Module 2. The foundations and principles of data analysis applied to physical performance

BARÇA INNOVATION HUB
Universitas

- The scientific method
- References

The scientific method

In the previous module, we have described the properties and functionalities of RStudio, that is, we have introduced a tool that will enhance our daily efficiency as sports scientists. We have also emphasized the importance of having an adequate question to answer during the process of using this tool. We must identify what is needed in our context, use the data in the best possible way, and translate it to provide valuable information to the different departments, with the ultimate goal of helping players fulfil that need or get hold of that determining element for their development (Bartlett & Drust, 2021).

We are referring to the scientific method, a method that allows us to be objective in decision-making and to follow an outlined and specific process to solve the problems that arise in our professional field and that will force us to meet certain quality standards to ensure that the results are useful and reliable.

In this module, we will strive to detail the process to follow from the moment a problem arises, focusing on the role played by data

analysis and describing the principles and requirements of this field. Furthermore, we will define its particularities regarding the context of physical performance.

French and Torres Ronda (2021) define the scientific method for sports science in a very visual way,

with a circle that will be repeated once the last step has been completed:

- Observe.
- Define.
- Build a hypothesis.
- Experiment.
- Interpret.
- Implement.

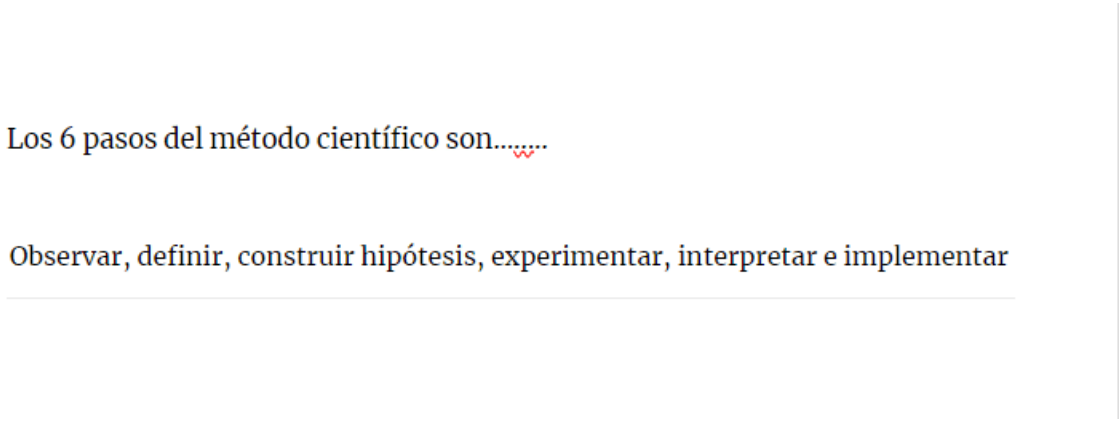
What role does data analysis play in this process/method?

As we can see above, data plays a fundamental role in the development of the scientific method. We will glue the first two steps that we have detailed above into one: observe and define. Taking into account our context, we have to define the objectives we want to achieve. These objectives may have

multiple performance-related orientations. We can also talk about different layers of complexity and objectives derived from them.

However, the choice of questions or observations depends on each context and each professional, and does not belong to the contents of this course. Here we will focus on how to be efficient in the use of data analysis to answer these questions. It should be noted that data or analysis is not the only means to get these answers, but we want to make the most of it, in order to get more insight into the complexity of our athletes' sports performance (Goes et al., 2021).

Simply put, data analysis is the transformation of recorded data into actionable and impactful information.



Data ualities

The growth and technological development of recent years has led to the emergence of concepts such as Big Data, which is often misunderstood, as it is often given interpretations that go beyond its original description. Big Data is a set of data that must meet a series of characteristics and, although not all of them are rigorously applicable to the scenario of sports performance, they can serve as a reference to describe "data" within our professional

context (Rein and Memmert, 2016). Originally, big data was described through “the three Vs":

- Volume: A large amount of data, which also requires some storage capacity (we can talk about databases or cloud systems from providers).
- Variety: the data we obtain come in different formats (text, numerical, time series).
- Velocity: the rate at which information is received is very high, which updates, the information we already have.

When considering our sports context, it is clear that these three points are met, although, in many cases, to a lesser extent than in other professional fields. Because of that reason we have to be less rigorous in our process if we want to get the most out of the projects we work on.

