Módulo 1. Analítica avanzada para determinar perfiles de rendimiento físico

BARÇA INNOVATION HUB
Universitas

En el curso anterior, en el módulo "Introducción a los modelos estadísticos", destacamos la importancia de los modelos como herramienta para establecer relaciones entre los datos de los que disponemos. Además, un modelo se aplica principalmente para poder utilizar los elementos que participan en esa relación y así estimar una variable respuesta o variable objetivo.

También describimos cómo los modelos estadísticos son, en su esencia, fórmulas matemáticas, las cuales permiten que el modelo sea fácilmente interpretable y que podamos conocer si las variables tienen mayor o menor importancia dentro del modelo, así como si afectan de manera positiva o negativa la variable respuesta.

Con la mayor capacidad computacional de la actualidad y el gran desarrollo en los métodos de análisis de datos, actualmente disponemos de herramientas más avanzadas cuyo objetivo es realizar predicciones más precisas. Estamos hablando del término, muy común en la actualidad, *Machine Learning.*

Los modelos de *Machine Learning* están diseñados a partir de relaciones matemáticas mucho más complejas y menos interpretables, las cuales se basan en encontrar

patrones en los datos para aprender de ellos y desarrollar un algoritmo que permita predecir nuevos datos y tomar decisiones.
Estamos utilizando terminología que puede estar muy alejada del día a día de un *sport scientist,* pero el objetivo de este módulo no es conocer los procesos matemáticos detrás de cada una de las técnicas de *Machine Learning*, sino mostrar la gran variedad de herramientas de *Machine Learning* que tenemos a disposición a través de RStudio para decidir si necesitamos alguna de ellas para responder preguntas o inquietudes en el ámbito del rendimiento físico. Además, como profesionales, tendremos que conocer cuáles son las características y las bases de uso de estas herramientas, pues esto nos permitirá decidir qué herramienta se ajusta mejor a nuestro objetivo, como sucedía con los modelos lineales.

Hay que destacar que, aunque tengamos la capacidad y las herramientas para utilizar técnicas de *Machine Learning,* en muchas ocasiones, los modelos más simples cumplirán el objetivo con creces y, teniendo en cuenta que son más fáciles de interpretar, serán una opción más adecuada.

1. Grupos de técnicas de Machine Learning

Las técnicas de *Machine Learning* se pueden dividir en dos grupos principales: las técnicas de aprendizaje supervisado y la de aprendizaje no supervisado.

- Técnicas de aprendizaje supervisado: son aquellas en las que tenemos una variable respuesta conocida. De la misma manera que en un modelo de regresión lineal tenemos una variable independiente que queremos estimar, las técnicas de aprendizaje supervisado pretenden predecir la variable respuesta a partir de variables dependientes.

- Técnicas de aprendizaje no supervisado: en este caso, no tenemos una variable respuesta, sino que utilizaremos técnicas de *Machine Learning* para conocer asociaciones entre los datos, relaciones entre ellos o grupos.

CONTINUAR

2. Algoritmos supervisados

Como acabamos de comentar, la característica principal de estas técnicas es que disponemos de una variable respuesta. El algoritmo va a utilizar esa variable respuesta para "aprender", es decir, realizar múltiples combinaciones y asociaciones para intentar predecir la variable y, en función de su éxito o fracaso, realizar modificaciones en el algoritmo para ajustarse a los resultados.

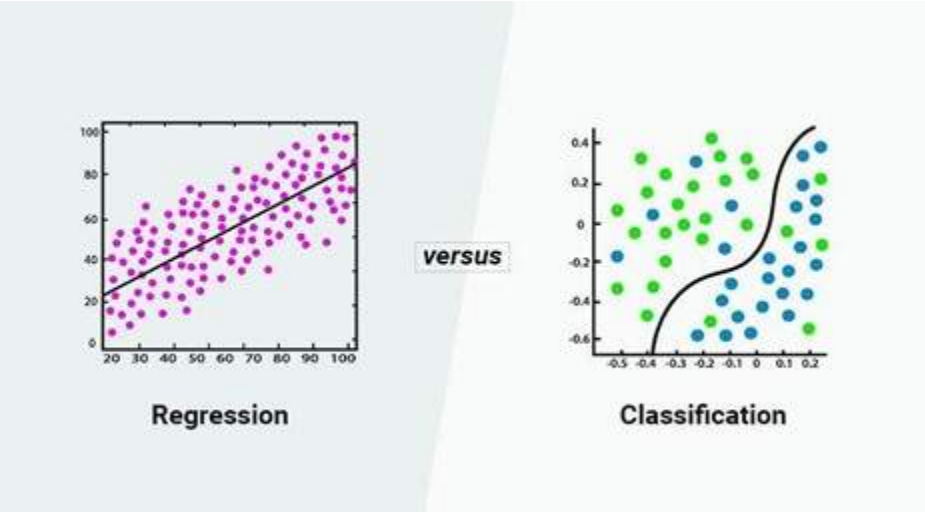
Los grupos de algoritmos supervisados se pueden resumir en lo siguiente:

