

Universidad de Buenos Aires  
Facultad de Ciencias Económicas  
Escuela de Negocios y Administración  
Pública

---

CARRERA DE ESPECIALIZACIÓN EN  
MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y  
ANÁLISIS DE DATOS EN ORGANIZACIONES

---

TRABAJO FINAL INTEGRADOR DE ESPECIALIZACIÓN'

---

**Big Data en el fútbol: Caracterización de  
jugadores según métricas avanzadas**  
Aplicación en jugadores de Argentina

---

**AUTOR: NELSON MENDOZA BARRIOS**

**MENTORA: SILVIA VIETRI**

**NOVIEMBRE - 2023**

---

## Resumen

El objetivo de este trabajo es hallar un método estadístico que permita detectar similitudes y diferencias entre los jugadores de un club en particular. La base de datos utilizada consta de diversas variables métricas registradas para cada uno de los jugadores que la integran, relacionadas con la posición que ocupan los mismos en el terreno de juego. A partir de estas características, se intenta identificar el estilo de juego de cada jugador, como, por ejemplo, ofensivo, defensivo, de distribución, entre otros.

Para abordar este problema, se empleó R Studio, donde se aplicó el Método de Análisis de Componentes Principales (ACP) y el Método de Conglomerados o *Cluster* para agrupar a los jugadores según las características identificadas en los resultados obtenidos. Con éxito, se pudo observar mediante la implementación del Análisis de Componentes Principales cómo los jugadores analizados se agrupan con relación a ciertas componentes que destacan sus mejores cualidades, lo que facilita la comprensión de los roles que desempeñan en el terreno de juego.

Este trabajo será de gran utilidad para comprender las diferentes facetas que presenta una posición en el terreno de juego y la influencia que puede tener en un esquema táctico. Además, brindará a los entrenadores, directores técnicos, y secretarios deportivos de entidades del rubro del deporte una solución para adquirir jugadores que se adapten a la estructura propuesta por su organización. Como resultado, se podrá optimizar de manera más eficiente el presupuesto disponible.

**Palabras clave: Métricas – Fútbol – Predicción – Análisis de Componentes Principales - Cluster**

## Índice

Introducción.....	- 1 -
1) Gestión de datos en contextos organizacionales. ....	- 3 -
1.1 – El futuro del futbol: Las agencias de scouting. ....	- 4 -
1.2 – Obtención y manejo de datos en una agencia de scouting. ....	- 6 -
1.3 – Big data: Una realidad y un desafío que mejoró el futbol. ....	- 7 -
1.4 – Breve explicación de Análisis de Componentes Principales.....	- 8 -
1.5 – Breve explicación de Análisis de Conglomerados .....	- 8 -
2) Descripción metodológica.....	- 8 -
2.1 – Recopilación de datos.....	- 9 -
2.2 – Procesamiento de datos. ....	- 10 -
2.3 – Caracterización de jugadores mediante Análisis de Componentes Principales -	11 -
2.4 – Caracterización de jugadores a partir de los resultados obtenidos por el método de Clusters. ....	- 20 -
2.4.1 – Método Ward.....	- 21 -
2.4.2 – Método K-Means.....	- 24 -
3) Conclusión.....	- 26 -
4) Bibliografía.....	- 27 -
5) Anexos.....	- 28 -
5.1 – Descripción de la base de datos.....	- 28 -
6) Reporte Trabajo Final Integrador de Especialización .....	- 30 -

## Introducción

Hoy en día, muchos se preguntan si existen métodos que permitan comparar a los jugadores, más allá de la observación específica de los mismos y los apuntes recopilados en el tiempo de acuerdo con cómo se desempeñaron, que sería la manera tradicional de detectar valores que puedan acoplarse y reforzar una plantilla de un club que está buscando mejorar. Hace más de una década, empresas de recopilación de datos ya vienen estudiando la posibilidad de reducir el trabajo mediante el uso de información que han recabado y, encontrando patrones de conducta en jugadores, las mismas se han encargado de trabajar con los clubes brindando una solución tanto táctica como presupuestaria, detectando jugadores que el club necesita por montos aceptables.

La primera recopilación de datos en el fútbol documentada se da en el año 1933, momento en que un hombre llamado Charles Reep decidió analizar ciertas acciones que rodean el juego para determinar quién ganaría en base a ciertas métricas que él consideraba importantes (Wilson, 2013). Reep sostenía que para hacer más goles se necesitaba hacer la menor cantidad de pases posibles y para esto era necesario jugadores de “juego directo”, ya que eso aseguraba atacar cuando el rival estaba desorganizado.

Avanzando a un análisis de nuestros tiempos sobre cómo llegar a la victoria en un partido de fútbol, hace poco más de 6 años, se determinó que es necesario que cada pase que se realice deje atrás a la mayor cantidad de jugadores rivales posibles, a este dato se le denominó “*Packed*”, una métrica avanzada desarrollada por la empresa de análisis de datos futbolísticos *Impect* (Biermann, 2019). Con esta métrica se pudo detectar la gran victoria de Alemania a Brasil en el Mundial del 2014 por 7 a 1 con el jugador con mayor índice en esta métrica siendo el alemán Toni Kroos. Al estudiar las demás métricas y al no encontrar que el equipo alemán haya superado en gran medida al equipo brasileño se optó por encontrar otras explicaciones al resultado tan llamativo, y es en este punto que *Impect* decide lanzar oficialmente la métrica “*Packed*” para explicar ciertos resultados. Determinar que jugadores podrían realizar esto se volvió un punto a ser estudiado por varios entrenadores.

Una métrica que se ha puesto muy de moda a la hora de analizar la calidad de remate de los jugadores de fútbol es el xG (Expected Goals) o goles esperados. Los goles esperados determinan tomando la posición, la forma y la dirección del remate asignándole un número de 0 a 1. Este número también se puede utilizar como porcentaje para explicar

que tan fácil o que tan difícil es anotar un gol, por ejemplo, los penales antes de ser pateados, y según varios penales ejecutados, obtienen un xG de 0.74 o 74% de probabilidad de ser gol, varios análisis se realizan sobre esta nueva métrica. Uno de ellos explica que si se divide la expectativa de gol (xG) total acumulado de un jugador en un partido jugado con todos los remates que realizó en el mismo encuentro, obtendremos su verdadero índice de peligrosidad en sus remates (Wahl, 2018). Esto hace que exista cierta forma de determinar si un jugador de ataque necesita o no ser reemplazado o necesita mejores compañeros para explotar más sus cualidades ofensivas, por nombrar un caso.

Sam Allardyce, director técnico inglés que ocupó en algún momento el banquillo del club Bolton Wanderers, ha logrado que su equipo anote entre el 45 al 50 por ciento de los goles en situaciones de balón parado, esto se llama así debido a que provienen de jugadas donde el balón no está corriendo, como en tiros libres, saques de esquina o laterales, es por eso que basó su búsqueda en jugadores que posean métricas elevadas en juego aéreo con éxito por citar un caso, lo cual le llevo a un trabajo de *scouting* sobre este aspecto (Stefan Szymanski, 2014). Mientras que el promedio de goles a balón parado de esa temporada era de solo 33%, su equipo excedía esta marca.

Ramon Rodríguez Verdejo, más conocido como Monchi, dejó una oración muy llamativa y concisa que explica la revolución de los datos en el fútbol, y cito:

“... El auténtico poder de los datos es el de cambiar nuestra relación con el juego.”

Y tiene razón, ya que, con el poder de los datos, él ha podido construir una red de reclutamiento de jugadores, lo que le permitió comprar jugadores con un potencial que no se ve a simple vista por montos bajos y luego vender a los mismos a clubes top por montos mucho más altos (tal es el reconocido caso del lateral brasileño Dani Alves), sosteniendo así la estructura económica del Club Sevilla (España) (Pinilla, 2019).

Los mercados de pases, como se denomina al acto de compra-venta de jugadores entre clubes, se realizan cada 6 meses aproximadamente, son momentos donde se hacen introspecciones para detectar que faltaría para mejorar la calidad de la plantilla (lista de jugadores de un club) y al encontrar la “pieza faltante” ver la forma de adquirirlo.

En estas operaciones se mueve muchísimo dinero, por dar un ejemplo según el portal *Transfermarkt*, el Chelsea Football Club de la liga de fútbol inglesa ha gastado en el

mercado de pases del verano europeo una asombrosa suma de €282.000.000 aproximadamente en de 8 jugadores, mientras que el Ajax de la liga de Países Bajos ha vendido prácticamente por la misma cantidad (unos €216.000.000 aproximadamente) en 7 jugadores (Transfermarkt, 2014). Esto nos marca diferentes aspectos que podrían formar parte de los objetivos de las entidades de este deporte.

El objetivo general es, mediante la utilización de Análisis de Componentes Principales (ACP) y el Análisis de Conglomerados o *Clusters*, efectuar una caracterización de jugadores, particularmente de la liga argentina de fútbol y mostrar en un análisis visual, cómo se agruparían dichos jugadores según ciertos criterios aplicados (Silvia Vietri, 2021). La base de datos con la que se trabajará corresponde al año 2022.

Este trabajo será de gran utilidad para conocer las distintas aristas que presenta una posición en el terreno de juego y el beneficio que puedan tener en un esquema táctico, brindando a los entrenadores y managers deportivos una solución a la hora de adquirir jugadores que se amolden a la estructura propuesta por la entidad deportiva.

