Trabajo de Final de Grado

Informe seguimiento II

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:

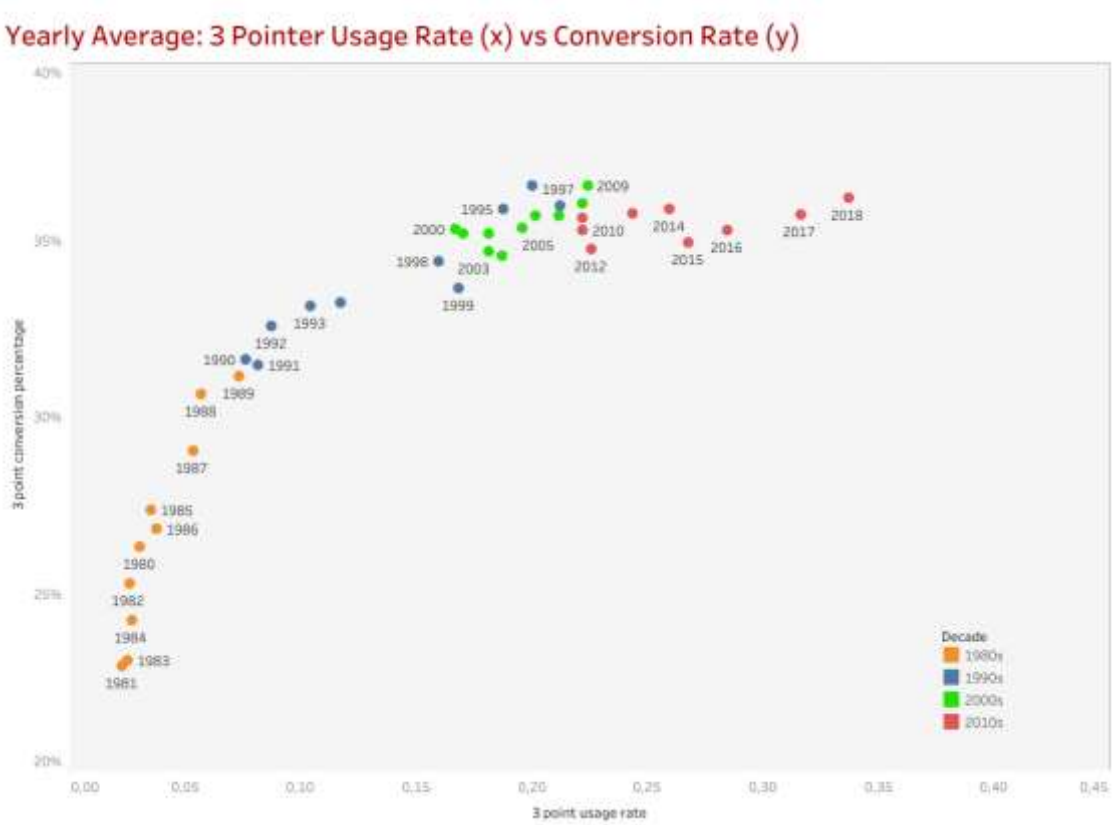


Ilustración 2: Porcentaje de uso del triple y su respectivo acierto en 4 épocas distintas *[4]*

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* de partidos para predecir la anotación.	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 la anotación sobre equipos y jugadores	Análisis de los resultados del experimento
8	22/04 – 28/04	Aplicar Regresión Logística al *dataset* de partidos para predecir equipo ganador	Código de una Regresión Logística y resultados del experimento

9	29/04 – 05/05	Aplicar Regresión Logística para predecir probabilidad de ganador sobre equipos y jugadores	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	Código del Random Forest y resultados del experimento
12	20/05 – 26/05	Aplicar Naive Bayes para predecir ganador del partido	Código del Naive Bayes y 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			

Seguimiento I

El objetivo de este apartado es analizar el trabajo realizado durante las primeras iteraciones, especificando tanto los resultados generados en cada una de las iteraciones como especificando el contenido de estos. A todo esto, aprovecharemos el análisis exhaustivo que realizamos semana tras semana para definir y explicar de forma explicita los problemas encontrados a lo largo del desarrollo del proyecto, las soluciones implementadas para conseguir los objetivos establecidos al inicio de éste y si existe modificaciones en la planificación inicial.

Dicho esto, entramos en materia analizando las diferentes iteraciones. En este caso, el primer módulo de trabajo ha girado en su mayoría, en torno a la generación de los diferentes datasets necesarios para poder realizar las predicciones correctamente. Todo y con eso, también ha dado tiempo a generar las primeras líneas de código referente a los modelos matemáticos que realizaran la predicción a posteriori.

Iteración 2 (11/03 – 17/03)

Estado	Tarea	Resultados
Planificación	Generar *dataset* de partidos junto con las cuotas y limpieza básica de datos	Fichero con datos de partidos y sus cuotas
Realidad	Generar *dataset* de partidos junto con las cuotas y limpieza básica de datos	Fichero con datos de partidos y sus cuotas

En esta iteración procedimos al inicio del primer módulo de trabajo. Esta primera semana, contemplada entre el 11/03/2019 y el 17/03/2019, perseguía generar correctamente el dataset de partidos en su totalidad. Por ello, este dataset debía contener tanto los partidos que se habían realizado año tras año, como la información básica de los enfrentamientos, así como las últimas cuotas aplicadas al posible ganador del evento.

Pese a la estimación de tiempo y carga de trabajo realizada al inicio de la planificación, el tiempo de ejecución para la correcta elaboración de esta iteración ha sido superior. A continuación, se puede observar los motivos por los cuales dicha tarea no ha sido finalizada con éxito.

