

Módulo 2. Bases y principios del análisis de datos aplicados al rendimiento físico



☰ Bases y principios del análisis de datos aplicados al rendimiento físico

☰ Referencias

Bases y principios del análisis de datos aplicados al rendimiento físico

El método científico

En el módulo anterior, hemos detallado, principalmente, cuáles son las propiedades y funcionalidades de RStudio, es decir, hemos hecho una introducción a una herramienta que nos permitirá ser más eficientes en nuestro día a día como *sport scientists*. Sin embargo, la idea que se ha recalcado durante el módulo anterior es que uno de los elementos fundamentales e imprescindibles del proceso de uso de dicha herramienta es tener una pregunta adecuada a la que responder. Debemos identificar cuál es la necesidad de nuestro contexto y usar los datos de la mejor manera posible y traducirlos, para aportar información de valor a los distintos departamentos, con objetivo final de ayudar a los jugadores a alcanzar esa necesidad o ese elemento determinante para su desarrollo (Bartlett y Drust, 2021).

Estamos hablando, fundamentalmente, del método científico, un método que permite ser objetivos en la toma de decisiones y seguir un

proceso marcado y específico para resolver los problemas que se plantean en nuestro marco profesional y que nos obligará a cumplir ciertos estándares de calidad para asegurarnos de que los resultados sean útiles y fiables.

En este módulo, trataremos de detallar cuál es el proceso que seguir desde el momento en el que se plantea un problema, centrándonos en el papel que juega el análisis de datos y describiendo cuáles son los principios y requerimientos de este ámbito. También definiremos cuáles son sus particularidades considerando el entorno del rendimiento físico.

French y Torres Ronda (2021) definen el método científico para ciencias del deporte de una manera muy visual, en la cual se trata el círculo que se repetirá una vez que se haya llegado al último paso:

- Observar.
- Definir.
- Construir una hipótesis.
- Experimentar.
- Interpretar.

- Implementar.

¿Qué papel juega el análisis de datos en este proceso/método?

Como podemos ver en los puntos anteriores, el dato juega un papel fundamental en el desarrollo del método científico. Vamos a englobar los dos primeros pasos que hemos detallado anteriormente en uno: observar y definir. Teniendo en cuenta el contexto en el que nos encontramos, deberemos definir cuáles son los objetivos que queremos conseguir. Estos objetivos pueden tener múltiples orientaciones relacionadas con el rendimiento. Podemos hablar también de distintas capas de complejidad y objetivos derivados de ellas.

Sin embargo, la elección de las preguntas u observaciones dependerá de cada contexto y cada profesional, y no pertenece a los contenidos de este curso, en el que nos centraremos en cómo podemos ser eficientes en la utilización del análisis de datos para resolver estas preguntas. Nuevamente, cabe destacar que el dato o el análisis no es el único medio para conseguir las respuestas, pero queremos sacar el máximo provecho de él, para que nos permita acercarnos a la complejidad del rendimiento deportivo de nuestros atletas (Goes *et al.*, 2021).

De manera simplificada, el análisis de datos es la transformación de los datos registrados en información aplicable y con impacto.

Los seis pasos del método científico son:

- Observar
- Definir
- Construir hipótesis
- Experimentar
- Interpretar
- Implementar
- Explicar
- Responder preguntas

SUBMIT

Cualidades del dato

El crecimiento y desarrollo tecnológico de los últimos años ha dado lugar a la aparición de conceptos como *big data*, muchas veces malinterpretado, ya que se le suele dar interpretaciones que van más allá de su descripción original. El *big data* es un conjunto de datos que debe cumplir una serie de características y, aunque no todas sean aplicables rigurosamente en el escenario del rendimiento deportivo, sí que nos pueden servir de referencia para describir al «dato» dentro de nuestro contexto profesional (Rein y Memmert, 2016). Originalmente, se describió al *big data* con las tres «v»:

- Volumen. Gran cantidad de datos, que también requieren cierta capacidad de almacenamiento (podemos hablar de bases de datos o sistemas en la nube de proveedores).
- Variedad: los datos que obtenemos son de distinto formato (texto, numérico, series temporales).

