Modelo predictivo de la posición de juego en futbolistas sub-20 mediante machine learning

Predictive model of playing position in soccer players u-20 150 through machine learning

Autores

Javier Gaviria Chavarro [1]
Óscar Hernán Jiménez Trujillo [2]
Nelson Daniel Marín [3]
Juan José Gomez [4]
Francisco Castro Valencia [5]
Miguel Ángel Gómez García [6]
Carlos Julio Grisales Guerra [7]

[1] Universidad Santiago de Cali
[2, 6] Institución Universitaria Escuela Nacional del Deporte
[3, 4, 5, 7] Universidad San Buenaventura (Colombia)

Autor de correspondencia:
Javier Gaviria Chavarro
javier.gaviria@gmail.com

Recibido: 07-03-26
Aceptado: 30-05-26

Cómo citar en APA

Gaviria Chavarro, J., Jiménez Trujillo, Óscar H., Marin, N. D., Gómez Peñaranda, J. J., Castro Valencia, F., Gómez García, M. Ángel, & Grisales Guerra, C. J. (2026). Modelo predictivo de la posición de juego en futbolistas sub-20 mediante machine learning. *Retos, 83*, 266-276. https://doi.org/10.47197/retos.v83.118967

Resumen

Introducción: La identificación de la posición de juego mediante variables físicas puede contribuir a optimizar la planificación del entrenamiento en futbolistas jóvenes. Objetivo El objetivo fue desarrollar un modelo supervisado de clasificación multiclase para predecir la posición de juego en futbolistas sub-20 utilizando variables antropométricas y de rendimiento físico.
Metodología: Se realizó un estudio cuantitativo, observacional, transversal y predictivo con 160 futbolistas sub-20 distribuidos equitativamente en cuatro posiciones (porteros, defensas, mediocampistas y delanteros). Se evaluaron variables antropométricas, velocidad, agilidad, fuerza, potencia, resistencia aeróbica y carga de entrenamiento. Se compararon cinco algoritmos supervisados de clasificación y su desempeño se evaluó mediante validación cruzada estratificada y métricas de clasificación.
Resultados: El modelo Gaussian Naïve Bayes presentó el mejor rendimiento (F1-macro = 0.728), alcanzando una precisión global del 78% en el conjunto de prueba. Las variables con mayor importancia predictiva fueron el sprint de 30 m, el cambio de dirección (COD505), el Agility T-Test, el sprint de 20 m, el Yo-Yo Test y el VO_2max, mientras que las variables antropométricas mostraron menor contribución.
Conclusiones: La velocidad, la agilidad y la capacidad aeróbica fueron los principales predictores de la posición de juego en futbolistas sub-20. Los modelos de aprendizaje automático constituyen una herramienta objetiva para identificar perfiles posicionales y apoyar la individualización del entrenamiento.

Palabras clave

Aprendizaje automático; fútbol; modelo predictivo; posición de juego.

Abstract

Introduction: Identifying playing positions through physical variables may contribute to optimizing training planning in young soccer players by supporting objective decision-making during talent identification and individualized training prescription. Objective: The aim of this study was to develop a supervised multiclass classification model to predict playing position in under-20 soccer players using anthropometric and physical performance variables.
Methodology: A quantitative, observational, cross-sectional, and predictive study was conducted with 160 under-20 soccer players equally distributed among goalkeepers, defenders, midfielders, and forwards. Anthropometric variables, speed, agility, strength, power, aerobic endurance, and training load were assessed using standardized field-based tests. Five supervised classification algorithms were compared, and their performance was evaluated using stratified cross-validation and standard classification metrics, including accuracy, precision, recall, and macro F1-score.
Results: The Gaussian Naïve Bayes model achieved the best performance (macro F1-score = 0.728), reaching an overall accuracy of 78% on the independent test set. The variables with the greatest predictive importance were the 30-m sprint, COD505 change-of-direction test, Agility T-Test, 20-m sprint, Yo-Yo test, and VO_2max, whereas anthropometric variables contributed less to the prediction.
Conclusions: Speed, agility, and aerobic capacity were the main predictors of playing position in under-20 soccer players. machine learning models represent an objective tool for identifying positional profiles, supporting individualized training programs, and providing useful information for player assessment and long-term athletic development.

Keywords

Machine learning; playing position; predictive model; soccer.

Introduction.

Over the years, physical training has undergone considerable evolution, leading to changes in the way training programs are designed and implemented. Traditionally, athletic preparation was primarily based on coaches' experience and manual data analysis (Moya et al., 2025). However, physical demands have become progressively rigorous requiring athletes to tolerate high-intensity workloads while also managing passive and active recovery periods (López Betancourt et al., 2026). This context requires strength and conditioning professionals to adequately quantify the weekly training load of each athlete in order to promote the necessary physiological adaptations, support recovery processes, and facilitate the achievement of optimal performance at the highest competitive level.

