Métrica RAPM: Evaluación del impacto ofensivo y defensivo de los jugadores de la Liga Nacional de Básquetbol en Argentina

Anteproyecto de Tesina

Estudiante: Simón Pedro Gazze

Director: Andrés Sosa

Co-directora: Cristina Cuesta

Fecha: 18 de febrero de 2026

UNIVERSIDAD NACIONAL
DE ROSARIO
FACULTAD DE CIENCIAS
ECONÓMICAS Y ESTADÍSTICA

Contenido

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 del año 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* (resumen estructurado de los resultados de un partido), el desarrollo de bases de datos *play-by-play* (registro de cada una de las acciones del partido) 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 desempempeño 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, ya que, 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.

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

[1]Término 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*).

[2]El *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)

La 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 ocurrian 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 problematica 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 está problemática, se plantea la métrica RAPM introducida por Sill (2010) en la cual se aborda el problema de la Multicolinealidad mediante estimaciones de los parametros realizadas con Métodos de Regularización, particularmente utilizando la **Regresión Ridge**. Esté 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 esta 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 estima 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 lo jugadores con menos de 250 minutos, Ilardi (2007) excluyen a lo jugadores con menos de 400 minutos, etc), 2) utilizar una gran cantidad de observaciones (Ilardi & Barzilai (2008) incorpora información de 5 temporadas), 3) plantear 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 multinomiales** 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

[3]Una 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 paises del mundo, ha empezado 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 procesos[4].

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 relevaron 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, 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 estadística o matemática.

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%2C%20gracias%20a%20esta%20colaboraci%C3%B3n%2C,el%20rendimiento%20deportivo%2C%20el%20sc

3 Objetivos

3.1 Objetivo general

El objetivo general de esta tesina es estimar el aporte individual de los jugadores de la Liga Nacional de Básquetbol de Argentina durante la temporada 2024/2025 mediante modelos RAPM, con el fin de construir un ranking que refleje su impacto en el desempeño de sus equipos.

3.2 Objetivos específicos

1) Generar una base de datos del tipo *play-by-play* para la temporada 2024/2025 de la Liga Nacional de Básquetbol de Argentina, mediante técnicas de *web scraping*.
2) 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) Desarrollar una metodología adecuada que permita solucionar el problema de los coeficientes extremos asociados a los jugadores con pocos minutos de juego (*LTPs Players*) durante la temporada.
4) Comparar modelos en base a distintos criterios de evaluación, para poder determinar cual de ellos proporciona mejores métricas, debiendo resumir de manera consistente y precisa el aporte ofensivo y defensivo de cada jugador en relación al desempeño de su equipo.

4 Metodología

En base a los objetivos planteados, 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 modelan 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,

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

para $i = 1, ..., n$;

donde:

n: es el número total de posesiones en el *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, β_k^o: coeficiente ofensivo del jugador k ($ORAPM$), β_k^d: coeficiente defensivo del jugador k ($DRAPM$).

Como $\hat{\beta}_k$ es una combinacion lineal de y_i , entonces la distribución de estos parámetros es conocida. Tenemos que:

$$\hat{\beta}_k \sim N(\beta_k, var[\hat{\beta}_k])$$

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, $y_i^M \sim Multinomial(n, \boldsymbol{\pi_i})$ 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:

$$logit(\pi_{ic}) = log(\frac{\pi_{ic}}{\pi_{i0}}) = \beta_{0c} + \sum_{k=1}^{K} \beta_{kc}^o x_{ik}^o + \sum_{k=1}^{K} \beta_{kc}^d x_{ik}^d \quad \text{con } c = \{1,2,3\}$$

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

Las probabilidades multinomiales se obtienen mediante las siguientes ecuaciones:

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

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 él está incluido en el quinteto y todos los demás jugadores en la alineación pertenecen al grupo de referencia (con RAPM igual a 0).