- Velocidad: el ritmo en el que se recibe la información es muy alto, lo que actualiza, nuevamente, la información que tenemos.

Cuando consideramos nuestro contexto deportivo, vemos que estos tres puntos se cumplen, por supuesto, en muchos casos, en menor medida que en otros ámbitos profesionales, pero no por ese motivo debemos ser menos rigurosos en nuestro proceso si queremos sacar el mayor rendimiento de los proyectos en los que trabajemos.

Más recientemente, se han añadido otras tres v al término *big data*:

- Veracidad. Calidad de los datos. Esta característica vendrá dada en medida por nuestro plan de recogida de datos y la elección de la herramienta a utilizar.
- Valor: no únicamente que la información aporte información relacionada con lo que queremos investigar, sino que sea práctica y realista su recogida.
- Visualización: concepto que habla de una de las últimas fases del análisis de datos, la comunicación; por lo tanto, deberemos completar una serie de pasos antes de poder llegar a esta parte final, pero es uno de los objetivos finales del dato, que puede tener

mayor impacto e influencia hacia los otros departamentos.

Como hemos comentado, los tres primeros puntos son altamente dependientes del contexto en el que nos encontremos; sin embargo, en los tres últimos puntos, podemos tener mayor impacto. Debemos ser meticulosos en garantizar que la veracidad y calidad de datos se cumplan. Debemos seleccionar qué datos están más relacionados con las preguntas que queremos responder, así como qué tipo de análisis queremos utilizar. Estos dos puntos aumentarán el valor de nuestros datos y harán que nuestro análisis tenga un mayor impacto. Ya en la parte final del proceso, deberemos escoger qué tipo de visualización es la más adecuada para la información que queremos compartir. Las herramientas como RStudio nos darán la capacidad de automatizar la creación de visualizaciones y también de escoger la mejor forma de presentación de estas.

¿Cuáles son las 3 V añadidas recientemente que describen al big data?

Veracidad

Valor

Visualización

Vivacidad

Variedad

SUBMIT

Calidad del dato

Conseguir «calidad del dato», es decir, lo que hemos descrito anteriormente sobre veracidad y valor del dato, pasa por lo que se conoce en el ámbito de la ciencia de datos como «higiene del dato», que describiremos con mayor detalle en el siguiente módulo. Sin embargo, vamos a destacar, en este apartado, aspectos teóricos fundamentales, relacionados con las herramientas de medida, que garantizan el primer filtro de calidad del dato y su obtención, para después seguir con el proceso de limpieza y análisis.

Las herramientas que debemos usar deben proporcionar medidas fiables, es decir, que sean reproducibles en situaciones similares. Impellizzeri y Marcora (2009) destacan dos tipos de fiabilidad:

- Fiabilidad absoluta, es decir, grado en que los registros varían por atleta.
- Fiabilidad relativa: grado en que los atletas mantienen su posición respecto al grupo en medidas repetidas.

Ambos aspectos son a considerar para decidir qué prueba o herramienta queremos utilizar. En otras palabras, y volviendo a hacer referencia a la importancia del planteamiento de la pregunta, según la investigación que queramos plantear, deberemos escoger herramientas que tengan mayor fiabilidad absoluta o relativa. Por ejemplo, si el objetivo es un estudio transversal de nuestro equipo, en el que queremos detectar diferentes perfiles condicionales entre nuestros jugadores, deberemos escoger herramientas con alta fiabilidad relativa. Sin embargo, si el objetivo es evaluar cambios en el perfil condicional a lo largo de la temporada, la herramienta a escoger deberá tener alta fiabilidad absoluta, para que sea capaz de detectar cambios reales a lo largo del período de estudio. Existen pruebas estadísticas para determinar dichas fiabilidades.

La segunda característica que destacar es la validez, esta propiedad se refiere a que el test esté midiendo un parámetro relevante a lo que pretendemos medir, de manera exacta y con precisión. Para determinar que una herramienta es válida, debemos comparar sus resultados con herramientas clasificadas como *gold standard*.

