y exteriores (ver Gráficos 6, 20, 33 y 48 del anexo). Sin embargo, siguiendo el mismo razonamiento y volviendo a analizar la primera idea de la que se partía, se puede comprobar cómo la especialización por parte de los jugadores de todas las posiciones ha hecho que la gran mayoría de los exteriores, así como más de la mitad de los interiores sean incluidos en el mismo *cluster*. A continuación, se presenta el dendograma generado a partir de un *Hierarchical Clustering* para el subset de tiros intentados de la temporada 2018/2019 (Gráfico 14). Se ha generado una división en 5 clusters con diferentes colores y adicionalmente se han incluido con líneas la partición que supondría para crear en 2 (línea superior) o 5 grupos (línea inferior).

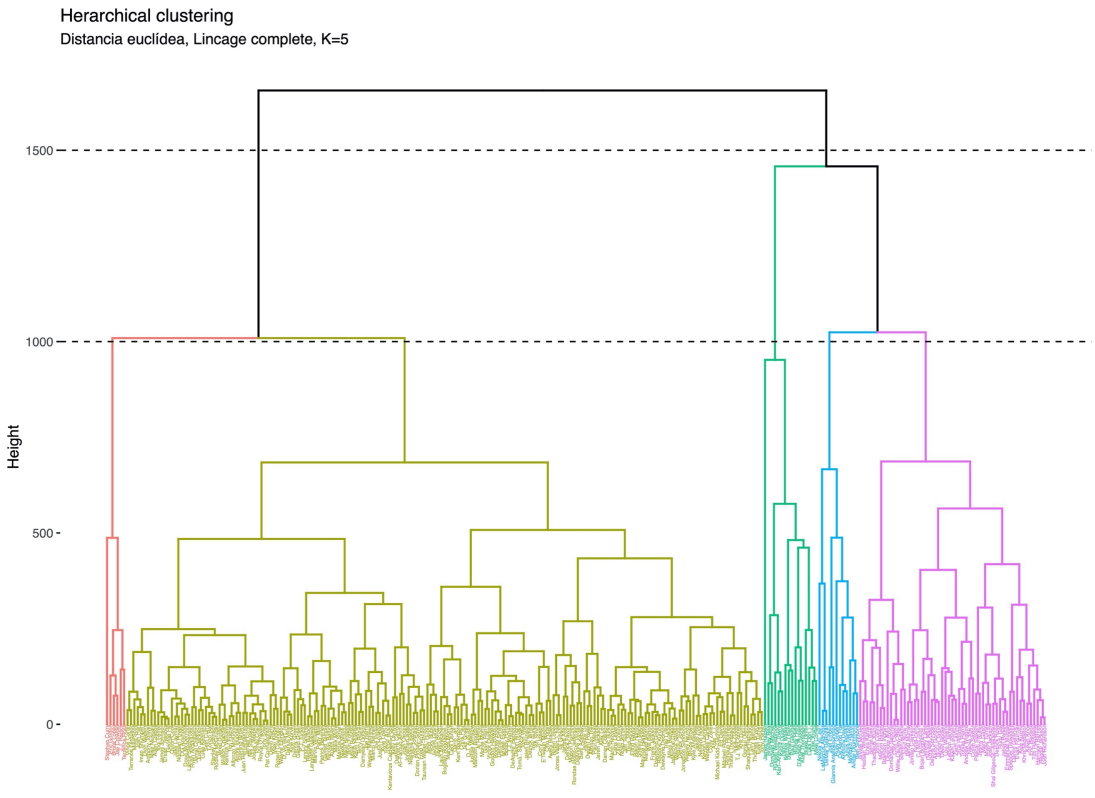


***Gráfico 14**: Dendograma generado a partir de un Hierarchical Clustering para el subset de tiros intentados de la temporada 2018/2019. Generado con R Studio.*

Al igual, se incluyen las tablas de asignación a *cluster* por posiciones y por categoría de exterior/interior (Tabla 5).

grupo real					
cluster	C	PF	PG	SF	SG
1	7	2	2	0	1
2	13	11	10	9	12
3	1	1	7	2	5
4	1	0	1	1	3
5	26	37	38	35	51

grupo real		
cluster	EXT	INT
1	48	35
2	129	64

Tabla 5*: Tablas de asignación para 5 clusters por posición y 2 clusters por categoría exterior/interior para el subset de tiros intentados de la temporada 2018/2019. Generado con R Studio.*

4.2. Análisis Comparativo De Las Posiciones Por El Tipo De Variables

El objetivo de esta línea de análisis es poder encontrar similitudes entre las posiciones de los jugadores a través de sus estadísticas en los tiros, esta vez empleando técnicas de clasificación. Para un ajuste mejor en la clasificación **también se realiza un estudio de las variables no relacionadas con el tiro, pero con un impacto importante en el rol que un jugador tiene asociado a su posición**.

Las bases de datos empleadas en este apartado también han sido extraídas de la web de *basketball-reference.* Se han seleccionado tres temporadas de referencia que serán las utilizadas para comprobar los modelos que han sido generados: 1994/1995, 2009/2010 y 2018/2019.

Se han generado 3 modelos y todos ellos constarán de datos de 4 temporadas. El primero, con los datos de las temporadas 1990/1991 – 1993/1994. El segundo, con las temporadas 2005/2006 – 2008/2009. Y el tercero, con las temporadas 2013/2014 – 2017/2018. Como en un conjunto de cuatro temporadas consecutivas la mayoría de los jugadores son los mismo, se ha seleccionado cada jugador de cada temporada como un individuo distinto.

La selección de los datos es la misma que para el apartado anterior. Así el set de datos utilizado en la formación del primer modelo ha quedado constituido por 919 individuos, el segundo por 1289 y el tercero por 1364. Las temporadas de referencia mantienen los mismos individuos que en el apartado de *clustering* (Sec. 4.1): 227, 252 y 276 respectivamente.

Todos los modelos generados serán utilizados para clasificar cada una de las temporadas de referencia. El objetivo es ver las diferencias que existen entre la clasificación con un modelo para su respectiva siguiente temporada y la comparación a la hora de clasificar las temporadas de años alejados en el tiempo.

Para el análisis de tiro se utilizarán las mismas las mismas variables que en el apartado anterior, incluyendo el porcentaje de acierto para cada tipo de tiro (*2P%, 3P%* y *FT%)* y excluyendo las variables de tiros de 2 puntos convertidos por minuto y tiros de 3 puntos convertidos por minuto (*2P/MIN* y *3P/MIN*).

Para el análisis no relacionado estrictamente con tiro se han seleccionado, además de los minutos disputados y la posición de cada jugador, las siguientes variables de estadística avanzada:

- ***Player Efficiency Rate*** (***PER***): Una medida de producción por minuto estandarizada de tal manera que el promedio de la liga es de 15 y tiene en cuenta un conjunto de apartados a los que puede contribuir un jugador.

- ***True Shooting Percentage*** (***TS%***): Una medida de tiro que trata de englobar los tiros de 2 puntos, de 3 puntos y tiros libres asignándoles diferentes pesos.

- ***3-Point Attempt Rate*** (***3PAr***): Proporción de triples que un jugador tira respecto del total de tiros de campo que hace (*3PA / (2PA + 3PA)*).

- ***Free Throw Attempt Rate*** (***FTr***): Proporción de tiros libres que por cada tiro de campo que realiza un jugador.

- ***Offensive Rebound*** **Percentage** (***ORB%***): Estimación del porcentaje de rebotes ofensivos que un jugador ha capturado respecto a la suma de rebotes ofensivos de su equipo más los rebotes defensivos de los rivales teniendo en cuenta los minutos que ha disputado el jugador.

- ***Defensive Rebound Percentage*** (***DRB%***): Estimación del porcentaje de rebotes ofensivos que un jugador ha capturado respecto a la suma de rebotes ofensivos de su equipo más los rebotes defensivos de los rivales mientras él ha estado en la pista.

- ***Total Rebound Percentage*** (***TRB%***): Estimación del porcentaje de rebotes defensivos que un jugador ha capturado respecto a la suma de rebotes defensivos de su equipo más los rebotes ofensivos de los rivales mientras él ha estado en la pista.

- ***Assist Percentage*** (***AST%***): Estimación del porcentaje de asistencias que ha repartido un jugador respecto al total del equipo mientras él ha estado en la pista.

Los primeros resultados que se van a utilizar son los correspondientes al tiro. Para empezar, se comprueba cómo se ajusta cada modelo a su temporada de referencia (la siguiente a sus datos para generarlo). Para ello se lleva a cabo su clasificación a través de las técnicas de *árboles de decisión* y *k-vecinos más cercanos* explicadas en la sección 3.2. Aunque se hayan aplicado ambas técnicas para todos los datos, sólo se mostraran las que consigan una mejor exactitud. El resto de las tablas, gráficos y código que no sea incluido en el cuerpo principal del trabajo se presentará en su anexo.

Para comprobar las variables estadísticamente significativas que intervienen en cada modelo, se han realizado *tests ANOVA* para cada uno de ellos. Al encontrar que la gran mayoría de las variables sí lo son para todos los casos, y que el incluir las que no lo son no implicaba un sobreajuste ni perjudicaba los resultados, no se ha excluido ninguna de ellas.

Al aplicar los modelos generados por ambas técnicas con sus respectivos datos de entrenamiento a la temporada 1994/1995 se comprueba que con los *árboles de decisión* se obtiene una mejor clasificación. Se presenta la matriz de confusión entre la posición en la que han sido clasificados los jugadores y su posición real (Tabla 6).

Con esta matriz de confusión la precisión (*accuracy*) del modelo es sólo del 33,92%. ¿Este valor tan bajo implica que el modelo no es suficientemente bueno? ¿Puede ser que las variables de tiro no impliquen un papel determinante en relación con la posición que ocupa un jugador? La respuesta a estas preguntas necesita un grado mayor de profundidad en el análisis. La interpretación debe refinarse teniendo en cuenta los resultados que hemos obtenido con las técnicas de *clustering*. No sería del todo justo para el modelo que su exactitud se midiese tan sólo con las posiciones que clasifica correctamente, y menos cuando se considera una tabla de 5x5.

```
          Reference
Prediction  C PF PG SF SG
        C  18 15  0  1  0
        PF 11  7  0  2  0
        PG  7 14 41 29 35
        SF  5  8  1  4  3
        SG  0  4  6  9  7

Overall Statistics

               Accuracy : 0.3392
                 95% CI : (0.2779, 0.4048)
    No Information Rate : 0.2115
    P-Value [Acc > NIR] : 5.801e-06

                  Kappa : 0.1698
```

Tabla 6*: Matriz de confusión entre la posición real y la predicción creada por el modelo basado en el árbol de decisión para las variables de tiro en la temporada 1994/1995. Generada con R Studio.*

Con el fin de encontrar una solución a este inconveniente se ha decidido tener en cuenta que la clasificación de una posición adyacente a otra no significa que sea errónea, por las similitudes que comparten estas posiciones. Para definir qué posiciones son adyacentes entre sí, se ha tenido en cuenta

la media del volumen de tiros de 3 puntos que realizan los jugadores de cada posición. Por lo tanto, el orden de las posiciones establecido a través de este parámetro sería: *Shooting Guard*, *Point Guard*, *Small Forward*, *Power Forward* y *Center*. De esta manera sería correcto clasificar a cualquier *SG* como *PG* y viceversa, a cualquier *PG* como *SF* y viceversa, etc. Para automatizar esta nueva matriz se diseñó en Excel una tabla que tuviese en cuenta todas estas condiciones además de, obviamente, la clasificación correcta de la posición del jugador.

Aplicando esto que se acaba de explicar a la anterior matriz de confusión se obtiene el siguiente resultado (Tabla 7).

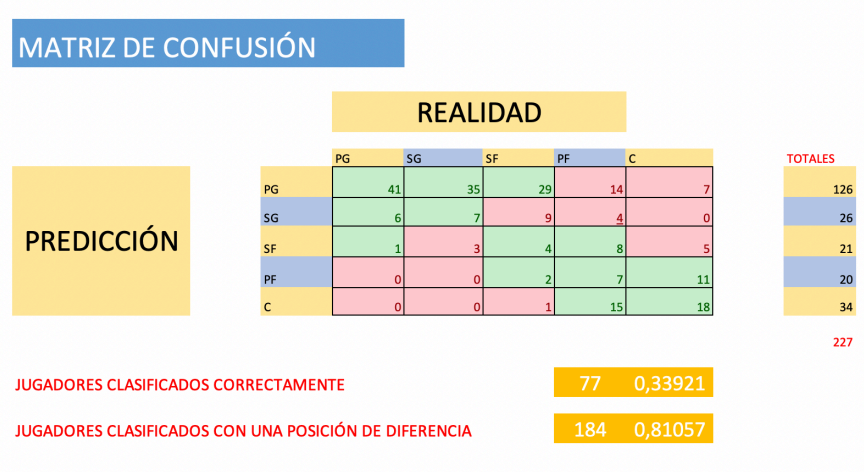


***Tabla 7**: Matriz de confusión ajustada 1 entre la posición real y la predicción creada por el modelo basado en el árbol de decisión para las variables de tiro en la temporada 1994/1995. Generada con Excel.*

La diferencia es realmente significativa, de una precisión del 33,92% con la matriz original, a un 81,06% que proporciona la nueva matriz ajustada. Y haciendo lo mismo para la temporada 2009/2010 se obtienen unos resultados parecidos, aunque mejores en la clasificación exacta de la posición del jugador (Tabla 8).

