EL PODER DE LOS DATOS EN LA NBA

Mateo Alís, Sergio Ortiz, Manuel Caballero, Joel Porcar, Maria Porta, Xavier Ventura

El objetivo de este trabajo es ver que estadísticas de jugadores de la NBA nos permite condensar decenas de métricas (puntos, rebotes, asistencias, etc.) en unas pocas componentes que explican la mayor parte de la variabilidad, facilitando la visualización de perfiles de juego, la detección de jugadores atípicos y la segmentación en tipos (tiradores, interiores, organizadores). Además de realizar un clustering de gran importancia para los GM de una franquicia que desee fichar a un jugador en mente, pero bien sea o porque no se pueden hacer cargo del contrato del jugador, o bien porque no se puede realizar ese traspaso, buscar alternativas para obtener el jugador más parecido que haya. Y por último, queríamos realizar un modelo de predicción utilizando el PLS-DA para predecir aquellos jugadores que pueden ser all-star en la liga ya que todo jugador que está en ella alguna vez en su vida a soñado con serlo.

Limpieza y Tratamiento de los datos

Para comenzar con el estudio de los jugadores de la NBA, primero cargamos los datos desde un archivo CSV con 679 filas. Este archivo contiene información sobre las estadísticas de los jugadores durante la temporada 2022-2023. Para facilitar el estudio de las variables, se clasificarán automáticamente como numéricas o categóricas.

Se observa que no tenemos valores nulos ni filas duplicadas en el dataset, pero se observa que aparecen 140 jugadores duplicados, lo que indica que hay jugadores que han jugado en más de un equipo durante la temporada 2022-2023. Esto es normal en la NBA, ya que los jugadores pueden ser traspasados entre equipos. Se observa que los jugadores que aparecen varias veces tienen en la columna Equipo el valor “TOT”, siendo el total de las estadísticas de ese jugador en todos los equipos en los que ha jugado (Ver rmarkdown en *ANEXO 1*.)

Por regla general cuanto más juega un jugador, más puntos debería anotar. Sin embargo, esto no siempre es cierto: hay jugadores que, aunque disputen muchos minutos, tienen un rol más pasador o defensivo, y otros que, jugando poco, son extremadamente eficientes ofensivamente. Por lo tanto, es importante tener en cuenta el número de minutos jugados y los puntos anotados por partido para evaluar el rendimiento de un jugador. Calculamos las estadísticas por minuto real, y ponderaremos por minutos jugados por partido con la función log(x+1) de manera que demos más peso a los que más minutos juegan sin exagerarlo. Con esto conseguimos que los jugadores que son muy productivos pero apenas juegan ya no se inflan y será una mejor escala para el análisis.

Posteriormente, se ha hecho una limpieza de los datos (Ver rmarkdown en *ANEXO 1*.)Se observa que los jugadores anómalos suelen tener muchos partidos y es muy

frecuente que sean jugadores All-Stars, lo que indica que no son errores de datos sino jugadores con un rendimiento excepcional. Se decide no eliminarlos, sin embargo, a los jugadores anómalos que han jugado menos de 8 partidos se ha decidido eliminarlos, por no ser una cifra muy alta para poder obtener conclusiones de esos jugadores.

Tras la limpieza, nuestra base de datos tiene 508 observaciones(jugadores) y 23 variables generalmente numéricas ya que estamos hablando de estadísticas, aun que hay otras categóricas como equipo y all-stars.

Análisis PCA

Aplicación del método

El objetivo de realizar un PCA a los jugadores de NBA es identificar patrones de rendimiento y eficiencia entre los jugadores de la NBA, permitiendo descubrir perfiles de juego diferenciados a partir de sus estadísticas individuales. Mediante la reducción de la dimensionalidad del conjunto de variables, se pretende representar de forma visual y simplificada las principales características que definen a los jugadores, facilitando la detección de estilos de juego, fortalezas y debilidades, así como la comparación objetiva entre ellos. Se recuerda que las variables han sido normalizadas por minuto y aplicado un escalado logarítmico ponderado por minutos jugados, de modo que ni los jugadores muy productivos con pocos minutos ni los habituales titulares inflen indebidamente el análisis. Gracias a esta transformación, el PCA representa visualmente las fortalezas, debilidades y perfiles diferenciados de cada jugador de manera más equilibrada y comparable.

Para poder realizar el PCA, es importante centrar y escalar los datos. Esto se hace para que todas las variables tengan la misma importancia en el análisis.

Selección de variables

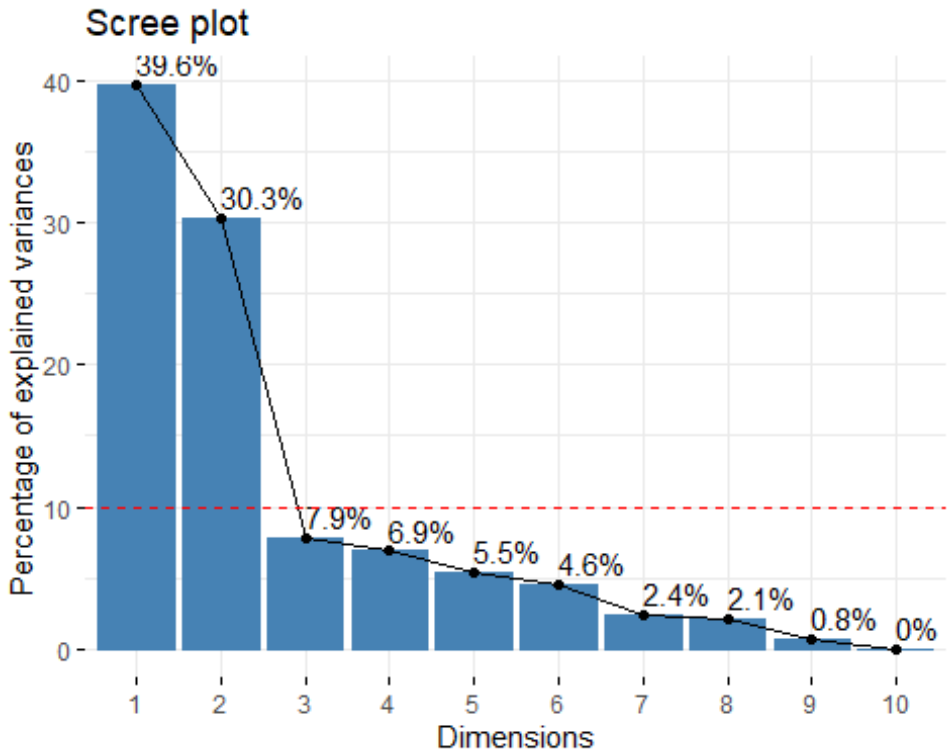
En cuanto a las variables que son interesantes para realizar el PCA, se observa que la variable rebotes no se puede usar porque es combinación lineal de las variables “ro” (rebotes ofensivos) y “rd” (rebotes defensivos). NO interesan las variables partidos_jugados ni partidos_titular, ya que se ha transformado las variables por minuto real y ponderado por minutos jugados por partido, si se incluyen, se corre el riesgo de que la primera componente principal distinga “titulares” de “suplentes” y no el rendimiento de los jugadores que es lo que se busca con el estudio.

Además, la variable all_stars no se usará porque es categórica, edad tampoco es interesante ni el nombre porque es identificativo. A partir de la matriz de correlaciones, se superime todas las métricas de tiros y anotaciones que están muy ligadas a “puntos_aj” (se ha eliminado tiros_anotados_aj y tiros_intentados_aj,tiros_2_anotados_aj y tiroa_2_intentados_aj), de modo que la primera componente no acabe midiendo únicamente volumen de anotación. Tampoco se conserva “triples_anotados_aj” —al compartir alta correlación con “triples_intentados_aj”, preferimos quedarnos con el volumen de intentos como mejor

indicador de rol ofensivo— ni “tl_anotados” —pues “tl_intentados_aj” ya refleja la agresividad al forzar faltas.

De esta manera solo se mantienen variables que aportan información específica sobre estilos de juego (intentos de triples y libres), defensa y creación de juego, lo que facilita identificar perfiles de rendimiento y distinguir roles en la pista.(Ver matriz de correlaciones en el *ANEXO 2*.)

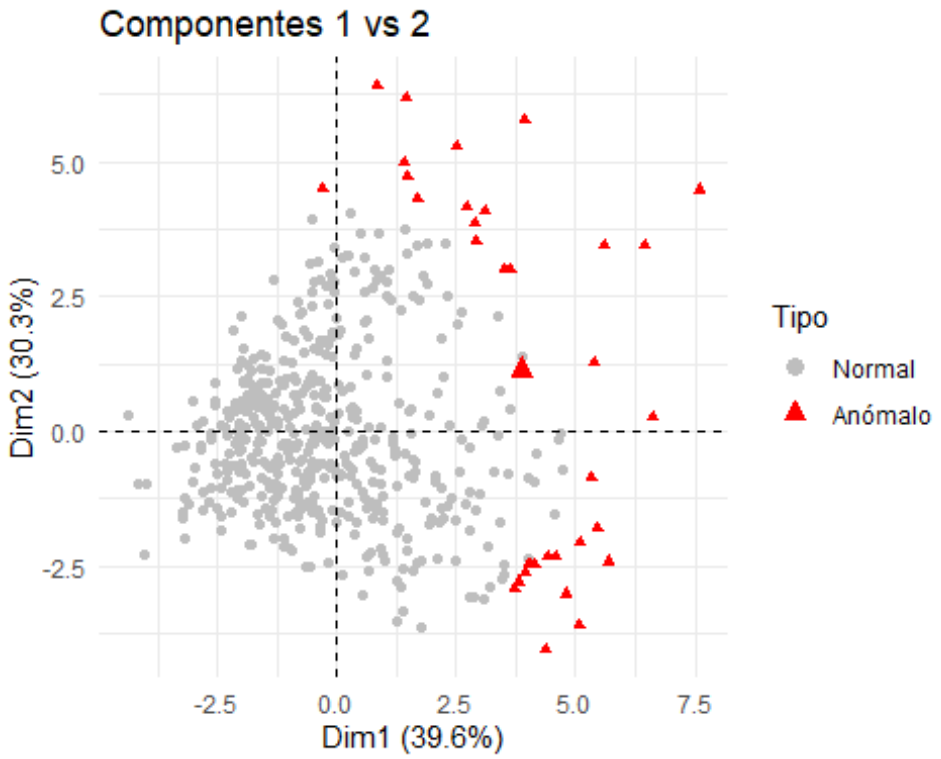
Se genera el modelo PCA para todas las posibles componentes principales (o un elevado número de ellas) y se selecciona el número “óptimo” de componentes principales (PCs).



