Module 1. Advanced analytics for the determination of physical performance profiles

BARÇA INNOVATION HUB
Universitas

- Unit 2.1 Advanced analytics for the determination of physical performance profiles
- Activities
- References

Unit 2.1 Advanced analytics for the determination of physical performance profiles

In the module "Introduction to statistical models" of the previous course, we highlighted the importance of models to establish relationships between the data we have available. Furthermore, a model is mainly applied to influence the elements of this relationship and therefore estimate a response variable or objective variable.

We also described statistical models, their essence, mathematical formulas (which allow the model to be easily interpretable and also allow us to know if variables have greater or lesser importance within the model, as well as if they positively or negatively affect the response variable).

With today's computational capacity and development in data analysis methods, there are more advanced tools aimed at making more accurate predictions. This is possible due to Machine Learning.

Machine Learning models are based on much more complex and less interpretable mathematical relationships, which are formulated on

finding patterns in the data to learn from them and develop an algorithm that enables the prediction of new data and decision-making.

These concepts may sound far from the day-to-day life of a sports scientist, but the objective of this module is not to learn about the mathematical processes behind each Machine Learning technique, but to show the wide variety of Machine Learning tools available through RStudio in order to decide if we need any of them to respond to questions or concerns in the field of physical performance. In addition, as professionals, we need to know the characteristics and bases of these tools, since this will allow us to decide which tool best suits our objective, in the same way as with linear models.

It is worth noting that, although we have the capacity and tools to use Machine Learning techniques, in many cases, the simplest models will meet most objectives and, considering they are easier to interpret, they may be a more suitable option.

Machine Learning Techniques Groups

Machine Learning techniques can be divided into two main groups: supervised learning techniques and unsupervised learning techniques.

- Supervised learning techniques: those in which the response variable is known. In the same way that in a linear regression model there is an independent variable that is to be estimated, supervised learning techniques aim to predict the response variable from dependent variables.
- Unsupervised learning techniques: in this case, there is not a response variable, but Machine Learning techniques are used to pinpoint associations or relationships between data or groups.

Supervised learning algorithms

As we have just mentioned, the main characteristic of these techniques is the presence of a response variable. The algorithm will use that response variable to "learn"; that is, make multiple combinations and associations to try to predict the variable and, depending on its success or failure, modify the algorithm to adjust to the results.

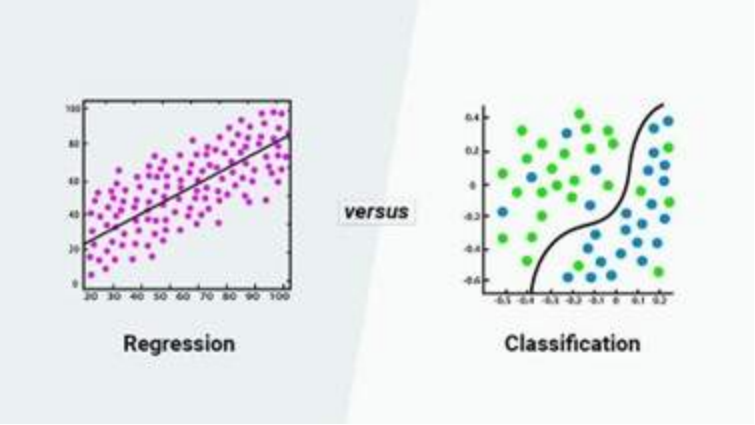
The supervised algorithm groups can be summarized as follows:

- Supervised regression algorithms: they have the same function as simple regression models in which a mathematical function is used to estimate the response variable, which is continuous or numerical. In the case of algorithms, the associations that the tool uses are more complex and more difficult to understand, but

the objective is the same. Providing the tool with variables and information to try to predict the response variable will provide us with an algorithm to predict results in future data.

- Supervised classification algorithms: in this case, the response variable is not a continuous or numerical variable, but has a binary or categorical nature. In these algorithms, the tool will use the input variables to extract patterns associated with the response variable.

Figure 1: Supervised regression and classification algorithms



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

We can easily find examples of both types of objectives in the context of the sport scientist. In the previous course we saw examples of regression to estimate the player's internal load based on internal load values and, as an example of classification, we saw the

estimation of whether a tennis serve is a direct serve depending on a series of input variables.

