Clasificación de Jugadores y Optimización de equipos de la NBA

Mauricio Alejandro Sierra Segovia

Facultad de Ciencias Físico Matematicas
Universidad Autonoma de Nuevo León

Abstract

El presente proyecto busca optimizar un equipo balanceado de la NBA con base en características específicas de jugadores, salarios y victorias esperadas. Para ello se utiliza la metodología de Componentes Principales para generar una reducción de variables, así como técnicas de clusterización para mejorar la clasificación de jugadores. Con base en estas características se intenta optimizar un equipo eficiente, que maximice el mayor número de victorias por el menor precio.
Keywords: Componentes Principales, Clúster, Optimización

Introducción

La optimización de equipos en la NBA se ha convertido en un enorme desafío que combina análisis estadístico, restricciones financieras y aunque no sea correcto muchisima intución de los directivos. Aunque el objetivo principal de las franquicias siempre es el campeonato, en los ultimos años se a visto una tendencia a buscar gastar lo menos posible. Esto a llevado a buscar maneras de calcular formas de maximizar el rendimiento colectivo del equipo por el menor precio. A día de hoy calcular la producción individual de los jugadores sigue siendo una tarea compleja debido a la naturaleza multidimensional del deporte: factores como eficiencia ofensiva, impacto defensivo no siempre se reflejan en métricas tradicionales. Aun más importante en los ultimos años hemos presenciado numerosos fracasos de construcción de rooster al intentar maximizar producción sin tomar en cuenta la sinergia colectiva. A estas dificultades se le suma el salary cap, un componente restrictivo que obliga a los equipos a equilibrar talento y presupuesto, limitando la posibilidad de simplemente acumular estrellas.

La literatura académica ha abordado este problema desde diferentes perspectivas. Por un lado, algunos estudios como **NBA Salaries Assessing True Player Value** han intentado establecer una relación entre la producción en cancha y el salario, buscando determinar si las compensaciones reflejan el verdadero valor aportado por los jugadores. Por otro lado, trabajos como **Using K-Means Clustering to Identify NBA Player Similarity** han explorado técnicas de machine learning, como el clustering, para agrupar jugadores según sus características y roles, ofreciendo herramientas para construir plantillas más eficientes. Estas aproximaciones evidencian que la optimización no depende únicamente de métricas individuales, sino de comprender patrones colectivos y restricciones económicas.

Descripción de los Datos

Descripción de Variables

Para el desarollo de este proyectro se obtuvo una base de datos de cerca de 600 jugadores de la NBA para la temporada 2024-2025 con diferentes metricas relacionadas al rendimiento, salarios y aportaciones a las victorias.

Primeramente se desarollo una base de datos por medio de un webscrapping al sitio de ESPN para obtener los salarios de los jugadores de la NBA. A su vez se descargo un conjuto de datos proporcionado por Stathead el cual incluia las métricas por jugador ”normalizadas” por cada 48 minutos. Esto nos ayudo a poder extrapolar la producción de todos los jugadores a un nivel comparable haciendo a un lado la cantidad de tiempo que juegan.

Para el procesamiento de estos datos se hicieron modificaciones a as cadenas de texto para quitar caracteres extraños, acentos o simbolos. Esto permitio poder hacer una union entre ambos sets de datos quedandonos con una base final de **436** jugadores distintos. Los registros perdidos se deben a que algunos jugadores presentaban diferencias muy especificas entre sus nombres en ambas bases o simplemente no se encuentran en alguna de las dos.

Entre nuestras variables de interes, encontramos datos de eficiencia, producción, impacto, ofensiva, defensa y hasta variables de disponiblidad. Todas las variables pueden verlas en la Tabla 1 con su nombre y decripción