- Algoritmos supervisados de regresión: tienen la misma función que los modelos de regresión simples en los cuales utilizamos una función matemática para estimar la variable respuesta, la cual es continua o numérica. En el caso de los algoritmos, como dijimos, las asociaciones que la herramienta utiliza son más complejas y de mayor dificultad de comprensión, pero el objetivo es el mismo. Dotar a la herramienta de variables e información para intentar

predecir la variable respuesta nos proporcionará un algoritmo para predecir resultados en datos futuros.

- Algoritmos supervisados de clasificación: en este caso, la variable respuesta que tenemos no es una variable continua o numérica, sino que tiene una respuesta binaria o categórica. Aquí, la herramienta utilizará las variables de entrada o input para obtener patrones que las asocian para determinar cuál es la variable respuesta.

Figura 1: Algoritmos supervisados de regresión y de clasificación



Fuente: Terra, 2024. https://tinyurl.com/3ef8zteb

Podemos encontrar fácilmente ejemplos de ambos tipos de objetivos en el contexto del *sport scientist*. Vimos en el curso anterior ejemplos de regresión para estimar la carga interna del jugador a partir de valores de carga interna y, como ejemplo de clasificación, vimos la estimación de si un servicio en tenis sería saque directo dependiendo de una serie de variables de entrada.

Otros ejemplos de objetivos clasificadores son los algoritmos de predicción de lesiones en los cuales, a partir de múltiples fuentes de información (historial lesivo, edad, carga, cualidades físicas, etc.), queremos estimar el riesgo de lesión del jugador. El resultado podría determinar qué jugadores tienen más éxito en una cantera deportiva para llegar al primer equipo, al cual podemos aportarle información como los años que el jugador ha formado parte del club, sus cualidades o test físicos, disponibilidad o lesiones, etc.

¿Cuáles de las siguientes afirmaciones sobre los algoritmos supervisados de regresión y clasificación son correctas?

- [ ] Los algoritmos de regresión se utilizan para predecir variables continuas o numéricas.
- [ ] Los algoritmos de clasificación se emplean para

predecir variables continuas.

- [ ] Los algoritmos de clasificación predicen variables con respuestas categóricas o binarias.
- [ ] Los algoritmos de regresión no usan funciones matemáticas para estimar la variable respuesta.
- [ ] Tanto los algoritmos de regresión como de clasificación se enfocan en predecir variables continuas.

SUBMIT

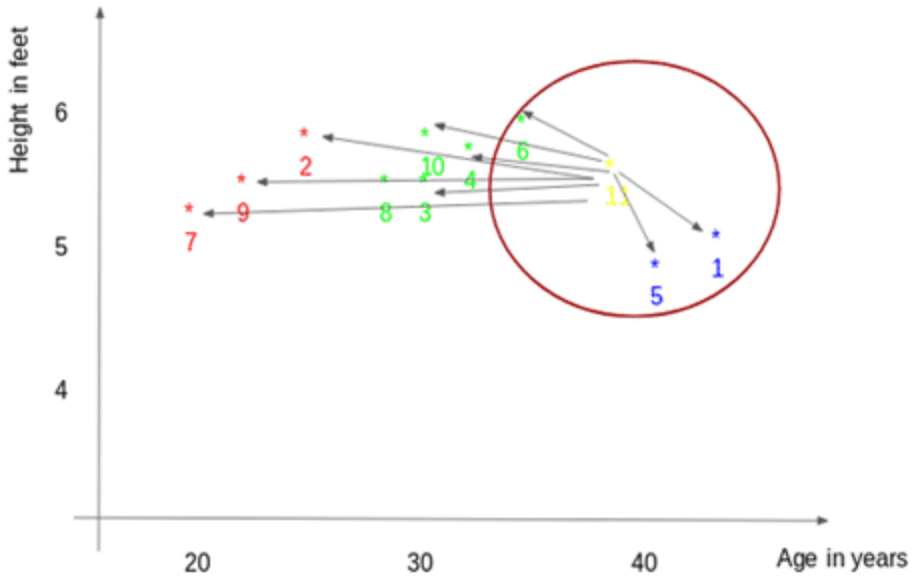
Como hemos descrito, hay múltiples algoritmos que permiten abordar los objetivos. Cada uno de ellos utilizará un proceso distinto para conseguir el objetivo de realizar predicciones precisas.