Aplicando tanto la regla del codo como la regla de la media (Kaiser), se concluye que es adecuado conservar dos componentes principales. La regla del codo muestra un cambio brusco en la pendiente tras la segunda componente, indicando que a partir de ahí la ganancia en varianza explicada es marginal. Por su parte, la regla de la media establece un umbral del 10%, superado claramente solo por las dos primeras componentes (36.3% y 31.9%), mientras que la tercera queda por debajo. Por tanto, conservar dos componentes permite simplificar la representación de los datos manteniendo un elevado porcentaje de varianza explicada (68.2%) y facilitando una interpretación visual clara.

Detección de anómalos

El estadístico T2 de Hotelling permite identificar valores anómalos extremos, que podrían estar condicionando el modelo, es decir, la creación de las PCs.



Al fijar el límite de T^2 al 95 %, se espera por azar que unos 25 de los 508 jugadores lo superen, pero en realidad se detecta nueve más de lo previsto, lo que sugiere la presencia de perfiles verdaderamente extremos. Sin embargo, muchos de esos "anómalos" han disputado numerosos partidos y varios son All-Stars, lo que apunta a que no se trata de datos erróneos sino de jugadores con roles titulares y de gran impacto, por lo que merece la pena analizarlos en profundidad(Ver en *ANEXO 12*).

Los jugadores con mayor T^2 son Giannis Antetokounmpo, Joel Embiid, Luka Doncic, Trae Young, Damian Lillard y Walker Kessler(Ver en *ANEXO 3*)

Lillard destaca por un rol ofensivo dominante (puntaje elevado, triples y tiros libres intentados, asistencias); Giannis muestra su control de la pintura (puntos, rebotes ofensivos y defensivos) y alto volumen de pérdidas; Embiid concentra desequilibrios en anotación interior (tiros libres, puntos, rebotes defensivos) y pérdidas; Trae Young exhibe picos en creación y definición de juego (asistencias, puntos, tiros libres) junto con pérdidas; y Walker Kessler representa el extremo opuesto como especialista defensivo de élite (tapones y rebotes muy altos, ofensiva reducida).

La SCR mide cuánto se desvían las estadísticas reales de los jugadores respecto a lo que el PCA puede reconstruir. La mayoría queda bien explicada (SCR baja), pero 10 jugadores superan el umbral al 95 % y 5 de ellos incluso el 99 %, cifras acordes con la probabilidad teórica esperada. Estos casos indican estilos de juego atípicos o métricas que no encajan bien en las dimensiones principales del modelo.(Ver en *ANEXO 11*)

Ben Simmons destaca por un desajuste en asistencias y un peso secundario de los tiros; Brook Lopez y Jaren Jackson Jr. flaquean en tapones; Jimmy Butler muestra

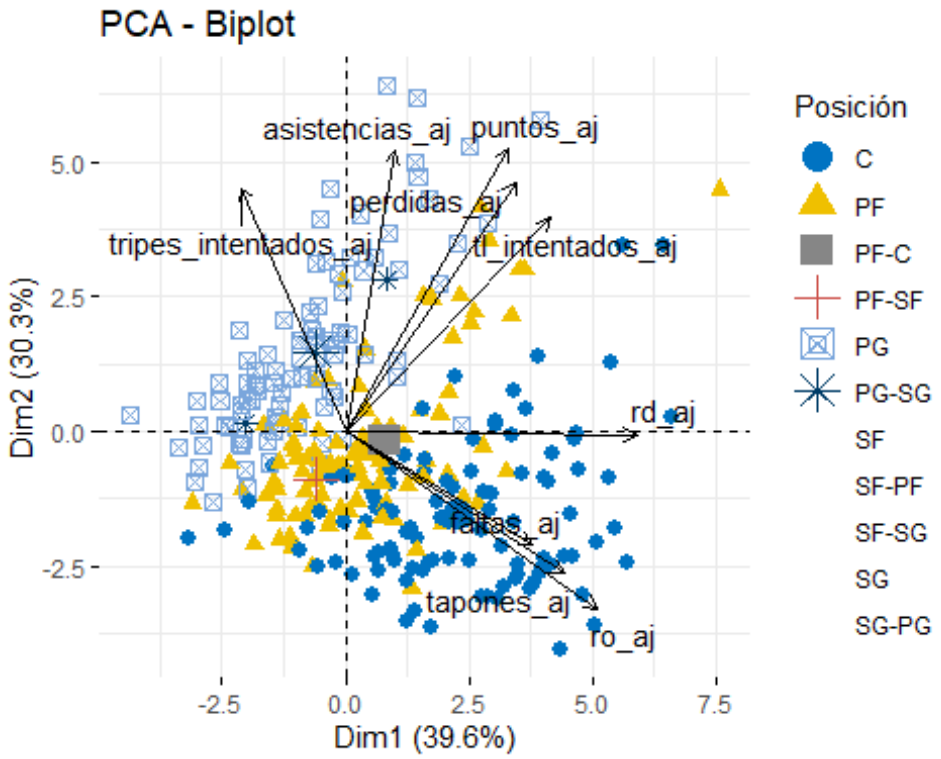
errores de modelo en libres y triples (con contribuciones adicionales de pérdidas y faltas); y Ryan Rollins, novato sin minutos de titular, queda señalado casi exclusivamente por pérdidas. Además, mientras los cuatro primeros son habituales titulares de la All-Star, Rollins apenas jugó y nunca arrancó de inicio, lo que explica su carácter atípico.(VER EN *ANEXO 4*)

Interpretación de los resultados del PCA

Las variables han sido coloreadas según su contribución a las dos primeras componentes principales del análisis PCA. En color azul se representan las variables auxiliares (partidos_jugados y partidos_titular), que no han sido utilizadas directamente en la obtención del modelo, pero se han proyectado sobre el nuevo espacio de componentes para facilitar su interpretación(Ver en *ANEXO 5*).

En el biplot de variables se observa dos ejes principales que capturan los estilos de juego en la NBA. En primer lugar, la Dimensión Ofensiva (PC1, 36,3 % de la varianza) agrupa de forma muy clara las métricas de volumen de ataque —puntos_aj, asistencias_aj y tl_intentados_aj— junto a la tasa de pérdidas_aj y las variables auxiliares partidos_jugados y partidos_titular. Esto indica que los jugadores que más minutos disputan y más participan en la anotación también asumen un mayor riesgo de pérdida de balón, algo lógico en quienes tienen un rol ofensivo dominante.

Por otro lado, la Dimensión Defensiva (PC2, 31,9 % de la varianza) está definida por rebotes ofensivos y defensivos (ro_aj, rd_aj), tapones_aj y faltas_aj, lo que refleja que quienes más luchan bajo los tableros y bloquean tiros suelen implicarse también en más infracciones. Al proyectar estas variables, constatamos que el análisis PCA nos permite diferenciar claramente entre perfiles ofensivos (Jugadores de alto volumen y creación de juego) y defensivos (Especialistas en rebote y protección del aro), ofreciéndonos una representación visual y simplificada de los extremos del rendimiento en la competición(Ver en *ANEXO 6*)



Dimensión Ofensiva (PC1 - 36,3% de la varianza): - Las variables puntos_aj, tl_intentados_aj, perdidas_aj y en menor medida asistencias_aj tienen vectores largos y apuntan en la misma dirección, indicando que los jugadores con valores altos en esta dimensión son los que más protagonismo tienen en ataque. - Este grupo incluye a jugadores de posiciones ofensivas como PG (Base) y SG (Escolta), que tienden a generar juego y acumular puntos, pero también a cometer más pérdidas debido a su alta participación en las jugadas.

Dimensión Defensiva (PC2 - 31,9% de la varianza): - Las variables ro_aj (rebotes ofensivos), rd_aj (rebotes defensivos), tapones_aj y faltas_aj están relacionadas con la actividad defensiva y de lucha por el balón. - Las posiciones de C (Pivot) y PF (Ala-Pivot) dominan en este plano, evidenciando su mayor implicación en acciones defensivas, rebote y protección del aro.

Las posiciones SF (Aleros) y jugadores polivalentes (SG-PG, SF-PF, etc.) aparecen más dispersos, reflejando su versatilidad tanto en defensa como en ataque.

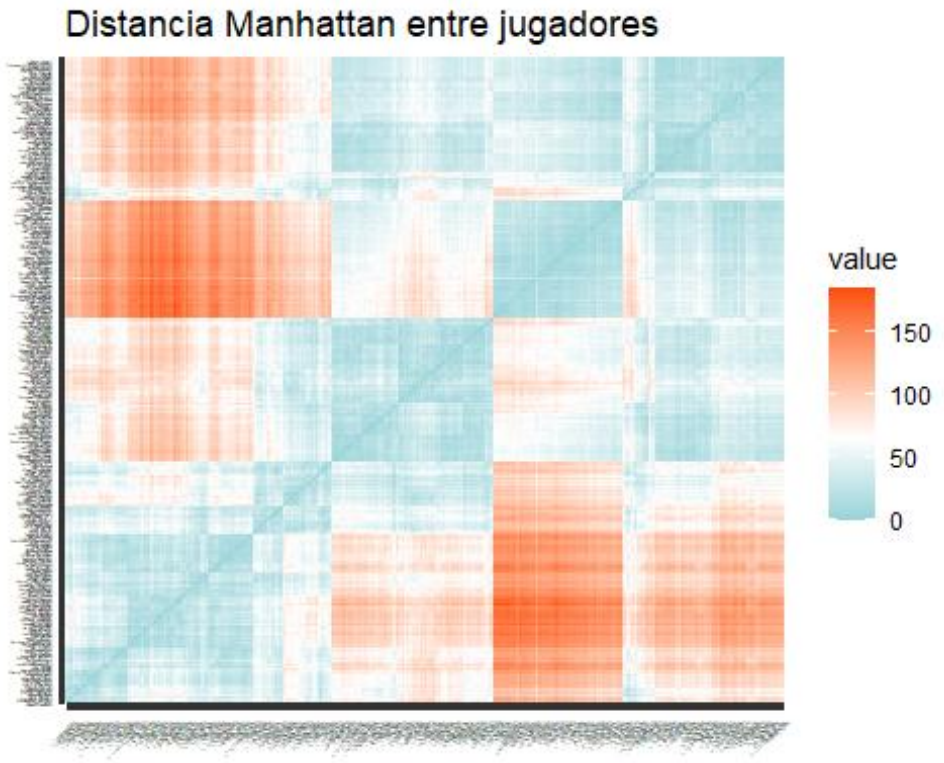
Clustering

Aplicación del método

El objetivo de realizar clustering es el de ayudar a un General Manager de la NBA quiere incorporar a un jugador que sea pieza clave de su equipo, pero bien por que el jugador cobra mucho dinero en la agencia libre o bien porque no han conseguido lograr un traspaso con su equipo no lo pueden traer. Debe tener otras alternativas que no

desvaríen mucho de sus pensamientos. Esta estrategia de ponderación se alinea con el rol de General Manager, minimizar el riesgo de incorporar perfiles rookies sin recorrido y asegurar que el candidato seleccionado aporte tanto en producción estadística como en veteranía y liderazgo desde el primer momento.

