UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf
Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

ESTUDIO ESTADÍSTICO DE LA NBA

Proyecto de la asignatura

Proyecto I, Comprensión de datos

Grado en Ciencia de Datos

Autores/as:

Hurtado Beneyto, Marc

Tarrasó Sorní, Aleixandre

Torres López, Pablo

Muedra Vela, Jorge

Zhan, Lianghao

Tutor:

Francisco Pedroche

1º Ciencia de Datos

Resumen

En este proyecto buscamos realizar un estudio estadístico de la NBA, desde sus inicios en 1940 hasta la actualidad, por lo que tratamos de mostrar la evolución y cómo se ha ido cambiando el juego, los equipos y los jugadores a lo largo del tiempo.

El objetivo principal de este estudio es analizar la liga en general, desde la competitividad de las diferentes temporadas y establecer correlaciones entre los premios individuales otorgados a los jugadores y su desempeño estadístico, y como realizar un análisis de componentes principales o PCA para reducir el número de variables y, por último, un modelo predictivo para el MVP.

Lo primero es examinar la evolución de los jugadores a lo largo de los años. Analizamos variables como el porcentaje de tiros de campo, los rebotes y las asistencias. Para ello, hemos utilizado técnicas de análisis estadístico para identificar tendencias y patrones en los datos.

También, hemos estudiado la evolución de los equipos a lo largo del tiempo. Aquí analizamos variables como el número de victorias, el promedio de puntos anotados y recibidos. Y también hemos identificado equipos dominantes en diferentes épocas.

Uno de los aspectos más interesantes de este estudio es la evaluación de la competitividad de cada temporada de la NBA. Para ello, hemos utilizado índices de competitividad, como pueden ser el HICB o la índice sigma (ver referencia [7]). Básicamente comparamos las temporadas entre si y determinamos cuáles han sido más o menos competitivas.

Además, realizamos un análisis de componentes principales para reducir el número de estas que determinan el desempeño de los jugadores y los equipos.

Añadido a todo esto, es el modelo predictivo del MVP lo hemos realizado mediante Python, más concretamente con la librería Pandas para obtener el DataSet necesario para entrenar al modelo, y sklearn otra librería donde básicamente se usa para hacer el modelo y obtener la predicción.

Por otra parte, hemos examinado las correlaciones entre los premios individuales otorgados a los jugadores, como el MVP (Jugador Más Valioso), el Jugador Defensivo del Año, el SMOY (El mejor sexto hombre), y su desempeño estadístico. El empleo de técnicas estadísticas para determinar qué variables tienen una mayor relación con la obtención de estos premios, nos ha ayudado a la hora de conseguir y analizar estos datos.

Por último, hemos identificado que jugadores dentro de los máximos exponentes de las diferentes estadísticas han sido más determinantes para las victorias del equipo.

Resum

En aquest projecte busquem dur a terme un estudi estadístic de l'NBA, des dels seus inicis el 1940 fins a l'actualitat, amb l'objectiu de mostrar l'evolució i com les tendències del joc, els equips i els jugadors s'han adaptat al llarg del temps.

L'objectiu principal d'aquest estudi és analitzar la lliga en general, des de la competitivitat de les diferents temporades fins a establir correlacions entre els premis individuals atorgats als jugadors i el seu rendiment estadístic, com ara realitzar un anàlisi de components principals (PCA) per reduir el nombre de variables i, finalment un model predictiu per al MVP.

El primer que fem és examinar l'evolució dels jugadors al llarg dels anys. Analitzem variables com el percentatge d'encert en tirs de camp, rebots i assistències. Per a fer-ho, hem utilitzat tècniques d'anàlisi estadística per identificar tendències i patrons en les dades.

També hem estudiat l'evolució dels equips al llarg del temps. Ací hem analitzat variables com el nombre de victòries, la mitjana de punts anotats i punts rebuts. I també, hem identificat equips dominants en diferents èpoques.

Un dels aspectes més interessants d'aquest estudi és l'avaluació de la competitivitat de cada temporada de l'NBA. Per a això, hem utilitzat índexs de competitivitat, com l'HICB o l'índex sigma (veure referència [7]). Bàsicament, hem comparat les temporades entre elles i hem determinat quines han estat més o menys competitives.

A més, realitzem una anàlisi de components principals per reduir el nombre de components principals que determinen el rendiment dels jugadors i dels equips.

Afegit a tot això, el model predictiu del MVP l'hem realitzat utilitzant Python, més concretament amb la llibreria Pandas per obtenir el conjunt de dades necessari per entrenar el model, i sklearn, una altra llibreria que bàsicament s'utilitza per a construir el model i obtenir la predicció.

D'altra banda, hem examinat les correlacions entre els premis individuals atorgats als jugadors, com ara MVP (Jugador Més Valuós), Jugador Defensiu de l'Any, Millor Sisé Home, i el seu rendiment estadístic. L'ús de tècniques estadístiques per determinar quines variables tenen una relació més gran amb l'obtenció d'aquests premis ens ha ajudat a obtenir i analitzar aquestes dades.

Finalment, hem identificat quins jugadors dins dels millors rendiments en les diferents estadístiques han estat més determinants per a les victòries de l'equip.

Abstract

In this project we seek to conduct a statistical study of the NBA, from its beginnings in 1940 to the present, we seek to show the evolution and how the trends of the game, teams and players have been adapting over time.

The main objective of this study is to analyze the league in general, from the competitiveness of the different seasons and establish correlations between the individual awards given to players and their statistical performance, such as performing a principal component analysis or PCA to reduce the number of variables and finally a predictive model for the MVP.

The first thing is to examine the evolution of the players over the years. We will analyze variables such as shooting percentage from the field, rebounds and assists. To do so, we have used statistical analysis techniques to identify trends and patterns in the data.

Also, we have studied the evolution of the teams over time. Here we analyzed variables such as number of wins, average points scored, and points received. And we have also identified dominant teams in different eras.

One of the most interesting aspects of this study is the evaluation of the competitiveness of each NBA season. For this purpose, we have used competitiveness indexes, such as the HICB or the sigma index (see reference [7]). Basically, we compared the seasons with each other and determined which ones have been competitive.

In addition, we perform a principal component analysis to reduce the number of principal components that determine the performance of players and teams.

Added to all this, is the predictive model of the MVP we have done it using Python, more specifically with the Pandas library to obtain the DataSet needed to train the model, and sklearn another library where basically it is used to make the model and obtain the prediction.

On the other hand, we have examined the correlations between individual awards given to players, such as MVP (Most Valuable Player), Defensive Player of the Year, SMOY (Best Sixth Man), and their statistical performance. The use of statistical techniques to determine which variables have a greater relationship with obtaining these awards has helped us in obtaining and analyzing this data.

Finally, we have identified which players within the top performers of the different statistics have been more determinant for the team's victories.

Tabla de contenidos

1. Introducción ……..p.7
    1.1 Motivación ……..p.7
    1.2 Objetivos ……..p.7
    1.3 Metodología ……..p.8
    1.4 Estructura ……..p.9

2. Estado del arte ……..p.9
    2.1 Crítica a estado del arte ……..p.9
    2.2 Propuesta ……..p.12

3. Preparación y comprensión de datos ……..……p.12

4. Alcance del proyecto ……..p.14

5. Calendario del proyecto ……..p.15

6. Conocimiento extraído ……..……p.16

    6.1 Variables ……..……..……..……..……..p.19

    6.2 Análisis descriptivo ……..……..……..……..…p.20

    6.3 Análisis estadístico ……..……..……..……..p.24

    6.4 Modelo predictivo.....p.44

7. Validación y despliegue ……..p.47

8. Conclusiones ……..p.47

9. Referencias ……..p.48

10. Anexos ……..p.50

1. Introducción

El baloncesto es uno de los deportes más populares del mundo y la NBA (National Basketball Association) es la liga de baloncesto profesional más importante a nivel mundial. La NBA cuenta con una gran cantidad de seguidores y aficionados en todo el mundo, generando una enorme cantidad de datos e información sobre sus jugadores, equipos y partidos. En este contexto, los proyectos de ciencias de datos aplicados a la NBA tienen un gran potencial para descubrir información valiosa y útil que puede ayudar a los equipos a tomar decisiones estratégicas y mejorar su rendimiento. En este proyecto, se busca utilizar técnicas de análisis de datos para entender mejor el desempeño de los jugadores y equipos de la NBA, así como para predecir el resultado de los partidos y el éxito de las estrategias utilizadas por los equipos. El objetivo final es proporcionar información útil para los entrenadores, jugadores y directivos de los equipos para mejorar su rendimiento y conseguir mejores resultados en la liga.

1.1. Motivación

Este proyecto de ciencias de datos sobre la NBA nos motiva al permitirnos aplicar nuestros conocimientos en un contexto real y relevante. Analizar los datos de esta reconocida liga de baloncesto nos brinda la oportunidad de tomar decisiones informadas, mejorar el rendimiento de los jugadores y desarrollar habilidades clave en ciencias de datos. Además, nos enfrentamos a desafíos reales y adquirimos experiencia valiosa para nuestra futura carrera profesional. En resumen, estamos emocionados por sumergirnos en este proyecto y explorar las infinitas posibilidades que nos ofrece.

1.2. Objetivos

El objetivo central de este proyecto es estudiar algunas métricas deportivas de la NBA para entender las tendencias y patrones en el desempeño de los equipos y jugadores. En el proyecto se analizan las correlaciones y se aplican métodos estadísticos para determinar las relaciones entre los cambios de reglas implementados en la NBA y su impacto en las estadísticas del juego.

1.3. Metodología

1. Definición del objetivo: En el presente proyecto se analiza el desempeño de los jugadores en función de diferentes variables, como puntos anotados, asistencias, rebotes, eficiencia en tiros, entre otros.

2. Recopilación de datos: Los datos provienen de diversas fuentes confiables, como bases de datos oficiales de la NBA, registros de partidos, estadísticas individuales y de equipo, entre otros.

3. Análisis descriptivo: Una vez que tenemos los datos, realizamos un análisis descriptivo para obtener una visión general de las variables seleccionadas. Esto implica, calcular medidas de tendencia central como promedio, mediana y moda, así como medidas de dispersión, como desviación estándar y rango. También creamos gráficos y visualizaciones para resumir y representar los datos de manera efectiva.

4. Análisis estadístico: En esta etapa, aplicamos técnicas estadísticas para obtener un mayor nivel de comprensión y profundidad en nuestro análisis. Vamos a realizar pruebas de hipótesis para determinar si existen diferencias significativas entre grupos de jugadores o equipos en relación con las variables seleccionadas. También utilizamos técnicas de correlación para evaluar las relaciones entre diferentes variables y su impacto en el rendimiento de los jugadores.

5. Interpretación de resultados: Una vez que hemos realizado el análisis descriptivo y estadístico, interpretamos los resultados obtenidos. Identificamos patrones, tendencias y relaciones relevantes que puedan ayudarnos a comprender mejor el desempeño de los jugadores en la NBA.

6. Conclusiones y recomendaciones: En esta etapa, extraemos conclusiones clave de nuestro análisis y formulamos recomendaciones basadas en los resultados obtenidos. Ahora identificamos qué variables tienen un mayor impacto en el rendimiento de los jugadores y recopilamos algunas curiosidades sobre la NBA.

1.4. Estructura

- Revisión del estado del arte: Se realiza una investigación sobre cómo se han abordado los objetivos planteados anteriormente y se identifican otros trabajos o referencias relacionados con estos objetivos.

- Conjunto de datos: Se describe el conjunto de datos utilizado en este proyecto, incluyendo su fuente y las características de su composición.
- Alcance del proyecto: Se establecen los requisitos, restricciones, supuestos y entregables del proyecto, así como los límites de lo que se aborda y lo que no se aborda.
- Calendario del proyecto: Se definen las fechas límite para cada etapa del proyecto.
- Resultados: Se presentan los resultados obtenidos de forma concisa.
- Conclusiones: Se comentan las conclusiones obtenidas en el estudio de manera resumida.
- Impacto: Se discute cómo los resultados y conclusiones del estudio pueden ser aplicados al contexto de la NBA.
- Propuestas de trabajo futuro: Se sugieren posibles líneas de investigación para continuar trabajando en el problema general o en objetivos específicos.

2. Estado del arte

2.1.- Crítica al estado del arte

El análisis de datos aplicado al ámbito de la NBA ha experimentado un crecimiento significativo en los últimos años. El avance de la tecnología y la disponibilidad de grandes volúmenes de datos han brindado nuevas oportunidades para comprender y mejorar el rendimiento de los jugadores y equipos en esta popular liga de baloncesto.

Tras realizar una búsqueda minuciosa de proyectos similares a través de la plataforma Google Scholar, hemos llegado a la conclusión de que estos dos eran los más similares a nuestra idea inicial del proyecto.

2.1.1.- La eficacia del lanzamiento a canasta en la NBA: Análisis multifactorial. Referencia [9]

Este estudio se ha realizado para analizar la eficacia de los lanzamientos de baloncesto en diferentes situaciones y características de los jugadores. Se ha encontrado que la eficacia en los lanzamientos de 1 punto es menor que en los de 2 y 3 puntos, y que los bases y aleros tienen una mayor eficacia que los pívots en los lanzamientos de 1 punto.

Además, encontramos que la eficacia en los lanzamientos de 2 y 3 puntos está relacionada con diferentes factores, como el período, el cuarto, la gesto forma, la presión defensiva, la zona del lanzamiento, el rol del jugador y la acción previa. Los resultados del estudio indican que los entrenadores deberían diseñar sesiones de entrenamiento más realistas y adecuadas a las características de la competición analizada, para mejorar la eficacia en los lanzamientos de baloncesto.

FIG 1. Esta imagen hace referencia a las diferentes zonas de tiro.

2.1.2 - Análisis de la eficiencia en los equipos de la NBA para las temporadas 2014/2015 y 2015/2016. Referencia [10]

La relación entre eficiencia y resultados deportivos es objeto de estudio en numerosos trabajos a través del análisis envolvente de datos (DEA). En este trabajo se utiliza un modelo DEA radial bajo orientación de entrada y rendimientos constantes a escala. Se toma como muestra los equipos de la NBA en las temporadas 2014-2015 y 2015-2016, diferenciando las dos fases que tiene la NBA (fase regular y fase Playoffs). Con estos datos se realiza tanto el cálculo de su eficiencia como el cálculo de la reducción que tiene que realizar cada equipo de cada entrada para ser eficiente. Además, se crea una muestra donde se unen ambas temporadas para obtener 60 unidades de objeto de análisis.

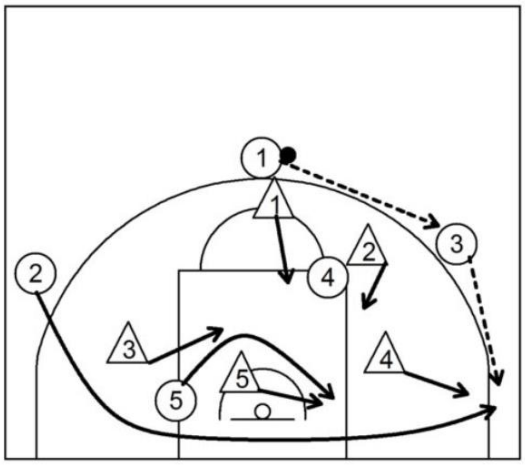


FIG 2. Esta imagen hace referencia al movimiento de los jugadores en ataque, y a las zonas donde suelen tirar.