Other examples of classifying objectives are injury prediction algorithms in which we aim to estimate the player's risk of injury on the basis of multiple sources of information (injury history, age, load, physical qualities, etc.). The result, based on information such as the years that the player has been part of the club, their qualities or physical tests, availability or injuries, etc., could determine which players would be more successful in a sports career to reach the first team.

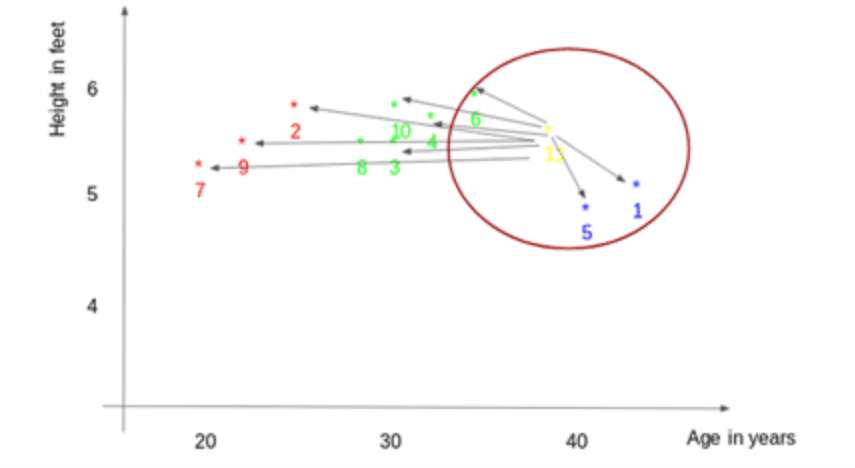
As we have seen, there are multiple algorithms that allow us to address different objectives. Each of them will use a different process to achieve the goal of making accurate predictions.

The most common regression algorithms include:

- Linear regression
- Regularized linear regression
- Decision tree regression
- Random forest regression
- Neural networks

- K-nearest neighbours

Figure 2: K-nearest neighbours algorithm



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

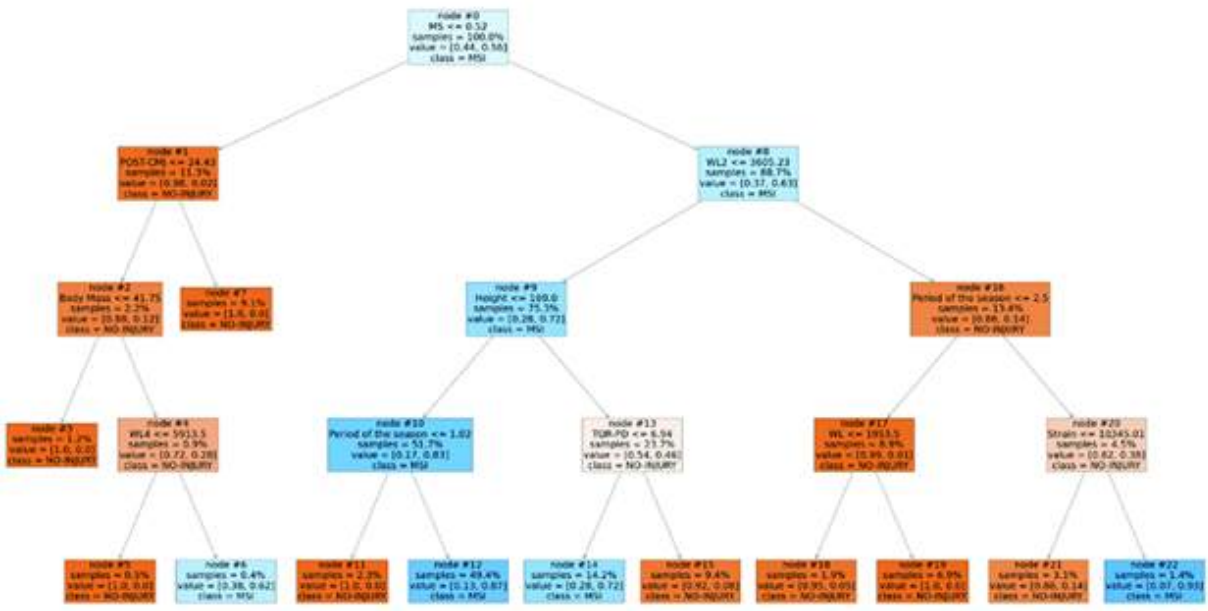
This graph shows a simple example where a K-nearest neighbours algorithm is applied. From 3 data variables - age, height and weight - we want to predict the weight of player 11 (in yellow). Simply put, to predict this value the algorithm uses the information of players with similar characteristics to those variables from player 11; and from the distance of these with player 11, the player's weight will be determined.

