

Anteproyecto de Tesina

Estudiante: Simón Pedro Gazze

Directores: Andrés Sosa, Ignacio Álvarez

Fecha: 2 de noviembre de 2025

Universidad Nacional de Rosario

Contenido

1 Introducción	2
1.1 Sports Analytics	2
1.2 Medidas de desempeño para jugadores de básquetbol	2
1.3 RAPM (Regularized Adjusted Plus Minus)	3
2 Motivación	4
3 Objetivos	5
3.1 Objetivo general	5
3.2 Objetivos específicos	5
4 Metodología	5
4.1 Modelos propuestos	5
4.1.1 Modelo de regresión lineal múltiple	5
4.1.2 Modelo logístico multinomial	6
4.2 EPTS y wEPTS	7
4.2.1 EPTS	7
4.2.2 wEPTS	8
4.3 Estimación de los RAPMs	8
4.3.1 Método de mínimos cuadrados	8
4.3.2 Método de máxima verosimilitud	8
4.4 Métodos de regularización	9
4.4.1 Regresión Ridge o Regularización L2	9
4.4.2 Lasso o Regularización L1	9
4.4.3 Elastic Net	10
4.5 Eficiencia computacional	10
4.6 Comparación de modelos	11
4.6.1 Consistencia	11
4.6.2 Validez	12
5 Aplicación	12
5.1 Generación y preprocesamiento de la base de datos	12
5.2 Estimación de los RAPMs	13
5.3 Comparación de los distintos enfoques	13
6 Cronograma de actividades	14
7 Bibliografía	15

1 Introducción

1.1 Sports Analytics

Sports Analytics o analítica en el deporte es un término que refiere a la aplicación de métodos estadísticos para describir, explicar y/o predecir fenómenos vinculados al rendimiento deportivo, tanto a nivel individual como colectivo con el objetivo de facilitar la toma de decisiones de entrenadores o staff técnico antes, durante y después de los partidos.

A pesar de que las primeras aplicaciones al análisis de datos al deporte se remontan a mediados del siglo XX, el impulso más significativo ocurrió a partir de los años 2000, en particular en el béisbol, con la popularización del enfoque conocido como *sabermetrics*¹. Desde entonces, la práctica se extendió a otros deportes como el básquetbol, el fútbol o el tenis; dando lugar a una nueva cultura de toma de decisiones basada en evidencia. Hoy en día, los equipos profesionales, las ligas y las federaciones emplean departamentos especializados en análisis de datos para optimizar estrategias de juego, evaluar el rendimiento de jugadores, prevenir lesiones y gestionar recursos económicos de manera más eficiente.

El **básquetbol** es uno de los deportes que más ha incorporado el análisis estadístico en su evolución reciente. Desde las primeras métricas derivadas del *box score*, el desarrollo de bases de datos *play-by-play* a partir de la década del 2000 hasta la incorporación más reciente de tecnologías de *player tracking*². A partir de esta información, se han desarrollado distintas herramientas estadísticas para poder estudiar la eficiencia de los jugadores, la eficiencia de los tiros al aro, el desempeño de los equipo, la importancia en la creación de espacios, entre otras cosas.

1.2 Medidas de desempeño para jugadores de básquetbol

Desde la consolidación del box-score en el básquetbol profesional, durante la década de 1950, empezaron a calcularse las primeras medidas de desempeño para los jugadores, como pueden ser los puntos, rebotes, asistencias, etc. Durante mucho tiempo estas métricas constituyeron una de las formas más tradicionales de evaluar el rendimiento individual en el básquetbol. Su amplia disponibilidad y fácil interpretación las convirtieron en el principal insumo para comparar jugadores, otorgar premios o negociar contratos. Estas medidas suelen agruparse en un conjunto de estadísticas denominadas *bottom-up*, las cuales se calculan en función de las acciones individuales de los jugadores, es decir, si un jugador realiza una acción que se asocia con un resultado positivo para el equipo, la calificación de ese jugador aumenta, mientras que la calificación de los demás jugadores, que no participaron en la jugada, permanece sin cambios. Pero estas medidas presentan algunas limitaciones importantes. Por ejemplo, sólo capturan una parte de los eventos relevantes del partido (acciones como una defensa asfixiante, buenas rotaciones defensivas o pases rápidos para iniciar una posesión no están contempladas) y no necesariamente reflejan el impacto real de un jugador en las victorias de su equipo, que es lo que importa a fin de cuentas.

Por otra parte, las medidas *top-down* se basan en el rendimiento del equipo en su conjunto, y el crédito por dicho rendimiento se distribuye entre los jugadores que estuvieron involucrados en el partido, sin importar qué acciones puntuales hayan realizado. Dentro de este conjunto de métricas encontramos una gran variedad de estadísticas muy utilizadas en la actualidad como el Plus/Minus (+/-), Win Shares, On-Off Rating, etc.

Con la idea de que lo más importante en un partido es **ganar**, y entendiendo que las estadísticas tradicionales son solo una medida imperfecta de muchas de las contribuciones que realizan los jugadores a la victoria, en este trabajo buscaremos “pensar más allá del box score” y nos centraremos en estadísticas *top-down*, más precisamente en la medida Plus/Minus. Esta estadística registra los cambios que se producen en el marcador durante los minutos en que el jugador de interés está en cancha, partiendo de la idea de que los equipos

¹Término acuñado por Bill James (1980) para referirse al análisis estadístico del béisbol, basado en la recopilación y estudio sistemático de datos con el fin de comprender y cuantificar el desempeño de los jugadores y equipos. El nombre proviene de SABR (Society for American Baseball Research).