Entre los algoritmos de regresión más comunes, encontramos:

- Regresión lineal
- Modelos lineales generalizados

- Árboles de regresión
- Bosques aleatorios
- Redes neuronales
- K vecinos más cercanos

Figura 2: Algoritmo k vecino más cercano



Fuente: Singh, 2024. https://tinyurl.com/265yxwun

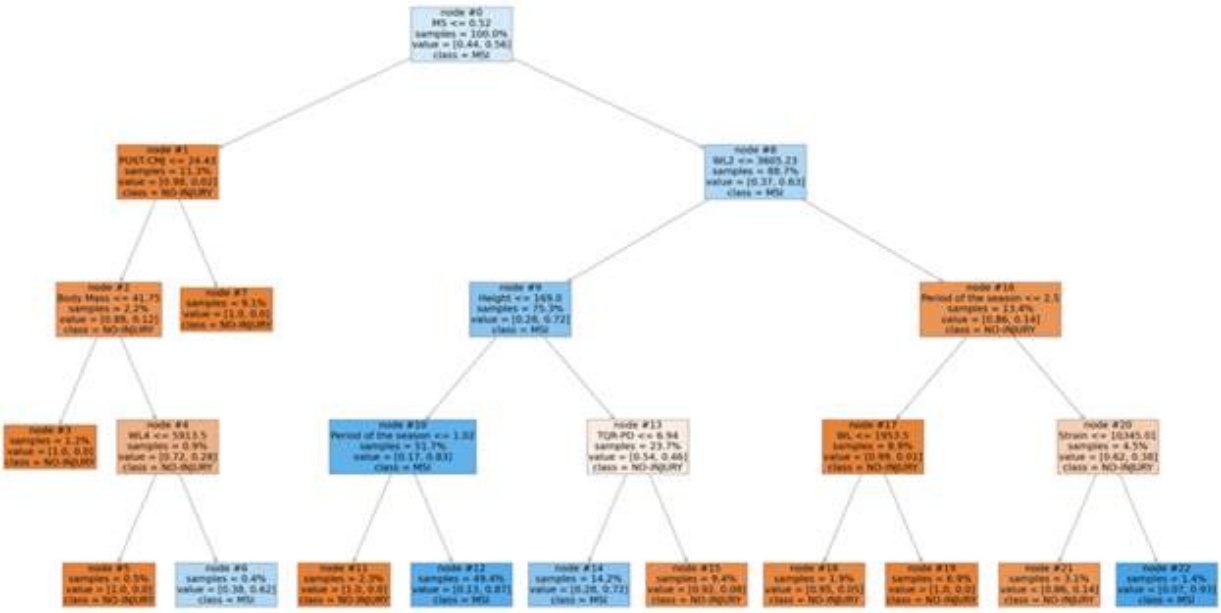
En este gráfico, vemos un ejemplo sencillo donde se aplica un algoritmo k vecino más cercano. Partimos de 3 variables de datos:

edad, altura y peso y queremos predecir el peso del jugador 11 (en amarillo). Para predecir este valor de manera muy simplificada, el algoritmo usa la información de los jugadores con características similares en las otras variables en el jugador 11; y a partir de la distancia de estos con el jugador 11, se determinará el peso del jugador.

Entre los algoritmos de clasificación más comunes, encontramos:

- Regresión logística
- Árboles de decisión
- Bosques aleatorios
- Redes neuronales
- K vecinos más cercanos

Figura 3: Modelo de clasificación con estructura de árbol



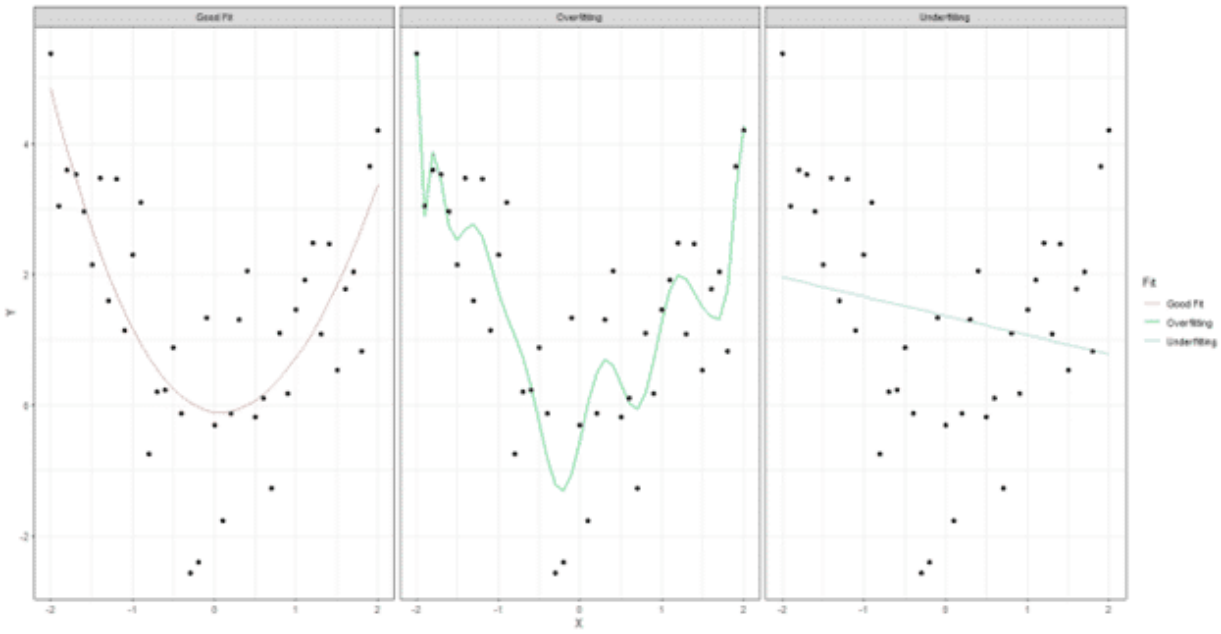
Fuente: Mandorino et al., 2021. https://tinyurl.com/38a7bfhw

