UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Máster en Ingeniería de Análisis de datos, Mejora de procesos y Toma de decisiones

Machine Learning en el mundo del fútbol

María del Pilar Malagón Selma

Directores:

Dra. Ana María Debón Aucejo

Dr. Alberto J. Ferrer Riquelme

Universidad Politécnica de Valencia, España
Septiembre 2019

Resumen

Este trabajo analiza el rendimiento en el terreno de juego de jugadores profesionales de las grandes ligas europeas de fútbol durante la temporada 2017-18 a partir de estrategias y técnicas de *Machine Learning* y *Big Data* como el análisis de componentes principales, análisis clúster, gráficos de contribuciones y gráficos de radar, entre otros métodos. Asimismo, se profundiza en el análisis de las posiciones idóneas de los futbolistas en el campo, descubriendo cuáles son las variables más representativas de cada demarcación a partir del modelo *Random Forest*. Esta investigación supone una propuesta metodológica novedosa en relación al análisis de datos en el fútbol gracias al uso de programas informáticos como R, *Aspen ProMV* y *Tableau* con la finalidad de facilitar el proceso de captación de talento en las direcciones deportivas de los clubes del fútbol profesional. En definitiva, este trabajo de investigación ofrece un conjunto de herramientas basadas en el análisis de datos para mejorar la toma de decisiones y reducir la incertidumbre a la hora de acometer la contratación de un nuevo jugador en el fútbol de élite.

Abstract

This research project analyses the performance of football players in the top 5 leagues across Europe during the 2017-18 season based on Machine Learning and Big Data strategies and techniques such as principal component analysis (PCA), cluster analysis, contributions graphs and radar charts, among other methods. Moreover, the analysis of the ideal positions of the players in the field is studied in order to discover which are the most representative variables for each position using the Random Forest model. This research is an innovative methodology proposal in relation to the data analysis in football thanks to the use of computer programs such as R, *Aspen ProMV* and *Tableau* in order to facilitate the process of talent recruitment for the sports departments of professional football clubs. In conclusion, this research project offers a set of tools based on data analysis to improve decision making and reduce uncertainty when it comes to hiring a new player in elite football.

Agradecimientos

A la Dra. Ana María Debón Aucejo y al Dr. Alberto J. Ferrer Riquelme, tutores de mi TFM, por su gran ayuda y colaboración para realizar este trabajo.

A mis amigos y comunidad por la motivación y los ánimos para realizar este trabajo.

A mi familia por su apoyo a lo largo de este curso en el que he realizado el Máster y el TFM.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Introducción

1.1. Motivación

En los últimos años ha surgido un nuevo proceso transformador e imparable: *La industria 4.0* o *Cuarta Revolución Industrial*. Este concepto alude a una nueva revolución que combina técnicas avanzadas de producción y operaciones con tecnologías inteligentes que se integrarán en las organizaciones, las personas y los activos (Deloitte 2019). Esta *industria 4.0* ha potenciado la irrupción definitiva del análisis de datos masivos (*Big Data*) en la mayoría de ámbitos y sectores de las sociedades modernas.

El *Big Data* se refiere al conjunto de datos cuyo tamaño supera la capacidad del software convencional de una base de datos para capturarlos, almacenarlos, tratarlos y analizarlos (Manyika *et al.* 2011). Este campo de estudio trata sobre la "capacidad de la sociedad de aprovechar la información de formas novedosas para obtener percepciones útiles o bienes y servicios de valor significativo" (Mayer-Schönberger and Cukier 2013), lo cual implica que la calidad de los datos sea tan importante como la cantidad de los mismos. Así, "los datos masivos suponen un paso importante en el esfuerzo de la humanidad por cuantificar y comprender el mundo" (Mayer-Schönberger and Cukier 2013). Este fenómeno conlleva nuevas percepciones capaces de transformar las relaciones entre los mercados, las organizaciones, los ciudadanos y los gobiernos (Marr 2017).

En el caso que nos ocupa, el *Big Data* y el deporte han gozado de una estrecha relación impulsada por el mercado estadounidense gracias al béisbol (Lewis 2003) y al baloncesto (Silver, 2014). Sin embargo, este nuevo nicho de mercado[1] (MarkeInhouse 2019) se había resistido para adentrarse en la industria del fútbol (*soccer*[2]).

Ha sido recientemente cuando los mercados anglosajones y alemanes, impulsados por sus respectivas ligas y federaciones nacionales, han promovido, estimulado y potenciado el uso de la ciencia de datos en los clubes creando departamentos y equipos de trabajo formados por matemáticos, ingenieros informáticos y estadísticos, entre otros (Kuper 2019).

[1] Porción de un segmento de mercado en la que los individuos que la forman poseen características y necesidades homogéneas que aún no están del todo cubiertas por la oferta del mercado.

[2] Anglicismo a para denominar al fútbol que se juega en Canadá y Estados Unidos.

Este avance tecnológico en la industria del fútbol no se entendería sin el papel fundamental de empresas como Opta (Opta Sports 2019), Wyscout (Wyscout 2019), Instat (InStat 2019) o Stats (Stats LLC 2019), proveedores de datos oficiales encargados de recoger, procesar y facilitar los datos a los clubes y federaciones para sus propios tratamientos y análisis de datos.

El deporte es un elemento clave en la vida social de millones de personas (Billings 2010) por su alta capacidad para involucrar a grandes audiencias (Boyle and Haynes 2004). En particular, el fútbol se ha convertido en el deporte más popular de todos en las sociedades modernas. Así lo demuestran los datos de audiencia televisiva de los acontecimientos futbolísticos más importantes del mundo: la Copa del Mundo y la UEFA Champions League. La edición de la Copa del Mundo de 2018 en Rusia, por ejemplo, tuvo un alcance total televisivo de 3.500 millones de personas (FIFA 2018). Asimismo, la Era Digital y, sobre todo, los medios de comunicación han desempeñado un papel decisivo en el proceso de transformación del fútbol en un espectáculo de masas capaz de generar grandes audiencias despertando sentimientos y emociones en sus públicos (Llopis 2013).

Bajo este contexto, la ciencia de datos y el deporte rey conforman un incipiente campo de estudio en boga. Según el periódico digital Gentside Spagne, la previsión de ingresos en España para la temporada 2018-2019 en la Liga de Fútbol Profesional ascendió hasta los 4.500 millones de euros, cifra muy superior a la del año 2012 donde los ingresos fueron de 2.228 millones de euros (Aguado 2019).

Este trabajo supone un primer acercamiento a la industria del fútbol profesional a partir de técnicas de uso de análisis de datos y *Machine Learning*[3] (SAS 2019). Se trata de un ámbito pionero cuyo principal objetivo consistirá en ayudar a los directivos de clubes, entrenadores y jugadores a adaptar los nuevos recursos tecnológicos a sus propias necesidades en el área deportiva. Entre otros retos y oportunidades, el uso del *Big Data* aplicado al rendimiento en el fútbol profesional podría permitirnos reducir riesgos en la toma de decisiones en la adquisición, renovación y rescisión de jugadores, así como facilitar el proceso de captación de talento o estudiar y analizar a los equipos rivales para anticiparnos y contrarrestar sus estrategias.

[3] El machine learning es un método de análisis de datos que automatiza la construcción de modelos analíticos. Es una rama de la inteligencia artificial basada en la idea de que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones con mínima intervención humana.

2. Objetivos

Como se ha mencionado en el punto anterior, las técnicas de Machine Learning y Big Data pueden ser un instrumento clave en la metodología y proceso de captación de talento en el fútbol profesional, sobre todo a la hora de acometer fichajes de jugadores de élite. Así, el objetivo general de este trabajo fin de máster es desarrollar herramientas basadas en técnicas de Machine Learning y Big Data que faciliten el proceso de captación de talento en un club de fútbol profesional.

Para conseguir estos propósitos se han concretado los siguientes objetivos específicos:

Objetivo 1. Crear una aplicación web que permita encontrar a jugadores similares a otros a partir de un perfil de futbolista determinado previamente.

A partir de la herramienta *Shiny* (Chang *et al.* 2019), integrada en *RStudio* (R. Team 2016), se desarrolla una interfaz de usabilidad intuitiva con el objetivo principal de descubrir perfiles de jugadores similares a otro dado en base a las características de rendimiento del futbolista.

Objetivo 2. Comparar jugadores a partir de determinadas variables de rendimiento más representativas de los jugadores.

A partir del programa informático *Tableau* (Peck 2014), se generan visualizaciones denominadas *radares* con el objetivo de identificar las fortalezas y debilidades de dos o más futbolistas. Este proceso pretende facilitar la toma de decisiones en la contratación de jugadores reduciendo el riesgo y la incertidumbre.

Objetivo 3. Analizar las variables más representativas de cada una de las posiciones que ocupan los jugadores sobre el terreno de juego.

A partir del software *Aspen ProMV* (Aspen Technology Inc. 2018) y *RStudio*, se estudia la disposición táctica de los jugadores en el campo para discernir las variables más características de cada posición. Este proceso tiene el objetivo principal de conocer cuáles son las variables más representativas de cada posición para mejorar la toma de decisión en la captación de jugadores.

2. Metodología

Este apartado describe el enfoque y los procesos metodológicos utilizados en esta investigación. Para la resolución de los objetivos expuestos con anterioridad se han utilizado diferentes técnicas de *Machine Learning* a partir de diferentes programas informáticos como *R* (Ihaka and Gentleman 1996), *Aspen ProMV* y *Tableau*.

2.1. Sofware utilizado

El programa *R* (R. C. Team 2008) es un sistema para el análisis estadístico y para la creación de gráficos. *R* es un software libre que a través de la definición de funciones posee una gran variedad de técnicas estadísticas y gráficas. Por su parte *RStudio* es un entorno de desarrollo integrado (IDE) para *R* que permite crear un espacio de trabajo, elaborar gráficos y utilizar una gran cantidad de paquetes estadísticos. Incluye una consola, editor de resaltado de sintaxis que admite la ejecución directa de código, así como herramientas para el trazado, el historial, la depuración y la gestión del espacio de trabajo (RStudio 2018). *RStudio* no se puede utilizar sin *R*.

Uno de los paquetes de *R* es *R Markdown* (Xie *et al.* 2018) el cual, se ha convertido en un ecosistema relativamente completo para la creación de documentos. Existen gran cantidad de tareas que se pueden realizar a través de *R Markdown* (Allaire *et al.* 2019):

- Compilar un único documento de *R Markdown* en un informe en diferentes formatos, como PDF, HTML o Word.
- Crear cuadernos en los que se puedan ejecutar directamente fragmentos de código de forma interactiva.
- Producir paneles con diseños flexibles, interactivos y atractivos.
- Crear aplicaciones interactivas basadas en *Shiny*.
- Generar sitios web y blogs.

Aspen ProMV, por su parte analiza datos de proceso interrelacionados para identificar el conjunto crítico mínimo de las variables que impulsan la calidad y el rendimiento de los procesos, identificando los puntos de ajuste óptimos. *Aspen ProMV* se puede utilizar para: (AspenTech 2013):

- Analizar la desviación de la calidad
- Analizar el rendimiento de cada unidad
- Analizar la capacidad de degradación de la producción
- Análisis multivariado (descubrimiento y optimización de variables clave)
- Análisis de la variabilidad por lotes

Finalmente, *Tableau* es un software que permite la visualización de datos de manera interactiva.

2.2. Shiny

Shiny es un paquete de *R* que facilita la creación de aplicaciones web interactivas directamente desde la propia interfaz de *RStudio*. La librería permite alojar aplicaciones independientes en una página web o incrustarlas en documentos *R Markdown*, así como crear paneles de control que permiten manipular tablas y gráficos de una manera sencilla e intuitiva. También puede extender sus aplicaciones con temas CSS, HTML, widgets y acciones de JavaScript (Chang *et al.* 2017).

La creación de la aplicación se divide en tres partes. En la primera parte, el usuario debe especificar cómo quiere que sea la apariencia de la *app* que está creando, es decir, se centra en el diseño de la aplicación. Este apartado se especifica a través del objeto “ui” como puede verse en la Figura 1.

```
ui <- fluidPage(
  setBackgroundColor("#fcfcfc"),
  titlePanel('Similitud entre jugadores'),
  sidebarLayout(
    sidebarPanel(
      p("Selecciona a un jugador para descubrir los perfiles de futbolistas mas similares"),
      br(),
      selectizeInput("Name",
                     label = 'Jugador',
                     choices = unique(df$Name),
                     options = list(placeholder = 'SELECCIONE JUGADOR',
                                    onInitialize = I('function() { this.setValue(""); }')))

    ),

    mainPanel(
      plotOutput(outputId = "similarityPlot")
    )
  )
)
```

Figura 1 Captura 1 del código de Shiny en R Studio.

En la segunda parte se introducen los cálculos realizados, es decir, la función "server" como se puede observar en la Figura 2 es la encargada de procesar y mostrar los resultados de nuestro análisis.

```
server <- function(input, output) {

  dat <- reactive({
    req(input$Name)

    col_index = which(names(df) == input$Name)

    tmp1 <- select(df, 1)
    tmp2 <- df[col_index]

    data <- cbind(tmp1, tmp2)
    names(data)[2]<-"similarity"
    data <- data[order(desc(data$similarity)),]

    data <- data[(data$Name != input$Name),]
    data <- data[c(1: 20), ]
    data
  }
  )
```

Figura 2 Captura 2 del código de Shiny en R Studio.

Finalmente, se introduce la llamada a la aplicación "*Shiny App*" que ejecuta y arranca todo el código introducido y muestra la aplicación resultante.

2.3. Aprendizaje no supervisado

El aprendizaje no supervisado es una técnica de *Machine Learning* que se distingue por trabajar sobre datos no "etiquetados", es decir, se desconoce a qué grupo o clase pertenecen los individuos y, por tanto, la variable respuesta es desconocida.

A lo largo de este trabajo se utilizarán diferentes técnicas propias del aprendizaje no supervisado para la ejecución del análisis exploratorio de los datos. En concreto, se utilizarán el análisis de componentes principales, más comúnmente denominado PCA, y el análisis clúster, en concreto, el algoritmo *k-means*.

2.3.1. Análisis de Componentes Principales

El análisis de componentes principales o PCA es una de las técnicas más clásicas de estadística que en el *Machine Learning* se clasifica como una técnica de aprendizaje no supervisado. Estos métodos suelen aplicarse como parte del análisis exploratorio de los datos. Por ejemplo, cuando nos interesa descubrir las relaciones entre las variables y observaciones de la base de datos, o la posible existencia de datos atípicos. (Wold *et al.* 1987)

Otra de las utilidades del PCA es que puede utilizarse para la reducción de la dimensionalidad a partir de los componentes principales. Esta técnica aprovecha la correlación entre las variables, para crear nuevas variables (componentes principales) que explican gran parte de la variabilidad. De este modo, se obtiene un número menor de variables las cuales no se encuentran correlacionadas entre sí. La primera componente principal será la que mayor varianza explique, a continuación, la segunda componente explicará la máxima variabilidad no recogida por la primera, y así sucesivamente.

El punto de partida en todos los análisis de datos multivariados es una matriz de datos **X**. Las *N* filas en la tabla son los individuos. Las columnas *K* son las variables y comprenden las mediciones realizadas en los individuos. La Figura 3 ofrece una descripción gráfica de la descomposición de la matriz de datos **X** en la matriz de componentes principales y en la matriz de residuos **E**, es decir, la parte de los datos que no se encuentra explicada en el modelo de Componentes Principales.

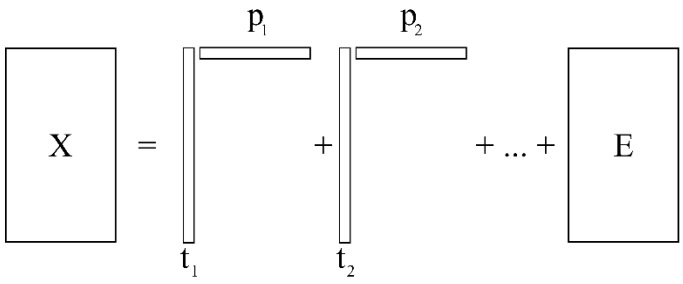

Figura 3 Descomposición matriz X en CP y E (Dunn 2019)

La base de datos original **X** se expresa mediante nuevas variables ortogonales, es decir, se produce una transformación a un nuevo sistema de coordenadas formado por los **T** scores y los **P** loadings. La proyección de **X** en un subespacio de dimensión *A* por medio de la matriz de proyección **P** da las coordenadas del individuo en el nuevo

subespacio, **T**. Las columnas en **T**, se denominan vectores de puntuación ($\mathbf{t_a}$) y las columnas en **P**, se llaman vectores de carga ($\mathbf{p_a}$). Estos últimos comprenden los coeficientes de dirección del hiperplano **CP**. Las desviaciones entre las proyecciones respecto a las coordenadas originales están recogidas por la matriz de residuos, **E**.

Los tamaños de los vectores **t**, y **p**, en una dimensión de PC no están definidos con respecto a una constante multiplicativa. Por lo tanto, es necesario anclar la solución de alguna manera. Esto generalmente se hace normalizando los vectores **p**, a la longitud 1.0. (Wold *et al.* 1987)

Para interpretar el modelo resultante del PCA se puede llevar a cabo el análisis individual de los factores que componen la matriz de componentes principales a través de diferentes gráficos realizados a partir del programa *Aspen ProMV*. El gráfico *SPE o SCR* (suma de cuadrados residual) indica el cuadrado de la distancia euclídea de la observación i-ésima a su proyección en el hiper-plano, formado por las direcciones principales, permitiendo identificar observaciones que rompen la estructura de correlación a través de los residuos. El gráfico de la T^2 de Hotelling (suma de cuadrados de los "scores" tipificados) calcula la medida de variación de cada muestra dentro del modelo desde su proyección a la media, a través del cuadrado de la distancia estimadade Mahalanobis.

Del mismo modo que se analizan los residuos, puede llevarse a cabo el estudio de las relaciones entre las variables y los individuos. El *Score Plot* permite visualizar la relación entre las observaciones dependiendo de las componentes principales seleccionadas. Este gráfico, en combinación con el *Loading plot*[4]*,* ofrece la relación de los individuos y las variables en el espacio de las X, creando el *Loadings-Bi plot*.

Finalmente, se puede utilizar el gráfico de contribuciones. Los gráficos de contribución son una técnica usada comúnmente para ayudar al diagnóstico de datos atípicos en los Procesos de Control Estadístico (SPC). A través del análisis de los gráficos de contribución puede conocerse cómo contribuye cada variable a cada individuo atípico en particular. (Van den Kerkhof *et al.* 2013)

[4] Gráfico que muestra la relación entre las variables en el espacio de las X

2.3.2. Análisis Clúster

El término clustering hace referencia a un amplio abanico de técnicas de análisis no supervisado cuya finalidad es encontrar patrones o grupos (clusters) dentro de un conjunto de observaciones. Las particiones se establecen de forma que las observaciones que están dentro de un mismo grupo son similares entre ellas y distintas a las de otros grupos (Rodrigo-Amat 2017).

En este trabajo se usará concretamente el método *k-means* (MacQueen 1967). Este método permite tanto clasificar por grupos a los individuos, como conocer la probabilidad de que cada individuo pertenezca a cada grupo. A su vez, la información recogida en el *cluster means* (información que ofrece la función *kmeans*) no sirve únicamente para determinar en qué grupos han sido clasificados los individuos, sino que a partir de los centroides también es posible averiguar qué variables han resultado representativas a la hora de determinar la pertenencia de los individuos a un grupo u otro (ver Anexo 9.7).

El método *k-meas* agrupa a los individuos en *k clusters*, siendo el analista el que determina el número de *clusters* o *K.* Sin embargo, la función k-means encuentra de forma automática el número óptimo de *clusters*, entendiendo como mejor número de *clusters* aquel cuya varianza interna sea lo más pequeña posible. Se trata por lo tanto de un problema de optimización, en el que se reparten las observaciones en *K clusters* de forma que la suma de las varianzas internas de todos ellos sea lo menor posible. Para poder solucionar este problema es necesario definir un modo de cuantificar la varianza interna.

Dos de las medidas más comúnmente utilizadas para cuantificar la varianza interna de un *cluster*:

- La suma de las distancias euclídeas al cuadrado entre cada observación (x_i) y el centroide (μ) de su *cluster*. Lo que equivale a la suma de cuadrados internos del *cluster*.
- La suma de las distancias euclídeas al cuadrado entre todos los pares de observaciones que forman el *cluster*, dividida entre el número de observaciones del *cluster*.

