Trabajo de Final de Grado

Informe seguimiento I

Machine Learning para la predicción de eventos en la NBA

Albert Villar Ortiz

Universidad Autónoma de Barcelona

ÍNDICE GENERAL

Introducción

La AI (*Artificial Intelligence*) es un campo cada vez más utilizado en nuestra sociedad para solucionar múltiples problemas del día a día: des de crear vehículos con la capacidad de conducir de forma autónoma hasta la generación de un sistema inteligente capaz de gestionar la domótica de tu residencia. Entrando más en detalle, el sector del *Machine Learning* está siendo poco a poco una parte indispensable para las grandes empresas que quieran obtener resultados concluyentes a partir de la enorme cantidad de datos que almacenan.

Pero no solo en el ámbito empresarial podemos encontrar esta tecnología, en los últimos años hemos podido ver como se han implementado sistemas de aprendizaje automático en muchos equipos deportivos con el fin de poder sacar provecho de forma más eficiente todas las estadísticas que registran día a día. La tendencia es tan clara, que podemos encontrar casos donde el análisis de datos ha propiciado un cambio brusco en el mercado o en la forma de realizar las cosas.

Entre todas las diferentes ligas que hay de baloncesto a nivel mundial, la NBA (*National Basketball Association*) [1] es tanto la más seguida como la que más dinero a recaudado. Tanto es así, que la NBA se ha convertido en la cuarta liga que más dinero ha generado en la temporada 2015-2016, por detrás de otras ligas todo poderosas como la NFL, logrando la increíble cifra de 4.8 billones de dólares [2].

De forma interna, la NBA está formada por un total de 30 franquicias subdivididas en dos conferencias: oeste y este. Además, cada una de ellas, esta subdividida en tres divisiones formadas por 5 equipos. Es importante destacar la estructuración de la NBA, pues en función de tu localización contaras con un calendario u otro. Si entramos aún más en detalle, podemos observar como cada equipo jugará:

- 4 veces contra los equipos que conviven en su misma división
- Entre 3 y 4 veces contra los equipos de las otras divisiones de su conferencia
- 2 veces contra los equipos que conviven en la otra conferencia

Al finalizar la temporada regular, todas las franquicias habrán disputado un total de 82 partidos divididos en partes iguales entre encuentros de local y visitante. Finalmente, los 8 mejores equipos de cada conferencia realizaran una eliminatoria al mejor de cinco partidos, obteniendo al fin dos equipos, cada uno campeón de su propia conferencia, que se enfrentaran por el título de la NBA.