Among the most common classification algorithms, we find:

- Logistic regression

- Decision tree classification
- Random forest regression
- Neural networks
- K-nearest neighbours

Figure 3: Tree-structured classification model



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

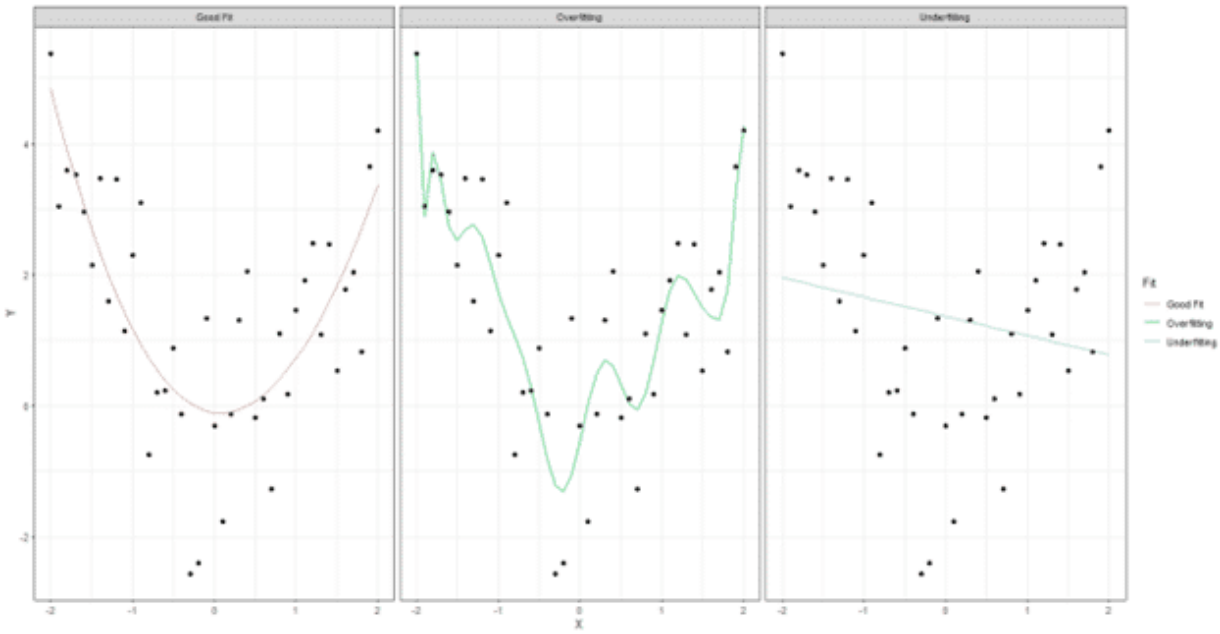
Mandorino et al. (2021) use a classification model that uses the structure of decision trees to predict muscle injuries in young football players. To this end, the model uses load variables and fatigue assessments. The result is as shown in the figure above. These charts allow us to describe the decision-making process used to determine if

an injury will occur. The process begins at the highest point of the graph: it indicates whether a condition is met or not; for example, if the values of a jump test are above a certain threshold. Depending on whether the answer is affirmative or negative, we will descend on the corresponding side of the decision tree, and so on until we reach one of the final "branches", which will determine the final prediction.

As we can see, on many occasions algorithms are common to both objectives - what will determine their use is the variable to be predicted, whether it is a continuous variable or a binary or categorical result.