2.1.3 - MÉTODOS ESTADÍSTICOS EN COMPETICIONES DEPORTIVAS DE BALONCESTO: LA NBA. Referencia [11]

Este documento es un trabajo de fin de grado (grado en estadística) y se enfoca en el análisis de las métricas estadísticas más conocidas y utilizadas por los analistas de la NBA. Además, se explica detalladamente cómo se llevan a cabo las experimentaciones para medir el rendimiento de los equipos y cómo se utilizan distintos métodos de predicción y procedimientos de agrupamiento para obtener resultados precisos.

El objetivo principal del trabajo es comprender cómo funcionan las métricas estadísticas en el baloncesto profesional y cómo pueden ser utilizadas para predecir el éxito y el rendimiento futuro de un equipo. Se espera que los resultados obtenidos sean útiles para entrenadores, jugadores, analistas y aficionados al baloncesto en general.

2.14 - A new method for comparing rankings through complex networks: Model and analysis of competitiveness of major European soccer leagues. Referencia [13]

En este estudio se presenta un nuevo método para comparar clasificaciones a través de redes complejas, específicamente en el contexto de la competitividad de las principales ligas de fútbol europeas. Se introduce el concepto de competidores efectivos y se presentan las propiedades estructurales principales del grafo de competitividad.

2.2.-Propuesta

La NBA es una de las ligas de baloncesto más importantes y competitivas del mundo. En este trabajo se propone realizar un análisis del estado del arte de la NBA, centrándose en el estudio de su evolución general y cómo las reglas del juego han cambiado a lo largo del tiempo. Además, se compara la competitividad de la NBA con otras ligas de baloncesto importantes. Para llevar a cabo este análisis, se realiza una revisión bibliográfica de fuentes secundarias, incluyendo artículos científicos, libros y publicaciones especializadas en baloncesto y estadísticas. Se estudian los cambios más significativos en las reglas de la NBA a lo largo de su historia y se analiza cómo estos han afectado al juego y a su competitividad. Además, se recopilan datos estadísticos para complementar el análisis. El objetivo final es proporcionar una visión completa y actualizada del estado del arte de la NBA y su competitividad en comparación con otras ligas.

3. Preparación y comprensión de datos

A continuación, se describen los pasos generales seguidos en esta etapa:

1. Recopilación de datos: Se identifican y recopilan diversas fuentes de datos relevantes para el proyecto. Estas pueden incluir bases de datos públicas de la NBA, registros estadísticos de partidos, datos de jugadores y equipos, así como datos de redes sociales y otras fuentes relacionadas con el baloncesto.

2. Exploración inicial: Se realiza una exploración inicial de los datos para comprender su estructura, características y posibles problemas. Esto implica examinar la distribución de variables, identificar datos faltantes o atípicos, y familiarizarse con las diferentes variables disponibles. En resumen, entender cómo funcionan los datos obtenidos.

3. Limpieza de datos: Se procede a la limpieza de los datos para eliminar cualquier inconsistencia, errores o datos no válidos. Esto puede incluir la eliminación de duplicados, la imputación de valores faltantes y la corrección de errores en los datos. Además, se aplican transformaciones necesarias, como la estandarización de formatos de datos o la codificación de variables categóricas.

4. Integración de datos: Si se tienen múltiples conjuntos de datos se realiza la integración de los mismos, asegurando que las variables correspondientes estén correctamente vinculadas. Esto puede requerir la identificación de claves de unión o la realización de operaciones de fusión para combinar los datos de manera adecuada.

5. Selección de variables: Se seleccionan las variables relevantes para el análisis en función de los objetivos del proyecto. Esto implica descartar variables no informativas o redundantes que no aporten valor al análisis. Además, se pueden generar nuevas variables derivadas a partir de las existentes si se considera necesario.

6. Análisis descriptivo: Se realiza un análisis descriptivo de las variables seleccionadas para obtener una comprensión profunda de los datos. Esto incluye el cálculo de medidas estadísticas descriptivas, como promedios, medianas, desviaciones estándar y percentiles. También se generan gráficos y visualizaciones para representar los datos de manera efectiva.

7. Análisis de estadístico: Se lleva a cabo un análisis de estadístico donde podemos diferenciar varios pasos como puede ser un análisis de correlaciones, para identificar posibles relaciones entre las variables. Esto permite comprender la interdependencia entre las diferentes características de los jugadores, equipos y otros aspectos relevantes. Se utilizan técnicas estadísticas y visualizaciones para evaluar la fuerza y la dirección de estas correlaciones. Como también un modelo predictivo para determinar quién será el MVP de esa temporada, un PCA para mostrar que muchas variables dentro de la liga se pueden unir y, por otro lado, unos análisis para determinar que temporada ha sido más competitiva. Por último, un análisis que determina la importancia de los jugadores para las victorias del equipo

8. Validación de datos: Se verifica la calidad y la coherencia de los datos después de la limpieza y transformación. Esto implica realizar comprobaciones adicionales, como la verificación de la consistencia de los datos con respecto a las reglas del dominio, la comparación con fuentes externas confiables o la realización de pruebas de integridad de datos.

4. Alcance del proyecto

4.1.- Requisitos:

Utilización de un mínimo de dos gráficos distintos para la comparación de unas mismas variables

4.2.- Restricciones:

- Tiempo:
- Solo contamos con tres meses
- Máxima desviación respecto a la línea base del cronograma de dos semanas
- Tiempo máximo de respuesta por parte del profesorado de una semana
- Actualización de los análisis por la aparición de nuevos datos y variables
- Capacidad máxima de almacenamiento de datos
- Trabajo en la nube (velocidad y ancho de banda del internet para el uso del Google Drive)

4.3.- Productos entregables:

-Datasets utilizados

-Códigos de programación

-Gráficos

-Informes semanales

4.2.- Límites de proyecto:

- Utilización de Python y Statgraphics.

- Uso de modelos estadísticos bivariantes.

- Uso de algoritmos de clusterización.

Se excluye:

- Actualización dinámica de tablas

4.3.- Criterios de éxito:

- Recibir buenas observaciones en las presentaciones y críticas constructivas

- Sensación de los miembros del equipo de haber aprendido

- Reuniones y puestas en común con al menos un 80 % de asistencia y trabajo hecho

- Alcance de conclusiones claras

- Cumplimiento de los objetivos planteados

4.4.- Asunciones:

- La muestra es representativa

- Toma de datos adecuada

4.5.- Alineación con ODS:

Este proyecto se alinea con varios Objetivos de Desarrollo Sostenible establecidos por las Naciones Unidas:

ODS 3: Vida saludable y bienestar: La NBA, como organización deportiva líder en el mundo, puede estar relacionada con el objetivo de desarrollo sostenible número 3, que busca garantizar una vida saludable y promover el bienestar para todas las personas en todas las edades. La NBA ha implementado diversas iniciativas para fomentar la actividad física y la salud, incluyendo el programa NBA FIT, que proporciona recursos gratuitos y planes de entrenamiento para fomentar la actividad física entre los jóvenes. Al analizar los datos relacionados con la NBA y las desigualdades en el deporte, el proyecto busca generar conocimiento que pueda contribuir a la mejora de la salud y el bienestar de los jugadores y las comunidades en general.

5. Calendario del proyecto

Febrero

El mes de febrero lo dividimos originalmente en dos fases:

- Investigación

- Obtención de los datos, descripción del problema y objetivo del análisis

Marzo

Durante el mes de marzo hemos hecho los siguientes puntos:

- Datos faltantes
- Datos atípicos

Abril

En el mes de abril hemos hecho los siguientes puntos:

- Análisis descriptivo

Mayo

El mes de mayo se divide fundamentalmente en 5 fases:

- Finalización de análisis estadístico
- Redacción de la memoria
- Inferencia
- PowerPoint
- Vídeo
- Finalizar nuestra página web

Junio

En el último mes decidimos hacer dos puntos:

-Realizar el modelo predictivo

-Maquetar el proyecto.

Nombre Actividad	**Fecha inicio**	**Duración en días**	**Fecha Fin**
Definir alcance del proyecto	27-feb	10	09-mar
Informe semanal3	07-mar	15	22-mar
Informe semanal4	23-mar	13	05-abr
Informe semanal5	18-abr	9	27-abr
Segunda presentación preliminar	04-may	14	18-may

Primera versión del informe final de la memoria	05-mar	9	14-mar
Presentaciones propuestas	22-feb	8	02-mar
Página web	07-abr	70	16-jun
Análisis descriptivo	01-may	16	17-may
Lanzamiento del proyecto	03-may	15	18-may
Análisis estadístico	07-may	21	28-may
Predicción del MVP	01-jun	14	15-jun
Maquetar el proyecto	15-jun	2	18-jun

FIG.3. Tabla fechas

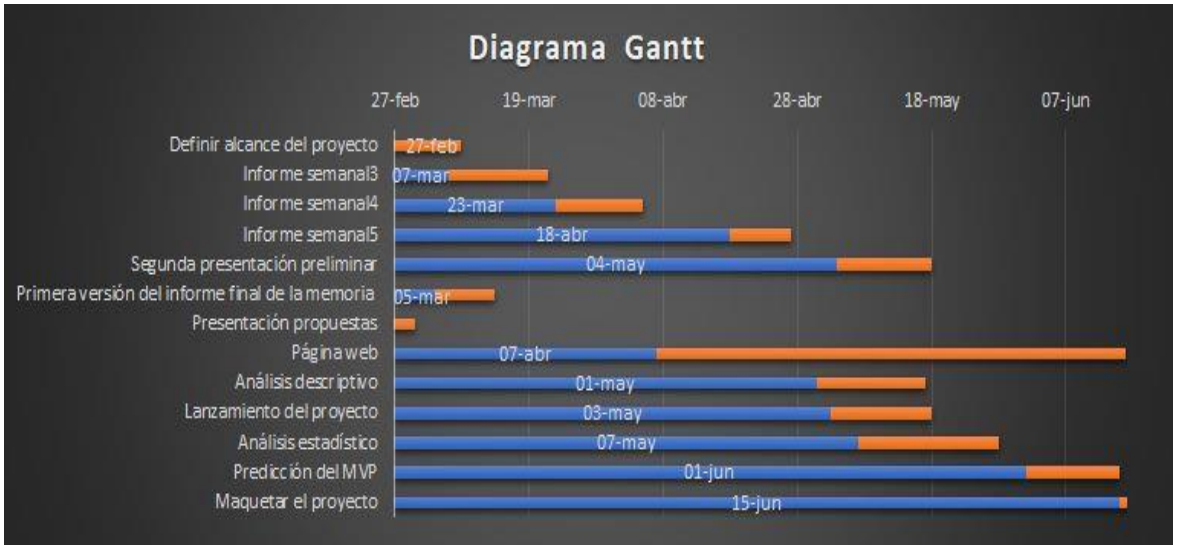


FIG.4. Diagrama de Grant

6. Conocimiento extraído

Durante el desarrollo de nuestro proyecto de análisis de estadísticas de jugadores y equipos de la NBA, hemos obtenido varios conocimientos y hemos logrado cumplir nuestros objetivos. A continuación, se presentan los resultados del proyecto:

- Búsqueda de jugadores y equipos: Hemos realizado con éxito las funciones de búsqueda de jugadores por nombre. Esto nos permite obtener y mostrar las estadísticas de los jugadores y equipos de manera precisa y eficiente.

- Generación de gráficos de evolución de estadísticas de equipos y jugadores: Hemos desarrollado diversas funciones que nos permite generar gráficos que representan la evolución de las estadísticas de todos los equipos de la liga. Además, de poder comparar jugadores y mostrar cómo han cambiado su juego según ha evolucionado su estilo de juego y de la liga en general. Estas funciones nos brindan la capacidad de visualizar y analizar de manera efectiva el rendimiento de los equipos y jugadores a lo largo del tiempo.

- Análisis de correlaciones en premios de la NBA: Mediante técnicas de WebScraping empleando las librerías de beautifulsoup, request y os. Además del uso de librerías para el pocesamiento y visualización de datos llamadas Pandas y Seaborn la cual permite mostrar de forma fácil las correlaciones. Hemos obtenido las tablas correspondientes a los premios MVP (Most Valuable Player), ROY (Rookie Of the Year), SMOY (Six Man Of the Year) y DPOY (Defesive Player Of the Year) de la NBA. Hemos realizado un análisis exhaustivo de las correlaciones existentes entre los votos obtenidos y las diferentes estadísticas de los jugadores. Hemos identificado correlaciones significativas entre los votos y estadísticas como Win Shares, Puntos, Partidos jugados, Minutos jugados, Rebotes y Asistencias.

- Análisis de componentes principales (PCA): Hemos aplicado el análisis de componentes principales (PCA) a los conjuntos de datos de jugadores y equipos de la NBA. Mediante el PCA, hemos descubierto las variables principales que explican la variabilidad en los datos. En el caso de los jugadores, encontramos que tres variables principales explican más del 88% de la variabilidad, y en el caso de los equipos, dos variables principales explican más del 77% de la variabilidad.

- Índices de competitividad en la NBA: Hemos calculado los índices Sigma y HICB para evaluar la competitividad en la NBA a lo largo de las temporadas. El índice Sigma nos proporciona una medida de la igualdad o equilibrio entre los equipos, mientras que el índice HICB refleja el nivel de concentración de puntos entre los equipos. Mediante el análisis de estos índices, hemos identificado las temporadas más competitivas y aquellas con menor competitividad.

- Predicción del MVP: Mediante librerías de Python como es el caso de pandas para obtener el DataSet (Conjunto de Estadísticas) que nos permite entrenar el modelo y, sklearn que permite crear la predicción. Hemos realizado un modelo predictivo que, dadas las estadísticas de todos los jugadores de una temporada, las analiza y calcula quien será el MVP de ese año.

6.1.- Variables

Nombre	Definición
MP	Minutos jugados
FGM	Tiros de campo anotados
FGA	Tiros de campo realizados
FG% / FG_PCT	Porcentaje de tiros de campo anotados
3PM / FG3M	Triples anotados
3P% / FG3_PCT	Porcentaje de triples
3PA / FG3A	Triples realizados
2PA	Tiros de dos puntos realizados
2PM	Tiros de dos puntos anotados
2P%	Porcentaje de tiro de dos puntos
FTM	Tiros libres anotados
FTA	Tiros libres realizados
FT% / FT_PCT	Porcentaje de tiros libres
ORB / OREB	Rebotes ofensivos capturados
DRB / DREB	Rebotes defensivos capturados
TRB / **REB**	Total de rebotes capturados
AST	Asistencias
STL	Robos
BLK	Tapones
TOV	Pérdidas
PF	Faltas personales
PTS	Puntos anotados
PLUS_MINUS	Valoración del partido

FIG.4. Tablas variables

6.2.- Análisis descriptivo

Hemos finalizado la función de búsqueda de jugadores. Esto nos permite obtener y mostrar las estadísticas de los jugadores de manera precisa y eficiente, acercándonos a nuestro objetivo de crear una plataforma completa para el análisis de estadísticas de jugadores de la NBA.

```python
def loc_jugadores(txt, retirado):
    activos = []
    retirados = []
    p = df1
    for fila in df1.iterrows():
        fila_l = list(fila)
        player = fila[1][2]
        if fila[1][5]:
            if txt.lower() in player.lower():
                activos.append(player)
        else:
            if txt.lower() in player.lower():
                retirados.append(player)
    if retirado:
        return retirados
    else:
        return activos

def stats_players(p):
    return df.loc[df['Player'] == p]
```

FIG.5. Función localizar jugadores

Además, hemos desarrollado una nueva función que nos permite generar gráficos que representan la evolución de las estadísticas de todos los equipos de la liga. Esta función nos brinda la capacidad de visualizar y analizar de manera efectiva el rendimiento de los equipos a lo largo del tiempo. Anteriormente, ya habíamos implementado la capacidad de organizar los equipos por conferencias y divisiones, lo que nos permite realizar análisis más detallados y comparativos dentro de la liga.