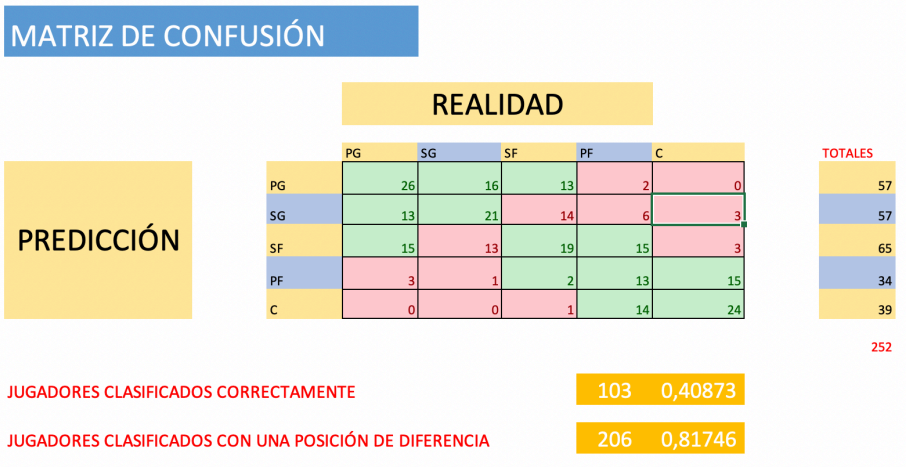


***Tabla 8**: Matriz de confusión ajustada 1 entre la posición real y la predicción creada por el modelo basado en el árbol de decisión para las variables de tiro en la temporada 2009/2010. Generada con Excel.*

De nuevo, los *árboles de decisión* proporcionan una mejor precisión (40,87%) que los k-vecinos más cercanos (36,11%). Además, si se comprueba el gráfico generado para este árbol (Gráfico 15) se obtiene una interpretación visual del proceso que se ha seguido para formar cada rama del árbol en base a unos valores específicos para cada variable.

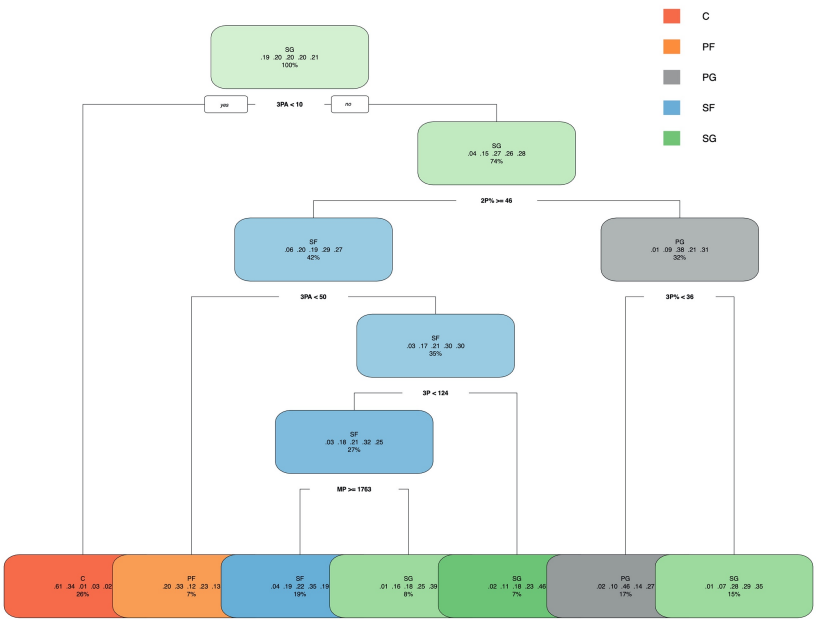


***Gráfico 15**: Gráfico del árbol de decisión creado a partir del modelo generado para la temporada 2009/2010 para las variables de tiro. Generado con R Studio.*

Para poder interpretarlo se tiene que ir de arriba abajo viendo los valores de las variables que dividen la proporción de datos. Así, por ejemplo, se podría afirmar que las proporciones para los jugadores que han intentado menos de 10 triples en toda la temporada el 61% de ellos han sido *Centers*, el 34% *Power Forwards*, el 1% *Point Guards*, el 3% *Small Forwards* y el 2% *Shooting Guards*. O si fijándose en más variables, las proporciones para los jugadores que han intentado más de 10 triples, su porcentaje en tiros de 2 puntos es inferior al 46% y su porcentaje en tiros de 3 puntos es superior al 36% el 1% de ellos han sido *Centers*, el 7% *Power Forwards*, el 28% *Point Guards*, el 29% *Small Forwards* y el 35% *Shooting Guards.*

Para la temporada 2018/2019 se ha comprobado que la precisión de la matriz generada con *árboles de decisión* (42,39%) no es mejor que la generada con el método de los *k-vecinos más cercanos* (42,75%), pero la matriz ajustada sí que lo es (75,36% frente a 71,74%) (Ver Tablas 40 y 42 del Anexo). Para este set de datos también se aplicó el árbol de decisión *C5.0* que dio peores resultados (39,49% de precisión). A pesar de ello se generó un gráfico de este procedimiento a una muestra del 2% de observaciones con el fin de ver su proceso de clasificación (Gráfico 16). En él se aprecia que la variable más determinante en el proceso de división es el de triples intentados por cada jugador.

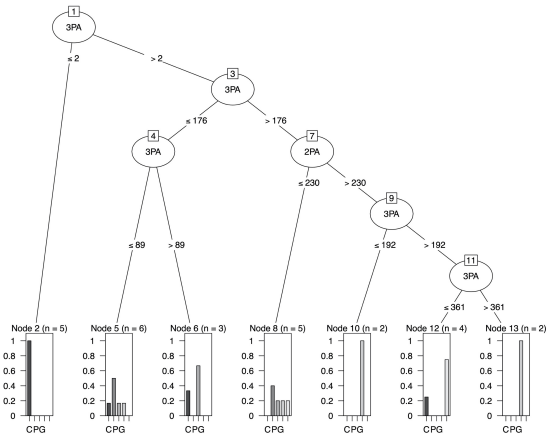


***Gráfico 16**: Gráfico del árbol de decisión C5.0 creado a partir del modelo generado para una muestra del 2% de jugadores de la temporada 2018/2019 para las variables de tiro. Generado con R Studio.*

Una vez analizada la clasificación que hace cada modelo para su temporada de referencia se ha procedido al análisis de cada modelo usando los datos de las otras temporadas. Así se llega a los siguientes resultados (Ver Tablas 26, 29, 35, 38, 44 y 47 del Anexo):

- El **modelo** generado para los datos de la temporada **1994/1995** clasifica (*árboles de decisión*) a los **jugadores** de la temporada **2009/2010** con una **precisión** del **38,49%** para la matriz normal **y** de **76,19%** para la ajustada.

- El **modelo** generado para los datos de la temporada **1994/1995** clasifica (*k-vecinos más cercanos*) a los **jugadores** de la temporada **2018/2019** con una **precisión** del **33,33%** para la matriz normal **y** de **66,66%** para la ajustada.

- El **modelo** generado para los datos de la temporada **2009/2010** clasifica (*árboles de decisión*) a los **jugadores** de la temporada **1994/1995** con una **precisión** del **40,97%** para la matriz normal **y** de **81,5%** para la ajustada.

- El **modelo** generado para los datos de la temporada **2009/2010** clasifica (*k-vecinos más cercanos*) a los **jugadores** de la temporada **2018/2019** con una **precisión** del **35,14%** para la matriz normal **y** de **63,41%** para la ajustada.

- El **modelo** generado para los datos de la temporada **2018/2**019 clasifica (*árboles de decisión*) a los **jugadores** de la temporada **1994/1995** con una **precisión** del **35,68%** para la matriz normal **y** de **79,3%** para la ajustada.

- El **modelo** generado para los datos de la temporada **2018/2019** clasifica (*árboles de decisión*) a los **jugadores** de la temporada **2009/2010** con una **precisión** del **38,1%** para la matriz normal **y** de **75,79%** para la ajustada.

Se presenta una tabla que resume todos estos casos (Tabla 9).

Datos Clasificados / Modelo utilizado	1994/1995 Normal	1994/1995 Ajustada	2009/2010 Normal	2009/2010 Ajustada	2018/2019 Normal	2018/2019 Ajustada
1994/1995			38,49%	76,19%	33,33%	66,66%
2009/2010	40,97%	81,50%			35,14%	63,41%
2018/2019	35,68%	79,30%	38,10%	75,79%		

***Tabla 9**: Tabla resumen de la precisión en la clasificación de los jugadores de una temporada con el modelo creado para otra diferente, usando la matriz de confusión normal y la ajustada con las variables de tiro. Generada con Excel.*

Como se puede observar con estos datos, la peor clasificación tiene lugar cuando se emplean los modelos de las temporadas 1994/1995 y 2009/2010 a los datos de la temporada 2018/2019 lo cual reafirma la hipótesis de que el mayor cambio ha tenido lugar después de la revolución en el tiro que se ha dado en los últimos años.

¿Por qué entonces el modelo generado para la temporada 2018/2019 sí clasifica más correctamente los datos de las otras temporadas? Esto puede deberse a que los cambios en los roles que se han producido a lo largo del tiempo no han podido ajustarse a que hoy en día los jugadores hagan los tiros que hacen, esto quiere decir que ningún modelo podía tener en cuenta, por ejemplo, que hoy los *Power Forwards* tiren prácticamente igual que los *Point Guards*, cuando sus roles en el juego son muy diferentes. Sin embargo, el modelo que puede generarse con los datos actuales sí puede reflejar que aún muchos jugadores siguen teniendo los patrones que se tenían en el pasado. En resumen, antes pocos jugadores

se comportaban como lo hacen los de ahora, mientras que aún hay en el presente bastantes jugadores que, aunque se hayan ajustado al juego actual, tiene patrones que se podían ver en el pasado. Una posible explicación a este fenómeno es que los "viejos" patrones de aprendizaje aún persisten en la formación de los jugadores dependiendo de la posición que ocupen desde edades tempranas.

De la misma forma que para el tiro, se ha realizado un análisis de las variables de estadística avanzada que se ha citado anteriormente. Para no saturar este apartado con el mismo tipo de información que la expuesta anteriormente se presenta la siguiente tabla donde quedan recogidos los resultados para cada temporada generados con su modelo y con el de las otras (Tabla 10) (Ver Tablas 51, 55, 59, 62, 67, 71, 76, 80 y 84 del Anexo).

Datos Clasificados / Modelo utilizado	1994/1995		2009/2010		2018/2019	
	Normal	Ajustada	Normal	Ajustada	Normal	Ajustada
1994/1995	60,79%	88,11%	59,13%	86,51%	49,64%	80,07%
2009/2010	60,79%	86,34	64,29%	90,08%	55,80%	86,96%
2018/2019	60,79%	87,67%	63,10%	90,87%	63,77%	84,06%

***Tabla 10**: Tabla resumen de la precisión en la clasificación de los jugadores de una temporada con el modelo creado para todas las demás, usando la matriz de confusión normal y la ajustada, con las variables de estadística avanzada. Generada con Excel.*

Cabe destacar que parte de los resultados sobre la precisión ajustada se ve desfavorecida por no tomar como adyacentes las posiciones de *Shooting Guard* y *Small Forward*. En el análisis de tiro tenía más sentido mantenerlas separadas, sin embargo, para esta parte de estadística avanzada que tiene que ver más con el rol que desempeña cada posición sí que puede tenerse en cuenta estas dos posiciones como adyacentes, lo que implicaría que cualquier *SG* clasificado como *SF* y viceversa sería correcto. Siguiendo el mismo razonamiento ahora cualquier *PG* clasificado como *SF* y viceversa es considerado como error, ya que sus roles son suficientemente distintos para no considerarlas posiciones adyacentes. De esta forma, un ejemplo de tabla para una temporada en particular quedaría así.

MATRIZ DE CONFUSIÓN

PREDICCIÓN \ REALIDAD	PG	SG	SF	PF	C	TOTALES
PG	42	2	0	0	0	44
SG	15	39	11	0	0	65
SF	0	9	31	11	3	54
PF	0	1	6	13	5	25
C	0	0	1	26	37	64
						252

JUGADORES CLASIFICADOS CORRECTAMENTE	162	0,64286
JUGADORES CLASIFICADOS CON UNA POSICIÓN DE DIFERENCIA	247	0,98016

***Tabla 11**: Matriz de confusión ajustada 2 entre la posición real y la predicción creada por el modelo basado en el árbol de decisión para las variables de estadística avanzada en la temporada 2009/2010. Generada con Excel.*

Realizando esto con todos los modelos y temporadas, éste es el resumen que se obtiene (Tabla 12) (Ver Tablas 52, 56, 60, 68, 72, 77, 81 y 85 del Anexo).

Modelo utilizado \ Datos Clasificados	1994/1995 Normal	1994/1995 Ajustada	2009/2010 Normal	2009/2010 Ajustada	2018/2019 Normal	2018/2019 Ajustada
1994/1995	60,79%	96,92%%	59,13%	97,22%%	49,64%	88,41%%
2009/2010	60,79%	96,92%	64,29%	98,02%	55,80%	93,48%
2018/2019	60,79%	93,39%%	63,10%	98,81%	63,77%	95,29%%

***Tabla 12**: Tabla resumen de la precisión en la clasificación de los jugadores de una temporada con el modelo creado para todas las demás, usando la matriz de confusión normal y la nueva ajustada, con las variables de estadística avanzada. Generada con Excel.*

Los porcentajes conseguidos con este ajuste se puede comprobar que son mucho más cercanos al 100% incluso para el caso en el que el modelo ha sido generado con los datos de referencia de la temporada 1994/1995 y probado con la temporada 2018/2019, que es, sin ninguna duda, donde se podría encontrar el mayor contraste. ¿Quiere esto decir que los jugadores, en las facetas no tan relacionadas con el tiro, no han cambiado tan significativamente? Se podría afirmar que los resultados no son concluyentes. Es cierto que los jugadores actualmente no son los que eran décadas atrás. Sin embargo, pese a que el juego haya ido evolucionando a lo largo de los años y así lo hayan hecho sus perfiles, los roles atribuidos a cada posición no han sufrido un cambio tan diferenciador.

