

TITULO: Un análisis estadístico del rendimiento de los jugadores de fútbol de La Liga en relación a sus salarios

Antonio Navarro Barrero



Resumen

Este proyecto tiene como objetivo principal el estudio de algunos métodos estadísticos en competiciones futbolísticas, centrándonos específicamente en los equipos de La Liga, principal competición a nivel nacional. Mediante tres métodos de predicción distintos, Regresión Lineal Múltiple, Árboles de Regresión y Método de k -NN vecinos, se pretenderá estudiar si existe relación entre distintas variables de rendimiento medidas sobre cada jugador frente al salario correspondiente de cada uno, analizando de esta manera si el valor de mercado por jugador es acorde a su nivel salarial. Para la construcción de estos modelos se han tomado datos sobre algunas variables de rendimiento para cada jugador a lo largo de las últimas 3 temporadas, con el objetivo de que sea un análisis actualizado sobre el valor de mercado actual de cada jugador.

Abstract

This project aims to study different statistical methods in football competitions, focusing specially on the teams of La Liga, the main national competition. Through three different prediction methods, Multiple Linear Regression, Regression Trees, and k-NN Nearest Neighbours, it will be pretended to investigate whether there is a relationship between different performance variables measured for each player and their corresponding salary. This analysis will help determine if the market value of a player matches with their salary level. For the construction of these models, data has been taken on some performance variables for each player over the last 3 seasons, with the aim of making it an updated analysis of the current market value of each player.

Índice

1 - Introducción	1
2 - Revisión bibliográfica	3
3 - Datos	6
4 - Modelos estadísticos utilizados	10
4.1 Regresión Lineal Múltiple	10
4.2 Árboles de regresión y Random Forest	12
4.3 Método de los k-vecinos más cercanos: KNN	13
5 - Resultados	15
5.1 Regresión Lineal Múltiple	15
5.2 Árboles de regresión y Random Forest	26
5.3 Método de los k-vecinos más cercanos: KNN	39
6 - Conclusiones	44
Bibliografía	45
Anexo 1 - Defensas	47
Anexo 2 - Mediocentros	48
Anexo 3 - Delanteros	53

1 - Introducción

A lo largo del siglo XXI, la sociedad ha sufrido una constante evolución en numerosos aspectos. Debido a esta evolución, se han ido implementando numerosas mejoras en los deportes en los últimos años gracias a este progreso, como se indica en Inuba (2022).

Por ello la elaboración de este trabajo, que consistirá en la realización de aplicaciones y métricas estadísticas en competiciones futbolísticas. Centraremos la atención en el deporte del fútbol, concretamente en la primera división de La Liga¹. De esta manera, se pretenderá entender de una forma mucho más clara, y visual en algunos casos, de cómo funcionan las distintas estadísticas que se emplean en este deporte. Por tanto, en este trabajo nos centraremos principalmente en el siguiente objetivo: Debido a las diferencias salariales entre los jugadores de un mismo equipo, se pretenderá estudiar cómo afectan distintas variables de rendimiento a la hora de explicar el salario de un jugador, con la intención de observar qué variables son las más importantes a la hora de explicar este salario, y a su vez, si por norma general existen diferencias entre los salarios dependiendo de la posición del jugador.

Se utilizará la página web de FBREF² para obtener las variables de rendimiento necesarias para construir los distintos modelos. Se trata de una página web de carácter estadístico deportivo, la cual abarca todo tipo de competiciones en distintos deportes. En nuestro caso, nos fijaremos únicamente en aquellas relacionadas con el fútbol, ya sea a nivel nacional o europeo.

Los salarios de cada jugador serán extraídos de Capology³, una página web que colabora con FBREF y que se especializa exclusivamente a la recopilación de datos sobre los contratos de cada jugador con cada equipo.

En el desarrollo de este trabajo se han construido diferentes modelos de predicción con una capacidad predictiva bastante buena. Los distintos modelos de predicción construidos comparten la característica de tener como una de las variables más importantes para determinar el salario de un jugador a la Popularidad.

1- La Liga: Hace referencia a la primera división de la liga española de fútbol. La Primera División de España o La Liga, es la máxima categoría del sistema de ligas de fútbol de España y la principal competición a nivel de clubes del país.

2- FBREF: Página web utilizada para obtener las variables de rendimiento de cada uno de los jugadores.
https://fbref.com/es/comps/12/Estadisticas-de-La-Liga#all_league_structure

3- Capology: Página web utilizada para obtener los salarios de todos los jugadores. <https://www.capology.com>
Como se explicará más adelante en la revisión bibliográfica, en Carrieri y otros (2018) se indica que a los directivos de los equipos de fútbol podría interesarles jugadores que tengan un nivel de popularidad elevado, con el objetivo de beneficiarse de este aspecto mediante campañas publicitarias. Tanto en los resultados que se obtuvieron en Carrieri y otros (2018), como en los obtenidos en este, la variable que mide el nivel de popularidad de un jugador está siempre presente como una de las variables más determinantes a la hora de determinar el salario o valor de mercado de un jugador.

Se han utilizado tres métodos de predicción distintos a lo largo del trabajo: la regresión lineal múltiple, los árboles de regresión o random forest y el método de *k*-NN vecinos. Estos dos primeros son los más utilizados en artículos científicos que realizan estudios similares, y por ende son los que mejores resultados ofrecen. Con estos dos primeros métodos se han construido modelos predictivos estables y con buena capacidad predictiva. Por otro lado, el método *k*-NN no ofrece resultados tan buenos como en los dos métodos anteriores, generando modelos con un error predictivo bastante superior.

Por norma general, las variables más importantes para determinar el salario de un jugador en los distintos modelos construidos son la Popularidad, el Equipo del jugador y los Pases en el Último Tercio del campo.

El principal problema encontrado a lo largo del desarrollo de este trabajo ha sido la recopilación de datos. Tanto las variables de rendimiento, el salario, el índice de popularidad y las posiciones de cada jugador han sido recogidas manualmente debido a la escasez de bases de datos que relacionen el salario frente al rendimiento. La mayoría de artículos que realizan estudios al respecto mencionan diversas páginas web que muestran el rendimiento por jugador, entre ellas FBREF, la utilizada en este trabajo, pero en todos ellos utilizan programas de Web Scraping. Estos programas se utilizan para minar información de páginas web, pero son de acceso privado luego es imposible acceder a ellos, y por otra parte requieren una compleja programación como para construirlos uno mismo. Por esta razón, mediante FBREF se han recogido jugador a jugador las variables de rendimiento deseadas. De igual forma que para el salario se usó Capology, Pages Views¹ se usará para el índice de popularidad y Transfermarkt² para las posiciones de cada jugador.

1- Pages Views: Página web utilizada para obtener el nivel de popularidad de todos los jugadores.
<https://pageviews.wmcloud.org/?project=en.wikipedia.org&platform=all-access&agent=user&redirects=0&start=2020-09-12&end=2023-04-24&pages=Cat|Dog>

2- Transfermarkt: página web utilizada para obtener las posiciones de juego de los jugadores en sus respectivos equipos. <https://www.transfermarkt.es>

2 - Revisión bibliográfica

Debido a la mejora tecnológica que comentábamos, el fútbol se ha adaptado recogiendo información para su propio beneficio. Tanto jugadores como entrenadores utilizan esta información para saber en qué aspectos destacan y en qué otros no rinden como quisieran, pero, si lo enfocamos a la hipótesis inicial ya planteada, como club, conocer el rendimiento de tus jugadores, junto con el salario de cada uno, te permite establecer una relación salario-rendimiento la cual se podría analizar para saber si tus jugadores están sobre o infravalorados. Por lo tanto, debido a la importancia a nivel mundial de las ligas europeas de fútbol, se han elaborado numerosos artículos estadísticos para el estudio de esta posible relación entre las dos variables mencionadas; y, como manera de conocimiento previo, se han seleccionado distintos artículos para entender mejor cómo estudian, qué métodos emplean y qué conclusiones obtienen los expertos en la materia.

Para comenzar, en Li y otros (2022), se pretende estudiar cómo influye en el salario distintas variables medidas sobre cada jugador. Para estudiar esta relación se aplican 2 métodos distintos. El modelo de regresión lineal múltiple se emplea como primer método, diferenciando 3 modelos distintos. El primer modelo, al cual llamaremos RLM1, únicamente tiene en cuenta las variables **Edad** y **Posición** como variables regresoras. El segundo modelo, RLM2 añade variables de rendimiento, como pueden ser **Goles**, **Asistencias**, **Regates** o **Bloqueos**, entre otras. Por último, para el tercer modelo, RLM3, se tienen en cuenta tanto las anteriores variables como los **Logros** de cada jugador, es decir, los resultados obtenidos a nivel de clubes en las distintas competiciones, creando una variable la cual pondera 3 competiciones distintas, y dependiendo del ranking que haya obtenido el club, tendrá una valoración que oscila del 0 al 10. El R^2 o coeficiente de determinación (proporción de varianza total explicada por la regresión) para estos tres modelos es de 0.481 para RLM1, 0.596 para RLM2 y de 0.606 para RLM3, lo cual, como indica el propio autor, se trata de un R^2 "mediocre". Por ello, pasamos al segundo método de predicción estadístico que se utiliza en el artículo, el Random Forest, el cual, a diferencia de los modelos anteriores, emplea algoritmos no lineales. Para este modelo de Random Forest se utilizarán las mismas variables empleadas para construir el RLM3. Como variables que más influyen sobre el salario están los mencionados Logros, Regates, Tiros, Goles o Bloqueos entre otras. En este modelo se obtuvo un valor de $R^2 = 0.948$, lo que nos indica que el modelo está altamente ajustado. Para finalizar el artículo, para los modelos RLM3 y RF, se realiza un análisis mediante intervalos de confianza para cada estimación del salario por jugador, para así poder comparar este intervalo salarial con el salario real, y determinar si un jugador está siendo bien valorado, sobre o infravalorado.

En Yaldo y otros (2017) se plantea como objetivo de estudio el no pagar en exceso a un jugador, porque podría generar conflictos dentro de la plantilla de un club. Esto es debido al siguiente motivo: se ha demostrado que la desigualdad salarial tiene un efecto negativo en el rendimiento de un equipo, ya que el rendimiento de un jugador "A" disminuye cuando la diferencia salarial frente a otro jugador "B" de similar rendimiento aumenta (Torgler y Schmidt, 2007). En el artículo nos encontramos con una base de datos de todos los jugadores en 8 ligas europeas. Aplicando distintos modelos de machine learning se obtienen distintas conclusiones. Por un lado, se

estudian a todos los jugadores en conjunto sacando en conclusión que las variables más importantes para determinar el rendimiento son: Capacidad de reacción, Finalización, Control de balón, etc. Por otra parte, se establece la hipótesis de que las variables de rendimiento son distintas dependiendo de la liga a la que pertenezca el jugador, obteniendo también una clasificación distinta por liga. Como ejemplo a destacar en la Liga española se tienen más en cuenta variables relacionadas con la habilidad y el control del balón, frente a la Bundesliga (liga alemana) que prefiere jugadores más completos físicamente (variables como la velocidad o la fuerza). Esta clasificación de influencia de variables por liga, podría ser de utilidad, por ejemplo, para los agentes de jugadores que quieran buscar un contrato adecuado para sus respectivos jugadores, ya que conociendo el rendimiento de sus propios jugadores se buscarían las ligas que mejor se adaptan a sus características.

En Carrieri y otros (2018) se nos ofrece otro punto de vista a la hora de realizar la selección de variables. En este artículo se añade una variable que mide el nivel de popularidad de cada jugador. Los autores explican que es de vital importancia añadir esta variable al estudio porque los dueños de los clubes pueden estar dispuestos a contratar jugadores con un índice de popularidad elevado con el fin de explotar su influencia con campañas de merchandising. Para ello, se emplea las búsquedas en Google Trends de cada jugador como índice de popularidad. Se utilizan modelos de regresión por cuantiles junto con regresiones por mínimos cuadrados. Como conclusión, el autor demuestra que el índice de popularidad afecta notablemente en la determinación del salario. Utilizando la regresión por cuantiles, se observa como a medida que aumenta el salario, aumenta de manera exponencial la influencia del índice de popularidad sobre el salario. Por consiguiente, los directivos deportivos estarían interesados en adquirir jugadores cuya popularidad sea elevada, al aumentar beneficios en sus campañas publicitarias, ventas de camisetas e incluso en el número de visualizaciones por partido en televisión.

Por ello, de la misma manera que emplean los autores en su artículo, se añadirá una variable a nuestra base de datos que nos indique el número de búsquedas que ha tenido un jugador en las últimas 3 temporadas.

El análisis del rendimiento deportivo y su relación con el salario no se limita únicamente al mundo del fútbol. En otros deportes, como el baloncesto, también se llevan a cabo estudios similares con el mismo objetivo de explicar esas remuneraciones en función del desempeño del jugador. Esto ocurre por ejemplo en la NBA (National Basket Association). En Zhao (2022), se construye un modelo de regresión lineal múltiple en el cual se explica el salario de los jugadores de la NBA en función de su rendimiento. Esto nos muestra que es crucial conocer las variables más influyentes en la determinación del salario de un jugador en cualquier deporte. Los estudios estadísticos realizados por los distintos autores mencionados no son solo aplicables al fútbol o al baloncesto, sino que también pueden extrapolados a otros deportes. Conocer estas variables influyentes y su impacto en la remuneración de los deportistas es fundamental para entender la dinámica económica de cualquier disciplina deportiva.

Por último, me gustaría resaltar un aspecto importante, a pesar de que no vaya a estudiarse en este trabajo. Existen artículos que analizan las capacidades ofensivas y defensivas de cada equipo, que de igual forma que en este trabajo se realiza recogiendo información sobre distintas variables de rendimiento sobre cada jugador. Gracias a estos estudios, surge la posibilidad de investigar el funcionamiento de las casas de apuestas, ya que, para determinar el resultado de un partido, se tienen en cuenta las capacidades de cada equipo. En estas casas de apuestas se estudia la relación entre el resultado de un partido y las características defensivas y ofensivas, combinadas con otras variables de rendimiento por equipo (como el promedio de goles o jugar como local o visitante, entre otras). Conocer este funcionamiento proporcionaría una perspectiva valiosa sobre cómo influyen estas características en la forma en que se pronostica y se apuesta en el fútbol en la sociedad actual. En Zebari y otros, (2021), como ejemplo relacionado con los aspectos mencionados, se construye un modelo de predicción de resultados futbolísticos mediante el uso de la distribución de Poisson, aplicado a La Liga española durante la temporada 2016-2017. De esta manera, se consigue analizar el funcionamiento de las casas de apuestas, en que se basan para tomar las decisiones que toman, a la vez de realizar una predicción de un posible enfrentamiento.

3 - Datos

Los datos utilizados para este modelo han sido extraídos de 3 fuentes distintas:

Primeramente, para analizar el rendimiento de cada jugador, se recopilaron distintas variables teniendo en cuenta la posición de cada uno, únicamente para los jugadores de campo, es decir, sin incluir a los porteros, debido a que no se puede construir un modelo de predicción con tan pocos datos, teniendo únicamente 2 porteros por equipo. Se realizará una división de los jugadores según su posición porque las funciones que desempeñan en el campo varían notablemente. Por ejemplo, las variables que se utilizan para medir el rendimiento de un defensa son distintas a las de un delantero, por lo que se dividió a los jugadores en: Defensas / Mediocentros / Delanteros. Estas variables fueron extraídas de FBREF.

A continuación, se añadió una variable con el salario de cada jugador, siendo esta la variable objetivo. La unidad de medida de la variable Salario en nuestras bases de datos es el millón de euros. Estas observaciones fueron extraídas de la página web Capology.

Finalmente, de la misma manera que indica Carrieri y otros (2018) sobre la importancia de conocer la popularidad de un jugador por los posibles intereses que conlleva, se ha agregado una variable que recoge el índice de popularidad por jugador. Debido a la dificultad que tiene recoger y medir en una sola variable la influencia de un jugador en distintos ámbitos (redes sociales, búsquedas en Internet, etc.) se ha empleado únicamente las búsquedas de cada jugador en la página web de Wikipedia, mediante una web llamada Page Views, que recoge la frecuencia total de búsqueda por jugador. De esta manera, se consigue un índice de popularidad proporcional al total de influencia en los diferentes ámbitos. Esta variable también está medida en unidades de millón.

Todas estas variables han sido recogidas manualmente. Cualquiera de ellas ha sido tomada como la suma de sus observaciones a lo largo de las últimas 3 temporadas, incluida la actual; es decir, desde la temporada 2020-2021 hasta la temporada 2022-2023.

Para un mejor entendimiento pasamos a explicar mediante una tabla las variables utilizadas en este modelo, dependiendo de la posición, y sus correspondientes abreviaturas:

Las tres posiciones comparten un elevado número de variables, las cuáles han sido reunidas en la Tabla 1.

Equipo	Equipo al que pertenece el jugador
Jugador	Nombre del Jugador
Posicion	Posición del Jugador
Salario	Salario anual de la temporada actual del jugador correspondiente
Popularidad	Índice de popularidad en base a Wikipedia (frecuencia total de búsquedas del nombre de cada jugador)
Fuera5GrandesLigas	Si el Jugador a jugado fuera de las 5 grandes ligas su valor será 1 (Sí), de lo contrario su valor será 0 (No)
PJ	Partidos Jugados
Titular	Número de Partidos en los que el jugador ha comenzado de Titular
Minutos	Minutos totales disputado
Goles	Goles
Ass	Asistencias de Gol
TarjA	Tarjetas Amarillas
TarjR	Tarjetas Rojas
FCom	Faltas Cometidas
FRec	Faltas Recibidas
BSR	Balones Sueltos Recuperados
DAereosG	Duelos Aéreos Ganados
DAereosP	Duelos Aéreos Perdidos
PAciertoPases	Porcentaje de Acierto en Pases
PasesUltTercio	Pases realizados en el último tercio del campo

Tabla 1: Tabla variables comunes. Elaboración propia

Además de estas variables que comparten las tres categorías, se han seleccionado otras acordes al rol que emplean los jugadores en cada posición. Para los defensas se recogen las variables destacables en la Tabla 2.

VReg	Número de veces que ha sido regateado
DisBloq	Disparos Bloqueados
Interc	Número de intercepciones de balón
Dpj	Número de despejes de balón
Err	Errores que provocan el tiro de un oponente

Tabla 2: Tabla variables exclusivas de Defensas. Elaboración propia

Para los mediocentros las variables destacadas se observan en la Tabla 3.

A-T	Acciones que conducen a un intento de tiro
A-Prg	Acciones que mueven el balón hacia al área del oponente
PerdidaBalon	Número de Pérdidas de Balón
VReg	Número de veces regateado
Interc	Intercepciones de balón

Tabla 3: Tabla variables exclusivas de Mediocentros. Elaboración propia

Por último, para los delanteros quedan recogidas sus variables a destacar en esta Tabla 4.

PasesClave	Pases Clave
A-T	Acciones que conducen a un intento de tiro
A-Prg	Acciones que mueven el balón hacia al área del oponente
RegInt	Número de regates intentados
RegExit	Número de regates exitosos
TotalDisp	Total de disparos
DaP	Disparos dirigidos a puerta

Tabla 4: Tabla variables exclusivas de Delanteros. Elaboración propia

Mediante estos datos, se construirán varios modelos de predicción, todos ellos explicando el salario en función de las variables de rendimiento. Por un lado, se creará un modelo para cada posición, teniendo en cuenta las variables destacadas en cada una de ellas con la finalidad de conocer si esas variables son significativas o no. Por otra parte, se construirá un cuarto modelo, para cada método de predicción, teniendo en cuenta a todos los jugadores de las distintas posiciones, pero utilizando únicamente las variables en común entre las 3 posiciones [Tabla 1].

Para una mayor facilidad a la hora de construir los distintos modelos de predicción vamos a agrupar a la variable Equipo según su rango salarial obteniendo así un menor número de categorías en la variable Equipo. Es decir, vamos a sumar los salarios de los jugadores de cada Equipo para observar como se distribuyen, obteniendo el resultado que vemos en la Figura 1.

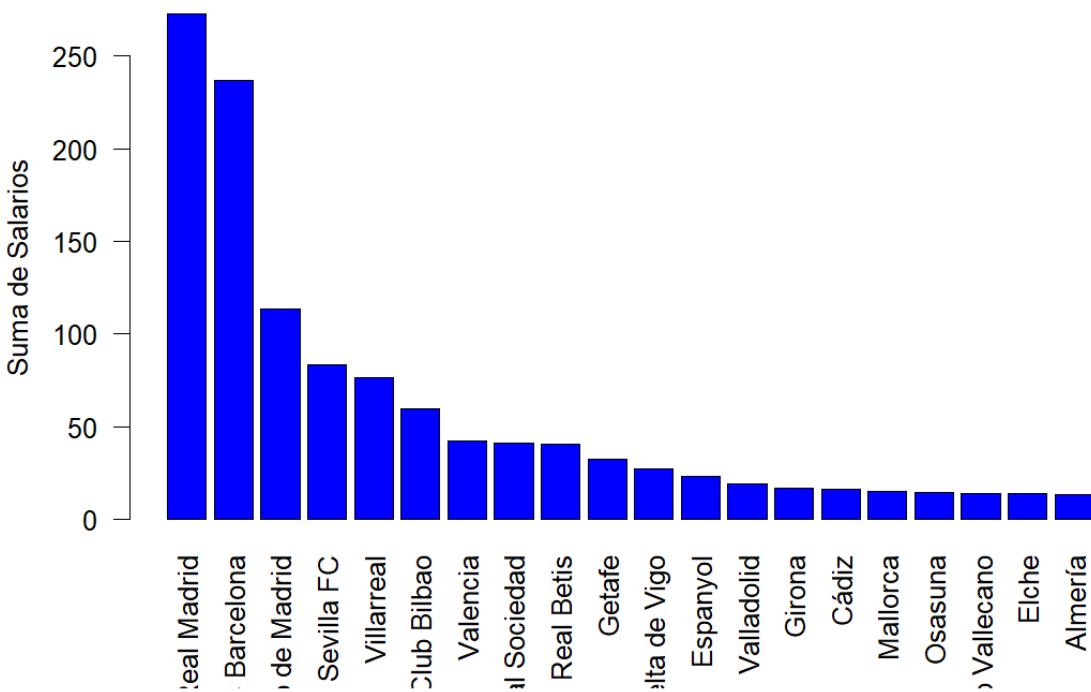


Figura 1: Gasto Salarial por Equipo. Elaboración Propia

Mediante esta Figura 1, podemos clasificar por rangos a los distintos equipos. Estos rangos estarán formados por:

- Rango Alto: Compuesto por Real Madrid y FC Barcelona
- Rango Medio Alto: Compuesto por Atlético de Madrid, Sevilla FC y Villarreal
- Rango Medio Bajo: Compuesto desde Athletic Club Bilbao hasta Real Betis
- Rango Bajo: Compuesto desde Getafe hasta Almería

4 - Modelos estadísticos utilizados

Primero de todo, con el objetivo de estudiar las posibles relaciones existentes entre las variables explicativas del modelo, se realizará un análisis de correlaciones de Pearson sobre las variables explicativas.