If they have enough information and the requirements for their use have been met, algorithms may be very efficient in predicting outcomes. This is due to their great mathematical capacity and complexity to find relationships in the data and to explore a large number of alternatives. However, this also poses an important limitation: overfitting. This concept refers to the fact that the algorithm will work successfully on the data we have provided, but it will not be as successful if we use that algorithm with new or future data. The main objective of using Machine Learning techniques is to have greater predictive capacity for the data in which we are interested; therefore, we must find the balance to have a certain degree of success in our predictions based on the information we provide the model with.

Figure 4: Differences between underfitting and overfitting



Source: Author's own production

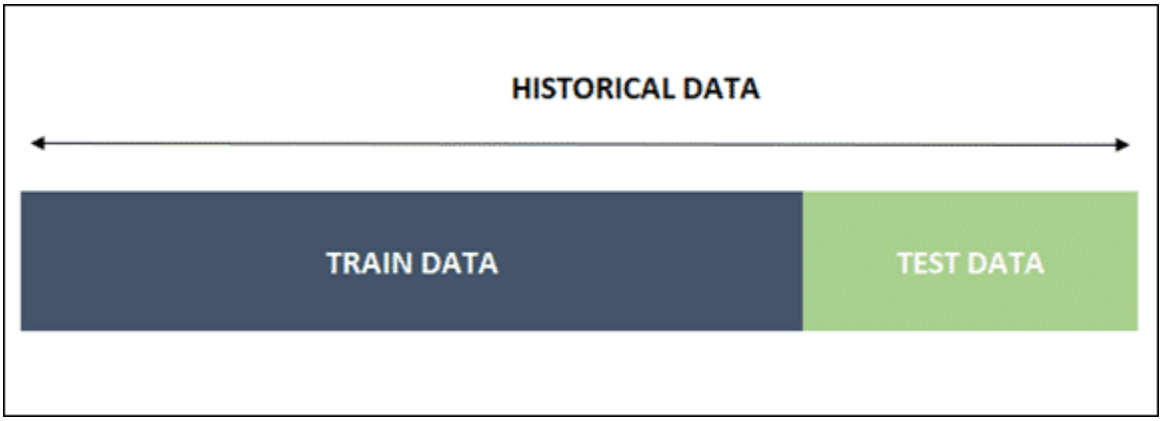
In the figure above, we can see the differences between underfitting, that is, very low predictive capacity, and overfitting: a model that is too specific on the current data.

To overcome this limitation, a process called training-test is used. This process consists of dividing the data we have for our model into two parts. Typically, 70% to 80% of the data will be part of the training data and the remaining percentage will be part of the test data.

In this way, an algorithm will be created from the training data and can be assessed based on the test data. The results on the test data

will show the performance of the algorithm and the technique used. If these results are not satisfactory, we can tweak the parameters or use other Machine Learning techniques to compare the results.

Figure 5: Dividing the total data into training data and test data



Source: Author's own production

Unsupervised learning algorithms

In the case of unsupervised algorithms, the objective will not be to predict a response variable, but to find patterns or associations between the data to achieve different purposes.

The unsupervised algorithm groups can be summarized into the following three categories:

- Unsupervised clustering algorithms: the objective of these algorithms is to find similarities between observations and

significant differences with other observations so as to determine groups of observations with similar characteristics. These algorithms are known as clustering.