Studies conducted in professional football players have reported differences in total distance covered per game, relative distance and maximal speed according to player position. Each position is associated with specific physical and physiological demands that athletes must meet in order to perform effectively during competition. For instance, center-forwards are generally characterized by lower volume and lower intensity demands, whereas midfielders tend to cover greater running volumes at moderate intensities and wingers are exposed to higher-intensity actions (Sánchez et al., 2026).

The physical and physiological variations among football athletes according to playing position provide a relevant basis for the development of classification models, as several variables may serve as discriminative indicators between positions. Among these variables, anthropometric and physical characteristics have been identified as particularly relevant. Additionally, some studies have reported that center-backs and wingers reach the highest peak speed during matches, whereas midfielders perform a greater number of sprints compared with other playing positions (Sánchez et al., 2024)

Machine learning techniques can be used to develop classification models by training algorithms on datasets or previous observations. In this process, the model identifies which variables are most closely associated with the target variable, and this learning procedure is largely automated. As a result, machine learning models may serve as predictive models, such as forecasting outcomes in future scenarios, or descriptive and explanatory purposes, by identifying how variables are structured and related within a given dataset (Alpaydin, 2020). In the football context, machine learning can synthetize information from repeated patterns in data and has been applied to estimate the probability of winning a match, predict fatigue-related responses in football athletes, identify variations in performance, and classify playing positions. Therefore, these techniques may contribute to optimizing training time and supporting decision-making in relation to problems that emerge during the training planning process,

Predicting playing position may provide valuable information for optimizing position-specific training in football. By identifying how physical capacities vary across competitive contexts, practitioners can individualize training loads and design technical-tactical tasks that better reflect each athlete's functional role and match-performance levels.

The integration of technology as a management tool plays a fundamental role in the work of strength and conditioning professionals, as it supports the assessment, control, and monitoring of athletes throughout the physical preparation process. Data collected from physical assessments provide relevant information about each athlete's performance status. Moreover, analyzing these data through predictive modeling may help practitioners anticipate possible performance-related responses and make more informed decisions regarding training load adjustment, recovery strategies, and technical-tactical interventions.

The prediction of playing position has gained increasing attention in sports science, as researchers have aimed to develop predictive models capable of classifying athletes' playing roles based on physical variables. Such models may provide a useful tool for optimizing training processes, identifying the physical demands associated with each position, monitoring and quantifying individual training loads, and tailoring training interventions to the specific requirements of each athlete's role on the field.

The aim of this study was to develop a supervised multiclass classification model to predict the playing position: PORTE (goalkeeper), DF (defender), MF (midfielder), DEL (forward), based on anthropometric and physical performance variables.

Method

The present study adopted a quantitative approach with a cross-sectional, observational, and predictive design. This design allowed researchers (A) to characterize the evolution of anthropometric and physical performance variables of under-20 (U-20) male soccer players through standardized assessments over a 3-week period, and (B) to develop a supervised multiclass classification model to predict the playing position of soccer players: PORTE (goalkeeper), DF (defender), MF (midfielder) and DEL (forward) using the collected variables. This approach provides a framework to analyze how variables behave during the assessment period and supports the development of models based on data sequences that result in classification of playing position (Hernández Sampieri et al., 2014). The research was conducted in different clubs' facilities in Cali, Valle del Cauca, Colombia. The field measurements were taken during a period of 3 weeks, outside the main competitive event.

Participants

A total of 160 male soccer players constituted the study population. All eligible players participated, meeting the inclusion criteria. In the U-20 category, participants were evenly distributed into four playing positions (40 forwards, 40 defenders, 40 midfielders, 40 goalkeepers) ensuring the representativeness of the sample, consistent with applied sport research where the population can be limited. Homogeneity in the number of participants per position facilitated the comparative analysis and the training of predictive models.

Procedure

To analyze the anthropometric and physical performance variables, previously validated assessments were selected (Shurley, 2019; Tatlıcı et al., 2022; Magee et al., 2021). Anthropometric variables included body weight (kg), height (cm), body mass index (BMI), body fat percentage (%). Physical variables related to agility and speed were assessed using the T-test (s), Illinois Agility Test (s), COD505 test (s), 10-m, 20-m, and 30-m sprint times (s) (Raya et al., 2013), whereas power and lower limb strength were assessed using CMJ height (cm), Relative Peak force (kg), and horizontal long jump (cm) (Jiménez-Reyes et al., 2011; Musham & Fitzpatrick, 2020; Sammoud et al., 2024). Aerobic capacity was measured through YoYo test (m), VO2max (ml $\cdot kg^{-1} \cdot min^{-1}$), RSA 6x20m (s), Fatigue Index (%). Finally, training load was quantified considering weekly training hours.