Este análisis de correlaciones se muestra mediante el uso de una matriz, en la cual se representa la relación lineal entre cada par de variables, determinando la fuerza y dirección de asociación entre cada par. Cada valor de la matriz puede variar entre -1 y 1, donde -1 indicaría una relación negativa perfecta, y por el contrario el valor 1 representa una relación positiva perfecta. Si su valor fuese 0 nos estaría indicando que no existe relación ninguna entre ese par de variables. Para entender mejor que relación representa cada uno de los posibles valores se expone la siguiente tabla:

$r = 1$	correlación perfecta
$0.8 < r < 1$	correlación muy alta
$0.6 < r <$ 0.8	correlación alta
$0.4 < r <$ 0.6	correlación moderada
$0.2 < r <$ 0.4	correlación baja
$0 < r < 0.2$	correlación muy baja
$r = 0$	correlación nula

Tabla 5: Rangos de correlación. Fuente STATS SOS

El cálculo de este valor sigue la siguiente fórmula: $r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$, donde:

- n es el tamaño de la muestra
- x_i/y_i son los valores de las variables en cada observación
- \bar{x}/\bar{y} son las medias de las variables x e y, respectivamente

Para el desarrollo de los distintos modelos mencionados se van a emplear diferentes técnicas de predicción con el fin de comparar los resultados y observar que modelo se ajusta mejor a los datos. Los métodos de predicción que se van a utilizar son:

- Regresión Lineal Múltiple - MLR
- Árboles de regresión / Random Forest (RF)
- Modelo de k-vecinos más cercanos – KNN

4.1 Regresión Lineal Múltiple

La Regresión Lineal Múltiple, MLR en inglés, permite generar un modelo lineal en el que el valor de una variable continua dependiente o respuesta, generalmente asociada a la letra Y, se determine a partir de un conjunto de variables independientes llamadas variables explicativas (X₁, X₂, X₃, ...). Esta técnica se puede emplear también para evaluar la influencia que tienen las variables explicativas sobre la variable respuesta.

Los modelos de regresión lineal múltiple siguen la siguiente ecuación:

$$Y = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n) + e$$

siendo:

- β_0 : la ordenada en el origen, el valor de la variable Y cuando los demás valen 0.
- β_i : el efecto promedio que tiene el incremento en una unidad de la variable explicativa X_i sobre la variable respuesta Y, manteniéndose constante los valores de las demás variables. Se conocen como coeficientes parciales de regresión.
- e : es la diferencia entre el valor observado y el estimado por el modelo, conocido como error.

Para poder determinar qué impacto tiene cada una de las variables en el modelo, es necesario calcular los coeficientes parciales estandarizados.

Estos modelos de regresión lineal requieren ciertas condiciones. Entre estas condiciones podemos encontrar la multicolinealidad de las variables explicativas, las cuales deben ser independientes entre sí, o bien la homocedasticidad de los residuos, que implica que la varianza de estos sea constante. Estas condiciones serán tomadas en cuenta a la hora de ajustar nuestro modelo.

Este tipo de regresión nos permite construir el modelo sujeto a distintos métodos de selección de variables. Estos son los métodos forward, backward y stepwise. Estos métodos son técnicas automatizadas que ayudan a identificar un subconjunto óptimo de variables explicativas a incluir en el modelo de regresión, en función de su capacidad para mejorar la capacidad predictiva del modelo.

En el método forward (selección hacia delante), se comienza con un modelo sin variables explicativas, al cual se le agrega aquella variable que produzca el mayor aumento sobre el ajuste del modelo. Este proceso se repite hasta que se llegue al número límite de variables deseadas.

De forma inversa, en el método backward (selección hacia atrás), se empieza con un modelo incluyendo todas las variables, quitando a este aquellas cuya eliminación supongan la menor disminución posible en el ajuste del modelo. Este método concluye

cuando no se pueden eliminar más variables sin disminuir significativamente la capacidad predictiva del modelo.

Por último se encuentra el método stepwise (selección paso a paso), el cual combina los enfoques del forward y backward seleccionando variable hacia adelante y hacia atrás en el mismo proceso. El método comienza con un modelo sin variables explicativas agregando a este aquella que aporte el mayor aumento sobre el ajuste del modelo. Luego en cada paso se consideran tanto las variables ya incluidas en el modelo como las que faltan por incluir, y se eliminan aquellas que no contribuyen significativamente a la capacidad predictiva del modelo. El proceso termina cuando no hay más variables candidatas a introducirse o eliminarse, o bien dado un numero de pasos preseleccionado.

Se estudiarán estos 3 métodos mencionados aplicados a la regresión lineal múltiple con el fin de analizar que método de los 3 es el que genera menor error a la hora de predecir.

Además de los métodos de selección de variables, tenemos también los métodos de selección de modelos. Entre estos, los más utilizados son:

- AIC (Akaike Information Criterion): Es una medida de calidad relativa del modelo en términos de ajuste y complejidad. Cuanto menos sea el AIC, mejor será la capacidad de ajuste del modelo y menor complejidad tendrá. Se calcula utilizando la siguiente fórmula: $AIC = 2k - 2\ln(L)$; donde k es el número de parámetros estimados del modelo y L es la verosimilitud del modelo, que indica qué tan bien el modelo se ajusta a los datos observados.
- BIC (Bayesian Information Criterion): Es similar al AIC, pero penaliza más fuertemente los modelos más complejos. Se calcula utilizando la fórmula: $BIC = -2\ln(L) + k * \ln(n)$, donde el nuevo parámetro n es el número de observaciones en los datos. Al tener en cuenta el tamaño de la muestra en $k*\ln(n)$, se penalizan aquellos modelos más complejos, luego es un criterio más estricto que el AIC para bases de datos de altas dimensiones.

A parte de estos dos criterios, existen otros que también son utilizados. Entre estos se encuentran el criterio del R-Cuadrado Ajustado, en el cual se penaliza inclusión de variables irrelevantes en el modelo, el criterio Cp de Mallows o el criterio SBC, este último es similar a los criterios AIC y BIC. Se estudiará, bajo las condiciones de cada combinación de criterios entre los métodos de selección de variables y los métodos de selección de modelos, que combinación es la que provoca un mejor ajuste a los datos observados de la variable Salario.

4.2 Árboles de regresión y Random Forest

Pasando al segundo método de predicción nos encontramos con los árboles de regresión. Los árboles de regresión (al estar tratando con una variable cuantitativa, como lo es el Salario, estamos hablando de árbol de regresión, de lo contrario, si fuese cualitativa estaríamos hablando de árbol de clasificación) constituyen una herramienta útil para la predicción de variables cuantitativas, de una manera sencilla y sin asunciones teóricas sobre los datos. Los árboles representan una segmentación de los

datos a partir de una serie de reglas simples, que se van aplicando de forma jerárquica y secuencial. De esta forma, se obtienen una serie de segmentos, llamados nodos, que contienen subconjuntos de la muestra. Los árboles de regresión constan de tres partes:

- **Nodos:** Puntos de división en el árbol. Estos se dividen en nodo raíz, que es el nodo superior del árbol y contiene la totalidad de los datos, y por otra parte los nodos internos, que son aquellos puntos que se encuentran entre el nodo raíz y las hojas.
- **Ramas:** Son las conexiones entre los nodos del árbol.
- **Hojas:** Son los nodos finales del árbol, no tiene sucesores. A estos nodos se les asigna un valor de predicción de forma que todas las observaciones que pertenezcan a ese nodo serán predichas a partir de dicho valor.

El proceso consiste en seleccionar en cada paso una característica o variable del conjunto de datos, que se utilizará para dividir las observaciones en subconjuntos más homogéneos y diferentes entre sí. El objetivo es ir encontrando en cada paso la variable que mejor diferencia las observaciones de acuerdo a la variable objetivo. Este proceso se repite hasta que se cumpla un criterio de parada. Algunos de estos criterios son:

- **Profundidad máxima del árbol:** Se establece un límite en la cantidad máxima de capas del árbol
- **Número mínimo de observaciones en una hoja:** Se especifica el umbral mínimo para la cantidad de observaciones necesarias en una hoja del árbol. Si el número de observaciones cae por debajo de este umbral en alguna de las hojas, no se realizarán más divisiones en ese camino, y se detiene la construcción del árbol en esa rama. Esto evita crear subdivisiones demasiado pequeñas que podrían ser poco representativas.

Estos dos son los criterios más utilizados, aunque también se podría utilizar tanto el *criterio de ganancia de información mínima*, que consiste en establecer un umbral mínimo para la ganancia de información, de forma que si la ganancia de información es inferior a este umbral, no se realiza la división y se detiene la construcción del modelo, como el *criterio de Índice de Gini*, que mide el grado de pureza de un nodo, de forma que si la reducción a la impureza al realizar una nueva división es menor a un umbral mínimo preseleccionado, se detiene la construcción del árbol en ese punto. Estos dos últimos criterios evitan respectivamente generar divisiones que no contribuyan de manera significativa a la capacidad predictiva del modelo y que las divisiones se realicen solo cuando haya una mejora sustancial en la pureza de las hojas resultantes.

Se tendrán en cuenta estos criterios con el fin de observar cuál de ellos es el que menor error comete a la hora de predecir nuestra variable objetivo.

Una vez construidos los árboles se estudiará la importancia de cada variable en el árbol. La importancia de las variables viene dada por la mejora producida en el criterio de división al realizar la partición de los nodos. Por este motivo, la magnitud de la importancia no se puede interpretar directamente, pero el ranking que se forma si.

En la segunda parte de este método, tenemos las técnicas de random forest. Se utiliza el random forest debido a que construir un único árbol de regresión puede acarrear distintos problemas, como la falta de generalización hacia nuevos datos o la sensibilidad de selección de variables predictoras. Por ello, con el random forest, se generan numerosos árboles de regresión distintos unos de otros en cuanto a las variables seleccionadas, de manera que las predicciones se realizan utilizando la información de todos ellos, mediante una media aritmética.

4.3 Método de los k-vecinos más cercanos: KNN

Este modelo se basa en la idea de que observaciones que tomen valores similares en las variables explicativas deben tomar también valores similares en la variable objetivo. De esta forma, a la hora de predecir el valor de una variable objetivo para una determinada observación, se buscan las k observaciones más próximas y se obtienen el valor de la predicción como la media de los valores de la variable objetivo.

El número de vecinos depende del conjunto de datos que se vaya a utilizar, pero lo habitual es probar un conjunto de valores k y seleccionar el que menor error produzca en el conjunto de datos de validación. La proximidad de los vecinos se mide a partir del cálculo de las distancias entre las observaciones.

Como condiciones que debe de cumplir el modelo, es muy importante estandarizar previamente las variables explicativas para que todas tengan el mismo peso a la hora de calcular dichas distancias. Otro aspecto es que hay que eliminar los NA o valores perdidos, ya que pueden afectar a los cálculos de las distancias.

Como se ha mencionado al comienzo de esta sección, para cada uno de los conjuntos de datos que disponemos (Defensas, Mediocentros, Delanteros y Jugadores_Combinados) se van a aplicar las 3 técnicas de predicción mencionadas, con el fin de estudiar cuál es el modelo que mejor se ajusta a los datos, midiendo así el rendimiento de cada modelo en términos de su capacidad predictiva. Por otra parte, estudiar la importancia de las variables, cuánto aporta cada una y de qué manera a la determinación del salario de un jugador.

5 - Resultados

Todos los resultados que aparecen a continuación han sido obtenidos mediante el uso de los programas SAS y R.

5.1 Regresión Lineal Múltiple

Como primer método estadístico a aplicar se encuentra la regresión lineal múltiple, pero como se ha comentado anteriormente, lo primero de todo es estudiar la relación entre nuestras variables explicativas. Realizando el análisis de correlaciones para la base de datos de los defensas obtenemos la Tabla 6: ver Anexo 1.

Para los defensas, vemos que la variable Salario presenta una correlación muy alta con la Popularidad del jugador, seguido de los Pases en el Último Tercio, y ya con una correlación menos fuerte con los Balones Sueltos Recuperados y el Porcentaje de Acierto en Pases. Debido a la complejidad de estudiar las relaciones entre las distintas variables manualmente vamos a pasar directamente al estudio de los respectivos modelos, donde se tienen en cuenta criterios de selección de variables.

Es común realizar primero una partición de los datos para construir el modelo de predicción utilizando el 80% de los datos, para luego más adelante comprobar la capacidad de ajuste del modelo construido con el 20% de datos restantes que son ajenos al modelo. De esta manera se comprueba si con datos ajenos al modelo, se siguen realizando buenas predicciones o en realidad el modelo no tiene buena capacidad de ajuste. Pero, debido a la escasez de datos que tienen las 3 bases de datos de Defensas, Mediocentros y Delanteros, oscilando entre 120 y 160 observaciones, varían mucho las conclusiones obtenidas, tanto en la selección de variables como en los estadísticos de ajuste del modelo. Este estudio se ha realizado mediante el uso de una macro creada en SAS, la cual repite el proceso de creación de un modelo de predicción un número determinado de veces mediante un bucle, en este caso está fijado en 100 iteraciones, para un valor aleatorio cada vez que realiza el proceso. Luego en cada vuelta que realiza el bucle, se selecciona una partición de los datos diferente y se construye el modelo con los datos seleccionados en esa vuelta del bucle. Ejecutando esta macro nos damos cuenta de que la selección de variables del modelo depende totalmente de la partición realizada, variando a su vez la capacidad de ajuste del

modelo. Por ejemplo, para los defensas, el R Cuadrado-Ajustado varía entre 70% y 85%, además de no seleccionar las mismas variables en los diferentes modelos. Esto nos dice que no es fiable realizar una partición de los datos.

Por ello, para los defensas, mediocentros y delanteros, se va a construir el modelo con todos los jugadores disponibles.

Construimos ahora un modelo de regresión lineal múltiple para los defensas sin tener en cuenta condiciones de selección de variables o modelos, para observar cómo se comportan las variables en su conjunto y realizar así una exploración inicial de los datos. Este modelo se resume en la siguiente tabla:

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	-2.302234627	B	1.93886249	-1.19	0.2371
Equipo Rango Alto	2.919079360	B	0.76382509	3.82	0.0002
Equipo Rango Medio Alto	0.157112750	B	0.47023739	0.33	0.7388
Equipo Rango Medio Bajo	0.768705768	B	0.35654107	2.16	0.0328
Equipo Rango Bajo	0.000000000	B			
Fuera5GrandesLigas No	0.217444264	B	0.42682352	0.51	0.6113
Fuera5GrandesLigas Si	0.000000000	B			
Popularidad	2.402725627		0.55926334	4.30	<.0001
PJ	-0.025745975		0.02104145	-1.22	0.2232
Titular	0.001285175		0.06168286	0.02	0.9834
Minutos	0.000169986		0.00068798	0.25	0.8052
Goles	0.042788635		0.06171282	0.69	0.4893
Ass	0.146294125		0.04762876	3.07	0.0026
TarjA	0.043721342		0.02836216	1.54	0.1255
TarjR	0.008711825		0.15029800	0.06	0.9539
FCom	-0.011710468		0.00745973	-1.57	0.1187
FRec	0.001581670		0.00551272	0.29	0.7746
BSR	-0.002029418		0.00296473	-0.68	0.4948
DAereosG	-0.005236217		0.00554466	-0.94	0.3466
DAereosP	-0.014044156		0.00944245	-1.49	0.1392
PAciertoPases	0.028400737		0.02543967	1.12	0.2662
PasesUltTercio	0.014061535		0.00310679	4.53	<.0001
VReg	-0.002663233		0.00998976	-0.27	0.7902
DisBloq	0.027310467		0.01735211	1.57	0.1178
Interc	0.004663421		0.00646533	0.72	0.4719
Dpj	0.002525488		0.00439171	0.58	0.5662
Err	-0.004748035		0.09892434	-0.05	0.9618

Tabla 7: Regresión sin condiciones para Defensas. Elaboración propia

Como se puede apreciar, el modelo presenta varias variables muy poco significativas, posiblemente debido a la existencia de relación entre algunas de las variables explicativas del conjunto de datos, por ello pasamos al análisis de los modelos de predicción teniendo en cuenta los distintos criterios de selección ya mencionados, obteniendo lo siguiente:

Modelos	StepW AIC	StepW BIC	StepW AdjRC	ForW AIC	ForW BIC	ForW AdjRC	BackW AIC	BackW BIC	BackW AdjRC
Nº Parámetros	9	9	9	9	9	9	9	9	11
R-Cuad. Ajust.	0.798	0.798		0.798	0.798		0.7986	0.7986	0.7987
MSE	2,316	2,316	2,316	2,316	2,316	2,316	2,316	2,316	2,3148

Tabla 8: Resultado de diferentes modelos para Defensas. Elaboración propia

Dados los siguientes resultados, vemos que, para cualquier combinación de criterios de selección de variables o modelos, el modelo construido es siempre el mismo, salvo para el modelo Backward AdjRC en el cual aumentamos a 11 parámetros, pero prácticamente la capacidad de ajuste no se ve afectada, luego nos quedamos con cualquiera de las otras combinaciones al dar lugar a un modelo más sencillo con la misma capacidad de ajuste. En este caso seleccionaremos el modelo Stepwise AIC. Este modelo es capaz de explicar el 79.86% de la variabilidad de la variable Salario mediante las variables explicativas, lo cual vemos a través del R-Cuadrado Ajustado. Este valor nos indica que el modelo tiene una capacidad considerablemente alta para predecir la variable Salario. Cuanto más se aproxime a 1 este valor, mejor capacidad predictiva tendrá el modelo. En cuanto al MSE, obtenemos un valor de 2.31, lo que supone un valor relativamente bajo, al compararlo con el rango en el que se mueve nuestra variable dependiente, que oscila entre 0.05 y 22.5, y por tanto nos indica que el modelo ajusta bastante bien, variando en media 2.31 millones de euros a la hora de predecir el salario de un jugador.

Este modelo está formado por las siguientes 9 variables:

$$\text{Salario} = \text{Equipo} + \text{Popularidad} + \text{Ass} + \text{TarjA} + \text{FCom} + \text{DAereosP} + \text{PAciertoPases} + \text{PasesUltTercio} + \text{DisBloq}$$

Parameter Estimates						
Parameter	DF	Estimate	Standardized Estimate	Standard Error	t Value	Pr > t
Intercept	1	-3.049101	0	1.697030	-1.80	0.0744
Equipo Rango Alto	1	2.777777	0.251144	0.701694	3.96	0.0001
Equipo Rango Medio Alto	1	-0.011093	-0.001202	0.411681	-0.03	0.9785
Equipo Rango Medio Bajo	1	0.764026	0.090816	0.331220	2.31	0.0224
Equipo Rango Bajo	0	0	0	.	.	.
Popularidad	1	2.511269	0.324865	0.511703	4.91	<.0001
Ass	1	0.140576	0.171261	0.038091	3.69	0.0003
TarjA	1	0.048150	0.106129	0.024792	1.94	0.0540
FCom	1	-0.014782	-0.145588	0.005367	-2.75	0.0066
DAereosP	1	-0.019346	-0.193908	0.006377	-3.03	0.0028
PAciertoPases	1	0.035806	0.075379	0.021872	1.64	0.1037
PasesUltTercio	1	0.012038	0.371902	0.002176	5.53	<.0001
DisBloq	1	0.031765	0.161431	0.012023	2.64	0.0091

Tabla 9: Modelo seleccionado para los Defensas. Elaboración propia

Analizando la variable categórica Equipo, vemos que el rango Medio Alto no es significativo. Al tener como referencia al rango Bajo, podemos decir que el rango al que pertenece el equipo de un jugador no afecta significativamente a su salario cuando comparamos a los jugadores pertenecientes al rango Bajo con los del rango Medio Alto, mientras que sí que existen diferencias entre el rango Bajo y los rangos Alto y Medio Bajo. Esto nos quiere decir que no hay diferencias significativas en términos de cómo se pagan a los jugadores en función de su rendimiento entre los equipos del rango Medio Alto y Bajo. Para los rangos significativos, Alto y Medio Bajo, tenemos unos parámetros estimados de 2.78 y 0.76 respectivamente. De estas estimaciones esperamos que los salarios de los jugadores pertenecientes al rango Alto sean 2.78 unidades de medida más altos con respecto a los jugadores del rango Bajo, manteniendo el resto de variables constantes. De igual manera ocurriría con el rango Medio Bajo, pero en este caso serían 0.76 unidades de medida superiores, manteniendo a su vez al resto de variables constantes. Por tanto, podemos decir que, para un mismo rendimiento, los jugadores son mejor pagados en los rangos Alto y Medio Bajo que en el rango Bajo, mientras que para los pertenecientes al rango Medio Alto no existen diferencias salariales para jugadores de mismo rendimiento con respecto al rango Bajo.

Analizando ahora las variables continuas, vemos que el modelo capta de manera adecuada aquellas variables que influyen de manera negativa en el rendimiento de un jugador. Tanto las faltas cometidas (FCom) como los Duelos Aéreos Perdidos (DAereosP), son variables que, en el mundo del fútbol, cuanto más aumenten peor rendimiento tendrá un jugador, y si nos fijamos en los parámetros estimados de nuestro modelo, ambas tienen parámetros negativos. Esto nos dice que, por ejemplo, analizando la variable FCom, por cada aumento unitario en las faltas cometidas manteniendo al resto de variables constantes, el salario de un jugador disminuye en 14.782€ ($1.000.000 \times -0,014782$). Fijándonos en el resto de variables, vemos que la variable que más aumenta el salario de un jugador por cada aumento unitario manteniendo al resto constantes es la Popularidad. Al estar tanto los Salarios como la Popularidad medidas en unidades de millón, podemos decir que, por cada millón que aumente un jugador en su índice de popularidad, se espera que su salario aumente en 2,511269 millones de euros.

Ahora que hemos obtenido un modelo de regresión que se ajusta bien a los datos, vamos a estudiar la importancia de cada variable en nuestro modelo. Para ello analizamos el gráfico de la progresión de los coeficientes en relación con el Salario.

