

Automatización en la recolección de datos Deportivos

AUTOR: Prada, Tadeo (Leg. Nº 60219)

Coluccio, Ignacio Nicolas (Leg. Nº 60288)

DOCENTE/S TITULAR/ES O TUTOR/ES O DIRECTOR: Rodríguez Varela, Juan Pablo

Gonzáles Rodríguez, Rubén Darío

Nosetti, María Inés

PROYECTO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE LICENCIADO/A Analítica
Empresarial y Social

BUENOS AIRES

PRIMER / SEGUNDO CUATRIMESTRE, 202X

Making Basketball Smart



82.02 PROYECTO FINAL - ANALÍTICA

AUTOMATIZACIÓN EN LA RECOLECCIÓN DE DATOS

DEPORTIVOS

ENTREGABLE FINAL

GRUPO 3: IBALL

TADEO PRADA LEGAJO: 60219

IGNACIO COLUCCIO LEGAJO: 60288

ÍNDICE

Objetivo del proyecto	3
Medición de Valor y KPIs a Impactar	3
Entregables y Outputs del Proyecto	3
Investigación de Metodologías y Publicaciones Académicas Relevantes	4
Abordaje del Problema y Plan de Trabajo	5
Datos a Utilizar	5
Herramientas a Utilizar	6
Business Case (Caso de Negocio)	6
Descripción de las bases	7
Jugadores Estadísticas	132
Estructura	132
Missings	198
Outliers	19
Correlaciones	220
Estadísticos descriptivos	242
Equipos Estadísticas	264
Estructura - estadísticos descriptivos	264
Outliers	29
Correlaciones	29
Planteo de Hipótesis	341
Enfoque de Solución	384
Metodologías a Implementar	406
Seteo de Experimentación	406
Desarrollo de la Solución	406
Clusters Generales	406
Clusters de Defensa	48
Clusters de Ataque	49
Perfiles de los Jugadores	49
Plantillas	541
Conclusiones	617
Puesta en Acción	628

OBJETIVO DEL PROYECTO

Es de público conocimiento que los datos están presentes en casi todos los ámbitos de la sociedad. Pero estos no significan nada si no son procesados y estudiados por el ojo conecedor, para hacer de los mismos información. La disponibilidad de los datos es importante pero la correcta utilización, y aplicación de los mismos es lo que genera la diferencia.

Esto puede aplicarse a los deportes, por la gran presencia de datos crudos generados, en especial en el basquet.

El basquet es uno de los deportes en los que más se utilizan los datos, siendo este un deporte que depende cada vez más y más de estadísticas, probabilidades y mejora de porcentajes.

Es por esto que se busca generar un diferencial desde el lado analítico y comprender el juego a través de estos números.

Es aquí donde viene el centro de nuestro proyecto. El cliente se presentó con dos problemáticas a solucionar con técnicas analíticas. Primero, al poseer tanta cantidad de estadísticas y métricas a calcular se invierte mucho tiempo en las etapas de obtención de datos y cálculo de dichas métricas. Tiempo que podría ser utilizado para entrenar a los jugadores para el próximo partido. Segundo, existe un estancamiento en la definición de posiciones a la hora de jugar. Históricamente el básquet tiene 5 posiciones que se distribuyen casi siempre del mismo modo. Esto hace que el armado de equipos siempre sea igual y la innovación sea baja.

La solución propuesta entonces se dividirá en dos partes. Inicialmente se pondrá énfasis en la automatización, ya que, en este momento el proceso de disponibilización de los datos se hace de manera manual lo que es extremadamente laborioso y lento. Se busca reducir el tiempo utilizado en estas tareas para así poder enfocarse en otros trabajos de más impacto y que no se puedan realizar por procesos automatizados. Luego, con técnicas de clustering, se buscará definir perfiles para cada uno de los jugadores, independientemente de cuál sea su posición original. Por encima de esto, a través de técnicas de programación lineal se armaran equipos que optimizaran los puntos realizados.

Los datos utilizados serán los del campeonato sudamericano de U-18, aunque esta es una gran posibilidad que se genere un proceso capaz de ser utilizado en otras competencias y por otros equipos.

MEDICIÓN DE VALOR Y KPIS A IMPACTAR

Estas serán las métricas utilizadas para poder medir el impacto del proyecto:

- Cantidad de tiempo ahorrado para el plantel técnico. Será medida en horas
- Número de equipos propuestos con técnicas analíticas.
- Efectividad de las plantillas propuestas. Se analizará viendo el desempeño de las plantillas en el juego y cómo impactan en el resultado. Se medirá con un diferencial de puntos al momento de entrar contra la cantidad de puntos al momento de salir.

ENTREGABLES Y OUTPUTS DEL PROYECTO

Los entregables a diseñar serán herramientas que el cuerpo técnico pueda utilizar de manera fácil e interactiva para poder prepararse para los partidos. Deben ser suficientemente fáciles para que un equipo, que no es experto en el uso de herramientas de este tipo pueda utilizarlo.

Por ende como entregables tendremos:

- Archivo .py que se encargue de scrapear, limpiar, juntar, promediar, y segmentar los datos. Los procesos de este entregable son:
 - Scraping de los datos: Se conecta por medio de una url y se inspecciona la página con el html, mediante la utilización de la librería "BeautifulSoup". De esta manera se descarga la información de cada partido a la plataforma "Spyder".
 - Limpiar los datos.
 - Se generan las bases con los datos de todos los partidos del torneo.
 - Se descarga la información al equipo.
- Archivo de Power BI. El mismo deberá ser actualizado únicamente, debido a que se buscará obtener una base que junte el total de los datos, y se sobrescriba cada vez que se corra el programa, permitiendo que siempre se mantenga el mismo archivo de origen y se filtre la información que desea verse directamente en la plataforma de microsoft. En este caso los pasos son:

- Conexión con los archivos obtenidos del paso anterior.
 - Definición de visualizaciones.
- Archivo .py para el armado de plantillas. Archivo en el cual se van a poder utilizar los clusters diseñados. Deberá ser simple de utilizar.
 - Tuneo de las restricciones
 - Comparación de resultados con el equipo actual.

INVESTIGACIÓN DE METODOLOGÍAS Y PUBLICACIONES ACADÉMICAS RELEVANTES

Luego de haber realizado una búsqueda sobre la temática en cuestión, se pudo acceder a una cantidad de publicaciones académicas que indican la situación actual con respecto al estudio de la temática.

Algunas de las conclusiones obtenidas después de su lectura son las siguientes:

- Hay un consenso general que las cinco posiciones definidas clásicamente para jugar de manera óptima están desactualizadas. (Narayan, 2019)
- Para intentar solucionar el ítem anterior se suelen utilizar metodologías de clusterización, tales como PCA, K-Means y Random Forest.
- La adquisición de los datos, el procesamiento y la limpieza de los datos es un proceso clave que suele ser el más determinante a la hora de obtener resultados.
- Gracias a la utilización de estadísticas avanzadas se puede obtener información que permita definir el valor de un jugador lo más objetivamente posible.

ABORDAJE DEL PROBLEMA Y PLAN DE TRABAJO

Se podrán distinguir tres distintas etapas en el proyecto,:

1. Recolección de Datos
2. Análisis
3. Visualización e Implementación.

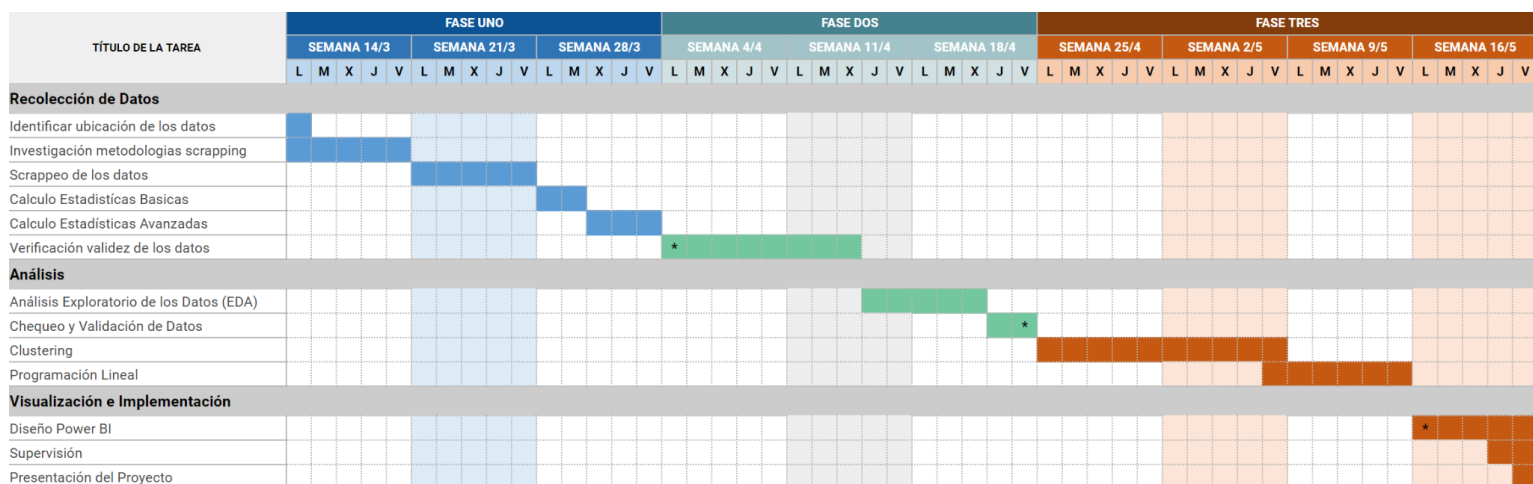


Figura 1: Gantt del Proyecto

Como puede verse, el proyecto está planeado con una metodología “Waterfall”. Sin embargo, debido a la naturaleza del problema y la cantidad de información y datos que devuelve, no se descarta la posibilidad de encontrar problemas en un punto avanzado del proyecto que deban ser resueltos más adelante. Otra posibilidad es que mediante la obtención de algún dato o la utilización de alguna herramienta, se puedan abrir posibilidades a la obtención de otros datos relevantes, por lo que la puerta no se cierra a información nueva. Por ende, si bien la base es una metodología “Waterfall”, tiene también algún comportamiento propio de la metodología “Agile”.

DATOS A UTILIZAR

Los datos serán obtenidos a través de metodologías de scrapping de la página oficial del South American U18 Championship. Los mismos son disponibilizados en la siguiente página: <https://www.fiba.basketball/sudamericano/u18/2022>

Es importante tener en cuenta que al ser datos scrapeados, será necesario un proceso de limpieza de datos exhaustivo para asegurar que los datos son correctos.

También es importante destacar que el proceso funciona para cualquier torneo oficial de FIBA, por lo que es posible utilizar esta solución en múltiples ocasiones.

HERRAMIENTAS A UTILIZAR

Como se mencionó, las herramientas a utilizar serán:

- Spyder: Permitirá hacer el procesamiento de los datos. Se crearán los clusters y las plantillas con programación lineal.
- Excel: Donde se almacenarán los datos.
- Power BI: Donde se visualizará la información.

BUSINESS CASE (CASO DE NEGOCIO)

Este proyecto debe considerarse un disparador.

Se define Data Automation como el proceso de carga, manejo y procesamiento de datos a través de herramientas automatizadas en lugar de realizar manualmente todas estas tareas. Este proceso es una tendencia cada vez más creciente en todas las industrias. No hay razón por la cual los deportes y el básquet sea la excepción.

Los procesos de Data Automation suelen estar asociados a dos principales ventajas:

- Tiempos reducidos: Al ser tareas realizadas automáticamente reducen los tiempos de trabajo para los humanos, quienes pueden disponer ese tiempo para tareas no automatizables, como entrenar jugadores.
- Mejor rendimiento y escalabilidad: Estos procesos presentan la información de manera consistente y uniforme habilitando la utilización de otras tecnologías tales como Machine Learning.

Del mismo modo, las técnicas de Machine Learning habilitan el descubrimiento de nuevos insights los cuales un humano quizás no podría identificar.

El caso de negocio de este proyecto es llevar todos estos procesos utilizados de manera frecuente en industrias como venta minorista o marketing al mundo del básquet. Y hacer de esta idea un ejemplo. Con el objetivo de cada vez implementar más y más tecnología en los deportes.

En este caso se comprobó que las ventajas consecuentes a la utilización de esta solución son:

- Disminución de horas hombre invertidas: Pasando de invertir más de 2 horas en lograr el resultado a cuestión de minutos.
- Eliminación de errores: Se minimiza la probabilidad de error en los cálculos, ampliamente diferente a la probabilidad presente cuando el proceso es realizado por un humano.
- Adaptación: Puede ser utilizado para cualquier torneo oficial de FIBA, generando así la necesidad de invertir una única vez en el proyecto.

- Escaso Mantenimiento: No debería ser necesario un mantenimiento de la solución en condiciones normales.
- Ventaja Competitiva: Al ser una solución completamente nueva, el resto de los seleccionados se ve en inferioridad de condiciones en estas temáticas.
- Mayor conocimiento del juego: El sistema de clusters permite entender el juego de tal manera que brinda una ventaja por sobre los competidores.
- Disparador de Posibilidades: La inversión en una solución de estas características puede ser la puerta de entrada a un mundo de posibilidades relacionadas al análisis de datos aplicados al deporte.

DESCRIPCIÓN DE LAS BASES

Una de las particularidades del proyecto es que la información viene de una página web. Por lo que una parte de la solución es generar un código en python, que se encargue de scrapear la información desde el HTML, la limpie, promedie los resultados de cada partido, calcule nuevas variables y la descargue.

Debido a todo este proceso, las bases de este proyecto son varias y se dividen en varias categorías.

La primera distinción es que hay bases para los equipos y para los jugadores. Esto se debe a que cada uno tiene estadísticas diferentes que impactan de manera diferente en los análisis a realizar.

La segunda distinción es la separación entre las estadísticas básicas y las estadísticas avanzadas. Estas no son disponibilizadas en la página, sino que son calculadas en el proceso de scraping y limpieza de los datos. Por consiguiente, las bases de datos a utilizar son las siguientes:

- Equipos Estadísticas Básicas:
 - Min: Cantidad de minutos disputados por un equipo.
 - Pts: Cantidad de puntos anotados por un equipo.
 - OREB: Cantidad de rebotes ofensivos agarrados por un equipo.
 - DREB: Cantidad de rebotes defensivos agarrados por un equipo.
 - REB: Cantidad de rebotes agarrados por un equipo.
 - AST: Cantidad de asistencias efectuadas por un equipo.
 - PF: Faltas cometidas por un equipo.
 - TO: Pérdidas de balón efectuadas por un equipo.
 - STL: Robos de balón efectuados por un equipo.
 - BLK: Bloqueos realizados por un equipo.
 - +/-: Puntos a favor de un equipo.
 - EFF: Eficiencia de un equipo. Tiene en cuenta la mayor parte de las estadísticas comunes.

- 2pa: Tiros de 2 puntos intentados.
 - 2pm: Tiros de 2 puntos encestados.
 - 2pp: Porcentaje de tiros de 2 puntos.
 - 3pa: Tiros de 3 puntos intentados.
 - 3pm: Tiros de 3 puntos encestados.
 - 3pp: Porcentaje de tiros de 3 puntos.
 - fga: Tiros de campo intentados.
 - fga: Tiros de campo encestados.
 - fgp: Porcentaje de tiros de campo.
 - fta: Tiros libres intentados.
 - ftm: Tiros libres encestados.
 - ftp: Porcentaje de tiros libres.
- Jugadores Estadísticas Básicas:
 - Min: Cantidad de minutos disputados por el jugador.
 - Pts: Cantidad de puntos anotados por el jugador.
 - OREB: Cantidad de rebotes ofensivos agarrados por el jugador.
 - DREB: Cantidad de rebotes defensivos agarrados por el jugador.
 - REB: Cantidad de rebotes agarrados por el jugador.
 - AST: Cantidad de asistencias efectuadas por el jugador.
 - PF: Faltas cometidas por el jugador.
 - TO: Pérdidas de balón efectuadas por el jugador.
 - STL: Robos de balón efectuados por el jugador.
 - BLK: Bloqueos realizados por el jugador.
 - +/-: Puntos a favor del jugador.
 - EFF: Eficiencia del jugador. Tiene en cuenta la mayor parte de las estadísticas comunes.
 - 2pa: Tiros de 2 puntos intentados.
 - 2pm: Tiros de 2 puntos encestados.
 - 2pp: Porcentaje de tiros de 2 puntos.
 - 3pa: Tiros de 3 puntos intentados.
 - 3pm: Tiros de 3 puntos encestados.
 - 3pp: Porcentaje de tiros de 3 puntos.
 - fga: Tiros de campo intentados.
 - fga: Tiros de campo encestados.
 - fgp: Porcentaje de tiros de campo.
 - fta: Tiros libres intentados.
 - ftm: Tiros libres encestados.
 - ftp: Porcentaje de tiros libres.
- Equipos Estadísticas Avanzadas:

- POS_eq: Es el espacio de juego que va desde que un equipo toma el control del balón hasta que el equipo contrario tome control del mismo.
- PLAYS_eq: Es el espacio de juego que va desde que un equipo posee el balón por cualquier motivo incluyendo REBOTE OFENSIVO hasta que deja de tener el balón.
- RITMO_eq: Es el promedio de posesiones entre ambos equipos ajustado a 40 minutos de juego.
- OER_eq: Ratio de Eficiencia Ofensiva. (Offensive efficiency rating) Puntos producidos por posesión.
- DER_eq: Ratio de Eficiencia Defensiva. (defensive efficiency rating) Puntos recibidos por posesión.
- eFG%: Porcentaje de tiro Efectivo. (efficiency field goal percentage): Es el porcentaje real de cancha de un equipo contemplando el punto extra que otorga el tiro de tres puntos para definir la real eficiencia.
- ORB%_eq: Porcentaje de Rebotes Ofensivos. Es el porcentaje de rebotes ofensivos capturados sobre la cantidad de rebotes disponibles en el aro del rival por un equipo.
- DRB%_eq: Porcentaje de Rebotes Defensivos. Es el porcentaje de rebotes defensivos capturados sobre la cantidad de rebotes disponibles en el propio aro por un equipo.
- AST/TO_eq: Ratio de Asistencias sobre Perdidas. Cantidad de asistencias en función de las perdidas que tiene un equipo.
- STL/TO_eq: Ratio de Recuperos sobre Perdidas. Cantidad de recuperos en función de las perdidas que tiene un equipo
- TS%_eq: Porcentaje de Tiro Verdadero. Es una medida de eficiencia de tiro de un equipo teniendo en cuenta tiros de 2P, 3P y TL.
- PPT2_eq: Puntos por Tiro de 2P. Es la eficiencia en los lanzamientos de 2P expresada en puntos obtenidos por cada lanzamiento de un equipo.
- PPT3_eq: Puntos por Tiro de 3P. Es la eficiencia en los lanzamientos de 3P expresada en puntos obtenidos por cada lanzamiento de un equipo.
- PPTL_eq: Puntos por Tiro Libre. Es la eficiencia en los lanzamientos de TL expresada en puntos obtenidos por cada lanzamiento de un equipo.
- %USO3P_eq: Porcentaje de uso de tiro de 3P. Es el porcentaje de jugadas finalizadas en tiro de 3P por un equipo

- TRB_eq: Porcentaje de Rebotes totales. (total rebound percentage) Es el porcentaje total de rebotes capturados sobre todos los rebotes disponibles en el partido por un equipo.
 - %USO2P_eq: Porcentaje de uso de tiro de 2P. Es el porcentaje de jugadas finalizadas en tiro de 2P por un equipo.
 - %USOTL_eq: Porcentaje de uso de Tiro Libre. Es el porcentaje de jugadas finalizadas en tiro libre por un equipo.
 - TOV%_eq: Porcentaje de Pelotas Perdidas. Es el porcentaje de balones perdidos en función de las jugadas que tuvo el jugador durante el juego.
 - FTR_eq: Ratio de Tiro Libre. Es la frecuencia con la que un equipo va a la línea de tiros libres en función de la cantidad de veces que toma un tiro de campo.
 - AST%_eq: Porcentaje de asistencias (Assists percentage) Es el porcentaje de canastos convertidos a través de pase gol.
 - %TS_eq: : Porcentaje de Tiro Verdadero. (True shooting percentage) Es una medida de eficiencia de tiro de un jugador o un equipo teniendo en cuenta tiros de 2P, 3P y TL.
 - BLK%_eq: Porcentaje de bloqueos. (Blocks percentage) Es el porcentaje de balones bloqueados por nuestro equipo en función de los tiros intentados del equipo oponente.
 - NET RATING: La resta entre eficiencia ofensiva y la eficiencia defensiva.
 - STL%_eq: Porcentaje de recuperos. (Steals percentage) Es el porcentaje de recuperos de nuestro equipo en función de las jugadas del equipo oponente.
- Jugadores Estadísticas Avanzadas:
 - PLAYS: Son las jugadas utilizadas por un jugador/a.
 - PPP:Ratio de Eficiencia Ofensiva. Puntos producidos por jugada del jugador/a.
 - eFG%: Porcentaje de tiro Efectivo. (efficiency field goal percentage): Es el porcentaje real de cancha de un jugador/a contemplando el punto extra que otorga el tiro de tres puntos para definir la real eficiencia.
 - USG%: Porcentaje de uso. Es el porcentaje de jugadas que un jugador/a utilizo sobre el porcentaje de jugadas que tuvo el equipo ajustado al tiempo de juego del jugador.
 - ORB%: Porcentaje de Rebotes Ofensivos. Es el porcentaje de rebotes ofensivos capturados sobre la cantidad de rebotes disponibles en el aro del rival por un jugador/a ajustado a los minutos que estuvo en cancha.