²El player tracking es un sistema tecnológico que utiliza cámaras y software de visión artificial en los estadios para registrar los movimientos de jugadores y pelotas en tiempo real en la cancha, produciendo datos detallados sobre su rendimiento.

deberían rendir mejor (reflejado en un Plus/Minus más alto) cuando sus buenos jugadores están en cancha que cuando no lo están. Sin embargo esta métrica posee una deficiencia clave: **la calificación de cada jugador va a depender en gran medida de la calidad de sus compañeros en el campo**. Por ejemplo: un jugador de rol dentro de un gran equipo va a tener un mayor Plus/Minus que una superestrella en un equipo con mal desempeño.

1.3 RAPM (Regularized Adjusted Plus Minus)

Esta deficiencia anteriormente mencionada se ha intentado solucionar implementando **modelos matemáticos** que incorporen los efectos tanto de los compañeros presentes en cancha como de los rivales en ese momento, buscando aislar el efecto real de los jugadores en la cancha. El primero en realizar una publicación sobre el tema fue Dan Rosenbaum (2004), el cual planteó un **modelo de regresión lineal múltiple** en base a observaciones provenientes de partidos de las temporadas 2002-2003 y 2003-2004 de la National Basketball Association o *NBA*. Como unidad de análisis se utilizaron segmentos de partidos donde no ocurrían sustituciones, la variable respuesta fue la diferencia en el marcador entre el equipo local y el visitante en ese segmento; y como variables explicativas se postularon variables dicotómicas para cada uno de los jugadores que hayan jugado al menos un partido en esas temporadas. Finalmente las métricas de desempeño se correspondían con los coeficientes estimados que acompañaban a las “Dummies” de cada jugador, dichas estimaciones se realizaron mediante la técnica de Mínimos Cuadrados Ordinarios. Estas estadísticas se conocen como Adjusted Plus Minus (APM).

A pesar de que estas métricas presentaban una solución innovadora a la problemática inicial, Rosenbaum señaló que las estimaciones obtenidas no resultaron muy precisas. El modelo propuesto presentaba estimaciones con mucho error, debido a un claro problema de **multicolinealidad** ya que muchos de los jugadores compartían una gran cantidad de minutos juntos en cancha, y se necesitaba de una gran cantidad observaciones de varias temporadas para que el modelo pudiera separar el efecto propio de cada jugador.

Para solucionar esta problemática, se plantea la métrica RAPM introducida por Sill (2010) en la cual se aborda el problema de la Multicolinealidad mediante estimaciones de los parámetros realizadas con Métodos de Regularización, particularmente utilizando la **Regresión Ridge**. Este método de regularización es el más abordado a lo largo de toda la literatura relevante sobre RAPM hasta la fecha.

Otro de los problemas recurrentes en estos modelos está relacionado con la inclusión de jugadores que participan muy pocos minutos en cancha (*LTPs Players*). Cuando un jugador juega muy poco, el modelo solo tiene unas pocas observaciones (segmentos) para evaluarlo, y si particularmente en esas observaciones el equipo tiene un muy buen desempeño (probablemente por casualidad); entonces el modelo no puede diferenciar si el efecto es propio del jugador y ajusta coeficientes extremos. Para solucionar estos casos se han contemplado las siguientes alternativas: 1)excluir a los jugadores con pocos minutos dentro de las temporadas (por ejemplo: Rosembaum (2004) excluye a los jugadores con menos de 250 minutos, Ilardi & Barzilai (2008) excluyen a los jugadores con menos de 300 minutos, etc), 2) utilizar una gran cantidad de observaciones (Ilardi & Barzilai (2008) incorpora información de 5 temporadas), 3) se plantean métodos de estimación de los parámetros que penalicen a este tipo de jugadores.

En los últimos años, muchos investigadores han publicado trabajos orientados a mejorar los resultados de las métricas RAPM. Particularmente en este trabajo usaremos como referencia central al artículo titulado: **Lasso multinomial performance indicators for in-play basketball data** (Damoulaki et.al , 2025), el cual presenta diferencias importantes respecto de los estudios previos. Dicho artículo utiliza **posesiones**³ como unidad de análisis, **puntos** en esas posesiones como variable respuesta y emplea **modelos logísticos** que representan de manera más adecuada la naturaleza de la respuesta.

Además de calcular estas métricas por primera vez para jugadores de Liga Nacional de Básquetbol de Argentina, se implementaran modificaciones con el objetivo de mejorar las estimaciones de nuestro modelo. Entre ellas, se propone asignar mayor peso a las observaciones que ocurren en momentos verdaderamente

³Una posesión comienza (Kubatko et al., 2007) cuando un equipo obtiene el control de la pelota y termina cuando ese equipo cede el control de la misma al equipo rival.

relevantes del partido, incorporar variables explicativas de interés y comparar diferentes métodos de estimación de los parámetros, entre otros ajustes.

2 Motivación

El creciente interés por la analítica deportiva que se ha evidenciado en los últimos años en distintos países del mundo, ha comenzado a captar la atención de las entidades deportivas en Argentina. Particularmente este año la Confederación Argentina de Básquetbol (CABB) y la Asociación de Clubes (AdC) anunciaron en octubre la ampliación de su alianza con Catapult (empresa australiana de tecnología deportiva), integrando equipamiento de wearables (GPS) y software de videoanálisis para las 19 franquicias de la Liga Nacional, las 34 de la Liga Argentina y las selecciones nacionales en todas las ramas. Mientras que en el ámbito educativo, el Instituto CAB (creado en 2023) realizó un convenio con Sports Data Campus (España) para ofrecer, en agosto de este año, el primer curso de *Big Data aplicado al básquetbol* en Argentina, con el objetivo de que entrenadores y staff técnico colaboren en proyectos reales de la CABB, aportando soluciones basadas en datos sobre rendimiento, scouting y optimización de procesos⁴.

Estas cuestiones previamente mencionadas dan indicio de que estamos ante un momento ideal para profundizar en el estudio y la aplicación del análisis de datos en el básquetbol. Se trata de un área que comienza a reconocerse como un componente fundamental tanto para mejorar el rendimiento deportivo de los equipos como para impulsar el desarrollo del negocio en torno al deporte. Sin embargo, el potencial de la analítica aplicada al básquetbol en Argentina aún pareciera encontrarse lejos de estar plenamente aprovechado.