Minimizar la suma total de varianza interna de forma exacta (encontrar el mínimo global) es un proceso muy complejo debido a la inmensa cantidad de formas en las que *N* observaciones se pueden dividir en *K* grupos. Sin embargo, es posible obtener una solución que, aun no siendo la mejor sea muy buena (óptimo local). (Rodrigo-Amat 2017)

En definitiva, mientras que el método *k-means* aportará luz sobre la clasificación de los grupos de individuos conformados, el uso del PCA permitirá conocer y profundizar de un modo más concreto en la relación existente entre estos.

2.4. Aprendizaje supervisado

El aprendizaje supervisado es el nombre que se le da a las técnicas de *Machine Learning* que se distinguen por trabajar sobre datos previamente clasificados. Este método conoce a priori el grupo o clase a la que pertenecen los individuos y, por tanto, la variable respuesta es conocida. La técnica de aprendizaje supervisado que se utilizará a lo largo del trabajo se llevará a cabo como algoritmo confirmatorio de los resultados obtenidos.

Esta técnica emplea conjunto de datos conocidos (el denominado conjunto de datos de entrenamiento) para realizar predicciones. El conjunto de datos de entrenamiento incluye datos de entrada y valores de respuesta. A partir de él, el algoritmo de aprendizaje supervisado busca crear un modelo que pueda realizar predicciones acerca de los valores de respuesta para un nuevo conjunto de datos. Con frecuencia se utiliza un conjunto de datos de prueba para validar el modelo (MathWorks 2019).

El *Random Forest* (Breiman 2001) es un algoritmo predictivo de aprendizaje no supervisado que usa la técnica de *Bagging*[5] (Rodrigo-Amat 2017)para combinar diferentes árboles donde cada árbol se construye con observaciones y variables seleccionadas de forma aleatoria de la base de datos analizada (Santana 2014).

En forma resumida sigue este proceso:

[5]El término bagging hace referencia al empleo del muestreo repetido (bootstrapping) con el fin de reducir la varianza de algunos métodos de aprendizaje estadístico, entre ellos los árboles de predicción

1. Se seleccionan individuos al azar (usando muestreo con reemplazo) para crear diferentes sets de datos.
2. Se crea un árbol de decisión con cada set de datos, obteniendo diferentes árboles, ya que cada set contiene diferentes individuos y diferentes variables en cada nodo.
3. Al crear los árboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad (es decir, sin podar).
4. Se predicen los nuevos datos usando el "voto mayoritario", donde se clasificará como "positivo" si la mayoría de los árboles predicen la observación como positiva.

A partir de este método es posible comprobar el error global[6] y el error desagregado, a partir de la matriz de confusión[7] (Zelada 2017) de las predicciones.

Además, a través del Gráfico de disminución media de Gini (Figura 21) que se define como una medida de la varianza total en el conjunto de las *K* clases (Gil 2018) y a partir del cual es posible estudiar la importancia de las variables a la hora de clasificar los individuos encontrándose ordenadas en orden descendente (de las variables más importantes a las menos importantes).

2.5. Gráficos de radares

Estos gráficos se utilizan para comparar miembros o individuos a través de un gráfico de una dimensión en función de diversas métricas o variables (Trajkovic 2015).

[6] Porcentaje de error total

[7] Matriz de confusión, matriz de datos realizada a partir de la separación de la base de datos en datos de entrenamiento y de prueba la cual se utiliza para evaluar el modelo. La Tabla 1 muestra la matriz de confusión que nos permite calcular el error para cada uno de los árboles obtenidos.

*Tabla **Error! Main Document Only.** Matriz de Confusión.*

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Los gráficos de radar se desarrollarán a través del programa *Tableau*, enfocado en la visualización interactiva de datos. Los gráficos de radar (John *et al.* 1983) utilizados se fundamentan en las fórmulas de trigonometría que se pueden observar en la Figura 4.

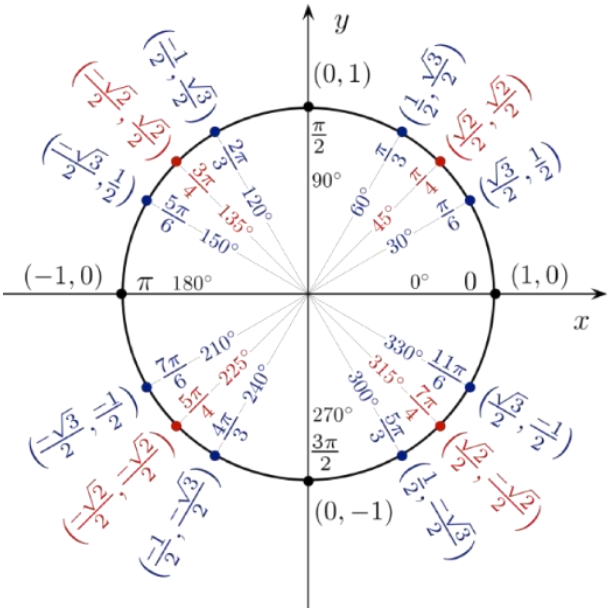

Figura 4 Fórmula para elaborar radares extraída de la web de Tableau.

Las diversas funcionalidades de *Tableau* permiten la elaboración de cálculos rápidos de las variables (percentiles, ránquines, sumatorios…), a través de los campos calculados, así como la personalización de gráficos y tablas seleccionando las diversas variables de nuestra base de datos. En el caso que nos ocupa, estas opciones nos ofrecen un amplio abanico de posibilidades a la hora de crear los gráficos de radar.

3. Análisis de la base de datos

Este bloque consta de dos partes que se corresponden a la descripción de la base de datos y al tratamiento de la misma para su posterior análisis.

3.1. Descripción de la base de datos

La base de datos con la que se contaba en un inicio estaba formada por 2.662 individuos y 74 variables explicativas, de las cuales el nombre del jugador, la posición, el equipo y la competición europea a la que pertenece (inglesa, alemana, italiana, francesa y española) son variables cualitativas. El resto de variables son cuantitativas y recogen la información sobre las acciones del juego, como, por ejemplo: número de pases ejecutados con éxito, número de goles, número de robos de balón, etc. (ver tabla en Anexo 9.1)

Es importante notar que no todas las variables se han utilizado para el desarrollo de todos los objetivos (la descripción de cómo se han seleccionado las variables se encuentra en el punto 3.2).

3.2. Tratamiento de los datos

Antes de comenzar el análisis, se ha llevado a cabo la limpieza y transformación de la base de datos.

En un principio, las variables explicativas recogían el número total de jugadas realizadas a lo largo de la temporada. Sin embargo, estas se han transformado para que indicaran el número total de acciones por partido, es decir, el número de veces que dicha variable explicativa es ejecutada cada 90 minutos. Para ello se ha aplicado sobre todas las variables la siguiente transformación:

$$(\text{Variable} / \text{Minutes played}) * 90\text{minutos}$$

Es importante indicar que el resultado de esta fórmula es acumulado, es decir, no se obtiene el número exacto de veces que cada jugada se realiza por partido, sino

cada 90 minutos. Por tanto, es posible que ese dato esté repartido a lo largo de varios partidos y no en uno único que se ha jugado en su totalidad.

En segundo lugar, se eliminaron todas las variables que tenían alta correlación entre ellas, ya que explicaban lo mismo. Por ejemplo: La variable de duelos aéreos representaba la suma de duelos aéreos ganados y duelos aéreos perdidos. Ante esto, se optó por eliminar la variable de duelos aéreos, repitiendo el proceso siempre que fuera necesario.

4. Resultados

En este apartado se presentan los resultados obtenidos en este trabajo fin de máster. Estos están relacionados con los objetivos mencionados anteriormente y se desarrollan de manera secuencial. Con todo ello se pretende abarcar desde las herramientas más generales a las más específicas que se pueden utilizar para facilitar el análisis y fichaje de un jugador.

4.1. Objetivo 1: Crear una aplicación web para encontrar jugadores similares a uno dado

Como se hizo alusión en la introducción, el objetivo general de este trabajo es el estudio de jugadores de fútbol de élite a partir del análisis de datos con la finalidad de facilitar el proceso de captación y fichajes en clubes profesionales.

En este contexto, el primer objetivo consiste en la creación de una aplicación web mediante el paquete de datos de *R* denominado *Shiny*. En concreto, el propósito es que la aplicación creada ofrezca la posibilidad de seleccionar a cualquiera de los jugadores incluidos en la base de datos, la cual se importa a través de la librería "*readxl"* (Wickham 2019). Una vez escogido un jugador, a través de un menú desplegable sencillo deben aparecer automáticamente los 20 jugadores más parecidos a este, ordenados de mayor a menor semejanza y, junto a ellos, un indicativo del porcentaje de similitud (Hamill 2019).

4.1.1. Desarrollo del modelo

Para desarrollar esta aplicación, en primer lugar, se debe proceder a la elección del modelo que se utilizará para ejecutar el programa. Puesto que el objetivo es el de agrupar a los jugadores más afines o similares, se debe emplear una técnica de aprendizaje no supervisado. En concreto, se va a utilizar el análisis de componentes principales (PCA) y después se generan los cluster a partir de las dos primeras componentes.

Antes de proceder al análisis se deben reescalar y centrar las variables. Este proceso es indispensable porque si una variable tiene una escala mucho mayor que el resto determinará, en gran medida, el valor de distancia/similitud obtenida al comparar las observaciones dirigiendo así la agrupación final (Rodrigo-Amat 2017). Por tanto, antes de comenzar el análisis, se reescala y se centran todas las variables de forma que la desviación típica sea de 1 y la media 0.

Una vez estandarizadas las variables, se ha realizado el análisis de componentes principales a partir de la función prcomp() y se ha determinado la necesidad de utilizar, al menos, 16 componentes principales si se desea tener un 90% de la variabilidad explicada. En la Figura 5 se puede apreciar un ejemplo de la visualización de los jugadores al proyectarse sobre dos componentes principales.

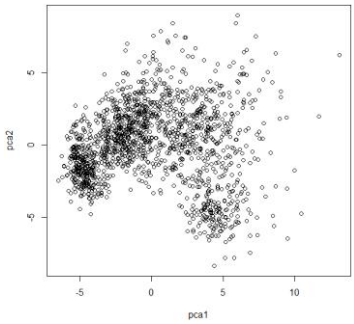

Figura 5 Ejemplo proyección 1 CP vs 2 CP

El modelo PCA proporciona información particular de las distancias entre los individuos permitiendo conocer el valor de cada jugador para cada componente principal extraída. Sin embargo, debe comprobarse el residuo de las observaciones de manera que se corrobore que ninguno de los datos es atípico[8] y, por lo tanto, se verifique que las distancias aportadas por el análisis de componentes principales son las idóneas.

Una vez obtenido el PCA, se procede a la interpretación del valor de los residuos a través del cálculo del error de predicción de las observaciones o SPE. Puesto que se cuenta con 1.529 observaciones, asumiendo como límite superior del SPE su percentil 99%, sería aceptable que unas 15 observaciones cayeran no lejos de dicho límite.

[8] Observaciones cuyo residuo elevado es detectado por el gráfico de SPE y provocan una ruptura en la estructura de correlación

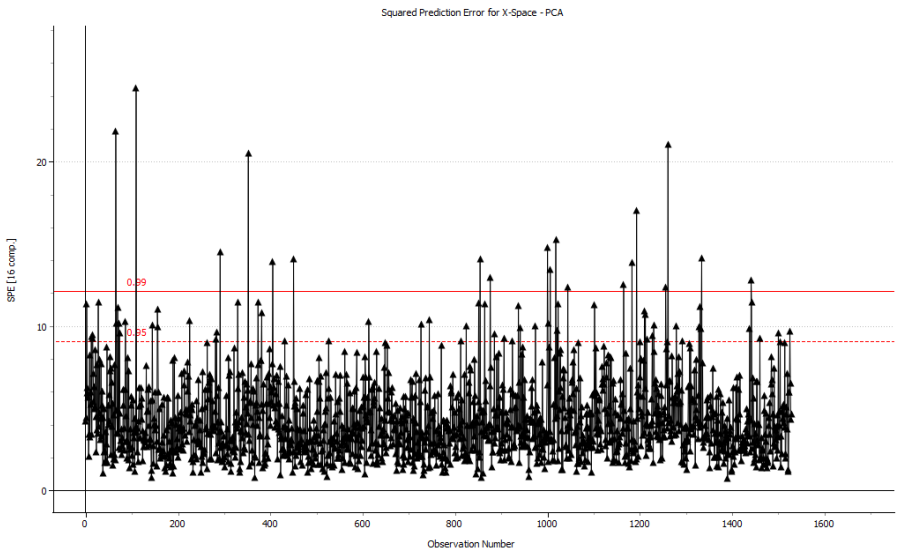

Figura 6 Gráfico SPE para los jugadores

A la luz del gráfico de SPE (Figura 6) se comprueba en primera instancia que son 19 las observaciones que han caído fuera del límite al 99%. Además, se observa que el SPE de cuatro de los jugadores algo elevado. Sin embargo, estos individuos finalmente no han sido eliminados, ya que, tras analizar el gráfico de contribución de los residuos de cada jugador sobre el espacio de las X, no se observó que sus acciones sobre el terreno de juego fueran atípicas respecto al resto de jugadores, más allá de que en la temporada analizada resaltaran en jugadas clave como: número de pases hacia atrás (Lorenzo Insigne), intercepciones de balón (Benoit Assou), goles convertidos (Ragnar Klavan) y pases clave (Eden Hazard). (ver Figuras en Anexo 9.2.)

Una vez analizado el gráfico de SPE, se procede a la creación de la aplicación web para encontrar jugadores similares a uno dado.

En primer lugar, se ha aplicado un filtro sobre la base de datos general debido a que muchos jugadores incluidos inicialmente no han participado apenas durante la temporada. Por tanto, se ha escogido a aquellos jugadores que han disputado al menos un 20% de los minutos totales jugados con el objetivo de reducir la distorsión de los resultados. La base de datos resultante es de 1.529 jugadores.

A partir de la matriz de *scores* del PCA con 16 componentes principales, se ha calculado la matriz de correlaciones entre individuos a través de la función *cor()*, perteneciente al paquete de datos de R *"PerformanceAnalytics"* (Peterson *et al.* 2019), de manera que la correlación entre los jugadores se encuentre recogida en el intervalo

de -1 y 1. De este modo, se obtiene la medida de similitud entre los jugadores como puede verse en la Tabla 2.

Tabla 1 Extracto matriz correlación

	Ashley Young	Luis Antonio Valencia	Alexis Sánchez	Juan Mata	Chris Smalling	Ander Herrera	Nemanja Matic	Romelu Lukaku
Ashley.Young	1.000000000	7.319650e-01	0.086929692	0.402464315	-0.03863235	0.185587258	0.379213546	-0.36965735
Luis.Antonio.Valencia	0.731965022	1.000000e+00	-0.134282304	0.415778759	0.22243421	0.194405175	0.565316659	-0.18204684
Alexis.Sánchez	0.086929692	-1.342823e-01	1.000000000	0.530982611	-0.58830975	-0.026155067	-0.140206275	0.45701172
Juan.Mata	0.402464315	4.157788e-01	0.530982611	1.000000000	-0.51159322	-0.021651315	0.033876919	0.22562822
Chris.Smalling	-0.038632353	2.224342e-01	-0.588309752	-0.511593223	1.00000000	-0.098351262	0.229339701	0.07485119
Ander.Herrera	0.185587258	1.944052e-01	-0.026155067	-0.021651315	-0.09835126	1.000000000	0.572551810	-0.61450397
Nemanja.Matic	0.379213546	5.653167e-01	-0.140206275	0.033876919	0.22933970	0.572551810	1.000000000	-0.44653750
Romelu.Lukaku	-0.369657350	-1.820468e-01	0.457011716	0.225628224	0.07485119	-0.614503967	-0.446537499	1.00000000

La aplicación creada en *Shiny* utilizará la matriz de correlaciones para ofrecer la información solicitada. De este modo, cada vez que se le pida a la aplicación quiénes son los jugadores más similares a uno en particular, la interfaz comparará automáticamente a todos los jugadores incluidos en la base de datos respecto al seleccionado. Finalmente, la aplicación devolverá un *ranking* de los jugadores más similares al especificado, que serán aquellos cuyo valor en la matriz de correlación sea más próximo a uno.

4.1.2. Similitud entre jugadores (*Shiny*)

Una vez configurado "*server*" con los datos obtenidos a partir de la matriz de correlación y tras el diseño de la aplicación a partir del paquete de datos de *R* "*shinyWidgets"* (Perrier 2019) mediante la función "*ui*" que ofrece la posibilidad de personalizar *widgets* de la aplicación (ver apartado 2.1.), se cargan todos los paquetes al servidor a través de la librería "*tidyverse"* (Wickham 2017), obteniendo como resultado:

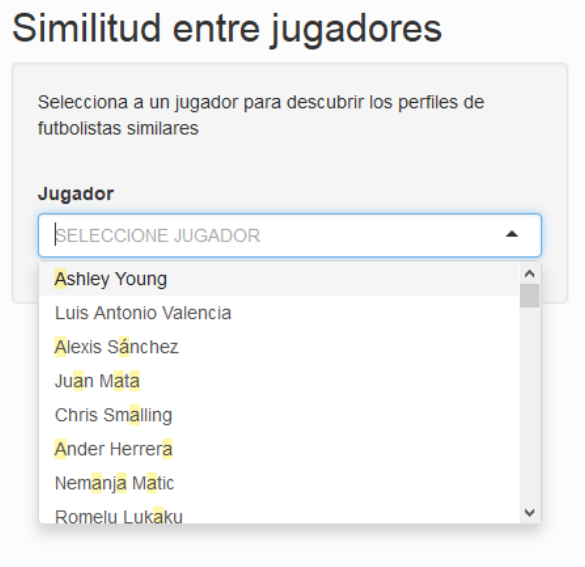

Figura 7 Vista aplicación Shiny

Al ejecutar la aplicación a partir de la función *ShinyApp* surge un panel de información y una caja o menú de selección como puede apreciarse en la Figura 7. Para realizar la búsqueda del jugador que se desea analizar se puede desplegar el menú y encontrarlo manualmente, o bien de forma automática introduciendo su nombre.

Una vez introducido el jugador en cuestión, el programa recupera la información que se muestra en la Figura 8 con Daniel Parejo a modo de ejemplo:

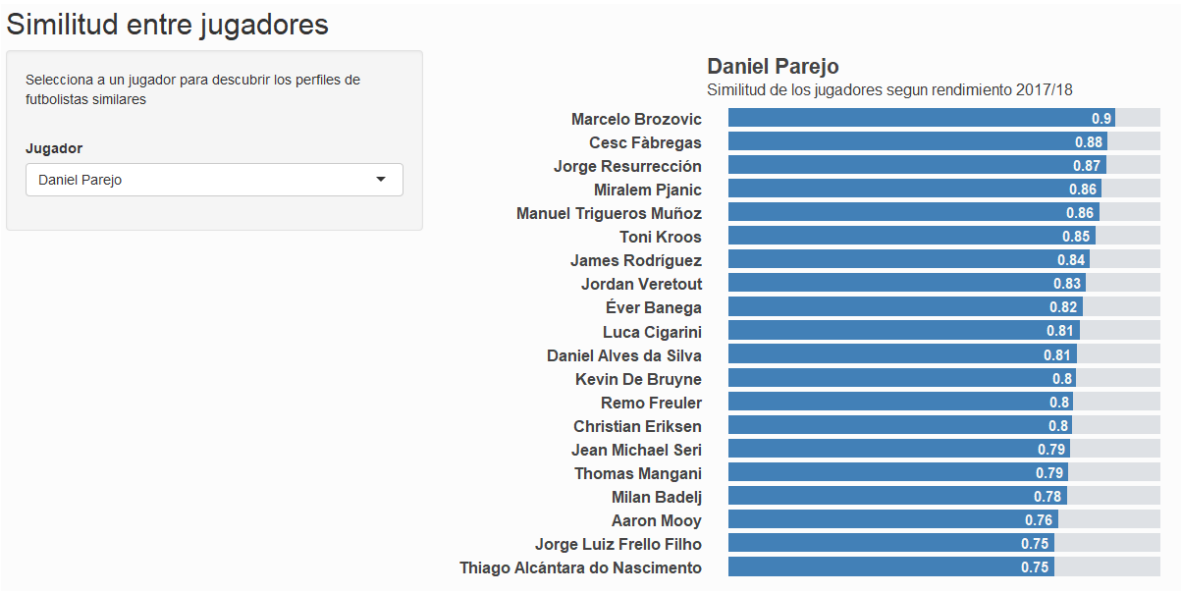

Figura 8 Aplicación Shiny. Salida Similitud entre jugadores. Daniel Parejo

La aplicación mostrará de forma predeterminada a los 20 jugadores más similares al solicitado, según lo calculado anteriormente en la matriz de correlación. Estos jugadores aparecen por orden descendente según su semejanza con el jugador seleccionado y junto a ellos la medida de similitud calculada.

