



Universidad de Valladolid

Facultad de Ciencias

TRABAJO FIN DE GRADO

Grado en Estadística

MÉTODOS ESTADÍSTICOS EN COMPETICIONES DEPORTIVAS DE BALONCESTO: LA NBA.

Autor: D. Javier Martínez García

Tutor/es:

Dr. Pedro César Álvarez Esteban

Dr. Hristo Inouzhe Valdes

INDICE

1.	INTRODUCCIÓN.....	7
1.1.	FORMATO DE LA NBA Y ESTADÍSTICAS.....	7
1.2.	CONTEXTO DEL TRABAJO	10
1.3.	OBJETIVOS.....	10
2.	MÉTRICAS PARA LOS EQUIPOS DE LA NBA	11
2.1.	POINT DIFFERENTIAL (+/-)	11
2.2.	PYTHAGOREAN WINS AND LOSES (PW/PL)	16
2.3.	EXPECTED WINNING PERCENTAGE.....	19
2.4.	SIMPLE RATING SYSTEM (SRS)	20
2.5.	NET RATING.....	23
2.5.1.	POSSESSIONS (poss).....	23
2.5.2.	OFFENSIVE RATING.....	24
2.5.3.	DEFENSIVE RATING.....	25
2.5.4.	NET RATING	27
2.6.	PACE.....	28
2.7.	TRUE SHOOTING PERCENTAGE (TS%).....	32
2.8.	FOUR FACTORS (EFG%, TOV RATE, OREB%, FT RATE).....	33
2.8.1.	EFFECTIVE FIELD GOAL PERCENTAGE (eFG%).....	34
2.8.2.	TURNOVER RATE (TOV%)	37
2.8.3.	OFFENSIVE and DEFENSIVE REBOUND PERCENTAGE (ORB% and DRB%)	38
2.8.4.	FREE THROW RATE	39
3.	K-MEANS CLUSTERING CON TODAS LAS VARIABLES	41
3.1.	EXPLICACIÓN TEÓRICA DEL PROCEDIMIENTO CLUSTER DE K-MEANS.....	42
3.2.	ANÁLISIS DE COMPONENTES PRINCIPALES (PCA).....	42
3.3.	REALIZACIÓN DEL MÉTODO K-MEANS.....	46
4.	MÉTODOS DE SELECCIÓN DE VARIABLES	52
4.1.	COEFICIENTE DE CORRELACIÓN DE PEARSON	52
4.2.	ALGORITMOS DE SELECCIÓN DE VARIABLES.....	55
4.2.1.	BACKWARD	55
4.2.2.	FORWARD.....	56
5.	MODELOS de regresión lineal múltiple	57
5.1.	EXPLICACIÓN TEÓRICA DE LA REGRESIÓN LINEAL MÚLTIPLE.....	57
5.2.	REGRESIÓN LINEAL MÚLTIPLE CON TODAS LAS VARIABLES.....	57
5.3.	ESTUDIO DE LOS “FOUR FACTORS” DE DEAN OLIVER	59
6.	K-MEANS CLUSTERING PARA LOS MODELOS DE REGRESIÓN	64

6.1.	<i>K-MEANS SOBRE MODELO “ATAQUE-DEFENSA”</i>	64
6.2.	<i>K-MEANS SOBRE MODELO DE “FOUR FACTORS”</i>	66
7.	<i>PREDICCIÓN DE PRESENCIA DE LOS EQUIPOS EN PLAY-OFFS UTILIZANDO REGRESIÓN LOGÍSTICA Y RANDOM FOREST</i>	68
7.1.	<i>EXPLICACIÓN TEÓRICA REGRESIÓN LOGÍSTICA</i>	68
7.2.	<i>REGRESIÓN LOGÍSTICA SOBRE EL MODELO “ATAQUE-DEFENSA”:</i>	68
7.3.	<i>REGRESIÓN LOGÍSTICA SOBRE EL MODELO “FOUR FACTORS”:</i>	70
7.4.	<i>EXPLICACIÓN TEÓRICA ALGORITMO RANDOM FOREST:</i>	71
7.5.	<i>RANDOM FOREST SOBRE MODELO “ATAQUE-DEFENSA”:</i>	72
7.6.	<i>RANDOM FOREST SOBRE MODELO “FOUR FACTORS”:</i>	74
8.	<i>RESULTADOS Y CONCLUSIONES</i>	77
9.	<i>BIBLIOGRAFÍA</i>	79
10.	<i>CÓDIGO</i>	81

AGRADECIMIENTOS

A mi tutor, Pedro César Álvarez, por su guía, consejo y ayudarme a enfocar el proyecto desde un punto de vista más experto. A José Belarmino Pulido, por su ayuda y por compartir sus conocimientos conmigo, también baloncestísticos.

A Hristo Inouzhe, por ayudarme a trazar un camino de este proyecto al principio.

A mis compañeros de proyecto Héctor Sáenz y Álvaro Berrio, por todo el trabajo complementario realizado.

Por último, a mis amigos y compañeros del Grado de Estadística, por todos estos años magníficos.

RESUMEN

El objetivo principal de este proyecto es el estudio de algunos **métodos estadísticos** dentro de las competiciones deportivas de baloncesto, más concretamente de la NBA (National Basketball Association). Para ello se analizarán distintas métricas estadísticas aplicadas a los distintos equipos participantes en todas las temporadas de la liga disputadas hasta la fecha.

Este proyecto tendrá como meta la comprensión de dichas métricas y comprobación de su funcionamiento a la hora de predecir el éxito y rendimiento futuro de un equipo.

Para llevar a cabo las experimentaciones anteriores, se realizarán explicaciones detalladas con ejemplos de las métricas más conocidas y usadas por los analistas de la NBA. También mediante distintos métodos de predicción y procedimientos de agrupamiento se nos permitirá conocer cuáles son las métricas o conjunto de métricas que mejor miden el éxito de un equipo. Este éxito en este proyecto estará definido tanto por el número total de victorias que se obtienen (a mayor número, mayor es dicho éxito), tanto como por la clasificación de un determinado equipo para los *Play-Offs* o posttemporada de la NBA.

ABSTRACT

The main target of this project is to analyse some **statistical methods** for some basketball competitions, more precisaly, the NBA. Because of that, some different statistical metrics from all the teams that have played in the history of the league will be analyse.

The main goal of this project will be the understanding of this metrics and the verification of its performance when we want to predict the future success and efficiency of the teams.

For those experimentations, some detailed explanations with the more used metrics by the NBA analysts will be done. Also, using different prediction methods and clustering procedures, will allow us to know what are the metrics or a set of them that better measure the success of a team. The success will be define in this project as the total number a victories a team get and as the classification of the teams to the *Play-Offs* of the NBA.

ACRÓNIMOS

Para facilitar la comprensión y lectura de esta memoria, se van a introducir algunos acrónimos que pueden aparecer en las explicaciones y ecuaciones.

Box score:

MP	Minutos jugados.	FTA	Tiros libres realizados.
FGM	Tiros de campo anotados.	FT%	Porcentaje en tiros libres.
FGA	Tiros de campo realizados.	ORB	Rebotes ofensivos capturados.
FG%	Porcentaje tiros de campo anotados.	DRB	Rebotes defensivos capturados.
3PM	Triples anotados.	TRB	Total de rebotes capturados.
3PA	Triples realizados.	AST	Asistencias.
3P%	Porcentaje en triples.	STL	Robos.
2PM	Tiros de dos puntos anotados.	BLK	Tapones.
2PA	Tiros de dos puntos realizados.	TOV	Pérdidas.
2P%	Porcentaje en tiros de dos puntos.	PF	Faltas personales.
FTM	Tiros libres anotados.	PTS	Puntos anotados.

Métricas:

W/L	Victorias/ Derrotas	NETRTG	Diferencia entre eficiencia ofensiva y defensiva (también como <i>NRtg</i>).
PW/PL	<i>Pythagorean Wins/ Loses</i>	POSS	Número de posesiones
MOV	Margen de victoria o diferencial de puntos (+/-)	TS%	<i>True shooting percentage</i> (otra forma de medir el porcentaje de tiro).
HCA	Ventaja de campo.	eFG%	Efectividad en los lanzamientos de campo (actualización de <i>FG%</i>).
SRS	<i>Simple Rating System</i> .	TOV%	Ratio de pérdidas (por 100 posesiones).
SOS	Dificultad del calendario.	OREB%	Porcentaje de rebotes ofensivos.
OFFRTG	Eficiencia ofensiva (aparece también como <i>ORtg</i>).	DRB%	Porcentaje de rebotes defensivos.
DEFRTG	Eficiencia defensiva (también como <i>DRtg</i>).	FT/FGA	Ratio de tiros libres.

1. INTRODUCCIÓN

Con el paso de los años y más en especial en las últimas décadas, la sociedad ha llevado a cabo un avance continuo. No ha sido menos para la tecnología y esto ha hecho que la forma de vida de las personas haya cambiado por completo y lo siga haciendo a gran velocidad.

Lo mismo ocurre con el baloncesto y en lo que a este trabajo se refiere, la estadística avanzada dentro de la NBA (*National Basketball Association*). Si hay algo que ha revolucionado el mundo del baloncesto sobre todo desde comienzo de siglo es el **nivel de análisis estadístico**, ya que actualmente se analiza prácticamente cualquier situación que puede llegar a ocurrir sobre una pista de baloncesto.

A raíz de este avance que se ha producido, la forma de entender el baloncesto NBA en la última década ha cambiado enormemente comparándolo con todo el período anterior de historia de la liga. Esto se debe, entre otras cosas, al auge, comprensión y expansión de la **estadística avanzada**, o como también se conoce, *Analytics*.

El uso de estadísticas más básicas o convencionales (puntos, rebotes, asistencias, robos...) se ha ido quedando al margen a medida que el baloncesto crecía en su análisis y se requería un mayor número de herramientas para entender y analizar al máximo posible todo lo que ocurría sobre una pista de baloncesto. Hace unas décadas, en la NBA no era posible contabilizar numerosas estadísticas y datos que hoy nos sirven de gran ayuda a la hora de entender el juego.

Debido al desarrollo de los sistemas informáticos avanzados ha sido posible la aplicación de una serie de **avances técnicos en el mundo de baloncesto**. Esto sumado a los métodos más convencionales o rudimentarios, nos va a dotar de más posibilidades y opciones a la hora de poner solución a los distintos problemas que se pueden ir encontrando [26].

Dentro de los deportes de competición, el análisis estadístico se encontraba en el béisbol de una forma más exhaustiva. Una adaptación de esta fórmula del éxito del uso de estadística avanzada en los deportes, más concretamente el béisbol, lo encontramos en la película *Moneyball*, donde Billy Beane, mánager de los Oakland Athletics, es capaz de conseguir un gran rendimiento de un equipo de nivel mediocre, a partir de técnicas que estaban adelantadas a la época.

En el caso del baloncesto, resulta más difícil el análisis estadístico debido a la complejidad del juego en equipo que conlleva este deporte. Actualmente la mayoría de equipos de la NBA hacen uso de esta herramienta y algunos de ellos tales como Boston, Houston, Oklahoma, Portland y San Antonio fueron los primeros en crear un departamento única y exclusivamente dedicado a monitorear y crear estadísticas [27].

1.1. FORMATO DE LA NBA Y ESTADÍSTICAS

La NBA o National Basketball Association, es la principal liga tanto estadounidense como mundial de baloncesto. Esta competición fue fundada en el año 1946 y en sus inicios contaba con 11 equipos de diferentes lugares del país. A lo largo de los años se han ido añadiendo más equipos a la competición y algunos ya existentes han cambiado de sede o de nombre. En la actualidad, la NBA cuenta con un total de 30 equipos, los cuales se distribuyen en dos Conferencias: Este y Oeste.

Dentro de cada conferencia, los equipos se repartirán en 3 divisiones:

CONFERENCIA ESTE:

- División Atlántica: Toronto Raptors, Boston Celtics, Philadelphia 76ers, Brooklyn Nets y New York Knicks.
- División Central: Milwaukee Bucks, Indiana Pacers, Chicago Bulls, Detroit Pistons y Cleveland Cavaliers.
- División Sur-Este: Miami Heat, Orlando Magic, Washington Wizards, Charlotte Hornets y Atlanta Hawks.

CONFERENCIA OESTE:

- División Nor-Oeste: Denver Nuggets, Utah Jazz, Oklahoma City Thunder, Portland Trail Blazers y Minnesota Timberwolves.
- División Pacífica: Los Angeles Lakers, Los Angeles Clippers, Sacramento Kings, Phoenix Suns y Golden State Warriors.
- División Sur-Oeste: Houston Rockets, Dallas Mavericks, Memphis Grizzlies, New Orleans Pelicans y San Antonio Spurs.



Figura 1: Mapa equipos NBA organizados en divisiones y conferencias (Pinterest, Kathy Hendricks).

La competición se va a encontrar dividida en dos fases. La primera fase hace referencia a la liga regular, donde cada equipo juega un total de 82 partidos divididos de forma similar entre local y visitante. Normalmente, en cada temporada esta fase comienza a finales del mes de octubre y finaliza a mediados del mes de abril. Hay algunas excepciones, que se tratan de los años de *lockout* o cierre patronal, cuando su inicio se llega a retrasar hasta finales del mes de diciembre.

Al finalizar esta primera fase, los 8 equipos que presenten un mejor porcentaje de victorias en cada conferencia serán clasificados para la segunda fase, denominada *Playoffs* o postemporada. Los 14 equipos restantes (7 de cada conferencia) quedarán definitivamente eliminados.

Esta segunda fase o Playoffs está formado por 4 Rondas: Primera Ronda, Semifinales de Conferencia, Finales de Conferencia y Finales de la NBA.

Cada Ronda se disputarán entre dos equipos al mejor de 7 partidos (primer conjunto que alcance las 4 victorias pasará de Ronda). El equipo que mejor clasificación haya conseguido dentro de su Conferencia, o en el caso de las Finales, de toda la NBA, tendrá ventaja de campo, lo que significa que los dos primeros partidos los disputará como local además de un hipotético séptimo y decisivo.

Los ocho primeros equipos de cada conferencia van a enfrentarse entre sí y el orden de dichos enfrentamientos para Primera Ronda es: #1 vs #8, #2 vs #7, #3 vs #6 y #4 vs #5.

En lo que a las estadísticas se refiere, estas son recogidas oficialmente por la propia NBA, y van a poder ser accesibles en la web dedicada a ellas: Stats.nba.com [1].

Existen otras webs cuya accesibilidad también es gratuita y que al igual que la anterior cuenta con una gran abundancia de datos. La más conocida y que más se va a utilizar y hacer referencia a lo largo de este trabajo es la de *Basketball-reference.com* [2]. Otras webs con estadísticas muy completas y métricas muy utilizables tales como *ESPN* [3] y *FiveThirtyEight.com* también se han tenido muy en cuenta, al igual que *nbastuffer.com* [5].

Para entender estas métricas, es necesario el conocimiento anterior algunas estadísticas básicas.

Rk	Team	G	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
1	Sacramento Kings*	82	244.9	38.2	85.0	.449	5.8	16.5	.354	32.4	68.5	.472	19.5	25.3	.771	12.0	33.0	45.0	22.6	9.7	5.3	14.9	19.5	101.7
2	Milwaukee Bucks*	82	241.2	38.0	82.9	.458	6.9	18.1	.379	31.1	64.8	.480	18.0	22.9	.787	11.9	30.5	42.4	22.5	8.2	4.7	13.7	23.5	100.7
3	Los Angeles Lakers*	82	242.7	37.9	81.5	.465	5.4	15.5	.344	32.6	66.0	.494	19.4	28.5	.683	13.2	31.5	44.7	23.0	6.9	6.0	14.4	22.8	100.6
4	Dallas Mavericks*	82	241.5	37.6	81.9	.459	6.3	16.5	.381	31.3	65.4	.479	18.9	23.8	.794	10.1	31.4	41.5	21.2	7.5	6.0	13.9	23.3	100.5
5	Toronto Raptors*	82	243.4	37.2	85.0	.437	5.2	14.2	.369	31.9	70.8	.451	18.1	24.2	.747	13.6	30.8	44.5	24.4	7.3	6.3	13.2	21.3	97.6
6	Orlando Magic*	82	243.4	36.7	83.8	.438	6.0	16.4	.364	30.8	67.4	.456	18.0	25.2	.714	13.0	29.9	42.9	22.0	8.2	5.9	15.2	25.3	97.5
7	Minnesota Timberwolves*	82	241.8	38.4	83.8	.458	3.9	11.0	.357	34.5	72.8	.473	16.6	21.2	.785	12.2	30.1	42.3	25.4	8.3	5.6	13.9	23.2	97.3

Figura 2: Ejemplo estadísticas básicas por partido de los equipos de la temporada 2000/01 (basketball-reference.com).

En la *Figura 2* encontramos el resumen estadístico por partido de los siete mejores equipos de la temporada 2000/01 estadísticamente hablando. En este ejemplo contamos con las estadísticas básicas más importantes y que resultan esenciales a la hora de entender el resto de las métricas y estadísticas que van a aparecer en este trabajo.

Como hemos dicho, todas las estadísticas que aparecen en la *Figura 2* están calculadas por partido para cada uno de los equipos. Aquí encontramos algunas básicas como la media de puntos por partido anotados por un determinado equipo (representado como *PTS*), los rebotes que captura de media (*TRB*, desglosados en ofensivos y defensivos, *ORB* y *DRB* respectivamente) o las asistencias medias que consigue repartir un equipo por partido (*AST*). Luego otras menos llamativas, pero igual de importantes como los robos medios realizados por un equipo (*STL*), los tapones medios realizados (*BLK*), las pérdidas medias que comete un equipo (*TOV*) o las faltas personales que realizan (*PF*).

El resto de estadísticas que aparecen en la *Figura 2* también son esenciales a la hora de entender este trabajo, pero al ser más complejas estas serán explicadas según vayan apareciendo.

1.2. CONTEXTO DEL TRABAJO

Este trabajo va a estar encuadrado en un proyecto común junto con otros dos más, cuyo principal objetivo es el análisis de datos correspondientes a la liga de baloncesto norteamericana NBA.

El primero de ellos hace referencia a la extracción y almacén de los datos. Para este procedimiento ha sido muy importante la colaboración de todos integrantes participantes en este proyecto, para que la persona encargada de esta parte tuviera conocimiento sobre qué datos iban a ser esenciales a la hora de continuar con el siguiente paso. Para esta primera etapa se han usado los servicios de *Cloudera* con *Hive* para el almacén de los datos. Esta tarea ha sido llevada a cabo por **Héctor Sáenz Niño**, estudiante del Grado en Ingeniería Informática.

El segundo trabajo tratará de la predicción de resultados de partidos en la NBA. Esta tarea va a ser llevada a cabo por **Álvaro Berrio Galindo**, estudiante del doble grado en Estadística e Informática, INDAT.

1.3. OBJETIVOS

El principal objetivo de este proyecto será el estudio de distintas métricas y estadísticas de los equipos de la NBA, para así poder compararlos y ver cuáles de ellas van a ser más precisas y nos van a proporcionar mejores resultados. Además, se llevará a cabo el estudio de modelos más recientes formados por la unión de estas métricas y que nos proporcionan mejores predicciones a la hora de medir el éxito y rendimiento de un equipo.

Por lo tanto, el objetivo principal de este proyecto tratará en la comprensión de dichas métricas y estadísticas y valorar su funcionamiento a la hora de predecir el rendimiento de los distintos equipos. Además, esto servirá a Álvaro como base para entender cuáles son las métricas más importantes y que debería prestar más atención a la hora de realizar su trabajo de predecir resultados en distintos partidos de esta competición, la NBA.

2. MÉTRICAS PARA LOS EQUIPOS DE LA NBA

La estadística avanzada en el baloncesto, y más concretamente en este trabajo en la NBA, hace referencia a analizar estadísticas a través de evidencia objetiva.

En este apartado se va a llevar a cabo la explicación de las métricas para los equipos NBA más utilizadas hasta el momento, poniendo ejemplos de cada una de ellas a partir de equipos pertenecientes a cualquier momento histórico de esta liga. Esto nos permitirá cuantificar su rendimiento y poder compararlos y a su vez, medir su éxito futuro.

2.1. POINT DIFFERENTIAL (+/-)

También conocida como margen de victoria de un equipo (Margin Of Victory, **MOV**). Esta métrica se corresponderá con la diferencia numérica entre los puntos anotados y los puntos recibidos de un equipo. Para entender las diferentes métricas que se van a ir mostrando tenemos que conocer el significado del término Récord dentro del vocabulario de la NBA. Este va a hacer referencia al balance de victorias-derrotas que obtiene un equipo de esta liga, es decir, al número de victorias y derrotas de dicho equipo esa temporada.

En el gráfico siguiente se mostrarán las mejores diferencias medias de puntos que han conseguido distintos equipos a lo largo de la historia de la NBA (*statmuse.com, 2020 [6]*):

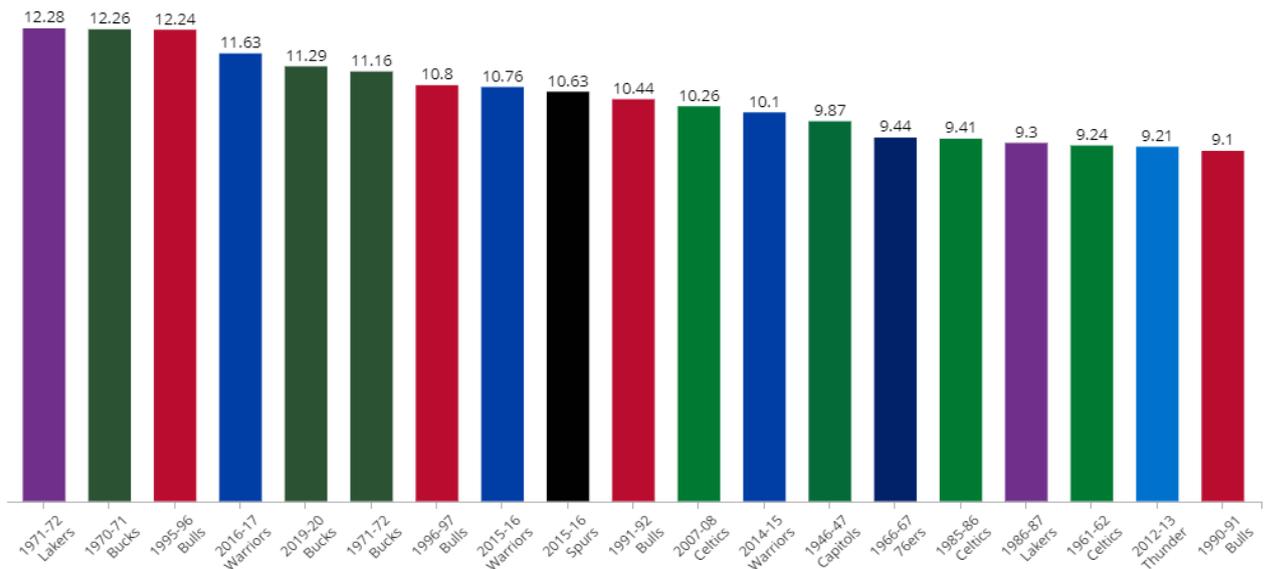


Gráfico 1. Mejores diferencias de puntos en la historia de la NBA.

Cuando observamos el *Gráfico 1*, podemos comprobar que el equipo que consiguió una mayor diferencia de puntos media durante una temporada son los Lakers de la temporada 1971/1972, equipo en el que destacan jugadores tales como Jerry West y Wilt Chamberlain, con un diferencial de 12.28.

En este ranking aparecen equipos ilustres de la historia de la NBA, tales como los Chicago Bulls de Michael Jordan (en cuatro temporadas diferentes, 95/96, 96/97, 91/92 y 90/91), los Warriors del mejor récord de la historia de la NBA (73-9 en la temporada 2015/2016) u otros como los LA Lakers de Magic Johnson de la temporada 86/87 y los Boston Celtics de Larry Bird de la temporada 85/86.

Teniendo en cuenta el ranking del *Gráfico 1*, tenemos diferentes datos de esos equipos en esas temporadas donde obtuvieron los mejores márgenes medios de puntos de la historia de la NBA (basketball-reference.com [2]):

EQUIPOS	POINT DIFFERENTIAL (MOV)	MEDIA PUNTOS ANOTADOS (RANKING NBA)	MEDIA PUNTOS RECIBIDOS (RANKING NBA)	LOGROS ESA TEMPORADA (RECORD V-D)
Lakers 71/72	12.28	121 (1º)	108.7 (7º)	Campeón (69-13)
Bucks 70/71	12.26	118.4 (1º)	106.2 (2º)	Campeón (66-16)
Bulls 95/96	12.24	103.1 (1º)	92.3 (6º)	Campeón (69-13)
Warriors 16/17	11.63	115.9 (1º)	104.3 (11º)	Campeón (67-15)
Bucks 19/20	11.29	118.6 (1º)	107.4 (8º)	Mejor récord de la temporada* (53-12)
Bucks 71/72	11.16	114.6 (4º)	103.5 (2º)	Finales de conferencia oeste (63-19)
Bulls 96/97	10.8	103.1 (1º)	92.3 (6º)	Campeón (69-13)
Warriors 15/16	10.76	114.9 (1º)	104.1 (19º)	Finales (73-9)
Spurs 15/16	10.63	103.5 (10º)	92.9 (1º)	Semifinales conferencia Oeste (67-15)
Bulls 91/92	10.44	109.9 (5º)	99.5 (3º)	Campeón (67-15)
Celtics 07/08	10.26	100.5 (11º)	90.3 (2º)	Campeón (66-16)
Warriors 14/15	10.1	110 (1º)	99.9 (15º)	Campeón (67-15)
76ers 66/67	9.44	125.2 (1º)	115.8 (3º)	Campeón (68-13)
Celtics 85/86	9.41	114.1 (16º)	104.7 (3º)	Campeón (67-15)
Lakers 86/87	9.3	117.8 (2º)	108.5 (12º)	Campeón (65-17)
Celtics 61/62	9.24	121.1 (3º)	111.9 (1º)	Campeón (60-20)
Thunder 12/13	9.21	105.7 (3º)	96.5 (9º)	Semifinales conferencia oeste (60-22)
Bulls 90/91	9.1	110 (7º)	101 (4º)	Campeón (61-21)

Figura 3: Mejores MOV de la historia de la NBA.

*Temporada en la que se produjo una pausa por el problema del coronavirus. Récord del equipo en el momento en el que paró dicha temporada.

Fijándose en los datos que tenemos en la *Figura 3*, podemos observar que la inmensa mayoría de los equipos cuyo margen de victoria media en una temporada figura entre los mejores de la historia, consiguieron o ser campeones de la NBA o llegaron a rondas muy avanzadas en los playoffs. Vemos también que los récords de victorias-derrotas de estos equipos son también muy positivos.

Cabe destacar que el denominador común que van a tener estos equipos, en general (podemos encontrar excepciones), es que cuando el margen de victoria es muy grande se trata de equipos muy anotadores y no necesariamente de equipos que encajan pocos puntos [6].

Como se anunció anteriormente, entre estos equipos con un diferencial de puntos muy elevados, podemos encontrar excepciones en los equipos que destacan por encajar pocos puntos y no necesariamente estar dentro del top 10 de equipos que en su temporada anotaron más puntos de media (como es el caso de los Celtics 07/08 o de los Spurs 15/16).

A lo largo de la historia ha habido grandes equipos eminentemente defensivos (recibían pocos puntos), y gracias a los datos con los que contamos (*basketball-reference.com* [2]) podemos ver que estos equipos no necesariamente contaban con un margen de victoria elevado, pero sí que

se trataba de equipos cuyas aspiraciones eran máximas en sus temporadas. Tenemos la siguiente tabla:

EQUIPOS	POINT DIFFERENTIAL (MOV)	PUNTOS ANOTADOS (RANKING NBA)	PUNTOS RECIBIDOS (RANKING NBA)	LOGROS ESA TEMPORADA (RECORD V-D)
Pistons 03/04	5.84 (4º)	90.1 (24º)	84.3 (1º)	Campeón (54-28)
76ers 00/01	4.28 (7º)	94.7 (15º)	90.4 (5º)	Finales (56-26)
Spurs 02/03	5.41 (3º)	95.8 (11º)	90.4 (3º)	Campeón (60-22)

Figura 4: Ejemplos MOV de grandes equipos defensivos.

La diferencia de puntos media de estos equipos es considerablemente pequeña si la comparamos con otros grandes equipos de la historia de la liga, sin embargo, estos sí que consiguieron grandes logros en sus respectivas temporadas. En resumen, vemos que los grandes equipos de la historia que destacan por recibir pocos puntos, su diferencial de puntos tiende a ser más pequeño, pero eso no quita que no obtengan grandes resultados.

A continuación, contamos con los datos de la diferencia de puntos para cada uno de los equipos en la temporada 2019/20 (hasta la pausa que se produjo debido a la pandemia del coronavirus, en marzo del 2020), además de los resultados también para cuando dichos equipos juegan en casa o de visitantes (teamrankings.com 2020 [7]):

EQUIPOS	MOV TEMPORADA 19/20	MOV PARTIDOS COMO LOCAL 19/20	MOV PARTIDOS COMO VISITANTE 19/20
Milwaukee Bucks 19/20	+11.3	+14.7	+8.2
LA Lakers 19/20	+7.4	+9.3	+5.6
LA Clippers 19/20	+6.5	+9.1	+4.0
Toronto Raptors 19/20	+6.5	+8.7	+4.2
Boston Celtics 19/20	+6.2	+8.2	+4.2
Dallas Mavericks 19/20	+6.0	+6.0	+6.1
Houston Rockets 19/20	+3.8	+6.8	+0.7
Miami Heat 19/20	+3.2	+9.3	-2.7
Utah Jazz 19/20	+3.2	+4.4	+2.0
Denver Nuggets 19/20	+3.0	+6.0	-0.2
Oklahoma City Thunder 19/20	+2.5	+3.9	+0.9
Philadelphia 76ers 19/20	+2.2	+10.4	-5.3
Indiana Pacers 19/20	+1.9	+4.5	-0.5
Brooklyn Nets 19/20	-0.6	+1.7	-3.0
New Orleans Pelicans 19/20	-0.8	+1.1	-2.8
Orlando Magic 19/20	-1.0	-0.4	-1.5
Memphis Grizzlies 19/20	-1.1	+0.6	-2.8
Phoenix Suns 19/20	-1.4	+0.1	-3.0
Portland Trail Blazers 19/20	-1.6	+1.9	-4.9
San Antonio Spurs 19/20	-1.8	+1.1	-4.3
Sacramento Kings 19/20	-1.9	-2.1	-1.8
Chicago Bulls 19/20	-3.1	-1.5	-4.8
Detroit Pistons 19/20	-3.6	-1.8	-5.2
Washington Wizards 19/20	-4.0	-0.7	-7.4
Minnesota Timberwolves 19/20	-4.3	-4.7	-3.7
New York Knicks 19/20	-6.5	-5.1	-7.8
Charlotte Hornets 19/20	-6.8	-5.0	-8.2
Cleveland Cavaliers 19/20	-7.9	-6.5	-9.6
Atlanta Hawks 19/20	-8.0	-2.8	-13.3

Golden State Warriors 19/20	-8.7	-8.7	-8.7
--------------------------------	------	------	------

Figura 5: MOV de los equipos de la temporada 2019/20.

