Universitat Oberta de Catalunya
uoc.edu

Learning to Rank aplicado al análisis avanzado del desempeño de jugadores en la NBA

Nombre Estudiante Iker Iturbe
Máster Universitario en Ciencia de Datos
Minería de datos y machine learning

Jerónimo Hernández González
Jordi Casas Roma

09 de junio del 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Learning to Rank aplicado al análisis avanzado del desempeño de jugadores en la NBA*
Nombre del autor:	*Iker Iturbe Azcorra*
Nombre del consultor/a:	*Jerónimo Hernández González*
Nombre del PRA:	*Jordi Casas Roma*
Fecha de entrega:	06/2019
Titulación:	*Máster universitario en Ciencia de Datos*
Área del Trabajo Final:	*Minería de datos y* machine learning
Idioma del trabajo:	*Castellano*
Palabras clave	Analytics, sports, basketball, learning to rank

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

Históricamente, la recogida de datos y el análisis de datos en los diferentes deportes se centra en estadísticas acumuladas anuales para comparar el desempeño de los diferentes jugadores.

Con el gran avance que se ha producido en la recogida y procesamiento de datos, existe la posibilidad de realizar análisis más avanzados. Serán análisis que nos permitan ponderar y realizar una clasificación, aplicando los conceptos del *learning to rank*, de los jugadores en función de aspectos que puedan ser influyentes a la hora de comparar su desempeño.

La hipótesis en que se basa este estudio sobre la NBA es que hay dos factores interrelacionados que influyen en el desempeño y que no suelen tomarse en consideración.

El primero, es el conocimiento del juego que permite a un jugador aplicar la estrategia correcta según se plantee un problema en forma de defensa adversaria. El segundo es la importancia del partido, ya que varía mucho según el momento de la temporada sea. En la NBA no existen descensos de categoría, la temporada regular es muy larga y en los *playoffs* las franquicias se juegan el trabajo de todo el año.

El objetivo de este estudio es conseguir un análisis estadístico que tenga en cuenta ambos factores para poder comparar los puntos fuertes y débiles de los jugadores. El resultado del estudio aportará información que permita a los entrenadores y directores deportivos realizar una rápida toma de decisiones en un mercado de fichajes muy cambiante.

Abstract (in English, 250 words or less):

Historically, data gathering and processing in sports has been based on yearly cumulative statistics to compare players' performance.

In the current Big Data Era, we have seen a large increase in the amount of data available. This enables to perform much more advanced statistical analyses by weighing the information about, for example, the defensive capabilities and performances of the opponents or the importance of the game.

There are two inter-related hypotheses on which this advanced study of the NBA player´s offensive performance is based. The first one is that the advanced knowledge of the game is what helps the player to make proper decisions when the opponent team poses a new problem by means of their defensive strategy.

And the second one is the importance of the game. In the particular case of the NBA, there is no relegation at the end of the league and the regular season is far too long compared to European leagues. All this means that the players' physical performance is much more important than the knowledge of the game, so that yearly cumulative data loses its significance in the comparison of players' performance.

The main aim of this study is providing a good statistical analysis taking into account these two factors to compare players' weak and strong points. The result will help coaches and managers when making quick and accurate decisions about recruiting their future staff in a very volatile market.

Índice de Tablas

Índice de Figuras

Índice de Contenidos

1. Introducción

1.1. Contexto y justificación del proyecto

Este proyecto se lleva a cabo con la finalidad de aportar una herramienta que permita a los entrenadores y directores deportivos de los equipos realizar la elección de los posibles fichajes según su desempeño.

Las características especiales de una competición como la NBA son dos: se juegan 82 partidos de liga regular por temporada (el doble que cualquier competición europea) y el resultado final depende de los *playoffs* (eliminatorias a 7 partidos) por el título.

Durante una temporada tan larga como es la de la NBA (hasta 110 partidos de temporada regular más *playoffs*), el cansancio físico influye en el desempeño de los jugadores a lo largo de la temporada, tal y como describen en su estudio (González et al., 2013)[1].

Dichas características, así como las que tienen que ver con las finanzas y el *marketing* de las franquicias (clubes), hacen que las estadísticas acumuladas anuales puedan estar desvirtuadas.

El estudio se servirá de técnicas de clasificación para tratar de construir un modelo que permita ponderar el peso de cada partido dentro de una temporada, teniendo en cuenta el rival al que se enfrenta y la importancia del partido.

El modelo de clasificación se construirá a partir de una serie de atributos (robos, tapones, puntos totales recibidos) que han sido destacados por Ibáñez et al. (2008)[2], muy importantes de cara a la evaluación de los resultados de una temporada.

Una vez elegidos los indicadores que nos permitan evaluar y comparar el desempeño de los jugadores, basándonos en las conclusiones del estudio de Javier García et al. (2013)[3].

La capacidad defensiva y la importancia del partido se ponderarán utilizando el cálculo del ratio de relación entre el precio del equipo local y el visitante de las casas de apuestas, tal y como se describe en detalle en apartados posteriores.

Como resultado final, que podemos analizar gráficamente en la figura una vez obtenido el modelo entrenado con la información recuperada mediante técnicas de *web scraping* de la página oficial de la NBA y que previamente se ponderará con el ratio de apuestas. Las etiquetas provendrán del archivo con la opinión entrenadores expertos (6 entrenadores superiores de baloncesto) comparando pares de jugadores.

Se procederá a detectar el resultado de quién es mejor entre los pares de todos los jugadores que juegan en la NBA más de 10 minutos por partido. Con las etiquetas generadas se sacará el ranking entrenado por los diferentes modelos y se evaluará su salida comparándolo con la información proveniente de BasketMondo. Para su análisis cuantitativo se incorporará a su vez la información general y salarial de los diferentes jugadores de forma que sirva para clasificar a los jugadores de cara a nuevas contrataciones.

1.2. Motivación personal

Como aficionado al baloncesto, he crecido observando a los mejores jugadores del mundo en una liga como la NBA. Tal y como ocurre en otros deportes, las políticas de fichajes se antojan muy importantes para las diferentes franquicias, y el poder comparar las estadísticas de los jugadores es complejo.

Como estudiante del máster en ciencia de datos, he querido aplicar los métodos de análisis estadístico para aportar mi granito de arena en el avance del conocimiento en la comparación del desempeño de jugadores de baloncesto.

Al tratarse de un deporte de equipo, es necesario que los jugadores entiendan las diferentes estrategias ofensivas y defensivas del juego. Para analizar esto puede ser muy relevante evaluar la capacidad de un jugador frente a los rivales defensivos más eficaces y en los momentos más importantes de la temporada. Esto es debido a que existen muchos jugadores con un gran entendimiento del juego, pero que tienen que programar su entrenamiento para obtener picos de rendimiento en partidos importantes. Lo hacen para poder ayudar al máximo al equipo en los momentos más importantes de la temporada y son jugadores que no obtienen medias anuales tan elevadas.

Por otro lado, hay grandes jugadores muy dotados físicamente que consiguen medias anuales muy altas, aunque muchas veces para que un equipo gane una liga hace falta mucho más.

Con ello surgió la idea de ponderar las estadísticas por partido, para tratar de conseguir que las estadísticas entre jugadores se puedan comparar y que puedan reflejar mejor el conocimiento del juego de algunos jugadores.

1.3. Objetivos del trabajo

El objetivo principal es realizar un estudio estadístico de clasificación avanzado, que pueda ayudar a los directivos y entrenadores como fuente de información de cara a plantear su estrategia en la política de fichajes y según las necesidades específicas de cada equipo. Para ello se analizará la información acumulada por temporada ponderada según las capacidades defensivas del equipo contrario y la importancia del partido. El resultado del proceso se relacionará con los salarios de los jugadores para tratar de establecer una clasificación de intereses de cara al mercado de fichajes.

Como objetivos secundarios o relacionados con el objetivo principal estarían:

- Analizar el estado del arte del estudio, es decir, ver la evolución de la línea de investigación que se ha seguido y ver el estado actual en el que se encuentra.
- Desarrollo iterativo del proyecto y evaluación del mismo.
- Documentación relativa al proceso utilizado y los resultados obtenidos.
- Documentación de la presentación del trabajo.
- Publicación de la memoria de trabajo.

De manera opcional, se establece el objetivo secundario de comparar los datos históricos en las diferentes temporadas obtenidos en el estudio. Mediante la comparación de datos históricos se tratará de encontrar tendencias o incluso un modelo que nos permita predecir el desempeño futuro de los jugadores e incluso de diferentes configuraciones posibles de equipos.

1.4. Enfoque y método seguido

La estrategia por pasos de este estudio para realizar un análisis avanzado del desempeño de los jugadores sería:

1. Obtener mediante técnicas de *web scraping* las estadísticas por jugador y partido de diferentes temporadas y calcular las medias por posición que ocupan en el campo.
2. Obtener información de las apuestas en las que se pondera la importancia del rival contrario y la importancia del partido dentro de la temporada. Descarga de datos del data set "NBA Historical Stats and Betting Data" de la plataforma Kaggle (2018).[1]
3. Ponderar el desempeño anual en función de la información obtenida en los apartados 2 y 3 calculando los totales anuales que puedan compararse.
4. Incorporar al estudio información de las lesiones para valorar el riesgo del fichaje y el salario para poder acondicionar la elección a las necesidades de cada equipo.
5. Establecer un modelo predictivo que relacione la información histórica obtenida en los pasos previos, para tratar de predecir mediante tareas de regresión el desempeño tanto de jugadores, como de posibles equipos. (OPCIONAL)

1.5. Planificación del Trabajo

La planificación de hitos del TFM viene definida por las siguientes entregas ya planificadas para la terminación del TFM y que consiste en:

1. Definición y planificación del trabajo final
2. Estado del arte o análisis de mercado del proyecto
3. Diseño e implementación del proyecto
4. Redacción de la memoria

[1] https://www.kaggle.com/ehallmar/nba-historical-stats-and-betting-data

5. Presentación y defensa del proyecto

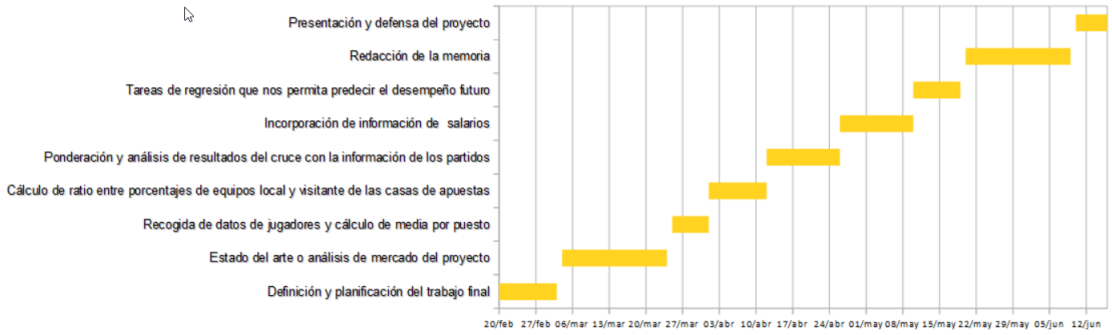
Tabla 1: Planificación de hitos

Fase	Fecha Inicio	Duración
Definición y planificación del trabajo final	20/02/19	11
Estado del arte o análisis de mercado del proyecto	04/03/19	20
Recogida de datos de jugadores y cálculo de media por puesto	25/03/19	7
Cálculo de ratio entre porcentajes de equipos local y visitante de las casas de apuestas	01/04/19	11
Ponderación y análisis de resultados del cruce con la información de los partidos	12/04/19	14
Incorporación de información de salarios	26/04/19	14
Tareas de regresión que nos permita predecir el desempeño futuro	10/05/19	9
Redacción de la memoria	20/05/19	20
Presentación y defensa del proyecto	10/06/19	6

Dentro del detalle de la actividad principal de diseño e implementación del proyecto, he establecido los siguientes hitos:

1. Recogida de datos de jugadores y cálculo de media por puesto.
2. Calculo del ratio de apuestas.
3. Ponderación y análisis de resultados del cruce con la información de las apuestas.
4. Incorporación de información de lesiones y salarios.
5. Tareas de regresión que nos permita predecir el desempeño futuro.
6. A continuación, el diagrama de Gantt y el detalle de fechas de cada hito.