En definitiva, esta herramienta fundamentada en el *Machine Learning* ofrece la posibilidad de encontrar de forma rápida y dinámica a los jugadores de mayor similitud entre sí. A partir de las variables de rendimiento de la base de datos, que describen el comportamiento de los jugadores sobre el terreno de juego, esta aplicación pretende facilitar el proceso de captación para fichar futbolistas en un club profesional con el fin de reemplazar, por ejemplo, a un jugador lesionado o que ha sido traspasado recientemente a otro club.

4.1.3. Visualización de los jugadores según las variables

Este apartado busca profundizar, de un modo más específico, en el estilo de los jugadores a partir de la información ofrecida por la aplicación.

Como se ha observado en la Figura 8 del apartado anterior, la aplicación *Shiny* ofrece los 20 jugadores más parecidos al referencial indicando su índice de similitud. Sin embargo, estos pueden comportarse de forma diferente sobre el terreno de juego por diversos aspectos. Estas diferencias y similitudes se analizarán a continuación.

El proceso de análisis se realiza a partir de las componentes principales extraídas en el PCA con la finalidad de estudiar la proyección de los individuos sobre el hiperplano. A través de este estudio se podrá comparar la distribución de los jugadores según las variables que sean más representativas según las componentes principales utilizadas. Este análisis se llevará a cabo con los mismos jugadores obtenidos en el ejemplo de la aplicación de *Shiny* (Figura 8) y se empleará el *Score Plot* en combinación con el *Loading Bi-Plot*.

En primer lugar, se analiza la información contenida en las diferentes componentes principales. De esta manera, se escogen aquellas que guardan la información más útil para el objetivo propuesto. Es decir, como los jugadores analizados son centrocampistas, las componentes principales seleccionadas para realizar el análisis serán aquellas que posean un mayor índice de variabilidad explicada de las variables más representativas para esta posición y lo mismo con el resto de posiciones.

Una vez realizado el análisis de todas las componentes, se escogen la 2, 3 y 7 (ver Figuras en Anexo 9.3). Como se observa en la segunda componente principal las variables más representativas son aquellas que describen particularmente acciones de gol. En el caso de la tercera componente principal, las variables más representadas son aquellas relacionadas con aspectos constructivos del juego como el acierto en los pases y en los centros. Finalmente, la séptima componente principal se ha escogido porque se incorporan variables que describen el carácter defensivo como, por ejemplo, las recuperaciones de balón.

Una vez realizado el PCA, y escogidas las componentes principales más representativas para el análisis, se ha utilizado el programa *Aspen ProMV* para visualizar cómo varía la proyección de los jugadores sobre el espacio de las X en función de las componentes utilizadas y de la distribución de las variables sobre el plano a través del Loading-Bi plot (Figuras 9 y 10)

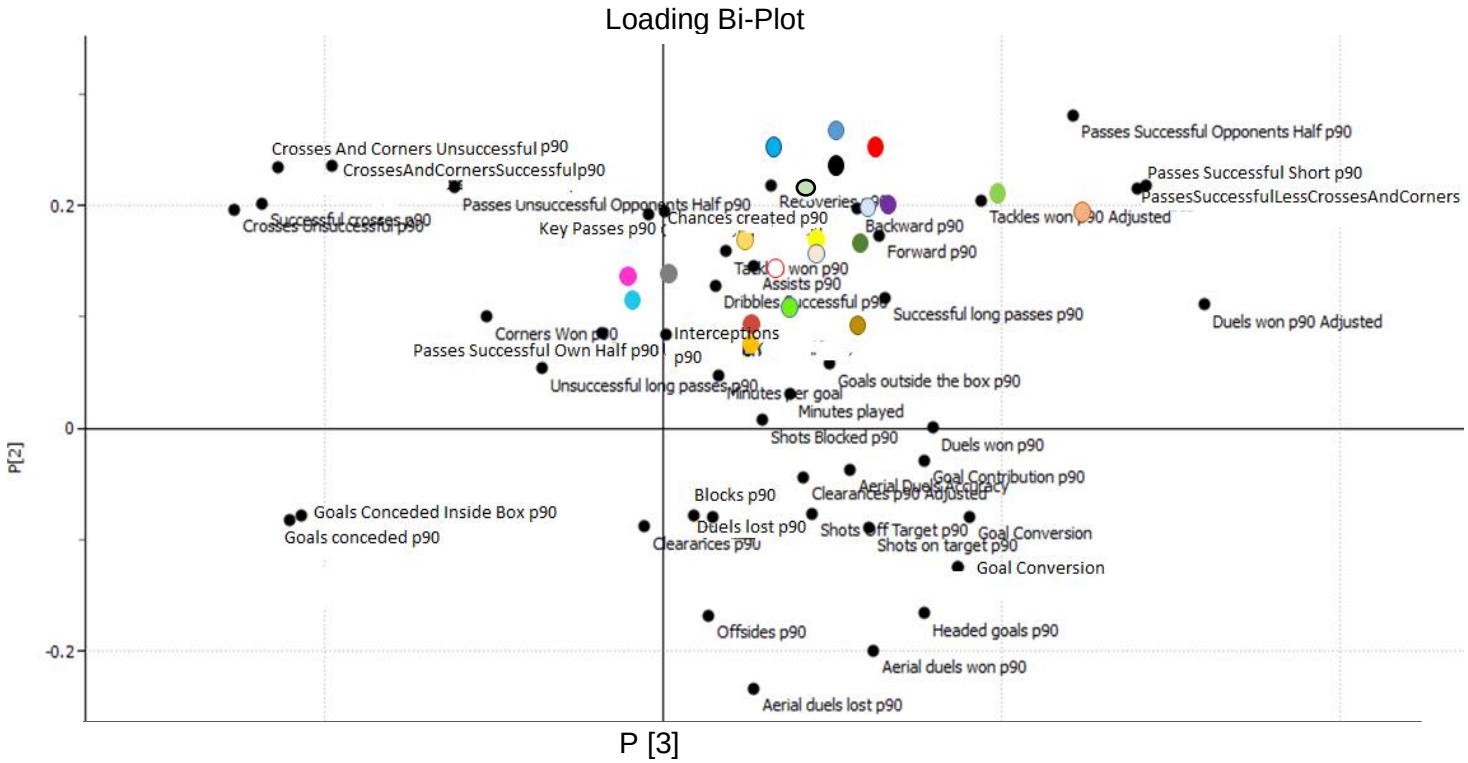

Figura 9 Loading Bi-Plot Proyección jugadores similares T3/T2

Daniel Parejo
ChristianEriksen
Thiago Alcántara
Toni Kroos
Marcelo Brozovic
Jorge Luiz Frello
Miralem Pjanic
Kevin de Bruyne
James Rodríguez
Cesc Fàbregas
Éver Banega
Jean Michael Seri
Daniel Alves
Manuel Trigueros
Jorge Resurrección
Milan Badelj
LucaCigarini
Thomas Mangani
Remo Freuler
Aaron Mooy
Jordan Veretout

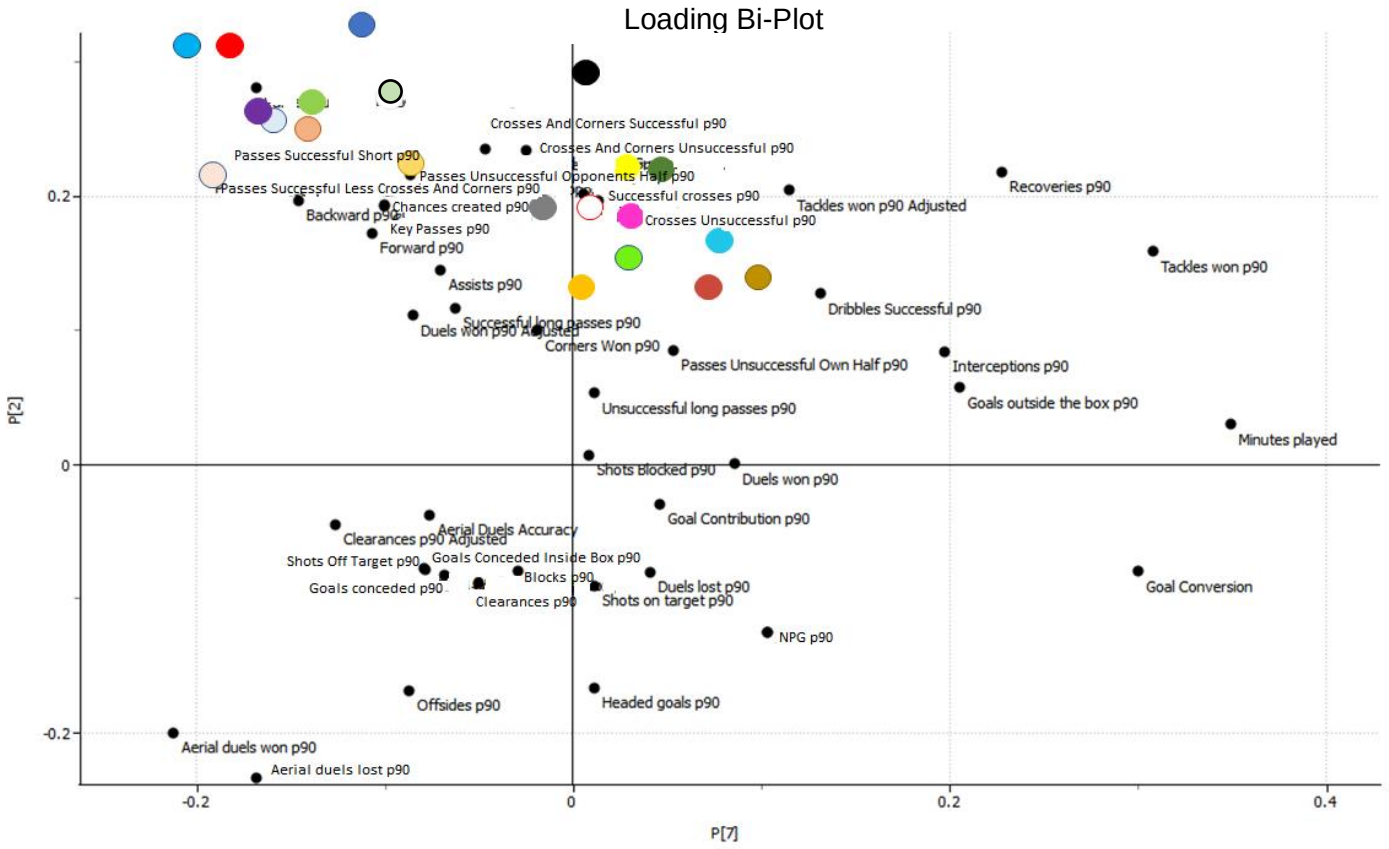

Figura 10 Loading Bi-Plot Proyección jugadores similares T7/T2

- Daniel Parejo
- ChristianEriksen
- Thiago Alcántara
- Toni Kroos
- Marcelo Brozovic
- Jorge Luiz Frello
- Miralem Pjanic
- Kevin de Bruyne
- James Rodríguez
- Cesc Fàbregas
- Éver Banega
- Jean Michael Seri
- Daniel Alves
- Manuel Trigueros
- Jorge Resurrección
- Milan Badelj
- LucaCigarini
- Thomas Mangani
- Remo Freuler
- Aaron Mooy
- Jordan Veretout

Si se comparan los gráficos de las Figura 9 y 10 se observa que la distribución de los jugadores sobre el plano varía según la componente principal utilizada.

En la Figura 9, se aprecia que un grupo de jugadores se encuentra desplazado hacia el extremo derecho superior, estos son: Thiago Alcántara, Jorge Luiz Frello, James Rodriguez, Kevin de Bruyne, Daniel Alves, Ever Banega, Marcelo Brozovic y Toni Kroos. Estos se encuentran cerca de las variables que determinan acierto en los pases, y en sentido opuesto a variables relacionadas con los pases fallados (estas son las variables más importantes en la tercera componente), indicando que ese grupo de jugadores tiende a tener más aciertos que fallos en los pases en relación a la media de los jugadores analizados. En el caso de los jugadores que se encuentran en el centro de la Figura 9, siendo estos: Jordan Veretout, Jorge Resurrección, Manuel Trigueros,

Milan Badelj, Remo Freuler y Luca CIgarini estos toman valores medios para aciertos y fallos en pases. Respecto a la segunda componente todos los jugadores están correlacionados negativamente con las variables que determinan los duelos aéreos y los goles de cabeza.

En la Figura 10, se observa que las variables que cobran mayor relevancia para la séptima componente son las defensivas, identificándose a aquellos jugadores que efectúan un mayor número de jugadas relacionadas con los robos de balón: Aaron Mooy, Jordan Veretout, Manuel Trigueros, Milan Badelj y Remo Freuler. De este modo, en este gráfico se diferencian dos bloques de jugadores en base a las variables que describen un juego defensivo (a la derecha del eje de abcisas) o creativo (a la izquierda de eje de abcisas).

En conclusión, este método de análisis es eficaz para reducir el número de individuos requeridos según los intereses del fichaje de un club específico tras haber identificado anteriormente a los 20 jugadores más similares a uno concreto. De este modo, el uso de las componentes principales permite agrupar a los jugadores por características del juego más específicas que el método desarrollado con *Shiny*, que únicamente ofrece el dato general de similitud entre jugadores.

4.2. Objetivo 2: Comparar jugadores

En el objetivo anterior se ha desarrollado una forma rápida y dinámica para encontrar perfiles de jugadores parecidos a uno específico, así como un método eficaz para agrupar a los jugadores por sus estilos de juego. Sin embargo, el grado de similitud entre dos futbolistas otorgado por la aplicación es una información global que revela, a grandes rasgos, quiénes son los jugadores que guardan un estilo de juego más parecido. Este resultado carece de matices que describan con precisión las particularidades y características del comportamiento de los individuos.

En este sentido, el segundo objetivo pretende complementar a la aplicación desarrollada con *Shiny* y al método realizado posteriormente. Para ello se ha creado una herramienta a partir de *Tableau* que compara a los jugadores resultantes en base a una serie de variables específicas. Estas variables se han escogido según las características del jugador de referencia seleccionado, aunque podrían adaptarse a las necesidades de cualquier club profesional que, hipotéticamente, requiriera de estos recursos.

Como ya se ha mencionado, para la comparación de jugadores se utilizará el mismo jugador de referencia que se empleó a partir de la salida del programa *Shiny* en el objetivo anterior: Daniel Parejo, futbolista del Valencia Club de Fútbol. En primer lugar, se crea una nueva base de datos en la que únicamente se incluyan aquellos jugadores cuya posición es la de centrocampista, ya que esta es la posición que ocupa el jugador referencial. A continuación, se utiliza el programa *Aspen ProMV* para llevar a cabo el análisis de componentes principales a partir del cual, se realizará el gráfico de contribuciones del jugador. Este gráfico aporta una valiosa información sobre las variables más representativas según el estilo de juego del individuo analizado.

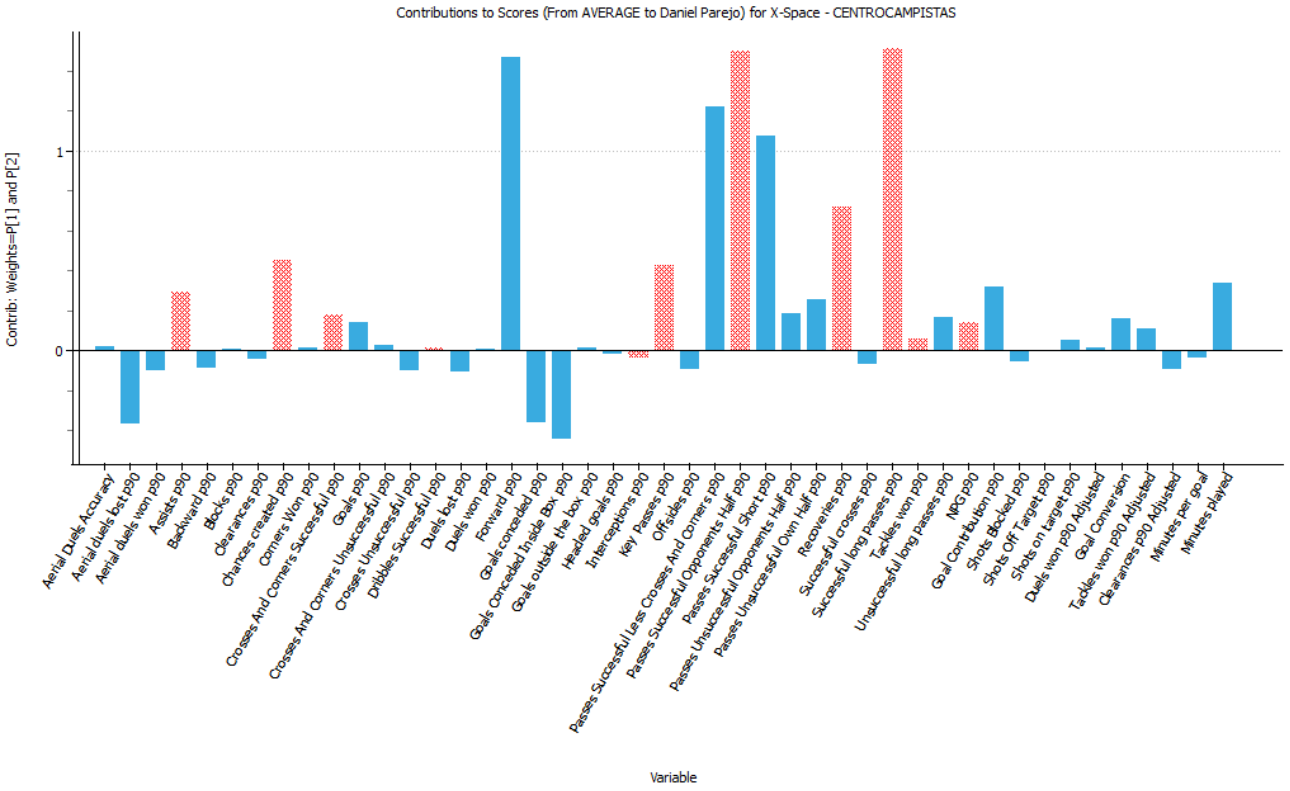

Figura 11 Contribuciones Parejos vs centrocampista medio

En el gráfico de contribuciones de la Figura 11 se observan las 45 variables utilizadas para comparar al jugador referencia Daniel Parejo frente al centrocampista medio. Valores positivos (negativos) de las barras indican que el jugador seleccionado toma valores mayores (menores) que el jugador medio. Las variables que se utilizarán para llevar a cabo el análisis comparativo entre los jugadores por ser las más representativas en el estilo de juego de Daniel Parejo están resaltadas en rojo: centros y saques de esquina ejecutados con éxito, pases exitosos en campo contrario, pases largos exitosos, pases clave, ocasiones creadas, asistencias, recuperaciones y entradas ganadas. Todas las variables escogidas indican jugadas ejecutadas con éxito. El análisis de las contribuciones permite analizar de un modo más específico a los jugadores recuperados por la aplicación para compararlos entre ellos. Esta comparación aporta un cariz descriptivo de mayor profundidad que sirve para discernir entre aquellos individuos que, pese a conservar grados de similitud muy elevados, tienen estilos de juego diferentes.

En segundo lugar, se comparan a los jugadores más parecidos al jugador seleccionado (en nuestro caso Daniel Parejo) obtenidos con la aplicación Shiny. El objetivo consiste en encontrar un sustituto lo más similar posible al referencial. Por tanto,

la comparación entre los jugadores se desarrolla a partir de las variables obtenidas mediante el gráfico de contribuciones. A modo de ejemplo, se añaden cuatro variables más que se podrían haber solicitado por la dirección deportiva de un club de fútbol profesional. En este caso las variables de rendimiento seleccionadas son: goles (sin incluir penaltis), interceptaciones, disparos a portería y regates completados con éxito.

Los jugadores comparados son Miralem Pjanic (Juventus) y Manuel Trigueros (Villarreal). Ambos futbolistas tienen un porcentaje de similitud con Daniel Parejo (Valencia CF) del 86%. Se comparan estos jugadores utilizando las variables que han resultado más representativas para el jugador referencial en el gráfico de contribuciones. Esta comparación se realiza a través de gráficos de radar desarrollados con el programa de visualización de datos *Tableau*, como muestran las Figuras 12 y 13.