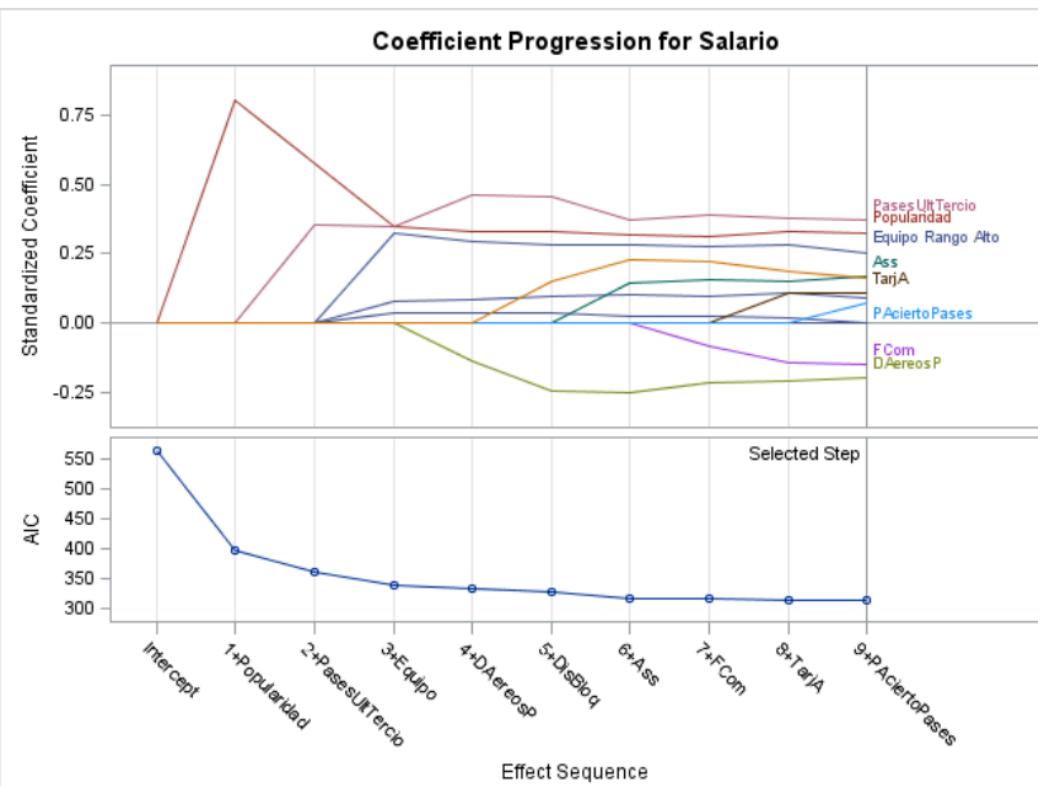


Figura 2: Evolución de importancia de variables para Defensas. Elaboración Propia

Con esta imagen observamos dos aspectos, por un lado, en la imagen de abajo, vemos lo que aporta cada variable en cada “step” que realiza el modelo para la reducción del AIC, y por otro lado, vemos en la imagen de arriba, la influencia de cada variable sobre la variable Salario. Estos coeficientes estandarizados representan el cambio en la variable dependiente (en unidades de desviación estándar) por cada cambio de una desviación estándar en la variable independiente, manteniendo constantes al resto de variables independientes. Vemos que la variable más influyente es PasesUltTercio, seguida de Popularidad, el Rango Alto en caso de pertenecer a este rango, y Ass. El resto de rangos de la variable Equipo están más cercanos a 0, siendo estos menos influyentes. De igual manera que veímos en los parámetros estimados del modelo, las variables FCom y DAereosP están representadas por debajo del 0, lo que nos indica nuevamente que tienen una influencia negativa y su importancia es ciertamente notoria al estar alejadas del 0.

Vemos que únicamente se ha incluido en este modelo la variable característica de los Defensas: Disparos Bloqueados, DisBloq (línea naranja). Su coeficiente de regresión estandarizado y su p-valor en el modelo, nos indican que es una variable de importancia media baja al presentar un coeficiente estandarizado de 0,20.

Para las bases de datos de los mediocentros y delanteros se ha aplicado el mismo procedimiento y enfoque que para los defensas. Por tanto, con el fin de no repetir el proceso completo para estas dos nuevas bases, se han adjuntado los resultados de los modelos construidos para los mediocentros y delanteros en sus respectivos anexos, Anexo 2 y Anexo 3, respectivamente. A continuación se incluirán únicamente las conclusiones obtenidas.

Para los mediocentros, como mejor modelo podríamos elegir al Backward BIC ya que es el que mejor relación complejidad-capacidad de ajuste tiene. Presenta una variabilidad explicada del 72,01% sobre el salario por parte de las 8 variables explicativas, con un MSE de 10,17. Podemos ver los distintos resultados obtenidos en la Tabla 12 del Anexo II.

Con los valores de los estadísticos obtenidos, podemos decir que este modelo se ajusta moderadamente bien a los datos, ya que tenemos un valor de R-Cuad.Aj. de 0.717, lo que nos indica que el modelo presenta una capacidad predictiva moderadamente buena. Si observamos el MSE, vemos que, sobre todo si lo comparamos con el obtenido para el modelo de los defensas, obtenemos un valor ciertamente más elevado. Esto podría explicarse por un menor porcentaje de variabilidad explicada frente al anterior modelo de los defensas y el aumento en el rango de los salarios para los mediocentros, los cuales oscilan entre 0.14 y 37.5, lo cual podría explicar esta subida del MSE. Este valor nos indica que el modelo tiene una precisión media, variando en un promedio de 10.17 millones de euros a la hora de predecir el salario de un mediocampista. Si analizamos sus parámetros estimados vemos lo siguiente: ver Tabla 13 del Anexo II.

Analizando la variable categórica Equipo, vemos que ahora, a diferencia que con los defensas donde el modelo no encontraba diferencias entre los rangos Medio Alto y Bajo, en este modelo no se encuentran diferencias significativas con el Rango Medio Bajo en referencia al Bajo, lo que es una conclusión más lógica al tener los equipos pertenecientes a estos dos rangos un gasto salarial en sus plantillas más similar al que tienen los equipos del rango Medio Alto y el Bajo.

En cuanto a las variables continuas, vemos que nuevamente la variable más significativa es PasesUltTercio, al tener el pvalor más pequeño del modelo, lo cual corroboraremos en el gráfico de la progresión de los coeficientes.

Como características principales a destacar en este modelo para los mediocentros observamos que en el modelo 3 de los 7 parámetros estimados nos ofrecen conclusiones contrarias a lo que sería lo lógico en el mundo del fútbol. Esto ocurre con las variables BSR (balones sueltos recuperados), DAereosP (duelos aéreos perdidos) y A_T (acciones para la creación de tiros). Para BSR y A_T obtenemos parámetros estimados negativos, cuando estas son dos variables que en la realidad se interpretarían como positivas, dado que al aumentar estas, mejor rendimiento tendrá un jugador. Por ejemplo, para BSR, por cada aumento en balones recuperados de un jugador, la variable salario se espera que disminuya en 13.104 euros, manteniendo al resto de variables constantes. Lo mismo para A_T, pero disminuyendo en 9.362 euros en su caso. Por otro lado, tenemos a DAereosP, cuyo parámetro estimado nos está indicando que por cada duelo aéreo más que pierda un jugador, su salario se espera que aumente en 25076 euros, cuando esta variable debería ser una variable negativa.

Pasando a analizar la progresión de la importancia de las distintas variables del modelo de los mediocentros, obtenemos el siguiente gráfico: ver Figura 3 Anexo II.

El ser PasesUltTercio la variable más influyente del modelo es una conclusión lógica al ser los mediocampistas, ya que en la mayoría de ocasiones, son los encargados de

dirigir el balón hacia la portería contraria, luego realizarán un mayor número de pases en el último tercio del campo rival.

Para los delanteros, observando los resultados obtenidos en la Tabla 16 del Anexo III, vemos que claramente el mejor modelo es el generado tanto por el Stepwise AIC/BIC como por el Forward AIC/BIC, dado que son el mismo modelo. Con este modelo se obtiene la mejor relación entre complejidad-capacidad de ajuste porque únicamente con 3 parámetros es capaz de explicar un porcentaje de variabilidad muy cercano al de los demás modelos, siendo necesarios en estos el uso de 6 a 9 parámetros. A su vez vemos que el MSE prácticamente no aumenta en comparación al resto de modelos. Por lo tanto, podemos concluir que este modelo se ajusta bastante bien a los datos y tiene una buena capacidad predictiva.

En este caso solo es necesario conocer el equipo, su popularidad y los pases clave que genera para predecir el salario. En este modelo la popularidad aporta 1.406.641 € por cada millón de unidades que aumente el índice de popularidad de un jugador, manteniendo a las otras 2 variables constantes. Por otra parte, por cada pase clave más que dé un delantero, siendo estos aquellos pases que sean determinantes para que una jugada acabe en gol, se espera que su salario aumente en 10.770 €.

Sobre las categorías de los equipos, obtendríamos las mismas conclusiones que para los mediocentros, únicamente es significativo pertenecer a los rangos Alto o Medio Alto para determinar el salario de un jugador, confundiendo por tanto a aquellos jugadores que pertenezcan a equipos del rango Medio Bajo, en referencia al rango Bajo.

Si comparamos estos resultados con los obtenidos para los mediocentros y los defensas, vemos que afecta en mayor medida el pertenecer al rango Alto siendo delantero que siendo mediocentro o defensa. Antes obteníamos unos parámetros estimados de 4,6 y 2,7, mientras que ahora obtenemos un valor de 8,78. Esto nos dice que son mejor pagados aquellos jugadores que son delanteros frente a los mediocentros o defensas, dentro del rango Alto.

Podemos observar el gráfico de la progresión de los coeficientes con relación al salario para estudiar la importancia de los parámetros, en la Figura 4 del Anexo III.

En este modelo la variable más determinante es pertenecer a un equipo de alto rango salarial. Estrechamente seguida se encuentra la variable Popularidad, siendo esta más determinante que en los dos anteriores modelos. Luego nos encontramos al rango Medio Alto, seguido de los PasesClave y ya más próxima al 0, siendo esta menos significativa, el rango Medio Bajo.

Finalmente llegamos al último modelo de regresión, en el cual estudiaremos como afectan las distintas variables de rendimiento comunes para el conjunto de jugadores.

En esta nueva base de datos tenemos una mayor cantidad de observaciones, donde antes nos movíamos entre 120 y 160 observaciones, ahora tenemos 414 observaciones. Si realizamos la misma macro mencionada anteriormente, los estadísticos de ajuste y la selección de variables son mucho más constantes que en las anteriores bases de datos.

Por ello vamos a dividir los datos en 2 ficheros, uno llamado entrenamiento con el cual vamos a construir nuestro modelo, y otro llamado prueba donde vamos a comprobar si nuestro modelo realiza buenas predicciones. Esta división estará formada por 83 observaciones que formarán el fichero prueba (20%) y las restantes el fichero entrenamiento (80%).

Analizando las correlaciones de las variables explicativas del fichero train se puede observar nuevamente que las variables con mayor fuerza de relación con la variable objetivo son Popularidad y PasesUltTercio.

Pasando a la construcción de la regresión lineal sin condiciones previas, vemos que:

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	0.628955936	B	1.97537521	0.32	0.7504
Equipo Rango Alto	5.316882543	B	0.67953311	7.82	<.0001
Equipo Rango Medio	0.919752473	B	0.47021233	1.96	0.0514
Equipo Rango Bajo	0.000000000	B	.	.	.
Popularidad	1.863771535		0.22503419	8.28	<.0001
Fuera5GrandesLigas No	0.489919096	B	0.42692068	1.15	0.2520
Fuera5GrandesLigas Si	0.000000000	B	.	.	.
PJ	0.010106673		0.01453965	0.70	0.4875
Minutos	-0.000368627		0.00063394	-0.58	0.5613
Titular	0.040093473		0.05127501	0.78	0.4348
Goles	0.005458152		0.02302093	0.24	0.8127
Ass	-0.022024503		0.04417999	-0.50	0.6185
TarjA	0.021626943		0.03168060	0.68	0.4953
TarjR	-0.158988631		0.17388972	-0.91	0.3613
FCom	-0.015469301		0.00660632	-2.34	0.0198
FRec	0.012186630		0.00473794	2.57	0.0106
BSR	-0.005919472		0.00229881	-2.58	0.0105
DAereosG	0.003782793		0.00554119	0.68	0.4953
DAereosP	-0.001492813		0.00654149	-0.23	0.8196
PAciertoPases	-0.019321991		0.02458866	-0.79	0.4326
PasesUltTercio	0.014728067		0.00199760	7.37	<.0001

Tabla 18: Regresión sin condiciones para todos los jugadores. Elaboración propia

El modelo presenta la mayoría de sus variables como no significativas luego debemos pasar a la construcción de modelos bajo los criterios con los que se han trabajado anteriormente, reflejando sus resultados en la Tabla 19:

Modelos	Stepwise AIC	Stepwise BIC	Stepwise AdjRC	Forward AIC	Forward BIC	Forward AdjRC	Backward AIC	Backward BIC	Backward AdjRC
Nº Parámetros	7	7	8	7	7	8	7	7	8
R-Cuad. Ajust.	0.725	0.725	0.7259	0.725	0.725	0.7259	0.7255	0.7255	0.7259
MSE	6,702	6,702	6,69220	6,702	6,702	6,69220	6,70297	6,7029	6,69220

Tabla 19: Resultado de diferentes modelos para todos los jugadores. Elaboración propia

En este caso se generan 2 modelos, un primero que se repite para los criterios AIC y BIC, y un segundo que se repite para el criterio AdjRC. Como mejor modelo podríamos

seleccionar al generado por los criterios AIC/BIC , ya que, si los comparamos con los del modelo AdjRC, disminuimos su complejidad al reducir de 8 a 7 parámetros sin prácticamente penalizar el ajuste al no verse apenas reducido el R-Cuadrado Ajustado ni aumentado el MSE.

Observando los estadísticos de capacidad de ajuste, podemos decir que el modelo se ajusta razonablemente bien a los datos, es decir, es un modelo aceptable.

Este modelo está formado por las siguientes variables:

$$\text{Salario} = \text{Equipo} + \text{Popularidad} + \text{FCom} + \text{FRec} + \text{BSR} + \text{DaereosG} + \text{PasesUltTercio}$$

Parameter Estimates						
Parameter	DF	Estimate	Standardized Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.148367	0	0.342213	-0.43	0.6649
Equipo Rango Alto	1	5.376865	0.322041	0.662385	8.12	<.0001
Equipo Rango Medio Alto	1	1.032767	0.074336	0.458786	2.25	0.0251
Equipo Rango Medio Bajo	1	0.270175	0.021753	0.380543	0.71	0.4782
Equipo Rango Bajo	0	0	0	.	.	.
Popularidad	1	1.948016	0.387195	0.187294	10.40	<.0001
FCom	1	-0.011588	-0.101913	0.004928	-2.35	0.0193
FRec	1	0.013379	0.138902	0.004002	3.34	0.0009
BSR	1	-0.004134	-0.133718	0.001772	-2.33	0.0203
DAereosG	1	0.004593	0.059371	0.002685	1.71	0.0881
PasesUltTercio	1	0.013999	0.415136	0.001867	7.50	<.0001

Tabla 20: Modelo seleccionado para todos los jugadores. Elaboración propia

En este modelo tenemos de nuevo a los rangos Alto y Medio Alto como únicos rangos significativamente diferentes del rango Bajo en cuanto a cómo pagan a sus jugadores en base a su rendimiento. Por tanto, con tal de eliminar la categoría Rango Medio Bajo al ser esta no significativa, vamos a agrupar como rango Bajo a los rangos Medio Bajo y Bajo, y por otro lado llamaremos a los equipos pertenecientes al rango Medio Alto como rango Medio.

Construimos de nuevo todos los modelos teniendo en cuenta la nueva estructura de la variable Equipo, y nuevamente obtenemos como mejor modelo a los generados por los criterios AIC y BIC, manteniendo los 7 parámetros de antes, aumentando el R-Cuadrado Ajustado a 0.726 y sin prácticamente reducir el MSE (se reduce en 0.02).

Luego finalmente obtenemos el siguiente modelo:

Parameter Estimates						
Parameter	DF	Estimate	Standardized Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.093671	0	0.333172	-0.28	0.7788
Equipo Rango Alto	1	5.256482	0.314831	0.639822	8.22	<.0001
Equipo Rango Medio	1	0.941338	0.067755	0.440002	2.14	0.0332
Equipo Rango Bajo	0	0	0	.	.	.
Popularidad	1	1.965800	0.390730	0.185468	10.60	<.0001
FCom	1	-0.011518	-0.101300	0.004923	-2.34	0.0199
FRec	1	0.013201	0.137057	0.003991	3.31	0.0010
BSR	1	-0.004082	-0.132035	0.001769	-2.31	0.0217
DAereosG	1	0.004647	0.060073	0.002682	1.73	0.0841
PasesUltTercio	1	0.014052	0.416694	0.001864	7.54	<.0001

Tabla 21: Modelo corregido para todos los jugadores. Elaboración propia

Ahora la variable Equipo no presenta categorías no significativas. Vemos que las variables Popularidad y PasesUltTercio, al igual que para los mediocentros y defensas, son las más significativas de entre las continuas. Nuestro modelo “final” capta bien la influencia de las distintas variables continuas, por ejemplo, al obtener un parámetro estimado para FCom negativo y para FRec positivo, salvo para BSR (balones sueltos recuperados). Esto puede deberse a que las grandes estrellas, que son generalmente los delanteros, son menos activos a lo largo de un partido y por tanto roban menos balones. Por ejemplo, Robert Lewandowski o Karim Benzema, ambos delanteros estrella del FC Barcelona y Real Madrid, no juegan con la intención de robar balones, sino suelen estar pendientes de estar bien colocados entre la defensa para que cuando se genere una ocasión de gol, estar perfectamente colocados. Esto si lo trasladamos al resto de delanteros, que son los jugadores que más cobran en media, es normal que salga un parámetro estimado negativo para BSR.

Ahora, si estudiamos el gráfico de la progresión de la importancia de cada variable, nos queda lo siguiente:

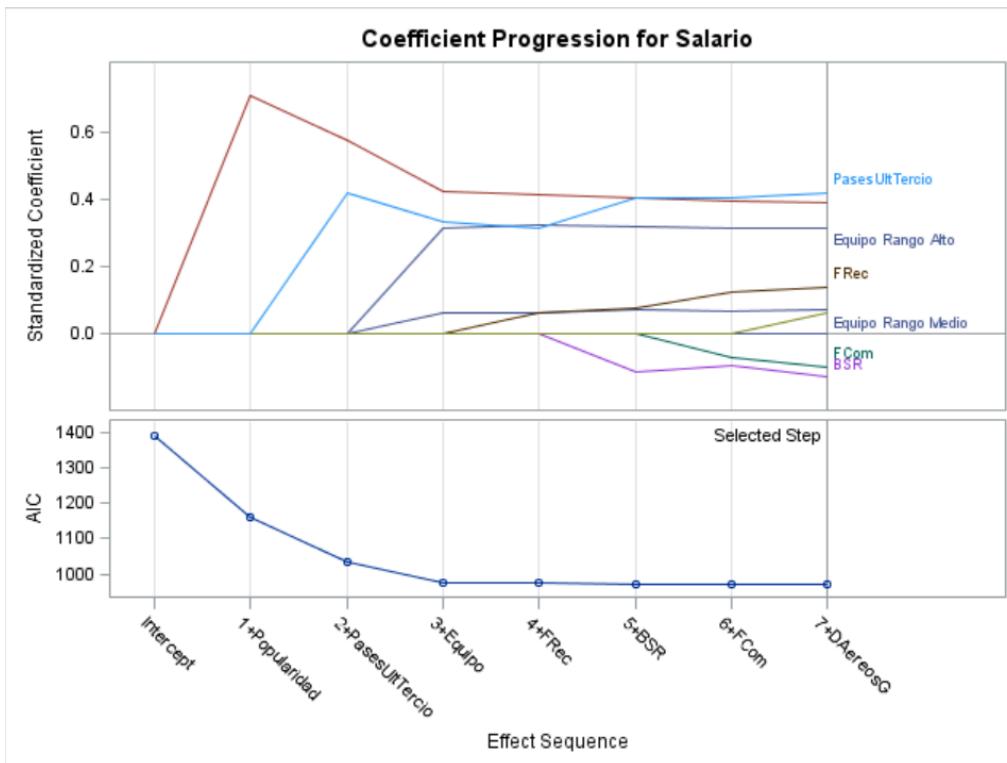


Figura 5: Evolución de importancia de variables para todos los jugadores. Elaboración propia

Mediante esta imagen corroboramos la conclusión que habíamos tomado sobre la importancia de las variables Popularidad y PasesUltTercio y vemos que el resto de variables son menos influyentes en comparación a estas, salvo pertenecer al rango Alto, categoría muy cercana a estas dos anteriores variables en cuanto a importancia.

El resto de variables son todas influyentes, pero de influencia media o media baja.