A continuación, se muestra una breve muestra de nuestras posibles gráficas cabe destacar que esto son ejemplos de las casi infinitas combinaciones

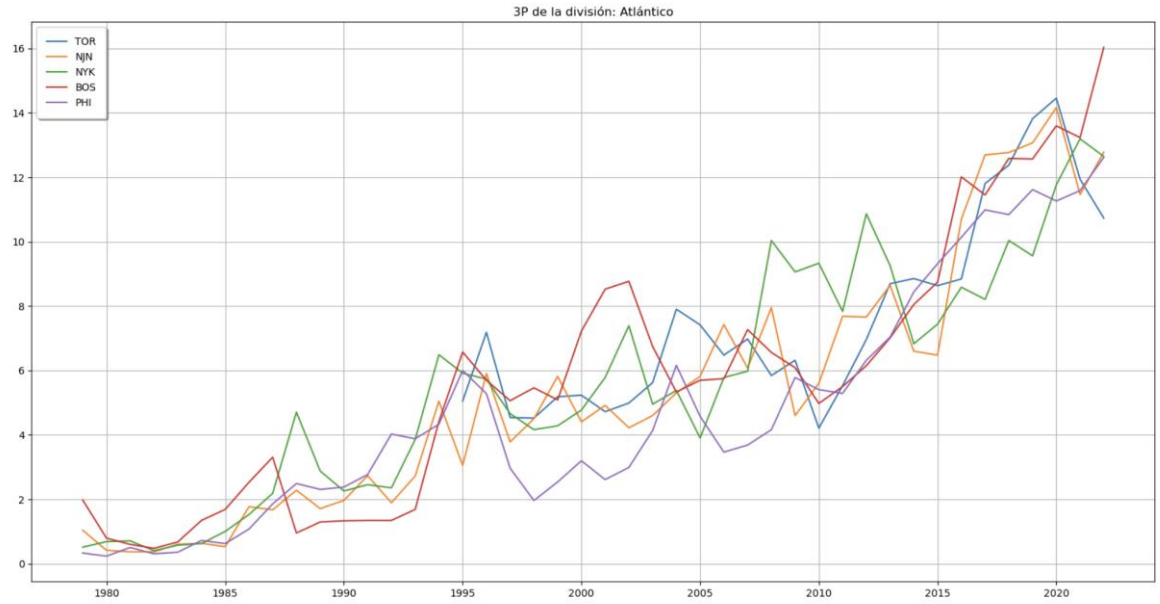


FIG.6. Gráfica de tres puntos

En el análisis de las gráficas presentadas se puede observar claramente la tendencia de aumento en los tiros de tres puntos dentro de la división atlántica de la competición. Esta tendencia también se refleja en el conjunto general de la liga.

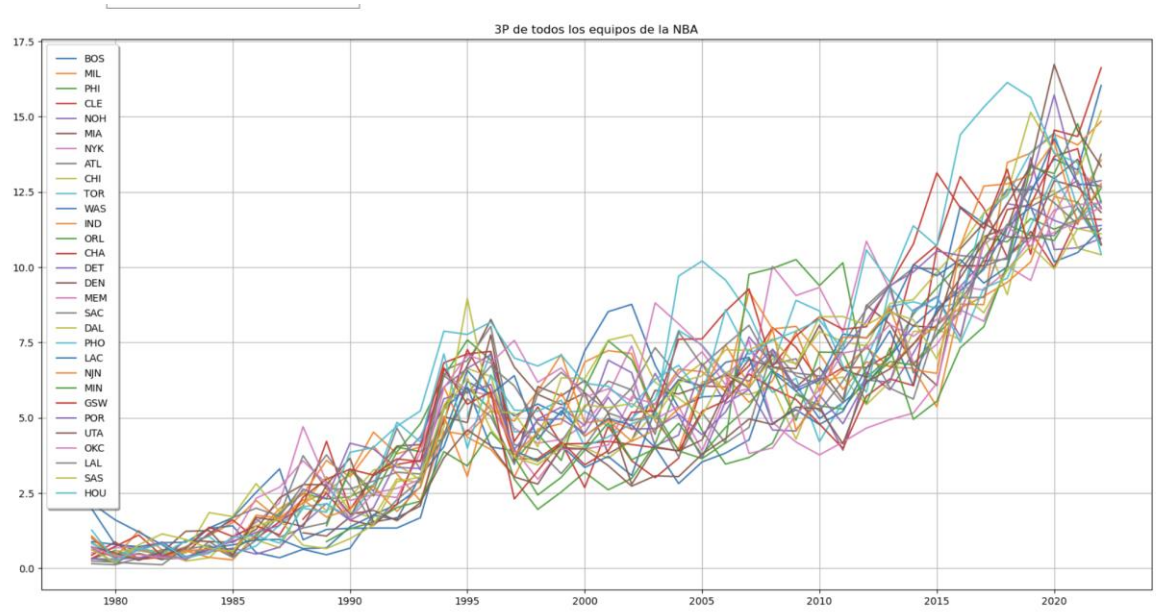


FIG.7. Gráfica de tres puntos de todos los equipos

En los últimos años, ha habido un aumento en los tiros de tres puntos en la NBA debido a su mayor eficiencia en términos de puntos por posesión. Los tiros de dos puntos se consideran menos eficientes en promedio, pero jugadores como Kevin Durant aún pueden encontrar éxito debido a sus habilidades individuales y movimientos ofensivos.

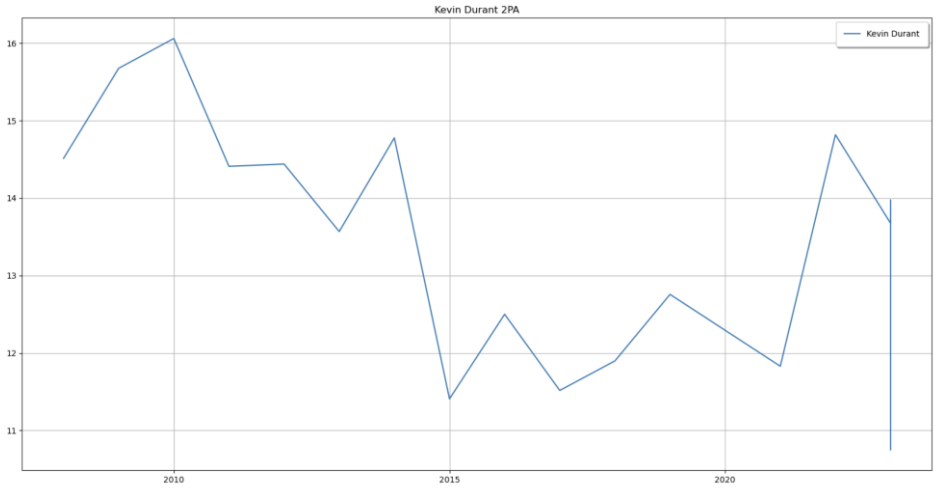


FIG.8.Kevin Durant tiros de dos anotados

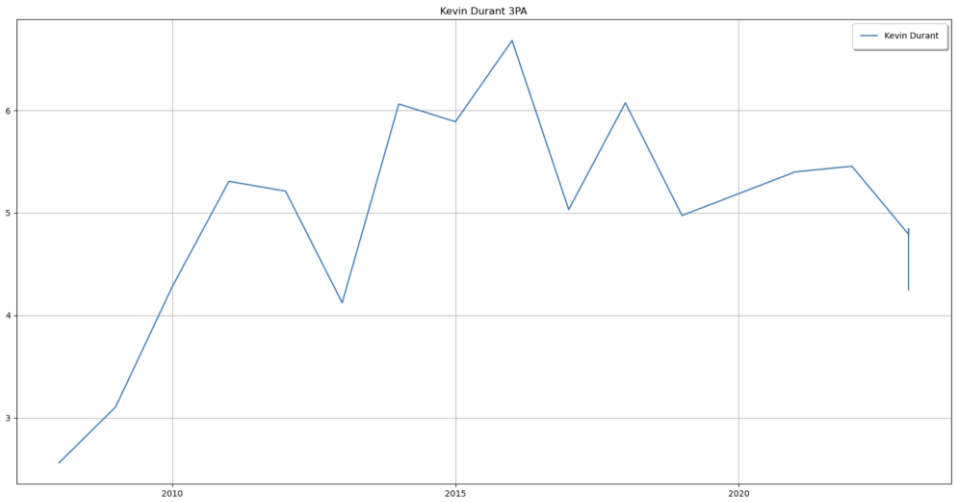


FIG.9.Kevin Durant tiros de tres puntos anotados.

Destacamos que pese a seguir en tendencia descendente los tiros de dos siguen siendo superiores a los tiros de tres por lo menos en este caso.

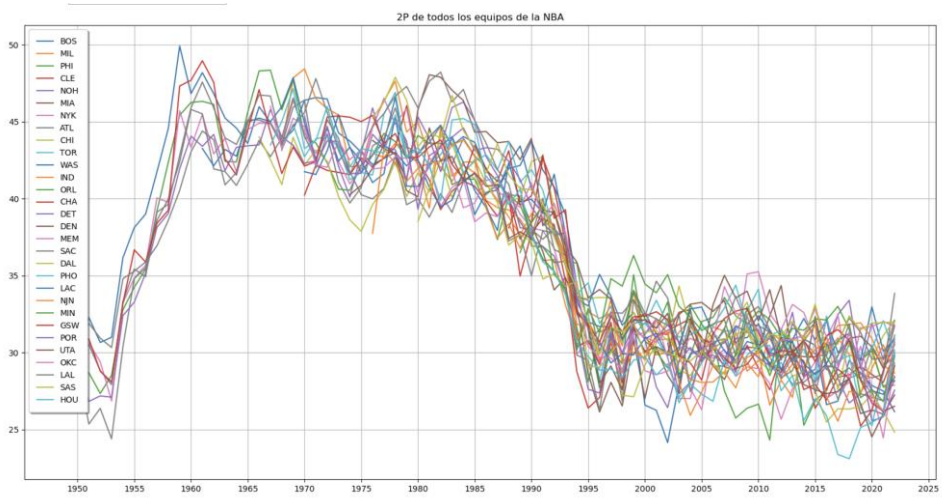


FIG.10.Evolución de los tiros de dos anotados por equipo en la historia.

Además de mostrar la evolución de los equipos o de un jugador determinado, podemos realizar comparaciones con los jugadores que deseemos, en este caso los tiros de tres de Stephen Curry y LeBron James.

FIG.11.Evolución de los tiros de tres de LeBron James.

Otro estilo de gráfica que hemos creado son los histogramas, los cuales de forma interactiva muestran a los diferentes equipos según la estadística que deseemos.

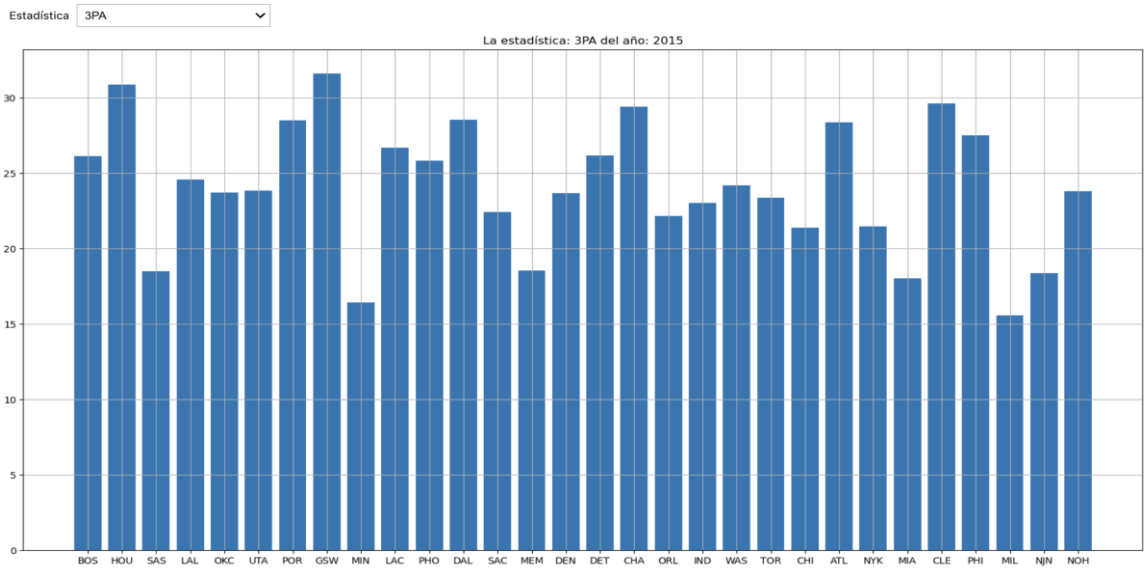


FIG.12.Los tiros de tres metidos por todos los equipos en el año 2015.

6.3.- Análisis estadístico

6.3.1.- Correlaciones en los premios

Hemos obtenido las votaciones de los cuatro premios más prestigiosos de la NBA (National Basketball Association): MVP (Most Valuable Player), ROY (Rookie Of the Year), SMOY (Six Man Of the Year) y DPOY (Defensive Player Of the Year) mediante la técnica de WebScraping. A partir de estos datos, hemos realizado análisis de correlación entre los puntos obtenidos por cada jugador y sus respectivas estadísticas. Todas estas correlaciones han sido obtenidas mediante el coeficiente de correlación de Pearson.

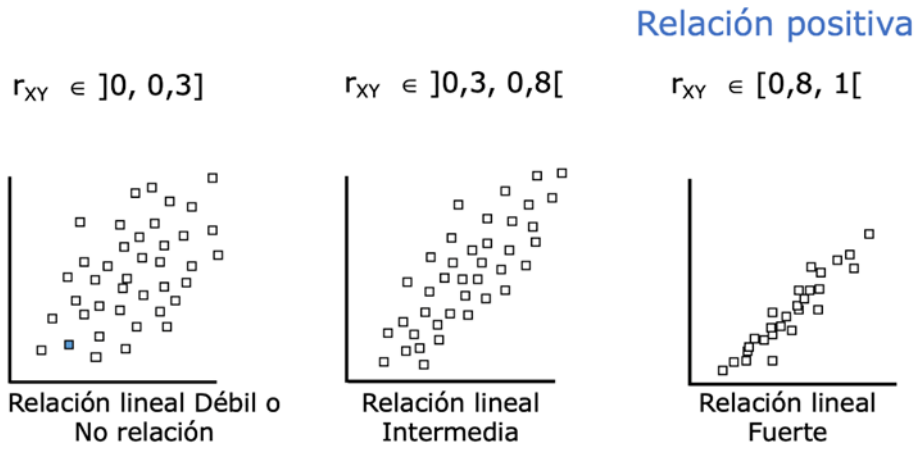


FIG.13.Correlación de Pearson.

Código de Python:

In [17]:
```
smoyC = smoy.corr()
mvpsC = mvps.corr()
royC = roy.corr()
dpoyC = dpoy.corr()
```

In [34]:
```
fig, ax = plt.subplots(nrows=1, ncols=1, figsize=(10, 10))
sns.heatmap(
    smoyC,
    annot      = True,
    cbar       = False,
    annot_kws = {"size": 8},
    vmin       = -1,
    vmax       = 1,
    center     = 0,
    cmap       = sns.diverging_palette(20, 220, n=200),
    square     = True,
    ax         = ax,
    fmt = '.2f'
)
ax.set_xticklabels(
    ax.get_xticklabels(),
    rotation = 45,
    horizontalalignment = 'right',
)
ax.set_title('Matriz de C0rrelación de SMOY')
ax.tick_params(labelsize = 10)
```

FIG.14.Código de Python.