Las interrogantes que plantea el presente trabajo son: ¿Podríamos elaborar un sistema de análisis de datos para detectar jugadores que mejor se adapten a nuestro estilo de juego? ¿Qué herramientas, que tengamos a nuestro alcance, podríamos utilizar para determinar similitudes y diferencias entre jugadores?

Para ello, el trabajo se ha dividido en varias secciones las cuales inician con el primer apartado, donde se describe la gestión de los datos en el contexto organizacional, continúa con el segundo apartado donde se describe la metodología utilizada para recopilar, procesar y analizar la información, luego el tercer apartado donde se define como se va a implementar el modelo de Análisis de Componentes Principales y el modelo de análisis de conglomerados o clusters, y termina con las conclusiones obtenidas, donde quedan expuestos todos los desarrollos y resultados revelados a partir de la investigación del presente trabajo.

## **1) Gestión de datos en contextos organizacionales.**

En esta sección, se describe el tipo de organización que va a ser desarrollado a lo largo del trabajo, indicando su objetivo, su composición y su modelo de negocios. Luego se

presenta la gestión de datos que va a atravesar la organización indicando el proceso de toma de decisiones y los datos que utiliza en ese proceso. Por último, se plantea la problemática a la que se enfrenta la organización en la actualidad.

### **1.1 - El futuro del fútbol: Las agencias de scouting.**

Cualquier club de fútbol del mundo que busque mejorar la calidad del plantel que posee tendrá como plan poder detectar de una manera más precisa qué tipo de jugadores necesita. Para ello, es posible recurrir a una agencia de *scouting* o reclutamiento que se encargue de buscar talentos en distintos países a través del uso de datos que se encuentran disponibles y se presentan habitualmente según lo que ocurre en el campo de juego.

Este tipo de agencias podría ser una organización de carácter privado e independiente a la que asisten y contratan clubes y federaciones para detectar talentos. Entre las tareas que cumple se encuentran: realizar un análisis completo del club, de su filosofía de juego, de su presupuesto, de su cultura, de fichajes anteriores, y otros factores que ayuden a crear los filtros a la hora de la búsqueda.

Con este propósito nacieron organizaciones como StatsBomb (StatsBomb, 2013) u Opta, cuyo trabajo principal era el de recopilar datos lo más exactos posibles sobre diferentes acciones de juego. Luego se han modernizado y mediante fórmulas y ecuaciones matemáticas, lograron remitir informes sobre los jugadores. Esto ha facilitado enormemente a los clubes el trabajo de búsqueda “*in situ*”, ya que eliminó los gastos masivos de traslado de los reclutadores a distintos países para observar nuevos talentos.

El objetivo de este tipo de organizaciones es la obtención constante de volúmenes extensos de datos de jugadores y sus acciones en el campo de juego, para ir agrupándolos según las necesidades de los clubes que participan en una liga específica.

Los clubes podrán solicitar informes de jugadores a la organización, la cual será preparada a medida para cada club, con diferentes nombres y estilos de juego con la finalidad de que se puedan adaptar al club lo más rápido posible y empezar a brindar los resultados que se esperan. La agencia de *scouting* se encargará de proveer material visual sobre los jugadores, pero no se hace responsable del análisis total, ya que los clubes cuentan con sus propios analistas de video, que deberán estudiar las imágenes para conocer aún más

al jugador, haciendo que la agencia se encargue de detectar jugadores exclusivamente con datos.

El modelo de negocio de la organización es lineal. La misma toma insumos de portales privados (de pago) que proveen las métricas de juego, las cuales son procesadas y analizadas con el fin de conocer a todos los equipos y puntualmente a los jugadores. Luego crea su propio modelo estadístico junto a otros análisis de datos obtenidos y ofrece a sus clientes los resultados junto a otros servicios personalizados.

La organización toma decisiones de negocio basadas en datos, algo que se denomina hoy en día como “*data driven decision making*” implicando que el resultado de los análisis de datos que se lleven a cabo formase parte de las estrategias a futuro. Se destacó una relación: cuantas más empresas se caracterizaron como basadas en datos, mejor se desempeñaron en medidas objetivas de resultados financieros y operativos. En particular, las empresas en el tercio superior de su industria en el uso de la toma de decisiones basada en datos fueron, en promedio, un 5 % más productivas y un 6 % más rentables que sus competidores (Brynjolfsson, 2012).

Como la mayoría de los trabajos de la organización requieren el uso de datos, el correcto procesamiento de estos, detectar patrones según los resultados y aplicar formulas algebraicas y conceptos de ciencia de datos se deduce que la organización recaerá totalmente en estos fundamentando más el concepto de “*data driven decision making*”, que las decisiones deben basarse en análisis y no por intuición (Fawcett, 2013). De esta forma, “comprender este proceso y sus etapas, ayuda a estructurar la resolución de problemas, la hace más sistemática y, por lo tanto, es menos propensa a cometer errores”.

La organización está compuesta por personas que tienen distintas funciones. A nivel general, se encuentran quienes poseen mayor conocimiento en extracción de datos, otros con mayor capacidad para procesamiento y análisis, y por último un área de creación de material visual. La combinación de estas funciones deriva en un informe detallado para los clientes de la agencia de *scouting*.

Las personas que forman parte de la organización ocuparan los cargos de *data engineers*, *data scientists* y analistas de videos. Los *data engineers* se encargarán de obtener la información, realizar limpieza de datos correspondiente, y dejar lo más pulido posible la



base de datos para que los *data scientists* puedan realizar los cálculos matemáticos correspondientes para lograr los objetivos organizacionales, así como modelos estadísticos que serán optimizados según las pruebas que se van realizando para lograr resultados más precisos para los clientes. Luego de detectar jugadores a seguir, los analistas de video buscaran contenido y crearan material visual para agregar a los informes, estos materiales serán recopilaciones de videos que tienen la finalidad de mostrar al jugador en situaciones de juego que correlacionen directamente con los resultados de las pruebas estadísticas, por ejemplo, si un jugador se ha destacado en acciones ofensivas, el video buscara reproducir exactamente eso.

### **1.2 - Obtención y manejo de datos en una agencia de scouting.**

La obtención de los datos se podría llevar a cabo mediante fuentes gratuitas o de pago existentes, algunas ya fueron citadas anteriormente, los cuales, generalmente, no están procesados ni poseen cierto grado de análisis de ninguna medida. La única función que se posee normalmente es de realizar filtros de los datos para hacer listas más cortas de jugadores ya sea por club, minutos jugados, país, periodo, etc

Las agencias necesitan precisión, entonces generalmente se recurre a la compra de información. Es necesario contar con las mejores herramientas posibles para llevar a cabo un análisis que pueda convencer a clubes millonarios que el jugador que se recomienda es crucial para un buen funcionamiento del equipo o una gran apuesta a futuro.

La organización busca que los informes sean entregados a los clientes con el mayor grado de veracidad posible para que los clubes usen de una mejor manera sus recursos económicos para mejorar su plantel. El principal potencial de las agencias de *scouting* es su modelo de caracterización de jugadores por posición, para agrupar jugadores que se sitúan en un mismo lugar en el terreno de juego, pero que cumplen roles distintos, esto ayudara a que las compras de jugadores sean más localizadas a la finalidad táctica del club y que la adaptación de los mismos surja de manera eficaz, y con resultados lo más inmediato posibles.

### 1.3 – Big data: Una realidad y un desafío que mejoró el fútbol.

El *big data*, se cree, ha sufrido una explosión entre el 2011 y el 2012, debido al aumento de información creada por segundo por páginas web dedicadas a distintos rubros, así también lo que las organizaciones podían producir. Hasta en los deportes se vio un crecimiento abismal en la información que se lograba recabar con el uso de GPS ubicados en indumentaria que se utiliza bajo las camisetas. El *big data* ha brotado por todas partes y la correcta utilización podría dar una ventaja competitiva, en el caso de la agencia de *scouting*, poder detectar talento lo antes posible para ofrecer informes a clubes, generaría confianza en el uso de datos para futuras transacciones. Así también la ignorancia en este campo producirá grandes riesgos para la organización y no las hará competitivas (Aguilar, 2013).

En el fútbol, el *big data* fue creciendo de manera lenta, hasta que, con la llegada de transmisión de partidos en vivo por televisión y la capacidad de grabarlos, se empezaron a desarrollar las primeras métricas oficiales como conteo de goles, faltas, pases para gol, registro de asistencia de fanáticos a los partidos, etc (Wilson, 2013). La aparición de Opta, denominado anteriormente STATS, y luego StatsPerform, abrió paso para otras organizaciones a apostar por recabar datos de situaciones de juego, con el objetivo de encontrar patrones, y entender mejor el fútbol.

Para poder elaborar un modelo estadístico que ayude a caracterizar a los jugadores es necesario contar con grandes volúmenes de datos, como mínimo de una temporada concluida, para poder establecer las comparaciones. Sin embargo, la organización se encuentra en un estado de sobrecarga de datos denominada “*big data*”, la cual conlleva mucho trabajo para extraer lo necesario.

Resulta difícil en muchos casos conseguir datos en portales gratuitos, esto ha obligado a consultoras de datos y agencias de *scouting* a realizar *partnerships* o alianzas con organizaciones que se dediquen al rubro del *big data*, y con el tiempo estas mismas organizaciones de recopilación de datos han decidido procesarlos y volverse ellos mismos los proveedores de informes, lo que obliga a las agencias a trabajar en diferentes aspectos y no solo en lo matemático, sino en lo deportivo y hasta en los comportamientos conductuales de los jugadores para buscar donde podrían adaptarse mejor.