Ahora una vez seleccionado nuestro mejor modelo con el fichero de entrenamiento, vamos a utilizarlo para realizar predicciones sobre el salario para el fichero que hemos dejado como prueba, obteniendo el siguiente resultado:

Obs	Jugador	Salario	prediction	error	28	Ronael Pierre Gabriel	1.200	2.4503	-1.25033
1	Ferland Mendy	10.420	10.2459	0.17413	29	Mouctar Diakhaby	0.800	0.9263	-0.12635
2	Álvaro Odriozola	7.290	6.2166	1.07337	30	Gabriel Paulista	5.330	2.0641	3.26588
3	Ronald Araújo	7.000	10.3208	-3.32082	31	Kalky	0.540	1.1249	-0.58492
4	Jordi Alba	20.830	13.1752	7.65480	32	Rodrigo Ely	0.600	0.1673	0.43266
5	Stefan Savic	4.500	4.5809	-0.08093	33	Juan Brandáriz	0.970	-0.1806	1.15064
6	Marcos Llorente	2.100	2.7254	-0.62540	34	Enzo Roco	0.155	0.1285	0.02646
7	Aitor Ruibal	0.440	0.5094	-0.06937	35	Eduardo Camavinga	8.330	15.0949	-6.76489
8	Luis Felipe	5.830	1.5677	4.26231	36	Saúl Ñíguez	12.000	3.7044	8.29564
9	Aïssa Mandi	4.790	2.1498	2.64024	37	Roberto Navarro	0.420	0.5086	-0.08859
10	Iñigo Martínez	5.000	2.6756	2.32439	38	Guido Rodríguez	2.680	4.1221	-1.44214
11	Daniel Vivian	1.180	0.1873	0.99272	39	William Carvalho	1.750	5.3884	-3.63836
12	Yuri Berchiche	4.200	2.4192	1.78078	40	Paul Akouokou	0.510	0.6096	-0.09958
13	Mikel Balenziaga	1.650	0.5651	1.08489	41	Rodríguez Sánchez	0.400	1.7601	-1.36012
14	David García	0.620	2.3620	-1.74199	42	Unai Vencedor	0.240	2.3408	-2.10085
15	Unai García	1.140	0.5088	0.63122	43	Iker Muniain	4.400	6.1369	-1.73693
16	Manuel Sánchez	1.260	0.4769	0.78314	44	Lucas Torró	0.450	2.4199	-1.96993
17	Óscar Mingueza	1.140	2.8506	-1.71061	45	Óscar Rodríguez	1.150	1.9361	-0.78607
18	Omar Alderete	1.200	0.8220	0.37800	46	Williot Swedberg	0.240	0.3711	-0.13113
19	Fabrizio Angileri	0.600	0.3583	0.24170	47	Iddrisu Baba	0.820	0.2510	0.56902
20	Marcao	4.170	2.4263	1.74375	48	Oliver Torres	7.450	4.7290	2.72096
21	Nemanja Gudelj	2.640	3.2933	-0.65327	49	Gonzalo Villar	2.500	2.7213	-0.22125
22	Nianzou	4.170	3.2644	0.90557	50	Martín Hongla	0.940	-0.2291	1.16914
23	Víctor Chust	0.240	0.3671	-0.12707	51	Álvaro Aguado	0.155	1.3164	-1.16141
24	Mamadou Mbaye	0.210	0.0461	0.16393	52	Anuar	1.500	0.1287	1.37129
25	Santiago Arzamendia	0.200	0.0386	0.16141	53	Nicolás González	0.600	1.9871	-1.38713
26	Lucas Olaza	1.000	1.0970	-0.09700	54	Hugo Guillamón	1.920	3.1928	-1.27277
27	Brian Oliván	0.540	-0.0672	0.60716	55	Gonzalo Melero	0.540	2.2673	-1.72728
56	César de la Hoz	0.380	-0.1932	0.57320					
57	Randy Ntejka	0.230	0.4071	-0.17705					
58	Mariano	8.330	6.8249	1.50510					
59	Ousmane Dembelé	12.000	12.4849	-0.48490					
60	Robert Lewandowski	20.830	26.5105	-5.68047					
61	Angel Correa	3.500	2.7381	0.76192					
62	Ander Barrenetxea	0.310	-0.2169	0.52686					
63	Umar Sadiq	1.500	0.4102	1.08978					
64	Jose Luis Morales	1.920	1.8103	0.10970					
65	William José	3.000	1.4016	1.59844					
66	Isí Palazón	0.170	2.7948	-2.62481					
67	Andrés Martín	0.160	0.4338	-0.27382					
68	Jørgen Strand Larsen	0.450	1.4388	-0.98884					
69	Haris Seferovic	1.500	1.1071	0.39288					
70	Valery Fernández	0.155	0.2128	-0.05775					
71	Bryan Gil	2.310	3.5659	-1.25592					
72	Borja Mayoral	4.790	2.2572	2.53275					
73	Brian Ocampo	0.540	-0.1513	0.69127					
74	Cyle Larin	0.420	0.6788	-0.25885					
75	Hugo Duro	2.890	0.9083	1.98168					
76	Nico Ribaudo	0.460	0.7316	-0.27158					
77	Aleix Vidal	0.220	1.0995	-0.87948					
78	Joselu	1.200	6.3398	-5.13978					
79	Largie Zamazani	0.155	0.3370	-0.18198					
80	Luis Javier Suárez	1.640	0.4408	1.19922					
81	Dyego Sousa	1.040	-0.0564	1.09640					
82	Pere Milla	0.340	1.7485	-1.40849					
83	Ezequiel Ponce	1.800	-0.0247	1.82470					

Tabla 22: Predicciones para jugadores externos a la construcción del modelo. Elaboración propia

Como vemos en la Tabla 22, se ha calculado para cada jugador del fichero test la predicción de su salario en base al modelo que hemos construido anteriormente, representando en una tabla a cada jugador con su respectivo salario observado y predicho, junto con el error que se ha cometido. Con estos errores por jugador calculamos el MSE total de la predicción realizada, obteniendo como se ve en la tabla 27 un MSE de 4.69.

Analysis Variable
: MSE
Sum
4.7139796

Tabla 23: Error cuadrático medio cometido para el modelo de todos los jugadores por regresión. Elaboración propia

Con el resultado del MSE obtenido en el fichero test podemos decir que nuestro modelo presenta buena capacidad predictiva para datos que no se han tenido en cuenta para la construcción del modelo.

Realizando un vistazo a este primer método de predicción, hemos observado que los modelos para los defensas, mediocentros y delanteros presentan una variable característica de sus respectivas posiciones. En estos modelos se han obtenido unos R-Cuadrado Ajustado de 79,86%, 72,01% y 81,45%, mientras que en el modelo final se obtuvo un R-Cuadrado Ajustado de 72,6%. Observando estos resultados podemos apreciar que en el caso de los defensas y los delanteros se ha obtenido un mayor porcentaje de variabilidad explicada. Por lo tanto, en los siguientes dos métodos de predicción, se aplicarán de nuevo las técnicas necesarias a las 4 bases de datos, pero se focalizará el estudio en los jugadores en su conjunto ya que es el conjunto de datos que nos permite realizar predicciones.

5.2 Árboles de regresión y Random Forest

Vamos a dividir esta sección en 2 partes. Primeramente, vamos a construir un árbol de regresión para cada una de las 4 bases de datos que disponemos, donde analizaremos diversos aspectos: qué modelo es el más indicado según las necesidades que se requieran, estudiar la importancia de las variables y las reglas que definen las hojas de los modelos seleccionados y por último, exclusivamente para la base de datos de todos los jugadores, realizar podas para simplificar los modelos más complejos. Una vez realizados estos aspectos, se utilizarán técnicas de random forest para comprobar si se consigue una mejoría en los modelos obtenidos.

Por la misma razón que en el método de regresión lineal múltiple, no se va a realizar ninguna partición de los datos para los Defensas, Mediocentros y Delanteros.

Para estas 3 bases de datos, vamos a comenzar por la construcción de 3 árboles diferenciados por el número de observaciones mínimas que va a presentar el árbol en cada nodo. Este aspecto se denomina minibucket. Los minibucket que se van a utilizar son del 10%, 5% y 2,5%. En la siguiente tabla quedan recogidos los resultados obtenidos en cada uno de los minibucket mencionados:

Defensas			
	10%	5%	2,5%
Nº Hojas	6	10	19
Nº Variables	3	6	12
R-Cuadrado	0,6261	0,65	0,8014
MSE	4,27	4	2,27

Tabla 24: Resultados según el minibucket para los Defensas. Elaboración propia

Mediocentros

	10%	5%	2,5%
Nº Hojas	6	9	16
Nº Variables	4	6	10
R-Cuadrado	0,524	0,7	0,83
MSE	17,16	10,8	6,02

Tabla 25: Resultados según el minibucket para los Mediocentros. Elaboración propia

	Delanteros		
	10%	5%	2,5%
Nº Hojas	4	12	18
Nº Variables	1	7	11
R-Cuadrado	0,5988	0,7558	0,769
MSE	9,37	5,7	5,39

Tabla 26: Resultados según el minibucket para los Delanteros. Elaboración propia

Como podemos ver, a medida que disminuimos el número de observaciones mínimo que se recoge en cada nodo, obtenemos modelos muy diferentes en cuanto al nivel de complejidad del modelo y su capacidad de ajuste. Por esta razón, se seleccionará el minibucket dependiendo de las necesidades que se precisen en cuanto a la relación complejidad-capacidad de ajuste.

Si se premia la sencillez del modelo, se seleccionaría el 10% de minibucket para los defensas, ya que al reducirlo al 5% prácticamente no hay una mejoría suficiente en su capacidad de ajuste cuando sí que aumenta notablemente su complejidad. Obtendríamos así un modelo con una capacidad de ajuste media, pero bastante sencillo al solo necesitar 3 variables. Para los mediocentros, al seleccionar el minibucket del 10%, se obtendría un modelo con una capacidad de ajuste poco suficiente, luego si observamos los resultados obtenidos para un 5% de minibucket, vemos que no aumenta en gran medida la complejidad del modelo, al solo aumentar en 2 variables con respecto al modelo anterior, y aumentar notoriamente su capacidad de ajuste. Por tanto, para los mediocentros seleccionaríamos el minibucket del 5% si necesitáramos un modelo sencillo, pero que nos asegure una capacidad de ajuste moderada. Finalmente, para los delanteros, vemos que con 1 única variable se es capaz de explicar casi el 60% de la variabilidad de la variable respuesta Salario para el minibucket del 10%. Este sería el modelo seleccionado si se necesitara un modelo sencillo, ya que si reducimos el número de observaciones mínimas al 5% se mejoraría bastante su capacidad de ajuste, pero aumentaría también notoriamente su complejidad.

Por otro lado, si se precisa de maximizar la capacidad de ajuste del modelo, sin tener en cuenta la complejidad del mismo, obtendríamos otras conclusiones. Para los defensas, es bastante notorio el cambio en la capacidad de ajuste que se obtiene al reducir el minibucket del 5% al 2,5%. Al hacerlo, aumenta el R-Cuadrado un 15% y el MSE se reduce a casi la mitad, luego este modelo nos ofrecería una muy buena capacidad de ajuste, siendo este el seleccionado si deseamos maximizar este aspecto. Pasando a los mediocentros, el modelo con el menor minibucket es el que mejor capacidad de ajuste presenta, superando el 80% de R-Cuadrado y reduciendo también casi a la mitad su error cuadrático medio, lo que nos indica que presenta una habilidad predictiva bastante buena. Por último, para los delanteros, vemos que reducir el

número mínimo de observaciones por nodo del 5% al 2,5% no provoca apenas mejoría en el modelo. Por tanto, sería recomendable utilizar el 5% de minibucket, ya que es un modelo con prácticamente la misma capacidad predictiva pero menor complejidad que el generado por el del 2,5%.

Es importante destacar que para bases de datos de tan baja dimensión, no es recomendable establecer un minibucket muy pequeño, ya que el número de observaciones mínimo por nodo será muy pequeño y podríamos tener varios problemas. Al permitir que cada nodo tenga muy pocas observaciones, el árbol puede terminar siendo demasiado complejo y adaptarse excesivamente a los datos. Esto puede dar lugar a un modelo que no generaliza bien a nuevos datos y que se ajusta demasiado a las peculiaridades específicas de la muestra. A esto se le denomina sobreajuste.

Otro aspecto que provoca establecer un minibucket pequeño para bases de datos con pocas observaciones es la pérdida de estabilidad / robustez del modelo. Los nodos del árbol generado en estos casos pueden ser más sensibles a variaciones en los datos, lo que daría lugar a un modelo menos fiable. Esto está directamente relacionado con el sobreajuste del modelo, ya que como se ha explicado antes, al adaptarse excesivamente a los datos de la muestra, si se provocan pequeñas variaciones en los datos, pueden generar grandes variaciones en el modelo final, obteniendo un modelo poco estable.

Si observamos los distintos modelos generados por los 3 minibuckets, vemos que se repiten ciertos patrones en su selección de variables. Para los defensas, si observamos la importancia de las variables de los distintos modelos obtenemos lo siguiente:

Árbol del 10% de minibucket

Popularidad	BSR	DAereosP
1142.521686	19.725589	4.134353

Árbol del 5% de minibucket

Popularidad	DisBloq	BSR	DAereosP	Equipo	Dpj
1142.521686	34.445747	19.725589	7.300146	4.900793	1.046437

Árbol del 2,5% de minibucket

Popularidad	PasesultTercio	DisBloq	Goles	BSR	DAereosG
1142.718829	244.187122	34.445747	30.284379	22.914833	6.925401
DAereosP	Tarja	VReg	PAciertoPases	FRec	Dpj

El resultado que nos ofrece el programa no nos permite sacar una conclusión en claro. Es verdad que se entiende que, a mayor valor, mayor importancia tiene la variable, pero no está claro del todo. Para clarificarlo, vamos a calcular los porcentajes relativos de cada variable y representarlos en una gráfica.

Árbol del 10% de minibucket

Popularidad	BSR	DAereosP
97.9543624	1.6911779	0.3544597

Árbol del 5% de minibucket

Popularidad	DisBloq	BSR	DAaerosP	Equipo	Dpj
94.42793111	2.84689619	1.63029426	0.60334760	0.40504420	0.08648666

Árbol del 2,5% de minibucket

Popularidad	PasesUltTercio	DisBloq	Goles	BSR	DAaerosG
76.54790103	16.35748988	2.30743519	2.02867546	1.53500782	0.46391547
DAaerosP	Tarja	VReg	PAciertoPases	FRec	Dpj
0.27695003	0.12172213	0.09870425	0.09630619	0.09579435	0.07009822

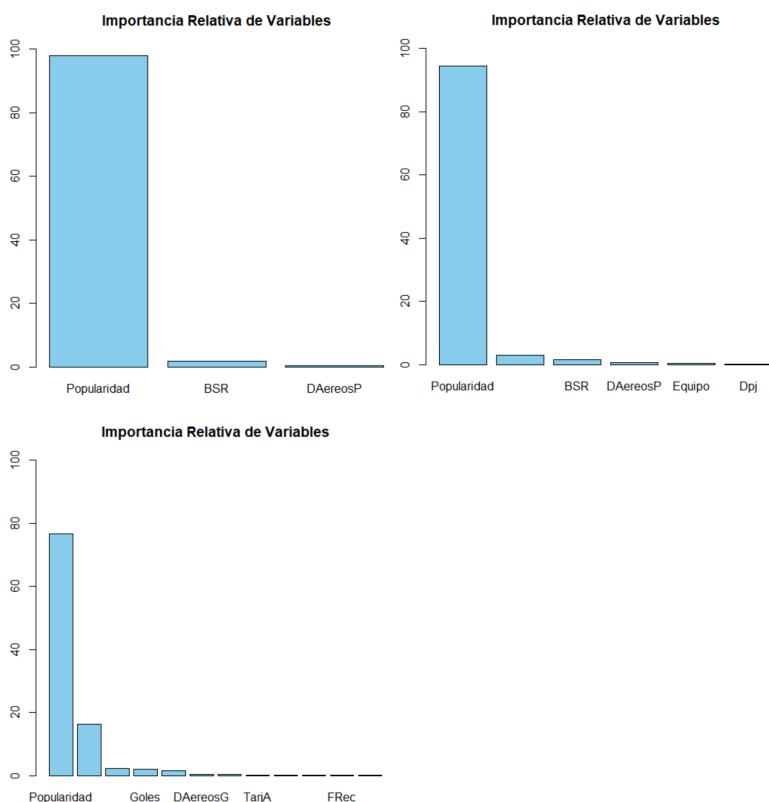


Figura 6: Importancia de variables en los árboles para los Defensas según el minibucket. Elaboración propia

Vemos que la variable Popularidad se repite como la más importante en los 3 modelos de defensas. El resto de variables son mucho menos importantes en comparación a la popularidad, salvo para el modelo del 2,5% que los PasesUltTercio adquieren mucha más importancia.

Si comparamos estos resultados con los obtenidos para los modelos de regresión lineal múltiple, vemos que la Popularidad en los árboles de regresión es mucho más importante que en los modelos de regresión lineal, en los cuales la variable PasesUltTercio era la más relevante.

Para comprender como funciona los árboles de regresión vamos a representar gráficamente el árbol más sencillo de los defensas, y posteriormente analizar sus reglas.

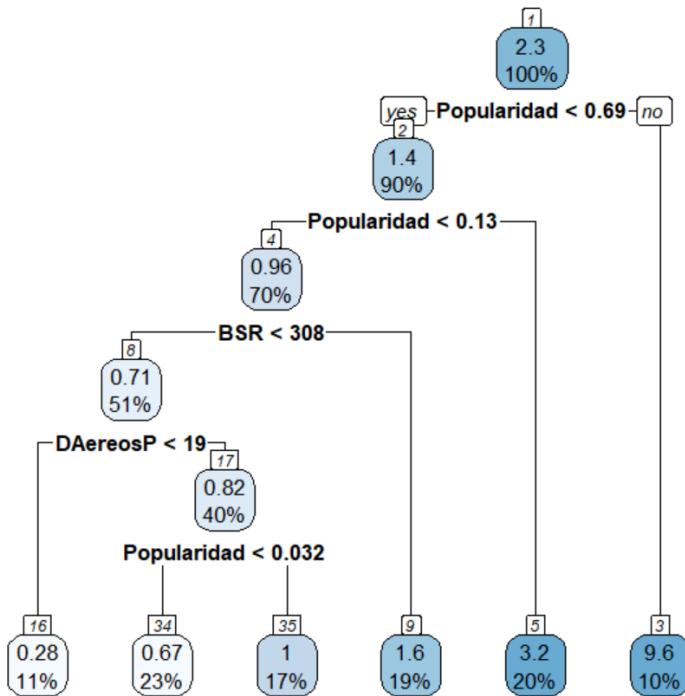


Figura 7: Árbol con 10% de minibucket sobre los Defensas. Elaboración propia

En esta Figura 7 vemos el árbol de regresión para el minibucket del 10%. Se observan los diferentes nodos de decisión (numerados por 1, 2, 4, 8 y 17) los cuales representan las condiciones que se evalúan para decidir qué camino seguir. Estas condiciones suelen estar definidas en forma de reglas lógicas, donde se especifica una relación entre una variable y un valor, ramificando de esta manera el árbol hacia diferentes nodos. Estas reglas se utilizan para dividir el conjunto de datos en subconjuntos. Las últimas ramificaciones se denominan hojas finales (numeradas en este caso por los números 16, 34, 35, 9, 5 y 3), las cuales contienen las predicciones realizadas.

La manera de seleccionar las variables de las condiciones de cada nodo se basa en observar que variable es la que produce una mayor mejora en el MSE. Por ejemplo, mediante la siguiente salida, vemos que en la hoja 1, la variable que hace disminuir en mayor medida al MSE es Popularidad. Concretamente provoca una mejoría del 54,6%.

```

Node number 1: 163 observations,      complexity param=0.5461535
mean=2.299141, MSE=11.42813
left son=2 (146 obs) right son=3 (17 obs)
Primary splits:
Popularidad      < 0.6858545 to the left,  improve=0.5461535, (0 missing)
Equipo           splits as RLLL, improve=0.5155281, (0 missing)
PasesUltTercio   < 251.5      to the left,  improve=0.3145119, (0 missing)
Goles            < 6.5       to the left,  improve=0.2942279, (0 missing)
PAciertoPases    < 84.26338  to the left,  improve=0.2601311, (0 missing)
  
```

De esta manera se seleccionan al resto de variables para cada nodo.

En la siguiente salida vemos las reglas de 2 de las hojas finales del árbol, sus respectivas predicciones y el porcentaje de observaciones que pertenecen a esa hoja final sobre el conjunto de datos.

```

Rule number: 34 [Salario=0.671486486486486 cover=37 (23%)]
  Popularidad< 0.6859
  Popularidad< 0.1266
  BSR< 307.5
  DAereoSP>=18.5
  Popularidad< 0.03173
  
```

Rule number: 3 [salario=9.62058823529412 cover=17 (10%)]
 Popularidad>=0.6859

Para la hoja final 34 vemos las condiciones que toman aquellos jugadores que pertenecen a esa hoja. Si un defensa tiene un valor de balones sueltos recuperados (BSR) menor a 307.5, los despejes aéreos perdidos mayores o iguales a 18.5 y una popularidad inferior a 0.03173 (medida en millones de unidades), el modelo predice que ese jugador tendrá un salario aproximado de 0.67 millones de euros. A esta hoja pertenecen un total de 37 defensas.

Por el contrario, vemos que para aquellos defensas que tienen una popularidad muy alta, concretamente mayor o igual a 0.6859, no se necesitan más variables para estimar su salario, obteniendo así una estimación de 9.62 millones de euros. A esta hoja pertenecen 17 defensas.

Para los mediocentros y delanteros se ha seguido la misma metodología y enfoque que para los árboles de regresión de los defensas. Los árboles obtenidos para estas dos bases de datos pueden verse en el anexo del trabajo. A rasgos generales, para los mediocentros la variable más importante es PasesUltTercio, seguida de Popularidad. Este resultado es el mismo que el obtenido en el modelo de regresión lineal, luego nos indica que es un resultado fiable. Estos resultados se pueden observar en las tablas 27 y 28 del anexo. Por otro lado, para los delanteros, la variable Equipo es la más importante del modelo, seguida de Popularidad y PasesClave, siendo este también el mismo resultado que obtuvimos para el modelo de regresión lineal, indicándonos que también es fiable que estas sean las variables más determinantes para determinar el valor de mercado de un delantero. A su vez, estos resultados los podemos observar en las tablas 29 y 30 del anexo.

Ahora pasamos a la construcción del árbol de regresión para la base de datos Jugadores_Combinados, realizando primeramente una partición de los datos, de manera que el 80% formen el fichero entrenamiento y el 20% restante el de prueba. Una vez realizado la partición, se construyen los distintos árboles de regresión en base a los distintos minibuckets. Los resultados de los modelos quedan recogidos en la siguiente tabla.

	Jugadores Combinados			
	10%	5%	2,5%	1%
Nº Hojas	6	11	21	47
Nº Variables	3	7	11	14
R-Cuadrado train	0,529	0,569	0,7	0,85
R-Cuadrado test	0,539	0,552	0,69	0,765

MSE	9,57	8,77	6,03	2,97
-----	------	------	------	------

Tabla 31: Resultados según el minibucket para todos los jugadores. Elaboración propia

Como podemos ver, el modelo mejora en cuanto a su capacidad predictiva a medida que reducimos el valor del minibucket.

De igual forma que para las bases de datos anteriores, existen 2 opciones de decisión a la hora de qué árbol escoger según las necesidades que se tengan. Si se precisa de un modelo sencillo, el cual sea capaz de explicar el salario de un jugador con el menor número de variables posibles, se tendrían en cuenta los modelos del 10% y del 5%. Entre estos dos, nos quedaríamos con el modelo del 10%, ya que, siguiendo el criterio de parsimonia, que nos dice que, dados dos modelos con una capacidad de ajuste similar, nos quedamos con el más sencillo de los dos. El aspecto negativo de seleccionar el modelo del 10% es que se queda demasiado corto en cuanto a su capacidad predictiva, al tener un R Cuadrado ligeramente superior al 50% y un valor del MSE ligeramente elevado. Por ello, sería más recomendable seleccionar el modelo del minibucket 2,5%, ya que, a pesar de ser un modelo algo más complejo que el anterior, obtenemos un R Cuadrado relativamente bueno, reducimos su MSE y se observa una clara estabilidad del modelo frente a datos externos al probar el modelo con los datos del fichero test y obtener prácticamente el mismo R Cuadrado que para el train. No se seleccionaría el modelo con el menor minibucket ya que, a pesar de obtener unos estadísticos de capacidad de ajuste mejores, este modelo es demasiado complejo al tener 47 hojas y estar formado por 14 de las 15 variables de la base de datos. Además, no es igual de estable que el anterior, al variar un 9% de su R Cuadrado test frente al del train, lo que nos estaría diciendo que el modelo podría estar sobreajustado a los datos de entrenamiento y por tanto ser menos estable.