Vamos a ver un ejemplo de validez en este caso muy común en el ámbito de la biomecánica, como es la creación de variables a partir del seguimiento de un movimiento. Gracias al uso de sensores de captación de movimiento o de sistemas de *tracking* de alta frecuencia, podemos disponer del seguimiento de ciertos segmentos del cuerpo durante gestos específicos deportivos. Una aplicación muy común en biomecánica es poder determinar ciertas variables o *features* en el movimiento o gesto deportivo específico que permitan identificar cuál es el movimiento óptimo para conseguir el mejor resultado. Por ejemplo, si estuviéramos analizando el saque de un tenista, podríamos obtener el seguimiento de los diferentes puntos articulares y, en consecuencia, ver el movimiento del brazo. De este seguimiento, se podrían obtener variables como altura máxima a la que el jugador lleva el brazo en la fase de armado, velocidades angulares antes del golpe, etcétera. Puede ser un objetivo detectar cuál de estas variables es más relevante para poder golpear la pelota de la manera más eficiente posible o en ciertas direcciones. Ajustar el gesto.

Si queremos que la información que obtenemos sea útil y aplicable, debemos usar las variables que nos den información válida, es decir,

que sean capaces de mostrar diferencias entre distintos jugadores, si partimos de la premisa de que los jugadores tienen características distintas (efecto que le dan a la pelota, velocidad...). Si cualquiera de las variables que hayamos calculado no muestra diferencias entre los distintos jugadores y, por lo tanto, no nos permiten aportar información a cómo afectan a las variables de salida (velocidad y efecto), no las consideraríamos válidas o útiles.

Una vez finalizada la recolección de datos siguiendo los estándares de calidad, debemos proceder a la correcta limpieza y tratamiento; estos dos pasos requieren mucho detalle y son altamente específicos al tipo de dato con el que tratamos, ya sea texto, numérico, etc. Este proceso lo trataremos en un módulo más adelante, ya que es suficientemente extenso y consiste, en gran parte, en el trabajo de un *sport scientist*. Así pues, suponiendo que hemos hecho el paso anterior, debemos seleccionar cuál es el análisis estadístico que mejor encaja con el problema que hemos planteado. Para eso, vamos a ver qué opciones de análisis tenemos (Houtmeyers *et al.*, 2021).

Tipos de análisis: descriptivo

Este tipo de análisis se considera el inicial o básico al empezar a tratar con datos. Nos permiten ver, de forma resumida, cual es el comportamiento de los datos que queremos analizar. Forma parte del