The data was obtained through standardized physical assessments, conducted over a 3-week period. Three weekly visits were scheduled at the different clubs' facilities. All assessments were conducted in the afternoon without prior training on the same day, to minimize fatigue effects and maintain consistent conditions. During each week, the players attended three assessment sessions planned on Monday, Wednesday, and Friday, ensuring at least one rest day between sessions to optimize recovery.

A weekly protocol was carried out across three consecutive weeks: a standardized warm-up that was performed prior to each session. The players used this protocol for 15 minutes to reduce the injury risk during the assessments and to prepare the neuromuscular system. The order was as follows:

Monday- visit 1: Body composition and anthropometric measurements were conducted. Body weight (kg), height (cm), body fat percentage (%) measured by bioelectrical impedance analysis. Body mass index was calculated from body weight and height. All assessments were performed at rest, following standardized protocols.

Wednesday-visit 2: Agility and speed trials were evaluated using 10-m, 20-m, and 30-m sprint time (s), COD505 test (s), Agility T-Test (s) Illinois Agility Test (s). The measurements were administered in a predetermined order with adequate rest periods between each trial to ensure optimal performance.

Friday- visit 3: Strength, power and endurance were assessed. The following variables were measured: long horizontal jump distance (cm), CMJ height (cm), relative peak force (kg), Yo-Yo test (m), VO_{2max} ($ml \cdot kg^{-1} \cdot min^{-1}$), repeated sprint ability (RSA) 6x20m (s) and fatigue index (%). Finally, a weekly questionnaire was administered to record training hours for each player.

Each test was monitored by the research team, who recorded the results in real time using digital data collection forms with automated calculation formulas. To ensure data comparability throughout the study period, assessments were conducted under standardized conditions. All records were coded following data security protocols maintaining confidentiality to only be used for data analysis purposes. Prior pilot testing was conducted to calibrate all equipment.

Ethical and methodological considerations

To promote data quality during the data collection process, predetermined guidelines were maintained across all assessment sessions. All assessments were conducted in the afternoon without prior training to reduce acute fatigue effects. Moreover, the researchers responsible for administering the tests received specific training to reduce measurement variability. Physiological and psychological states were expected to vary across days (e.g., accumulated fatigue, academic stress or exercise-induced stress, lack of sleep), for this reason, these factors were captured through repeated measurements. These variations were subsequently used as predictors within the model.

The study was carried out following international ethical guidelines, particularly those established in the Declaration of Helsinki. Before starting the research, each player received a clear explanation of the study purposes, procedures involved and possible risks or benefits of participating. Only the subjects who signed the informed consent participated in this study. To protect the players' identity, all data were coded numerically and stored in password-protected files, accessible only to the research team.

Data analysis

The statistical analysis was structured into four phases: data preprocessing, model training, performance evaluation, and interpretative analysis. All procedures were conducted in the Python 3.10 environment using the scikit-learn library (Pedregosa et al., 2011), NumPy (Harris et al., 2020), and Pandas (McKinney, 2010). The aim was to develop a supervised multiclass classification model to predict players' playing position: PORTE (goalkeeper), DF (defender), MF (midfielder), and DEL (forward) based on anthropometric and physical performance variables.

The dataset was examined to ensure consistency and completeness, with no extreme outliers or missing values identified in the analyzed variables. All quantitative predictor variables were standardized using the z-score method through the StandardScaler class, in order to homogenize measurement scales and prevent variables with larger magnitudes from disproportionately influencing model fitting.

The playing position variable, defined as the dependent variable, was treated as a multinomial categorical outcome. The dataset was split into a training set (80%) and a testing set (20%), preserving the original proportion of the four categories through stratified sampling (train_test_split, stratify = y). As each playing position contained the same number of observations (n = 40), no oversampling or synthetic balancing techniques were required, ensuring class balance without altering the original data structure.

Five supervised algorithms were implemented for multiclass classification:
(1) Multinomial Logistic Regression,

(2) Gaussian Naïve Bayes,

(3) Decision Tree,

(4) Support Vector Machines (SVM),

(5) K-Nearest Neighbors (KNN).

Each model was implemented within a pipeline integrating data standardization and classifier fitting. Hyperparameter optimization was performed using exhaustive grid search with five-fold stratified cross-validation (StratifiedKFold), selecting the F1-score as the primary evaluation metric, as recommended for assessing the balance between precision and recall in multiclass scenarios (Sokolova & Lapalme, 2009).