La obtención de los datos en referente a los enfrentamientos entre dos equipos a partir de la librería *NBAStatR*, seguían el siguiente formato:

- Los datos en relación con un partido estaban mezclados con los datos estadísticos relacionados a un enfrentamiento. Esto según la estructura de los datos implementada al principio del proyecto, provocaba que los datos estadísticos debían de desaparecer de éste dataset y ser tratados de forma individual creando un dataset propio.
- La información en relación con un partido estaba repartida en múltiples filas (no constantes) puesto que los datos venían especificados por jugador. Así que, para un mismo partido, existía N filas del equipo local y otras N filas del equipo visitante. Esto, siguiendo la misma idea que el punto anterior, con la estructura ideada al inicio del proyecto, provocaba que tuviéramos que reestructurar todo el dataset consiguiendo así representar un enfrentamiento en una única fila.

Una vez solucionado todo lo relacionado con la estructura de los datos en los enfrentamientos de la NBA, procedimos a la obtención de las cuotas establecidas por diferentes casas de apuestas, con el fin de poder determinar que probabilidad asignaron a cada evento justo antes de iniciar el partido. Dado que entre las diferentes funcionalidades que ofrecía la librería utilizada para obtener los datos hasta la fecha no incluía ninguna función con la cual poder obtener dichas cuotas por cada enfrentamiento, nos vimos obligados a buscar y analizar diferentes fuentes con el fin de poder decidir cual era la más idónea. Finalmente, se decantó por utilizar uno de los portales más importantes en relación con las cuotas deportivas como es oddsportal.com.

Es aquí, donde mediante el apartado resultados archivados obtendremos las cuotas de cada uno de los enfrentamientos. Para obtener dicha información debíamos generar un código automatizado que recorriera año tras año todas las cuotas y almacenase dicha información para después poder fusionar dichos datos con los datos obtenidos de los enfrentamientos, es decir, debíamos realizar una tarea de WebScrapping. Esta situación provocaba aún más retrasos en la entrega de la iteración 1 por los siguientes motivos:

- La poca experiencia que teníamos para obtener datos mediante WebScrapping provocaba que el tiempo necesario para poder generar el código que recopilara los datos fuera mucho mayor de lo esperado.
- Con estos datos en relación con las cuotas de cada evento deportivo, no teníamos manera de poder relacionar unas cuotas y un partido con cada uno de nuestros enfrentamientos, pues trabajaban con id’s diferentes y no permitía hacer una relación directa enfrentamiento – cuotas. Por ello, decidimos añadir a nuestro código un apartado en el que, por cada conjunto de cuotas en relación con un evento, buscábamos mediante el día, la hora y los equipos enfrentados en la base de datos de enfrentamientos para así conseguir asociar los eventos correctamente y poder añadirle la información de las cuotas de forma correcta.

Finalmente, después de este análisis podemos determinar que la tarea planificada para la iteración 2 (primera iteración del módulo de seguimiento 1), ha sido finalizada parcialmente:

- El dataset que contempla toda la información en relación con un enfrentamiento (información básica, es decir, sin datos estadísticos) ha sido generado con éxito y puede verse la versión final en el siguiente directorio del GitHub:
    - 'Codigo/python/dataFinal/games/games.xlsx'

  Además, dichos datos han pasado un pequeño filtro de limpieza, reescribiendo valores nulos para evitar posteriores problemas en la predicción con el modelo de Machine Learning, modificando la información representada como *string* por numerales con el mismo objetivo que el punto anterior, etc.
- Los datos sobre las cuotas ofrecidas por las casas de apuestas en cada evento deportivo no han sido recopilados con éxito a tiempo. Por lo tanto, esta tarea no ha sido finalizada.

Por lo que respecta a la planificación el retraso existente en esta iteración ha provocado que los objetivos establecidos inicialmente en la iteración 13 hayan sido modificados. Dicha semana se aprovechará para acabar de obtener todos los datos en relación con las cuotas deportivas para a posterior poder utilizar esos datos para la predicción, así como conocer cuanto beneficio podríamos extraer a partir de nuestro modelo.

Iteración 3 (18/03 – 24/03)

Estado	Tarea	Resultados
Planificación	Determinar atributos significativos del *dataset* partidos	Fichero con atributos de partidos definitivos
Realidad	Determinar atributos significativos del *dataset* partidos	Fichero con atributos de partidos definitivos

En esta iteración se proponía seguir trabajando sobre los datos de enfrentamientos. Ésta segunda semana, contemplada entre el 18/03 y el 24/03, buscaba dejar definida la estructura final de los datos sobre los enfrentamientos, intentando definir cuales eran los atributos más importantes con los cuáles realizar las posteriores predicciones.

En este caso, la estimación de tiempo necesaria para poder realizar esta tarea sea adecuó correctamente a la realidad y, por lo tanto, se puede entender esta tarea finalizada con éxito. A continuación, vamos a reflejar el trabajo realizado para conseguir este objetivo y las pequeñas decisiones tomadas al respecto.

Para poder definir que atributos eran los más importantes y por lo tanto los que nos iban a ofrecer mejores resultados en nuestras predicciones, hemos analizado qué diferentes métodos existen actualmente para poder determinar una metodología constante y aplicarla en el dataset de enfrentamientos. Tras analizar las diferentes vertientes hemos decidido utilizar dos de los métodos más utilizados por la comunidad:

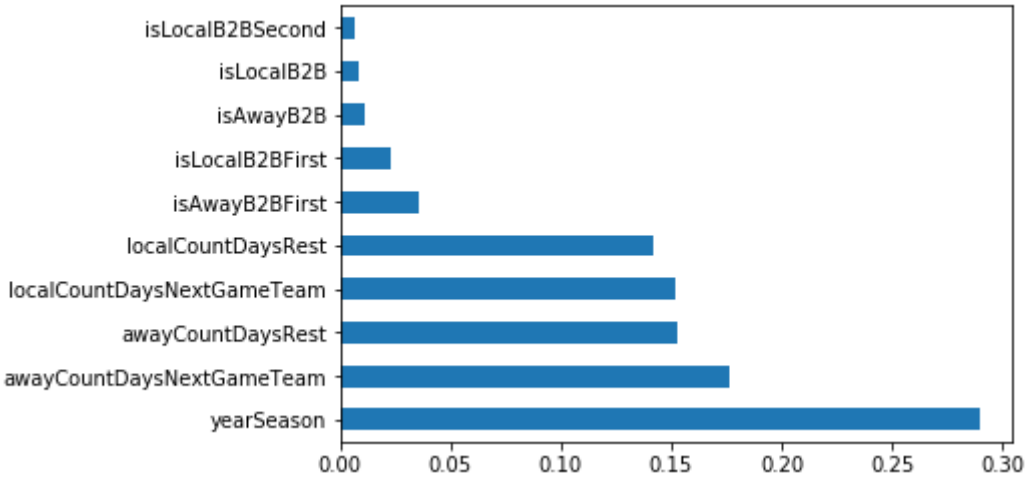
- Univariate Selection
- Feature Importance

De entre todas las opciones que tenemos a nuestro alcance las dos especificadas anteriormente parecen ser las más utilizadas comúnmente a la hora de seleccionar las características más importantes de un dataset [13].

Para ello, ejecutaremos dichos métodos para cada uno de nuestros datasets y de esta manera también podremos comparar los resultados obtenidos utilizando diferentes algoritmos de selección de atributos.

Tras ejecutar las diferentes metodologías hemos obtenido unos resultados que se pueden ver reflejados en el documento Datasets del repositorio de GitHub en la carpeta de Documentación. En éste, por cada atributo, podemos conocer su descripción y sus resultados en los diferentes métodos ejecutados.

Es por ello por lo que en las posteriores predicciones que realicemos utilizaremos esta información para determinar que atributos entraran en la ejecución del modelo.



Como podemos observar en el anterior ejemplo gráfico, las características que tienen más influencia en el resultado del partido (ganador simple) son los días de descansos de ambos equipos, así como los días restantes hasta el siguiente enfrentamiento. Todo y con eso, parece impactante que el dato que esté mas relacionado con la salida sea la temporada en la que se ha jugado el enfrentamiento.

Si nos paramos a analizar dichos resultados podemos llegar a una conclusión razonable. La NBA esta configurada con el objetivo de buscar la igualdad y equilibrar la competición año tras año. Por ese motivo, un equipo que se encuentre entre el top 5 equipos de la liga puede caer hasta formar parte de los 5 peores equipos de liga por diversos motivos el año siguiente (pérdida de su jugador estrella, falta de espacio salarial para completar el equipo provocando que éste baje su rendimiento…).

Estos son los resultados obtenidos a partir del método de Feature Importance, que son muy parecidos a los obtenidos mediante la otra metodología (univariate selection). Para visualizar todos los resultados, acceder al documento Datasets del repositorio de GitHub en la carpeta de Documentación.

Iteración 4 (25/03 – 31/03)

Estado	Tarea	Resultados
Planificación	Aplicar Regresión Lineal al *dataset* partidos para predecir la anotación de cada equipo	Código de una Regresión Lineal y resultados del experimento
Realidad	Aplicar Regresión Lineal al *dataset* partidos para predecir la anotación de cada equipo	Código de una Regresión Lineal y resultados del experimento

En esta iteración dejamos de lado todo lo relacionado con el preprocesamiento de datos. Esta tercera semana contemplada entre el 25/03 y el 31/03, persigue generar el primer modelo de predicción que utilizaremos en nuestro proyecto: la regresión lineal.

Tras analizar la mejor forma con la que representar una regresión lineal y visualizar diferentes ejemplos existentes en Internet, hemos sido capaces de generar el código que entrenará y predecirá los resultados de cada uno de nuestros datasets. Para ello, nos hemos guiado de algunos ejemplos y tutoriales que podemos encontrar en el portal towardsdatascience.com o incluso en la propia librería de sklearn [14].

Pero este apartado era tan solo era una de las dos secciones que se estimaron hacer para esta iteración número 4. Además, se esperaba poder aplicar dicho modelo a los datos obtenidos semanas antes con el objetivo de poder obtener los primeros resultados prediciendo la anotación de cada uno de los equipos. Para ello, accedimos a los datos almacenados en la primera semana del módulo de seguimiento 1 y eligiendo los atributos más importantes gracias al análisis realizado la semana anterior ejecutamos nuestro modelo y observamos los resultados obtenidos:

Localidad	MAE
Local	5.78
Visitante	9.61

Como podemos observar, los resultados obtenidos con la regresión lineal sobre estos primeros datos parecen ser bastante decentes. El valor que nos proporciona el uso del algoritmo MAE (Mean Absolute Error) nos dice la diferencia media entre el valor real y el valor predicho por el algoritmo. Por lo tanto, podemos entender que la predicción de anotaciones de equipos locales difiere aproximadamente de 6 puntos respecto a la realidad, y como visitante obtenemos unos peores resultados al estar entorno a los 10 puntos de diferencia.

Actualmente no conocemos otro proyecto que haya desarrollado la predicción de anotación por equipo y por lo tanto no podemos comprobar si estos resultados están por encima del estado del arte o no. Por ello, esperaremos a obtener los resultados con todos los otros datasets, y realizaremos la comparación en este mismo proyecto.