- DRB%: Porcentaje de Rebotes Defensivos. Es el porcentaje de rebotes defensivos capturados sobre la cantidad de rebotes disponibles en el propio aro por un jugador/a ajustado a los minutos que estuvo en cancha.
- TRB%: Porcentaje de Rebotes Totales. Es el porcentaje de rebotes totales capturados sobre la cantidad de rebotes disponibles por un jugador/a ajustado a los minutos que estuvo en cancha.
- AST/TO: Ratio de Asistencias sobre Perdidas. Cantidad de asistencias en función de las perdidas que tiene un jugador/a.
- STL/TO: Ratio de Recuperos sobre Perdidas. Cantidad de recuperos en función de las perdidas que tiene un jugador/a.
- TS%: Porcentaje de Tiro Verdadero. Es una medida de eficiencia de tiro de un jugador/a teniendo en cuenta tiros de 2P, 3P y TL.
- PPT2: Puntos por Tiro de 2P. Es la eficiencia en los lanzamientos de 2P expresada en puntos obtenidos por cada lanzamiento de un jugador/a.
- PPT3: Puntos por Tiro de 3P. Es la eficiencia en los lanzamientos de 3P expresada en puntos obtenidos por cada lanzamiento de un jugador/a.
- PPTL: Puntos por Tiro Libre. Es la eficiencia en los lanzamientos de TL expresada en puntos obtenidos por cada lanzamiento de un jugador/a.
- PPTC: Puntos por Tiro de Campo. Es la eficiencia en los lanzamientos de TC expresada en puntos obtenidos por cada lanzamiento de un jugador/a.
- %USO3P: Porcentaje de uso de tiro de 3P. Es el porcentaje de jugadas finalizadas en tiro de 3P por un jugador/a.
- %USO2P: Porcentaje de uso de tiro de 2P. Es el porcentaje de jugadas finalizadas en tiro de 2P por un jugador/a.
- %USOTL: Porcentaje de uso de Tiro Libre. Es el porcentaje de jugadas finalizadas en tiro libre por un jugador/a.
- TOV%: Porcentaje de Pelotas Perdidas. Es el porcentaje de balones perdidos en función de las jugadas que tuvo el jugador durante el juego.
- FTR: Ratio de Tiro Libre. Es la frecuencia con la que un jugador va a la línea de tiros libres en función de la cantidad de veces que toma un tiro de campo.
- AST%: Porcentaje de asistencias. Es el porcentaje de canastos convertidos a través de pase gol.
- STL%: Porcentaje de recuperos. Es el porcentaje de recuperos de un jugador en función de las jugadas del equipo oponente ajustado a los minutos que ese jugador/a estuvo en cancha.

- o BLK%: Porcentaje de bloqueos. Es el porcentaje de balones bloqueados por un jugador/a en función de los tiros intentados del equipo oponente ajustado a los minutos que el jugador estuvo en cancha.
- o HOB: Numero de canastos en los que un jugador/a esta involucrado como anotador o pasador.

Es importante destacar que todas estas estadísticas son calculadas por partido para comprender el progreso a lo largo de la competición.

DESCRIPCIÓN TÉCNICA DE LAS BASES

Se hará un análisis y descripción técnica de las bases. Por esto se hace referencia a un análisis de la estructura, los estadísticos descriptivos y los missings de cada una de las bases.

JUGADORES ESTADÍSTICAS

ESTRUCTURA

Ambas bases de jugadores están compuestas por 442 registros, que hacen referencia a 96 jugadores, los cuales participaron del torneo. Hay 25 variables consideradas “Estadísticas Avanzadas” y 24 consideradas “Estadísticas Básicas”. Con el fin de facilitar el análisis se hizo un “join” entre ambas bases, por lo que se puede decir que la base de “Jugadores” contiene 52 variables que definen el juego de cada jugador.

En cuanto al tipo de dato, son todos datos numéricos excepto la columna “Players” que contiene el nombre de cada jugador.

A continuación se puede visualizar una imagen del resumen de la base:

```

Data columns (total 54 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Players              96 non-null     object
1   PLAYS                96 non-null     float64
2   PLAYS_eq             96 non-null     float64
3   PPP                  96 non-null     float64
4   eFG%                 96 non-null     float64
5   USG%                 96 non-null     float64
6   ORB%                 96 non-null     float64
7   DRB%                 96 non-null     float64
8   TRB%                 96 non-null     float64
9   AST/TO               96 non-null     float64
10  STL/TO               96 non-null     float64
11  TS%                  96 non-null     float64
12  PPT2                 96 non-null     float64
13  PPT3                 96 non-null     float64
14  PPTL                 96 non-null     float64
15  PPTC                 96 non-null     float64
16  %USO3P               96 non-null     float64
17  %USO2P               96 non-null     float64
18  %USOTL               96 non-null     float64
19  TOV%                 96 non-null     float64
20  FTR                  96 non-null     float64
21  AST%                 96 non-null     float64
22  POS_op               96 non-null     float64
23  STL%                 96 non-null     float64
24  BLK%                 96 non-null     float64
25  HOB                  96 non-null     float64
26  Unnamed: 0           96 non-null     int64
27  #                    96 non-null     int64
28  Min                  96 non-null     float64
29  Pts                  96 non-null     float64
30  OREB                 96 non-null     float64
31  DREB                 96 non-null     float64
32  REB                  96 non-null     float64
33  AST                  96 non-null     float64
34  PF                   96 non-null     float64
35  TO                   96 non-null     float64
36  STL                  96 non-null     float64
37  BLK                  96 non-null     float64
38  +/-                  96 non-null     float64
39  EFF                  96 non-null     float64
40  2pm                  96 non-null     float64
41  2pa                  96 non-null     float64
42  2pp                  96 non-null     float64
43  3pm                  96 non-null     float64
44  3pa                  96 non-null     float64
45  3pp                  96 non-null     float64
46  fgm                  96 non-null     float64
47  fga                  96 non-null     float64
48  fgp                  96 non-null     float64
49  ftn                  96 non-null     float64
50  fta                  96 non-null     float64
51  ftp                  96 non-null     float64

```

Figura 2: Estructura de la base

Como puede observarse todos los datos son numéricos.

A continuación se realizó un resumen de las variables de esta base. Estos datos son muy importantes a la hora de definir objetivos. Esto se debe a que permiten comprender las capacidades de los jugadores sea cual sea la competición y definir los objetivos acorde con estos valores, siendo lo más objetivos posible. Por ejemplo: si la solución es utilizada en un torneo donde compiten selecciones mayores, los valores promedio serán mayores que si compiten seleccionados menores de edad, es por esto que deben definirse distintos “benchmarks” para cada competición.

Variable	mean	min	50%	75%	max
PLAYS	8,24775	0,33333333	7,236	11,006	24,24
PLAYS_eq	95,5940903	88,74	93,9783333	99,8675	106,1
PPP	0,60339948	0	0,602	0,79625	1,1872

eFG%	0,33767344	0	0,3387	0,437	0,8334
USG%	0,19456602	0,05593545	0,200080	0,22710929	0,32269636
ORB%	0,0602434	0	0,0411	0,08395	0,4635
DRB%	0,11591181	0	0,10943333	0,15265	0,32
TRB%	0,08740608	0	0,0845	0,1141625	0,193
AST/TO	0,66432639	0	0,5167	0,94255	3,0132
STL/TO	0,55289045	0	0,5	0,8687	1,5334
TS%	0,36679809	0	0,3675	0,4718	0,8422
PPT2	0,68936111	0	0,7058	0,8836	1,6666
PPT3	0,46424271	0	0,3625	0,76625	1,6
PPTL	0,71574965	0	0,6818	1,1364	1,894
PPTC	0,67537569	0	0,6773	0,87365	1,6666
%USO3P	0,23299878	0	0,2074	0,3582	0,71575
%USO2P	0,47327639	0,09533333	0,4823	0,586875	1
%USOTL	0,07857344	0	0,068275	0,1072	0,3206
TOV%	0,19470625	0	0,1903	0,25075	0,80966667
FTR	0,17868628	0	0,1259	0,25355	1,02
AST%	0,10849149	0	0,085	0,16205	0,423
POS_op	79,3422153	73,92	78,152	84,352	86,9866667

STL%	0,01131076	0	0,008875	0,0166	0,07033333
BLK%	0,01634861	0	0,009875	0,0234	0,1465
HOB	0,13097641	0	0,11679602	0,17760343	0,41884907
Min	17,5777922	2,41666667	17,2966667	23,325	33,2966667
Pts	5,61875	0	5,35	7,65	16,2
OREB	1,15277778	0	1	1,6	4
DREB	2,49652778	0	2,45	3,6	7
REB	3,64930556	0	3,6	5	8,6
AST	1,08576389	0	0,7	1,6	4,8
PF	1,49583333	0	1,5	2	3,6
TO	1,50416667	0	1,45	2	4,2
STL	0,83211806	0	0,775	1,2	3
BLK	0,45208333	0	0,25	0,6	3,2
+/-	-0,07760417	-17	-2,3	4,05	23,2
EFF	5,52760417	-4,4	5,325	8,8	18,8
2pm	1,59965278	0	1,45	2	5,6
2pa	3,99322917	0,25	3,6	5,25	12,6
2pp	0,34467096	0	0,35285714	0,44180556	0,83333333
3pm	0,4953125	0	0,2	0,8	3

3pa	2,05017361	0	1,8	3,2125	9
3pp	0,15474781	0	0,12083333	0,25541667	0,53333333
fgm	2,09496528	0	1,8	2,8	6,8
fga	6,04340278	0,33333333	5,325	8,4	17,4
fgp	0,30413442	0	0,30636364	0,3950919	0,83333333
ftm	0,93350694	0	0,8	1,4	3,2
fta	1,59131944	0	1,2	2,2	6
ftp	0,3149324	0	0,3	0,5	0,83333333

Tabla 1: Estadísticos de la base

Gracias a la **Tabla 1** se puede llegar a las siguientes conclusiones iniciales:

- No hay jugadores que jueguen todo el partido.
- Hay jugadores que no anotaron un tiro en todo el torneo.
- El que más cantidad de tiros anotados en promedio anotó 7 tiros.
- El que más cantidad de tiros anotados en promedio intentó 17 tiros.
- El porcentaje de tiro de campo medio es de 0,30 y el 75% de los jugadores tiró menos de 0,40. Esto es propio en los equipos de este rango etario, dónde las capacidades físicas y técnicas no están desarrolladas al completo.
- El 75% de los jugadores anota menos de 8 puntos por partido mientras que el que más anotó en promedio tuvo un poco más de 16 puntos.
- Al haber 96 jugadores y 8 equipos, puede afirmarse que no hubieron jugadores que no ingresaron en ninguno de los partidos, ya que las plantillas están conformadas por 12 jugadores.
- El HOB refiere al porcentaje de canastas en las que un jugador forma parte. Puede verse que el 75% de los jugadores tiene un 0,18 de HOB, mientras que el mayor tiene un HOB de 0,44. Puede ser relevante diferenciar al jugador con ese último valor para definir la modalidad de juego del equipo al que pertenece y hacer un correcto planteo de la táctica del partido, si se enfrentaran con este.
- Un uso similar se le puede dar a las variable de Uso:

- Por un lado el USG% que define el porcentaje de jugadas de las que participó un jugador/a, la cual define cómo utiliza el equipo a cada jugador.
 - Por otro lado el %USO3P, %USO2P y %USOTL, ya que definen de una correcta manera la predisposición de cada jugador a cada tipo de jugada.
- En cuanto al PPP, que resulta muy útil para evaluar la contribución ofensiva del jugador, podemos saber que el 75% de los jugadores está por debajo del 0.8, por lo que un jugador con un valor mayor a eso debe ser considerado como una verdadera amenaza en el costado ofensivo.
- Si hablamos de variables como AST%, STL%, BLK%, TOV%, son muy útiles a la hora de comparar jugadores entre el mismo equipo, ya que son ponderadas por el tiempo que cada jugador estuvo en cancha. Por esto sería útil diferenciarlas por equipo, como se hace en el archivo de Power Bi y como se demuestra a continuación. Con la diapositiva extraída del archivo Power Bi a entregar, se puede hacer esta distinción por equipo y por partido, para hacer tanto “zoom in” en los datos como sea necesario. Como puede verse, Patricio Giralt es el líder en %BLK, Nicolas Stenta en %STL (ROB), Dylan Bordon en %AST y Lee Ayalyia en %TOV (TO), lo cuál es negativo.

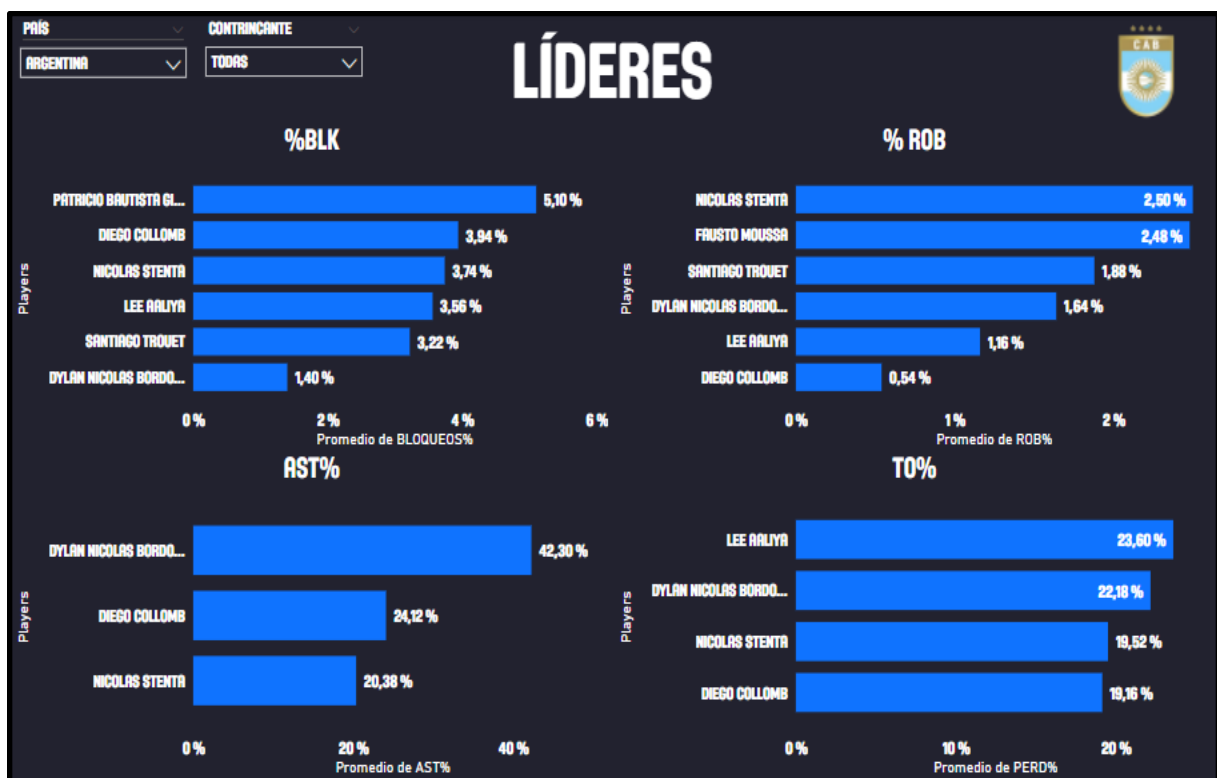


Figura 3: Análisis del %BLK, %ROB, %AST y %TO para Argentina

El análisis de la **Figura 3** será repetido tantas veces como sea necesario y estará incorporado en la solución. Para hacerlo se pueden visualizar gráficos del tipo de la **Figura 4** a continuación. Lo que se muestra en esta son las mismas variables que la **Figura 3** pero de todo el torneo. Además se agrega una línea que hace referencia al valor promedio de cada una de las variables, permitiendo hacer el análisis hecho previamente de manera mucho más veloz y simple.