Mandorino et al. (2021) utilizan un modelo de clasificación que usa la estructura de los árboles de decisión para predecir lesiones musculares en jugadores jóvenes de fútbol. Para ello, el modelo utiliza variables de carga y valoraciones de fatiga. El resultado es el que se muestra en la figura anterior. Estos gráficos permiten describir cuál es el proceso de toma de decisiones para determinar si se producirá una lesión. El proceso parte desde el punto más alto del gráfico: se indica si una condición se cumple o no, por ejemplo, si los valores de un test de salto están por encima de cierto umbral; dependiendo de si la respuesta es afirmativa o negativa, descenderemos por el correspondiente lado del árbol de decisión, y así consecutivamente hasta llegar a una de las "ramas" finales, que determinará la predicción final.

Como vemos, los algoritmos son, en muchas ocasiones, comunes en los dos objetivos; lo que determinará el uso es la variable que queremos predecir, si se trata de una variable continua o de un resultado binario o categórico.

Si disponen de la cantidad de información suficiente y se ha cumplido con los requisitos para su utilización, los algoritmos pueden ser muy eficientes en la predicción de resultados. Esto se debe a la gran capacidad y complejidad matemática para encontrar relaciones en los datos y explorar una gran cantidad de alternativas. Sin embargo, también plantea una limitación importante: el *overfitting*. Este concepto se refiere a que el algoritmo funcionará con éxito en los datos que hemos aportado, pero no tendrá el mismo éxito si utilizamos ese algoritmo con datos nuevos o futuros. El objetivo principal de utilizar técnicas de *Machine Learning* es tener mayor capacidad de predicción en los datos en que estemos interesados, por lo tanto, debemos encontrar el equilibro ideal para tener un cierto grado de éxito en nuestras predicciones a partir de la información que proporcionamos al modelo.

Figura 4: Diferencias entre *underfitting* y *overfitting*



Fuente: elaboración propia.

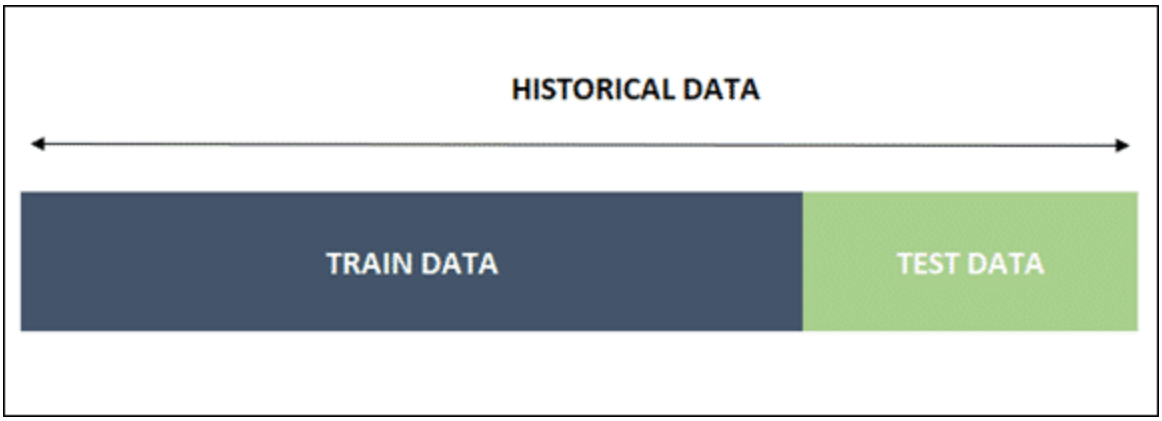
En el gráfico anterior, vemos este aspecto representado: las diferencias entre *underfitting*, es decir, muy baja capacidad de predicción, y *overfitting*, modelo demasiado específico de los datos actuales.