El modelo del 2,5% tiene la siguiente estructura:

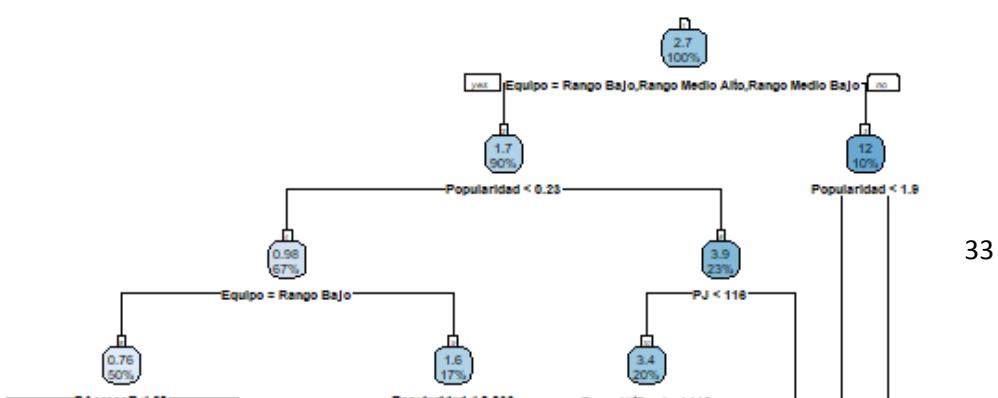


Figura 8: Árbol con 2.5% de minibucket sobre todos los jugadores. Elaboración propia

Para explicar cómo ha realizado las divisiones, vamos a centrarnos en la primera y última hoja del árbol, coincidiendo estas con los jugadores peores y mejores valorados según el modelo.

Para la primera hoja o subconjunto de jugadores peor valorados, se han elegido las siguientes reglas:

```
Rule number: 64 [Salario=0.243 cover=15 (5%)]
Equipo=Rango Bajo,Rango Medio Alto,Rango Medio Bajo
Popularidad< 0.2266
Equipo=Rango Bajo
DAereosP< 27.5
TarjR< 0.5
FCom< 36
```

En este subconjunto de jugadores han sido agrupados el 5% de ellos.

En cambio, para la última hoja o para el subconjunto de los jugadores mejores valorados, se han elegido las siguientes:

```
Rule number: 7 [Salario=18.615 cover=12 (4%)]
Equipo=Rango Alto
Popularidad>=1.911
```

El 4% de los jugadores pertenecen a este subconjunto.

La importancia de las variables que han sido seleccionadas en este modelo se resume en la siguiente tabla:

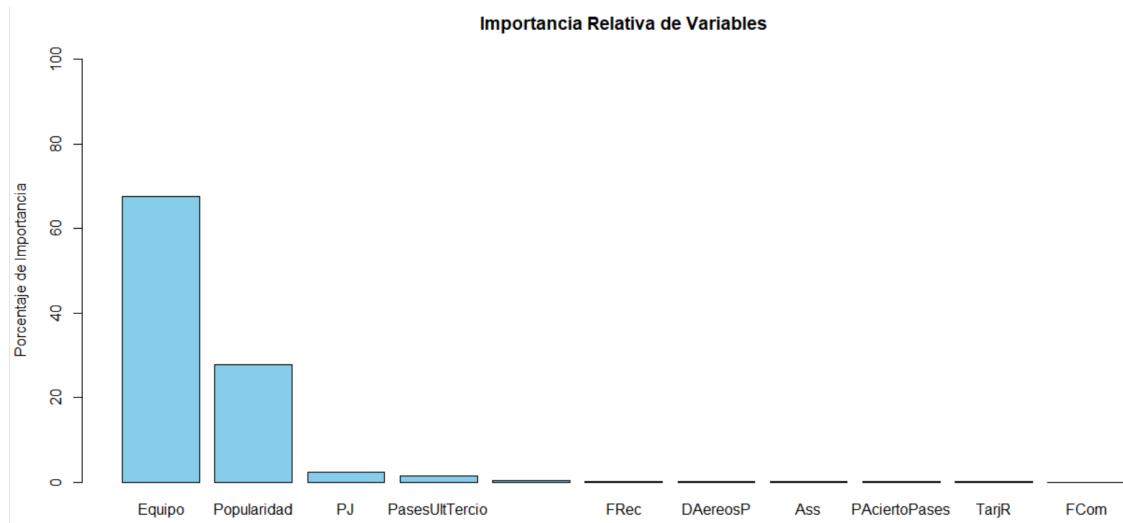


Figura 9: Importancia de variables para todos los jugadores. Elaboración propia

El Equipo es la variable más importante, con un 67.5%, seguido de la Popularidad, con un porcentaje del 27.7%. El resto de variables tienen una importancia mucho menor en el modelo.

Este resultado es ciertamente diferente al obtenido por el modelo de regresión lineal, en el cual la variable PasesUltTercio era la variable más significativa. La Popularidad se mantiene en la misma importancia para los dos modelos, siendo la segunda más importante, pero en este árbol, se le da mucha más importancia a pertenecer a un equipo en un rango determinado.

Ahora, una vez explicado el funcionamiento del árbol, vamos a realizar la poda en el mismo. Este proceso trata de estudiar la posibilidad de eliminar alguna(s) hoja(s) del árbol construido con el objetivo de reducir el posible sobreajuste.

En la teoría, para realizar la poda de un árbol se debe generar la secuencia de valores críticos de *alpha* que dan lugar a un crecimiento/decrecimiento del árbol. Este valor determina la complejidad del árbol, oscilando entre 0 y 1 (el 0 se corresponde a un árbol de profundidad máxima y el 1 a un árbol sin divisiones). La secuencia de valores de *alpha* queda recogida en la siguiente tabla:

	CP	nsplit	rel error	xerror	xstd
1	4.699156e-01	0	1.0000000	1.0072555	0.2508668
2	1.159332e-01	1	0.5300844	0.6000117	0.1337405
3	7.013599e-02	2	0.4141512	0.6046502	0.1453730
4	1.640425e-02	3	0.3440152	0.4858403	0.1261565
5	8.399378e-03	4	0.3276109	0.4811255	0.1264041
6	4.894635e-03	5	0.3192115	0.4769984	0.1259069
7	4.839589e-03	6	0.3143169	0.4806758	0.1256493
8	3.623595e-03	7	0.3094773	0.4814599	0.1256231
9	2.403318e-03	8	0.3058537	0.4843213	0.1262312
10	2.332272e-03	9	0.3034504	0.4860720	0.1262944
11	1.580895e-03	10	0.3011181	0.4856008	0.1262598
12	9.223041e-04	11	0.2995372	0.4853814	0.1262501
13	5.412341e-04	12	0.2986149	0.4844758	0.1262502
14	4.287568e-04	13	0.2980737	0.4841766	0.1262509
15	3.346375e-04	14	0.2976449	0.4849154	0.1262400
16	2.419898e-04	15	0.2973103	0.4853380	0.1262387
17	1.628787e-04	16	0.2970683	0.4853488	0.1262432
18	1.338140e-04	17	0.2969054	0.4855312	0.1262419
19	6.605776e-05	18	0.2967716	0.4859771	0.1262378
20	5.124879e-05	19	0.2967056	0.4860838	0.1262380
21	0.000000e+00	20	0.2966543	0.4860582	0.1262378

Tabla 32: Secuencia de valores de alpha para el árbol del 2.5%. Elaboración propia

Dado que en ocasiones no resulta fácil identificar el valor de alpha indicado para cortar el árbol, vamos a recurrir a la siguiente representación gráfica, donde nos permite observar la información de la tabla, junto con la evolución de la aportación a la reducción del error relativo a medida que se le van añadiendo hojas al árbol.

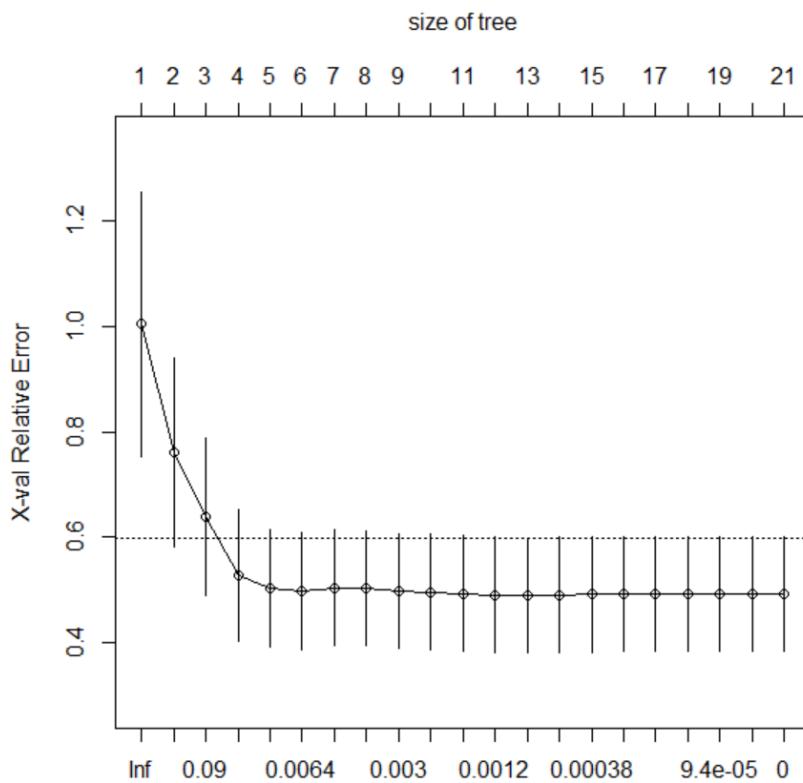


Figura 10: Evolución de la aportación a la reducción del error relativo según el tamaño del árbol. Elaboración propia

En esta Figura 10, vemos que a partir de un árbol de 4 hojas no se produce una mejora significativa en la reducción del error. Por lo tanto, vamos a podar el árbol de tal manera que nos queden 4 hojas finales.

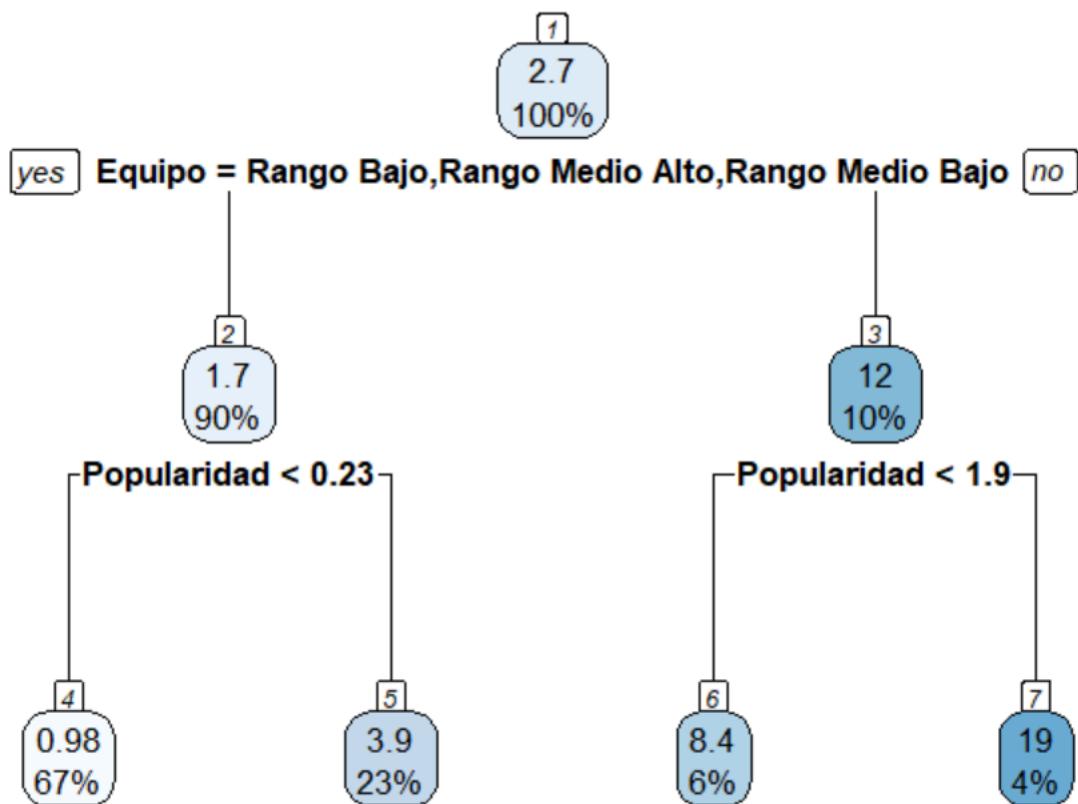


Figura 11: Árbol podado para todos los jugadores. Elaboración propia

Como vemos el árbol se ha reducido bastante en dimensión, simplificando el modelo hasta tener 2 variables y únicamente 4 hojas finales, en comparación a las 21 hojas y 11 variables que tenía anteriormente. Nuevamente las variables seleccionadas son el Equipo y la Popularidad del jugador, obteniendo los siguientes porcentajes de importancia.

Equipo	Popularidad
71.63514	28.36486

Los estadísticos de ajuste de este nuevo árbol podado son los siguientes:

Nº Hojas	4
Nº Variables	2
R-Cuadrado train	0,655
R-Cuadrado test	0,655
MSE	6,9

Tabla 33: Estadísticos de ajuste para el árbol podado. Elaboración propia

Como vemos, obtenemos un R Cuadrado en ambos ficheros del 65.5%, únicamente un 5% inferior al árbol antes de ser podado. A su vez, si relacionamos la sencillez que hemos obtenido al podarlo frente al anterior árbol, pasamos de 11 a 2 variables y 21 a 4 hojas. Esto nos dice que la poda ha logrado simplificar el modelo sin comprometer

significativamente su capacidad de ajuste. El modelo se ha vuelto más fácil de interpretar y se ha reducido el riesgo de sobreajuste a los datos de entrenamiento.

Ahora vamos a pasar a la segunda parte de este método de predicción, donde vamos a aplicar técnicas de random forest con el objetivo de ver si conseguimos mejorar el modelo.

Se comienza creando un primer modelo de RF con los valores por defecto de mtry y ntree, con un tamaño de hoja equivalente al 5% de los datos para posteriormente modificarlo y ver cómo afecta este aspecto a los resultados obtenidos. Obtenemos por tanto la siguiente información de este primer modelo RF:

```
call:  
randomForest(formula = Salario ~ . - Posicion - Titular - Minutos -  
Jugador, data = Jugadores_Combinados_train, nodesize = ceiling(0.05 *  
nrow(Jugadores_Combinados_train)))  
    Type of random forest: regression  
    Number of trees: 500  
No. of variables tried at each split: 5  
  
    Mean of squared residuals: 7.015389  
    % Var explained: 65.52
```

Vemos que se han construido un total de 500 árboles y se han seleccionado aleatoriamente 5 variables en cada división. Este RF da lugar a un modelo con un R^2 de 65,52% estimado mediante las observaciones OOB. Esto quiere decir que el modelo es capaz de explicar el 65,52% de la variabilidad en la variable Salario, utilizando datos no empleados durante el entrenamiento (OOB = Out of Bag).

Ahora vamos a analizar el efecto que tiene el parámetro ntree sobre el R^2 .

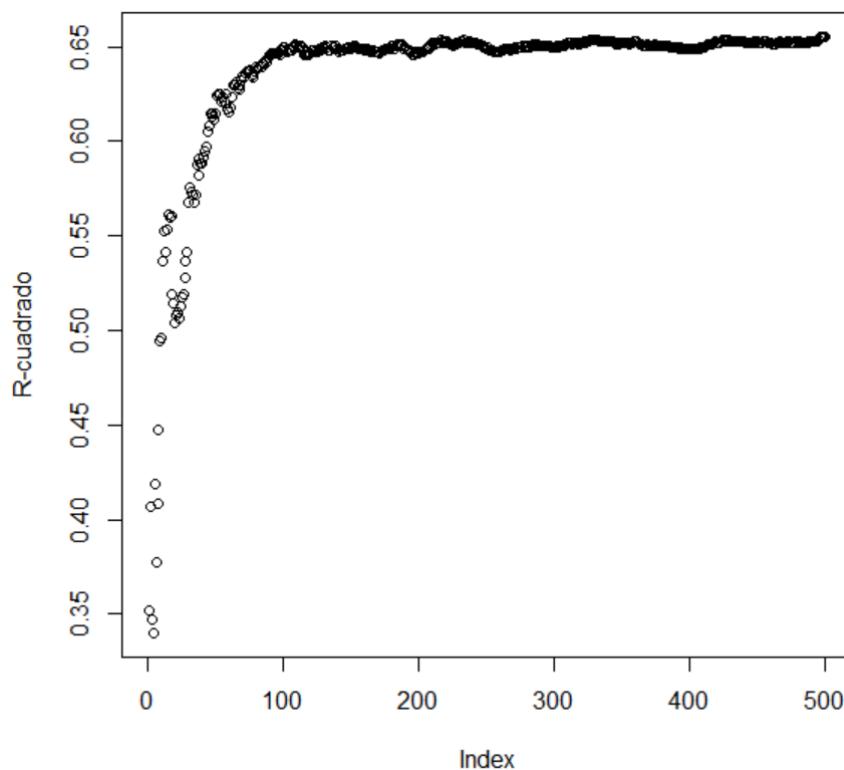


Figura 12: Evolución de la aportación al incremento del R-Cuadrado según el número de árboles. Elaboración propia

Vemos que prácticamente a partir de los 200 árboles no se produce ninguna mejora significativa en el aumento del R^2 , volviéndose estable este valor.

Si modificamos el número mínimo de observaciones por hoja, reduciéndolos a 2,5% y a 1% no se produce ningún cambio en su R^2 , luego nos quedamos con el 5% para el minibucket ya que generará modelos más simples, con menor riesgo de sobreajuste.

El último parámetro que nos falta por analizar es el `mtry`. Algunos autores sugieren probar 4-5 valores diferentes entre 2 y el número total de variables predictoras. Mediante el uso de un bucle, pasamos a analizar qué valor es el más apropiado para nuestro modelo, obteniendo la siguiente gráfica:

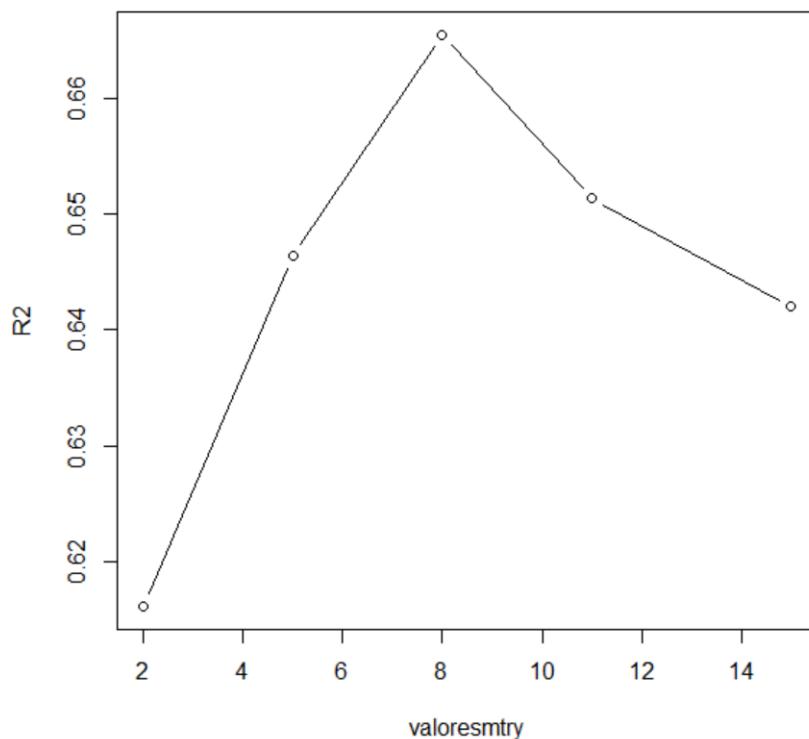


Figura 13: Evolución de la aportación al incremento del R-Cuadrado según el `mtry`. Elaboración propia

Observamos que el valor del `mtry` que maximiza el R^2 es 8. Si imprimimos el árbol obtenido por este valor del `mtry` obtenemos la siguiente salida:

```
call:
randomForest(formula = salario ~ . - Posicion - Titular - Minutos -
Jugador, data = Jugadores_Combinados_train, nodesize = ceiling(0.05 *
nrow(Jugadores_Combinados_train)), ntree = 200, mtry = valoresmtry[i])
      Type of random forest: regression
                  Number of trees: 200
No. of variables tried at each split: 8

Mean of squared residuals: 6.806172
  % Var explained: 66.55
```

Vemos que la capacidad de ajuste del modelo ha sufrido una mínima mejoría, al pasar de 65% de R^2 a 66% y reducir ligeramente su MSE.

