***Actividad 1: Propuesta proceso KDD completo*.**

Análisis estadístico de la NBA para mejorar la toma de decisiones de los equipos.

*Miguel Houghton López*

8/12/2021



**Motivación y objetivos**

La NBA es la **mayor liga de baloncesto del mundo**. Es una de las más lucrativas del mundo y mueve aproximadamente **5000 millones de dólares al año**. Los contratos de los jugadores pueden variar entre aproximadamente 1 millón de dólares al año (mínimo salarial de rookies) y casi 50 millones al año (Stephen Curry es el jugador mejor pagado con 46 millones al año).

Dentro de esta liga, existe un sorteo a principio de la temporada, el **Draft de la NBA**, donde se escogen los jugadores universitarios o internacionales sin experiencia en la liga para que marquen el futuro del equipo con su talento.

Este Draft, es **crucial** puesto que el rumbo de un equipo durante los siguientes años puede ser marcado por un jugador seleccionado en este sorteo. Por ello, la necesidad de **realizar una decisión acertada** por cada equipo a la hora de escoger a los jugadores provenientes de ligas universitarias es esencial.

Durante el paso de los años, se ha podido ver como muchas **estrellas** **actuales de la NBA**, **no fueron seleccionadas en la que debería haber sido su posición adecuada** durante su sorteo del Draft. Por ejemplo, superestrellas como *Stephen Curry* (elección 7ª del Draft de 2009) o *Demar Derozan* (elección 9ª del Draft de 2009) son ejemplos de **jugadores que fueron infravalorados** por varios equipos la noche del Draft.

La otra cara de la moneda muestra como **equipos seleccionaron en posiciones altas a jugadores** prometedores con carreras finalmente mediocres o **que no llegaron a triunfar en la NBA**. Jugadores como *Jonny Flynn* (elección 6ª del Draft de 2009) o *Enes Kanter* (elección 3ª del Draft de 2011) son ejemplos de **errores** en sus elecciones de sus respectivos Draft.

Pero hay un **factor aún más complejo** de detectar por los equipos y que **les afecta más gravemente**. Los jugadores al cabo de dos años optan a **renovación de contrato**, y pueden **pasar de un contrato** como novatos de 1 millón de euros al año **a contratos multimillonarios de larga duración**.

El **error a la hora de renovar a estos jugadores puede ser desastroso** para un equipo, puesto que pueden gastar gran parte de su presupuesto en **jugadores sin futuro** y tener grandes dificultades para deshacerse de jugadores así, **con grandes contratos y de larga duración**.

**Este error es bastante común** entre los equipos, puesto que las **estadísticas** de los jugadores pueden ser **buenas durante los dos primeros** **años**, y luego acabar siendo jugadores mediocres a la larga.

El **objetivo** de este trabajo va a ser **investigar qué patrones** podemos observar en la gran de cantidad de **estadísticas** de los jugadores para **determinar su futuro** y así permitir que los **equipos puedan tomar mejores decisiones** a la hora de fichar y renovar jugadores con el objetivo de formar las mejores plantillas posibles.

Estos patrones permitirán definir **cuáles son las estadísticas** de los jugadores **más importantes a tener en cuenta** y así poder tomar la decisión correcta de **renovar a un jugador o traspasarle** una vez terminada su época de novato (los dos primeros años, con un contrato sustancialmente menor).

Para poder llevar esto a cabo, se deberán estudiar **determinados grupos de jugadores pertenecientes a la misma clase del Draft** para así poder compararlos y comprobar con las estadísticas de hoy en día **qué jugadores fueron infravalorados o sobrevalorados**.

Además, **las estadísticas que se deben estudiar de cada agrupación de jugadores serán las de sus primeras temporadas** (de 2 – 4 temporadas) para así observar el comportamiento de los jugadores y poder **determinar si la renovación de un jugador podría haber sido evitada.**

Consecuentemente, comprobar si una **gran cantidad de dinero ahorrado** o **la formación de una mejor plantilla** podría haber sido una realidad de haber tenido en cuenta las conclusiones que se pretenden obtener de este trabajo.

**Base de Datos**

El origen principal de los datos viene de un conjunto de datasets de estadísticas históricas de la NBA encontrados en el siguiente enlace de kaggle:

<https://www.kaggle.com/drgilermo/nba-players-stats>

En este link vienen un total de 3 archivos CSV, de los cuales utilizaremos dos de ellos para formar nuestra base de datos:

* **Seasons\_Stats.csv**: archivo con estadísticas de todos los jugadores durante todas las temporadas desde 1950 hasta 2017.
* **Player\_data.csv**: archivo con un resumen de todos los jugadores que participaron en la NBA en el periodo anteriormente mencionado.