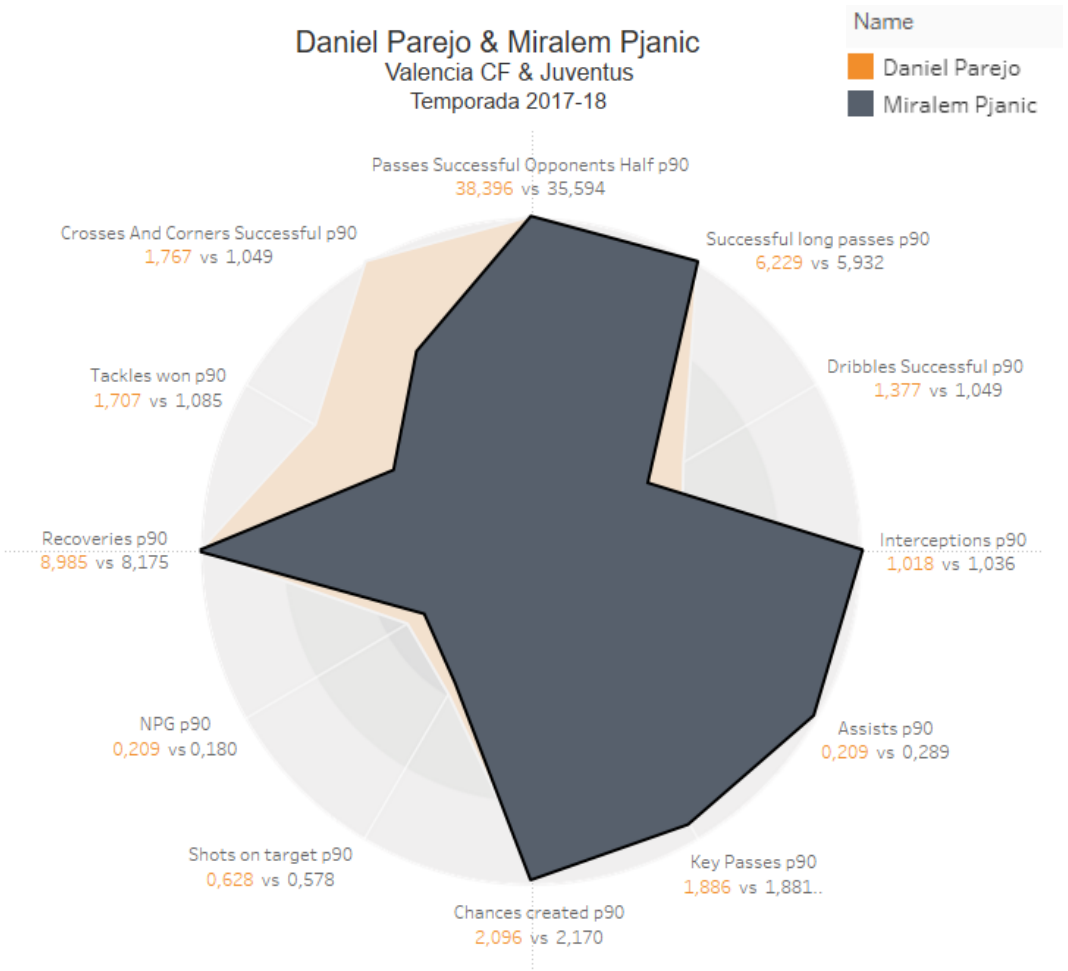

Figura 12 Radar Daniel Parejo vs Miralem Pjanic

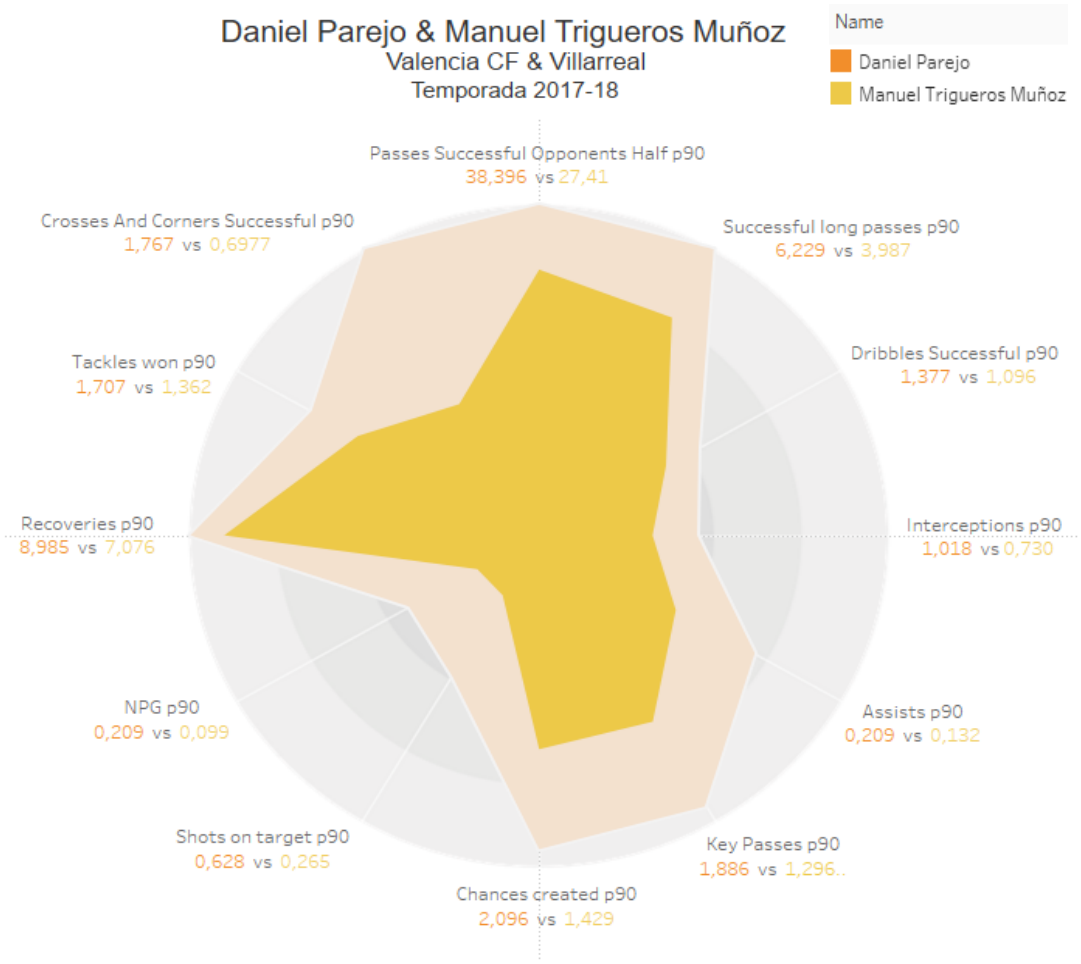

Figura 13 Radar Daniel Parejo vs Manuel Trigueros Muñoz

Este proceso corrobora que, aunque el porcentaje de similitud respecto a Daniel Parejo es equivalente (Figura 8), las características del estilo de juego entre los jugadores analizados poseen matices diferenciales. Las diferencias observadas en el gráfico de radar (Figuras 12 y 13) se deben a que se utilizaron 45 variables para realizar el análisis de componentes principales y obtener los porcentajes de similitud entre los jugadores. Sin embargo, en el gráfico de radar únicamente se han tenido en cuenta 12 variables consideradas como las más representativas en el estilo de juego de Daniel Parejo. Por ello, al comparar a Parejo con Miralem Pjanić y Manuel Trigueros (Figuras 12 y 13) se pueden apreciar diferencias importantes. De los dos gráficos de radar se concluye que, en las características seleccionadas, Miralem Pjanic es más similar a Parejo que Manuel Trigueros.

Tras este primer análisis se concluye que la aplicación desarrollada es una herramienta útil y valiosa como punto de partida a la hora de encontrar a los jugadores más similares entre sí. Sin embargo, este proceso requiere de un análisis complementario más específico donde se detallen con mayor precisión las cualidades y

el estilo de juego en cuestión. Como se ha visto, los jugadores obtenidos por la aplicación *Shiny* pueden tener un porcentaje de similitud muy elevado, pero diferir cuando se ahonda en sus características específicas. Con todo, el gráfico de radar se muestra como una herramienta complementaria de gran utilidad por su facilidad de uso y de comprensión para visualizar los datos.

4.3. Objetivo 3: Análisis de posiciones

A lo largo de este estudio se ha trabajado con una base de datos compuesta por 1.529 jugadores clasificados en tres posiciones: defensa, mediocentro o delantero, pero los futbolistas se distribuyen en posiciones más específicas en el terreno de juego. La base de datos original, por ejemplo, no diferencia entre defensas centrales y defensas laterales.

El objetivo de este apartado es conocer las variables más representativas para cada posición más específica con la finalidad de facilitar la tarea a ojeadores[9] (Pérez 2018) y analistas para que presten especial atención a las acciones del juego (variables) que definen mayoritariamente una posición concreta a la hora de seguir a un jugador para su posible contratación. Por todo ello, y a modo de ejemplo, se llevará a cabo el análisis de la demarcación de defensa de manera que el análisis de datos permita diferenciar entre centrales y laterales. Este modelo de análisis servirá para los sucesivos estudios del resto de demarcaciones en casos futuros.

4.3.1. Análisis gráfico de contribuciones

En el caso que nos ocupa, las posiciones que se analizarán son las de defensa central y defensa lateral (Lombardi 2018):

Centrales: se refiere a aquellos jugadores que desempeñan la mayoría de sus acciones en el centro de la defensa, cerca del área de su propio equipo. Los jugadores que ocupan esta posición se caracterizan por tener inteligencia táctica y liderazgo, complexión física fuerte, estatura alta y, por tanto, buen dominio del juego aéreo.

Laterales: se refiere a aquellos jugadores que desempeñan sus acciones, generalmente, en las bandas en la zona defensiva. Esta posición la suelen ocupar futbolistas con suficiente despliegue físico, pues su deber también consiste en ofrecer apoyo al equipo en fase ofensiva. Por tanto, su rol se basa en brindar amplitud al terreno de juego tanto con responsabilidad ofensiva como defensiva.

[9] El ojeador es la persona que observa los partidos de fútbol con el objetivo de identificar talento, ya sea en jugadores jóvenes o no, que puedan ser incorporados al equipo.

Para llevar a cabo el objetivo propuesto es necesario comparar a los futbolistas que compiten en las demarcaciones que van a ser analizadas. Para realizar esta comparación se seleccionan los jugadores a partir de las valoraciones realizadas por el videojuego FIFA 18 que muestra a los mejores jugadores según sus actuaciones en la última temporada (EASPORTS 2018). De este modo, se han elegido para realizar el estudio a los mejores defensas centrales y laterales de la temporada 2017/2018.

En primer lugar, a través del programa *Aspen ProMV* se realiza un análisis de componentes principales utilizando únicamente la base de datos que contiene a los futbolistas que juegan en la posición de defensa. Seguidamente, se elabora un gráfico de contribuciones con los jugadores que han sido identificados como centrales y laterales frente al defensa promedio, respectivamente. A través del análisis de este gráfico será posible identificar las variables que destacan en ambas posiciones al comparar los gráficos de los jugadores que compiten en estas demarcaciones. De este modo, se comprobará si existen diferencias en el estilo de juego y, en su caso, se determinarán qué variables resultan características para cada tipo de defensa, tanto para centrales como para laterales.

A continuación, se ha llevado a cabo el análisis de las actuaciones de los futbolistas escogidos frente a las del defensor medio a través del gráfico de contribuciones con la finalidad de comparar las variables que han resultado representativas en los defensas centrales respecto a las de los defensas laterales. En las figuras 14 y 15 se puede ver un ejemplo de los resultados aportados por el defensa central Sergio Ramos y el lateral Marcelo, ambos jugadores del Real Madrid.

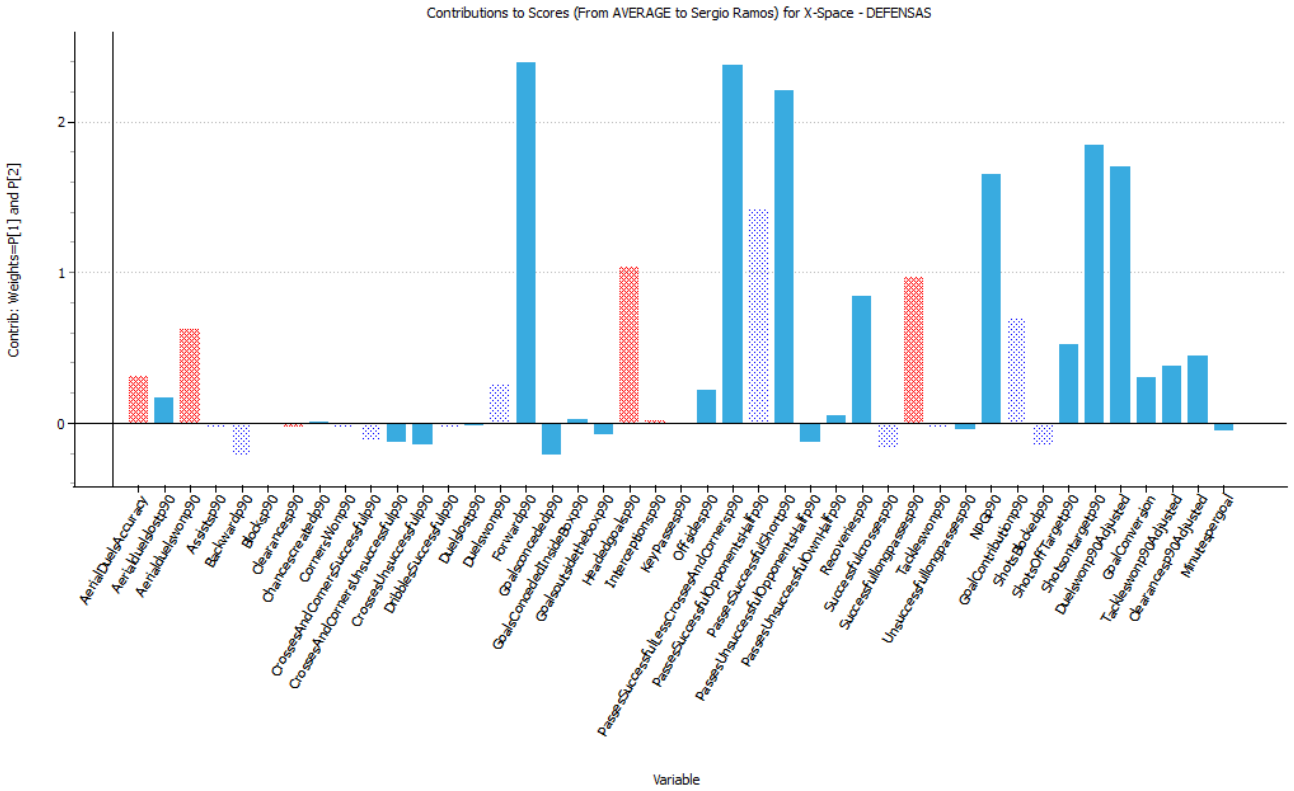

Figura 14 Contribuciones Sergio Ramos vs defensa promedio

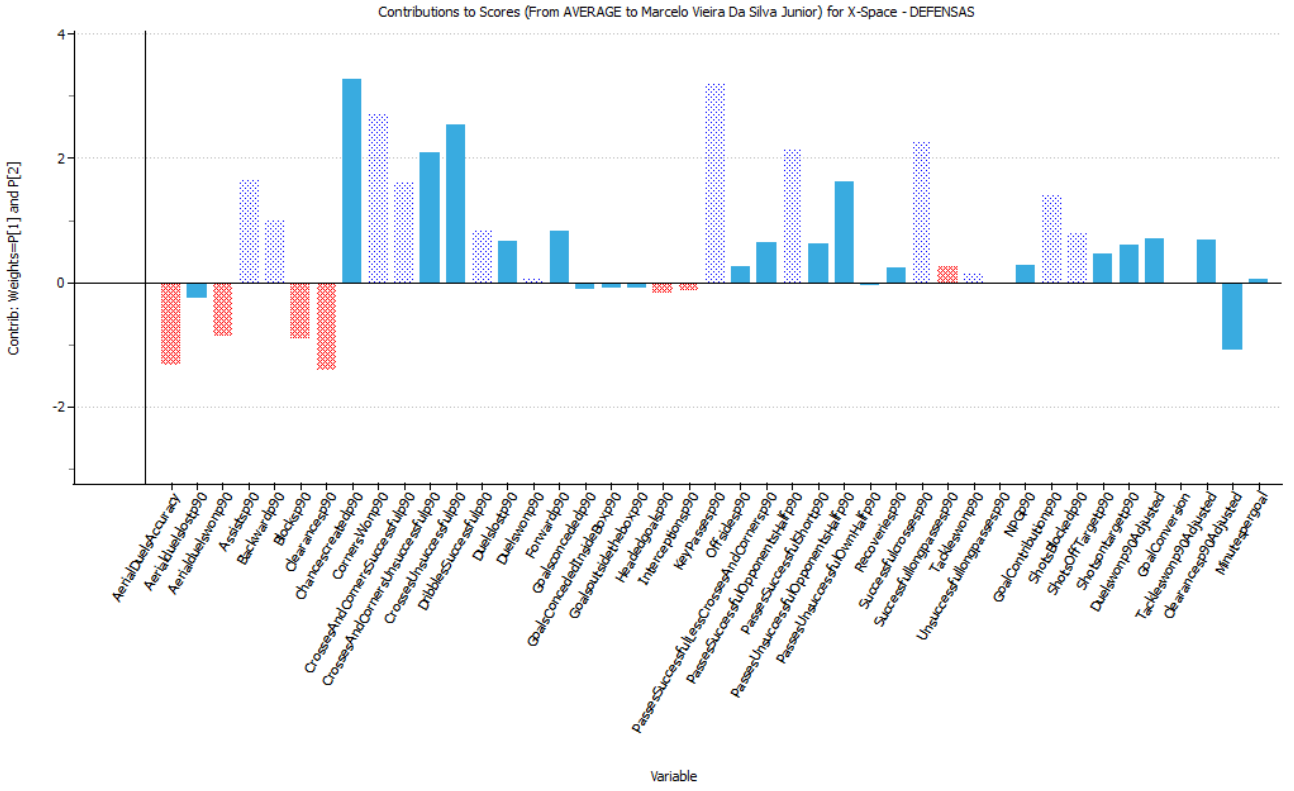

Figura 15 Contribuciones Marcelo vs defensa promedio

Si se comparan los jugadores de forma individual, como en las Figuras 14 y 15 en los casos de Sergio Ramos y Marcelo frente al defensor promedio, se puede evidenciar que existen diferencias importantes entre ambos jugadores. En rojo están resaltadas las variables que han resultado representativas para los centrales analizados y en azul celeste aquellas que lo han sido para los laterales.

Este análisis comparativo se ha repetido con varios jugadores (ver Figuras en Anexo 9.5) para comprobar que el resultado no fuera fruto del azar. De este modo, se ha confirmado que los jugadores que compiten en las mismas posiciones tienen un comportamiento similar. Esta conclusión ha sido obtenida a partir de la visualización de los gráficos de contribución de los centrales (ver Figuras 14, 31 y 32) y los laterales (ver Figuras 15, 34, 36).

Si se observa de un modo más detenido las variables que resultan representativas en cada jugador se advierte que, en el caso de los defensas centrales, estos sobresalen por la cantidad de jugadas defensivas que ejecutan con éxito, así como por la cantidad de duelos aéreos que disputan (Figura 14, 33 y 35). En contra, los laterales se muestran como jugadores más creativos y ofensivos siendo importantes aquellas variables que tienen que ver con jugadas que terminan en gol o han podido serlo, los duelos uno contra uno y la cantidad de centros y pases ejecutados en campo contrario (ver Figuras 15, 32, 34).

Otra manera de desarrollar el análisis de los jugadores mediante el gráfico de contribuciones es la de comparar directamente a los defensas centrales y laterales. En la Figura 16, se presenta el gráfico de contribución de los defensas centrales frente a los laterales.