En este marco, y con el fin de caracterizar el estado de situación del análisis de datos en el básquet argentino, se relevan los proyectos e iniciativas actualmente activos en el país vinculados a la estadística y la tecnología aplicada al deporte.

- Facundo Salas y Sebastián Fiol son entrenadores de básquetbol Nivel 3 Eneba, especialistas en estadísticas avanzadas y analistas de datos en equipos profesionales. Ambos son propietarios de la firma *CHAS All Stats* y analistas de datos de Zárate Básquet, equipo perteneciente a la Liga Nacional de Básquetbol de Argentina.
- *Básquet Advance*, dirigido por el entrenador Cristian Sánchez, ofrece informes pre/post partido con estadísticas avanzadas y tendencias clave para entrenadores y scouts.

Ambos proyectos ofrecen herramientas muy importantes y de gran valor tanto para los equipos como para el desarrollo del análisis de datos en el deporte. Pero particularmente en ambos casos, los que llevan adelante estos proyectos son entrenadores de básquetbol, los cuales a pesar de que pueden contar con una sólida formación técnica, no son profesionales de ciencias estadísticas o matemáticas.

Por este motivo, resulta fundamental la incorporación de profesionales en carreras de ciencias exactas, que trabajen en conjunto con los entrenadores y cuerpos técnicos, para de esta manera poder ofrecer nuevos análisis más rigurosos y profundos, así como mejorar las herramientas actualmente disponibles. Esta correspondencia entre el conocimiento técnico-deportivo y el enfoque estadístico permitiría potenciar el desarrollo del básquet argentino y contribuir a elevar su nivel competitivo en un futuro.

⁴<https://www.argentina.basketball/ver/noticia/la-confederacion-argentina-de-basquet-y-sports-data-campus-firman-un-acuerdo-de-colaboracion#:~:text=que%20gracias%20esta%20colaboraci%C3%B3n%2C,el%20rendimiento%20deportivo%2C%20el%20scouting>

3 Objetivos

3.1 Objetivo general

Realizar el cálculo de RAPMs para los jugadores de la Liga Nacional de Básquetbol en Argentina para la temporada 2024/2025. Se buscará obtener un ranking de jugadores que pueda estimar de manera adecuada el aporte realizado por cada uno de los jugadores cuando están en cancha, para que sus equipos sean mejores.

3.2 Objetivos específicos

- 1) **Generación de la base de datos:** Utilizar técnicas de web scraping para construir una base de datos del tipo *play-by-play* para la temporada 2024/2025 de la Liga Nacional de Básquetbol de Argentina.
- 2) **Cálculo de ratings ofensivos (ORAPM) y ratings defensivos (DRAPM):** Plantear un modelo que permita diferenciar el aporte ofensivo y defensivo de cada uno de los jugadores de la liga, obteniendo finalmente rankings que identifiquen a los mejores jugadores en cada rubro.
- 3) **Tratamiento de los jugadores con pocos minutos de juegos (LTPs Players):** Buscar maneras adecuadas para solucionar el problema de los coeficientes extremos asociados a los jugadores con pocos minutos de juego durante la temporada.
- 4) **Determinar una metodología objetiva para evaluar modelos:** Definir criterios de evaluación para poder determinar que modelo proporciona mejores métricas, las cuales deben resumir de manera consistente y precisa el aporte ofensivo y defensivo de cada jugador en relación al desempeño de su equipo.
- 5) **Comparación de modelos en base a la distribución de la respuesta:** Comparar modelos gaussianos de referencia o *benchmarks models* planteados en la literatura relevante con modelos multinomiales que representen de manera más adecuada la naturaleza de nuestra variable de interés.
- 6) **Comparación de modelos en base a la forma de estimación de los parámetros:** Comparar modelos con distintas formas de regularización para estimar los parámetros, comparando los modelos de referencia que utilizan regresión ridge con modelos que utilicen las técnicas: lasso y elastic net.
- 7) **Evaluación del efecto de variables explicativas en el modelo:** Se incorporarán variables exógenas para evaluar si las mismas mejoran los resultados obtenidos por el modelo.

4 Metodología

En base a los objetivos planteados en esta tesina de grado, se compararán resultados provistos por modelos con distintas características, con el fin de poder sintetizar la contribución ofensiva y defensiva de los jugadores de manera adecuada. Para poder identificar que modelo es el que brinda mejores resultados, también se definirá un procedimiento adecuado para evaluarlos.

Los distintos modelos utilizados, las formas de estimar los parámetros y los métodos de comparación de modelos se definen a continuación.

4.1 Modelos propuestos

4.1.1 Modelo de regresión lineal múltiple

Los modelos más comunmente utilizados en la literatura relevante sobre RAPMs son los modelos de regresión lineal múltiple, los cuales van a modelar el comportamiento de la variable respuesta *Puntos en la posesión i*

(y_i) en función de $P = 2K$ variables explicativas dicotómicas, correspondientes a las apariciones ofensivas y defensivas de K jugadores.

El componente aleatorio supone que las respuestas y_i tienen variancias constantes σ^2 , o que las variancias son proporcionales a pesos conocidos y positivos w_i , los cuales brindan la posibilidad de asignar más peso a algunas observaciones que a otras. De esta manera tenemos que,

$$\begin{cases} y_i \sim N(\mu_i, \sigma^2/w_i) \\ \mu_i = \beta_0 + \sum_{j=1}^K \beta_j^o x_{ji}^o + \sum_{j=1}^K \beta_j^d x_{ji}^d \end{cases}$$

para $i = 1, \dots, n$;

Donde:

n : es el número total de posesiones en nuestro dataset, K : es el número total de jugadores en consideración, y_i : es el número de puntos realizados en la posesión i (los puntos en una posesión pueden variar de 0 a 6, siendo los casos en que se anotan 5 o 6 puntos, jugadas extremadamente atípicas), μ_i = es la cantidad esperada de puntos anotados en la posesión i , x_{ik}^o : indicadora ofensiva del jugador k en la posesión i (1 si el jugador k estaba jugando en ataque en la posesión i , 0 en otro caso), x_{ik}^d : indicadora defensiva del jugador k en la posesión i (-1 si el jugador k estaba jugando en defensa en la posesión i , 0 en otro caso), β_0 : Intercepto del modelo (usualmente representa al jugador de referencia, definido como aquel con coeficientes RAPM iguales o cercanos a cero), β_k^o : coeficiente ofensivo del jugador k (*ORAPM*), β_k^d : coeficiente defensivo del jugador k (*DRAPM*).