Table 1: Set de Datos

Métrica	**Descripción**
WS (Win Shares)	Número de Victorias Producidas (Estimación)
USGBPM (Box Plus Minus)	Número de diferencia de puntos del jugador en cancha
VORP (Value Over Replacement)	Número de puntos que genera por encima de su cambio
PER (Player Eficienci Rating)	Medida estandarizada de Producción
Age	Edad
2PA	Tiros de 2 Puntos por Partido
2PP	Porcentaje de Tiros de 2 Puntos Anotados
3PA	Tiros de 3 Puntos por Partido
3PP	Porcentaje de Tiros de 3 Puntos Anotados
FTA	Tiros Libres por Partido
FTP	Porcentaje de Tiros Libres Anotados
ORB	Rebotes Ofensivos
DRB	Rebotes Defensivos
AST	Asistencias
STL	Robos
BLK	Tapones
TOV	Pérdidas
PF	Faltas Personales
PTS	Puntos
Pos	Posición
Salary	Salario por Temporada
Reliability	Porcentaje de Partidos de Temporada Regular en los que juega

Caracteristicas encontradas

Al procesar estas variables se calcularon sus medias y varianzas asi como tambien se les aplico la Prueba de normalidad de Shapiro-Wilk a todas ellas. En esta prueba se concluyo que ninguna de las varibales era parametica. Graficamos los histogramas y los qq plots para ver la distribución de nuestras variables. la gran mayoria de ellas presentan resultados como los vistos en las graficas 1 donde paracen ser normales pero cuentan con una cola larga y más pronunciada en uno de sus extremos que como puede ser apreciado tanto en el QQ plot como el histograma nos confirma que nuestras variables no se distribuyen de manera normal.

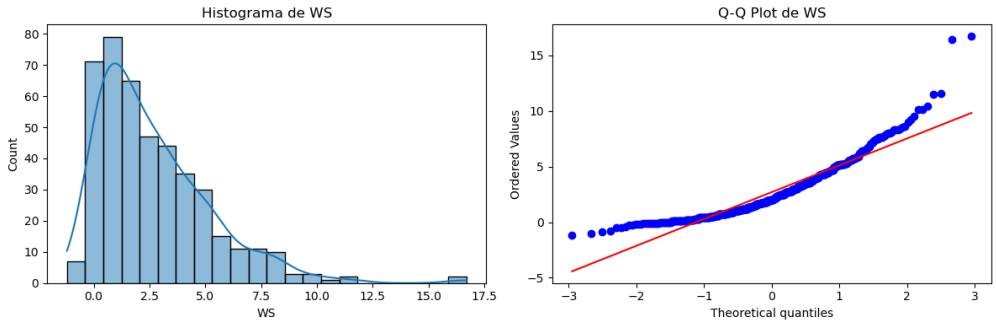

Figure 1: Graficos de distribución WS

Por ultimo podemos ver el Grafico 2 que presenta las correlaciones calculadas para los datos. Entre las relaciones más fuertes, se observa que Win Shares (WS) mantiene una alta correlación positiva con Puntos (PTS) (0.78) y con Valor sobre reemplazo (VORP) (0.85), lo que indica que jugadores que generan más victorias tienden a tener mayor producción ofensiva y un impacto global superior. Asimismo, BPM y VORP presentan una correlación muy elevada (0.91), reflejando que ambas métricas capturan dimensiones similares del rendimiento. En cuanto a eficiencia, PER muestra correlaciones significativas con WS (0.71) y PTS (0.67), sugiriendo que la eficiencia individual está estrechamente ligada tanto al aporte en victorias como a la anotación. Por otro lado, el Salary exhibe correlaciones moderadas con métricas de impacto como WS (0.53) y VORP (0.57), lo que indica que el salario tiende a reflejar el rendimiento. Respecto a estadísticas tradicionales (es decir las que no llevan formulas de por medio), destaca la relación entre Intentos de tiro libre (FTA) y PTS (0.76), lo que confirma la importancia de generar faltas para incrementar la producción ofensiva (caracteristica muy criticada entre los mejores anotadores a la vez que muy valoradas). También se observa que Asistencias (AST) correlacionan positivamente con BPM (0.45), sugiriendo que la creación de juego contribuye al impacto global. Finalmente, algunas correlaciones negativas son relevantes como lo es la Edad con USG (-0.33) y PTS (-0.32), lo que indica que el uso ofensivo y la anotación tienden a disminuir con la edad.