To estimate model stability, repeated stratified cross-validation (5×2 Repeated Stratified K-Fold) was applied, allowing assessment of performance variability across different data partitions. F1-score, macro precision, macro recall, and balanced accuracy were computed as comparison metrics. The model exhibiting the highest mean performance and the lowest standard deviation was selected as the final model.

The Gaussian Naïve Bayes model demonstrated the best overall performance and was therefore selected for final evaluation. Its performance was assessed on the independent test set, computing both global and class-specific metrics, including F1-score, precision, recall, and accuracy. A classification report and the corresponding confusion matrix were also generated to quantify correct classifications for each playing position (DEL, DF, MF, and PORTE).

The discriminative capacity of the model was evaluated through Receiver Operating Characteristic (ROC) curve analysis under a one-vs-rest multiclass framework (Fawcett, 2006), obtaining the area under the curve (AUC) for each class as well as micro and macro averaged AUC values. The precision of these estimates was determined using 95% confidence intervals (95% CI), computed via the bootstrap method with 5000 resamples (Efron & Tibshirani, 1993). Additionally, pairwise AUC comparisons between classes (e.g., MF vs. DF) were conducted using two-sided resampling-based tests.

The relevance of predictor variables was estimated using the permutation feature importance method, which evaluates the mean decrease in macro-averaged F1-score when the values of a given variable are randomly permuted (Breiman, 2001; Fisher et al., 2019). This nonparametric approach allows to identify the variables with the greatest impact on the model's predictive capacity without modifying its internal structure.

Variables were ranked according to their standard deviation and mean contribution, with those related to agility, aerobic capacity, and linear speed identified as the most influential. The relative magnitude of each predictor's importance was visualized using bar plots, enabling clear interpretation of the hierarchical contribution of the variables.

Results

A total of 160 athletes were analyzed and evenly distributed across the following playing positions: PORTE (40 goalkeepers) DF (40 defenders), MF (40 midfielders), DEL (40 forwards). Anthropometric and physical performance variables are presented in the table below.

Table 1. Anthropometric and physical performance variables.

Variable	Mean	Std. deviation	Minimum	Maximum
Height (cm)	180.77	7.11	164	198.5
Body Weight (kg)	73.04	6.38	58.9	89.5
Body fat (%)	10.29	2.28	6	18
BMI	22.41	2.23	18.3	30
Sprint 10 (m/s)	1.8	0.08	1.6	2.02
Sprint 20 (m/s)	3.08	0.11	2.81	3.3
Sprint 30 (m/s)	4.43	0.16	4.05	4.8
Agility T test (s)	10.04	0.52	9.03	11.5
COD505 (s)	2.36	0.09	2.17	2.6
Illinois Agility (s)	16.17	0.68	14.54	17.69
CMJ height (cm)	38.93	6.82	25	55
Long horizontal jump (cm)	230.92	16.97	180	273
YoYo test (m)	2520.68	368.12	1510	3200
VO2max (ml/kg/min)	56.09	3.82	48	66
Relative Peak Force (kg)	28.86	4.61	20	39.8
RSA 6x20m_mean_s	6.99	0.27	6.2	7.67
Fatigue Index %	4.24	1.02	1.72	8
Weekly training hours	6.56	1.11	5	8

Source: own elaboration.

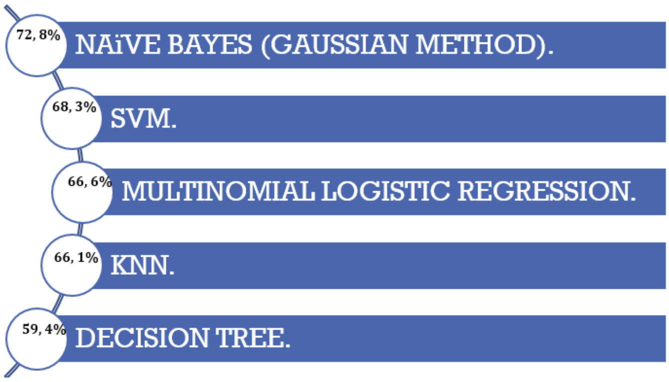
Height averaged 180.77 ± 7.10 cm (range: 164.0–198.5 cm), while body weight averaged 73.04 ± 6.38 kg and body fat percentage reached 10.28 ± 2.28 %. BMI ranged from 18 to 30 kg/m^2 with a mean of 22.41 ± 2.23.

Sprint performance averaged 4.43 ± 0.16 s over 30m, 3.08 ± 0.11 s over 20m and 1.80 ± 0.08 s over 10m. In agility assessments, COD505 recorded a mean time of 2.36 ± 0.09 s and Agility T-Test yielded 10.03 ± 0.52 s whereas the Illinois Agility Test showed 16.17 ± 0.68 s.