4.3. Análisis De Las Principales Variables Que Afectan A Un Tiro

En este apartado **se analiza la importancia de las distintas variables que afectan en una situación de tiro, tanto en el apartado ofensivo como el defensivo, en su efectividad**. Por ello se ha recurrido a técnicas de regresión logística, y aunque el objetivo principal no es predecir si un tiro en concreto va a tener éxito, se utilizarán los modelos generados para clasificar los mismos datos para ver cuán importantes son las variables empleadas en cada modelo.

La base de datos empleada para ello ha sido extraída de la web de *NBASavant* y consta de 50.000 tiros realizados en la fase regular de la temporada 2015/2016 (el total de la temporada fueron aproximadamente 212.540), la cual ha sido seleccionada por ser la más reciente que cuenta con información sobre las coordenadas (*x,y*) en las que se ha realizado cada tiro. Este *set* de datos cuenta con las siguientes variables:

- ***Name***: Jugador que realiza el tiro.
- ***Period***: Cuarto en el que se ha realizado el tiro.
- ***Minutes Remaining***: Minutos que le faltan al cuarto para acabar cuando se ha realizado el tiro.
- ***Seconds Remaining***: Segundos que le faltan al minuto para acabar cuando se ha realizado el tiro.
- ***Shot Result***: Éxito (made) o fracaso (missed) del tiro. (Ésta será la variable independiente en el modelo)
- ***Action Type***: Tipo de tiro que se ha realizado.
- ***Shot Type***: Si un tiro ha sido de 2 puntos o 3 puntos.

- ***Shot Distance***: Distancia a la canasta desde la posición donde se ha realizado el tiro (en pies).
- ***x****:* Coordenada *x* desde donde se ha realizado el tiro.
- ***y****:* Coordenada *y* desde donde se ha realizado el tiro.
- ***Dribbles***: Número de botes que ha realizado el tirador desde que recibe el balón hasta que realiza el tiro.
- ***Touch Time***: Segundos que pasan desde que el tirador recibe el balón hasta que se realiza el tiro.
- ***Defender Name***: Jugador del equipo rival más cercano al jugador que realiza el tiro.
- ***Defender Distance***: Distancia del jugador del equipo rival más cercano al jugador que realiza el tiro (en pies).
- ***Shot Clock***: Segundos que le faltan a la posesión para finalizar.

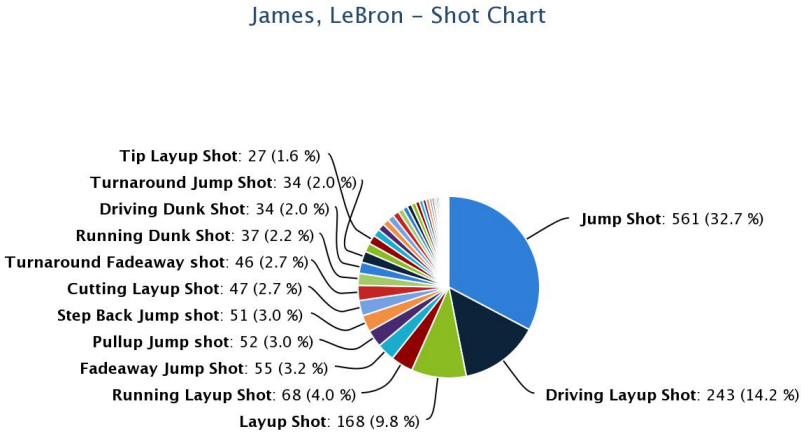


Gráfico 17*: Diagrama de sectores con los tiros que realizó LeBron James en la temporada 2015/2016 por el tipo de tiro. Fuente: NBASavant.*

Desafortunadamente, una de las variables que a priori se puede pensar que es más significativa, el defensor más cercano al tirador, no se encuentra recogido para todos los tiros en la base de datos extraída. Por ello se eliminaron todas las instancias para los que no se tenía información de esta variable. Esto dio lugar a un *set* de datos de 25.940 tiros con 15 variables.

Como el método principal que se va a utilizar es el de regresión logística múltiple y éste sólo tiene en cuenta la linealidad de las variables se han definido 3 variables nuevas referentes a las coordenadas desde donde se realiza el tiro:

- $\boldsymbol{x^2}$: Cuadrado de la coordenada *x* desde donde se ha realizado el tiro.
- $\boldsymbol{y^2}$: Cuadrado de la coordenada *y* desde donde se ha realizado el tiro.
- $\boldsymbol{x * y}$: Producto de las coordenadas *x* e *y* desde donde se ha realizado el tiro.

Con el fin de encontrar más variables que pudieran tener un efecto importante en este análisis se recurrió a la web de estadísticas oficiales de la propia NBA y de ella se obtuvieron por un lado 4 variables referentes a las características físicas de cada jugador:

- ***Shooter Height***: Altura del tirador (en pies).
- ***Shooter Weight***: Peso del tirador (en libras).
- ***Defender Height***: Altura del defensor más cercano (en pies).
- ***Defender Weight***: Peso del defensor más cercano (en libras).

y por otro lado 18 variables referentes al volumen de tiros y su porcentaje de acierto por distancias:

- ***Less Than 5 FT FGM***: Tiros conseguidos por partido a menos de 5 pies de la canasta por el jugador que realiza el tiro.
- ***Less Than 5 FT FGA***: Tiros intentados por partido a menos de 5 pies de la canasta por el jugador que realiza el tiro.
- ***Less Than 5 FT FG%***: Porcentaje de acierto del tirador que realiza el tiro para tiros a menos de 5 pies de la canasta.
- ***5-9 FT FGM***: Tiros conseguidos por partido entre 5 y 9 pies de la canasta por el jugador que realiza el tiro.
- ***5-9 FT FGA***: Tiros intentados por partido entre 5 y 9 pies de la canasta por el jugador que realiza el tiro.
- ***5-9 FT FG%***: Porcentaje de acierto del tirador que realiza el tiro para tiros entre 5 y 9 pies de la canasta.
- ***10-14 FT FGM***: Tiros conseguidos por partido entre 10 y 14 pies de la canasta por el jugador que realiza el tiro.
- ***10-14 FT FGA***: Tiros intentados por partido entre 10 y 14 pies de la canasta por el jugador que realiza el tiro.
- ***10-14 FT FG%***: Porcentaje de acierto del tirador que realiza el tiro para tiros entre 10 y 14 pies de la canasta.
- ***15-19 FT FGM***: Tiros conseguidos por partido entre 15 y 19 pies de la canasta por el jugador que realiza el tiro.
- ***15-19 FT FGA***: Tiros intentados por partido entre 15 y 19 pies de la canasta por el jugador que realiza el tiro.
- ***15-19 FT FG%***: Porcentaje de acierto del tirador que realiza el tiro para tiros entre 15 y 19 pies de la canasta.
- ***20-24 FT FGM***: Tiros conseguidos por partido entre 20 y 24 pies de la canasta por el jugador que realiza el tiro.
- ***20-24 FT FGA***: Tiros intentados por partido entre 20 y 24 pies de la canasta por el jugador que realiza el tiro.
- ***20-24 FT FG%***: Porcentaje de acierto del tirador que realiza el tiro para tiros entre 20 y 24 pies de la canasta.
- ***25-29 FT FGM***: Tiros conseguidos por partido entre 25 y 29 pies de la canasta por el jugador que realiza el tiro.

- ***25-29 FT FGA***: Tiros intentados por partido entre 25 y 29 pies de la canasta por el jugador que realiza el tiro.

- ***25-29 FT FG%***: Porcentaje de acierto del tirador que realiza el tiro para tiros entre 25 y 29 pies de la canasta.

Por lo que el *set* final de datos con el que se realizó el análisis quedó constituido por 25.931 tiros con 40 variables asociadas a cada tiro. Todo el análisis al igual que en los apartados anteriores se ha realizado con *R Studio* y su código será incluido en los anexos de este trabajo.

Además de analizar todos los tiros en conjunto, se extrajo un *subset* que contaba con los tiros realizados a más de 5 pies de distancia, ya que existe una gran diferencia entre ellos y todo lo que son mates (*dunks*) y entradas cerca del aro (*layups*), y por lo tanto las variables más importantes que intervienen en ellos también tienen que ser diferentes.

Primero se analizará el conjunto de todos los tiros empleando todas las variables disponibles. Ajustando los datos a un modelo lineal generalizado (*GLM*) con función de enlace binomial se genera un modelo de regresión logística múltiple. El modelo que emplea todas las variables (modelo general) es el siguiente (Tabla 13).

Como se puede comprobar la mayoría de las variables son significativas, sobre todo las referentes a qué tipo de tiro se ha realizado y las que tienen que ver con la posición de los jugadores y la canasta en el momento del tiro. Sin embargo, se puede ver que pocas variables referentes al volumen y acierto en el tiro no tienen un impacto tan importante en el modelo.

Esto puede llevar a plantearse que el hecho de existir tantos niveles dentro de la variable categórica del tipo de tiro que se ha realizado, y que ésta aporta a su vez mucha información sobre la distancia a la que se está realizando el tiro podría por sí sola constituir un predictor lo suficientemente preciso en comparación con el modelo general.

Pasa lo mismo con la distancia a la canasta del tirador en el momento del tiro. Es indudablemente la variable más importante, así que se debería comprobar su poder predictor aislado del resto de variables para poder comprobar de una forma más fiable la importancia de las otras variables.