La importancia de las variables del modelo RF quedan recogidas en la

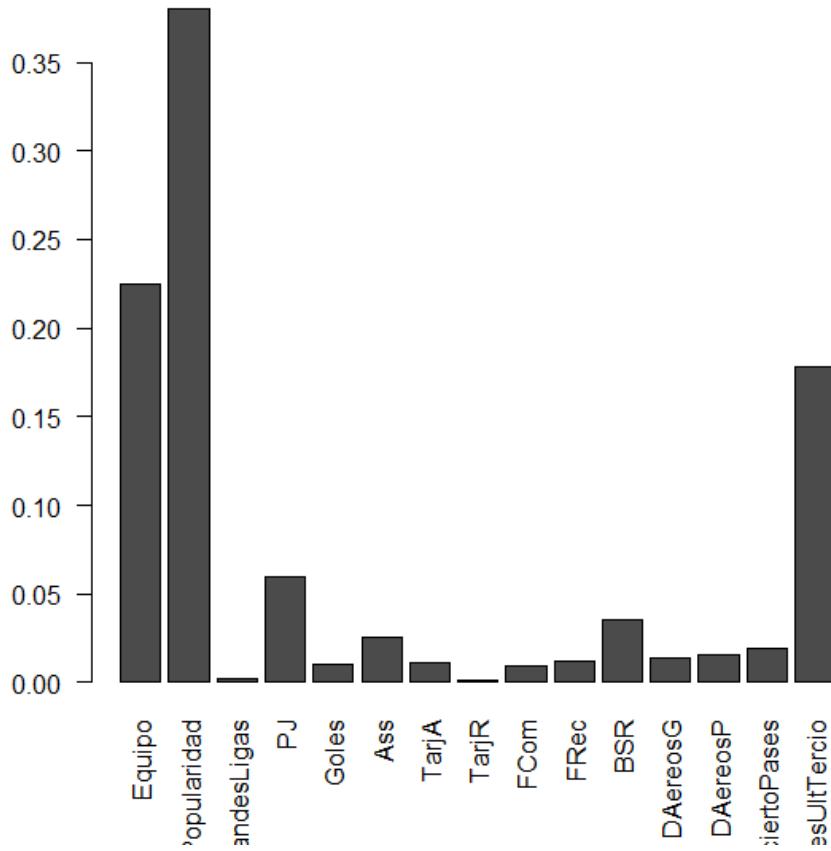


Figura 14: Importancia de variables del modelo Random Forest. Elaboración propia

Vemos que las variables más importantes son Popularidad, seguido de Equipo y PasesUltTercio. En el modelo de regresión lineal, la variable más importante era PasesUltTercio, mientras que en este pasa a estar en tercer lugar de importancia. En este caso la variable Popularidad aporta casi el 40% de la información del modelo, lo que nos indica que es un factor muy determinante a la hora de analizar el valor de mercado de un jugador.

Por último, vamos a comprobar cómo se comporta este modelo en los ficheros train y test que dividimos anteriormente, obteniendo lo siguiente:

Árbol RF	
R-Cuadrado train	0,9
R-Cuadrado test	0,82

Tabla 34: Estadísticos de ajuste obtenidos en los distintos ficheros con el modelo RF. Elaboración propia

Vemos que el modelo construido de RF obtiene unos estadísticos de ajuste, tanto en el train como en el test, muy buenos, lo que nos indica que el modelo generado presenta una gran capacidad de ajuste de generalización a nuevos datos.

5.3 Método de los k-vecinos más cercanos: KNN

Como último método de predicción tenemos el método KNN, mediante el cual realizaremos predicciones aprovechando similitudes entre observaciones de datos, en nuestro caso jugadores. Esto significa que las predicciones que se generan provienen directamente de las observaciones del entrenamiento, sin estimar ningún modelo, realizando para ello estimaciones parciales de forma local para cada observación. A este tipo de métodos se les atribuye el nombre de *Aprendizaje Perezoso*.

Por esta razón, este tipo de métodos son mucho menos sensibles a la falta de datos comparado con la regresión lineal y los árboles de regresión. Por ello, vamos a dividir las 4 bases de datos en entrenamiento y prueba. Artículos como Sotojo y otros (2023) o Cheamanunkul y otros (2014), corroboran que el método de k -NN vecinos es aplicable a bases de datos de pequeña dimensión. Se empleará el cálculo del MSE como forma de comparar con los modelos anteriores si este método se equivoca más, menos o en la misma medida.

En esta sección, vamos a contemplar 2 maneras de aplicar el método de los k -NN vecinos. La primera consistirá en analizar las correlaciones de las variables independientes frente a la variable objetivo Salario de manera que las 2 variables independientes que estén más relacionadas con la variable objetivo se seleccionaran como coordenadas para realizar las predicciones.

Por otro lado, ya que se dispone de un mayor número de variables, se estudiará la posibilidad de realizar una reducción de las variables independientes mediante el cálculo de componentes principales y aquellos dos que expliquen un mayor porcentaje de variabilidad explicada de la variable Salario, serán los seleccionados como coordenadas.

Estas dos opciones de desarrollo se realizarán para un bucle de 100 muestras en el cual se calculará en MSE medio de esas 100 muestras para así concluir qué opción es más viable para cada base de datos. Se realiza esta diferenciación por dos motivos. El primero para ver si aprovechando un mayor número de variables mediante los componentes principales se consigue reducir el error cometido. El segundo es que la primera opción de desarrollo es adecuada para aquellas bases de datos que tengan 2 variables muy correlacionadas con la variable objetivo, pero, por ejemplo, en el caso de los delanteros, solo hay 1 variable que tenga un coeficiente de correlación de Pearson superior a 0,6. Por tanto si se selecciona una variable con una correlación media con respecto a la objetivo afectará a la capacidad de predicción del modelo.

Comenzando con la primera opción de desarrollo para los defensas, mediante la Tabla 10 mostrada en el método de regresión, vemos que las 2 variables más correlacionadas con el Salario son Popularidad y PasesUltTercio. Vamos a utilizar 3 como número de vecinos por la siguiente cuestión: números de vecinos pequeños, como podrían ser 1 o 2 vecinos, provocan estimaciones con poca varianza, pero gran sesgo; y por el contrario números de vecinos grandes, como pueden ser 7 u 8, provocan estimaciones con poco sesgo, pero gran varianza. Por tanto, de entre los valores intermedios que quedan: 3, 4 o 5 vecinos; elegimos 3 como número de

vecinos. Para explicar por qué se seleccionan 3 vecinos vamos a utilizar la distribución espacial de las observaciones de una de las muestras seleccionadas aleatoriamente.

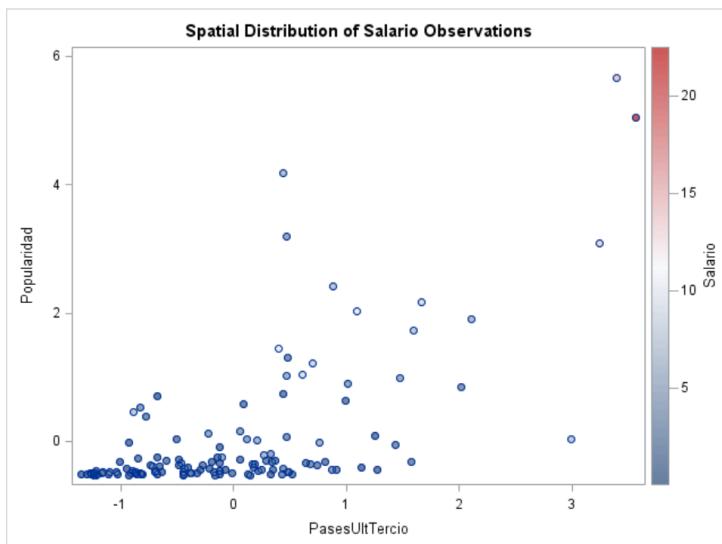


Figura 15: Distribución espacial de las observaciones. Elaboración propia

Como vemos en esta Figura 15, la mayoría de las observaciones de la muestra de entrenamiento se acumulan en la parte inferior izquierda. Estas observaciones representan niveles de popularidad y recuento de número de pases muy bajos, relacionadas a su vez con Salarios bajos (algunas de las observaciones tienen valores negativos por estar estandarizados los datos, queriendo decir que se encuentran por debajo de la media en su respectiva variable). Por otro lado, si analizamos la distribución de las observaciones con mayor salario (cuanto más rojo sea el punto mayor salario), vemos que no hay muchas observaciones similares entre sí. Esto es porque existen pocos defensas con salarios muy altos, solo el 3,7% de los jugadores superan los 10 millones salariales. Por tanto, si escogemos 4 o 5 vecinos, a la hora de predecir el salario para los defensas que pertenezcan al test, se van a tomar observaciones muy alejadas del train para predecir el salario de los jugadores del test, cometiendo por tanto un error mayor cuanto más vecinos se utilicen.

Mediante el uso de una macro que repite el proceso 100 veces, obtenemos el MSE medio de las 100 muestras aleatorias para la base de datos de los defensas. Las variables independientes han sido estandarizadas previamente, para que las distancias sean equiparables. Obtenemos en este proceso un MSE para las 100 muestras de 4,48.

Una vez calculado el error cometido tomando como coordenadas aquellas 2 variables que estén más correlacionadas con el salario, pasamos a analizar si el implementar componentes principales reduce el error cometido. Para nuestra base de datos de defensas, obtenemos la siguiente tabla de autovalores.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	9.52782737	6.76099377	0.4537	0.4537
2	2.76683359	0.46015123	0.1318	0.5855
3	2.30668236	1.15208990	0.1098	0.6953
4	1.15459246	0.33606193	0.0550	0.7503
5	0.81853053	0.06537777	0.0390	0.7893
6	0.75315275		0.0359	0.8251

Tabla 35: Autovalores para la matriz de correlación de los Defensas. Elaboración propia

Con esta Tabla 35 vemos el porcentaje acumulado de la varianza de las variables originales que queda explicado para cada número de componentes. Como en el método k -NN se requieren 2 coordenadas, se seleccionarán aquellas dos componentes que mayor proporción individual de varianza expliquen, luego nos quedamos con las componentes 1 y 2, las cuales explican casi un 60% de varianza de las variables originales.

Realizando de nuevo el bucle de 100 muestras aleatorias estableciendo como coordenadas a las componentes 1 y 2, se obtiene un MSE medio de 9,7. El error cometido en este caso es casi el doble que en el anterior. Esto puede deberse a que las componentes que obtenemos para la matriz de correlación de esta base de datos solo explican el 58,55% de la variabilidad de las variables independientes. Esto significa que hay una parte significativa de variabilidad que no se está considerando al utilizar solo estas 2 componentes, siendo por tanto razonable el aumento del MSE debido a esa falta de información / pérdida de variabilidad explicada.

Por tanto, como coordenadas para realizar el método k -NN para los defensas se seleccionan las variables Popularidad y PasesUltTercio, al ser estas las que menor error cometan.

Pasando al análisis de los mediocentros, obtenemos que para la primera opción de desarrollo se deberían seleccionar como coordenadas de nuevo a las variables Popularidad y PasesUltTercio. Podemos comprobar las correlaciones de esta base de datos en la Tabla 10. Realizando de nuevo el bucle, obtenemos un MSE medio para las 100 muestras aleatorias de 12,3.

Una vez calculado el error medio cometido con las 2 variables más correlacionadas con la objetivo, pasamos a analizar los autovalores que se generan para la matriz de datos de los mediocentros.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	9.87424705	7.17527753	0.4702	0.4702
2	2.69896952	0.62739992	0.1285	0.5987
3	2.07156960	0.91120242	0.0986	0.6974
4	1.16036719	0.04358158	0.0553	0.7526
5	1.11678561	0.33405606	0.0532	0.8058
6	0.78272955		0.0373	0.8431

Tabla 36: Autovalores para la matriz de correlación de los Mediocentros. Elaboración propia

Como vemos en esta Tabla 36, con las 2 primeras componentes se consigue explicar el 59,87% de la variabilidad de las variables independientes, valor ligeramente superior al obtenido para los defensas.

Pasamos por tanto a la realización del método *k*-NN con estas 2 componentes como coordenadas, obteniendo un MSE medio para las 100 muestras de 22,74. Este valor indica un error altísimo, y por tanto, no sería recomendable utilizar este método para los mediocentros.

Pasamos ahora al análisis de la tercera base de datos, los delanteros. Si observamos su matriz de correlación, en la Tabla 14, vemos que las variables más relacionadas con el Salario son Popularidad y A_T. Nuevamente, realizando el bucle para las 100 muestras, obtenemos un MSE medio de 10,95.

Con la segunda opción de desarrollo, se obtiene la siguiente tabla de autovalores.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	10.8870907	7.5601068	0.4734	0.4734
2	3.3269839	1.0084889	0.1447	0.6180
3	2.3184950	0.8180105	0.1008	0.7188
4	1.5004845	0.6250743	0.0652	0.7840
5	0.8754103	0.0793602	0.0381	0.8221
6	0.7960501		0.0346	0.8567

Tabla 37: Autovalores para la matriz de correlación de los Delanteros. Elaboración propia

Vemos ahora que con 2 componentes conseguimos aumentar el porcentaje de variabilidad de las variables independientes con respecto a las bases de datos anteriores, concretamente hasta un 62%. Utilizando estas dos primeras componentes como coordenadas, obtenemos un MSE medio de 12,5.

Para la última base de datos, los Jugadores_Combinados, observamos que las variables más correlacionadas con el Salario son Popularidad y PasesUltTercio. Si utilizamos estas dos variables como coordenadas, obtenemos un MSE medio para las 100 muestras aleatorias de 8,5.

Ahora, si pasamos al empleo de todas las variables de la base de datos mediante el uso de componentes principales, obtenemos la siguiente tabla:

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	6.47040903	4.16640996	0.4044	0.4044
2	2.30399907	0.25650306	0.1440	0.5484
3	2.04749601	0.79736387	0.1280	0.6764
4	1.25013213	0.24921068	0.0781	0.7545
5	1.00092145	0.37153492	0.0626	0.8171
6	0.62938653		0.0393	0.8564

Tabla 38: Autovalores para la matriz de correlación de todos los jugadores. Elaboración propia

En este caso con 2 componentes principales, obtenemos un 55% de variabilidad explicada de las variables independientes. De nuevo, utilizando estas dos primeras componentes como coordenadas, obtenemos un MSE medio de 12,5.

Vemos por tanto que, comparando estos resultados con los obtenidos en métodos anteriores (regresión lineal y árboles de regresión), este método ofrece predicciones menos precisas, luego podemos concluir que este método no se ajusta bien para nuestras bases de datos, y por tanto no es un método capaz de explicar correctamente el valor de mercado de un jugador mediante variables de rendimiento.

6 - Conclusiones

Si observamos los resultados obtenidos por los tres modelos que se han utilizado, vemos que obtenemos los mejores resultados mediante el método Random Forest, de igual manera que ocurre en Li y otros (2022).

Es interesante destacar que ninguno de los modelos construidos a lo largo de este trabajo presenta como variable de gran importancia los Goles del jugador. Este aspecto suele ser una métrica fundamental para evaluar el rendimiento de un jugador en la realidad, pero, de igual forma que ocurre en algunos de los artículos mencionados con anterioridad, como en Li y otros (2022) o Yaldo y otros (2017), esta variable Goles no obtiene tanta importancia como se pensaría en un principio. Por tanto, parece que en el contexto de este estudio, teniendo en cuenta nuestros resultados y los de otros artículos, la cantidad de goles no es necesariamente el indicador más relevante para predecir el salario de un jugador.

La Popularidad y los PasesUltTercio son dos variables que se repiten como las más importantes en la construcción de los distintos modelos del trabajo. Como vimos en Carrieri y otros (2018), también se concluía que la popularidad de un jugador era un aspecto determinante para determinar su valor de mercado o salario. En cambio, no podemos comparar con ningún artículo si los PasesUltTercio son también importantes para otros autores porque en otros artículos no se ha recogido los datos de la misma forma que en este trabajo, con respecto a esta variable PasesUltTercio.

Concluimos por tanto que los resultados obtenidos han sido muy positivos al obtener un modelo de RF con estadísticos de ajuste muy elevados tanto en el fichero de prueba como en el de entrenamiento, y además, hemos identificado patrones en la estructura de nuestros modelos que coinciden con los resultados de otros estudios mencionados.

Bibliografía

Carrieri, V., Principe, F. y Raitano, M., 2018. What makes you 'super-rich'? New evidence from an analysis of football players' wages. *Oxford Economic Papers*, 70(4), 950-973.

<https://academic.oup.com/oep/article-abstract/70/4/950/5048427?redirectedFrom=fulltext&login=false>

Cheamanunkul, S., & Freund, Y. (2014, December). Improved kNN rule for small training sets. In 2014 13th International Conference on Machine Learning and Applications (pp. 201-206). IEEE.

https://www.researchgate.net/publication/269396235_Improved_kNN_Rule_for_Small_Training_Sets

Inuba, 2022. Tecnología en el deporte y sus aplicaciones.

<https://inuba.com/blog/tecnologia-avances-deporte-innovacion/>

Li, C., Kampakis, S., & Treleaven, P. (2022). Machine Learning Modeling to Evaluate the Value of Football Players. arXiv preprint arXiv:2207.11361..
<https://arxiv.org/abs/2207.11361>

Método de árboles de regresión -> Rodrigo, J. A. "Árboles de decisión, random forest, gradient boosting y C5.0", Febrero 2017.

https://cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_c50

Método de *k*-NN vecinos más cercanos. -> Leif E. P. (2009) K-nearest neighbor.

Scholarpedia, 4(2):1883. http://www.scholarpedia.org/article/K-nearest_neighbor

Método de regresión lineal múltiple -> Abuín, J. R. "Regresión lineal múltiple." IdEyGdM-Ld Estadística, Editor 32 (2007).

http://humanidades.cchs.csic.es/cchs/web_UAE/tutoriales/PDF/Regresion_lineal_multiple_3.pdf

Sutojo, T., Rustad, S., Akrom, M., Syukur, A., Shidik, G. F., & Dipojono, H. K. (2023). A machine learning approach for corrosion small datasets. *npj Materials Degradation*, 7(1), 18.<https://www.nature.com/articles/s41529-023-00336-7>

STATS SOS -> Rangos para el coeficiente de correlación de Pearson.

<https://statssos.online/2015/03/10/pero-que-linda-relacion-tienen-la-correlacion-de-pearson/>

Torgler, B., & Schmidt, S. L. (2007). What shapes player performance in soccer? Empirical findings from a panel analysis. *Applied Economics*, 39(18), 2355-2369.<https://www.tandfonline.com/doi/abs/10.1080/00036840600660739>

Yaldo, L., & Shamir, L. (2017). Computational estimation of football player wages. *International Journal of Computer Science in Sport*, 16(1).<https://sciendo.com/article/10.1515/ijcss-2017-0002>

Zebari, G. M., Zeebaree, S., Sadeeq, M. M., & Zebari, R. (2021). Predicting Football Outcomes by Using Poisson Model: Applied to Spanish Primera División. *Journal of Applied Science and Technology Trends*, 2(04), 105-112.https://www.researchgate.net/profile/Subhi-Zeebaree/publication/357359461_Predicting_Football_Outcomes_by_Using_Poisson_Model_Applied_to_Spanish_Primera_Division/links/61caf740e669ee0f5c6c0204/Predicting-Football-Outcomes-by-Using-Poisson-Model-Applied-to-Spanish-Primera-Division.pdf

Zhao, Y. (2022, December). Model Prediction of Factors Influencing NBA Players' Salaries Based on Multiple Linear Regression. In 2022 2nd International Conference on Economic Development and Business Culture (ICEDBC 2022) (pp. 1439-1445). Atlantis Press.<https://www.atlantis-press.com/proceedings/icedbc-22/125983656>