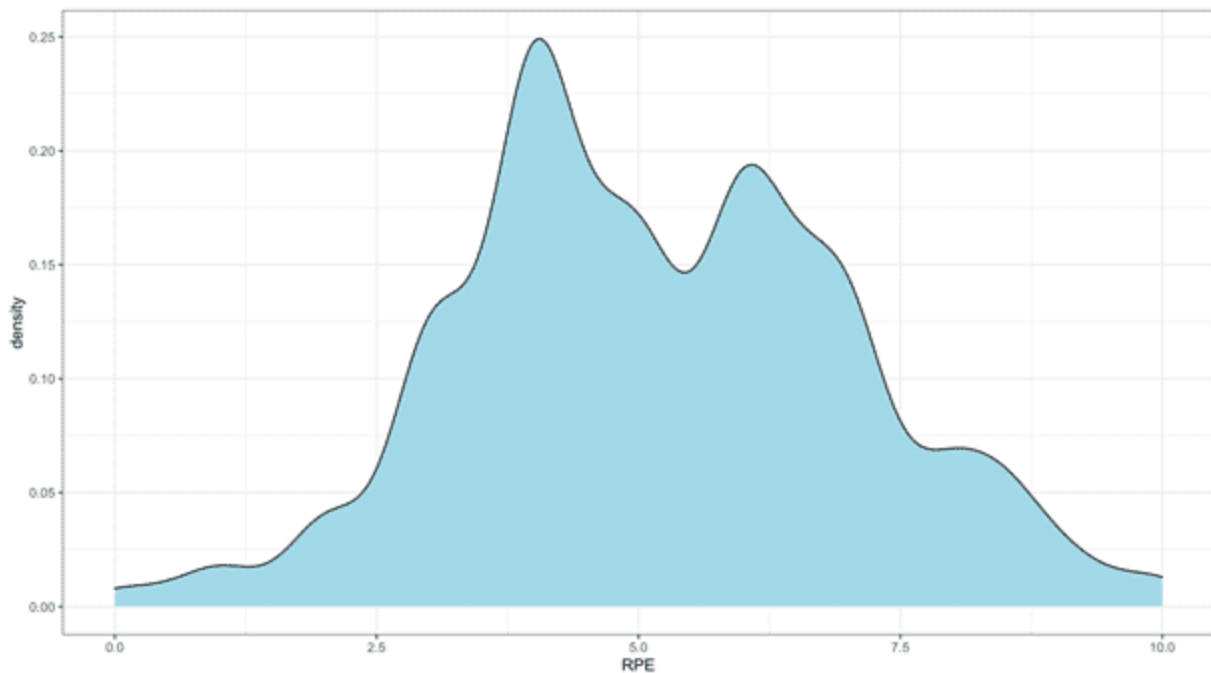
proceso de exploración de los datos en muchos casos, antes de que sean tratados o modificados. Aun así, una vez que se tengan los datos «finalizados», este tipo de análisis puede ser de mucho valor por sí mismo y ofrece una gran eficiencia, ya que no son análisis costosos respecto al tiempo dedicado en ellos. Hay que tener en cuenta que este tipo de análisis únicamente muestra las características de los datos que estamos analizando, y, por lo tanto, deberemos ser muy cautelosos a la hora de generalizar.

Para describir cuáles son los tipos de análisis descriptivos, utilizaremos ejemplos específicos.

- **Medidas de frecuencia:** nos permiten saber cómo se distribuyen nuestros datos de acuerdo con una métrica, como, por ejemplo, si queremos saber qué respuestas son las más comunes en el cuestionario RPE los días después de partido para nuestros jugadores. En la imagen a continuación, podemos ver cómo las respuestas se concentran en dos valores (picos más altos) el 4 y el 7, aproximadamente, lo que muestra que puede haber dos tipos de sesiones en las cuales hay dos grupos de jugadores que perciben el entrenamiento más intenso y el otro de menor exigencia. En el caso de usar un histograma como en la imagen, esto permite evaluar también la forma y distribución de los datos, lo que aporta información

sobre si hay valores sesgados hacia una dirección u otra.

Figura 1: Medidas de frecuencia



Fuente: Elaboración propia.