Los mejores equipos de la liga siempre presentan una diferencia de puntos media positiva durante la temporada y, corroborándolo con datos de otras temporadas (basketball-reference.com [2]), normalmente los equipos con mejores resultados a lo largo de la temporada presentan una diferencia de puntos media durante la temporada superior a 2.0. En el caso opuesto se encuentran los peores equipos, que presentan un récord negativo y cuyo MOV también será negativo.

Como puede intuirse y salvo excepciones, los equipos consiguen un mayor margen de victoria media en los partidos que juegan en casa con respecto a los partidos que actúan como visitantes (nbastuffer.com, “Home Court Advantage”).

Esto tiene una estrecha relación con una métrica muy conocida (nbastuffer.com [5]) denominada **Ventaja de Campo (Home Court Advantage, HCA)**. Este índice se corresponde con el efecto en el resultado que tienen diferentes factores que deberían tener un efecto positivo con el equipo que juega en casa y generalmente un efecto negativo con los equipos que juegan como visitantes. En la *Figura 5* vemos que se producen excepciones, y es que hay equipos cuyos mejores resultados se obtienen como visitantes (en la temporada 19/20, Dallas Mavericks, Sacramento Kings y Minnesota Timberwolves). En el caso contrario, observamos equipos donde la ventaja de campo es un efecto muy destacado, siendo el caso de Philadelphia 76ers el más marcado, ya que tiene un diferencial de puntos medio de +10.4 en sus partidos como local, mientras que cuando son visitantes esa diferencia media disminuye hasta -5.3 puntos. La ventaja de campo normalmente está valorada entre 3 y 5 puntos positivamente para el equipo que actúa como local.

Teniendo ahora las clasificaciones de la NBA, en este caso de nuevo de la temporada 19/20, conocidas como NBA Standings (nba.com), obtenemos:

EQUIPOS	RÉCORD V-D TEMPORADA 19/20	RÉCORD V-D PARTIDOS COMO LOCAL 19/20	RÉCORD V-D PARTIDOS COMO VISITANTE 19/20
Milwaukee Bucks 19/20	53-12	28-3	25-9
LA Lakers 19/20	49-14	23-8	26-6
LA Clippers 19/20	44-20	25-7	19-13
Toronto Raptors 19/20	46-18	23-9	23-9
Boston Celtics 19/20	43-21	23-9	20-12
Dallas Mavericks 19/20	40-27	19-15	21-12
Houston Rockets 19/20	40-24	22-10	18-14
Miami Heat 19/20	41-24	27-5	14-19
Utah Jazz 19/20	41-23	21-10	20-13
Denver Nuggets 19/20	43-22	25-8	18-14
Oklahoma City Thunder 19/20	40-24	20-13	20-11
Philadelphia 76ers 19/20	39-26	29-2	10-24
Indiana Pacers 19/20	39-26	21-11	18-15
Brooklyn Nets 19/20	30-34	18-14	12-20
New Orleans Pelicans 19/20	28-36	13-19	15-17
Orlando Magic 19/20	30-35	16-15	14-20
Memphis Grizzlies 19/20	32-33	18-15	14-18
Phoenix Suns 19/20	26-39	13-22	13-17
Portland Trail Blazers 19/20	29-37	18-14	11-23
San Antonio Spurs 19/20	27-36	16-14	11-22

Sacramento Kings 19/20	28-36	14-17	14-19
Chicago Bulls 19/20	22-43	14-20	8-23
Detroit Pistons 19/20	20-46	11-22	9-24
Washington Wizards 19/20	24-40	16-16	8-24
Minnesota Timberwolves 19/20	19-45	8-24	11-21
New York Knicks 19/20	21-45	11-22	10-23
Charlotte Hornets 19/20	23-42	10-21	13-21
Cleveland Cavaliers 19/20	19-46	11-25	8-21
Atlanta Hawks 19/20	20-47	14-20	6-27
Golden State Warriors 19/20	15-50	8-26	7-24

Figura 6: Ejemplo de ventaja de campo (HCA) temporada 2019/20.

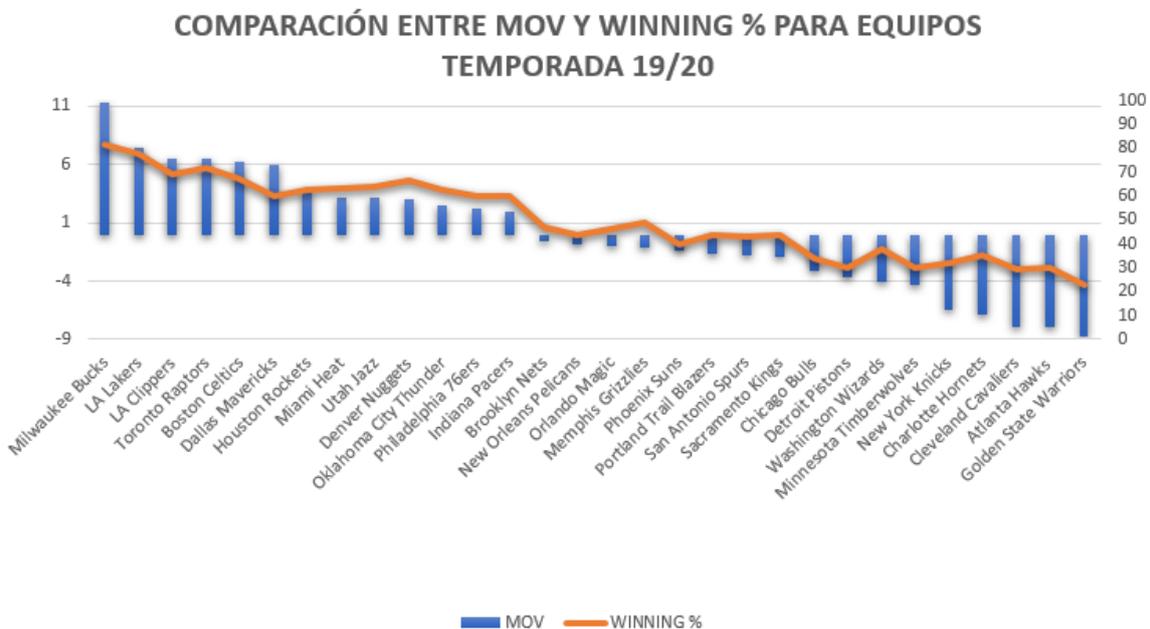


Gráfico 2: Comparación entre MOV y WINNING % para los equipos de la temporada 19/20.

A partir de los datos de la Figura 5 y comparando los resultados con la Figura 4, entendemos que esta métrica (*point differential*, MOV) es una mejor predictora de actuaciones futuras de los equipos de la liga que de sus propios récords de victorias-derrotas (nbastuffer.com, “Point differential”). Apoyándonos ahora en el Gráfico 2, los mejores equipos de la competición (aquellos con un mejor récord y que se encuentran a la izquierda en dicho gráfico), tienen los mayores valores de MOV, mientras que los peores equipos (en este caso los situados más hacia la derecha), tienen MOV pequeños, en la mayoría de los casos con valores negativos. Sin embargo, en dicho Gráfico 2 también vemos la existencia de equipos con peor récord que otros equipos, pero que su MOV es considerablemente superior con respecto al otro equipo y viceversa. Un ejemplo muy claro es la comparación entre Dallas Mavericks y Utah Jazz, donde los primeros a pesar de su valor de MOV casi doblar al de los segundos (+6.0 contra +3.2), estos presentan un Winning % ligeramente inferior al de los segundos, lo que quiere decir que el récord de Dallas Mavericks es peor que el de Utah Jazz.

2.2. PYTHAGOREAN WINS AND LOSES (PW/PL)

Esta métrica es una estimación del que “debería” ser el récord de victorias-derrotas de un equipo en una temporada, basándose únicamente en el número de puntos que un equipo anota y recibe de media (basketball-reference.com [2]).

Bill James lo inventó para el baseball y más tarde esta métrica fue adaptada al baloncesto.

Daryl Morley fue el primero que adaptó este concepto al baloncesto, usando 13.91 como valor del exponente. Este valor varía en función de distintos estadísticos que lo usen, por ejemplo John Hollinger utilizó esta fórmula dando un valor de 16.5 al exponente. El número de PW y PL va a ser muy cercano al número total de victorias-derrotas de un equipo durante la liga regular (captaincalculator.com, “Basketball. Pythagorean Win Percentage Calculator”).

Para calcular esta métrica vamos a dar al exponente un valor aproximado de 14, por lo que la fórmula quedará de la siguiente manera (basketball-reference.com, “Glosary. Pythagorean Wins and Loses”):

$$\text{EXPECTED PYTHAGOREAN WINNING \%} = \frac{\text{PUNTOS ANOTADOS}^{14}}{\text{PUNTOS ANOTADOS}^{14} + \text{PUNTOS ENCAJADOS}^{14}}$$

Vamos a ver cómo funciona esta métrica para los equipos participantes en la temporada 96/97 (basketball-reference.com, “Miscellaneous stats”):

EQUIPOS	MEDIA DE PUNTOS ANOTADOS	MEDIA DE PUNTOS RECIBIDOS	RECORD (VICTORIAS-DERROTAS)	PORCENTAJE DE VICTORIAS (W/L %)	EXPECTED PYTHAGOREAN WINNING %	PYTHAGOREAN WINS/LOSES (PW/PL)
Chicago Bulls 96/97	103.1	92.3	69-13	84.1%	82.48%	68-14
Utah Jazz 96/97	103.1	94.3	64-18	78%	77.71%	64-18
Seattle Supersonics 96/97	100.9	93.2	57-25	69.5%	75.24%	62-20
Miami Heat 96/97	94.8	89.3	61-21	74.4%	69.78%	57-25
New York Knicks 96/97	95.4	92.2	57-25	69.5%	61.72%	50-32
Houston Rockets 96/97	100.6	96.1	57-25	69.5%	65.49%	54-28
Atlanta Hawks 96/97	94.8	89.4	56-26	68.3%	69.45%	57-25
LA Lakers 96/97	100.0	95.7	56-26	68.3%	64.91%	53-29
Detroit Pistons 96/97	94.2	88.9	54-28	65.9%	69.23%	57-25
Charlotte Hornets 96/97	98.9	97.0	54-28	65.9%	56.75%	46-36
Portland Trail Blazers 96/97	99.0	94.8	49-33	59.8%	64.72%	53-29
Orlando Magic 96/97	94.1	94.5	45-37	54.9%	48.51%	40-42
Washington Bullets 96/97	99.4	97.7	44-38	53.7%	56.01%	46-36
Cleveland Cavaliers 96/97	87.5	85.6	42-40	51.2%	57.62%	47-35

Minnesota Timberwolves 96/97	96.1	97.6	40-42	48.8%	44.60%	37-45
Phoenix Suns 96/97	102.8	102.2	40-42	48.8%	52.05%	43-39
Indiana Pacers 96/97	95.4	94.4	39-43	47.6%	53.68%	44-38
LA Clippers 96/97	97.2	99.5	36-46	43.9%	41.89%	34-48
Sacramento Kings 96/97	96.4	99.8	34-48	41.5%	38.10%	31-51
Milwaukee Bucks 96/97	95.3	97.2	33-49	40.2%	43.13%	35-47
Toronto Raptors 96/97	95.5	98.6	30-52	36.6%	39.00%	32-50
Golden State Warriors 96/97	99.6	104.4	30-52	36.6%	34.10%	28-54
New Jersey Nets 96/97	97.2	101.8	26-56	31.7%	34.36%	28-54
Dallas Mavericks 96/97	90.6	97.0	24-58	29.3%	27.78%	23-59
Philadelphia 76ers 96/97	100.2	106.7	22-60	26.8%	29.32%	24-58
Denver Nuggets 96/97	97.8	104.1	21-61	25.6%	29.44%	24-58
San Antonio Spurs 96/97	90.5	98.3	20-62	24.4%	23.91%	19-63
Boston Celtics 96/97	100.6	107.9	15-67	18.3%	27.27%	22-60
Vancouver Grizzlies 96/97	89.2	99.4	14-68	17.1%	18.01%	15-67

Figura 7: Expected Pythagorean Winning% y PW/PL para los equipos de la temporada 1996/97.

Si calculamos el error promedio entre el verdadero porcentaje y la estimación de este apartado, obtenemos un error medio del 3.75%, lo que corresponde a una diferencia media entre ambas de 3.076 victorias (aproximadamente 3). Contando con que hablamos de una liga como es la NBA donde se disputan un total de 82 partidos cada liga regular por parte de cada equipo, esta estimación va a ser buena y los resultados que obtenemos a partir de ella son muy parecidos al verdadero récord del equipo.

Los equipos donde dicha estimación ha tenido peores resultados son Charlotte Hornets 96/97 (8 victorias de diferencia), New York Knicks 96/97 y Boston Celtics 96/97 (7 victorias de diferencia) y otros tales como Seattle SuperSonics 96/97, Orlando Magic 96/97, Cleveland Cavaliers 96/97 e Indiana Pacers 96/97 (cuya diferencia está en 5 victorias). Para el resto de equipos participantes en esa temporada, dicha diferencia es de 4 o menos victorias.

Esta es la métrica que se usa en la web basketball-reference.com. Sin embargo cabe destacar la existencia de otras métricas útiles para la estimación del porcentaje de victorias. Una de ellas es **Projected Winning Percentage (PW%, nbastuffer.com)**:

A partir de esta métrica va a ser posible estimar el porcentaje de victorias de un equipo teniendo en cuenta únicamente la diferencia de puntos media o point differential (MOV). La fórmula va a ser la siguiente:

$$PW\% = \frac{(\text{MOV} * 2.7) + 41}{82}$$

- En esta fórmula incluimos el valor 2.7 ya que cada punto de diferencia de media que consigue el equipo se va a trasladar en 2.7 victorias más durante la temporada.

- El 82 va a hacer referencia al número de partidos que se disputan durante la temporada regular en la NBA, mientras que 41 suele ser el punto de corte en cuanto a número de victorias se refiere de un equipo para acceder a los *PlayOffs* de la NBA.

De modo idéntico al anterior, vamos a estimar el número de victorias de los equipos participantes en la temporada 1996/97, a partir de los datos de basketball-reference.com, "Season Stats":

EQUIPOS	MOV (MARGIN OF VICTORY)	RECORD (VICTORIAS- DERROTAS)	PORCENTAJE DE VICTORIA (W/L %)	PROJECTED WINNING PERCENTAGE (PW%)	PROJECTED WINS/LOSSES (PW-PL)
Chicago Bulls 96/97	10.80	69-13	84.1%	85.56%	70-12
Utah Jazz 96/97	8.79	64-18	78%	78.94%	65-17
Seattle Supersonics 96/97	7.68	57-25	69.5%	75.29%	62-20
Miami Heat 96/97	5.49	61-21	74.4%	68.1%	56-26
New York Knicks 96/97	3.12	57-25	69.5%	60.27%	49-33
Houston Rockets 96/97	4.48	57-25	69.5%	64.75%	53-29
Atlanta Hawks 96/97	5.44	56-26	68.3%	67.91%	56-26
LA Lakers 96/97	4.27	56-26	68.3%	64.06%	53-29
Detroit Pistons 96/97	5.24	54-28	65.9%	67.25%	55-27
Charlotte Hornets 96/97	1.87	54-28	65.9%	56.16%	46-36
Portland Trail Blazers 96/97	4.17	49-33	59.8%	63.73%	52-30
Orlando Magic 96/97	-0.35	45-37	54.9%	48.85%	40-42
Washington Bullets 96/97	1.62	44-38	53.7%	55.33%	45-37
Cleveland Cavaliers 96/97	1.84	42-40	51.2%	56.06%	46-36
Minnesota Timberwolves 96/97	-1.48	40-42	48.8%	45.13%	37-45
Phoenix Suns 96/97	0.66	40-42	48.8%	52.17%	43-39
Indiana Pacers 96/97	0.98	39-43	47.6%	53.23%	44-38
LA Clippers 96/97	-2.35	36-46	43.9%	42.26%	35-47
Sacramento Kings 96/97	-3.38	34-48	41.5%	38.87%	32-50
Milwaukee Bucks 96/97	-1.89	33-49	40.2%	43.78%	36-46
Toronto Raptors 96/97	-3.12	30-52	36.6%	39.73%	33-49

Golden State Warriors 96/97	-4.71	30-52	36.6%	34.49%	28-54
New Jersey Nets 96/97	-4.56	26-56	31.7%	34.98%	29-53
Dallas Mavericks 96/97	-6.35	24-58	29.3%	29.09%	24-58
Philadelphia 76ers 96/97	-6.54	22-60	26.8%	28.46%	23-59
Denver Nuggets 96/97	-6.28	21-61	25.6%	29.32%	24-58
San Antonio Spurs 96/97	-7.88	20-62	24.4%	24.05%	20-62
Boston Celtics 96/97	-7.33	15-67	18.3%	25.86%	21-61
Vancouver Grizzlies 96/97	-10.23	14-68	17.1%	16.31%	13-69

Figura 8: Projected Winning % y Projected Wins/Losses para los equipos de la temporada 1996/97.

En este caso el error promedio entre el verdadero porcentaje y el Projected Winning %, es del 3.593%, lo que corresponde a una diferencia de 2.946 victorias, que si lo aproximamos quedará en 3. Este resultado es muy parecido al que obteníamos con la métrica anterior. Además, vemos que estas dos estimaciones cometen errores muy parecidos en los mismos equipos.

En referencia a lo expuesto anteriormente, vemos que tanto *Expected Pythagorean Wins* como el *Project Winning Percentage* son dos métricas que estiman de una forma muy parecida el porcentaje de victoria de un determinado equipo, obteniendo resultados ligeramente mejores en la segunda.

2.3. EXPECTED WINNING PERCENTAGE

A partir de esta métrica nos va a ser posible obtener la estimación del porcentaje de conseguir la victoria cuando un equipo A se enfrenta a un equipo B. La fórmula para dicha estimación es la siguiente (nbastuffer.com):

$$WINNING PERCENTAGE A = \frac{PW\%(A) * (1 - PW\%(B))}{PW\%(A) * (1 - PW\%(B)) + PW\%(B) * (1 - PW\%(A))}$$

De la fórmula anterior cabe destacar que el *PW%* hace referencia *al Expected Pythagorean Winning Percentage* del equipo al que se refiere en cada caso, pero también se podría hacer con el *Projected Winning Percentage* ya que daría resultados parecidos. Debido a que en páginas como basketball-reference.com utiliza de manera más habitual la primera, es la que vamos a utilizar nosotros.

Seguidamente vamos a mostrar un ejemplo de la métrica *Expected Winning Percentage* gracias a datos que vamos a obtener de la página basketball-reference.com. En consecuencia vamos a tomar como referencia a los Portland Trail Blazers y los Denver Nuggets de la temporada 2018/19, donde para empezar, vamos a calcular el *Expected Pythagorean Winning %* para ambos equipos:

EQUIPOS	MEDIA PUNTOS ANOTADOS	MEDIA PUNTOS RECIBIDOS	EXPECTED PYTHAGOREAN WINNING %
---------	-----------------------	------------------------	--------------------------------

Portland Trail Blazers 18/19	114.7	110.5	62.77%
Denver Nuggets 18/19	113	108.9	62.65%

Figura 9: Comparación del Expected Pythagorean Winning entre Blazers y Nuggets 2018/19.

Viendo la *Figura 9* nos damos cuenta de que el Expected Pythagorean Winning % entre ambos equipos es prácticamente igual. Aún así vamos a calcular el *Expected Winning %* de ambos equipos si se enfrentaran contra el otro:

- EXPECTED WINNING % (Portland Vs Denver) = 50.128%
- EXPECTED WINNING % (Denver Vs Portland) = 49.81%

Debido a los resultados obtenidos, lo normal sería un resultado muy igualado entre ambos equipos, ya que el porcentaje de victoria esperado en enfrentamientos entre ellos está muy igualado (probabilidad esperada de que Portland gane a Denver 0.32% superior a que ocurra lo contrario).

Durante la liga regular de esa temporada 18/19, ambos equipos se enfrentaron en un total de cuatro ocasiones. Cada uno de los dos equipos disputaron dos partidos de ellos en su casa, es decir, como locales y otros dos como visitantes. Los resultados de esos partidos fueron:

En los partidos en Portland:

- 113-112 a favor de Denver Nuggets (30/11/2018).
- 108-115 a favor de Portland Trail Blazers (07/04/2019).

En los partidos en Denver:

- 113-116 a favor de Denver Nuggets (13/01/2019).
- 110-119 a favor de Denver Nuggets (05/04/2019).

Como se suponía, los enfrentamientos entre ambos equipos se resolvieron de forma muy igualada, donde Denver Nuggets consiguió la victoria en ambos partidos que disputó como local y además consiguió sacar una victoria en Portland.

2.4. SIMPLE RATING SYSTEM (SRS)

Continuando con el tema del diferencial de puntos o margen de victoria (*MOV*), Justin Kubatko realiza una propuesta (basketball-reference.com, 2008) donde añade una nueva métrica denominada *Simple Rating System (SRS)*, cuya idea sale de Doug Grien (pro-football-reference.com).

En relación al *SRS*, esta nueva métrica va a tener en cuenta tanto el margen de victoria media de un equipo (*MOV*) como el ***Strength of Schedule (SOS)***, que en castellano hace referencia a la dificultad en el calendario que tiene que afrontar un determinado equipo.

Con respecto al *SOS* de un equipo, además de tener en cuenta los partidos que han jugado sus oponentes hasta ese momento, también habrá que tener en cuenta la localización del partido, es decir, habría que tener en consideración el factor de la “ventaja de campo” (nbastuffer.com, “Analytics-101, Strength of Schedule”).

Una forma simple de calcular el *SOS* (basketball-reference.com) es calcular la media del diferencial de la eficiencia de los partidos que han jugado sus oponentes, ajustándolo también

a la localización del partido (ventaja de campo). Esta métrica nos indicará cuanto de mejores y peores han sido los rivales de un equipo respecto a la media (considerando a esta el valor 0). Por lo que si un equipo tiene un valor *SOS* de 0.08, significa que sus rivales fueron, de media, 0.08 puntos mejores que la media.

Volviendo al *Simple Rating System (SRS)*, vamos a explicarlo mediante un ejemplo con datos extraídos de basketball-reference.com. Para ello vamos a tomar los datos de la temporada 19/20 antes de ser pausada por la pandemia del coronavirus, donde cogemos a los 6 equipos que presentan un *SRS* mayor:

EQUIPOS	MARGEN DE VICTORIA (MOV)	DIFICULTAD DEL CALENDARIO (SOS)	SIMPLE RATING SYSTEM (SRS)
Milwaukee Bucks 19/20	11.29	-0.85	10.44
LA Lakers 19/20	7.41	0.34	7.75
LA Clippers 19/20	6.52	0.22	6.74
Toronto Raptors 19/20	6.45	-0.57	5.88
Dallas Mavericks 19/20	6.04	-0.20	5.84
Boston Celtics 19/20	6.17	-0.48	5.69

Figura 10: *SOS* y *SRS* mejores equipos temporada 2019/20.

Respecto a los diferentes valores que tenemos, comprobamos que los Bucks además de ser el equipo con una mayor diferencia media de puntos, tenía un calendario más asequible, ya que el valor de *SOS* que les corresponde es de -0.85, lo que indica que sus rivales, de media, han sido 0.85 puntos peores que la media. Así que este equipo será el que además presente un *SRS* mayor con un valor de 10.44, lo que nos dice que Milwaukee ha sido de media 10.44 puntos mejor que la media del resto de equipos.

Por ejemplo, si a continuación comparamos el *SRS* de Bucks y Lakers (10.44 contra 7.75), deduciremos que los Bucks han sido 2.69 puntos mejores que los Lakers en lo que va de temporada. Si además, ahora tenemos en cuenta el “factor cancha” que cuenta 3 puntos positivos al equipo local, esta métrica favorecería a los Bucks en 5.69 puntos cuando juegan en Milwaukee contra Lakers, mientras que a su vez favorecería a estos últimos en 0.31 puntos si juegan en Los Angeles contra este mismo equipo.

Podrían realizarse numerosos rankings y listas a partir de esta métrica, pero tomamos el siguiente ejemplo (Justin Kubatko, [basketball-reference](http://basketball-reference.com), 2008), en el que se toma el equipo con mejor *SRS* cada temporada junto con el campeón esa misma temporada, cogiendo como punto de partida la temporada 1980/81. Además, como este ejemplo fue ilustrado en el año 2008, vamos a alargarlo hasta la temporada 2018/19 (la cual fue la última que se disputó en su totalidad):

TEMPORADA	EQUIPO CON MAYOR SIMPLE RATING SYSTEM (SRS) ESA TEMPORADA	SRS	LOGROS CONSEGUIDOS POR EL EQUIPO
1980/81	Philadelphia 76ers	7.76	Finales de Conferencia Este
1981/82	Boston Celtics	6.35	Finales de Conferencia Este
1982/83	Philadelphia 76ers	7.53	Campeones
1983/84	Boston Celtics	6.42	Campeones
1984/85	Milwaukee Bucks	6.70	Semifinales Conferencia Este
1985/86	Boston Celtics	9.06	Campeones
1986/87	LA Lakers	8.32	Campeones

1987/88	Boston Celtics	6.15	Finales de Conferencia Este
1988/89	Cleveland Cavaliers	7.95	Primera Ronda de Playoffs
1989/90	Phoenix Suns	7.09	Finales de Conferencia Oeste
1990/91	Chicago Bulls	8.57	Campeones
1991/92	Chicago Bulls	10.07	Campeones
1992/93	Seattle Supersonics	6.66	Finales de Conferencia Oeste
1993/94	Seattle Supersonics	8.68	Primera Ronda de Playoffs
1994/95	Seattle Supersonics	7.91	Primera Ronda de Playoffs
1995/96	Chicago Bulls	11.80	Campeones
1996/97	Chicago Bulls	10.70	Campeones
1997/98	Chicago Bulls	7.24	Campeones
1998/99	San Antonio Spurs	7.12	Campeones
1999/00	LA Lakers	8.41	Campeones
2000/01	San Antonio Spurs	7.92	Finales de Conferencia Oeste
2001/02	Sacramento Kings	7.61	Finales de Conferencia Oeste
2002/03	Dallas Mavericks	7.91	Finales de Conferencia Oeste
2003/04	San Antonio Spurs	7.51	Semifinales Conferencia Oeste
2004/05	San Antonio Spurs	7.84	Campeones
2005/06	San Antonio Spurs	6.69	Semifinales Conferencia Oeste
2006/07	San Antonio Spurs	8.35	Campeones
2007/08	Boston Celtics	9.30	Campeones
2008/09	Cleveland Cavaliers	8.68	Finales de Conferencia Este
2009/10	Orlando Magic	7.12	Finales de Conferencia Este
2010/11	Miami Heat	6.76	Finales de la NBA
2011/12	Chicago Bulls	7.43	Primera Ronda de Playoffs
2012/13	Oklahoma City Thunder	9.15	Semifinales de Conferencia Oeste
2013/14	San Antonio Spurs	8.00	Campeones
2014/15	Golden State Warriors	10.01	Campeones
2015/16	Golden State Warriors	10.38	Finalistas de la NBA
2016/17	Golden State Warriors	11.35	Campeones
2017/18	Houston Rockets	8.21	Finales de Conferencia Oeste
2018/19	Milwaukee Bucks	8.04	Finales de Conferencia Este

Figura 11: Equipos con mayor SRS cada temporada de la NBA.

En los últimos 39 años, el líder durante la liga regular respecto al valor de *SRS*, ha sido campeón de la NBA en 17 ocasiones, lo que corresponde un 43.6% de las veces.

Considerando nuevamente este período de tiempo de la NBA, vemos que los equipos que consiguieron un mayor *SRS* fueron los Chicago Bulls de la temporada 1995/96, con un valor de 11.80 y los Warriors de la temporada 2016/17, con 11.35.

De los equipos que presentaron mayor *SRS* en sus respectivas temporadas y que no consiguieron alzarse con el título de campeones, la mayoría de ellos generalmente consiguieron llegar lejos en los *Play-Offs* y todos ellos consiguieron clasificarse para dicha postemporada.

Ahora tomamos como referencia la temporada 2006/07 y los equipos que consiguieron acceder a las finales de la NBA, San Antonio Spurs y Cleveland Cavaliers:

EQUIPOS	MARGEN DE VICTORIA (MOV)	DIFICULTAD DEL CALENDARIO (SOS)	SIMPLE RATING SYSTEM (SRS)
San Antonio Spurs 06/07	8.43	-0.08	8.35
Cleveland Cavaliers 06/07	3.83	-0.5	3.33

Figura 12: Comparación MOV, SOS y SRS entre Spurs y Cavaliers 2006/07.

Comprobando la *Figura 12*, vemos que los San Antonio Spurs consiguieron un *SRS* de 8.35, lo que indica que estos fueron 8.35 puntos mejores que la media, mientras que dicho valor para Cleveland Cavaliers fue 3.33. En resumen, durante la liga regular la métrica *SRS* nos muestra que los San Antonio Spurs fueron 5.02 (8.35 – 3.33) puntos mejores que los Cavaliers.

2.5. NET RATING

Esta métrica se corresponde con la diferencia entre el Offensive Rating (*OFFRTG*) y Defensive Rating (*DEFRTG*).

Para entender bien su funcionamiento, vamos a ahondar primero en la explicación de un factor esencial para su comprensión. Se trata de las posesiones (*Possessions, poss, nba.com*). Esta última resulta esencial para a posteriori entender otras analíticas básicas tales como el ritmo de juego (*pace*) o las ya nombradas anteriormente *Offensive Rating* y *Defensive Rating* y por consiguiente el *Net Rating*.

2.5.1. POSSESSIONS (poss)

O posesiones en castellano, es denominado de dicha manera ya que engloba el tiempo completo que un equipo está en posesión del balón. Según *fansided.com*, “Nylon-Calculus-101-possessions”[8] no debe confundirse el término jugadas con el término posesiones. Las jugadas van a estar separadas entre ellas por cualquier tiro de campo realizado, incluyendo los tiros libres, además de las pérdidas de balón. Si se captura un rebote ofensivo, se extenderá una posesión, pero no ocurre lo mismo con las jugadas, ya que si es capturado el rebote en ataque, se iniciaría una nueva jugada.

Para considerar que una posesión ha llegado a su final, consideraremos que esto ocurre siempre que se anote un lanzamiento, un tiro libre, con un rebote y con un pérdida. Un rebote defensivo siempre va a dar por finalizada una posesión, pero los rebotes ofensivos tienen que ser correctamente considerados, ya que estos alargan una posesión.