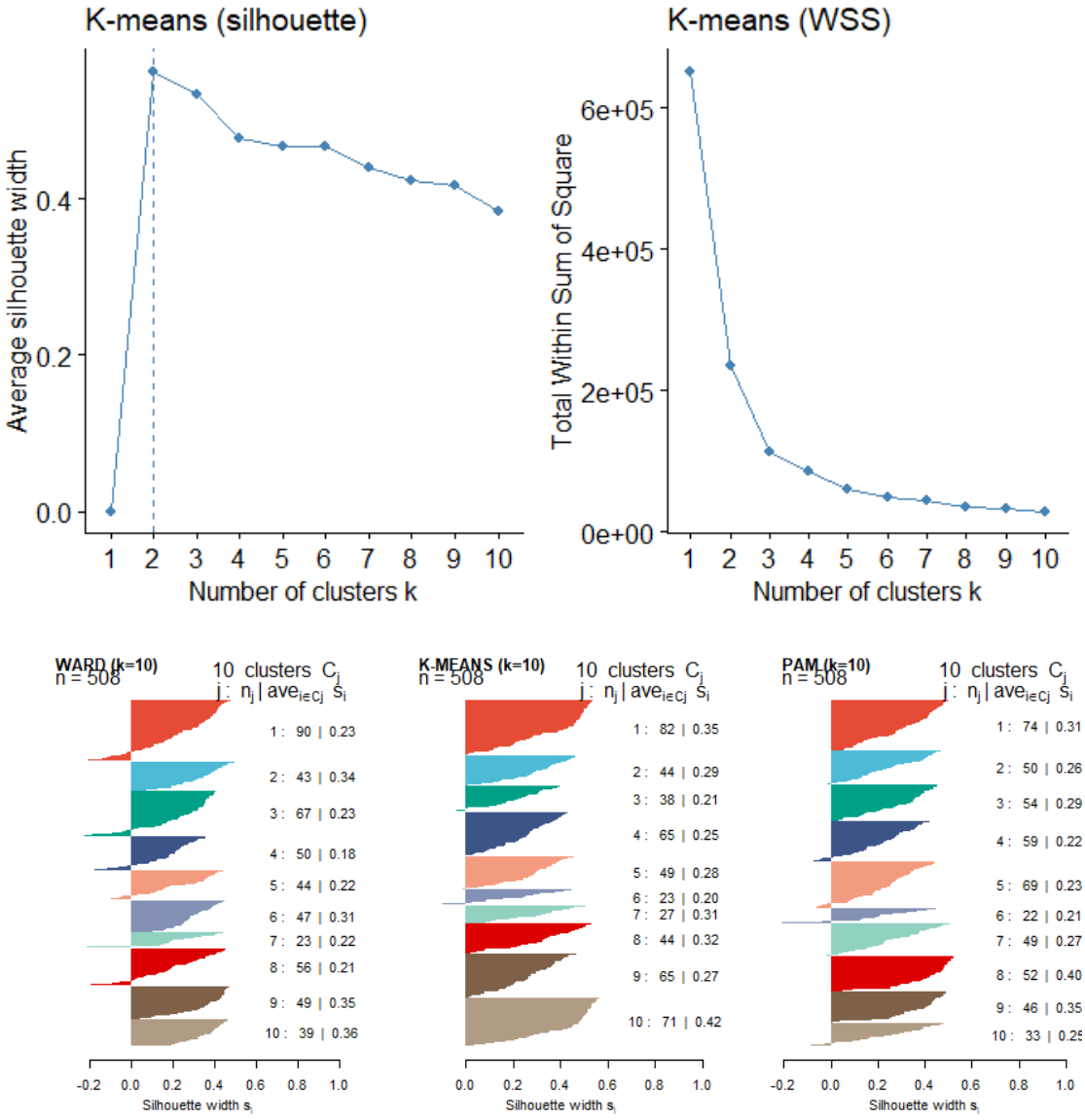
Se utilizará una medida de distancia basada en la distancia de Manhattan, ya que al sumar diferencialmente las discrepancias absolutas en cada estadística (puntos, asistencias, rebotes, experiencia, titularidades), nos ofrece un coste de sustitución lineal y completamente desglosable. Por ejemplo, si dos jugadores difieren en 2 asistencias y 3 rebotes, su distancia aumentará exactamente en 5 unidades. De esta forma la agrupación que se haga nos beneficia al buscar jugadores que tengan un estilo de juego muy similar.



El mapa de color muestra que los jugadores se agrupan es diversos grupos, que es lo que buscamos.

En primer lugar, se aplicará modelos jerárquicos, utilizando el método de la media que es el que utilizaremos, pero también hemos realizado Ward y K-medoides(Ver *Anexo 7* y *Anexo 8*). Se empieza estimando el número óptimo de clusters:

Resultados numéricos y gráficos



Se observa como utilizando K-medias vemos como cada individuo está perfectamente clasificado en su cluster menos alguno que otro. Por ello se utilizará el método de K-means a la hora de extraer nuestros clusters.

Por otro lado, se ve como el número optimo de clusters que nos salían con los 3 métodos era 2. En este caso este no serviría de nada, ya que el objetivo es clasificar jugadores por estilo de juego similar que tengan además de experiencia, partidos jugados, etc. Lo más importante sería que hubieran cuantos más grupos mejor para poder centrarnos en roles de jugadores específicos.(*ANEXO 7 Y ANEXO 8*) Lo que sería bueno para nosotros ya que si seleccionaremos 2 o 3 clusters agruparíamos nuestros 500 jugadores en 3 grupos, lo que no tendría sentido en nuestro objetivo. Por lo que nos vamos a decantar por seleccionar un total de 10 clusters para que hayan grandes discrepancias entre jugadores, que es lo que se busca. Por lo que, si

quisiéramos ver un jugador alternativo a otro, tendríamos que ir al cluster de ese jugador y ver que otros jugadores hay en el cluster que juegen en su misma posición. Ya que si por ejemplo, nuestro jugador fuera un base y escogiéramos un pivot no tendría sentido la elección.

Disusión de los resultados del clustering

Tras realizar la separación de jugadores en clusters(desde 23 hasta 80 observaciones por cluster) se ha obtenido las siguientes conclusiones de cada uno:

CLUSTER 1- En este grupo se puede apreciar aquellos jugadores que son recientes estrellas de la nba y jugadores que están a poco de convertirse en ellas. Encontrando jugadores que han sido all-star alguna vez en su carrera o que se han quedado a las puertas de serlo. Jugadores con gran impacto en cancha para su equipo, que se consideran estrellas de este, normalmente la segunda espada.

CLUSTER 2- Este grupo está formado por aquellos jugadores de rotación de un equipo. Un equipo suele jugar con uno 12/13 jugadores de media por partido en la temporada regular, y en este grupo se encuentran esos jugadores de banquillo que son de gran importancia para el equipo ya que dan descanso a los jugadores principales, obviamente no son estrellas, pero son necesarios ya que es importante los jugadores de banquillo.

CLUSTER 3- Son aquellos jugadores que tienen su rol especifico en el equipo, que salen y juegan sus 25 minutos desde el banquillo principalmente, pero que si algún partido por diversas situaciones deben jugar de titular lo hacen sin problema cumpliendo a la perfección. Son jugadores que si se encontraran en otro equipo y tuvieran mayor numero de oportunidades podrían hacerse un hueco en el quinteto inicial sin problema ya que están capacitados para ello.

CLUESTER 4- En este caso se habla de jugadores de rol(rotación de banquillo) de una importancia un tanto menor que en el cluster 2, ya que no contribuyen tanto al juego de su equipo. Estos no tienen tanto impacto en rebotes ni puntos como en el cluster 2. Pero podrían pertenecer si su equipo le diera un poco más de importancia en el juego.

CLUSTER 5- Son jugadores que participan en algún que otro partido normalmente cuando se busca dar descanso a alguna parte de la plantilla(en la NBA se juega cada 2 días normalmente y hay mucha sobrecarga). Estos jugadores normalmente suelen ser veteranos en la liga y gusta mucho a los equipos tenerlos bajo su poder. Son aquellos jugadores que ya han vivido mucho en la liga y se acostumbran a jugar sus 20-25 partido al año no jugando muchos minutos, pero lo principal es su veteranía a la hora de afrontar las temporadas y ayudar al equipo.

CLUSTER 6- En este cluster tienen presencia jugadores interiores(SF-Alero,PF-AlaPivot y C-Pivot) que tienen un gran peso en su equipo pero son propensos a lesionarse dada su poca participación a lo largo de la temporada. Estos jugadores son considerados de gran importancia en el quinteto inicial como normalmente tercer mejor jugador, pero cogido con pinzas ya que un jugador pensado para esto debe tener un gran físico y capacidad para estar presente en todos los partidos. Un GM debe tener

mucho cuidado con estos jugadores ya que puede haber un gran desembolso económico en ellos y después que estén gran parte de la temporada lesionados.

CLUSTER 7-Jugadores que no paran quietos en pista. Siempre están en constante movimiento involucrado en jugadas. Son aquellos jugadores que se asocian con todo el equipo, que las jugadas pasan por ellos aun sin ser estrellas. No se quedan en la esquina esperando a que les llegue el triple como un jugador de rotación, sino que a pesar de salir bastante desde el banquillo(aun que también pueden jugar de titular) tienen un rol fundamental para que cuando no están las estrellas el baloncesto de su equipo se juegue de forma fluida.