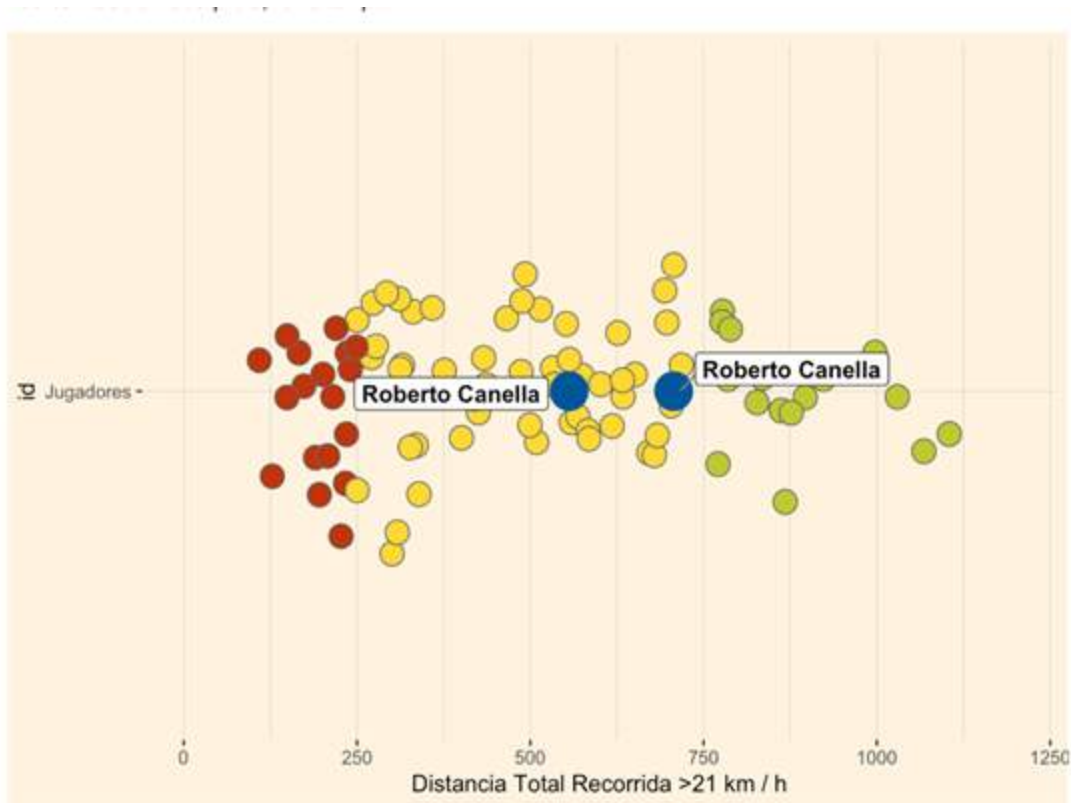
- Medidas de tendencia central: describe cuál es la respuesta central; por ejemplo, si queremos saber cuál es la distancia recorrida promedio por un defensa central en un partido, calcularemos la media de todas las distancias totales en los partidos que el jugador haya completado el tiempo reglamentario.

Las medidas de tendencia central más comunes son las siguientes:

- Media.
 - Mediana.
 - Moda.
- Medidas de dispersión: permite conocer el rango en el que se distribuyen los datos, su variabilidad normal. Si, considerando el ejemplo anterior, en lugar de tener un único valor describiendo las demandas de partido, queremos saber en qué rango se mueve el jugador habitualmente, podemos calcular la desviación estándar para ver qué valores son habituales y detectar fácilmente qué partidos han sido de mayor exigencia. Esta medida sirve también para identificar posibles valores extremos u *outliers*. Las medidas de tendencia central más comunes son las siguientes:
 - Desviación estándar.
 - Rango.
 - Rango intercuartil.

- Medidas de posición: permiten clasificar, ordenar según el valor analizado a los atletas dentro de la fuente de datos analizada. En el gráfico a continuación, se muestra dónde se sitúan dos jugadores (azul) en comparación con el resto de los jugadores de su misma posición en la liga, usando la variable de distancia a alta velocidad. El color de los puntos muestra también cuáles son los jugadores dentro del rango de dispersión estándar (amarillo) y cuáles están por arriba (verde) o debajo (rojo).