Figura 1: Diagrama de Gantt



1.6. Breve descripción de los otros capítulos de la memoria

El presente trabajo comienza con una introducción al contexto donde se enmarcará el mismo, la utilización de técnicas de *Learning to Rank* para la clasificación de jugadores según su desempeño durante la liga regular. A continuación, se justifica el motivo o interés por el cual se ha decidido a realizar este tipo de estudio, para continuar con la descripción de los objetivos principales y secundarios propuestos.

Seguidamente se indica el enfoque y el método que se seguirá en el estudio, así como la planificación. Dentro de la propia planificación se describen

brevemente los recursos y el tiempo de dedicación previsto para cada una de las fases, finalizando así la parte introductoria del trabajo.

En el tercer capítulo se describe cómo ha estado marcado el estado del arte en el campo del *data mining* asociado al baloncesto. Empezando por una breve introducción acerca de la recogida de datos (*scouting*). Continuando con los aspectos de la evaluación del desempeño, el preprocesado de la información y las técnicas de análisis de datos en proyectos relacionados con el deporte. Para acabar y continuando con el enfoque de ir de lo general a lo particular, se describe el estado del arte del *data mining* aplicado al baloncesto.

A continuación, se detallan las herramientas y métodos más importantes que se utilizan en el estudio.

En el capítulo dedicado a las fuentes de datos del desempeño de los jugadores de la NBA, se indican cuáles serán las utilizadas para el entrenamiento y evaluación del modelo, descrito más adelante.

Seguidamente se realiza una definición del modelo para realizar la clasificación de jugadores utilizando técnicas de *learning to rank*, así como las razones para su elección justificada.

Una vez definido el modelo, se describirán las transformaciones necesarias para la preparación de los datos definidos en las fuentes de datos para poder aplicar los diferentes algoritmos de *machine learning*. Posteriormente se visualizarán y evaluarán los resultados de los diferentes algoritmos y la justificación de la elección de estos.

Finalmente, se mostrarán las conclusiones del estudio, las propuestas de mejora y el posible trabajo futuro a partir del estudio, así como la bibliografía relacionada con el estudio.

2. Análisis del ámbito y estado del arte

2.1. El arte del *scouting*

La metodología observacional ha sido utilizada a lo largo de los años para la recogida de datos sobre el desempeño de los jugadores de baloncesto en las diferentes ligas de baloncesto de élite de todo el mundo.

Como método observacional para la recogida de datos[4], entendemos aquel "*procedimiento científico encaminado a articular una percepción deliberada de la realidad manifiesta con su adecuada interpretación captando su significado, de forma que mediante un registro objetivo, sistemático y específico de la consulta generada espontáneamente en un determinado contexto, una vez que se ha sometido a una adecuada codificación y análisis, nos proporciona resultados válidos dentro de un marco específico de conocimiento*".

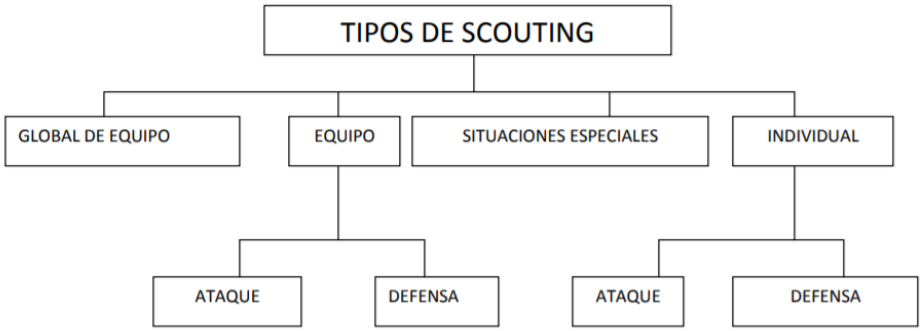
Los requisitos de cumplimiento básico exigidos por la metodología observacional son los expuestos en la figura 2 por Anguera et al. (2000)[5]:

Figura 2: Requisitos de la Metodología Observacional

1. **Espontaneidad del comportamiento**
2. Realizado en **contextos naturales** (dentro del ámbito del deporte y de la actividad física serían aquellos contextos donde se produce habitualmente la actividad, el terreno de juego o la cancha)
3. Que se trate de un estudio prioritariamente **idiográfico** (centrado en el individuo)
4. Requiere de la **elaboración de instrumentos ad hoc**
5. Que se garantice una **continuidad temporal**
6. Eligiendo un **tipo de perceptividad** del comportamiento, puede ser
    a. En grado máximo (observación directa, o de conductas manifiestas),
    b. En grado parcial (observación indirecta, o de conductas encubiertas
7. Determinar **características propias** del objeto de estudio y el tamaño de las unidades estudiadas.

Dentro de los tipos de *scouting* utilizados en las diferentes ligas de baloncesto del mundo, podemos realizar una clasificación de las diferentes metodologías de *scouting* descritas en la figura 3 por Aitor Antía Martín-Carrillo et al. (2014)[6]:

Figura 3: Tipologías de *scouting*



La recogida de datos en los últimos años ha mejorado sustancialmente con los avances en la tecnología y fotografía digital, tal y como nos describe en el estudio (Bishop et al., 2003)[7].

2.2. *Data mining* en la evaluación del desempeño en los deportes

Los principales atributos de interés dentro del *data mining* para el análisis de datos del desempeño en los deportes son los grados, tiempos y marcas (denominado RTS en la literatura en inglés, proveniente de *Rankings, time and scores*).

Por otro lado, en los diferentes deportes se establecen una serie de medidas de los atributos específicos de cada deporte y que vienen definidos por las principales reglas, regulaciones, tácticas, estrategias, desempeño, condiciones y habilidades relacionadas con cada deporte en particular.

En la siguiente figura 4 descrita por Ofoghi et al. (2013) podemos observar las relaciones entre los requerimientos del análisis del desempeño en deportes y los principales métodos de la minería de datos (*Clustering*, clasificación, modelado de relaciones, minería de reglas)[8]:

Figura 4: Requerimientos del *Data mining* en el Deporte

The mapping between sports performance analysis requirements and major data mining methods and the mapping between the sports performance analysis requirements and the data mining technique characteristics

Sports performance analysis requirements	Data mining methods			
	clustering	classification	relationship modeling	rule mining
performance pattern discovery	√	√	–	–
performance prediction	–	√	√	–
real-time decision-making	–	–	√	√
demand analysis	√	√	–	–

El análisis del desempeño en los diferentes deportes de élite implica técnicas de preprocesado de la información (específicas para deporte) y técnicas de análisis de datos (más relacionadas con el tipo de problema a analizar).

2.3. El preprocesado de la información en proyectos de *data mining* relacionados con los deportes

Las técnicas de preprocesado en el ámbito de la minería de datos relacionadas con el deporte, como las descritas en el estudio de Bhandari et al. (1997)[9], pueden incluir, según la naturaleza del problema, los siguientes pasos:

- **Filtrado**: se trata de la selección de registros de datos en función de la naturaleza del problema y que puede incluir atributos como la competición, el año, el puesto del jugador, etc.
- **Formateo**: se trata de convertir la información en un formato que pueda ser interpretado por un software de analítica de datos específico.
- **Extracción**: se trata de encontrar nueva información que se base en la información recolectada.
- **Conversión estructural:** se trata de convertir partes específicas de los datos en un formato que permita un análisis de datos más preciso. Por ejemplo, la creación de un atributo que convierta las posiciones en una disciplina de los juegos olímpicos en si se consiguió o no una medalla o un diploma olímpico.
- **Conversión descriptiva:** se trata de convertir partes específicas de los datos en atributos que describan la naturaleza del problema de mejor forma. Un buen ejemplo de esto es el cálculo de estadísticas por minuto de cara a comparar los datos entre jugadores que podemos observar en el estudio de Sampaio et al. (2015)[10].

Los tres primeros pasos no son específicos del deporte, mientras que la conversión estructural depende del método analítico utilizado y del volumen de datos a analizar. Por su parte, la conversión descriptiva está fuertemente conectada al conocimiento específico del deporte y sus características.

Dentro de las tareas de preprocesado habrá que tener en cuenta la influencia de los siguientes aspectos:

1. El número de eventos en una competición
2. El número de participantes en la competición. En el caso de los deportes de equipo, habrá que tener en cuenta el desempeño del resto de participantes.
3. Duración del evento deportivo.
4. El criterio de éxito específico definido para cada deporte.

2.4. Técnicas de análisis de datos relacionados con los deportes

Desde el punto de vista de la ciencia del deporte, existen los siguientes objetivos:

1) Encontrar patrones del desempeño que describan cómo un atleta o equipo pueden aumentar sus opciones de ganar una competición particular.
2) Predecir el desempeño de atletas o equipos en función del desempeño en anteriores competiciones o entrenamientos.
3) Un selector de respuestas en tiempo real que tenga en cuenta las acciones y reacciones que se dan entre equipos en un evento en particular.
4) Encontrar según las demandas de un deporte en particular los atletas que mejor cumplen una serie de características.

Desde el punto de vista de la analítica de datos, los diferentes métodos se diferencian por las siguientes características:

1) Interpretabilidad de los resultados por profesionales no expertos en la materia.
2) Precisión de los resultados obtenidos.
3) Flexibilidad del método para ajustarse a nuevos parámetros o juegos de datos.

En la siguiente figura 5 descrita por Ofoghi et al. (2013), podemos observar la relación entre los diferentes objetivos y la evaluación de las características que las describen[8]:

Figura 5: Técnicas del *Data mining*

	Data mining technique characteristics		
	interpretability	precision	flexibility
performance pattern discovery	high	moderate	moderate
performance prediction	low	high	high
real-time decision-making	very high	high	very low
demand analysis	moderate	moderate	moderate

2.5. Estado del arte del *data mining* en el baloncesto

Diferentes estudios de Trninic et al. (1999)[11] y (2000)[12], han analizado los atributos a medir en el análisis del desempeño ofensivo. En estos estudios se hace una clasificación de atributos defensivos y ofensivos para analizar el desempeño de los jugadores por rol dentro del equipo. Dichos atributos son el control del balón, habilidades de pase, penetración con bote, lanzamientos de 3 puntos, lanzamientos de 2 puntos, tiros libres, faltas recibidas, eficacia en los bloqueos, ataque sin balón, rebotes ofensivos, lanzamiento en transiciones ofensivas y juego en múltiples roles.

El estudio de Gerald Mangine et al. (2014)[13] es más innovador en este aspecto, ya que analiza la capacidad de los jugadores para hacer un seguimiento de la información visual que se da en un partido de baloncesto.

En el estudio de Sampaio del 2015[10], realizaron un análisis estadístico en el que analizaron el desempeño de los jugadores *All-Star* frente al resto de jugadores. Para ello, analizaron información sobre los puntos por partido, minutos por partido y las acciones de juego descritas por Trninic[11] en su estudio, añadiendo información sobre el posicionamiento de los jugadores en la cancha.

Como conclusión de dicho estudio, se detecta la necesidad de normalizar los datos de entrada del desempeño por jugador para poder comparar los datos entre jugadores mediante la obtención de las medias por minuto. Esto se debe a que la diferencia de minutos que juegan unos jugadores frente a otros puede hacer que las medias por partido no puedan compararse.

Actualmente, uno de los mayores avances en la evaluación del desempeño de jugadores, es la incorporación de datos del posicionamiento en la cancha, tal y como refleja el estudio de Goldsberry et al. (2012)[14].

La falta de disponibilidad de información sobre el posicionamiento de los jugadores ha hecho que este apartado quede fuera del ámbito del presente estudio.

Como línea de trabajo a futuro, proponemos incorporar datos del posicionamiento de los jugadores y el balón en el campo tanto para el análisis del desempeño de los jugadores como para la clasificación de los equipos según su defensa.

3. Materiales y métodos

3.1. Anaconda

Para el desarrollo del proyecto se ha elegido Anaconda como entorno integrado de desarrollo. Anaconda es una distribución gratuita y de código abierto que permite programar aplicaciones en Python y R aplicando técnicas de ciencia de datos y aprendizaje automático.