Como $\hat{\beta}_j$ es una combinación lineal de y_j , entonces la distribución de estos parámetros es conocida. Tenemos que:

$$\hat{\beta}_j \sim N(\beta_j, \text{var}[\hat{\beta}_j])$$

Tanto los parámetros de regresión (β) como la variancia de la respuesta (σ^2) son desconocidos, por lo que deben estimarse a través de los datos.

Este modelo es de sencilla aplicación y ha sido ampliamente utilizado por distintos analistas que trabajan con este tipo de métricas. Sin embargo, la variable respuesta en este caso es discreta, tomando valores enteros entre 0 y 6, lo que sugiere que el supuesto de normalidad del modelo de regresión lineal múltiple no resulta apropiado para representar adecuadamente la naturaleza de la variable de interés.

4.1.2 Modelo logístico multinomial

En consecuencia a lo comentado anteriormente, podría ser más razonable emplear un modelo que contemple la distribución discreta de los datos. Por lo que se procede a la implementación de una regresión logística multinomial para modelar el resultado en puntos de cada posesión.

Recordando que y_i es igual a la cantidad de puntos en una posesión, nuestra nueva variable respuesta resulta:

$$y_i^M = \begin{cases} y_i, & \text{para } y_i \in \{0, 1, 2\} \\ 3, & \text{para } y_i \geq 3 \end{cases}$$

Donde, $y_i^M \sim \text{Multinomial}(n, \boldsymbol{\pi}_i)$ con $\boldsymbol{\pi}_i = (\pi_{i0}, \pi_{i1}, \pi_{i2}, \pi_{i3})$ y $\sum_{c=0}^3 \pi_{ic} = 1$.

Resultando $\pi_{i0}, \pi_{i1}, \pi_{i2}$ y π_{i3} las probabilidades respectivas de no anotar, de anotar 1 punto, de anotar 2 puntos y de anotar 3 puntos o más en la posesión i .

Luego construimos el modelo logístico multinomial con categoría de referencia, eligiendo como categoría de referencia a 0:

$$\text{logit}(\pi_{ic}) = \log\left(\frac{\pi_{ic}}{\pi_{i0}}\right) = \beta_{0c} + \sum_{j=1}^K \beta_{jc}^o x_{ji}^o + \sum_{j=1}^K \beta_{jc}^d x_{ji}^d \quad \text{con } c = \{1, 2, 3\}$$

Dicho modelo describe el efecto de los jugadores en cada uno de los $c - 1 = 3$ logits, es decir, va a permitir calcular el aporte ofensivo y defensivo del jugador para cada tipo de anotación.

Las probabilidades multinomiales se obtienen mediante las siguientes ecuaciones:

$$\pi_{ic} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_i)}{1 + \sum_{c=1}^3 \exp(\mathbf{x}_i^T \boldsymbol{\beta}_i)}$$

4.2 EPTS y wEPTS

4.2.1 EPTS

En el enfoque mencionado anteriormente los coeficientes $\{\beta_1^o, \beta_2^o, \beta_3^o\}$ y $\{\beta_1^d, \beta_2^d, \beta_3^d\}$ representan las contribuciones de cada jugador en los distintos tipos de anotación. Aunque puede resultar interesante y permita realizar distintos análisis, el objetivo final de este estudio es obtener una métrica que unifique la contribución ofensiva y defensiva en una única medida general.

Para la contribución ofensiva, esto puede lograrse calculando la cantidad de puntos esperada en una posesión que anota el equipo de dicho jugador cuando el está incluido en el quinteto y todos los demás jugadores en la alineación pertenecen al grupo de referencia (con RAPM igual a 0).

Para la contribución defensiva de un jugador puede evaluarse mediante la cantidad esperada de puntos (por posesión) concedidos por el equipo de dicho jugador cuando él no está en la alineación y todos los demás jugadores en cancha pertenecen al grupo de referencia.

A esta nueva métrica la denominamos EPTS (*Expected Points per Team possession with the player Scoring contribution*)

$$EPTS_k^r = E(PTS_i \mid X_{ik}^r \neq 0, \mathbf{X}_{i/k}^r = 0, \mathbf{X}_i^{\bar{r}} = 0), \quad r \in \{o, d\}$$

Estas dos medidas, $EPTS^o$ y $EPTS^d$, serán funciones de los coeficientes ofensivos y defensivos del modelo multinomial ajustado. Se obtienen mediante:

$$EPTS_k^r = \pi_{k1}^r + 2\pi_{k2}^r + 3,01\pi_{k3}^r, \quad r \in \{o, d\}$$

Donde π_{kc}^o son las probabilidades estimadas mediante un modelo de regresión logística multinomial de anotar uno, dos o tres puntos o más, respectivamente, por posesión del equipo del jugador k cuando él está incluido en la alineación y todos sus compañeros de equipo en la alineación y todos los oponentes son jugadores incluidos en el grupo de referencia (es decir, aquellos con coeficientes de regresión iguales a cero).

De manera similar, π_{kc}^d son las probabilidades de anotar uno, dos o tres puntos o más, respectivamente, por posesión por parte del equipo oponente del jugador k cuando dicho jugador no está incluido en la alineación, bajo el mismo escenario.