Figura 2: Medidas de posición



Tipos de análisis: diagnóstico

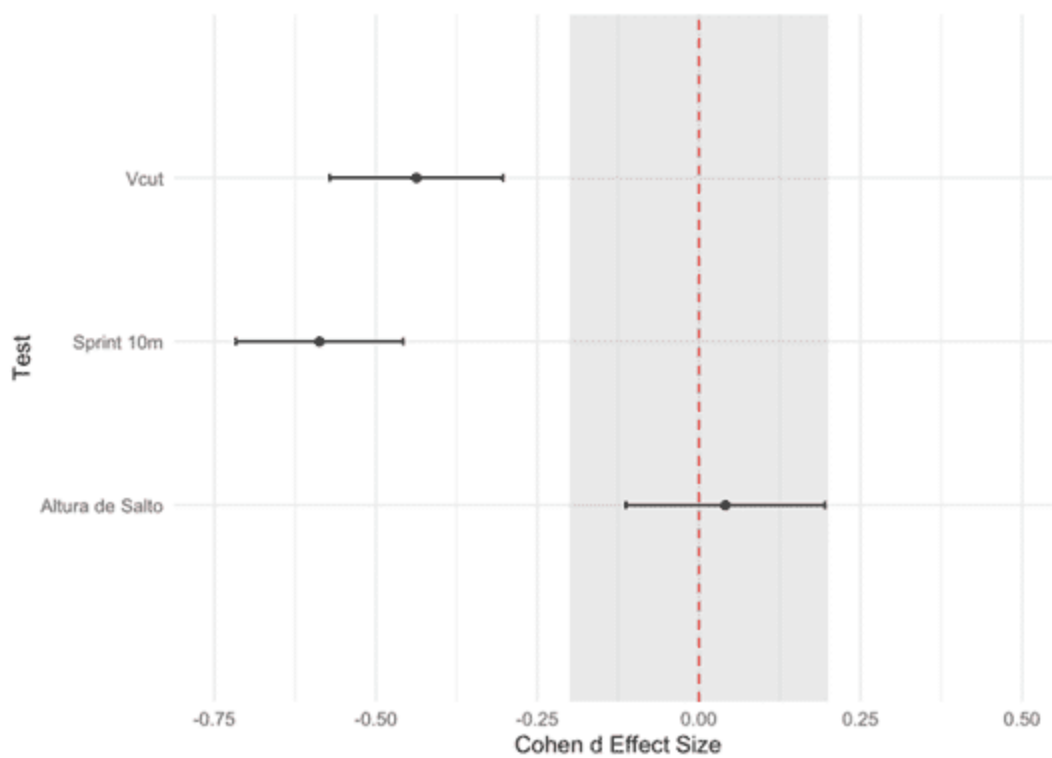
Este análisis pretende hacer una comparación de medidas. Son los tipos de análisis que podemos utilizar para objetivar si hay diferencias reales, estadísticamente significativas entre dos grupos, o dos momentos de la temporada y, por lo tanto, determinar si las intervenciones que estamos haciendo son las correctas. Sin embargo, la identificación de las características de las medias es fundamental para determinar qué tipo de análisis estadístico corresponde.

- Comparación de medidas/proporciones de una muestra: se usa cuando queremos comparar los valores promedio de nuestra fuente de datos con un valor de referencia conocido. Por ejemplo, si queremos comparar nuestro grupo de jugadores en los resultados del *test* 30-15 con los valores de referencia/normativos en el mismo grupo de edad.
- Comparación de medidas/proporciones de grupos independientes: es el mismo ejemplo que el anterior, pero, en este caso, en lugar de comparar con un valor de referencia, comparamos a dos equipos de

categorías inferiores, para determinar si hay diferencias entre ellos.

- Comparación de medidas repetidas: en este caso, queremos determinar si ha habido un cambio antes y después de hacer una intervención. En la siguiente imagen, se puede ver el cambio, representado por el tamaño del efecto en tres pruebas distintas para un grupo de jugadores determinado.