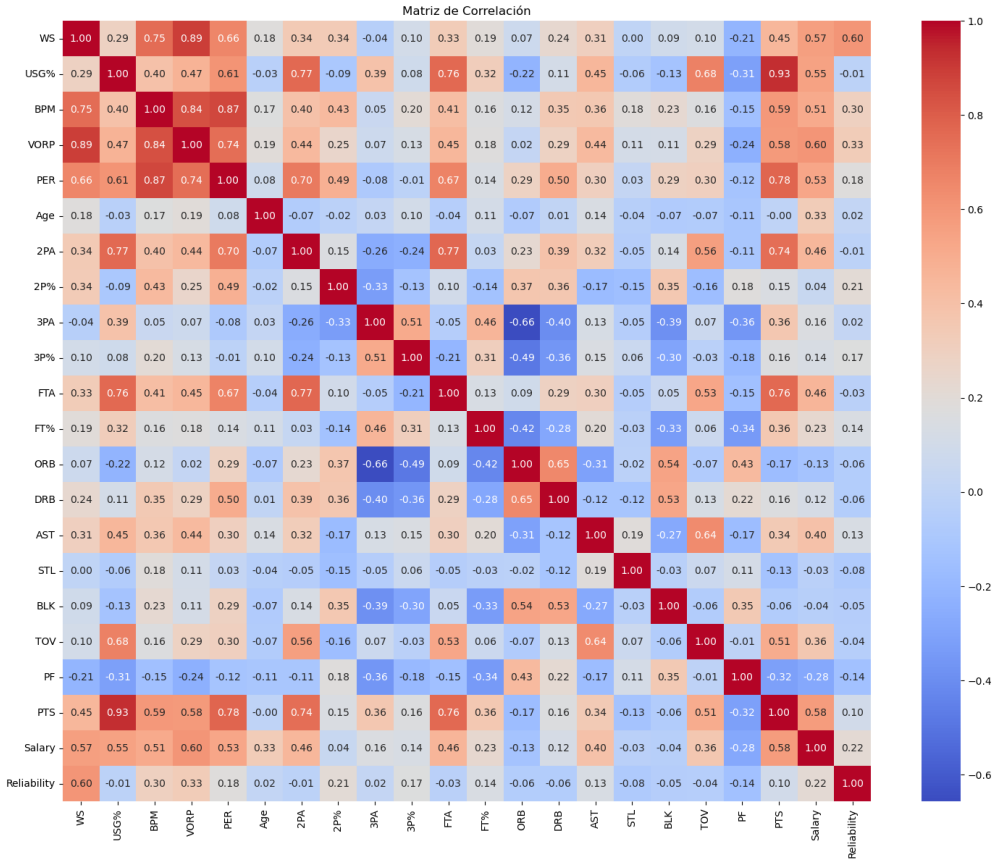

Figure 2: Graficos de Correlación

Metodología

El enfoque metodológico se divide en dos etapas principales:

- Agrupamiento de jugadores según sus características.
- Optimización de la composición del equipo para maximizar el rendimiento (medido en Win Shares), minimizar el costo salarial y garantizar diversidad de estilos de juego.

Dado que el objetivo central del análisis está vinculado a la primera etapa, se explicará en detalle el proceso de agrupamiento, mientras que la segunda se abordará de manera general.

Primera Etapa: Agrupamiento de Jugadores

El objetivo de esta fase es identificar grupos homogéneos de jugadores en función de sus características estadísticas y métricas avanzadas. Para ello, se emplearon técnicas de reducción de dimensionalidad y posteriormente algoritmos de clustering.

Reducción de Dimensionalidad mediante Análisis de Componentes Principales (PCA)
El PCA se utilizó inicialmente para transformar el conjunto original de variables en un número reducido de componentes que capturen la mayor varianza posible. Este método permite:

Objetivo: Simplificar la estructura de los datos y facilitar la agrupación, evitando redundancia entre variables altamente correlacionadas.

Para la aplicación del modelo se generaron los componentes principales. Una vez aplicada la metodología se obtuvieron los siguientes resultados.