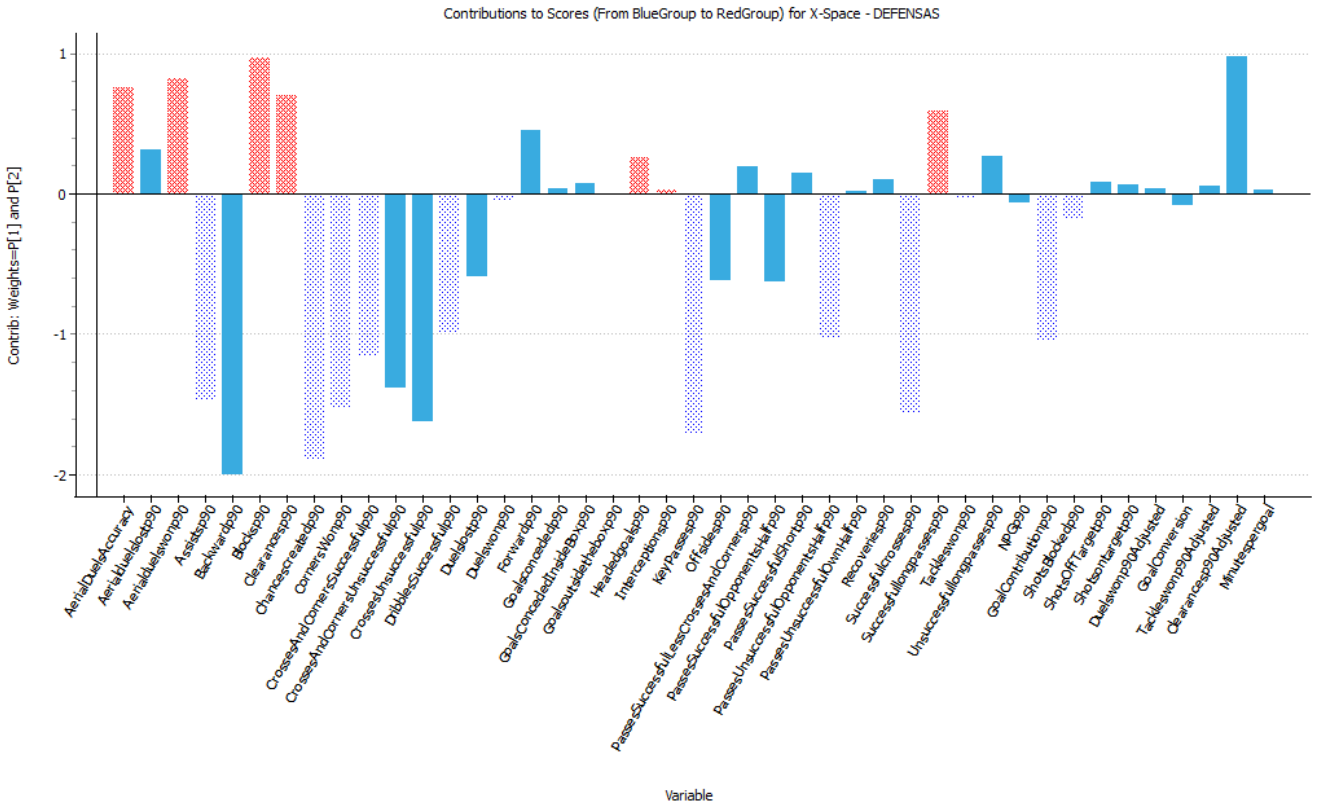

Figura 16 Contribuciones Centrales vs Laterales

A la luz del gráfico de contribuciones de la Figura 16, se observa que las variables que resultaban representativas tanto para los defensas centrales como para los defensas laterales continúan siendo las mismas. En el caso de los centrales: duelos aéreos ganados, porcentaje de duelos aéreos ganados, disparos bloqueados, despejes, goles de cabeza, intercepciones y pases largos ejecutados con éxito. En el caso de los laterales: número de asistencias, oportunidades de gol creadas, córners ganados (cabeceados con éxito), corners y centros ejecutados con éxito, regates exitosos, duelos ganados, pases clave, pases exitosos en campo contrario, centros ejecutados con éxito, entradas a ras de suelo ganadas, número total de asistencias y goles y disparos a puerta bloqueados.

4.3.2. Análisis exploratorio

Como se mencionó al inicio del objetivo, la base de datos con la que se cuenta no diferencia entre defensas centrales y defensas laterales. Por tanto, es necesario utilizar un método de aprendizaje no supervisado porque, como se dijo al inicio del trabajo, estos son utilizados para clasificar a los individuos en grupos. Este análisis exploratorio tiene como objetivo descubrir si los individuos se clasifican de un modo diferente en caso de utilizar todas las variables o, únicamente, las que han resultado representativas (ver Tabla 3).

Por tanto, una vez analizadas ambas posiciones, se procede a la creación de una segunda base de datos en la que sólo se tengan en cuenta las variables seleccionadas, es decir, aquellas más representativas para la posición de defensa central y lateral en cada caso. Las variables escogidas se muestran en la Tabla 3.

Tabla 2 Variables representativas centrales vs laterales

Centrales	**Laterales**
AerialDuelsAccuracy	Assistsp90
Aerialduelswonp90	Chancescreatedp90
Blocksp90	CornersWonp90
Clearancesp90	CrossesAndCornersSuccessfulp90
Headedgoalsp90	DribblesSuccessfulp90
Interceptionsp90	Duelswonp90
Successfullongpassesp90	KeyPassesp90
	PassesSuccessfulOpponentsHalfp90
	Successfulcrossesp90
	Tackleswonp90
	GoalContributionp90
	ShotsBlockedp90

Todas las variables seleccionadas corresponden a jugadas exitosas puesto que la intención es descubrir las variables positivas que ayuden a fichar jugadores. Por tanto, se considera que la búsqueda de un jugador para su posterior contratación no se ejecuta por sus errores, sino por sus aciertos. Para ello, la base de datos en la cual se encuentran todas las variables se ha filtrado para que únicamente se consideren las variables que representan las jugadas exitosas. Así, además de la base de datos original con todas las variables, se dispone de una segunda base de datos con las variables seleccionadas.

El análisis exploratorio se ha ejecutado a partir del análisis clúster. Más concretamente, se ha utilizado el método *k-means* para el estudio de ambas bases de datos con la finalidad de comprobar cómo varía la clasificación entre grupos de los individuos dependiendo de las variables utilizadas (todas las variables o las variables representativas). En el método *k-means* es el analista quien debe escoger el valor de *k* (número de grupos). Sin embargo, su valor no se indicará directamente, sino que se calculará de modo que se compruebe si las variables escogidas son adecuadas para agrupar el número de posiciones requerido, es decir a laterales y centrales. El cálculo se ha realizado a partir del paquete de *R "e1071"* (Meyer 2019), que, a través de la iteración de diferentes centroides, aplica la función *k-means* y devuelve la suma total de los errores internos (ver Figura en Anexo 9.6).

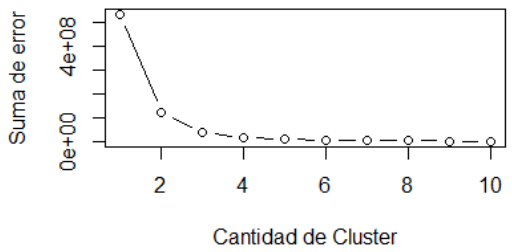

Figura 17 K óptimo (todas las variables)

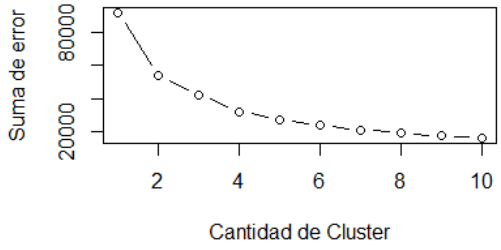

Figura 18 K óptimo (variables representativas)

A la luz de estos resultados ofrecidos por las Figuras 17 y 18 se comprueba que, tanto si se utilizan todas las variables como si se utilizan únicamente aquellas consideradas representativas, se reconoce la existencia de dos grupos de jugadores diferentes.

Por tanto, una vez seleccionado *k=2* se procede al estudio del error global y desagregado (ver 2.4). En primer lugar, se realiza un clúster mediante la función de *R kmeans*. Esta función encuentra una cantidad de soluciones de agrupamiento de *k-medias* y, finalmente, clasifica a cada individuo dentro de cada grupo buscando la suma mínima total dentro del grupo de distancias al cuadrado (Fox 2017). Por ejemplo, en este caso cada individuo se habrá asignado al clúster 1 o 2: lateral o central, respectivamente.

En segundo lugar, se observa el *cluster means* que indica la media para hallar los centroides por variables. Esta información es aportada por la función *kmeans* y, a partir de ella, es posible determinar qué posiciones han tomado los valores 1 y 2 (ver Anexo 9.7). El análisis de los clústeres se ha llevado a cabo en base a la información aportada por el gráfico de contribuciones (Figura 16). Los laterales realizan un mayor

número de pases claves, centros, regates y generan más oportunidades de gol, mientras que los centrales ganan un mayor número de duelos aéreos y tienen una mayor incidencia en jugadas defensivas. A partir de esta información se ha comprobado que, en el caso de la base de datos con todas las variables, los laterales estaban agrupados en el clúster 2 y los centrales en el 1, mientras que en la base de datos en la que únicamente se utilizan las variables más representativas sucede lo contrario.

Finalmente, para visualizar de una forma rápida y sencilla los resultados del análisis, estos se muestran a través de la matriz de confusión. En el caso de realizar el clúster con todas las variables el resultado es el siguiente:

Tabla 3 Matriz de confusión (todas las variables)

	Cluster	
posicion	Central	Lateral
Central	215	128
Lateral	189	77

En la diagonal de se pueden ver los jugadores bien clasificados. A partir de estos resultados que muestra la Tabla 4 se observa que se producen más errores a la hora de clasificar a los laterales. En concreto, se han clasificado 189 laterales como centrales y 128 centrales como laterales.

```
> # % correcto
> 100 * sum(diag(mc)) / sum(mc)
[1] 47.94745
> # % incorrecto
> (err <- 100*(1-sum(diag(mc))/sum(mc)))
[1] 52.05255
> err
[1] 52.05255
```

Figura 19 Porcentaje de acierto y error clúster todas las variables

A través de la Figura 19 se comprueba que en el caso del error global aproximadamente la mitad de los jugadores han sido erróneamente clasificados. Si se realiza el clúster únicamente con las variables consideradas representativas el resultado que se obtiene es el siguiente:

Tabla 4 Matriz de confusión (variables representativas)

posicion	cluster Central	Lateral
Central	291	52
Lateral	106	160

Una vez calculada la matriz de confusión se puede apreciar en la Tabla 5 que la tasa de acierto es mayor. La diagonal, que indica los aciertos, tiene una mayor cantidad de jugadores bien clasificados. A su vez se puede comprobar que se repite el mismo problema que en el caso anterior, puesto que se han clasificado 106 laterales como centrales y 52 centrales como laterales.

```
> # % correcto
> 100 * sum(diag(mc)) / sum(mc)
[1] 74.05583
> # % incorrecto
> (err <- 100*(1-sum(diag(mc))/sum(mc)))
[1] 25.94417
```

Figura 20 Porcentaje de acierto y error clúster variables representativas

En el caso del error global se comprueba a luz de lo expuesto por la Figura 20 que este ha pasado de ser más de la mitad a una cuarta parte.

Tras el análisis exploratorio se concluye que la eliminación de las variables que no eran representativas para ninguna de las posiciones analizadas tiene un efecto positivo a la hora de realizar la agrupación de los individuos en grupos, es decir, se ha eliminado ruido.

4.3.3. Validación de los resultados

Finalmente, se va a llevar a cabo la confirmación de los resultados obtenidos en el apartado anterior donde se concluía, a partir del análisis *cluster,* que la utilización de las variables más representativas (Tabla 3) para las posiciones de defensa central y lateral permitía una mejor clasificación de los jugadores en sus respectivas posiciones. Esta confirmación se efectuará a partir del método supervisado *Random Forest*. El uso de esta técnica es otro método que permite establecer si es posible identificar la posición

específica de un jugador mediante la observación o análisis de un número reducido de variables. Para realizar esta técnica de *Machine Learning* ha sido necesario clasificar manualmente a los defensas en centrales y laterales porque, como se dijo al inicio del trabajo, las técnicas de aprendizaje supervisado se distinguen por trabajar sobre datos previamente clasificados. Se ha escogido esta técnica para la validación del modelo puesto que el objetivo consiste en conocer las variables que aportan una mayor diferenciación entre posiciones, información que este algoritmo también aporta. Al igual que en el apartado anterior se analizarán ambas bases de datos.

Para llevar a cabo el proceso de validación se ha utilizado la misma base de datos con la que se trabajó previamente al realizar el análisis exploratorio de los datos, con la única diferencia de que en esta ocasión los jugadores se encuentran etiquetados, es decir, a priori se conoce si su posición es la de defensas lateral o central. En primer lugar, se debe comenzar por dividir la base de datos en un 70% para entrenar el modelo y un 30% para validarlo. Esta división se realiza a partir del paquete de *R* "*dplyr"* (Wickham 2019) a través de la función *sample_frac()*, que selecciona filas de forma aleatoria. Además, para que los individuos sean seleccionados de manera aleatoria, se utiliza la función *set.seed()* para que el proceso de selección de individuos no comience en el primer individuo.

En primer lugar, se analiza el gráfico de disminución media de Gini de la base de datos con todas las variables, de manera que se comprueben cuáles son las variables más importantes para clasificar a los jugadores según su posición específica, defensa central o lateral. La Figura 21 se obtiene a partir de la función *varImplot()* del paquete de datos "*randomForest"* (Liaw and Wiener 2002), y ofrece el resultado tras el ajuste realizado por un bosque aleatorio.

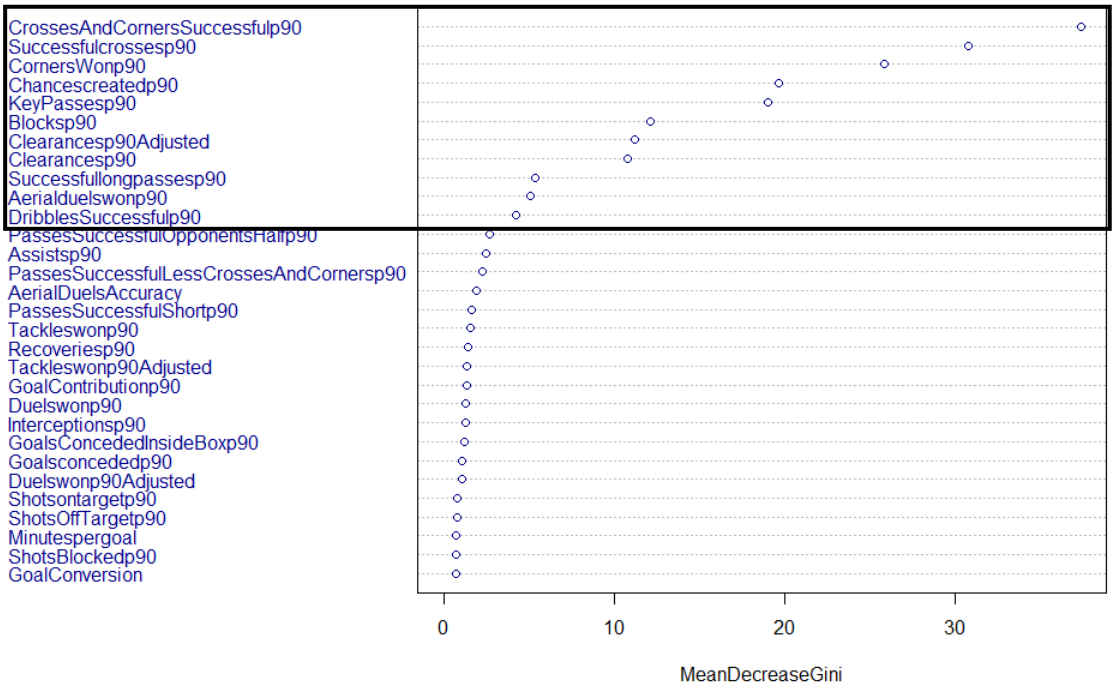

Figura 21 Gráfico disminución media de Gini

A la luz de los resultados presentados por la Figura 21 se puede comprobar que las variables de mayor importancia a la hora de clasificar a los jugadores según su posición, es decir, aquellas cuyo señalizador se encuentra más desplazado a la derecha y más alejado del 0, coinciden en su totalidad con las variables consideradas como representativas, estas son: córners y centro exitosos, córners ganados (cabeceados con éxito), córners y centros ejecutados con éxito, oportunidades creadas, pases clave, pases bloqueados/taponados, despejes de balón, pases largos exitosos, duelos aéreos ganados y regates exitosos. A su vez, se observa que todas las variables, a excepción de una, consideradas están asociadas al estilo de juego de los defensas centrales.

En segundo lugar, se procede al entrenamiento del modelo a partir de los datos de entrenamiento mediante la función *randomForest()* que implementa el algoritmo de bosque aleatorio. Una vez ejecutada esta fase, se utiliza el set de datos de prueba para validar el modelo. De este modo, el algoritmo será capaz de clasificar a los jugadores a partir de lo “aprendido” en base al set de datos de entrenamiento. Finalmente, se crea la matriz de confusión con los datos de validación para que se compruebe de un modo más visual la tasa de acierto del algoritmo (ver Figura en Anexo 9.8).

Tabla 5 Matriz de confusión (todas las variables)-Validación

	Position	
predicciones	Central	Lateral
Central	99	2
Lateral	6	76

Como se puede en la Tabla 6, al utilizar el método supervisado, es decir, al haber entrenado el modelo antes de realizar la predicción de las posiciones, los resultados mejoran de una manera evidente, aunque el error al clasificar a los laterales siga siendo ligeramente mayor que al clasificar a los centrales.

```
> # % correcto
> 100 * sum(diag(mc)) / sum(mc)
[1] 95.62842
> # % incorrecto
> (err <- 100*(1-sum(diag(mc))/sum(mc)))
[1] 4.371585
> err
[1] 4.371585
```

Figura 22 Porcentaje de acierto y error todas las variables-Validación

Si se observa el porcentaje de error global en la Figura 22, este ha pasado de ser aproximadamente del 50% (Figura 19) a no alcanzar el 5%.

Finalmente, se valida el set de datos creado únicamente con las variables consideradas representativas para comprobar si el uso de un número menor de variables perjudica o, por el contrario, beneficia a la predicción de las posiciones.

Tabla 6 Matriz de confusión (variables representativas)-Validación

	Position	
predicciones	Central	Lateral
Central	99	3
Lateral	6	75

Si se observa la matriz de confusión de la Tabla 7 se comprueba que, al igual que sucedía anteriormente, el uso de métodos de aprendizaje supervisado supone un mayor acierto a la hora de clasificar.

```
> 100 * sum(diag(mc)) / sum(mc)
[1] 95.08197
> # % incorrecto
> (err <- 100*(1-sum(diag(mc))/sum(mc)))
[1] 4.918033
> err
[1] 4.918033
```

Figura 23 Porcentaje de acierto y error variables representativas-Validación

El error global es casi del 5% según lo visto en la Figura 23. Este resultado es mucho menor que el obtenido con el método de aprendizaje no supervisado (Figura 20) y muy similar al obtenido utilizando todas las variables.

Otro método que podría haber sido utilizado para comprobar si es posible conocer la posición específica de los jugadores, a partir de un grupo de variables concretas y que también ofrece información sobre las variables con mayor poder discriminatorio, es el PLS *(Partial Least Squares-Discriminant Analysis*). No obstante, este método se descartó para la predicción de las posiciones de los jugadores sobre el terreno de juego, puesto que la existencia de fuertes relaciones no lineales en los datos afectaba a la calidad de la predicción, siendo esta un poco mayor del 60%.

En conclusión, el método utilizado para determinar las variables más representativas de cada posición es válido. En el caso del aprendizaje no supervisado el porcentaje de acierto para clasificar a los jugadores entre grupos es mayor si únicamente se utilizaban las variables más representativas. Por su parte, el resultado es prácticamente idéntico en el caso del aprendizaje supervisado para ambas bases de datos. Este método es una manera de facilitar la tarea tanto a ojeadores como analistas. Al reducir el número de variables, estos pueden centrarse en el estudio de las acciones del juego específicas para cada posición, optimizando así el tiempo de búsqueda y análisis de un jugador. Este proceso puede extrapolarse a otras posiciones específicas como mediocentros o delanteros.

5. Limitaciones del problema

A lo largo de este trabajo de investigación se han encontrado diferentes limitaciones externas a la voluntad del investigador que han influido en el desarrollo del estudio.

En primer lugar, el fútbol profesional forma parte de una industria privada y hermética cuya información más sensible no es de carácter público. A pesar de su nivel mediático gracias a la repercusión de la prensa convencional, es evidente que alrededor del fútbol existe una restricción en el acceso a cierta información confidencial como, por ejemplo, en lo referente a sueldos y contratos de los futbolistas. Este tipo de limitaciones informativas se manifiesta en las bases de datos sobre el rendimiento de los jugadores, partes médicos, lesiones, etc.

En el caso que nos ocupa, la base de datos empleada para el desarrollo de este trabajo es incompleta o, de algún modo, sustancialmente mejorable. Como se ha mencionado en el apartado de metodología, la distribución táctica de los jugadores sobre el terreno de juego suele clasificarse en siete posiciones diferenciadas: portero, defensa central, defensa lateral, mediocentro defensivo, mediocentro, extremo y delantero. Sin embargo, la base de datos utilizada clasifica a los jugadores en base a tres posiciones (defensa, mediocentro y delantero). Este hecho ha originado una limitación importante para analizar a los jugadores como, por ejemplo, en el caso del análisis de las posiciones.