Para ello se han llevado a cabo dos modelos de regresión logística simple y se comparará su poder clasificador con el modelo general.

```
Coefficients:
                                      Estimate  Std. Error z value             Pr(>|z|)
(Intercept)                         -1.47345901  0.51733151  -2.85             0.00440 **
period                              -0.01587363  0.01206954  -1.32              0.18645
minutes_remaining                   -0.00372143  0.00403024  -0.92              0.35581
seconds_remaining                    0.00049379  0.00079135   0.62              0.53264
action_typeAlley Oop Layup shot      1.23966438  0.15647105   7.92 0.0000000000000233 ***
action_typeCutting Dunk Shot        -1.69596497  0.27988430  -6.06 0.0000000136527026 ***
action_typeCutting Finger Roll Layup Shot -0.47047045  0.28974496  -1.62              0.10443
action_typeCutting Layup Shot        0.39218386  0.13696730   2.86              0.00419 **
action_typeDriving Bank Hook Shot    0.69206975  0.37862311   1.83              0.06757 .
action_typeDriving Bank shot         0.84842677  0.16049560   5.29 0.0000001241998029 ***
action_typeDriving Dunk Shot        -0.14723712  0.17102481  -0.86              0.38929
action_typeDriving Finger Roll Layup Shot  0.66580383  0.14270065   4.67 0.0000030751137945 0 ***
action_typeDriving Floating Bank Jump Shot  0.85182363  0.19393533   4.39 0.0000112153893227 ***
action_typeDriving Floating Jump Shot  1.35741277  0.14909095   9.10 < 0.0000000000000002 ***
action_typeDriving Hook Shot         1.28564408  0.15665419   8.21 0.0000000000000023 ***
action_typeDriving Jump shot         1.40418451  0.83299368   1.69              0.09185 .
action_typeDriving Layup Shot        1.19048474  0.12684299   9.39 < 0.0000000000000002 ***
action_typeDriving Reverse Dunk Shot  1.29321252  0.65341659   1.98              0.04780 *
action_typeDriving Reverse Layup Shot  1.22277184  0.14648155   8.35 < 0.0000000000000002 ***
action_typeDunk Shot                 0.21531857  0.13773982   1.56              0.11800
action_typeFadeaway Bank shot        0.72262324  0.28394104   2.54              0.01093 *
action_typeFadeaway Jump Shot        1.50849653  0.15290025   9.87 < 0.0000000000000002 ***
action_typeFinger Roll Layup Shot    1.07295134  0.17024177   6.30 0.0000000029285410 ***
action_typeFloating Jump shot        1.53037747  0.13893530  11.02 < 0.0000000000000002 ***
action_typeHook Bank Shot            0.71857103  0.25641254   2.80              0.00507 **
action_typeHook Shot                 1.70873312  0.13335473  12.81 < 0.0000000000000002 ***
action_typeJump Bank Shot            1.21534836  0.14881598   8.17 0.0000000000000032 ***
action_typeJump Shot                 2.13757524  0.16311430  13.10 < 0.0000000000000002 ***
shot_type3PT Field Goal             -0.02855239  0.13555537  -0.21              0.83317
shot_distance                        0.08257823  0.01300662   6.35 0.0000000021680348 ***
x                                    0.00002350  0.00031575   0.07              0.94067
y                                   -0.00222142  0.00099800  -2.23              0.02602 *
`x^2`                               -0.00002173  0.00000592  -3.67              0.00024 ***
`y^2`                               -0.00000764  0.00000273  -2.79              0.00523 **
`x*y`                               -0.00000529  0.00000344  -1.54              0.12355
dribbles                            -0.03422541  0.01017061  -3.37              0.00077 ***
touch_time                           0.05488819  0.01178084   4.66 0.0000031758614147 ***
shot_clock                          -0.01340053  0.00241408  -5.55 0.0000000284075757 ***
shooter_height                       0.07193449  0.05632238   1.28              0.20153
shooter_weight                      -0.00092426  0.00076591  -1.21              0.22753
defender_distance                   -0.11292633  0.00907499 -12.44 < 0.0000000000000002 ***
defender_height                      0.05585805  0.04458086   1.25              0.21022
defender_weight                      0.00360891  0.00062441   5.78 0.0000000074812558 5 ***
`LESS THAN 5FT FGM`                 -0.04658429  0.10283610  -0.45              0.65055
`LESS THAN 5FT FGA`                  0.03815788  0.06079088   0.63              0.53021
`LESS THAN 5FT FG%`                 -0.01632683  0.00409725  -3.98 0.0000675312274972 ***
`5-9 FT FGM`                        -0.07027896  0.22619456  -0.31              0.75603
`5-9 FT FGA`                        -0.07422384  0.10672780  -0.70              0.48677
`5-9 FT FG%`                        -0.00599434  0.00214326  -2.80              0.00516 **
`10-14 FT FGM`                      -0.26051873  0.25207163  -1.03              0.30136
`10-14 FT FGA`                       0.19120672  0.11743874   1.63              0.10350
`10-14 FT FG%`                      -0.00117839  0.00168052  -0.70              0.48318
`15-19 FT FGM`                       0.16952529  0.15777540   1.07              0.28261
`15-19 FT FGA`                      -0.14302492  0.06776098  -2.11              0.03480 *
`15-19 FT FG%`                      -0.00233177  0.00144730  -1.61              0.10715
`20-24 FT FGM`                       0.01159200  0.16664838   0.07              0.94454
`20-24 FT FGA`                       0.11821071  0.06507161   1.82              0.06927 .
`20-24 FT FG%`                       0.00091702  0.00118378   0.77              0.43854
`25-29 FT FGM`                       0.14113766  0.18749101   0.75              0.45159
`25-29 FT FGA`                      -0.16959251  0.07569396  -2.24              0.02506 *
`25-29 FT FG%`                      -0.00166607  0.00103974  -1.60              0.10907
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabla 13*: modelo de regresión logística múltiple a partir de todas las variables disponibles. Generada con R. Studio.*

Como la predicción de dichos modelos se basa en un valor de probabilidad generado por el mismo, se utilizarán 3 tipos de *thresholds* distintos (0.5, 0.4 y 0.6) con el fin de obtener una mejor precisión en la clasificación. La siguiente tabla (Tabla 14) compara el modelo con todas las variables, el generado sólo con el tipo de tiro y el generado sólo con la distancia utilizando cada uno de los *thresholds* (Ver Tablas 102-110 del Anexo).

Modelo	Realidad / Predicción	Threshold 0.4 Encestado	0.4 Fallado	0.5 Encestado	0.5 Fallado	0.6 Encestado	0.6 Fallado
Todas las variables	Encestado	9821	3109	13184	5958	15246	8447
	Fallado	6243	6758	2880	3909	818	1420
	Precisión	63,94%		65,92%		64,27%	
	Sensibilidad	61,14%		82,07%		94,91%	
	Especificidad	68,49%		39,62%		14,39%	
Tipo de tiro	Encestado	11210	4439	12501	5536	15654	9096
	Fallado	4854	5428	3563	4331	410	771
	Precisión	64,16%		64,91%		63,34%	
	Sensibilidad	69,78%		77,82%		97,45%	
	Especificidad	55,01%		43,89%		7,81%	
Distancia	Encestado	12409	5740	14559	8013	15361	8880
	Fallado	3655	4127	1505	1854	703	987
	Precisión	63,77%		63,29%		63,04%	
	Sensibilidad	77,25%		90,63%		95,62%	
	Especificidad	41,83%		18,79%		10,00%	

< 20%

20% - 60%

>60%

***Tabla 14**: Tabla resumen de las matrices de confusión y su precisión generadas con los modelos de regresión logística de todas las variables, la variable "Action Type" y la variable "Shot Distance". Generada con Excel.*

Si se observan las precisiones de cada modelo para los diferentes *threshold* que se han seleccionado se puede comprobar que todos tienen una precisión de entre el 61% y 66% aproximadamente. Pero no sería estrictamente correcto tomar esto como válido sin ver los datos al detalle, ya que para los *thresholds* 0.5 y 0.6 la gran mayoría de tiros son clasificados como encestados, lo cual beneficia a la precisión por constar de muchos más tiros encestados que fallados (16.071 frente a 9.869).

Por ello, para cada modelo particular se deberá tener en cuenta no sólo la precisión de la matriz de confusión, sino también su sensibilidad y especificidad, y en base a esto elegir el *threshold* para el que la matriz de confusión tiene unas métricas más consistentes. Para la tabla anterior el que cumple mejor con esta condición es el de 0.4.

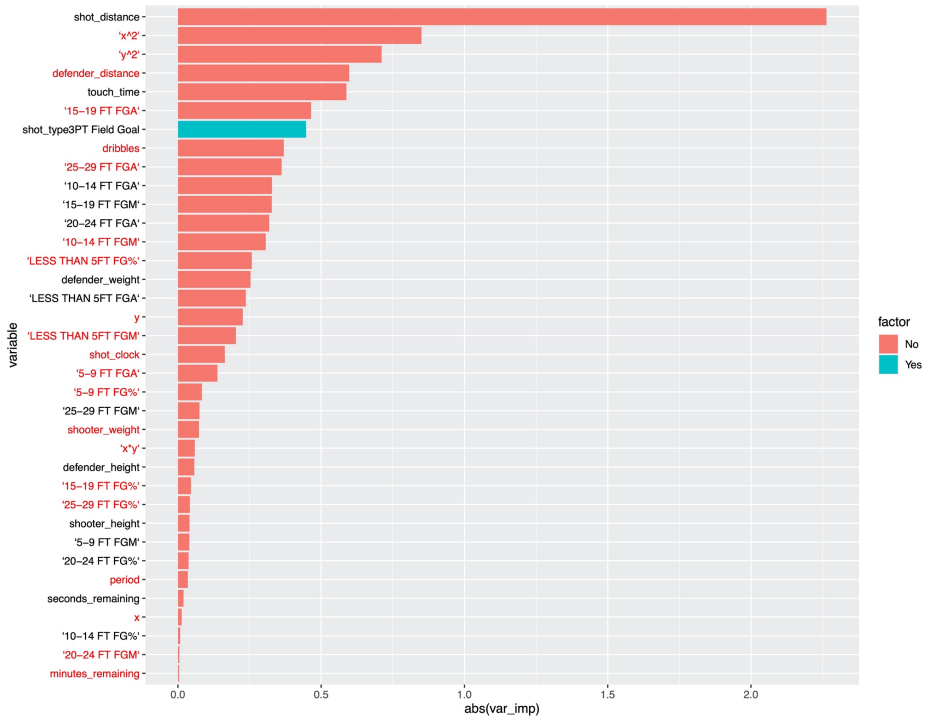


***Gráfico 18**: Representación gráfica de la importancia de las variables en el modelo con todas las variables excepto el tipo de tiro. Generado con R Studio.*

Una vez viendo que la precisión más alta que se puede conseguir está en torno al 65% el objetivo es encontrar el mínimo número de variables que consiguen una precisión parecida con una sensibilidad y especificidad consistentes. Por lo tanto, se comprueba la importancia de las variables en el modelo habiendo eliminado la variable del tipo de tiro y se obtiene el Gráfico 18. Las variables que aparecen con un color de fuente rojo indican que sus coeficientes son negativos.

Como es lógico, en base a la información analizada previamente, la distancia del tiro es, con mucha diferencia, la variable más importante en el modelo. Tampoco se tendrá en cuenta la variable si un tiro es de 2 puntos o de 3 puntos ya que, aunque no aporta excesiva información sobre la distancia, sí que crea una diferencia importante entre estos tipos de tiro. Entonces, ¿qué número mínimo del resto de variables producen resultados igual de correctos (o por lo menos parecidos) y cuáles son?

Se obtiene que el modelo mínimo que cumple con estas premisas es el formado con las variables recogidas en la Tabla 15.

```
Coefficients:
                     Estimate  Std. Error z value             Pr(>|z|)
(Intercept)        -0.00652487  0.16629673   -0.04                 0.97
defender_distance  -0.14460058  0.00837881  -17.26 < 0.0000000000000002 ***
`x^2`               0.00003896  0.00000190   20.54 < 0.0000000000000002 ***
`y^2`               0.00002981  0.00000176   16.97 < 0.0000000000000002 ***
dribbles           -0.06400551  0.00957088   -6.69       0.000000000023 ***
defender_weight     0.00461801  0.00051470    8.97 < 0.0000000000000002 ***
touch_time          0.11901772  0.01110412   10.72 < 0.0000000000000002 ***
`LESS THAN 5FT FG%` -0.02662394  0.00209425  -12.71 < 0.0000000000000002 ***
```

Tabla 15*: Modelo de regresión logística múltiple con el menor número de variables posibles para obtener los resultados apropiados. Generada con R Studio.*

Para el cual si se comprueba su matriz de confusión con un *threshold* de 0.4 (el más apropiado por sus métricas) se obtiene la siguiente tabla (Tabla 16).

Todas las métricas de esta matriz de confusión son lo suficientemente consistentes y parecidas a las del modelo con todas las variables como para poder decir que es útil.

Con lo cual, se puede afirmar con cierta seguridad que las variables que más influyen en el éxito o fracaso de un tiro, de más a menos influyentes (ver Gráfico 70 del Anexo), y excluyendo la distancia a la canasta y el tipo de tiro son:

1. ***Touch Time*** *(X_6)*
2. $\boldsymbol{y^2}$ *(X_3)*
3. ***Defender Distance*** *(X_1)*
4. $\boldsymbol{x^2}$ *(X_2)*
5. ***Dribbles*** *(X_4)*
6. ***Less Than 5 FT FG%*** *(X_7)*
7. ***Defender Weight*** *(X_5)*

```
Confusion Matrix and Statistics

          Reference
Prediction  made missed
    made   11211   4993
    missed  4853   4874

               Accuracy : 0.62
                 95% CI : (0.614, 0.626)
    No Information Rate : 0.619
    P-Value [Acc > NIR] : 0.397

                  Kappa : 0.192

 Mcnemar's Test P-Value : 0.161

            Sensitivity : 0.698
            Specificity : 0.494
         Pos Pred Value : 0.692
         Neg Pred Value : 0.501
             Prevalence : 0.619
         Detection Rate : 0.432
   Detection Prevalence : 0.625
      Balanced Accuracy : 0.596