Anaconda permite a los científicos de datos:

- Descarga rápida de más de 1500 paquetes en Python/R de ciencia de datos.
- Manejo de librerias, dependencias y entornos con Conda.
- Desarrollar, entrenar y testear modelos de aprendizaje automático y *deep learning*.
- Analizar información con escalabilidad y rendimiento.
- Visualizar los resultados.

3.2. Python

El lenguaje de programación elegido para el proyecto en Python. Es un lenguaje de programación de propósito general, multi-paradigma (incluye diferentes estilos de programación), interpretado y de alto nivel.

Python permite a los programadores utilizar diferentes estilos de programación para crear programas simples o complejos, obtener resultados más rápido y escribir código de forma sencilla, tratando de acercarse al lenguaje natural.

Fue creado a finales de los 80 por Guido Van Rossum, aunque hoy en día existe una *Python Software Foundation* que se encarga del desarrollo de la misma.

Hay dos atributos que hacen que el tiempo de desarrollo en Python sea menor que otros lenguajes de programación:

- Se trata de un lenjuaje interpretado que realiza la compilación en segundo plano y evita el compilado previo a la ejecución de un programa. Como es un lenguaje de alto nivel, trata de maximizar la abstracción del código para facilitar su lectura y comprensión.
- El código en Python es mucho más corto que en otros lenguajes y por lo tanto el tiempo de desarrollo es menor.

3.3. Técnicas de *web scraping*.

De entre las técnicas descritas por Zao Bo (2017)[15], existen varios métodos entre los que cabe destacar los siguientes para el proyecto que nos ocupa:

- Extraer la información del HTML en bruto, como Beatiful Soup o Pyquery.

- Utilizar un entorno de pruebas de software para aplicaciones basadas en la web, como Selenium.

3.4. Técnicas de selección de atributos.

En la construcción y optimización de los modelos de clasificación, es necesario realizar un análisis sobre la posibilidad de poder realizar una selección de un subconjunto de atributos del conjunto inicial.

Dicha necesidad es especialmente importante cuando las fuentes de datos cumplen las carácterísticas del *big data*, definidas en el estudio de M. Chen et al.(2014)[16]:

- Volumen. Gran cantidad de datos que hoy en día genera la sociedad en su conjunto.
- Variedad. Diferentes tipologías de datos, provenientes de diferentes fuentes.
- Velocidad. Diariamente se generan nuevos datos y su crecimiento es exponencial en muchos casos.

Las ventajas esperadas con las técnicas de selección de atributos son las siguientes:

- Mejora del desempeño predictivo del modelo.
- Contrucción de modelos de manera más eficiente.
- Mejora del entendimiento de los modelos generados.

De acuerdo con Guyon et al. (2003), existen 3 técnicas principales entre los métodos de selección de atributos:

- **Métodos de filtro**. Se trata de utilizar métodos estadísticos para filtrar aquellos atributos que no sean relevantes o sean redundantes, antes de aplicar el algoritmo de aprendizaje. *Ventajas:* sencillos y rápidos en su ejecución, independientes del modelo. *Desventajas:* se realizan sin tener en cuenta el modelo predictivo y por tanto no tienen en cuenta las relaciones entre las variables.
- **Métodos *wrapper* o envolventes***.* Exploran el conjunto completo de atributos para poder asignar a cada uno una puntuación. Dicha puntuación varía en función del peso que tienen en la predicción dentro del método elegido para realizar la clasificación. *Ventajas:* mejores resultados que los métodos de filtro. *Desventajas:* computacionalmente demandante y específicas del método de clasificación utilizado.
- **Métodos *embedded* o empotrados**. Realizan la búsqueda de un subconjunto óptimo de atributos durante la construcción de la función de clasificación. *Ventajas:* toman en cuenta las relaciones entre las variables y el modelo de clasificación así como las relaciones entre atributos, y son computacionalmente menos demandantes que los métodos *wrapper*. *Desventajas*: conceptualmente complejos y no se adaptan a las características de algunos métodos, pudiendo caer en óptimos locales.

3.5. Algoritmos en el ámbito del *data mining*

Los métodos generales del *data mining* según los apuntes de la asignatura "Modelos avanzados de minería de datos" del máster en ciencia de datos" (UOC,2018) son:

✓ **Regresión logística.** Se trata de un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. *Ventajas*: este método es muy interesante para entender la influencia de las diferentes variables independientes en la variable categórica objetivo. *Desventajas:* solo funciona bien cuando la variable a predecir es binaria; todas las variables predictoras son independientes y en la información de entrada no hay valores faltantes.

✓ **Naive Bayes.** Se basa en el teorema de Bayes y la asunción de la independencia de cada par de características. *Ventajas:* requiere un juego de datos de entrenamiento para estimar los parámetros del modelo predictivo, y es realmente rápido en comparación con métodos más sofisticados. *Desventajas*: el algoritmo es conocido por ser un mal estimador.

✓ **Gradiente estocástico descendente.** Es un algoritmo simple y muy eficiente para ajustar modelos lineales. Permite ajustar diferentes niveles de pérdida y penalizaciones para la clasificación. *Ventajas:* eficiencia y facilidad de implementación. *Desventajas:* requiere el ajuste de una serie de parámetros y es sensible a las diferencias de escalado en los datos de entrada.

✓ **K-Nearest Neighbours (KNN).** Es un tipo de aprendizaje que no incluye una fase de entrenamiento, ya que, a la hora de predecir, KNN utiliza toda la información para construir el modelo. *Ventajas:* produce algoritmos simples y robustos. *Desventajas:* se necesita calcular el valor de K y el cálculo tiene un coste de procesamiento muy elevado.

✓ **Arboles de decisión**. Se trata de un método que, dados una serie de atributos y las etiquetas con su clasificación, genera una serie de reglas que pueden servir para clasificar la información de futuras ocurrencias. *Ventajas:* son simples de entender y visualizar y sirven tanto para datos continuos, como categóricos. *Desventajas:* puede llegar a ser difícil de entender cuando las reglas derivadas de la clasificación son muy complejas. Su excesiva especialización en los datos de entrenamiento puede no generalizar bien para nuevos datos que se introduzcan en el cálculo del modelo.

✓ **Random Forest.** Este algoritmo es un meta-estimador que utiliza una serie de árboles de decisión con características diferentes y que se entrena con partes del conjunto original de datos. La estimación se realiza como una media de los diferentes estimadores, para evitar el sobreentrenamiento. Como entrada del algoritmo, se utilizan partes del juego de datos original y con el

mismo número de casos, utilizando reemplazos en los mismos. *Ventajas:* producen resultados más precisos y evitan el sobreentrenamiento. *Desventajas:* la predicción en tiempo real es muy lenta, son difíciles de implementar y el algoritmo es complejo.

✓ **Máquinas de soporte vectorial.** Es un algoritmo que representa los datos de entrenamiento como puntos en el espacio que, al estar separados en categorías, maximizan el espacio vacío entre las diferentes categorías. *Ventajas:* es eficiente incluso para espacios dimensionales amplios. *Desventajas*: el algoritmo no provee directamente las probabilidades estimadas, sino que hay que calcularlo mediante una costosa validación cruzada.

✓ **Redes neuronales artificiales.** Representan un modelo matemático compuesto por una agrupación de unidades computacionales simples que están interconectadas a través de un sistema de conexiones o enlaces para transmitir la información. Las diferentes neuronas artificiales pueden realizar el procesamiento de la información en paralelo. La información de entrada atraviesa la red neuronal sometiéndose a diversas transformaciones y produciendo diferentes salidas. En los enlaces, el valor de salida de la neurona anterior es multiplicado por un valor de peso. Los pesos en los enlaces pueden incrementar o inhibir el estado de activación de las neuronas adyacentes. De la misma forma puede existir un umbral que modifique el valor resultante a la salida de la neurona o establezca el límite para su propagación a las neuronas que estén interconectadas y que se llama función de activación. *Ventajas*: sobresalen en áreas donde la detección de soluciones o características es difícil de expresar con la programación convencional. Son especialmente adecuados para el modelado de relaciones. *Desventajas*: son algoritmos complejos de desarrollar.

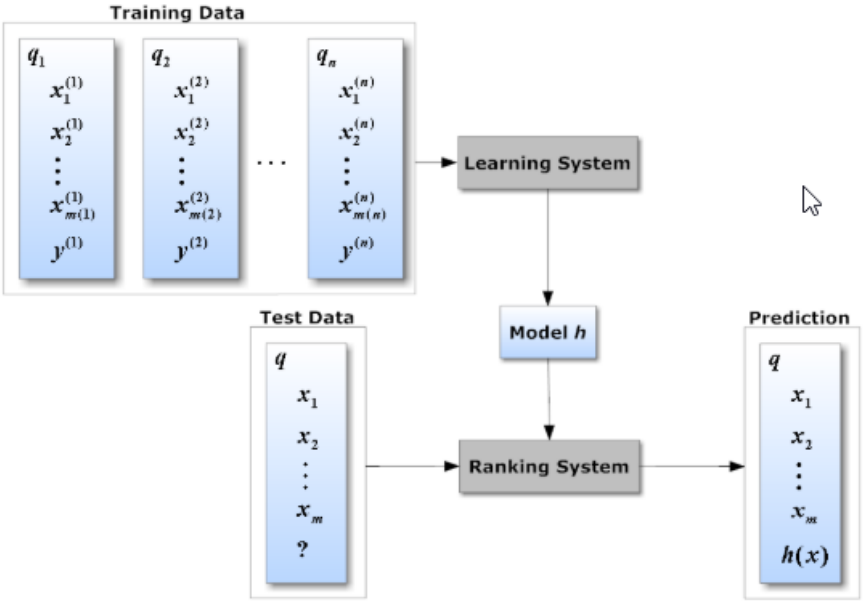
3.6. Algoritmos generales de clasificación en el ámbito del *data mining*

Dentro de la rama de algoritmos de clasificación, el problema descrito en el proyecto entraría dentro de lo que se ha llamado *learning to rank*. *Learning to rank* es un tipo de modelo matemático entrenado con técnicas de aprendizaje supervisado que permite realizar tareas de clasificación de una serie de objetos en función de otra serie de atributos de dichos objetos (vector de características).

La infraestructura propia del *learning to rank* nos permite construir los modelos, así como ajustar los parámetros mediante el uso de información etiquetada a mano por expertos en la materias. Incluso permite utilizar información etiquetada de diferentes fuentes y de esta forma combinar diferentes estimadores de relevancia de manera óptima.

La aproximación general para este tipo de problemas se compone de dos pasos que consisten en la fase de entrenamiento y la fase de testeo, tal y como vemos en la siguiente descripción del proceso de *learning to rank* aplicado a sistemas de recuperación de la información de Catarina Moreira (2019)[17]:

Figura 6 Arquitectura del Learning to Rank



En esta infraestructura, $q_i(i = 1, ..., n)$ se corresponde con la colección de n consultas para el paso de entrenamiento, $x^{(i)} = \{x_j^{(i)}\}_{j=1}^{m(i)}$, siendo m el número de documentos asociados a la consulta q_i que conforma el vector de carácterísticas. $y^i(i = 1, ..., n)$ se corresponde con el set de juicios de relevancia de la *query* en cuestión.

Cuando se aplica un método de aprendizaje especifíco a unos datos de entrenamiento determinados, el sistema combina los diferentes estimadores de relevancia de una manera óptima, aprendiento el correspondiente modelo de *ranking h*. Durante el proceso de aprendizaje, una función de pérdida se aplica para medir las inconsistencias entre el espacio de hipótesis h y la realidad terrenal y.

En el paso de testeo, el modelo de *ranking* aprendido se aplica a una *query* nueva de forma que se pueda ordenar los documentos según la relevancia aplicada a la necesidad de información, devolviendo el *set* del *ranking* de documentos ordenados por *query*.

Tal y como describe Hang Li et al. (2011), las técnicas de *learning to rank* se han utilizado con éxito en sistemas de recuperación de información, en procesamiento del lenguaje natural y en minería de datos.

