Machine Learning para la predicción de eventos en la NBA

Albert Villar Ortiz

Resumen— En este proyecto se ha analizado la viabilidad actual de algunos de los métodos más utilizados en el aprendizaje automatizado, sobre uno de los ámbitos más importantes en nuestro día a día: el deporte. Entrando en detalle, tras haber centralizado toda la información y/o estadística avanzada relacionada con la liga americana de baloncesto por excelencia, la NBA, se han generado un conjunto de modelos matemáticos con los que poder predecir el equipo ganador de cada evento, así como la anotación de cada uno de los dos equipos enfrentados. Entre estos, podemos encontrar: la regresión lineal y logística, el algoritmo de Random Forest (tanto clasificador como regresor) y, finalmente, la máquina de vectores de soporte (SVM).

Palabras clave— Aprendizaje automatizado, Baloncesto, NBA, Predecir, Ganador del partido, Anotación por equipo.

Abstract— This project has analyzed the current viability of some of the most used methods in machine learning, in one of the most important areas nowadays: sports. Going into detail, after having gathered all the information and / or advanced statistics related to the Americanbasketball league by excellence, the NBA, a set of mathematical models has been generated to predict the winning team of every event, as well as the score of each team. Among these, we can find: the linear and logistic regression, the Random Forest algorithm (both classifier and regressor) and, finally, the support vector machine (SVM).

Index Terms— Machine Learning, Basketball, NBA, Predict, Winner of the match, Points by team

1 Introducción

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 de 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. Es por ello, que este proyecto pretende aplicar dicha tecnología sobre una liga: la NBA.

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 anualmente en los últimos datos que se tiene registro (2017-2018), 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. A posteriori, 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

Llegados a este punto, este proyecto busca unir estos dos mundos elaborando un *dataset* (teniendo en cuenta tan solo los datos de temporada regular) propio para predecir tanto el ganador de un partido como la supuesta anotación por parte de cada uno de los equipos. Para lograr dicho reto, se utilizarán diversos algoritmos de *Machine Learning*, los cuales serán analizados por separado, para finalmente determinar cuál funciona mejor en esta casuística.

2 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 los 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 cuatro épocas distintas

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 [3].
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, e incluso la probabilidad de que el balón vaya a un lado u otro del campo al fallar, mediante lo que Rajiv Maheswaran (director de Second Spectrum) llama: Ingeniería de los Puntos [4].

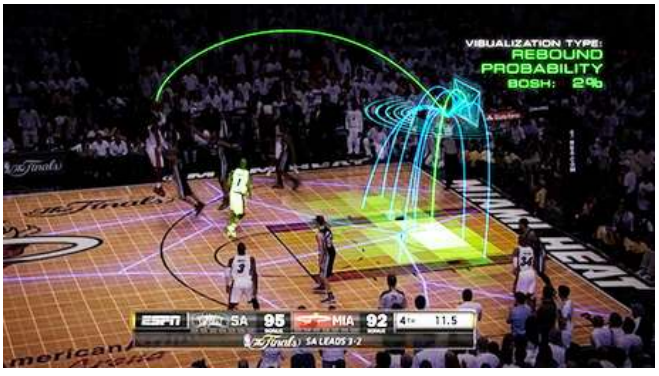


Ilustración 3: Cálculo de probabiliades de rebote en directo de Second Spectrum

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 [5], el cuál concluyeron que la tecnología que proporcionaba mejores resultados para la predicción de ganadores en la NBA era la regresión logística, 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 [6], 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% [7]. 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.

3 Objetivos

La realización de este proyecto ha perseguido múltiples objetivos generales en función de que beneficio ofreciese. Algunos de ellos se consideraron objetivos primordiales y por lo tanto necesarios para la correcta evolución del proyecto, por lo que recibieron prioridad máxima durante todo el desarrollo del producto. Pese a su priorización, a aquellos puntos con mayor probabilidad de fracaso se le asignaron planes de contingencia, con la idea de asegurarnos de que el proyecto pese a los diferentes imprevistos que pudieran surgir cumpliera con su cometido.