Estas probabilidades vienen dadas por:

$$\pi_{kc}^r = \frac{\exp(\mu_{kc}^r)}{1 + \exp(\mu_{k1}^r) + \exp(\mu_{k2}^r) + \exp(\mu_{k3}^r)} \quad \text{con } \mu_{kc}^r = \beta_{0c} + \beta_{kc}^r$$

4.2.2 wEPTS

Una versión mejorada y ponderada de los EPTS se calcula en *Damoulaki et.al* (2025), la cual tiene en cuenta la participación del jugador a lo largo de la temporada. En la cual la métrica anteriormente presentada se pondera según la proporción de posesiones en las que el jugador de interés estuvo en cancha. Por lo tanto, el EPTS ponderado (wEPTS) se define como:

$$\begin{aligned} wEPTS_k^r &= E(Pts_i \mid \mathbf{X}_{i \setminus k}^r = \mathbf{0}, \mathbf{X}_i^{\bar{r}} = \mathbf{0}, T_i = t_k) \\ &= P(X_{ik}^r \neq 0 \mid T_i = t_k) E(Pts_i \mid X_{ik}^r \neq 0, \mathbf{X}_{i \setminus k}^r = \mathbf{0}, \mathbf{X}_i^{\bar{r}} = \mathbf{0}) \\ &\quad + P(X_{ik}^r = 0 \mid T_i = t_k) E(Pts_i \mid X_{ik}^r \neq 0, \mathbf{X}_{i \setminus k}^r = \mathbf{0}, \mathbf{X}_i^{\bar{r}} = \mathbf{0}) \\ &= W_k^r EPTS_k^r + (1 - W_k^r) EPTS_0 \end{aligned}$$

donde $r \in \{o, d\}$. Además, el peso W_k^r se estima como:

$$\hat{W}_k^r = \frac{n_k^r}{n_{t_k}^r} = \frac{\sum_{i=1}^n \mathcal{I}(X_{ik}^r \neq 0)}{\sum_{i=1}^n \mathcal{I}(T_i^r = t_k)}.$$

Ahora tenemos que $\mathcal{I}(A)$ es una variable indicadora que vale 1 en caso de que A sea verdadero y 0 en caso contrario, T_i^k es el equipo en ataque ($r = o$) o en defensa ($r = d$) en la posesión i, t_k es el equipo del jugador k , n_k^r es el número de posesiones en las que el jugador k participó en el quinteto en ataque o defensa en la temporada y $n_{t_k}^r$ es el número de posesiones en las que el equipo del jugador k está en ataque o defensa en la temporada.

4.3 Estimación de los RAPMs

4.3.1 Método de mínimos cuadrados

En el modelo de regresión lineal múltiple, vamos a estimar los parámetros buscando minimizar las diferencias entre las observaciones de nuestro dataset y las estimaciones del modelo,

Función a minimizar: $S = \sum_{i=1}^n w_i(y_i - \mu_i)^2$

Los estimadores de mínimos cuadrados de β_j se definen como aquellos valores de β_j que minimizan la suma de cuadrados S , y se denotan como $\hat{\beta}_0, \dots, \hat{\beta}_p$.

Obteniendo de esta manera los siguientes coeficientes estimados: $\hat{\beta}_j = \frac{\sum_{i=1}^n w_i x_{ij}^* y_i}{\sum_{i=1}^n w_i (x_{ij}^*)^2}$,

para $j = 0, \dots, p$, donde x_{ij}^* da los valores de la j-ésima variable explicativa x_j después de haber sido ajustada por todas las demás variables explicativas x_0, \dots, x_p excepto x_j . La variable explicativa ajustada x_j^* es aquella parte de x_j que no puede ser explicada mediante una regresión sobre las otras variables explicativas (Dunn y Smyth, 2018).

4.3.2 Método de máxima verosimilitud

En el modelo logístico multinomial, los parámetros $\beta_0, \beta_1, \dots, \beta_p$ se estiman utilizando simultáneamente las (c-1) ecuaciones logit, de manera que se maximice la log-verosimilitud:

Función a maximizar: $L_m(\boldsymbol{\beta}, \mathbf{y}) = \sum_{i=1}^n w_i \sum_{j=1}^c y_{ij} \log(\pi_{ij})$,

Como esto no tiene una solución cerrada, debe recurrirse a algoritmos iterativos como *Newthon-Rapshon* o *Fisher-Scoring* para obtener los resultados.

4.4 Métodos de regularización

Como se ha mencionado anteriormente, el modelo planteado presenta variables predictoras altamente correlacionadas debido a que muchos de los jugadores comparten gran parte del tiempo en cancha, por lo que el ajuste con mínimos cuadrados ordinarios o mediante máxima verosimilitud se torna inestable. Con el objetivo de solucionar este problema se realizaran estimaciones mediante métodos de regularización (o *shrinkage*), los cuales agregan un término de penalización a la función de pérdida para de esta manera controlar la magnitud de los coeficientes estimados. Esta técnica va a provocar que las estimaciones sean sesgadas, pero a cambio de una notable reducción en la variancia.

Las técnicas que se utilizarán en esta investigación son las siguientes:

4.4.1 Regresión Ridge o Regularización L2

Este método incorpora una penalización cuadrática sobre los coeficientes del modelo.

Para el **modelo de regresión lineal múltiple** los coeficientes estimados $\hat{\beta}$ se obtienen minimizando la función de perdida (S) modificada de la siguiente manera:

$$S^* = \sum_{i=1}^n w_i(y_i - \mu_i)^2 + \lambda \sum_{j=1}^p \beta_j^2 = S + \lambda \sum_{j=1}^p \beta_j^2$$

Donde $\lambda \geq 0$ es un *hiperparámetro* que controla la intensidad de la penalización.