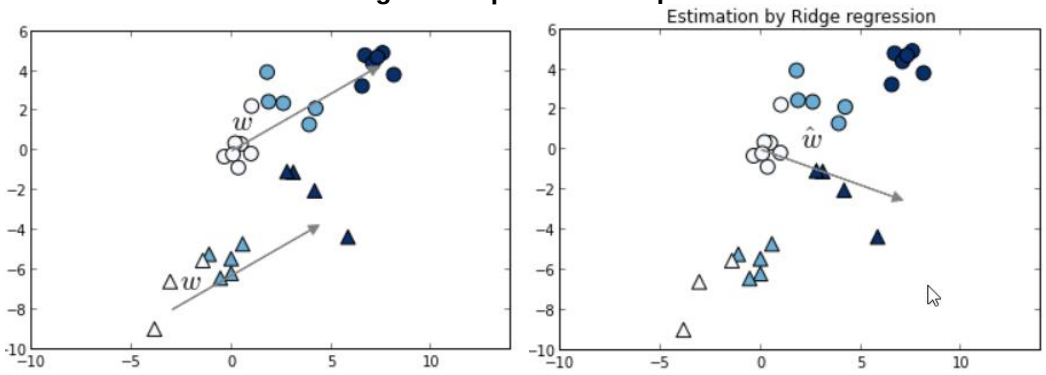
Hay que diferenciar entre la clasificación ordinal, en la que se trata de asignar una clasificación dentro de varios grupos a un producto, como puede ocurrir en un sistema de asignación de *ratings* para una película y el *learning to rank*, en el que lo importante es que los diferentes objetos a comparar estén correctamente ordenados entre sí.

Existen tres tipos de aproximaciones descritas tambien en (Hang Li et al., 2011) para la solución de problemas de clasificación:

✓ **Aproximación Pointwise.** Es un algoritmo en el que cada objeto tiene una puntuación numérica u ordinal. Se trata de algoritmos de clasificación y regresión ordinales. *Ventajas:* eficiencia y facilidad de implementación. *Desventajas:* solo son útiles para ejemplos muy simples.

✓ **Aproximación Pairwise.** Es un algoritmo en el que el problema de clasificación es transformado en un problema de clasificación con dos clases. Las etiquetas se asignan comparando los atributos entre los diferentes pares de objetos, tal y como podemos observar en la figura 2 aportada por Fabian Pedregosa et al. (2012).[18] *Ventajas:* produce algoritmos simples y robustos. *Desventajas:* necesita de un procesado previo de la información y de una buena estrategia de entrenamiento del modelo.

Figura 7: Aproximación pairwise

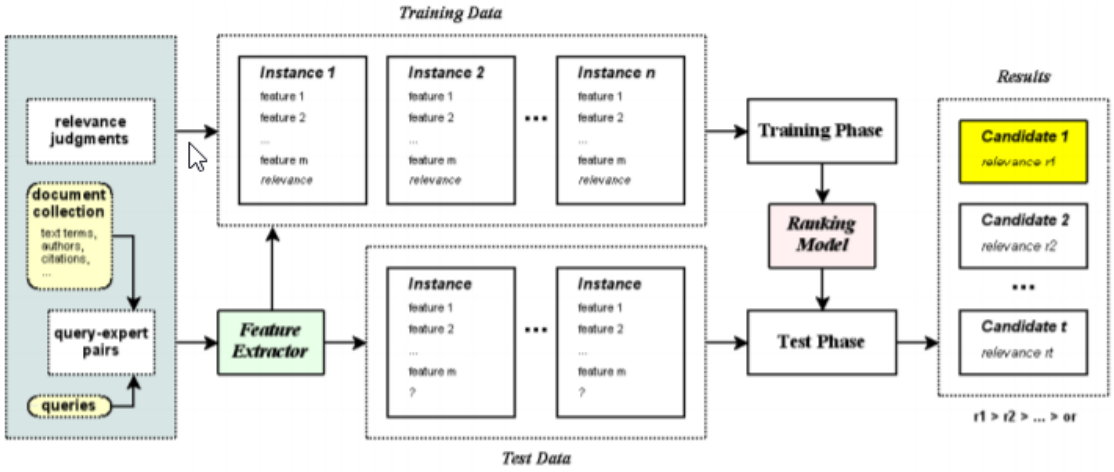


✓ **Aproximación Listwise.** Es un algoritmo que utiliza listas de clasificación como instancias de clasificación y predicción, de forma que la información de grupo puede incorporarse para ajustar el modelo. *Ventajas:* mantiene características de grupo. *Desventajas:* cálculo complejo y costoso.

El prototipo de *learning to rank* se compone de un grupo de consultas $Q = \{q_1, \ldots, q_{|Q|}\}$ y un grupo de expertos $E = \{e_1, \ldots, e_{|E|}\}$, cada uno asociado a documentos específicos que describen la experiencia de los expertos. Entonces se crea un corpus de entrenamiento para los problemas de *learning to rank* compuesto por pares (*query*-experiencia), $(q_i, e_j) \in Q \times E$ en el que los expertos e_j asignan a q_i una etiqueta a mano en función del juicio sobre la experiencia de las *queries* que hayan concluido. Está etiqueta creada por el experto consiste únicamente en información binaria sobre si es relevante para la *query* o no.

Podemos observar el prototipo en la figura 8, recogida del estudio anteriormente referenciado de *learning to rank* aplicado a sistemas de recuperación de la información de Catarina Moreira (2019):[17]

Figura 8: Prototipo de Learning to Rank IR



3.7. Matriz de confusión

Para la evaluación de los modelos resultantes, calcularemos la matriz de confusión de los diferentes algoritmos. Para ello se realizarán los siguientes recuentos descritos en la Tabla 2 para ver cómo se está comportando el modelo en sus predicciones:

Tabla 2: Matriz de confusión[2]

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
Predicted Values	Negative (0)	FN	TN

En la matriz podemos ver los resultados de una serie de conteos para ver cuándo la predicción de quién es mejor jugador coincide con la información etiquetada de las encuestas. Los recuentos consistirán en:

- **Verdaderos positivos** (TP=*True Positive*). Detecciones y recuentos de los datos reales positivos que coinciden con el resultado de la predicción de la etiqueta de mejor jugador de los pares.
- **Falsos negativos** (FN=*False Negative*). Son los llamados errores de tipo 2. La predicción que nos arroja el modelo para esos pares de

[2] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

jugadores es falsa, mientras que según la opinión de los expertos es verdadera.

- **Falsos positivos** (FP=*False Positive*). Son los llamados errores de tipo 1. La predicción que nos arroja el modelo para esos pares de jugadores es verdadera, mientras que según la opinión de los expertos es falsa.
- **Verdaderos negativos** (TN=*True Negative*). Detecciones y recuentos de los datos reales negativos que coinciden con el resultado de la predicción de la etiqueta de mejor jugador de los pares.

A partir de los anteriores recuentos, se pueden calcular las siguientes métricas para evaluar los modelos propuestos en el estudio:

- Tasa de verdaderos positivos (*Recall*), tambien llamada sensibilidad o exhaustividad. Probabilidad de que, dado un caso real positivo, el modelo realice una detección positiva a su vez.

  Recall = TP / (TP+FN)

- Valor de predicción positiva (*Precision*). Probabilidad de que, dada una detección positiva, la realidad sea una etiqueta positiva también.

  Precision= TP / (TP+FP)

- Exactitud (*Accuracy*). Porcentaje total de aciertos del modelo.

  Accuracy = (TP+TN) / (TP+TN+FP+FN)

- F1-score. Media de precisión que determina un valor único ponderado de las métricas *precision* y *recall*.

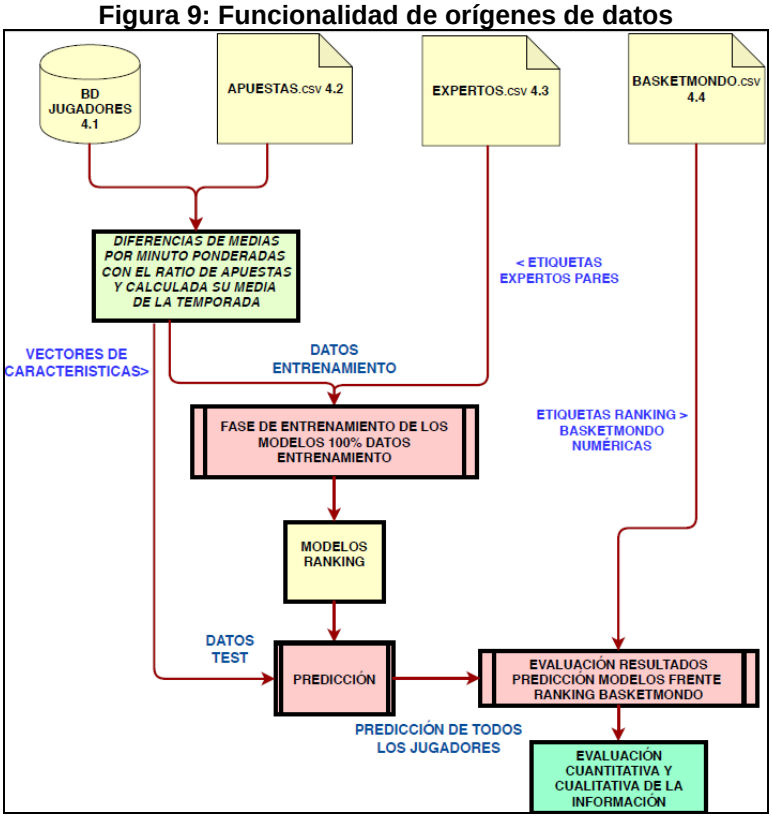
  F1-*score* = 2*((Precision * Recall) / (Precision + Recall))

- AUC. Significa área debajo de la curva ROC, siendo esta la gráfica que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación. Es una representación de dos parámetros, las tasas de verdaderos positivos y la de falsos positivos.

4. Bases de datos.

Lo primero que planteamos es un esquema general de la función de cada origen de datos y que sirva para su comprensión en detalle planteada en los siguiente subapartados. El esquema general de las funciones de cada origen de datos dentro del proyecto se representa en la figura 9:

Figura 9: Funcionalidad de orígenes de datos



Como podemos observar, la información proveniente de los datos de jugadores y de las apuestas de utiliza como vector de atributos para entrenar el modelo. Las etiquetas de los expertos se utilizan así mismo para el entrenamiento del modelo. Y por último, la información de BasketMondo se utiliza para la evaluación de los resultados obtenidos por los diferentes modelos de clasificación.

A continuación se presenta una descripción detallada de cada uno de los orígenes de datos.

4.1. Base de datos jugador por partido para una temporada regular - JUGADORES

La principal fuente del estudio proviene de la página oficial de estadísticas de la nba: https://stats.nba.com/. Los datos se han obtenido utilizando un entorno de pruebas de software para aplicaciones basadas en la web, llamado *Selenium.*

La necesidad de obtener la información de todos los partidos de la liga regular con estadísticas por jugador coincide con el acceso a la web y más en concreto a los apartados "*Teams/Box Scores*". Una vez que podemos ver los datos, se pueden realizar filtros y, en concreto, habrá que utilizar el filtro "*Season Type*" (tipo de temporada) e introducir "*Regular Season*" (temporada regular).