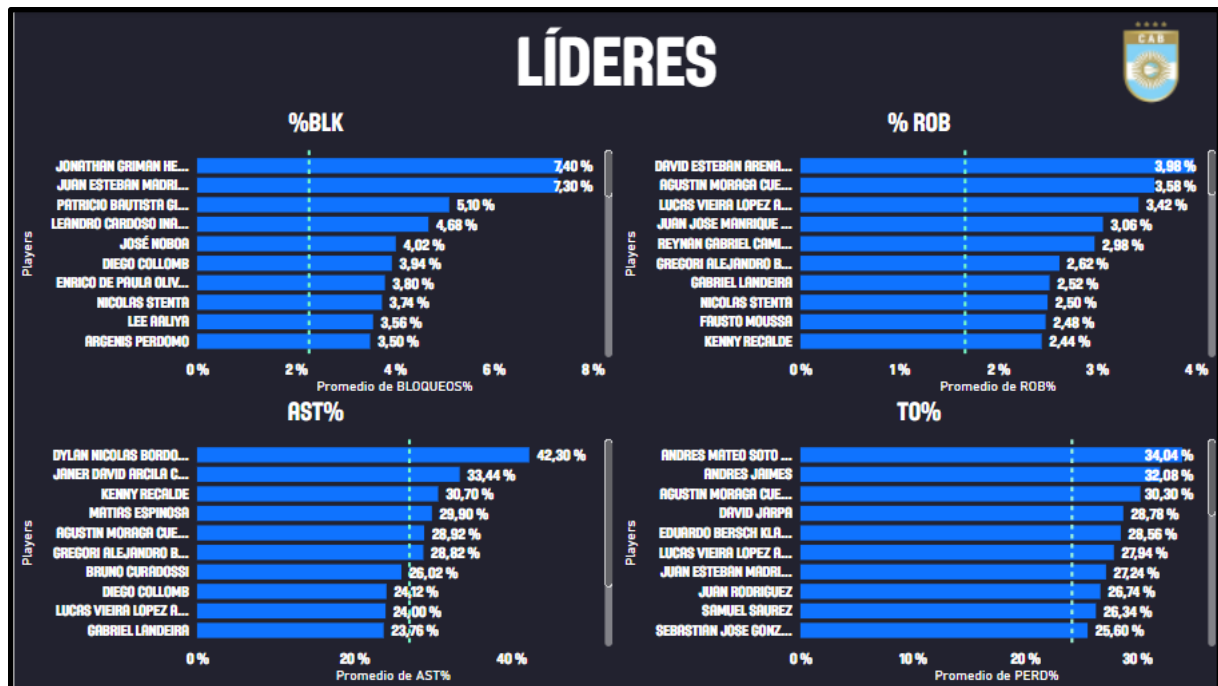


Figura 4: Análisis del %BLK, %ROB, %AST y %TO para todo el torneo

MISSINGS

Como se dijo previamente, uno de los pasos del proyecto es corroborar que los datos sean correctos. Es por esto, que debe confirmarse que no hayan valores faltantes en ninguna de las bases. Haciendo uso de la librería “missingno” pudo visualizarse este gráfico que presenta cuántos y dónde están los valores “missings” en una base.

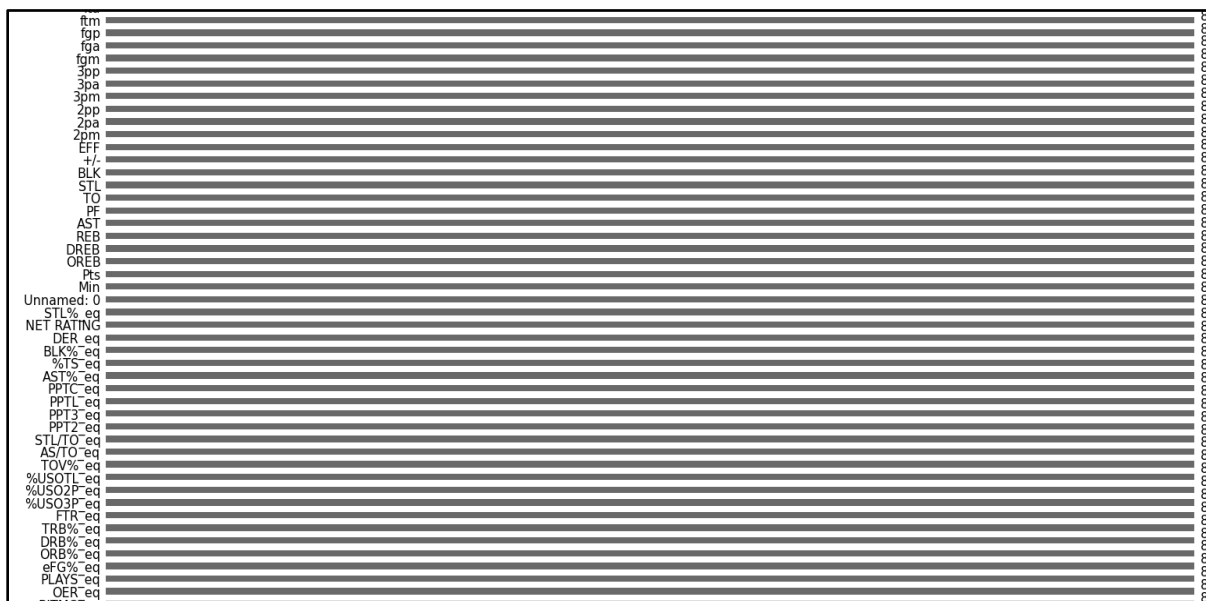


Figura 5: Descripción de Outliers de Datos Individuales

Si bien no se aprecia en detalle en la **Figura 5**, queda confirmado que la base no presenta valores faltantes.

Es importante resaltar que el scraping es una parte muy relevante del proyecto. Debido a la modalidad en la que los datos son disponibilizados, el hecho de tener la posibilidad de realizar la búsqueda en internet y poder hacer uso de los datos de manera casi instantánea, representa un diferencial y abre un gran abanico de posibilidades en cuanto a análisis de los mismos. Este diferencial se acrecienta por el ámbito en el que se necesitan los datos, ya que quienes harán uso de estos, no suelen ser conocedores de estas tecnologías.

En este proceso de scraping y disponibilidad de los datos, se realiza la limpieza de los mismos y la preparación de las bases que luego serán utilizadas. Es por esto que puede afirmarse con cierta confianza que los datos, a menos que ocurra un imprevisto, serán correctos, y no presentarán valores faltantes ni nulos.

OUTLIERS

En cuanto al tratamiento de outliers, se llevó a cabo la siguiente metodología:

1. Análisis de Gráficos de caja y bigote.
2. De encontrarse valores atípicos en valores considerados “estadísticas básicas”, se procedió a corroborar los datos con los de la página web.
3. De encontrarse valores atípicos en valores considerados “estadísticas avanzadas”, se procedió a corroborar las fórmulas en el código de python, comparándolas con el glosario de estadísticas disponibilizado por el cliente.
4. Si los datos y/o fórmulas fueran correctas, los valores atípicos se mantuvieron como estaban, ya que caracterizan al jugador tanto como los no atípicos.

A continuación pueden observarse algunos ejemplos de variables y su distribución de los datos para cada uno de los partidos.

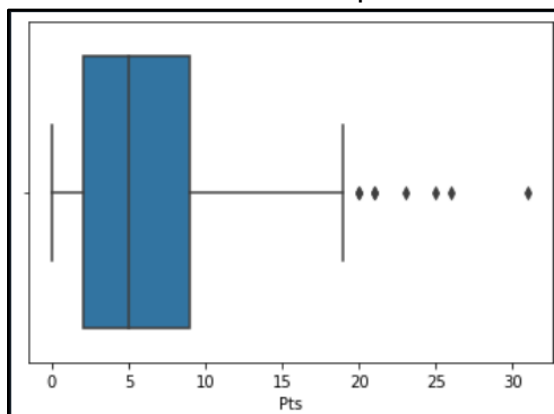


Figura 6: Caja y Bigotes de Puntos

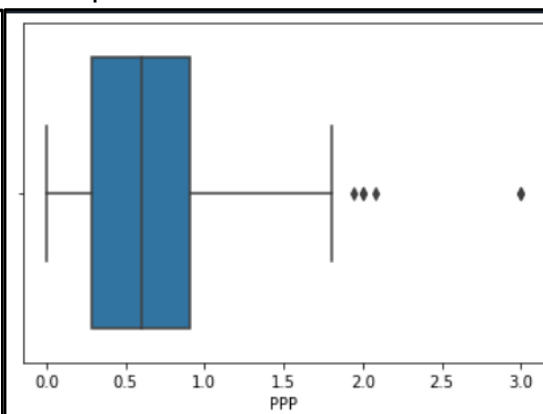


Figura 7: Caja y Bigotes de PPP

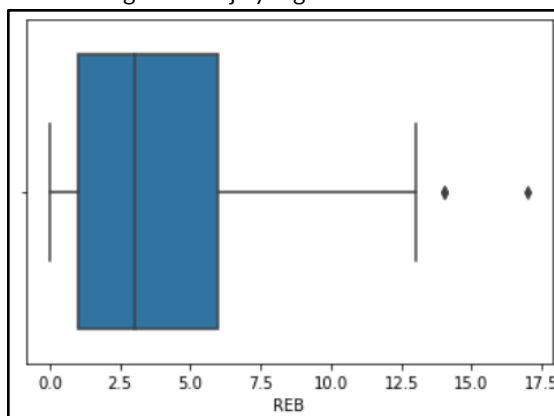


Figura 8: Caja y Bigotes de Rebotes

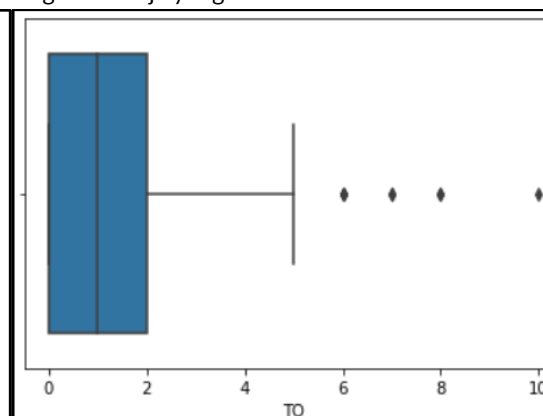


Figura 9: Caja y Bigotes de Pérdidas

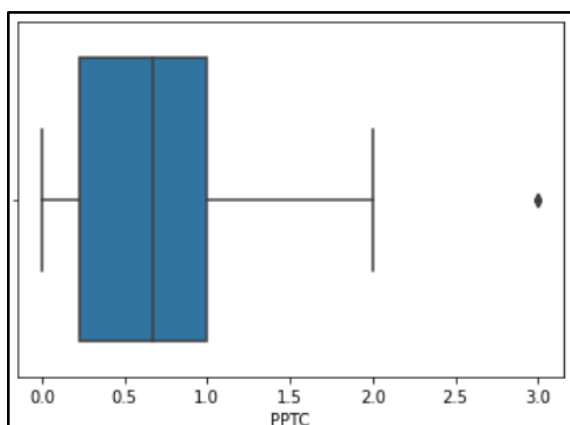


Figura 10: Caja y Bigotes de Puntos por Tiro de Campo

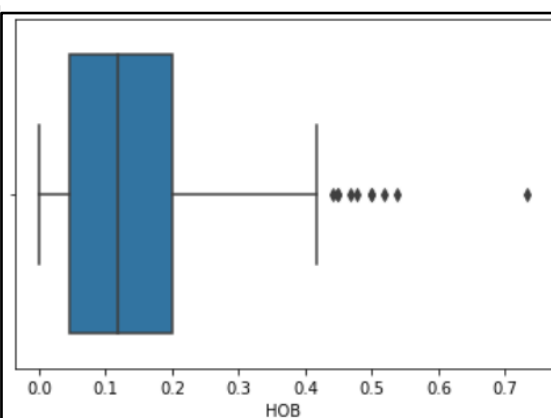


Figura 11: Caja y Bigotes de HOB

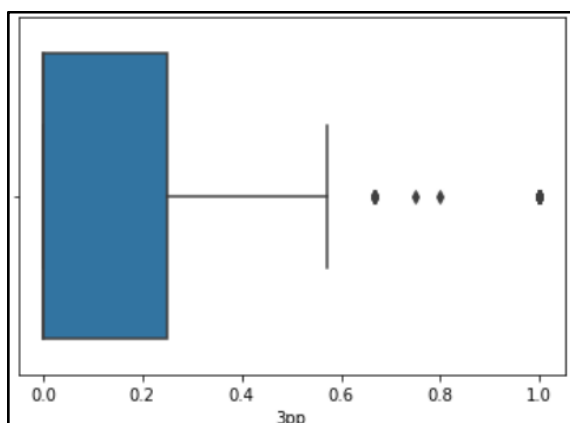


Figura 12: Caja y Bigotes de Porcentaje de 3

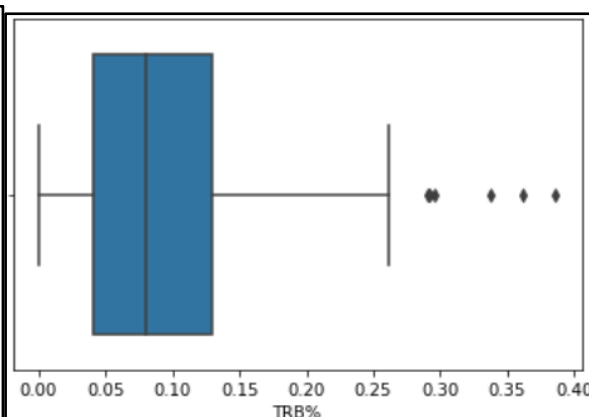


Figura 13: Caja y Bigotes de Porcentaje de Rebotes Totales

Como puede observarse en estos casos y para el resto, hay valores que se consideran atípicos. Sin embargo, no hubo ningún inconveniente con la extracción de los datos ni con las fórmulas del archivo de python, por lo que, como se definió en la metodología a implementar, no se eliminó ni se modificó ningún dato.

CORRELACIONES

En este apartado se realizará un análisis de la correlación entre las variables que definen al jugador. De esta manera se podrán identificar 2 datos altamente relevantes:

- Por un lado, conocer cuáles son las variables con una alta correlación, para poder evitar que ensucien el proceso de segmentación de datos, a desarrollar más adelante en el proceso.
- Por otro lado, es relevante conocer la correlación entre las variables para comprender el comportamiento de las mismas y su relación con el juego. Esto quiere decir, por ejemplo, detectar que existe una alta correlación inversa entre la cantidad de minutos de cada jugador, con el porcentaje de tiros de campo, puede dar a conocer que los jugadores están siendo exigidos más de la cuenta.

La correlación entre variables puede observarse en la **Figura 14**:

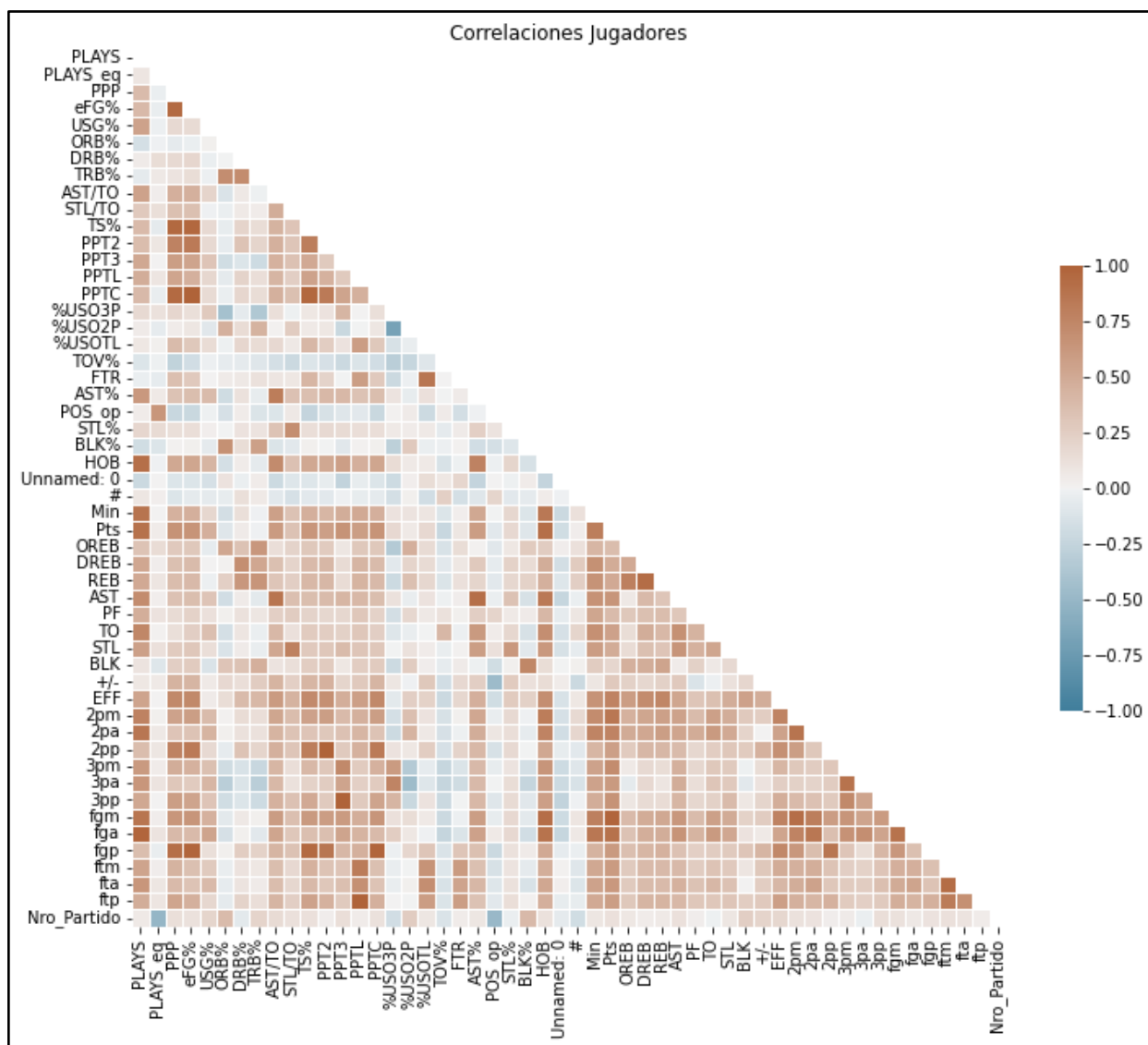


Figura 14: Tabla de correlaciones entre variables

Algunas de las conclusiones obtenidas son las siguientes:

- El rebote defensivo tiene una mayor correlación con el rebote total que el rebote ofensivo. Esto puede deberse a que en teoría es mayor la cantidad de rebotes defensivos de un jugador que la de rebotes ofensivos.
- Puntos está altamente correlacionado con Plays y con HOB, indicando que mientras más participa el jugador mayor será la cantidad de puntos anotados.
- Hay una alta correlación entre PPP y eFG%, lo cual tiene sentido porque ambos refieren a la eficiencia del jugador en ofensiva.
- Sorpresivamente hay una relación de 0.77 entre el porcentaje de tiro de 3 y la cantidad de tiros de 3 intentados. Si bien este valor no refiere una correlación perfecta, da la pauta para un posible análisis entre estas, ya que si se trasladara al juego se podría decir que mientras más se tire de 3, mejores porcentajes se obtendrán. A su vez, hay un 0.65 de

correlación entre este porcentaje y la cantidad de puntos, definiendo que en teoría, mientras más se tire de 3 puntos, más puntos se anotarán.