Siguiendo el modelo de *nba.com*, se define como el número total de posesiones llevadas a cabo por un equipo de media en un partido, y se calcula de la siguiente manera:

$$POSESIONES = FGA + TOVs + 0.44 * FTA - (OFFENSIVE REBOUNDS)$$

- *FGA* hace referencia al número total de tiros de campo realizados por un equipo.
- *TOVs* indica el total de pérdidas que ha cometido un equipo.
- *FTA* muestra el total de tiros libres llevados a cabo por un equipo. Este valor es multiplicado por 0.44 ya que se ha calculado que alrededor del 44% de los tiros libres van a comenzar una nueva posesión (*nbastuffer.com*).

Esta es una métrica básica para entender el resto de estadísticas relacionadas con el baloncesto. Hay que tener en cuenta que al final de una temporada, los equipos más rápidos (aquellos cuyas posesiones duran menos y por ende realizan un mayor número de ellas por partido) tendrán un total de posesiones superior a los equipos que son mas lentos; sin embargo, independientemente de lo anterior, ambos equipos siempre van a tener el mismo número total de posesiones realizadas que el acumulado de posesiones de sus rivales.

Una vez entendido a lo que hace referencia el término posesiones, gracias a ello vamos a pasar a explicar dos nuevas métricas: *Offensive Rating* y *Defensive Rating*.

2.5.2. OFFENSIVE RATING

Representado como *OFFRTG* en *nba.com* y *ORTg* en *basketball-reference.com*.

Mide el número de puntos de media que anota un equipo por cada 100 posesiones (*nba.com*, “glossary” y *ESPN*).

$$OFFRTG = 100 * \left(\frac{PUNTOS ANOTADOS}{POSESIONES} \right)$$

Métrica más exacta para comparar el ataque de distintos equipos que si nos fijamos únicamente en los puntos por partido que anotan, ya que ajusta al mismo número de posesiones, tomando 100 habitualmente, debido a que se ha calculado que de media un equipo gasta 100 posesiones durante un partido (*bleacherreport*, “Offensive and Defensive Rating”).

Si queremos comprobar cuáles han sido los mejores ataques a lo largo de la historia de la NBA, no nos vale únicamente con comparar los *OFFRTG* de todos lo equipos, si no que tenemos que ajustarlo al *OFFRTG* medio del resto de los equipos en sus distintas temporadas (*OFFRTG+*), ya que con el paso de los años las tendencias de juego han ido cambiando (*bleacherreport.com*, “Ranking the NBAs best 20 offenses of all time” [9]), estudio recogido hasta al año 2014, pero lo alargamos hasta la temporada pasada, la 2018/19:

EQUIPOS	OFFRTG	OFFRTG MEDIO DE LA LIGA	OFFRTG+	RÉCORD	LOGROS
Dallas Mavericks 2003/04	112.1	102.9	108.94	52-30	1ª ronda de Playoffs
Phoenix Suns 2004/05	114.5	106.1	107.2	62-20	Finales Conferencia
Dallas Mavericks 2001/02	112.2	104.5	107.37	57-25	Semifinales de Conferencia
Utah Jazz 1997/98	112.7	105.0	107.33	62-20	Finales de la NBA
Chicago Bulls 1996/97	114.4	106.7	107.22	69-13	Campeones de la NBA
Sacramento Kings 2003/04	110.3	102.9	107.19	55-27	Semifinales de Conferencia
Phoenix Suns 2009/10	115.3	107.6	107.16	54-28	Finales de Conferencia
Chicago Bulls 1995/96	115.2	107.6	107.06	72-10	Campeones de la NBA
Phoenix Suns 2006/07	113.9	106.5	106.95	61-21	Semifinales de Conferencia
Denver Nuggets 1981/82	114.3	106.9	106.92	46-36	1ª ronda de Playoffs

Milwaukee Bucks 1970/71	103.9	97.2	106.89	66-16	Campeones de la NBA
Dallas Mavericks 2002/03	110.7	103.6	106.85	60-22	Finales de Conferencia
Boston Celtics 1987/88	115.4	108.0	106.85	57-25	Finales de Conferencia
Chicago Bulls 1991/92	115.5	108.2	106.75	67-15	Campeones de la NBA
Los Angeles Lakers 1986/87	115.6	108.3	106.74	65-15	Campeones de la NBA
Los Angeles Lakers 1997/98	111.9	105.0	106.57	61-21	Finales de Conferencia
Utah Jazz 1996/97	113.6	106.7	106.47	64-18	Finalistas de la NBA
LA Clippers 2014/15	112.4	105.6	106.39	56-26	Semifinales de Conferencia
Indiana Pacers 1998/99	108.7	102.2	106.36	33-17	Finales de Conferencia
Seattle Supersonics 1997/98	111.6	105.0	106.29	61-21	Semifinales de Conferencia

Figura 13: Mejores OFFRTG+ de la historia de la NBA.

*OFFRTG+ : ajustamos el Offensive Rating a la temporada en la que se consiguió. Es decir, se divide el OFFRTG de ese equipo esa temporada entre el OFFRTG medio del resto de equipos y se multiplica por 100.

A lo largo de la historia encontramos OFFRTG más elevados que los que aparecen en esta tabla, pero como dijimos anteriormente, hay que tener en cuenta la tendencia de juego en cada temporada. Por ejemplo, los Dallas Mavericks de la temporada 2003/04 que son considerados como el mejor ataque de la historia, su valor de OFFRTG no es de los más elevados, pero su OFFRTG+ es mayor que el resto de equipos de la historia de la NBA. Esto es debido a la gran potencia defensiva que existió ese año en la NBA, es decir, el valor del ratio defensivo esa temporada fue más bajo que en otras. Al OFFRTG+ lo consideramos la mejor métrica inter-era para comparar ataques (*bleacherreport.com* y *basketball-reference*).

De los 20 mejores ataques de la historia de la NBA, únicamente 6 de ellos resultaron campeones de la NBA, lo que supone un 30%. Vemos que algunos de ellos incluso quedaron eliminados en primera ronda de *Play-Offs*, siendo Denver Nuggets de la temporada 81/82 uno de los casos más interesantes, ya que a pesar de ser top 10 en equipos atacantes, su récord esa temporada fue 46-36, lo que es de media el récord de un equipo de mitad de tabla de la NBA.

2.5.3. DEFENSIVE RATING

Representado como DEFRTG en *nba.com* y DRTg en *basketball-reference.com*.

Al contrario que el anterior, esta métrica mide el número de puntos que recibe un equipo por cada 100 posesiones (*NBA.com* y *ESPN*).

$$DEFRTG = 100 * \left(\frac{PUNTOS\ RECIBIDOS}{POSESIONES} \right)$$

A continuación, vamos a realizar el mismo estudio que antes, pero en este caso, se mostrarán las 20 mejores defensas de la NBA. Como en el caso anterior, no podemos comparar distintos DEFRTG de distintas épocas. Si no que tendremos que calcular el DEFRTG+ para cada uno ajustando a los DEFRTG medios de los equipos para cada una de sus temporadas (*bleacherreport.com*, "Ranking the NBAs best defenses of all time" [10]), en este ejemplo. Al

igual que ocurría en el *OFFRTG*, el estudio también está realizado hasta el año 2014, por lo tanto se va a alargar hasta la temporada pasada, la 2018/19:

EQUIPOS	DEFRTG	DEFRTG MEDIO DE LA LIGA	DEFRTG+	RÉCORD	LOGROS
Boston Celtics 1963/64	83.8	94.6	112.89	59-21	Campeones de la NBA
Boston Celtics 1964/65	84.2	93.6	111.16	62-18	Campeones de la NBA
Boston Celtics 1961/62	85.1	93.6	109.99	60-20	Campeones de la NBA
Boston Celtics 1962/63	87.4	95.9	109.73	58-22	Campeones de la NBA
Minneapolis Lakers 1951/52	79.3	86.9	109.58	40-26	Campeones de la NBA
San Antonio Spurs 2003/04	94.1	102.9	109.35	57-25	Semifinales de Conferencia
Boston Celtics 1960/61	84.5	92.1	108.99	57-22	Campeones de la NBA
Boston Celtics 2007/08	98.9	107.5	108.7	66-16	Campeones de la NBA
New York Knicks 1992/93	99.7	108.0	108.32	60-22	Finales de Conferencia
New York Knicks 1993/94	98.2	106.3	108.25	57-25	Semifinales de Conferencia
Detroit Pistons 2003/04	95.4	102.9	107.86	54-28	Campeones de la NBA
Golden State Warriors 2014/15	101.4	109.1	107.59	67-15	Campeones de la NBA.
San Antonio Spurs 1998/99	95.0	102.2	107.58	37-13	Campeones de la NBA
San Antonio Spurs 2015/16	99.0	106.4	107.59	67-15	Semifinales de Conferencia
Boston Celtics 1965/66	88.3	94.9	107.46	54-26	Campeones de la NBA
Indiana Pacers 2013/14	99.3	106.7	107.45	56-26	Finales de Conferencia
San Antonio Spurs 2004/05	98.8	106.1	107.39	59-23	Campeones de la NBA
Boston Celtics 1959/60	84.9	91.1	107.3	59-16	Campeones de la NBA
Boston Celtics 1968/69	89.1	95.5	107.18	48-34	Campeones de la NBA
New York Knicks 1969/70	92.4	99.0	107.14	60-22	Campeones de la NBA

Figura 14: Mejores DEFRTG+ de la historia de la NBA.

*DEFRTG+: para calcular esta métrica vamos a tomar el DEFRTG medio de la temporada y lo dividiremos entre el DEFRTG del equipo en cuestión para más tarde multiplicar el resultado por 100 y así obtenemos el DEFRTG+. Para entenderlo, si un equipo permite que le anoten 10 puntos menos que la media cada 100 posesiones entonces tendrá un DEFRTG+ de 110. Cuando dicho valor de un equipo es 100, significa que la defensa está perfectamente ajustada a la media (*bleacherreport.com*).

Observando la *Figura 14*, vemos que en 8 ocasiones aparecen los Boston Celtics de la década de los 60, además de ocupar las cuatro primeras posiciones. Esto es debido a cómo de dominante Bill Russell fue en la época (*bleacherreport.com*).

Russell disputó un total de 13 temporadas en la NBA, ganando 11 anillos, y cuya peor temporada en cuanto a *DEFRTG+* de su equipo fue en la 1967/68, donde dicho valor ocupa el puesto No. 87 de mejores *DEFRTG+* de la historia de la NBA (*basketball-reference.com* y *bleacherreport.com*).

En este caso, de las consideradas mejores defensas de la historia de la NBA en cuanto al *DEFRTG+*, vemos que 15 equipos de ellos resultaron campeones de la competición en sus respectivas temporadas, lo que supone un alto 75%.

Como ya se explicó anteriormente, la temporada 2003/04 se caracterizó por ser una temporada eminentemente defensiva, cuyo *DEFRTG* medio de todos los equipos fue 102.9, lo que supone el valor más bajo desde 1979. A partir de esta temporada se llevaron a cabo diferentes cambios en las normas de juego, para favorecer a los ataques, como por ejemplo la prohibición del contacto con las manos del defensor al atacante. A pesar de todo esto, en el ranking anterior aparecen dos equipos de esta temporada, lo que supone un mérito mayor. Y es que los San Antonio Spurs y Detroit Pistons consiguieron un *DEFRTG* de 94.1 y 95.4 respectivamente. Incluso el valor de 94.1 que consiguieron los San Antonio Spurs esa temporada se corresponde con el valor más bajo que ha logrado un equipo de la liga desde que se llevo a cabo la fusión (en 1976 se produjo la unión de las ligas ABA y NBA de baloncesto norteamericanas), superando a otros tales como los 2007/08 Boston Celtics, Detroit Pistons de su misma temporada o los New York Knicks de los años 90 (*bleacherreport.com*, “Best NBAs Defenses” [10]).

2.5.4. NET RATING

Representado como *NETRTG* en *nba.com* y *NRtg* en *basketball-reference.com*.

Diferencia numérica entre el *Offensive Rating* y el *Defensive Rating*.

$$NET\ RATING = OFFRTG - DEFRTG$$

Esta métrica nos va a permitir comparar a unos equipos con otros englobando tanto el ataque de cada uno de ellos como su defensa. También mostrará cómo un equipo lo ha ido haciendo a lo largo de una temporada y nos servirá como un buen predictor para sucesos futuros (*pivotanalysis.com*, “Net-Ranking” [11] y *nbastuffer.com* [5]).

A partir de datos obtenidos de *basketball-reference.com* vamos a mostrar los 10 mejores *NETRTG* conseguidos por distintos equipos, tomando como punto de partida la temporada 1988/89 hasta la actual, 2019/20:

EQUIPOS	NETRTG (OFFRTG – DEFRTG)	RÉCORD	LOGROS
Chicago Bulls 1995/96	+13.4 (115.2 – 101.8)	72-10	Campeones de la NBA
Chicago Bulls 1996/97	+12.0 (114.4 – 102.4)	69-13	Campeones de la NBA
Golden State Warriors 2016/2017	+11.6 (115.6 – 104.0)	67-15	Campeones de la NBA
Boston Celtics 2007/08	+11.3 (110.2 – 98.2)	66-16	Campeones de la NBA
San Antonio Spurs 2015/16	+11.3 (110.3 – 99.0)	67-15	Semifinales de Conferencia Oeste
Chicago Bulls 1991/92	+11.0 (115.5 – 104.5)	67-15	Campeones de la NBA
Golden State Warriors 2015/16	+10.7 (114.5 – 103.8)	73-9	Finalistas de la NBA
Golden State Warriors 2014/15	+10.2 (111.6 – 101.4)	67-15	Campeones de la NBA
Cleveland Cavaliers 2008/09	+10.0 (112.4 – 102.4)	66-16	Finals de Conferencia
Oklahoma City Thunder 2012/13	+9.8 (112.4 – 102.6)	60-22	Semifinales de Conferencia

Figura 15: Mejores *NETRTG* de la historia de la NBA.

De los diez equipos con un mayor *NETRTG* desde la temporada 1988/89, seis de ellos consiguieron alzarse con el título, lo que supone un elevado 60%.

2.6. PACE

Métrica que hace referencia a una estimación del número de posesiones de un equipo por 48 minutos. Sencillamente, mide la velocidad a la que un equipo juega. Este quizás sea el evaluador que mejor va a poner en contexto todas las estadísticas que podemos encontrar en el *boxscore* (*nbamaniacs.com*, “*Nba, un nuevo enfoque analítico*” [12]).

En el contexto histórico, se ha podido observar y comprobar que por normal general los grandes equipos a lo largo de la historia de la liga cuyo ritmo o *pace* sea menor, suelen clasificarse como eminentemente defensivos, anotando una menor cantidad de puntos de media por partido que el resto de grandes equipos y a su vez recibiendo también una menor cantidad de puntos de media (*sportingcharts.com*, “*NBA dictionary, pace-factor*”). Esto lo vamos a ver más tarde con algún ejemplo.

La fórmula que se usa para calcular ritmo medio (*pace*) de un equipo es la siguiente (*nba.com*, *basketball-reference.com*, *ESPN*, *sportingcharts.com*):

$$PACE = 48 \times \frac{TEAM POSSESSIONS + OPPONENT POSSESSIONS}{2 * \frac{TEAM MINUTES PLAYED}{5}}$$

* TEAM MINUTES PLAYED, hace referencia al total de minutos jugados por cada uno de los jugadores del equipo.

En el siguiente gráfico vamos a mostrar como ha ido cambiando el ritmo de juego a lo largo de la historia de la NBA, tomando como punto de partida la temporada 1973/74. Para ello vamos a tomar el ritmo medio de todos los equipos para cada temporada (*basketball-reference.com*, “*Play-index*”):

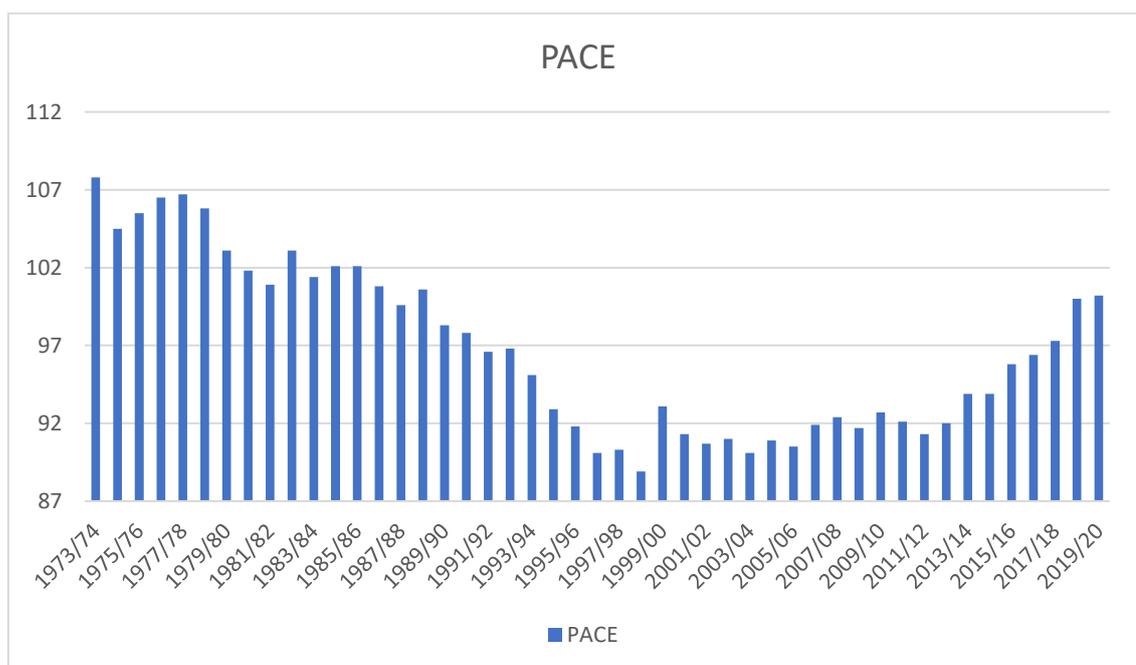


Gráfico 3: *PACE* medio de todos los equipos cada una de las temporadas de la NBA, desde 1980/81 a 2019/20.

*Tomamos como inicio para este gráfico la temporada 1973/74 ya que fue la primera en la que se empezó a contabilizar el ritmo de juego (pace). Esto es debido a que, para calcular el ritmo de juego medio de un equipo, necesitamos diferenciar entre rebotes ofensivos y defensivos y antes de dicha temporada no se contaban de forma separada (*lineups.com*)

Como se puede ver, el ritmo de juego se encuentra creciendo progresivamente desde 2012 aproximadamente, aunque vemos que entre las temporadas 2017/18 y 2018/19 se produjo un aumento considerablemente grande, y es que entre ambos años los equipos llevan a cabo de media 2.7 posesiones más (de 97.3 posesiones medias por partido se pasa a 100 en la siguiente temporada).

De igual modo, vemos que entre la temporada 1998/99 y 1999/00 se produce el aumento más grande en cuanto a ritmo de juego se refiere. Esto es debido al ritmo de juego tan bajo que hubo en la temporada 1998/99 (88.9, el más pequeño de la historia de la NBA desde que se contabiliza esta métrica), y que fue debido al lockout de esa temporada, el cual provocó que el calendario estuviera más condensado y lo equipos disminuyeran su ritmo de juego (*lineups.com*, “*Why NBA game pace is at historic high*” [13]).

En el siguiente gráfico mostraremos datos (*basketball-reference.com*) que nos ilustrarán el número de equipos de cada temporada cuyo valor de ritmo de juego o pace fue superior a 100:

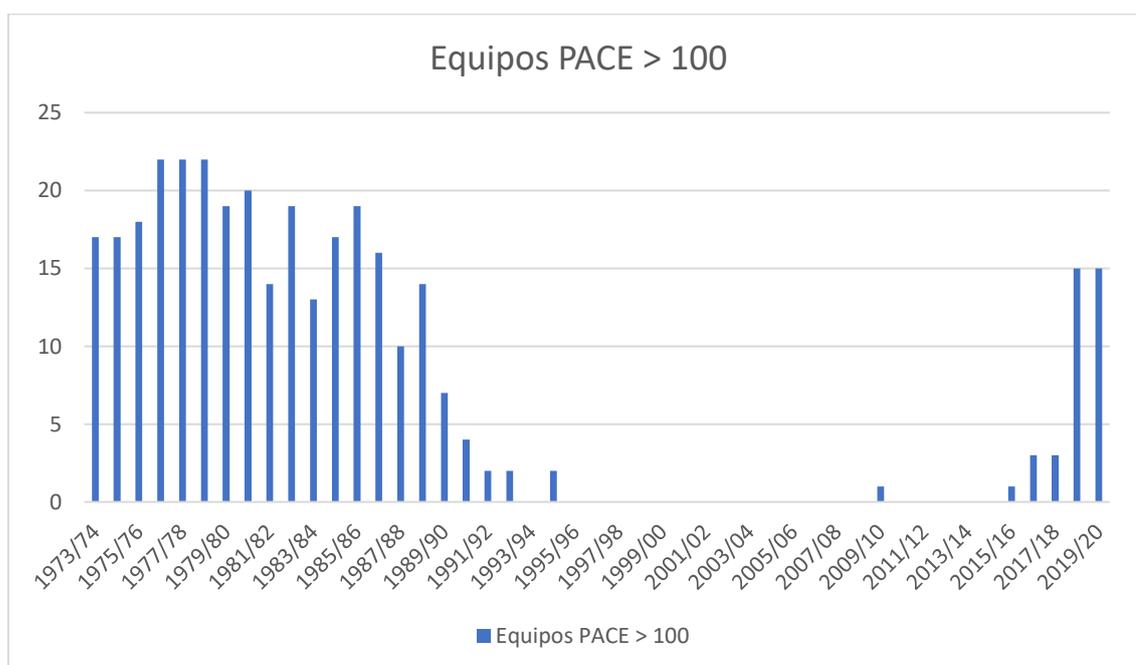


Gráfico 4: Número de equipos por temporada con PACE > 100.

Darse cuenta que solamente en la temporada 2018/19 hay más equipos (15) con ritmo de juego superior a 100 que el total de equipos con esta característica combinados desde la temporada 1991/92, lo que nos muestra una clara tendencia de que el valor de esta métrica se encuentra en crecimiento en los últimos años.

Durante las dos primeras décadas que aparecen en el *Gráfico 4*, vemos que la tendencia al ritmo de juego es elevada. Aquí encontramos equipos que destacan principalmente por su ataque, tales como los Boston Celtics 1975/76 y LA Lakers 1981/82, cuyo ritmo de juego era 106.9 y 103.1, lo que supone valores muy elevados, además de anotar 106.9 y 103.1 puntos por partido respectivamente.

En las siguientes dos décadas vemos que la tendencia de juego de la liga lleva a un ritmo de juego muy bajo, donde encontramos campeones tales como los Detroit Pistons 2003/04, con un ritmo de juego medio de 87.9, anotando 90.1 y recibiendo únicamente 84.3 puntos por partido; o los Chicago Bulls de los 90s, que consiguieron sus 6 títulos con un ritmo medio de 92.1.

En resumen, vemos que cuando los índices del ritmo de juego son altos, va a primar el ataque, mientras que cuando prima un baloncesto defensivo, es decir, cuando se lleva a cabo un baloncesto más metódico y organizado, con posesiones más largas y dando una gran importancia a la defensa, los índices van a ser pequeños.

Observando el *Gráfico 3* y *Gráfico 4* vemos la clara tendencia al aumento del ritmo de juego medio en las últimas temporadas, donde encontramos por ejemplo a los Golden State Warriors 2015/16, cuyo ritmo de juego fue 99.6, anotando 113.2 puntos por partido.

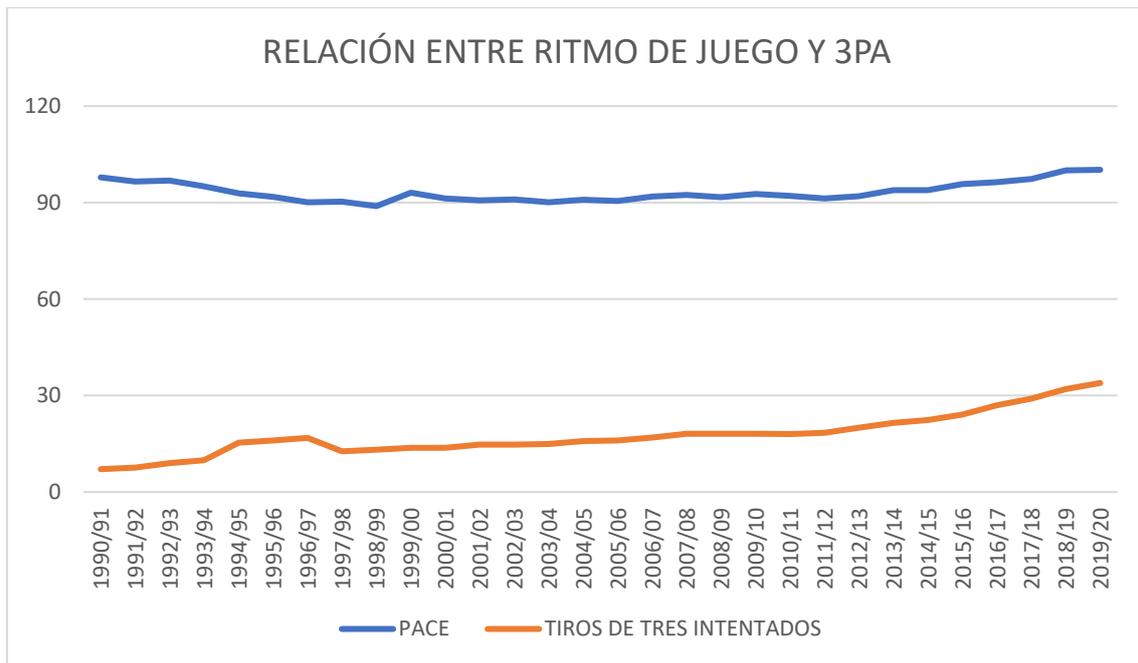


Gráfico 5: Relación entre *PACE* y media de lanzamientos de 3 desde la temporada 1990/91.

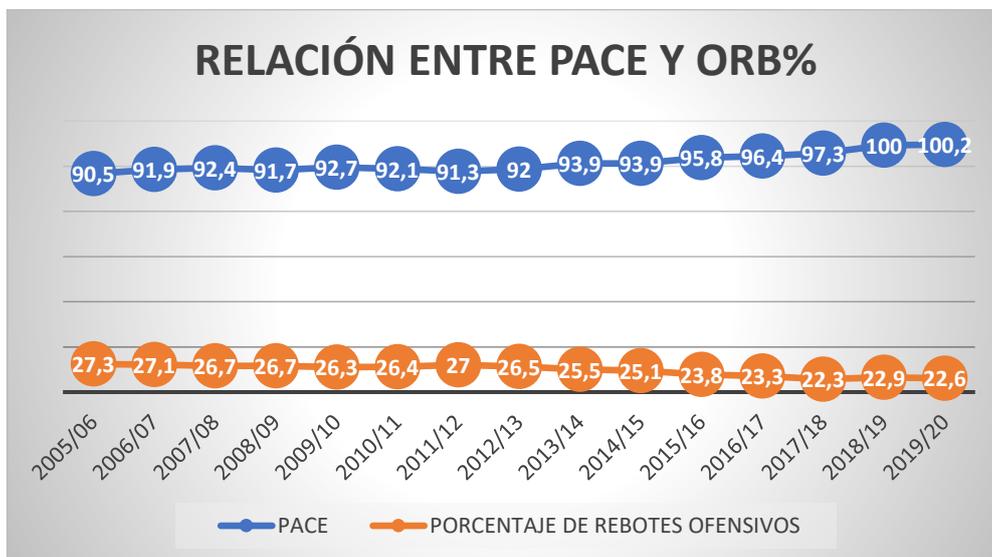


Gráfico 6: Relación entre ritmo de juego y ORB% desde la temporada 2005/06:

*ORB%: esta métrica se corresponde con el porcentaje estimado de la cantidad de rebotes ofensivos capturados sobre el total disponible de un jugador (basketball-reference.com y nba.com).

Distintos trabajos en el análisis de estadísticas NBA ponen de manifiesto la relación existente entre este índice y otros tales como el 3PA y el ORB%. Esta relación la podemos observar en los Gráficos 5 y 6.

En estos ejemplos, vemos que el aumento en el ritmo de juego de estos últimos años está muy relacionado positivamente con el 3PA, ya que a mayores lanzamientos de tres de media por partido, mayor será el ritmo de juego. Con el ORB% pasa el caso contrario y es que el aumento en el ritmo de juego de los últimos años coincide con la disminución del ORB%, es decir, estos últimos años los equipos tienden a cargar menos el rebote ofensivo.

Un ejemplo claro de que el ritmo de juego o PACE de un equipo no va a ser un factor determinante a la hora de medir el éxito de estos mismo, se puede ver comparando a los Denver Nuggets de la temporada 1990/91 y los San Antonio Spurs 2004/05:

EQUIPOS	OFFRTG	DEFRTG	NETRTG	PACE	PTS/ PARTIDO	OPP. PTS/ PARTIDO	eFG%	RÉCORD V-D
Denver Nuggets 1990/91	105.2	114.7	-9.5	113.7	119.9	130.8	45.7%	20-62
San Antonio Spurs 2004/05	107.5	98.8	+8.7	88.9	96.2	88.4	49.2%	59-23

Figura 16: Comparación PACE de Denver Nuggets 1990/91 contra San Antonio Spurs 2004/05.

En este caso se presenta uno de los ejemplos más extremos que se pueden ver en cuanto al ritmo de juego o PACE, ya que tenemos a los Denver Nuggets 1990/91, que es uno de los equipos más rápidos de la historia de la NBA y los San Antonio Spurs 2004/05, que en contraposición, tiene uno de los valores en el PACE más pequeños. Pero un mayor número de posesiones por partido de un determinado equipo no significa la mejoría de ese equipo respecto a otro con un valor pequeño, si no que lo que realmente importa es cuánto de eficiente es ese equipo en las posesiones que lleva a cabo.

Observando la Figura 16, vemos que un equipo que es eficiente ofensivamente es aquel que consigue anotar más puntos a partir de las posesiones que realiza, no tener en cuenta los puntos totales. Esto se ve muy claramente en este ejemplo, ya que a pesar de que Denver Nuggets 1990/91 anotaban 23.7 puntos más por partido, que resulta un valor extremadamente elevado, San Antonio era un equipo más eficiente en ataque con un OFFRTG de 107.5 contra 105.2 de Denver. El eFG%; métrica la cual será explicada más adelante, de San Antonio es considerablemente superior que el de Denver Nuggets, lo que nos dice que los primeros son más eficientes en sus ataques.

En cuanto a la defensa podemos observar que ocurre algo parecido y es que el DEFRTG de Denver Nuggets esa temporada se corresponde con el peor que ha obtenido ningún equipo en la historia de la liga (114.8, recibiendo además 130.8 puntos por partido).

2.7. TRUE SHOOTING PERCENTAGE (TS%)

Esta métrica es una medida de la eficiencia en el tiro y que va a tener en cuenta los lanzamientos de campos realizados de dos y tres puntos, además de los tiros libres.