Por ello los objetivos de 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, diferenciando entre tres características básicas:
  ○ Equipos
  ○ Jugadores
  ○ Partidos
• Determinar los aspectos estadísticos que más influyen en los resultados de un partido
• Diseñar y desarrollar un modelo capaz de predecir el equipo ganador, los puntos anotados en un partido y el margen de victoria.

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

Y, finalmente, los objetivos de carácter personal se pueden ver reflejados en los siguientes apartados:
• 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.

4 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 se pueden ver reflejadas en los siguientes puntos:
• R
• Python
• Anaconda Navigator
• RStudio
• Microsoft Excel

Dada la situación y la necesidad del proyecto, se ha definido que nuestra metodología debe de ser ágil [9].
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 [10] 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:

• 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.

Ilustración 4: Representación de la pizarra Kanban del proyecto

• 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.

• Medir el tiempo en completar una tarea
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.

5 Desarrollo

Primeramente, para este proyecto se ha generado un dataset propio donde se ha recopilado diferente información relacionada con la NBA. Como objetivo, este proyecto pretende encontrar información en relación con tres cam-

pos: equipos, jugadores y partidos. Para realizar este apartado, se ha utilizado una librería externa programada en R (NBAStatR), que nos ha permitido extraer ligeramente la problemática de obtener información de diferentes fuentes a partir de un mismo punto. El conjunto de funciones y posibles datos que podemos obtener mediante esta librería pueden ser visualizados a partir de su propia documentación [10].

El uso de contenido de terceros ha provocado que en buena parte del proyecto se haya dedicado tiempo para estructurar de forma correcta los datos con el fin de obtener representada la información tal y como la necesitamos de cara a la ejecución de los modelos. Tanto es así, que al princpio para cada partido existían tantas entradas como jugadores habían participado en él, y provocaba que la dimensión de información por cada partido no fuera la misma, generando así una problemática a la hora de aplicarlo a cualquier algoritmo matemático. Por ello, se reestructuraron los datos representando cada partido en una única fila y eliminando aquella información que no se consideraba propia del dataset, o fuera del alcance de este proyecto (estadísticas individuales por jugador, múltiples formas de identificar un mismo partido, información personal de cada jugador…).
Además, cabe destacar, que también se procedió a recopilar las cuotas que habían asignado las casas de apuestas a cada evento, sin utilizar ninguna librería externa, es decir, mediante la extracción directa de páginas web utilizando técnicas de Web Scrapping. En este caso, se decidió utilizar el portal de apuestas por excelencia como es oddsportal. Dicho proceso de obtención de datos se realizó con éxito, pero finalmente, al analizar los datos y intentar fusionarlos con los otros tipos de datasets creados se llegó a la conclusión que los datos no coincidían y que, por lo tanto, no había una manera correcta de unir la información y por lo tanto utilizar las probabilidades que nos aportan las bookies para nuestras predicciones.
Esta problemática también ha provocado que uno de los objetivos que perseguía este proyecto, como es el hecho de poder cuantificar los beneficios o pérdidas que generábamos a partir de los diferentes modelos del proyecto, no haya podido ser acabado con éxito. De esta manera conoceremos cual es el porcentaje de acierto, pero sin llegar a saber si ese porcentaje es suficiente para poder generar beneficios en alguna casa de apuestas

Tras la obtención de los datos, se procedió al proceso de selección de características, con el objetivo de detectar cuales eran los atributos que mas influenciaban el resultado de nuestras predicciones. Dicho apartado es uno de los más importantes pues depende del algoritmo que utilicemos podremos obtener características más o menos fiables a la hora de hacer las predicciones. Pese a ello, dado que el conjunto de datos no estaba creado, y los objetivos de este proyecto no pretendía generar los mejores resultados si no realizar un análisis comparativo, se decidió utilizar dos algoritmos sencillos y muy utilizados comúnmente como son: univariate selection y feature importance.