Se ha llevado a cabo un análisis exhaustivo de las correlaciones existentes en los premios otorgados en la NBA. Para este estudio, se han seleccionado cuatro premios de gran relevancia e importancia: el MVP (Most Valuable Player), que reconoce al mejor jugador de la temporada; el ROY (Rookie Of the Year), que premia al jugador más destacado en su primer año; el SMOY (Six Man Of the Year), que distingue al mejor jugador suplente; y el DPOY (Defensive Player Of the Year), que se otorga al jugador con mayor destacado desempeño defensivo.

Mediante el empleo de técnicas de WebScraping y el procesamiento de datos utilizando la biblioteca Panda en Python se han obtenido las tablas correspondientes a cada uno de estos premios, estas tablas se han mostrado mediante la biblioteca Seborn:

6.3.1.1.-Tabla de MVPs (Most Valuable Players):

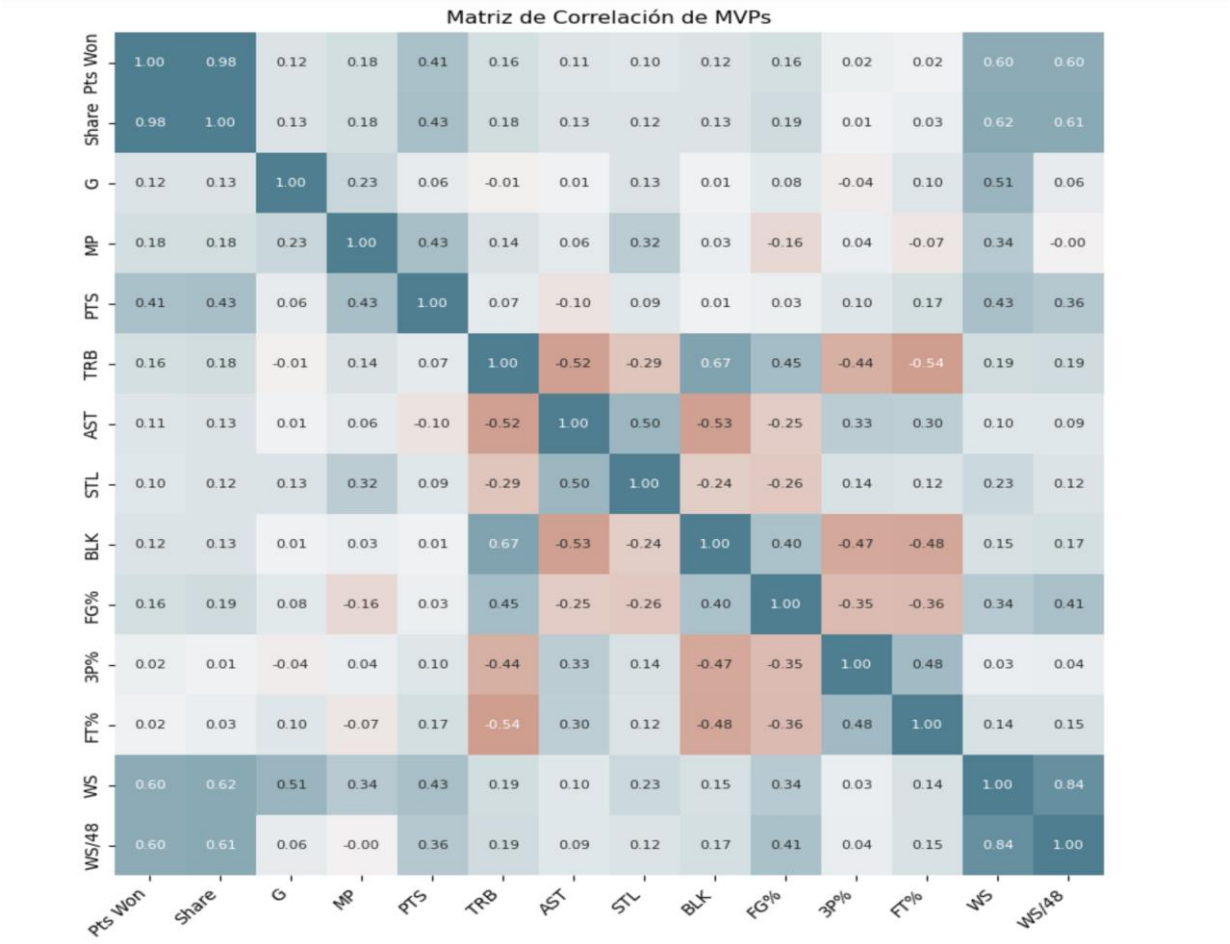


FIG.15.Tablas de MVPs

Durante nuestro análisis, hemos identificado correlaciones significativas entre los votos obtenidos en las votaciones y las diferentes estadísticas de los jugadores. En particular, hemos observado relaciones destacadas con la columna 'Share', que indica el porcentaje de los votos ganados ese año. Las correlaciones más relevantes son las siguientes:

- Win Shares (WS) y Win Shares per-48 minutes (WS/48):

Estas dos estadísticas muestran la influencia del jugador en las victorias del equipo. La alta correlación con 'Share' sugiere que los jugadores con un mayor impacto en la victoria del equipo tienden a recibir más votos en las votaciones.

- Puntos (PTS):

La estadística de puntos anotados también muestra una correlación significativa con 'Share'. Esto indica que los jugadores que destacan en la capacidad anotadora son valorados positivamente por los votantes.

Además, hemos observado correlaciones más débiles en las siguientes estadísticas:

- Partidos jugados (G):

Aunque la relación es menos fuerte, el número de partidos jugados muestra una correlación positiva con 'Pts. Won'. Esto sugiere que la disponibilidad y la continuidad en el juego pueden influir en la consideración de los votantes.

- Minutos jugados (MP):

La relación con 'Pts Won' también es débil pero positiva. Los jugadores que reciben más minutos de juego tienden a obtener más votos en las votaciones.

- Rebotes (TRB) y asistencias (AST):

Estas estadísticas muestran correlaciones débiles con 'Pts Won'. Aunque no tienen un impacto tan significativo, aún pueden contribuir en cierta medida a la evaluación de los votantes.

6.3.1.2.- ROY (Rookie of the Year):

En el caso del premio al Rookie del Año, hemos observado que las correlaciones entre las estadísticas de los jugadores y los votos obtenidos en las votaciones son más dispersas y menos definidas en comparación con otros premios. Sin embargo, se han identificado algunas correlaciones destacadas, siendo las más altas las siguientes:

- Asistencias (AST):

Existe una correlación relativamente alta entre el número de asistencias realizadas por los novatos y los votos obtenidos. Esto sugiere que los jugadores que destacan en la creación de juego y distribución de balón tienden a ser considerados positivamente por los votantes en la elección del Rookie del Año.

- Puntos (PTS):

La estadística de puntos anotados también muestra una correlación significativa con los votos obtenidos. Esto indica que los novatos que demuestran una habilidad destacada para anotar puntos son valorados positivamente en la votación.

- Minutos jugados (MP):

La correlación con los votos obtenidos es relativamente alta en esta estadística. Los novatos que reciben más minutos de juego tienen más posibilidades de demostrar su talento y contribuir en el campo, lo que puede influir en la consideración de los votantes.

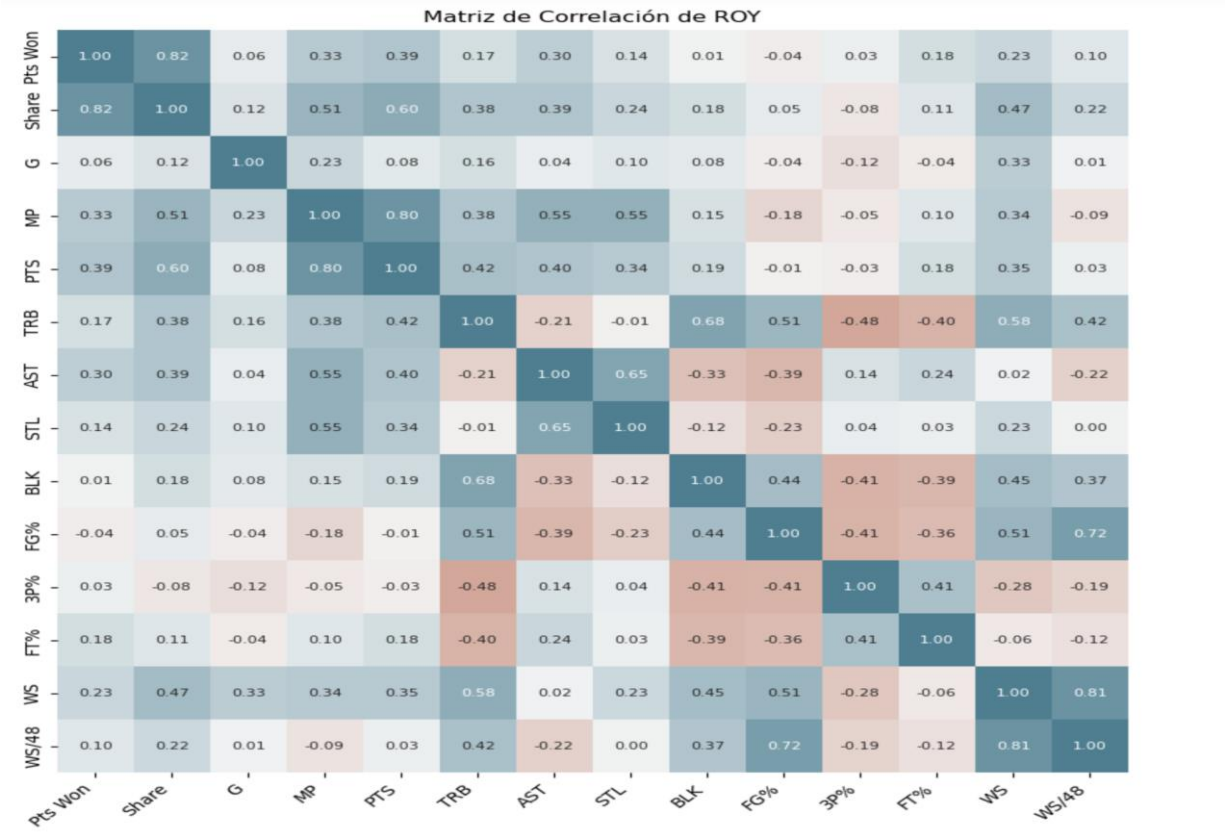


FIG.16.Tablas de ROY

6.3.1.3.- Mejor Sexto Hombre (Six Man of the Year):

Durante nuestro análisis del premio al Mejor Sexto Hombre del Año, hemos encontrado que las correlaciones entre las estadísticas de los jugadores y los votos obtenidos en las votaciones son menos claras y más dispersas en comparación con otros premios. Aunque no se observa una relación clara y definida, se han identificado algunas correlaciones destacadas, siendo las más altas las siguientes:

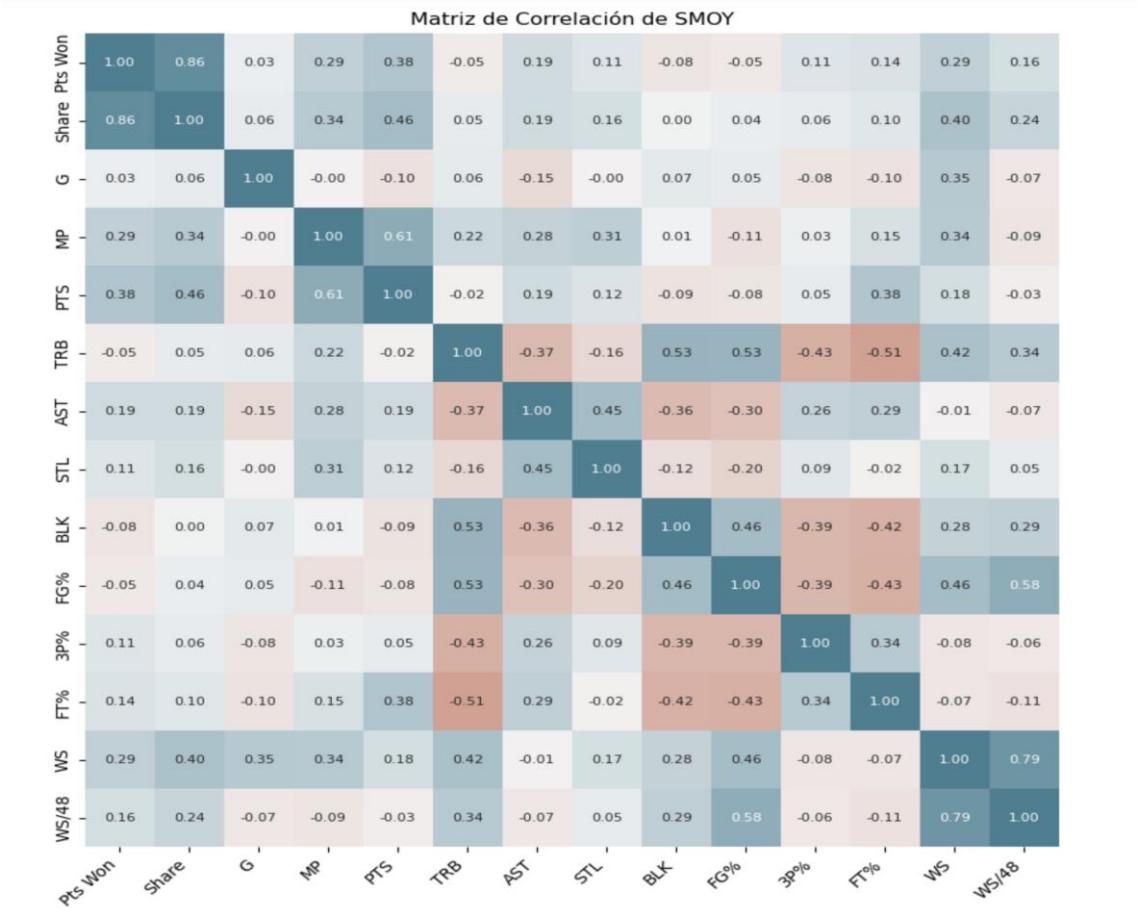


FIG.17.Tablas de Six Man of the Year

- Asistencias (AST):

Existe una correlación significativa entre el número de asistencias realizadas por los jugadores y los votos obtenidos en la elección del Mejor Sexto Hombre. Esto sugiere que los jugadores que destacan en la creación de juego y la distribución de balón desde el banquillo son valorados positivamente por los votantes.

- Puntos (PTS):

La estadística de puntos anotados también muestra una correlación significativa con los votos obtenidos. Esto indica que los jugadores que demuestran una habilidad destacada para anotar puntos saliendo desde el banquillo son considerados positivamente en la votación.

- Minutos jugados (MP):

La correlación con los votos obtenidos es relativamente alta en esta estadística. Los jugadores que tienen un impacto significativo en el juego a pesar de tener menos tiempo en la cancha son valorados positivamente en la elección del Mejor Sexto Hombre.