A continuación, en la Tabla 3, se describen en inglés y castellano los atributos que describen el desempeño de los jugadores que se han recogido vía técnicas *web scrapping*:

Tabla 3: Descripción de atributos del desempeño de JUGADORES

Atributo	Descripción Castellano	Descripción Inglés
min	minutos jugados	Minutes Played
pts_weight_A_B	Puntos anotados	Points
fgm_weight_A_B	Tiros de campo lanzados (2 Pts. + 3 Pts.)	Field Goals Made
fga_weight_A_B	Tiros de campo anotados (2 Pts. + 3 Pts.)	Field Goals Attempted
fg%_weight_A_B	Porcentaje de tiros de campo anotados (2 pts. + 3 pts.)	Field Goal Percentage
3pm_weight_A_B	Tiros de 3 pts. anotados	3 Point Field Goals Made
3pa_weight_A_B	Tiros de 3 pts. lanzados	3 Point Field Goals Attempted
3p%_weight_A_B	Porcentaje de tiros de 3 pts. anotados	3 Point Field Goals Percentage
ftm_weight_A_B	Tiros libres anotados	Free Throws Made
fta_weight_A_B	Tiros libres lanzados	Free Throws Attempted
ft%_weight_A_B	Porcentaje de tiros libres anotados	Free Throw Percentage
oreb_weight_A_B	Rebotes ofensivos	Offensive Rebounds
dreb_weight_A_B	Rebotes defensivos	Defensive Rebounds
reb_weight_A_B	Rebotes totales (Def. + Ofe.)	Rebounds
ast_weight_A_B	Asistencias	Assists
stl_weight_A_B	Robos	Steals
blk_weight_A_B	Tapones	Blocks
tov_weight_A_B	Recuperaciones	Turnovers
pf_weight_A_B	Faltas personales	Personal Fouls
+/-_weight_A_B	Más-Menos.	Plus-Minus

El atributo "Más-Menos" contiene el número de puntos que ha anotado el equipo menos los puntos recibidos estando el jugador en cancha.

Para una temporada, se dispone de información sobre 512 jugadores y sobre su desempeño en los 82 partidos que componen la temporada.

4.2. Fichero de datos de apuestas deportivas por partido - APUESTAS

Por otro lado, para realizar la ponderación de las estadísticas por partido, se ha incorporado un fichero de apuestas deportivas, en el que se recogen los índices americanos de diferentes casas de apuestas. Con la media de dichos

valores, se saca una media por partido que utilizaremos para ponderar la importancia del partido dentro de la liga.

El *data set* "*NBA Historical Stats and Betting Data*" proviene de la plataforma Kaggle (2018). De dicho enlace se descargará el archivo "nba-historical-stats-and-betting-data.zip" que contiene diferentes archivos .csv. En nuestro caso, combinaremos los archivos "nba_betting_money_line.csv" y "nba_games_all.csv".

La Tabla 4 contiene la descripción de cada uno de los atributos necesarios para realizar la ponderación y fusión de ambos archivos con informaciones de apuestas:

Tabla 4: Descripción de atributos de datos de APUESTAS

nba_betting_money_line.csv		*nba_games_all.csv*	
Atributos	**Descripción**	**Atributos**	**Descripción**
game_id	Identificativo del partido	**game_id**	Identificativo del partido
price1	Precio apuestas equipo 1	matchup	*String* de 3 caracteres con el equipo 1 y 2
price2	Precio apuestas equipo 2	is_home	Indicador primer equipo (1 - juega en casa 0 - juega fuera de casa)

El cruce de la información de ambas tablas se realiza por identificativo de partido (game_id). El campo *matchup* contiene un *string* formado por las iniciales de tres caracteres alfabéticos de los equipos de la NBA con dos formatos:

- HOU vs. UTA
- TOR @ WAS

Con la descomposición de los mismos, conseguimos los nombres de los equipos. Debe realizarse un filtro, ya que los partidos están repetidos en orden inverso; es decir, existen apuestas para los partidos "HOU vs. UTA", pero también para los "UTA vs. HOU" de la misma fecha. Es aquí donde hay que utilizar el campo *is_home* para poder filtrar solo uno de los dos partidos repetidos.

Otra de las cosas importantes a tener en cuenta es el formato de las apuestas contenido en los campos *price1* y *price2*. Dichas apuestas están en formato americano y por lo tanto tendrán que ser transformadas al estándar decimal (europeo).

Para una temporada se disponen de datos de los 82 partidos que jugaron las diferentes franquicias durante la temporada regular.

4.3. Encuestas de comparación de desempeño de jugadores NBA a expertos entrenadores con título superior - EXPERTOS

Se trata de un fichero .csv con una estructura simple, consistente en tres campos:

- *PlayerA*: Jugador A.
- *PlayerB*: Jugador B.
- Mejor (1=jugador A mejor desempeño que jugador B, 0= jugador A mejor desempeño que jugador B)
- Encuestador: *String* de tres caracteres con iniciales de nombre, 1er apellido, 2º apellido.

El archivo consiste en información sobre 639 comparaciones entre pares de jugadores teniendo en cuenta su desempeño. Se trata de la etiqueta de los expertos que constituye información sobre 1278 jugadores (comparaciones x 2 jugadores), lo cual supone tres comparaciones por jugador para cada uno de los 426 jugadores. Se ha hecho coincidir con el número de jugadores que juegan más de 10 minutos y que se explicará posteriormente en el apartado correspondiente al preprocesado de información.

La información de los expertos servirá para etiquetar la información transformada de los jugadores en la fase de entrenamiento.

4.4. JSON Precios final temporada jugadores en aplicación BasketMondo - BASKETMONDO

El fichero JSON de los precios de los jugadores a final de temporada se ha obtenido por cortesía de BasketMondo[3]. Se trata de una aplicación de ocio, en la que los diferentes usuarios tienen un presupuesto para fichar jugadores y cuya puntuación al final se basa en las estádisticas del apartado 4.1 al final de cada jornada. El funcionamiento del juego consiste en ir fichando y vendiendo a jugadores para maximizar la puntuación acumulada por jornadas al final de la temporada regular.

Del fichero se seleecionan para el estudio los campos *fullname* (nombre completo del jugador), *actual_price* (precio ajustado de final de temporada teniendo en cuenta las veces que se ha comprado y vendido un jugador), y *c_price* (precio histórico del jugador).

Se filtran aquellos jugadores que tienen precio actual o histórico cuyo valor es igual al precio preestablecido por defecto por BasketMondo, ya que de otro modo no habría la posilibilidad de calcular el *ranking* final de Basketmondo de dichos jugadores.

Con dicha información se realizará la evaluación de las predicciones del modelo contruido para el estudio. 4.1.

[3] http://www.basketmondo.com/ Aplicación gratuita de ocio de gestion de franquicias NBA

5. Preprocesado de datos

5.1. Filtros sobre orígenes de datos

5.1.1. Filtros de datos sobre orígenes de datos JUGADORES.csv - *+10 minutos*

La base de datos del desempeño de jugadores por partido para una temporada regular (JUGADORES) recogida en el apartado 4.1 del documento, se filtra de forma que solo se tengan en cuenta los jugadores con más de 10 minutos en los partidos que hayan jugado durante la temporada regular. El resultado son 427 jugadores de los 512 inicales.

5.1.2. Filtros sobre orígenes de datos BASKETMONDO.csv – *Precio disponible*

La base de datos con los precios de los jugadores en BasketMondo (BASKETMONDO) recogida en el apartado 4.4 del documento, se filtra de forma que solo se tengan en cuenta los jugadores que tienen un precio actual o histórico diferente de los valores asignados por defecto por la aplicación. El resultado son 423 jugadores de los 512 inicales.

5.2. Preprocesado de datos

5.2.1. Preprocesado de datos proveniente del fichero APUESTAS.csv – *Apuestas decimal*

La base de datos con los índices de media de cinco casas de apuestas vienen expresados en el sistema americano de apuestas. Dichas apuestas están expresadas de la siguiente forma:

- *1er carácter*: **+** signica que la cantidad expresada a continuación se corresponde a la cantidad de dolarés que recibe el apostador por cada 100 dólares que apueste; **-** significa que la cantidad expresada a continuación se correponde a la cantidad de dolarés que debe apostar el apostador para recibir 100 dólares.
- *Cantidad* en dólares.

El sistema de apuestas decimal que se usa en Europa, Australia y Canadá consiste en un índice con el que se multiplica la cantidad apostada y se obtiene la cantidad que el apostador se llevará en caso de que se dé el resultado de la apuesta.

Por lo tanto, habrá que convertir las apuestas al formato decimal para poder calcular un ratio que nos sirva para ponderar la importancia de los partidos dependiendo del equipo contrario y del momento de la temporada regular que sea.

Para la convesión en formato decimal utilizaremos la siguiente fórmula para todos los precios:

- apuestas_decimal(-) = 100 + (100 / abs(row[price1]))
- apuestas_decimal(+) = 100 + row[price1]

Una vez calculados los índices decimales, el ratio de apuestas para ponderar los datos del desempeño se calculará mediante la fórmula:

- ratio_apuestas= ((apuestas_decimal_local - apuestas_decimal_visitante) / apuestas_decimal _visitante)

5.2.2. Preprocesado de datos de ponderación del desempeño en función del ratio de apuestas (*PONDERACIÓN*)

De los datos filtrados de jugadores para cada partido, se multiplican todos los atributos por el ratio de apuestas obtenido en el apartado anterior 5.2.1.

5.2.3. Preprocesado de datos para la normalización del desempeño por minuto y cálculo acumulado de la media de todos los partidos de cada uno de los atributos (*MEDIAS MINUTO*)

De los datos ponderados de jugadores para cada partido, se dividen todos los atributos por el número de minutos en cada partido para obtener las estádisticas en una escala que se pueda comparar.

Una vez calculados los atributos por minuto, se calculará la media de la temporada.

5.2.4. Preprocesado de datos de diferencias de atributos entre pares de jugadores (*DIFERENCIA DE PARES*) *(datos* training *encuestas)*

Partiendo del fichero de EXPERTOS.csv, se cruzará la información de los jugadores para coger las medias de sus atributos y se cálculará la diferencia entre pares de jugadores. La salida de este proceso compondrá los datos de entrada de los modelos.

5.2.5. Preprocesado de datos de diferencias de atributos entre pares de jugadores (*DIFERENCIA DE PARES*) *(datos* training *todos los jugadores)*

Partiendo de la lista de jugadores, se generarán pares de jugadores de forma que todos crucen con todos. La salida se cruzará con la información de los jugadores para coger las medias de sus atributos y se cálculará la diferencia entre pares de jugadores. La salida de este proceso compondrá los datos de test para la predicción de los modelos.

5.2.6. Preprocesado de datos de ajuste de parámetros

Se ejecutará este preprocesado y sus subapartados 30 veces como parte del ajuste de parámetros.

5.2.6.1. Preprocesado de datos para dividir en datos de entrenamiento (70%) - datos de test (30%)

Del total de registros provenientes del fichero EXPERTOS.csv en el que hemos añadido las diferencias de cada atributo entre los jugadores que forman el par, ahora se escogen un 70% para los datos de entrenamiento y un 30% para los datos de test.

5.2.6.2. Preprocesado de datos para la optimización de parámetros – *GridSearchCV*

Partiendo de los datos de entrenamiento, se entrena el modelo y se obtienen los resultados asciados a dicho entrenamiento. En el caso del presente, se calcularán las siguiente métricas:

- *Accuracy*
- *Recall*
- *Precision*
- F1
- AUC

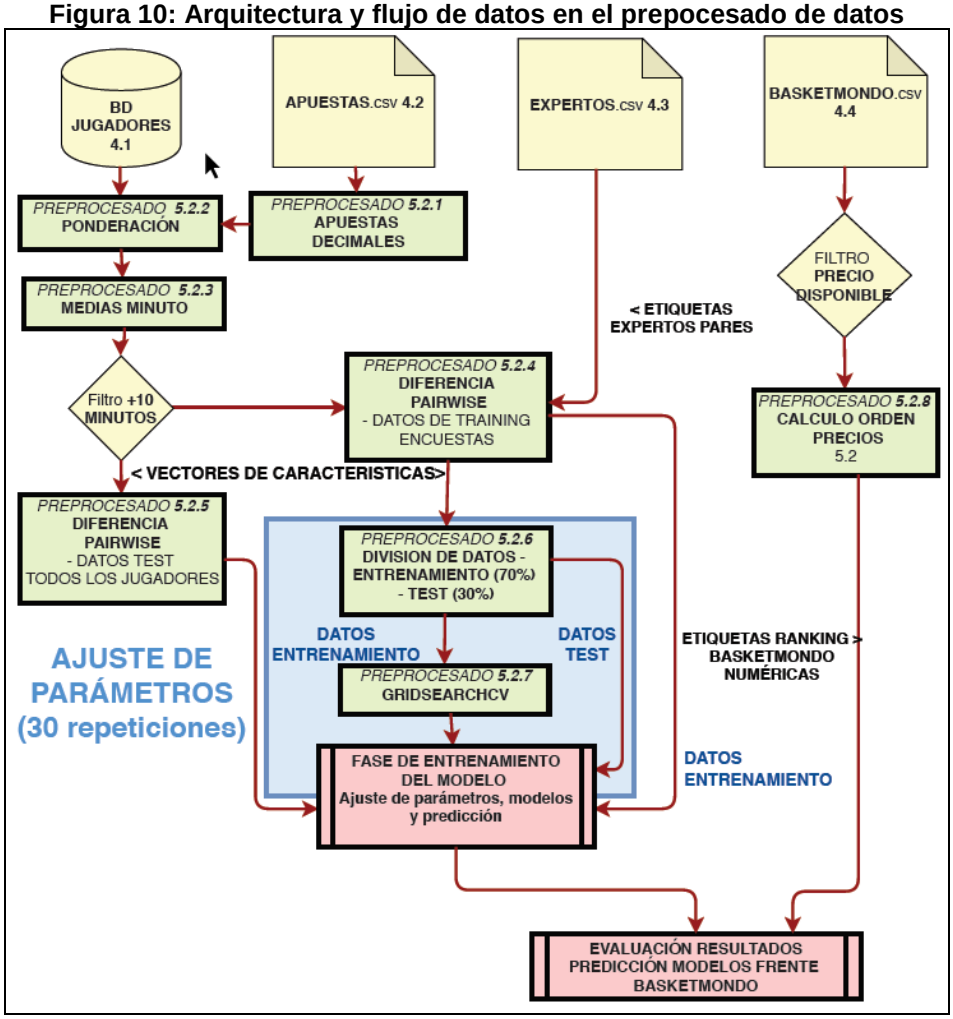
5.2.7. Arquitectura y flujo de datos en el reprocesado de datos

El la figura 10 podemos ver la arquitectura del prepocesado de datos y el flujo de los datos que servirán de entrada para construir los modelos con diferentes algoritmos.

En la figura podemos observar cómo se relacionan las diferentes fuentes de datos entre sí y los filtros existentes. En azul, podemos ver que los pasos 5.2.6 y 5.2.7 que forman parte del preprocesado de datos, están dentro del bucle para ajustar los parámetros del modelo. Por su parte, la fuente de datos de BasketMondo solo se utiliza en la evaluación del modelo.

Para el entrenamiento de los modelos del ajuste de parámetros, se cogerán todas las diferencias entre los desempeños de pares de jugadores que vienen etiquetados de las encuestas a expertos.

Figura 10: Arquitectura y flujo de datos en el prepocesado de datos



6. Diseño experimental

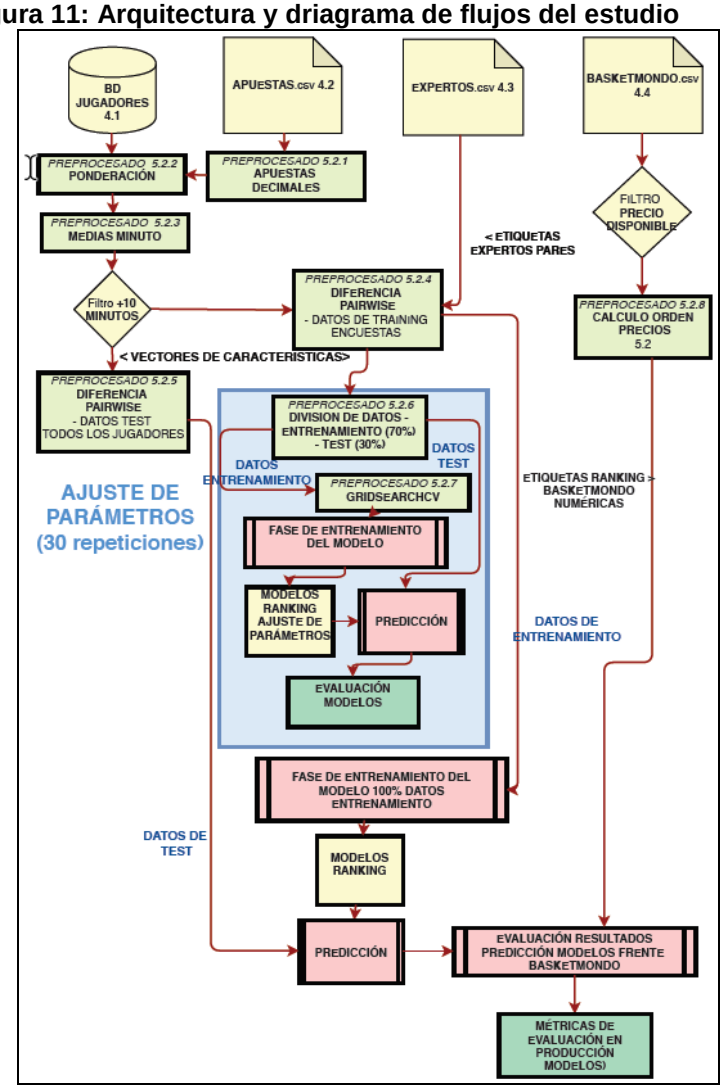
En primer lugar, conviene explicar que el diseño experimental se ha realizado para diferentes algoritmos de *learning to rank*, cuyas implementaciones podemos encontrar en la documentación de Scikit-learn.[4]

Los algoritmos escogidos debido a la naturaleza del problema son:

- **SVM** – Support Vector Machines – Máquinas de soporte vectorial.
- **KNN** – K Nearest Neighbors – K Vecinos más cercanos.
- **Random Forest** – Bosque de árboles.
- **Naive Bayes**.

Cada uno de los algoritmos será una implementación que seguirá la arquitectura y el diagrama de flujo definidos en el siguiente esquema general de la figura 11:

Figura 11: Arquitectura y driagrama de flujos del estudio



[4] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

A todos ellos se les aplica el mismo diseño, con algunas transformaciones extra debido a que algunas de las implementaciones de los diferentes algoritmos reciben tipos de datos diferentes las unas de las otras.

Además, en el caso de Naive Bayes no hay ajuste de parámetros, ya que el algoritmo no tiene parámetros que ajustar.

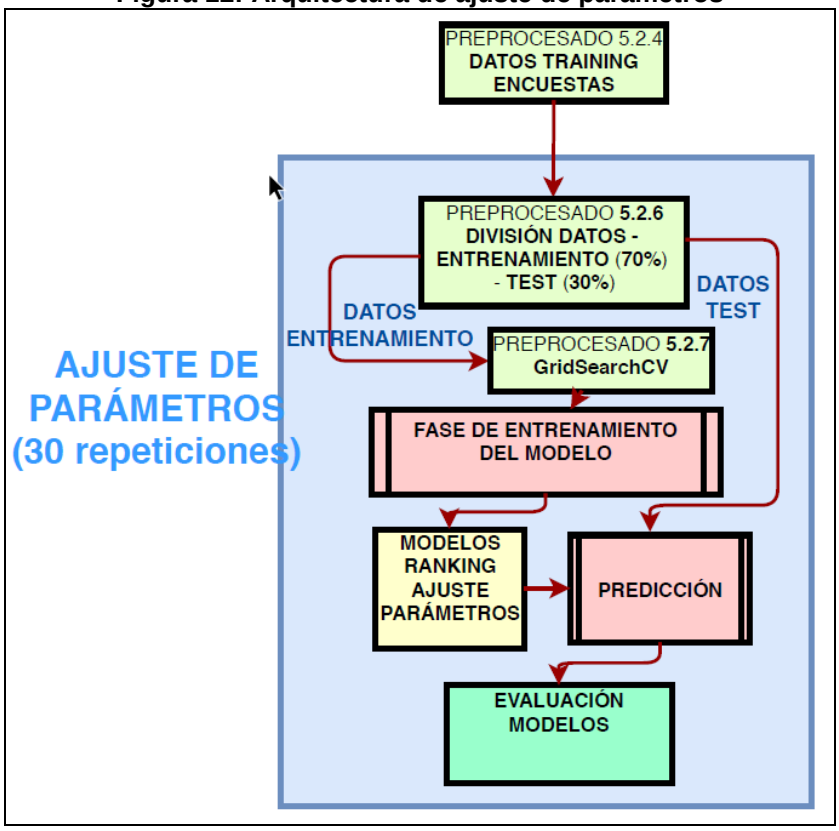
6.1. Ajuste de parámetros de los modelos

El proceso de ajuste de parámetros consistirá en la ejecución de 30 repeticiones de una serie de pasos que detallamos a continuación:

- Preprocesado de datos para modelos de ajuste de parámetros. Ya mencionado en el capítulo anterior.
- Entrenamiento del modelo, tomando el 70% de los datos de diferencias entre pares de jugadores con etiquetas provenientes de las encuestas.
- Predicción del 30% restante de datos de las encuestas que no se han utilizado para entrenar el modelo.

La arquitectura del ajuste de parámetros podemos observarla en la siguiente figura 12:

Figura 12: Arquitectura de ajuste de parámetros



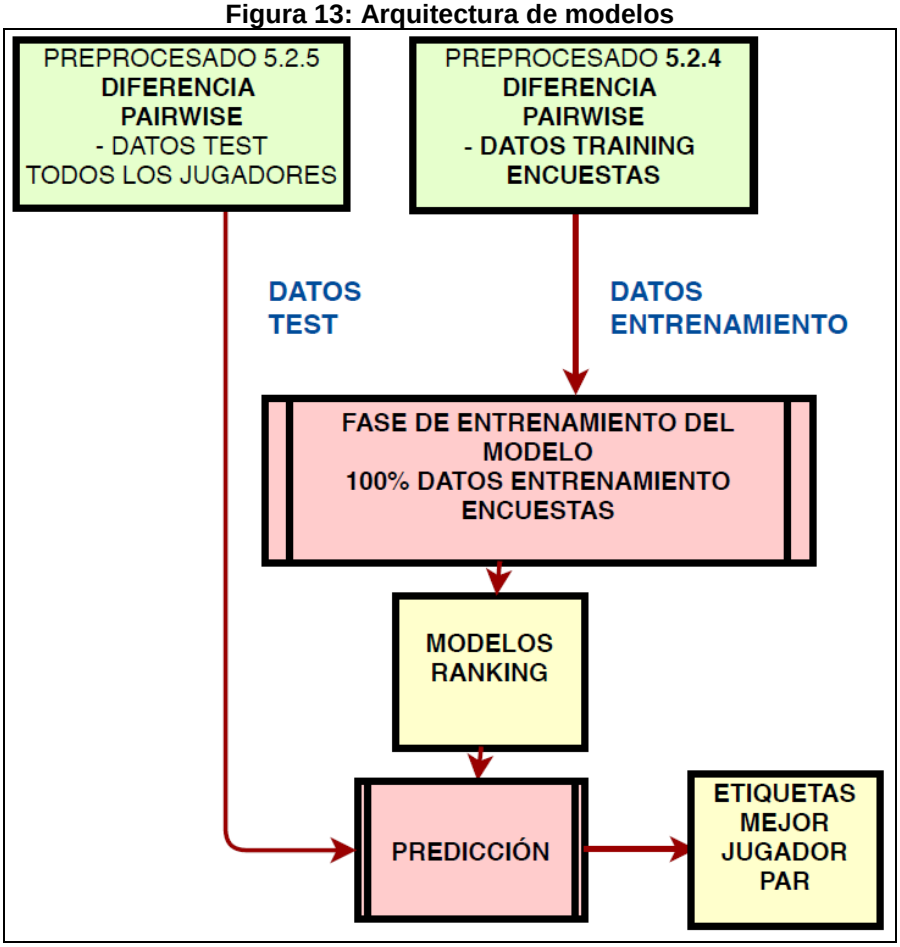
Durante las diferentes ejecuciones, se dispondrá el ajuste del parámetro Refit del GridSearchCV a Exactitud (Refit='Accuracy'), para que el modelo vaya recalculándose y encuentre los parámetros óptimos.

6.2. Diseño de modelos

Como datos de entrenamientos de los modelos, se utilizarán todos los datos de diferencias de jugadores con información etiquetada que proviene de las encuestas a expertos.