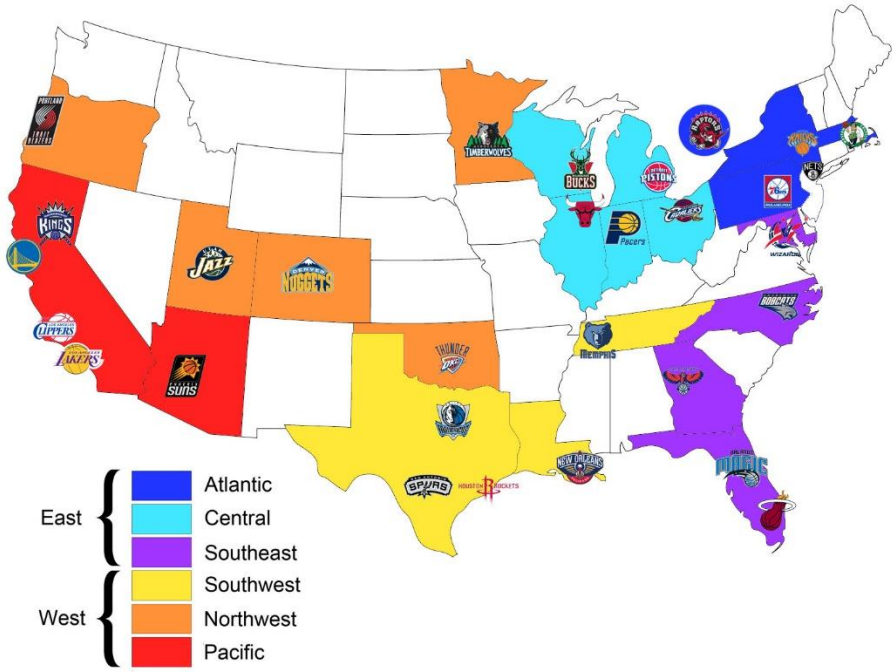

Ilustración 1: Representación gráfica de la estructuración de la NBA *[3]*

Finalmente, este proyecto busca unir estos dos mundos elaborando un *dataset* propio para predecir tanto el ganador de un partido como la supuesta anotación por parte de cada uno de los equipos. Para lograr dicha hazaña, se utilizarán diversos algoritmos de *Machine Learning*, los cuales serán analizados para determinar cuál funciona mejor en esta casuística.

Estado del arte

El uso de *Machine Learning* en el ámbito deportivo, no es tan solo un objetivo propiciado por los seguidores, puesto que a nivel interno las franquicias también utilizan dichos sistemas para optimizar el uso de sus estadísticas.

Tanto es así, que en el último año hemos podido detectar una tendencia muy significativa en la manera de jugar en la NBA, propiciada por el análisis de datos. Año tras año, el porcentaje de intentos en el lanzamiento de 3 puntos ha aumentado considerablemente como se puede ver representado en esta gráfica:

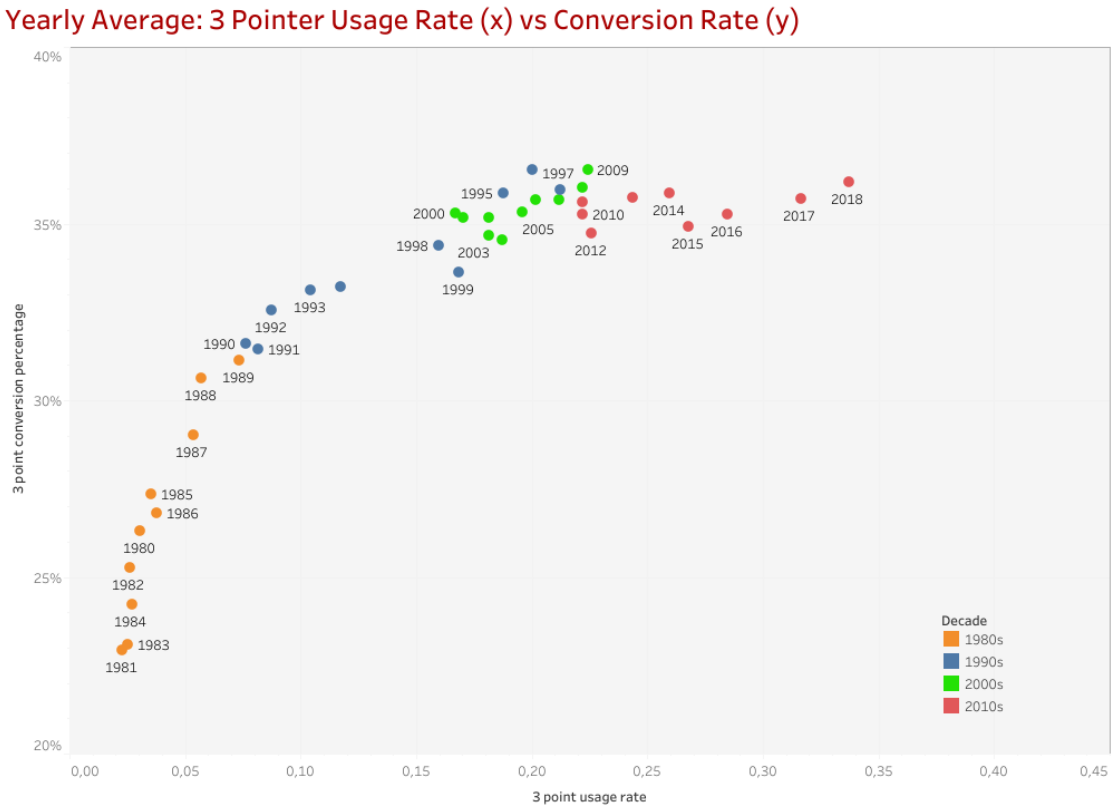

Ilustración 2: Porcentaje de uso del triple y su respectivo acierto en 4 épocas distintas *[4]*