A efectos del ejercicio de la organización, se trabajará con métricas disponibles del portal *Wyscout*, y se tomarán las métricas que mayormente describan una posición específica para demostrar el punto del modelo estadístico el cual será aplicado.

#### **1.4 - Breve explicación de Análisis de Componentes Principales**

El Análisis de Componentes Principales tiene como objetivo condensar o resumir información de un conjunto determinado de variables en otras variables llamadas componentes que determinan las dimensiones necesarias para representar adecuadamente los datos (Peña, 2002). Con ellos podemos hacer gráficos de los datos en pocas dimensiones, con mínima pérdida de información, para entender su estructura subyacente.

#### **1.5 - Breve explicación de Análisis de Conglomerados**

El análisis de componentes principales se puede combinar con la elaboración de conglomerados o *clusters*, los cuales pueden definirse a partir de distintos tipos de agrupamiento: agrupamiento jerárquico y/o agrupamiento no jerárquico. Podemos definir el *Clustering* como una forma de análisis exploratorio de datos donde las observaciones se dividen en grupos significativos que comparten características comunes entre sí (Gorenshteyn, 2022). El término *clustering* hace referencia a técnicas de aprendizaje no supervisado cuya finalidad es encontrar patrones o grupos dentro de un conjunto de observaciones. Agrupando el conjunto de datos basándose en la similitud de los valores de sus atributos. La distancia euclidiana forma parte del análisis de *clusters* ya que mediante esta medida se puede determinar la distancia similitud entre dos puntos en un espacio euclidiano (Bourbaki, 1987). Es ampliamente utilizada en diferentes áreas, como la estadística, el aprendizaje automático y la geometría.

## **2) Descripción metodológica.**

En esta sección se describe cada parte del proceso metodológico para alcanzar el objetivo propuesto por la organización o agencia de *scouting*, la cual se compuso por cómo se llevó a cabo la recopilación de datos, el análisis de estos datos de manera individual y posteriormente la aplicación de Análisis de Componentes Principales (ACP) como una vía de *machine learning* para hallar diferencias o similitudes entre jugadores, que si bien comparten la misma posición en el terreno de juego, pueden cumplir roles distintos.

## 2.1 – Recopilación de datos.

Los datos sobre jugadores de la liga inglesa fueron obtenidos del portal Wyscout, que se encarga de la creación de bases de datos con distintas métricas que miden las acciones en el terreno de juego, como también de proveer de material visual de las ya mencionadas acciones (Wyscout, 2022).

El relevamiento de datos incluye la temporada 2022. No se optó por un análisis de la temporada actual debido a que continua en proceso. La página web contiene datos de jugadores de todo el mundo, pero la idea del trabajo es optar por seleccionar una liga y comparar jugadores en la misma.

La base contiene 52 jugadores de diferentes clubes de la liga argentina, todos ocupan la posición de mediocampistas, nombre que se da a los jugadores que se sitúan en el centro del terreno de juego y poseen carácter mixto, ya que algunos podrían cumplir más roles defensivos u ofensivos que otros, lo que refuerza la idea de buscar caracterizar a los jugadores en esta posición según las métricas encontradas. Se empleó 11 atributos seleccionados específicamente para denotar las acciones más importantes que se realizan durante un partido de fútbol. La base de datos puede contener valores en cero, esto no significa que sean faltantes, solo que esas acciones no fueron realizadas en ningún momento por los jugadores.

Las 11 variables seleccionados para la realización del trabajo son: Equipo (categórica), Duelos defensivos/90 (numérica), Entradas/90 (numérica), Intercepciones/90 (numérica), Goles/90 (numérica), Remates/90 (numérica), Regates/90 (numérica) Pases recibidos/90 (numérica), Pases/90 (numérica), Asistencias/90 (numérica), Jugadas Claves/90 (numérica).

Al utilizar las métricas es necesario recurrir a una ecuación matemática que consiste en extraer las acciones de juego por partido jugado, esto significa que primeramente se observan las acciones totales y los minutos jugados totales, se dividen entre sí y el resultado se multiplica por 90 (que corresponde a los 90 minutos jugados reglamentariamente). Con esta ecuación podemos comparar de mejor manera a jugadores con más minutos, con jugadores que han jugado una menor cantidad de tiempo. Todo esto refuerza la posibilidad de llevar a cabo el experimento.

En el anexo 5.1 – “Descripción de la base de datos” se presenta la estructura del conjunto de datos, en donde se explica las características de las variables, a fin de dar más contexto a lo que se expone en este estudio.

## 2.2 – Procesamiento de datos.

Para poder iniciar el trabajo de caracterización de jugadores, es necesario conocer las distintas posiciones de los 11 jugadores que ingresan al terreno de juego, con esta información podemos empezar a buscar si un jugador que se sitúa en posición ofensiva posee las mismas cualidades que su par de otro club, es por eso por lo que un buen manejo de las variables es clave porque entenderemos su comportamiento. Algunos ejemplos de variables de comportamiento son goles, asistencias, duelos defensivos, remates, jugadas claves, etc. Estas métricas generalmente son analizadas con un enfoque por partido jugado, la cual nace de un cálculo matemático que busca dividir las acciones totales con los minutos totales jugados y multiplicarlos por 90 (tiempo de juego de un partido de fútbol). Esta fórmula matemática permite comparar de una manera más efectiva a jugadores que hayan jugado muchos más minutos que otros. A partir de lo mencionado, se realizará lo que se conoce como Análisis de Componentes Principales que buscará detectar cargas factoriales de las métricas, las correlaciones y los grupos que se formaran por el método de *clustering*.

Los datos fueron procesados y analizados en el lenguaje de programación R Studio (R Core, 2020) en su versión 3.6.3 correspondiente al 29 de febrero de 2020, con las librerías *pasteqs*, *dplyr*, *textshape*, *stats*, *corrplot*, *FactoMineR* y *factoextra* para llevar a cabo el Análisis de Componentes Principales (ACP), el análisis de *Clusters* o conglomerados así como la detección de grupos se realizara con la librería *NbClust* y las visualizaciones se llevarán a cabo con *ggplot2*, también disponible en la versión de R mencionada anteriormente.

Se trabajó primeramente en estudiar las estadísticas descriptivas de la base de datos, observando solamente las variables continuas pertenecientes a las acciones de juego por partido de 52 jugadores.

Para poder lograr un análisis completo de las variables es crucial calcular una matriz de correlaciones que determine que métricas se asocian mayormente entre sí. Esto es de gran

ayuda para las personas que no estén en mayor medida involucradas con el fútbol, muchos científicos de datos, estadísticos, matemáticos, etc, no necesariamente conocen el deporte, o lo ven muy poco, es por eso que con una matriz de correlaciones se pueden determinar, por ejemplo, qué se necesita más para realizar goles, o cómo se correlaciona realizar un pase con una jugada clave, brindando también una visión diferente del fútbol con números.

Para la investigación se optó por un análisis exhaustivo de jugadores a los que denominamos “mediocampistas” o “centrocampistas”. Estos jugadores se sitúan comúnmente entre los defensores, extremos y atacantes. Podrían tener características ofensivas (participan mayormente en la generación de goles para su equipo), defensivas (optando más por un arduo trabajo de recuperación del balón y cese de avance rival) o de distribución del balón (enfocándose mayormente en recibir el balón y distribuir a distintos sectores del terreno de juego). Es una posición que posee muchas funcionalidades dependiendo de lo que el director técnico (persona encargada del estilo de juego del equipo) esté buscando según lo que crea conveniente.

Es por lo anteriormente mencionado que un análisis detenido de jugadores en esta posición podría, de cierta forma, dar una ventaja competitiva a los *scouts* de los clubes a la hora de buscar reemplazantes. La correcta detección de las características de los mediocampistas del fútbol argentino podría significar horas y horas de búsqueda para aquel reemplazante perfecto solicitado por el director técnico tras la salida de un jugador clave en dicha posición.

### **2.3 – Caracterización de jugadores mediante Análisis de Componentes**

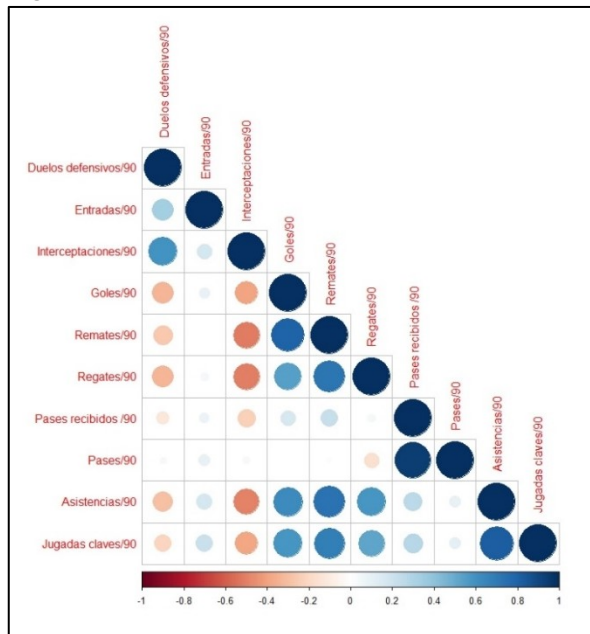
#### **Principales**

El proceso para caracterizar jugadores, e ir agrupándolos según su estilo de juego o las ordenes tácticas que reciban, inició con la matriz de correlaciones, en donde se buscó observar el grado de correlación de las variables que pueda dar un indicio de que variables métricas son más similares entre sí.