Iteración 5 (01/04 – 07/04)

Estado	Tarea	Resultados
Planificación	Generar *dataset* de equipos y jugadores y limpieza básica de datos	Fichero con datos de equipos y jugadores
Realidad	Generar *dataset* de equipos y jugadores y limpieza básica de datos	Fichero con datos de equipos y jugadores

En esta iteración nos volvemos a ver las caras con la obtención y el preprocesamiento de los datos. Esta cuarta semana contemplada entre el 01/04 y el 07/04, perseguía acabar de generar todos los datasets previstos des del inicio del proyecto. Para ello, debíamos conseguir recopilar toda la información referente a los equipos en cada uno de los enfrentamientos, así como todos aquellos datos relevantes sobre los jugadores que pueden existir en un enfrentamiento.

Si hacemos referencia al dataset relacionado con los datos esperados sobre los jugadores, podemos determinar que esta parte de la tarea ha sido completada con éxito. Se ha conseguido recopilar toda la información esperada menos el quinteto inicial de cada equipo, puesto que esta información no se ve mostrada con las funcionalidades de la librería utilizada hasta ahora. Todo y con eso, se considera que esa información puede ser secundaría y por lo tanto podemos dar por correcto el dataset generado con los datos de los jugadores. De nuevo la estructura y los datos obtenidos pueden ser visualizados en el fichero Datasets en la carpeta de documentación del repositorio de GitHub.

En cambio, si hacemos foco en el dataset que esperaba recopilar todos aquellos datos de los equipos respecto a sus enfrentamientos previos, podemos determinar que la tarea no ha sido completada con éxito. Los motivos del retraso de esta tarea se pueden resumir en los siguientes puntos:

- Para poder crear correctamente el dataset de equipos en relación con los enfrentamientos, se debe de tener en cuenta para cada partido los resultados obtenidos con anterioridad para poder realizar una media y tener las estadísticas y resultados previos. Esto ha sido más complicado de lo esperado y por lo tanto ha provocado que la tarea no pudiese ser finalizada en el tiempo estimado.
- Esta semana han aparecido diversas situaciones personales inesperadas que han hecho que tenga algo menos de tiempo para la elaboración de esta iteración.

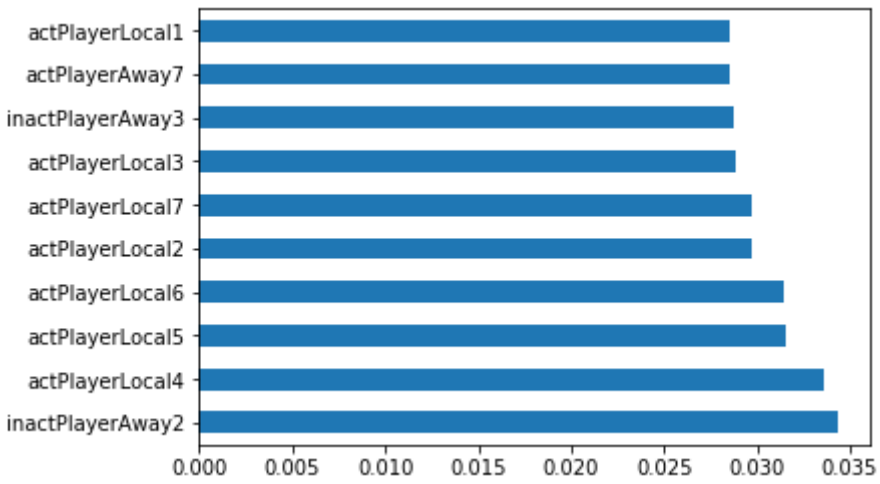
Con lo comentado anteriormente, podemos definir esta tarea como completada parcialmente, y el retraso existente ha provocado un seguido de cambios en la planificación. Estos cambios se pueden observar en la iteración 13, en la cuál además de finalizar la obtención de datos de cuotas deportivas, también se aprovechará para finalizar los datasets de equipo/enfrentamiento, permitiendo así conseguir toda la información que se pretendía analizar en este proyecto.

Iteración 6 (08/04 – 14/4)

Estado	Tarea	Resultados
Planificación	Determinar atributos significativos del *dataset* equipos y jugadores	Fichero de equipos y jugadores con atributos definitivos
Realidad	Determinar atributos significativos del *dataset* equipos y jugadores	Fichero de equipos y jugadores con atributos definitivos

Esta sexta iteración volveremos a trabajar con los métodos de feature detection explicados en iteraciones anteriores. Esta semana contemplada entre el 08/04 y el 14/04, pretendía realizar el mismo análisis ejecutado sobre los datos básicos de los enfrentamientos, pero esta vez sobre los datos de equipos y jugadores obtenidos la semana anterior. Dado que solo se consiguió generar el dataset de jugadores, dicho análisis tan solo se pudo ejecutar sobre los datos de los jugadores y se espera poder realizarlo en etapas posteriores a los datos de los equipos que nos falta.

Los resultados obtenidos se pueden visualizar en su totalidad en el fichero de datasets del apartado de documentación del repositorio de github. Pese a ello, como podemos observar las características que más influyen en el resultado final de un encuentro son la siguientes:



Como podemos observar, los atributos que más influencian en el resultado de un partido son los jugadores que han estado activos en el encuentro por parte del equipo local. Si analizamos estos resultados en detalle, podemos encontrar unas conclusiones del por qué de esta situación. Esta comprobado por algunos estudios realizados en todo el mundo [link a ese estudio] que el equipo local gana en el 70% de las veces. Por ese motivo, podemos entender que el factor clave para que un equipo gane, reside en el estado de los jugadores del equipo local que de forma estadística se llevan más porcentaje de victoria.

Finalmente, si tenemos que definir un resultado final del trabajo de esta iteración, podemos decir que la tarea ha sido realizada parcialmente por lo explicado anteriormente (la dependencia de los datos esperados a de la semana anterior). Esto a provocado un cambio en la planificación de la semana 14, dado que este análisis de atributos se deberá de realizar en ese momento por tal de poder realizar los mismos pasos con todos los datasets.