- Un jugador puede tomar la decisión de hacer las siguientes jugadas tiro de 2 puntos (%USO2P), tiro de 3 puntos (%USO3P), asistencia (%AST), tiro libre (%USOTL) o pérdida (%TOV). Las variables de %USO2P y %USO3P tienen una correlación de -0.68, definiendo que las 2 tienen la mayor probabilidad de ocurrencia.
- Se ve una correlación negativa entre +/- y %TOV, así como entre +/- y TO. Esto confirma la creencia que mientras más pérdidas tenga un jugador, menos probabilidades de alcanzar la victoria tendrá.
- AST (asistencias) y %AST (porcentaje de asistencias) tienen una alta correlación.
- Se puede observar una correlación negativa entre las variables relativas al tiro de 3 puntos (3pa, 3pp y 3pm) y las variables relativas a los rebotes (ORB%,DRB%,TRB%). Este insight permite caracterizar a los jugadores de manera más correcta, diferenciando al tipo de jugador especializado en rebotar el balón y el tipo que es encargado de lanzar el balón desde más allá de la línea de 3.

ESTADÍSTICOS DESCRIPTIVOS

Este apartado es uno de los más relevantes y va de la mano con las conclusiones obtenidas en el apartado de “Estructura de la Base”. Esto se debe a que este análisis tampoco será llevado a cabo una única vez con el fin de definir cuál será la base en el modelo de clustering y optimización desarrollado más adelante en este informe, sino que será una de las partes más importantes de la solución y estará presente partido a partido.

Una de las ventajas de esta solución es permitir al cliente realizar un análisis de la información, lo más rápido y simple posible. Es por esto que uno de los entregables del proyecto consiste en un archivo de Power BI que permita conseguir este objetivo. Este archivo tiene múltiples pestañas dónde se puede visualizar distintos tipos de datos. A continuación se hará un resumen de las pestañas disponibles en relación a las bases de jugadores:

- Análisis Individual - 1 pestaña por jugador
- Pelotas Individual - Para un equipo específico y total del torneo
- Uso del Balón - Para un equipo específico y total del torneo
- Efectividad Individual - Para un equipo específico y total del torneo
- Rebotes Individual - Para un equipo específico y total del torneo
- Puntos por tiro - Para un equipo específico y total del torneo

- Líderes Efectividad - Para un equipo específico y total del torneo
- Líderes All Around - Para un equipo específico y total del torneo
- Líderes Rebotes - Para un equipo específico y total del torneo
- Líderes Puntos por Tiro - Para un equipo específico y total del torneo

Este análisis permite conocer cuáles son los puntos dónde se debe mejorar, cuáles son los jugadores o equipos con mejor rendimiento, cómo se comportan los equipos frente a distintos rivales, cuál es la evolución de los equipos o jugador a lo largo de cada torneo, entre otras cosas.

Algunos de los puntos claves que se pueden obtener de la base son:

- Ignacio Ortega Naumchik es el que mejor porcentaje de True Shooting con un 84,22%, haciéndolo el jugador más eficiente en lanzamientos de campo. Por el contrario, Jesus Andrés Caraballo Martinez, Oscar Moya, Rommel Cevallos y Santiago Moreno son los peores con un 0% de True Shooting.
- Llamativamente, Santiago Moreno es el que mejor porcentaje de Robos tiene, con un 7,03%. Este insight es muy relevante a la hora de definir al jugador.
- A su vez, Jesus Andrés Caraballo Martinez es el que mejor porcentaje de Bloqueos tiene, con un 14,65%, mientras que el segundo en orden tiene un 7,78%.
- La alineación Argentina tiene al jugador con mayor porcentaje de asistencias. Dylan Bordon ocupa este lugar con un 42,30%.
- La mayor amenaza en el costado defensivo es Jonhatan Herrera, ya que es el primero en PPP con 1,18 puntos producidos por jugada, y un eFG% de 61,70%. Sin embargo, en la segunda y tercera posición de la columna de PPP y en la tercera de eFG% encontramos a 3 jugadores brasileños, lo cual podría tener una buena relación con que este equipo haya sido campeón en la competición. Para la plantilla argentina, el mejor en ambos ámbitos fue Fausto Mussa, en la quinta posición de PPP y en la segunda posición de eFG%.
- Los 3 mejores reboteadores del torneo fueron: Silvio Neto (Brasil), Enrico Oliveira (Brasil) y Patricio Giralt (Argentina). Se podría formular la hipótesis de que una de las variables esenciales a tener en cuenta si se quiere lograr la victoria es la de los rebotes, ya que los 3 jugadores pertenecen a las selecciones que se encontraron en la final.
- El mayor anotador del torneo con 16 puntos por partido fue Joseph Caicedo, seguido por Dylan Bordon (Argentina), Janer Castillo y Reinan Dos Santos con 15 puntos.

Es importante destacar que para obtener todas estas conclusiones se utilizaron distintos filtros que permiten evitar la descontextualización de los datos. Esto quiere decir, por ejemplo, que un jugador que jugó únicamente un partido y anotó el único tiro que lanzó, no sea considerado como un mejor lanzador que alguien con un promedio

de 65% de tiros de campo y 10 lanzamientos promedio en 20 minutos jugados por partido. Estas distintas segregaciones fueron definidas en conjunto con el cliente en distintas reuniones. Sin embargo son modificables en la plataforma Power BI, en relación a la competición a analizar.

EQUIPOS ESTADÍSTICAS

Para lo que respecta a las estadísticas de equipo, el desarrollo será mucho más limitado, debido a que la parte de segregación de jugadores y de maximización de variables de la solución no incluyen estas bases, por la poca cantidad de equipos que participaron del torneo.

Es por esto que nos limitaremos a la descripción de las bases y no al planteo de hipótesis.

ESTRUCTURA - ESTADÍSTICOS DESCRIPTIVOS

En relación a las bases de totales, se respetan las mismas consideraciones que para los individuales.

Ambas bases de totales están compuestas por 40 registros, que hacen referencia a 8 equipos, los cuales participaron del torneo. Hay 25 variables consideradas "Estadísticas Avanzadas" y 24 consideradas "Estadísticas Básicas". Con el fin de facilitar el análisis se hizo un "join" entre ambas bases, por lo que se puede decir que la base de "Totales" contiene 52 variables que definen el juego de cada equipo.

En cuanto al tipo de dato, son todos datos numéricos excepto la columna "País" y "País_Contrincante" que contiene el nombre de cada equipo y contra quién jugó.

A continuación se puede visualizar una imagen del resumen de la base:

#	Column	Non-Null Count	Dtype
0	Players	96 non-null	object
1	PLAYS	96 non-null	float64
2	PLAYS_eq	96 non-null	float64
3	PPP	96 non-null	float64
4	eFG%	96 non-null	float64
5	USG%	96 non-null	float64
6	ORB%	96 non-null	float64
7	DRB%	96 non-null	float64
8	TRB%	96 non-null	float64
9	AST/TO	96 non-null	float64
10	STL/TO	96 non-null	float64
11	TS%	96 non-null	float64
12	PPT2	96 non-null	float64
13	PPT3	96 non-null	float64
14	PPTL	96 non-null	float64
15	PPTC	96 non-null	float64
16	%USO3P	96 non-null	float64
17	%USO2P	96 non-null	float64
18	%USOTL	96 non-null	float64
19	TOV%	96 non-null	float64
20	FTR	96 non-null	float64
21	AST%	96 non-null	float64
22	POS_op	96 non-null	float64
23	STL%	96 non-null	float64
24	BLK%	96 non-null	float64
25	HOB	96 non-null	float64
26	Unnamed: 0	96 non-null	int64
27	#	96 non-null	int64
28	Min	96 non-null	float64
29	Pts	96 non-null	float64
30	OREB	96 non-null	float64
31	DREB	96 non-null	float64
32	REB	96 non-null	float64
33	AST	96 non-null	float64
34	PF	96 non-null	float64
35	TO	96 non-null	float64
36	STL	96 non-null	float64
37	BLK	96 non-null	float64
38	+/-	96 non-null	float64
39	EFF	96 non-null	float64
40	2pm	96 non-null	float64
41	2pa	96 non-null	float64
42	2pp	96 non-null	float64
43	3pm	96 non-null	float64
44	3pa	96 non-null	float64
45	3pp	96 non-null	float64
46	fgm	96 non-null	float64
47	fga	96 non-null	float64
48	fgp	96 non-null	float64
49	ftm	96 non-null	float64
50	fta	96 non-null	float64
51	ftp	96 non-null	float64

Figura 15: Estructura de la base

Como puede observarse en la **Figura 15** todos los datos son numéricos.

A continuación se realizó un resumen de las variables de esta base. Al igual que en el caso de las variables de cada jugador, las de los totales son muy importantes a la hora de definir objetivos. Esto se debe a las mismas razones que las variables individuales pero de manera más general. Por ejemplo, si se conoce que un equipo tiene un muy buen ritmo de juego, se planteará el partido de distinta manera que si es un equipo con un ritmo más lento. A su vez, estas variables son muy relevantes a la hora de reconocer problemas, como en el caso de un equipo con muchas pérdidas.

Variable	mean	min	50%	75%	max
POS_eq	79,124	74,24	77,584	81,592	85,464
RITMO_eq	78,6099	74,4	77,382	81,0989	85,472
OER_eq	0,819575	0,6706	0,7813	0,8997	1,0654

PLAYS_eq	95,424	89,04	94,044	98,33	103,864
eFG%_eq	0,39055	0,3208	0,3821	0,4171	0,476
ORB%_eq	0,343425	0,2804	0,3243	0,3495	0,5098
DRB%_eq	0,656575	0,6178	0,6467	0,6578	0,7472
TRB%_eq	0,5	0,449	0,4812	0,5083	0,6336
FTR_eq	0,159925	0,0908	0,1444	0,2117	0,258
%USO3P_eq	0,24555	0,1976	0,2412	0,2749	0,2924
%USO2P_eq	0,481	0,4268	0,4849	0,50685	0,5302
%USOTL_eq	0,084025	0,0572	0,0744	0,1045	0,1228
TOV%_eq	0,189325	0,1686	0,1895	0,2004	0,217
AS/TO_eq	0,72995	0,477	0,7356	0,8871	1,0266
STL/TO_eq	0,557125	0,4306	0,5698	0,6087	0,68
PPT2_eq	0,806175	0,6298	0,7923	0,87445	0,9836
PPT3_eq	0,743025	0,6568	0,7021	0,78525	0,932
PPTL_eq	1,324	1,0832	1,2817	1,4261	1,5658
PPTC_eq	0,78095	0,6418	0,7638	0,83385	0,9522
AST%_eq	0,506	0,4324	0,4989	0,54525	0,5944
%TS_eq	0,4179	0,3434	0,4131	0,4576	0,502
BLK%_eq	0,077725	0,0368	0,063	0,0991	0,141

DER_eq	0,819575	0,578	0,8393	0,87935	0,9702
NET RATING	1,73E-17	-0,2818	-0,0755	0,1159	0,4874
STL%_eq	0,099325	0,0784	0,1043	0,1111	0,12
Min	201,25	200	200	201,25	205
Pts	64,8	51,6	63,3	69,25	82,2
OREB	16,3	14	14,7	16,3	24,6
DREB	31,7	28	30,9	32,55	39,2
REB	48	43,6	46,2	46,7	63,8
AST	12,475	9	13,2	14,6	15,8
PF	17,025	14	17,5	17,95	21,6
TO	18,1	15,8	18,8	19,05	19,6
STL	9,5	7,2	10,1	10,7	11,8
BLK	5,25	2,4	4,2	6,9	9
+/-	1,11E-16	-24,6	-5,9	8,55	39,2
EFF	64,5	40,2	60,6	79,8	102
2pm	18,4	14,4	18,8	20,5	21,6
2pa	45,85	39,8	45,3	48,35	55
2pp	0,40306	0,314866	0,396106	0,437223	0,491831
3pm	5,725	4,8	5,6	5,85	8

3pa	23,4	18,4	22,9	25,4	29,6
3pp	0,247687	0,218841	0,234055	0,261769	0,310735
fgm	24,125	20	23,9	26,15	29,6
fga	69,25	62,4	69,3	72,95	77,4
fgp	0,348878	0,288501	0,342289	0,376572	0,429322
ftm	10,825	6	9,6	14,55	15,6
fta	18,35	12	15,9	22,6	26,4
ftp	0,582562	0,476667	0,563908	0,627423	0,689032

Tabla 3: Análisis estadístico de la base de Equipos

Si se hace referencia a las tablas de totales, puede decirse que los valores son relativamente más parejos entre equipos. Puede verse que no hay tanta variación en los datos. De esta manera, un buen objetivo a plantearse sería estar por encima de la media en cada una de estas variables.

Eso no quita que haya estadísticas que hay que tener en cuenta. Por ejemplo:

- El país con mayor eficiencia (EFF) tiene 102 puntos, 37,5 puntos por encima del promedio. Eso significa o que ese equipo tiene pocas veces las pelotas y siempre convierten o que dominan el juego y siempre convierten, ambas ideas que hay que tener en cuenta a la hora de prepararse para ese equipo.
- Otra métrica a tener en cuenta es la del máximo valor de rebotes, con más de 15 rebotes más que el promedio de los equipos.
- Una de las métricas más relevantes, por su relación con el resultado final de cada partido, es el de +/- . Podemos ver que el 75% de los equipos tiene un valor menor a 9 de +/- , mientras que el mayor tiene casi 40. Esta clara diferencia se puede ver relacionada en el resultado del torneo ya que ese valor pertenece al campeón, Brasil.
- Puede verse que en %BLK hay una buena diferencia entre el 75% de los equipos y el mayor. En la parte de visualización de la solución se puede apreciar esto de manera más clara. A continuación pueden verse las mismas variables que se mostraron en el segmento de jugadores, pero para los equipos. En amarillo se puede reconocer al equipo argentino.

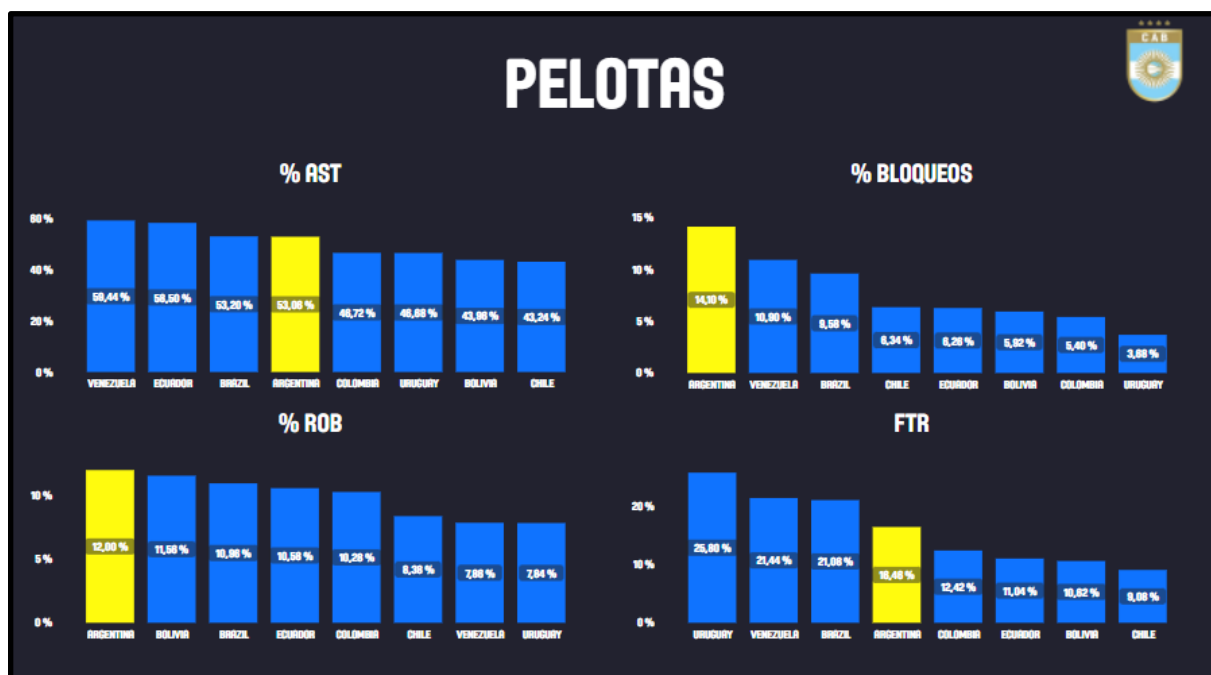


Figura 16: Análisis del %BLK, %ROB, %AST y %TO para el total de los equipos

- Otras variables donde el equipo argentino está primero son las relacionadas con la eficacia de tiro, definiendo al equipo argentino como un equipo que cuida mucho la selección de tiro. Esto va de la mano con que Argentina tiene los menores valores en Ritmo y PLAYS, con 74 y 89 respectivamente, lo que confirma que el equipo tiene un ritmo lento con jugadas largas y pases de balón. Esto puede verse en la siguiente imagen



Figura 17: Análisis del eFG%, PPT2, PPT3, PPTL y PPTC para el total de los equipos

- El campeón, Brasil tiene los mejores resultados en Eficiencia Ofensiva (OER) y los 3 tipos de rebotes. Así como el segundo puesto en eFG%, PPT2, PPT3 y PPTC. Esto permite postular a estas variables como las más importantes en pos de una victoria.

OUTLIERS

En cuanto al tratamiento de outliers, se llevó a cabo la misma metodología que para la base de Jugadores. De la misma manera, no fue necesario eliminar ningún valor.

CORRELACIONES

En este apartado se realizará un análisis de la correlación entre las variables que definen al equipo. El objetivo es el mismo que en el análisis de la tabla de Jugadores pero sin el apartado de quitar variables para el próximo modelado.