More recently, three other Vs have been added to the concept “Big Data”:

- Veracity. Data quality. This feature will be given by our data collection plan and the tool choice.
- Value: this not only has to do with information being related to what we want to investigate, but that it is practical and realistic to collect.
- Visualization: a concept that refers to one of the last phases of data analysis: communication. Therefore, we will need to complete a series of steps before we can reach this final part, but it is one of the final objectives of data, which can have a greater impact and influence on the other departments.

As we have mentioned, the first three points are highly context-dependent; however, in the last three points, we can have a greater impact. We must meticulously guarantee that the veracity and quality standards of data are met. We must select the data that are most related to the questions we want to answer, as well as the type of analysis we want to use. These two points will increase the value of our data and will allow our analysis to have a greater impact. In the final part of the process, we must choose the most appropriate type of visualization for the information we want to share. Tools such as RStudio enable us to automate the creation of visualizations and also to choose the best way to present them.

¿Cuáles son las 3 "V" añadidas recientemente que describen la Big Data?

- [ ] Veracidad
- [ ] Valor
- [ ] Visualización
- [ ] Vivacidad

SUBMIT

Data quality

Achieving "data quality", which has to do with the veracity and value of data, involves what is known in the field of data science as "data hygiene", which will be described in greater detail in the next module. However, in this section, we will highlight fundamental theoretical aspects related to measurement tools which guarantee the first data quality filter and the way it is get. Then, we will continue with the cleaning and analysis process.

The tools we use must provide reliable measures, that is, they must be reproducible in similar situations. Impellizzeri and Marcora (2009) highlight two types of reliability:

- Absolute reliability, that is, the degree to which records vary by athlete.
- Relative reliability: The degree to which athletes maintain their position in relation to the group in repeated measures.

Both aspects are to be considered when deciding which test or tool to use. In other words, and going back to the importance of posing the question, depending on the research we want to carry out, we must choose tools that have greater absolute or relative reliability. For example, if the objective is a cross-sectional study of our team in which we want to detect different conditional profiles among our players, we should choose tools with high relative reliability. However, if the objective is to assess changes in the conditional profile throughout the season, the tool must have high absolute

reliability, so that it spots real changes throughout the study period. There are statistical tests to determine such types of reliability.

The second characteristic to highlight is validity. This property refers to the fact that the test must accurately and precisely measure a parameter relevant to what we intend to measure. To determine that a tool is valid, we must compare its results with those of tools classified as gold standard.

Let us look at an example of validity which is very common in the field of biomechanics: the creation of variables on the basis of tracking a movement. With the use of motion capture sensors or high-frequency tracking systems, we can track certain parts of the body during specific sports movements. This is commonly applied in biomechanics to determine certain variables or features in a specific sports movement in order to identify the optimal movement to achieve the best result. For example, if we were to analyse a tennis player's serve, we could get the tracking of the different joints and, in consequence, see the movement of the arm. Variables such as the maximum height at which the player stretches their arm in the preparation phase, angular velocities before the strike, etc., could be obtained through this monitoring. The objective could be to detect which of these variables is most relevant so as to hit the ball as efficiently as possible or in certain directions Adjust the movement.

If we want the information we get to be useful and applicable, we must use variables that provide valid information, that is, which could be used to show differences between players, if we agree that players have different characteristics (effect they have on the ball, speed, etc.). If any of the variables we have calculated do not show differences between the different players and, therefore, do not allow us to provide information on how they affect the output variables (speed and effect), we would not consider them valid or useful.

Once the data collection has been completed following quality standards, we must proceed to correctly cleaning and treating it. These two steps require a lot of detail and are highly specific to the type of data we are dealing with, whether it is text, numeric, etc. We will deal with this process in a subsequent module, as this has to do, to a large extent, with the sports scientist's job. So, assuming that we have performed the previous step, we must select the statistical analysis that best fits the problem we have posed. To do this, we will look at the analysis options available (Houtmeyers et al., 2021).