Seguimiento II

Una vez entregado el primer informe de seguimiento, el cuál consistía en las primeras 6 iteraciones de nuestro proyecto, continuamos con el segundo módulo de trabajo que estaba formado por 6 iteraciones. Éste segundo apartado del proyecto dejaba de lado todo el tratado de datos, el cual ya había sido realizado en las etapas anteriores, y se hacía más foco al desarrollo de todos los otros modelos matemáticos que se tenían previsto en la planificación del proyecto, así como la predicción de éstos sobre los diferentes datasets generados.

Para poder entender y definir correctamente todas las tareas realizadas, seguiremos el mismo formato que hemos utilizado para la explicación del módulo de seguimiento 1.

Iteración 7 (15/04 – 21/04)

Estado	Tarea	Resultados
Planificación	Aplicar Regresión Lineal para predecir la anotación sobre equipos y jugadores	Análisis de los resultados del experimento
Realidad	Reestructuración del dataset de jugadores para poder aplicarle los modelos matemáticos.	Dataset de jugadores reestructurado y resultados del experimento

En esta iteración se volvía al modelo de regresión lineal generado en algunas iteraciones anteriores. Esta semana contemplada entre el 15/04 y el 21/04, perseguía aplicar el modelo de regresión lineal sobre los datasets creados semanas anteriores relacionados con datos de equipos y jugadores.

Debido a que el dataset de equipos no había podido ser finalizado por lo motivos explicados en apartados anteriores, solo podíamos aplicar la selección de características sobre los datasets en relación a los jugadores.

A la hora de aplicar y poder visualizar los resultados, llegamos a la conclusión de que la estructura que habíamos generado para almacenar los jugadores activos e inactivos de ambos equipos no era idónea, pues se almacenaban como un array de strings que después el regresor era incapaz de utilizar para la predicción.

Visto esa problemática que imposibilitaba el hecho de realizar la tarea planificada para esta semana, se decidió aprovechar la semana para reestructurar los datos de los jugadores con el objetivo de poder aplicar los diferentes modelos en etapas anteriores.

Una vez realizado el cambio en la planificación y haber estudiado la forma óptima para almacenar los datos, refactorizamos el código que generaba el dataset. Posteriormente, procedimos a aplicar la regresión lineal sobre el dataset de jugadores por tal de conseguir unos nuevos resultados. Lo que obtuvimos fue lo siguiente:

Localidad	MAE
Local	0,000000238
Visitante	10,45

Como podemos observar, los resultados que hemos obtenido en la planificación de la anotación del equipo local no son realistas y se debe a algún error que actualmente desconocemos. Se intentará investigar el porqué de dicho valor, intentando detectar si está ocurriendo overfitting en el entrenamiento del modelo. Por lo que respecta a la predicción de la anotación del equipo visitante podemos observar como el resultado es algo peor a lo obtenido con los otros modelos hasta la fecha.

Iteración 8 (22/04 – 28/04)

Estado	Tarea	Resultados
Planificación	Aplicar Regresión Logística al *dataset* de partidos para predecir equipo ganador	Código de una Regresión Logística y resultados del experimento
Realidad	Aplicar Regresión Logística al *dataset* de partidos para predecir equipo ganador	Código de una Regresión Logística y resultados del experimento

En esta iteración dejamos de lado de nuevo todo lo relacionado con el preprocesamiento de datos. Esta octava semana contemplada entre el 22/04 y el 28/04, persigue generar el segundo modelo de predicción que utilizaremos en nuestro proyecto: la regresión logística.

Tras analizar la mejor forma con la que representar una regresión logística y visualizar diferentes ejemplos existentes en Internet, hemos sido capaces de generar el código que entrenará y predecirá los resultados de cada uno de nuestros datasets. Para ello, nos hemos guiado de algunos ejemplos y tutoriales que podemos encontrar en el portal towardsdatascience.com o incluso en la propia librería de sklearn [15].

Pero este apartado era tan solo era una de las dos secciones que se estimaron hacer para esta iteración número 8. Además, se esperaba poder aplicar dicho modelo a los datos obtenidos semanas antes con el objetivo de poder obtener los primeros resultados prediciendo la anotación de cada equipo. Para ello, accedimos a los datos almacenados en la primera semana del módulo de seguimiento 1 y eligiendo los atributos más importantes gracias al análisis realizado algunas semanas antes ejecutamos nuestro modelo y observamos los resultados obtenidos:

Ganador	Accuracy
Local = 0 || Visitante = 1	58.55

Como podemos observar, los resultados obtenidos con la regresión logística sobre estos primeros datos dejan mucho que desear. El motivo es simple, estamos trabajando con datos básicos de un enfrentamiento sin tener en cuenta ningún dato estadístico, que al fin y al cabo son las características que más información nos van a aportar a la hora de poder realizar una predicción con éxito.

Todo y con eso, es importante conocer los resultados y analizarlos, pese a que los datos no sean los más importantes, para así a posteriori poder comparar y detectar como influyen cada uno de los tipos de datos al resultado final de un partido de la NBA.

Iteración 9 (29/04 – 05/05)

Estado	Tarea	Resultados
Planificación	Aplicar Regresión Logística para predecir probabilidad de ganador sobre equipos y jugadores	Análisis de los resultados del experimento
Realidad	Aplicar Regresión Logística para predecir probabilidad de ganador sobre los datos de jugadores	Análisis de los resultados del experimento

En esta iteración se volvía al modelo de regresión logística generado en algunas iteraciones anteriores. Esta semana contemplada entre el 29/04 y el 05/05, perseguía aplicar el modelo de regresión logística sobre los datasets creados semanas anteriores relacionados con datos de equipos y jugadores.