CLUSTER 8- Son los jugadores que llevan el peso del partido para su equipo aquellos jugadores que pueden anotar 20/25 puntos cada noche sin problema para ayudar a que su equipo consiga la victoria. Aquellos que siempre tienen la pelota en sus manos controlando el juego de su equipo, unos anotadores natos, pero además son asistidores y grandes defensores que siempre están en contacto con la pelota y que su facilidad para anotar está a años luz con el resto de jugadores. Aquí sería jugadores diferenciales para su equipo que muchos desearían tener.

CLUSTER 9- En este caso son aquellos jugadores que se consideran de banquillo. De esos que suelen jugar pero sus minutos son los que normalmente conocidos como de residuos, cuando un partido ya bva perdiendo o ganando de mucho, estos jugadores se encargan de dar descanso a los titulares para evitar posibles lesiones y así ganan experiencia en cancha por si se les necesitara en un momento clave de la temporada. Son jugadores que siempre deben de tener los equipos por cualquier imprevisto que haya, ya que deben cumplir como suplente si se les necesita o si el equipo se encuentra dentro de una mala racha de lesiones. Por lo que se trata de jugadores que se adaptan perfectamente a su rol y saben lo que es estar muchos partido en el banquillo hasta que les llega su momento para cumplir.

CLUSTER 10- Trata de jugadores que no juegan directamente, o si lo es no es relevante. Muchos equipos tienen a estos jugadores, que normalmente suelen estar en rodaje con el equipo sobre todo a la hora de realizar los entrenamientos. Son muy importantes en su rol que es el de que los titulares se sientan incomodos cuando entrenan contra ellos. Normalmente suelen considerarse jugadores jóvenes que están en desarrollo con un equipo para ver si a base de entrenamientos convence a los entrenadores y consigue hacerse un hueco en el banquillo del equipo. Son los llamados jugadores de desarrollo.

Ahora, se hace hincapié en encontrar un jugador que sea similar a T.J.McConnell, este es un jugador bastante importante desde el banquillo, que lleva la producción de su equipo desde esa segunda unidad(banquillo) y que puede realizar sin problemas su rol y ayudar al máximo al equipo.

Davion Mitchell, Devonte’ Graham o Malcolm Brogdon entre otros serían jugadores a por los que ir en lugar de T.J. McConnell. Se considera de gran acierto estos jugadores ya que comparten estilos de juego de lo más similares entre ellos, que es lo que buscábamos, todo se consideran esos jugadores de banquillo capaces de generar para

sus compañeros y encargados de tener el control de la pelota en aquellos momentos donde los titulares descansan en el partido.

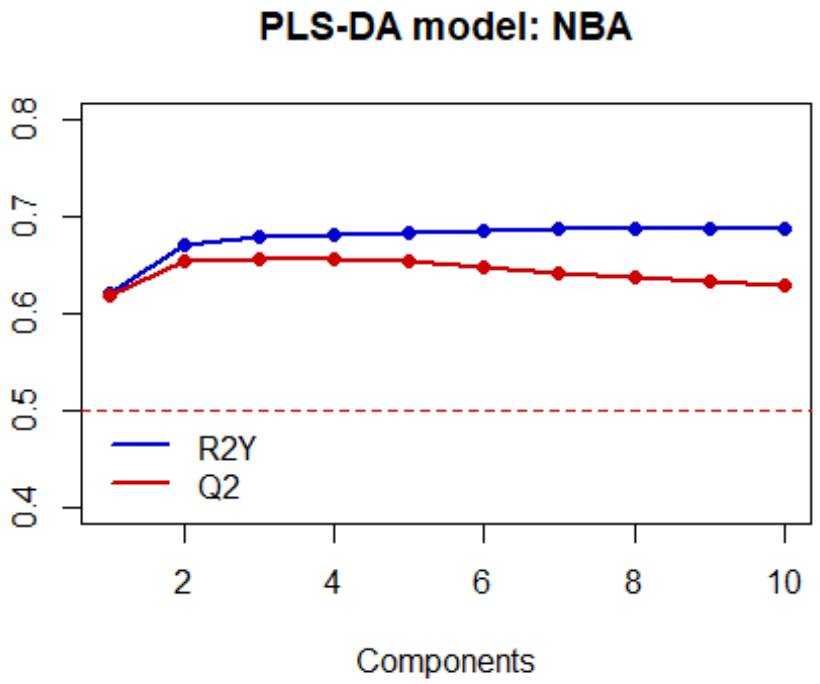
Ha sido una decisión muy acertada seleccionar un gran número de clusters. Los jugadores proporcionados son muy similares al que se buscaba, por tanto el objetivo que teníamos lo hemos cumplido a la perfección y ya podríamos como GM de una franquicia tomar decisiones muy importantes en un equipo para la creación optima de su plantilla sin desvariar muy de los pensamientos que se tenían en un primer momento.

Análisis PLS

Aplicación del método

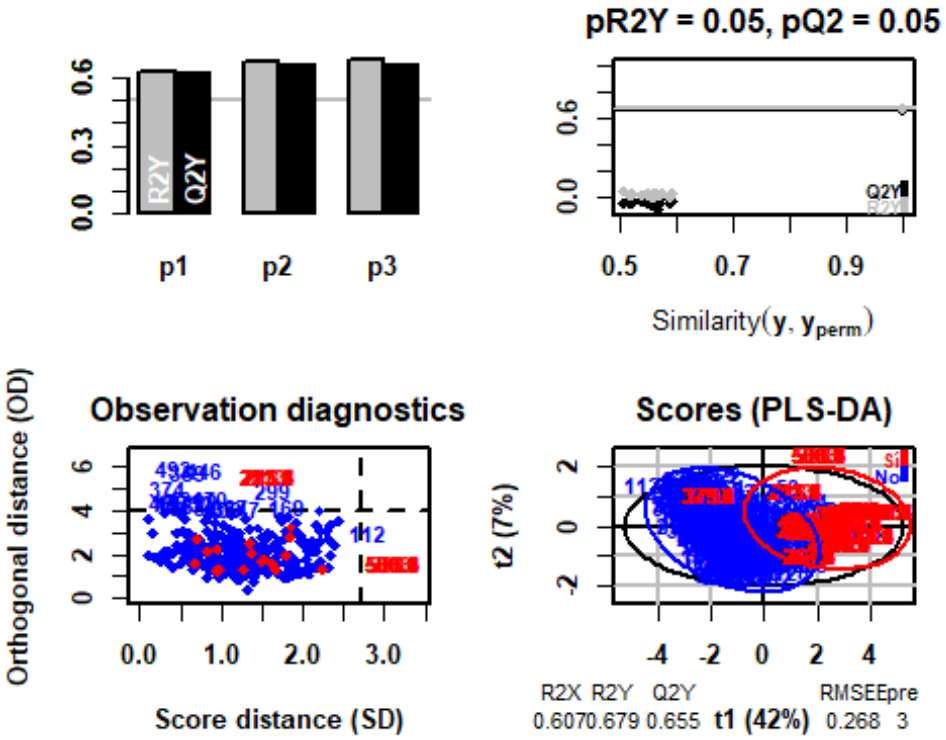
Con 480 “No” frente a 28 “Sí” en el conjunto de entrenamiento, la proporción original es casi de 17:1 (unos 5 % de All-Stars). Si llevásemos esto directamente a 1:1, se tendría que replicar casi 452 jugadores All-Star (de 28 a 480), corriendo el riesgo de sobre ajustar el modelo a muy pocas observaciones genuinas de “Sí”. Por eso conviene optar por un balance intermedio que aumente la representación de All-Stars sin inflar en exceso la muestra con duplicados idénticos.

Un ratio de 2:1 (“dos No” por cada “Sí”) lleva a apuntar a unos 240 All-Stars en lugar de 28 (porque 480/2 ≈ 240), de modo que se tendría que duplicar 212 casos de “Sí” (240 − 28). En la práctica, esto crea un total de 720 observaciones (480 No + 240 Sí). Con este grado de upsampling, nuestro modelo PLS-DA ve un tercio de la muestra como All-Stars, lo cual es suficiente para que aprenda los patrones característicos de esos jugadores sin recurrir a miles de copias idénticas. En el contexto de la NBA, donde los All-Stars representan una minoría muy reducida pero de gran importancia (queremos detectarlos con alta sensibilidad), un ratio de 2:1 ofrece un buen compromiso: mejora la capacidad para identificar rasgos distintivos de All-Stars sin sacrificar tanta variabilidad como para sobre ajustar. Si tras esto se observa que todavía demasiado sesgo (por ejemplo, muchos falsos positivos o falsos negativos).

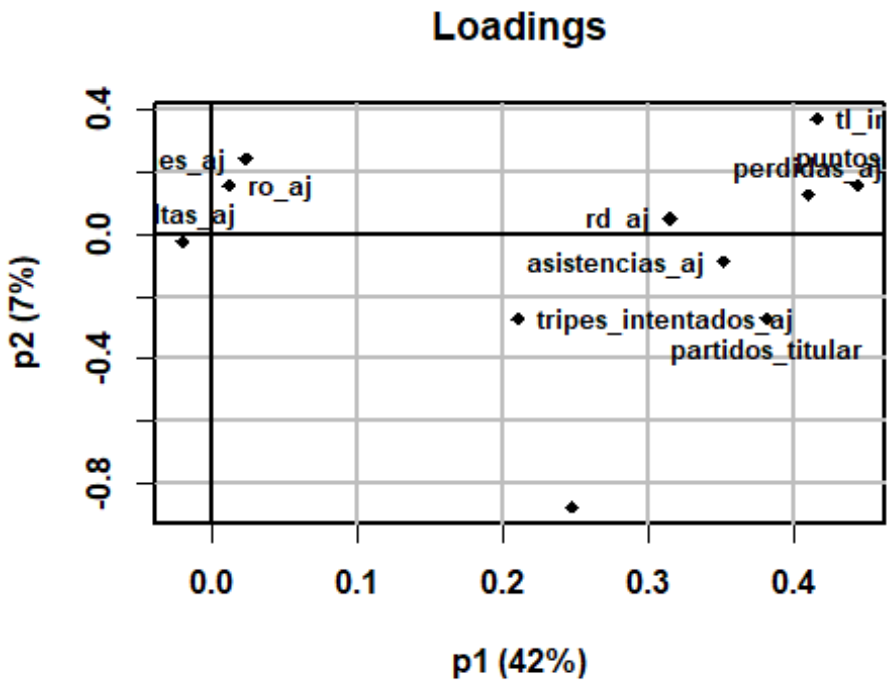


En este contexto, elegir tres componentes Latentes resulta adecuado porque a partir de la tercera componente la ganancia en R^2Y se estabiliza cerca de 0.68–0.70 y el valor de Q^2 comienza a descender levemente. Al seleccionar cinco componentes comprobamos que tenemos un alto poder explicativo (R^2Y) sin sacrificar significativamente la capacidad predictiva (Q^2), evitando así sobre ajustar el modelo mientras capturamos la mayor varianza útil para discriminar All-Stars.