La explicación científica nos la puede proporcionar Daryl Morey, director general de la franquicia Houston Rockets, el cuál analizo los datos y detecto que los lanzamientos con mejor valor de retorno son los mates y los lanzamientos de 3 puntos, siendo pues los lanzamientos lejanos de dos puntos el peor lanzamiento posible [5].

Pero no solo eso, sino que actualmente existen sistemas muy sofisticados capaz de calcular la probabilidad de cada uno de los lanzamientos en riguroso directo, mediante lo que Rajiv Maheswaran (director de Second Spectrum) llama: Ingeniería de los puntos [6].

Además de las franquicias, múltiples seguidores han desarrollado diversas investigaciones aprovechando la gran cantidad de datos que genera la NBA, con el objetivo de predecir eventos. Uno de los proyectos más significativos recibe el nombre de NBA Oracle, desarrollado por tres estudiantes de Carnegie Mellon University [7], el cuál concluyeron que la tecnología que proporcionaba mejores resultados para la predicción de ganadores en la NBA era realizar una regresión lineal, obteniendo una media de *accuracy* del 70% y un máximo de 73%.

Siguiendo la misma vía que este estudio, así como el realizado por Renato Amorim Torres [8], se ha decantado por utilizar aproximadamente entre 5-7 años para formar nuestro *dataset*. Además, a diferencia de dichos estudios, los cuales solo utilizaron o las estadísticas de cada uno de los equipos o el historial de partidos entre dichos equipos, en este proyecto se incluirá también los datos estadísticos de cada uno de los jugadores, con el objetivo de analizar si visualizamos una mejora o un decremento en el porcentaje de acierto.

Finalmente, Jorge Morate Vázquez, desarrolló un extenso proyecto en el cual consiguió la cifra más alta ahora, 74% [9]. Aunque tan solo utilizó como *dataset* una temporada, su método más efectivo, *Random Forest*, será utilizado en este proyecto para comprobar si su metodología funciona de la misma forma al augmentar el conjunto de datos.

Objetivos

La realización de este proyecto persigue múltiples objetivos generales. Para cada uno de ellos, se le asignará una prioridad, y en los casos en los que exista alguna posibilidad de fracaso, se le asignará un plan de contingencia alternativo. Por ello los puntos a realizar en este trabajo son:

- Analizar la viabilidad de los métodos actuales de aprendizaje computacional para la predicción de eventos deportivos.
- Generar un *dataset* propio con todas las estadísticas existentes sobre la NBA, diferenciado en tres características básicas:
    - Equipo
    - Jugadores
    - Partidos
        - PLAN DE CONTINGENCIA: En el caso en el que algún aspecto no pueda ser recopilado de forma correcta, se valorará el utilizar otras librerías como *nba_py*, o incluso generar un programa propio que pueda proporcionar esos datos.
- Diseñar y desarrollar un modelo capaz de predecir el equipo ganador, los puntos anotados en un partido y el margen de victoria
    - PLAN DE CONTINGENCIA: En el caso en el que la generación de algún modelo se complique y se dedique más tiempo de lo planificado y eso haga retrasar la realización de las siguientes tareas, se eliminará dicho modelo de la planificación y a cambio se aumentará el conjunto de datos de nuestro *dataset* teniendo en cuenta la opinión de un estadístico estadounidense, que realiza análisis estadísticos con datos en crudo [10].
- Determinar los aspectos estadísticos que más influyen en los resultados de un partido
- PLAN DE CONTINGENCIA: Por último, en el caso en el que las tareas se finalicen antes del tiempo establecido, se contemplaría la posibilidad de aumentar el *dataset* con la opinión del estadístico anteriormente comentado o por las posiciones de tiro de cada uno de los equipos.