Mientras que 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 ETPS (*Expected Points per Team possession with the player Scoring contribution*)

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

Estas dos medidas, $ETPS^o$ y $ETPS^d$, 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:

$$wEPTS_k^r = W_k^r \, EPTS_k^r + (1 - W_k^r) \, EPTS_0$$

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, se van a estimar los parámetros por mínimos cuadrados ponderados, siendo la función a minimizar:

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

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

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

para $j = 0, ..., p$, donde x_{ik}^* son los valores de la k-ésima variable explicativa x_k después de haber sido ajustada por todas las demás variables explicativas $x_0, ..., x_p$ excepto x_k. La variable explicativa ajustada x_k^* es aquella parte de x_k 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 $\beta_0, \beta_1, ..., \beta_p$ se estiman utilizando simultáneamente las (c-1) ecuaciones logit, de manera que se maximice la log-verosimilitud:

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

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 provoca 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 coficientes estimados $\hat{\boldsymbol{\beta}}$ 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:

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

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, presentada por Robert Tibshirani (1996), 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 está 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 referencia[5] 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:

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

Para el **modelo logístico multinomial** resulta:

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

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**, introducido por Hui Zou y Trevor Hastie (2005), la cual combina las penalizaciones L1 (Lasso) y L2 (Ridge).

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

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

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

Para el **modelo logístico multinomial** resulta:

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

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 se obtienen datos de 378[6] partidos. A su vez, en cada uno de estos partidos se registraron aproximadamente 80 posesiones por equipo, obteniendo 160 posesiones totales. De manera que el *dataset* final contará con aproximadamente 60.000 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 60.000 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, 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).

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`.

[6] Dos (2) partidos no se disputaron por motivos climáticos, y se repartieron los puntos

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 correspondera con cada una de las jugadas o acciones que sucedan dentro del partido, por lo cual posteriormente sera 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 corresponderan con las estimaciones de los parámetros de los modelos planteados.

Tanto el modelo de regresión lineal múltiple como el modelo logístico multinomial, seran ajustados utilizando la libreria `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 (James et al., 2021).

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 (Sæbø, O., & Hvattum, L.M., 2019; Hvattum, L.M., 2019) pero aplicada al básquetbol, donde cada uno de los distintos modelos serán comparados en función de la validez y consistencia 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

Cuadro 1: Cronograma de actividades.

Actividades	Fecha estimada
Definición del tema de investigación.	Julio 2025 - Septiembre 2025
Revisión bibliográfica y recopilación de antecedentes.	Julio 2025 – Octubre 2025
Generación de la base de datos mediante técnicas de *web scraping*.	Octubre 2025 - Diciembre 2025
Limpieza y preprocesamiento de los datos.	Diciembre 2025 - Enero 2026
Implementación de los modelos y análisis preliminar de los resultados.	Enero 2026 – Febrero 2026
Comparación de modelos y elaboración de conclusiones finales.	Febrero 2026 - Marzo 2026
Revisión general del trabajo y realización de modificaciones.	Marzo 2026 - Abril 2026
Redacción del informe final.	Marzo 2026 - Abril 2026
Presentación de la tesina.	Mayo 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. (2025). *Lasso multinomial performance indicators for in-play basketball data.* Comput Stat 40, 2157–2181.

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in R. 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. (2007). *Adjusted Plus-Minus: An idea whose time has come.* Blog: 82games. http://www.82games.com/ilardi1.htm. Consultado el 17 de febrero de 2026.

Ilardi, S., & Barzilai, A. (2008). *Adjusted Plus-Minus ratings: New and improved for 2007–2008.* Blog: 82games. http://www.82games.com/ilardi2.htm. Consultado el 17 de febrero de 2026.

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. Consultado el 17 de febrero de 2026.

Sæbø, O., & Hvattum, L.M. (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.

Zou, H., & Hastie, T. (2005). *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society: Series B, 67(2), 301–320.