El primer método es uno de los mas simples a aplicar, puesto que el algoritmo univariate selection funciona en base a pruebas estadísticas como el chi-cuadrado, seleccionando las características que tienen la relación más fuerte con una variable predicha. Desafortunadamente, la prueba solo es fiable si las variables son completamente independientes. Todo y con eso, dado su simplitud se ha decidido incorporar en este proyecto.
La prueba del chi-cuadrado lo que busca es verificar si las frecuencias observadas en cada categoría son compatibles con la independencia entre ambas variables. Para evaluarla, se calculan los valores que indicarían la independencia absoluta, lo que se denomina frecuencias esperadas, comparándolos con las frecuencias de la muestra. Podemos observar pues dicho concepto en la siguiente fórmula donde fo es la frecuencia observada mientras que la variable fe representa la frecuencia esperada.

$$\alpha^2 calc = \sum \frac{(fo - fe)^2}{fe}$$

El segundo método utilizado es el feature importance. En este algoritmo se identificará una característica como importante si al mezclar sus valores aumenta el error del modelo, por que querrá decir que el modelo se basó en la característica para la predicción. Mientras tanto, el algoritmo identificará una característica como no importante si al barajar sus valores el error del modelo se mantiene sin cambios, lo que querrá decir que en este caso el modelo ignoró la característica a la hora de realizar la predicción [11].
Finalmente, contamos con el tercer método de selección de características, el más simple de todos. Para aplicarlo tan solamente tenemos que utilizar la totalidad de los atributos de cada uno de los datasets. Con toda esa información, veremos si el modelo es capaz de discernir entre la información importante o no.
Se espera que el hecho de utilizar la totalidad del conjunto de datos no provoque que los modelos reciban demasiado ruido que imposibilite el hecho de que puedan realizar la predicción de forma correcta.

Una vez seleccionados los atributos de cada conjunto de datos desarrollamos los modelos matemáticos que realizarían las predicciones tanto del ganador del partido como la anotación de cada uno de los equipos. El primer modelo que desarrollamos fue el regresor lineal con el objetivo de poder predecir la anotación de cada equipo que disputa un partido. Como en nuestro proyecto vamos a trabajar con múltiples atributos sobre cada predicción, el modelo matemático acabará siendo un regresor lineal con múltiples parámetros, siendo la fórmula base la siguiente:

Lo mismo sucede con el regresor que nos va a permitir clasificar quien será el ganador en cada partido (regresor logístico), el cuál también utilizará múltiples parámetros en cada predicción.

Además de los dos tipos de regresores especificados anteriormente, también se ha seguido el modelo de Random Forest, el cuál basa sus predicciones en árboles de decisiones. Éste no deja de ser un mapa de los posibles resultados de una serie de decisiones relacionadas. Por lo general, comienzan con un único nodo y luego se ramifica en resultados posibles.

En este modelo, como en el regresor, también contaremos con dos variantes (en función de si la salida es un valor numérico o una clasificación).

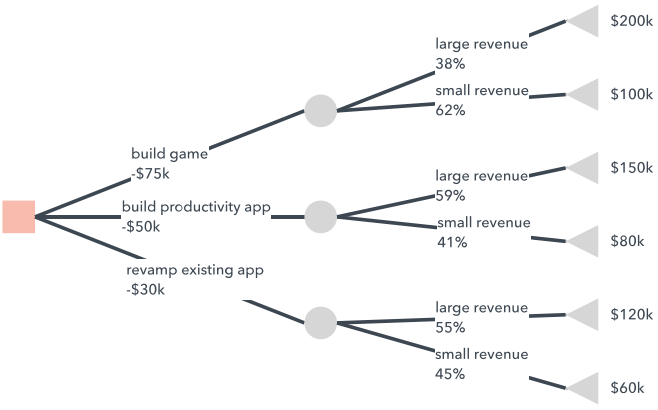


Ilustración 5: Ejemplo de árbol de decision con el que se basa el algoritmo de Random Forest

Por último, encontramos el modelo de Naive Bayes que es el último modelo matemático que se ha utilizado en este proyecto. En Naive Bayes Clasificador podemos ver la distribución P(X | Y) como una descripción de cómo generar instancias aleatorias X condicionadas en el atributo de destino Y. Bajo esta definición se da por supuesto que todos los atributos de X son condicionalmente independientes, dada la Y.

$$p(X|Y) = \frac{p(Y|X) * p(X)}{p(Y)}$$

Dentro de este modelo existen diversos tipos de Naive Bayes. En este caso, se ha utilizado el Naive Bayes Gaussiano, por su facilidad a la hora de implementarlo y también por ser el más común entre todos los modelos de Naive Bayes.