Además, la elaboración de este proyecto también busca un seguido de objetivos específicos, tanto personales como técnicos. En relación a los de carácter personal, podemos observar los siguientes:

- Adquirir habilidades propias de gestión de proyecto, tales como la planificación, la priorización o la documentación.
- Detectar si la Inteligencia Artificial es el ámbito en el que realmente me quiero especializar.
- Poner a prueba mi autogestión.

Y, finalmente, los objetivos específicos de carácter técnico se pueden ver reflejados en los siguientes puntos:

- Aumentar mis conocimientos actuales sobre aprendizaje computacional.
- Conocer el funcionamiento y la usabilidad de una API para la generación de *datasets*.
- Aprender cuales son los métodos más utilizados para el análisis de resultados.

Metodología

Durante el desarrollo del proyecto utilizaremos un conjunto de herramientas, accesibles para todo el mundo, que nos permitirán a cada etapa finalizar con éxito nuestros objetivos. Éstas pueden verse reflejadas en la siguiente tabla, donde se define, además, la versión en uso y el motivo de su existencia en este proyecto.

Herramienta	Motivo
R	Utilizaremos dicho lenguaje para la generación del *dataset* mediante la librería *NBAStatR*.
Python	Utilizaremos dicho lenguaje para la creación del modelo computacional que predecirá los resultados en diferentes eventos.
RStudio	Este IDE será el utilizado para trabajar el lenguaje R
Anaconda	Este IDE será el utilizado para trabajar el lenguaje Python
Microsoft Excel	Este programa será utilizado para almacenar los datos en un formato en específico.

Una vez establecidas las herramientas que formaran nuestro marco de trabajo, es necesario especificar cuál será la metodología que se utilizará durante el proyecto y que servirá como vía de organización.

Dada la situación y la necesidad del proyecto, se ha definido que nuestra metodología debe de cumplir un requisito esencial para lograr un producto de calidad, de forma eficiente y gestionando los cambios de forma sencilla: la estrategia utilizada debe de ser ágil [11].

Bajo dicha premisa, observamos que hoy en día conviven muchas metodologías ágiles, cada una con sus respectivas características. La dificultad, pero, no recae en definir qué estrategia es mejor, si no detectar en que situaciones es más efectiva utilizar una u otra. En este sentido, aunque la reina en este ámbito no deja de ser SCRUM, finalmente se ha decantado por utilizar la estrategia Kanban [12] adaptada ligeramente para este proyecto.

Kanban fue creada por Toyota con el objetivo de controlar el avance del trabajo en una línea de producción, aunque en los últimos años se ha utilizado en la gestión de proyectos de desarrollo software. Las principales reglas de esta estrategia son:

1) **Visualizar el trabajo y las fases del ciclo de producción**

Al igual que SCRUM, Kanban se basa en el desarrollo incremental dividiendo el trabajo en partes. Éstas se pueden observar de forma visual en una pizarra, conociendo así el estado de cada tarea en todo momento.

2) **Determinar el límite de tareas en curso**

Quizás una de las características principales de Kanban es el hecho de limitar el número de tareas que se pueden realizar en paralelo. Este aspecto viene dado por el hecho de que esta estrategia busca generar resultados de forma incremental, y por lo tanto, su intención es finalizar tareas dando más valor al producto antes de iniciar unas de nuevas.

3) **Medir el tiempo en completar una tarea**

Todo y los puntos anteriormente especificados, la necesidad de este proyecto de generar unos resultados de forma continuada, propicia que se deba personalizar esta metodología añadiéndole, además, el trabajo iterativo propio de una estrategia SCRUM.

Cabe destacar, que el hecho de añadir dicha característica no implica que nos encontremos bajo una estrategia SCRUMBAN, dado que aspectos tan importantes como *Daily* o *Sprint Planning* no existirán durante el desarrollo de este proyecto.