Su objetivo es calcular con una mayor precisión el tiro de un jugador o del conjunto de los jugadores de un equipo, más que el porcentaje en sí (*nba.com stats*).

La fórmula para valorar esta métrica es la siguiente:

$$TS\% = \frac{\text{Puntos Anotados}}{2 \times (FGA + 0'44 \times FTA)}$$

*FGA: tiros de campo intentados.

*FTA: tiros libres intentados.

Esta fórmula calcula el porcentaje de tiro que pondera el valor de los tiros de 2 puntos, 3 puntos y los tiros libres. Esto se ve ya que si por ejemplo un jugador anota un triple sumará un punto más al total de ellos que si esa misma canasta la hubiese anotado de dos, sin embargo, en ambos casos, sumará un tiro de campo realizado. Por lo tanto esta métrica da un mayor valor a los tiros de 3 que de 2.

Si a continuación nos fijamos en la fórmula, vemos que el total de tiros libres realizados se multiplica por 0.44. Este coeficiente se relaciona con el ratio de tiros libres provenientes de jugadas de 2 o de 3 puntos, además de los lanzamientos cuando se produce una jugada de falta y canasta (se llama and-1 en la NBA). Por ejemplo, si todos los tiros libres provienen de posesiones de 2 puntos, el coeficiente sería de 0.5, mientras que sería de 0.333 si a su vez todos esos tiros libres hubiesen provenido de jugadas de 3 puntos. Para las jugadas de and-1, será 0. Este coeficiente cambia en función de la liga que tomemos e incluso de la era, ya que las tendencias de juego cambian (unaneuronaparacadadeporte.com).

A continuación, se muestra un gráfico de cómo ha ido evolucionando el TS% a lo largo de los últimos 30 años, mostrando cada año al campeón de la temporada, junto con su TS% y su 3Par. El 3Par, hace referencia al porcentaje de lanzamientos de 3 puntos entre el total de lanzamientos de campo realizados (datos de basketball-reference.com):

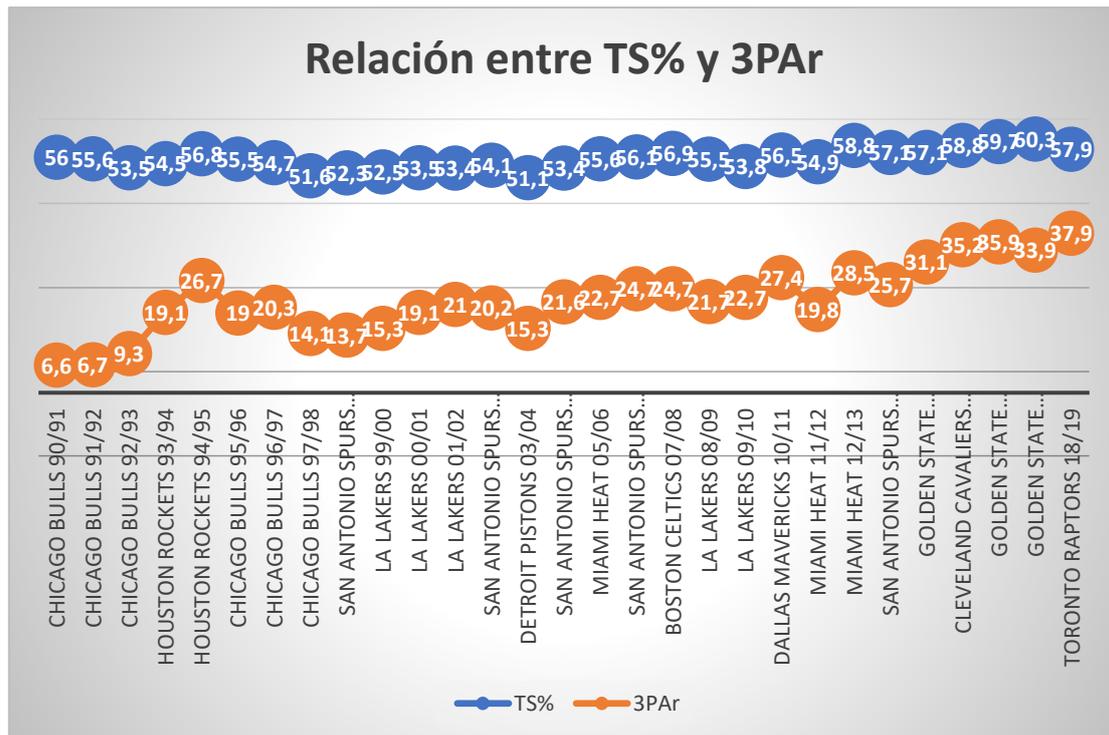


Gráfico 7: Relación entre TS% y 3PAr.

*3PAr: hace referencia al ratio de tiros de 3 que realiza un equipo. Este lo calcularemos dividiendo el número total de triples anotados entre el total de tiros de campo anotados.

Claramente vemos una tendencia al aumento del número de tiros de tres, y al ser este una propiedad importante del TS%, observamos que cuanto mayor es el número de lanzamientos de tres medios, normalmente provoca que el porcentaje de esta métrica aumente ligeramente.

El cambio más extremo se comprueba comparando los Chicago Bulls 97/98, quienes tenían un 3PAr de 14.1% y un TS% del 51.6%; con los Golden State Warriors 17/18, cuyo 3PAr era superior al doble del anterior (33.9%) y el TS% era 60.3%; una diferencia considerablemente grande.

2.8. FOUR FACTORS (EFG%, TOV RATE, OREB%, FT RATE)

Desarrollada por Dean Oliver, analista de datos y entrenador de baloncesto que ha trabajado para Seattle Supersonics y Denver Nuggets entre otros. Entre 2002 y 2004 desarrolló la teoría que estaba a punto de cambiar la forma en la que los números y el baloncesto se asocian. De acuerdo a la teoría de los "Cuatro Factores" o "Four Factors", esos aspectos estarán basados en cómo podría finalizar una posesión (statathlon.com, "Four Factors Basketball Success" [14]).

Los 4 factores que conforman esta métrica son:

- *Effective Field Goal Percentage (eFG%).*
- *Turnover Percentage (TOV%).*
- *Offensive Rebound Percentage (ORB%).*
- *Free Throw Rate (FT RATE).*

Además Oliver asignó a cada uno de estos factores un peso determinado basándose en el hecho de que una buena combinación entre ellos presenta una alta correlación con ganar partidos. Tanto el tiro, como mantener la posesión del balón sin pérdidas, rebotar y acudir a la línea de

personal son habilidades bastante diferentes entre ellas. Los pesos que se asignan a cada uno de los factores son: *eFG%* (40%), *TOV%* (25%), *ORB%* (20%) y *FT RATE* (15%) (*hoopshabit.com*, “*NBA advanced four factors winning*” [15]).

Sin embargo, a pesar de que Dean Oliver denominó a esta métrica “*Four Factors*”, este proceso estará formado realmente por ocho factores. Esto se debe a que no es únicamente el hecho de anotar puntos, si no también el de parar al equipo contrario de que los anote. Por lo tanto, esto nos dice que debemos coger esos mismos “*Four Factors*” y calcular la versión de los equipos contrarios para cada equipo (*squared2020.com*, “*Introduction to Oliver’s Four Factors*” [16]).

Para todos los factores basta con calcular la versión para el rival, excepto en el caso de los rebotes, en el que habrá que calcular el *Defensive Rebound Percentage (DRB%)*.

Por lo tanto los cuatro nuevos factores que tenemos son los siguientes:

- *Opponent Effective Field Goal Percentage.*
- *Opponent Turnover Percentage.*
- *Defensive Rebound Percentage (DRB%).*
- *Opponent Free Throw Rate.*

Explicamos cada uno de los factores:

2.8.1. EFFECTIVE FIELD GOAL PERCENTAGE (eFG%)

Se trata del factor más importante (al que mayor peso se le otorga, 40%) de los “*Four Factors*” de Dean Oliver.

Esta métrica es similar al *Field Goal Percentage (FG%)*, lo único que con *eFG%* ponderaremos también según los tiros de tres puntos.

Primero vamos a explicar como funciona la métrica ***Field Goal Percentage (FG%)***. Esta se define como la cantidad de tiros anotados entre la cantidad de tiros intentados:

$$FG\% = \frac{FGM}{FGA}$$

Mientras tanto, *eFG%* se define como la cantidad de tiros de campo anotados ponderados según su valor con respecto a los tiros de campo intentados:

$$eFG\% = \frac{FGM + (0.5 \times 3PM)}{FGA}$$

En esta métrica, a diferencia del *FG%*, no se le va a dar la misma importancia a los tiros de 2 puntos que a los tiros de 3, ya que ni valen lo mismo, ni es igual de fácil o difícil anotar uno u otro. Un aspecto donde puede fallar esta métrica es que siempre se otorga a los lanzamientos de tres puntos una mayor importancia que a los de dos, sin tener en cuenta cómo fue la defensa en ese momento (*Player Tracking*), o el momento en el que se produjo (*unaneuronaparacadadeporte.com*, “*Estadísticas NBA. FG% Vs eFG%*” [17]).

En referencia a lo expuesto anteriormente vamos a poner un ejemplo para diferenciar y así entender la diferencia entre *FG%* y *eFG%* [17]:

- Primer caso: Imaginemos un equipo que realiza 60 tiros, anota 42 lanzamientos, de los que 30 son de dos puntos y 12 son de tres puntos. Aplicamos ambas fórmulas:

$$FG\% = \frac{FGM}{FGA} = \frac{42}{60} = 0'7 \Rightarrow 70\%$$

$$eFG\% = \frac{FGM + (0'5 \times 3PM)}{FGA} = \frac{42 + (0'5 \times 12)}{60} = 0'8 \Rightarrow 80\%$$

- Segundo caso: En este caso el equipo realiza el mismo número de lanzamientos, es decir, 60. Pero ahora anotará 15 tiros de dos puntos, y 27 tiros de tres puntos. Aplicamos las fórmulas:

$$FG\% = \frac{42}{60} = 0'7 \Rightarrow 70\%$$

$$eFG\% = \frac{42 + (0'5 \times 27)}{60} = 0'925 \Rightarrow 92.5\%$$

La diferencia que observamos entre ambas métricas es evidente. Con el *FG%* hemos obtenido el mismo porcentaje para ambos equipos, ya que estos realizan y anotan el mismo número de tiros de campo. En el caso contrario el *eFG%* cambia entre dichos equipos y en el caso del que ha anotado más triples, va a conseguir un porcentaje bastante mayor que el que ha anotado más tiros de dos, aún habiendo anotado la misma cantidad de tiros de campo.

Cabe destacar, que en cuanto a la comparación entre el *TS%* y el *eFG%*, el primero generalmente obtiene un porcentaje significativamente mayor, ya que también tiene en cuenta los tiros libres, y esta es la estadística que mayor porcentaje de acierto presenta. Esto lo podemos ver en el *Gráfico 8*.

Dentro de los “*Four Factors*” se puede discutir fácilmente que se trata de la estadística más importante ya que en el transcurso de un partido la mayoría de los puntos son anotados en lanzamientos de campo. Si tu equipo presenta un elevado *eFG%*, y consigues que el oponente tenga un *eFG%* bajo, probablemente tanto tu ataque como tu defensa estén realizando un buen trabajo y tus probabilidades de conseguir la victoria sean mayores (*breakthroughbasketball.com*, “*What is Effective Field Goal Percentage? And why you should use it*” [18]).

En contraposición a lo anterior, una similitud entre las dos métricas anteriores es que al tener ambas en cuenta los lanzamientos de tres, los equipos que lancen más triples con un mayor acierto, conseguirán un mayor porcentaje tanto de *TS%* como de *eFG%*.

A partir de datos obtenidos de la página *basketball-reference.com*, vamos a ver las comparaciones entre la dos métricas anteriores para cinco equipos de la temporada 19/20:

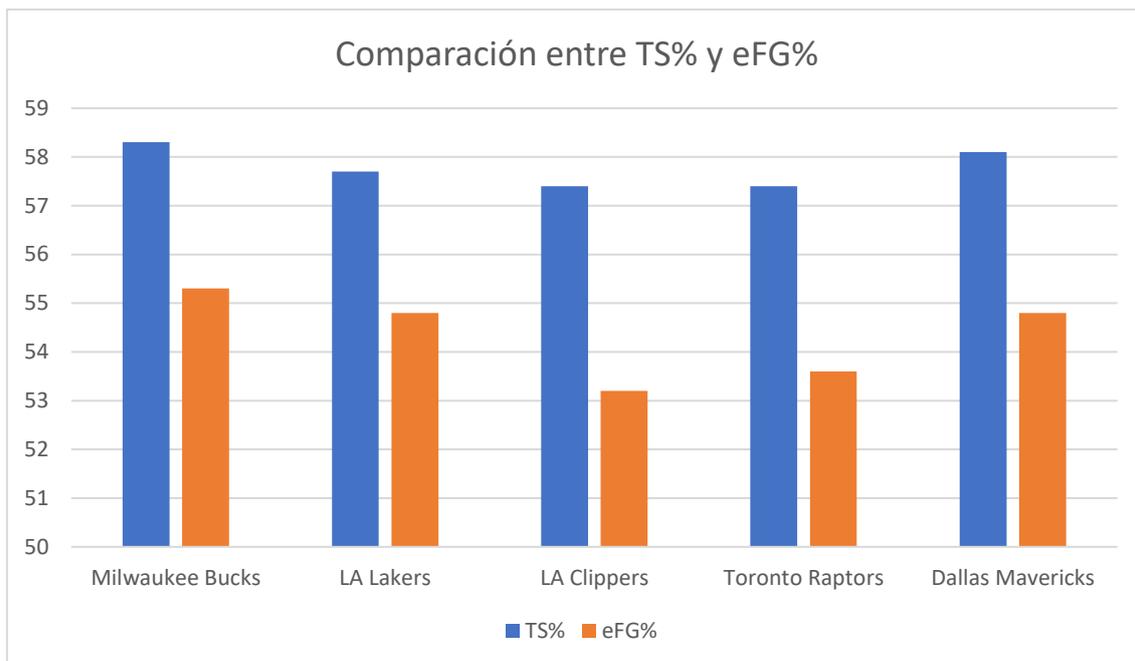


Gráfico 8: Comparación entre TS% y eFG%.

Un ejemplo de cómo funciona este factor tomando datos de *basketball-reference.com* de la temporada 18/19:

EQUIPOS	RÉCORD V/D DE LA TEMPORADA	eFG% / Ranking de equipos de esa temporada	Oponentes eFG% / Ranking de equipos de esa temporada
Golden State Warriors 18/19	57-25	56'5% / 1º	50'8% / 3º
Milwaukee Bucks 18/19	60-22	55% / 2º	50'3% / 1º
Toronto Raptors 18/19	58-24	54'3% / 3º	50'9% / 4º
Houston Rockets 18/19	53-29	54'2% / 4º	52'5% / 18º

Figura 17: eFG% y Opp. eFG% para los mejores equipos temporada 2018/19.

Se puede señalar que para la *Figura 17* anterior hemos seleccionado a los 4 mejores equipos durante la temporada en cuanto al eFG%. Estos 4 equipos consiguieron 4 de los 5 mejores récords de la liga esa temporada (sólo Houston Rockets superado por Denver Nuggets quienes consiguieron un récord de 54-28), además de que en dicha tabla encontramos a los dos finalistas de esa temporada (Golden State Warriors y Toronto Raptors) y 3 de los 4 finalistas de ambas conferencias (únicamente Houston Rockets no alcanzó las finales de la Conferencia Oeste ya que cayó eliminado por Golden State Warriors en las Semifinales de la Conferencia Oeste, en una serie que se decidió en el sexto partido).

Después, observamos que exceptuando Houston Rockets, el resto de estos equipos consigue dejar a sus oponentes en porcentajes de eFG% considerablemente bajos (3 de los 4 mejores equipos de la temporada en ese sentido).

Concluimos que, en referencia a lo que expusimos anteriormente, generalmente los equipos con mayores aspiraciones en una temporada de la NBA, presentan un mayor eFG% y además consiguen dejar a sus oponentes en un valor bajo de esta misma métrica. En este sentido, entendemos el por qué de Dean Oliver de considerar a este factor el más importante dentro de los "Four Factors".

2.8.2. TURNOVER RATE (TOV%)

Cabe destacar que dentro de los “Four Factors”, este va a ser al que Dean Oliver atribuye el segundo mayor peso (25%). Esta métrica se define en la página *basketball-reference.com* como una “estimación del número de pérdidas por cada 100 posesiones”. La fórmula es la siguiente:

$$TOV\% = \frac{TOV}{(FGA + 0'44 \times FTA + TOV)} \times 100$$

Vamos a realizar un análisis para esta métrica teniendo en cuenta los equipos de la temporada 2018/19. Esta idea está tomada a partir de un estudio similar que realizó Alex Boeder para la temporada 2011/12 (*nba.com/bucks*, “Advanced Stats Series Part II: Turnover Ratio” [19]):

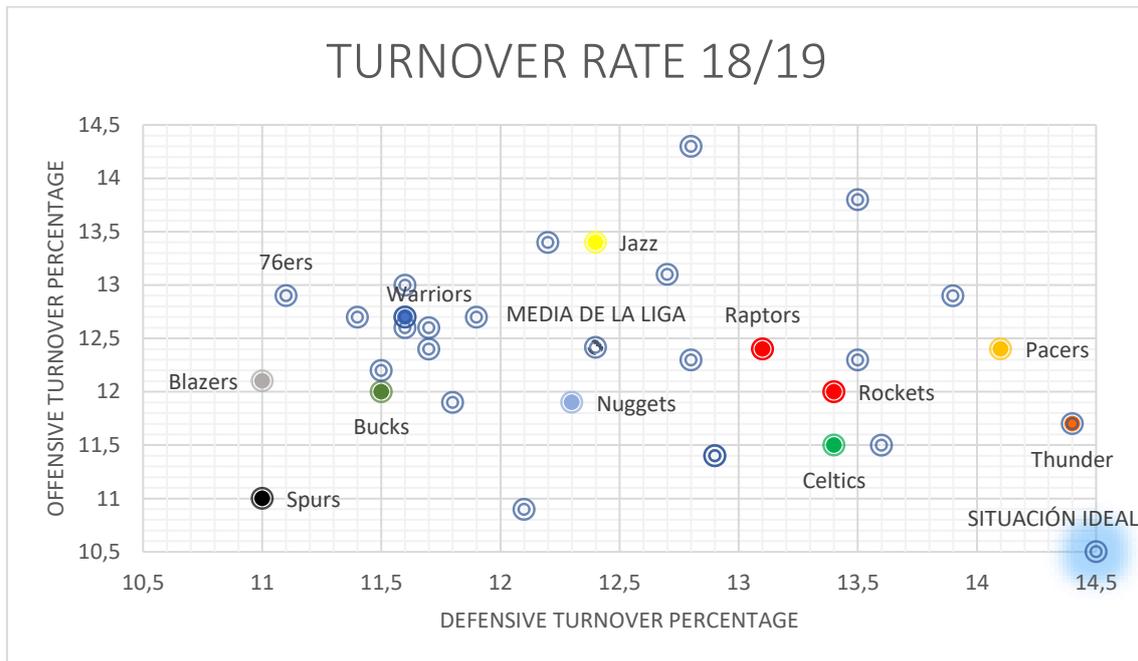


Gráfico 9: Turnover Rate temporada 18/19.

Respecto al *Gráfico 9*, sólo han sido marcados los 12 mejores equipos en cuanto a récord de la temporada, el resto de equipos no tienen una etiqueta asignada.

El punto ideal de cualquier equipo en cuanto al *TOV%* sería encontrarse en la parte inferior derecha del gráfico (etiqueta denominada “SITUACIÓN IDEAL”), es decir, tener el menor *TOV%* en ataque posible (lo que significaría cometer pocas pérdidas en sus ataques) y a su vez provocar que el *TOV%* del equipo contrario sea elevado (forzar el mayor número de pérdidas, lo que significa entre otras cosas recuperar muchos balones).

Del mismo modo, los equipos que más nos pueden llegar a llamar la atención son:

- 1º) San Antonio Spurs: fue el equipo que mejor cuidó la bola de toda la liga, con un *TOV%* de 11%, pero a su vez, fue el que menos % consiguió que obtuviesen sus oponentes. Si nos fijamos en sus estadísticas en *basketball-reference.com*, tuvieron una media de 12.1 pérdidas por partido y de 6.1 recuperaciones, ocupando las posiciones 1ª y 30ª respectivamente; esto apoyará nuestra idea inicial.

2º) Boston Celtics y Oklahoma City Thunder: ambos son los equipos que más cercanos se encuentran a la que consideramos situación ideal en cuanto a esta métrica. Vemos que estos dos equipos cuidaron muy bien de la bola (*TOV%* en ataque de 11.5 y 11.7 respectivamente) y al mismo tiempo (sobre todo en el caso de Thunder) fueron de los equipos que forzaron al oponente a un mayor *TOV%* (Thunder es el mejor equipo en esta categoría).

3º) Otros equipos como Indiana Pacers destacaron por su defensa forzando a los rivales a altos porcentajes de *TOV%* (14.1%), sin embargo, a la hora de cuidar el balón se sitúan en un valor muy cercano a la media.

4º) En cuanto a los dos mejores equipos, tomando como referencia el récord de la temporada, tanto Golden State Warriors como Milwaukee Bucks no destacan en este factor, ya que sus valores llegan a estar incluso en valores menos convenientes que la media del resto de equipos. A pesar de esto, cabe recordar que con esta métrica únicamente tenemos en cuenta un único factor de los denominados “*Four Factors*” de Dean Oliver y el cual tiene un peso de 25% sobre el total. Cuando antes nos fijamos en el factor que se considera más importante o con un mayor peso a la hora de predecir victorias (*eFG%*), vimos que estos dos equipos eran los mejores de la liga (*Figura 17*).

2.8.3. OFFENSIVE AND DEFENSIVE REBOUND PERCENTAGE (ORB% AND DRB%)

Se trata de la tercer factor dentro de los “*Four Factors*” de Dean Oliver. Este será el tercero en importancia, ya que se le asigna un peso del 20%. En este caso, ORB% hace referencia al tercer factor ofensivo, mientras que DRB% será el tercero defensivo.

Entender la estadística de rebotes en el mundo del baloncesto es esencial a la hora de un mejor trabajo y así poder encontrar debilidades en los equipos y ver qué funciona y qué no funciona. Según Jeff Haefner (*breakthroughbasketball.com*, “*Why every basketball coach should look at Rebounding Percentage instead of Totals*” [20]), normalmente a las estadísticas que más importancia se da en este aspecto son al número total de rebotes y al margen de rebotes (diferencia entre el total de rebotes de un equipo y su oponente), sin embargo, estas estadísticas son incompletas. Vamos a poner un ejemplo:

Imaginemos que un equipo lanza con un porcentaje del 30%, mientras que su oponente hace lo propio con un 60% de efectividad. Ahora pensemos que el primer equipo captura 20 rebotes ofensivos, mientras que el segundo equipo únicamente captura 14. Esto nos hace creer que el primer equipo fue más consistente a la hora de rebotear ofensivamente, ya que obtuvo un total de 6 rebotes ofensivos más que el segundo equipo, pero esto no va a ser así.

En contraposición a lo que podemos llegar a creer en un primer momento, el caso es que el segundo equipo llevó a cabo un mejor trabajo a la hora de rebotear ofensivamente. En ese sentido, el segundo equipo no tuvo tantas oportunidades ya que lanzaron de campo con un mayor porcentaje, mientras que el primer equipo, que lanzó con un porcentaje mucho menor, disfrutó del doble de oportunidades más que el segundo equipo para capturar rebotes ofensivos. Esto mismo ocurre en el caso de los rebotes defensivos.

Por lo tanto, como apuntaba Jeff Haefner, las estadísticas relacionadas con los rebotes que deberíamos tener en cuenta son *DFR%* y *ORB%*, cuyas fórmulas según la web *basketball-reference* son:

- *Defensive Rebounding Percentage (DRB%):*

$$DRB\% = \frac{DRB}{DRB + ORB}$$

- *Offensive Rebounding Percentage (ORB%):*

$$ORB\% = \frac{ORB}{ORB + DRB}$$

2.8.4. FREE THROW RATE

Es el cuarto factor dentro de los “*Four Factors*” de Dean Oliver y, a pesar de ser el que menos peso tiene a la hora de predecir victorias (15%) y ser el menos conocido; es un factor esencial, como nos indica Efran Khan (*bleacherreport.com*, “*Advanced NBA Stats for Dummies: How to understand the new sports math*” [21]).

Según Efran Khan, este factor es un indicador de la eficiencia ofensiva. En el baloncesto moderno, los tiros libres son una de las formas más eficientes a la hora de anotar, junto con los lanzamientos de tres puntos y los tiros bajo el aro. Generalmente, cuanto mayor sea el número de tiros libres que un equipo lanza por cada lanzamiento de campo que realiza, mayor será la eficiencia ofensiva para ese equipo. Esto también sirve para la faceta defensiva, ya que el *Free Throw Rate* del equipo rival se corresponde con el octavo factor (el 4º defensivo).

A este factor se le representa como *FT/FGA* en la web de baloncesto *basketball-reference.com*, y su fórmula es:

$$FT\ RATE = \frac{FTM}{FTA}$$

En relación a lo leído de Jason Irby (*theseason.gc.com*, “*Basketball Free Throw Rate*”), cuando los entrenadores hablan sobre las faltas en acciones de tiro, el objetivo principal es siempre intentar recibir las máximas posibles. El sentido general de esta métrica es aumentar el número de tiros libres intentados y es que, esta métrica es más importante que el propio porcentaje de tiros libres de un equipo (tiros libres anotados dividido entre el total de tiros libres lanzados).

Vamos a poner dos ejemplos tomando datos de la web *basketball-reference.com* en relación a la temporada 2018/19:

- Tengamos en cuenta a los Indiana Pacers. En términos del *eFG%*, Indiana consiguió estar en el top 10 de los mejores equipos esa temporada, pero en cuanto al *Offensive Rating*, su clasificación resultó ser la 18ª.

Esto sería un dato que no nos cuadra, ya que cuando hablábamos del *eFG%*, veíamos que era una métrica que la relacionábamos muy directamente con el suceso de un equipo. En consecuencia a la anterior discrepancia, sería muy difícil de explicar y la única manera que encontramos es si nos fijamos en el *Free Throw Rate* de Indiana ese año; vemos que es únicamente de 18.2%, lo que supone la 24ª marca de la liga.

Indiana es un equipo el cual produce muy buenos lanzamientos y consigue anotarlos con asiduidad, pero a no ser que no consigan acudir a la línea de tiro libre más a menudo, permanecerán en la mitad-baja de la clasificación en cuanto a eficiencia ofensiva (*basketball-reference.com* y *bleacherreport.com*).

- Otro ejemplo que vamos a mostrar es si comparamos al mejor equipo de la temporada en cuanto a *Free Throw Rate* con el peor. El mejor en cuanto a esta métrica serán Los

Angeles Clippers (25.8%), mientras que el peor los Orlando Magic (16.8%). Esto significa que por cada 100 lanzamientos de campo realizados por ambos equipos, los Clippers anotarán media de 9 tiros libres más, lo que supone 9 puntos a mayores.

Para esta métrica hay que tener en cuenta que en la página web *basketball-reference.com*, se diferencia entre *FTr* y *FT/FGA*, donde *FT/FGA* hace referencia a la métrica que ha sido explicada en este apartado, mientras que el *FTr* es similar, pero en vez de tomar el número de tiros libres anotados, tomará el número de tiros libres realizados, por lo que el valor de esta será algo superior al anterior.

Para los “*Four Factors*” tomaremos el llamado *FT/FGA*.

3. K-MEANS CLUSTERING CON TODAS LAS VARIABLES

Primero y gracias a los datos que nos proporciona la web basketball-reference.com, vamos a tomar las principales estadísticas que se han explicado anteriormente en este proyecto. Para ello vamos a tener en cuenta los 30 equipos de la NBA en las últimas 5 temporadas, por lo que el estudio se realizará sobre un total de 150 equipos.

El objetivo principal en este apartado tras la aplicación del método de *k-means*, será la agrupación de esos 150 equipos en distintos grupos en función de las semejanzas en las características intrínsecas de las variables de cada uno de dichos equipos y gracias a ello poder agruparlos en mejores y peores, o diferenciar también entre equipos más ofensivos o más defensivos y ver que rendimiento tanto presente como futuro pueden llegar a conseguir.

Como se ha indicado, los datos pertenecen a la página web basketball-reference.com y dentro de ella, tomamos las estadísticas que se encuadran dentro de la categoría de *Miscellaneous Stats* de los 30 equipos cada temporada repartidos en un total de 5 temporadas (tomamos como punto de partida la temporada 2015/16 y llegamos hasta la 2019/20):

Rk Team	Age	W	L	PW	PL	MOV	SOS	SRS	ORtg	DRtg	NRtg	Pace	FTr	3PAr	TS%	Offense Four Factors				Defense Four Factors			
																eFG%	TOV%	ORB%	FT/FGA	eFG%	TOV%	DRB%	FT/FGA
1 Milwaukee Bucks*	26.9	60	22	61	21	8.87	-0.82	8.04	113.8	105.2	+8.6	103.3	.255	.419	.583	.550	12.0	20.8	.197	.503	11.5	80.3	.162
2 Golden State Warriors*	28.4	57	25	56	26	6.46	-0.04	6.42	115.9	109.5	+6.4	100.9	.227	.384	.596	.565	12.6	22.5	.182	.508	11.7	77.1	.205
3 Toronto Raptors*	27.3	58	24	56	26	6.09	-0.60	5.49	113.1	107.1	+6.0	100.2	.247	.379	.579	.543	12.4	21.9	.198	.509	13.1	77.1	.190
4 Utah Jazz*	27.3	50	32	54	28	5.26	0.03	5.28	110.9	105.7	+5.2	100.3	.295	.394	.572	.538	13.4	22.9	.217	.507	12.4	80.3	.189
5 Houston Rockets*	29.2	53	29	53	29	4.77	0.19	4.96	115.5	110.7	+4.8	97.9	.279	.519	.581	.542	12.0	22.8	.221	.525	13.4	74.4	.210

Figura 18: Miscellaneous Stats Temporada 2018/19 (basketball-reference.com)

Para llevar a cabo este estudio, vamos a utilizar el lenguaje de programación *R*, usando a su vez la aplicación *RStudio*.