Para corregir esta limitación, se sigue el proceso llamado entrenamiento-test. Este proceso consiste en dividir los datos que tenemos para nuestro modelo en dos partes. Habitualmente, entre un 70 % y un 80 % de los datos formarán parte de los datos de entrenamiento y el porcentaje restante de los datos test.

De esta manera, se creará un algoritmo a partir de los datos de entrenamiento y podrá ser evaluado en los datos test. Los resultados en los datos test marcarán el rendimiento del algoritmo y la técnica

utilizada. Si estos resultados no son satisfactorios, podemos realizar modificaciones en los parámetros o utilizar otras técnicas de *Machine Learning* para comparar los resultados.

Figura 5: División del total de datos en datos train y datos test



Fuente: elaboración propia

CONTINUAR

Actividades

¿Cuáles de las siguientes afirmaciones sobre el modelo de clasificación utilizado por Mandorino et al. (2021) son correctas?

- [ ] El modelo de clasificación se basa en la estructura de redes neuronales.
- [ ] El modelo utiliza árboles de decisión para predecir lesiones musculares en jugadores de fútbol.
- [ ] Las variables utilizadas por el modelo son la carga y valoraciones de fatiga.
- [ ] El proceso de toma de decisiones en el árbol de decisión no depende de si las condiciones se cumplen o no.
- [ ] El modelo no proporciona una predicción final al llegar a las ramas finales.

SUBMIT

CONTINUAR

3. Algoritmos no supervisados

En el caso de los algoritmos no supervisados, el objetivo no será predecir una variable respuesta, sino encontrar patrones o asociaciones entre los datos para conseguir distintas finalidades.

Los grupos de algoritmos supervisados se pueden resumir en los siguientes tres:

- Algoritmos no supervisados de agrupación: el objetivo de estos algoritmos es encontrar aspectos comunes entre observaciones y diferencias significativas con otras observaciones y, a partir de los resultados, determinar grupos de observaciones con características similares. Estos algoritmos se conocen como *clustering*.

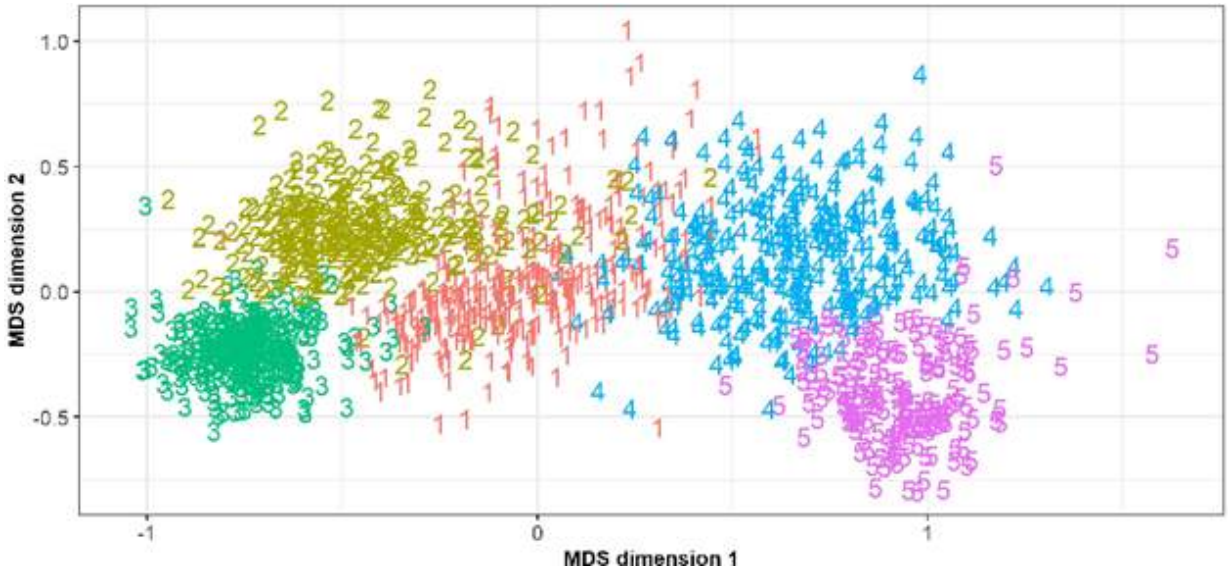
- Algoritmos no supervisados de reducción dimensional: el objetivo de estos algoritmos es transformar datos en los que tenemos muchas variables en un resumen con variables compuestas

que muestran las relaciones entre ellas y, a su vez, no aportan información repetida. Esto reducirá el número de variables totales; ayudará en su visualización y su uso para otros algoritmos de *Machine Learning*.