Resultados numéricos y gráficos



Los cuatro paneles ilustran que, con tres componentes, nuestro PLS- DA capta eficazmente la variabilidad de la dicotomía All- Star/no All- Star sin sobreajuste (R^2Y se estabiliza en ≈0,66–0,68 y Q^2 alcanza ≈0,66), y que este poder discriminante no es fruto del azar (los modelos con etiquetas permutadas quedan muy por debajo de los valores reales, p = 0,05). El diagnóstico de observaciones muestra que casi ningún jugador excede los umbrales de distancia de scores (SD) u ortogonal (OD), lo que indica ausencia de outliers extremos, y el score plot de t_1 vs. t_2 revela una separación clara: los All- Stars se sitúan en valores altos de t_1 y los no All- Stars en la parte izquierda del espacio.



En este diagrama de loadings se aprecia cómo cada estadística de jugador se proyecta sobre las dos primeras componentes latentes: la primera (p1, 42 %) agrupa variables ofensivas fuertes como “tl_intentos_aj”, “puntos” y “perdidas_aj”, lo que indica que los All-Stars suelen tener alto volumen de tiro y anotación (y, por la agresividad ofensiva, más pérdidas), mientras que “asistencias_aj” y “partidos_titular” también cargan positivamente en p1, señalando que estar de titular y repartir asistencias caracterizan ese perfil. La segunda componente (p2, 7 %) resalta acciones defensivas (“robos_aj”, “tapones_aj”) en valores positivos, diferenciando a los jugadores que, además de anotar, aportan robos y bloqueos. En conjunto, estos loadings muestran que los All-Stars se distinguen tanto por su impacto ofensivo (p1) como por contribuciones defensivas (p2).

La Componente 1 es la que mejor discrimina la clase(R = 0.7878); quienes obtienen valores altos de t1 se proyectan en un u1 alto (grupo All-Star), y quienes tienen t1 bajo se ubican en u1 bajo (No All-Star). La Componente 2 aporta información adicional de discriminación (más suavemente, R=0.3614), detectando matices que no quedaron en t1, pero no es tan potente para clasificar por sí sola.Ç La Componente 3 ya no es relevante para separar las clases(R=0.1669); solo matiza pequeñas diferencias residuales que no capturaron las dos primeras componentes(Ver *anexo 9*)

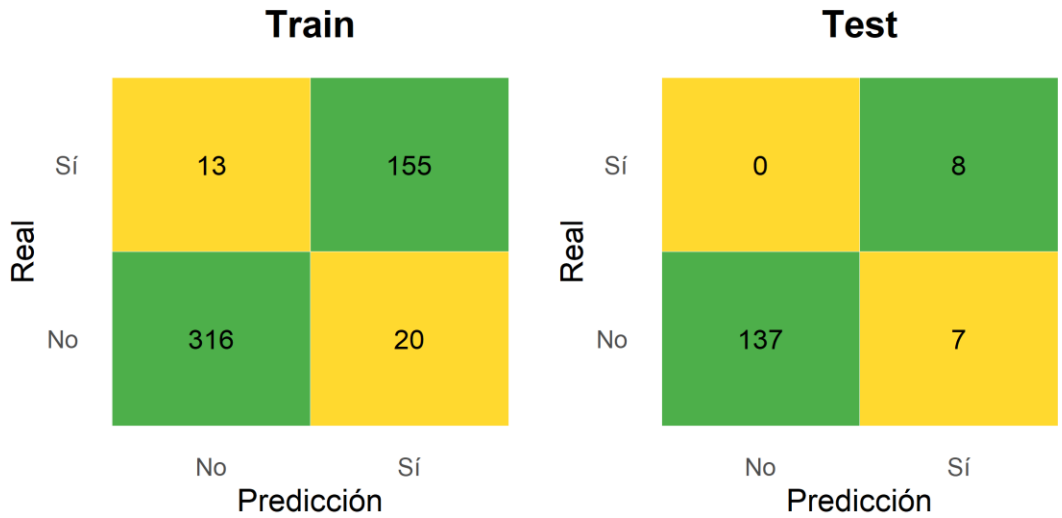
Las variables con VIP > 1 (“puntos_aj”, “tl_intentos_aj”, “perdidas_aj”, “partidos_titular” y “asistencias_aj”) son claramente las más relevantes para distinguir All-Stars de no All-Stars. “puntos_aj” (≈ 1.53) y “tl_intentos_aj” (≈ 1.44) reflejan que el volumen de anotación y la agresividad ofensiva—más intentos de tiro generando pérdidas (“perdidas_aj” ≈ 1.22) son rasgos distintivos de un All-Star. Además, figurar como

“partidos_titular” (≈ 1.30) y repartir “asistencias_aj” (≈ 1.05) señalan el rol de líder en la cancha, típico de quienes reciben votos o son vistos como piezas clave en su equipo(*Ver en ANEXO 10*).

Discusión de los resultados del PLS-DA

Se observa un buen ajuste en entrenamiento: Con un 93.45 % de accuracy, el modelo aprende a distinguir correctamente en su propio conjunto de entrenamiento tanto All-Stars como No- All-Stars. Hay que tener cuidado de sobreajuste: Estos valores provienen del mismo conjunto con el que se ha entrenado (upsampleado). Para verificar que el modelo no está memorizando patrones del train, hay que corroborar su desempeño sobre el conjunto de test original (sin duplicados).

En resumen, en test el modelo no pierde ningún All-Star (sensibilidad perfecta), pero “paga el precio” de incluir algunos no All-Stars como falsos positivos, resultando en una precisión moderada sobre los “Sí” predichos. Estos falsos positivos se podrían considerar all-stars en el caso en el que el criterio de los entrenadores/periodistas basados en las estadísticas tuviera el 100% del peso en la votacion. En realidad, esta votacion se realiza al 50% entre entrenadores/periodistas y el voto del público. Por ello nuestro modelo al tener solamente en cuenta las estadísticas, los jugadores que predecimos sí que podrían ser considerados all star, pero debido a los fans de los equipos y sus jugadores votan por fanatismo y no por meritocracia decantando la balanza por los jugadores más mediticos.



Con el upsampling 2:1 logramos un PLS-DA que separa bien a los All-Stars sin quedar sesgado hacia la clase mayoritaria. Con solo dos componentes captamos casi toda la señal: la primera explica el 42 % de R^2X y se correlaciona 0,79 con la etiqueta, mientras la segunda aporta la varianza residual necesaria. Los VIP confirman lo que esperábamos: puntos ajustados e intentos de tiro son los rasgos decisivos, y enseguida aparecen pérdidas, partidos como titular y asistencias. En entrenamiento obtenemos una balanced accuracy del 93 % (92 % de sensibilidad y 94 % de especificidad); en prueba, detectamos todos los All-Stars y mantenemos una especificidad del 95 %. El

coste de no dejar escapar ningún All-Star es un aumento de falsos positivos, de modo que el valor predictivo positivo cae al 53 %, pero en conjunto el modelo generaliza bien y refleja con claridad que el volumen ofensivo es el sello principal de la élite NBA.

Coclusiones

A lo largo de este trabajo hemos encadenado tres piezas analíticas PCA, k-means y PLS-DA que, combinadas, proporcionan una visión profunda y operativa del ecosistema NBA. En primer lugar, el PCA nos permitió reducir más de veinte variables por jugador a dos ejes que concentran casi el 70 % de la varianza: un eje ofensivo, dominado por puntos, intentos de tiro y asistencias, y un eje defensivo, gobernado por rebotes, tapones y faltas. Este espacio latente describe con claridad el continuo ataque/defensa y, gracias a la estandarización por minuto y un ponderado logarítmico de los minutos jugados, elimina el sesgo de uso para comparar con justicia tanto titulares como suplentes productivos. Además, las distancias T^2 y la SCR revelaron que las observaciones “extremas” corresponden a perfiles reales y valiosos (All-Stars de uso altísimo u especialistas defensivos), por lo que conviene conservarlas para representar los polos del rendimiento.

Sobre ese mismo espacio aplicamos clustering k-means, obteniendo agrupaciones coherentes con la intuición de entrenadores y analistas: bases creadores de tiro exterior, alas anotadores polivalentes, grandes protectores del aro, especialistas de rol, etc. Esta segmentación nos permitió desarrollar la función get_sustitutos(), que identifica candidatos de la misma posición y clúster para reemplazar a un jugador determinado. De este modo aportamos una herramienta práctica de scouting: basta con un nombre para obtener alternativas de perfil similar, útil en rotaciones, traspasos o ajustes por lesión.