Table 2: Resultados de CP

Componente	Eigenvalor	Varianza Explicada	Varianza Acumulada
PC1	5.844	29.15	29.15
PC2	4.105	20.48	49.63
PC3	2.059	10.27	59.90
PC4	1.427	7.12	67.02
PC5	1.046	5.22	72.24

En la tbala 2 (Se omiten el resto de Componemtes) podemos ver el número de componentes que se obtuvo de hacer la prueba. Dada la cantidad de componentes se considero que era muy pequeño para ayudar a la metodología de clusters a clasificarlos correctamnete. Además de ello se detectaron problemas como que los datos de origen no cumplían con la normalidad multivariada segun la prueba mardia. Aunque no es un problema que desacredite el modelo si lo vuelve menos efectivo. Además, el estudiar los logs de los 5 componentes se concluyo que los componentes principales no eran realemnte interpretables, lo que haria los resultados un poco dudosos. Por ello se decidio descartar esta metodología y pasar directamente a la clusterisación.

Clustering
Para aplicar esta metodología primeramente aplicamos la nomarilzación de los datos, esto no hizo que se agruparan de manera normal pero si, garantizando que todas las variables tengan la misma escala y evitaran sesgos por magnitudes diferentes.

Determinación del número óptimo de clústeres
Se emplearon pruebas y métricas específicas para evaluar la calidad del agrupamiento buscando optimizar el mejor número de clusters.

Se utilizo el **Método del Codo** el cual analiza la variación explicada en función del número de clústeres, buscando el punto donde la mejora marginal se reduce significativamente.

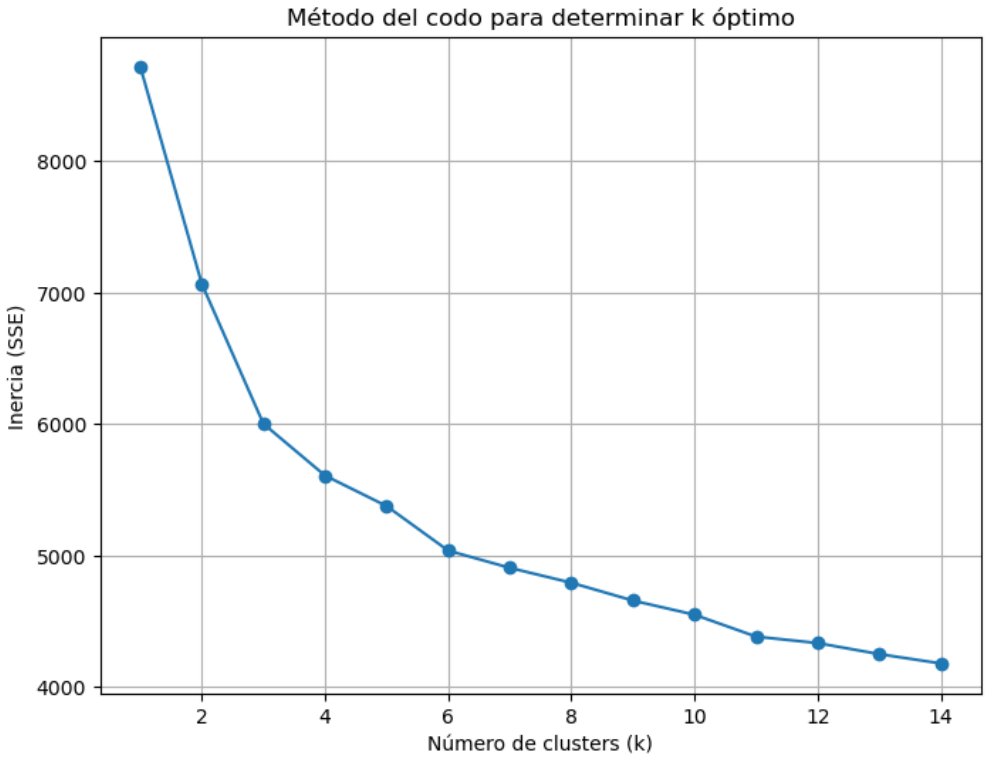

Figure 3: Graficos de Codo

En este caso podemos ver 2 reducciones marginales importantes una en el punto 3 y otra en el punto 6.

Calinski-Harabasz y Davies-Bouldin, Complementan la evaluación de la compacidad y separación de los grupos. Calinski-Harabas mide la dispersión de los clusters y podemos interpretarlo como que a mayor valor, mayor dispersion entre los clusters. Mientras que Davies-Bouldin mide la dispersión interna del cluster lo que nos hace buscar un valor bajo pues este representa una mejo compacidad.