Anexo 1 - Defensas

Pearson Correlation Coeffi Prob > r under H0											
	Salario	Popularidad	PJ	Titular	Minutos	Goles	Ass	TarjA	TarjR	FCom	FRec
Salario	1.00000	0.80247	0.39463	0.36085	0.37425	0.43913	0.39638	0.15318	0.02112	0.10450	0.12185
Salario		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0509	0.7890	0.1843	0.1213
Popularidad	0.80247	1.00000	0.35100	0.30277	0.31102	0.45499	0.31883	0.04287	-0.05836	0.06698	0.10143
Popularidad		<.0001		<.0001	<.0001	<.0001	<.0001	0.5889	0.4749	0.2181	0.1976
PJ	0.39463	0.35100	1.00000	0.95215	0.95436	0.49035	0.39081	0.63937	0.13417	0.71970	0.57256
PJ		<.0001		<.0001	<.0001	<.0001	<.0001	0.0677	<.0001		<.0001
Titular	0.36985	0.30277	0.95215	1.00000	0.99588	0.47674	0.32108	0.68692	0.24019	0.71485	0.55078
Titular		<.0001		<.0001	<.0001	<.0001	<.0001	0.0001	0.0020	<.0001	<.0001
Minutos	0.37425	0.31102	0.95436	0.99588	1.00000	0.47874	0.32321	0.67012	0.22564	0.70743	0.54095
Minutos		<.0001		<.0001	<.0001	<.0001	<.0001	0.0001	0.0037	<.0001	<.0001
Goles	0.43913	0.45499	0.49035	0.47674	0.47874	1.00000	0.33047	0.30158	0.06168	0.34538	0.24475
Goles		<.0001		<.0001	<.0001	<.0001	<.0001	0.0001	0.4341	<.0001	0.0016
Ass	0.39638	0.31883	0.39081	0.32105	0.32321	0.33047	1.00000	0.11433	-0.16809	0.20879	0.33491
Ass		<.0001		<.0001	<.0001	<.0001	<.0001	0.1462	0.0320	0.0075	<.0001
TarjA	0.15318	0.04287	0.63937	0.68692	0.67012	0.30158	0.11433	1.00000	0.35563	0.69141	0.38566
TarjA		0.0509	0.5889	<.0001	<.0001	<.0001	<.0001	0.1482		<.0001	<.0001
TarjR	0.02112	-0.05836	0.13417	0.24019	0.22594	0.06168	-0.16809	0.35563	1.00000	0.15728	0.04947
TarjR		0.7890	0.4749	0.0877	0.0202	0.0037	0.4341	0.0320	<.0001	0.0450	0.5306
FCom	0.10450	0.06698	0.71970	0.71485	0.70743	0.34538	0.20879	0.69141	0.15728	1.00000	0.62742
FCom		0.1843	0.2181	<.0001	<.0001	<.0001	0.0075	<.0001	0.0450		<.0001
FRec	0.12185	0.10143	0.67268	0.56078	0.64095	0.24475	0.33491	0.38568	0.04947	0.62742	1.00000
FRec		0.1213	0.1978	<.0001	<.0001	0.0016	<.0001	0.5306	<.0001		
BSR	0.47915	0.40237	0.76786	0.75734	0.75499	0.36688	0.38689	0.48239	0.12825	0.57372	0.51402
BSR		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1028	<.0001		<.0001
DAereosG	0.21698	0.19829	0.51742	0.59281	0.59137	0.26238	-0.17393	0.51689	0.36739	0.46881	0.17328
DAereosG		0.0054	0.0112	<.0001	<.0001	0.0007	0.0284	<.0001	<.0001	<.0001	0.0270
DAereosP	0.23514	0.19663	0.61361	0.65306	0.64832	0.23485	0.03136	0.52424	0.26208	0.51262	0.29111
DAereosP		0.0025	0.0119	<.0001	<.0001	0.0025	0.0743	<.0001	0.0007	<.0001	0.0002
PAciertoPases	0.40835	0.36137	0.20748	0.20606	0.20922	0.25519	-0.18141	0.13755	0.18193	0.07792	-0.00491
PAciertoPases		<.0001	<.0001	0.0079	0.0083	0.0074	0.0100	0.0205	0.0800	0.0201	0.3228
PasesUltTercio	0.72180	0.63367	0.68718	0.66788	0.67117	0.41795	0.38449	0.39932	0.10127	0.42131	0.32186
PasesUltTercio		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1983	<.0001	<.0001	<.0001
VReg	0.19005	0.17523	0.48880	0.40010	0.38528	0.23301	0.47811	0.25862	-0.04794	0.51513	0.52953
VReg		0.0151	0.0253	<.0001	<.0001	0.0028	<.0001	0.0008	0.5434	<.0001	<.0001
DisBlok	0.33388	0.24149	0.48674	0.56634	0.56042	0.22405	-0.18371	0.46984	0.38006	0.29846	0.06386
DisBlok		<.0001	0.0019	<.0001	<.0001	0.0040	0.0189	<.0001	<.0001	0.0001	0.4181
Interc	0.28488	0.20089	0.72868	0.77671	0.76342	0.32300	0.14832	0.60398	0.32167	0.67810	0.51642
Interc		0.0008	0.0102	<.0001	<.0001	<.0001	0.0568	<.0001	<.0001	<.0001	<.0001
Dpj	0.22570	0.14299	0.55618	0.65092	0.64650	0.21303	-0.15981	0.53421	0.34663	0.40834	0.12147
Dpj		0.0038	0.0668	<.0001	<.0001	0.0063	0.0416	<.0001	<.0001	<.0001	0.0943
Err	0.29902	0.24341	0.31935	0.34645	0.35199	0.20519	-0.00821	0.17350	0.24928	0.12746	0.08303
Err		0.0001	0.0017	<.0001	<.0001	0.0008	0.0172	0.0268	0.0013	0.1049	0.2920
clients, N = 163 Rho=0											
	BSR	DAereosG	DAereosP	PAciertoPases	PasesUltTercio	VReg	DisBlok	Interc	Dpj	Err	
Salario	0.47915	0.21698	0.23514	0.40835	0.72180	0.19005	0.33388	0.26488	0.22570	0.29802	
Salario		<.0001	0.0054	0.0025	<.0001	0.0151	<.0001	0.0008	0.0038	0.0001	
Popularidad	0.40237	0.19829	0.19663	0.36137	0.63367	0.17523	0.24149	0.20089	0.14299	0.24341	
Popularidad		<.0001	0.0112	0.0119	<.0001	0.0253	0.0119	0.0102	0.0888	0.0017	
PJ	0.76786	0.51742	0.61361	0.20748	0.68718	0.68718	0.48860	0.48874	0.72898	0.55618	0.31635
PJ		<.0001	<.0001	<.0001	0.0079	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Titular	0.75734	0.59281	0.65306	0.20606	0.66788	0.67117	0.40010	0.56534	0.77671	0.65092	0.34645
Titular		<.0001	<.0001	<.0001	0.0083	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Minutos	0.75499	0.59137	0.64632	0.20922	0.67117	0.38528	0.56042	0.76342	0.64850	0.35199	
Minutos		<.0001	<.0001	<.0001	0.0074	<.0001	<.0001	<.0001	<.0001	<.0001	
Goles	0.36688	0.26238	0.23485	0.25519	0.41795	0.23301	0.22405	0.32300	0.21303	0.20519	
Goles		<.0001	0.0007	0.0025	0.0010	<.0001	0.0028	0.0040	<.0001	0.0063	
Ass	0.38689	-0.17393	0.03315	-0.18141	0.38449	0.47811	-0.18371	0.14832	-0.15981	-0.15681	-0.00821
Ass		<.0001	0.0264	0.6743	0.0205	<.0001	0.0189	0.0568	0.0416	0.9172	
TarjA	0.48239	0.51689	0.52424	0.13755	0.39932	0.25982	0.46984	0.60398	0.53421	0.17350	
TarjA		<.0001	<.0001	<.0001	0.0800	<.0001	0.0008	<.0001	<.0001	<.0001	
TarjR	0.12825	0.36739	0.26208	0.18193	0.20120	0.10127	-0.04794	0.38008	0.32167	0.34663	0.24928
TarjR		0.1028	<.0001	0.0007	0.0201	0.1983	0.5434	<.0001	<.0001	<.0001	0.0013
FCom	0.57372	0.48681	0.51282	0.07792	0.42131	0.51513	0.29648	0.57810	0.40834	0.12748	
FCom		<.0001	<.0001	<.0001	0.3228	<.0001	0.0001	0.0001	<.0001	<.0001	0.1049
FRec	0.51402	0.17328	0.29111	-0.00491	0.32168	0.52953	0.06388	0.51642	0.13147	0.08303	
FRec		<.0001	0.0270	0.0002	0.9504	<.0001	0.0001	0.4181	<.0001	0.0943	
BSR	1.00000	0.61348	0.75921	0.14496	0.84706	0.63952	0.54113	0.63024	0.66608	0.38251	
BSR		<.0001	<.0001	0.0049	<.0001	<.0001	0.0001	<.0001	<.0001	<.0001	
DAereosG	0.61348	1.00000	0.85417	0.28240	0.53482	0.13687	0.77601	0.60274	0.86474	0.30649	
DAereosG		<.0001	<.0001	<.0001	0.0003	<.0001	0.0815	<.0001	<.0001	<.0001	
DAereosP	0.75921	0.85417	1.00000	0.20388	0.61594	0.38462	0.71551	0.60585	0.81865	0.31197	
DAereosP		<.0001	<.0001	<.0001	0.0091	<.0001	0.0001	0.0001	<.0001	<.0001	
PAciertoPases	0.14496	0.28240	0.20388	1.00000	0.31958	-0.19659	0.41444	0.23361	0.23351	0.25335	0.37210
PAciertoPases		0.0649	0.0003	0.0091	<.0001	<.0001	0.0119	0.0001	0.0027	0.0011	<.0001
PasesUltTercio	0.84705	0.53482	0.61594	0.31958	1.00000	0.48308	0.51023	0.52809	0.53097	0.38388	
PasesUltTercio		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
VReg	0.63952	0.13687	0.38462	-0.19659	0.48308	1.00000	0.02640	0.7380	0.35440	0.15253	0.05089
VReg		<.0001	0.0615	<.0001	0.0119	<.0001	<.0001	0.7380	<.0001	0.0519	0.5168
DisBlok	0.54113	0.77601	0.71551	0.41444	0.51023	0.20640	1.00000	0.51852	0.87839	0.47012	
DisBlok		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
Interc	0.63024	0.80274	0.60585	0.23361	0.52620	0.35440	0.51852	1.00000	0.60004	0.22639	
Interc		<.0001									

Tabla 6: Análisis de correlaciones para los Defensas. Elaboración propia.

Anexo 2 - Mediocentros

2.1. Método de Regresión Lineal

Pearson Correlation Coef Prob > r under H												
	Salario	Popularidad	PJ	Titular	Minutos	Goles	Ass	TarjA	TarjR	FCom	FReo	
Salario	1.00000	0.71082	0.48545	0.50249	0.50497	0.21180	0.43872	0.21327	0.02121	0.16511	0.34544	
Popularidad	0.71082	1.00000	0.39120	0.36902	0.37003	0.25904	0.43284	0.11301	-0.01522	0.10988	0.27682	
PJ	0.48545	0.39120	1.00000	0.90556	0.92494	0.39514	0.57542	0.56778	0.15072	0.61880	0.62265	
Titular	0.50249	0.36902	0.90556	1.00000	0.99288	0.40942	0.62381	0.58585	0.14938	0.64820	0.68724	
Minutos	0.50497	0.37003	0.92494	0.99296	1.00000	0.41088	0.61180	0.58171	0.14399	0.63824	0.67043	
Goles	0.21180	0.26904	0.39814	0.40942	0.41088	1.00000	0.54688	-0.00449	0.17708	0.21173	0.43071	
Ass	0.0197	0.0197	0.0197	0.0197	0.0197	0.0197	1.00000	<.0001	0.9610	0.0520	0.0167	
Ass	0.43872	0.43284	0.57542	0.62381	0.61180	0.54986	1.00000	0.13329	0.08882	0.26813	0.58975	
TarjA	0.21327	0.11301	0.56776	0.58585	0.58171	-0.00449	0.13329	1.00000	0.32671	0.81955	0.50225	
TarjA	0.0188	0.2172	0.0001	0.0001	0.0001	0.0001	0.0001	0.1450	0.0003	<.0001	<.0001	
TarjR	0.02121	-0.01522	0.15072	0.14938	0.14399	0.17708	0.06882	0.32671	1.00000	0.32088	0.23857	
TarjR	0.8174	0.8884	0.0988	0.1020	0.1151	0.0520	0.4533	0.0003	0.0003	0.0003	0.0084	
FCom	0.16511	0.10956	0.01880	0.01880	0.01880	0.01880	0.01880	0.21173	0.28813	0.81955	0.00000	
FCom	0.0703	0.2311	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0003	<.0001	
FRec	0.34544	0.27882	0.82283	0.68724	0.67043	0.43971	0.58975	0.50225	0.23857	0.68302	1.00000	
FRec	0.0001	0.0021	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0084	<.0001	<.0001	
BSR	0.53832	0.39074	0.76257	0.79580	0.81047	0.22092	0.46091	0.49165	0.08932	0.47509	0.50211	
BSR	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0149	<.0001	0.3299	<.0001	<.0001	
DAereosG	0.25098	0.12385	0.43197	0.48808	0.48848	0.04351	0.07421	0.49927	0.15356	0.56572	0.31518	
DAereosG	0.0055	0.1766	0.3135	0.0001	0.0001	0.0001	0.0001	0.0001	0.0026	<.0001	0.0004	
DAereosP	0.21928	0.09240	0.46684	0.51409	0.51018	0.14950	0.15400	0.48343	0.16011	0.56938	0.39409	
DAereosP	0.0157	0.3135	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0794	<.0001	<.0001	
PAciertoPases	0.41081	0.37397	0.21485	0.20144	0.21035	-0.10232	0.06691	0.10285	-0.01205	-0.04793	0.00782	
PAciertoPases	<.0001	<.0001	0.0180	0.0287	0.0208	0.0208	0.2641	0.4569	0.2816	0.8956	0.8047	0.9339
PasesUltTercio	0.73929	0.56889	0.72887	0.74469	0.75744	0.25230	0.62485	0.38913	0.08529	0.34806	0.50207	
PasesUltTercio	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0052	<.0001	0.3523	<.0001	<.0001	
A_T	0.56514	0.56934	0.69659	0.71190	0.71888	0.47818	0.80287	0.19840	-0.03218	0.26492	0.56564	
A_T	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0291	0.7260	0.0033	<.0001	
A_Prg	0.38768	0.48069	0.46057	0.47162	0.48683	0.54888	0.62470	0.06447	0.13471	0.21323	0.53107	
A_Prg	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.4824	0.1407	0.0169	<.0001	
PerdidaBalon	0.18171	0.19124	0.33785	0.37087	0.35243	0.46492	0.46790	0.13795	0.28555	0.34865	0.56671	
PerdidaBalon	0.0481	0.0356	0.0002	0.0001	0.0001	0.0001	0.0001	0.1313	0.0032	<.0001	<.0001	
VReg	0.37799	0.29437	0.08049	0.63342	0.62869	0.26313	0.48779	0.42301	0.18229	0.50808	0.54630	
VReg	<.0001	0.0010	<.0001	<.0001	<.0001	0.0035	<.0001	0.0454	<.0001	<.0001	<.0001	
Interc	0.41485	0.26586	0.84309	0.71498	0.72369	0.09298	0.22522	0.80945	0.29959	0.56831	0.44599	
Interc	<.0001	0.0032	<.0001	<.0001	<.0001	0.3104	0.0130	<.0001	0.0008	<.0001	<.0001	
Sufficientes, N = 121 0: Rho=0												
	BSR	DAereosG	DAereosP	PAciertoPases	PasesUltTercio	A_T	A_Prg	PerdidaBalon	VReg	Interc		
Salario	0.53832	0.25095	0.21926	0.41081	0.73929	0.56514	0.38768	0.18171	0.37799	0.41485		
Salario	<.0001	0.0065	0.0157	<.0001	<.0001	<.0001	<.0001	0.0461	<.0001	<.0001		
Popularidad	0.39074	0.12385	0.09240	0.37393	0.56889	0.56934	0.48069	0.19124	0.29437	0.26586		
Popularidad	<.0001	0.1766	0.3135	<.0001	<.0001	<.0001	<.0001	0.0356	0.0010	0.0032		
PJ	0.76257	0.43107	0.46084	0.21485	0.72887	0.60859	0.46057	0.33785	0.60849	0.64308		
PJ	<.0001	<.0001	<.0001	0.0180	<.0001	<.0001	<.0001	0.0002	<.0001	<.0001		
Titular	0.79580	0.48805	0.51409	0.20144	0.74469	0.71190	0.47162	0.37087	0.63342	0.71488		
Titular	<.0001	<.0001	<.0001	0.0267	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		
Minutos	0.81047	0.49648	0.51018	0.21035	0.75744	0.71588	0.46663	0.35243	0.62899	0.72369		
Minutos	<.0001	<.0001	<.0001	0.0206	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		
Goles	0.22092	0.04351	0.14960	-0.10232	0.25230	0.47818	0.54688	0.48492	0.28313	0.09298		
Goles	0.0149	0.0356	0.0356	0.1017	0.02641	0.0052	<.0001	<.0001	0.0001	0.0035	0.3104	
Ass	0.46091	0.07421	0.15400	0.06691	0.62465	0.80287	0.62470	0.46796	0.46779	0.22522		
Ass	<.0001	0.4188	0.0917	0.4569	<.0001	<.0001	<.0001	<.0001	<.0001	0.0001	0.1310	
TarjA	0.49165	0.49627	0.48343	0.10285	0.36913	0.19840	0.06447	0.13795	0.42301	0.80945		
TarjA	<.0001	<.0001	<.0001	0.2616	<.0001	<.0001	0.0291	0.4624	0.1313	<.0001	<.0001	
0.08832	0.15358	0.16011	-0.01205	0.08529	-0.03218	0.13471	0.28555	0.18229	0.29068			
0.3299	0.0928	0.0794	0.8956	0.3523	0.7280	0.1407	0.0032	0.0454	0.0008			
0.47509	0.56572	0.59836	-0.04753	0.34806	0.26492	0.21323	0.34685	0.50808	0.56381			
0.50211	0.31518	0.39409	0.00762	0.50297	0.56564	0.53107	0.56671	0.54630	0.44599			
1.00000	0.06459	0.64508	0.31273	0.85608	0.74442	0.40998	0.29238	0.60277	0.63785			
BSR	<.0001	0.0005	0.0005	<.0001	<.0001	<.0001	<.0001	0.0011	<.0001	<.0001		
DAereosG	0.64549	1.00000	0.85854	0.03746	0.42363	0.23931	0.10409	0.16908	0.37203	0.59175		
DAereosG	<.0001	<.0001	0.0001	0.0001	0.0001	0.0002	0.2559	0.0037	<.0001	<.0001		
DAereosP	0.64508	0.85654	1.00000	-0.06980	0.39838	0.31688	0.24157	0.36241	0.53401	0.47735		
DAereosP	<.0001	<.0001	0.0004	0.4468	<.0001	0.0004	0.0076	<.0001	<.0001	<.0001		
0.31273	0.03748	-0.06980	1.00000	0.43432	0.20794	0.10884	-0.03351	0.15319	0.28517			
0.85608	0.42363	0.36838	0.43432	1.00000	0.83808	0.47108	0.27081	0.65442	0.53521			
0.74442	0.23931	0.31688	0.20794	0.83808	1.00000	0.62091	0.33895	0.58110	0.27453			
0.40998	0.10409	0.24157	0.10804	0.47108	0.62091	1.00000	0.84528	0.64314	0.29837			
0.29238	0.06908	0.36241	-0.03351	0.20794	0.10884	0.62091	<.0001	0.0001	0.0016			
0.60927	0.37203	0.53401	0.15319	0.65442	0.58110	0.64314	0.69953	1.00000	0.53997			
0.83785	0.50117	0.47735	0.28517	0.53521	0.27453	0.29837	0.32701	0.53997	1.00000			
0.83785	<.0001	<.0001	<.0001	0.0015	<.0001	0.0010	0.0003	<.0001	<.0001			

Tabla 10: Análisis de correlaciones de Mediocentros. Elaboración propia.

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	4.826864642	B	6.79747687	0.71	0.4794
Equipo Rango Alto	4.197935546	B	1.81555545	2.31	0.0229
Equipo Rango Medio Alto	3.162037810	B	1.25117255	2.53	0.0131
Equipo Rango Medio Bajo	-0.529020476	B	0.89053971	-0.59	0.5539
Equipo Rango Bajo	0.000000000	B	.	.	.
Popularidad	1.639412228		0.69987229	2.34	0.0212
Fuera5GrandesLigas No	1.333580883	B	1.08859259	1.23	0.2236
Fuera5GrandesLigas Si	0.000000000	B	.	.	.
Minutos	-0.000933654		0.00144123	-0.65	0.5186
PJ	-0.010459846		0.03380770	-0.31	0.7577
Titular	0.130459317		0.10672961	1.22	0.2246
Goles	0.081699267		0.08268443	0.99	0.3256
Ass	0.009752554		0.12025630	0.08	0.9355
TarjA	0.042103018		0.07347638	0.57	0.5680
TarjR	-0.546420722		0.36072458	-1.51	0.1331
FCom	-0.018832127		0.01492962	-1.26	0.2102
FRec	0.013736409		0.01006780	1.36	0.1756
BSR	-0.008091339		0.00578874	-1.40	0.1654
DAereosG	-0.024616873		0.01490830	-1.65	0.1020
DAereosP	0.043332075		0.01901750	2.28	0.0249
PAciertoPases	-0.078924310		0.08195954	-0.96	0.3380
PasesUltTercio	0.034781143		0.00516531	6.73	<.0001
A_T	-0.021518351		0.00739729	-2.91	0.0045
A_Prg	0.011813634		0.01014206	1.16	0.2470
PerdidaBalon	-0.007545457		0.01639024	-0.46	0.6463
VReg	-0.034948747		0.01672897	-2.09	0.0393
Interc	0.002810009		0.01623653	0.17	0.8630

Tabla 11: Regresión sin condiciones. Elaboración propia.

Modelos	StepW AIC	StepW BIC	StepW AdjRC	ForW AIC	ForW BIC	ForW AdjRC	BackW AIC	BackW BIC	BackW AdjRC
Nº Parámetros	6	5	15	6	5	15	9	8	15
R-Cuad. Ajust.	0.708	0.705	0.7281	0.708	0.705	0.7281	0.7228	0.7201	0.7281
MSE	10,60	10,72	9,8885	10,60	10,72	9,8885	10,0792	10,176	9,8885

Tabla 12: Resultados de diferentes modelos para Mediocentros. Elaboración propia.

Parameter Estimates						
Parameter	DF	Estimate	Standardized Estimate	Standard Error	t Value	Pr > t
Intercept	1	-1.034718	0	0.809807	-1.28	0.2040
Equipo Rango Alto	1	4.739072	0.235876	1.645013	2.88	0.0048
Equipo Rango Medio Alto	1	2.926277	0.140086	1.131417	2.59	0.0110
Equipo Rango Medio Bajo	1	-0.435078	-0.030165	0.761912	-0.57	0.5691
Equipo Rango Bajo	0	0	0		-	-
Popularidad	1	1.906654	0.256325	0.643355	2.96	0.0037
Titular	1	0.055134	0.251036	0.023599	2.34	0.0213
FCom	1	-0.020417	-0.175030	0.008876	-2.30	0.0233
BSR	1	-0.013104	-0.398660	0.004499	-2.91	0.0043
DAereosP	1	0.025076	0.172730	0.011364	2.21	0.0294
PasesUltTercio	1	0.026220	0.834526	0.004047	6.48	<.0001
A_T	1	-0.009362	-0.279457	0.003298	-2.84	0.0054

Tabla 13: Modelo seleccionado para los Mediocentros. Elaboración propia.

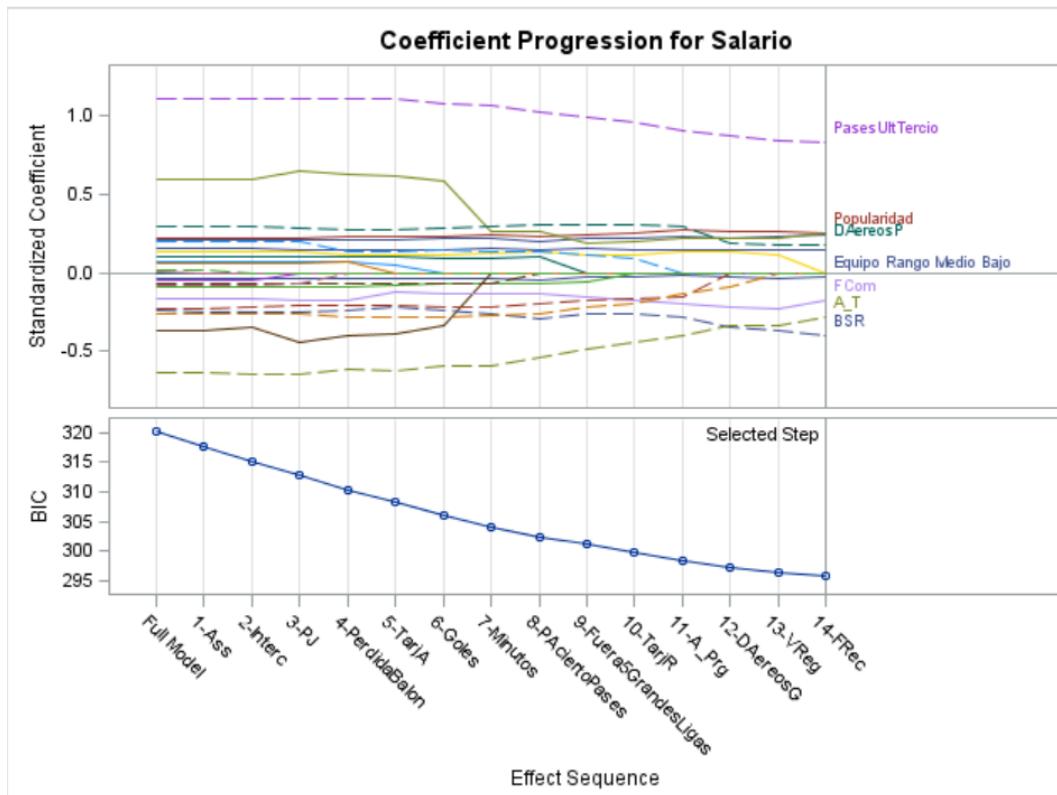


Figura 3: Evolución de la importancia de variables para Mediocentros. Elaboración propia.