En la Figura N° 1 podemos observar que se van formando grupos variables correlacionadas como, por ejemplo, la variable de “Pases/90” y “Pases recibidos/90” contienen un grado alto de correlación, debido a que los jugadores que ocupan la posición

de “mediocampista” reciben tantos pases como los que terminan dando durante un partido de futbol. Otros ejemplos de variables que correlacionan en un alto grado y dan explicaciones coherentes referente al futbol son “Duelos defensivos/90” e “Intercepciones/90”, debido a que ambas poseen carácter defensivo, se realizan cuando el jugador no goza de la posesión del balón. También tenemos el caso de “Jugadas claves/90” y “Asistencias/90” así como “Remates/90” y Goles/90” que son de carácter de ofensivo y nos dan la pauta de que para que se dé uno, debe haberse ejecutado el anterior, por ejemplo, para que se dé un gol, se tuvo que realizar un remate, mismo caso que las asistencias, en donde debe existir una jugada clave previa.

**Figura 1: Matriz de correlaciones de variables.**



**Fuente: Elaboración propia en R Studio.**

Para continuar con el experimento, se trabajó en el Análisis de Componentes Principales que tiene como objetivo agrupar las variables según la carga factorial que poseen. Primeramente, se escalaron y se centraron las variables métricas, para luego evaluar mediante un “summary” los resultados.

En la Tabla N° 1 podemos observar la importancia de cada componente, estas se calculan según: la desviación estándar, la proporción de varianza y la varianza acumulada.

**Tabla 1: Importancia de las componentes.**

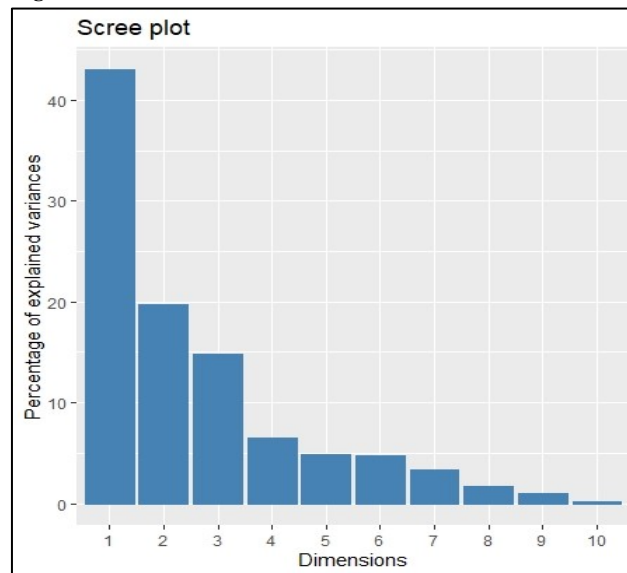
Importance of components:	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	2.0754	1.4056	1.2168	0.80444	0.70058	0.68957	0.5831	0.41346	0.31437	0.11365
Proportion of Variance	0.4308	0.1976	0.1481	0.06471	0.04908	0.04755	0.0340	0.01709	0.00988	0.00129
Cumulative Proportion	0.4308	0.6283	0.7764	0.84110	0.89018	0.93773	0.9717	0.98883	0.99871	1.00000

**Fuente: Elaboración propia en R Studio.**

Se pudo observar que, con 3 componentes, explicamos el 77.6% de nuestras variables, algo que resulta satisfactorio para continuar con el análisis. Otro criterio que se empleó para optar por la utilización de 3 componentes fue el criterio de “Kaiser” que indica que hay que conservar las componentes principales cuyos valores propios son mayores que la unidad, como se puede ver en la Tabla N° 1 la desviación estándar cae a 0.80444 en la componente principal 4, desechando así la opción de tomar esta componente.

También se decidió correr un *Scree Plot*, que tiene como finalidad indicar en un gráfico la varianza de cada componente. Como podemos ver en la Figura N° 2, las varianzas explicadas de las 3 primeras componentes son muy elevadas entre sí, a diferencia de la cuarta componente en adelante, se supuso así que la mejor decisión es tomar 3 componentes para continuar con el análisis.

**Figura 2: Scree Plot**

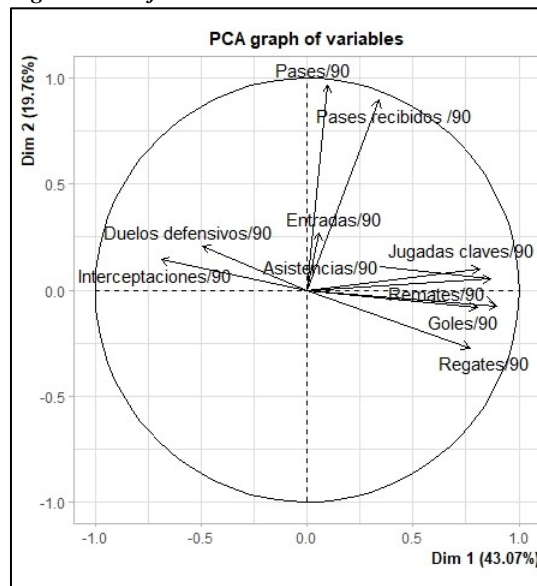


**Fuente:** *Elaboración propia en R Studio.*

El gráfico de variables, indicado en la Figura N° 3, se forma mediante las correlaciones de las variables presentadas en la base de datos, como forma parte del Análisis de Componentes Principales, se decidió profundizar en su composición y en que representa. Las correlaciones se miden según las distancias en el plano, por ejemplo, Regates/90 y duelos defensivos/90 poseen una correlación negativa (forman entre sí un ángulo de 180°), mientras que goles y remates prácticamente se superponen. Si el ángulo determinado por dos variables, por ejemplo, en este caso Pases/90 y Pases Recibidos/90 es de 90 °, significa que las variables no están correlacionadas o su correlación es cero, por lo tanto, no se relacionan linealmente.



Figura 3: Grafico de variables de APC.



Fuente: Elaboración propia en R Studio.

Posterior al análisis de los primeros gráficos derivados del ACP, se decidió asociar las variables con mayor carga factorial en las primeras 3 componentes. Se observa que las variables con mayor valor absoluto en las cargas factoriales definen a cada componente, las cuales se ven representadas en las siguientes tablas con sus respectivas cargas factoriales y la explicación de la denominación en función a las variables que las integran.

La Componente Principal N° 1 ha sido denominada “Acciones directas de gol”, ya que se conforma totalmente por variables que actúan en la obtención de un gol de manera directa, tales como “Regates/90”, “Remates/90”, “Asistencias/90”, “Jugadas Clave/90” y, lógicamente, “Goles/90”. Como se observa, todas las cargas factoriales son negativas.

Tabla 2: Componente Principal 1 – Acciones directas de gol.

VARIABLE	CARGA FACTORIAL
REMATES/90	-0.43235850
ASISTENCIAS /90	-0.41884209
JUGADAS CLAVE/90	-0.39192126
GOLES/90	-0.38836042
REGATES/90	-0.36978509

Fuente: Elaboración propia en Word.

Luego, vemos que la Componente Principal N° 2 ha tenido un enfoque diferente a la Componente Principal 1, como podemos ver en la Tabla 3. Esta se compone de variables como: “Pases Recibidos/90” y “Pases/90” cuyas cargas se registran con signo negativo. La Componente Principal 2 posee entonces un carácter que nos indica la capacidad de un jugador para recibir el balón y continuar distribuyéndolo por el campo de juego, recibió la denominación de “Acciones de distribución de balón”.

**Tabla 3: Componente Principal 2 – Acciones de Distribución de balón.**

<b>VARIABLE</b>	<b>CARGA FACTORIAL</b>
PASES RECIBIDOS/90	-0.63777184
PASES/90	-0.68888083

*Fuente: Elaboración propia en Word.*

Por último, hemos detectado que la Componente Principal N° 3 corresponde a acciones completamente opuestas a la Componente Principal N° 1. Esta componente está compuesta por acciones que el jugador realiza sin la tenencia del balón con el fin de resguardar su propio arco y evitar recibir goles. Son acciones conocidas como defensivas tales como “Duelos defensivos/90”, “Entradas/90” e “Intercepciones/90”, por lo que se denominó a esta componente como “Acciones defensivas” por la predominación de acciones de carácter defensivo. Esta componente también contiene cargas factoriales negativas.