       'Positive' Class : made
```

Tabla 16*: Matriz de confusión para el mínimo número de variables con el que obtener resultados oportunos y un threshold de 0.4. Generada con R Studio.*

Entonces la función generada con este modelo que devuelve la probabilidad de pertenecer al grupo 1 (fallar el tiro) viene dada por:

$$\hat{p}(Y=1|X_i)=\frac{e^{-0.006525-0.1446X_1+0.000039X_2+0.0000298X_3-0.064X_4+0.0046X_5+0.119X_6-0.0266X_7}}{1+e^{-0.006525-0.1446X_1+0.000039X_2+0.0000298X_3-0.064X_4+0.0046X_5+0.119X_6-0.0266X_7}} \quad i=1,2,\dots,7$$

Así, por ejemplo, para un tiro realizado por un jugador que tiene un porcentaje de acierto del 45.9% en tiros a menos de 5 pies, a 5.1 pies de distancia del defensor más cercano cuyo peso es de 240 libras, desde las coordenadas $x=\pm 189$ ($x^2=35721$) y $y=\pm 144$ ($y^2=20736$), habiendo transcurrido 1.1 segundos desde que el tirador ha recibido el balón y sin haber dado ningún bote antes de realizar el tiro, la probabilidad que devuelve el modelo de que ese tiro sea errado viene dada por:

$$\hat{p}(Y=1|X_i)=\frac{e^{-0.006525-0.1446*5.1+0.000039*35721+0.0000298*20736-0.064*0+0.0046*240+0.119*1.1-0.0266*45.9}}{1+e^{-0.006525-0.1446*5.1+0.000039*35721+0.0000298*20736-0.064*0+0.0046*240+0.119*1.1-0.0266*45.9}}=0.783$$

Con este resultado cualquier *threshold* elegido anteriormente haría que la predicción fuera clasificada como fallada. Además, el tiro real de este ejemplo fue fallado, o sea que la predicción es acertada, lo cual tiene mucho sentido ya que la probabilidad generada por el modelo de que ese tiro fuese fallado es muy alta.

Del *subset* de datos para tiros a más de 5 pies de la canasta se tiene un conjunto de 7782 tiros de los cuales 3655 fueron encestados y 4127 fueron fallados. La primera diferencia que se aprecia pues, es que para este análisis se cuenta con un mayor número de tiros errados que encestados, lo cual sin embargo no es algo impredecible ya que es lógico que los tiros más cercanos a la canasta hayan sido los que han tenido más acierto.

Procediendo de la misma forma que para el conjunto de todos los datos, el primer paso ha consistido en comprobar el poder predictor de todas las variables en su conjunto y ver si, alguna de ellas por sí sola, consigue generar un modelo que clasifique igual de bien los datos como lo hace el modelo general.

Se obtiene que, al igual que el anterior análisis, la variable del tipo de tiro es un predictor lo suficientemente correcto en comparación con el modelo general. Sin embargo, para este conjunto de datos, la variable de la distancia a la canasta no consigue predicciones tan precisas como lo hacía con todos los tiros. El resumen de todo este análisis (ver Tablas 132-140 del Anexo) queda recogido en la Tabla 17.

Modelo	Realidad / Predicción	Threshold 0.4		Threshold 0.5		Threshold 0.6	
		Encestado	Fallado	Encestado	Fallado	Encestado	Fallado
Todas las variables	Encestado	429	233	1628	1268	3053	2955
	Fallado	3226	3894	2027	2859	602	1172
	Precisión	55,55%		57,66%		54,29%	
	Sensibilidad	11,74%		44,54%		83,53%	
	Especificidad	94,35%		69,28%		28,40%	
Tipo de tiro	Encestado	289	166	1135	918	3259	3398
	Fallado	3366	3961	2520	3209	396	729
	Precisión	54,61%		55,82%		51,25%	
	Sensibilidad	7,91%		31,05%		89,17%	
	Especificidad	95,98%		77,76%		17,66%	
Distancia	Encestado	0	0	617	667	3541	3917
	Fallado	3655	4127	3038	3460	114	210
	Precisión	53,03%		52,39%		48,20%	
	Sensibilidad	0,00%		16,88%		96,88%	
	Especificidad	100,00%		83,84%		5,09%	

< 20%
20% - 50%
>50%

***Tabla 17**: Tabla resumen de las matrices de confusión y su precisión generadas con los modelos de regresión logística de todas las variables, la variable "Action Type" y la variable "Shot Distance" para el subset de datos realizados a más de 5 pies de distancia. Generada con Excel.*

Además, si se comprueba la importancia de las variables excluyendo únicamente el tipo de tiro se obtiene el Gráfico 19.

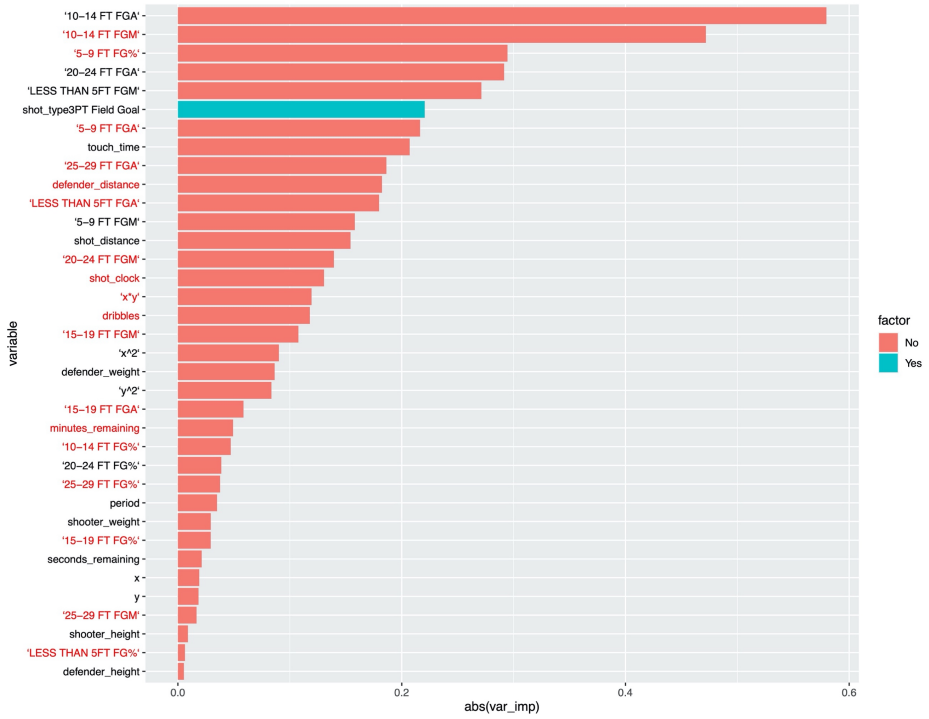


***Gráfico 19**: Representación gráfica de la importancia de las variables en el modelo con todas las variables excepto el tipo de tiro para el subset de datos realizados a más de 5 pies de distancia. Generado con R Studio.*

Pese a que las dos variables más importantes del modelo "*14-19 FT FGA*" y "*14-19 FT FGA*" parece que presentan una diferencia significativa con el resto de predictores, igual que pasa con la variable de la distancia del tiro a la canasta, no son capaces de producir clasificaciones lo suficientemente precisas por si solas (ver Tablas 141-143 del Anexo).

En el proceso de encontrar el mínimo número de predictores que consigan replicar el modelo general lo más fielmente posible se concluyó que el modelo que mejor se ajustaba a ello era el siguiente (Tabla 18).

```
Coefficients:
                       Estimate Std. Error z value     Pr(>|z|)
(Intercept)             0.75282    0.13615    5.53 0.0000000321 ***
shot_distance           0.02132    0.00622    3.43      0.00061 ***
`10-14 FT FGM`         -0.89508    0.29839   -3.00      0.00270 **
`10-14 FT FGA`          0.41426    0.13627    3.04      0.00237 **
`5-9 FT FG%`           -0.01745    0.00287   -6.08 0.0000000012 ***
shot_type3PT Field Goal 0.25114    0.11784    2.13      0.03308 *
`25-29 FT FGA`         -0.05981    0.01994   -3.00      0.00270 **
defender_distance      -0.04896    0.01337   -3.66      0.00025 ***
```

***Tabla 18**: Modelo regresión logística múltiple con el menor número de variables posibles para obtener los resultados apropiados para el subset de datos realizados a más de 5 pies de distancia. Generada con R Studio.*

En concreto, si se comprueba su poder predictor con un *threshold* de 0.5 (el más apropiado por sus métricas) la matriz de confusión resultante es la siguiente (Tabla 19).

Al igual que pasaba con el anterior análisis y comprobando los resultados obtenidos para este conjunto de datos, se puede afirmar con cierta confianza que las variables que juegan un papel más importante en el éxito o fracaso de un tiro a más de 5 pies de distancia son las siguientes.

Se presentan de mayor a menor importancia (ver Gráfico 75 del Anexo) y se ha excluido únicamente la variable del tipo de tiro por su alto poder predictor de forma aislada.

1. ***10-14 FT FGA*** *(X_3)*
2. ***10-14 FT FGM*** *(X_2)*
3. ***5-9 FT FG%*** *(X_4)*
4. ***Shot Distance*** *(X_1)*
5. ***Shot Type (3PT Field Goal)*** *(X_5)*
6. ***Defender_Distance*** *(X_7)*
7. ***25-29 FT FGA*** *(X_6)*

```
Confusion Matrix and Statistics

          Reference
Prediction made missed
    made   1280   1081
    missed 2375   3046

               Accuracy : 0.556
                 95% CI : (0.545, 0.567)
    No Information Rate : 0.53
    P-Value [Acc > NIR] : 0.00000318

                  Kappa : 0.09

 Mcnemar's Test P-Value : < 0.0000000000000002

            Sensitivity : 0.350
            Specificity : 0.738
         Pos Pred Value : 0.542
         Neg Pred Value : 0.562
             Prevalence : 0.470
         Detection Rate : 0.164
   Detection Prevalence : 0.303
      Balanced Accuracy : 0.544