En segundo lugar, el fútbol es un deporte complejo donde intervienen multitud de factores como los estados de ánimo de los futbolistas, las relaciones interpersonales entre los miembros de la plantilla y el cuerpo técnico, la toma de decisiones de los árbitros y las condiciones climatológicas, entre otros aspectos destacados. Por tanto, los resultados son limitados puesto que los números nos ayudan a explicar una parte importante de los hechos descriptivos, pero no alcanzan a profundizar en aspectos puramente cualitativos que pueden influir en el desempeño de un futbolista sobre el campo.

Esta limitación anterior, propia de los métodos cuantitativos, se podría haber complementado con entrevistas en profundidad a expertos de la industria del fútbol como entrenadores, analistas, ojeadores y directores deportivos, entre otros. Sus opiniones y reflexiones podrían haber resultado de gran interés para profundizar sobre los análisis de los resultados presentados en este trabajo de investigación.

6. Futuras investigaciones

A la vista de lo analizado y de las limitaciones encontradas en la elaboración de este trabajo, en este apartado se recomienda seguir el estudio respecto a dos líneas de investigación.

Por un lado, resultaría interesante complementar el análisis desarrollado en este trabajo con los testimonios de analistas de datos y encargados de los departamentos de Big Data de clubes de fútbol profesional, pues sus experiencias y conocimientos podrían aportar una valiosa información sobre el estado actual del dato en la industria del fútbol. Asimismo, y como se ha hecho alusión anteriormente, sería recomendable profundizar en el análisis del juego con entrenadores, ojeadores y otros miembros de las direcciones deportivas encargados del proceso de búsqueda y captación de jugadores. Esto permitiría añadir matices cualitativos que servirían para enriquecer el análisis de datos elaborado en este trabajo.

Por otro lado, se considera relevante para futuras investigaciones sobre *Machine Learning* y la industria del fútbol profundizar sobre el valor de mercado de los futbolistas en relación a su rendimiento actual y estado de forma. El diseño de una aplicación de características similares a la propuesta en este trabajo, pero añadiendo información sobre los sueldos y la duración de los contratos de los jugadores, podría permitir llevar a cabo una estimación y predicción del valor de mercado de los mismos. El resultado de este proceso serviría para medir y evaluar la sobrevaloración e infravaloración de los jugadores en comparación con los valores del mercado real, por ejemplo, a partir de la información ofrecida por la web especializada *Transfermarkt* (Axel Springer SE 2019).

7. Conclusiones

Este apartado proporciona a modo de resumen las principales ideas y los hallazgos más relevantes extraídos a partir del proceso de análisis de datos llevado a cabo en la elaboración de este trabajo de investigación.

El filósofo alemán Martin Heidegger, pesimista al valorar las consecuencias de un dominio técnico del mundo, afirmó en una de sus conferencias más populares que los poderes y las consecuencias derivadas de los avances de la ciencia, y de las nuevas tecnologías, “hace ya tiempo que han desbordado la voluntad y capacidad de decisión humana porque no han sido hechas por el hombre” (Heidegger 2002). El autor se refiere a que los efectos de muchos de los grandes avances de la humanidad creados por el ser humano no fueron concebidos en su origen como en lo que posteriormente se transformaron. Heidegger nos invita a reflexionar sobre los peligros del mundo técnico para que se contemple una visión crítica ante el imperio del pensamiento puramente científico.

Por tanto, hemos de reconocer las limitaciones de las técnicas de *Machine Learning* propias de los procesos computacionales. Las herramientas fundamentadas en el análisis de datos desarrolladas en este trabajo se han concebido como un apoyo complementario a los profesionales del fútbol encargados de tomar decisiones en el proceso de fichajes de jugadores, pero también a los responsables de los cuerpos técnicos que deben decidir entre alinear a un jugador u otro en cada partido. Así, se han empleado diferentes técnicas de análisis de datos cuya finalidad no es la de sustituir la toma de decisión de las personas por algoritmos, sino la de complementar y reforzar las decisiones facilitando el proceso de búsqueda de jugadores, además de reducir tiempo y costes en los organigramas de los clubes de fútbol profesionales.

De esta manera, a partir de una aplicación propia desarrollada en *Shiny*, un paquete de R destinado a la construcción sencilla e intuitiva de interfaces interactivas con datos, podemos encontrar quiénes son los jugadores que guardan una mayor similitud entre sí usando el análisis de componentes principales (PCA) en base a variables de rendimiento que describen cómo se comportan los jugadores sobre el campo. Esta herramienta permite encontrar opciones de reemplazo a un futbolista determinado en una plantilla de manera rápida y dinámica, pues pone de manifiesto quiénes son, a grandes rasgos, los jugadores más parecidos a otro concreto.

No obstante, este proceso ha requerido de un análisis de mayor profundidad que identifique con precisión los rasgos diferenciales entre los jugadores similares, pues estos no tienen un estilo de juego completamente homogéneo. Para ello se ha recurrido a diferentes técnicas multivariantes de análisis de datos que nos permiten comparar futbolistas en base a las acciones específicas que llevan a cabo sobre el terreno de juego. Este análisis comparativo se presenta como un recurso de apoyo a la aplicación diseñada, pues permite ahondar en aspectos más específicos del juego como las recuperaciones, los duelos aéreos, los pases exitosos y las ocasiones creadas, entre otras muchas variables. Por tanto, a partir de visualizaciones atractivas como los gráficos de radar podemos encontrar aquellos jugadores que se adecúan con una mayor exactitud a características concretas demandadas por un directivo, analista del juego o entrenador.

Finalmente, se ha llevado a cabo un análisis más específico de las posiciones que ocupan los jugadores sobre el terreno de juego para conocer cuáles son las variables más representativas de cada posición. Este método tiene la finalidad de facilitar el trabajo a ojeadores y analistas reduciendo el número de variables a tener en cuenta a la hora de buscar y estudiar el comportamiento de un perfil de jugador requerido por el club.

La metodología utilizada en este trabajo es puramente cuantitativa. Aunque todo lo expuesto supone un método novedoso durante el proceso de captación de talento complementario al ojo humano, no se debe perder de vista la complejidad de un deporte colectivo y mediático como el fútbol. En primer lugar, a diferencia de otros deportes donde las reglas de juego establecen una limitación de las posesiones (baloncesto) o las acciones se desarrollan en periodos cortos de tiempo (tenis, béisbol…), el fútbol es un juego de habilidades abiertas y acciones intermitentes donde todos los jugadores están en movimiento a la vez. Esto último dificulta sobremanera el proceso de cuantificar todas las acciones, así como el análisis táctico y técnico del juego, ya sea en directo o a través de vídeo. En segundo lugar, su carácter mediático es garantía de audiencias millonarias, pues la incertidumbre del resultado final lo convierte en un evento atractivo para los medios de comunicación, particularmente, para las televisiones que adquieren los derechos de explotación de las imágenes por ingentes cantidades de dinero. Esa idea de incertidumbre del resultado final del partido contribuye a convertirlo en un deporte difícilmente medible y cuantificable. En esta línea, en el fútbol intervienen una serie de factores externos de carácter cualitativo tales como aspectos psicológicos, emocionales o sentimentales que complejizan cualquier toma de decisión en la captación y contratación de un jugador.

En el deporte intervienen multitud de aspectos y muchos de ellos no son medibles ni cuantificables, siendo precisamente esto la esencia no sólo del fútbol, sino de todos los deportes. No parece posible fichar jugadores basándose en datos únicamente. El fútbol es un deporte compuesto por personas y, por ello, se encuentra supeditado a las limitaciones propias de los seres humanos: estados de ánimo, estados físicos, sentido de pertenencia al grupo, aspectos extradeportivos y familiares, entre otras. Si únicamente se prestara atención a los datos, se acabaría perdiendo la capacidad de superación. Jugadores como Paco Alcácer o Samuel Eto'o tal vez nunca se hubieran convertido en estrellas (Szwarc 2018). Los futbolistas serían clasificados únicamente en dos grupos: aptos y no aptos. Además, son muchos los casos mediáticos de fichajes frustrados que, tras triunfar previamente en un equipo, no consiguieron destacar cuando fueron traspasados a un nuevo club. Esto puede deberse, en parte, a la capacidad del ser humano para aprender nuevos idiomas, adaptarse a nuevos entornos, culturas, religiones y otros ámbitos sociales.

La certeza de que la realidad humana debe estar por encima de cualquier dato o tecnología es evidente. Sin embargo, no tendría sentido despreciar la ayuda que ofrecen las nuevas tecnologías y, en especial el análisis de datos, a los clubes de fútbol profesional. Como afirmó el astrofísico Carl Sagan: *"La ciencia no es perfecta, con frecuencia se utiliza mal, no es más que una herramienta, pero es la mejor herramienta que tenemos, se corrige a sí misma, está siempre evolucionando y se puede aplicar a todo"* (Sagan 1970).

Asimismo, el filósofo francés, Gilbert Simondon, da las claves de la relación entre el hombre y la tecnología, lo que bien se podría aplicar al análisis de datos en el mundo del fútbol. Son esos momentos cuando los datos no pueden dar respuesta a todo, pero también cuando el ser humano necesita de un conocimiento global que sólo se puede otorgar por la tecnología, que el hombre y la máquina se unen en una simbiosis perfecta:

"Una herramienta puede ser bella en la acción cuando se adapta tan bien al cuerpo que parece prolongar de manera natural y amplificar en alguna forma sus caracteres estructurales; un puñal sólo es bello en la mano que lo sostiene; por lo mismo, una herramienta, una máquina o un conjunto técnico son bellos cuando se insertan en un mundo humano y lo recubren al expresarlo" (Simondon 2008).

8. Bibliografía

Aguado, M. *GENTSIDE.* 2019. https://www.esgentside.com/futbol/que-ingresos-genera-la-liga-espanola-de-futbol-en-un-ano_art18542.html (último acceso: 13 de 07 de 19).

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng J., Chang, W. and Iannone, R. *Rmarkdown: Dynamic Documents for R. R package version 1.14.* 2019. https://rmarkdown.rstudio.com.

Aspen Technology Inc. 2018. http://ir.aspentech.com/static-files/a6b44c77-4f5d-4462-a1b0-3b96c5b417e2

AspenTech. *Aspen Plus. Getting Started Building and Running a Process Model.* Burlintong: ASPEN Technology, 2013.

Axel Springer SE. *Transfermarkt.* 2019. https://www.transfermarkt.es/ (último acceso: 25 de 7 de 2019).

Billings, A. C. *La comunicación en el deporte.* Barcelona: UOC, 2010.

Boyle, R. and Haynes, R. *Football in the new media age.* London: Routledge, 2004.

Breiman, L. «Random forests. Machine learning.» *Kluwer Academic Publishers*, 2001: 5-32.

Chang, W., Cheng, J., Allaire, J., Xie, Y. and McPherson, J. *Shiny from R Studio.* 2017. https://shiny.rstudio.com/ (último acceso: 15 de 7 de 2019).

—. *Shiny: Web Application. Framework for R. R package version 1.3.2.* 2019. https://CRAN.R-project.org/package=shiny (último acceso: 14 de 7 de 2019).

Deloitte. *¿Qué es la Industria 4.0? Davos y la Industria.* 2019. https://www2.deloitte.com/es/es/pages/manufacturing/articles/que-es-la-industria-4.0.html (último acceso: 12 de 07 de 2019).

Dunn, K. *Process Improvement Using Data.* 2019. https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/some-properties-of-pca-models (último acceso: 1 de 9 de 2019).

EASPORTS. *FIFA 18 – Valoraciones de Jugadores. Los 100 Mejores.* 2018. https://www.easports.com/es/fifa/fifa-18-player-ratings-top-100#60-41 (último acceso: 20 de 7 de 2019).

FIFA. *2018 FIFA World Cup Russia. Global broadcast and audience summary.* 2018. https://resources.fifa.com/image/upload/2018-fifa-world-cup-russia-global-broadcast-and-audience-executive summary.pdf?cloudid=njqsntrvdvqv8ho1dag5 (último acceso: 13 de 7 de 2019).

Fox, J. *KMeans. Agrupación de K-medias usando múltiples semillas aleatorias.* s.f. https://www.rdocumentation.org/packages/RcmdrMisc/versions/2.5-1/topics/KMeans 2017 (último acceso: 20 de 7 de 2019).

Gil, C. *Árboles de decisión y métodos de ensemble.* 2018. (último acceso: 15 de 7 de 2019).

Hamill, R. *Player Similarity Scores.* 2019. https://github.com/RayHamill/Football/tree/master/similarity_scores (último acceso: 20 de 6 de 2019).

Heidegger, M. *Serenidad.* Barcelona: Serbal, 2002.

Ihaka, R. and Gentleman, R. «R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics.» *Revista de Estadística Computacional y Gráfica* 5 (1996): 299–314.

InStat. *InStat.* 2019. http://instatsport.com/ (último acceso: 13 de 7 de 2019).

John, C., Cleveland, W., Kleiner, B., & Tukey, P. *Graphical methods for data analysis.* Pacific Grove, California: CRC Press, 1983.

Kuper, S. *How FC Barcelona are preparing for the future of football. Financial Times.* 2019. https://www.ft.com/content/908752aa-3a1b-11e9-b72b-2c7f526ca5d0 (último acceso: 13 de 7 de 19).

Lewis, M. *Moneyball: The art of winning an unfair game.* New York: W.W.Norton, 2003.

Liaw, A. and Wiener, M. «Classification and Regression by randomForest.» *R News* 2(3) (2002): 18-22.

Llopis Goig, R. «Identificación con clubes y cultura futbolística en España. Una aproximación sociológica.» *RICYDE.Revista Internacional de Ciencias Del Deporte* 9, nº 33 (2013): 236-251.

Lombardi, F. *¿Cuáles son las diferentes posiciones en el fútbol?* 2018. https://tuzonadefutbol.com/posiciones-en-el-futbol/ (último acceso: 2019 de 7 de 16).

Long, FH. «Proteomic and Metabolomic Approaches to Biomarker Discovery. Chapter 19 --Multivariate Analysis for Metabolomics and Proteomics Data.» 2013: 299-311.

MacQueen, J.B. «Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability.» 1 (1967): 281-297.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A. *Big data: The next frontier for innovation, competition, and productivity.* Washington D.C.: McKinsey Global Institute, 2011.

MarkeInhouse. *Mercado competitivo a través de 5 ejemplos de nicho de mercado.* 2019. https://www.marketinhouse.es/5-ejemplos-de-nicho-de-mercado/ (último acceso: 13 de 7 de 2019).

Marr, B. *Big data en la práctica: cómo 45 empresas exitosas han utilizado análisis de big data para ofrecer resultados extraordinarios.* Zaragoza: Teell, 2017.

MathWorks. *Aprendizaje Supervisado. Técnica de Machine Learning para crear modelos predictivos a partir de datos de entrada y respuesta conocidos.* 2019. https://es.mathworks.com/discovery/supervised-learning.html (último acceso: 15 de 7 de 2019).

Mayer-Schönberger, V. and Cukier, K. *Big data: la revolución de los datos masivos.* Madrid: Turner, 2013.

—. *Big data: la revolución de los datos masivos.* Madrid: Turner, 2013.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-0.1.* 2019. https://CRAN.R-project.org/package=e1071 (último acceso: 20 de 7 de 2019).

Opta. *Opta's event definitions.* 2018. https://www.optasports.com/news/opta-s-event-definitions/ (último acceso: 6 de 6 de 2019).

Opta Sports. *Opta.* 2019. https://www.optasports.com/ (último acceso: 13 de 7 de 2019).

Peck, G. *Tableau 8: the official guide.* New York: McGraw-Hill Education, 2014.

Pérez, D. *Objetivo Analista.* 2018. https://objetivoanalista.com/scouting-scout-analista-ojeador/ (último acceso: 17 de 7 de 2019).

Perrier, P., Meyer, F., and Granjon, D. *ShinyWidgets: Custom Inputs Widgets for Shiny. R package version 0.4.8.* 2019. https://CRAN.R-project.org/package=shinyWidgets (último acceso: 17 de 7 de 2019).

Peterson, Brian G. and Carl, P. *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis. R package version 1.5.3.* 2019. https://CRAN.R-project.org/package=PerformanceAnalytics (último acceso: 17 de 7 de 2019).

Rodrigo-Amat, J. *Árboles de predicción: bagging, random forest, boosting y C5.0. Bagging.* 2017. https://rpubs.com/Joaquin_AR/255596 (último acceso: 15 de 7 de 2019).

—. *Clustering y heatmaps: aprendizaje no supervisado.* 2017. https://rpubs.com/Joaquin_AR/310338 (último acceso: 15 de 7 de 2019).

—. *Clustering y heatmaps: aprendizaje no supervisado.* 2017. https://rpubs.com/Joaquin_AR/310338 (último acceso: 15 de 7 de 2019).

Rodrigo-Amat, R. *Clustering y heatmaps: aprendizaje no supervisado. Medidas de distancia. Escala de las variables.* 2017. https://rpubs.com/Joaquin_AR/310338 (último acceso: 17 de 7 de 2019).

RStudio. *RStudio.* 2018. https://www.rstudio.com/products/rstudio/ (último acceso: 14 de 7 de 2019).

Sagan, C. *Cosmos.* National Geographic. 1970.

Santana, E. *Machine Learning con R.* 2014. http://apuntes-r.blogspot.com/2014/11/ejemplo-de-random-forest.html (último acceso: 15 de 7 de 2019).

SAS. *Know, SAS. The Power to.* 2019. https://www.sas.com/es_es/insights/analytics/machine-learning.html (último acceso: 13 de 7 de 2019).

Silver, N. *La señal y el ruido: cómo navegar por la maraña de datos que nos inunda, localizar los que son relevantes y utilizarlos para elaborar predicciones infalibles.* Barcelona: Península, 2014.

Simondon, G. *El modo de existencia de los objetos técnicos.* Buenos Aires: Prometeo, 2008.

Stats LLC. *STATS.* 2019. https://www.stats.com/ (último acceso: 13 de 7 de 2019).

Szwarc, D. *90MiN.* 2018. https://www.90min.com/es/posts/6200112-cinco-jugadores-que-pasaron-de-suplentes-a-estrellas (último acceso: 20 de 7 de 2019).

Team, R Core. «R: A language and environment for statistical computing.» Vienna, 2008.

Team, RStudio. *RStudio: Integrated Development for R. RStudio.* Inc., Boston, MA. 2016. http://www.rstudio.com/. (último acceso: 14 de 7 de 2019).

Trajkovic, J. *Use radar charts to compare dimensions over several metrics.* 2015. https://www.tableau.com/about/blog/2015/7/use-radar-charts-compare-dimensions-over-several-metrics-41592 (último acceso: 16 de 7 de 2019).

Van den Kerkhof, P.,Vanlaer, J., Gins, G. and Van Impe, Jan F.M. «Contribution plots for Statistical Process Control: analysis of the smearing-out effect.» Zürich, Switzerland: 2013 European Control Conference (ECC), 2013.

Wickham, H. *Tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1.* 2017. https://CRAN.R-project.org/package=tidyverse (último acceso: 17 de 7 de 2019).

Wickham, H., and Bryan, J. *readxl: Read Excel Files. R package version 1.3.1.* 2019. https://CRAN.R-project.org/package=readxl (último acceso: 16 de 7 de 2019).

Wickham, H., François, R., Henry, L. and Müller, K. *dplyr: A Grammar of Data Manipulation. R package version 0.8.0.1.* 2019. https://CRAN.R-project.org/package=dplyr (último acceso: 20 de 7 de 2019).

Wold, S., Esbensen, K., and Geladi, P. *Principal component analysis. Chemometrics and intelligent laboratory systems.* Vol. 2, cap. (1-3), 37-52. 1987.

Wold, S., Esbensen, K., and Geladi, P. *Principal component analysis. Chemometrics and intelligent laboratory systems.* Vol. 2, cap. (1-3), 37-52. 1987.

Wyscout. *Wyscout.* 2019. https://wyscout.com/ (último acceso: 13 de 7 de 2019).

Xie, Y., Allaire, J. and Grolemund, G. *R Markdown: The Definitive Guide.* Boca Raton, Florida: CRC Press, 2018.

Zelada, C. *Evaluación modelos de clasificación. Matriz de confusión.* 2017. https://rpubs.com/chzelada/275494 (último acceso: 15 de 7 de 2019).