Cuando $\lambda = 0$ no hay penalización y la estimación será igual a la obtenida por mínimos cuadrados; y a medida que λ aumenta la penalización se vuelve más dominante y los coeficientes tienden a 0. Para encontrar el valor óptimo de λ se utilizarán técnicas de **validación cruzada**.

Mientras que en el caso del **modelo logístico multinomial**, se agrega un parámetro de penalización a la función de log-verosimilitud, de manera tal que la log-verosimilitud penalizada a maximizar resulta:

$$L_m^*(\beta, \mathbf{y}) = \sum_{i=1}^n w_i \sum_{j=1}^c y_{ij} \log(\pi_{ij}) + \lambda \sum_{j=1}^p \beta_j^2 = L_m(\beta, \mathbf{y}) + \lambda \sum_{j=1}^p \beta_j^2$$

Y el *hiperparámetro* λ se selecciona nuevamente mediante validación cruzada.

4.4.2 Lasso o Regularización L1

A diferencia de la regresión Ridge, en la cual los coeficientes pueden acercarse a 0 pero nunca llegan exactamente a ese valor, la técnica Lasso tiende a forzar algunos parámetros a ser cero, con lo cual también se realiza una selección de las variables/jugadores más influyentes. La intención de aplicar esta técnica esta basada en que puede ser útil para resolver el problema de los coeficientes extremos que presentan los jugadores que juegan pocos minutos. Ademas esto permite una interpretación más significativa del intercepto del modelo implementado, ya que ahora representa la contribución promedio de un jugador de referencia⁵ en cada posesión.

La penalización en este caso se basa en la suma de los valores absolutos de los coeficientes.

Para el **modelo de regresión lineal múltiple** resulta:

$$S^* = \sum_{i=1}^n w_i(y_i - \mu_i)^2 + \lambda \sum_{j=1}^p |\beta_j| = S + \lambda \sum_{j=1}^p |\beta_j|$$

⁵Por jugador de referencia entendemos a cualquier jugador que pertenece al grupo de jugadores con coeficientes iguales a cero.

Para el **modelo logístico multinomial** resulta:

$$L_m^*(\beta, \mathbf{y}) = \sum_{i=1}^n w_i \sum_{j=1}^c y_{ij} \log(\pi_{ij}) + \lambda \sum_{j=1}^p |\beta_j| = L_m(\beta, \mathbf{y}) + \lambda \sum_{j=1}^p |\beta_j|$$

En este caso valores grandes de λ incrementan la penalización, llevando a que un mayor número de coeficientes se reduzcan exactamente a cero.

4.4.3 Elastic Net

Cuando existen predictores altamente correlacionados, **ridge** reduce la influencia de todos ellos a la vez de forma proporcional, mientras que **lasso** tiende a seleccionar uno de ellos, dándole todo el peso y excluyendo al resto. En presencia de correlaciones fuertes esta selección puede variar mucho ante pequeñas perturbaciones , por lo que las soluciones de lasso pueden ser más inestables.

Para encontrar un equilibrio entre ambos métodos se plantea la penalización **elastic net**, la cual combina las penalizaciones L1 (Lasso) y L2 (Ridge).

Para el **modelo de regresión lineal múltiple** resulta:

$$S^* = \sum_{i=1}^n w_i (y_i - \mu_i)^2 + \lambda [\alpha \sum_{j=1}^p |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^p \beta_j^2] = S + \lambda \sum_{j=1}^p [\alpha |\beta_j| + \frac{(1-\alpha)}{2} \beta_j^2]$$

Para el **modelo logístico multinomial** resulta:

$$L_m^*(\beta, \mathbf{y}) = \sum_{i=1}^n w_i \sum_{j=1}^c y_{ij} \log(\pi_{ij}) + \lambda \sum_{j=1}^p [\alpha |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^p \beta_j^2] = L_m(\beta, \mathbf{y}) + \lambda \sum_{j=1}^p [\alpha |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^p \beta_j^2]$$

Donde, $\lambda \geq 0$ sigue siendo el parámetro de regularización que controla la fuerza de la penalización, y se agrega un nuevo hiperparámetro $\alpha \in [0, 1]$ que controla la combinación entre lasso y ridge. Si $\alpha = 1$ se obtiene lasso y si $\alpha = 0$ se obtiene ridge.

4.5 Eficiencia computacional

En el presente estudio trabajaremos con la base de datos *play-by-play* correspondiente a la temporada 2024/2025 de la Liga Nacional de Básquetbol en Argentina. A lo largo de dicha temporada, los 20 equipos participantes disputaron 38 partidos cada uno bajo el formato todos contra todos, por lo que obtendremos datos de 379⁶ partidos. A su vez, en cada uno de estos partidos se registraron aproximadamente 80 posesiones, por lo que el dataset contará con aproximadamente 30.000 observaciones (posesiones).

Por otro lado, cada equipo cuenta con alrededor de 15 jugadores, lo que representará 30 variables dicotómicas por equipo (representando las presencias ofensivas y defensivas de cada uno de los jugadores de la liga). Esto representa un total de 600 variables explicativas en el dataset final. Por lo tanto, la base de datos se compondrá de una matriz cercana a las 30.000 filas y 600 columnas.

En el estudio de *Damoulaki et.al (2025)*, la liga analizada (NBA) presenta una mayor cantidad de partidos, más equipos y consecuentemente más jugadores, obteniendo una base de datos final compuesta con 322.852 posesiones y 717 jugadores (1434 variables dicotómicas). Debido al gran tamaño de los datos, los autores identificaron problemas de eficiencia computacional al ajustar un modelo de regresión logística multinomial, optando en su lugar por estimar tres modelos logísticos binomiales, uno para cada tipo de anotación. Si bien se sabe que las estimaciones basadas en el enfoque de regresiones logísticas separadas son menos eficientes,

⁶Un (1) partido no se disputó por motivos climáticos, y se repartieron los puntos

con errores estándar mayores, en comparación con el enfoque directo de regresión logística multinomial. Pero esta pérdida de eficiencia es menor cuando la categoría de referencia domina en los datos, como en este caso (Agresti, 2013).