Vamos a leer los datos para los 150 equipos de la NBA que participaron en las últimas 5 temporadas y lo estructuraremos de la siguiente forma:

TEAM	AGE	W	L	PW	PL	POINTS	Opp.POINTS	MOV	SOS	SRS	Ortg	DRtg	NRtg	Pace
BUCKS 19/20	29.2	53	12	52	13	118.6	107.4	11.29	-0.85	10.44	112.6	101.9	10.7	105.0
LAKERS 19/20	29.6	49	14	45	18	114.3	106.9	7.41	0.34	7.75	113.0	105.6	7.4	100.9
CLIPPERS 19/20	27.4	44	20	44	20	116.2	109.7	6.52	0.22	6.74	113.6	107.2	6.4	101.8
RAPTORS 19/20	26.6	46	18	44	20	113.0	106.5	6.45	-0.57	5.88	111.6	105.2	6.4	100.6
MAVERICKS 19/20	26.2	40	27	45	22	116.4	110.3	6.04	-0.20	5.84	116.7	110.6	6.1	99.1

FTr	X3PAr	TS.	eFG.	TOV.	ORB.	FT.FGA	Opp.eFG.	Opp.TOV.	DRB.	Opp.FT.FGA
0.263	0.423	0.583	0.553	12.8	20.7	0.196	0.486	12.2	81.7	0.172
0.268	0.355	0.577	0.548	13.2	24.6	0.196	0.509	13.8	78.4	0.202
0.293	0.370	0.574	0.532	12.7	24.0	0.232	0.503	12.3	77.3	0.210
0.256	0.418	0.574	0.536	12.8	21.6	0.205	0.502	14.6	76.1	0.200
0.257	0.461	0.581	0.548	11.3	23.5	0.198	0.519	10.9	77.4	0.172

Figura 19 : Miscellaneous Stats de los equipos NBA de la temporada 2015/16 a la actual.

En esta *Figura 19* se adjuntan dos estadísticas básicas que no aparecían en la *Figura 18* y que también hemos obtenido gracias a basketball-reference.com. Estas son la cantidad de puntos anotados y recibidos de media por partido por cada uno de los equipos. Las variables AGE (edad media de los jugadores), W/L, PW/PL no van a ser tenidas en cuenta en este análisis.

Podemos observar que para la identificación de cada uno de los equipos mostramos el nombre del equipo con la temporada en la cual participa, así podemos diferenciar mismos equipos en distintas temporadas.

A cada equipo se le va a asignar un número, que se corresponderá con la fila que ocupan en el archivo de datos que se ha utilizado.

En los pasos venideros, para poder representar el resultado de nuestro procedimiento de *k-means*, debido a que tenemos más de dos dimensiones, esa representación se hará realizando un *Análisis de Componentes Principales (PCA)* y representándolo con las dos primeras componentes, todo esto intentando que la pérdida de información que se produce al realizar este proceso sea mínima. Por lo tanto, el *PCA* es un método de reducción dimensional.

3.1. EXPLICACIÓN TEÓRICA DEL PROCEDIMIENTO CLUSTER DE K-MEANS

El algoritmo *k-medias* o *k-means* es uno de los métodos de agrupamiento o *clustering* más conocidos en el Machine Learning [24]. Este va a pertenecer a los modelos de clasificación no supervisados ya que no se conocen de antemano las clases a las que van a pertenecer los datos, si no que va a ser trabajo del propio modelo el encontrar semejanzas entre ellos y así poder agruparlos a partir de las características de sus variables. Para la realización de este método se pueden tener en cuenta distintos tipos de distancias (euclídea, Manhattan...) lo que hace que este método sea muy adaptable a escenarios muy diversos.

En la realización de este ejercicio hemos utilizado la distancia euclídea. Esta se define como la distancia la cual va a unir dos puntos diferentes. Se calcula utilizando el teorema de Pitágoras. La fórmula de esta distancia para un espacio euclídeo n-dimensional es la siguiente:

Cada punto está definido por un vector de n coordenadas :

$$p = (p_1, p_2, \dots, p_n) \text{ y } q = (q_1, q_2, \dots, q_n)$$
$$d_{euc} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$$

En términos generales, el algoritmo de *k-means clustering* va a resolverse de la siguiente manera:

- Primero vamos a definir ***k centroides***.
- Después se calculará la distancia de cada punto a cada uno de los *k* centroides que tenemos y se asignará a aquel al que la distancia es la menor.
- A la vez, se va a actualizar la posición de los *k* centroides, calculándose así la posición promedio de todos los puntos que pertenecen a cada grupo.
- A continuación, repetirá los dos pasos anteriores hasta que ninguno de los centroides cambien de posición, por lo que las asignaciones de puntos entre grupos no cambia.

3.2. ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

El proceso de *Análisis de Componentes Principales* o *PCA*, básicamente, es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información [25]. Por lo tanto, nos va a permitir “reducir” la información que aportan numerosas variables a tan solo un número reducido de componentes.

Este proceso va a identificar aquellas direcciones en las que la varianza sea mayor. Sabemos que la varianza de una variable está medida siempre en su misma escala y elevándolo al cuadrado, por lo que si antes de realizar el proceso de PCA no normalizamos, las variables cuyas magnitudes sean mayores (véase puntos anotados de media o recibidos en comparación con otras como *eFG%*) dominarán sobre el resto.

Una de las funciones de R que realiza el *Análisis de Componentes Principales o PCA* es *prcomp*. Recordamos que para llevar a cabo este análisis ya lo hacemos con los datos estandarizados.

Al realizar este proceso obtendremos 20 componentes principales. En la *Figura 20* siguiente se van a mostrar los pesos de cada una de las variables en las 5 primeras componentes principales:

	PC1	PC2	PC3	PC4	PC5
POINTS	-0.264608095	-0.328422624	-0.073652230	0.018565128	0.15021783
Opp. POINTS	0.074363153	-0.462177020	-0.066763202	0.003271664	0.09593513
MOV	-0.369520716	0.134352712	-0.008456378	0.017252752	0.06111944
SOS	0.195401300	-0.114501918	-0.114674773	-0.147349857	0.08262650
SRS	-0.367718441	0.130223957	-0.017901446	0.006443152	0.06994566
ORtg	-0.336393037	-0.159846565	-0.065053674	-0.118724019	0.10213925
DRtg	0.196860114	-0.377159717	-0.052682510	-0.153982089	0.01848763
NRtg	-0.369413733	0.134872770	-0.011591772	0.016985127	0.06055574
Pace	-0.080582736	-0.379333597	-0.048340091	0.162204794	0.14468513
FTr	-0.046506472	0.054855320	-0.559164897	-0.260268961	-0.23992791
X3PAR	-0.166056721	-0.271119385	0.050184137	0.126658516	-0.21377004
TS.	-0.325750929	-0.187392754	-0.082441317	0.095259907	-0.11204151
eFG.	-0.320724203	-0.197532711	0.016248468	0.170984747	-0.03142375
TOV.	0.097876940	0.001173807	-0.257696956	0.393458185	0.15713845
ORB.	-0.002391315	0.104231461	-0.232782690	-0.183922074	0.79341093
FT. FGA	-0.070916306	0.032007875	-0.544770914	-0.294806764	-0.31435339
Opp. eFG.	0.184663583	-0.341838378	0.054648874	-0.057782078	-0.06505733
Opp. TOV.	0.033185202	0.134524263	-0.111567966	0.558084375	-0.13822090
DRB.	-0.124629556	0.033415634	0.317290194	-0.316405372	0.01757301
Opp. FT. FGA	0.124779033	0.043810543	-0.333728851	0.317236774	0.15472369

Figura 20: Pesos en las 5 primeras componentes principales en modelo con todas las variables.

En los pesos para la primera componente principal o *PC1* de la *Figura 21*, algunas variables como el *MOV*, el *SRS* y el *NETRTG* (*NRtg* en el análisis llevado a cabo en R) son las que reciben un mayor peso, mientras que otras como el *ORB%* tienen un peso menor.

En la siguiente figura se muestra un resumen del análisis anterior que nos va a permitir ver la importancia de cada una de las componentes:

COMPONENTES	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
DESVIACIÓN ESTÁNDAR	2.581	2.063	1.531	1.211	1.033	1.001	0.872	0.847	0.812	0.73
PROPORCIÓN DE VARIANZA	0.333	0.213	0.117	0.073	0.053	0.05	0.038	0.035	0.033	0.027
PROPORCIÓN ACUMULADA	0.333	0.546	0.663	0.736	0.79	0.84	0.878	0.914	0.947	0.973

COMPONENTES	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20
DESVIACIÓN ESTÁNDAR	0.664	0.268	0.085	0.074	0.035	0.031	0.015	0.011	7.35e-4	7.7e-16
PROPORCIÓN DE VARIANZA	0.022	0.004	0.0004	0.00028	6e-05	5e-05	1e-05	1e-05	0	0
PROPORCIÓN ACUMULADA	0.995	0.9992	0.9995	0.9998	0.9999	0.99999	0.99999	1	1	1

Figura 21: Importancia de cada una de las componentes principales en modelo con todas las variables.

A partir de la *Figura 21* vemos que las dos primeras componentes explican un 33.31 y en 21.29% de la varianza, lo que hace un total de 54.60% de varianza acumulada. Si seguimos con el resto de componentes vemos que la cantidad de varianza que explican cada una de ellas es muy pequeña, llegando a ser prácticamente nula a partir de la componente principal número 12.

Esto lo apoyamos con el siguiente gráfico:

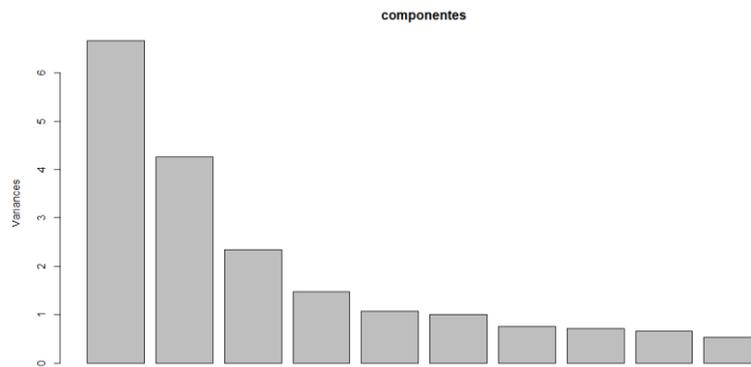


Gráfico 10: Diagrama de barras de la varianza explicada por las componentes principales para el modelo con todas las variables.

En el *Gráfico 10* se ha realizado un diagrama de barras de los autovalores, es decir, los valores que se obtienen al elevar al cuadrado cada una de las desviaciones típicas que se obtuvieron anteriormente en el *PCA*.

Las dos primeras componentes representarán un 54.60% de la varianza total, por lo que vamos a realizar una representación *biplot* (gráfico con ambas componentes).

Mediante la función *predict* calcularemos los valores tanto para la primera como la segunda componente principal para cada uno de los 150 equipos de las últimas 5 temporadas de la NBA, que será el paso anterior a representar cada uno de los equipos en el gráfico en función de las dos primeras componentes principales (PC1 y PC2):

[1]	-6.7822879367	-3.7138972823	-3.5182151746	-3.5915480177	-4.3717664621
[6]	-3.1844831502	-3.0365323636	-2.9325835396	-1.4138357481	-1.9641555846
[11]	-3.1775652519	-1.3520135929	-1.2826344422	-0.2966364173	0.6481057240
[16]	-0.0530282277	-0.1246020834	-0.1222491643	-0.6393604498	-0.1463116246
[21]	0.6086258148	1.3712063080	1.6731529317	0.7805009278	0.5021868249
[26]	3.1886439465	3.1070846787	2.9752612488	3.0729962783	4.1404113156
[31]	-5.8501199593	-4.4355461744	-3.7845962483	-3.2177540164	-3.1434251663
[36]	-2.4110989007	-1.7116334666	-2.6029816850	-1.2656584368	-1.5269490672
[41]	-2.5743640143	-1.6571125859	-1.3088652135	-0.6663258135	-0.8316500215
[46]	0.3091228990	0.1173702519	0.2950248965	0.0192327747	0.6816899206
[51]	-0.2356276739	-0.1928594245	0.3875757243	1.7914519273	0.1949142127
[56]	1.9175224583	3.6869439940	4.34405666124	4.2943380438	4.2156775461
[61]	-5.3682833682	-4.0337837487	-4.1222183815	-1.7239531314	-2.2983593027
[66]	-0.9304580553	-1.3744716556	-0.4841097803	-0.8465501745	-0.9474747743
[71]	-1.1640261970	-1.2599856119	-0.2852315756	-1.9826648831	-0.0850648731
[76]	-0.3559674335	0.0005899089	-0.5958943381	0.4825669255	-0.0654646254
[81]	0.6653993047	1.6636412560	2.1965412742	1.2158045713	2.2737399029
[86]	2.6458615722	4.0028004408	4.4562161223	3.4876676415	4.6073093672
[91]	-6.0927438186	-2.7401322963	-3.7046936725	-2.4540593286	-1.6251229404
[96]	-1.7754311187	-2.7033761115	-1.6593300937	-0.8078928987	0.4690252945
[101]	1.3931830550	0.1919476355	-0.9219036740	0.8102507002	-0.3165085277
[106]	0.2080153811	0.1873668068	0.3675700550	1.2366340679	0.8582710594
[111]	1.5905644717	1.3993573815	2.3860862108	2.2637235938	2.4143766276
[116]	3.3441834073	3.1491273509	3.8694864824	3.6743347003	2.4836094617
[121]	-5.7380152520	-4.0015550226	-3.0404039518	-2.5218771580	-1.5830571925
[126]	-1.5601824471	-0.7832434173	-0.0140455447	-1.0628243138	0.6139448828
[131]	0.5863637906	0.1225571750	0.2581503169	0.7818059355	0.3303928283
[136]	0.8240198615	1.0023207665	1.8794497934	2.0495811744	3.2049496442
[141]	1.7051599031	2.5132549002	2.5884618573	2.3942367048	2.4798605728
[146]	3.5473421218	4.3788000867	4.6371636856	5.9916105993	5.9428951804

Figura 22: Valor en la PC1 de cada uno de los 150 equipos de la NBA en modelo de todas las variables.

[1]	-0.42513098	-0.27413243	-0.91577915	0.07171957	-2.47213657
[6]	0.10650415	-3.38221201	-1.55647331	-0.52672173	-1.10439392
[11]	-1.32590083	-0.37471621	-0.25920984	-3.99218837	-0.20530607
[16]	-2.65941893	-2.18544337	-3.42873387	-1.20094844	-3.33317574
[21]	-2.04517165	-3.63067599	-0.80303765	-1.75641905	-4.38998613
[26]	-1.14534913	-1.20058066	-4.28977473	-3.43170759	-2.67785696
[31]	-1.42109096	-2.44750405	-0.74359030	-0.15427630	-2.15213000
[36]	-1.34594517	-0.26551678	-0.65195767	-0.67308253	0.93053957
[41]	-1.60565678	-1.62368408	-2.12717858	0.03218647	-1.50379240
[46]	0.65922766	0.04495622	-2.67427049	-1.54791346	-2.23074589
[51]	-3.39329738	-2.02892267	-2.09061632	0.52031475	-3.58474568
[56]	-4.01934054	-1.96276618	-3.30545163	-1.76226409	-2.93152462
[61]	-0.51825336	0.79706213	-1.11355766	2.21109925	0.96853327
[66]	1.26104037	2.46337889	2.88307300	1.54103574	-0.14906872
[71]	-1.23474890	-1.45751129	0.86026672	-2.28735150	0.48333127
[76]	-0.62565105	1.58129880	-0.07976154	1.10852323	0.01570105
[81]	-0.34947954	-0.16386642	0.13095854	-1.37625596	-0.47374894
[86]	-0.90202874	0.82868970	0.40794309	-0.98720472	-1.64411846
[91]	0.47312294	3.42709521	-1.75770657	0.55991056	3.31909512
[96]	2.12333972	-1.05190158	0.36785952	0.18082795	1.40191660
[101]	3.25053065	2.34622640	-1.80643249	2.78528853	0.81597692
[106]	-0.22044826	1.12866458	1.20506222	0.44343930	2.07590505
[111]	2.41705790	0.60106102	1.61440883	0.12001606	0.67293569
[116]	-0.68130498	0.48496270	-1.18204645	0.49323899	-1.50313356
[121]	0.73494380	5.55875855	2.39615274	3.12058986	3.01891530
[126]	3.81641452	3.24398395	3.31501815	2.69861645	4.78133403
[131]	3.74000056	4.12798306	1.38053872	3.21461561	0.80026663
[136]	1.89653179	1.54200486	2.75249520	2.18892275	3.66896319
[141]	0.28407839	3.21740140	1.76896635	1.58849423	0.77783410
[146]	2.75477054	0.95443470	1.14818020	1.12068903	0.97820206

Figura 23: Valor en la PC2 de cada uno de los 150 equipos de la NBA en modelo de todas las variables.

A partir de los datos obtenidos en la Figura 22 y Figura 23 vamos a representar cada uno de los 150 equipos en función de las dos primeras componentes en un *biplot*:

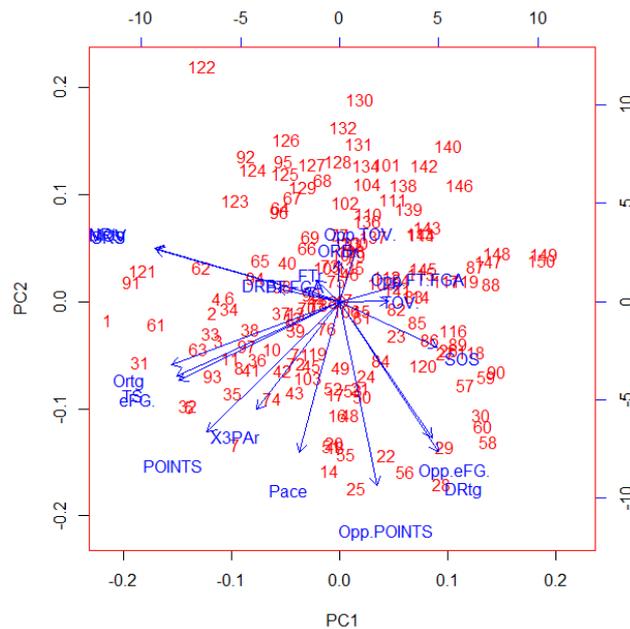


Gráfico 11: Gráfica de los 150 equipos sobre los 2 primeros ejes principales con la técnica biplot para modelo con todas las variables.

Observando el Gráfico 11 veremos las distintas tendencias que han ido teniendo los distintos equipos a lo largo de los años. Esto es debido a que además de tener en el gráfico las coordenadas de los 150 equipos, también tendremos las coordenadas de las distintas variables representadas mediante vectores.

En la parte central-izquierda del Gráfico 11 encontramos equipos que por lo general han sido referencia a lo largo de estas temporadas, como los Bucks 19/20, los Warriors de las temporadas 16/17, 18/19 y 15/16 o los Rockets 17/18 (números 1, 91, 32, 121 y 61 respectivamente). Según

el *Gráfico 11*, la tendencia de estos equipos, entre otras muchas cosas, era la de presentar un gran *OFFRTG*, además de un *eFG%* y un *TS%* elevados, que son algunas de las variables que contaban con mayor peso en la primera componente, junto con otras como *MOV* o *SRS*, en las que estos equipos también destacan obteniendo grandes valores.

Si ahora nos movemos a la parte inferior derecha encontramos equipos que no fueron importantes en el transcurso de las últimas 5 temporadas como son por ejemplo los Cavaliers 19/20, Suns 18/19 o los Timberwolves 18/19 (números 29, 58, 50 respectivamente). Estos equipos vemos que su *DEFRTG* presenta valores elevados, además de sus rivales presentar elevados porcentajes de tiro y anotar un número elevado de puntos. Estas variables presentaban un gran peso en la segunda componente principal (eje de la Y en el *Gráfico 11*), lo que provoca entre otras cosas que estos equipos se encuentren en la parte inferior del gráfico.

A partir de lo expuesto anteriormente y gracias a el *Gráfico 11*, vamos a poder ir haciéndonos una idea de como el *clustering* de *k-means* va a clasificar los distintos equipos en grupos en función de las características intrínsecas de sus variables (en función de los distintos valores que obtuvieron para las distintas *Miscellaneous Stats*).

En el *Gráfico 11* encontramos alguna observación que no sabríamos muy bien donde encuadrarla. La más destacada de estas es la 122 (que hace referencia a los San Antonio Spurs 15/16), quienes consiguieron obtener un *NETRTG* muy elevado (vimos en la explicación de cada una de estas métricas que se trató del quinto mejor valor de la historia de la liga). Este equipo se caracterizó por destacar defensivamente (su valor de 99.0 de *DEFRTG* es el mejor en este periodo de 5 temporadas y consiguieron dejar a sus rivales en un bajo 47.7% en *eFG%*). Además sus valores de *eFG%* y *TS%* fueron altos, junto con grandes resultados en variables como *MOV* y *SRS*, además de provocar numerosas pérdidas a los equipos rivales (alto *Opp. TOV%*). Al tratarse de un equipo tan destacado defensivamente pero también tratarse de un equipo muy bueno en las variables ofensivas, hace que se encuentre considerablemente alejado de los considerados grandes equipos de este periodo de 5 temporadas.

Una vez finalizado el *PCA*, continuamos con el **método *k-means clustering***.

3.3. REALIZACIÓN DEL MÉTODO K-MEANS

Para empezar la aplicación de este método, vamos a necesitar la cantidad *k* de centroides que va a ser utilizada, pero esta no es conocida de antemano, por lo que es necesario el aplicamiento de alguna técnica para determinar dicho valor.

Una de las técnicas que se utilizan para obtener el número óptimo *k* de *clusters* es el llamado **Método del Codo** o **Elbow Method**. Esta es la solución para cuando no tenemos información adicional en la que basarnos, ya que, por ejemplo, si quisiésemos separar los distintos equipos en ofensivos y defensivos sí que sabríamos el valor que daríamos a *k*, ya que querríamos dividir el conjunto de equipos en dos únicas categorías. En este caso queremos estimar el número óptimo de categorías en la que dividir a los equipos en función de sus características. Para ello se aplica el algoritmo de *k-means* para un rango determinado de valores de *k* (ese rango se lo tenemos que marcar nosotros). A continuación, hay que identificar el valor a partir del cual la reducción de la suma total de la varianza *intra-cluster* va a dejar de ser sustancial. Este es el que conocemos como *Método del Codo*.

En *R* existe una función denominada *fviz_nbclust* que nos permite automatizar el proceso completo para la técnica anterior y además nos proporciona una representación de los resultados que obtenemos:

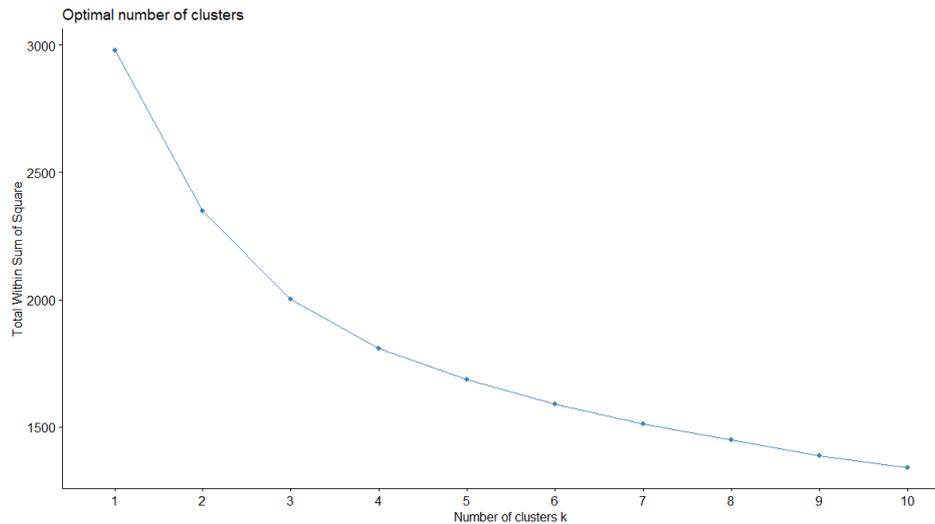


Gráfico 12: Representación de la suma total de cuadrados *intra-cluster* para modelo con todas las variables.

Una vez realizado este análisis se considera que a partir de 4 *clusters* la reducción en la suma de cuadrados internos empieza a estabilizarse, por lo que tomamos $k = 4$. A partir del *Gráfico 12* podríamos pensar que lo mejor sería coger $k = 5$, pero al probar a realizar el análisis con dicho valor, había algunos grupos que quedaban muy mezclados entre ellos.

Para este análisis utilizamos la función de *R* denominada *fviz_cluster*, la cual se encuentra en el paquete *factorextra*. El número de variables es superior a dos, lo que indica que la dimensionalidad de nuestro análisis es superior a dos, así que esta función lleva automáticamente a cabo un *análisis de componentes principales* o *PCA*. Por lo tanto, los resultados obtenidos quedarán representados sobre un espacio de dimensión dos, siendo ambas dimensiones las dos componentes principales que explicábamos en el subapartado anterior. El resultado que obtenemos tras aplicar el procedimiento clustering de *k-means* con $k = 4$ es:

NÚMERO IDENTIFICADOR DEL EQUIPO	EQUIPOS	CENTROIDE AL QUE SE ASIGNA
23	BULLS 19/20	1
26	KNICKS 19/20	1
27	HORNETS 19/20	1
28	HAWKS 19/20	1
29	CAVALIERS 19/20	1
30	WARRIORS 19/20	1
54	GRIZZLIES 18/19	1
57	BULLS 18/19	1
58	SUNS 18/19	1
59	KNICKS 18/19	1
60	CAVALIERS 18/19	1
82	MAVERICKS 17/18	1
83	KNICKS 17/18	1
84	NETS 17/18	1
85	MAGIC 17/18	1
86	HAWKS 17/18	1
87	GRIZZLIES 17/18	1

88	KINGS 17/18	1
89	BULLS 17/18	1
90	SUNS 17/18	1
112	PELICANS 16/17	1
113	MAVERICKS 16/17	1
114	KINGS 16/17	1
115	KNICKS 16/17	1
116	SUNS 16/17	1
117	76ERS 16/17	1
118	LAKERS 16/17	1
119	MAGIC 16/17	1
120	NETS 16/17	1
141	KINGS 15/16	1
143	NUGGETS 15/16	1
144	WOLVES 15/16	1
145	PELICANS 15/16	1
146	BUCKS 15/16	1
147	SUNS 15/16	1
148	NETS 15/16	1
149	LAKERS 15/16	1
150	76ERS 15/16	1
40	PACERS 18/19	2

46	HEAT 18/19	2
64	JAZZ 17/18	2
66	THUNDER 17/18	2
67	CELTICS 17/18	2
68	SPURS 17/18	2
69	BLAZERS 17/18	2
73	PACERS 17/18	2
75	WIZARDS 17/18	2
77	HEAT 17/18	2
79	PISTONS 17/18	2
92	SPURS 16/17	2
95	JAZZ 16/17	2
96	RAPTORS 16/17	2
100	THUNDER 16/17	2
101	GRIZZLIES 16/17	2
102	HEAT 16/17	2
104	BULLS 16/17	2
105	HORNETS 16/17	2
107	BUCKS 16/17	2
108	PACERS 16/17	2
110	HAWKS 16/17	2
111	PISTONS 16/17	2
122	SPURS 15/16	2
123	THUNDER 15/16	2
124	CAVALIERS 15/16	2
125	CLIPPERS 15/16	2
126	RAPTORS 15/16	2
127	HAWKS 15/16	2
128	CELTICS 15/16	2
129	HORNETS 15/16	2
130	JAZZ 15/16	2
131	PACERS 15/16	2
132	HEAT 15/16	2
133	BLAZERS 15/16	2
134	PISTONS 15/16	2
135	ROCKETS 15/16	2
136	MAVERICKS 15/16	2
137	WIZARDS 15/16	2
138	BULLS 15/16	2
139	MAGIC 15/16	2
140	GRIZZLIES 15/16	2
142	KNICKS 15/16	2
9	NUGGETS 19/20	3
10	THUNDER 19/20	3
12	76ERS 19/20	3
13	PACERS 19/20	3
14	PELICANS 19/20	3
15	MAGIC 19/20	3
16	SUNS 19/20	3
17	GRIZZLIES 19/20	3

18	BLAZERS 19/20	3
19	NETS 19/20	3
20	SPURS 19/20	3
21	KINGS 19/20	3
22	WOLVES 19/20	3
24	PISTONS 19/20	3
25	WIZARDS 19/20	3
36	BLAZERS 18/19	3
37	NUGGETS 18/19	3
39	THUNDER 18/19	3
42	SPURS 18/19	3
43	CLIPPERS 18/19	3
44	MAGIC 18/19	3
45	NETS 18/19	3
47	PISTONS 18/19	3
48	KINGS 18/19	3
49	MAVERICKS 18/19	3
50	WOLVES 18/19	3
51	PELICANS 18/19	3
52	HORNETS 18/19	3
53	LAKERS 18/19	3
55	WIZARDS 18/19	3
56	HAWKS 18/19	3
70	WOLVES 17/18	3
71	NUGGETS 17/18	3
72	PELICANS 17/18	3
74	CAVALIERS 17/18	3
76	CLIPPERS 17/18	3
78	HORNETS 17/18	3
80	BUCKS 17/18	3
81	LAKERS 17/18	3
99	WIZARDS 16/17	3
103	NUGGETS 16/17	3
106	BLAZERS 16/17	3
109	WOLVES 16/17	3
1	BUCKS 19/20	4
2	LAKERS 19/20	4
3	CLIPPERS 19/20	4
4	RAPTORS 19/20	4
5	MAVERICKS 19/20	4
6	CELTICS 19/20	4
7	ROCKETS 19/20	4
8	JAZZ 19/20	4
11	HEAT 19/20	4
31	BUCKS 18/19	4
32	WARRIORS 18/19	4
33	RAPTORS 18/19	4
34	JAZZ 18/19	4

35	ROCKETS 18/19	4
38	CELTICS 18/19	4
41	76ERS 18/19	4
61	ROCKETS 17/18	4
62	RAPTORS 17/18	4
63	WARRIORS 17/18	4
65	76ERS 17/18	4

91	WARRIORS 16/17	4
93	ROCKETS 16/17	4
94	CLIPPERS 16/17	4
97	CAVALIERS 16/17	4
98	CELTICS 16/17	4
121	WARRIORS 15/16	4

Figura 24: Equipos asignados a cada centroide tras procedimiento k-means sobre modelo con todas las variables.