Figura 3: Análisis diagnóstico

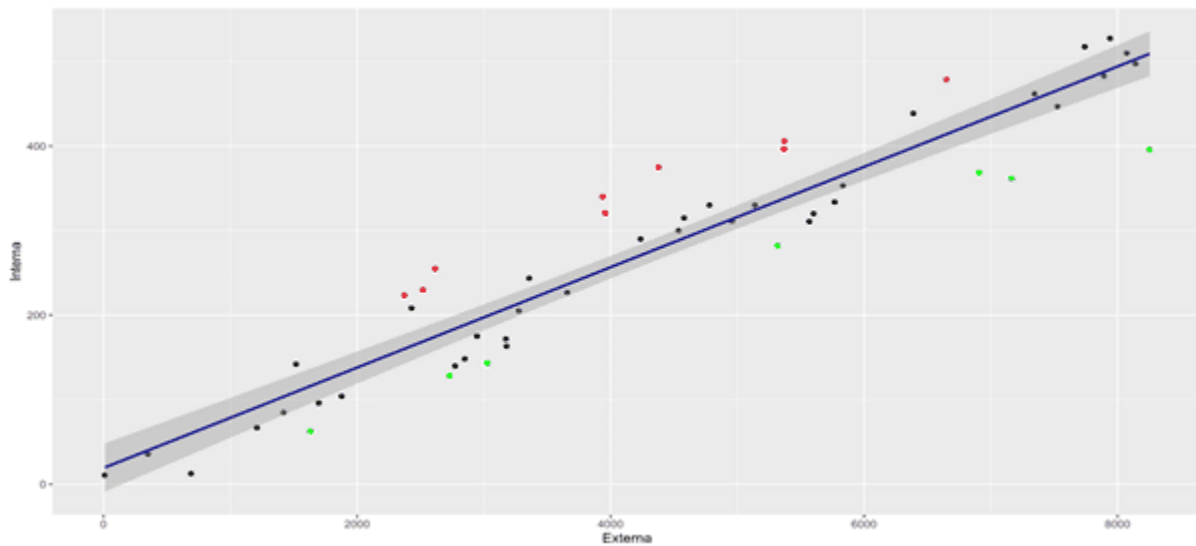


Tipos de análisis: predictivo

El análisis predictivo pretende usar los datos de los que disponemos, encontrar relaciones entre variables para poder extrapolar dichas relaciones en situaciones futuras. Hay que tener siempre en cuenta que estas predicciones se tratan de estimaciones, por lo que hay que considerar que hay un error asociado a las relaciones o los modelos estadísticos que utilicemos. Cuando comuniquemos estos resultados, una buena práctica es comunicar cuál es el error asociado a cada una de las estimaciones que estamos haciendo.

En el siguiente ejemplo, podemos ver la relación entre dos variables: una indica la carga interna del jugador y la otra la carga locomotora o externa. Conociendo esta relación entre las variables, seremos capaces de prescribir una cierta carga externa (controlable) según la carga interna que deseemos conseguir, la cual llevará a una adaptación específica.

Figura 4: Análisis predictivo



Fuente: Elaboración propia.

Tipos de análisis: prescriptivo

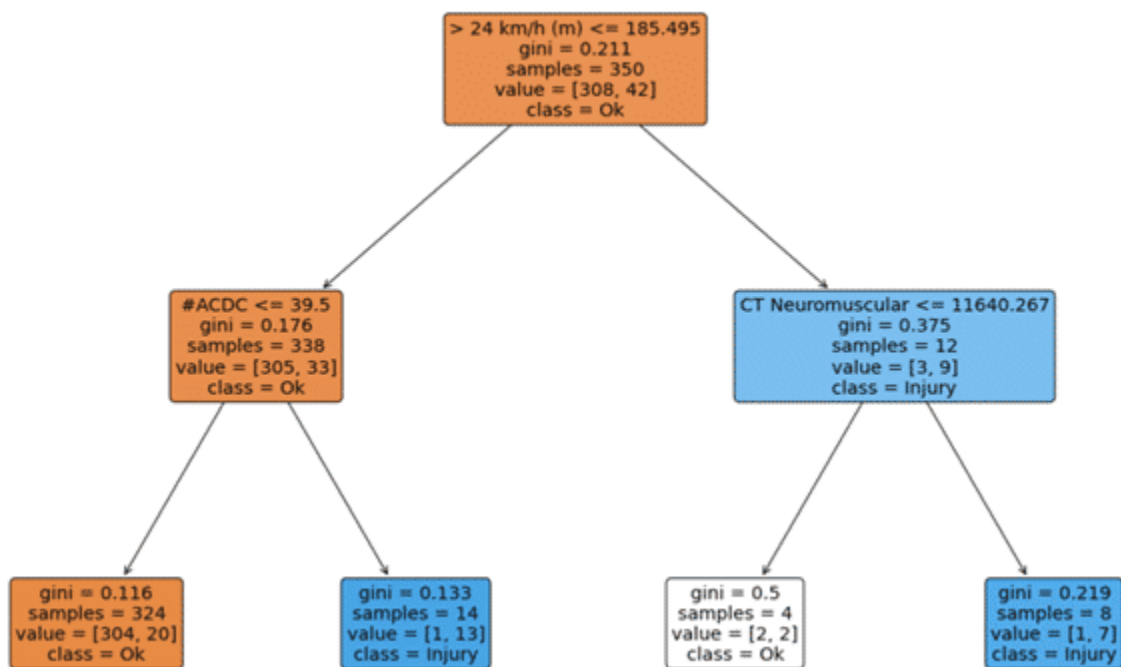
Este tipo de análisis es el más complejo; en su esencia, pretende determinar qué acción tomar en el contexto en el que nos encontramos, en otras palabras, cuál es la mejor solución para el problema que se plantea. La diferencia con un modelo predictivo es que utiliza analítica más avanzada, la cual proporciona probabilidades de que un resultado suceda y cuáles son las consecuencias o los derivados de una acción u otra.