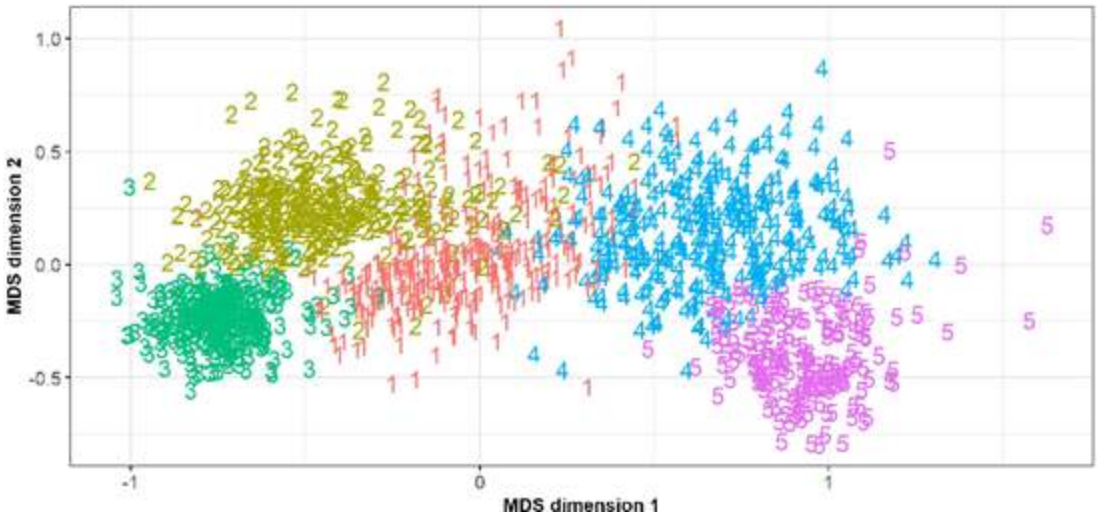
- Unsupervised dimensionality reduction algorithms: the objective of these algorithms is to transform data in which we have many variables into a summary with composite variables that show the relationships between them and do not provide repeated information. This will reduce the number of total variables; it will help in its visualization and its use by other Machine Learning algorithms.

- Unsupervised association algorithms: the goal of these algorithms is to find relationships between events that occur together. Simply put, when one event occurs, it is determined what is likely to occur in the next event. The most common example is Netflix recommendations, which are based on the movies or series we have already watched.

There are also multiple applications of these algorithms in sports performance. In the following graph (Figure 6) by Akhanli and Henning (2022), a clustering technique is used to find players with similarities using different performance metrics. The algorithm classifies all the players analysed (each of the numbers) into groups according to their similarities (5 groups in this case). These methods can be very useful in departments such as Scouting, to spot players in

databases who have similar performance to that of profiles we are looking for.

Figure 6: Multidimensional representation of clustering data

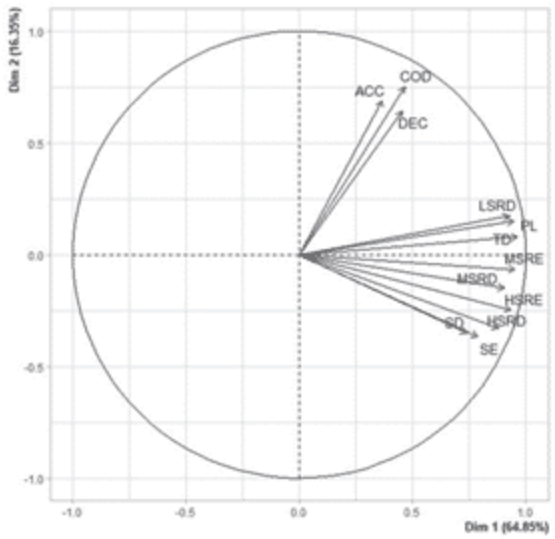


Source: Akhanli & Henning. (2022), p. 15.

An example of the application of dimensionality reduction algorithms can be easily found in the context of the Sport Scientist. If we use GPS load monitoring systems, these devices give us information on more than 100 calculated variables. We can opt for different methodologies to reduce the number of variables for our analyses and visualizations, such as establishing the correlations between them to determine which ones are providing the same type of information or using the literature to determine which aspects of the load are represented by each of them (mechanical, locomotive, metabolic load,

etc.). However, if we want to use all the information for a more complex project (such as injury prediction), we can choose to use a dimensional reduction algorithm which will show which of the variables explain the response variable to a greater extent and the percentage of information provided by each of them. These algorithms are known as principal component analysis (PCA). The principal components are a summary of relationships between variables (with different weighted values) that aim to explain the response variable.

Figure 7: PCA graph for the two main extracted components



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

The visualization prepared by Nosek et al. (2023) shows the two main components of the dimensionality reduction algorithm, which, in this

case, explain approximately 80% of the information. The first of them (horizontal axis) is made up of 8 variables - almost all of them are associated with volume (total distance, player load, etc.); and the second component (vertical axis) is made up of 3 variables related to mechanical load (accelerations, change of direction, decelerations). Using this technique, we no longer need 13 variables to explain the training load but we can use only two main components that provide us with this summarized information.