Types of analysis: descriptive analysis

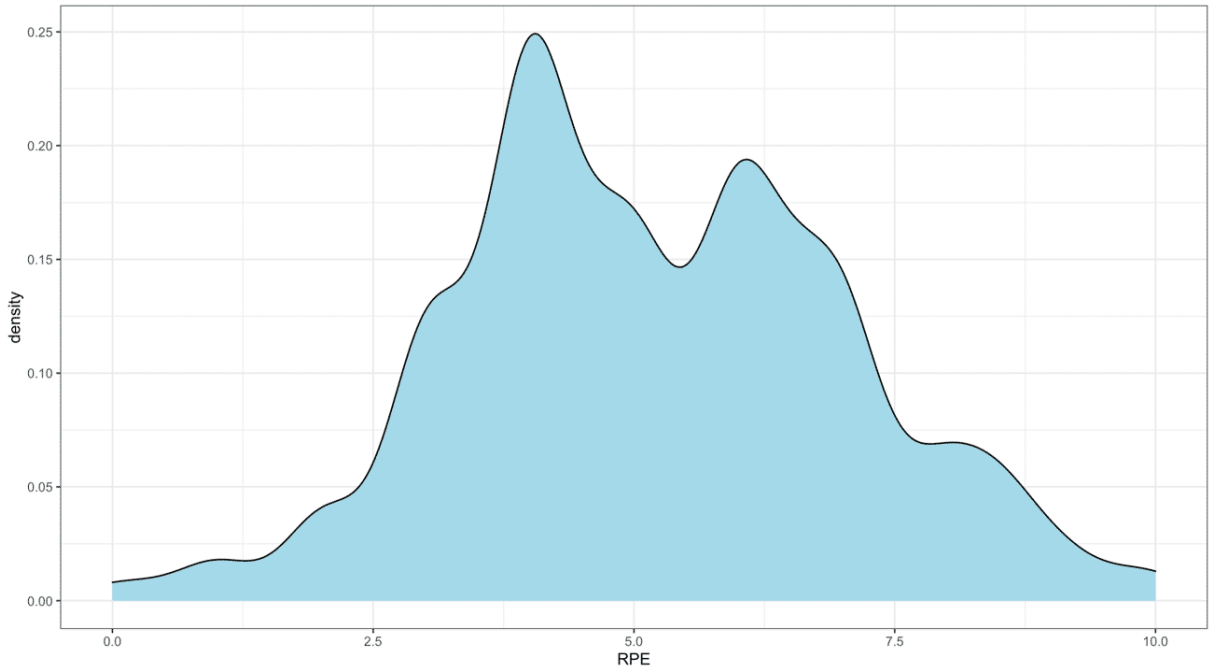
This type of analysis is considered the initial or basic one when starting to work with data. It enables us to see the behaviour of the data we want to analyse in a summarized way. It is part of the process of exploring data in many cases, before it is processed or transformed. Nonetheless, once the data is "finalized", this type of analysis can be of great value and efficiency in itself since it is not generally expensive compared to the time devoted to it. It should be noted that this type of analysis only shows the characteristics of the data we are analysing, and, therefore, we must be very cautious when coming up with generalizations.

To describe the different types of descriptive analysis, we will resort to specific examples.

- **Frequency measurements:** they allow us to know how our data is distributed according to a metric, such as, for example, if we want to know which answers are the most common among our players in the RPE questionnaire on the days after the match. In the image below, we can see how the responses are concentrated in two values (higher peaks):

approximately 4 and 7, which shows that there can be two types of sessions in which two groups of players perceive differently: one as as more intense and the other as less demanding. Using a histogram, as in the image, also allows for the shape and distribution of the data to be evaluated, which provides information on whether there are values inclining towards one direction or another.

Figure 1: Frequency measurements



Types of analysis: descriptive analysis

- Central tendency measurements: these describe what the mean response is. For example, if we want to spot the average distance covered by a central defender in a match, we will calculate the average of all the total distances in the games that the player has

completed during a given time. The most common measures of central tendency are as follows:

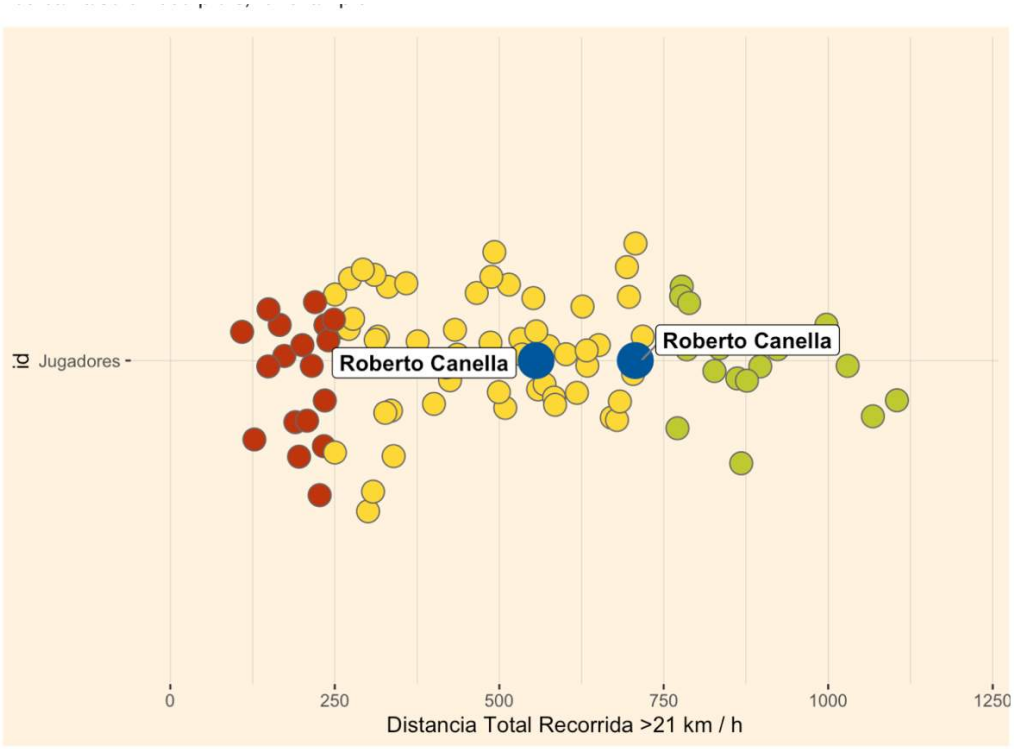
  - Mean.
  - Median.
  - Mode.

- Measurements of dispersion: these allow us to know the range in which the data is distributed, its normal variability. If, considering the previous example, instead of having a single value describing the match demands, we want to know the range in which the player usually moves, we can calculate the standard deviation to see which values are standard and easily detect which matches have been more demanding. This measurement is also used to identify possible outliers. The most common measures of central tendency are as follows:

  - Standard deviation.
  - Range.
  - Quartile deviation.

- Position measurements: these allow us to classify, and arrange athletes according to the value analysed within the data source. The graph below shows where two players (blue) stand compared to the rest of the players at the same position in the league, using the high-speed distance variable. The colour of the dots also shows which players are within the standard dispersion range (yellow) and which are above (green) or below (red).

Figure 2: Position measurements



Types of analysis: diagnostic analysis

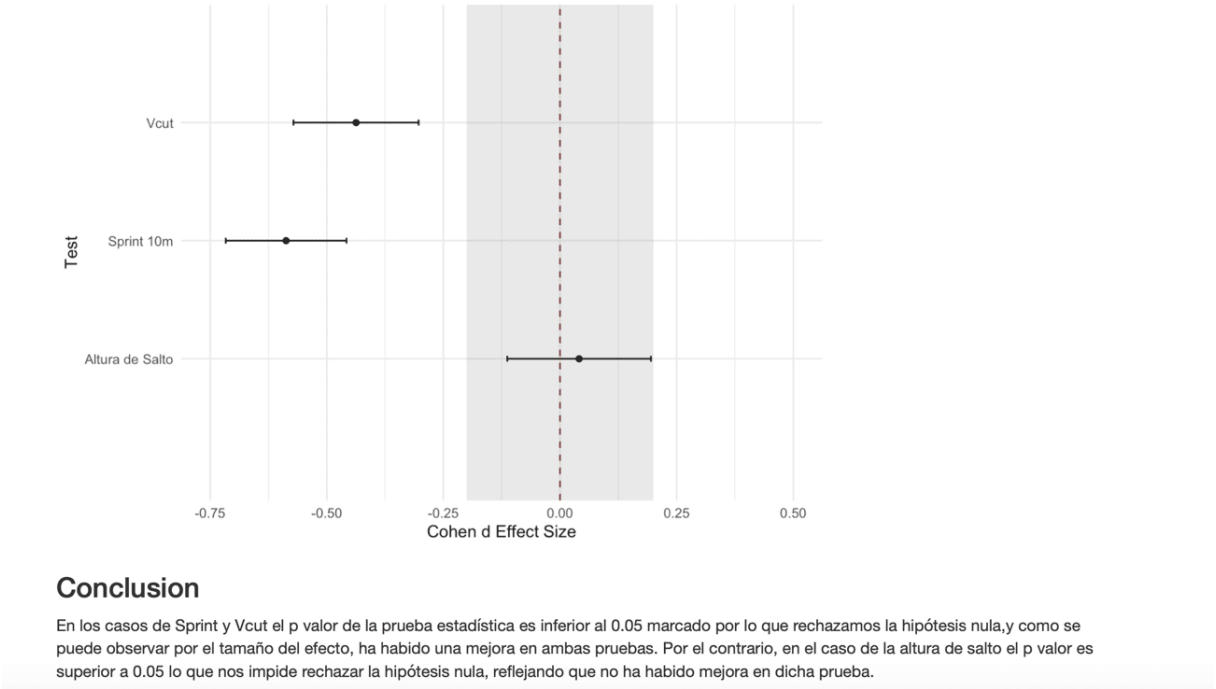
This analysis aims at comparing measures. It can be used to objectively state whether there are real, statistically significant differences between two groups or two moments of the season and, therefore, determine if the interventions we are making are the right ones. However, identifying the characteristics of the measure is essential to determine which type of statistical analysis is appropriate.