En el presente estudio, dado que el volumen de datos es considerablemente menor, se evaluará la opción de realizar la estimación conjunta del modelo multinomial. En el caso de que las limitaciones computacionales sigan apareciendo, se considerará la posibilidad de emplear el motor de procesamiento *Apache Spark*, que permite el manejo eficiente de grandes volúmenes de datos, o recurrir a recursos en la nube (por ejemplo, Google Cloud Platform).

4.6 Comparación de modelos

Una vez que se obtengan los ratings de jugadores a través de las distintas técnicas mencionadas, resulta sumamente importante definir un método para evaluarlos, y de esta manera poder identificar cual es el que ofrece mejores resultados.

En el ámbito del fútbol el profesor Lars Magnus Hvattum plantea una solución muy interesante para esta problemática, la cual no parece haber sido implementada en básquetbol hasta la fecha. En el artículo *Modelling the financial contribution of soccer players to their clubs* (Sæbø & Hvattum, 2019) y en videos de su página de youtube Football Player Ratings se definen 2 criterios diferentes para medir los distintos sistemas de rankings, buscando evaluar la **consistencia** y la **validez** de los mismos.

4.6.1 Consistencia

Según Hvattum, un buen modelo producirá resultados similares independientemente de que partidos hayan sido incluidos en el dataset. Bajo este enfoque, si un mismo sistema de rankings brinda estimaciones de RAPMs similares al ser aplicado a distintas muestras del dataset (de tamaños similares), entonces dicho modelo será consistente.

Para evaluar esto, se divide el conjunto de datos completo en dos particiones con cantidades similares de posesiones. Luego, se calculan los valores de RAPM para cada partición y se estima la correlación de Pearson entre ambos conjuntos de coeficientes. Cuanto mayor sea este coeficiente, mayor será la consistencia del modelo.

Procedure for measuring the repeatability of ratings (reliability)

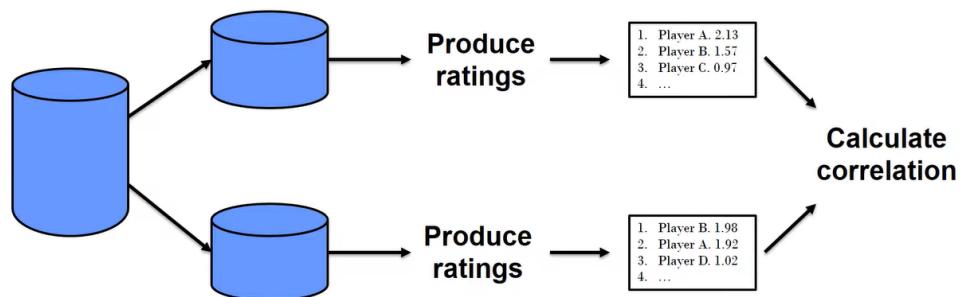


Figure 1: Procedimiento para evaluar la consistencia de los modelos RAPM.

Fuente: Hvattum (2023), *How to evaluate football player ratings?* [Video]. YouTube.

4.6.2 Validez

Hvattum define que un modelo es válido si el conjunto de estimaciones que se obtienen para los jugadores puede representar correctamente el aporte que estos realizan a sus equipos, en términos de obtener victorias. Para poder identificar esto realiza el supuesto de que los ratings de los jugadores tienen una conexión directa con el desempeño de sus equipos, es decir, que los equipos que tengan un rating promedio más alto deberían tener una mayor probabilidad de ganar los partidos. Esta forma de evaluar los modelos ya había sido propuesta por uno de los artículos fundacionales de esta temática (Ilardi & Barzilai, 2008), pero no ha sido explorada en profundidad en artículos relacionados al básquetbol.

La idea general se basa en plantear un modeo de regresión logística binario, donde cada observación corresponde a un partido completo, la variable respuesta es el resultado del encuentro (1 si el equipo local gana, 0 si pierde) y como única variable explicativa tendremos la diferencia entre los promedios de los ratings RAPM de los jugadores de ambos equipos (es decir, la diferencia entre los ratings promedios del local y visitante).

Para ello, se estiman los valores de RAPM utilizando los datos de, por ejemplo, la primera mitad de la temporada. Posteriormente, se ajusta el modelo logístico con un conjunto de partidos posteriores y, finalmente, se evalúa la capacidad predictiva del modelo sobre los partidos restantes, empleando métricas de desempeño como el AUC, Accuracy, etc.

Measuring the ability of ratings to predict match outcomes (validity)

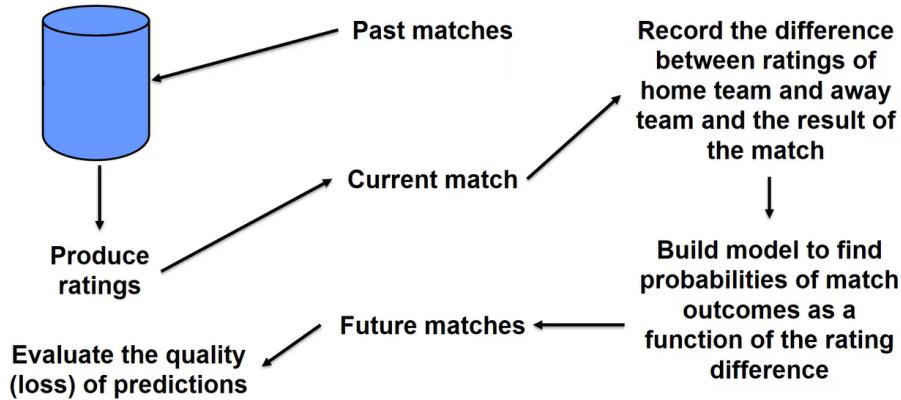


Figure 2: Procedimiento para evaluar la validez de los modelos RAPM.