6.3.1.4.- Jugador Defensivo del Año (Defensive Player Of the Year):

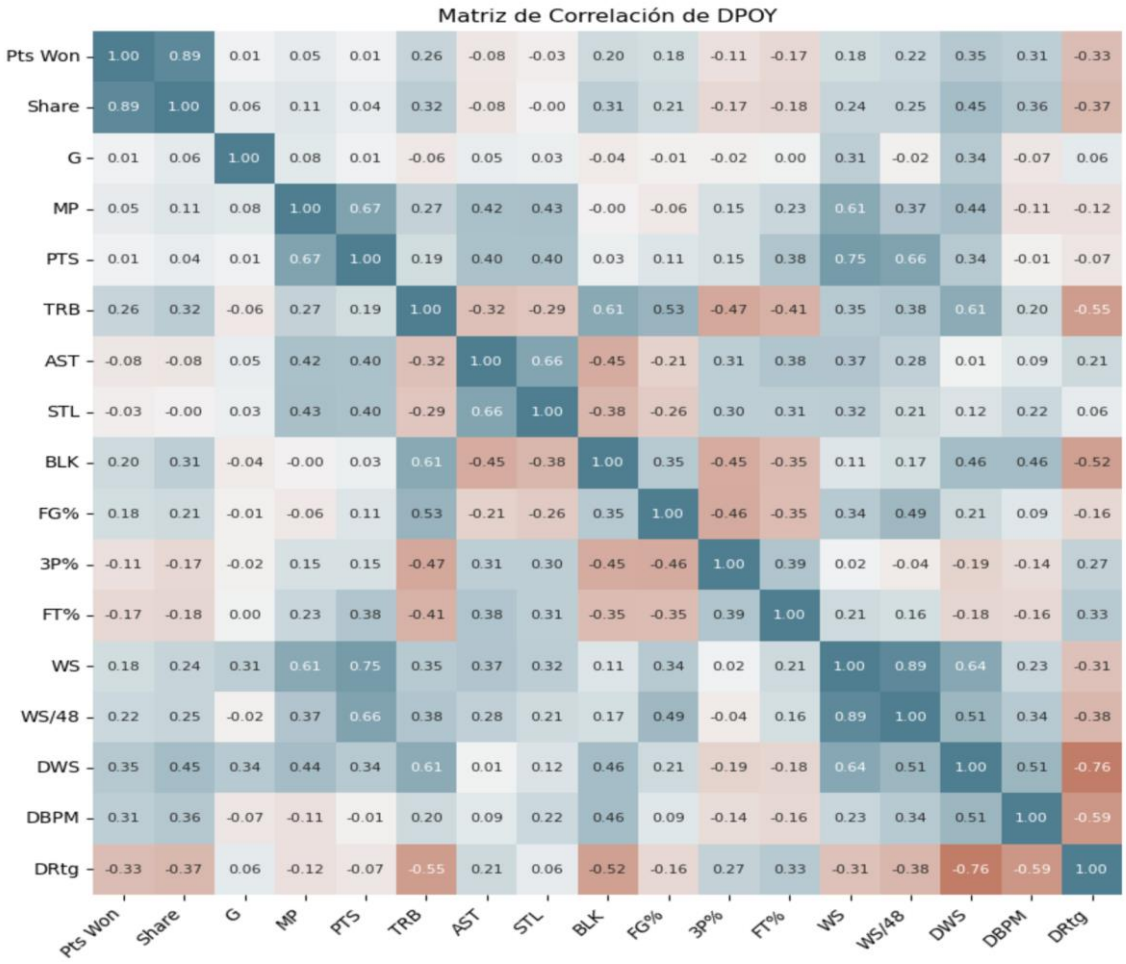


FIG.18.Tablas de DPOY

Es interesante observar que, en el premio al Jugador Defensivo del Año se encuentran correlaciones más destacadas en las estadísticas defensivas como los rebotes y los tapones. Además, se observa una relación negativa con los robos, lo cual puede deberse a que la mayoría de los ganadores son jugadores de posición interior o pívots, quienes tienden a enfocarse más en la protección del aro y los rebotes defensivos en lugar de los robos de balón.

Sin embargo, también es importante destacar que algunos jugadores exteriores han logrado ganar el premio a pesar de no ser tan propensos a los robos de balón. Marcus Smart y Gary Payton son ejemplos de especialistas defensivos en el robo que han obtenido este reconocimiento. Esto demuestra que, aunque no sea una correlación generalizada, los jugadores exteriores con habilidades destacadas en el robo pueden ser considerados para el premio al Jugador Defensivo del Año.

Nos parece sorprendente no obtener relaciones muy claras, pero entendemos que puede ser debido a que en la NBA se le da una gran importancia a la narrativa de ese año, por ejemplo, si un jugador realiza una campaña con los mejores números no siempre gana el premio, puede ser que se le dé a otro porque pertenece a un equipo

más importante. Eso nos parece bastante mal, pero es sabido por todos que en la NBA hay mercados grandes y otros más pequeños, y a nivel comercial es mejor que gane un jugador que otro. Pero no podemos determinar la veracidad de que le otorguen un premio a un jugador por nombre. Esto es solo una teoría respaldada por varios especialistas y analistas en la liga. Que podemos apoyar, al encontrar casos en los cuales jugadores con mejores estadísticas no han ganado el premio.

6.3.2.- PCA

Además, hemos aplicado el Análisis de Componentes Principales (PCA, por sus siglas en inglés) a los conjuntos de datos que incluyen información sobre los jugadores y los equipos de la NBA a lo largo de los años. El PCA es una técnica estadística que nos permite reducir la complejidad de espacios de datos con múltiples dimensiones mientras se conserva la información relevante.

Al realizar el PCA en el conjunto de datos de los jugadores, hemos descubierto que solo tres variables principales pueden explicar más del 88% de la variabilidad presente en el conjunto de datos. Esto implica que la mayoría de la información se concentra en estas tres variables clave.

De manera similar, al aplicar el PCA al conjunto de datos de los equipos, hemos encontrado que solo dos variables principales pueden explicar más del 77% de la variabilidad en el conjunto de datos. Esto indica que la información crítica sobre los equipos se puede resumir en estas dos variables principales.

Podemos determinar que esto depende a la gran cantidad de variables (estadísticas) que se parecen, por ejemplo, 3PA y 3P es innegable que un jugador con muchos triples intentados también tendrá un mayor número de triples metidos. Al igual que, si un jugador es capaz de anotar de forma eficiente, es decir, con grandes porcentajes tiros de dos puntos, será más propenso a realizar más intentos que otro que no se caracteriza por eso.

Esto que observamos en los jugadores ocurre igual en los equipos, por ejemplo, los Houston Rockets en la temporada 2017-18 se caracterizaban por un gran número de intentos de tres puntos, de hecho, jugaban a intentar meter muchos triples, entonces este equipo cuando se realice el análisis de componentes principales las variables de tiros de tres puntos serán muy parecidas.

Para más información ver el anexo [7]

6.3.3.- IDC (Índice De Competitividad)

Para evaluar la competitividad en la NBA, es común utilizar índices que proporcionen una medida de igualdad o equilibrio entre los equipos. Uno de estos índices

ampliamente utilizados es el índice Sigma (σ) (referencia [7]). El índice Sigma se basa en el porcentaje de victorias de los equipos y el número total de equipos en la liga.

La fórmula para calcular el índice Sigma es la siguiente:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(w_i - \frac{1}{2}\right)^2},$$

FIG.19. Fórmula sigma

Es decir, el índice 'sigma' es la desviación estándar de la proporción de victorias que obtiene un equipo en una temporada en particular

Donde:

- n es el número total de equipos en la liga.
- Wi es el porcentaje de victorias del equipo i.
- Wmean es el promedio de los porcentajes de victorias de todos los equipos.

El índice Sigma proporciona una medida de dispersión de los porcentajes de victorias de los equipos. Cuanto menor sea el valor de Sigma, mayor será la competitividad y equilibrio entre los equipos. Por otro lado, un valor de Sigma más alto indica una mayor desigualdad y dominio de ciertos equipos en la liga.

Mediante el cálculo del índice Sigma para cada temporada, hemos generado del índice sigma una tabla que contiene los valores de estos índices respectivos. Esta tabla nos proporciona una visión estructurada y sistemática de la competitividad en la NBA a lo largo de las diferentes temporadas.

Utilizando esta tabla, hemos creado una gráfica sencilla pero efectiva que nos permite visualizar de forma instantánea las temporadas más competitivas. Esta representación gráfica nos permite identificar rápidamente aquellos períodos en los que la competitividad ha sido más intensa, al observar los valores más bajos de los índices.

Código en Python: (Con más detalle en el Notebook anexo [9])

```
import math as m

l = []

for y in years:
    x = wrate_df['Wrate'+str(y)].loc[wrate_df['Wrate'+str(y)].notnull()]
    l.append(m.sqrt(((x-0.5)**2).sum()/x.count()))
```

```
sigma = pd.DataFrame(l, index=years, columns=['Sigma'])
```

```
import math as m

l1 = []

for y in years:
    x = ptsrate_df['PTSrate'+str(y)].loc[ptsrate_df['PTSrate'+str(y)].notnull()]
    l1.append(100*x.count()*(x**2).sum())
```

```
hicb = pd.DataFrame(l1, index=years, columns=['HICB'])
```

```
hicb.sort_values(by = 'HICB')
```

FIG.20. Código de implementación

	Sigma
1956	0.050556
1955	0.057266
1954	0.081579
1950	0.092468
1976	0.096026
...	...
1952	0.176654
1972	0.181333
1949	0.185705
1997	0.186137
1996	0.187646

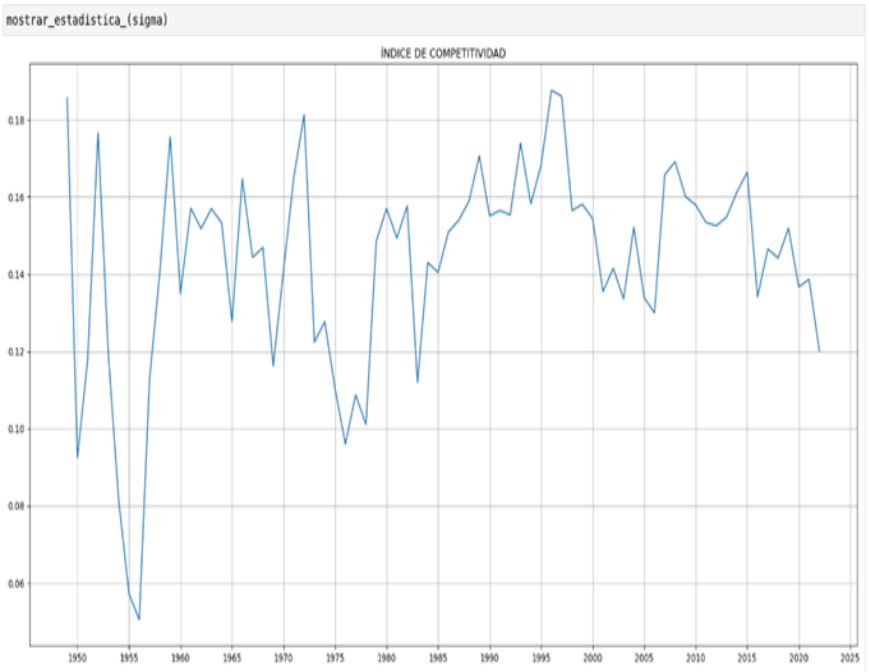


FIG.21. Gráfica sigma

Tras analizar los valores de los índices Sigma en la tabla y la gráfica generadas, hemos identificado que las temporadas más competitivas en la NBA han sido las siguientes: 1956-57, 1955-56, 1954-55, 1950-51 y 1976-77. Esto puede deberse a que había pocos equipos por lo tanto tenían victorias muy parecidas, además cuando revisas como han quedado en las series de playoff se muestra una gran igualdad, es decir, no ha habido una gran superioridad de los equipos. En la NBA moderna se encuentra una sigma muy parecida en las temporadas, esto puede ser que los equipos no cambian mucho, salvo

unos casos muy específicos, por tanto, el rendimiento de estos cambia progresivamente, por eso mismo no hay unas grandes variaciones en temporadas continuas.

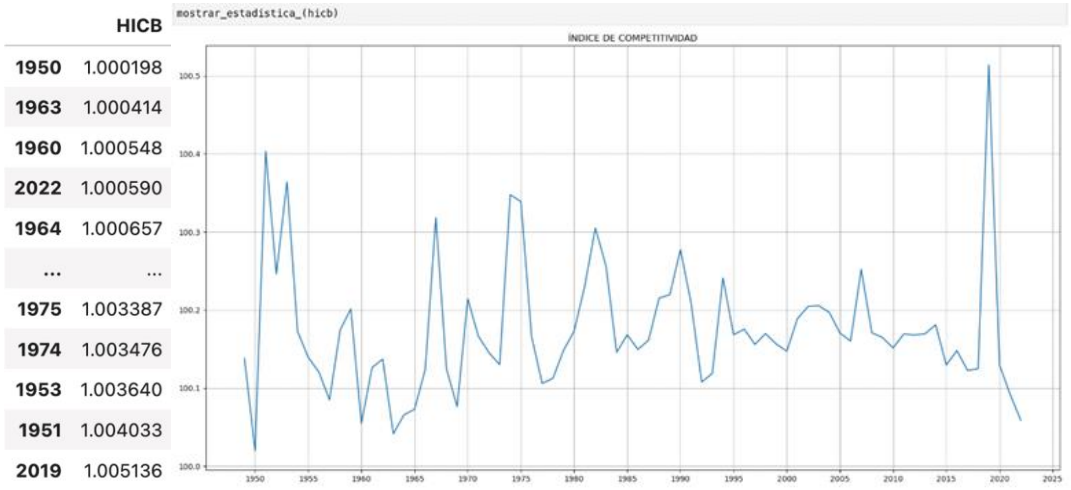
El índice Herfindahl-Hirschman (HHI) o HICB (H-index of competitive balance) (referencia [7]) es otro indicador utilizado para evaluar la competitividad en la NBA. Este índice se basa en el número de equipos (n) y el porcentaje de puntos obtenidos por cada equipo (Si) en relación con el total de puntos en una temporada determinada.

La fórmula para calcular el HICB es la siguiente:

$$\mathrm{HICB} = 100\, n \sum_{i=1}^{n} s_i^2,$$

FIG.22.Fórmula HICB

Al calcular el índice HICB para cada temporada, obtendremos un valor que refleja el nivel de concentración de puntos entre los equipos y, por lo tanto, proporciona una medida de la competitividad relativa de esa temporada. Al igual que, en el índice Sigma un valor más alto del índice HICB indica una menor competitividad, lo cual implica que algunos equipos son dominantes en relación con los demás.



	HICB
1950	1.000198
1963	1.000414
1960	1.000548
2022	1.000590
1964	1.000657
...	...
1975	1.003387
1974	1.003476
1953	1.003640
1951	1.004033
2019	1.005136

FIG.23.Gráfica HICB

Según los índices Sigma y HICB, podemos observar que la temporada 1950-51 muestra una gran competitividad en la NBA.

Al examinar detalladamente las estadísticas y el desarrollo de la temporada, se observa una pequeña diferencia entre los equipos en términos de rendimiento. Esto se refleja en los valores cercanos en el índice Sigma, lo que indica una distribución equilibrada de victorias entre los equipos.

Además, en las eliminatorias de los playoffs, no hubo muchos casos de "swaps", es decir, equipos que eliminaron a sus rivales sin perder ningún partido. Esto sugiere una mayor competitividad y enfrentamientos reñidos en los playoffs.

En cuanto a las finales, se determina en los últimos momentos del último partido posible.

Si nos centramos en los valores extremos en el índice Sigma, como los observados en 1954, 1955 y 1956, es importante destacar que estos años pueden caracterizarse por un rendimiento muy similar entre los equipos.