- Comparison of measurements/proportions of a sample: This is used when we want to compare the average values of our data source with a known reference value. For example, if we want to compare our group of players in the 30-15 test results with the baseline/normative values in the same age group.
- Comparison of measurements/proportions of independent groups: this is the same example as above, but, in this case, instead of comparing with

a reference value, we compare two teams of lower categories, to determine if there are differences between them.

- Repeated measures comparison: in this case, we want to determine if there has been a change before and after an intervention. In the image below, you can see the change, represented by the extent of the effect in three different tests for a given group of players.

Figure 3: Diagnostic analysis



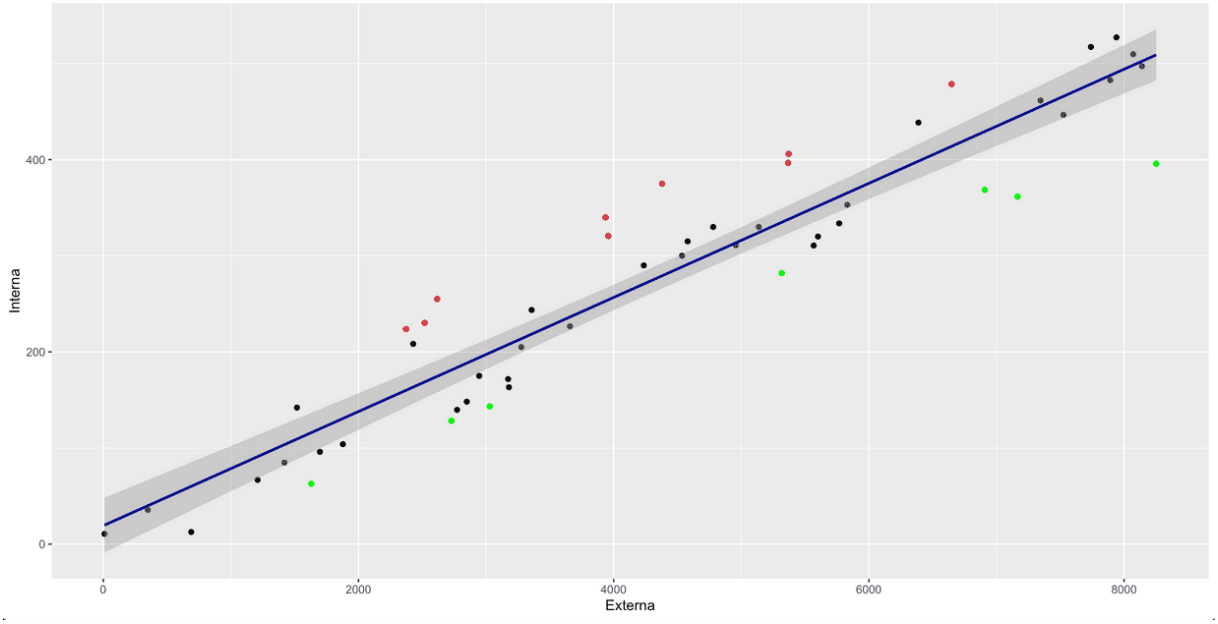
Types of analysis: predictive analysis

Predictive analysis aims to use the data we have to find relationships between variables in order to extrapolate these relationships in future situations. Please note these predictions are estimates - it must be considered that there is always an error associated to the relationships or statistical models that we use. When we communicate these results, it is a

good practice to communicate what the error associated with each of the estimates we are making is.