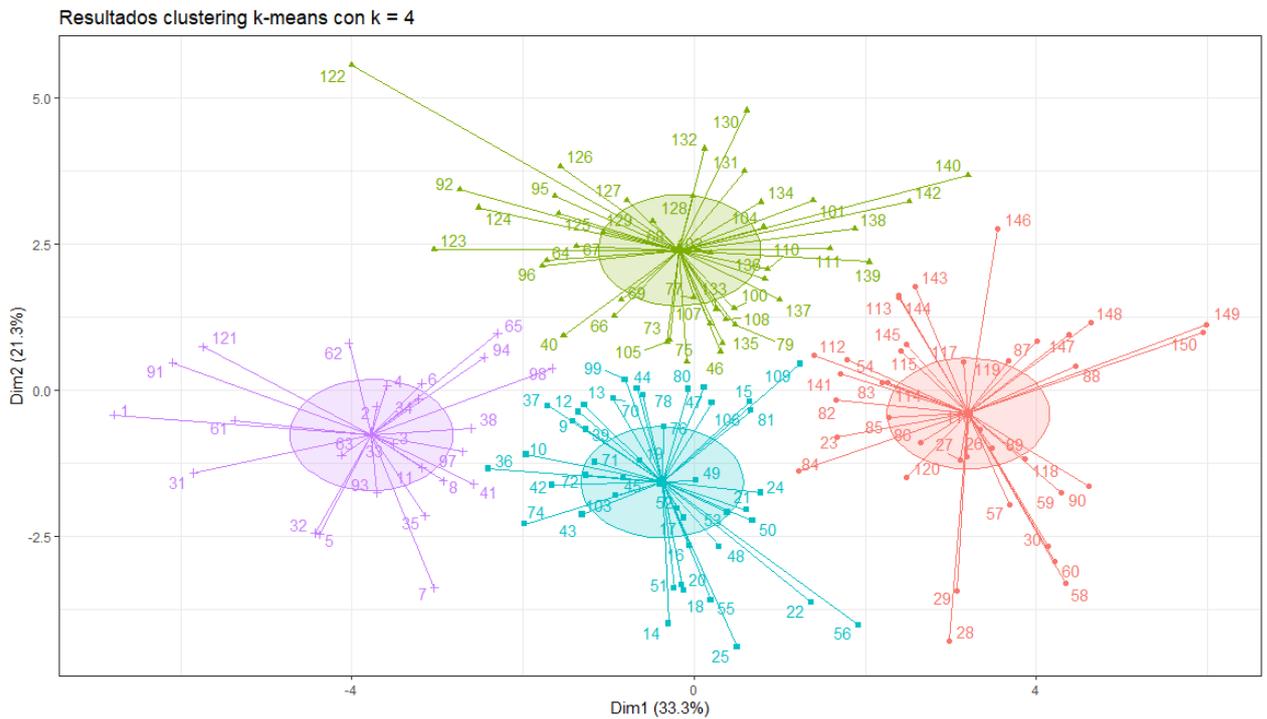


Gráfico 13: Resultados del clustering k-means con k = 4 sobre modelo con todas las variables.

En la *Figura 24* se muestra una tabla donde aparecen los 150 equipos utilizados en este estudio. Además en dicha tabla encontramos el número que se le asigna a cada uno de esos equipos, a parte del centroide al que pertenecen según este análisis.

Se utilizó $k = 4$, por lo que el número de grupos en el que se van a dividir los equipos será 4. Vamos a ver qué equipos son asignados a cada grupo:

1º) **Centroide 4** (color morado, 26 equipos): En este grupo encontramos los mejores equipos de estas 5 temporadas de la NBA. 3 de los últimos 4 campeones de la liga se encuentran aquí (a excepción de los Cleveland Cavaliers campeones en la temporada 2015/16), además 10 de los mejores clasificados en cuanto al récord de victorias-derrotas durante la temporada 2019/20 también pertenecen a este grupo. Todos los equipos asignados a este centroide han sido o son candidatos al título o a obtener posiciones altas en la clasificación de cara a los *Play-Offs* en sus respectivas temporadas. Hay algunos ejemplos como puede ser Dallas Mavericks 2019/20 que quizás por récord en liga regular no se debería encontrar dentro de este grupo donde se clasifican a los 26 mejores equipos de estas temporadas (proyección de 49 victorias durante la temporada 2019/20), pero que

estadísticamente hablando y mirando a los valores que obtienen para las diferentes variables, estos son mucho mejores en comparación a dicho récord del que hablábamos.

2º) **Centroide 1** (color rojo, 38 equipos): Los equipos que se asignan a este grupo son los que peor rendimiento han ofrecido durante estas temporadas, es decir, equipos que nunca se han clasificado para *Play-offs* en este periodo de tiempo y que aparecen en las últimas posiciones en las clasificaciones. Como veíamos en el *Gráfico 11*, los equipos que se encuentran en este grupo destacan por tener un *DEFRTG* elevado, además de sus rivales obtener un *eFG%* también alto. Son los equipos con un mayor *TOV%* además de ser los que provocan un mayor número de tiros libres realizados por sus rivales, dos de los que considerábamos factores esenciales dentro de los “*Four Factors*” de Dean Oliver.

En los dos centroides o grupos restantes los equipos que se clasifican tras el análisis no están demasiado claros, ya que hay equipos que debido a su récord podrían clasificarse con los mejores equipos, pero que lo consiguen a partir de otros estilos o tendencias de juego, lo que provoca que tengan valores buenos en variables diferentes a los equipos que se clasificaban en el centroide 4.

Aún así, explicamos los equipos que encontramos en estos dos grupos restantes:

3º) **Centroide 2** (color verde, 43 equipos): El rango de equipos dentro de este grupo es algo variado, pero generalmente se trata de equipos que obtienen posiciones altas en *Play-offs* o que se encuentran en la lucha por conseguir plazas para disputar la posttemporada. Muchos de estos equipos se clasifican en este grupo ya que, a pesar de ser grandes equipos y muchos de ellos pelear por el campeonato (en este grupo se encuentran los Cleveland Cavaliers campeones en la temporada 2015/16), no consiguieron ser estadísticamente hablando tan dominantes como lo fueron los equipos que se asignaban al Centroide 4, aunque consiguieron buenos récords y por ende, buenas clasificaciones en sus respectivas temporadas.

Otro dato que nos llama la atención es que en este grupo no hay asignado ningún equipo de la temporada 2019/20, lo que nos hace apoyar la idea que se exponía anteriormente en este proyecto y es la clara tendencia en los últimos años al aumento en el ritmo de juego o *PACE*. La variable *PACE* es una de las que cuenta con un mayor peso dentro de la segunda componente y por lo tanto los equipos de esta última temporada, que por media suelen tener un mayor valor en el ritmo de juego, aparecen en lugares más bajos dentro del gráfico y son asignados a otro grupo por lo general.

Dentro de este grupo y gracias al *Gráfico 11*, estos equipos destacaban respecto a equipos de otros grupos en que conseguían mejores valores en las estadísticas más defensivas, como forzar a los rivales a un mayor número de pérdidas o a fallar más tiros (mayor valor de *Opp.TOV%* y menor de *Opp.eFG%* respectivamente).

Un ejemplo claro era el que mencionábamos anteriormente de los San Antonio Spurs 2015/16, quienes siendo uno de los mejores equipos en estas 5 temporadas, consiguió dicho éxito destacando en diferentes facetas a lo que lo hicieron otros grandes equipos durante este periodo.

4º) **Centroide 3** (color azul, 43 equipos): A este grupo también son asignados grandes equipos que ocupan plazas altas dentro de sus respectivas temporadas o que lucharon por obtener una plaza en la posttemporada (encontramos a los Cleveland Cavaliers 2017/18 finalistas de la NBA esa temporada), sin embargo, es el grupo que peor clasifica, ya que el

rango de equipos en cuanto a récord en sus temporadas y por lo tanto, clasificación, es muy variado.

Un factor esencial es el que nombrábamos anteriormente, y es que la tendencia de las últimas temporadas de un aumento en los lanzamientos de tres (lo que provoca un aumento de la variable *3PAR*) y en el ritmo de juego o *PACE*, genera que muchos equipos, la mayoría de las últimas dos o incluso tres temporadas, que deberían clasificarse dentro de los peores equipos en temporadas se asignen dentro de este grupo. Algunos ejemplos de ello son los Minnesota Timberwolves o Detroit Pistons de la temporada 2019/20.

En resumen, este *análisis cluster de k-means* distingue muy bien entre los mejores y los peores equipos de estas 5 temporadas, sin embargo, el claro cambio en la tendencia de juego en los últimos años nos lleva a problemas a la hora de clasificar algunos equipos. A pesar de esto, tenemos cuatro grupos bien marcados, uno para los mejores equipos, otro para los peores y los dos últimos para buenos equipos, aunque con características distintas estadísticamente hablando y que en algún caso incluyen a aspirantes al título de la NBA.

4. MÉTODOS DE SELECCIÓN DE VARIABLES

El objetivo principal de este capítulo, teniendo en cuenta que más tarde se va a llevar a cabo la exploración de modelos de predicción del éxito de los distintos equipos, basados en métodos de regresión lineal múltiple, será eliminar el **sobreajuste** que se produce en los datos cuando tenemos el modelo formado por todas las variables. Ahí es donde entran los métodos de selección de este capítulo, donde eliminaremos aquellas variables que son las principales causantes de dicho sobreajuste y que además, una vez eliminadas, el modelo resultante pierda la menor calidad de ajuste posible.

Para estos modelos, se tomará como variable respuesta el total de victorias conseguidas por cada equipo en sus respectivas temporadas, que nos permite medir el éxito de cada uno de ellos en función de si obtienen más o menos victorias. Tanto esta variable como el resto de variables regresoras las encontramos en la *Figura 19*.

Hay que tener en cuenta que la temporada 2019/20 hasta el momento de realización de este trabajo, debido al problema del coronavirus, los equipos no llegaron a disputar la totalidad de los 82 partidos de liga regular; por lo tanto, al tener en cuenta el número de victorias de los equipos, siempre va a ser menor que otras temporadas. Para solucionar dicho error hemos tomado el Porcentaje de Victoria de todos los equipos esa temporada y con dicho valor hemos calculado el hipotético número de victorias totales que hubieran conseguido los equipos si se hubiesen disputado los 82 partidos de liga regular.

Los métodos que vamos a aplicar son los siguientes:

4.1. COEFICIENTE DE CORRELACIÓN DE PEARSON

Se trata de la covarianza estandarizada. Evalúa el grado de relación entre dos variables cuantitativas. La relación que se quiere estudiar entre ambas variables que se toman es lineal y además deben ser cuantitativas. Este coeficiente va a tomar valores entre $[-1,+1]$, donde el +1 hace referencia a una correlación lineal positiva perfecta y el -1 una correlación lineal negativa perfecta.

Este método nos permite eliminar aquellas variables que nos van a producir redundancia en nuestros datos, es decir, aquellas que nos producen un sobreajuste.

A la hora de elegir cuáles van a ser las variables que vamos a eliminar de nuestro modelo, nos decantaremos por aquellas cuya correlación con otras variables es alta, que es lo que produce la redundancia. Además, este método nos permite ver aquellas variables cuya relación con la variable respuesta es mayor.

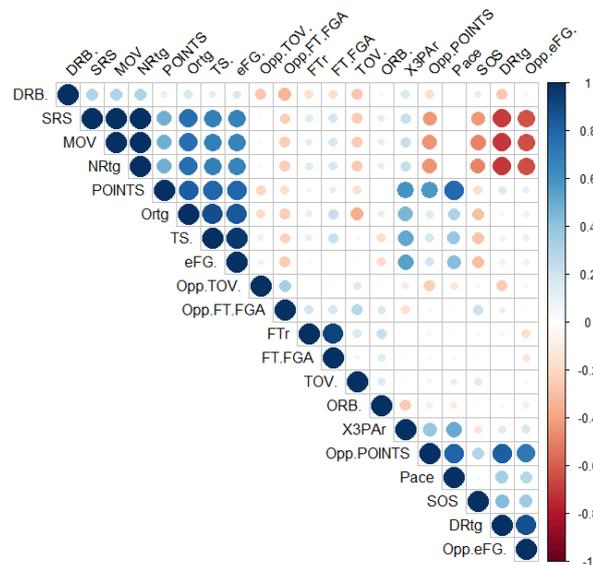


Gráfico 14: Correlación entre las distintas variables.

Algunas variables tales como *POINTS* y *Opp. POINTS* están muy relacionadas con el *OFFRTG* y el *DEFRTG*, ya que están intrínsecas en el cálculo de estas últimas y no tienen en cuenta otra variable esencial como es el *PACE* o ritmo de juego; por lo que las vamos a eliminar.

Otras tales como el *SRS*, *MOV* o el *NETRTG* (en el caso de este último al tratarse de la diferencia entre *OFFRTG* y *DEFRTG* lo eliminamos por redundancia) también están muy relacionadas con otras variables (altas correlaciones), por lo que al igual que las anteriores, las eliminamos.

En relación al *FTr* (*Free Throw Rate*), este hace referencia al total de tiros libres entre el total de tiros de campo realizados. En el *Gráfico 14* observamos que esta variable está altamente correlada con *FT/FGA* (denominada *FTr* en los “*Four Factors*” de Dean Oliver, pero en la web de *basketball-reference.com* hace una distinción entre ambas), calculándose esta como el número total de tiros libres anotados entre el total de tiros de campo realizados. Es un hecho que los tiros libres son el lanzamiento de campo donde se consigue un mayor porcentaje, por lo que la diferencia entre ambas variables va a ser mínima y la correlación entre ambas, al ser semejantes, es alta, por lo que eliminaremos también la primera.

Con el *SOS*, que nos indica la dificultad en el calendario que va a afrontar un equipo durante una temporada, también podría ser eliminada, ya que podemos asumir que tomando la temporada entera los equipos tendrán un valor de esta métrica similar. Sin embargo, debido a que los equipos juegan con más frecuencia durante la temporada contra equipos de su propia conferencia y división y gracias al nivel poco parejo que se viene dando entre ambas conferencias en los últimos años (encontramos más equipos de nivel en la conferencia Oeste con respecto a la Este), sí que la vamos a tener en cuenta.

Una vez eliminadas, vamos a ver la correlación existente entre el resto de variables y la variable respuesta, la cual se trata del total de victorias de cada equipo en sus respectivas temporadas.

VARIABLE	CORRELACIÓN	VARIABLE	CORRELACIÓN
SOS	-0.487	TOV%	-0.25
OFFRTG	0.734	ORB%	0.114
DEFRTG	-0.675	FT/FGA	0.171
PACE	-0.028	Opp. eFG%	-0.628

3PAr	0.195	Opp. TOV%	0.028
TS%	0.672	DRB%	0.274
eFG%	0.656	Opp. FT/FGA	-0.241

Figura 25: Correlación entre variables y variable respuesta total de victorias.

Según la *Figura 25* podemos ir haciéndonos una idea de qué variables son las que mayor relación presentan con la variable respuesta, ya sea positiva o negativa esta relación.

Encontramos algunas tales como *OFFRTG* que presenta una correlación alta con la variable respuesta. Esto es un resultado que cabía esperar, ya que esta variable mide cómo de bien ataca un equipo y es esencial. Además, en su fórmula tiene intrínsecas otras métricas como el número de puntos anotados o el ritmo de juego. En el caso del *DEFRTG* ocurre lo mismo pero en este caso la correlación va a ser alta pero negativa, teniendo en cuenta otras métricas tales como el total de puntos encajados y el ritmo de juego.

En cuanto a las variables respectivas al tiro (tanto *TS%*, como *eFG%*, además de la correspondiente defensiva a esta última, *Opp. eFG%*) van a presentar una alta correlación con la variable respuesta. La métrica *eFG%* está implícita dentro de *TS%*, ya que esta última tiene en cuenta los mismos factores además de los tiros libres. En el *Gráfico 14* vemos que existía una gran correlación entre estas dos, por lo que, para evitar la redundancia, vamos a eliminar también la variable *TS%*, por el simple hecho de que *eFG%* es uno de los factores de la métrica “*Four Factors*”, de ahí esta decisión.

Algunas otras como el *DRB%* o el *3PAr* van a tener una correlación positiva débil.

Del mismo modo, hay algunas cuya correlación es prácticamente nula, como por ejemplo el *PACE* o el *Opp. TOV%*.

Sin embargo, esta correlación de *Pearson* puede tener algunas limitaciones a la hora de seleccionar las variables. Esto se debe a que puede ocurrir que la relación que muestran dos variables se deba a otra tercera que influye sobre las dos anteriores, denominando a este fenómeno *confounding*. Por lo tanto, si el valor de la correlación existente entre dos variables es distinto al valor de la correlación parcial entre esas dos mismas variables cuando se está controlando una tercera, significa que esta última va a influir en las dos primeras.

Esto ocurre en el caso de que haya métricas que en su propio cálculo incluyen a otras métricas, o las métricas que se usan están muy relacionadas con otras de este mismo análisis. Un ejemplo, es en el caso en el que queremos estudiar la relación existente entre la variable respuesta número total de victorias de los equipos y el *OFFRTG* de ellos. Podemos sospechar que esta relación esté influenciada por el *eFG%* y es que, cuanto mayor sea el porcentaje en el lanzamiento de un equipo, mayor será su total de puntos anotados por posesión, lo que provoca que aumente su *OFFRTG* y, a su vez, mayor será por norma general el número de victorias que logran.

En relación a este último ejemplo, en la *Figura 25* vemos que la relación entre el número de victorias y el *OFFRTG* es alta (0.734) y significativa, ya que tiene un p-valor pequeño. Sin embargo, cuando se estudia la relación entre ambas, bloqueando la variable *eFG%*, a pesar de que dicha relación sigue siendo significativa (p-valor pequeño), la correlación pasa a ser baja (0.445).

4.2. ALGORITMOS DE SELECCIÓN DE VARIABLES

Para la realización de los distintos algoritmos, hemos elegido de entre los distintos criterios globales que disponíamos el *Criterio de Información Akaike (AIC)*. El objetivo principal será el de encontrar un modelo cuyo *AIC* sea el más pequeño posible, ya que en ese caso ideal, habría una verosimilitud grande y pocos parámetros. El *AIC* se va a definir de la siguiente manera:

$$AIC = -2 \times \log Lik + k \times n_{par}$$

Donde: *logLik* se corresponderá con el valor de *log-verosimilitud* del modelo para el vector de parámetros, *k* es un valor de penalización por el exceso de parámetros y n_{par} es el número de parámetros del modelo.

A la hora de interpretar esto hay que recordar que el mejor modelo es aquel que tiene un *logLik* elevado y un *AIC* bajo.

Dentro de estos algoritmos podemos encontrar los siguientes métodos:

4.2.1. BACKWARD

Consiste en la selección de variables hacia atrás. La idea general es partir del modelo más complejo, con todas las variables y el cual incorporara todos los efectos que pueden influir en la respuesta. En cada etapa se irá eliminando la variable menos influyente y que provoque que el valor del *AIC* descienda, hasta que ya no sea posible la eliminación de ninguna variable más.

Con la función *stepAIC* y el modo *direction = "backward"* podemos realizar este proceso de selección.

Coeficientes:

TÉRMINOS	COEFICIENTES	ERROR ESTÁNDAR	P-VALOR
Constante	40.927	0.253	< 2e-16 ***
OFFRTG	8.873	0.287	< 2e-16 ***
DEFRTG	-7.728	0.258	< 2e-16 ***
3PAr	-0.569	0.289	0.051

R^2 ajustado = 0.9375

P-valor = 2.2e-16

Figura 26: Final del método de selección de variables Backward sobre modelo con todas las variables.

***Con asterisco marcadas aquellas variables con p-valor significativo.

Vemos que el modelo de regresión múltiple que nos queda tras haber aplicado el método Backward es el formado por las variables: *OFFRTG*, *DEFRTG* y *3PAr*.

El R^2 ajustado resultante del modelo es de 93.75%. Este nos permite obtener el grado de intensidad o efectividad que van a presentar las variables independientes (las tres con las que nos quedamos en este modelo) en explicar la variable dependiente (número total de victorias por temporada). Aún habiendo eliminado un gran número de variables con respecto a nuestro modelo inicial, la cantidad de variabilidad explicada es muy alta (valor de R^2 ajustado cercano a 1), por lo que nos sigue quedando un muy buen modelo.

Tanto *OFFRTG* como *DEFRTG* van a tener un p-valor pequeño, por lo que con una confianza del 95% rechazamos la hipótesis nula de nulidad de ambos coeficientes. En el caso del *3PAr*, el p-valor es muy cercano a 0.05, por lo que al igual que antes tomando una confianza del 95% rechazamos la hipótesis.

4.2.2. FORWARD

Eliminación hacia delante. En este caso con el criterio global del *AIC*, partiremos del modelo más sencillo e iremos añadiendo términos siempre y cuando el valor del *AIC* del modelo no vaya aumentando. Iremos añadiendo las variables restantes más significativas hasta que ya no queden más variables o hasta que no se cumpla la condición que exponíamos antes.

Con la misma función *stepAIC* que en el caso anterior y con el modo *direction = "forward"*, llevaremos a cabo este método. La única diferencia es que en este caso vamos a tener que marcar qué modelo va a ser el más simple, que de inicio se trata del modelo de regresión lineal simple, sin ninguna variable. También tendremos que marcar nuestro "*horizonte*" y objetivo, que se trata del modelo de regresión múltiple el cual va a contener todas las variables.

Al realizar este método, obtenemos un modelo el cual contaba exactamente con las mismas variables que cuando lo hacíamos mediante el método *backward*.

5. MODELOS DE REGRESIÓN LINEAL MÚLTIPLE

El objetivo de este capítulo será probar distintos modelos de regresión y ver qué tal predicen estos el número total de victorias. Para ello, nos vamos a ayudar de los resultados que obteníamos en los métodos de selección de variables del apartado anterior. También estudiaremos el modelo formado por las variables que conforman los “*Four Factors*” de Dean Oliver. Para estos ejemplos tomaremos la variable número total de victorias de cada equipo en su respectiva temporada como variable respuesta.

En nuestro caso vamos a usar modelos de **regresión lineal múltiple**.

5.1. EXPLICACIÓN TEÓRICA DE LA REGRESIÓN LINEAL MÚLTIPLE

Esta nos va a permitir generar un modelo lineal en el que el valor de la variable respuesta (Y) se determina a partir de un conjunto de variables independientes o predictoras (X1, X2, X3,...). Esta se corresponde con una extensión de la regresión lineal simple, ya que esta última estima la variable respuesta a partir de una única variable predictora. Además de para predecir el valor de la variable respuesta, estos modelos también pueden usarse para evaluar la influencia que van a tener los predictores sobre ella.

Estos modelos tienen la siguiente ecuación:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni}) + e_i$$

Donde: el parámetro β_0 hace referencia a la ordenada en el origen, es decir, es el valor de la variable respuesta Y cuando el resto de los predictores es 0. β_i se corresponde con el efecto promedio el cual provoca el incremento en una unidad de la variable predictora X_i sobre la variable respuesta Y, mientras el resto de las variables se mantienen constantes. A estos parámetros los vamos a conocer como coeficientes parciales de la regresión. Por último, e_i es el residuo o error, que se corresponde con la diferencia entre el valor observado y el estimado por el modelo

5.2. REGRESIÓN LINEAL MÚLTIPLE CON TODAS LAS VARIABLES

Primero, vamos a llevar a cabo el modelo tomando todas las variables, eliminando únicamente aquellas que considerábamos redundantes (que nos producen un sobreajuste) cuando hacíamos el análisis de selección de variables a partir de las correlaciones de *Pearson* entre ellas.

Coefficientes:

TÉRMINO	COEFICIENTES	ERROR ESTÁNDAR	P-VALOR
Constante	40.927	0.257	2e-16 ***
SOS	-0.154	0.308	0.618
OFFRTG	7.435	2.166	0.0007 ***
DEFRTG	-8.763	2.875	0.0027 ***
PACE	-0.357	0.341	0.2978
3PAr	-0.553	0.377	0.145
eFG%	1.578	1.95	0.42
TOV%	-0.459	0.908	0.615
ORB%	0.244	0.752	0.746
FT/FGA	0.115	0.514	0.823

Opp. eFG%	1.028	2.587	0.692
Opp. TOV%	-0.793	1.317	0.548
DRB%	-0.636	0.839	0.449
Opp. FT/FGA	0.256	0.64	0.694

R^2 ajustado = 0.935

P -valor = < 2.2e-16

Figura 27: Modelo de regresión lineal múltiple con todas las variables.

Como explicábamos antes, cuando tenemos el modelo tomando todas las variables regresoras, va a haber una alta colinealidad entre varias de ellas. Una forma de ver esta **multicolinealidad** dentro del modelo es que teniendo un R^2 ajustado alto y un p -valor pequeño, sólo contamos con dos variables significativas, además de la constante β_0 (*OFFRTG* y *DEFRTG* son las únicas variables regresoras que tienen un p -valor significativo y que además coinciden con las variables que nos quedamos en el modelo después de la selección de variables). De la misma manera, también podemos verlo en el *Gráfico 14*, donde la correlación entre muchas de esas variables es muy alta. Para solucionar este problema aplicábamos dos **métodos de selección de variables** (*Backward* y *Fordward*), donde en ambos obteníamos el mismo modelo con los mismos regresores.

Siguiendo el resultado de esos dos métodos, obteníamos el modelo de regresión múltiple que se muestra en la *Figura 26* y explicado justo debajo de dicha figura. Obteníamos un modelo con tres variables regresoras cuya fórmula era la siguiente:

$$n^{\circ} \text{ de victorias} = \beta_0 + \beta_1 * \text{OFFRTG} + \beta_2 * \text{DEFRTG} + \beta_3 * \text{3PA}$$

$$\text{donde } \beta_0 = 40.9267, \beta_1 = 8.8728, \beta_2 = -7.7288 \text{ y } \beta_3 = -0.5698$$

Una de las variables resultantes de aplicar los métodos de selección es el *OFFRTG*. Dentro de las variables que se eliminan están los cuatro factores ofensivos, pertenecientes a los “*Four Factors*” de Dean Oliver, estas son el *eFG%*, *TOV%*, *ORB%* y *FT/FGA*. Si realizamos un modelo de regresión múltiple tomando el valor de *OFFRTG* de los equipos como **variable respuesta**, y estos 4 factores como variables regresoras nos queda:

Coefficientes:

TÉRMINO	COEFICIENTES	ERROR ESTÁNDAR	P-VALOR
Constante	27.682	1.359	< 2e-16 ***
eFG%	159.142	2.097	< 2e-16 ***
TOV%	-1.497	0.044	< 2e-16 ***
ORB%	0.487	0.019	< 2e-16 ***
FT/FGA	33.784	1.93	< 2e-16 ***

R^2 ajustado = 0.9801

P -valor = < 2.2e-16

Figura 28: Modelo de regresión múltiple con 4 factores ofensivos.

Vemos que la variabilidad explicada en este modelo es del 98.01% y además las 4 variables son significativas. Por lo tanto, estas van a explicar muy bien el *OFFRTG* (prácticamente un 100%), lo que apoya aquello que decíamos anteriormente sobre los “*Four Factors*” y es que estas cuatro variables explican muy bien cómo de bueno va a ser el ataque de los diferentes equipos.

Para el *DEFRTG* va a ocurrir exactamente lo mismo que con el anterior, pero en este caso tenemos que tener en consideración las variables de los “*Four Factors*” consideradas defensivas

(recordar que los “Four Factors” estaban formadas en realidad por 8 métricas distintas, cuatro ofensivas y cuatro defensivas). En este caso, los **cuatro factores defensivos** van a ser: *Opp. eFG%*, *Opp. TOV%*, *DRB%* y *Opp. FT/FGA*. El modelo resultante, con el *DEFRTG* como variable respuesta, será el siguiente:

Coefficientes:

TÉRMINOS	COEFICIENTES	ERROR ESTÁNDAR	P-VALOR
Constante	76.665	1.735	< 2e-16 ***
Opp. eFG%	158.367	1.531	< 2e-16 ***
Opp. TOV%	-1.359	0.028	< 2e-16 ***
DRB%	-0.493	0.017	< 2e-16 ***
Opp. FT/FGA	27.991	1.417	< 2e-16 ***

R^2 ajustado = 0.9891

P -valor = < 2.2e-16

Figura 29: Modelo de regresión múltiple con 4 factores defensivos.

En este caso ocurre lo mismo que cuando intentábamos explicar el *OFFRTG*, y es que tomando los 4 factores defensivos se consigue explicar casi al 100% la variabilidad de la variable respuesta *DEFRTG*.

Por ello y volviendo al modelo de la *Figura 26*, el número de victorias de los distintos equipos durante la liga regular, se puede estimar muy bien a partir de las variables *OFFRTG*, *DEFRTG* (que a su vez podemos estimar muy bien a partir de los “Four Factors”, tanto defensivos como ofensivos) y el *3PAr* de los equipos, aunque este último tendrá una menor significancia con respecto a los otros dos. Es decir, para medir el éxito de un equipo (midiéndolo según el número total de victorias que obtiene durante la liga regular) como cabía esperar se valora muy positivamente que este tenga un buen ataque (alto valor en el *OFFRTG*) y a su vez se valora negativamente que tenga una mala defensa (cuanto menor valor el *DEFRTG* mejor). También tiene influencia, aunque de manera poco significativa y negativa, el porcentaje en tiros de tres que un equipo realiza con respecto al número total de tiros de campo que realiza (*3PAr*).

Para apoyar lo anterior vamos a hacer un análisis de Inflación de Varianza (**VIF**):

VARIABLES	OFFRTG	DEFRTG	3PAr
VALOR DEL VIF	1.282	1.038	1.301

Figura 30: Valores del VIF de las variables presentes en el modelo “ataque-defensa” tras selección de variables.

Con los valores del *VIF* de la *Figura 30* estamos estudiando cuanto se “hincha” la varianza si eliminásemos esa variable de nuestro modelo. Deberíamos empezar a preocuparnos en valores que fuesen superiores a 4 o 5, por lo que tras haber realizado el proceso de selección de variables y quedarnos con estas tres últimas, desaparece la multicolinealidad que sí que teníamos anteriormente.

5.3. ESTUDIO DE LOS “FOUR FACTORS” DE DEAN OLIVER

Cuando hablábamos de los “Four Factors” de Dean Oliver, se trata de la filosofía que este plasmó en su libro publicado en 2002 “*Basketball on Paper*”, que consistía en medir la importancia de cuatro factores o estrategias en el éxito de un equipo.

Estos cuatro factores se correspondían con anotar efectivamente, es decir, con buenos porcentajes; cuidar el balón en ataque, lo que hace referencia a cometer el menor número de

pérdidas posibles; coger el mayor número posible de rebotes y por último intentar acudir a la línea de tiros libres lo máximo posible.

Para estos cuatro factores, Oliver propuso usar *eFG%*, *TOV%*, *ORB%* y *FT/FGA* respectivamente.

Sin embargo, también hay que tener en cuenta la defensa de un equipo cuando queremos medir su éxito, por lo que también habría que tener en cuenta estos mismos cuatro factores pero llevándolos al apartado defensivo.

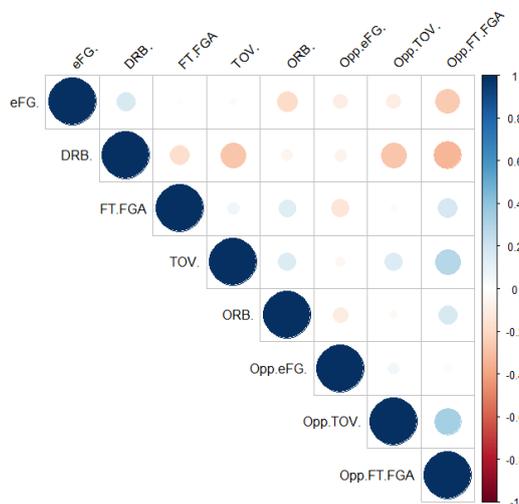
Para ello, se tiene en cuenta *Opp. eFG%*, *Opp. TOV%*, *DRB%* y *Opp. FT/FGA* y como consecución de todo esto, los famosos “Four Factors” serían en realidad ocho.