6 Resultados

Tras el desarrollo del proyecto se ha obtenido un conjunto de datos único en relación a la NBA, puesto que se ha estructurado dicha información de cara a que sea aplicada a diferentes modelos matemáticos para su posterior predicción. Dichos datos se pueden ver reflejados en diferentes Excels en función del tipo de dato que tratase. La base de información que se buscaba (equipos, jugadores y partidos) han sido recogidos y tratados con éxito, mientras que el último conjunto de datos (predicciones y probabilidades), han sido obtenidos con éxito, pero no han podido ser unidos junto con la otra información generada debido a demasiadas inconcluencias. Los datos de las cuotas y de los partidos no coincidían con los datos que teníamos des de un principio, más existían bastantes eventos sin cuota ni probabilidad, generando así un dataset incompleto. Es por ello que, pese a que tenemos esa información almacenada, no ha sido fusionado con todos los otros datos del proyecto, y por lo tanto, no han formado parte del experimento.

Tras aplicar los distintos modelos a los diferentes conjuntos de datos desarrollados en este mismo proyecto, se han obtenido varios resultados que se van a poder ver representados de la siguiente manera: por cada atributo a predecir (ganador de cada evento y anotación de cada uno de los equipos) se ha generado una tabla comparativa donde podemos observar que resultados hemos obtenido en las múltiples combinaciones tanto de datos como de modelos que hemos definido en este proyecto. Con esta idea lo que se pretende es poder visualizar de forma clara y concisa todo el conjunto de pruebas que se ha llevado a cabo a la hora de analizar nuestros modelos. Para cada modelo y conjunto de datos podemos ver hasta 3 diferentes features selection siendo la primera el uso del univariate selection, la segunda el modelo de feature importance y por último el uso de todas las características al completo.

Predicción del ganador del partido (Accuracy)						
	Logistic R.		RF Clas.		NB Clas.	
Equipos	FS1	58.55	FS1	58.44	FS1	41.86
	FS2	58.55	FS2	58.44	FS2	41.86
	FS3	58.55	FS3	58.55	FS3	41.86
Jugadores	FS3	**63.19**	FS3	62.77	FS3	61.36
Partidos	FS1	58.55	FS1	52.76	FS1	58.55
	FS2	58.55	FS2	53.44	FS2	58.55
	FS3	58.55	FS3	52.60	FS3	58.55
Total	FS1	51.40	FS1	55.05	FS1	41.60
	FS2	58.55	FS2	54.58	FS2	41.60
	FS3	**63.19**	FS3	54.27	FS3	41.60

Como podemos ver en los resultados especificados en la tabla anterior, las predicciones han salido algo mas bajas de lo esperado tras el análisis realizado en el apartado del estado del arte. Todo y con eso, hemos obtenido un máximo interesante mediante el uso del modelo de regresión logística alcanzando un porcentaje de acierto del 63.19%. Este máximo se ha alcanzado en dos modelos de datos distintos: en el conjunto de datos de jugadores y en el conjunto de datos total.

El hecho de que el conjunto de datos que más influye en la victoria de un equipo sea los jugadores que han participado en cada partido, tiene mucho sentido. En cada equipo hay un seguido de jugadores que forman parte del core de éste, y que por lo tanto el estar o no estar disponible la estrella de un equipo, es vital para el resultado final.

A posteriori, se ha intentando investigar cuales pueden ser los motivos que han provocado que los resultados no se fueran más aproximados a los vistos en otros proyectos, y se han encontrado algunos puntos que podrían ser replanteados de cara a intentar aplicar los modelos de forma óptima.