       'Positive' Class : made
```

Tabla 19*: Matriz de confusión para el mínimo número de variables con el que obtener resultados oportunos y un threshold de 0.5 para el subset de datos realizados a más de 5 pies de distancia. Generada con R Studio.*

Entonces la función generada con este modelo que nos devuelve la probabilidad de pertenecer al grupo 1 (fallar el tiro) viene dada por:

$$\hat{p}(Y=1|X_i)=\frac{e^{0.75282+0.02132X_1-0.895X_2+0.41426X_3-0.01745X_4+0.25114X_5-0.0598X_6-0.04896X_7}}{1+e^{0.75282+0.02132X_1-0.895X_2+0.41426X_3-0.01745X_4+0.25114X_5-0.0598X_6-0.04896X_7}} \quad i=1,2,\dots,7$$

La diferencia que se encuentra ahora es que uno de los predictores es una variable categórica (*Shot Type*) lo que indica que si para esta variable el tiro sobre el que se hace la predicción es de 3 puntos el valor que toma $X_5=1$ y si es de 2 puntos el valor que toma $X_5=0$.

Así, por ejemplo, para un tiro de 2 puntos realizado a una distancia de 7 pies a la canasta por un jugador que tiene un porcentaje de acierto del 40.8% en tiros entre 5 y 9 pies de distancia a la canasta, que intenta 2.2 tiros de entre 10 y 14 pies de distancia a la canasta por partido y que consigue 0.9 de ellos, que no intenta ningún tiro de entre 25 y 29 pies de distancia a la canasta y cuyo defensor más cercano se encuentra a 2.2 pies de él la probabilidad que devuelve el modelo de que ese tiro sea errado viene dada por:

$$\hat{p}(Y=1|X_i)=\frac{e^{0.75282+0.02132*7-0.895*0.9+0.41426*2.2-0.01745*40.8+0.25114*0-0.0598*0-0.04896*2.2}}{1+e^{0.75282+0.02132*7-0.895*0.9+0.41426*2.2-0.01745*40.8+0.25114*0-0.0598*0-0.04896*2.2}}=0.5469$$

Por lo tanto, eligiendo un threshold de 0.5 la predicción para este tiro será clasificada como tiro fallado, lo cual concuerda con la realidad del tiro de este ejemplo ya que también fue fallado.

5. CONCLUSIONES

Tras todo el análisis llevado a cabo en este trabajo se pueden extraer una serie de conclusiones que apoyan las ideas desarrolladas en la sección 1.6 sobre el papel de la estadística en el juego.

Por un lado, y con relación al multianálisis realizado a través de las técnicas de ***clustering*** (4.1) se encuentran suficientes evidencias para poder afirmar que las posiciones que desempeñan los jugadores hoy en día, en cuanto a lo que al tiro se refiere, han sufrido un enorme cambio. De una baja especialización en los tiros de 3 puntos para la inmensa mayoría de jugadores, se ha pasado a que actualmente ocurra el efecto contrario, ahora es más difícil encontrar jugadores que no hayan incorporado el tiro de 3 puntos a su repertorio. Incluso para jugadores interiores como los *Power Forwards,* ahora es común que sus porcentajes de tiro sean más similares a las de los jugadores de

cualquier posición exterior que a las de sus otros compañeros interiores, los *Centers*. De hecho, si se tuviera que diferenciar entre dos grandes grupos como pueden ser interiores y exteriores, y teniendo sólo en cuenta las estadísticas de tiro, tendría más sentido incluir a los *PF* como exteriores a pesar de formar así un grupo que contenga 4 posiciones y otro con sólo 1 posición. No obstante, ya que las diferencias que existen entre posiciones han dejado de ser tan significativas, **actualmente a la hora de diferenciar a los jugadores por su naturaleza de tiro, más que hablar de sus posiciones concretas, sería de mayor utilidad ver en qué cluster están situados en base a sus variables de tiro.**

Por otro lado, en lo referente al análisis utilizando técnicas de ***clasificación*** (4.2) se pueden deducir dos conclusiones. Primeramente, respecto a las variables de tiro, usando los datos de los jugadores actuales se obtiene una buena clasificación de los jugadores de años anteriores (siempre que se use el criterio con la matriz ajustada), sin embargo, utilizando los datos de jugadores del pasado se obtiene una peor clasificación de los jugadores de hoy en día. Esto puede traducirse en que, aunque hoy existan jugadores que en su posición sigan teniendo los patrones de tiro que tenían sus antecesores, la mayoría de ellos han ido evolucionando y por ello el ser clasificados correctamente se ha convertido en una misión mucho más complicada al existir muchas similitudes entre las distintas posiciones. No obstante, si en vez de tener en cuenta las variables de tiro se utilizan las variables de estadística avanzada que tienen que ver más con el conjunto de todos los aspectos del juego en los que puede aportar un jugador, se observa como la clasificación de jugadores es mucho más precisa y consistente. **Esto significa que, pese a que el juego haya cambiado enormemente, las posiciones siguen ligadas en general a roles concretos y aportan en las diferentes facetas del juego de una manera más similar a como lo hacían anteriormente.**

Por último, tras el análisis visto en la sección de ***regresión*** (4.3) lo más evidente que puede ser observado es que los tiros realizados a más de 5 pies y el conjunto de todos los tiros encuentran diferencias muy significativas en cuanto a las variables que son más importantes en su efectividad. Para el conjunto de todos los tiros se observa que tanto el tipo de tiro como la distancia son variables suficientemente importantes como para servir de predictores por sí solos; y del resto de variables se comprueba que las más importantes son la distancia que hay entre el defensor más cercano y el tirador, las coordenadas (utilizando sus cuadrados), el número de botes realizados y el tiempo desde que el tirador recibe el balón hasta que efectúa el tiro, el peso del defensor y su porcentaje de acierto en los tiros de menos de 5 pies. Mientras tanto, para los tiros realizados a más de 5 pies de distancia el tipo de tiro sigue siendo un predictor lo suficientemente consistente, mientras que la distancia desde la que se realiza el tiro, pese a ser una variable de gran importancia, por sí sola no logra ser un predictor tan fiable, lo cual puede deberse a que, como se presentó en la sección 1.6.1, para todas las zonas del campo que no están inmediatamente delante de la canasta se encuentran porcentajes de acierto bastante similares. Y las variables que juegan un papel más importante en la efectividad de un tiro de estas características se observa que, aparte de la distancia ya mencionada, son la distancia del defensor más cercano (al igual que para el conjunto de todos los tiros), el volumen de tiros intentados y anotados en zonas de entre 10 y 14 pies de distancia a la canasta, el volumen de tiros intentado en zonas de entre 25 y 29 pies de distancia a la canasta, el porcentaje de acierto para tiros de entre 5 y 9 pies de distancia a la canasta y si el tiro es de 2 o 3 puntos. **Todo ello sugiere que para los tiros muy cercanos a la canasta los porcentajes de acierto de las diferentes zonas no son realmente relevantes, sino que lo más significativo son las variables que determinan la naturaleza del tiro y, sin embargo, para los tiros más lejanos, ya entran en juego las variables que determinan cómo de bueno es un tirador (o por lo menos si efectúa muchos lanzamientos) en las diferentes zonas del campo.**

6. BIBLIOGRAFÍA

- Abbas, N. M. (13 de agosto, 2019). NBA Data Analytics: Changing the Game. Recuperado de https://towardsdatascience.com/nba-data-analytics-changing-the-game-a9ad59d1f116

- Amat, J. (agosto, 2016). Regresión logística simple y múltiple. Recuperado de https://rpubs.com/Joaquin_AR/229736

- Amat, J. (septiembre, 2017). Clustering y heatmaps: aprendizaje no supervisado. Recuperado de https://rpubs.com/Joaquin_AR/310338

- Baghal, T. (2012). Are the “four factors" indicators of one factor? an application of structural equation modeling methodology to NBA data in prediction of winning percentage. Journal of Quantitative Analysis in Sports, 8(1).

- Berrendero, J. R. Clasificación lineal, cuadrática y de vecinos más próximos. Recuperado de https://rpubs.com/joser/ClasificacionSupervisada

- Bocskocsky, A., Ezekowitz, J., & Stein, C. (2014, March). The hot hand: A new approach to an old ‘fallacy’. In 8th Annual MIT Sloan Sports Analytics Conference (pp. 1-10).

- Bosco, J. (23 de abril, 2018). Árboles de decisión con R – Clasificación. Recuperado de https://rpubs.com/jboscomendoza/arboles_decision_clasificacion

- Calvo, D. (3 de octubre, 2016). Árboles de Clasificación en R. Recuperado de http://www.diegocalvo.es/arboles-de-clasificacion-en-r/

- Cervone, D., D’Amour, A., Bornn, L., & Goldsberry, K. (2014, February). POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data. In Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA, USA (Vol. 28, p. 3).

- Cervone, D., D’Amour, A., Bornn, L., & Goldsberry, K. (2016). A multiresolution stochastic process model for predicting basketball possession outcomes. Journal of the American Statistical Association, 111(514), 585-599.

- Delgado, R. (16 de junio, 2018). Introducción a los modelos de clasificación con R. Recuperado de https://rpubs.com/rdelgado/397838

- Erculj, F. & Štrumbelj, E. (2015). Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball. PloS one. 10. e0128885. 10.1371/journal.pone.0128885.

- Fewell, Jennifer & Armbruster, Dieter & Ingraham, John & Petersen, Alexander & Waters, James. (2012). Basketball Teams as Strategic Networks. PloS one. 7. e47445. 10.1371/journal.pone.0047445.

- Franks, A., Miller, A., Bornn, L., & Goldsberry, K. (2015, February). Counterpoints: Advanced defensive metrics for nba basketball. In 9th Annual MIT Sloan Sports Analytics Conference, Boston, MA.

- Goldsberry, K. (2012, March). Courtvision: New visual and spatial analytics for the nba. In 2012 MIT Sloan sports analytics conference (Vol. 9, pp. 12-15).

- Goldsberry, K. (02 de mayo, 2019). How Mapping Shots In The NBA Changed It Forever. Recuperado de https://fivethirtyeight.com/features/how-mapping-shots-in-the-nba-changed-it-forever/

- Goldsberry, K., & Weiss, E. (2013). The Dwight effect: A new ensemble of interior defense analytics for the NBA. Sports Aptitude, LLC. Web, 1-11.

- Gómez, M.A., Alarcón, F. & Ortega, E. (2015). Analysis of shooting effectiveness in elite basketball according to match status. Revista de Psicologia del Deporte. 24. 37-41.

- Jones, E. S. (2016). Predicting outcomes of NBA basketball games (Doctoral dissertation, North Dakota State University).

- Lucey, P., Bialkowski, A., Carr, P., Yue, Y., & Matthews, I. (2014, February). How to get an open shot: Analyzing team movement in basketball using tracking data. In Proceedings of the 8th annual MIT SLOAN sports analytics conference.

- MacCann, Z. (18 de mayo, 2012). Player tracking transforming NBA analytics. Recuperado de https://www.espn.com/blog/playbook/tech/post/_/id/492/492

- Maheswaran, R., Chang, Y. H., Henehan, A., & Danesis, S. (2012, February). Deconstructing the rebound with optical tracking data. In Proceedings of the 6th annual MIT SLOAN sports analytics conference.

- McIntyre, A., Brooks, J., Guttag, J., & Wiens, J. (2016). Recognizing and analyzing ball screen defense in the nba. Ann Arbor, 1001, 48104.

- Meng, A. (05 de noviembre de 2018). How the Three-Point Line Changed the NBA and the Game of Basketball. Recuperado de https://nycdatascience.com/blog/r/how-has-the-three-point-line-changed-the-nba-and-the-game-of-basketball/

- Morate Vázquez, J. (2016). Predicción de equipo ganador en el baloncesto (Bachelor's thesis).

- Mudric, M. (26 de julio, 2019). How The NBA Data And Analytics Revolution Has Changed The Game. Recuperado de https://www.smartdatacollective.com/how-nba-data-analytics-revolution-has-changed-game/

- Nistala, A., & Guttag, J. (2019). Using Deep Learning to Understand Patterns of Player Movement in the NBA. MIT Sloan Sports Analytics Conference.

- Oh, M. H., Keshri, S., & Iyengar, G. (2015, February). Graphical model for baskeball match simulation. In Proceddings of the 2015 MIT Sloan Sports Analytics Conference, Boston, MA, USA (Vol. 2728).

- Piette, J., Anand, S., & Zhang, K. (2010). Scoring and shooting abilities of NBA players. Journal of Quantitative Analysis in Sports, 6(1).

- Shea, S. The 3-Point Revolution. Recuperado de https://shottracker.com/articles/the-3-point-revolution

- Sangüesa, A. A., Moeslund, T. B., Bahnsen, C. H., & Iglesias, R. B. (2017, November). Identifying basketball plays from sensor data; towards a low-cost automatic extraction of advanced statistics. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 894-901). IEEE.

- Skinner, B., & Guy, S. J. (2015). A method for using player tracking data in basketball to learn player skills and predict team performance. PloS one, 10(9).

- Thinking Machines Data Science. A whole new ball game: Quantifying changes in NBA basketball over the past 30 years. (19 de octubre, 2018). Recuperado de https://stories.thinkingmachin.es/nba-in-30-years/

- Weil, S. “The Importance of Being Open: What Optical Tracking Data Can Say About NBA Field Goal Shooting”, in MIT Sloan Sports Analytics Conference, 2011.

SVMMARY

The NBA is the professional basketball league of the USA and Canada which, without doubt, is the most important one in the world. The league is currently composed of two Conferences of fifteen teams each. These in turn are divided in three divisions each, but nowadays and since 2014 these divisions have no major significance. The competition system is divided in two big blocks: The Regular Season and The Playoffs.

The Regular Season usually begins in the third or fourth week of October. It consists of a total of 1230 matches, each team plays 82 matches. It ends between the second and third week of April leaving each team with a balance of wins and losses that is called their "record". This record is what determines the seeding of the teams, and therefore the matches for the next phase: The Playoffs. However, the classification is taken into account by Conference. From the fifteen teams of each Conference, the eight with the best record are classified.

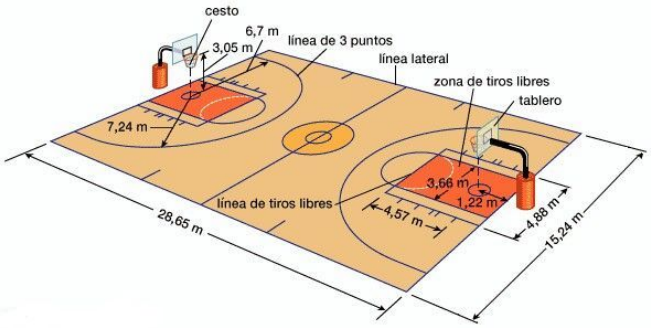
Once the eight teams from each conference have reached The Playoffs, the fight for the title begins. In each conference there are quarter-finals, semi-finals and Conference finals. The Conference champions measure their strength in the NBA final to see who will emerge as the NBA champion.

There have been many significant changes over the years, but this paper focuses on the most important ones from a statistical perspective and their potential impact on the NBA. It should be noted that the NBA has its own regulations, different from those established by the FIBA, the International Basketball Federation.

First of all, some concepts about the NBA have been explained, but at no point has basketball been discussed as such. Basketball is a ball sport in which, in a limited time game, two teams face off to see which one is capable of scoring more points. It is not possible for a game to end in a draw. If the score ends in a draw at the end of regular time, extra periods of time, known as overtimes, will be added until one of the two teams has finished with more points.

A game in the NBA is divided into four periods of 12 minutes each, and each extra period, if necessary, would be of 5 minutes. Although the game is played in a five-on-five format, each team can call up a maximum of 13 players and make any desired changes during the course of the game.