9. Anexo

9.1. Base de datos

Tabla 7 Base de datos (Opta 2018)

Competición	English Premier League, French Ligue1, German Bundesliga, Italian Serie A, Spanish La Liga
Equipo	Nombre del equipo en el que juega
Name	Nombre del jugador
Position	Posición en la que juega: centrocampista, delantero y defensa
Arial duels accuracy	Porcentaje de acierto en los duelos aereos, es decir, una vez ganados hayan llegado a su destino, ya sea el pase a un jugador o el tiro a portería
Aerial duels lost p90	Número de duelos aéreos perdidos
Aerial duels won p90	Número de duelos aéreos ganados
Assists p90	Número de asistencias, es decir, pases que han terminado en gol
Backward p90	Número de pases hacia atrás por temporada
Blocks p90	Número de pases bloqueados/taponados
Clearances p90	Número de veces que el jugador ha despejado el balón, mandándolo fuera del campo o dentro del mismo, pero aun así enviándolo lejos de la portería propia

Chances created p90	Número de pases o pases finales antes de un disparo que lleva al receptor de la pelota a intentar un gol
Corners won p90	Número de corners ganados, es decir, que han sido cabeceados, no siendo necesarios que hayan terminado en gol
Crosses and corners successful p90	Número de corners y centros exitosos, es decir, centros y corners que han sido rematados por un jugador del propio equipo, independientemente de que este haya terminado en gol o no
Goals p90	Número de goles marcados
Crosses and corners unsuccessful p90	Número de corners y centros fallidos, es decir, centros y corners que no han sido rematados por un jugador del propio equipo
Crosses unsuccessful p90	Número de centros fallidos, es decir, centros que no han sido rematados por un jugador del propio equipo
Dribbles successful p90	Número de regates completados con éxito
Duels lost p90	Número de disputas perdidas entre dos jugadores de equipos rivales donde no existe un poseedor del balón definido
Duels won p90	Número de disputas ganadas entre dos jugadores de equipos rivales donde no existe un poseedor del balón definido
Forward p90	Número de pases hacia delante
Goals conceded p90	Número de goles que ha recibido el equipo mientras este estaba en el campo

Goals conceded inside box p90	Número de goles que ha recibido el equipo desde fuera del área mientras este estaba en el campo
Goals outside the box p90	Número de goles marcados por el jugador desde fuera del área
Headed goals p90	Número de goles marcados de cabeza
Interceptions p90	Número de intercepciones, es decir, veces que el jugador ha evitado que un pase llegue a su destino
Key passes p90	Número de pases propiciadores de remate
Offsides p90	Número de fueras de juego
Passes successful less crosses and corners p90	Número de pases exitosos, sin contar corners y centros, es decir, pases que han tenido como destinatario a un jugador del propio equipo
Passes successful opponents half p90	Número de pases exitosos en el campo contrario
Passes successful short p90	Número de pases cortos completados con éxito, es decir, que han llegado al jugador de destino sin ser interceptados
Passes unsuccessful opponents half p90	Número de pases fallidos en el campo contrario
Passes unsuccessful own half p90	Número de pases fallidos en campo propio
Recoveries p90	Número de recuperaciones del balón

Successful crosses p90	Número de centros exitosos, es decir, centros y que han sido rematados por un jugador del propio equipo, independientemente de que este haya terminado en gol o no
Successful long passes p90	Número de pases largos completados con éxito, es decir, que han llegado al jugador de destino sin ser interceptados
Tackles won p90	Número de entrada a ras del suelo que se realizan con una pierna por delante de la otra para robar el balón al oponente limpiamente.
Unsuccessful long passes p90	Número de pases largos fallidos, es decir, pases que no han llegado a su destinatario
NPG p90	Número de goles marcados (No penaltis)
Goal contribution p90	Número total de asistencias y goles
Shots blocked p90	Número de disparos a portería bloqueados/taponados
Shots off target p90	Intento de gol, pero el tiro fue muy desviado y no fue entre los tres palos, no hubo intervención del portero ni de otro jugador
Shots on target p90	Un intento de gol que requirió intervención para evitar que fuera gol o lo fue / tiro que entraría sin ser desviado
Duels won p90 adjusted	Número de disputas ganadas entre dos jugadores de equipos rivales donde no existe un poseedor del balón definido. Ajustado a la posesión del equipo
Goal conversion	Porcentaje de número totales de goles respecto a los remates totales realizados

Tackles won p90 adjusted	Número de entrada a ras del suelo que se realizan con una pierna por delante de la otra para robar el balón al oponente limpiamente. Ajustado a la posesión del equipo
Clearances p90 ajusted	Número de veces que el jugador ha despejado el balón, mandándolo fuera del campo o dentro del mismo, pero aun así enviándolo lejos de la portería propia. Ajustado a la posesión del equipo
Minutes per goal	Número de goles por minuto, es decir, cada cuanto tiempo marca un gol el jugador de media
Minutes played	Minutos totales jugados