Esto lo podemos ver gráficamente en las siguientes gráficas:

```
1 Comparación_Anual(1954, 'W')
```

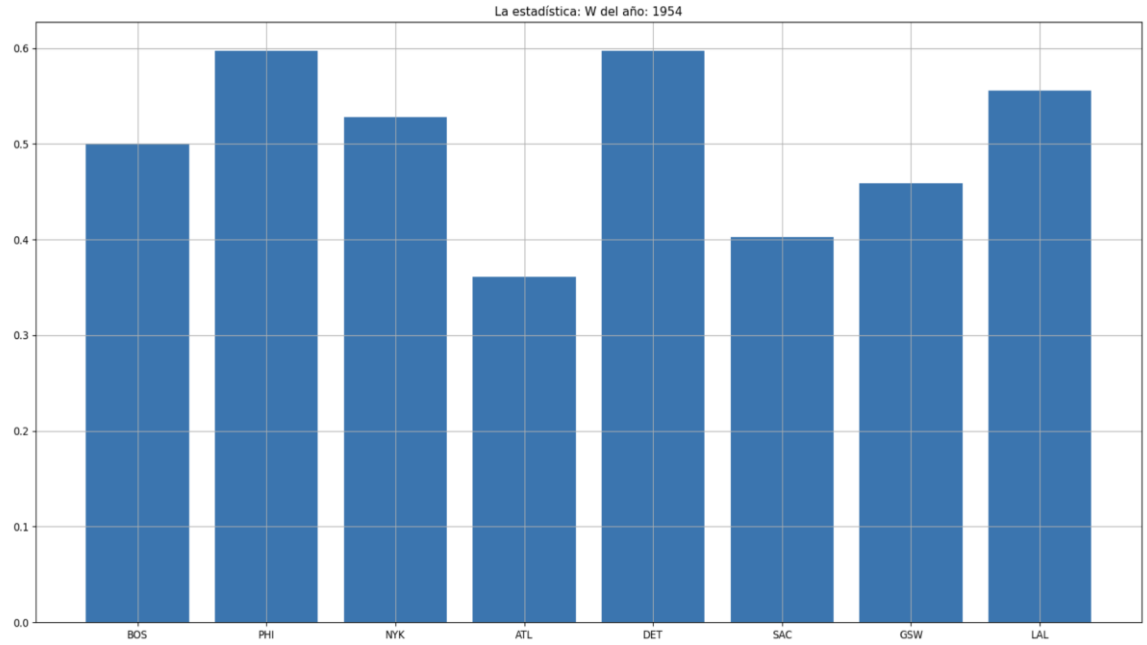


FIG.24. Gráfica victorias 1954

```
1 Comparación_Anual(1955, 'W')
```

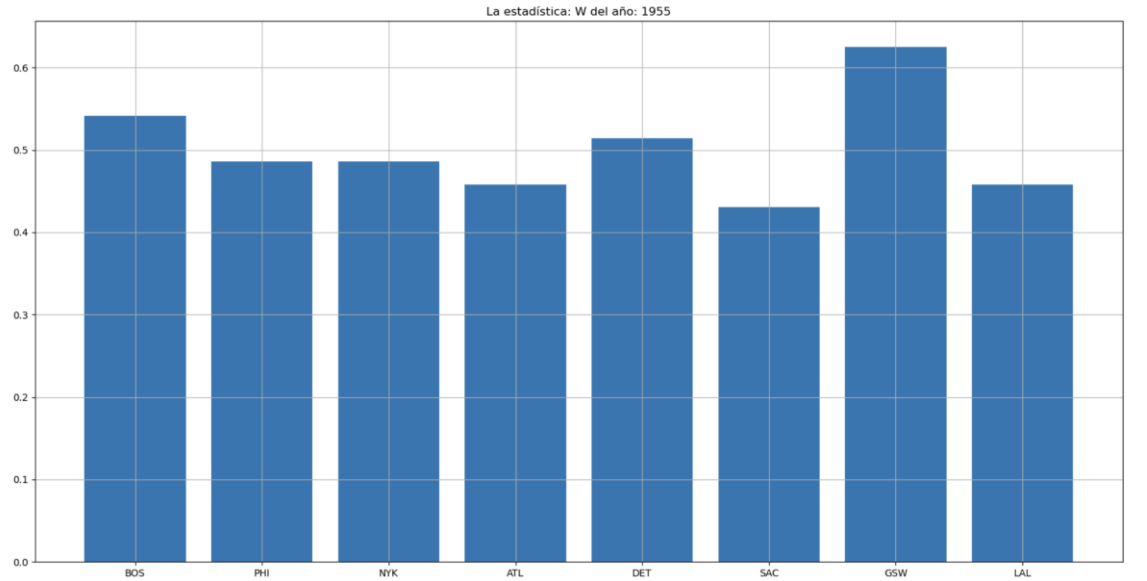


FIG.25. Gráfica victorias 1955

```
1 Comparación_Anual(1956, 'W')
```

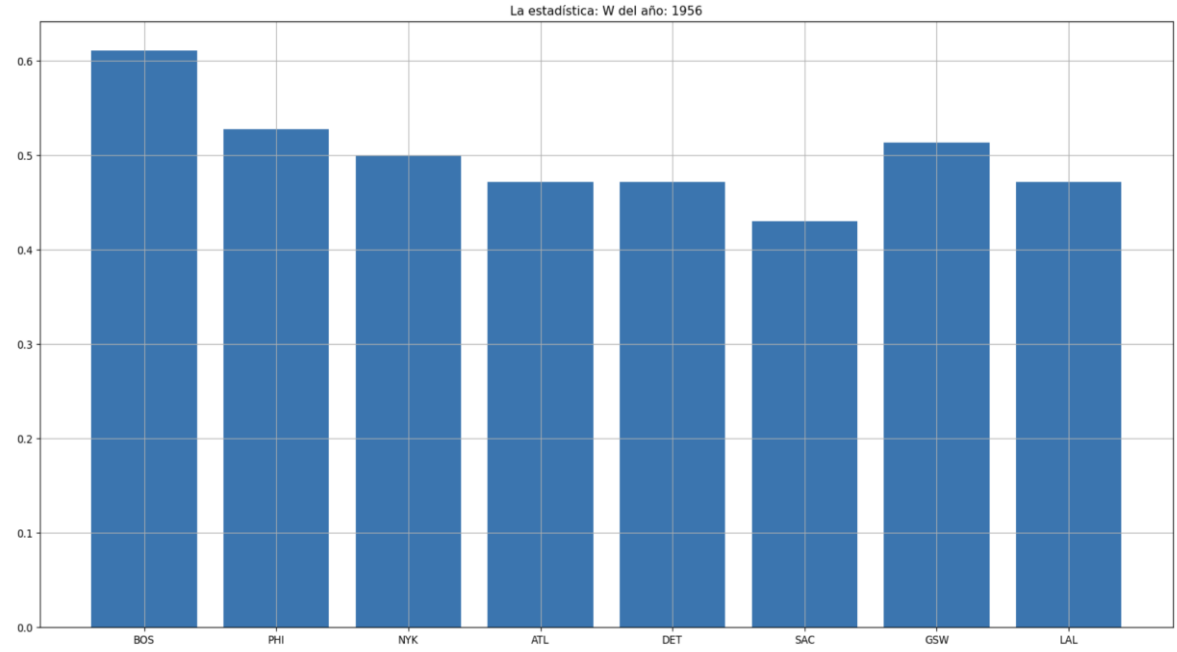


FIG.26. Gráfica victorias 1956

Estas gráficas como pone en el título son las victorias de todos los equipos esos años.

Durante la época más moderna, se observa una tendencia de valores más similares en los índices Sigma y HICB, con algunos altibajos. Esto indica una competitividad razonablemente equilibrada entre los equipos en general.

Sin embargo, también se pueden identificar algunas temporadas específicas que se destacan por tener una menor competitividad, según el índice HICB. En particular, se observa que la temporada 2019-20 muestra un valor de HICB que indica una menor competitividad.

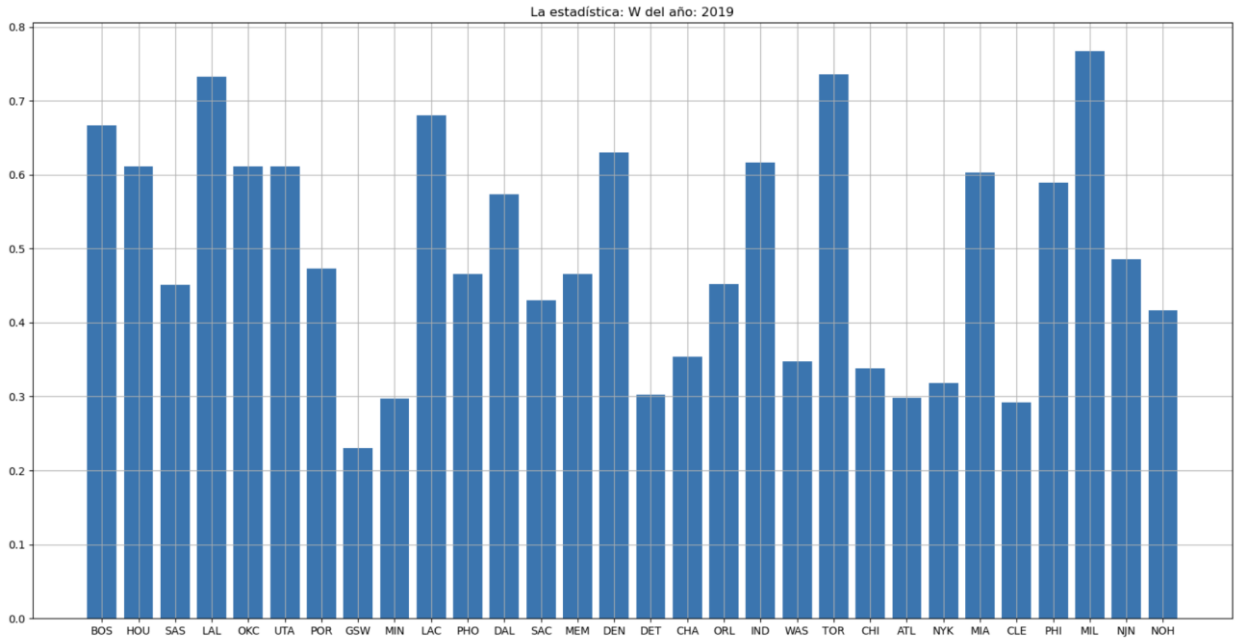


FIG.27. Gráfica victorias 2019

Además, se mencionan las temporadas 1996-97 y 1997-98 como ejemplos de temporadas con pocas victorias de diferencia entre los equipos. Estos años muestran una alta competencia y una menor disparidad en términos de resultados entre los equipos participantes.

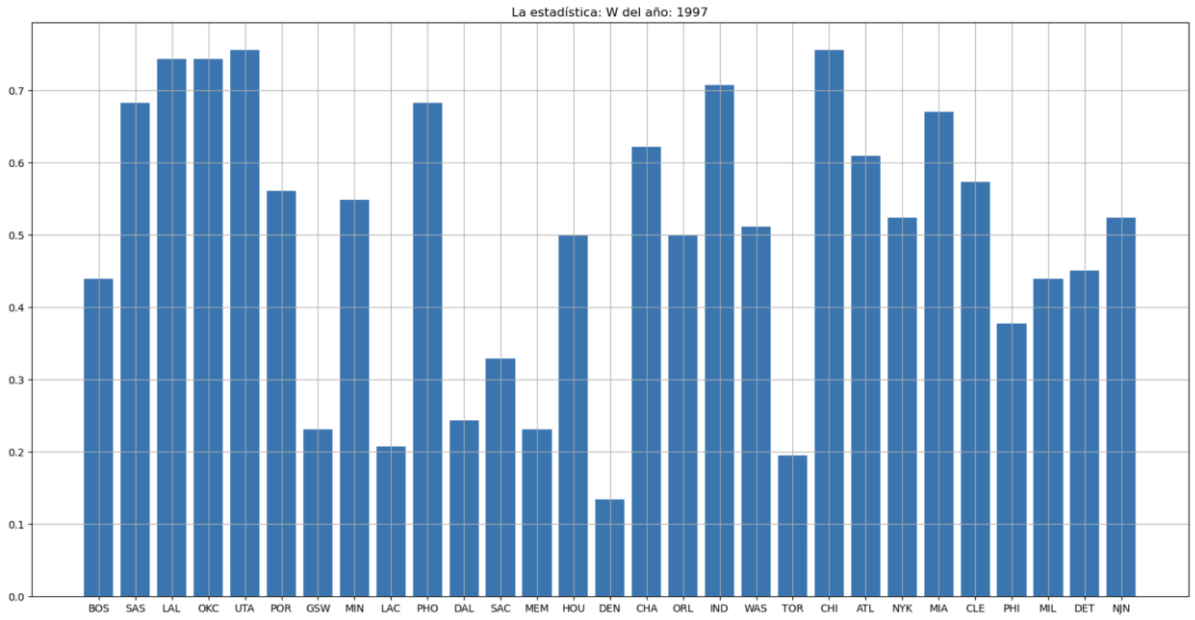


FIG.28. Gráfica victorias 1997

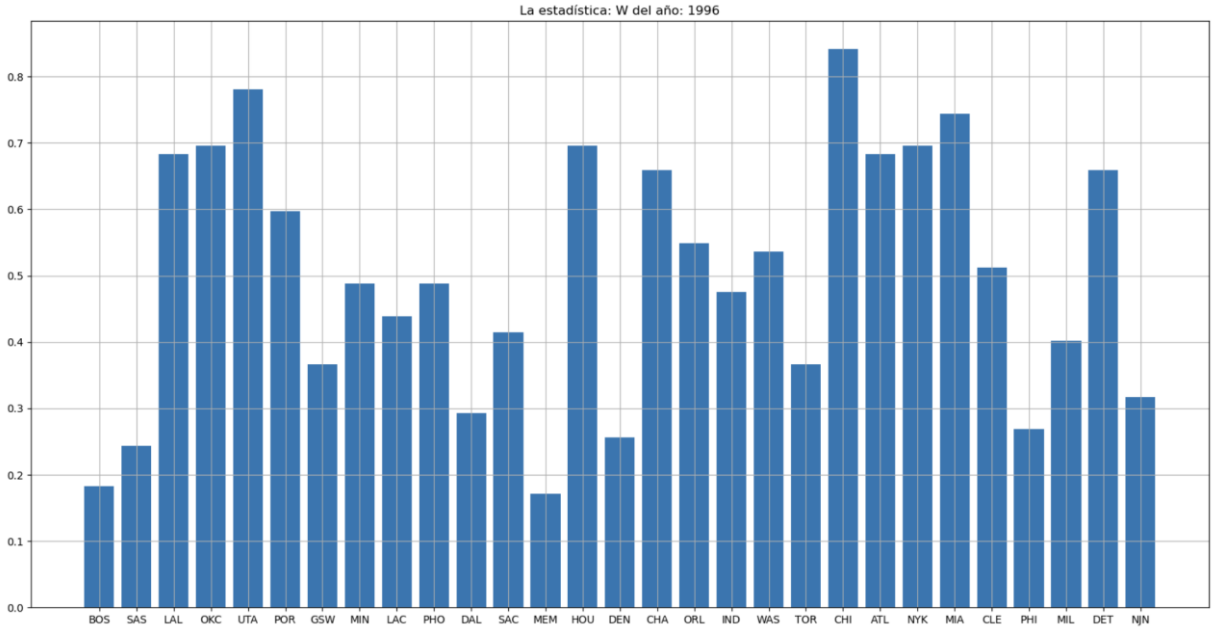


FIG.29. Gráfica victorias 1996

Para más información ver el anexo [10]

6.3.4.- Jugadores Importantes

Es innegable que en la NBA hay, ha habido y habrá jugadores determinantes y superiores a la media. Por esta razón, resulta interesante poder analizar algunas estadísticas clave para evaluar su impacto en el rendimiento de sus equipos. Por ejemplo, vamos a mostrar un breve análisis para el caso de Stephen Curry, quien ha sido reconocido como la estrella de su equipo durante muchos años, nos preguntamos si esto se respalda con las estadísticas.