- Algoritmos no supervisados de asociación: el objetivo de estos algoritmos consiste en encontrar relaciones de eventos que ocurren juntos. A modo simplificado: cuando ocurre un evento, se determina qué es probable que pase en el siguiente. El ejemplo más común son las recomendaciones de Netflix, que están basadas en las películas o las series que ya hemos visto.

Encontramos, también, múltiples aplicaciones de estos algoritmos en el rendimiento deportivo. En el siguiente gráfico (figura 6) de Akhanli y Henning (2022), se utiliza una técnica de *clustering* para encontrar jugadores con similitudes utilizando distintas métricas de rendimiento. Vemos cómo el algoritmo clasifica a todos los jugadores analizados (cada uno de los números) en grupos según su similitud (5 grupos en este caso). Estos métodos pueden ser de mucha utilidad en departamentos como *scouting*, para detectar jugadores en bases de datos que tengan rendimientos similares a perfiles que estamos buscando.

Figura 6: Representación multidimensional de los datos con clustering

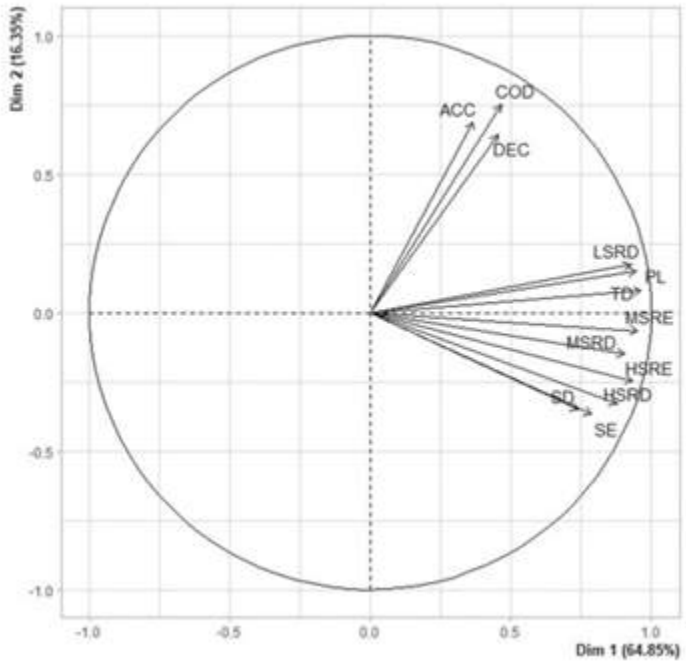


Fuente: Akhanli y Henning. (2022), p. 15.

Un ejemplo de aplicación de algoritmos de reducción dimensional lo podemos encontrar fácilmente en el contexto del Sport Scientist. Si utilizamos sistemas de monitorización de la carga GPS, estos dispositivos nos dan información de más de 100 variables calculadas. Podemos optar por diferentes metodologías para reducir el número de variables para nuestros análisis y visualizaciones, como establecer las correlaciones entre ellas para determinar cuáles están aportando el mismo tipo de información o utilizar la bibliografía para determinar qué aspectos de la carga están representando cada una de ellas (carga mecánica, locomotora, metabólica, etc.). Sin embargo, si

queremos utilizar toda la información posible para un proyecto de mayor complejidad (por ejemplo, la predicción de lesiones), podemos optar por usar un algoritmo de reducción dimensional, el cual mostrará cuáles de las variables explican en mayor medida la variable respuesta y el porcentaje de información que aporta cada una de ellas. Estos algoritmos son conocidos como análisis de componentes principales (PCA). Los componentes principales son un resumen de relaciones entre variables (con diferentes ponderaciones) que pretenden explicar la variable respuesta.

Figura 7: Gráfico del PCA para los dos componentes principales extraídos



Fuente: Nosek et al. (2023), p. 7.

La visualización de Nosek et al. (2023) muestra los dos componentes principales del algoritmo de reducción dimensional, que, en este caso, explican aproximadamente el 80 % de la información. El primero de ellos (eje horizontal) se compone de 8 variables, casi todas ellas se tratan de variables asociadas con el volumen (distancia total, *player load,* etc.); y el segundo componente (eje vertical) contiene 3 variables, en este caso, relacionadas con la carga mecánica (aceleraciones, cambio de dirección, deceleraciones). Utilizando esta técnica, hemos pasado de necesitar 13 variables para explicar la carga de entrenamiento a utilizar únicamente dos componentes principales que nos aporta esa información resumida.

En el caso de Weaving et al. (2019), utilizan también PCA para resumir la carga del jugador. Como se puede ver en el gráfico a continuación (figura 8), sus dos componentes principales están relacionados con la carga a alta velocidad y la carga global. Podemos utilizar estos componentes para visualizar la carga de forma resumida a lo largo de la temporada del jugador y, tal como se muestra en el gráfico, evaluar la demanda condicional de la sesión teniendo en cuenta los componentes principales.