The goal of each team is clear: to get more points than your opponent, this can be done in multiple ways. The court is structured in such way that the baskets obtained from shots made within the three-point line located 7.24 meters from the rim (6.71 meters in the corners) add 2 points, while those obtained from shots made beyond this line add 3 points. It is also possible to obtain 1 point baskets if they come from the line of free throws after a foul has occurred that forces a player to go to this line.



During the time that a match lasts, hundreds of interesting actions occur, both defensive and offensive. Many of them are included in the official statistics of the match, which give rise to the so-called boxscore: points, assists, rebounds, field goal percentages, etc. Many others are not objectively measurable and therefore are not collected anywhere, despite being important in the game and convenient to analyze: correct positioning, decision making, timing of actions, and so on.

Why do we need to go deeper into the analysis of all these variables and look beyond them? The answer is simple, and it is in the very nature of statistics: the more information you have about an event, and the better the analysis of that information, the more accurate results you can obtain.

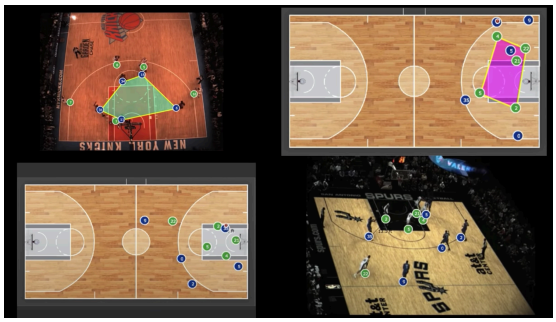
In the traditional statistics you only see actions that take very little of the total time of a game, everything behind is left to a more advanced statistic that not too many people are able to handle or interpret, but that is, without a doubt, enormously advantageous in terms of obtaining results.

Of course, a professional coach doesn't need to know anything about statistics to make an intuitive and sufficiently thorough analysis of the game, but he may be missing important facts that cannot be discovered without getting hold of that statistic he doesn't know. What was once sufficient is no longer so. It's no longer enough to know where victories come from to boost them and defeats to correct them. Now we must take into account all the possible factors involved in these processes and be able to optimize them to the maximum.

In the 2009/2010 season, the NBA reached an agreement with SportVU, a data acquisition and processing company, and began implementing a video system based on six cameras in the main stadiums of the league, capable of capturing the movement of each player, as well as the ball and the actions that took place on the court including their locations, at 25 images per second. At the end of the 2013/2014 season, all the competition halls were equipped with this data collection system.

Taking into account that a match without overtimes lasts 48 minutes - the equivalent of 2880 seconds - and that for every second you have 25 times all the information about what is happening on the court, this translates into a total of 72,000 images in which you have the coordinates of each player on the court (*x, y; t*) and those of the ball (*x, y, z; t*).

All this from a single game and, as explained above, each team during The Regular Season plays a total of 82 games. In this way, an infinite amount of data is obtained that requires a very refined and effective analysis in order to extract useful information from it.

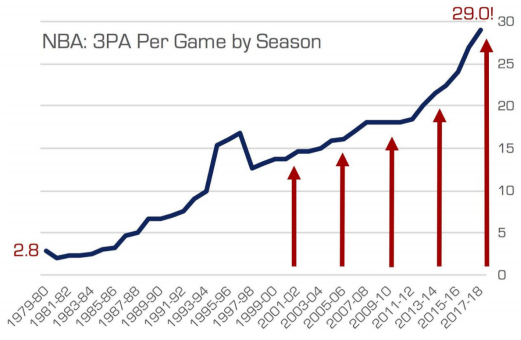


After the 2016/2017 season, the official provider of data tracking changed from *SportVU* to *Second Spectrum*, although its systems are not very different.

All these developments have led to a number of significant changes in the game with a view to improving performance and efficiency. However, the greatest impact of this new use of technology has been seen most clearly in two aspects that are directly correlated and which will be explained later on: the three-point line and the players.

Undoubtedly, the biggest revolution was the introduction of the three point line which has produced change in perspective over time. It hasn't always existed, but since it was introduced in the 1979/1980 season its impact on the game has not stopped growing.

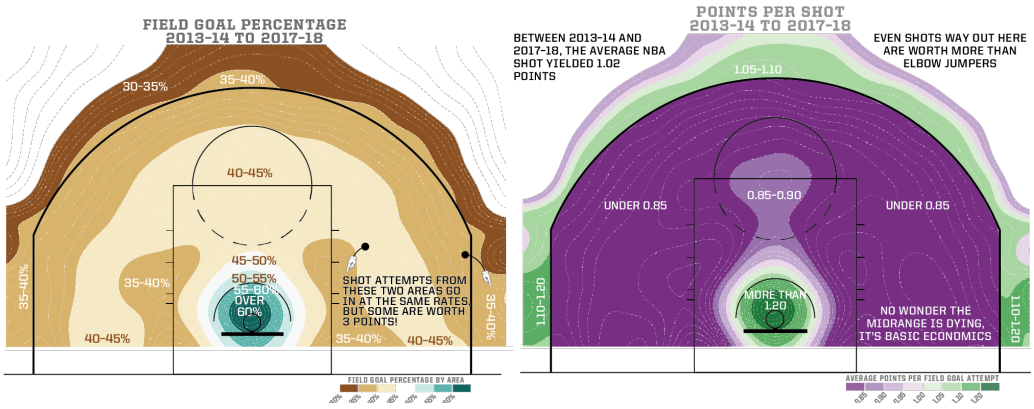
In the first season where the three-point line is found, each team attempted an average of 2.8 three pointers per game. Today that number rises to 29 attempts per game.



Most people would believe that this is due to a significant improvement in the success rate shots since the three point line was stablished in the early years, it was not something that the players were used to. And yes, there has been an improvement, no doubt about it. While the average success rate of 2 point shots has only increased by 2%, the success rates of shots behind the 3 point line have improved by 10%. This may not seem like a significant change at first, but it is one of the most important factors in this revolution.

So, what else is behind the improvement in the success of these shots? Mathematics.

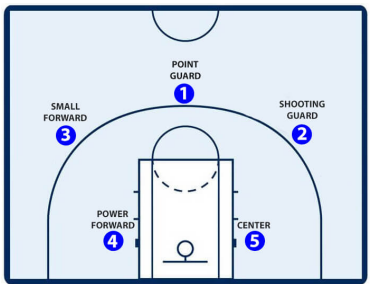
If we look at the field goal percentages per shot in the different zones of the court and apply the expected value of the points to be achieved for each zone it is easy to see how the 3 point shots are more valuable than the rest of the shots that are not right next to the basket.



This transformation and the impact of 3-point shots in the league has translated into a need for players to adapt to this new style of play in which long-distance shooting is so highly rewarded.

The traditional profiles of basketball players are defined by two essential aspects: physical characteristics and their role within the team. The positions on a basketball team that players can occupy are:

- Point Guard (PG).
- Shooting Guard (SG).
- Small Forward (SF).
- Power Forward (PF).
- Center (C).



The theory on paper states that the first three (PG, SG and SF) form the outside part of the lineup and the other two (PF and C) the inside part. Also, and depending on the physical characteristics of height, wingspan, speed and power, the traditional profile says that the outside players are smaller and have better ball control and shooting skills, while the inside players are bigger, stronger and responsible for capturing rebounds and scoring baskets near the hoop. But all this is, as mentioned, theory.

It is important to distinguish the concepts of position and role of a player, as it will be treated extensively throughout this work. A player's position is defined as mentioned above by a number of characteristics both physical and skills and is a way of classifying the player into a larger group where players in the same group have similar characteristics. Role is a more diffuse concept that refers to the function of the player within the team and is somewhat more particular to each player regardless of his position.

This paper has two main goals:

1. The first is to analyze the evolution that NBA players have undergone over time and especially, how this evolution has been reflected in their positions and in the roles associated with these positions.

2. The second will be based on an analysis of all the variables that can be of interest at the time of a shot, in order to discover which are the most significant and that can serve as indicators of their success. All of this taking into account that not all the shots are the same and therefore, they should not be affected by the same variables.

First, we will analyze the evolution over the years of the roles by position of the players. For this purpose, clustering techniques will be used. What we want to check is if there are, or have been, patterns that differentiate the players by their positions taking as reference their shots. Briefly, if the players of each position share a similar shooting pattern.

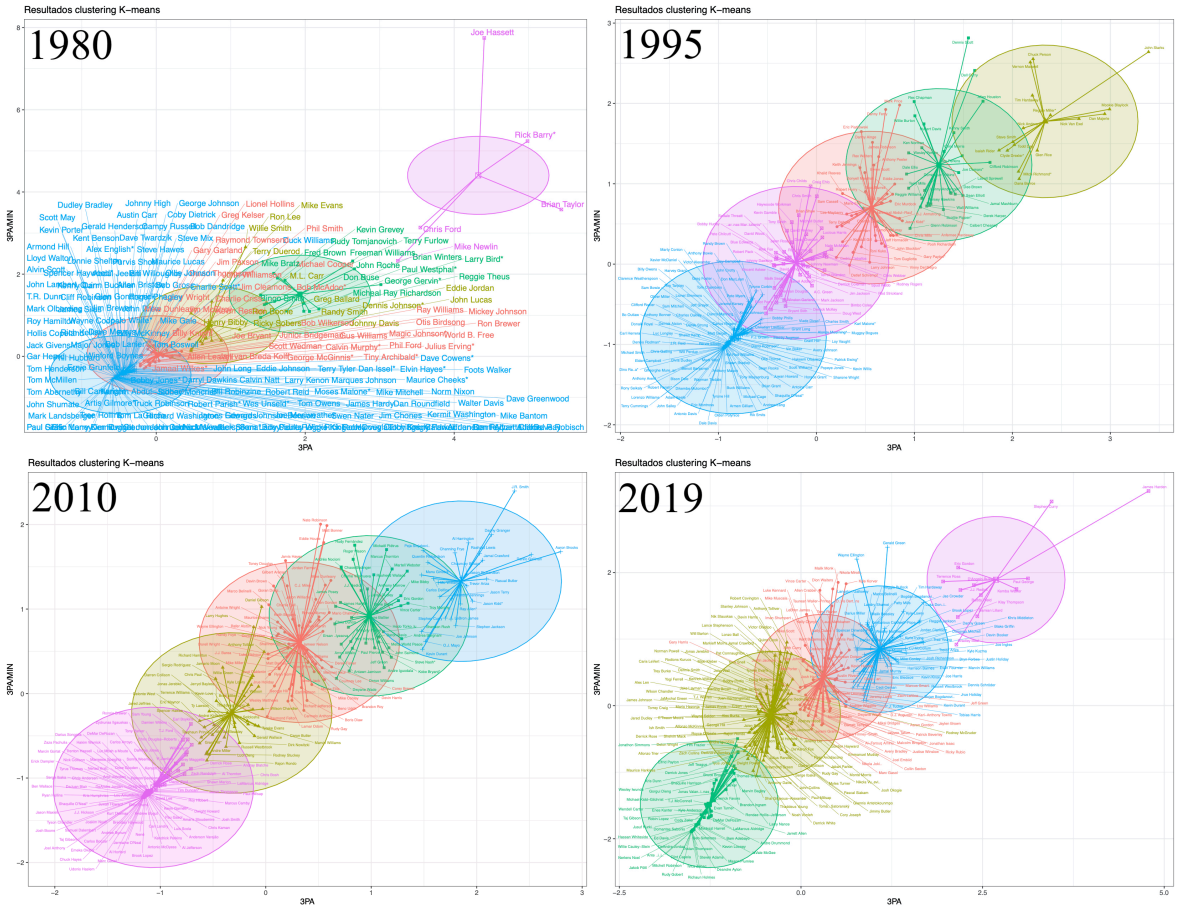
The database used for this section has been extracted from the basketball-reference website, one of the most reliable and popular directories on USA basketball statistics. The 1979/1980, 1994/1995, 2009/2010 and 2018/2019 seasons have been selected, and for each of them the shooting variables of all its players (in season totals) have been collected and new variables have been defined to be standardized by the minutes played by each player. It is a simple way to homogenize the data, but adequate because it is clear that there are many biases that can be determined by the playing time.

A number of modifications have also been made in order to form a more robust dataset to work with. Firstly, all players who have played less than 1000 minutes in a single season have been removed, which would be the equivalent of the least important players in the team. For players who have been part of several teams during the same season, the overall of the season has been taken into account. Finally, even though the players named with two positions constitute a very small percentage, only the position where they have played more minutes has been taken into account.

Thus, a total of 197 players for the 1979/1980 season, 227 players for the 1994/1995 season, 252 players for the 2009/2010 season and 276 players for the 2018/2019 season have been taken into account.

In addition, for each season 4 different subsets have been established combining the available variables of each season considered: Subset 1 (All shots), Subset 2 (Shots per minute), Subset 3 (Attempted shots), Subset 4 (Three pointers).

The following is an analysis of the first event of interest, which follows on from the introduction: How has the 3-point shot affected the players? For this analysis we will use the subset of three pointers for each season and for the clustering technique a k-means with k = 5.



It can be seen that, with the course of time and up to 2010, the groups are gradually becoming more homogeneous. Finally, in 2019 we can see the opposite effect to the low volume of 1980, now we can see a group that stands out from the rest because of its high expertise in this section.

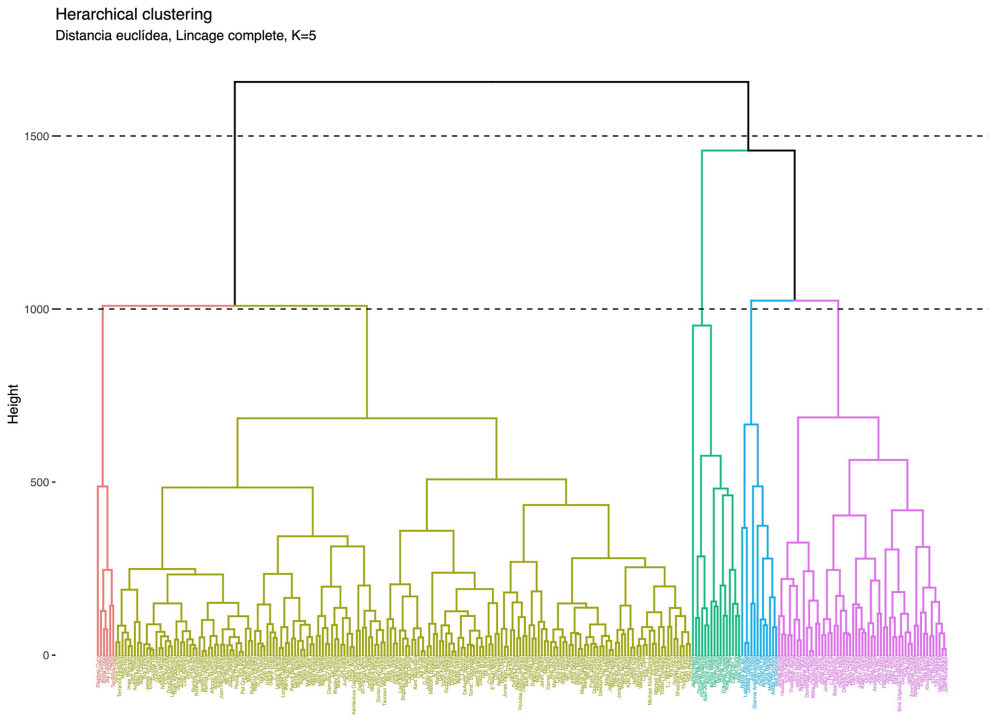
With this we can see that, as explained in the introduction, the change in the perspective of the 3-point shot has forced players to adapt to this style of play, but can we also find similarities in the shooting variables of players from the same position?

The subset of shots per minute for the 1994/1995 season is the one that most accurately meets this objective. It can be seen that the clusters formed are much more homogeneous in terms of the number of individuals that are part of each one. Moreover, if they are analyzed, it can be seen that, although they do not correctly classify each specific position, the difference between inside and outside players is very well appreciated. The first two clusters are mostly composed of inside players (PF and C). The third is the most heterogeneous and may contain the most versatile players who can play different positions. And finally, the last two clearly include the outside players (PG, SG and SF). The assignment table of the clusters is as follows:

	grupo real				
cluster	C	PF	PG	SF	SG
1	13	13	3	3	4
2	15	11	2	3	2
3	13	15	8	10	3
4	2	4	18	22	19
5	2	7	26	11	23

Finally, and to keep on with this topic, we present an analysis by means of a Hierarchical Clustering. It examines how the branches of the tree are formed depending on the number of clusters generated and whether the branches reproduce the inside and outside roles.

Analyzing the results of the different seasons, none of them fulfils this idea of pattern separation based on the shots attempted between inside and outside players. However, following the same reasoning and analyzing again the first idea from which we started, we can see how the specialization of the players of all the positions has made that the great majority of the outside players, as well as more than half of the inside players, are included in the same cluster.



In order to be able to find similarities between the players' positions through their shots, an analysis was carried out using classification techniques. For a better adjustment in the classification, a study of the variables not related to the shot, but with an important impact in the role that a player has associated to his position, was also made.

The databases used in this section have also been extracted from the basketball-reference website. Three reference seasons have been selected and will be used to check the models that have been generated: 1994/1995, 2009/2010 and 2018/2019.

Three models have been generated and all of them will consist of data from four consecutive seasons. The first one, with data from seasons 1990/1991 - 1993/1994. The second, with the 2005/2006 - 2008/2009 seasons. And the third, with the seasons 2013/2014 - 2017/2018. As in a set of four consecutive seasons most of the players are the same, each player of each season has been selected as a different individual.

The conditions of data selection are the same as for the previous section. Thus, the data set used in the formation of the first model was made up of 919 individuals, the second of 1289 and the third of 1364. The reference seasons maintain the same individuals as in the clustering section 227, 252 and 276 respectively.

All the models generated will be used to classify each of the reference seasons. The objective is to see the differences that exist between the classification with a model for its respective season and the comparison when classifying the seasons of distant years in time.

For the shooting analysis, the same variables will be used as in the previous section, including the field goal percentage for each type of shot and excluding the variables of shots per minute.

For the analysis not strictly related to shooting, in addition to the minutes played and the position of each player, different advanced statistical variables have been selected.

The first results to be used are those corresponding to the shooting. To begin with, we check how each model fits into its reference season (the one following its data to generate it). To do this, the classification is carried out using the decision tree and k-nearest neighbors techniques. Although both techniques have been applied for all data, only those that achieve the best accuracy are shown.

When applying the models generated by both techniques with their respective training data to the 1994/1995 season, it is verified that a better classification is obtained with the decision trees. The confusion matrix between the position in which the players have been classified and their actual position is presented.

With this confusion matrix the accuracy of the model is only 33.92%. Does this low value imply that the model is not good enough? Could it be that the shooting variables do not imply a determining role in relation to the position of a player? The answer to these questions needs a greater insight into the analysis. The interpretation must be refined taking into account the results that we have obtained with the clustering techniques. It would not be entirely fair to the model if its accuracy were measured only by the positions it correctly ranks, and even less so when considering a 5x5 table.