Además Dean Oliver en su libro propone un peso para cada uno de estos factores, lo que quiere decir que cada uno de estos factores va a tener un peso diferente a la hora de medir el éxito de un equipo. Los pesos que propone son los siguientes:

- 40% para la efectividad en el tiro (*eFG%* y *Opp. eFG%*).
- 25% para el cuidado del balón (*TOV%* y *Opp. TOV%*).
- 20% para los rebotes (*ORB%* y *DRB%*).
- 15% para acudir a la línea de tiros libres (*FT/FGA* y *Opp. FT/FGA*).

Debido a que estos “Four Factors” fueron publicados en un libro el cual fue publicado en el año 2002, vamos a realizar un estudio para ver si debido al cambio en la tendencia y en el juego que hemos explicado en este trabajo y que se viene produciendo durante las últimas décadas, también va a afectar al peso o importancia que se otorga a cada uno de estos a la hora de predecir cómo de bien lo ha hecho un determinado equipo.

Primero se mostrará la correlación entre estas variables, para poder ver su linealidad.



Gráfica 15: Correlación de Pearson entre variables del modelo “Four Factors”.

La correlación entre estas variables no es considerablemente alta entre ningún par de ellas, por lo que no deberíamos encontrar sobreajuste en este modelo.

A continuación, vamos a realizar un modelo de regresión lineal múltiple, donde vamos a tomar como variable respuesta el número de victorias de cada uno de los equipos durante la liga regular y como variables regresoras los valores que obtuvieron estos mismos equipos en cada uno de los ocho factores que tenemos en cuenta en los “Four Factors” al final de cada temporada. Para ello vamos a tener en cuenta los mismos datos que en los ejercicios anteriores,

es decir, tomamos los 150 equipos que han participado en las cinco últimas temporadas de la NBA.

Coefficientes:

TÉRMINOS	COEFICIENTES	ERROR ESTÁNDAR	P-VALOR
Constante	-53.92	21.9	0.015 ***
eFG%	394.87	15.51	< 2e-16 ***
TOV%	-372.12	33.55	< 2e-16 ***
ORB%	127.92	13.36	< 2e-16 ***
FT/FGA	73.53	14.15	7.02e-07 ***
Opp. eFG%	-382.98	16.11	< 2e-16 ***
Opp. TOV%	291.03	29.26	< 2e-16 ***
DRB%	89.69	18.15	2.15e-06 ***
Opp. FT/FGA	-70.55	15.54	1.2e-05 ***

R^2 ajustado = 0.9247

P -valor = < 2.2e-16

Figura 31: Modelo de regresión múltiple para los "Four Factors".

Por lo que la fórmula para este modelo va a ser la siguiente:

$$n^{\circ} \text{ de victorias} = \beta_0 + \beta_1 * eFG\% + \beta_2 * TOV\% + \beta_3 * ORB\% + \beta_4 * FT/FGA + \beta_5 * Opp. eFG\% + \beta_6 * Opp. TOV\% + \beta_7 * DRB\% + \beta_8 * Opp. FT/FGA$$

$$\text{donde: } \beta_0 = -53.92, \beta_1 = 394.87, \beta_2 = -372.12, \beta_3 = 127.92, \beta_4 = 73.53, \beta_5 = -382.98, \beta_6 = 291.03, \beta_7 = 89.69 \text{ y } \beta_8 = -70.55.$$

A partir de la *Figura 31* vemos que es un buen modelo a la hora de predecir el número de victorias que va a obtener cada equipo. Se observa que todas las variables regresoras son significativas, además el valor del R^2 ajustado es elevado (cercano al 93%).

En relación a los coeficientes que obtenemos (los valores de los distintos β) vamos a ver el peso que se otorga a cada una de las variables regresoras:

- Para eFG%: media = $\frac{394.87 + 382.98}{2} = 388.925$
- Para el TOV%: media = $\frac{372.12 + 291.03}{2} = 331.575$
- Para los rebotes: media = $\frac{127.92 + 89.69}{2} = 108.805$
- Para el FTr (FT/FGA): media = $\frac{73.53 + 70.55}{2} = 72.04$

La suma total de estas medias es = 901.345.

Por lo tanto, los pesos que se otorgan a cada uno de los factores en este modelo son:

- 43.15% para la efectividad en el tiro (eFG% y Opp. eFG%).
- 36.79% para el cuidado del balón (TOV% y Opp. TOV%).
- 12.07% para el total de rebotes (ORB% y DRB%).
- 8% para acudir a la línea de tiros libres (FT/FGA y Opp. FT/FGA).

Gracias a la identificación de estos pesos podemos compararlos con los que Dean Oliver daba en su modelo (40/25/20/5). El porcentaje en el tiro continua siendo el factor diferencial a la hora de medir el éxito de un equipo, tanto en ataque como a provocar bajos porcentajes en los lanzamientos del rival. La importancia de las pérdidas de balón de un equipo se ve incrementada en casi un 50% respecto a los pesos iniciales, mientras que para los rebotes y para el hecho de acudir a la línea de los tiros libres disminuye casi un 50%, por lo que en el modelo inicial, la

importancia de estos dos últimos valores estaba inflada si tenemos en cuenta estas últimas cinco temporadas. También nos llama la atención que el rebote defensivo está valorado menos que el rebote ofensivo.

Como resumen, vemos que un claro cambio en la tendencia de juego que se ha producido los últimos años provoca cambios en el peso que vamos a dar a cada uno de los factores. Ya explicábamos anteriormente (cuando hablamos en la parte de explicación de las métricas del tercer factor de los "Four Factors") que en los últimos años a pesar del aumento del *3PAr*, se produce una clara disminución del *ORB%*, ya que los equipos tienden a realizar el balance defensivo antes incluso que cargar el rebote ofensivo. La efectividad en el tiro sigue siendo el factor diferencial, al que, con respecto a las últimas temporadas se le suma el cuidado de la bola. Otros como el *FT/FGA* tienen poca influencia.

Destacamos también que en todos los factores, tendrán mayor importancia aquellos que son ofensivos con respecto a los defensivos, por lo que normalmente los mejores equipos destacan principalmente por su ataque respecto a su defensa (aunque pueden existir excepciones).

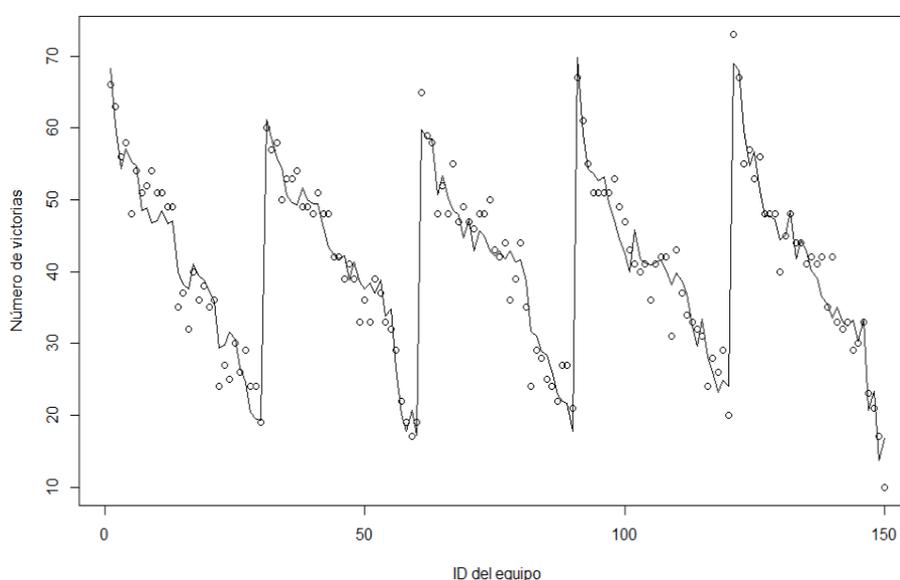


Gráfico 15: Comparación entre victorias reales y predichas para modelo "Four Factors".

En el *Grafico 15* encontramos la comparación entre el número real de victorias que obtuvieron los 150 equipos pertenecientes a las últimas cinco temporadas y el número predicho a partir de las variables correspondientes a la métrica "Four Factors". Estarán representadas mediante puntos las victorias reales y mediante la línea las victorias predichas.

Los equipos que presentan una peor predicción, siendo esta una diferencia superior a 7 victorias entre las victorias reales y las predichas son los siguientes: Mavericks 19/20, Nuggets 19/20, Cavaliers 17/18, Mavericks 17/18, Wolves 16/17 y Grizzlies 15/16 (cuyos id's asignados eran 5, 9, 74, 82, 109 y 140).

Encontramos un total de 112 equipos cuya diferencia es igual o menor que cuatro victorias. Esto significa que para casi un 75% de los equipos tendremos muy buenas predicciones utilizando el modelo con las variables presentes en los "Four Factors" de Dean Oliver.

A pesar de que este modelo no es perfecto, es muy ajustado como se ha demostrado en los hechos anteriores. A la hora de predecir el número de victorias de un equipo usando los “*Four Factors*”, se considerará mucho más eficiente que si cogiésemos cualquier otra estadística de equipo (statathlon.com/ “*Four Factors in basketball success*”).

Ahora, tomamos el modelo de la *Figura 26*, que estaba basado en la eficiencia ofensiva y defensiva (usa *OFFRTG* y *DEFRTG* como variables regresoras y eliminamos el *3PA* del modelo, ya que este ejercía muy poca influencia).

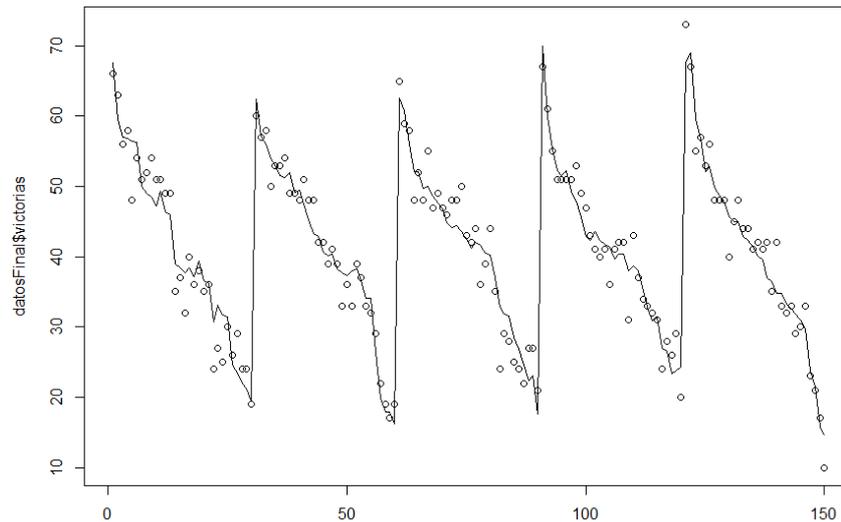


Gráfico 16: Comparación entre victorias reales y predichas para el modelo del OFFRTG y DEFRTG.

En este modelo llegamos a encontrar casos donde la predicción falla en un número superior a ocho victorias en un par de casos (Mavericks 19/20 y Mavericks 17/18), lo que supone ya diferencias considerablemente grandes. A pesar de esto, los resultados que se obtienen son muy parecidos al modelo de “*Four Factors*” anterior.

En resumen, tras los resultados obtenidos hemos podido comprobar que los ocho factores que se encuadran dentro de los “*Four Factors*”, van a estar altamente asociados con el éxito de un determinado equipo. Los pesos que vamos a dar a cada una de estas variables han ido y seguirán cambiando a lo largo de los años, ya que la tendencia y el estilo de juego de los equipos de la liga también lo hace. Gracias a este modelo podemos obtener muy buenos resultados, ya que por ejemplo encontramos 64 equipos los cuales la diferencia en ambos tipos de victorias es inferior a dos. En cuanto al modelo que tomaba como variables el OFFRTG y DEFRTG, obtenemos resultados muy parecidos en las predicciones, así que como era de suponer, también supone un muy buen ajuste de los datos.

6. K-MEANS CLUSTERING PARA LOS MODELOS DE REGRESIÓN

Vamos a realizar este procedimiento de agrupamiento de datos para los modelos que estudiábamos en el apartado anterior y así compararlos con el que realizábamos anteriormente en el *Apartado 3*. Por ello, vamos a tener en cuenta tanto el modelo que vamos a denominar “*ataque-defensa*” y que está formado por las variables *OFFRTG* y *DEFRTG* y el modelo de los “*Four Factors*”, conformado por las 8 variables que componen esta métrica.

6.1. K-MEANS SOBRE MODELO “ATAQUE-DEFENSA”

Únicamente va a tener en cuenta dos variables, el ratio ofensivo y el defensivo (*OFFRTG* y *DEFRTG* respectivamente).

Recordar que los id's de los equipos se corresponde con la posición que ocupan estos en el archivo Excel, donde se encuentran ordenados desde la temporada actual (2019/20) hasta la primera temporada que tomamos en el estudio (2015/16).

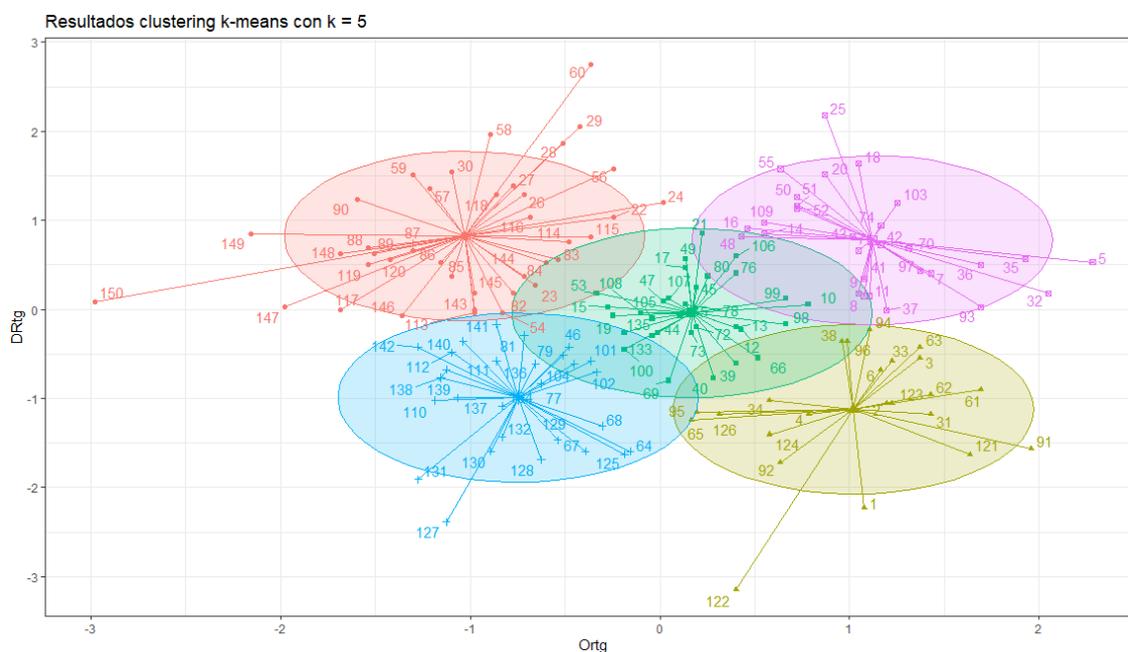


Gráfico 17: Resultado k-means clustering con k = 5 sobre modelo *ataque-defensa*.

La situación idónea para cualquier equipo dentro de este gráfico debería ser estar situados en la parte inferior derecha del mismo, lo que quiere decir que tendrían un *OFFRTG* alto y dejarían al equipo rival en un valor del mismo bajo, o lo que es lo mismo, dicho equipo tendría un *DEFRTG* bajo.

Vamos a explicar cada uno de los centroides y los equipos que se asignan a cada uno de ellos:

- **Centroide 1** (color rojo): Dentro de este grupo se encuadran los peores equipos dentro de estas 5 temporadas. Como vemos en el *Gráfico 17*, estos equipos se encuentran situados en la parte superior izquierda, lo que significa que además de presentar valores bajos en el *OFFRTG*, su valor de *DEFRTG* es elevado. Ningún equipo dentro de este grupo ha conseguido la clasificación para los *Play-offs* dentro de sus respectivas temporadas, algunos de los equipos que encontramos son: Warriors 19/20, Cavaliers 18/19, Suns

- 17/18, Nets 16/17 y 76ers 15/16 (peores equipos estadísticamente en cada temporada y cuyos id's asociados son 30, 60, 90, 120 y 150).
- **Centroide 3** (color verde): En este grupo encontramos equipos que a pesar de no destacar estadísticamente en ninguna de las dos variables sobre el resto, son buenos equipos donde la mayoría de ellos consiguieron plaza en los *Play-offs* en sus respectivas temporadas. Sin embargo, ningún equipo dentro de este grupo ha conseguido ser campeón o finalista de la NBA en estas 5 temporadas. A pesar de esto y como comentábamos, encontramos muy buenos equipos tales como: 76ers 19/20, Thunder 18/19, Blazers 17/18, Celtics 16/17 y Rockets 15/16 (con id's: 12, 39, 69, 98 y 135).
 - **Centroide 2** (color oro): Se trata del grupo que tiene a los mejores equipos y es que incluye aquellos con muy buena defensa y muy buen ataque. A diferencia de lo que ocurría cuando aplicábamos este método sobre el modelo que contenía a todas las variables, los Spurs 15/16, que se trataba de la mejor defensa de estas 5 temporadas y una de las mejores de la historia, sí que estará clasificado dentro de este grupo de los mejores equipos, y es que, a pesar de ser un equipo defensivo, su ataque también se encontraba entre los mejores. También los Cavaliers 15/16, campeones esa temporada, ahora estarán clasificados en este grupo de equipos punteros. A este centroide además vamos a asignar: Bucks 19/20, Raptors 18/19, Warriors 17/18, Clippers 16/17 y Thunder 15/16 (con id's: 1, 33, 63, 94 y 123).
 - **Centroide 5** (color morado): Aquí encontramos los mejores ataques en cuanto al valor del *OFFRTG* se refiere. Otra característica es que estos equipos también van a tener un valor del *DEFRTG* elevado. Por lo general, encontramos equipos de un nivel muy alto (mínimo *Play-offs* muchos de ellos) y que al tener tan buenos ataques no necesitan valores más bajos en el *DEFRTG*; ejemplos muy claros de esto son por ejemplo los Warriors 18/19, Cavaliers 16/17 y 17/18 (todos ellos subcampeones de la NBA, cuyos *id's* son: 32, 97 y 74). Además encontramos equipos clasificados en este grupo pero que no llegan al nivel de los mencionados anteriormente y es que, aunque su *OFFRTG* es alto, su *DEFRTG* lo es todavía más. Algun ejemplo de estos son los Suns o Wizards de la temporada 19/20 o los Wolves 17/18 (con id's: 16, 25 y 70).
 - **Centroide 4** (color azul): En contraposición al centroide 5, en este caso los equipos se caracterizan por ser eminentemente defensivos, es decir, obtienen valores de *DEFRTG* bajos y valores de *OFFRTG* medios-bajos. Ahora también encontramos equipos que a pesar de que tienen un *OFFRTG* bajo, su *DEFRTG* es todavía menor, por lo que son muy buenos equipos en sus respectivas temporadas. Ejemplos muy claros del caso anterior son por ejemplo los Heat 18/19, los Jazz o Celtics de la temporada 17/18 o los Clippers 15/16 (con id's: 46, 64, 67 y 125).

Si comparamos este método *clustering k-means* cuando utilizábamos el modelo con todas las variables (en el *Apartado 3*) con el modelo que llamábamos “ataque-defensa” (únicamente tomando el *OFFRTG* y el *DEFRTG* como variables), llegamos a la conclusión de que en el segundo caso vemos las cosas mucho más claras. Y es que, a pesar de que ocurre algo semejante a lo que ocurría en el primer caso, que es que este método lo que mejor va a diferenciar es a los mejores equipos de los peores; en este segundo caso el resto de grupos están mucho más definidos y diferenciados entre ellos.

6.2. K-MEANS SOBRE MODELO DE "FOUR FACTORS"

En este apartado volvemos a realizar el mismo procedimiento del caso anterior pero tomando el modelo de los "Four Factors" es decir, tenemos 8 variables, 4 ofensivas (eFG%, TOV%, ORB% y FT/FGA) y 4 defensivas (Opp. eFG%, Opp. TOV%, DRB% y Opp. FT/FGA).

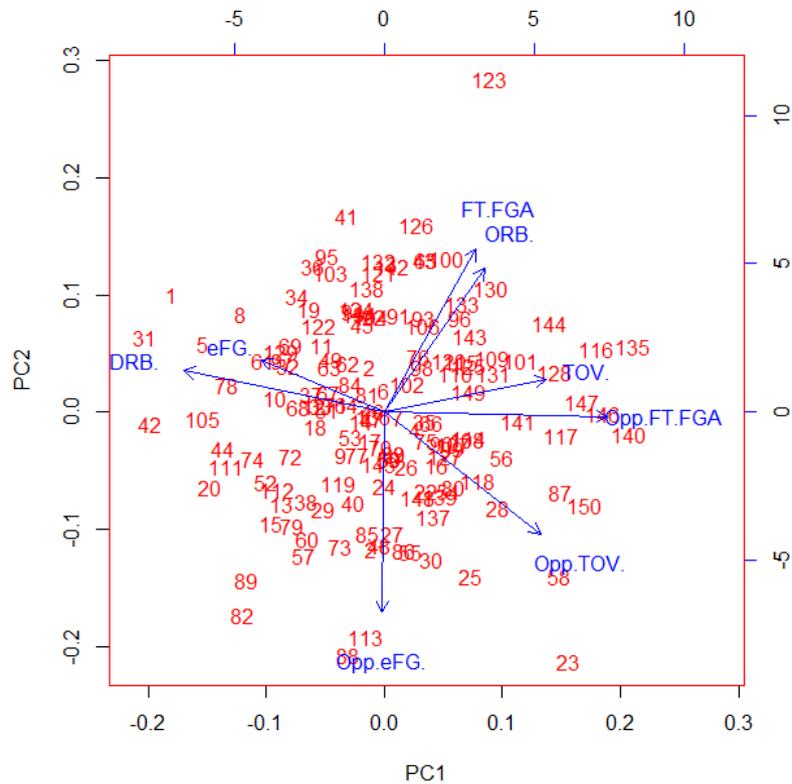


Gráfico 18: Análisis de componentes principales (PCA) del modelo "Four Factors".

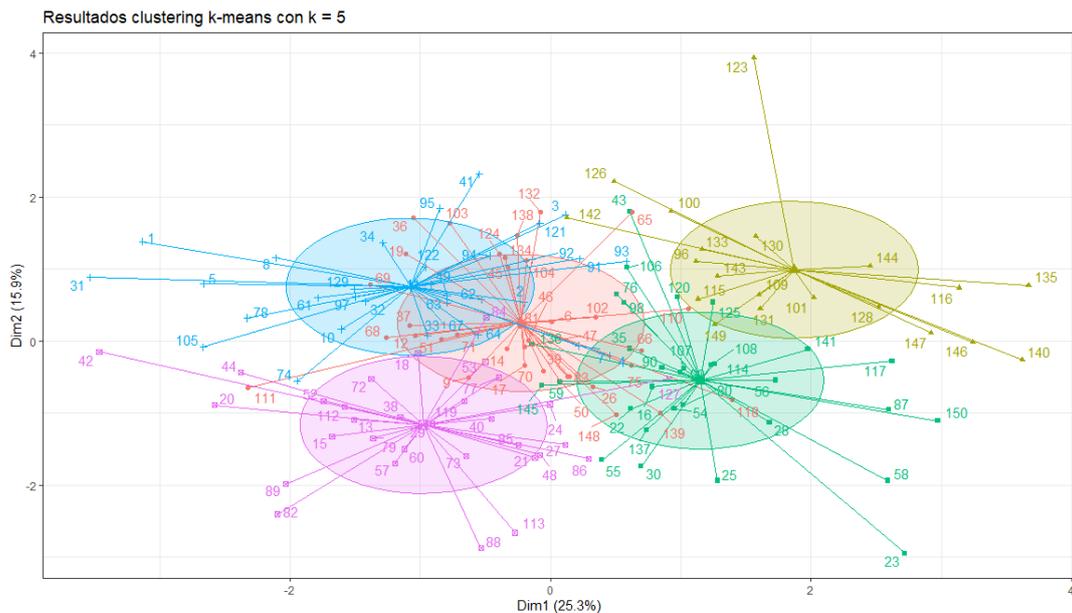


Gráfico 19: Resultado k-means clustering con k = 5 sobre modelo "Four Factors".

En la *Gráfico 18* vemos que los mejores equipos tienden a tener altos porcentajes de tiro y cometer pocas pérdidas, los cuales son los factores que más influenciarán positivamente en el éxito de un equipo. Además también se clasifican en el grupo de grandes equipos aquellos que destacan por su defensa, es decir, que dejan al equipo contrario en bajos porcentajes de tiro.

Con este modelo podemos decir que clasifica mejor que el modelo cuando utilizábamos todas las variables con las que contábamos, pero no nos da resultados tan claros como cuando tomábamos el modelo de *“ataque-defensa”*; a pesar de esto, se asemejan bastante, ya que distingue a los mejores equipos de los peores muy bien, y luego al igual que pasaba cuando tomábamos el modelo de ataque-defensa, distinguimos a los mejores equipos (buenos ofensivamente y defensivamente) de otros buenos equipos pero que destacaban más en las facetas ofensivas o en las defensivas.

Como conclusión de este apartado y apoyando lo visto en el anterior (cuando realizábamos modelos de regresión lineal múltiple) los *“Four Factors”* es una muy buena métrica para medir cómo de bueno es el ataque y la defensa de un equipo y por lo tanto, el éxito que va a tener ese equipo. Esta métrica nos permite saber qué puede pasar durante la posesión de un equipo y a partir de ahí medir el impacto de todas las posesiones en el número de victorias de los equipos.

Encontrar el balance entre todas estas variables debería ser el objetivo de los equipos para así conseguir victorias en la NBA, sabiendo que el hecho de anotar y el control de la bola son los factores más influyentes. Sin embargo, no solo debemos tener en cuenta las dos anteriores, si no que tomando todos los factores nos permitirá ver cuáles son las debilidades de un equipo determinado que le impiden conseguir victorias con asiduidad y mejorar en esas facetas en las que es peor respecto al resto de la liga. Otra conclusión que sacamos de todo esto y que nos puede sorprender es la gran importancia que tiene el control de la bola, es decir, perder el menor número de balones posibles, además de forzar al rival al mayor número posible de pérdidas de balón.

7. PREDICCIÓN DE PRESENCIA DE LOS EQUIPOS EN PLAY-OFFS UTILIZANDO REGRESIÓN LOGÍSTICA Y RANDOM FOREST

El objetivo de este apartado es comprobar cómo de bien se puede predecir y clasificar la presencia de los distintos equipos en los *Play-Offs* a partir de los modelos que hemos tomado como esenciales, tanto el de “ataque-defensa”, como el de los “Four Factors”. En este caso la variable respuesta será la “presencia del equipo en *Play-Offs*” y esta va a ser categórica, es decir, tomará dos valores, 1 en el caso de que el equipo se clasifique a la postemporada y 0 en el caso contrario.

Para ello, vamos a usar los métodos de clasificación de **regresión logística** y el de **random forest**.

7.1. EXPLICACIÓN TEÓRICA REGRESIÓN LOGÍSTICA

Se trata de un problema cuantitativo que se utiliza en problemas donde la variable respuesta va a tomar valores dentro de un conjunto finito en función de variables independientes o predictoras. Estos predictores pueden ser tanto continuos como categóricos.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_ix_i$$

$$\text{logit}(Y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_ix_i$$

A continuación el valor de la probabilidad de Y se puede obtener con la inversa del logaritmo natural:

$$p(Y) = \frac{e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_ix_i}}{1 + e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_ix_i}}$$

Así obtenemos la probabilidad de que Y sea cada uno de los valores que puede adoptar dentro de su conjunto finito de datos.

7.2. REGRESIÓN LOGÍSTICA SOBRE EL MODELO “ATAQUE-DEFENSA”:

Tomando el modelo de “ataque-defensa”, eso es, aquel conformado únicamente por las variables regresoras *OFFRTG* y *DEFRTG*, vamos a llevar a cabo una regresión logística.

Con la función de *R glm* se realiza automáticamente el procedimiento de regresión logística, tomando como variable respuesta la variable categórica denominada *PLAYOFFS*, la cual toma dos únicos valores, 0 si el equipo no se clasificó para *Play-Offs* su temporada y 1 si sí que lo hizo.

Coefficientes:

TÉRMINOS	COEFICIENTES	ERROR ESTÁNDAR	P-VALOR
Constante	17.3312	16.058	0.28
OFFRTG	1.1526	0.241	1.77e-06 ***
DEFRTG	-1.3075	0.261	5.64e-07 ***

AIC = 72.458

Figura 32: Regresión logística sobre modelo “ataque-defensa”.

Ambos p-valores de estos regresores, tomando una confianza del 95% van a ser significativos (valor inferior a 0.05). Esto quiere decir que ambos regresores van a ser buenos predictores a la

hora de determinar si un equipo, en cuanto a estas estadísticas se refiere, va a conseguir clasificarse para *Play-Offs*.

La probabilidad de esta clasificación de cada uno de los equipos del conjunto de datos es:

$$\text{logit}(\text{clasificación a Play - Offs}) = 17.3312 + 1.1516 * \text{OFFRTG} - 1.3075 * \text{DEFRTG}$$

$$P(\text{clasificación a Play - Offs}) = \frac{e^{17.3312 + 1.1516 * \text{OFFRTG} - 1.3075 * \text{DEFRTG}}}{1 + e^{17.3312 + 1.1516 * \text{OFFRTG} - 1.3075 * \text{DEFRTG}}}$$

En este estudio se va a utilizar un *threshold* de 0.5. Esto consiste en que si la probabilidad predicha de que un equipo participante en la últimas 5 temporadas de la NBA se clasifique a *Play-Offs* sea superior a 0.5, se asigna al nivel 1 y se hará lo propio en el caso contrario, es decir, si dicha probabilidad predicha es inferior a 0.5 se asignará al nivel 0.

OBSERVACIONES/ PREDICCIONES	0	1
0	60	10
1	7	73

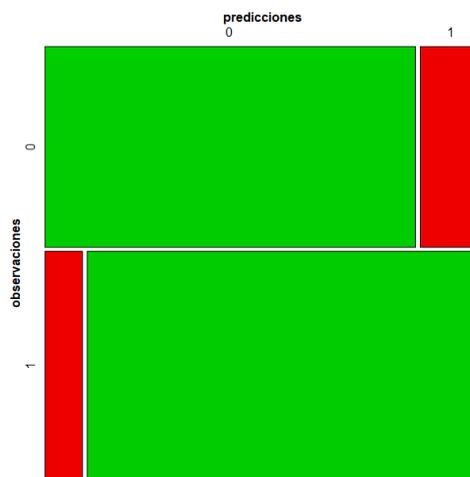


Figura 33: Matriz de confusión y su mosaico del *threshold* de la regresión logística sobre el modelo de “ataque-defensa”.