Para responder a esta pregunta, hemos llevado a cabo un análisis que nos permite determinar la importancia o contribución que ha tenido este jugador en las victorias de su equipo. A continuación, presentaremos los resultados obtenidos:

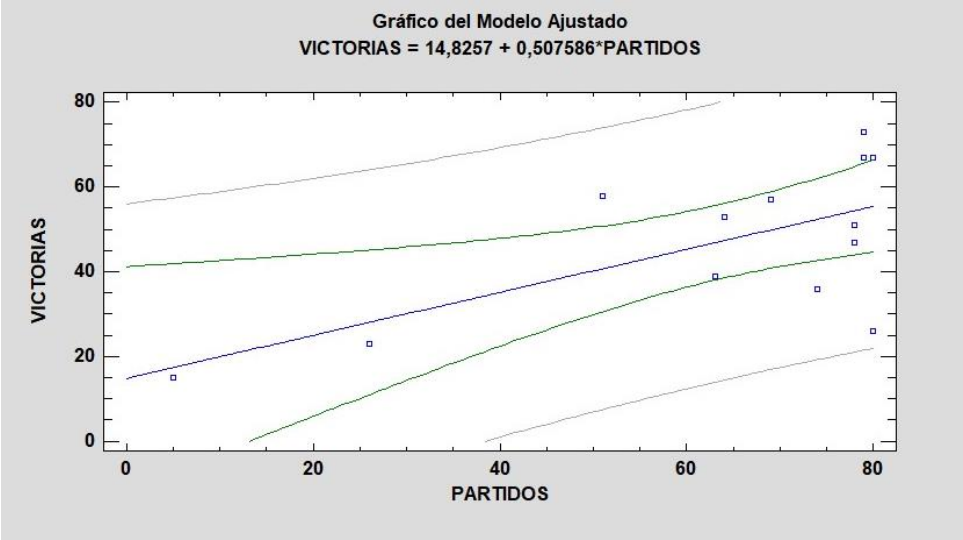


FIG.30. Gráfica modelo ajustado Stephen Curry

Esta gráfica representa la importancia que ha tenido Stephen Curry en los Warriors. En este análisis, hemos examinado todos los partidos que ha jugado a lo largo de su carrera con ese equipo y hemos evaluado su impacto en las victorias del equipo. El coeficiente de correlación obtenido es de 0.672632, lo que indica una relación moderadamente fuerte entre las variables. Con base en este valor, podemos afirmar que Curry ha tenido una gran importancia en el equipo.

Pero obviamente, Curry no es el único jugador relevante. Por lo tanto, hemos realizado el mismo estudio para los máximos exponentes en otras estadísticas, es decir, hemos realizado esto mismo con los tres líderes a nivel estadístico, de, por ejemplo, puntos, rebotes... Además de la de los mejores jugadores según 'The Athletic' referencia [12].

Aquí mostramos la tabla con los 20 mejores jugadores con su coeficiente de correlación, al igual que con Curry.

TOP	Jugadores	C. correlación
1	Michael Jordan	0,454196
2	LeBron james	0,510993
3	Kareem Abdul -Jabbar	0,485778
4	Bill Russell	0,742383
5	Magic Johnson	0,670952
6	Wilt Chamberlain	0,494565
7	Larry Bird	0,862637
8	Shaquille O'Neal	0,339566

9	**Tim Duncan**	0,459782
10	**Kobe Bryant**	0,631836
11	**Hakeem Olajuwon**	0,593328
12	**Oscar Robertson**	0,396427
13	**Kevin Durant**	0,354948
14	**Jerry West**	0,291697
15	**Stephen Curry**	0,672632
16	**Karl Malone**	0,223676
17	**Kevin Garnett**	0,330377
18	**Moses Malone**	0,472891
19	**Julius Erving**	0,583197
20	**David Robinson**	0,427854

FIG.31. Tabla Mejores jugadores coeficientes

Dado estos coeficientes, vamos a intentar determinar porque se han dado estos índices, no vamos a revisarlo todos debido al gran número y que muchas explicaciones se repetirían, por lo tanto, hemos decidido hacerlo en los dos más altos y en el caso de los dos más bajos. Empezando por los de abajo nos encontramos con:

Karl Malone con 0,22367:

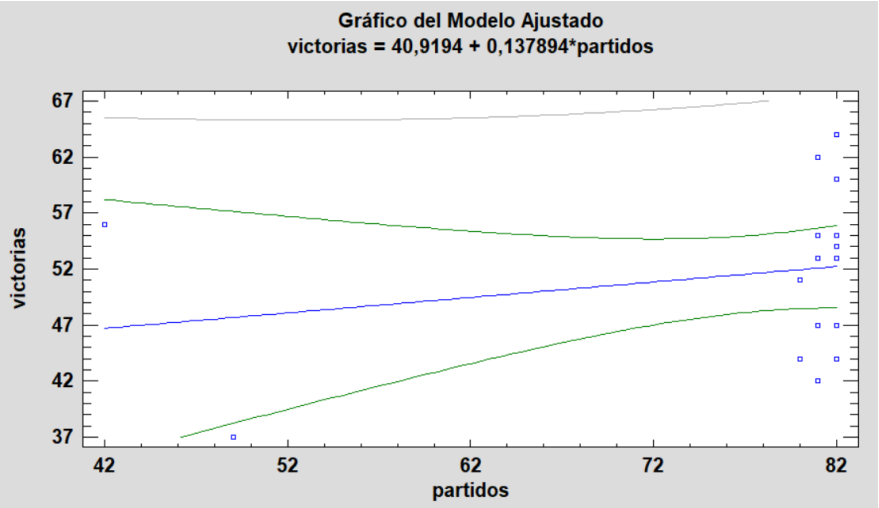


FIG.32. Gráfica modelo ajustado Karl Malone

El estadístico R-Cuadrada indica que el modelo ajustado explica 5,00309% de la variabilidad en victorias. El coeficiente de correlación es igual a 0,223676, indicando una relación relativamente débil entre las variables. Lo que quiere decir este dato es que este jugador no ha sido importante para sus equipos. Es cierto que Karl Malone es ampliamente reconocido como un jugador destacado y se le considera uno de los mejores jugadores de baloncesto de todos los tiempos. Sin embargo, al observar un coeficiente de correlación de 0,22, podemos inferir que su influencia en los resultados no fue tan significativa. Existen varias razones posibles para este coeficiente relativamente bajo. Una de ellas podría ser el hecho de que Malone compartía la pista con John Stockton, quien es el máximo asistente en la historia de la NBA. La presencia de Stockton como una fuerza dominante en la ofensiva del equipo podría haber disminuido la contribución individual de Malone en términos de victorias. En este sentido, ambos jugadores podrían tener coeficientes de correlación más bajos debido a su colaboración conjunta en el éxito del equipo.

Otro jugador con un coeficiente muy bajo es Jerry West con 0’29, esto nos parece el caso más sorprendente de todos, ya que es un jugador super destacado a nivel histórico, incluso es el logo de la liga. Esto puede deberse a distintos factores, uno de ellos puede ser a que compartía pista con Elgin Baylor, otra superestrella y además de Wilt Chamberlain, muchos consideran este como el primer Big Three de la historia, es decir, una unión de tres estrellas de la liga en el mismo equipo.

Por otro lado, encontramos los jugadores con un índice de correlación muy alto como Larry Bird que tiene de coeficiente 0,862637 por lo tanto lo que significa que es el jugador que más importancia ha tenido para su equipo. Esto debido a que en los años que este jugador vestía la camiseta de los Boston Celtics fueron unos de los mayores exponentes de la NBA. En todos los años Bird siempre fue el mayor exponente del equipo y una de las grandes estrellas de la NBA llegando a ganar tres MVPs seguidos cosa que solo lo pueden decir unos pocos.

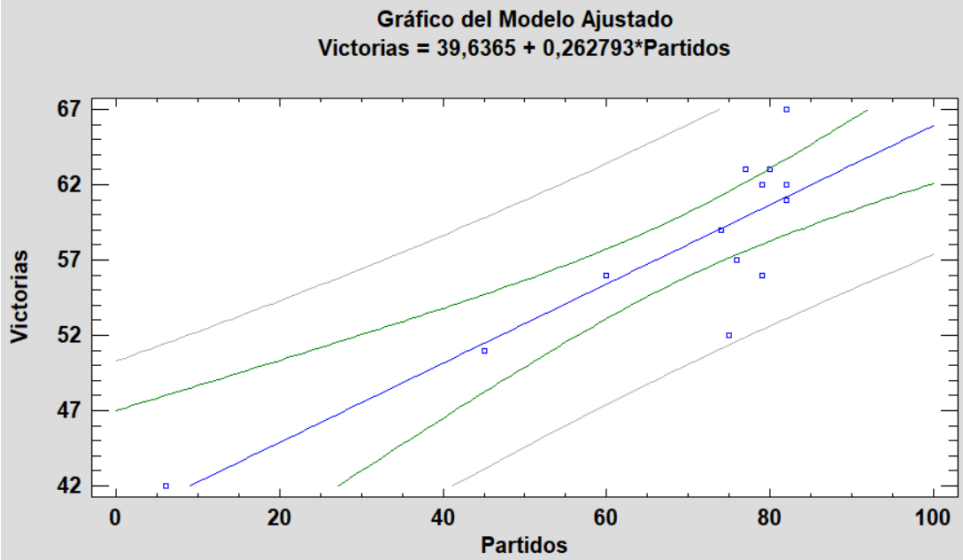


FIG.33. Gráfica modelo ajustado Jerry West

Otro jugador con un índice muy alto, de hecho, el más alto es Bill Russell, el jugador con más anillos en la historia de la liga, con un total de once en trece temporadas, y obviamente como en todos los mostrados en esa lista fue una estrella, la más grande en los 50, y como es obvio el máximo referente del equipo, pese a ser un jugador con pocos puntos por partido para lo que es ser estrella en la NBA, Russell afectaba de otras formas al juego siendo dominante en todos los aspectos del juego como pueden ser rebotes, tapones, robos... En esa época no se contabilizaban robos y tapones, pero según expertos podría haber realizado varios quíntuple-doble (es que un jugador consiga en un único partido superar en los 5 valores cuantificables de referencia (puntos, rebotes, asistencias, tapones y robos) el doble dígito en sus datos estadísticos individuales, es decir, que haga al menos 10 puntos, 10 rebotes...).

El estadístico R-Cuadrada indica que el modelo ajustado explica 55,1132% de la variabilidad en victorias. El coeficiente de correlación es igual a 0,742383, indicando una relación moderadamente fuerte entre las variables. Este jugador sí que ha sido importante para su equipo..

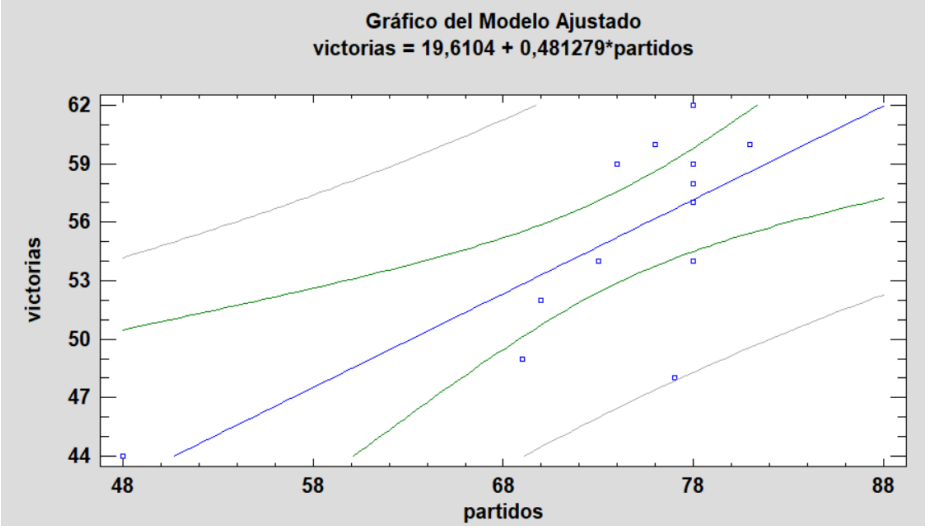


FIG.34. Gráfica modelo ajustado Bill Russell

6.4.- Predicción del MVP

La última parte del trabajo es la creación de un modelo predictivo para determinar quién será el MVP de una temporada a partir de los datos de las temporadas anteriores. Este modelo recibe como parámetros de entrenamiento un dataset, donde cada fila es la información de un jugador una temporada. La información incluye las estadísticas individuales de ese jugador ese año y también las estadísticas globales de su equipo ese año. Para construir el dataset de entrenamiento, hemos tenido que limpiar los datasets de estadísticas de los equipos, de los jugadores y la información de los MVPs. Una vez limpios los hemos unido en un dataset final con toda la información necesaria.

```
1 df_final
```

	Player	Pos	Age	G	GS	MP	FG	FGA	FG%	3P	...	T-STL	T-BLK	T-TOV	T-PF	T-PTS	Rank	First	Pts Won	Pts Max	Share
0	Kareem Abdul-Jabbar	5	36	80	80.0	2622.0	716	1238	0.578	0.0	...	695.0	481.0	1537.0	1931	9696	4	3.0	153.0	760.0	0.201
1	Michael Cooper	2	27	82	9.0	2387.0	273	549	0.497	38.0	...	695.0	481.0	1537.0	1931	9696	0	0.0	0.0	0.0	0.000
2	Calvin Garrett	3	27	41	0.0	478.0	78	152	0.513	2.0	...	695.0	481.0	1537.0	1931	9696	0	0.0	0.0	0.0	0.000
3	Magic Johnson	1	24	67	66.0	2567.0	441	780	0.565	6.0	...	695.0	481.0	1537.0	1931	9696	3	5.0	305.0	760.0	0.401
4	Eddie Jordan	1	29	3	0.0	27.0	4	8	0.500	0.0	...	695.0	481.0	1537.0	1931	9696	0	0.0	0.0	0.0	0.000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
19073	Daniel Theis	5	29	21	6.0	393.0	67	112	0.598	10.0	...	521.0	430.0	1095.0	1542	9671	0	0.0	0.0	0.0	0.000
19074	Brodric Thomas	2	25	12	0.0	60.0	8	18	0.444	2.0	...	521.0	430.0	1095.0	1542	9671	0	0.0	0.0	0.0	0.000
19075	Derrick White	2	27	26	4.0	713.0	94	230	0.409	34.0	...	521.0	430.0	1095.0	1542	9671	0	0.0	0.0	0.0	0.000
19076	Grant Williams	4	23	77	21.0	1875.0	205	432	0.475	106.0	...	521.0	430.0	1095.0	1542	9671	0	0.0	0.0	0.0	0.000
19077	Robert Williams	5	24	61	61.0	1804.0	271	368	0.736	0.0	...	521.0	430.0	1095.0	1542	9671	0	0.0	0.0	0.0	0.000

19078 rows × 64 columns

FIG.34. Dataset final para entrenar el modelo

Limpiamos los DataFrame

```
# obtenemos el año a partir de la temporada
df_team["Year"] = [int(s[:4]) for s in df_team["Season"]]

# años con información del MVP
years = set(df_mvp["Year"])

# eliminamos la informacion de jugadores y equipos sin informacion del MVP
df_team = df_team[df_team["Year"].isin(years)]
df_player = df_player[df_player["Year"].isin(years)]

# limpiamos nombres de los jugadores
df_player["Player"] = [p.replace('*', '') for p in df_player["Player"]]

# corregimos los nombre de equipo
dic_teams = {
    'BRK':'NJN' ,
    'CHH': 'CHA',
    'CHO': 'CHA',
    'KCK':'SAC',
    'NOK':'NOH',
    'NOP':'NOH',
    'SDC':'WAS',
    'SEA':'OKC',
    'TOT':'TOT',
    'VAN':'MEM',
    'WSB':'WAS',
}
df_player["Team"] = df_player["Tm"].replace(to_replace=dic_teams)
df_mvp["Team"] = df_mvp["Tm"].replace(to_replace=dic_teams)

# cambiamos posición por un número
dic_posiciones = {
    'C': 5,
    'PF': 4,
    'SF': 3,
    'SG': 2,
    'PG': 1,
}
df_player["Pos"] = df_player["Pos"].replace(to_replace=dic_posiciones)
```

FIG.35. Código de limpieza de los datasets.