9.2. Contribuciones de los residuos al espacio de las X

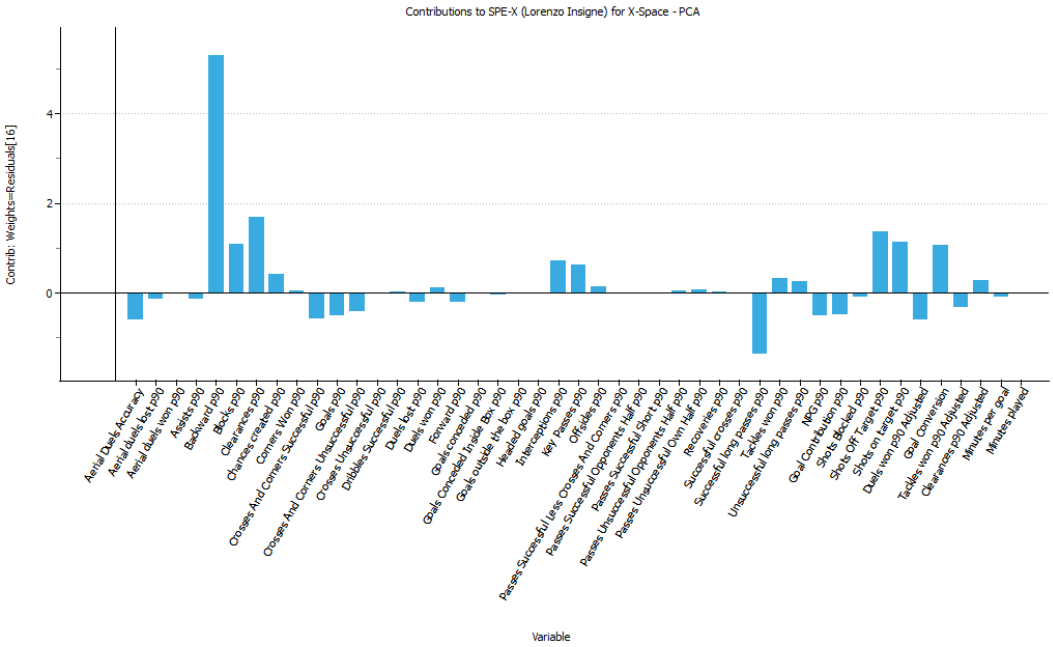

Figura 24 Lorenzo Insigne

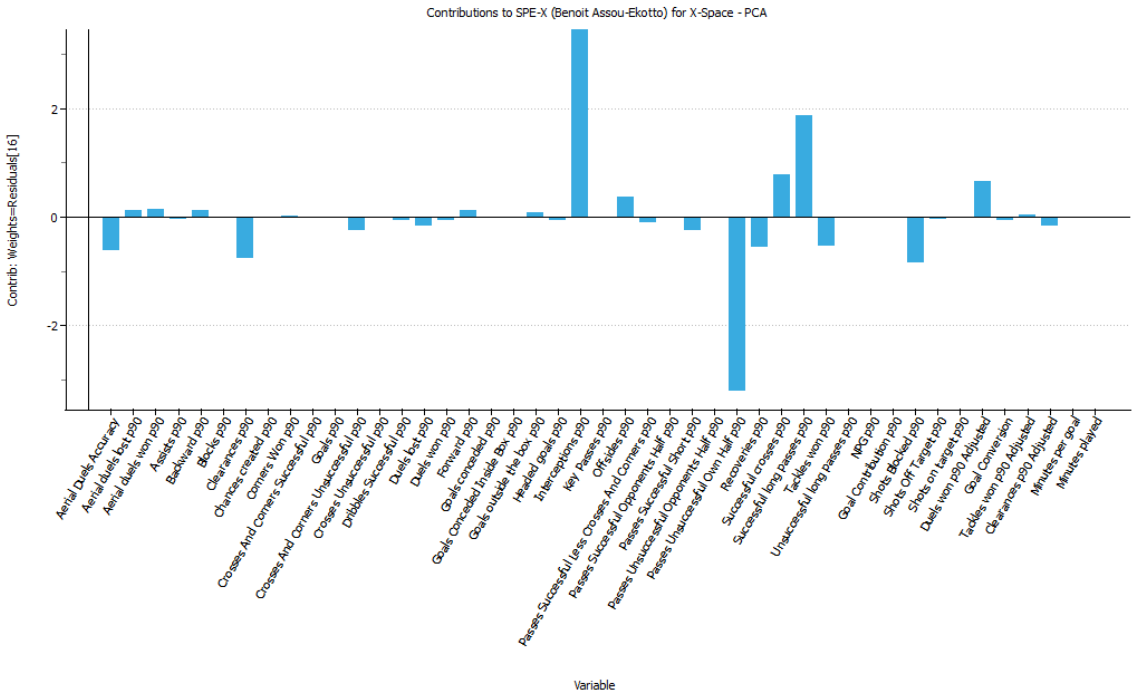

Figura 25 Benoit Assou

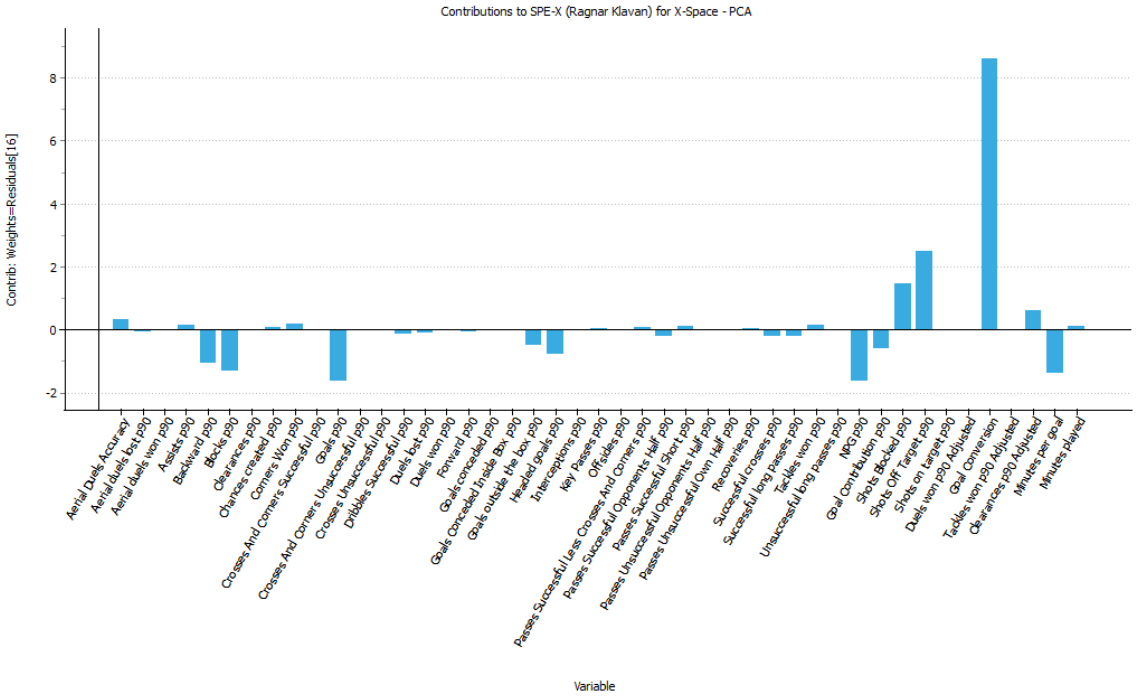

Figura 26 Ragnar Klavan

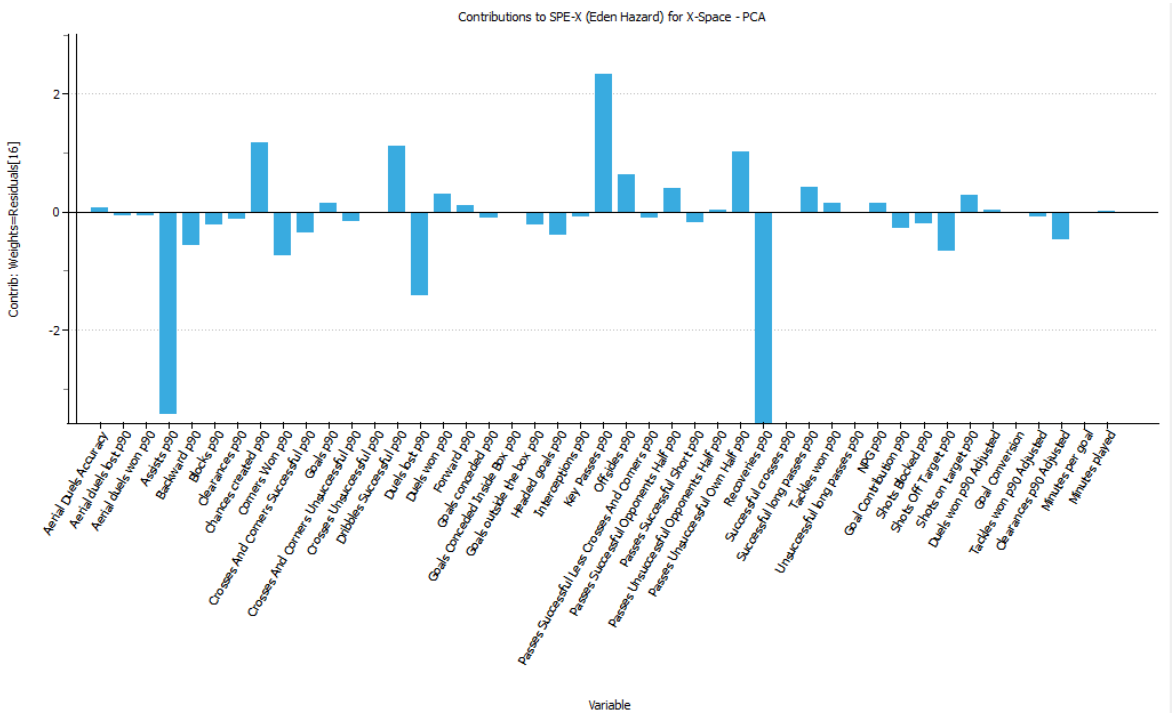

Figura 27 Eden Hazard

9.3. Contribución de las variables a los componentes principales

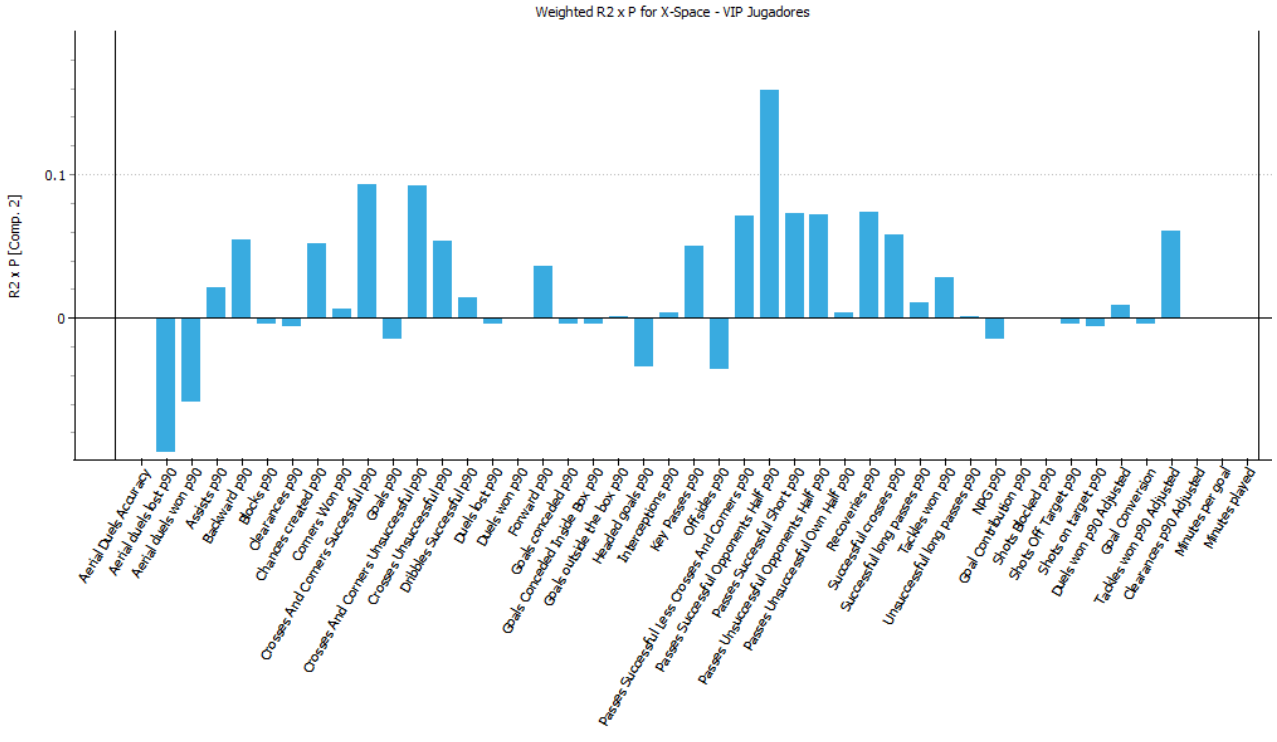

Figura 28 Pesos de la 2ª CP

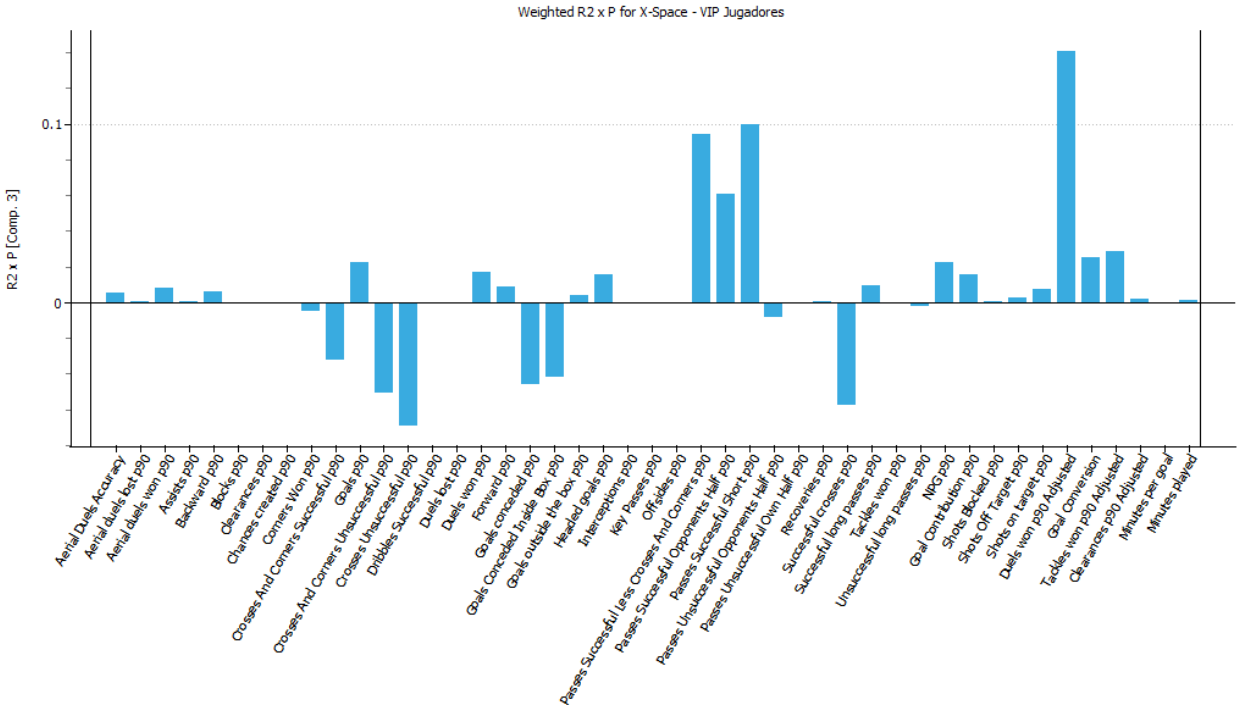

Figura 29 Pesos de la 3ª CP

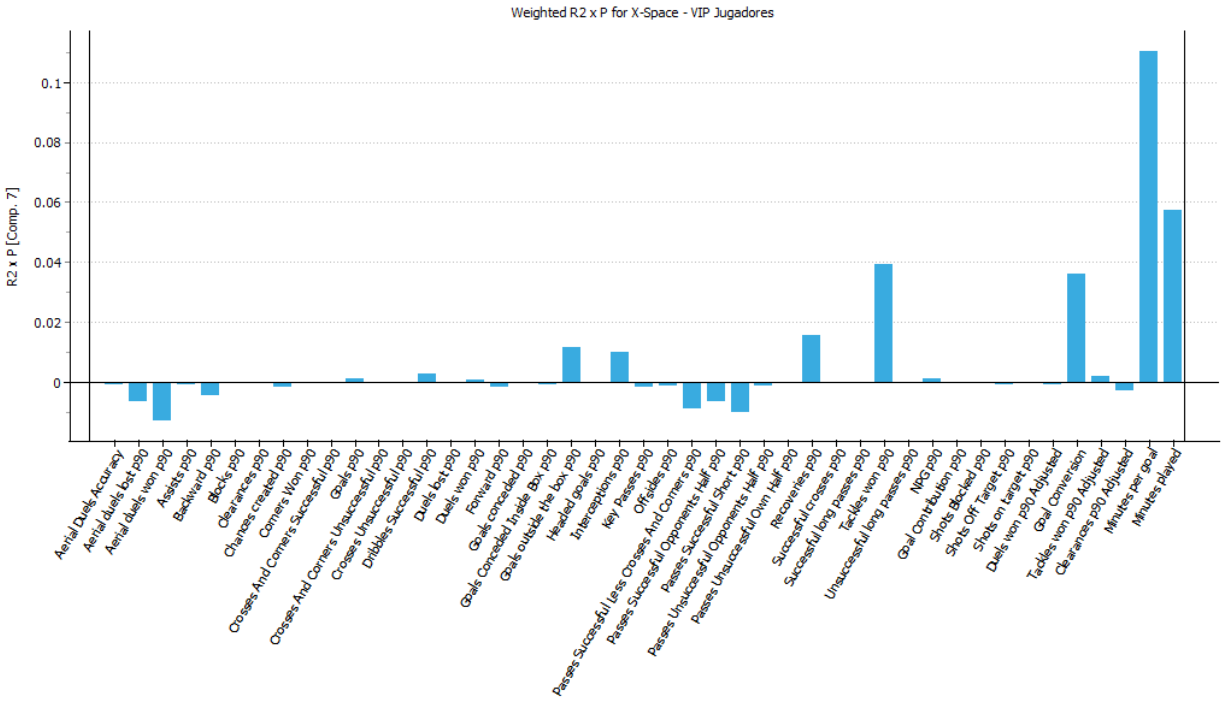

Figura 30 Pesos de la 7ª CP

9.4. Defensa: Lateral vs Central

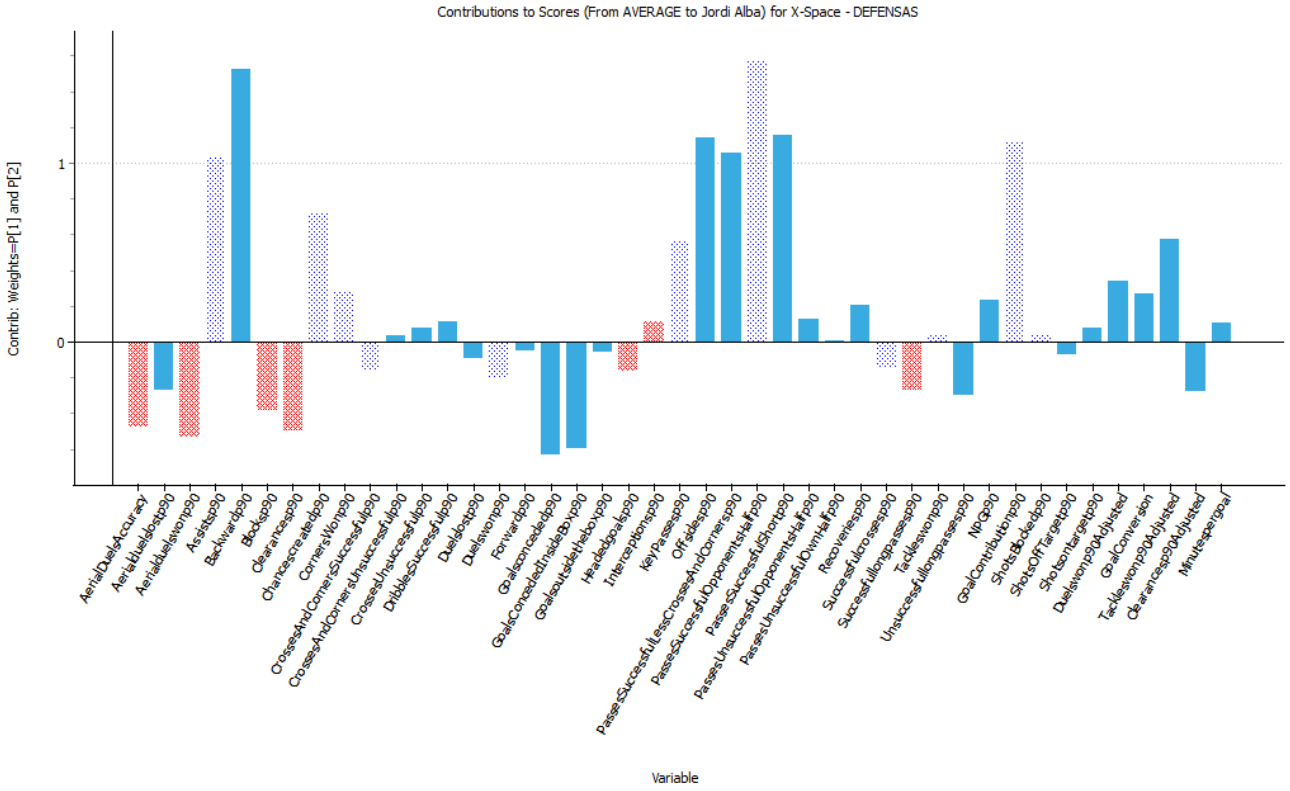

Figura 31 Contribuciones Jordi Alba vs defensa promedio

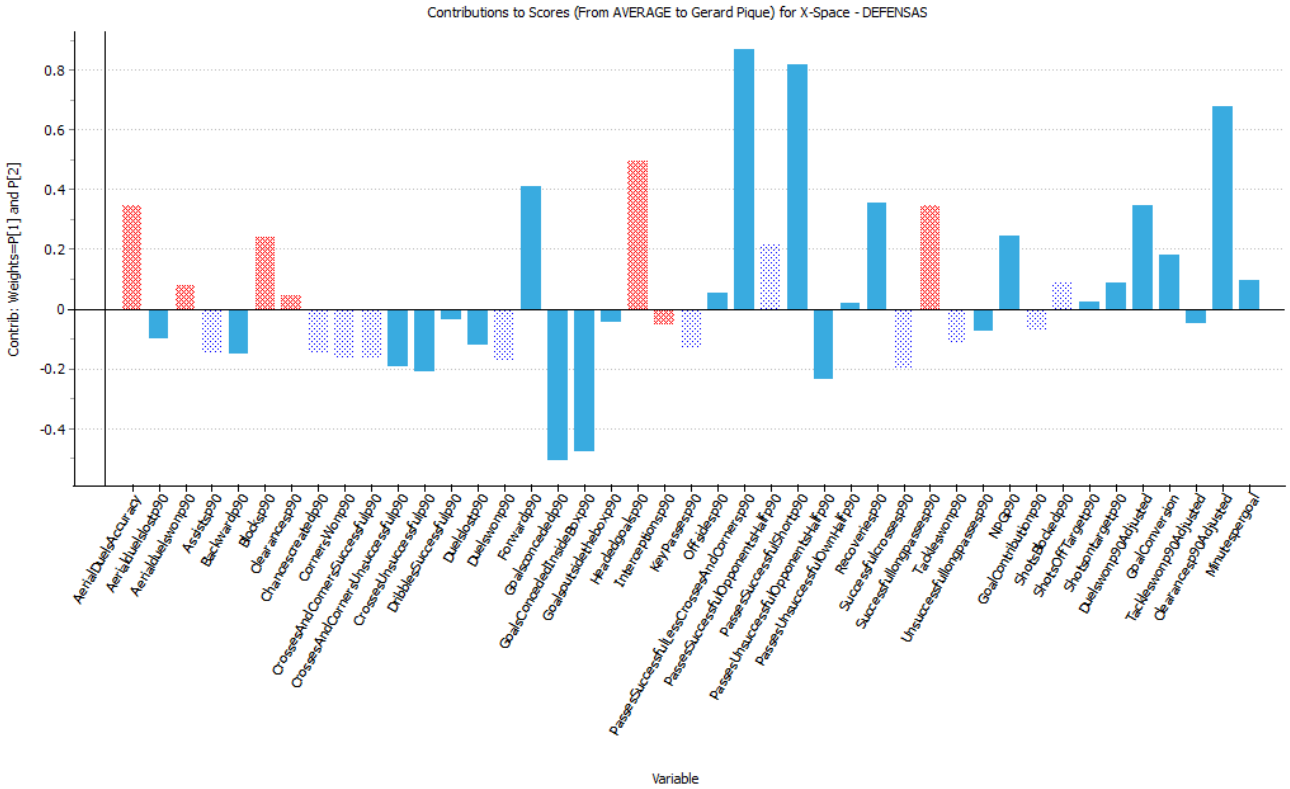

Figura 32 Contribuciones Gerard Piqué vs defensa promedio

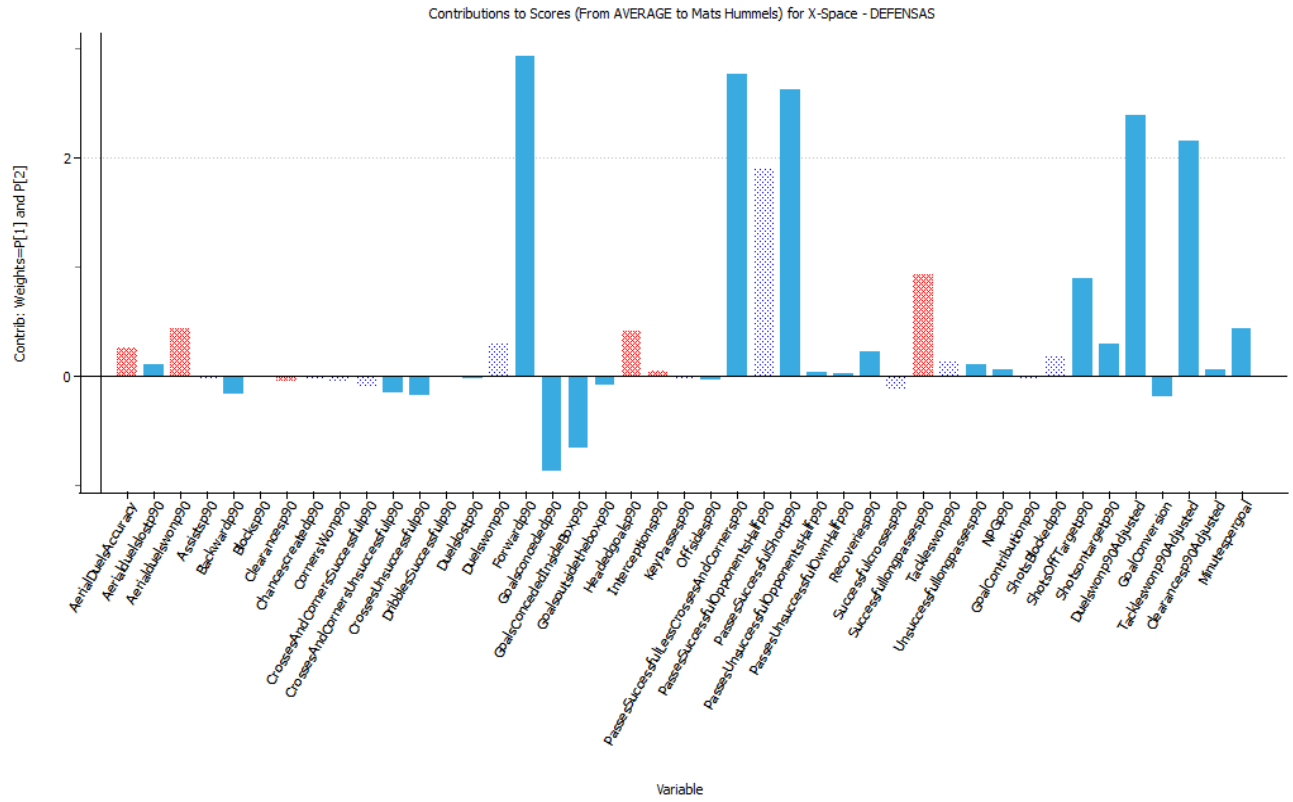

Figura 33 Contribuciones Mats Hummels vs defensa promedio

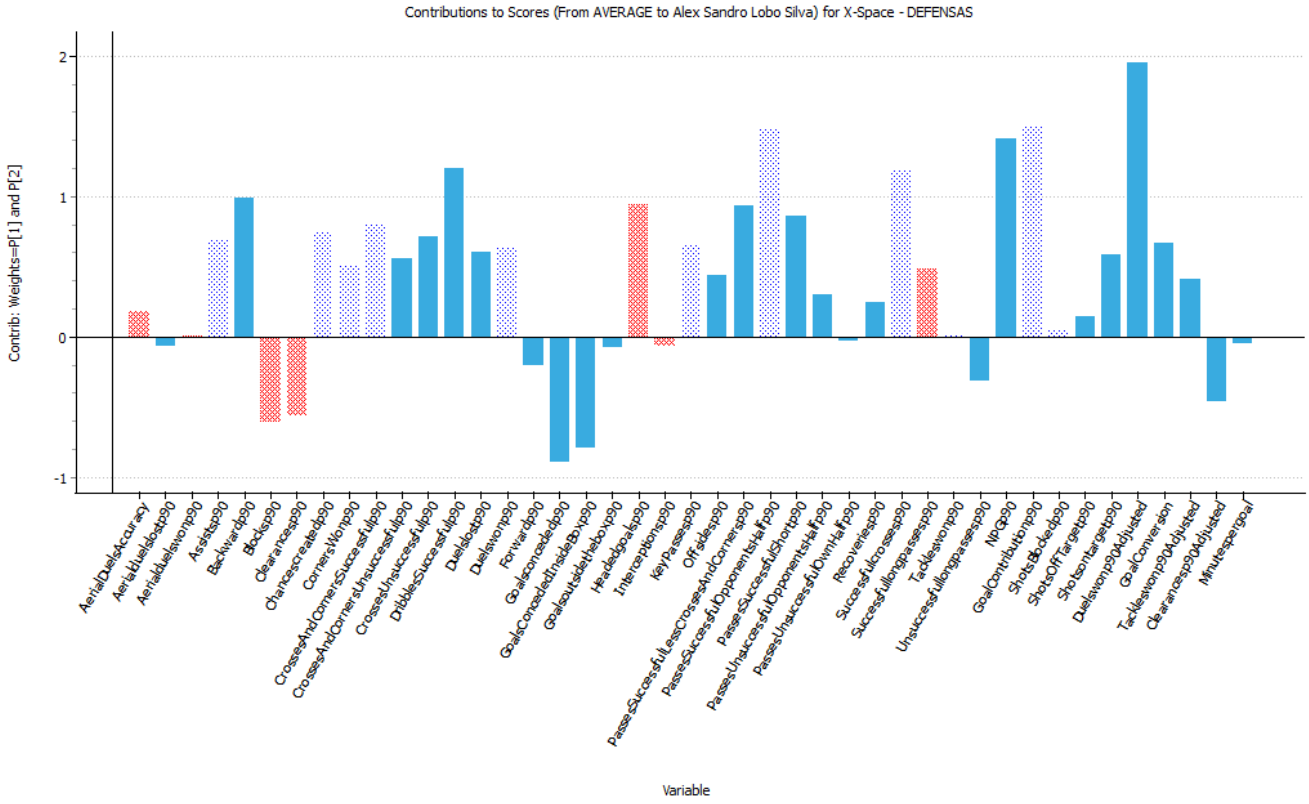

Figura 34 Contribuciones Alex Sandro vs defensa promedio

9.5. Cálculo K-óptimo

```
library("e1071")

set.seed(123)
Errores <-NULL
K_Max   <-10

for (i in 1:K_Max)
{
  Errores[i] <- sum(kmeans(newplayers[-1], centers=i)$withinss)
}

plot(1:K_Max, Errores, type="b",
     xlab="Cantidad de Cluster",
     ylab="Suma de error")
```

Figura 35 Código R cálculo y visualización k-óptimo

9.6. Cluster Means

```
Cluster means:
  AerialDuelsAccuracy Aerialduelswonp90 Assistsp90 Blocksp90 Clearancesp90 Chancescreatedp90 CornersWonp90
1            56.30488          2.069945 0.05609831 0.5325157      3.985081         0.5211377     0.3135127
2            55.95545          1.868566 0.04679315 0.4798140      3.767924         0.4806559     0.3101280
  CrossesAndCornersSuccessfulp90 DribblesSuccessfulp90 Duelswonp90 GoalsconcededInsideBoxp90
1                      0.3365154             0.5054483    5.158181         1.335303                  1.156335
2                      0.3289555             0.5789067    5.082656         1.355171                  1.167938
  Goalsoutsidetheboxp90 Headedgoalsp90 Interceptionsp90 KeyPassesp90 Offsidesp90
1           0.005839487     0.02227428         1.563897    0.4650394  0.07215472
2           0.002312220     0.01196994         1.569028    0.4338627  0.05666571
  PassesSuccessfulLessCrossesAndCornersp90 PassesSuccessfulOpponentsHalfp90 PassesSuccessfulShortp90 Recoveriesp90
1                                 37.88310                          15.62807                34.70031      5.214819
2                                 36.52234                          15.18572                33.66996      5.265115
  Successfulcrossesp90 Successfullongpassesp90 Tackleswonp90     NPGp90 GoalContributionp90 ShotsBlockedp90
1            0.2627997                3.182795      1.278241 0.05379293          0.10989124       0.1234531
2            0.2875302                2.852378      1.293292 0.02736046          0.07415361       0.1030505
  ShotsOffTargetp90 Shotsontargetp90 Duelswonp90Adjusted GoalConversion Tackleswonp90Adjusted Clearancesp90Adjusted
1         0.2855115        0.1664477            5.070197      14.629772              1.274060              3.751240
2         0.2322835        0.1194463            4.973626       5.277345              1.259379              3.538006
  Minutespergoal
1      1878.8868
2       151.1802
```

Figura 36 Salida función kmeans (todas las variables)

```
Cluster means:
  AerialDuelsAccuracy Aerialduelswonp90 Assistsp90 Blocksp90 Clearancesp90 Chancescreatedp90 CornersWonp90
1            45.49245          1.176865 0.06614912 0.3271883      2.769646         0.7419032     0.4692531
2            61.72317          2.341924 0.04126190 0.5885306      4.413143         0.3620522     0.2269022
  CrossesAndCornersSuccessfulp90 DribblesSuccessfulp90 Duelswonp90 Headedgoalsp90 Interceptionsp90
1                      0.5745816             0.7935122    4.889071     0.006113536          1.462372
2                      0.2016937             0.4263744    5.225030     0.020418170          1.623333
  KeyPassesp90 PassesSuccessfulOpponentsHalfp90 Successfulcrossesp90 Successfullongpassesp90 Tackleswonp90
1    0.6757541                         17.20760             0.4631786                2.407551      1.422193
2    0.3207903                         14.33444             0.1809630                3.260536      1.216686
  GoalContributionp90 ShotsBlockedp90
1          0.09479770       0.13751115
2          0.08158351       0.09518364
```

Figura 37 Salida función kmeans (variables representativas)

9.7. Random Forest

```
set.seed(256)

library(dplyr)

players_entrenamiento <- sample_frac(newplayers, .7)
players_prueba <- setdiff(newplayers, players_entrenamiento)

library(randomForest)
#mostrar las variables
head(newplayers)

#Entrenamiento del modelo
rf.position<-randomForest(factor(Position) ~., data=players_entrenamiento)

print(rf.position)
importance(rf.position)

# Plot variable importance
varImpPlot(rf.position, main="",col="dark blue")
help(varImpPlot)

# Hacer predicciones set de prueba
predicciones <- predict(rf.position, players_prueba)
# Matriz de confusión
(mc <- with(players_prueba,table(predicciones, Position)))
```

Figura 38 Código Validación Random Forest