A la hora de seleccionar las características de cada uno de los datasets, hemos utilizado dos modelos matemáticos sencillos y muy utilizados por la comunidad, pero no se ha realizado un análisis exhaustivo de con que tipo de datos estamos trabajando y, como consecuencia, que algoritmo de feature selection son los más apropiados para este proyecto.
Además de este punto, a la hora de aplicar los diferentes modelos matemáticos se han utilizado unos parámetros generales y comunes sin entrar mucho en detalle. Por ello, un punto que a lo mejor ha provocado estos resultados, es el hecho de no haber optimizado al máximo los modelos intentando encontrar los majores parámetros según el conjunto de datos que se utilizaba.
Por último, el hecho de haber aumentado la cantidad de datos y atributos que utilizamos para la predicción respecto a otros proyectos, así como los años utilizados, pueden haber provocado que hayámos incluido más ruido de lo habitual sobre el modelo y que éste no haya sido capaz de detectar correctamente la relación entre los datos y la salida.
Pese a todo esto, si entramos más en detalle con los resultados, podemos observar como los mejores resultados para cada uno de los modelos ha sido sobre el conjunto de datos de jugadores, y estos difieren ligeramente tal y como podemos comprobar en la siguiente gráfica:

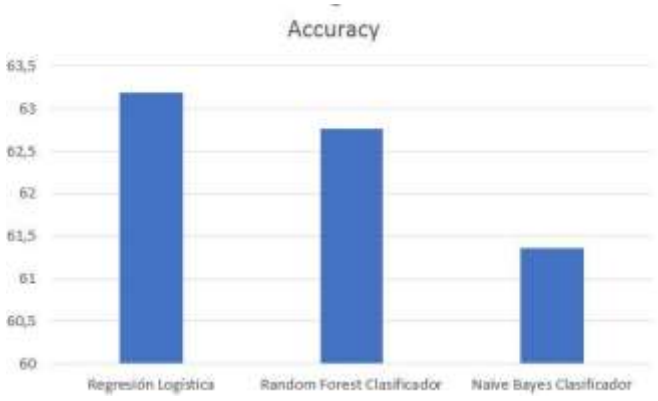


Ilustración 6: Gráfica de los mejores resultados para cada uno de los modelos matemáticos utilizados.

Por último, si comparamos los resultados obtenidos en este proyecto con los obtenidos en otros analizados en estado del arte, podemos ver que este resultado es el peor de todos. Todo y con eso, hay que tener en cuenta que el marco de trabajo ha sido distinto, puesto que los años no son los mismos e incluso en alguno de los proyectos tan solo se ha utilizado una única temporada. Por lo que, para poder realizar una comparación correcta, se tendrían que analizar los modelos generados en, exactamente, las mismas conclusiones. Una comparación básica puede ser observada en la siguiente gráfica:

Ilustración 7: Comparativa de resultados entre modelos matemáticos comentados a lo largo de este proyecto

A continuación, vamos a analizar los resultados obtenidos para la anotación del equipo local. Para medir la calidad de las predicciones en este caso se utilizará el MAE (Mean Absolute Error).

Predicción de la anotación del equipo local (MAE)				
	Linear R.		RF Regr.	
Equipos	FS1	7.50	FS1	7.5
	FS2	7.50	FS2	7.5
	FS3	7.50	FS3	7.5
Jugadores	FS3	9.94	FS3	7.49
Partidos	FS1	**7.34**	FS1	7.49
	FS2	**7.34**	FS2	7.49
	FS3	**7.34**	FS3	7.49
Total	FS1	9.84	FS1	7.49
	FS2	9.84	FS2	7.49
	FS3	9.84	FS3	7.49

En este caso, no contamos con ningún proyecto realizado con anterioridad con el cual poder realizar comparaciones y determinar si los resultados son suficientemente correctos o no. Pese a ello, cabe decir que los resultados obtenidos han sido satisfactorios e incluso mejor de lo que se esperaba al inicio del proyecto. He obtenido el mejor resultado utilizando de nuevo un regresor, esta vez el regresor lineal, obteniendo un error medio absoluto de 7.34 puntos respecto la anotación final del equipo local.
A diferencia de la predicción del ganador, a la hora de predecir la anotación de los equipos los datos que mas influyen no son respecto a los jugadores si no a los partidos que habían jugado con anterioridad.

Pese a los buenos resultados, sería una buena idea aplicar algunos de los cambios especificados anteriormente con la predicción de ganador del partido, para intentar observar si es posible alcanzar un error medio aproximado de 5, dado que según mi punto de vista éste sí seria un resultado más que correcto dado los datos que participan en este proyecto.