Debido a que el dataset de equipos no había podido ser finalizado por lo motivos explicados en apartados anteriores, solo podíamos aplicar dicho modelo matemático sobre los datasets en relación a los jugadores.

Los resultados que hemos obtenido se pueden ver reflejados en la siguiente tabla representativa:

Dataset	Ganador	Accuracy
Jugadores	Local = 0 || Visitante = 1	63.19
Equipos	Local = 0 || Visitante = 1	-

Como podemos observar, los resultados obtenidos con la regresión logística sobre estos primeros datos estadísticos nos proporcionan un porcentaje de acierto mayor tal y como habíamos pronosticado semanas antes. El hecho de poder contar con información relacionada sobre quien ha disputado el partido y quien no, ha proporcionado al regresor mucha más visión sobre cada equipo y por lo tanto una mayor precisión de predicción.

El hecho de no poder aplicar dicho modelo a los datos relacionados con los equipos ha provocado que dicha tarea se traslade a la iteración 14 junto con otras tareas pendientes del mismo tipo.

Iteración 10 (06/05 – 12/05)

Estado	Tarea	Resultados
Planificación	Aplicar Random Forest para predecir ganador del partido con el dataset de partidos	Código del Random Forest y resultados del experimento
Realidad	Aplicar Random Forest para predecir ganador del partido con el dataset de partidos	Código del Random Forest y resultados del experimento

En esta iteración volvemos a trabajar con los modelos. Esta décima semana contemplada entre el 06/05 y el 12/05, persigue generar el tercer modelo de predicción que utilizaremos en nuestro proyecto: Random Forest.

Tras analizar la mejor forma con la que representar un Random Forest y visualizar diferentes ejemplos existentes en Internet, hemos sido capaces de generar el código que entrenará y predecirá el ganador del partido. Para ello, nos hemos guiado de algunos ejemplos y tutoriales que podemos encontrar en el portal towardsdatascience.com o incluso en la propia librería de sklearn [16].

Pero este apartado era tan solo era una de las dos secciones que se estimaron hacer para esta iteración número 10. Además, se esperaba poder aplicar dicho modelo a los datos obtenidos semanas antes con el objetivo de poder obtener los primeros resultados prediciendo el equipo ganador del partido y la anotación de cada uno de los equipos. Para ello, accedimos a los datos almacenados en la primera semana del módulo de seguimiento 1 y eligiendo los atributos más importantes gracias al análisis realizado algunas semanas antes ejecutamos nuestro modelo y observamos los resultados obtenidos:

Ganador	Accuracy
Local = 0 || Visitante = 1	53.18

Localidad	MAE
Local	5.78
Visitante	9.85

Como podemos observar, los resultados obtenidos con el algoritmo de Random Forest sobre estos primeros datos son muy parecidos a los obtenidos con la regresión lineal/logística con el mismo conjunto de datos. Pese a esa similitud, podemos observar como la predicción al ganador del partido se mantiene superior con el regresor logístico, mientras que a la hora de predecir la anotación solo vemos una diferencia en el MAE del visitante, que sitúa al Random Forest por detrás de los regresores utilizados en iteraciones anteriores.

Todo y con eso, veremos que sucede cuando se utilice un conjunto de datos más completo, puesto que según diferentes fuentes (especificadas en el estado del arte de este proyecto), sitúan el algoritmo de Random Forest como uno de los mejores a la hora de predecir el ganador del partido con un máximo de 74% de accuracy.

Iteración 11 (13/05 – 19/05)

Estado	Tarea	Resultados
Planificación	Aplicar Random Forest para predecir ganador del partido con los datasets de equipos y jugadores	Análisis de los resultados del experimento
Realidad	Aplicar Random Forest para predecir ganador del partido con los datasets de jugadores	Análisis de los resultados del experimento

En esta iteración se seguiría trabajando con el modelo de Random Forest generado en algunas iteraciones anteriores. Esta semana contemplada entre el 13/05 y el 19/05, perseguía aplicar el modelo de Random Forest los datasets creados semanas anteriores relacionados con datos de equipos y jugadores.

Debido a que el dataset de equipos no había podido ser finalizado por lo motivos explicados en apartados anteriores, solo podíamos aplicar dicho modelo matemático sobre los datasets en relación a los jugadores.

Los resultados que hemos obtenido se pueden ver reflejados en la siguiente tabla representativa:

Dataset	Ganador	Accuracy
Jugadores	Local = 0 || Visitante = 1	63.19
Equipos	Local = 0 || Visitante = 1	-

Dataset	Localidad	MAE
Jugadores	Local	7.49
	Visitante	9.40
Equipos	Local	-
	Visitante	-

Como podemos observar, los resultados obtenidos con el Random Forest sobre estos primeros datos estadísticos nos proporcionan un porcentaje de acierto mayor tal y como habíamos pronosticado semanas antes. El hecho de poder contar con información relacionada sobre quien ha disputado el partido y quien no, ha proporcionado al regresor mucha más visión sobre cada equipo y por lo tanto una mayor precisión de predicción.

Por lo que respecta a las predicciones sobre la anotación de cada uno de los equipos podemos observar como nos ha proporcionado una ligera mejoría en cuanto al equipo visitante, pero también una ligera perdida de precisión respecto al equipo local.

Iteración 12 (20/05 – 26/05)

Estado	Tarea	Resultados
Planificación	Aplicar Naive Bayes para predecir ganador del partido en todos los datasets	Código del Naive Bayes y resultados del experimento
Realidad	Aplicar Naive Bayes para predecir ganador del partido en todos los datasets	Código del Naive Bayes y resultados del experimento

En esta iteración volvemos a trabajar con los modelos matemáticos, esta vez con el último del proyecto. Esta decimosegunda semana contemplada entre el 20/05 y el 26/05, persigue generar el cuarto y último modelo de predicción que utilizaremos en nuestro proyecto: Naive Bayes.