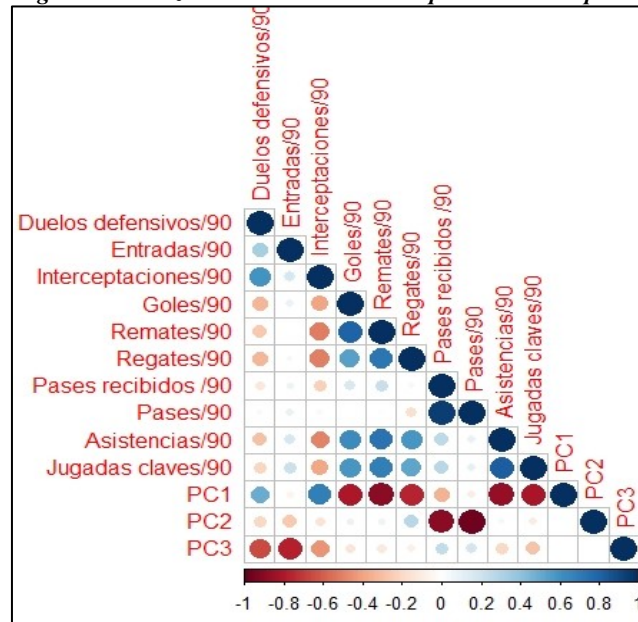
**Tabla 4: Componente Principal 3 – Acciones Defensivas.**

<b>VARIABLE</b>	<b>CARGA FACTORIAL</b>
ENTRADAS/90	-0.63481894
DUELOS DEFENSIVOS/90	-0.54190575
INTERCEPTACIONES/90	-0.36218233

*Fuente: Elaboración propia en Word.*

Una vez culminada la caracterización de las componentes, se decidió corroborar la correlación de las componentes formadas con las variables, realizando nuevamente la Matriz de Correlaciones con las Componentes Principales (anexadas a la base de datos). Se puede confirmar en la Figura N° 4 que las componentes formadas correlacionan fuertemente, y de manera negativa, con las variables seleccionadas para cada componente.

Figura 4: Matriz de Correlación con Componentes Principales

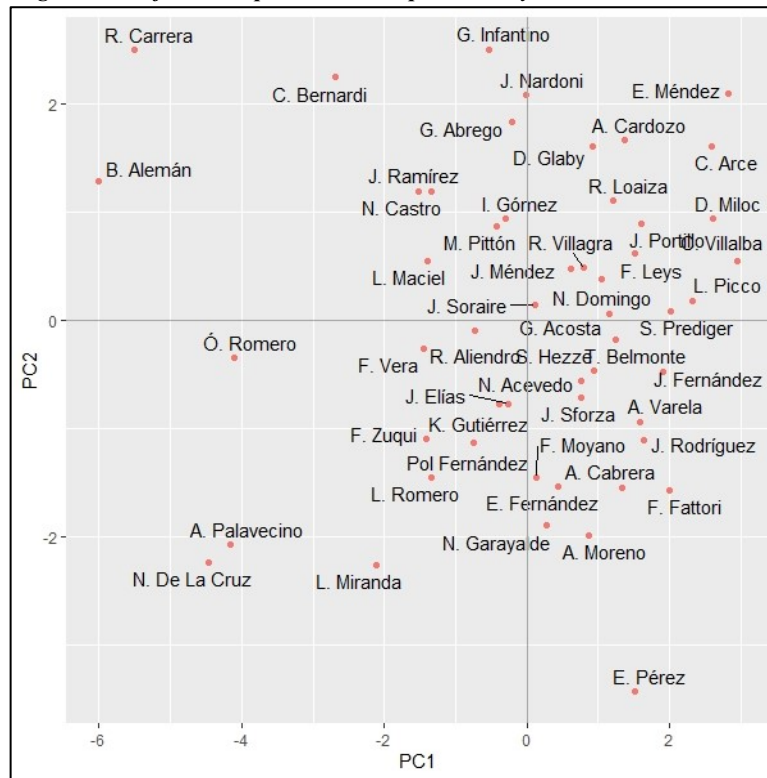


Fuente: Elaboración propia en R Studio.

En la siguiente sección se determinó el espacio que ocupan las observaciones (grupo de jugadores) en los gráficos de dispersión, cuyas coordenadas se fijan según su distancia entre las componentes. Se presentan entonces los gráficos de la Componente 1 y 2, 1 y 3, y 2 y 3 teniendo en cuenta las coordenadas que corresponden a cada individuo en cada componente principal. Es clave mencionar que, debido a que las cargas factoriales han arrojado valores negativos, se supone que la posición de las observaciones posee una representación inversa, por ejemplo, mientras que una observación (Jugador) se sitúe más hacia la derecha o hacia arriba en un gráfico (teniendo un valor superior a 0 o positivo) significa que es más débil en esa componente.

En el primer caso, se analizó cómo se ubicaban las observaciones cuando se compararon las componentes 1 y 2 como se puede ver en la Figura N° 5, las cuales representan **ACCIONES DE GOL DIRECTAS Y ACCIONES DE DISTRIBUCIÓN DE BALÓN.**

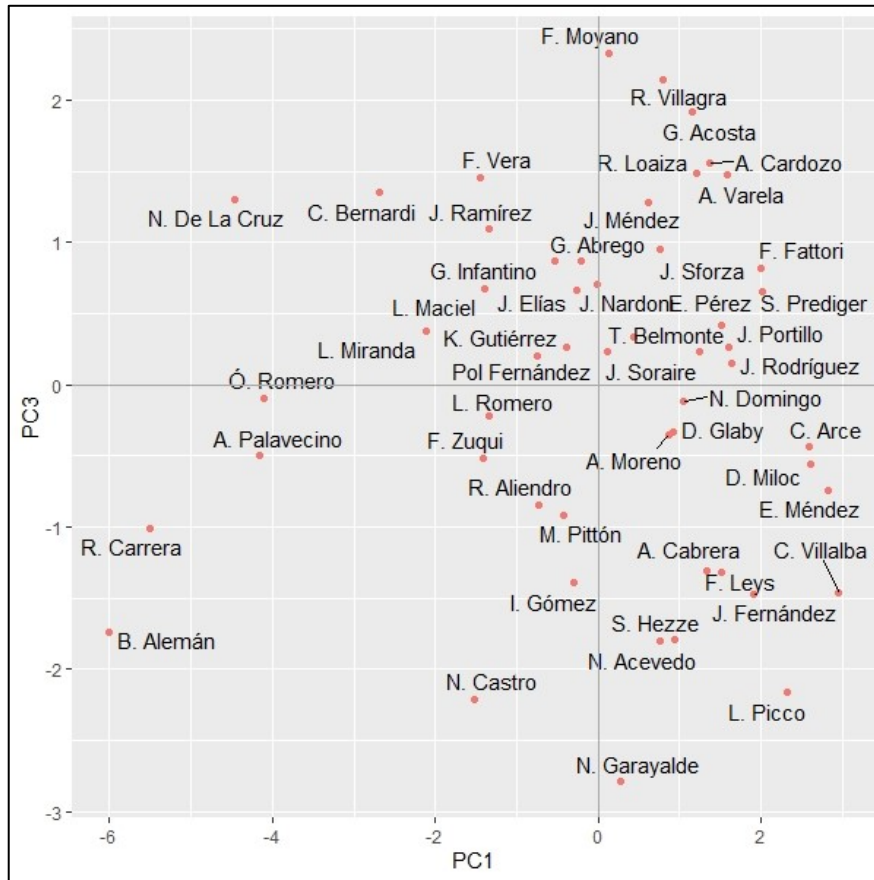
Como podemos observar, el jugador R. Carrera se sitúa en lo más alto del eje Y, que representa la Componente Principal N° 2, esto indica que el jugador no goza de una capacidad de distribución con relación a sus pares, pero si se mira el eje X se puede detectar que el jugador es uno de los más ofensivos, siendo el segundo mediocampista con más goles por partido, más remates por partido y cuarto mediocampista con más asistencias y jugadas claves por partido.

**Figura 5: Grafico de dispersión de Componentes 1 y 2**

Fuente: Elaboración propia en R Studio.

En el segundo caso tenemos el grafico de dispersión de las Componentes Principales 1 y 3, estas son **LAS ACCIONES DIRECTAS DE GOL Y LAS ACCIONES DEFENSIVAS**. Según lo que se puede detectar en el eje Y de la Figura N° 6, identificamos a F. Moyano como el jugador que posee el carácter defensivo más débil, mientras que N. Garayalde como el jugador con mejor cualidad defensiva. En cuanto al eje X, observamos los resultados previos de la Figura N° 5, la diferencia es que observando nuevamente al ejemplo del grafico anterior, R. Carrera, vemos que posee también un grado defensivo por arriba de lo normal, lo que lo convierte en un jugador que se adaptaría muy bien a esquemas donde realizaría muchas acciones defensivas y de gol.