A partir de este *threshold* de 0.5 obtenemos la clasificación de la *Figura 33*. De los 70 equipos que no lograron una plaza en la postemporada de sus respectivas temporadas, el modelo clasifica correctamente a 60 de ellos, lo que es un porcentaje muy elevado. En cuanto a los equipos que sí que consiguieron dicha plaza, del total de 80 este modelo clasificará correctamente a 73, lo que supone un porcentaje incluso superior al anterior (91.25% de bien clasificados contra un 85.71%). Generalizando, el modelo es capaz de clasificar correctamente $\frac{60 + 73}{150} = 88.66\%$ de las observaciones.

No se debe olvidar que los errores los cuales han sido calculados son de entrenamiento, por lo que no van a ser generalizables a nuevas observaciones. Para que esta estimación que obtenemos llegue a ser más realista, hay que calcular el error del test.

Para mejorar la calidad de predicción de estos modelos, vamos a utilizar la **validación cruzada k-fold**. Esta consiste en tomar los datos originales y a partir de ellos crear dos conjuntos separados, uno para test y otro para entrenamiento. Para este proceso únicamente necesitamos un valor de *k*, que es el número de subconjuntos en los que se va a dividir el conjunto de

entrenamiento. Al momento de realizar dicho entrenamiento, se va a tomar cada subconjunto k como conjunto de prueba del modelo, mientras que el resto se tomará como conjunto de entrenamiento.

Repetimos este proceso un total de k veces, y en cada iteración se va a ir seleccionando un conjunto de prueba diferente. Cuando se finalizan las iteraciones, se va a calcular la precisión y el error para cada uno de los modelos que se han generado y se halla el promedio de todas ellas, para así conseguir la precisión y error final.

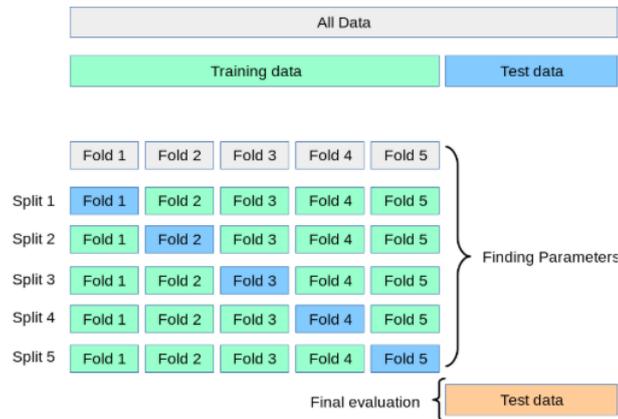


Figura 34: Esquema del función de la validación cruzada k-fold (scikit-learn.org).

A partir de la función de R *createFolds* vamos a llevar a cabo la aplicación de la validación cruzada, y finalmente lo vamos a entrenar cada modelo obtenido, tomando $k = 10$ subconjuntos.

Realizando este procedimiento de ***k-fold validación cruzada*** sobre la regresión logística del modelo de “ataque-defensa” obtenemos una precisión final del 84.16%, lo que sigue siendo un valor muy elevado, por lo que las variables *OFFRTG* y *DEFRTG* serán muy buenas clasificadoras a la hora de distinguir a los equipos los cuales se han clasificado o no a los *Play-Offs*.

7.3. REGRESIÓN LOGÍSTICA SOBRE EL MODELO “FOUR FACTORS”:

Repetimos el proceso del apartado anterior pero en este caso tomando el modelo de los “*Four Factors*”.

Coefficientes:

TÉRMINOS	COEFICIENTES	ERROR ESTÁNDAR	P-VALOR
Constante	-26.39	25.82	0.307
eFG%	187.79	39.47	1.96e-06 ***
TOV%	-203.56	58.36	0.00049 ***
ORB%	53.44	19.51	0.0065 ***
FT/FGA	56.71	18.51	0.0022 ***
Opp. eFG%	-230.03	46.57	7.86e-07 ***
Opp. TOV%	176.71	46.64	0.00015 ***
DRB%	41.09	22.94	0.073
Opp. FT/FGA	-16.71	17.43	0.337

AIC = 81.459

Figura 35: Regresión logística sobre modelo “Four Factors”.

$$\begin{aligned} \text{logit}(\text{clasificación a Play - Offs}) &= -26.39 + 187.79 * eFG\% - 203.56 * TOV\% + 53.44 * ORB\% + 56.71 \\ &* FT/FGA - 230.03 * Opp.eFG\% + 176.71 * Opp.TOV\% + 41.09 * DRB\% \\ &- 16.71 * Opp.FT/FGA \end{aligned}$$

En este caso y tomando también una confianza del 95%, encontramos que la mayoría de variables son significativas. Las únicas dos variables que no cumplen lo anterior son *DRB%* y *Opp.FT/FGA*, de las que ya exponíamos en otros apartados de este proyecto y basándonos en lo visto en el libro “*Basketball On Paper*” de Dean Oliver que se trataba de las dos que menos peso iban a tener asignado a la hora de predecir el éxito de un equipo.

OBSERVACIONES/ PREDICCIONES	0	1
0	63	7
1	7	73

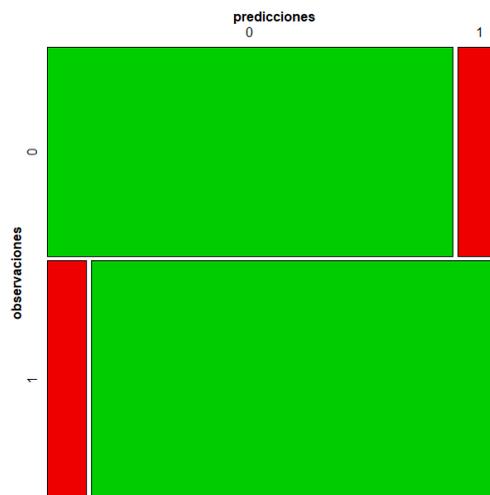


Figura 36: Matriz de confusión y su mosaico de la regresión logística sobre el modelo “Four Factors”.

A partir de nuevo de haber realizado un threshold de 0.5, vemos que este modelo clasifica de igual manera en cuanto a la tasa de acierto a los equipos que se clasifican y a los que no además, generalizando a todos los datos, lo hace ligeramente mejor que el modelo del apartado anterior (91.25% de acierto en los equipos clasificados y 90% en los equipos no clasificados). Este modelo es capaz de clasificar correctamente el 90.66% de los datos de entrenamiento, lo que supone un mayor valor que en el caso del modelo de “ataque-defensa”.

Aplicando validación cruzada k-fold obtenemos una precisión final del 86.66%, lo que también supone un valor superior al ejemplo anterior.

7.4. EXPLICACIÓN TEÓRICA ALGORITMO RANDOM FOREST:

También conocido como “Bosques Aleatorios”, se trata de una combinación de árboles predictores tal que cada uno de los árboles va a depender de un vector aleatorio que ha sido probado independientemente y que tiene la misma distribución en cada uno de estos. Cada uno de los árboles que lo conforman se van a construir de la siguiente forma:

- Un conjunto de **entrenamiento con N casos**. Una muestra de esos N casos se va a tomar con reemplazamiento, que se tratará del conjunto de entrenamiento para construir el árbol i .
- Existiendo M variables de entrada, se va a especificar un número $m < M$ tal que para cada nodo, m variables se van a seleccionar aleatoriamente de M . El valor m se va a mantener constante durante toda la generación del bosque.
- Cada árbol crece hasta su máxima extensión y no se realiza proceso de poda.

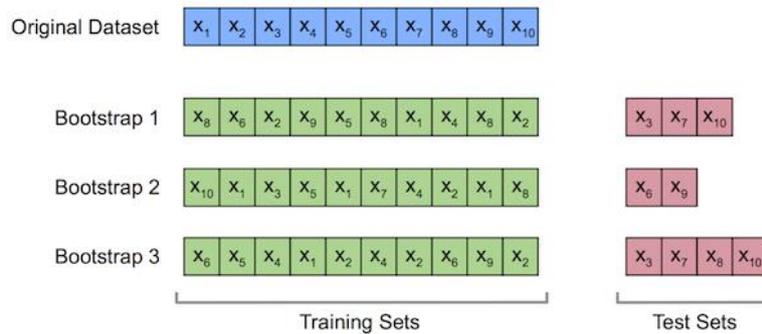


Figura 37: Organización del ensamblador Random Forest (bookdown.org, “Ensambladores Random Forest”).

El proceso en el que se realiza el muestreo de los datos se va a denominar *bootstrap*. Va a haber un tercio de los datos que no se van a usar para entrenamiento, por lo que se van a usar para test, y este es el conjunto que se va a denominar *out of bag (OOB) samples* y el error estimado en estos *OOB samples* se conoce como *OOB error*.

Para estas predicciones conocidas como *OOB* y que se quedaron fuera del árbol, se hará una predicción y se calculará el error de la predicción. Esto se realiza en cada iteración para calcular el error nombrado anteriormente, el *OOB error*.

7.5. RANDOM FOREST SOBRE MODELO “ATAQUE-DEFENSA”:

Recordar que el modelo “ataque-defensa” es aquel con el que se pretendía predecir el número total de victorias de un equipo utilizando únicamente dos variables (el *OFFRTG* y el *DEFRTG*). En este caso como variable respuesta contamos con una variable categórica “*play*” la cual nos indica la clasificación o no de un determinado equipo para la postemporada (“1 = equipo clasificado para los *Play-Offs*” y “0 = equipo no clasificado para los *Play-Offs*”).

Para llevar a cabo este procedimiento tenemos que dividir los datos en conjunto de entrenamiento y en un conjunto test (*out of bag samples*). Por normal general, se suele añadir un 70% de los datos al primer grupo y el 30% de los resultantes al segundo. Sin embargo, vamos a coger como conjunto entrenamiento los datos correspondientes a los equipos que participaron en las cuatro temporadas anteriores (desde la 2015/16 a la 2018/19) y como conjunto test los correspondientes a la temporada en curso en este momento en la NBA (2019/20). De esta manera el reparto quedará en un 80% y 20%.

El *out of bag error (OOB error)* es un método el cual nos sirve como medida para la predicción del error de algoritmos como el *Random Forest*. Este error se corresponde con la predicción media del error en cada muestra de entrenamiento X_i , pero usando únicamente árboles los cuales no tenían X_i en su muestra *Bootstrap*.

Mediante la función *randomForest* de R se lleva a cabo el procedimiento del algoritmo *Random Forest*.

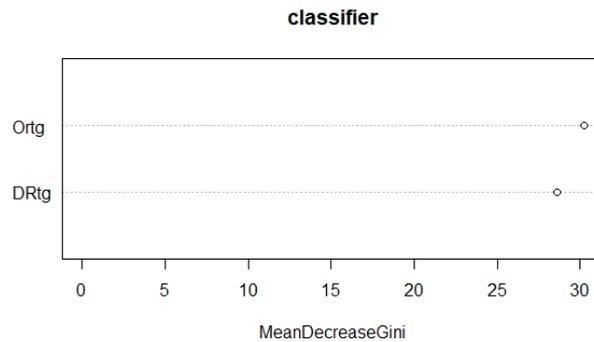


Gráfico 20: Importancia de cada una de las variables en el modelo “ataque-defensa” para el algoritmo Random Forest.

El coeficiente *Mean Decrease Gini* es una medida de la contribución de cada una de las variables a la homogeneidad de cada uno de los nodos y hojas del *Random Forest* resultante. Cuanto mayor sea dicho valor, mayor va a ser la importancia de la esa variable. Por lo tanto, en el *Gráfico 20* podemos observar que el *OFFRTG* va a ser ligeramente más importante.

Matriz de confusión:

OBSERVACIONES/ PREDICCIONES	0	1	ERROR EN CADA CLASE
0	45	11	0.196
1	9	55	0.141

Número de árboles = 500

OOB error estimado = 16.67%

Figura 38: Algoritmo Random Forest sobre modelo “ataque-defensa”.

Por defecto, al aplicar esta función en R el número de árboles de decisión totales que conformarán el “bosque” será de un total de 500. Como se trata de un problema de clasificación, utilizaremos una matriz de confusión para evaluar el funcionamiento de nuestro modelo.

El valor del *OOB error* en este caso es 16.67%, que es el error que se obtiene tras añadir cada nuevo caso. Según la matriz de confusión, de los 120 equipos que conformaban el conjunto de entrenamiento, 100 de ellos estarán clasificados correctamente, mientras que para el resto, 11 se corresponden como falsos equipos de *Play-Offs* (equipos que se clasifican como equipos de *Play-Offs* pero en realidad no lo son) y 9 como falsos equipos no clasificados para *Play-Offs* (es decir, equipos que sí que lograron acceder a los *Play-Offs*, pero que en el modelo se clasifican como que no).

Ahora, vamos a usar el modelo anterior para predecir en nuestro conjunto test si los equipos en dicho conjunto (participantes en la temporada 2019/20) son equipos de *Play-Offs* o no lo son. Para llevar a cabo dicha validación en nuestro conjunto test, vamos a utilizar la función *pred* de R.

VALORES REALES/VALORES PREDICHOS	0	1
0	12	2
1	3	13

Figura 39: Validación del modelo Random Forest de “ataque-defensa” sobre el conjunto test (temporada 2019/20).

En la *Figura 39* tenemos la matriz de confusión que nos queda una vez validado el modelo. Vamos a ver dos valores asociados, como son la sensibilidad y la especificidad los cuales nos van a mostrar el funcionamiento del modelo. Lo ideal sería que una vez validado el modelo, en la matriz de confusión sólo obtuviésemos verdaderos positivos y verdaderos negativos.

La sensibilidad nos va a indicar la capacidad de nuestro estimador para clasificar como equipos de *Play-Offs* aquellos equipos que realmente sí que consiguieron una plaza en la postemporada, mientras que la especificidad nos dará la capacidad para clasificar los equipos que no consiguieron llegar a *Play-Offs*, como equipos que se quedaron fuera de dicha postemporada.

Por lo tanto obtenemos los siguientes valores:

$$sensibilidad = \frac{VP}{VP + FN} = \frac{13}{13 + 3} = 0.8125 = 81.25\%$$

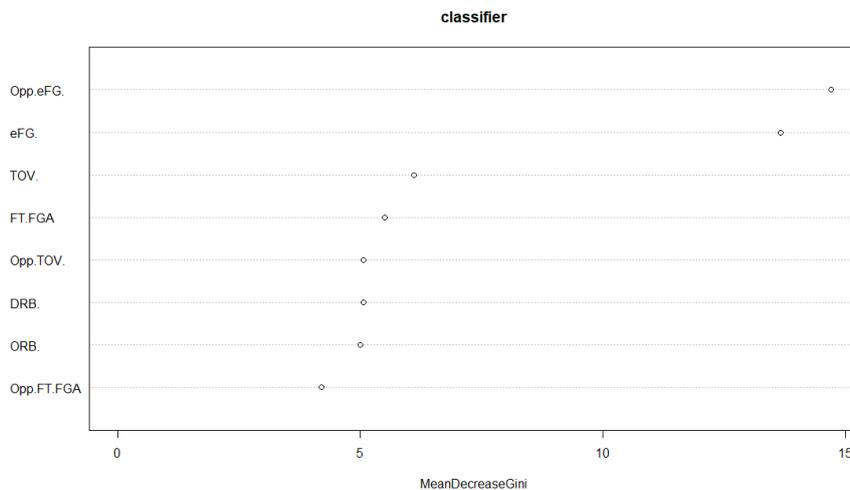
$$especificidad = \frac{VN}{VN + FP} = 0.8571 = 85.71\%$$

El modelo va a clasificar ligeramente mejor los equipos que no consiguieron una plaza en los *Play-Offs* con respecto a los que sí que lo hicieron, pero aún así, el valor tanto para la sensibilidad como la especificidad es elevado, por lo que usando el modelo de “ataque-defensa” sí que va a ser un buen predictor a la hora de determinar la presencia de un equipo en la postemporada.

Si realizamos el procedimiento de *validación cruzada k-fold*, la precisión que obtenemos es de 83.33%, lo que supone una disminución con respecto a cuando aplicábamos este mismo procedimiento sobre este mismo modelo de “ataque-defensa” utilizando regresión logística (84.16%).

7.6. RANDOM FOREST SOBRE MODELO “FOUR FACTORS”:

A continuación vamos a realizar el mismo proceso que realizábamos en el subapartado anterior, pero en este caso tomando el modelo de los “*Four Factors*”, es decir aquel formado por las variables: *eFG%*, *Opp. eFG%*, *TOV%*, *Opp. TOV%*, *ORB%*, *DRB%*, *FT/FGA* y *Opp. FT/FGA*.



VARIABLES	VALOR MEAN DECREASE GINI
eFG%	13.672

TOV%	6.109
ORB%	4.995
FT/FGA	5.503
Opp. eFG%	14.708
Opp. TOV%	5.065
DRB%	5.058
Opp. FT/FGA	4.201

Gráfico 20: Importancia de cada una de las variables en el modelo “Four Factors” para el algoritmo Random Forest.

En cuanto a la importancia de cada variable para clasificar entre equipos clasificados y no clasificados para *Play-Offs*, obtenemos los siguientes pesos:

- *eFG%* (se tiene en cuenta el *eFG%* y el *Opp. eFG%*): 47.85%
- *TOV%* (tanto *TOV%* como *Opp. TOV%*): 18.84%
- Rebotes (*ORB%* y *DRB%*): 16.95%
- *FT/FGA* (*FT/FGA* y *Opp. FT/FGA*): 16.36%

Para calcular estos pesos se toman los valores del *Mean Decrease Gini* para cada una de estas variables y se divide entre el total de todas.

Recordar que en este caso se ha tomado el modelo de “*Four Factors*” pero tomando como variable respuesta la variable categórica “*PLAYOFFS*”, la cual toma el valor “1” para indicar que dicho equipo consiguió clasificarse y el valor “0” para lo contrario.

Los pesos en este caso y utilizando el algoritmo *Random Forest* varían en relación al modelo inicial expuesto por Dean Oliver en su libro “*Basketball on Paper*”. El porcentaje en lanzamiento adquiere todavía más importancia que inicialmente, llegando a un peso cercano al 50%, mientras que las otras 3 variables se mantienen con pesos muy parecidos, donde sorprende ver la bajada en el cuidado de la bola (*TOV%*) y la gran subida en el *FT/FGA*, que triplica su peso inicial.

Matriz de confusión:

OBSERVACIONES/ PREDICCIONES	0	1	ERROR EN CADA CLASE
0	40	16	0.286
1	10	54	0.156

Número de árboles = 500

OOB error estimado = 21.67%

Figura 40: Algoritmo Random Forest sobre modelo “Four Factors”.

El *OOB error* en este caso es de 21.67%, que, aunque es ligeramente superior al anterior, sigue siendo un valor considerablemente pequeño. Esto es algo que cabía esperar y que ya ocurría anteriormente y es que el *OFFRTG* y el *DEFRTG* son dos métricas muy completas, que tienen implícitas muchas y muy buenas métricas de ataque y defensa y a su vez otras como el *Pace* o ritmo de juego.

VALORES REALES/ VALORES PREDICHOS	0	1
0	12	2
1	1	15

Figura 41: Validación del modelo Random Forest de “Four Factors” sobre el conjunto test (temporada 2019/20).

La sensibilidad y especificidad de este algoritmo sobre este modelo son:

Sensibilidad = 93.75%

Especificidad = 85.71%

En relación a la temporada 2019/20 y que tomamos como muestra para el test, vemos que a diferencia de lo que ocurría antes, este algoritmo con este modelo va a clasificar mejor a los equipos que sí que lograron clasificarse para la posttemporada con respecto a los que no lo hicieron, pero en ambos casos lo hace con tasas de acierto muy elevadas.

Este modelo también va a ser muy buen predictor a la hora de clasificar en cuanto a las estadísticas obtenidas por los equipos en las variables de los “*Four Factors*” su presencia o no en los *Play-Offs*.

Aplicando **validación cruzada k-fold** la predicción que obtenemos es de 84.66%, lo que supone una disminución a lo que obteníamos antes aplicando regresión logística sobre este mismo modelo

En este apartado hemos estudiado el funcionamiento de dos métodos de clasificación como son la *regresión logística* y el *Random Forest o Bosque Aleatorio*. El objetivo era ver la capacidad que tenían ambos, a partir de las variables presentes en los dos modelos que hemos estudiado en apartados anteriores (“*ataque-defensa*” y “*Four Factors*”) de predecir la presencia o no de los distintos equipos en los *Play-Offs* de la NBA.

En relación a los resultados que se han ido obteniendo, ambos métodos han tenido muy buen funcionamiento, con tasas de acierto muy elevadas (siempre superiores al 80%), existiendo una ligera mejora en la regresión logística con respecto al *Random Forest*, por lo que en este caso el primer método va a clasificar algo mejor.

Diferenciando entre ambos modelos (“*ataque-defensa*” y “*Four Factors*”) y a diferencia de lo que podíamos argumentar en otros apartados, el segundo modelo va a proporcionar mejores predicciones a la hora de clasificar a los equipos.

8. RESULTADOS Y CONCLUSIONES

Este proyecto lleva a cabo un elaborado análisis de las métricas estadísticas más usadas dentro de la liga norteamericana de baloncesto, la NBA. Podría servir como base para extender este trabajo para el estudio de otras ligas de baloncesto del mundo, como puede ser la Euroliga (compiten los mejores equipos de Europa), o la liga Española (ACB), entre otras, teniendo en cuenta que los estilos de juego cambian radicalmente en función de la liga que analicemos.

En cuanto al análisis de dichas métricas se llevó a cabo un estudio de cada una de ellas, con ejemplos de numerosas temporadas y equipos a lo largo de la historia de la NBA. También se explicó su cálculo y funcionamiento, facilitando la comparación de distintos equipos, ya sea de la misma temporada o incluso de diferentes, lo que nos permitía ver y analizar la influencia que tenían los cambios de tendencias y normas a lo largo de las temporadas de la NBA en cada una de estas métricas.

Más tarde, se realizó uno de los métodos de agrupación, el *k-means clustering*. Para este primer caso, incluiríamos en el método todas las variables que fueron estudiadas en los apartados anteriores, aún sabiendo del sobreajuste que se producía en los datos debido a la similitud entre muchas de ellas, para ver cómo agrupaba los distintos equipos incluidos en la muestra en función de las estadísticas que conseguían estos. Los resultados que obteníamos no eran concluyentes, ya que a pesar de que diferenciaba bien entre buenos y malos equipos, no obteníamos una buena distinción en el resto.

En lo relativo a los modelos de regresión múltiple, confirmamos el sobreajuste en los datos cuando teníamos el modelo formado por todas las variables. Debido a esto, realizamos el estudio del modelo que obteníamos tras aplicar los algoritmos de selección de variables, que era el formado únicamente por el *OFFRTG*, *DEFRTG* y *3PAr* (aunque este último con una influencia mínima), además del modelo formado por las variables de los “*Four Factors*” de Dean Oliver. En ambos casos, se obtenían muy buenos ajustes al número total de victorias que obtenían los distintos equipos durante la liga regular.

A raíz de lo anterior repetimos el *k-means* pero en este caso con los nuevos modelos. Los resultados que obteníamos eran mejores que antes, ya que las distinciones de los distintos grupos que se formaban estaban más claras.

Finalmente, queríamos predecir la presencia de los distintos equipos en *Play-Offs* y para ello se usó regresión logística y *random forest*. Los resultados obtenidos eran muy buenos, ya que tras utilizar validación cruzada, en ambos modelos la tasa de acierto superaba el 80%, siendo ligeramente superior en el caso de la regresión logística.

En general, este trabajo nos ha demostrado que a partir de las métricas y estadísticas con las que se cuenta en la actualidad, podemos tener un amplio conocimiento de lo que ocurre sobre una pista de baloncesto y analizar qué es en lo que está fallando o qué le falta a un equipo para alcanzar el máximo rendimiento posible.

De este trabajo se podrían llevar a cabo distintas implementaciones en un futuro, ya que a pesar de que se llevan recogiendo estadísticas en la NBA desde el año 1943, es en los últimos años cuando más avances se han realizado, hasta el punto de la existencia de sistemas de cámaras que realizan seguimientos segundo a segundo de los jugadores y del movimiento del balón. Esto llevará a la aparición de nuevas métricas que aún no conocemos.

Otra implementación muy interesante de este trabajo ya la comentábamos al inicio de este apartado y sería la extensión de estos análisis a otras ligas de baloncesto, para medir su funcionamiento.

9. BIBLIOGRAFÍA

- [1] nba.com. <https://stats.nba.com/> Fecha de acceso: 28/08/2020.
- [2] Basketball Reference. <https://www.basketball-reference.com/>. Fecha de acceso: 28/08/2020.
- [3] ESPN stats. Fecha de acceso: 20/06/2020.
- [4] nbamaniacs, "Tabla de clasificación NBA". <https://www.nbamaniacs.com/tabla-clasificacion-nba/>. Fecha de acceso: 16/06/2020.
- [5] Davor Fernández, 01/02/2019, "¿Cómo funciona la NBA?. Tope salarial, traspasos, contratos y Draft". <https://medium.com/@kanelonmonger/c%C3%B3mo-funciona-la-nba-tope-salarial-traspasos-contratos-y-draft-f025c01c8064#:~:text=Si%20no%20hubiera%20m%C3%A1ximos%2C%20probablemente,del%20total%20del%20salary%20cap.>
- [5] "Basketball Team Evaluation Metrics". <https://www.nbastuffer.com/analytics-101/team-evaluation-metrics/>
- [6] "Largest MOV in NBA history". <https://www.statmuse.com/nba/ask/which-team-has-the-highest-point-differential-per-game-in-a-season.> Fecha de acceso: 18/04/2020.
- [7] teamrankings.com. Fecha de acceso = 20/06/2020.
- [8] Justin, 20/06/2016, "Nylon Calculus 101: Possessions". <https://fansided.com/2015/12/21/nylon-calculus-101-possession/>
- [9] Adam Fromal, 04/09/2014, "Ranking the NBA's 20 Best Offenses of All Time". <https://bleacherreport.com/articles/2185102-ranking-the-nbas-20-best-offenses-of-all-time>
- [10] Adam Fromal, 12/09/2014, "Ranking the NBA's 20 Best Defenses of All Time". <https://bleacherreport.com/articles/2185159-ranking-the-nbas-20-best-defenses-of-all-time>
- [11] Pivot Analysis, 01/05/2019, "Offensive and Defensive Ratings - Net Rating. What do they tell us?". <https://www.pivotanalysis.com/post/net-rating#:~:text=The%20Pivot%20Analysis%20application%20will,they%20crash%20the%20offensive%20boards.>
- [12] Carles Ortiz, 17/09/2012, "NBA: un nuevo enfoque analítico". <https://www.nbamaniacs.com/articulos/nba-un-nuevo-enfoque-analitico/>
- [13] Kelly Scaletta, 02/03/2020, "Why NBA pace is at historic high". <https://www.lineups.com/articles/why-nba-game-pace-is-at-historic-high/>
- [14] Konstantinos Kotzias, 09/03/2018, "The Four Factors of Basketball as a Measure of Success". <https://statathlon.com/four-factors-basketball-success/>
- [15] Mika Honkasalo, 23/06/2014, "NBA Advanced Stats: The Four Factors Of Winning". <https://hoopshabit.com/2014/07/23/nba-advanced-stats-four-factors-winning/>
- [16] Justin Jacobs, 05/09/2017, "Introduction to Oliver's Four Factors". <https://squared20.com/2017/09/05/introduction-to-olivers-four-factors/>
- [17] Vero Arroyo, 19/06/2015 "Estadísticas NBA: FG% vs eFG%". <https://unaneuronaparacadadeporte.com/2015/07/19/estadisticas-nba-fg-vs-efg/#:~:text=La%20diferencia%20es%20evidente.,m%C3%A1s%20tiros%20de%20%20puntos.>
- [18] Jeff Haefner, 08/04/2020, "Whats is Effective Field Goal Percentage? And why you should use it?". <https://www.breakthroughbasketball.com/stats/effective-field-goal->

[percentage.html#:~:text=The%20Effective%20Field%20Goal%20Percentage,%E2%80%9Chigh%20level%20key%20indicator%E2%80%9D.](#)

[19] Alex Boeder, 02/11/2012, "Advanced Stats Series: Turnover Percentage". <https://www.nba.com/bucks/features/boeder-121102>

[20] Jeff Haefner, 11/03/2016, "Why every basketball coach should look at Rebounding Percentage instead of Totals". <https://www.breakthroughbasketball.com/stats/rebounding-stats.html#:~:text=Why%20Every%20Basketball%20Coach%20Should%20Look%20at%20Rebounding%20Percentage%20Instead%20of%20Totals,-By%20Jeff%20Haefner&text=Tracking%20and%20using%20statistics%20properly,players%20will%20boost%20their%20performance.>

[21] Ebran Khan, 18/10/2013, "Advanced NBA Stats For Dummies: How To Understand the New Sports Math". <https://bleacherreport.com/articles/1813902-advanced-nba-stats-for-dummies-how-to-understand-the-new-hoops-math>

[22] 06/03/2016, "Guía de estadísticas avanzadas NBA". <https://rompimientodefensivo.com/2016/03/06/guia-de-estadisticas-avanzadas-nba-plus-minus/>

[23] Justin Kubatko, 10/03/2008, "The Simple Rating System". <https://www.basketball-reference.com/blog/indexba52.html?p=39>

[24] Abdul Yunus, "K-means Clustering Using R", rpubs.com. Fecha de acceso: 28/06/2020.

[25] Joaquín Amat Rodrigo, "Análisis de Componentes Principales (Principal Component Analysis, PCA)". rpubs.com. Fecha de acceso: 26/06/2020.

[26] Sergio Rabinal, 18/03/2020, "Estadística avanzada: diccionario de términos, qué es, qué significa...", <https://es.nba.com/news/estadistica-avanzada-diccionario-de-terminos-que-es-que-sirve-nba-analytics-2019-sabermetrics/1gc73ududvomy1s3mqiahg238c>

[27] Patricia Pérez, 20/06/2012, "Las matemáticas transforman la NBA". https://tendencias21.levante-emv.com/las-matematicas-transforman-la-nba_a12206.html

[28] Dean Oliver, 2002: "Basketball on paper: rules and tools for performance análisis". Potomac Books, Inc., 2011

10. CÓDIGO

Para la implementación de este trabajo se ha utilizado el lenguaje de programación R. Los distintos paquetes y librerías utilizadas a lo largo del proyecto están nombradas según van apareciendo a lo largo de la memoria.

Se podrá acceder a este código, además de a los datos que han sido utilizados en su ejecución en el repositorio del siguiente [enlace](#).