Figura 8: Carga a alta velocidad y carga global

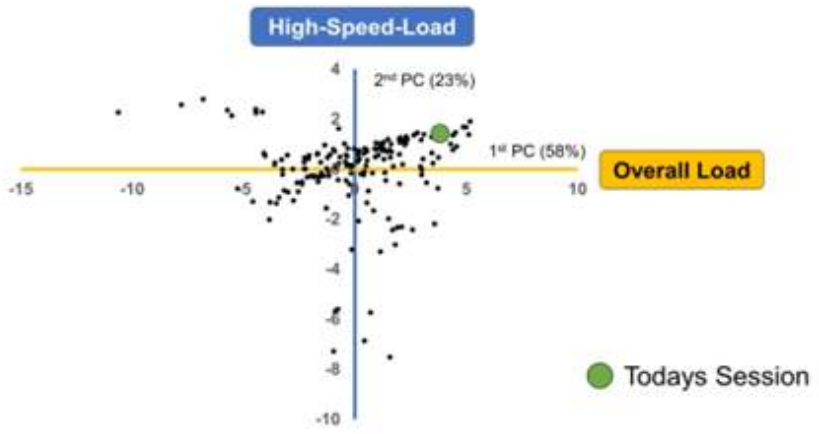
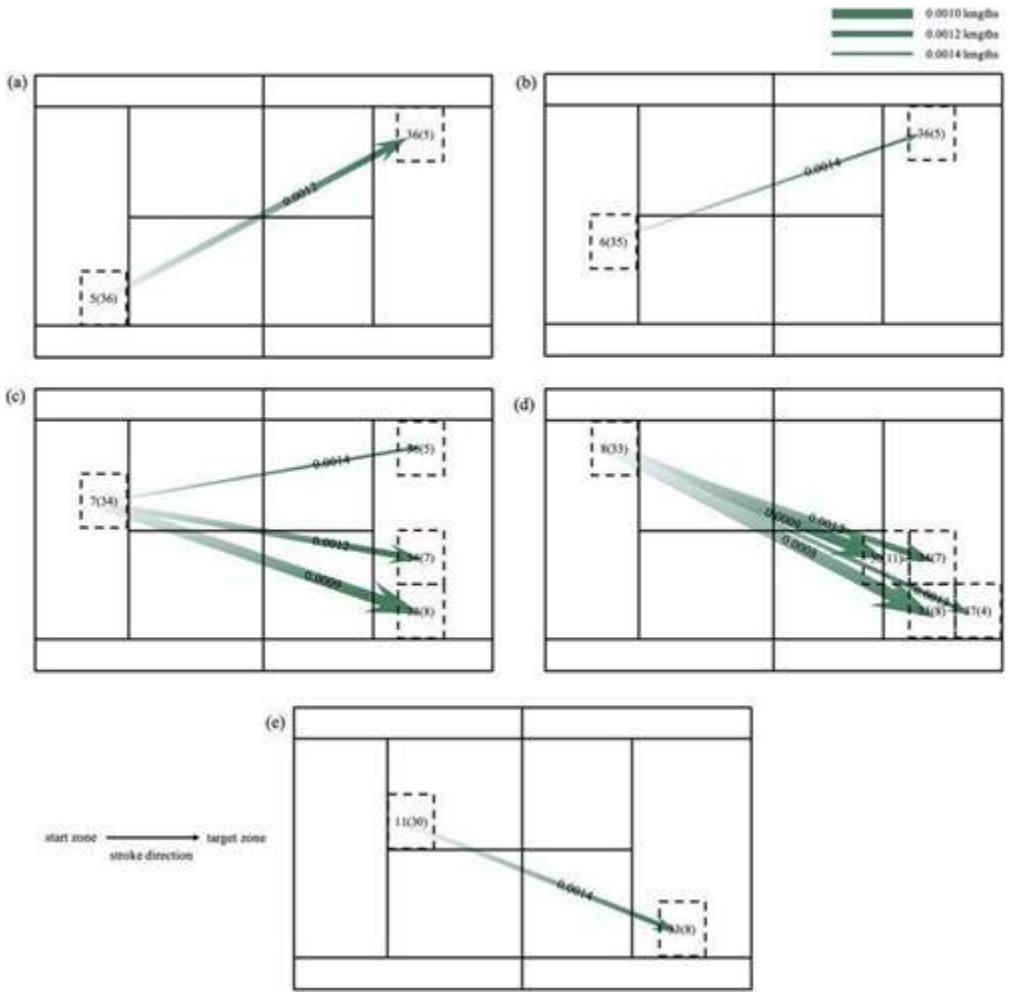


Figure 4. Scatterplot of the first (x axis) and second (y axis) principal component scores for the 169 training sessions for an individual player. The green data point highlights today's session in the context of every other training session completed.

Fuente: Weaving et al. (2019), p. 7.