Vamos a utilizar un ejemplo ficticio, en el que se intentan predecir lesiones a partir de variables de carga locomotora. La predicción de lesiones ha sido un ámbito de investigación durante muchos años y

su enfoque es extremadamente complejo. El ejemplo a continuación es, únicamente, una simplificación.

Este ejemplo muestra un árbol de decisión en el que, por cada parámetro que cumple el jugador, aparecen dos criterios para determinar si ocurrirá una lesión o no.

Figura 5: Análisis prescriptivo



Fuente: Elaboración propia.

Análisis avanzados

RStudio tiene la posibilidad de aplicar todas las técnicas mencionadas anteriormente, pero, además, posee la capacidad de aplicar métodos estadísticos avanzados, lo que lo diferencia como herramienta muy útil para estos análisis. En este apartado, únicamente, mencionaremos técnicas avanzadas que existen, ya que en otro módulo del curso se verán con más de detalle.

Teniendo en cuenta las propiedades de los datos que comentamos con anterioridad (volumen, velocidad y variedad) y la complejidad en el deporte, cada vez es más necesario utilizar este tipo de análisis avanzados para avanzar en nuestras investigaciones.

- **Técnicas avanzadas supervisadas**
 - Regresiones.
 - Clasificación.

- **Técnicas avanzadas no supervisadas**
 - Clasificación.
 - Reducción dimensional.
 - Asociación.

¿Qué tipo de análisis pretende usar los datos de los que disponemos para encontrar relaciones entre variables y poder extrapolar dichas relaciones en situaciones futuras?

- El análisis predictivo
- El análisis diagnóstico
- El análisis prescriptivo

SUBMIT

CONTINUAR

Referencias

Bartlett, J. D. y Drust, B. (2021). A Framework for Effective Knowledge Translation and Performance Delivery of Sport Scientists in Professional Sport. *European Journal of Sport Science*, 21(11), 1579-1587.

French, D. y Torres Ronda, L. (2021). *NSCA's Essentials of Sport Science*. Human Kinetics.

Goes, F. R., Meerhoff, L. A., Bueno, M. J. O., Rodrigues, D. M., Moura, F. A., Brink, M. S., Elferink-Gemser, M. T., Knobbe, A. J., Cunha, S. A., Torres, R. S., y Lemmink, K. A. P. M. (2021). Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, 21(4), 481-496.

Houtmeyers, K. C., Jaspers, A., Figueiredo, P. (2021). Managing the Training Process in Elite Sports: From Descriptive to Prescriptive Data Analytics. *Int J Sports Physiol Perform*, 16(11):1719-1723.

Impellizzeri, F. M. y Marcora, S. M. (2009). Test validation in sport physiology: lessons learned from clinimetrics. *Int J Sports Physiol Perform*, 4(2):269-77.

Rein, R., y Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus* 5, 1410. <https://doi.org/10.1186/s40064-016-3108-2>.

CONTINUAR