Long horizontal jump distance averaged 230.92 ± 16.97 cm, and CMJ height reached 38.93 ± 6.82 cm. Yo-Yo test performance reached a mean distance of 2520.68 ± 368.12 m, while VO_2max averaged 56.09 ± 3.82 ml·kg^{-1}·min^{-1}. Relative Peak Force averaged 28.86 ± 4.61 kg, and RSA performance yielded a mean time of 6.99 ± 0.27, and the fatigue index mean was 4.24 ± 1.02 %. Weekly training volume was 6.56 ± 1.11 hours.

Supervised algorithms were compared to determine which model exhibited the highest discriminative ability to predict playing position based on the selected predictors: Multinomial Logistic Regression, Gaussian Naïve Bayes, Decision Tree, Support Vector Machine (SVM), and K-nearest Neighbors (KNN). Model evaluation was conducted using five-fold stratified cross-validation, considering F1- macro score as the primary metric, complemented by balanced accuracy, precision and macro recall.

Figure 1. Models performance.



Source: authors' own elaboration (2025).

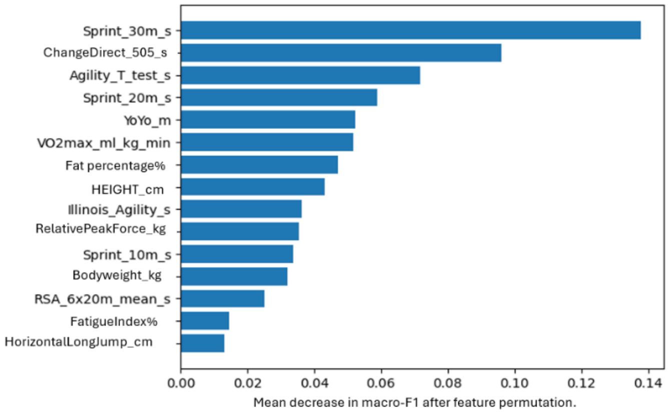
The Gaussian Naïve Bayes model achieved the best overall performance, with a mean F1-macro score of 0.728 ± 0.018, a balanced accuracy of 0.732, and a mean precision of 0.737. In contrast, the SVM (0.683 ± 0.037), Multinomial Logistic Regression (0.666 ± 0.037), k-Nearest Neighbors (0.661 ± 0.021), and Decision Tree (0.594 ± 0.081) models showed inferior performance.

Model stability was further confirmed through repeated stratified cross-validation (5×2), yielding an average F1-macro score of 0.721 ± 0.052 and a balanced accuracy of 0.730. These results demonstrated consistent behavior across different data partitions, outperforming both the SVM (0.675 ± 0.052) and Logistic Regression (0.671 ± 0.060) models.

On the independent test set, the model achieved an overall accuracy of 78%, with an F1-macro score of 0.78, a precision of 0.79, and a recall of 0.78. Position-specific F1-scores were 0.88 for midfielders (MF), 1.00 for goalkeepers (PORTE), 0.67 for forwards (DEL), and 0.59 for defenders (DF), indicating superior performance for positions with more distinct performance profiles.

Permutation feature importance analysis identified the variables that contributed most to the model's overall performance. The 30-m sprint time emerged as the most influential predictor (0.138 ± 0.039), followed by the COD505 change-of-direction test (0.096 ± 0.038) and the Agility T-Test (0.072 ± 0.032). A second tier of predictors included the 20-m sprint time (0.059 ± 0.042), Yo-Yo test distance (0.052 ± 0.019), and VO_2max (0.052 ± 0.042). Body fat percentage (0.047 ± 0.021) and height (0.043 ± 0.024) also contributed relevant information. Variables such as the Illinois Agility Test (0.036 ± 0.024), relative peak force (0.036 ± 0.029), 10-m sprint time (0.034 ± 0.023), and body mass (0.032 ± 0.025) showed lower influence, although they complemented the model's classification capacity.

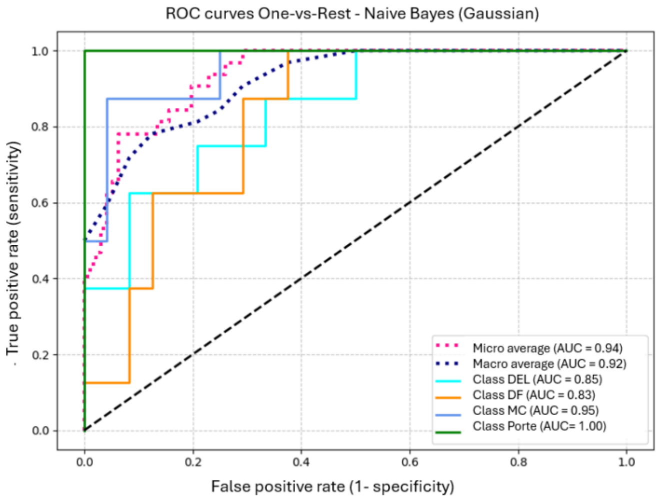
Figure 2. Permutation feature importance analysis.