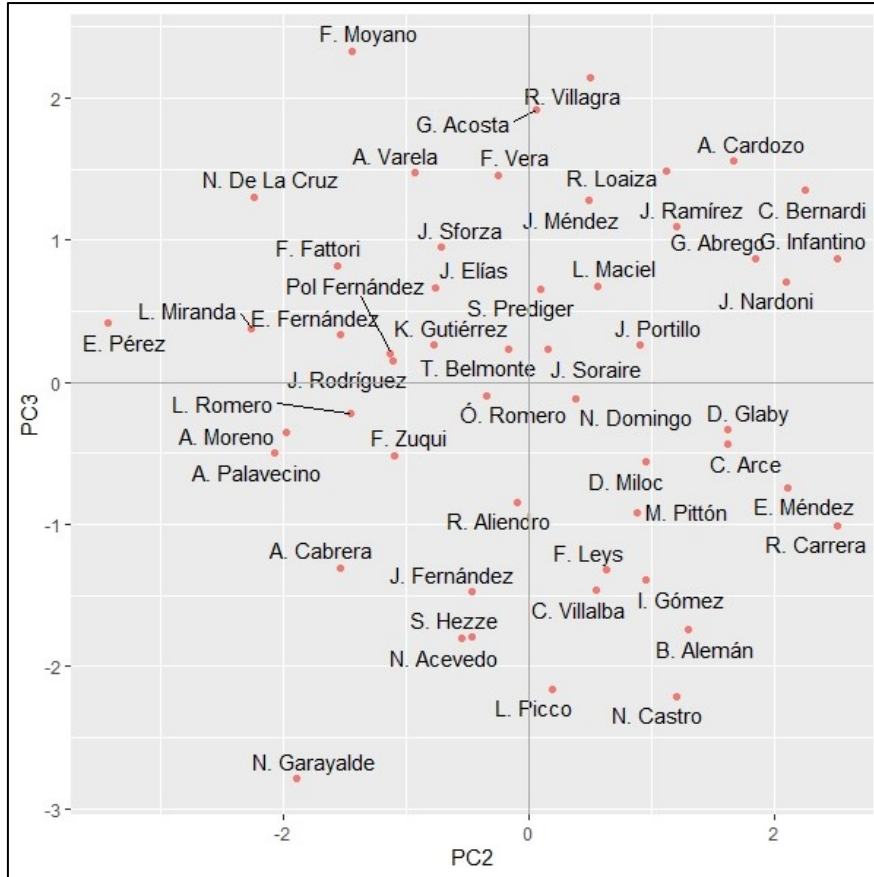
Figura 6: Grafico de dispersión de Componentes 1 y 3



Fuente: Elaboración propia en R Studio.

En la Figura N° 7, podemos observar que el grafico de dispersión de las Componentes Principales 2 y 3 relacionó **LAS ACCIONES DE DISTRIBUCIÓN DE BALÓN CON LAS ACCIONES DEFENSIVAS**, para determinar que jugadores son los que más aportan al juego defensivo del equipo y posterior distribución. Para citar los casos con mayor peso en cada componente, tenemos al jugador E. Pérez como el mejor exponente para representar la Componente Principal 2, mientras que el jugador N. Garayalde es el mejor representante de la Componente Principal 3, y sobre este punto se detectó que el jugador que combina de mejor manera ambas componentes es el citado anteriormente, ya que se sitúa más próximo a los valores negativos más altos en ambas componentes.

**Figura 7: Grafico de dispersión de Componentes 2 y 3**



*Fuente: Elaboración propia en R Studio.*

Algo que se ha detectado al finalizar la realización de los 3 gráficos de dispersión que han combinado las Componentes Principales seleccionadas es jugador que se ha situado en lo más cercano a las coordenadas (0,0), es el caso del jugador J. Soraire. Para estos casos, se concluyó que esta observación es etiquetada como un jugador cuyo rendimiento no sobresalen en ninguna de las categorías mencionadas (ataque, defensa o distribución), esto podría verse como algo positivo si es que se está buscando un jugador con características mayormente generales y no específicas, en torno a las métricas analizadas para cumplir distintos roles desde la posición de “mediocampista”.

Como este análisis se ha basado en rendimiento futbolístico, se cree que, no necesariamente es algo negativo situarse cercano a las coordenadas (0,0). Esto indica que el jugador, si bien no se destaca en ninguna Componente Principal, tampoco posee números por debajo de otros.

## 2.4 – Caracterización de jugadores a partir de los resultados obtenidos por el método de Clusters.

A la hora de estudiar que jugadores podrían ser agrupados según sus características de juego, lo primordial sería realizar un análisis de *Cluster* para determinar quiénes son similares entre sí con el fin de establecer posibles reemplazos para nuestro plantel. Como primera medida se estudian los análisis de *cluster* jerárquico y no jerárquico para establecer el número de agrupaciones que se apliquen y posteriormente la explicación de cada uno de ellos.

Para la realización de los *clusters* es necesario entender cómo se realizan las agrupaciones de observaciones resultantes del análisis. El Análisis de *Cluster* Jerárquico va generando grupos en cada fase del proceso de agrupación buscando el número de clases, grupos o clúster que hacer una agrupación óptima. Es capaz de fijar por si solos el número de *clusters*, por ello se pueden utilizar de forma exploratoria. Dentro del Análisis de *Cluster* Jerárquico podemos detectar diferentes métodos, como método del vecino más cercano, vecino más lejano, agrupación de vinculación promedio (*average*), Ward, Centroide y Mediana.

Posteriormente se podría aplicar un análisis no jerárquico, ya que estos categorizan elementos según un numero de *cluster* dado, que podría ser el resultante del análisis jerárquico. El método no jerárquico más popular el K medias o *K-Means*.

Basándose en los resultados previos de las Componentes Principales, se ha determinado que es necesario realizar un análisis preciso para determinar de qué manera se agruparan los datos. Mediante la utilización de la herramienta de estadística R Studio se llevó a cabo un análisis con la función *NbClust*, esta función permite a R Studio efectuar un cálculo rápido, según ciertos parámetros, para establecer el número de grupos. Estos parámetros son, por ejemplo, que tipo de distancia emplear (euclidiana, máxima, binaria, etc), un rango de componentes principales según la base de datos a utilizar, y que método aplicar (*K-means*, *Ward*, vecino más cercano, etc), y, por último, el *index* o índice, que determina finalmente la cantidad de *clusters*, en la cual se incluye normalmente a todos los índices posibles y dependiendo de la mayoría propuesta se establece el número de *clusters*.

Se trabajó con el método de Ward (Jerárquico) y con el método de K-Means (No Jerárquico) con la finalidad de caracterizar de la mejor manera las observaciones presentadas en este trabajo.

### 2.4.1 - Método Ward.

Para esto, se ejecutó la función NbClust de R Studio con la siguiente configuración: se solicitó distancia euclidiana, con un mínimo de 2 y un máximo de 10 *clusters*, el método empleado fue el *ward.D* (no se eleva al cuadrado las diferencias resultantes tras la actualización de los *clusters*), y se han empleado todos los índices posibles. Se sugirió trabajar con 2 *clusters*, como lo indica la Tabla N° 5. Tras la ejecución de la función, se procede a estudiar los promedios de las 10 variables para detectar dentro de cada grupo cuales serían las variables con mayor promedio de la media.

**Tabla 5: Resultado de NbClust con método ward.D**

```
*****
* Among all indices:
* 10 proposed 2 as the best number of clusters
* 2 proposed 3 as the best number of clusters
* 7 proposed 4 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 2 proposed 9 as the best number of clusters
* 4 proposed 10 as the best number of clusters

***** Conclusion *****

* According to the majority rule, the best number of clusters is 2

*****
```

**Fuente: Elaboración propia en R Studio.**

**Tabla 6: Promedios de variables en 2 grupos.**

Group	1	2	Duelos defensivos/90	Entradas/90	Interceptaciones/90	Goles/90	Remates/90	Regates/90	Pases recibidos /90	Pases/90	Asistencias/90	Jugadas claves/90
1	1	7.782000	0.9520	4.782000	0.063000	0.9765000	1.611500	32.68000	49.07350	0.715000	0.202000	
2	2	7.861562	0.8825	4.910625	0.049375	0.8665625	1.835313	19.53625	34.37719	0.531875	0.160625	

**Fuente: Elaboración propia en R Studio.**

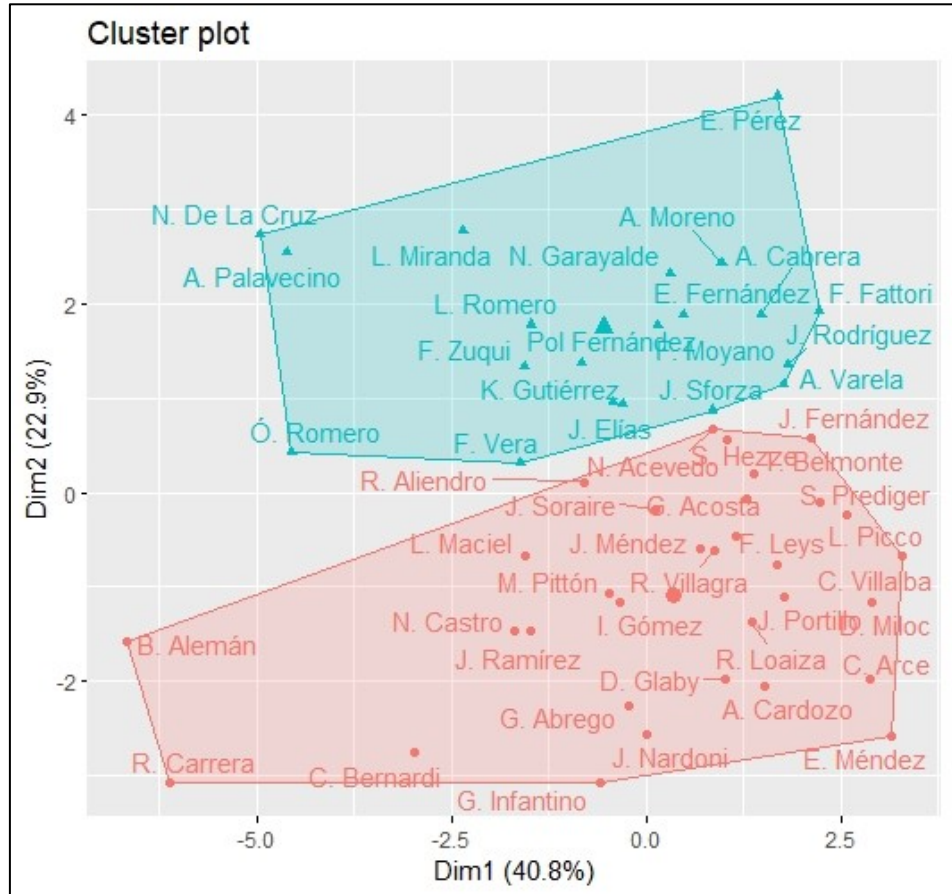
Se realizó un análisis de cómo han quedado los promedios más altos, según la Tabla N° 6 para definir como se agruparon las observaciones y que carácter tienen esos grupos.