Se procederá a realizar el entrenamiento del modelo y con dicho modelo se realizará la predicción para las diferencias entre todos los jugadores, como podemos ver en la figura 13:

Figura 13: Arquitectura de modelos



Como datos de test, se cogerán todos los pares de jugadores de la lista de jugadores que haya salido del proceso de filtrado de 10 minutos. Serían 426 jugadores en pares = 181 476 registros. Tales pares estarán duplicados, por lo que, al quitar la mitad, quedarán 90 739 registros, sobre los que se realizará la predicción para obtener las etiquetas correspondientes y poder realizar el *ranking* de jugadores.

7. Evaluación de resultados

7.1. Evaluación de resultados en el ajuste de parámetros de los modelos

Como el diseño experimental se ha realizado para diferentes algoritmos de *learning to rank*, obtendremos las siguientes métricas ya definidas para cada uno de los diferentes algoritmos. Podemos observar los resultados en la tabla 5 para el ajuste de parámetros:

Tabla 5: Evaluación de resultados en ajuste de parámetros

	Ajuste Parámetros (30 véces)		
	Accuracy	**RMSE**	***Ejecución (seg.)***
SVM	0.743568	0.253398	170
KNN	0.687413	0.349190	315
Naive Bayes	NO		
Random Forest	0.703596	0.307119	593

Podemos observar cómo SVM consigue los mejores resultados en cuanto a la exactitud media de las ejecuciones (*Accuracy*) y en cuanto al error cuadrático medio (RMSE).

En cuanto al coste en tiempo de los diferentes algoritmos, cabe destacar que que Random Forest penaliza mucho su rendimiento frente al resto de algoritmos. Con un juego de datos de entrenamiento superior, habría que tenerlo en cuenta.

7.2. Evaluación de resultados de los modelos frente a la realidad BASKETMONDO

Para la evaluación de los resultados del *ranking* de nuestro modelo frente al *ranking* en función de los precios de BasketMondo, se realizarán comprobaciones a nivel cuantitativo y cualitativo.

7.2.1. Rendimiento cuantitativo

Para medir el rendimiento cuantitativo de los diferentes modelos y teniendo en cuenta la naturaleza de la correlación, que es un *ranking* con empates, se han establecido las siguientes métricas:

- *Spearman's Rank-Order Correlation.* Es una medida no paramétrica de correlación de *rankings*. Expresa cómo de bien la relación entre dos atributos puede hacerse usando una función monotónica.

- *Kendall rank correlation coefficient.* Es una medida estadística que mide la asociación ordinal entre dos cantidades ya medidas. Para ello utiliza el cálculo de la propia variable tau.

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

Los resultados para cada algoritmo los podemos observar a continuación en la tabla 6 de evaluación de resultados:

Tabla 6: Evaluación de resultados BasketMondo

	METRICAS CORRELACIÓN BASKETMONDO		
		KENDALL	
	SPEARMAN´S	TAU	P_value
SVM	-0.639	-0.445686	7.10465e-32
KNN	-0.409	-0.287009	3.84359e-14
Naive Bayes	NO		
Random Forest	-0.510	-0.358429	3.48362e-21

De los resultados podemos concluir que ambos *rankings* (modelo de encuentas y BasketMondo) están correlacionados y además de una forma negativa, lo cual indicaría que están ordenados en sentido inverso el uno del otro.

7.2.2. Rendimiento cualitativo

Para medir el rendimiento cualitativo de los diferentes modelos y teniendo en cuenta que la evaluación se realiza con el *ranking* de precios de mercado que nos suministró BasketMondo, que es un *ranking* con empates, se han establecido las siguientes métricas:

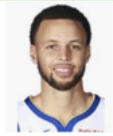
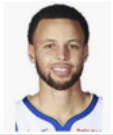
- Conteo por jugador del número de pares que el jugador gana y que se han detectado como resultado de la predicción de los modelos desarrollados en pasos anteriores.

Con los cinco primeros jugadores de dichos conteos y del *ranking* de BasketMondo se ha construido la tabla 7 de resultados cualitativos de los modelos frente al *ranking* de BasketMondo.

Podemos concluir que hay resultados muy similares y vemos cómo unos cuantos jugadores coinciden con BasketMondo a pesar de la ponderación con el ratio de apuestas que se ha realizado sobre el modelo.

El detalle de las conclusiones discutidas con algunos de los expertos encuestados se incluye a continuación en el apartado de discusión de resultados.

Tabla 7: Evaluación de resultados cualitativos BasketMondo

SVM Ranking			KNN			Random Forest			BasketMondo		
1		Name: **Montrezl Harrell** Position: PF / C (BM:50) **Age**: 25 **Height**: 2.03 m Weight: 109 kg **Salary**: 6 million USD	1		Name: **James Harden** Position: SG / PG Age: 29 Height: 1.96 m Weight: 99.8 kg Salary: 28.3 million USD	1		Name: **Josh Jackson** Position: SF / SG (RBM=170) Age: 25 Height: 2.03 m Weight: 91 kg Salary: 6.063 million USD	1		Name: **Stephen Curry** Position: PG Age: 31 Height: 1.91 m Weight: 86 kg Salary: 37.46 million USD
2		Name: **DeAndre Jordan** Position: C (RBM=41) Age: 30 Height: 2.11 m Weight: 120 kg Salary: 21.17 million USD	2		Name: **Chris Paul** Position: PG (RBM=24) **Age**: 34 Height: 1.83 m Weight: 79 kg **Salary**: 24.27 million USD	2		Name: **Greg Monroe** Position: PF / C (RBM=379) Age: 29 Height: 2.11 m Weight: 120 kg Salary: 16.41 million USD	2		Name: **James Harden** Position: SG / PG Age: 29 Height: 1.96 m Weight: 99.8 kg Salary: 28.3 million USD
3		Name: **Joel Embiid** Position: PF / C **Age**: 25 Height: 2.13 m Weight: 113 kg **Salary**: 6.1 million USD	3		Name: **Kevin Durant** Position: SF Age: 30 Height: 2.06 m Weight: 109 kg Salary: 26.54 million USD	3		Name: **Jared Dudley** Position: SF /PF (RBM=255) **Age**: 33 Height: 2.01 m Weight: 108 kg Salary: 10.47 million USD	3		Name: **Giannis Antetokounmpo** Position: F / G Age: 24 Height: 2.11 m Weight: 110 kg Salary: 24.16 million USD
4		Name: **Ben Simmons** Position: PG / F (RBM=13) **Age**: 25 Height: 2.08 m Weight: 104 kg **Salary**: 5.903 million USD	4		Name: **Eric Gordon** Position: SG (RBM=109) Age: 30 Height: 1.93 m Weight: 98 kg Salary: 12.39 million USD	4		Name: **Chris Paul** Position: PG (RBM=24) **Age**: 34 Height: 1.83 m Weight: 79 kg **Salary**: 24.27 million USD	4		Name: **Kevin Durant** Position: SF Age: 30 Height: 2.06 m Weight: 109 kg Salary: 26.54 million USD
5		Name: **Josh Jackson** Position: SF / SG (RBM=170) **Age**: 25 Height: 2.03 m Weight: 91 kg **Salary**: 6.063 million USD	5		Name: **Stephen Curry** Position: PG Age: 31 Height: 1.91 m Weight: 86 kg Salary: 37.46 million USD	5		Name: **James Harden** Position: SG / PG Age: 29 Height: 1.96 m Weight: 99.8 kg Salary: 28.3 million USD	5		Name: **Joel Embiid** Position: PF / C Age: 25 Height: 2.13 m Weight: 113 kg Salary: 6.1 million USD

Position: G = PG = Guard (Base), SG = Shooting guard (Base anotador), F = SF = Forward (Alero), PF= Power Forward (Ala-Pívot), C = Center (Pívot)

8. Discusión de resultados

Dentro de este apartado me dispongo a explicar las conclusiones a las que he podido llegar una vez concluidos y estructurados los resultados del proyecto.

Para la interpretación del mismo, se ha incluido un código de colores. En cuanto a las **carácterísticas de cada jugador,** *rojo* significa que es un buen valor para ese atributo, *verde* significa que no lo es. En cuanto al **color de fondo del número en el *ranking***, *azul* significa que son fichajes potenciales por varios factores, y *amarillo* que son jugadores que, a pesar de que su desempeño es bueno, tienen otra característica (edad, sueldo, etc..) que hace que no sea recomendable su fichaje o que haya que estudiarlo con detenimiento. El *blanco* significa que coincide con el *ranking* de BasketMondo y por lo tanto sus resultados arrojan una correlación que se ajusta con los resultados de nuestro estudio.

El primer paso ha sido compartir la tabla 7 de resultados cualitativos del *top* 5 que se ha mostrado en el apartado anterior con algunos de los expertos.

Las conclusiones de mi interpretación de los resultados por jugador del *top* 5 en mi calidad de entrenador superior y estudiante del mastér de ciencia de datos son:

1. **Montrezl Harrell** – Solo aparece en el *ranking* SVM, en primera posición. Jugador joven, explosivo y con mucha resistencia física. Uno de los jugadores más intensos de la liga. Tal vez no tenga más oportunidades de fichar como estrella de ningún equipo grande debido a que su estatura de 2,03 m. esta bastante por debajo de la media en la NBA hoy en día. Su desempeño esta fuera de toda duda y, si sigue en esa línea, se convertirá sin duda en un gran 6º hombre de cualquier franquicia que pueda optar al anillo de la NBA. No hay duda de que es un fichaje a explorar, porque además es lo bastante joven como para poder mejorar su rango de tiro. Destacar además que su *ranking* en BasketMondo es de 50, como podemos ver al final del campo *position* (RBM = 50)[5]. Los usuarios pueden estar viendo solo su presencia en el campo, pero la intensidad y explosividad con las que juega le pueden llevar a un desempeño muy destacable.

2. **DeAndre Jordan** – Solo aparece en el *ranking* SVM. Es un jugador en el final de su carrera y su salario ya es alto porque es un jugador que sabe sacar partido de su experiencia en el juego para maximizar sus resultados. Seguramente además está interesado en hacer buenos partidos con equipos importantes para poder conseguir algún cambio de franquicia que le pueda ser provechoso en el final de su carrera, bien sea para ganar títulos o bien para firmar un buen contrato como 6º hombre de una franquicia importante. Es un jugador interesante pero

[5] Se incluye leyenda en campo position para los jugadores que no están en el top 5 de Basketmondo.

estará involucrado en carámbolas de cambios de jugadores y derechos de Draft de la NBA (elección de jugadores nuevos en la liga). En el *ranking* de BasketMondo, sale en la posición 41. Sin duda le beneficia que es pívot y que su experiencia le hace sacar el máximo partido a su esfuerzo.

3. **Joel Embiid** – Aparece en uno de los *rankings* y en BasketMondo. Es una de las coincidencias en la evaluación de resultados. Es uno de los pívots con más proyección de la liga y uno de los jugadores con más dobles-dobles (doble dígito en cualquier atributo estadístico de los recogidos en el estudio). Es más que una promesa y sin duda firmará uno de los grandes contratos con alguna franquicia de la NBA. Ciertamente, si se está en disposición económica de ficharlo, se recomienda el estudio de su fichaje.

4. **Ben Simmons** – Aparece también solo en uno de los *rankings*. Es un jugador de características muy especiales para su puesto. Es un base, joven, de 2,08 m., explosivo, finalizador cerca del aro, no tanto de tiro exterior, y buen reboteador. Es otro jugador que compite en dobles-dobles con Embiid, aunque mucho menos rodado e irregular en sus actuaciones. Es una apuesta de futuro para ser base titular y firmar un gran contrato con cualquier franquicia ganadora de la NBA. Aún no ha terminado de explotar para demostrar su fiabilidad de cara a un gran contrato pero, a nada que mejore su tiro exterior, puede convertirse en una de las refencias de la liga. Es un jugador cuyo fichaje definitivamente merece la pena explorar.