Planificación Inicial

Iteración	Fecha	Tarea	Resultados
1	04/03 – 10/03	Redactar Informe Inicial	Informe Inicial
INFORME INICIAL			
2	11/03 – 17/03	Generar *dataset* de partidos junto con las cuotas y limpieza básica de datos	Fichero con datos de partidos y sus cuotas
3	18/03 – 24/03	Determinar atributos significativos del *dataset* partidos	Fichero con atributos de partidos definitivos
4	25/03 – 31/03	Aplicar Regresión Lineal al *dataset* partidos para predecir equipo ganador	Código de una Regresión Lineal y resultados del experimento
5	01/04 – 07/04	Generar *dataset* de equipos y jugadores y limpieza básica de datos	Fichero con datos de equipos y jugadores
6	08/04 – 14/04	Determinar atributos significativos del *dataset* equipos y jugadores	Fichero de equipos y jugadores con atributos definitivos
INFORME SEGUIMIENTO I			
7	15/04 – 21/04	Aplicar Regresión Lineal para predecir probabilidad de ganador	Análisis de los resultados del experimento
8	22/04 – 28/04	Aplicar Regresión Lineal para predecir anotación de cada equipo	Análisis de los resultados del experimento
9	29/04 – 05/05	Aplicar Regresión Logística para predecir probabilidad de ganador	Análisis de los resultados del experimento

10	06/05 – 12/05	Aplicar Regresión Logística para predecir anotación de cada equipo	Análisis de los resultados del experimento
11	13/05 – 19/05	Aplicar Random Forest para predecir ganador del partido	Análisis de los resultados del experimento
12	20/05 – 26/05	Aplicar Naive Bayes para predecir ganador del partido	Análisis de los resultados del experimento
Informe Seguimiento II			
13	27/05 – 02/06	Comparación de resultados	Análisis de los resultados
14	03/06 – 09/06	Añadir el concepto de valor	Algoritmo que determina si la predicción tiene valor
15	10/06 – 16/06	Cálculo de beneficios netos	Análisis de beneficios
PROPUESTA INFORME FINAL			
16	17/06 – 23/06	Optimización del código	Código más escalable
17	24/06 – 30/06	Acabar Dossier Propuesta de Presentación	Dossier finalizado Presentación TFG
DOSSIER			

Planificación actual

Iteración	Fecha	Tarea	Resultados
Estado actual	14/04	-	Dataset de partidos, equipos y jugadores generado. Código de regresión lineal generado.
INFORME SEGUIMIENTO I			
7	15/04 – 21/04	Aplicar Regresión Lineal para predecir probabilidad de ganador	Análisis de los resultados del experimento
8	22/04 – 28/04	Aplicar Regresión Lineal para predecir anotación de cada equipo	Análisis de los resultados del experimento
9	29/04 – 05/05	Aplicar Regresión Logística para predecir probabilidad de ganador	Análisis de los resultados del experimento
10	06/05 – 12/05	Aplicar Regresión Logística para predecir anotación de cada equipo	Análisis de los resultados del experimento
11	13/05 – 19/05	Aplicar Random Forest para predecir ganador del partido	Análisis de los resultados del experimento
12	20/05 – 26/05	Aplicar Naive Bayes para predecir ganador del partido	Análisis de los resultados del experimento
Informe Seguimiento II			
13	27/05 – 02/06	Comparación de resultados	Análisis de los resultados

14	03/06 – 09/06	Añadir el concepto de valor	Algoritmo que determina si la predicción tiene valor
15	10/06 – 16/06	Generación datasets con cuotas	Dataset completo
PROPUESTA INFORME FINAL			
16	17/06 – 23/06	Cálculo de beneficios netos	Análisis de beneficios
17	24/06 – 30/06	Acabar Dossier Propuesta de Presentación	Dossier finalizado Presentación TFG
DOSSIER			

Análisis de seguimiento I

Como hemos podido ver en los dos puntos anteriores de este mismo informe, la planificación presentada inicialmente variará ligeramente después de estas primeras iteraciones. Esto es debido a los diversos imprevistos que han aparecido durante el desarrollo del proyecto, provocados en parte, por la falta de experiencia a la hora de predecir el tiempo de desarrollo de cada tarea, así como de problemas en el entorno de configuración de alguna de las herramientas utilizadas. La explicación más explícita de los cambios se puede resumir en los siguientes puntos:

Los problemas/imprevistos encontrados:

- A la hora de configurar el entorno de trabajo en el IDE seleccionado (*PyCharm*) se encontraron muchos problemas debido a las versiones previamente instaladas en el ordenador. Tal es así, que fue imposible poder instalar las librerías necesarias para la elaboración de este proyecto, provocando así el hecho de cambiar de entorno de trabajo en Python de forma inmediata.
- Pese al uso de librerías externas que permitían extraer los datos del *dataset* de una forma algo mas sencilla, el estado de estos provocó que el tratamiento posterior fuera más extenso, aumentando el tiempo invertido en esa tarea hasta el doble de lo esperado.

- El hecho de definir la estructura que seguiría nuestros datos también generó ligeramente un retraso más a nuestra predicción, puesto que la falta de experiencia en el uso de algunas herramientas de *Machine Learning*, provocaron dudas en cual sería el formato óptimo para almacenar los datos.

Una vez analizado el estado y los motivos por los cuales no hemos sido capaces de seguir estrictamente la planificación inicial, decidimos realizar un seguido de cambios para ajustar nuestro trabajo futuro dado nuestro estado actual y así completar de forma eficiente nuestros objetivos iniciales. Para ello, decidimos realizar los siguientes cambios:

1. Las cuotas de los partidos, así como el cálculo del beneficio que se generaría y el cálculo del valor, se encontrarán directamente al final del proyecto, puesto que son los apartados que más difieren a nuestros objetivos y, por lo tanto, los que menos prioridad tienen.
   Tanto es así que, si en estas próximas iteraciones se genera más retraso de lo esperado, serán estas tareas las que desaparecerán de nuestro marco de trabajo.
2. Dado el punto anterior, la generación de los modelos utilizaran los datos estrictamente deportivos extraídos a partir de la librería *NBAStatR*, dejando así de lado toda aquella información relacionada con apuestas deportivas.

Bibliografía

[1] Wikipedia, «Wikipedia,» 07 03 2019. [En línea]. Available: https://es.wikipedia.org/wiki/National_Basketball_Association. [Último acceso: 09 03 2019].

[2] «howmuch,» 1 Julio 2016. [En línea]. Available: https://howmuch.net/articles/sports-leagues-by-revenue. [Último acceso: 08 03 2019].

[3] Bulls, «reddit,» 2015. [En línea]. Available: https://www.reddit.com/r/nba/comments/1nq0r8/heres_a_map_i_made_of_all_nba_teams_organised_by/. [Último acceso: 04 03 2019].

[4] velvetbrain, «velvetbrain,» [En línea]. Available: http://www.velvetbrain.net/nba/nba_3pt_trends/. [Último acceso: 05 03 2019].

[5] T. Economist, «YouTube,» 04 12 2018. [En línea]. Available: https://www.youtube.com/watch?v=oUvvfHkXyOA. [Último acceso: 07 03 2019].

[6] R. Maheswaran, «YouTube,» 6 07 2015. [En línea]. Available: https://www.youtube.com/watch?v=66ko_cWSHBU. [Último acceso: 08 03 2019].

[7] H. W. M. P. Matthew Beckler, «NBA Oracle,» Matthew Beckler, Pittsburgh, 2009.

[8] R. A. Torres, «Prediction of NBA games based on Machine Learning Methods,» Renato Amorim Torres, Wisconsin, 2013.

[9] J. M. Vázquez, «Predicción de Equipo Ganador en el,» Jorge Morate Vázquez, Madrid, 2016.

[10] «masseyratings,» [En línea]. Available: https://www.masseyratings.com/nba/games. [Último acceso: 10 03 2019].

[11] «obs-edu,» [En línea]. Available: https://www.obs-edu.com/es/blog-project-management/metodologias-agiles/5-motivos-por-los-que-implementar-una-metodologia-de-desarrollo-agil. [Último acceso: 05 03 2019].

[12] J. Garzas, «javiergarzas,» 22 11 2011. [En línea]. Available: https://www.javiergarzas.com/2011/11/kanban.html. [Último acceso: 06 03 2019].