Source: authors' own elaboration (2025).

The model performance was evaluated using ROC curves under a multiclass one-vs-rest approach. The micro- and macro- averaged AUC values were 0.94 and 0.92, respectively, indicating excellent overall discriminative ability. At the class level, AUC values were 1.00 for goalkeepers, 0.95 for midfielders, 0.85 for forwards and 0.83 for defenders.

Figure 3. ROC curves.



Source: authors' own elaboration.

AUC confidence intervals (95% CI) were calculated using bootstrapping resampling (5000 iterations), confirming the stability of the model: macro-average AUC = 0.92 (CI 95%: 0.84-0.98) and micro-average AUC = 0.94 (95% CI: 0.88-0.98).

At the class level, goalkeepers achieved an AUC of 1.00 (CI 95%: 1.00-1.00), midfielders 0.95 (95% CI: 0.86-1.00), forwards 0.85 (CI 95%: 0.68-0.98) and defenders 0.83 (95% CI: 0.66-0.95). A pairwise comparison between midfielders and defenders revealed a statistically significant difference (ΔAUC = 0.125; 95% CI:0.0188–0.2704; p = 0.0148), indicating a higher discriminative ability of the model for midfielders.

Discussion

The application of machine learning algorithms for positional classification in football enables the identification of performance patterns associated with the physical and functional demands of each role on the field. In the present study, the integration of physical and anthropometric variables provided adequate predictive capacity, with the goalkeeper position showing the highest classification accuracy within the model. These findings are consistent with recent studies that have used machine learning techniques to analyze performance in professional and youth football players (Yagin et al., 2023; Sneha et al., 2024; Moya et al., 2025). However, unlike previous studies that have primarily focused on variables obtained through GPS systems or competitive match metrics, the present study incorporated physical and anthropometric tests conducted in a functional assessment context, allowing for a more comprehensive approach to the player profile.

From a physical conditioning perspective, these findings reinforce the importance of monitoring conditional capacities during stages of athletic development and specialization. Variables such as speed, agility, and aerobic endurance not only allow practitioners to quantify the athlete's physical status, but also support the individualization of training processes according to the specific demands of each playing position. In this regard, authors such as Cometti (2007) and Azcárate and Yanci (2016) have emphasized that the physical demands of modern football require differentiated profiles according to the tactical function performed during play.

Anthropometric variables remain a relevant component in the characterization of football players, particularly in positions where stature and body composition may provide functional advantages. Nevertheless, in the present study, these variables showed a secondary contribution compared with physical performance variables. These results are consistent with those reported by Azcárate and Yanci (2016), who found that the most evident differences between playing positions were mainly related to speed and power rather than isolated body characteristics. Even so, recent research such as that of Medina Curimilma (2025) highlights that anthropometric characteristics remain useful for understanding biological individuality and guiding personalized training strategies.

Among the physical variables assessed, the 30-m sprint was one of the tests with the greatest capacity to discriminate between playing positions. This finding is consistent with studies conducted in youth football, where linear speed and acceleration capacity are considered key factors for differentiating positional profiles (Çetin & Koçak, 2022). Similarly, Yagin et al. (2023) reported that biomechanical variables related to acceleration and displacement show high sensitivity for the automatic classification of football players using predictive models.

In contrast, Santacruz Marín et al. (2025) did not find statistically significant differences in sprint tests between playing positions. However, it is important to consider that their study was conducted in an under-15 population with a considerably smaller sample size, which may influence the stability of the results due to differences associated with the process of biological maturation. In under-20 categories, such as the population analyzed in the present study, physical capacities tend to show a higher degree of consolidation, allowing more differentiated profiles to be observed according to competitive role.

Regarding the predictive models used, the present study incorporated five supervised classification algorithms: Naïve Bayes, support vector machine, k-nearest neighbors, multinomial logistic regression, and decision tree. This methodological diversity allowed different classification approaches to be compared and the best-performing model to be selected. In line with previous studies that have used machine learning algorithms to classify or predict football players' positions based on anthropometric, motor fitness, and biomechanical variables (Sneha et al., 2024; Yagin et al., 2023), the use of multiple models in the present study expanded the analytical possibilities and reduced the risk of interpreting the findings from a single statistical perspective.