Tras analizar la mejor forma con la que representar un Naive Bayes y visualizar diferentes ejemplos existentes en Internet, hemos sido capaces de generar el código que entrenará y predecirá el ganador del partido. Para ello, nos hemos guiado de algunos ejemplos y tutoriales que podemos encontrar en el portal towardsdatascience.com o incluso en la propia librería de sklearn [17].

Pero este apartado era tan solo era una de las dos secciones que se estimaron hacer para esta iteración número 12. Además, se esperaba poder aplicar dicho modelo a los datos obtenidos semanas antes con el objetivo de poder obtener los primeros resultados prediciendo el equipo ganador del partido. Para ello, accedimos a los datos almacenados en la primera semana del módulo de seguimiento 1 y eligiendo los atributos más importantes gracias al análisis realizado algunas semanas antes ejecutamos nuestro modelo y observamos los resultados obtenidos:

Dataset	Ganador	Accuracy
Partidos	Local = 0 || Visitante = 1	41.86
Jugadores	Local = 0 || Visitante = 1	61.41

Como podemos observar, los resultados obtenidos con el algoritmo de Naive Bayes sobre estos primeros datos dejan mucho que desear utilizando el dataset más básico (partidos). Los motivos pueden ser los mismos que en los otros modelos o incluso que no estemos utilizando el modelo de Naive Bayes más idóneo para este conjunto de datos puesto que estamos utilizando un NB Gaussiano.

En cambio, si hacemos foco a los resultados obtenidos con el dataset de jugadores activos e inactivos de cada partido, podemos observar como hemos conseguido subir hasta el 61% de acierto. Como en iteraciones anteriores hacia referencia, el estado de los jugadores en la NBA es muy importante siendo decisivo el hecho de que un jugador este lesionado o no sobre todo si hablamos de las estrellas de cada franquicia.

Planificación restante

Después de la finalización de los dos primeros módulos de nuestro proyecto, siendo esto la gran parte de nuestro proyecto, aún quedan un total de 5 iteraciones por realizar para dar por acabado la totalidad de nuestro trabajo de fin de grado. Dado los problemas que han ido apareciendo de forma imprevista y que han provocado que la planificación inicial se haya visto modificada ligeramente, procedemos a definir cual es la planificación actual restante para las semanas que quedan de trabajo. Los cambios que se han desarrollado se han ido especificando semana tras semana, con el objetivo de intentar especificar cuales han sido los motivos que han llevado a cabo dichos cambios.

Dicho esto, los siguientes pasos a desarrollar serán los siguientes:

Iteración	Fecha	Tarea	Resultados
13	27/05 – 02/06	Obtener correctamente el dataset de equipos/H2H y las cuotas de los partidos y crear un dataset global	Dataset de equipos o h2h y cuotas junto con un dataset global con todos los datos.
14	03/06 – 09/06	Aplicar todos los modelos a los nuevos datasets.	Resultados finales de nuestros modelos
15	10/06 – 16/06	Cálculo de beneficios netos	Análisis de beneficios
PROPUESTA INFORME FINAL			
16	17/06 – 23/06	Añadir concepto de valor y/o información extra	Código que optimiza el beneficio y análisis de estos

17	24/06 – 30/06	Acabar Dossier Propuesta de Presentación	Dossier finalizado Presentación TFG
DOSSIER			

Conclusiones iniciales

Tras el desarrollo de los dos primeros módulos de trabajo, así como de la obtención de buena parte de los datos que se tenían previstos para el uso en este proyecto, hemos podido obtener unos resultados iniciales con los que poder comparar y extraer unas conclusiones.

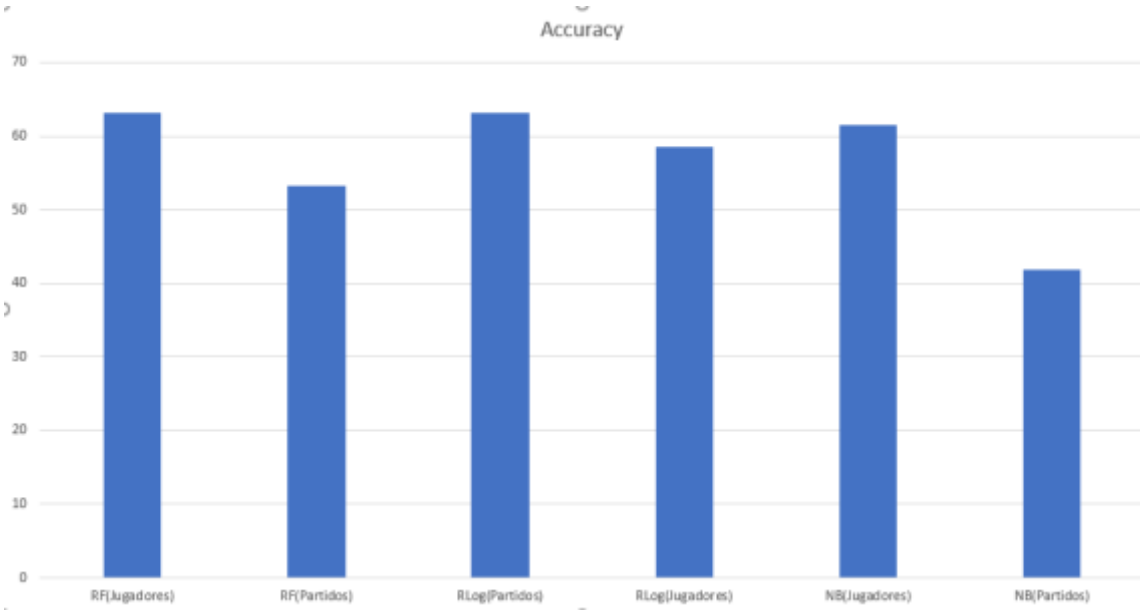
Cabe destacar que las conclusiones definitivas serán extraídas una vez se hayan obtenido todos los módulos de datos previstos, así como de haber aplicado los diferentes modelos matemáticos a todos ellos con el fin de poder hacer una comparación con mayor profundidad y, esperemos, con unos resultados más satisfactorios.