```
          Reference
Prediction  C PF PG SF SG
       C  18 15  0  1  0
       PF 11  7  0  2  0
       PG  7 14 41 29 35
       SF  5  8  1  4  3
       SG  0  4  6  9  7

Overall Statistics

               Accuracy : 0.3392
                 95% CI : (0.2779, 0.4048)
    No Information Rate : 0.2115
    P-Value [Acc > NIR] : 5.801e-06

                  Kappa : 0.1698
```

In order to find a solution to this problem, it has been decided to take into account that the classification of an adjacent position to another does not mean that it is wrong, because of the similarities that these positions share. In order to define which positions are adjacent to each other, we have taken into account the average volume of 3 point shots that the players of each position take. Therefore, the order of the

positions established through this parameter would be: Shooting Guard, Point Guard, Small Forward, Power Forward and Center. In this way it would be correct to classify any SG as PG and vice versa, any PG as SF and vice versa, etc. To automate this new matrix, a table was designed in Excel that would take into account all these conditions in addition to, obviously, the correct classification of the player's position.

Applying this to the results for each model applied to each season, the following summary is obtained (See Tables 26, 29, 35, 38, 40, 44 and 47 of the Annex and Tables 7 and 8 of the main work).

Model applied \ Data to be Classified	1994/1995 Normal	1994/1995 Fitted	2009/2010 Normal	2009/2010 Fitted	2018/2019 Normal	2018/2019 Fitted
1994/1995	33,92%	81,06%	38,49%	76,19%	33,33%	66,66%
2009/2010	40,97%	81,50%	40,87%	81,75%	35,14%	63,41%
2018/2019	35,68%	79,30%	38,10%	75,79%	42,39%	75,36%

As can be seen with these data, the most inaccurate classification takes place when the models of the 1994/1995 and 2009/2010 seasons are used to classify the 2018/2019 data which reaffirms the hypothesis that the biggest change has taken place after the revolution in shooting that has occurred in recent years.

In the same way as for shooting, an analysis of the variables of advanced statistics has been carried out, as mentioned above. In addition, in the shooting analysis it made more sense to maintain the order SG-PG-SF-PF-C, however, for this part of advanced statistics that has more to do with the role played by each position it would be more convenient to use the order PG-SG-SF-PF-C, which would imply that any SG classified as SF and vice versa would be correct. Following the same reasoning now any PG classified as SF and vice versa is considered as an error, since their roles are sufficiently different not to consider them as adjacent positions.

Making this adjustment with all models and seasons, this is the summary that is obtained (See Tables 52, 56, 60, 68, 72, 77, 81 and 85 of the Annex).

Model applied \ Data to be Classified	1994/1995 Normal	1994/1995 Fitted	2009/2010 Normal	2009/2010 Fitted	2018/2019 Normal	2018/2019 Fitted
1994/1995	60,79%	96,92%%	59,13%	97,22%%	49,64%	88,41%%
2009/2010	60,79%	96,92%	64,29%	98,02%	55,80%	93,48%
2018/2019	60,79%	93,39%%	63,10%	98,81%	63,77%	95,29%%

Finally, the importance of the different variables that affect a shooting situation, both in the offensive and defensive section, in its effectiveness was analyzed. Therefore, logistic regression techniques have been used, and although the main objective is not to predict if a particular shot is going to be successful, the models generated will be used to classify the same data to see how important the variables used in each model are.

The database used for this purpose has been extracted from the NBASavant website and consists of 50,000 shots made in the regular phase of the 2015/2016 season. This data set has variables referring to the shooter and his shooting variables for different distances, the closest defender, the physical characteristics of both and the shooting situation.

Therefore, the final set of data with which the analysis was performed was made up of 25,931 shots with 40 variables associated with each shot.
In addition to analyzing all the shots as a whole, a subset was extracted that counted the shots made from more than 5 feet away, since there is a big difference between them and everything that is dunks and layups, and therefore the most important variables involved in them have to be different as well.

First the set of all the shots will be analyzed using all the available variables. By fitting the data to a generalized linear model (GLM) with binomial link function, a multiple logistic regression model is generated. This model will be called the general model

As the variables "Action Type" and "Distance" seem to have a great predictive power, the models generated only with each of these variables have been checked. For this purpose, two simple logistic regression models have been carried out and their classifying power is compared with the general model (See Tables 102-110 of the Annex). This is the summary of them.

		Threshold					
		0.4		0.5		0.6	
Model	Actual / Prediction	Made	Missed	Made	Missed	Made	Missed
All features	Made	9821	3109	13184	5958	15246	8447
	Missed	6243	6758	2880	3909	818	1420
	Accuracy	63,94%		65,92%		64,27%	
	Sensitivity	61,14%		82,07%		94,91%	
	Specificity	68,49%		39,62%		14,39%	
Action Type	Made	11210	4439	12501	5536	15654	9096
	Missed	4854	5428	3563	4331	410	771
	Accuracy	64,16%		64,91%		63,34%	
	Sensitivity	69,78%		77,82%		97,45%	
	Specificity	55,01%		43,89%		7,81%	
Distance	Made	12409	5740	14559	8013	15361	8880
	Missed	3655	4127	1505	1854	703	987
	Accuracy	63,77%		63,29%		63,04%	
	Sensitivity	77,25%		90,63%		95,62%	
	Specificity	41,83%		18,79%		10,00%	

< 20%
20% - 60%
>60%

If you look at the accuracies of each model for the different thresholds that have been selected, you can see that they all have an accuracy of between 61% and 66% approximately. But it would not be strictly correct to take this as valid without looking at the data in detail, since for thresholds 0.5 and 0.6 the vast majority of shots are classified as made, which benefits the accuracy of the matrix since the dataset consists of many more made shots than misses (16,071 versus 9,869).

Therefore, for each particular model, not only the accuracy of the confusion matrix must be taken into account, but also its sensitivity and specificity. Based on this, the threshold is adjusted to ensure the confusion matrix has more consistent metrics. For the table above, the one that best meets this condition is 0.4.

Once we see that the highest accuracy that can be achieved is around 65%, the goal is to find the minimum number of variables that achieve similar accuracy with consistent sensitivity and specificity.

It is obtained that the minimum model that fulfills these premises is the one formed with the variables collected in the following table.

```
Coefficients:
                    Estimate  Std. Error z value             Pr(>|z|)
(Intercept)       -0.00652487  0.16629673   -0.04                 0.97
defender_distance -0.14460058  0.00837881  -17.26 < 0.0000000000000002 ***
`x^2`              0.00003896  0.00000190   20.54 < 0.0000000000000002 ***
`y^2`              0.00002981  0.00000176   16.97 < 0.0000000000000002 ***
dribbles          -0.06400551  0.00957088   -6.69        0.000000000023 ***
defender_weight    0.00461801  0.00051470    8.97 < 0.0000000000000002 ***
touch_time         0.11901772  0.01110412   10.72 < 0.0000000000000002 ***
`LESS THAN 5FT FG%` -0.02662394  0.00209425  -12.71 < 0.0000000000000002 ***
```

Therefore, it can be stated with some certainty that the variables that most influence the success or failure of a shot, from more to less influential (see Graph 70 in the Annex), and excluding the distance to the basket and the type of shot are: Touch Time (X_6), y^2 (X_3), Defender Distance (X_1), x^2 (X_2), Dribbles (X_4), Less Than 5 FT FG% (X_7) and Defender Weight (X_5).

Then the function generated with this model that returns the probability of belonging to group 1 (miss the shot) is given by:

$$\hat{p}(Y = 1|X_i) = \frac{e^{-0.006525-0.1446X_1+0.000039X_2+0.0000298X_3-0.064X_4+0.0046X_5+0.119X_6-0.0266X_7}}{1+e^{-0.006525-0.1446X_1+0.000039X_2+0.0000298X_3-0.064X_4+0.0046X_5+0.119X_6-0.0266X_7}} \quad i = 1,2,\dots,7$$

From the subset of data for shots over 5 feet from the basket there is a total of 7782 shots of which 3655 were made and 4127 were missed. The first difference that can be appreciated is that for this analysis we have a greater number of missed shots than those that were made, which nevertheless is not something unpredictable since it is logical that the shots closer to the basket have been the ones that have had more success.

Proceeding in the same way as for all the data, the results obtained are as follows. The table compares the general model with those using only the "Action Type" and "Distance" variables (See Tables 132-140 in the Annex).

Model	Actual / Prediction	Threshold 0.4 Made	Threshold 0.4 Missed	Threshold 0.5 Made	Threshold 0.5 Missed	Threshold 0.6 Made	Threshold 0.6 Missed
All features	Made	429	233	1628	1268	3053	2955
	Missed	3226	3894	2027	2859	602	1172
	Accuracy	55,55%		57,66%		54,29%	
	Sensitivity	11,74%		44,54%		83,53%	
	Specificity	94,35%		69,28%		28,40%	
Action Type	Made	289	166	1135	918	3259	3398
	Missed	3366	3961	2520	3209	396	729
	Accuracy	54,61%		55,82%		51,25%	
	Sensitivity	7,91%		31,05%		89,17%	
	Specificity	95,98%		77,76%		17,66%	
Distance	Made	0	0	617	667	3541	3917
	Missed	3655	4127	3038	3460	114	210
	Accuracy	53,03%		52,39%		48,20%	
	Sensitivity	0,00%		16,88%		96,88%	
	Specificity	100,00%		83,84%		5,09%	

< 20%
20% - 50%
>50%

The model with the minimum number of variables that achieve sufficiently correct results.

```
Coefficients:
                      Estimate Std. Error z value     Pr(>|z|)
(Intercept)            0.75282    0.13615    5.53 0.0000000321 ***
shot_distance          0.02132    0.00622    3.43      0.00061 ***
`10-14 FT FGM`        -0.89508    0.29839   -3.00      0.00270 **
`10-14 FT FGA`         0.41426    0.13627    3.04      0.00237 **
`5-9 FT FG%`          -0.01745    0.00287   -6.08 0.0000000012 ***
shot_type3PT Field Goal  0.25114    0.11784    2.13      0.03308 *
`25-29 FT FGA`        -0.05981    0.01994   -3.00      0.00270 **
defender_distance     -0.04896    0.01337   -3.66      0.00025 ***
```

The variables according to their importance from highest to lowest in this model (See Graph 75 in Annex) are: 10-14 FT FGA (X_3), 10-14 FT FGM (X_2), 5-9 FT FG% (X_4), Shot Distance (X_1), Shot Type (3PT Field Goal) (X_5), Defender_Distance (X_7) and 25-29 FT FGA (X_6).

The function generated with this model that returns the probability of belonging to group 1 (miss the shot) is given by:

$$\hat{p}(Y = 1|X_i) = \frac{e^{0.75282+0.02132X_1-0.895X_2+0.41426X_3-0.01745X_4+0.25114X_5-0.0598X_6-0.04896X_7}}{1 + e^{0.75282+0.02132X_1-0.895X_2+0.41426X_3-0.01745X_4+0.25114X_5-0.0598X_6-0.04896X_7}} \; i = 1,2,\dots,7$$

The difference discovered is that one of the predictors is a categorical variable (Shot Type). This indicates that, if for this variable the shot on which the prediction is made is a three pointer, then $X_5 = 1$ but if it is a 2 point shot, then $X_5 = 0$.

After all the analysis carried out in this paper, a number of conclusions can be drawn that support the ideas developed in the introduction.

On the one hand, in relation to the multi-analysis carried out through the clustering techniques. To differentiate between two large groups such as inside and outside players (only taking into account the shooting variables) it is more accurate to include the PF as an outside player, despite building a group with 4 positions and another with only 1 position. However, since the differences between positions are no longer so significant, nowadays when differentiating players by their shooting nature, rather than talking about their specific position, it would be more useful to see in which cluster they are located based on their shooting variables.

On the other hand, regarding the analysis using classification techniques two conclusions can be deduced. Although today they are players in their position that continue to have the shooting patterns that their predecessors had, most have been evolving. Therefore, to be classified correctly it has become much more complicated since many similarities exist between the different positions. Nevertheless, if instead of taking into account the shooting variables you use the advanced statistical variables you can see that, despite the fact that the game has changed enormously, the positions are still linked in general to specific roles and contribute to the different aspects of the game in a more similar way than before.

Finally, after the analysis seen in the regression section the most evident thing that can be observed is that the shots made at more than 5 feet and the set of all the shots find very significant differences in the variables that are more important in their effectiveness. The results suggest that for shots very close to the basket the field goal percentages of the different zones are not really relevant, but what is more significant are the variables that determine the nature of the shot. However, for the more distant shots, the variables that determine how good a shooter is (or at least if he makes many shots) in the different zones of the field already play such a bigger role.