The Naïve Bayes model showed the most stable and accurate performance in the study, achieving high values of precision and discrimination. Although this is a relatively simple model, the results suggest that the selected variables have sufficient capacity to differentiate playing positions. This finding partially agrees with previous studies in which machine learning models such as XGBoost and Random Forest demonstrated high predictive performance for football player classification and positional analysis

(Sneha et al., 2024; Yagin et al., 2023). However, in the present study, Naïve Bayes performed better, possibly due to the sample size and the combination of physical and anthropometric variables included in the analysis.

The ROC curves showed that the goalkeeper position achieved the highest predictive capacity within the model, with values close to 1. This behavior may be related to the physical and functional specificity of this position, as the demands placed on goalkeepers differ considerably from those of outfield players. High values were also observed for midfielders and forwards, whereas defenders showed lower discriminative capacity. These results suggest that some positions present more clearly defined physical profiles than others, thereby facilitating their automatic classification through machine learning algorithms.

These results are consistent with those reported by Sporiš et al. (2009), who found that football players occupying different positional roles present distinct physical and physiological profiles. In particular, midfielders showed higher values of relative oxygen consumption, maximal heart rate, maximal running speed, and blood lactate compared with defenders and attackers. Therefore, aerobic endurance appears to play a relevant role in midfielders due to the high locomotor and spatial coverage demands characteristic of this position.

Furthermore, multidirectional agility tests, such as the Illinois Agility Test, the T-Test, and the COD 505, demonstrated high sensitivity for discriminating lateral and offensive positions. Previous studies have indicated that change-of-direction ability and reaction capacity are relevant indicators of sport-specific performance in competitive football (Keiner et al., 2021; Horníková et al., 2021; Sammoud et al., 2024). In the present study, these variables showed considerable predictive weight within the Naïve Bayes model, particularly in players performing offensive transition roles and tactical width functions.

The findings are also related to those reported by Vidal-Maturana et al. (2024), who identified associations between isometric strength, jump power, and repeated-sprint ability in high-performance football players. Although strength and jump variables showed lower predictive weight in the present study, they remain relevant components within the comprehensive assessment of physical performance.

Overall, the results suggest that linear speed, agility, and aerobic capacity are the main indicators for differentiating playing positions in under-20 football players, whereas anthropometric and strength-related variables provide complementary information. The integration of machine learning techniques with traditional physical tests represents a useful tool for the objective assessment of performance, the identification of positional profiles, and the individualization of training loads within contemporary physical conditioning. Finally, future studies should incorporate technical-tactical variables, GPS-derived metrics, and factors related to biological maturation in order to strengthen the accuracy and applicability of predictive models in developmental and competitive football.

Conclusions

The results support the conclusion that speed, agility and aerobic endurance were the variables with the greatest contribution to distinguishing playing position in U-20 soccer players. The Naïve Bayes model showed the most stable and accurate performance, indicating that the selected variables provided sufficient information for a reliable classification model. Sprint time and change of direction tests revealed clear distinction among positions, highlighting that goalkeepers exhibited a more clearly defined physical profile compared with the other player groups.

It also became evident that anthropometric variables showed a lower contribution to the predictive process, highlighting the greater relevance of physical performance over body characteristics for establishing positional profiles in this category. This study specifically applies machine learning techniques to the field of sports performance, demonstrating that these approaches can complement existing assessment methods and support decision-making related to playing role identification and training planning.

Finally, it is suggested that future research should incorporate technical–tactical indicators, data derived from GPS systems, and variables related to biological development in order to enhance predictive power and develop more comprehensive models for performance analysis in youth soccer.

Funding

The study did not receive external funding. All procedures, equipment and measurement were provided by the research team's own resources