5. **Josh Jackson** – Aparece en el *top* 5 de dos de los *rankings*. No obstante, su posición en el *ranking* de BasketMondo es una de las más bajas de este *top* 5, el 170. Seguramente esto es debido a que está en una de las franquicias menos atractivas de la liga y a su juventud, ya que nos encontramos con un jugador de rachas. Parece que su desempeño, desparpajo y motivación a la hora de competir contra los grandes equipos hace que aparezca en esta clasificación. Sin duda, le ha beneficiado la ponderación de las apuestas, pero por el salto tan grande que ha dado en la clasificación (seguro que debido también a su conocimiento del juego y su explosividad), debe ser explorado para un posible fichaje.

6. **James Harden** – Gran extrella de la NBA. Indudablemente, uno de los mejores jugadores de la liga por su regularidad en las últimas temporadas. Anotador incontrolable, le perjudica el hecho de pertenecer a los equipos que acaban la liga regular en buena posición. Por lo tanto, el resultado de la ponderación con el ratio de apuestas le perjudica. Aun así, sale en dos de los *rankings*. Es uno de los resultados positivos esperados con el *top* 5 de BasketMondo.

7. **Chris Paul** – Otra de las grandes estrellas en los últimos años. En este caso se encuentra en el final de su carrera y eso le pesa en su

clasificación de BasketMondo (24). La novedad es algo que afecta al realizar nuevos fichajes en BasketMondo a la hora de mejorar la clasificación total. Este jugador, al tener un precio caro y estar haciendo menos puntos en los totales acumulados y medias por partido en las últimas temporadas, se percibe como en tendencia negativa de cara a sumar puntos en la clasificación de BasketMondo (suma puntos a cada jugador por hecho estadístico considerado). Sin embargo, desde el punto de vista de fichajes, se trata de un base anotador en el final de su carrera, pero que sabe contemporizar para estar en los partidos importantes (definido por uno de los expertos como ”jugador que cuando toca estar, siempre está”). Uno de los jugadores más veteranos y con mejor conocimiento del juego que hay en la liga.

8. **Kevin Durant** – Sin duda otra de las estrellas que coincide en los resultados de un modelo y BasketMondo. Jugador determinante en la liga. Anotador, con un físico envidiable, polivalente y con una inteligencia en el juego fuera de toda duda. Y además le ayuda que es otro competidor de la liga de dobles-dobles, con rendimientos en puntos y rebotes muy destacados dentro de la liga.

9. **Eric Gordon** – Base veterano, anotador y con experiencia. Aparece en uno de los *rankings* a pesar de que en BasketMondo solo ocupa la posición 109. No es jugador atractivo de cara a liderar un equipo ganador. Es otro perfil de jugador de complemento que en un momento dado podría entrar de titular en un equipo ganador y al que su experiencia le ayuda a la hora de maximizar su desempeño. Jugador inteligente, es de los jugadores que casi nunca resta.

10. **Stephen Curry** – Estrella que ha revolucionado la liga. A su alrededor se ha formado el equipo que está dominando la NBA en los últimos años. Base anotador, especialmente desde larga distancia, ya que su físico no le permite prodigarse mucho cerca del aro. Muy hábil e inteligente, es capaz de generar opciones para los compañeros. Otra coincidencia con BasketMondo que muestra la correlación de los *rankings*.

11. **Greg Monroe** – Pívot atlético, explosivo y veterano, en el culmen de su carrera. Jugador que ya ha arrojado buenos datos estadísticos en temporadas pasadas. Claro beneficiado de su entendimiento del juego y del tipo de atributos que se utilizan en la medida del desempeño que son atributos en su mayoría de carácter ofensivo (tiros de 2, 3 y libres, rebotes, asistencias). No le gusta defender y es por eso que tanto su ranking de BasketMondo (379), como sus minutos totales jugados estén muy por debajo que su desempeño ofensivo. Jugador a estudiar para rotaciones, su veteranía en un entorno motivador(franquicia que quiera ganar la liga) como hombre interior de recambio sería muy atractiva, especialmente si se implica en labores defensivas (tiene capacidad).

12. **Jared Dudley** – Jugador intenso, veterano, que salta a la cancha desde el banquillo para dar recambio a jugadores más jóvenes. Su capacidad a

la hora de leer el juego le permite analizar las situaciones que se están dando en el partido antes de entrar y maximizar sus actuaciones para ayudar al equipo. Su ranking en BasketMondo (255) está muy por debajo del desempeño obtenido de uno de los modelos. Es probable que sea por el carácter del jugador, con cierta tendencia a verse involucrado en peleas dentro de la NBA. Se recomienda explorar su contratación, más allá de que sea supeditada a claúsulas para minimizar los riesgos de contratación.

Como discusiones del modelo en general y del proceso de aprendizaje cabría descatar:

- Existen coincidencias importante y correlación entre los resultados de los modelos y el *ranking* de BasketMondo. Stephen Curry, James Harden, Kevin Durant, Joel Embiid. 80% de acierto, todos excepto Giannis Antetokounmpo. Y eso que solo comparamos en detalle 5 de los 426 jugadores que componen el estudio, aunque eso si, para 3 modelos.

- El modelo entrenado con una mayoría de atributos ofensivos, beneficia a los pívots. Se trata de un problema en la recogida de datos en el baloncesto, que está muy ligada tambien a las estrategias de marketing. Los dobles-dobles son mucho más fáciles por los jugadores con buenas condiciones físicas. 50% de jugadores resultantes de los modelos son altos (+2,08m.) y atléticos.

- La estrategia de ponderar del ratio de apuestas (mide la importancia del rival y la importancia del partido, dentro de la temporada), ha resultado efectiva, ya que hay coincidencia de jugadores se han destacado en el *ranking* por su capacidad para jugar bien contra los equipos buenos (Eric Gordon, Chris Paul, Greg Monroe, DeAndre Jordan, etc..). Más allá de que haya que ajustar el ratio en sucesivas iteraciones del proceso para ir mejorandolo.

- En BasketMondo se premia la regularidad. Al tratarse de un sistema simple de conteo de puntuación y haber un mercado de fichajes en el que las estrellas son muy caras (todo el mundo los quiere), una de las estrategias ganadoras es la de maximizar la regularidad de los jugadores de banquillo. Además tiene más en cuenta el desempeño anual acumulado o por épocas, que el desempeño por minuto que hemos utilizado en el estudio para poder comparar jugadores.

- Por último, el hecho de que la correlación de los modelos sea negativa está influido por la ponderación de los ratios de apuestas. Creo que es un indicio que habrá que investigar de que esa ponderación sea muy grande y haya que ir ajustándola para mejorar el modelo.

Para terminar, destacar que más allá de la bondad del modelo, se ha conseguido avanzar en la construcción y entendimiento un modelo que tuviera en cuenta los equipos contrarios y la importancia del partido y además se ha

conseguido coincidencias importantes frente a un conocimiento estructurado de dichos parámetros aportado por BasketMondo, objetivos principales del proyecto.

9. Conclusiones

El estudio realizado ha demostrado la ejecución de un sistema de *learning to rank* aplicado al análisis avanzado del desempeño de jugadores en la NBA y apoyado en la información que suministraban los modelos basados en tres implementaciones de algoritmos de clasificación.

Los resultados obtenidos a nivel cualitativo han sido muy aceptables. Sin embargo, se puede mejorar el proceso añadiendo más fuentes de información y ajustando el proceso de aprendizaje.

Las líneas futuras en cuanto a añadir más fuentes de información para seguir evolucionando el conocimiento en esta materia podrían incluir las siguientes:

- Información estadística histórica.
- Añadir atributos defensivos.
- Añadir información de lesiones.
- Añadir información de cómo interactúan unas estadísticas con las estadísticas del resto del componentes del equipo.
- Información sobre el posicionamiento de los jugadores.
- Añadiendo más volumen de información etiquetada para el entrenamiento del modelo.

El ajuste del proceso de aprendizaje podría venir por el ajuste de la ponderación del ratio de apuestas en función de los calculos de la correlación del ranking resultante de los modelos frente al ranking de BasketMondo.

10. Bibliografía.

[1] Adam Gonzalez, Jay Hoffman, Joseph Rogowski, William Burgos, Edwin Manalo, Keon Weise, Maren Fragala, Jeffrey Stout (2013) Performance Changes in NBA Basketball Players Vary in Starters vs. Nonstarters Over a Competitive Season. [Fecha de consulta 18 de Marzo de 2019]

[2] Sergio J. Ibáez, Jaime Sampaio, Sebastian Feu, Alberto Lorenzo, Miguel A. Gómez & Enrique Ortega (2008) *Basketball game-related statistics that discriminate between teams' season-long success*, European Journal of Sport Science, 8:6, 369-372.

[3] Javier García, Sergio J. Ibañez, Ra仔 Martinez De Santos, Nuno Leite, Jaime Sampaio (2013) Identifying Basketball Performance Indicators in Regular Season and Playoff Games, Journal of Human Kinetics volume 36/2013, 161-168.

[4] Anguera, M.T., Hernández-Mendo, A. La metodología observacional en el ámbito del deporte Revista de Ciencias del Deporte, 2013. 9 (3).

[5] Anguera MT Blanco A, Losada JA., Hernández A (2000) La metodología observacional en el deporte: conceptos básicos. Revista Digital - Buenos Aires 2000 - Año 5 - N°24 Disponible en: http://www.efdeportes.com/ (Fecha de consulta junio 2014)

[6] Aitor Antía Martín-Carrillo, Ignacio Refoyo Román (2014) El scouting como método de observación en el baloncesto. Trabajo de Fin de Grado en ciencias del deporte de la Universidad Politécnica de Madrid.

[7] Bishop D. (2003) Performance analysis: What is performance analysis, and how can it be integrated within the coaching process to benefit performance? Peak Performance 4-7. Available from http://www.pponline.co.uk/encyc/540 sports-performance-analysis-coaching-and-training-39.

[8] Ofoghi, B., Zeleznikow, J., MacMahon, C., & Raab, M. (2013). Data mining in elite sports: a review and a framework. Measurement in Physical Education and Exercise Science, 17(3), 171-186. https://www.tandfonline.com/doi/abs/10.1080/1091367X.2013.805137

[9] Bhandari, I., Colet, E., Parker, J. et al (1997) Data Mining and Knowledge Discovery. 1:121. https://doi.org/10.1023/A:1009782106822

[10] Sampaio J, McGarry T, Calleja-González J, Jiménez Saiz S, Schelling i del Alcazar X, et al. (2015) Exploring Game Performance in the National Basketball Association Using Player Tracking Data. PLOS ONE 10(7): e0132894.

[11] S. Trninic and D. Dizdar (1999) System of the Performance Evaluation Criteria Weighted per Positions in the Basketball Game. Coll. Antropol. 24 (2000) 1: 217–234

[12] S. Trninic, D. Dizdar & B.Dezman (2000) Empirical Verification of the Weighted System of Criteria for the Elite Basketball Players Quality Evaluation. Coll. Antropol. 24 (2000) 2: 443–465.

[13] Mangine, Gerald T.1; Hoffman, Jay R.1; Wells, Adam J. Gonzalez, Adam M., Rogowski, Joseph P., Townsend, Jeremy R., Jajtner, Adam R.1, Beyer, Kyle S.1, Bohner, Jonathan D.1, Pruna, Gabriel J.1, Fragala, Maren S., Stout, Jeffrey R. (2014) Visual Tracking Speed Is Related to Basketball-Specific Measures of Performance in NBA Players. Journal of Strength and Conditioning Research: September 2014 - Volume 28 - Issue 9 - p 2406–2414

[14] Goldsberry K, Weiss E. The Dwight Effect: A New Ensemble of Interior Defense Analytics for the NBA. MIT Sloan Sports Analytics Conference 2012. 2012. [Fecha de consulta 18 de Marzo de 2019]

[15] Zhao, Bo. (2017). Web Scraping. 10.1007/978-3-319-32001-4_483-1.

[16] Chen, M., Mao, S. & Liu, Y. Mobile Netw Appl (2014) Big Data: A Survey 19: 171. https://doi.org/10.1007/s11036-013-0489-0

[17] Catarina Alexandra Pinto Moreira (2011) Learning to Rank Academic Experts Instituto Superior Técnico, Technical University of Lisbon, Portugal

[18] Fabian Pedregosa (2012) Learning to rank from medical imaging data. Disponible en: http://fa.bianp.net/archives.html [Fecha de consulta 27 de Mayo de 2019]