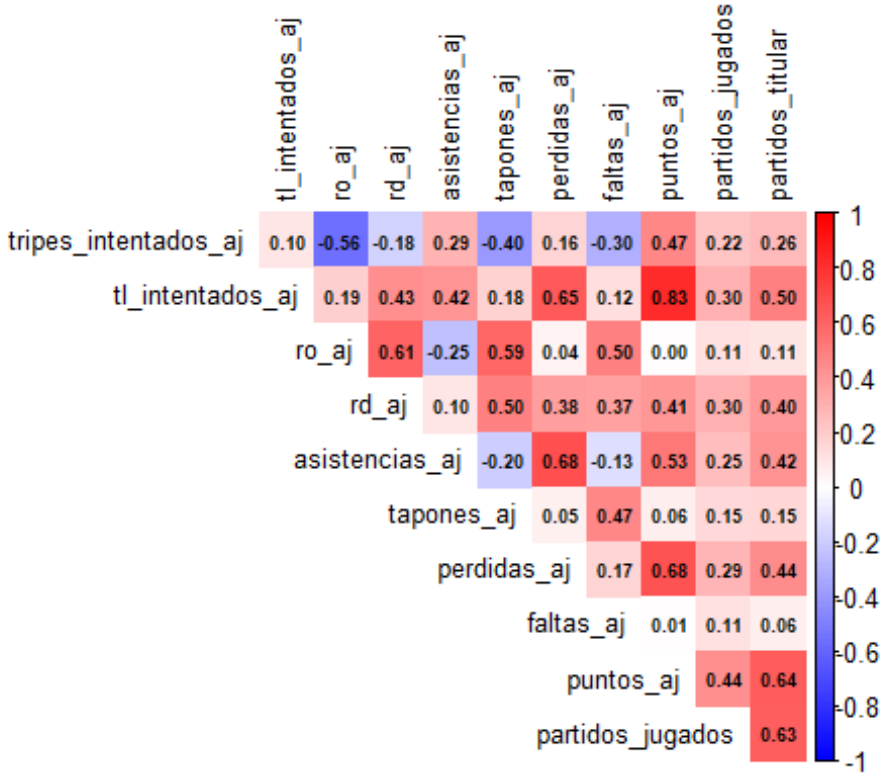
Por último, pusimos a prueba la capacidad discriminante de las mismas variables mediante un PLS-DA dirigido a clasificar jugadores en All-Stars y no All-Stars. Con solo tres componentes alcanzamos $R^2Y \approx 0{,}68$ y $Q^2 \approx 0{,}66$. El modelo señala como motores principales de la condición All-Star el volumen ofensivo (puntos y tiros libres) y la creación (asistencias), mientras que las pérdidas y ciertos indicadores defensivos actúan como contrapeso.

Anexos

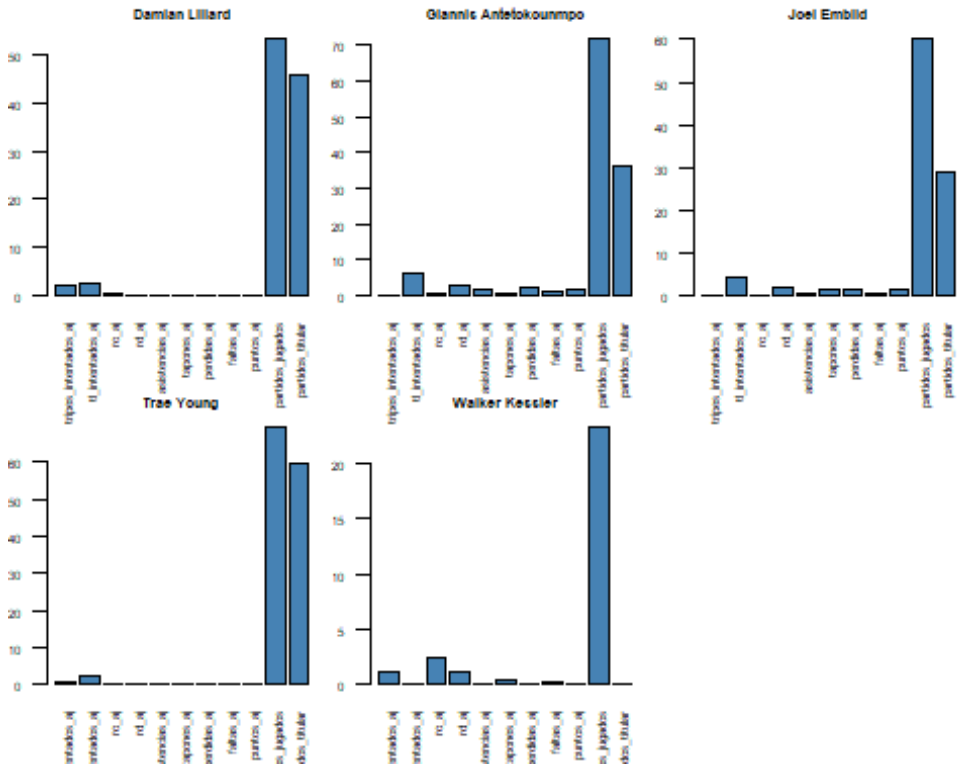
Anexo 1:

https://github.com/cofrian/NBA_MDP/blob/main/limpieza_tratamiento_datos.Rmd

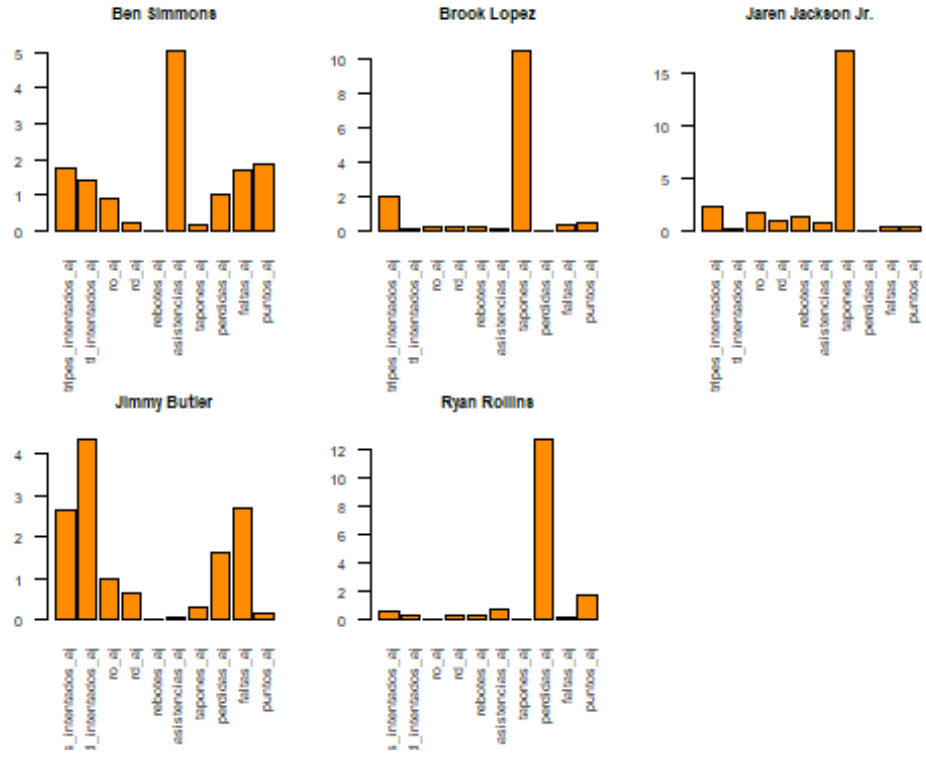
Anexo 2:



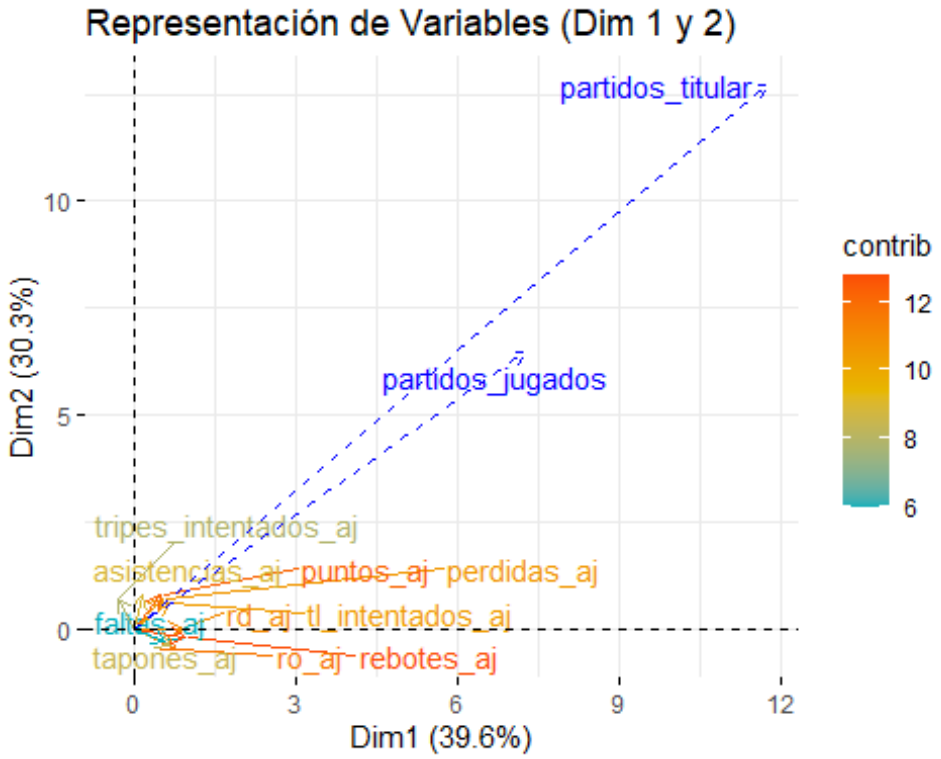
Anexo 3:



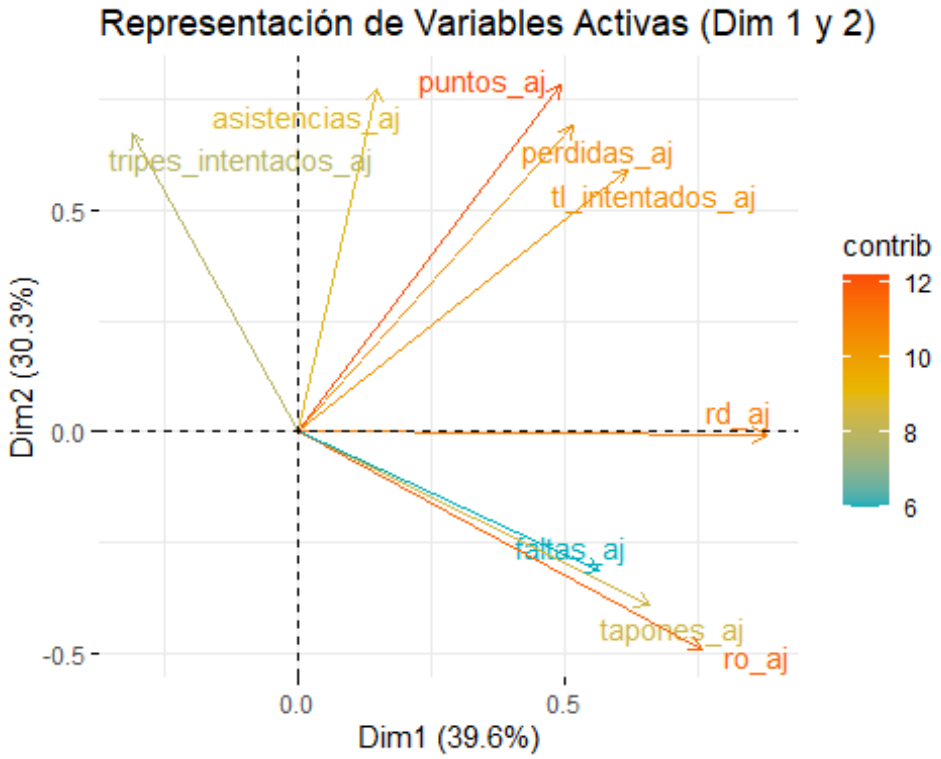
Anexo 4:

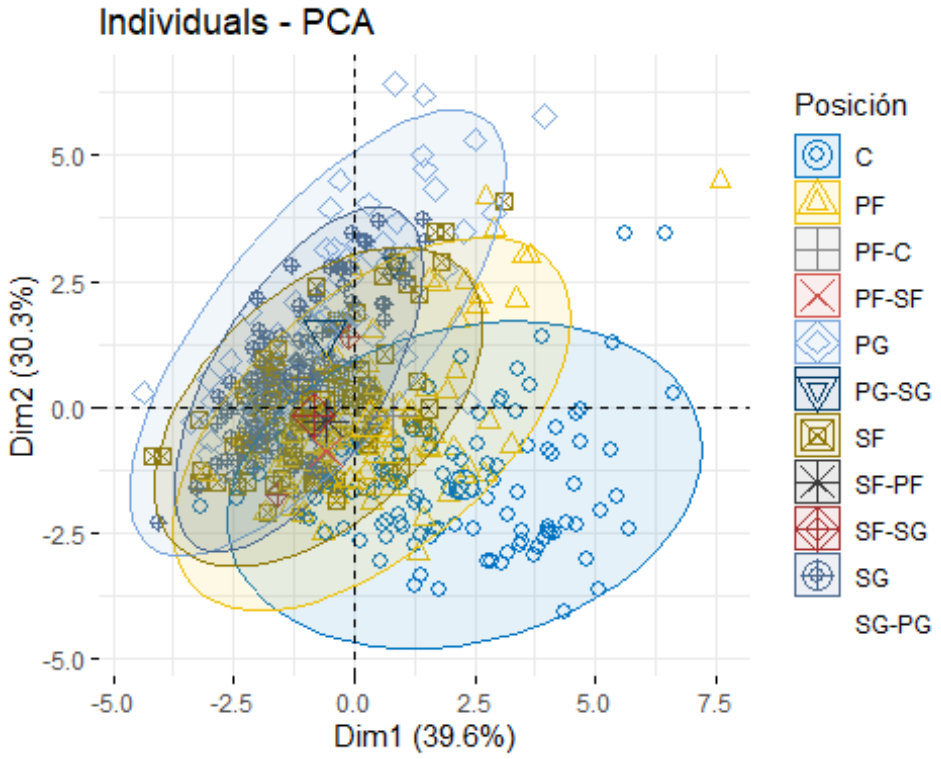


Anexo 5 Gráfico de loadings y scores:

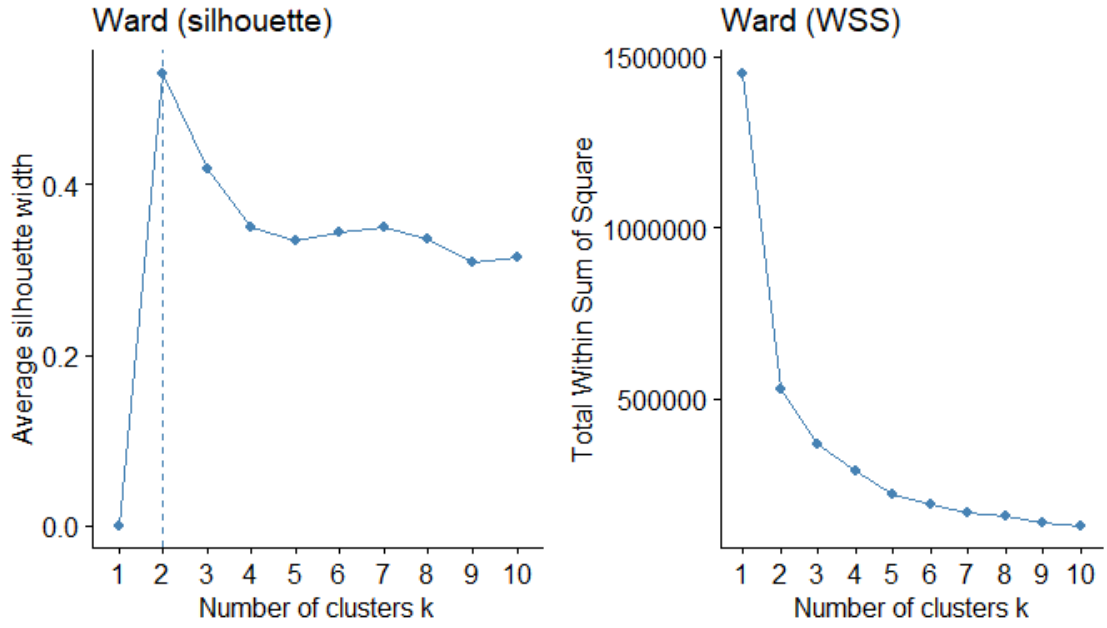


Anexo 6 Gráfico de loadings y scores:

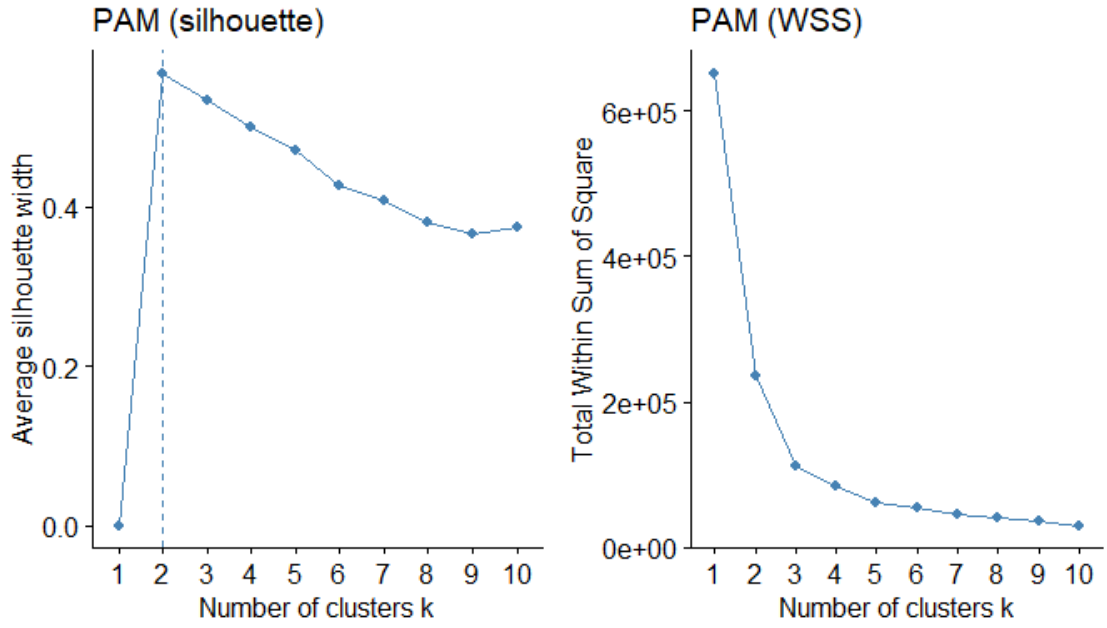




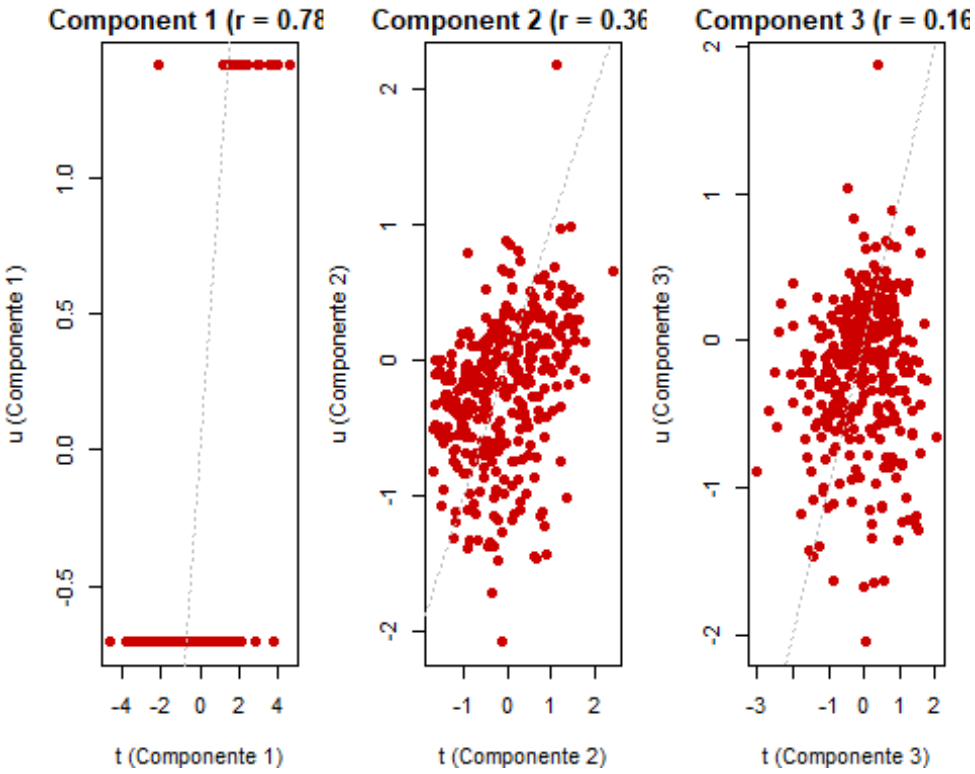
Anexo 7: Modelo jerárquico de Ward



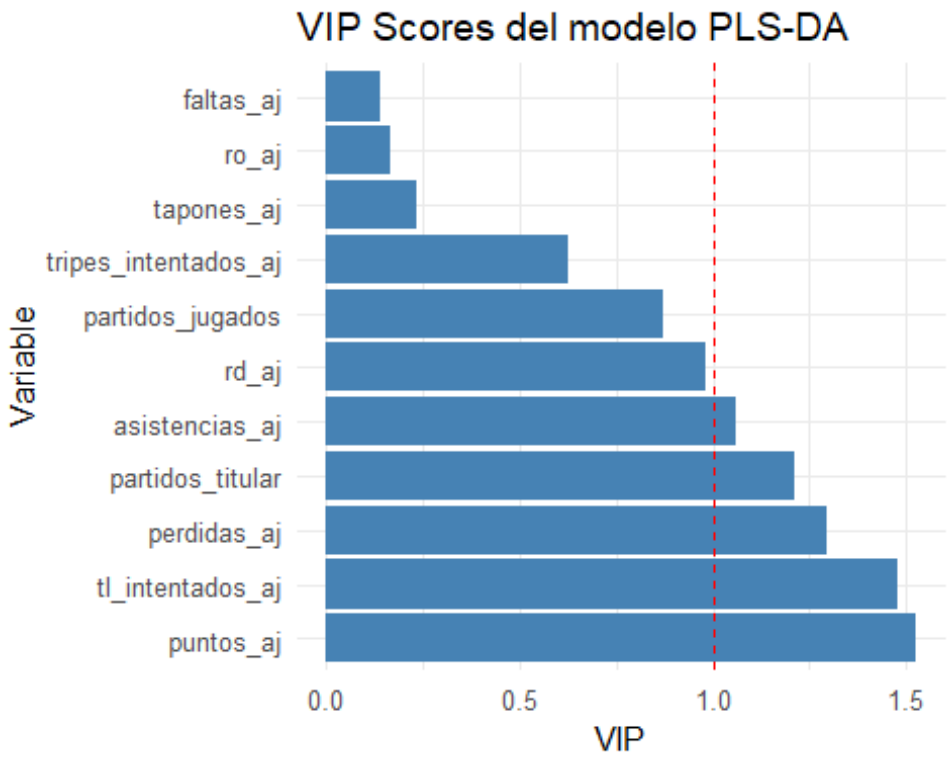
Anexo 8: Modelo jerárquico de Ward



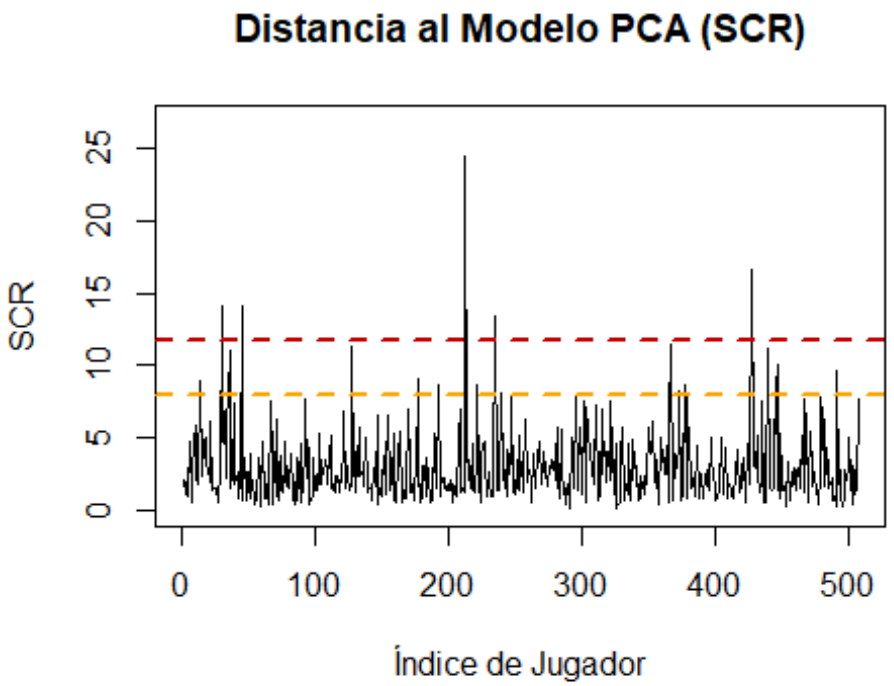
Anexo 9:



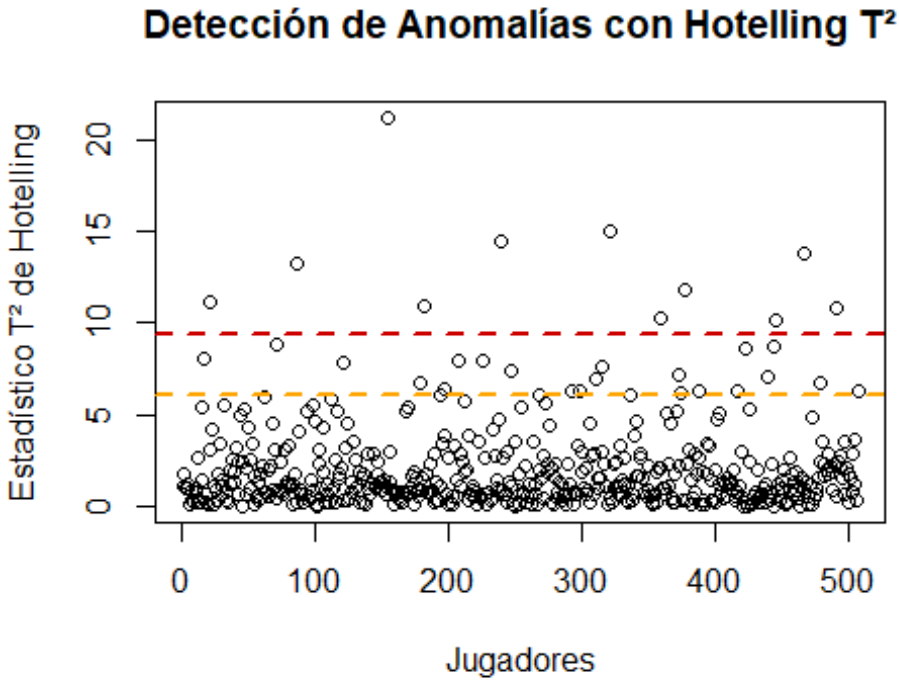
Anexo 10:



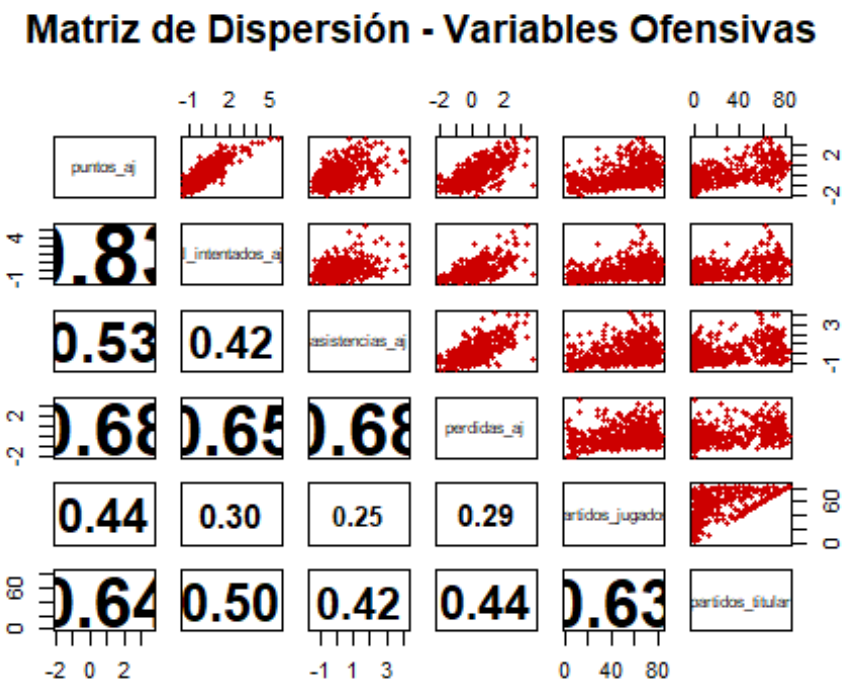
Anexo 11:

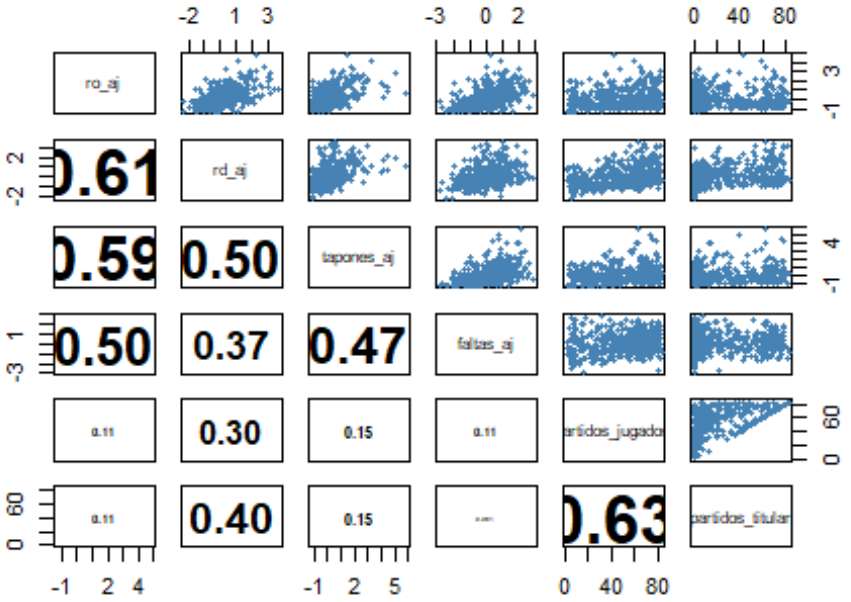


Anexo 12:



Anexo 13:


Matriz de Dispersión - Variables Ofensivas
puntos_aj
l_intentados_aj
asistencias_aj
perdidas_aj
0.44
0.30
0.25
0.29
0.50
0.42
0.44
0.53
0.42
0.44
0.50
0.30
0.25
0.29
0.50
0.42
0.44
0.63


Matriz de Dispersión - Variables Defensivas
ro_aj
rd_aj
tapones_aj
faltas_aj
0.61
0.59
0.50
0.50
0.37
0.47
0.11
0.30
0.15
0.11
0.11
0.40
0.15
0.63