Haciendo referencia al conjunto de tareas que hemos ido realizando a lo largo de estas semanas, es importante remarcar como conclusión que sin duda todo lo relacionado con la obtención de datos, así como su preprocesamiento previo a la ejecución de modelo, ha sido la parte más compleja con una gran diferencia. El hecho de tener los datos con la estructura deseada des del inicio (véase documento datasets en la carpeta documentación del GitHub de nuestro proyecto), con el objetivo de poder tratar cada tipo de dato como un módulo que podemos añadir o eliminar a la hora de hacer las pruebas, y dada también la estructura con la que obteníamos los datos des de un inicio, ha provocado que el tiempo planificado para cada una de las tareas difiriera bastante de la realidad.

Todo y con eso, el hecho de saber adaptarse a esas situaciones y haber replanificado el proyecto, intentando buscar la forma óptima y sencilla con la que aplicar cada modelo, ha provocado que pese a los retrasos los objetivos planteados al inicio del proyecto vayan a ser cumplidos con éxito.

Si entramos en detalle con los resultados obtenidos en la ejecución de este proyecto, podemos fraccionar nuestras conclusiones según la característica que quisiéramos predecir.

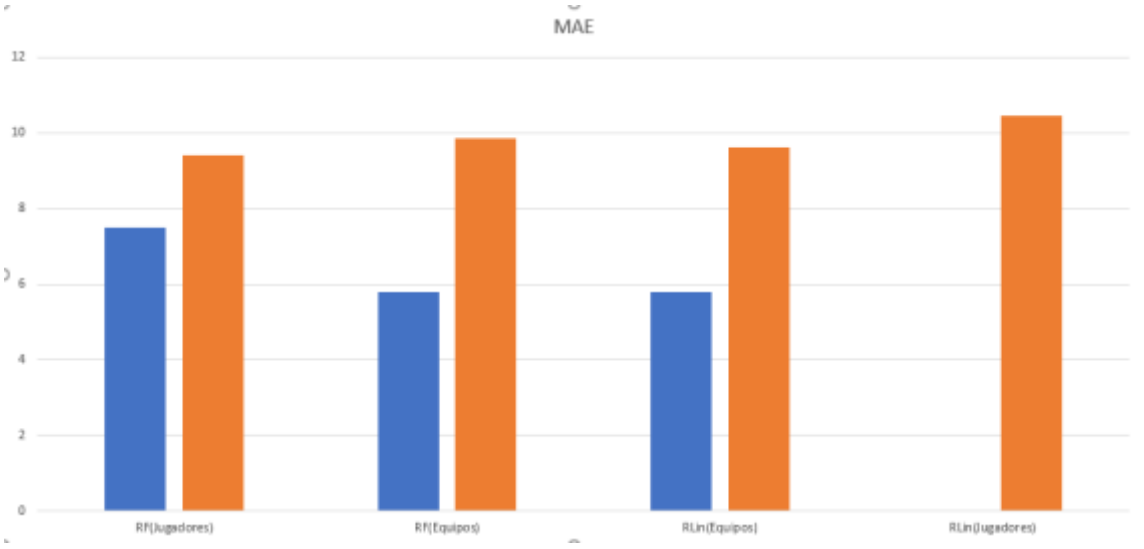
Por lo que respecta al ganador del partido, hemos obtenido resultados distintos según el modelo, pero también según la información que hayamos utilizado para la predicción.



Como podemos observar, todos aquellos modelos que han utilizado el dataset de jugadores, han obtenido mejores resultados que aquellos que tan solo han tenido en cuenta el dataset básico de enfrentamiento. Esto nos demuestra la importancia a la hora de seleccionar la información que utilizamos para realizar la predicción. A parte de esta conclusión, podemos ver como tanto el Regresor Logístico como el Random Forest obtienen el mismo accuracy, 63%, siendo éste el máximo hasta la fecha, por encima del algoritmo de Naive Bayes con los mismos datos.

Quedará en las conclusiones finales ver como influencian los datos de equipos a las predicciones y también visualizar como funcionan los algoritmos con todos los datos extraídos de este proyecto.

Si ahora entramos en detalle con los resultados obtenidos mediante la predicción de anotación por cada equipo, podemos observar los siguientes resultados parciales:



El mejor resultado en el equipo local, lo obtenemos con el modelo de Random Forest, de nuevo sobre el conjunto de datos sobre Jugadores, mientras que el equipo visitante tiene su mejor porcentaje de acierto con el Random Forest y el conjunto de datos básico de partidos.

En estos resultados también observamos un valor que tiende a 0, seguramente por el uso de overfitting en el entrenamiento del modelo de regresión lineal. En etapas posteriores se analizará dicho error y se intentará modificar para obtener un resultado válido.

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].

[13] R. Shaikh, «towardsdatascience.com,» 27 10 2018. [En línea]. Available: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e. [Último acceso: 19 03 2019].

[14] scikit-learn.org, «scikit-learn.org,» [En línea]. Available: https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html. [Último acceso: 26 03 2019].

[15] scikit-learn.org, «scikit-learn.org,» [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. [Último acceso: 24 04 2019].

[16] scikit-learn.org, «scikit-learn.org,» [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Último acceso: 06 05 2019].

[17] scikit-learn.org, «scikit-learn.org,» [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html. [Último acceso: 21 06 2019].