Para el Grupo 1, se ha detectado que Jugadas claves/90, Entradas/90, Goles/90, Remates/90, Pases recibidos/90, Pases/90, Asistencias/90 poseen un promedio mayor, se cree así que los jugadores del Grupo 1 poseen mayor incidencia en anotar los goles de sus respectivos equipos, optando por un enfoque más ofensivo. En cambio, en el Grupo 2, se ha detectado que Duelos defensivos/90, Interceptaciones/90, Regates/90 poseen un promedio mayor, indicando así que estos jugadores aportan a su equipo un carácter más defensivo y de recuperación de balón.



La representación de las observaciones en sus respectivos grupos se puede visualizar en la Figura N° 8, en donde se puede notar que es muy pronunciada la separación entre las observaciones.

*Figura 8: Grafico de Clusters según método jerárquico de Ward.*



*Fuente: Elaboración propia en R Studio.*

El Grupo 1 se compone por 20 observaciones (38%), mientras que el Grupo 2 se queda con 32 observaciones (62%). Es común que el grupo con mayor incidencia ofensiva (Grupo 1) posea menos observaciones ya que, en una posición sensible como la de mediocampo, sean menos los jugadores que abarquen mayor cantidad de acciones determinantes para que su equipo logre quedarse adelantarse en el marcador en la búsqueda de la victoria, siendo que estos no son, por naturaleza, los que definen un partido. Mientras que el grupo mayoritario (Grupo 2), tenga mayor trabajo defensivo y de recuperación, algo que comúnmente corresponde a jugadores de dicha posición.

Finalmente, se expone en la Tabla N° 10 las observaciones con sus respectivos grupos resultantes del análisis de *clusters*.



1821 Universidad  
de Buenos Aires

**Tabla 10: Observaciones en sus respectivos grupos según Ward.**

<b>JUGADOR</b>	<b>GRUPO</b>
N. De La Cruz	1
A. Varela	1
F. Vera	1
A. Moreno	1
A. Palavecino	1
K. Gutiérrez	1
J. Sforza	1
Ó. Romero	1
J. Rodríguez	1
L. Romero	1
Pol Fernández	1
L. Miranda	1
N. Garayalde	1
F. Zuqui	1
J. Elías	1
F. Fattori	1
F. Moyano	1
E. Pérez	1
A. Cabrera	1
E. Fernández	1
S. Hezze	2
R. Villagra	2
T. Belmonte	2
J. Ramírez	2
J. Méndez	2
G. Infantino	2
J. Nardoni	2
G. Abrego	2
R. Carrera	2
R. Aliendro	2
A. Cardozo	2
M. Pittón	2
C. Villalba	2
C. Bernardi	2
B. Alemán	2
I. Gómez	2
J. Portillo	2
R. Loaiza	2
G. Acosta	2
L. Maciel	2
L. Picco	2
F. Leys	2
N. Castro	2



1821 Universidad de Buenos Aires

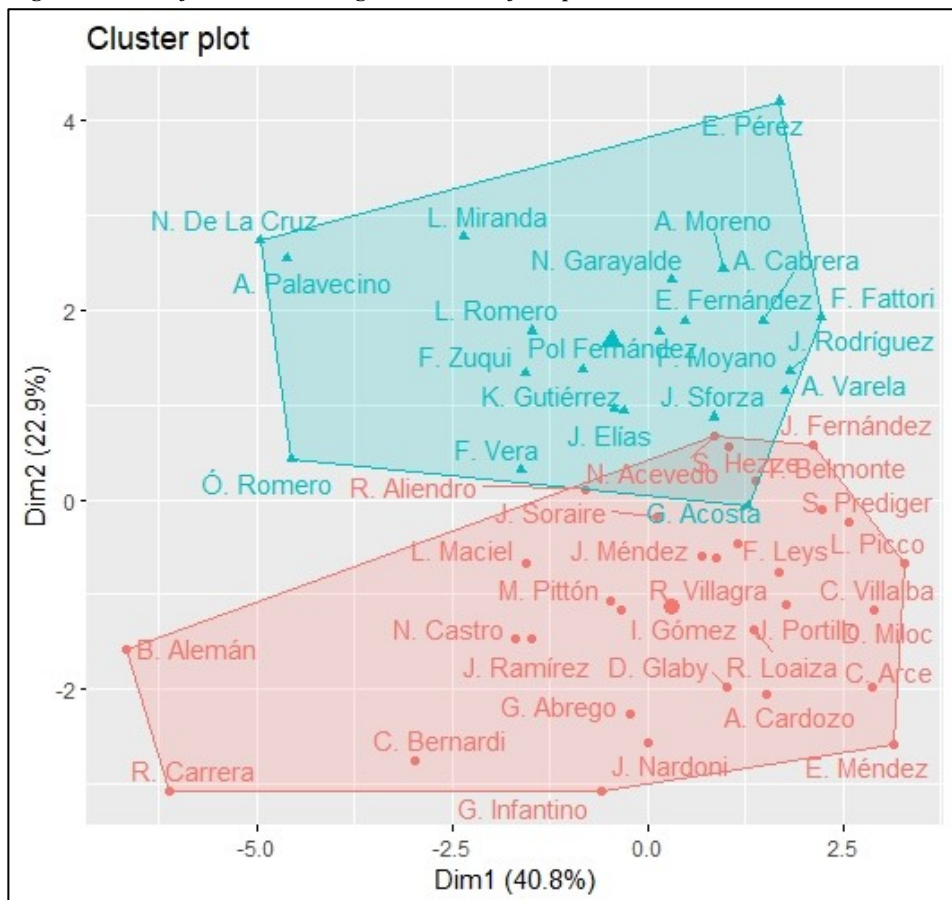
J. Soraire	2
D. Miloc	2
D. Glaby	2
N. Acevedo	2
E. Méndez	2
S. Prediger	2
N. Domingo	2
J. Fernández	2
C. Arce	2

Fuente: Elaboración propia en R Studio.

### 2.4.2 – Método K-Means.

Luego se pasó al método no jerárquico con el objetivo de detectar si hubo variaciones en la composición de individuos en los grupos establecidos previamente. Para lograr el análisis de K-Medias o *K-Means*, se empleó la función *kmeans* de la biblioteca de R Studio, la cual trabaja con la base de datos con la cual se llevó a cabo este estudio y una selección de *clusters* extraído del análisis jerárquico por el método de Ward que son 2 *clusters*.

Figura N° 9: Grafico de clusters según método no jerárquico de K-means



Fuente: Elaboración propia en R Studio.

Los resultados arrojados, como se puede ver en la Figura N° 9, son muy similares en cuanto a su composición, con una única diferencia que se detectó en una observación. El jugador G. Acosta se ha cambiado de grupo, formando parte ahora del Grupo 1, el cual incluye jugadores más determinantes a la hora de anotar goles, y no sorprende ya que se ubica muy cercano a la media en cuanto a “Asistencias/90” y “Pases Recibidos/90”, y si por encima de la media en cuanto a “Pases/90”, lo cual supone que es una pieza clave en su equipo a la hora de distribuir el balón en zonas cercanas al arco del equipo rival, justificando así el cambio de grupo tras el análisis de K-medias.

Entonces la distribución de observaciones quedaría de la siguiente manera: 21 observaciones corresponden al Grupo 1, mientras que 31 observaciones corresponden al Grupo 2. A continuación se expone en la Tabla N° 11 la distribución final de observaciones por grupos.

**Tabla 11: Observaciones en sus respectivos grupos según Kmedias**

<b>JUGADOR</b>	<b>GRUPO</b>
N. De La Cruz	1
A. Varela	1
F. Vera	1
A. Moreno	1
A. Palavecino	1
K. Gutiérrez	1
J. Sforza	1
Ó. Romero	1
J. Rodríguez	1
L. Romero	1
Pol Fernández	1
L. Miranda	1
N. Garayalde	1
F. Zuqui	1
J. Elías	1
F. Fattori	1
F. Moyano	1
E. Pérez	1
A. Cabrera	1
E. Fernández	1
G. Acosta	1
R. Villagra	2
T. Belmonte	2
J. Ramírez	2
J. Méndez	2
G. Infantino	2



1821 Universidad  
de Buenos Aires

.UBAeconómicas posgrado

ENAP Escuela de Negocios y Administración Pública

J. Nardoni	2
G. Abrego	2
R. Carrera	2
R. Aliandro	2
A. Cardozo	2
M. Pittón	2
C. Villalba	2
C. Bernardi	2
B. Alemán	2
I. Gómez	2
J. Portillo	2
R. Loaiza	2
S. Hezze	2
L. Maciel	2
L. Picco	2
F. Leys	2
N. Castro	2
J. Soraire	2
D. Miloc	2
D. Glaby	2
N. Acevedo	2
E. Méndez	2
S. Prediger	2
N. Domingo	2
J. Fernández	2
C. Arce	2

Fuente: *Elaboración propia en R Studio.*

### 3) Conclusión

En el contexto del fútbol, y en efecto en cualquier disciplina deportiva, la búsqueda constante de la victoria impulsa a los clubes a explorar métodos que garanticen la identificación y adquisición de talentos excepcionales. En este sentido, la aplicación de análisis de datos emerge como una herramienta de gran valía para desvelar a aquellos jugadores que a menudo escapan a la detección convencional debido a una caracterización superficial. Es común que los cazatalentos basen sus evaluaciones en la intuición y la experiencia previa, desestimando los matices y necesidades precisas del equipo en cuestión.