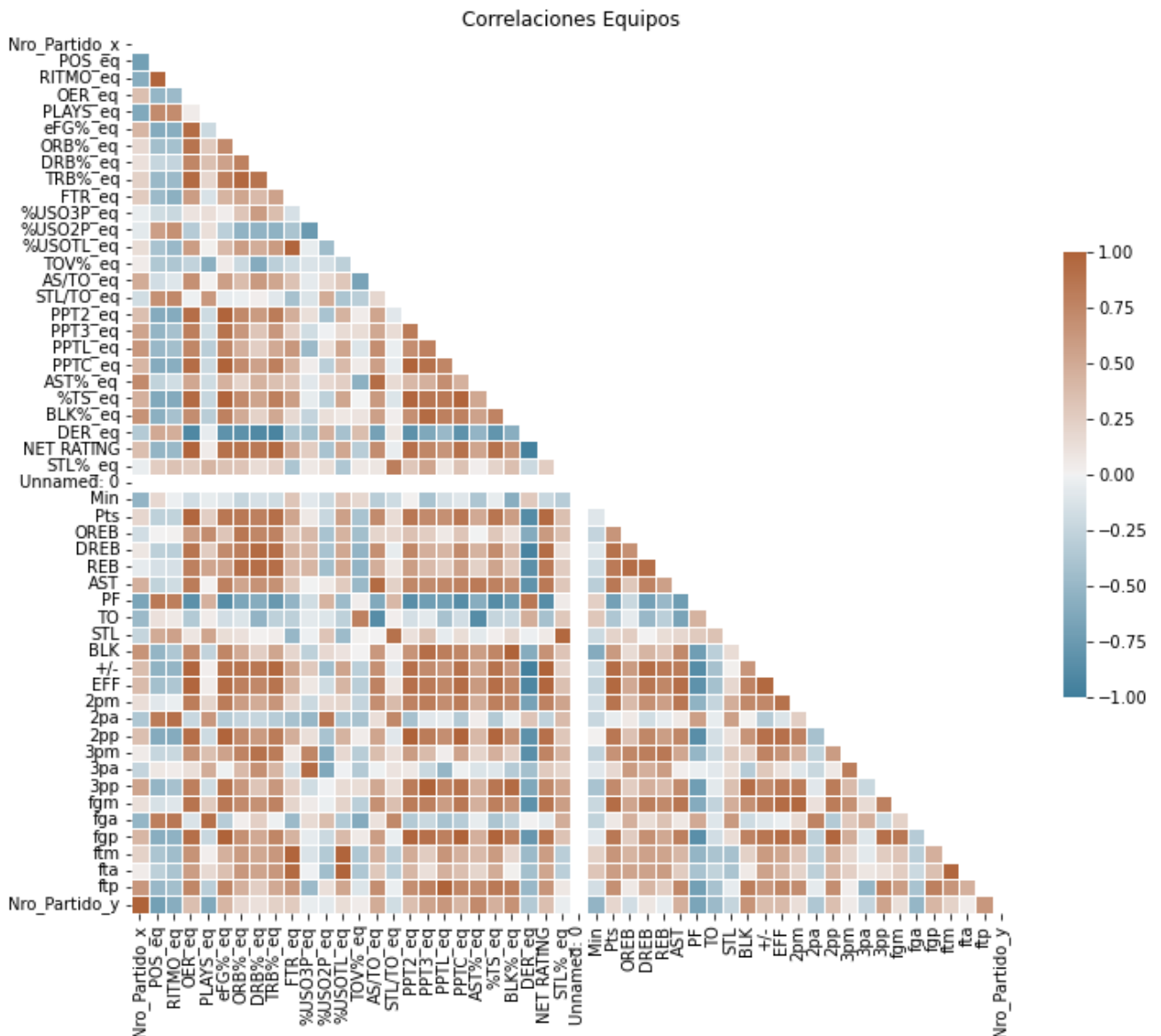


Figura 18: Matriz de correlación

Algunas de las conclusiones de este análisis son:

- Las relaciones están mucho más marcadas que en la base de Jugadores. Esto puede deberse a la gran diferencia de registros que se encuentran entre las bases.
- La eficiencia defensiva tiene una clara correlación indirecta con muchas de las variables de tiro y con los puntos, dando a conocer que una mientras más se mejore la ofensiva más se puede descuidar a la defensiva.
- Existe una correlación casi perfecta entre el Net Rating y las variables de porcentajes de rebotes.
- Se respetan algunas de las características de los Jugadores como la de Uso de 2 puntos y de 3.

PLANTEO DE HIPÓTESIS

Habiendo analizado y visualizado la información presentada en los puntos anteriores, se propondrán algunas posibles hipótesis.

Como se demostró al ver algunas de las distribuciones de las variables de los jugadores, no todos los jugadores son iguales. Si bien eso es evidente, es importante recalcar que si puede identificarse de mejor manera qué jugador debería jugar en cada posición, será posible mejorar el resultado de nuestros partidos.

Es importante saber que el básquet es un deporte basado en posiciones. Así se definió hace más de 100 años y así sigue siendo. Sin embargo, con el correr del tiempo, a medida que el juego fue cambiando, las reglas cambiaron, las capacidades atléticas también, las 5 posiciones, planteadas en base a características físicas básicas como altura y peso, fueron mutando.

Existen equipos basados en la defensa y el ataque rápido que se basan en velocidad, equipos con alta capacidad de anotar que no tienen mucha altura, y equipos que tienen un juego más lento con gran altura y mucha más eficiencia. Estas son algunas de las características nuevas presentes en el juego de los últimos años.

Es por esto que es importante definir a los jugadores por sus capacidades de juego y no por las físicas, y así determinar según el entrenador, cuál es la mejor posición para cada jugador, y cómo hacer que cada uno se complemente de la mejor manera.

Así es como se planteó la hipótesis de que *no se está jugando con los jugadores óptimos para cada posición en cada momento*.

Sin embargo, esto no es precisamente fácil de medir y si bien hay muchas estadísticas para comparar, es relevante también si estos cambios impactan de manera significativa al desempeño del equipo.

De ahí surge la idea de diseñar rankings para las estadísticas. Se diseñó una tabla que en vez de tener las estadísticas para cada equipo, tiene el ranking comparativo de esa estadística contra el resto de los equipos. De esta manera, si un equipo es el equipo con más puntos, en vez de presentar la cantidad de puntos, se presenta la posición en la que se encuentra con respecto al resto, en este caso, en la primera posición.

Esto se ve reflejado de manera más clara en la **Tabla 4**.

Variable	Argentina	Bolivia	Brasil	Chile	Colombia	Ecuador	Uruguay	Venezuela
POS_eq	8	1	5	7	2	3	6	4
RITMO_eq	8	1	5	6	2	3	7	4
OER_eq	2	7	1	8	4	6	5	3
PLAYS_eq	8	1	2	7	3	4	5	6
eFG%_eq	1	8	2	7	4	6	5	3
ORB%_eq	2	5	1	6	7	8	3	4
DRB%_eq	6	8	1	7	4	2	5	3
TRB%_eq	2	8	1	7	5	6	4	3
FTR_eq	4	7	3	8	5	6	1	2
%USO3P_eq	6	7	1	2	5	3	4	8
%USO2P_eq	5	1	8	6	3	4	7	2
%USOTL_eq	4	6	2	8	5	7	1	3
TOV%_eq	1	5	8	2	4	6	3	7
AS/TO_eq	4	7	2	8	5	3	6	1
STL/TO_eq	4	1	5	8	3	2	7	6
PPT2_eq	1	8	2	7	3	6	5	4
PPT3_eq	1	5	2	7	4	6	8	3
PPTL_eq	2	7	3	8	6	5	4	1

PPTC_eq	1	8	2	7	4	6	5	3
AST%_eq	4	7	3	8	5	2	6	1
%TS_eq	1	8	2	7	4	6	5	3
BLK%_eq	1	6	3	4	7	5	8	2
DER_eq	7	1	8	3	4	5	2	6
NET RATING	2	8	1	7	4	5	6	3
STL%_eq	1	2	3	6	5	4	8	7
Min	3	3	3	3	1	3	1	3
Pts	2	7	1	8	4	6	5	3
OREB	4	2	1	6	8	6	3	5
DREB	4	8	1	7	3	5	6	2
REB	5	3	1	8	3	7	6	2
AST	3	6	1	8	5	4	7	2
PF	7	1	8	5	2	3	4	6
TO	2	1	6	3	4	7	4	8
STL	2	1	5	6	3	4	8	7
BLK	1	4	3	4	7	4	8	2
+/-	2	8	1	7	4	5	6	3
EFF	2	7	1	8	4	5	6	3

2pm	2	5	1	8	3	6	7	4
2pa	6	1	5	7	3	4	8	2
2pp	1	8	2	7	3	6	5	4
3pm	3	6	1	4	4	2	6	8
3pa	7	6	1	3	4	2	5	8
3pp	1	5	2	7	4	6	8	3
fgm	2	6	1	7	3	5	7	4
fga	7	1	2	6	4	3	8	5
fgp	1	8	2	7	4	6	5	3
ftm	4	6	2	8	5	6	1	3
fta	4	6	2	8	5	7	1	3
ftp	2	7	3	8	6	5	4	1

Tabla 4: Rankings de los países

En la **Tabla 4** se puede ver que Argentina está segundo en puntos, quinto en rebotes o primero en bloqueos. En la **Figura 19** puede visualizarse cuáles variables son las que requieren mejoras.

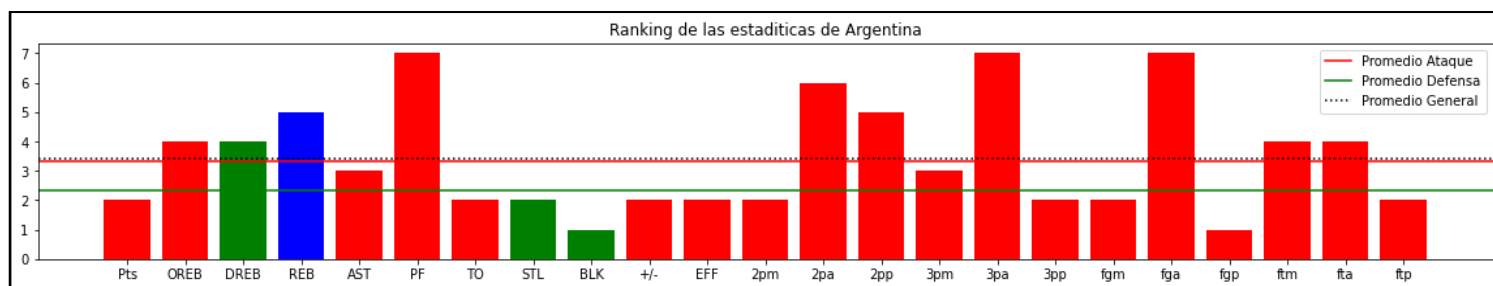


Figura 19: Ranking de Argentina

Se pueden observar todas las posiciones de estadísticas básicas para el equipo de Argentina. En rojo están las variables de ataque, en verde las de

defensa y en azul la única variable general. Por encima, se muestran los promedios de los rankings de ataque, de defensa y el general.

Este será uno de los mecanismos que se utilizarán para poder medir el impacto de los cambios en el equipo. Al mismo tiempo permitirá detectar insights de manera eficaz y veloz. Sin mucho esfuerzo, puede identificarse que en PF (Personal Fouls), 3pa (Three Pointers Attempted) y FGA (Field Goals Attempted) la Argentina está séptima, es decir anteúltima. Esto permite asumir que es un equipo que prioriza la selección de buenos tiros, por sobre el ritmo de juego. Sin embargo, si también se observa el ranking en TO (pérdidas de balón) puede llegarse a la conclusión que el equipo está teniendo problemas en el cuidado del balón, desperdiciando posesiones y así intentando menos tiros de campo (FGA).

De esta manera, mediante el análisis de las estadísticas individuales se categorizará a los jugadores, y teniendo en cuenta esa información se procederá a buscar maximizar ciertas variables, comparando cuál es el impacto que tiene esta maximización en los rankings del equipo.

ENFOQUE DE SOLUCIÓN

Como fue mencionado anteriormente en el apartado **OBJETIVO DEL PROYECTO**, el objetivo analítico para este proyecto está relacionado con la estructura de posiciones en el basketball y las complejidades a la hora de encasillar a todos los jugadores en 5 posiciones fijas. La propuesta es determinar como una metodología de clusterización puede resultar en presentar posiciones diferentes para los jugadores, buscando así una manera más dinámica de entender cómo debería jugar cada jugador.

Para poder realizar este análisis, se atravesará por varias etapas:

1. Etapa de Filtrado:

Debido a la cantidad de variables que se correlacionan entre sí, se realizará un filtrado de las variables utilizadas. Esto apunta a proveer mejores resultados ya que el modelo tendrá que ajustarse a una menor cantidad de variables, permitiendo así además una interpretación más sencilla de los datos.

2. Etapa de Ataque:

Tratando de encontrar patrones de ataque en los jugadores, también se realizan clusters que solo incluyan las variables de ataque.

3. Etapa de Defensa:

Del mismo modo que para las variables de ataque, se realizarán clusters que contengan sólo las variables de defensa.

4. Etapa de armado de potenciales equipos:

Habiendo hecho todo esto, se obtendrán perfiles para poder determinar un estilo de juego para cada uno de los jugadores. Haciendo uso de estos perfiles se

buscará armar potenciales equipos dependiendo del estilo necesario. Para poder realizar esto, se hará uso de técnicas de programación lineal para poder así optimizar los recursos disponibles.

Es importante recalcar que los puntajes serán tomados en su valor original y aquel jugador que jugó más minutos seguramente posea mayor cantidad de puntos. Esto se hace así ya que se considera que la cantidad de minutos jugados y sus resultados son importantes. No es lo mismo un jugador que consistentemente juega a alto nivel que un jugador que jugó dos minutos y quizás tuvo mucha suerte. Se busca mantener esa distinción realizada por los entrenadores para no sesgar la base con potenciales golpes de suerte.

METODOLOGÍAS A IMPLEMENTAR

La principal metodología a aplicar será K-Means.

Dado que todas las variables son de tipo numérico, el algoritmo podrá desempeñarse de manera efectiva con cualquier input que sea disponibilizado. Como los datos tienen valores muy disímiles entre sí, como minutos jugados con valores cercanos a 10000 y eFG% (Effective Field Goals %) con valores de 0,14, para evitar cualquier tipo de sesgos a la hora de crear clusters se normalizarán y estandarizarán todas las variables. Si bien es un proceso simple, es altamente esencial a la hora de arrojar resultados congruentes.

SETEO DE EXPERIMENTACIÓN

A la hora de medir la bondad de ajuste de los clusters, se utilizarán varias metodologías.

Para poder determinar de manera preliminar la cantidad de clusters a crear se utilizarán metodologías tales como la del codo.

Luego, para medir qué tan bien se ajustan los grupos creados a la base, se utilizará el valor de Silhouette el cual mide qué tan bien se ajusta cada punto al grupo que se le asigna. Cuanto mayor sea dicho valor, mejor van a representar nuestros grupos a la base, y del mismo modo se va a poder representar mejor a cada jugador.

Con respecto a los equipos armados, se compara cómo se desempeñan contra los resultados actuales del equipo de Argentina. Para hacer eso se realizará una extrapolación haciendo uso de los minutos jugados por cada jugador

DESARROLLO DE LA SOLUCIÓN

CLUSTERS GENERALES

Se comienza la realización de los clusters haciendo uso de los datos agregados a lo largo de todo el torneo para cada jugador, así midiendo el desempeño de cada uno de una manera más holística. Se inició haciendo grupos que incluyeron todas las variables disponibles para cada jugador, sin embargo, al haber una gran correlación

entre algunas variables el modelo obtuvo malos resultados. Por ejemplo, la variable 2pm (Two points made) y la variable %USO2P (Estadística que mide la cantidad de jugadas de dos puntos en las cuales está involucrado el jugador) están altamente relacionadas y consumían capacidad de cómputo y solamente oscurecían o dificultan la comprensión de los grupos sin agregar valor verdadero. Cuando se terminaron de determinar cuáles son las variables que valen la pena utilizar para armar grupos, la lista fue la siguiente:

- PLAYS
- PPP
- USG%
- ORB%
- DRB%
- AST/TO
- STL/TO
- TS%
- PPT2
- PPT3
- PPTL
- %USO3P
- %USO2P
- %USOTL
- TOV%
- FTR
- AST%
- STL%
- BLK%
- HOB
- Min
- Pts
- OREB.
- DREB
- PF
- STL
- BLK
- +/-
- EFF
- 2pa:
- 2pp
- 3pa
- 3pp
- fga
- fgp

- fta
- ftp

Esto significa que las variables no utilizadas son las siguientes: 2pm, 3pm, AST, eFG%, FPT, PPTC, REB, TO, TRB%. Todas ellas se descartan por una correlación demasiado alta con alguna de las otras variables.

Continuando, haciendo uso de tanto de la metodología del codo como el Silhouette, se determinó que la cantidad de clusters que mejor se ajustaba a la base era la de 3 grupos diferentes. La metodología para la elección de grupos, consistió tanto en los valores arrojados por el análisis del método Silhouette y el método del codo, como en el nivel de detalle que se le quiso dar a cada grupo. Se podría haber determinado que la cantidad de clusters sea 5 para representar los 5 jugadores presentes en un equipo al mismo tiempo pero se prefirió que cada grupo represente de mejor manera a la base que un número arbitrario pero útil de igual manera.

En este caso, por ejemplo, el número de clusters con mejor valor en el método de Silhouette era 2. Sin embargo, se consideró que definir a los jugadores únicamente en 2 grupos carece de valor, por lo que se optó por desarrollar 3 grupos.

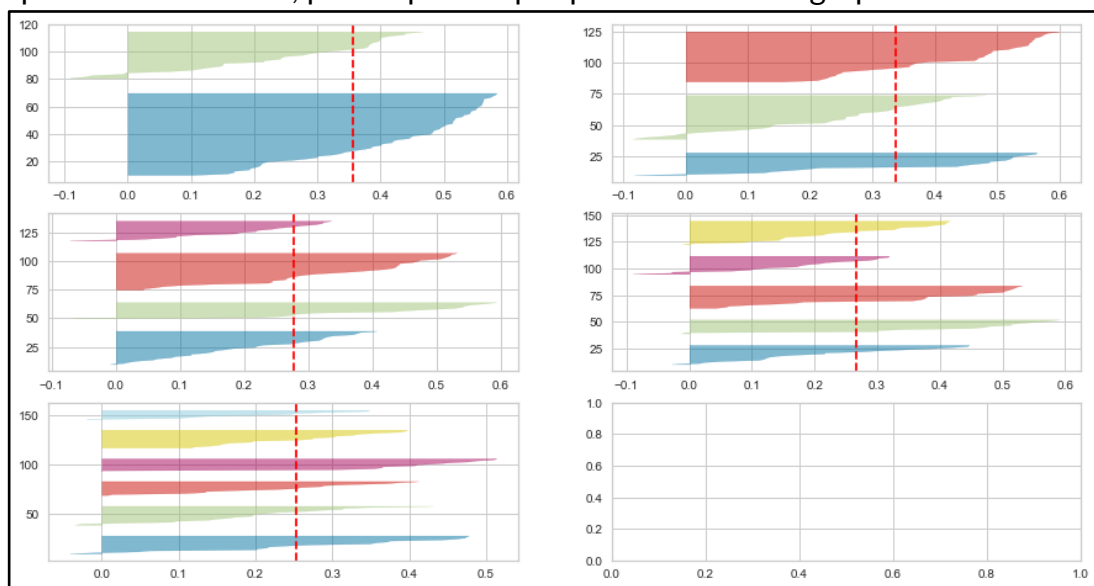


Figura 10: Silhouette para clusters de N dos a seis

El valor del método del Silhouette, fue de 0.338. Si bien está lejos de ser un valor perfecto, se considera que es un valor aceptable por la naturaleza del problema y la cantidad acotada de registros.