Los **datos de mayor interés** de los anteriores archivos seleccionados serán (en negrita su nombre de columna en base de datos):

* *Porcentajes de tiro*: porcentajes de tiro de 3, de tiro de 2 y de tiros libres (**FT%**, **3P%**,**2P%**)
* *Volumen de tiro*: cantidad de tiros realizados por cada uno de los jugadores (**FTA**, **2P**, **3P**)
* *Estadísticas de juego no relativas al tiro*: asistencias, rebotes defensivos y ofensivos (**AST**, **ORB**, **DRB**)
* *Estadísticas defensivas*: Robos y tapones (**STL**y **BLK**)
* *Pérdidas* (**TOV**)
* *Posición de los jugadores* (**Pos**)
* *Año de comienzo en la NBA y edad con la que comenzaron* (**start\_year**y **age**)

La tabla con la que trabajaremos será el resultado de **combinar ambos datasets**, para así obtener no solo estadísticas de los jugadores en todas las temporadas, si no también añadirles su año de comienzo en la NBA.

Por supuesto, esta tabla será la base sobre la que se realizarán todas las tareas de limpieza, transformación y demás acciones.

Esta tabla original estará compuesta por un total de**54 columnas** y aproximadamente **24700 filas**. Esta tabla contendrá información con estadísticas de todo tipo las cuales trabajaremos posteriormente para obtener sólo las más relevantes.

**Salida buscada del proceso KDD**

La salida buscada del proceso KDD son patrones que permitan confirmar la razón por la cual ciertos jugadores han resultado ser estrellas de la NBA después de haber sido infravalorados en su elección del Draft y encontrar a su vez patrones que permitan detectar anomalías en ciertos jugadores que hayan resultado mediocres con el paso del tiempo.

Por ello, los tipos de salida principales de este proyecto son:

* **Predicción** de las **estadísticas** que los jugadores vayan a conseguir **a** **lo largo en su carrera** observando su **comportamiento** durante los **primeros años**.
* **Pronóstico** para determinar **qué jugadores terminarán justificando una renovación** o un contrato y **qué jugadores terminarán con carreras sin éxito en la NBA**.
* La **segmentación** se deberá ver a la hora de **distinguir jugadores de distintas generaciones y observar el comportamiento en relación a su clase del Draft**. También se podrán ver segmentaciones en cuanto a tipos de jugadores (posición de juego).

**Descripción detallada del proceso KDD**

*Selección (explicado previamente en ‘Base de datos’)*

Como hemos comentado previamente en el apartado *Base de Datos,*para llevar a cabo este trabajo se deberán extraer datos en bruto de estadísticas de todo tipo de jugadores en diversas temporadas de la NBA.

En concreto, se seleccionan dos datasets y se combinan para formar la base de datos original.

*Data Cleaning*

Las tareas de limpieza de datos se hacen notables al tener una **gran cantidad de datos de distintas épocas** y al **no obtener un origen oficial de la NBA** (los datasets originales vienen de kaggle).

El primer aspecto esencial a tener en cuenta en la limpieza de datos es el factor de que el **tiro de tres no se hizo oficial en la NBA hasta el año 1980**.

Este punto de inflexión es bastante **importante** ya que se ven **afectadas unas 5700 filas de 25000**. Las columnas referentes tanto a **volumen como a porcentajes de este tipo de tiros** deberían ser **eliminadas**.

**No obstante**, dada la **importancia** **de estas columnas en el baloncesto moderno**, se ha decidido que se **eliminarán los registros enteros en los que no se tenga en cuenta estas estadísticas debido a su importancia**. Esto implica que el trabajo se realizará con **estadísticas únicamente posteriores a 1980** y el **número** **de filas se reducirá a unas 19000**.

Por otra parte, existen algunas **columnas con el tipo de datos incorrecto**. Es el caso de las **columnas tipo porcentaje**, los cuales deberían ser **tipo decimal**,y en el origen de datos se han creado como texto.

Por lo demás, la única tarea de limpieza restante ha sido **modificar ciertos valores desconocidos o nulos** a un **valor por defecto** para evitar fallos en los cálculos (por ejemplo, si un jugador una temporada no tira triples y queremos hacer una media de tiros de tres en sus primeros años, el valor de esa columna no debe ser nulo, sino un 0).