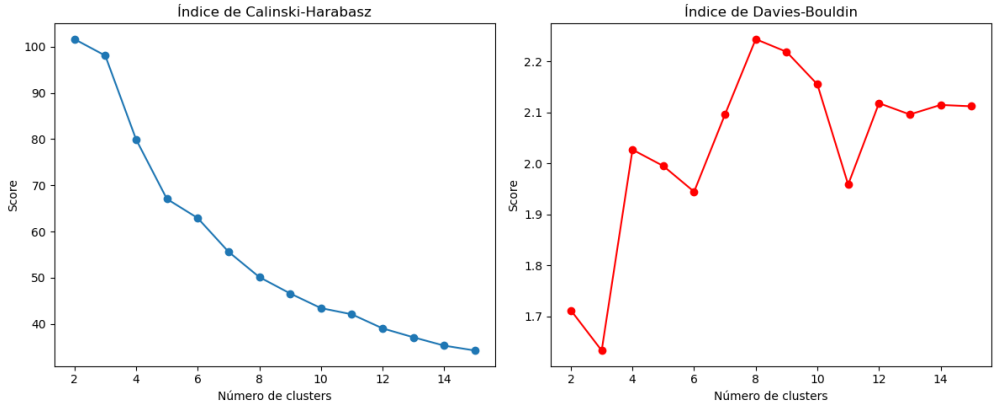

Figure 4: Graficos Calinski-Harabasz y Davies-Bouldin

En este caso nuestros resultados encuentran el punto más alto de Calinski en el punto 3 y Davies en el punto 3. Ya con los resultados de estas 3 pruebas concluimos que el punto optimo seria usar 3 clusters.

Con esto aplicamos clustering directamente sobre los datos normalizados, sin reducción previa, para preservar la riqueza de las características originales.

Subsecuentemente sacamos las medias de todas las variables en cada uno de los clusters para detectar cuales son las caracteristicas de los grupos formados:

- **Big Men** jugadores con alto porcentaje de acieto cerca de la canasta, muchos rebotes, tapones y por ende más faltas pues suelen jugar de manera más fisica.
- **Creador Ofensivo** Son los jugadores con todas las caracteristicas ofensivas en la parte alta de la curgva, los más eficientes, mejores porcentajes, más tiros, más puntos y más asistencias. Su principal carcteristica negativa son sus perdidas de balón
- **3 and D** Son jugadores con alto porcentaje de triples y buenas estadisticas de defensa, su rol suele ser espaciar la pintura al ser una amenaza de tiro, generan que la cancha se abra más para los Creadores Ofensivos

Aqui podemos ver un ejemplo de como luce la clusterización obtenida, donde 0 son Big Men, 1 Creadores ofensivos y 2 son los 3 and D.