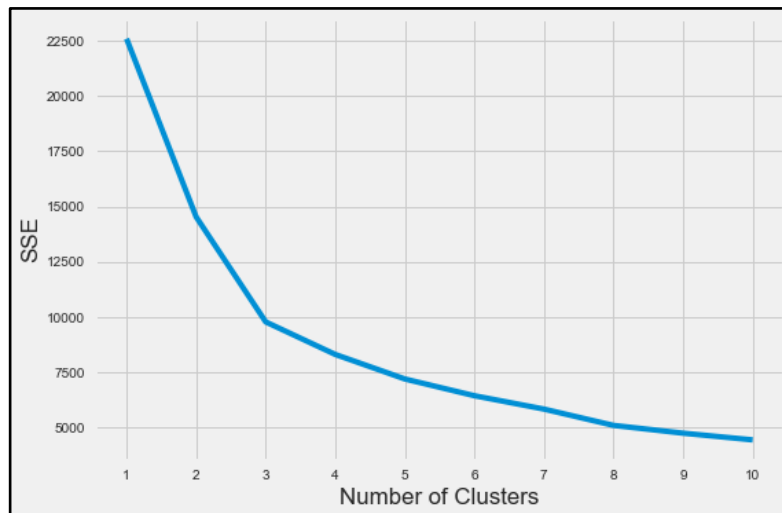


Figura 21: Gráfico de codo para el cluster general

En conjunto con esta decisión, el valor óptimo de clusters indicado por el método del codo fue de 3, lo que confirmó nuestra iniciativa.

Es importante dejar en claro que los gráficos y el análisis de los métodos demostrados previamente se repitió para cada una de las segregaciones que continúan.

Además de los métodos mencionados para la selección del número de clusters, se realizó el análisis de la distribución de los jugadores.

Para plasmar su distribución en el plano, se utilizó el método de PCA, para “condensar” la totalidad de las variables en unas únicas 2, que expliquen lo máximo posible la realidad. Habiendo utilizado ese método, las variables se graficaron sobre 2 ejes, para poder comprender la separación de los grupos.

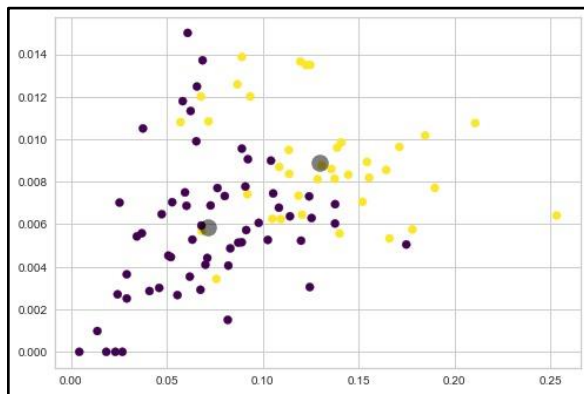


Figura 22: Separación de puntos para 2 grupos

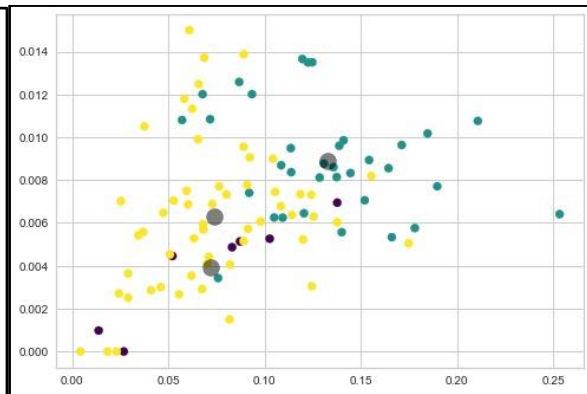


Figura 23: Separación de puntos para 3 grupos

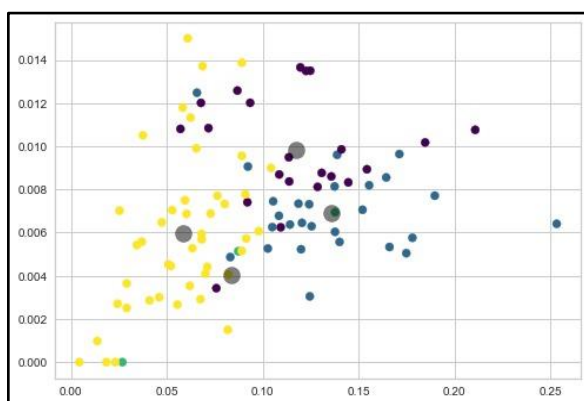


Figura 24: Separación de puntos para 4 grupos

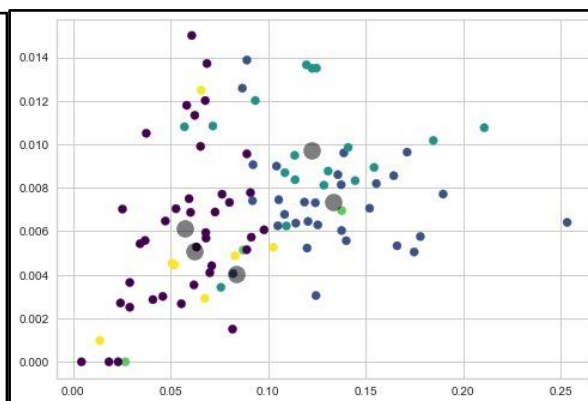


Figura 25: Separación de puntos para 5 grupos

Como puede observarse el número de grupos que posee una separación más clara es el 2 (**Figura 22**), seguido del 3 (**Figura 23**). Sin embargo, se consideró que la definición de únicamente 2 grupos era muy acotada para incorporar las totalidad de las características del deporte. Sumado a esto, puede observarse en las **figuras 24 y 25** que de utilizar más de 3 grupos, sería demasiada la segregación, lo cual tampoco es útil para una correcta descripción.

Por todas estas razones se procedió a generar 3 divisiones.

Luego de correr el código, y obtener la segregación de jugadores, se utilizó la plataforma Power BI para representar y analizar de mejor manera los resultados.

En los gráficos siguientes, se pueden demostrar diferencias entre grupos.

Viendo la **Figura 26**, el Cluster 0 es el grupo de jugadores que tiene el peor índice de +/-, la menor cantidad de minutos jugados, la menor cantidad de participación en jugadas y la mayor cantidad de jugadores. De modo contrario, el Cluster 2 es aquel que representa a los jugadores que más tiempo suelen pasar en la cancha y que más se suelen involucran en las jugadas. Al mismo tiempo, el Cluster 2 posee un +/- levemente mejor que el del Cluster 0 aunque claramente por debajo del del Cluster 1. Mientras tanto, el Cluster 1, juega en promedio menos minutos que el Cluster 2 y también participa en menos jugadas. Es, sin embargo, el grupo con la menor cantidad de jugadores, con tan solo 18. Con esta vista preliminar de cómo se componen los clusters con algunas variables básicas puede comenzar a hacerse una idea de cómo se puede representar a los distintos jugadores.



Figura 26: Comparación de clusters para +/-, PLAYS, Min y Players

Pasando a analizar las variables de tiro visibles en la **Figura 27**, se puede ver que el Cluster 0 es el que peor se desempeña en todas las métricas, es decir, son los que menos puntos hacen y también los que menos puntos intentan hacer.

Por esto se considera listo el análisis de ese grupo y se procede a analizar los dos restantes.

Se puede ver que los jugadores pertenecientes al Cluster 2 son aquellos que más intentos de tiro tienen, esto tiene sentido con aquello que se vió en las variables anteriores, ya que son los que más tiempo juegan en cancha, permitiendo así, realizar más intentos. Sin embargo, los miembros del Cluster 1 son aquellos que más puntos realizan tanto desde fuera del área como dentro del área. En otras palabras, si bien disparan con menos frecuencia que los jugadores del Cluster 2, tienen mayor eficacia y convierten sus tiros en puntos con más frecuencia. Con este análisis hecho puede decirse que los jugadores pertenecientes al Cluster 0 no suelen destacar mucho en ninguna métrica. Si bien son jugadores de alto calibre, no son el grupo con el mejor desempeño. Ahora los clusters 1 y 2 ya empiezan a diferenciarse entre sí, apuntando a un mejor desempeño del Cluster 1.



Figura 27: Comparación de clusters para 2pa, 2pp, 3pa, 3pp

Algunas métricas que indican cómo es el desarrollo del juego son los Steals, las Asistencias, los Rebotes, los Turnovers y los Bloqueos. Estas estadísticas si bien son básicas, definen muy bien el juego de cada jugador, tanto en ataque como en defensa.

Como se puede observar en la **Figura 28** y la **Figura 29**, los integrantes del Cluster 1 lideran en todas estas métricas, incluso en TOV donde se prefiere que sea un número lo más bajo posible. Son los jugadores que mejor se desenvuelven en la mayoría de los aspectos del partido.



Figura 28: Comparación de clusters para STL%, AST%, DRB% y BLK%

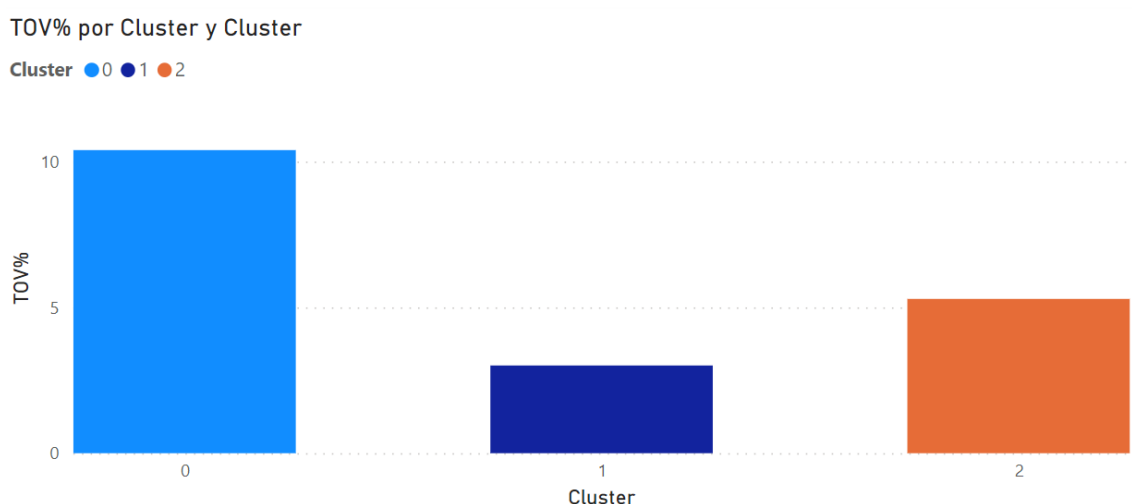


Figura 29: Comparación de clusters para TOV%

Por último, se analizará algunas variables sobre la eficiencia de los jugadores y su participación en el juego. La **Figura 30** muestra que el Cluster 2 es aquel con el mejor HOB y el mejor %USG, es decir, son aquellos que más hacen uso de la pelota y más participan de las jugadas en su totalidad, terminen en gol o no. Mientras que el Cluster 1 lidera las métricas de eficiencia y eficiencia ofensiva, en otras palabras, son aquellos que, cuando hacen uso de la pelota, son mejores en convertir la jugada ya sea en gol o en evitar que el oponente convierta.

En todas estas métricas los jugadores del Cluster 0 son aquellos que peor se desenvuelven.



Figura 30: Comparación de clusters para HOB, EFF, PPP y USG%

Habiendo dicho todo esto, en un nivel general, quedan tres grupos claramente definidos:

- Cluster 0: Jugadores Promedio. Si bien pertenecen a un grupo de alto nivel de jugadores, es poco probable que sean aquellos que

desestabilizan un partido o determinan quién gana. Esto no quiere decir que sean malos. Es muy común en el deporte que haya jugadores muy capaces pero que cumplen cierto rol en el equipo, o entran en momentos indicados, o se dedican a una única cosa. En la NBA hay muchos jugadores de este estilo, muy capaces y dominantes en cualquier otra liga, pero que pasan bastante tiempo sentados en el banco debido al uso que el entrenador decide hacer de ellos.

- Cluster 1: Jugadores Clave o Diferenciales. Son aquellos jugadores que determinan el fluir de una partida. Potenciales MVPs (Most Valuable Player). Siempre buscando lograr una diferencia y convertir cada partido en una ganancia. Este tipo de jugadores son determinantes. Siempre será útil tener diferenciados a estos jugadores, ya sean de nuestro equipo o del contrario. No quiere decir que sean los que tienen que tener el total control del balón pero sí quienes toman mejores decisiones.
- Cluster 2: Jugadores Confiables y Armadores. Jugadores Sólidos. Todo equipo los necesita y sin ellos los Jugadores Claves no brillarían tanto. Sus estadísticas quizás sufren un poco por su tiempo en cancha, pero por el bien del equipo. Estos jugadores son muy importantes para el plantel, hacen bien su trabajo en ambas partes del campo y tienen un buen rendimiento físico, por ende su extenso tiempo en cancha. Son fundamentales para el desarrollo del partido, pero quizás no son determinantes en momentos críticos.

CLUSTERS DE DEFENSA

Buscando hacer más profundidad en el sector defensivo de los equipos se buscó hacer clusters que solamente incluyan las estadísticas defensivas. El número de clusters óptimo para ajustarse a estas estadísticas es 2 de acuerdo al método del codo y al método Silhouette, con un valor de 0.452.

Las estadísticas seleccionadas fueron las siguientes (la explicación de cada una está desarrollada previamente):

- DRB%
- STL/TO
- POS_op
- STL%
- DREB
- BLK
- STL

Como se puede ver en la **Figura 31**, el Cluster 1 se desempeña de mejor manera que el Cluster 0 en todas las métricas. Puede resultar engañoso, pero tener un menor índice de Turnovers (TO) es mejor, ya que Turnovers hace referencia a cuando el equipo pierde la posesión del balón, por lo cual se busca que ese valor sea el menor posible.

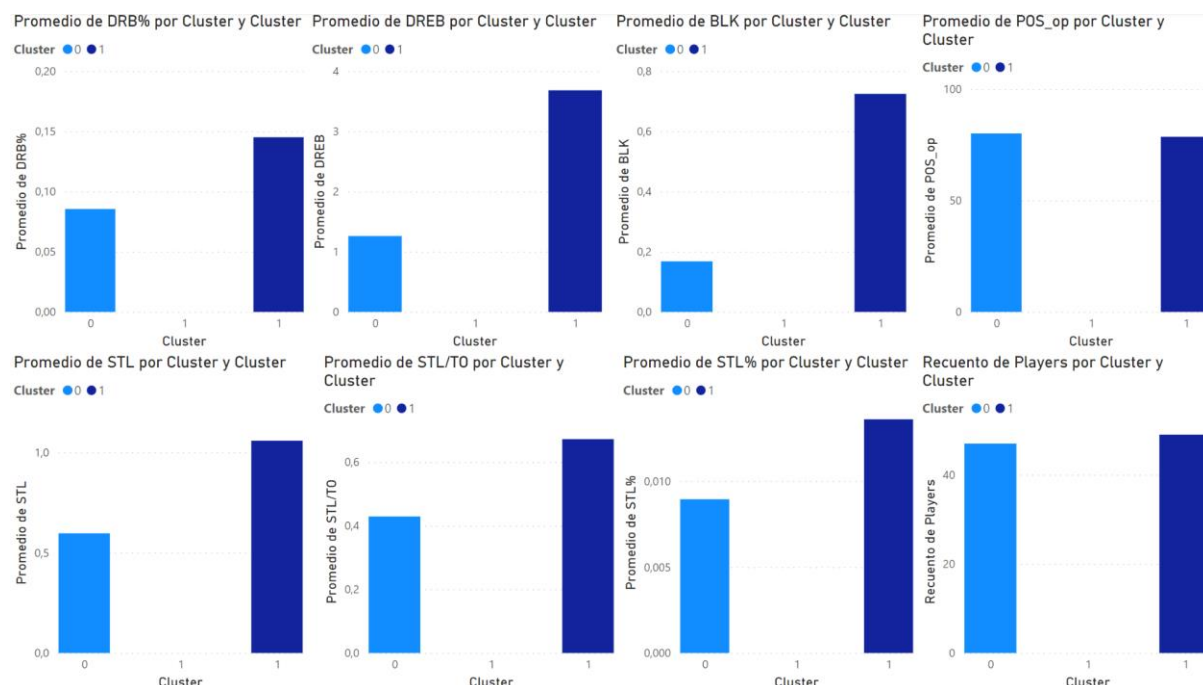


Figura 31: Comparación de clusters para DRB%, DREB, BLK, POS_op, STL, STL/TO, STL% y Players

Como puede observarse en la **Figura 31**, es clara la diferencia entre los grupos, lo que permite definir que el grupo 1 es el que mejores valores presenta, por ende los jugadores de ese grupo son más importantes para el equipo que los del grupo 0.

El costado defensivo del campo es tan importante como el costado ofensivo, por ende es muy útil conocer el desempeño de los jugadores en este nivel. Como se dijo el jugador del cluster 1 hace un mejor uso de su tiempo en cancha que el del grupo 0, y genera diferencia en este costado.

Además hay una clara diferencia en la variable bloqueos, lo que podría indicar a los jugadores del cluster 1 como jugadores de mayor altura así como los que tienen un mejor salto vertical o un mayor entendimiento del juego, características muy relevantes en el costado defensivo.

CLUSTERS DE ATAQUE

Del mismo modo que con las estadísticas defensivas, se busca hacer un “zoom” en las variables de ataque. De acuerdo al método del codo y al Silhouette con un valor de 0.343, la cantidad de Clusters que mejor representan a la base son 3. Las métricas utilizadas para dicho proceso son las siguientes:

- PLAYS
- PLAYS_eq
- PPP
- eFG%
- USG%
- ORB%
- TRB%
- AST/TO
- TS%
- PPT2
- PPT3
- PPTL
- PPTC
- %USO3P
- %USO2P
- %USOTL
- TOV%
- FTR
- AST%
- BLK%
- HOB
- Min
- Pts
- OREB
- REB
- AST
- PF
- TO
- +/-
- EFF
- 2pm
- 2pa
- 2pp
- 3pm
- 3pa
- 3pp
- fgm
- fga
- fgp
- ftm
- fta
- ftp

De acuerdo con la mayoría de las estadísticas presentes en la **Figura 32**, los Clusters 1 y 2 se desempeñan de manera bastante similar, mientras que el Cluster 0 estaría teniendo un desempeño peor que el resto de los clusters. Esto lleva a preguntarse, dónde reside el diferencial entre el Cluster 1 y el Cluster 2.