*Data Transformation*

En esta fase se llevarán a cabo las tareas de **escalado de los datos** y **clasificación** por **subconjuntos** de datos.

En cuanto a **clasificación (segmentación)**, dependiendo de **qué clase del Draft se quiera estudiar** (si queremos averiguar qué jugadores fueron infravalorados y cuáles sobrevalorados en una determinada clase del Draft), se deberán crear **agrupaciones de datos filtrando por el año de comienzo de la carrera de cada jugador**.

Además, puesto que el objetivo del trabajo conlleva estudiar las primeras temporadas de cada uno de los jugadores en la NBA (antes de una posible renovación de contrato), se deberán hacer **agrupaciones de datos con solo las estadísticas de las primeras temporadas** de cada uno de los jugadores seleccionados previamente.

Por otra parte, en cuanto a escalado de los datos, es **muy importante** que el **modelo tenga en cuenta a todas las estadísticas por igual**, ya que por ejemplo un porcentaje de tiro es un número menor que 1 y es igual o más importante que una cifra de asistencias (que puede ser un valor mayor que 100).

Por lo que, una vez realizada la agrupación de datos en subconjuntos, deberemos proceder al escalado de estos datos (**normalización**)

*Data Reduction*

En el trabajo este apartado se puede ver reflejado a la hora de **escoger las características más relevantes en el juego de un jugador**.

Por ejemplo, el modelo a desarrollar deberá tener en cuenta **estadísticas esenciales** como las **mencionadas en el apartado ‘Base de datos’ donde se mencionan los datos de mayor interés**.

Otros datos como **partidos jugados, minutos u otras columnas también vistas que no son relevantes ya que no son relativas estrictamente al juego de los jugadores** deberán ser excluidas para mejorar la eficiencia del modelo y para que los **resultados se ajusten lo máximo posible a lo que genera valor en el proyecto** (estudiar el comportamiento en partido de los jugadores).

*Data Mining*

La etapa de minería de datos estará compuesta por la **aplicación de algoritmos** para la obtención de los mejores resultados posibles.

El algoritmo que mejor encajaría en este desarrollo sería un **Clustering / Nearest Neighbour** ya que lo que se busca en este trabajo son **patrones que distingan algunos jugadores de otros**.

La generación de grupos por distintas estadísticas, permitirán ver **patrones comunes entre ciertos grupos de jugadores**.

El **mejor resultado** de este trabajo podría ser la **agrupación de jugadores que actualmente son superestrellas distinguidos de agrupaciones de jugadores que actualmente son malos jugadores** basándonos en datos de los primeros años de las carreras de dichos jugadores.

El proceso se basaría en **escoger estadísticas** para agrupar como por ejemplo de **tiro, defensivas, malas, ofensivas no relativas al tiro**... y aplicar **algoritmos de clustering** para obtener agrupaciones de jugadores de la misma clase del Draft según sus **similitudes en el juego**.

**Información adicional y trabajo futuro**

Comentar como información adicional que los **datos estadísticos podrían ser representados en herramientas de visualización** y ser útiles para cada uno de los equipos a la hora por ejemplo de **estudiar una renovación / fichaje o estudiar las estadísticas de un equipo rival**.

Como **trabajo futuro**, podríamos añadir la posibilidad de **analizar más en profundidad la justificación de los salarios** cobrados por los jugadores en cada caso. Añadiendo las cifras de salarios a nuestra base de datos original, se podría hacer una **correlación entre las estadísticas analizadas en este trabajo con las cifras económicas**.

Además, mi intención es que **esta base componga una parte de mi TFM**, el cual podrá complementar esta base de datos con datos actuales obtenidos mediante **scrapping** (debido a la dificultad de obtención de datos oficiales de la NBA).

La intención de mi TFM será **predecir otro tipo de factores que puedan mejorar el posicionamiento de cada equipo** como por ejemplo un caso relacionado al estudiado en este trabajo: la **predicción de recuperación de lesiones de los jugadores**. De la misma manera que en este trabajo realizamos una predicción del comportamiento de un jugador observando sus estadísticas de sus primeros años, la intención es **predecir si un jugador llegará a obtener el mismo rendimiento observando sus estadísticas de juego previas a su lesión**.

Esta idea deberá tener en cuenta muchos **factores como físicos o la constancia en las estadísticas** de los jugadores y se **tendrá que comprobar con casos reales de jugadores** **que han mantenido su nivel / han mejorado tras una lesión y jugadores que han caído en picado tras una lesión grave en sus estadísticas**.