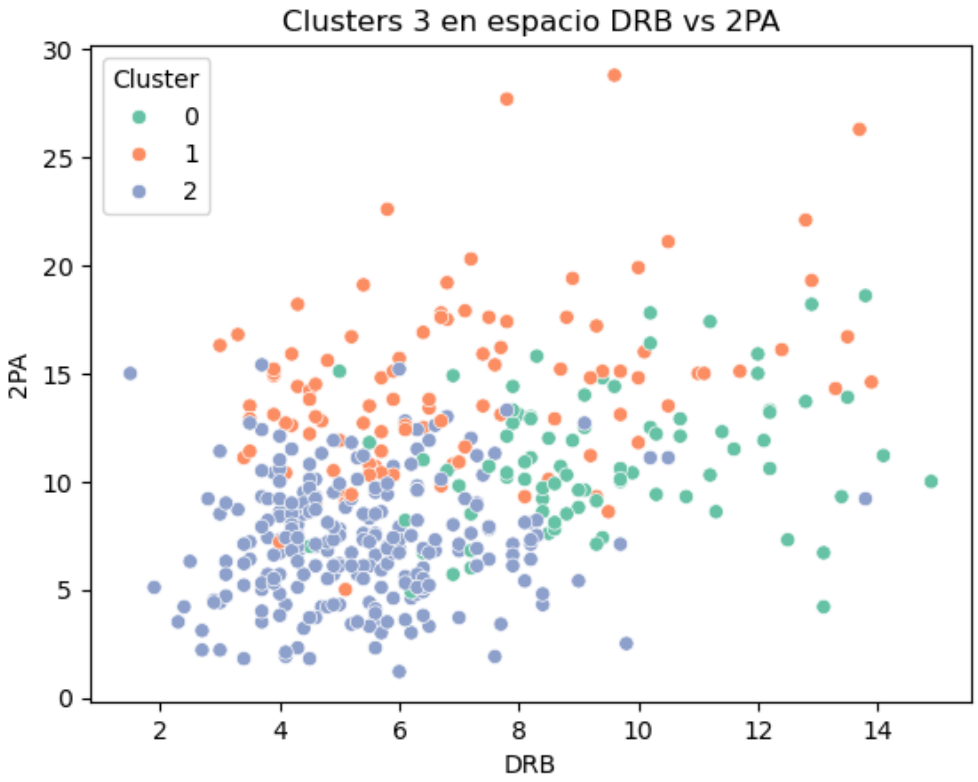

Figure 5: Resultados Clusterizados

Optimización
Por ultimo y con base en estos resultados, decidimos de manera arbitraria las condiciones que agregaremos al modelo de Optimización:

- **Objetivo principal:** Maximizar la suma de Victorias Esperadas
- Salary Floor - La suma de los salarios tiene que ser mayor a 126.529 millones
- Salary Cap - La suma de los salarios tiene que ser menor a 140.588 millones
- Solo se pueden seleccionar 3 Creadores ofensivos
- Solo se pueden seleccionar 5 Big Men
- Solo se pueden seleccionar 7 3 and D

El resultado de este modelo fue un equipo de 15 jugadores con una suma de victorias proyectadas de 115.9 (Solo se juegan 82 partidos) y un costo total de 136.689 millones.

Este roster esta conformado por:

Table 3: RPlantilla Optimizada

Jugador	**Aporte a Victorias**	**Parte del Salario**
Shai Gilgeous-Alexander	0.14	0.26
Jarrett Allen	0.10	0.15
Evan Mobley	.08	0.08
Payton Pritchard	0.07	0.05
Alperen Sengun	0.07	0.04
Christian Braun	0.07	0.02
Amen Thompson	0.07	0.07
Onyeka Okongwu	0.06	0.10
Luke Kornet	0.06	0.02
Chris Paul	0.05	0.08
Walker Kessler	0.05	0.02
Cason Wallace	0.04	0.04
Toumani Camara	0.04	0.01
Dyson Daniels	0.04	0.04
Keon Ellis	0.04	0.02

Resultados

Los resultados arrojan cosas muy interesnates en primera solo 5 jugadores tienen un porcentaje del salario contra el total mayor a su porcentaje de aporte a las victorias. Pero dos de ellos tienen aportes de al menos el doble.

Pero la conclusión más importante que podemos obtener es que aparte de que para los conocedores del deporte parece ser un rooster muy balanceado, los jugadores en contratos de novatos (Duran 4 años) parecen ser los más valiosos pues parecen ser los que tienen una relación más acorde entre su producción y su salario.

Conclusiones

Aunque la cantidad de clusters que obtuvimos no fue la que esperabamos los resultados parecen ser bastante buenos para ser una primera versión del estudio. En un futuro podriamos hacer analisis más profundos usando metricas más avanzadas como spots de tiro, número de pantallas y estadisticas más acorde al rol en si de cada uno de los jugadores. Aplicando este mismo metodo podriamos encontrar aun más diferencia en los etilos de juego.

Tambien seria muy interesante aplicar estos casos con restricciones más fuertes, como ver que jugadores no estan en el mercado, revisar solo jugadores que esten finalizando contrato. O bajarlo a nivel colegial y usarlo para hacer scouting de posibles jugadores a seleccionar en el draft.

Las posibilidades de estudios subsecuentes son bastante amplias.

Bibliography

- Ghirardo, M. (2013). NBA Salaries: Assessing True Player Value. Cal Poly. https://digitalcommons.calpoly.edu/statsp/35/
- McHun, A. (2020, August 3). Using K-means clustering to identify NBA player similarity. Medium. https://medium.com/@allenmchun/using-k-means-clustering-to-identify-nba-player-similarity-2b33f11e3aa7