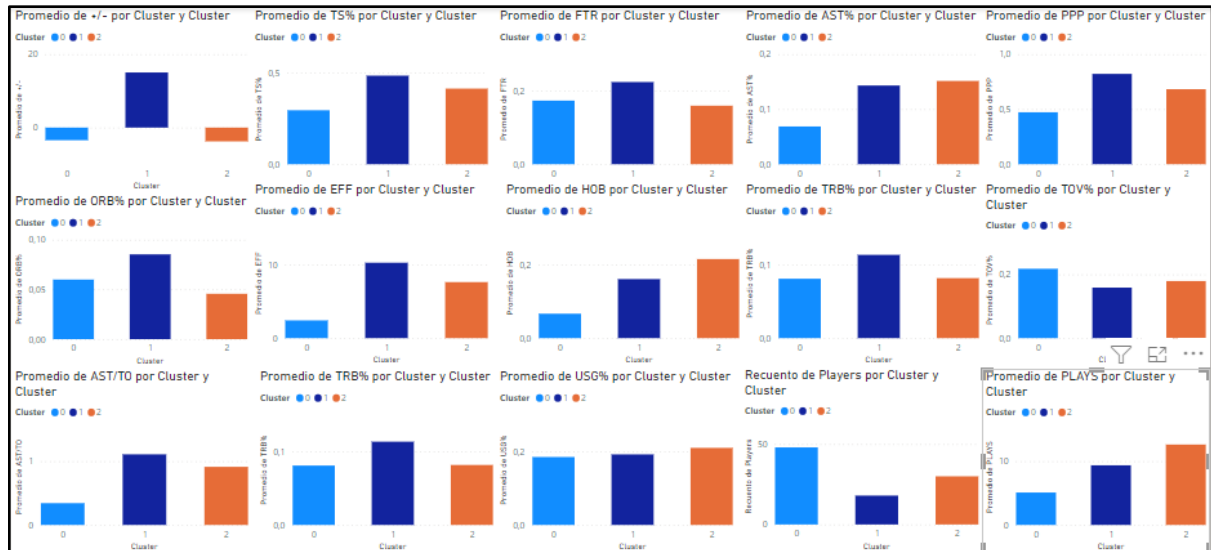


Figura 32: Comparación de clusters para variables de ataque no relacionadas con puntos

Haciendo un foco en las estadísticas de puntos, puede observarse en la **Figura 33** que el Cluster 1 se desempeña de mejor manera en las estadísticas asociadas a 2 puntos mientras que el Cluster 2 se desempeña de mejor manera en las estadísticas asociadas a 3 puntos. Está es una diferencia bastante importante, ya que permite separar a los jugadores que juegan dentro de la línea de 3 puntos de aquellos que juegan desde afuera.

Además el grupo 1 tiene mejores porcentajes en tiros de 2 puntos, indicio de que el tiro es más cercano al aro. También tiene un mayor porcentaje de uso de 2 puntos, indicando la preferencia del jugador por este tiro, que si bien no parece ser tan grande la diferencia hay que tener en cuenta que son valores porcentuales. Por último pero no menos importante, los puntos por tiro de 2 puntos son mucho mayores en este grupo, lo cual viene de la mano con las variables anteriores y ayuda a corroborar la hipótesis previamente planteada.

El grupo 2, es todo lo contrario.

Otro dato importante es el hecho que el grupo 2 toma muchos más tiros que el 1, tanto de 2 como de 3 puntos. Esto indica que el grupo 2 tiene mayor contacto con el balón. Esto es típico de un jugador perimetral, y esta tendencia se acrecienta a medida que avanza el tiempo.

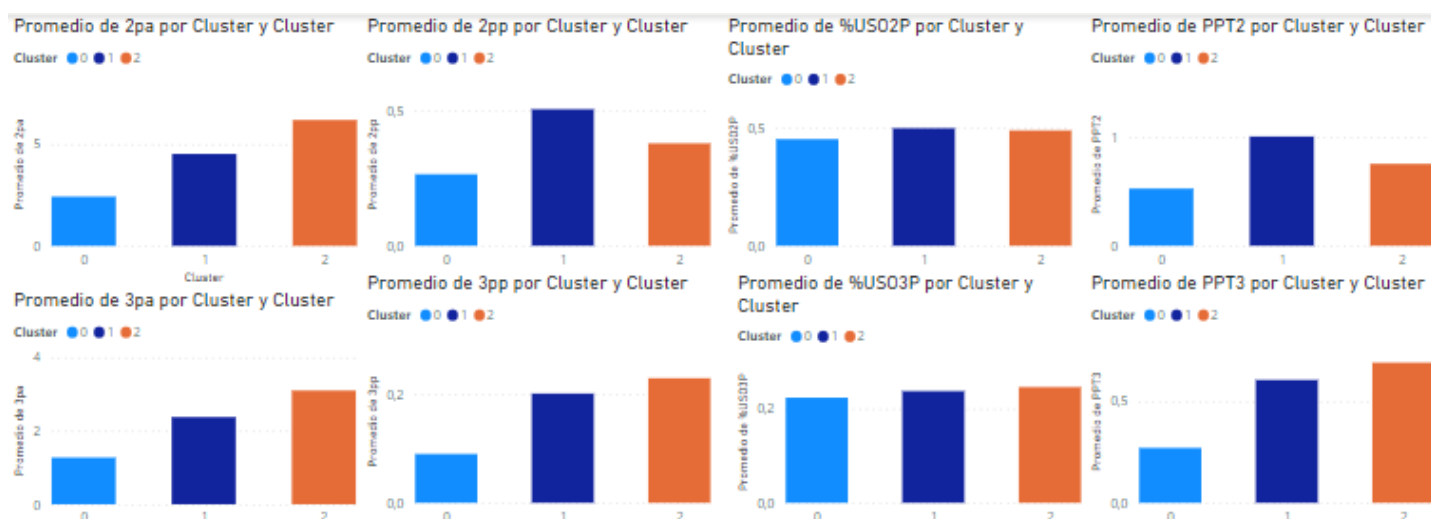


Figura 33: Comparación de clusters para 2pa, 2pp, %USO2P, PPT2, 3pa, 3pp, %USO3P y PPT3

Este análisis deriva en las siguientes definiciones de clusters:

- Cluster 0: Atacantes de Rol: Juegan en momentos específicos y tienen roles específicos. No suelen ser el centro de juego.
- Cluster 1: Atacantes Interiores: Aquellos jugadores que juegan dentro de la línea de 3 puntos. Un ejemplo de este tipo de jugador podría ser Tim Duncan o Jokic.
- Cluster 2: Atacantes Perimetrales: Son aquellos jugadores que juegan desde fuera de la línea de 3 puntos, buscando hacer triples o habilitar a sus compañeros. Un ejemplo de este tipo de jugador podría ser Stephen Curry .

PERFILES DE LOS JUGADORES

Todo este desarrollo de los clusters puede ser muy lindo, pero si no se logran perfilar a los jugadores, realmente no tiene mucho sentido. Haciendo una combinación de todas estas categorizaciones se presentan los perfiles que se determinarían de acuerdo al proceso analítico mencionado.

Es importante poder definir a los jugadores por su juego y dejar de encasillarlos por sus características físicas.

A continuación se presenta una comparación entre las etiquetas definidas por el sistema de cada uno de los jugadores de la selección argentina, y su verdadera posición asignada por el cuerpo técnico. Esto se realizará solo para los jugadores argentinos aunque el trabajo se podría realizar para todos los jugadores.

Players	País	Cluster General	Cluster Ataque	Cluster Defensa	Posición
Diego Collomb	Argentina	Jugador Promedio	Atacante Periférico	Defensor Clave	BASE
Dylan Nicolas Bordon Dri	Argentina	Jugador Clave	Atacante Interior	Defensor Clave	BASE
Fausto Moussa	Argentina	Jugador Promedio	Atacante Promedio	Defensor Promedio	ESCOLTA
Francisco Espinosa	Argentina	Jugador Promedio	Atacante Promedio	Defensor Promedio	ALERO
Ivan Pratto	Argentina	Jugador Promedio	Atacante Promedio	Defensor Promedio	PIVOT
Juan Respaud	Argentina	Jugador Promedio	Atacante Promedio	Defensor Promedio	BASE
Lee Aaliya	Argentina	Jugador Clave	Atacante Interior	Defensor Clave	ALA PIVOT
Nicolas Stenta	Argentina	Jugador Clave	Atacante Interior	Defensor Clave	ESCOLTA
Patricio Bautista Giralt	Argentina	Jugador Promedio	Atacante Promedio	Defensor Clave	PIVOT
Santiago Trouet	Argentina	Jugador Clave	Atacante Interior	Defensor Clave	ALERO
Tiziano Prome	Argentina	Jugador Promedio	Atacante Promedio	Defensor Promedio	ALA PIVOT
Ignacio Ortega Naumchik	Argentina	Jugador Promedio	Atacante Promedio	Defensor Clave	PIVOT

Tabla 4: Perfiles de los jugadores argentinos

Ya desde el comienzo se puede ver que dos jugadores que estén asignados a la misma posición no son representados por los mismos clusters. Si se compara el caso de los dos primeros jugadores Diego Collomb y Dylan Nicolas Bordon Dri, se conoce que ambos están categorizados bajo la posición de Base. Sin embargo, si se tienen en cuenta los clusters, Collomb se desempeña como un Atacante Periférico mientras que Bordon como un Atacante Interior. Además de que Bordon es un jugador clave mientras que Collomb no lo es.

Así es como se van a determinar algunos potenciales perfiles. La manera de analizarlos será por el orden de posiciones originales para resaltar las diferencias.

- Base

- Diego Collomb: Jugador defensivo, que juega bien desde las afueras del área.
- Dylan Nicolas Bordon Dri: Jugador desestabilizante. Si bien juega desde dentro del área a la hora de atacar, también es un excelente defensor
- Juan Respaud: Jugador Promedio
- Escolta
 - Fausto Moussa: Jugador Promedio
 - Nicolas Stenta: Se comporta similar a Bordon a pesar de estar categorizados como posiciones diferentes
- Alero
 - Francisco Espinosa: Jugador Promedio
 - Santiago Trouet: Al igual que Stenta y Bordon, es un atacante interno que defiende de manera muy efectiva.
- Pivot
 - Ivan Pratto: Jugador Promedio
 - Patricio Bautista Giralt: Un gran defensor. A la hora de atacar quizás no destaca mucho, pero evita que el equipo oponente convierta sus jugadas en puntos.
 - Ignacio Ortega Naumchik: Similar a Giralt
- Ala Pivot
 - Lee Aaliya: Se une al grupo de nuestros grandes jugadores. Junto con Stenta, Trouet y Bordon.
 - Tiziano Prome: Jugador Promedio

Es muy importante destacar que esta segregación brinda información basada únicamente en lo ocurrido en el último torneo. La gran capacidad de la solución será ampliada notablemente a medida que se le agreguen variables y más partidos a la base. Por ejemplo, si se conociera la distancia recorrida de un jugador por partido, se ampliaría más la búsqueda y sería posible dar mucho más detalle del juego de cada jugador. Esto ocurre con miles de nuevas variables posibles.

PLANTILLAS

Habiendo hecho todo este análisis y conociendo al plantel desde otro punto, se da con la posibilidad de seguir agregando valor. Para esto se decidió desarrollar un modelo de programación lineal, que permitirá optimizar cualquier variable deseada, y brindará el quinteto ideal para cumplir con ese objetivo.

Sin embargo se considera que por la naturaleza del básquetbol, no es suficiente con maximizar una variable para buscar obtener buenos resultados, ya que pueden dejarse muchas cosas de lado.

Es por esto que también se tomó la decisión de añadir restricciones al modelo, que permitirán maximizar cualquier variable, pero sin dejar de lado puntos clave del juego.

Este modelo agrega una gran cantidad de valor, ya que mientras mayor sea el conocimiento del juego, mayor sea la cantidad de datos y mejor sea la implementación de la herramienta, mejores resultados se pueden obtener.

Se pueden crear nuevas variables, nuevas alternativas, nuevas restricciones, entre otras opciones.

Al momento, la variable a maximizar elegida fue la cantidad de puntos anotados.

Para demostrar la capacidad del modelo se desarrollaron una serie de alternativas, presentadas en el anexo, que permitieron conocer diferentes alternativas de plantillas.

Este es el resultado final de los 5 jugadores que maximizan los puntos contra los valores que ocurrieron:

Variable	Modelo	Realidad
Pts	80,11363636	72,25
REB	38,18181818	42,6
AST	16,70454545	14,7
TO	19,09090909	19,75
STL	12,95454545	11,05
BLK	8,522727273	9
+/-	88,63636364	73,8
EFF	94,77272727	87,05
2pp	4,587391775	5,09555556
3pp	2,284090909	2,49416667
fgp	3,901225586	5,0044333

Tabla 5: Comparación del modelo sin restricciones contra la realidad

Si bien algunos valores son menores que en la realidad, los valores más relevantes para el análisis son mucho mejores. Por ejemplo los puntos y el +/- tienen una gran diferencia que puede llegar a significar un campeonato.

Este quinteto incluye a los siguientes jugadores:

- Dylan Nicolas Bordon Dri
- Fausto Moussa
- Lee Aaliya
- Nicolas Stenta
- Santiago Trouet

Sin embargo, no es sensato enviar 5 jugadores al campo, solo teniendo en cuenta la cantidad de puntos que pueden anotar, porque el juego no solo se trata de eso.

Por ende se desarrollaron restricciones que van de la mano con las definiciones de los jugadores realizadas previamente. De esta manera es posible obtener un equipo considerado “completo”, y cubrir todos los aspectos del deporte.

El grupo ideal incluiría 2 jugadores del cluster 2 de ataque, 2 del 1 de ataque y uno a definir por el algoritmo. Además, al no interferir los clusters entre sí, se incluirían 2 jugadores como mínimo del cluster general 1 y 2 jugadores como mínimo del cluster de defensa 1.

Sin embargo, no existen 2 jugadores que sean parte del cluster de ataque 2 en la plantilla argentina, por lo que se modificó la restricción como 1 jugador.

Habiendo seteado esas restricciones, puede observarse que la plantilla que maximiza el puntaje anotado es la misma. Estas son muy buenas noticias, ya que dan a entender que el análisis y la segregación de jugadores realizada previamente, tiene algún tipo de confianza empírica.

Por último, se incluyeron más niveles en el análisis de nuestro quinteto ideal, y se desarrollaron 2 restricciones más. La primera, se asegura que el jugador haya tenido como mínimo 10 minutos de juego promedio en el torneo, y la segunda se asegura que el balance +/- del jugador sea de 3, lo cual brindaría al equipo la ventaja sobre el oponente.

Por ende esta alternativa tiene las siguientes restricciones:

- Mínimo 10 minutos
- Mínimo +/- de 3
- Mínimo 2 del cluster 1 de ataque
- Mínimo 1 del cluster 2 de ataque
- Mínimo 2 del cluster 1 de defensa
- Mínimo 2 del cluster 1 general

Esta formación devuelve los siguientes resultados (comparados con las otras alternativas):

	Alternativa Sin Restricciones	Realidad	Alternativa Ideal (Con Restricciones)
Pts	80,11363636	72,25	77,2987538
REB	38,18181818	42,6	40,1650387
AST	16,70454545	14,7	16,1670596
TO	19,09090909	19,75	18,945773
STL	12,95454545	11,05	12,1252947
BLK	8,522727273	9	9,34658134
+/-	88,63636364	73,8	73,2569889
EFF	94,77272727	87,05	95,2340856
2pp	4,587391775	5,09555556	5,08568702
3pp	2,284090909	2,49416667	2,53452341
fgp	3,901225586	5,0044333	4,6443066

Tabla 6: Comparación Realidad, Modelo sin Restricciones, Modelo con Restricciones

Con la "Alternativa Ideal(Con Restricciones)", se deja de perder en contra de la realidad en partes muy relevantes, que son por ejemplo, rebotes, EFF, bloqueos, pérdidas, entre otros.

Por ende esta alternativa se presenta como una mejor solución para un juego "all around" o balanceado.

Lo curioso de esta alternativa es que, al analizar los datos del torneo, los jugadores que pertenecen a la misma, estuvieron un 80% del torneo en cancha. Esto significa que de alguna manera u otra, el resultado obtenido es similar al del cuerpo técnico, sin tener todos los años de experiencia. Esto no quiere decir que se pueda utilizar este método como un director técnico, sino que con el conocimiento de alguien experimentado mezclado con este tipo de tecnologías se podría llegar a resultados nunca antes vistos.

El quinteto que representa esta alternativa es:

- Diego Collomb

- Dylan Nicolas Bordon Dri
- Lee Aaliya
- Nicolas Stenta
- Santiago Trouet

A continuación se puede observar la lista de rankings planteados en la hipótesis, pero teniendo en cuenta a los valores obtenidos gracias al modelo de programación lineal en el caso de Argentina.

Variable	Argentina	Argentina Upgrade	Variación
POS_eq	9	1	8
RITMO_eq	9	4	5
OER_eq	2	3	-1
PLAYS_eq	9	3	6
eFG%_eq	2	1	1
ORB%_eq	3	1	2
DRB%_eq	7	1	6
TRB%_eq	3	1	2
FTR_eq	5	1	4
%USO3P_eq	7	1	6
%USO2P_eq	6	1	5
%USOTL_eq	5	1	4
TOV%_eq	2	1	1
AS/TO_eq	5	1	4

STL/TO_eq	5	1	4
PPT2_eq	2	1	1
PPT3_eq	2	1	1
PPTL_eq	3	1	2
PPTC_eq	2	1	1
AST%_eq	5	1	4
%TS_eq	2	1	1
BLK%_eq	2	1	1
DER_eq	8	1	7
NET RATING	2	0	2
STL%_eq	2	1	1
Pts	3	2	1
OREB	4	9	-5
DREB	4	5	-1
REB	5	9	-4
AST	4	1	3
PF	8	7	1
TO	3	1	2
STL	3	1	2

BLK	2	1	1
+/-	3	1	2
EFF	3	2	1
2pm	3	1	2
2pa	7	2	5
2pp	2	1	1
3pm	4	3	1
3pa	8	6	2
3pp	2	1	1
fgm	3	1	2
fga	8	4	4
fgp	2	1	1
ftm	4	8	-4
fta	4	8	-4

Tabla 7: Comparación Ranking Plantilla Planteada vs Plantilla Original

Como puntos claves a analizar en estos resultados pueden encontrarse:

- Una de las hipótesis planteadas fue que las variables más importantes en pos de una victoria podrían ser Eficiencia Ofensiva (OER) y los 3 tipos de rebotes, eFG%, PPT2, PPT3 y PPTC. En este caso puede verse que el equipo argentino propuesto obtendría el primer lugar en todas estas variables menos en OER, donde obtendría el tercer lugar. Esto puede servir como indicio de las consecuencias que puede traer el uso de este modelo de programación.

- Se mejoran los porcentajes de tiro y se limita el uso del tiro de 3 puntos.
- Se obtendría el primer lugar en AS/TO y en STL/TO indicando un buen cuidado del balón. Esto va de la mano con el 4 lugar en Ritmo, indicando que el juego se mantiene relativamente lento, aunque se mejoró por sobre la realidad.}

Como puede reconocerse con un simple análisis, los resultados son mejores. Si bien el resultado no es 100% aplicable por temas de cansancio y rendimiento de los jugadores, sería imprudente no seguir invirtiendo tiempo y conocimientos en el desarrollo de este modelo.

CONCLUSIONES

Con el objetivo de llegar a conclusiones concisas se hará un veloz recorrido de lo realizado.

Como primer paso, se obtienen los datos de la web, con el método de scraping desarrollado por nosotros.

Luego, realizando grupos donde se separan a todos los jugadores por todas sus métricas, se busca encontrar patrones que permitan diferenciarlos. Al hacer esto se despoja a los jugadores de los prejuicios instalados en el deporte, basados más que nada en el físico, y se puede definirlos por estilo de juego.

Como resultado se obtienen grupos con Jugadores Promedio, Jugadores Confiables y Jugadores Clave. Sin embargo, esto no fue suficiente y se busco separarlos por su desempeño ofensivo y defensivo. Con esto se separó a los Buenos Defensores de aquellos Defensores Promedio y del mismo modo, a los Atacantes Promedio de aquellos que se desempeñan de manera efectiva. Dentro de los jugadores efectivos en ataque se separan Atacantes Internos de Atacantes Perimetrales.