2.2 Método de Árboles de Regresión

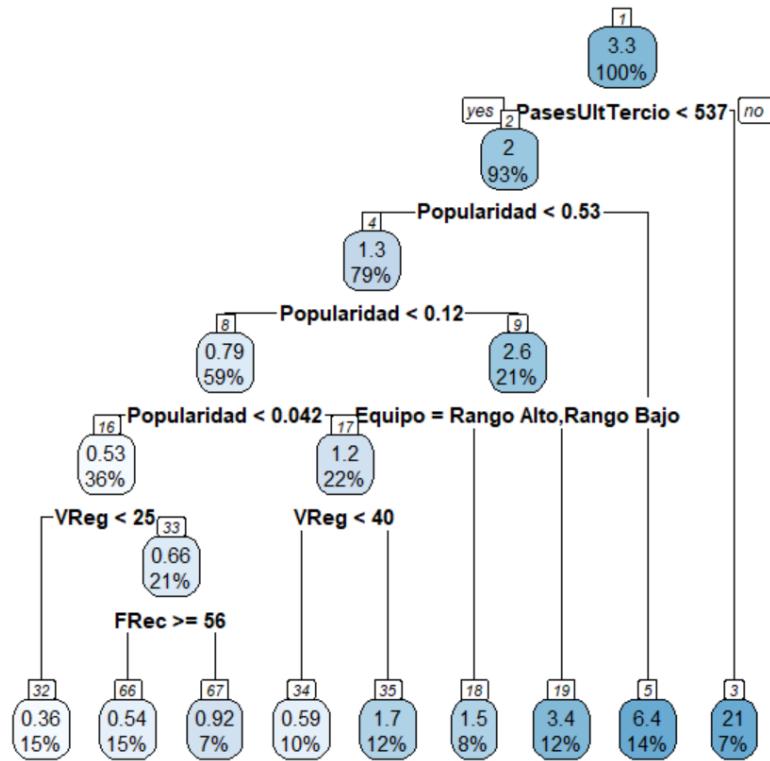


Tabla 27: Árbol de regresión con 5% de minibucket para Mediocentros. Elaboración propia.

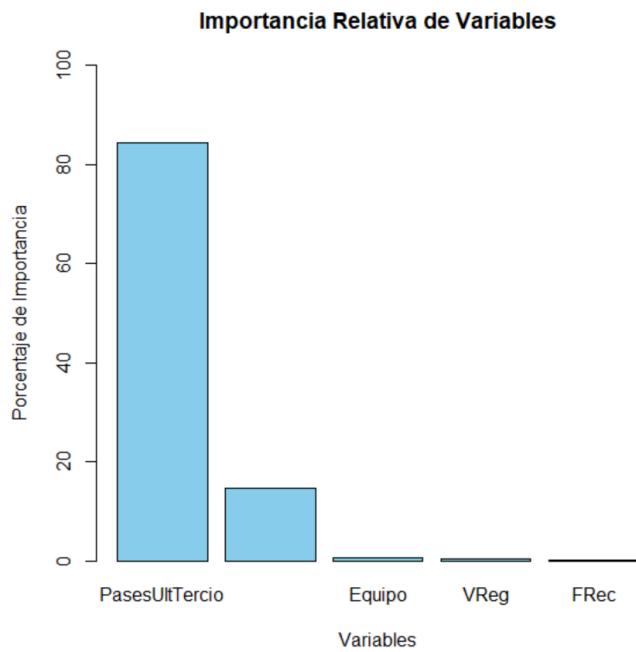


Tabla 28: Importancia de variables del árbol del 5% para los Delanteros. Elaboración propia.

Anexo 3 - Delanteros

	Salario	Popularidad	PJ	Titular	Minutos	Goles	Ass	TarjA	TarjR	FCom	FRec	BSR	
Salario	1.00000	0.80777	0.25091	0.30330	0.31223	0.49902	0.48741	-0.08031	-0.07890	-0.03964	0.19310	0.18775	
Popularidad	0.80777	<.0001	1.00000	0.22807	0.32789	0.33948	0.04505	0.46761	-0.07404	-0.02800	-0.01206	0.12096	0.12401
Populardad	<.0001			0.0091	0.0001	<.0001	<.0001	<.0001	0.4025	0.7890	0.8917	0.1704	0.1598
PJ	0.25091	0.22807	1.00000	<.0001	0.85190	0.88201	0.52089	0.61273	0.32108	0.01284	0.93721	0.6254	0.00539
PJ	0.0040				<.0001	<.0001	<.0001	<.0001	0.0002	0.8847	<.0001	<.0001	<.0001
Titular	0.30330	0.32789	0.86190	1.00000	0.06178	0.07540	0.71688	0.38980	-0.00224	0.56978	0.69224	0.63980	
Titular	0.0005	0.0001	<.0001		<.0001	<.0001	<.0001	<.0001	0.9798	<.0001	<.0001	<.0001	
Minutos	0.31223	0.33948	0.88201	0.99178	1.00000	0.69018	0.72034	0.37230	-0.00290	0.58479	0.69229	0.63878	
Minutos	0.0003	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	0.9739	<.0001	<.0001	<.0001	
Goles	0.46902	0.64505	0.52089	0.67540	0.69018	1.00000	0.54084	0.15322	0.04159	0.35051	0.34396	0.18212	
Goles	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0618	0.6384	<.0001	<.0001	0.0654	
Ass	0.48741	0.48781	0.61273	0.71688	0.72034	0.54084	1.00000	0.13802	-0.08548	0.18554	0.49147	0.53082	
Ass	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1174	0.2800	0.0348	<.0001	<.0001	
TarjA	-0.08031	-0.07404	0.32108	0.36930	0.37230	0.15322	0.13802	1.00000	0.21740	0.68283	0.43833	0.23134	
TarjA	0.3637	0.4025	0.0002	<.0001	<.0001	0.0818	0.1174	0.0130	<.0001	<.0001	0.0081		
TarjR	-0.07990	-0.02900	0.01284	-0.00224	-0.00290	0.04159	-0.09544	0.21740	1.00000	0.20852	0.07037	-0.12126	
TarjR	0.3662	0.7690	0.8847	0.9798	0.9739	0.9384	0.2800	0.0130	0.0184	0.4263	0.1693		
FCom	-0.03984	-0.12028	0.53721	0.55978	0.56479	0.36051	0.18554	0.68283	0.20852	1.00000	0.58082	0.27105	
FCom	0.6543	0.8917	<.0001	<.0001	<.0001	<.0001	<.0001	0.0246	0.0184	<.0001	0.0017	0.0017	
FRec	0.19310	0.12095	0.62542	0.69224	0.69229	0.34389	0.49147	0.43833	0.07037	0.58082	1.00000	0.53330	
FRec	0.0277	0.1704	<.0001	<.0001	<.0001	<.0001	<.0001	0.4283	<.0001			<.0001	
BSR	0.18775	0.12401	0.60538	0.63890	0.63878	0.16212	0.53082	0.23134	-0.12126	0.27195	0.53530	1.00000	
BSR	0.0324	0.1568	<.0001	<.0001	<.0001	0.0654	<.0001	0.0681	0.1693	0.0017	<.0001		
DAerosG	-0.01923	0.00869	0.26810	0.32885	0.33460	0.03081	0.00765	0.07983	0.01209	0.40456	0.27702	0.14282	
DAerosG	0.8281	0.9128	0.0200	<.0001	<.0001	0.0003	0.0312	0.0066	0.8814	<.0001	0.0014	0.1050	
DAerosP	-0.02378	0.01131	0.35158	0.39167	0.39348	0.33072	0.04471	0.18341	0.04354	0.49769	0.31248	0.23248	
DAerosP	0.7884	0.8984	<.0001	<.0001	<.0001	0.0001	0.0135	0.0367	0.6228	<.0001	0.0008	0.0078	
PAciertoPases	0.37178	0.32252	0.13780	0.07150	0.09025	0.11311	0.21187	-0.09524	-0.06996	-0.16254	0.02348	0.16326	
PAciertoPases	<.0001	0.0002	0.1185	0.4188	0.3072	0.2001	0.0155	0.2811	0.4290	0.0832	0.7909	0.0645	
PasesUltTercio	0.39838	0.35531	0.50373	0.58194	0.57523	0.33942	0.59240	0.11591	-0.13182	0.13109	0.41810	0.72521	
PasesUltTercio	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1691	0.1349	0.1371	<.0001	<.0001	
PasesClave	0.53879	0.47698	0.62579	0.70147	0.70208	0.46088	0.79426	0.19049	-0.17597	0.21653	0.51191	0.76088	
PasesClave	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0299	0.0452	0.0133	<.0001	<.0001	
A_T	0.56022	0.49338	0.64528	0.71924	0.72347	0.48708	0.78410	0.20286	-0.15088	0.25235	0.55015	0.78977	
A_T	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0208	0.0867	0.0038	<.0001	<.0001	
A_Prg	0.52274	0.43068	0.59477	0.65827	0.66131	0.37689	0.76335	0.18662	-0.17169	0.16595	0.51843	0.79677	
A_Prg	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0335	0.0508	0.0592	<.0001	<.0001	
RegInt	0.37858	0.26355	0.51347	0.53784	0.54723	0.26149	0.65246	0.22341	-0.14901	0.24526	0.53979	0.72398	
RegInt	<.0001	0.0024	<.0001	<.0001	<.0001	0.0027	<.0001	0.0106	0.0908	0.0049	<.0001	<.0001	
RegExit	0.39288	0.28877	0.50354	0.53169	0.53345	0.26478	0.65127	0.19483	-0.14280	0.22753	0.49833	0.73008	
RegExit	<.0001	0.0009	<.0001	<.0001	<.0001	0.0023	<.0001	0.0283	0.1058	0.0092	<.0001	<.0001	
TotalDisp	0.46175	0.53927	0.66426	0.79148	0.80551	0.85207	0.61497	0.26774	0.05427	0.50893	0.53524	0.30392	
TotalDisp	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0021	0.5397	<.0001	0.0007	0.0004	
DaP	0.50098	0.60749	0.62711	0.75691	0.77298	0.91241	0.60459	0.24739	0.07497	0.47000	0.48297	0.24859	
DaP	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0045	0.3968	<.0001	<.0001	0.0047	
Ficientes, N = 130 0: Rho=0													
DAerosG	0.01923	-0.02378	0.37178	0.39836	0.53679	0.56022	0.52274	0.37858	0.39268	0.48175	0.50098		
DAerosG	0.8281	0.7884	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
Popularidad	0.00089	0.01131	0.32252	0.38531	0.47688	0.49338	0.43058	0.28385	0.28877	0.53927	0.60749		
Popularidad	0.9129	0.8984	0.0002	<.0001	<.0001	<.0001	<.0001	<.0001	0.0024	0.0009	<.0001	<.0001	
PJ	0.26810	0.35158	0.13780	0.50373	0.62579	0.64528	0.59477	0.51347	0.50354	0.66429	0.62711		
PJ	0.0020	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
Titular	0.32885	0.39167	0.07150	0.58194	0.70147	0.71924	0.65287	0.53784	0.53169	0.79148	0.75691		
Titular	0.0001	<.0001	0.4188	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
Minutos	0.33480	0.39348	0.09025	0.57823	0.70208	0.72347	0.66131	0.54723	0.53845	0.80551	0.77298		
Minutos	<.0001	<.0001	0.3072	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
Goles	0.30981	0.33072	0.11311	0.33942	0.40808	0.48708	0.37889	0.26149	0.26478	0.85207	0.91241		
Goles	0.0003	0.0001	0.2001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0023	0.0023	<.0001	<.0001	
Ass	0.00785	0.04471	0.21187	0.59240	0.70428	0.78410	0.76338	0.65526	0.65127	0.81497	0.80459		
Ass	0.9312	0.6135	0.0155	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
TarjA	0.07983	0.18341	-0.09524	0.11591	0.19049	0.20268	0.18662	0.22341	0.19483	0.28774	0.24739		
TarjA	0.3666	0.3067	0.2811	0.1891	0.0299	0.0208	0.0335	0.0106	0.0263	0.0201	0.0045		
TarjR	0.01209	0.04354	-0.06998	-0.13182	-0.17597	-0.15085	-0.17698	-0.14901	-0.14280	0.05427	0.07497		
TarjR	0.8914	0.6228	0.4290	0.1349	0.0452	0.0887	0.0508	0.0908	0.1058	0.5397	0.3988		
FCom	0.40456	0.49709	-0.15254	0.13109	0.21653	0.25235	0.16598	0.24526	0.22753	0.50893	0.47000		
FCom	<.0001	<.0001	0.0832	0.1371	0.0133	0.0038	0.0592	0.0049	0.0092	<.0001	<.0001		
FRec	0.27702	0.31248	0.02348	0.41810	0.51191	0.55015	0.51843	0.53979	0.49833	0.53524	0.48297		
FRec	0.0014	0.0003	0.7909	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
BSR	0.14382	0.23248	0.16282	0.72521	0.76088	0.79877	0.76867	0.72398	0.73008	0.30392	0.24859		
BSR	0.1050	0.0078	0.0845	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0004	0.0047		
DAerosG	1.00000	0.87987	-0.39386	0.17673	0.11608	0.13795	0.01884	-0.00158	-0.00518	0.37583	0.31303		
DAerosG	<.0001	<.0001	0.0443	<.0001	0.1885	0							

Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	-0.891881507	B	2.56483996	-0.35	0.7287
Equipo Rango Alto	8.950947140	B	1.11708992	8.01	<.0001
Equipo Rango Medio Alto	1.943870463	B	0.70011790	2.78	0.0065
Equipo Rango Medio Bajo	0.602838991	B	0.55616253	1.08	0.2809
Equipo Rango Bajo	0.000000000	B		.	.
Popularidad	1.257250277		0.22869332	5.50	<.0001
Fuera5GrandesLigas No	0.568649533	B	0.61522416	0.92	0.3575
Fuera5GrandesLigas Si	0.000000000	B		.	.
Minutos	-0.000325647		0.00102089	-0.32	0.7504
PJ	-0.009597932		0.02132435	-0.45	0.6536
Titular	0.050146525		0.07120576	0.70	0.4829
Goles	-0.003985355		0.03485777	-0.11	0.9092
Ass	-0.021128028		0.06618608	-0.32	0.7502
TarjA	-0.044404449		0.04726989	-0.94	0.3497
TarjR	0.055254725		0.32715214	0.17	0.8662
FCom	-0.002347893		0.00952808	-0.25	0.8058
FRec	0.015043855		0.00665537	2.26	0.0259
BSR	-0.007906375		0.00493758	-1.60	0.1124
DAereosG	0.001245121		0.00499573	0.25	0.8037
DAereosP	0.001048718		0.00622044	0.17	0.8664
PAciertoPases	0.027737639		0.03484628	0.80	0.4279
PasesUltTercio	-0.009924307		0.00862655	-1.15	0.2526
PasesClave	-0.014095412		0.02910965	-0.48	0.6293
A_T	0.011564863		0.00948124	1.22	0.2253
A_Prg	0.003117473		0.00541393	0.58	0.5660
RegInt	-0.018337438		0.01198972	-1.53	0.1292
RegExit	0.021989423		0.02106784	1.04	0.2990
TotalDisp	-0.011492775		0.01316039	-0.87	0.3845
DaP	0.006164147		0.03534258	0.17	0.8619

Tabla 15: Regresión sin condiciones para los Delanteros. Elaboración propia.

Modelos	StepW AIC	StepW BIC	StepW AdjRC	ForW AIC	ForW BIC	ForW AdjRC	BackW AIC	BackW BIC	BackW AdjRC
Nº Parámetros	3	3	8	3	3	8	8	6	9
R-Cuad. Ajust.	0.814	0.814		0.814	0.814				
MSE	4,368 7	4,368 7	4,18517	4,368 7	4,368 7	4,18517	4,13975	4,2074	4,12670

Tabla 16: Resultados de diferentes modelos para Delanteros. Elaboración propia.

Parameter Estimates						
Parameter	DF	Estimate	Standardized Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.421487	0	0.312955	1.35	0.1805
Equipo Rango Alto	1	8.780910	0.505513	0.899228	9.76	<.0001
Equipo Rango Medio Alto	1	1.703689	0.121719	0.605330	2.81	0.0057
Equipo Rango Medio Bajo	1	0.555359	0.045952	0.482632	1.15	0.2521
Equipo Rango Bajo	0	0	0	.	.	.
Popularidad	1	1.406641	0.443611	0.162345	8.66	<.0001
PasesClave	1	0.010770	0.096923	0.005144	2.09	0.0383

Tabla 17: Modelo seleccionado para Delanteros. Elaboración propia.

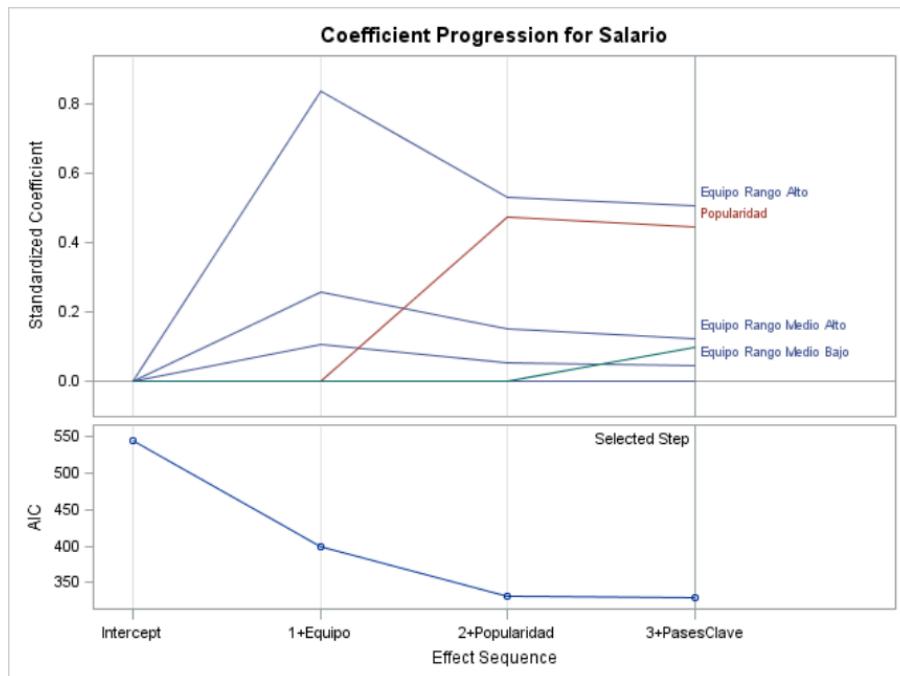


Figura 4: Evolución de importancia de variables para Delanteros. Elaboración propia

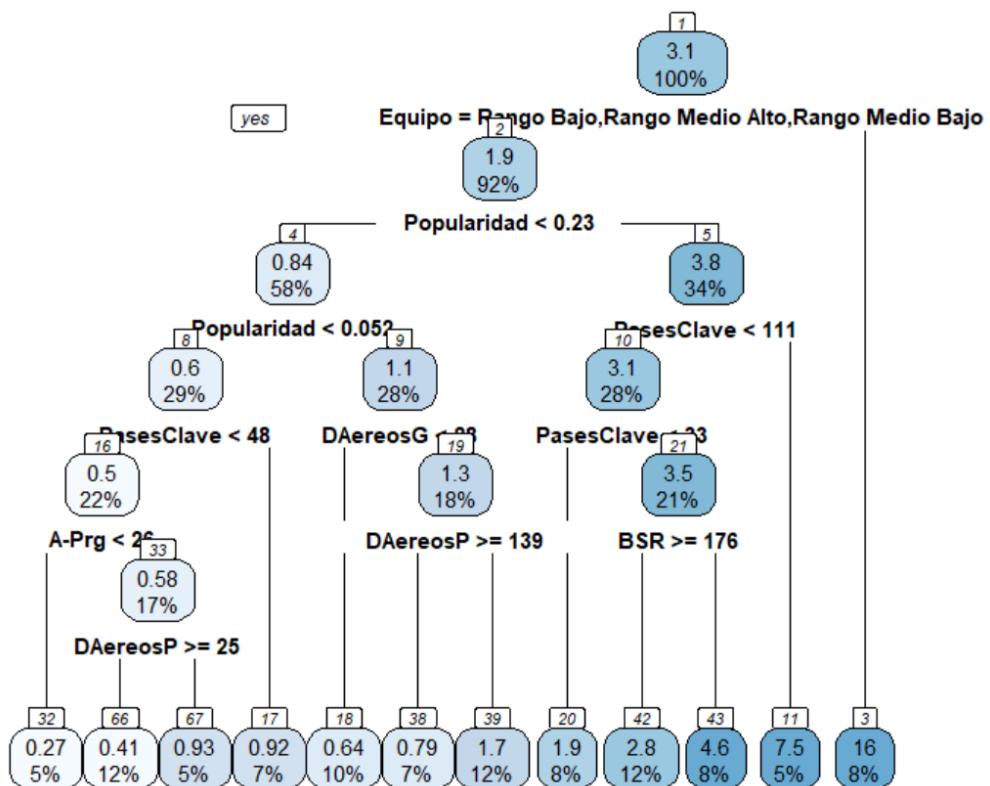


Tabla 29: Árbol de regresión del 5% de minibucket para los Delanteros. Elaboración propia.

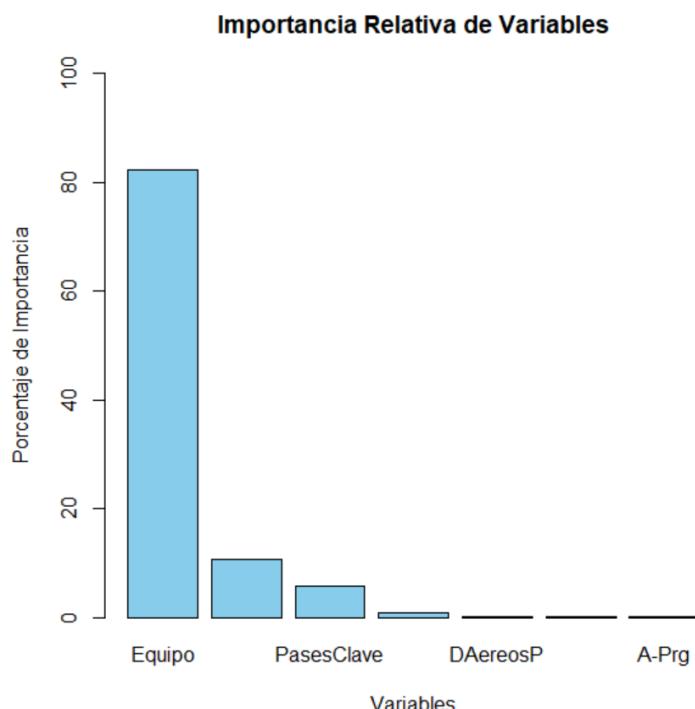


Tabla 30: Importancia de variables del árbol del 5% para los Delanteros. Elaboración propia.