References

Alpaydin, E. (2020). Introduction to machine learning (4th ed.). MIT Press.
Azcárate Jiménez, U., & Yanci Irigoyen, J. (2016). Perfil físico en futbolistas de categoría amateur de acuerdo a la posición que ocupan en el campo. Revista Española de Educación Física y Deportes, (415), 21–37. https://doi.org/10.55166/reefd.v0i415.504
Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
Çetin, O., & Koçak, M. (2022). Repeated sprint ability of youth football players in the same age category according to playing position and competition level. Montenegrin Journal of Sports Science and Medicine, 11(1), 59–63. https://doi.org/10.26773/mjssm.220307
Cometti, G. (2007). La preparación física en el fútbol. Editorial Paidotribo.
Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. Chapman & Hall/CRC. https://doi.org/10.1201/9780429246593
Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*, 1–81.
Horníková, H., & Zemková, E. (2021). Relationship between physical factors and change of direction speed in team sports. Applied Sciences, 11(2), 655. https://doi.org/10.3390/app11020655
Jiménez-Reyes, P., Cuadrado-Peñafiel, V., & González-Badillo, J. J. (2011). Análisis de variables medidas en salto vertical relacionadas con el rendimiento deportivo y su aplicación al entrenamiento. Cultura, Ciencia y Deporte, 6(17), 113–119. https://doi.org/10.12800/ccd.v6i17.38
Keiner, M., Kapsecker, A., Stefer, T., Kadlubowski, B., & Wirth, K. (2021). Differences in squat jump, linear sprint, and change-of-direction performance among youth soccer players according to competitive level. Sports, 9(11), 149. https://doi.org/10.3390/sports9110149
McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a
Medina-Curimilma, W. E. (2025). La importancia de la antropometría en el rendimiento deportivo del fútbol. MQRInvestigar, 9(1), e58. https://doi.org/10.56048/MQR20225.9.1.2025.e58
Moya, D., Tipantuña, C., Villa, G., Calderón-Hinojosa, X., Rivadeneira, B., & Álvarez, R. (2025). Machine learning applied to professional football: Performance improvement and results prediction. Machine Learning and Knowledge Extraction, 7(3), 85. https://doi.org/10.3390/make7030085
Musham, C., & Fitzpatrick, J. F. (2020). Familiarisation and reliability of the isometric mid-thigh pull in elite youth soccer players. Sport Performance & Science Reports, 85, 1–4.
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
Raya, M. A., Gailey, R. S., Gaunaurd, I. A., Jayne, D. M., Campbell, S. M., Gagne, E., Manrique, P. G., Muller, D. G., & Tucker, C. (2013). Comparison of three agility tests with male servicemembers: Edgren Side Step Test, T-Test, and Illinois Agility Test. Journal of Rehabilitation Research and Development, 50(7), 951–960. https://doi.org/10.1682/JRRD.2012.05.0096
Sammoud, S., Negra, Y., Bouguezzi, R., Ramirez-Campillo, R., Moran, J., Bishop, C., & Chaabene, H. (2024). Effects of plyometric jump training on measures of physical fitness and lower-limb asymmetries in prepubertal male soccer players: A randomized controlled trial. BMC Sports Science, Medicine and Rehabilitation, 16, 37. https://doi.org/10.1186/s13102-024-00821-9
Santacruz Marín, C. A., España Chacua, W. A., Gaviria Chavarro, J., Motato Rodríguez, L. A., & Galeano Virgen, J. D. (2025). Modelo predictivo del sprint basado en salto contramovimiento e índice de masa corporal en futbolistas. Sportis Scientific Journal, 11(4), 1–19. https://doi.org/10.17979/sportis.2025.11.4.11646

Sneha, S., Prithvi, B. S., Niranjanamurthy, M., Kiran Kumar, H. K., & Dayananda, P. (2024). Machine learning based assessment of elite football players based on anthropometric and motor fitness parameters with regard to their playing positions. SN Computer Science, 5, 974. https://doi.org/10.1007/s42979-024-03261-x

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sporiš, G., Jukić, I., Ostojic, S. M., & Milanović, D. (2009). Fitness profiling in soccer: Physical and physiologic characteristics of elite players. Journal of Strength and Conditioning Research, 23(7), 1947–1953. https://doi.org/10.1519/JSC.0b013e3181b3e141

Vidal-Maturana, F., Gajardo-Cid, N., Merino-Muñoz, P., Hermosilla-Palma, F., Valdés-Badilla, P., & Herrera-Valenzuela, T. (2024). Relación entre potencia y fuerza isométrica con la capacidad de repetir sprint en futbolistas profesionales varones. SPORT TK-Revista Euroamericana de Ciencias del Deporte, 13(Supl. 3), 8. https://doi.org/10.6018/sportk.634671

Yagin, F. H., Hasan, U. C., Clemente, F. M., Eken, O., Badicu, G., & Gulu, M. (2023). Using machine learning to determine the positions of professional soccer players in terms of biomechanical variables. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology. Advance online publication. https://doi.org/10.1177/17543371231199814

Authors' details

Javier Gaviria Chavarro	javier.gaviria@gmail.com	Autor
Óscar Hernán Jiménez Trujillo	oscarjimenez601@gmail.com	Autor
Nelson Daniel Marín	ndmarinm@correo.usbcali.co	Autor
Juan José Gómez Peñaranda	jjgomezp1@correo.usbcali.edu.co	Autor
Francisco Castro Valencia	fcastrov@correo.usbcali.edu.co	Autor
Miguel Ángel Gómez García	gomezgarciamiguel440@gmail.com	Autor
Carlos Julio Grisales Guerra	cjgrisalesg@usbcali.edu.co	Autor