Predicción de la anotación del equipo visitante (MAE)				
	Linear R.		RF Regr.	
Equipos	FS1	9.46	FS1	9.63
	FS2	9.46	FS2	9.63
	FS3	9.46	FS3	9.63
Jugadores	FS3	10.22	FS3	**9.44**
Partidos	FS1	9.49	FS1	9.8
	FS2	9.49	FS2	9.8
	FS3	9.49	FS3	9.8
Total	FS1	9..98	FS1	9.8
	FS2	9..98	FS2	9.8
	FS3	9..98	FS3	9.8

Tal y como pasó con los experimentos realizados sobre el equipo local, tampoco contamos con ejemplos de otros proyectos que hayan predicho la anotación del equipo visitante. En este caso, el mínimo error se ha obtenido utilizando el modelo de Random Forest, con un resultado ligeramente mejor que el del regresor lineal. Inicialmente, sorprende que el error sea superior por algo más de 2 puntos respecto al análisis realizado con el equipo local. Podríamos intentar entender ese fenómeno, con el hecho de que generalmente es más sencillo jugar en tu campo que en el campo de tu rival. Por este motivo, los equipos visitantes pueden generar resultados más dispersos, provocando así que el modelo cometa muchos más errores en sus predicciones.

Al igual que sucede con el equipo local, sería interesante de cara a pasos futuros, intentar alcanzar un error medio aproximado a 5. Pero dado que según los experimentos realizados podemos llegar a la conclusión de que el error siempre será superior que el del equipo local, habría que intentar buscar aproximadamente el 7 de MAE.

7 Conclusiones

Generar un conjunto de datos estable y bien estructurado es una de las tareas más complicadas dentro del sector de la inteligencia artificial, y más cuando no hay unas reglas de ordenación ni estructuración preestablecidas. Todo y con eso, los datos que formaban parte del core de nuestro proyecto han sido recopilados con éxito, aunque el apartado de cuotas y probabilidades no hayan podido ser relacionadas por el hecho de obtener la información des de otra fuente distinta.

Además, en este proyecto se han analizado los métodos de Machine Learning más utilizados en este ámbito y se ha podido observar como los resultados variaban en función de los datos utilizados en cada momento. En este proyecto, se ha alcanzado un máximo en la predicción del ganador con el método de Regresión Lineal, obteniendo un accuracy de 63.19 utilizando el conjunto de datos tanto de jugadores como el conjunto de datos total. Por otro lado, la predicción de los puntos por cada equipo ha alcanzado sus mejores resultados con otros modelos diferentes: a manos del Regresor Lineal, se ha obteniendo finalmente un error medio de 7.34 en la predicción de la anotación del equipo local y un error medio de 9.44 para el equipo visitante, pero, esta vez, a manos del Random Forest.

Cabe destacar, que los objetivos tanto técnicos como personales establecidos al inicio del proyecto, han sido completados con éxitos, por lo que se ha concluído que este ámbito tecnológico es en el que quiero especializarme, así como el hecho de ser capaz de organizarme de forma adecuada frente a un proyecto de desarrollo iterativo.

Finalmente, queda como líneas futuras el analizar si los algoritmos de feature selection son los apropiados, identificar si los parámetros de nuestros modelos son los óptimos para nuestro conjunto de datos, hasta poder aproximarnos a los resultados que han obtenido otras personas. Una vez alcanzado este punto, el objetivo sería aumentar el conjunto de datos e intentar especificar cuales son las mejores predicciones (predicciones con probabilidades superiores a un threshold, etc), con el objetivo de poder obtener un mayor porcentaje de acierto.

Agradecimientos

Quiero agradecer a todas y cada una de las personas que día a día han estado apoyándome y brindándome la confianza necesaria para poder afrontar con positividad y entusiasmo todas las situaciones e imprevistos.

Además, también quiero aprovechar para agradecer a todos los profesores y profesionales del sector que me han proporcionado los conocimientos necesarios para poder llevar a cabo proyectos de este calibre, y me han guiado cuidadosamente hasta alcanzar este punto.