Este estudio se orientó a la caracterización de 52 jugadores que ocupan posiciones y regiones similares en el campo de juego. La finalidad principal radicaba en discernir entre aquellos jugadores que, pese a desempeñar roles aparentemente análogos, aportan de

manera dispar en el engranaje colectivo, empleando enfoques distintos para contribuir al éxito del equipo.

A través de una reducción dimensional, se logró sintetizar la información en tres componentes principales. Estos componentes, a saber, acciones ofensivas, distribución del balón y acciones defensivas, encapsulan las facetas críticas del fútbol. El fruto de esta reducción dimensional se traduce en una representación visual de las preferencias y tendencias de los jugadores, plasmadas en un plano bidimensional mediante diversas combinaciones de dichos componentes.

Esta metodología arrojó una perspicacia más aguda acerca del impacto relativo de los jugadores en diferentes aspectos del juego. Mediante la aplicación de técnicas de agrupamiento, fue posible perfilar a los jugadores con mayor precisión, segmentándolos en dos categorías distintas: aquellos cuyo enfoque radica en potenciar las jugadas de gol y quienes se inclinan hacia la recuperación de la posesión.

En síntesis, se infiere que una adopción efectiva de la metodología de análisis expuesta en este trabajo por parte de los departamentos de reclutamiento de los clubes deportivos permitirá establecer criterios más selectivos para la identificación de jugadores que se alineen de manera óptima con el estilo de juego preconcebido por el cuerpo técnico. Asimismo, la correcta ejecución de este enfoque contribuirá a minimizar gastos superfluos en jugadores de renombre, ya que es plausible que entre los jugadores más jóvenes se encuentren las características esenciales para la consecución de triunfos en el campo de juego.

#### **4) Bibliografía**

- Aguilar, L. J. (2013). *Big Data - Analisis de grandes volúmenes de datos en organizaciones*. Ciudad de Mexico: Alfaomega Grupo Editor.
- Analyst, J. W. (14 de 04 de 2022). *Evolving Expected Goals (xG)*. Obtenido de <https://theanalyst.com/eu/2022/04/evolving-expected-goals-xg/>
- Biermann, C. (2019). *Football Hackers: The Science and Art of a Data Revolution*. Londres: Blink Publishing.
- Bourbaki, N. (1987). *Topological Vector of Spaces. Chapters 1-5*. Heidelberg: Springer-Verlag.
- Brynjolfsson, A. M. (Octubre de 2012). *Big Data: The Management Revolution*. Obtenido de <https://hbr.org/2012/10/big-data-the-management-revolution>

- Fawcett, F. P. (13 de 2 de 2013). *Data Science and its Relationship to Big Data and Data-Driven Decision Making*. Obtenido de <https://www.liebertpub.com/doi/full/10.1089/big.2013.1508>
- Gorenshteyn, D. (2022). *Datacamp*. Obtenido de Cluster Analysis in R: [www.datacamp.com](http://www.datacamp.com)
- Peña, D. (2002). *Análisis de datos multivariantes*. Madrid: McGraw Hill.
- Pinilla, D. (2019). *El Metodo Monchi: El sistema de trabajo del exitoso director deportivo*. Buenos Aires: Librofutbol.com.
- R Core, T. (2020). *R: A language and environment for statistical*. Obtenido de <https://www.R-project.org>
- Silvia Vietri, S. D. (2021). *Unidad 3: Analisis de Componentes Principales*. Buenos Aires: Universidad de Buenos Aires.
- StatsBomb. (2013). *StatBomb Data Champions*. Obtenido de <https://statsbomb.com/es/>
- Stefan Szymanski, S. K. (2014). *Soccernomics: Why England Loses, Why Germany and Brazil Win, and Why the U.S., Japan, Australia, Turkey -- and Even Iraq -- Are Destined to Become the Kings of the World's Most Popular Sport*. Nueva York: Nation Books.
- Transfermarkt. (2014). *Transfermarkt*. Obtenido de <https://www.transfermarkt.es/>
- Wahl, G. (2018). *Football 2.0: How the world's best play the modern game*. Nueva York: Crown Archetype.
- Wilson, J. (2013). *Inverting The Pyramid: The History Of Soccer Tactics*. Nueva York: Nation Books.
- Wyscout. (2022). *Wyscout*. Obtenido de <https://platform.wyscout.com/app/?>

## 5) Anexos

### 5.1 – Descripción de la base de datos.

En la Tabla N° 12 se exponen las variables de la base de datos seleccionada y se explicará en qué consisten para contextualizar la información de las secciones anteriores.

*Tabla 12: Variables de la base de datos.*

VARIABLE	DESCRIPCIÓN
<b>Nombre</b>	Nombre del jugador
<b>Equipo</b>	Club en el cual el jugador presta servicios durante el periodo analizado.
<b>Duelos defensivos/90</b>	Los duelos son intentos (exitosos o no) de un equipo de frenar el avance del rival, impidiendo que lleguen a su arco.

<b>Intercepciones/90</b>	Esta acción consiste en cortar un pase rival, para la recuperación de la posesión del balón. Métrica por partido jugado.
<b>Entradas/90</b>	Representan las barridas realizadas por un jugador (exitosa o no) para recuperar la posesión de balón.
<b>Goles/90</b>	Acción de mandar el balón a las redes del arco, generalmente del rival, esto suma un gol o un punto al equipo que lo logre, para que se contabilice el balón debe cruzar la línea de meta. Métrica por partido jugado.
<b>Remates/90</b>	Acción de patear el balón con destino a gol, en ocasiones el remate puede ir fuera del terreno de juego, así como en los postes o directamente a las redes. Métrica por partido jugado
<b>Regates/90</b>	Acción de llevar el balón en los pies e intentar eludir a un rival.
<b>Pases recibidos/90</b>	Acción de recibir el balón por parte de un compañero. Métrica por partido jugado.
<b>Pases/90</b>	Acción de entregar el balón a un compañero, esto puede ser con éxito, en caso de que nuestra entrega finalmente de con un compañero de equipo, o sin éxito si el balón es interceptado por un rival o sale de los límites del terreno de juego. Métrica por partido jugado.
<b>Asistencias/90</b>	Acción de dar un pase a un compañero y que este, subsecuentemente anote un gol. Métrica por partido jugado.
<b>Jugadas claves/90</b>	Acción de dar un pase a un compañero, y que con este pase exista una probabilidad estimativa y no calculada, que el compañero haga un gol. Métrica por partido jugado.

*Fuente: Elaboración propia*



## 6) Reporte Trabajo Final Integrador de Especialización

“Big Data en el fútbol: Caracterización de jugadores según métricas avanzadas”  
Aplicación en jugadores de Argentina

**Autor: Nelson Mendoza Barrios**

Como mentora del presente Trabajo Final Integrador de Especialización del alumno Nelson Mendoza, transmito mi opinión sobre la investigación realizada.

Con respecto al tema de análisis, considero que aborda una problemática relacionada con la gestión de datos en organizaciones, en este caso organizaciones relacionadas con la compra y venta de jugadores de fútbol; tema relevante en la actualidad.

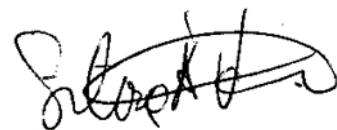
El alumno analiza, a partir de un conjunto de datos de variables registradas en un grupo de jugadores de fútbol de la liga argentina, distintos aspectos relacionados fundamentalmente con el estilo de juego y rendimiento de cada uno.

Los objetivos, tanto el general como los específicos son coherentes y acordes a la hipótesis propuesta.

En la investigación se trata, por un lado, de reducir el número de variables a registrar para identificar condiciones distintivas en los jugadores, y por otro, hallar patrones de conducta en los jugadores que permitan caracterizarlos, generando de esta forma, una alternativa beneficiosa desde el punto de vista de la táctica y también desde lo económico. Caracterizar a los jugadores permite predecir equipos que funcionen exitosamente desde la óptica de los resultados de juego, y también detectar jugadores que puedan resultar valiosos en el equipo, generando erogaciones de dinero que sean aceptables para los clubes.

En el trabajo se describen aspectos teóricos que dan marco a la investigación y se articulan contenidos de las asignaturas Métodos de Análisis Multivariado (con la aplicación de distintos métodos trabajados durante la cursada, como Método de Componentes Principales y Método de Conglomerados) y Taller de Programación (con la aplicación de RStudio).

La coherencia del enfoque planteado, el uso de algoritmos de Minería de Datos, la pertinencia de las referencias bibliográficas y el correcto análisis de los resultados obtenidos, permiten señalar a este trabajo como un aporte relevante del tema tratado, evidenciando la posibilidad de emprender líneas de investigación futuras en el área deportiva.



Dra. Silvia Vietri