Finalmente, en el estudio de Yunjing Zhou et al. (2023), vemos un ejemplo de algoritmo de asociación aplicado al rendimiento deportivo. El objetivo de este estudio fue determinar los patrones de los golpes en un partido de tenis según dónde se realizó el último golpe. Este tipo de estudios tiene una clara aplicación al análisis de partido y rivales, y puede suponer una gran ventaja competitiva si se desarrolla de manera correcta.

Figura 9: Visualización de los patrones obtenidos mediante el Algoritmo de asociación



Fuente: Zhou et al. (2023), p. 9.

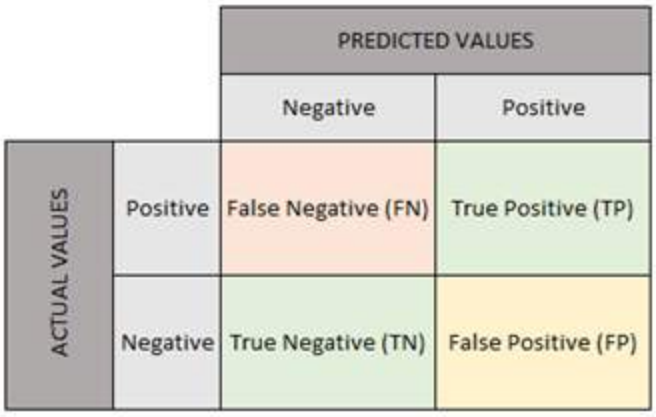
CONTINUAR

4. Evaluación del rendimiento de los algoritmos

De la misma manera que es necesario evaluar los modelos lineales que desarrollamos en el módulo anterior, cada uno de los grupos de técnicas de *Machine Learning* tendrá su evaluación correspondiente. Al tratarse de modelos más complejos, sus métricas de evaluación también pueden ser múltiples y con matices, pero aquí están las características más importantes por evaluar:

- Modelos de clasificación
    - Precisión: proporción de identificaciones positivas que fueron realmente correctas.
    - *Recall*: proporción de positivos reales que fueron identificados correctamente.
    - F1: combinación de ambos aspectos.

Figura 10: Matriz de confusión



Fuente: elaboración propia.

- Modelos de agrupación:
    - *Silhouette Score:* medida de separación y densidad de los grupos detectados.
- Modelos de reducción dimensional:
    - Proporción de varianza explicada: porcentaje de varianza que explica cada uno de los componentes principales.

CONTINUAR

Actividades

¿Cuáles de las siguientes afirmaciones sobre la evaluación de modelos de clasificación son correctas?

- [ ] La precisión mide la proporción de identificaciones positivas que fueron realmente correctas.
- [ ] El *recall* mide la proporción de positivos reales que fueron identificados incorrectamente.
- [ ] La métrica F1 es una combinación de precisión y *recall*.
- [ ] El *recall* mide la proporción de identificaciones positivas que fueron realmente correctas.
- [ ] La precisión y el *recall* son métricas irrelevantes para evaluar modelos de clasificación.

SUBMIT

CONTINUAR

Referencias

Akhanli, S. E., y Hennig, C. (2022). Clustering of football players based on performance data and aggregated clustering validity indexes. *Journal of Quantitative Analysis in Sports*, 19, 103-123. **https://www.researchgate.net/publication/360098833_Clustering_of_football_players_based_on_performance_data_and_aggregated_clustering_validity_indexes**

Mandorino, M., Figueiredo, A., Cima, G., y Tessitore, A. (2021). Predictive Analytic Techniques to Identify Hidden Relationships between Training Load, Fatigue and Muscle Strains in Young Soccer Players. *Sports*. 10(3)**. https://doi.org/10.3390/sports10010003**

Nosek, P., Andrew, M., Sormaz, M., Drust, B., y Brownlee, T. (2023). The use of principal component analysis for reduction of training load data in professional soccer. *Kinesiology*, 55(2), 202-212. **https://hrcak.srce.hr/file/446882**

Singh, A. (1 de agosto de 2024). **https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/**

Terra, J. (23 de julio de 2024). https://www.simplilearn.com/regression-vs-classification-in-machine-learning-article

Weaving, D., Beggs, C., Dalton Barron, N., Jones, B., y Abt, G. (2019). Visualizing the Complexity of the Athlete-Monitoring Cycle Through Principal-Component Analysis. *International Journal of Sports Physiology and Performance,* 14(9), 1304-1310. https://doi.org/10.1123/ijspp.2019-0045

Zhou, Y., Zong, S., Cao, R., Gómez, M. Á., Chen, C., y Cui, Y. (2023). Using network science to analyze tennis stroke patterns. *Chaos, Solitons & Fractals,* 170, 113305. https://doi.org/10.1016/j.chaos.2023.113305

CONTINUAR