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

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

[H. W. M. P. Matthew Beckler, «NBA Oracle,» Matthew

[5] Beckler, Pittsburgh, 2009.

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

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

[8] «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].

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

[10] A. Bresler, «asbcllc.com,» [En línea]. Available: http://asbcllc.com/nbastatR/reference/index.html. [Último acceso: 25 06 2019].

[11] «christophm.github.io,» [En línea]. Available: https://christophm.github.io/interpretable-ml-book/feature-importance.html. [Último acceso: 15 06 2019].

[12] 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].

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

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

APÉNDICE

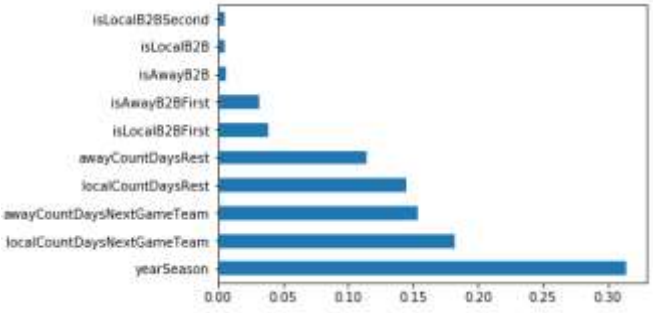
A1. CARACTERÍSTICAS SELECCIONADAS PARA LAS PREDICCIONES

En este apartado vamos a poder ver una parte de las pruebas de selección de características que hemos realizado a lo largo del proyecto. Se puede ver en profundidad en el fichero 'Datasets' del dossier.

Dataset equipo con ganador del partido

La selección de características en este dataset respecto a la salida de ganador ha sido la siguiente para cada algoritmo:

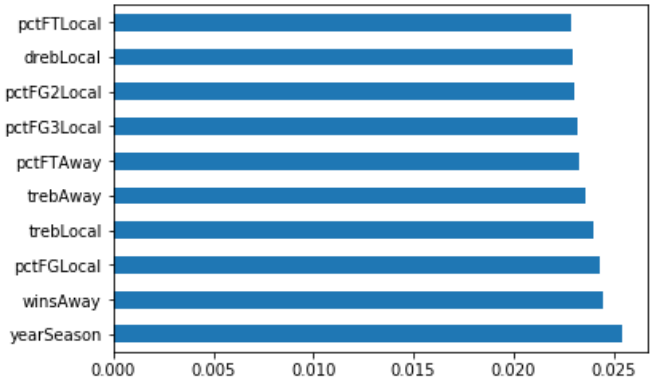
	Specs	Score
11	awayCountDaysRest	115.490781
5	localCountDaysRest	104.858064
12	awayCountDaysNextGameTeam	69.823887
6	localCountDaysNextGameTeam	33.924562
4	isLocalB2BSecond	7.749662
10	isAwayB2BSecond	5.734226
2	isLocalB2B	3.960224
8	isAwayB2B	3.273963
9	isAwayB2BFirst	0.113970
0	yearSeason	0.006382



Dataset partidos con ganador del partido

La selección de características en este dataset respecto a la salida de ganador ha sido la siguiente para cada algoritmo:

	Specs	Score
46	winsAway	223.847759
45	winsLocal	178.581321
28	fg3aAway	16.742112
19	blkLocal	14.172979
22	ptsLocal	13.906216
44	ptsAway	11.874628
37	drebAway	10.975616
27	fg3mAway	10.832497
6	fg3aLocal	9.169565
9	fg2aLocal	8.954497



Dataset total con ganador del partido

La selección de características en este dataset respecto a la salida de ganador ha sido la siguiente para cada algoritmo:

	Specs	Score
91	inactPlayerAway11	3.192117e+07
92	inactPlayerAway12	1.605234e+07
30	actPlayerLocal3	1.293235e+07
47	inactPlayerLocal8	1.253536e+07
29	actPlayerLocal2	1.091816e+07
31	actPlayerLocal4	1.065150e+07
39	actPlayerLocal12	9.554232e+06
46	inactPlayerLocal7	9.296589e+06
44	inactPlayerLocal5	9.032648e+06
45	inactPlayerLocal6	8.974005e+06