In the following example, we can see the relationship between two variables: one indicates the player's internal workload and the other indicates the locomotive or external workload. Knowing this relationship between the variables, we will be able to prescribe a certain external (controllable) workload according to the internal workload we want to achieve, which will lead to a specific adaptation.

Figure 4: Predictive analysis



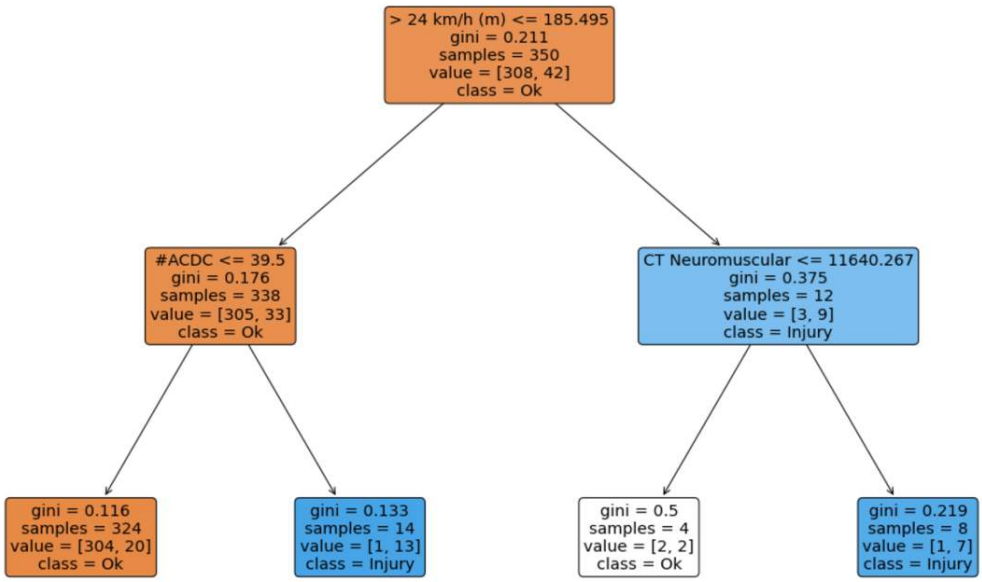
Types of analysis: prescriptive analysis

This type of analysis is the most complex one. Essentially, it aims to determine what action to take in our context, that is, the best solution to the problem that we have spotted. The difference with a predictive model is that it uses more advanced analytics, which provide probabilities of an outcome happening and the consequences or derivatives of one action or another.

We will use a made-up example in which we will try to predict injuries based on locomotor workload variables. Injury prediction has been a field of research for many years and its approach is extremely complex. The example below is quite simplified.

This example shows a decision tree in which, for each parameter that the player meets, two criteria appear to determine whether an injury will occur or not.

Figure 5: Prescriptive analysis



Advanced Analyses

RStudio makes it possible to apply all the techniques mentioned above, as well as advanced statistical methods, which is why it stands out as a very useful tool for these types of analyses. In this section, we will only mention

some advanced techniques, since we will look into them on more detail in another module of the course.

Considering the properties of the data discussed above (volume, velocity and variety) and the complexity in the field if sport, it is increasingly necessary to use this type of advanced analysis to advance in our research.

- **Advanced supervised techniques**
  - Regression.
  - Classification.
- **Advanced unsupervised techniques**
  - Classification.
  - Dimensional reduction.
  - Association.

¿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?

Type your answer here

SUBMIT

CONTINUE

References

Bartlett, J. D. and 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. and 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., and 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. and Marcora, S. M. (2009). Test validation in sport physiology: lessons learned from clinimetrics. *Int J Sports Physiol*

Perform, 4(2):269-77.

Rein, R., and 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.