Una vez realizada la limpieza y la creación del dataset de entrenamiento, es hora de crear el modelo usando la librería de Python, sklearn. Esta librería es muy utilizada en el campo del aprendizaje automático y nos ha ayudado a crear un modelo predictivo de forma fácil y eficiente. Para hacer la predicción hemos utilizado un modelo basado en máquinas de vectores de soporte (*support-vector machines*, SVM) que asigna a cada jugador una puntuación. El jugador que mayor puntuación obtiene se considera el MVP. Nosotros utilizamos las estadísticas desde el año 1984 hasta el hasta el 2020 para entrenar y las de los años 2021 y 2022 para comprobar lo bien que funciona. En las dos temporadas hemos obtenido un mean square error, que es el error que hemos usado, bastante pequeño, aproximadamente del 0,0024 en las dos temporadas.

El modelo ha conseguido obtener, o sea predecir, quién fue el MVP de la temporada 2021-22.

```
1 list(reversed(list(sorted(final))))[:15]
```

```
[(0.15516148954714212, 'Nikola Jokić'),
 (0.13925653227190868, 'Luka Dončić'),
 (0.13620784472973088, 'Stephen Curry'),
 (0.13469213621200268, 'Giannis Antetokounmpo'),
 (0.12888127205723363, 'Russell Westbrook'),
 (0.11999262042933476, 'Bradley Beal'),
 (0.11710377572845251, 'Julius Randle'),
 (0.1154988876672319, 'Damian Lillard'),
 (0.11439979734556402, 'Jayson Tatum'),
 (0.11173775608249933, 'Trae Young'),
 (0.10695768959431759, 'Zion Williamson'),
 (0.10479104500756117, 'Devin Booker'),
 (0.10405391990821675, 'Joel Embiid'),
 (0.0913044115544551, 'Zach LaVine'),
 (0.09046551154700097, 'Donovan Mitchell')]
```

El modelo ha acertado el MVP 2021-2022

Nikola Jokić

FIG.36. Predicción del MVP de la temporada 2021-22

En cambio, el de la temporada 2022-23 no lo predice correctamente ya que el modelo determina que debería ser Nikola Jokić pero el verdadero MVP según las votaciones fue Joel Embiid. Aunque el modelo se acerca bastante ya que hubo una gran polémica sobre el premio otorgado a Embiid y muchos expertos determinaron que debería haber sido el serbio (como predice nuestro modelo) el ganador.

```
1 list(reversed(list(sorted(final))))[:15]
```

```
[(0.17573093666864423, 'Nikola Jokić'),
 (0.16723016614488503, 'Giannis Antetokounmpo'),
 (0.1591324586700322, 'Joel Embiid'),
 (0.14359398326391112, 'Trae Young'),
 (0.14226858554841235, 'Jayson Tatum'),
 (0.14001664742348538, 'DeMar DeRozan'),
 (0.13640442878726175, 'Luka Dončić'),
 (0.11711398441617213, 'Karl-Anthony Towns'),
 (0.11478008728399977, 'LeBron James'),
 (0.11002542123030098, 'Kevin Durant'),
 (0.1056018520476246, 'Devin Booker'),
 (0.10363194224932035, 'Ja Morant'),
 (0.10201083574906242, 'Stephen Curry'),
 (0.10140667385494781, 'Russell Westbrook'),
 (0.10121838729168672, 'Jaylen Brown')]
```

El modelo NO ha acertado el MVP 2022-2023

Joel Embiid, aunque hubo mucha polémica y muchos expertos opinaron que debería ser Nikola Jokić

;-)

FIG.37. Predicción del MVP de la temporada 2021-22

7. Validación y despliegue

Para validar los resultados obtenidos en este proyecto y comprobar la efectividad de los análisis realizados, hemos realizado etapas de validación. En primer lugar, hemos verificado la integridad de los datos utilizados mediante técnicas de limpieza y preprocesamiento. Limpiamos los datos faltantes y los valores atípicos para asegurar la calidad de los datos.

Posteriormente, hemos aplicado técnicas de análisis exploratorio de datos (EDA) para examinar la distribución de las variables y buscar posibles relaciones entre ellas. Se han utilizado gráficos para visualizar y evaluar la coherencia de los resultados. Se han realizado pruebas estadísticas para respaldar las conclusiones y validar las hipótesis planteadas.

En cuanto al despliegue del proyecto, hemos creado una aplicación web interactiva para presentar y visualizar los resultados obtenidos. Esta aplicación permite a los usuarios explorar los datos y obtener información relevante sobre las estadísticas de los jugadores y el rendimiento de los equipos en la NBA. Además, se han implementado gráficos interactivos y filtros para facilitar la navegación y la comprensión de los resultados.

La aplicación web ha sido desarrollada utilizando tecnologías como HTML y CSS. Se ha alojado en un servidor web para que esté accesible en línea y se ha proporcionado un enlace para que los interesados puedan acceder a ella y explorar los resultados del proyecto.

En resumen, hemos validado los resultados obtenidos mediante técnicas de limpieza de datos y análisis estadístico. Además, hemos desplegado una aplicación web interactiva para mostrar de manera accesible y amigable los resultados del proyecto.

8. Conclusiones

En este proyecto, se ha llevado a cabo un análisis exhaustivo de datos de jugadores y equipos en la NBA. A continuación, se presentan las conclusiones más significativas obtenidas a partir de los puntos 5 y 6 del proyecto:

El análisis estadístico realizado se centra en las correlaciones entre las estadísticas de los jugadores y los votos obtenidos en los premios de la NBA, como el MVP, el ROY, el SMOY y el DPOY. Se observaron correlaciones significativas entre las votaciones y las siguientes estadísticas:

- Para el MVP: Win Shares (WS), Win Shares per-48 minutes (WS/48) y Puntos (PTS) se muestran correlaciones destacadas con los votos obtenidos.
- Para el ROY: Asistencias (AST), Puntos (PTS) y Minutos jugados (MP) se muestran correlaciones significativas con los votos obtenidos.
- Para el SMOY: Asistencias (AST), Puntos (PTS) y Minutos jugados (MP) se muestran correlaciones destacadas con los votos obtenidos.
- Para el DPOY: Rebotes (TRB), Tapones (BLK) y Robos (STL) se muestran correlaciones significativas con los votos obtenidos.

Además, se ha empleado el Análisis de Componentes Principales (PCA) para reducir la complejidad de los datos de jugadores y equipos, identificando variables clave que explican la variabilidad en los conjuntos de datos.

En cuanto a la competitividad en la NBA, se han calculado índices como el índice Sigma y el índice Herfindahl-Hirschman (HICB). Hemos encontrado que las temporadas más competitivas fueron 1956-57, 1955-56, 1954-55, 1950-51 y 1976-77 según el índice Sigma. El análisis del HICB también ha mostrado que la temporada 1950-51 fue altamente competitiva.

En resumen, el análisis estadístico realizado nos ha revelado las correlaciones entre las estadísticas de los jugadores y los votos en los premios de la NBA, así como la evaluación de la competitividad en diferentes temporadas.

En cuanto al modelo predictivo, observamos que se acerca bastante y podría llegar a ser útil, para implementarlo en trabajos futuros o simplemente para la temporada siguiente.

En cuanto a los trabajos futuros, este proyecto sienta las bases para realizar análisis más profundos y sofisticados en el campo del baloncesto y la NBA. Algunas posibles áreas de investigación futura podrían incluir el análisis de datos en tiempo real durante los partidos, el desarrollo de modelos predictivos para predecir el rendimiento de los jugadores o el estudio de estrategias de juego específicas utilizadas por los equipos.

En relación con otras asignaturas, este proyecto tiene una fuerte conexión con las asignaturas de Estadística y Análisis de Datos. Se han aplicado diversas técnicas estadísticas para analizar los datos y se han utilizado métodos de visualización de datos para presentar los resultados de manera clara y concisa. Además, el proyecto también está relacionado con asignaturas de Ciencias del Deporte, ya que proporciona información valiosa para comprender el rendimiento de los jugadores y los equipos en el contexto del baloncesto profesional.

9. Referencias

[1] *Mandić, R., Jakovljević, S., Erčulj, F., & Štrumbelj, E. (2019). Trends in NBA and Euroleague basketball: Analysis and comparison of statistical data from 2000 to 2017. PLOS ONE, 14(10), e0223524. https://doi.org/10.1371/journal.pone.0223524*

[2] Groothuis, P. A., & Hill, J. C. (2004). Exit Discrimination in the NBA: A Duration Analysis of Career Length. Economic Inquiry, 42(2), 341-349. https://doi.org/10.1093/ei/cbh065

[3] Gómez, M. I., Gasperi, L., & Lupo, C. (2016). Performance analysis of game dynamics during the 4th game quarter of NBA close games. International Journal of Performance Analysis in Sport, 16(1), 249-263. https://doi.org/10.1080/24748668.2016.11868884

[4] Rabinal, S. (2022, 14 enero). Los cambios en las reglas de la NBA en la historia a través de 10 jugadores (parte I). Sporting News Spain. https://www.sportingnews.com/es/nba/news/cambios-reglas-nba-historia-10-jugadores/dc99ziwdv4kn1xqa2e7nfm95i

[5] NBA Salary Cap History | Basketball-Reference.com. (s. f.). Basketball-Reference.com. https://www.basketball-reference.com/contracts/salary-cap-history.html

[6] NBA Contracts Summary | Basketball-Reference.com. (s. f.). Basketball-Reference.com. https://www.basketball-reference.com/contracts/

[7]Criado, R., García, E., Pedroche, F. F., & Romance, M. (2013). A new method for comparing rankings through complex networks: Model and analysis of competitiveness of major European soccer leagues. Chaos, 23(4), 043114. https://doi.org/10.1063/1.4826446

[8]Rubio, J., & Fotografía, A. (2022, 24 diciembre). Los jugadores que obligaron a cambiar las reglas en la NBA. Diario AS. https://as.com/baloncesto/fotorrelato/los-jugadores-que-obligaron-a-cambiar-las-reglas-en-la-nba-f/

[9] S.J. Ibáñez, J. García, S. Feu, I. Parejo, M. Cañadas (2008, septiembre). La eficacia del lanzamiento a canasta en la NBA: Análisis multifactorial.

https://ccd.ucam.edu/index.php/revista/article/view/132/123

[10] José Luis Osorio Guajardo (2017). Análisis de la eficiencia en los equipos de la NBA para las temporadas 2014/2015 y 2015/2016.

https://zaguan.unizar.es/record/62511/files/TAZ-TFG-2017-2661.pdf

[11] D. Javier Martínez García (2008). Métodos estadísticos en competiciones deportivas de baloncesto: NBA.

https://uvadoc.uva.es/bitstream/handle/10324/43847/TFG-G4618.pdf

[12] The Athletic NBA Staff (2022, Febrero 23). NBA 75: Top 75 NBA players of all time, from MJ and LeBron to Lenny Wilkens

https://theathletic.com/3137873/2022/02/23/the-nba-75-the-top-75-nba-players-of-all-time-from-mj-and-lebron-to-lenny-wilkens/

[13] Criado, R., García, E., Pedroche, F. F., & Romance, M. (2013). A new method for comparing rankings through complex networks: Model and analysis of competitiveness of major European soccer leagues. Chaos, 23(4), 043114.

https://doi.org/10.1063/1.4826446

10. Anexos

[1] Búsqueda de curiosidades y datos interesantes.

https://upvedues my.sharepoint.com/:w:/r/personal/jmuevel_upv_edu_es/_layouts/15/Doc.aspx?sourcedoc=%7B69F68272-7AE1-4838-BECC-401A70679434%7D&file=Proy.docx&action=default&mobileredirect=true&cid=8aa402d9-7d53-4064-a6fc-985b95e5705f

[2] Informe semanal 3

https://upvedues-my.sharepoint.com/:w:/g/personal/atarsor_upv_edu_es/EUJxbw1TzolHrUWx31TSVSwBlLwufL0GGKSez3eAAjPeNg?e=ChtcHI

[3] Informe semanal 4.

https://upvedues-my.sharepoint.com/:w:/r/personal/mhurben_upv_edu_es/_layouts/15/Doc.aspx?sourcedoc=%7B0CF037FA-838B-49D7-B84C-

42DB20AF0DC8%7D&file=informeSemanal4_NBA.docx&action=default&mobileredirect=true&DefaultItemOpen=1&login_hint=ATARSOR%40upv.edu.es&ct=1687104326056&wdOrigin=OFFICECOM-WEB.START.EDGEWORTH&cid=d51899e5-7858-45b1-af27-72c15b4acff7&wdPreviousSessionSrc=HarmonyWeb&wdPreviousSession=3f7e433a-e85c-497a-a6de-e3b0095825ab

[4] Informe semanal 5.

https://upvedues-my.sharepoint.com/:w:/r/personal/mhurben_upv_edu_es/_layouts/15/Doc.aspx?sourcedoc=%7BA0E2BAA4-E946-4B6B-9DAE-F379F79B5BEA%7D&file=Informe%20Semanal%205.docx&action=default&mobileredirect=true&DefaultItemOpen=1&login_hint=ATARSOR%40upv.edu.es&ct=1687104261107&wdOrigin=OFFICECOM-WEB.START.EDGEWORTH&cid=fdb50c4a-162a-482b-ab43-524778dcc599&wdPreviousSessionSrc=HarmonyWeb&wdPreviousSession=3f7e433a-e85c-497a-a6de-e3b0095825ab

[5] Informe semanal 8

https://upvedues-my.sharepoint.com/:w:/g/personal/atarsor_upv_edu_es/EQtN5tfOfqNPgpLYtwH5kU0BQuJqiTiXHZ0zytJlyfHjWQ?e=wbl2S8

[6] Presentación

https://upvedues-my.sharepoint.com/:p:/g/personal/atarsor_upv_edu_es/ETcf6r1U2MhCsli6QkwSAYQBvw3tesX8OzAzvO9fIMi8jw?e=uGxwAZ

[7] PCA

https://upvedues-my.sharepoint.com/:w:/g/personal/mhurben_upv_edu_es/EcKMKBwAiVxOpC6Ok5l77rMBbgD4d2XNj_mxbV2naNm4yA?e=0mIGeG

[8] Correlación premios

https://upvedues-my.sharepoint.com/:w:/g/personal/mhurben_upv_edu_es/EUYKu-xEbrtBrSPilw2J3HsBeIISN-Xjo7DArCZBZaCzeQ?e=r5zJMr

[10] Índice de competitividad

https://upvedues-my.sharepoint.com/:w:/g/personal/mhurben_upv_edu_es/EfAyIDT_HN9NlNU9a5UJBfYBElt3OafByCtVujHveQbyVw?e=iMydYL

[11] Video del proyecto

https://www.youtube.com/watch?v=NwrkyGaT5yw

[12] Presentación

Presentación2.pptx