Weaving et al. (2019) also use PCA to summarize the player's load. As can be seen in the graph below (Figure 8), its two main components are related to high-speed load and global load. We can use these components to visualize the load in a summarized way throughout the player's season and, as shown in the graph, evaluate the conditional demand of the session taking into account the main components.

Figure 8: High-speed load and global load

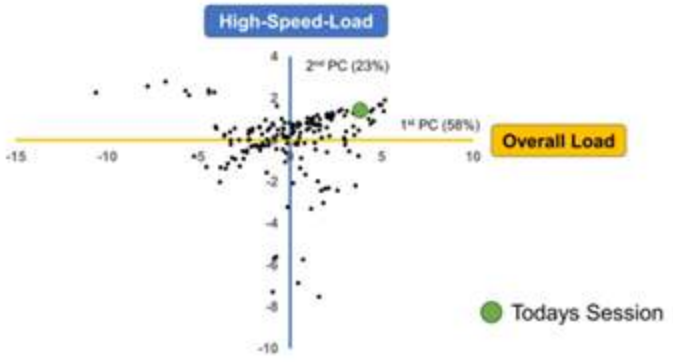
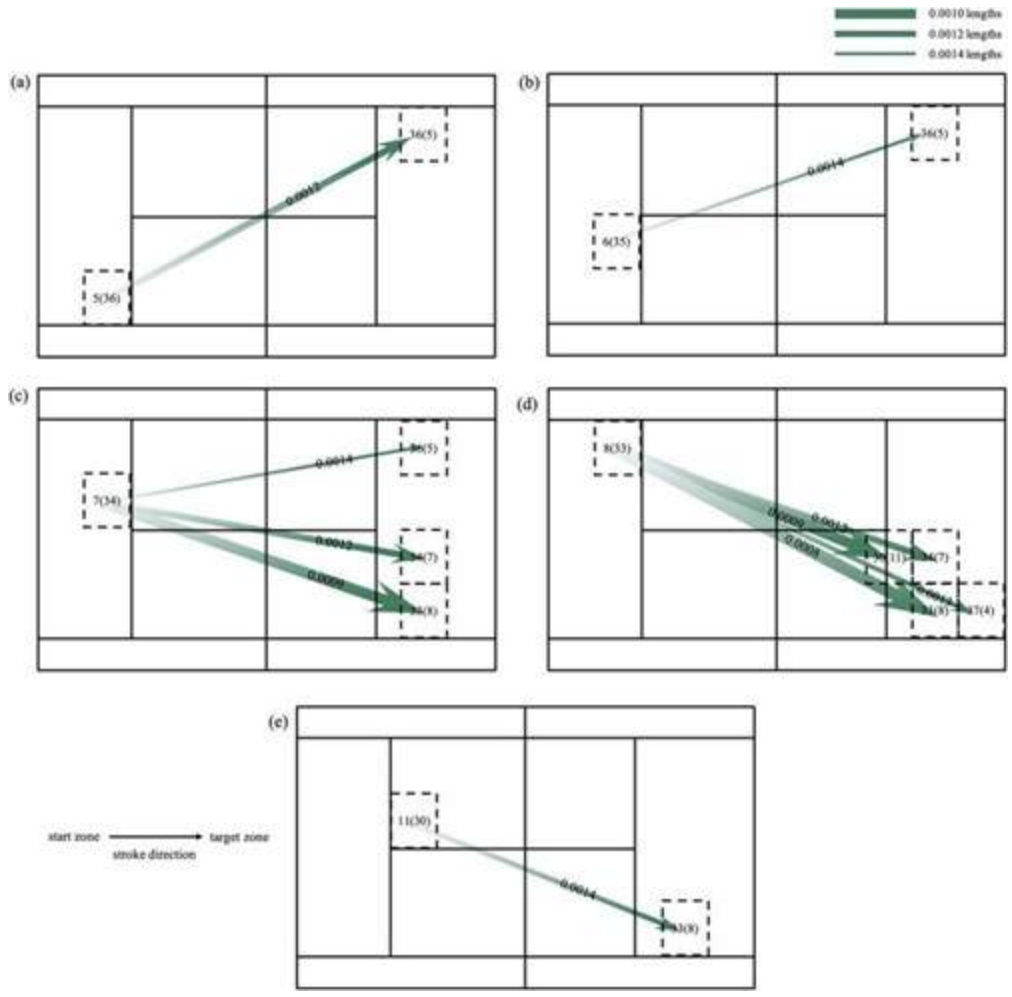


Figure 4. Scatterplot of the first (x axis) and second (y axis) principal component scores for the 169 training sessions for an individual player. The green data point highlights today's session in the context of every other training session completed.