Fuente: Hvattum (2023), *How to evaluate football player ratings?* [Video]. YouTube.

5 Aplicación

5.1 Generación y preprocesamiento de la base de datos

Una parte central de esta tesina estará dedicada a la construcción de una base de datos *play-by-play* correspondiente a la temporada 2024/2025 de la Liga Nacional de Básquetbol de Argentina. Este tipo de base registra todas las acciones ocurridas dentro de un partido. Dado que esta información no se encuentra disponible de forma accesible, será necesario recurrir a técnicas de web scraping para su obtención. Los datos necesarios se extraerán del Sitio Oficial de la Liga Nacional de Básquet, utilizando librerías de Python como Selenium y BeautifulSoup.

En primera instancia, se automatizará el proceso de identificar cada uno de los enlaces de los 379 partidos de temporada regular y posteriormente se iterará sobre los mismos para obtener la información de cada una de las jugadas del partido. Cabe destacar que en nuestro dataset final sin procesar cada fila se corresponderá con cada una de las jugadas o acciones que sucedan dentro del partido, por lo cual posteriormente será necesario realizar un proceso de preprocesamiento de los datos que permita agrupar dichas acciones en **posesiones**, entendidas como el conjunto de jugadas ocurridas antes de que cambie el control de la pelota.

A través de este proceso, las variables que se registrarán son: la cantidad de puntos que se anotaron en esa posesión, los jugadores en cancha en ese momento, la condición de localía del equipo en posesión de la pelota, el momento del partido (minutos restantes), el marcador en ese momento de partido y los equipos involucrados.

Finalmente, luego de un proceso de limpieza y estructuración de los datos, se obtendrá una base que pueda ser utilizada en la implementación de los modelos estadísticos mencionados en las secciones anteriores.

5.2 Estimación de los RAPMs

En esta segunda etapa, se utilizará la base de datos ya preprocesada para obtener los ratings ofensivos y defensivos de cada uno de los jugadores que hayan jugado al menos 1 partido a lo largo de la temporada analizada, estos ratings se corresponderán con las estimaciones de los parámetros de los modelos planteados.

Tanto el modelo de regresión lineal múltiple como el modelo logístico multinomial, serán ajustados utilizando la librería `glmnet` del software estadístico R. Esta librería permite la estimación de parámetros mediante técnicas de regularización como Ridge, Lasso y Elastic Net, lo que resulta particularmente útil para manejar colinealidad entre variables y para la selección automática de predictores.

Por otro lado, en caso de presentarse limitaciones computacionales que dificulten el ajuste de los modelos sobre la totalidad de los datos, se considerará la implementación de los mismos en Apache Spark a través de la librería `sparklyr`, lo que permitirá distribuir el procesamiento y manejar grandes volúmenes de información de manera eficiente.

5.3 Comparación de los distintos enfoques

En esta última etapa, se procederá a la evaluación y comparación de los modelos ajustados, siguiendo la metodología propuesta por Hvattum pero aplicada al básquetbol. Cada uno de los distintos modelos será evaluado en función de la consistencia que presenten y la validez de los mismos. Los resultados se presentarán mediante gráficos comparativos que faciliten la visualización de las diferencias entre enfoques y estará acompañado de un análisis personal basado en la experiencia propia sobre los jugadores de la liga, permitiendo contextualizar los resultados y destacar aspectos relevantes que puedan no ser capturados únicamente por los modelos estadísticos.

6 Cronograma de actividades

Table 1: Cronograma de actividades.

Actividades	Fecha.estimada
Definición del tema de investigación.	Junio 2025 - Agosto 2025
Revisión bibliográfica y recopilación de antecedentes.	Junio 2025 – Septiembre 2025
Generación de la base de datos mediante técnicas de *web scraping*.	Septiembre 2025 - Noviembre 2025
Limpieza y preprocesamiento de los datos.	Noviembre 2025 - Diciembre 2025
Implementación de los modelos y análisis preliminar de los resultados.	Diciembre 2025 – Enero 2026
Comparación de modelos y elaboración de conclusiones finales.	Enero 2026 - Febrero 2026
Revisión general del trabajo y realización de modificaciones.	Febrero 2026 - Marzo 2026
Redacción del informe final.	Febrero 2026 - Marzo 2026
Presentación de la tesina.	Abril 2026

7 Bibliografía

- Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). Wiley Series in Probability and Statistics. Wiley, New York.
- Damoulaki, A., Ntzoufras, I. & Pelechrinis, K. *Lasso multinomial performance indicators for in-play basketball data*. Comput Stat 40, 2157–2181 (2025).
- Dunn, P. K., & Smyth, G. K. (2018). *Generalized Linear Models with Examples in R* (2nd ed.). Springer.
- Hvattum, L. M. (2019). *A comprehensive review of plus-minus ratings for evaluating individual players in team sports*. International Journal of Computer Science in Sport, 18, 1–23.
- Ilardi, S., & Barzilai, A. (2007). *Adjusted Plus-Minus: An idea whose time has come*. Blog: 82games. <http://www.82games.com/ilardi1.htm>
- Ilardi, S., & Barzilai, A. (2008). *Adjusted Plus-Minus ratings: New and improved for 2007–2008*. Blog: 82games. <http://www.82games.com/ilardi2.htm>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (ISLR2).
- Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. (2007). *A starting point for analyzing basketball statistics*. Journal of Quantitative Analysis in Sports, 3, Article 1.
- Rosenbaum, D. (2004). *Measuring how NBA players help their teams win*. Blog: 82games. <http://www.82games.com/comm30.htm>
- Sæbø, O., & Hvattum, L. (2019). *Modelling the financial contribution of soccer players to their clubs*. Journal of Sports Analytics, 5, 23–34.
- Sill, J. (2010). *Improved NBA adjusted +/- using regularization and out-of-sample testing*. Proceedings of the 2010 MIT Sloan Sports Analytics Conference.
- Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.