Toda esa información permitió perfilar a los jugadores de una manera diferente a la que era utilizada previamente.

Con esta información, se armaron distintos modelos de programación lineal que devolvieran alternativas de equipos que maximicen alguna variable en particular, sin dejar de lado otros aspectos del juego.

Como fue mencionado previamente, el equipo final armado con estas metodologías es similar al utilizado por el cuerpo técnico en el torneo, lo que presenta que los resultados obtenidos no solo son comprobables, sino que también sería válido experimentar otras opciones de armado de equipos dependiendo de lo que las situaciones requieran. Del mismo modo, estas técnicas se pueden utilizar para tratar de identificar cual es el equipo más probable de uno de los equipos oponentes,

proveyendo una ventaja técnica a Argentina sin requerir horas y horas de análisis manual.

Se cree que el uso y el constante desarrollo de este sistema puede traer muy buenos resultados, por la capacidad y potencial de la tecnología utilizada. Alguien con más experiencia en el deporte, que tenga mejor comprensión del juego o mismo un mayor poder de decisión sobre el equipo podría hacer gran uso de esta herramienta.

Otro punto clave, es que sea cual sea el equipo que se quiera analizar, será posible siempre y cuando sea subido a una página web de similares características. De hecho este sistema fue probado también con los datos de jugadores de la selección mayor.

Por último, queda decir que tanto el proceso de ETL que disponibiliza los datos, como el modelo de clusterización y el de optimización serán integrados en el mismo archivo de código, para ser ejecutados una única vez y así obtener los resultados que vayan directamente al mismo archivo de Power BI, integrando todas las etapas del proceso en una misma solución.

PUESTA EN ACCIÓN

Hace ya un par de semanas, los entregables fueron entregados de manera final al cliente. El cual está muy impresionado con lo logrado y está haciendo uso de todo lo desarrollado en el torneo Pre-Mundial de América, donde el seleccionado obtuvo el cuarto puesto y logró el objetivo de la clasificación el siguiente año.

En palabras de nuestro cliente: "Felicitaciones y gracias porque hoy funciono de primera todo lo entregado. Ya corrimos todo, ya entregamos todo y pude dormir unas horas más así que estoy re contento".

No solo el cliente está muy contento con el trabajo realizado sino también mostró un gran interés por las plantillas armadas. Se nos solicitó que le dieramos una mano con la creación de distintas planillas que maximicen distintas cosas que puntos como PPP.

Diciendo esto puede afirmarse que la solución es aplicable, útil y relevante, además de que las posibilidades de mejora de la misma son inimaginables.

BIBLIOGRAFÍA

Atkinson, B. (2019, December 12). Guide to building a college basketball machine learning model in Python. Medium. Retrieved April 3, 2022, from <https://towardsdatascience.com/guide-to-building-a-college-basketball-machine-learning-model-in-python-1c70b83acb51>

Jain, S., & Kaur, H. (2017). Machine learning approaches to predict basketball game outcome. IEEE Xplore. Retrieved April 3, 2022, from <https://ieeexplore.ieee.org/document/8344688/>

Narayan, S. (2019). Applications of machine learning: Basketball strategy (thesis). Massachusetts Institute of Technology.

Predict basketball player efficiency ratings with machine learning and Visual Studio Code | Lwdg. (2020). Microsoft Developer. Retrieved April 3, 2022, from <https://www.youtube.com/watch?v=R6S05RkWI10&list=PLlrXD0HTieHgJdiA08EVViP8D6hfDRXx8&index=14>.

Weiner, J. (2021, January 5). Predicting the outcome of NBA games with Machine Learning. Towards Data Science. Retrieved April 3, 2022, from <https://towardsdatascience.com/predicting-the-outcome-of-nba-games-with-machine-learning-a810bb768f20>

ANEXO

Las alternativas utilizadas para probar el modelo de optimización fueron:

Alternativa 1

País: Argentina

PPP: Mayor a 0.5

MasMenos : Mayor a 3

Alternativa 2

País: Argentina

PPP: Mayor a 0.5

Alternativa 3

País: Argentina

Alternativa 4

País: Argentina

MasMenos : Mayor a 1

Alternativa 5

País: Argentina

MasMenos : Mayor a 5

Alternativa 6

País: Argentina

Cluster Ataque: Entre 2 y 5 del cluster 1

Alternativa 7

País: Argentina

MasMenos : Mayor a 1

Cluster Ataque: Entre 2 y 5 del cluster 1

Alternativa 8

País: Argentina

MasMenos : Mayor a 5

Cluster Ataque: Entre 2 y 5 del cluster 1

Alternativa 9

País: Argentina

Cluster General: Entre 2 y 5 del cluster 1

Alternativa 10

País: Argentina

MasMenos : Mayor a 5

Cluster General: Entre 2 y 5 del cluster 1

Alternativa 11

País: Argentina

Cluster General: Entre 2 y 5 del cluster 1

Cluster Ataque: Entre 2 y 5 del cluster 1

Cluster Ataque: Entre 2 y 5 del cluster 2

Cluster Defensa: Entre 2 y 5 del cluster 1

En este link están los archivos que contienen los resultados de cada alternativa:

<https://drive.google.com/drive/folders/1CIBdNsyHCHicCc--43kQHWgMeF0eEjmE?usp=sharing>

ANEXO II: PRESENTACIÓN

I-BALL

Sports Analytics



Coluccio Ignacio y Prada Tadeo

CONTENIDO



01

PROBLEMÁTICA

REQUISITOS



03



02

SOLUCIÓN

DEMO



04



01

PROBLEMÁTICA

PROBLEMÁTICA

APROVECHAR LOS DATOS

CONOCER AL RIVAL

RÁPIDO

CONFIABLE

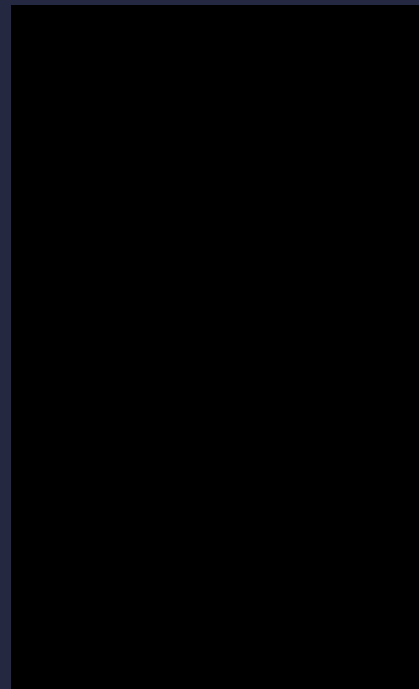
FÁCIL DE UTILIZAR

ADECUADO AL USO



¿POR QUÉ DATOS?

- **ELECCIÓN DE JUGADORES**
- **ESTRATEGIA**
- **RECUPERACIÓN**
- **ALIMENTACIÓN**





“Players are still forced into the characterization of the traditional five positions: point guard, shooting guard, small forward, power forward, and center, and metrics such as True Shooting Percentage and Expected Shot Quality are just beginning to become well-known.”

— **Narayan, Santhosh.**
(MIT) 2019



02

REQUISITOS

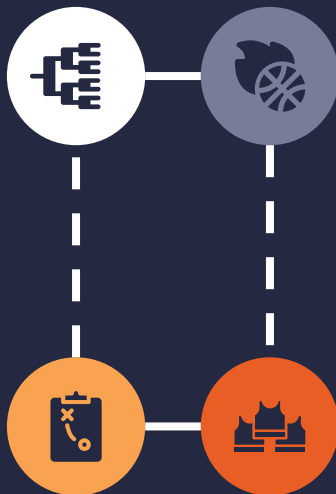
¿CÓMO SERÁ UTILIZADO?

SCOUTING

INDICADORES

**DEFINICIÓN DE
OBJETIVOS**

**INICIAR Y
RESPONDER
PREGUNTAS**

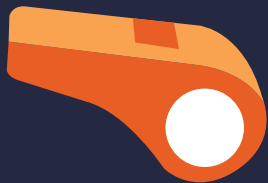




03

SOLUCIÓN

¿QUÉ SE TUVO EN CUENTA?



CUMPLIR

- Velocidad
- Simplicidad
- Familiaridad
- Exportabilidad

USO MÚLTIPLE



VALOR

¿EN QUÉ CONSISTE?



ETL

Extracción de los datos
de los partidos a una
planilla de manera
automática



ANÁLISIS

Descripción de los datos
y la búsqueda de
Insights



A.I

Perfilamiento de
jugadores y creación de
equipos

ETL



EXTRACCIÓN

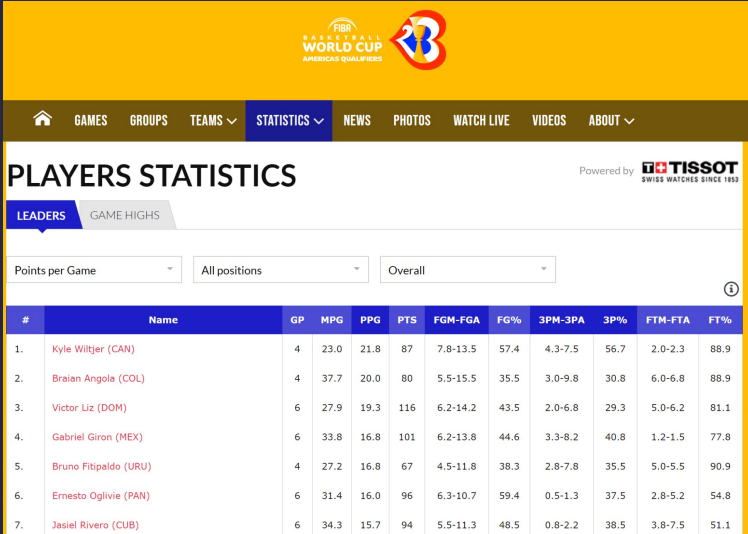
LIMPIEZA

**CÁLCULO DE
ESTADÍSTICAS
AVANZADAS**

DISPONIBILIZACIÓN

FEDERACIÓN INTERNACIONAL DE BALONCESTO (FIBA)

Página de la cual se podrán
recolectar todos los datos
disponibles para crear nuestros
análisis



PLAYERS STATISTICS											
Powered by 											
LEADERS											
Points per Game											
All positions											
Overall											
#	Name	GP	MPG	PPG	PTS	FGM-FGA	FG%	3PM-3PA	3P%	FTM-FTA	FT%
1.	Kyle Wiltjer (CAN)	4	23.0	21.8	87	7.8-13.5	57.4	4.3-7.5	56.7	2.0-2.3	88.9
2.	Braian Angola (COL)	4	37.7	20.0	80	5.5-15.5	35.5	3.0-9.8	30.8	6.0-6.8	88.9
3.	Victor Liz (DOM)	6	27.9	19.3	116	6.2-14.2	43.5	2.0-6.8	29.3	5.0-6.2	81.1
4.	Gabriel Giron (MEX)	6	33.8	16.8	101	6.2-13.8	44.6	3.3-8.2	40.8	1.2-1.5	77.8
5.	Bruno Fitipaldo (URU)	4	27.2	16.8	67	4.5-11.8	38.3	2.8-7.8	35.5	5.0-5.5	90.9
6.	Ernesto Oglivie (PAN)	6	31.4	16.0	96	6.3-10.7	59.4	0.5-1.3	37.5	2.8-5.2	54.8
7.	Jasiel Rivero (CUB)	6	34.3	15.7	94	5.5-11.3	48.5	0.8-2.2	38.5	3.8-7.5	51.1

ANÁLISIS



**CONOCIMIENTO
DEL RIVAL**



**CONOCIMIENTO
PROPIO**



**PARTIDO A
PARTIDO**



EVOLUTIVO



**DESCRIPCIÓN DE
JUGADORES**



EXPORTABLE

A.I



CLUSTERING

Metodología utilizada para
crear los nuevos perfiles
para los jugadores



PROGRAMACIÓN LINEAL

Busca la configuración
óptima de jugadores para
crear equipos

CLUSTERING

JUGADOR

CLUSTER GENERAL

Crea clusters usando
todas las métricas
disponibles

**CLUSTER
DEFENSIVO**

Crea clusters usando las
métricas defensivas

CLUSTER OFENSIVO

Crea clusters usando las
métricas ofensivas

PROGRAMACIÓN LINEAL

Se ingresa la variable que se busca optimizar y las limitaciones.
El programa busca la configuración óptima de jugadores que maximicen estas variables.



EQUIPO PROPUESTO



JUGADOR	CLUSTER GENERAL	CLUSTER DEFENSIVO	CLUSTER OFENSIVO
Dylan Bordon Dri	Clave	Clave	Interior
Diego Collomb	Promedio	Clave	Periferico
Lee Aaliya	Clave	Clave	Interior
Nicolas Stenta	Clave	Clave	Interior
Benjamin Marchiaro	Clave	Clave	Interior
Santiago Trouete	Clave	Clave	Interior



“El francotirador Benjamín Machiaro abrió el camino con 18 puntos en poco más de 15 minutos de juego, disparando $\frac{3}{4}$ desde más allá del arco. Santiago Trouet y Diego Collomb sumaron 10 puntos cada uno para un equipo que dominó todas las categorías, incluidos los rebotes (55-32), los puntos en la pintura (42-16), los puntos de segunda oportunidad (12-7) y los puntos perdidos (25-32). 11.”

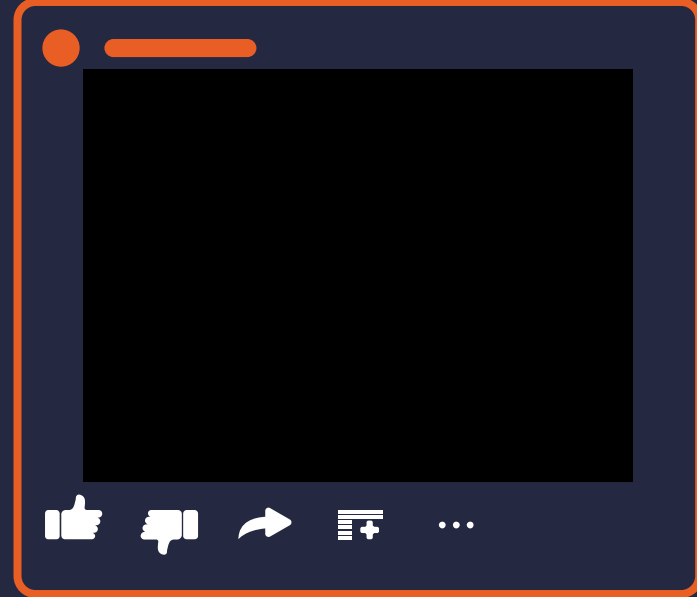
— **Página FIBA**



04

DEMO

DEMOSTRACIÓN



¿Qué ofrecemos?

15 - 20
hs

\$250.000

Por torneo

< 7

Minutos

1°

Posición alcanzada de usar
nuestra opción

¿DÓNDE SE UTILIZÓ?

Pre-Mundial

U-18

Mundial

U-17



Clasificatorio

U-18



Selección

Mayor





Consultas?

DISPOSICIÓN PERMANENTE N° 700
ANEXO III

MODELO DE AUTORIZACIÓN PARA LA PUBLICACIÓN DE DOCUMENTOS
EN EL REPOSITORIO INSTITUCIONAL

El/ los autor/es firmante/s autoriza/n al Instituto Tecnológico de Buenos Aires (ITBA) a poner a disposición del público la obra detallada en el presente documento, a solo fin de divulgación de la producción científico-académica de la Universidad. El trabajo será de consulta libre y gratuita en el Repositorio Institucional ITBA, a través de Internet, y en todos aquellos repositorios digitales en los que participe la Universidad.

Esta autorización representa la cesión al ITBA, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, de los derechos de reproducción, distribución y comunicación pública por cualquier medio o soporte de la obra.

Asimismo, se autoriza la transformación de la obra, sin producir cambios en el contenido, siempre que sea necesaria para permitir su preservación y uso en formato electrónico, incluyendo la realización de copias digitales y migraciones de formato necesarios para la seguridad, resguardo y preservación a largo plazo de la misma.

1. Datos del/os Autor/es¹

Apellido y Nombre: Ignacio Nicolás Coluccio
DNI: 42497577
Legajo: 60288
E-mail: icoluccio@itba.edu.ar

Apellido y Nombre: Tadeo Prada
DNI: 42193996
Legajo: 60219
E-mail: tprada@itba.edu.ar

2. Datos de la obra

Título completo del trabajo: Automatizacion en la recoleccion de Datos Deportivos
.....
.....
Palabras clave: Clusters, Basketball, AUTomatizacion, Programacion Lineal
.....
Carrera: Licenciatura en Analitica Empresarial y Social

3. Tipo de obra:

- Tesis doctorado []
- Tesis maestría []
- Trabajo final de Especialización []
- Proyecto Final de Grado ☒

¹ Deberán firmar todos los autores de la obra.

4. Autorizo la publicación del:

Texto completo ☒

A partir de su aprobación/presentación ☒

Dentro de los 6 meses posteriores a su aprobación/presentación []

Otro plazo mayor (detallar/justificar):

5. NO autorizo []

Si Ud. Se encuentra comprendido en el caso de que su producción esté protegida por derechos de Propiedad Industrial y/o acuerdos previos con terceros que implique la confidencialidad² de los mismos, indique por favor el motivo:

El período de confidencialidad o el secreto del trámite finaliza el:

El/los autor/es declara/n que la autorización realizada no infringe derechos de terceros y libera/n al ITBA de todo tipo de responsabilidad (sea civil, administrativa o penal) que pudiera surgir frente a cualquier reclamo o demanda referida a la obra por parte de terceros, asumiendo dicha responsabilidad de forma exclusiva.

Acepta/n y toma/n conocimiento de que en caso que la obra sea inédita perderá la condición de tal con su publicación en la web.

Lugar y fecha: Jueves 18 de agosto del año 2022, Buenos Aires, Argentina.

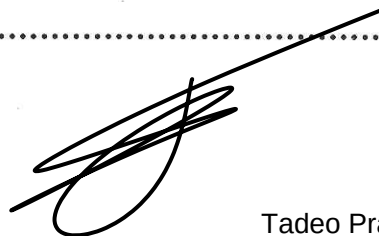
Firma y aclaración del/os autor/es

.....

.....



Ignacio Nicolas Coluccio



Tadeo Prada

A ser completado por el Departamento de Grado /Posgrado/Doctorado:

Nro. de Acta Calificación:

Jurado (Apellido, Nombre):

.....

.....

.....

Fecha de defensa/aprobación:/...../.....

Firma y sello
(Director de Departamento)

² NOTA: Se incluyen acuerdos con terceros, políticas institucionales, leyes, reglamentaciones, decisiones unilaterales, etc. En cualquiera de estos casos, se deberá acompañar copia del acuerdo de confidencialidad, del acuerdo que contiene cláusulas de confidencialidad o de la solicitud de derecho de propiedad industrial, o documentación correspondiente.