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

Finally, in the study performed by Yunjing Zhou et al. (2023), we see an example of an association algorithm applied to sports performance. The aim of this study was to determine the patterns of strokes in a tennis match based on where the last stroke was made. This type of study has a clear application to the analysis of matches and opponents, and can be a great competitive advantage if it is correctly carried out.

Figure 9: Visualization of the patterns got by the association algorithm



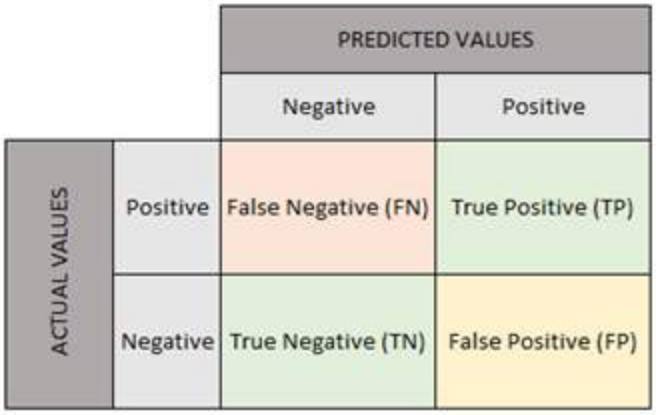
Source: Zhou et al. (2023), p. 9.

Assessing algorithms performance

In the same way that it is necessary to evaluate the linear models that we developed in the previous module, each of the groups of Machine Learning techniques have their corresponding assessment. As they are more complex models, there are multiple and nuanced evaluation metrics, but below there are the most important characteristics to assess:

- Classification models
    - Accuracy: rate of positive identifications that were actually correct.
    - Recall: rate of actual positives that were correctly identified.
    - F1: the combination of both aspects.

Figure 10: Confusion matrix



Source: Author's own production

- Grouping models:
    - Silhouette Score: a measure of separation and density of the groups detected.
- Dimensionality reduction models:

- A rate of explained variance: a percentage of variance that explains each of the main components.

CONTINUE

Activities

Considering supervised regression and classification algorithms, which of the following statements are correct?

- [ ] Regression algorithms are used to predict continuous or numerical variables.
- [ ] Classification algorithms are used to predict continuous variables.
- [ ] Classification algorithms predict variables with categorical or binary nature.
- [ ] Regression algorithms do not use mathematical functions to estimate the response variable.
- [ ] Both regression and classification algorithms focus on predicting continuous variables.

SUBMIT

Which of the following statements are correct regardinng the classification model used by Mandorino et al. (2021)?

- [ ] The classification model is based on the structure of neural networks.
- [ ] The model uses decision trees to predict muscle injuries in football players.
- [ ] The variables used by the model are the load and fatigue ratings.
- [ ] The decision-making process in the decision tree does not depend on whether the conditions are met or not.
- [ ] The model does not provide a final prediction when it reaches the final branches.

SUBMIT

Which of the following statements about assessing classification models are correct?

- [ ] Accuracy measures the rate of positive identifications that were actually correct.
- [ ] The recall measures the rate of actual positives that were incorrectly identified.
- [ ] The F1 metric is a combination of accuracy and recall.
- [ ] Recall measures the rate of positive identifications that were actually correct.
- [ ] Accuracy and recall are irrelevant metrics for evaluating classification models.

SUBMIT

References

Akhanli, S. E., & 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., & 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., & 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. (August 1, 2024). https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

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

Weaving, D., Beggs, C., Dalton Barron, N., Jones, B., & 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., & 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

CONTINUE