Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 15 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i15.52857

PAPER

Predictive Model for Physical Performance in Athletics: Correlation between Anthropometric Data and Cardiorespiratory Capacity in Students from a Private School

Jairo Samir Cornejo Vega[1], Genesis Andrea Ortiz Gomez[1], Everardo Sánchez Puche[2], Christian Ovalle[1,3](✉)

[1]Universidad Privada del Norte, Lima, Perú

[2]Corporación Universitaria Latinoamericana, Barranquilla, Colombia

[3]Universidad Tecnológica del Perú, Lima, Perú

dovalle@utp.edu.pe

ABSTRACT

In the current educational context, physical education and student sports development face challenges marked by continuous technological evolution. This study proposes a predictive model supported by machine learning and artificial intelligence (AI), establishing a connection between cardiorespiratory capacity (VO2max) and student anthropometric data. With a sample of 179 students aged 13 to 18, the model-building process included preparing and partitioning a dataset, training, and evaluation under the CRISP-DM methodology. A multiple linear regression model was applied, incorporating weight, age, height, sex, and body mass index (BMI) to analyze their relationship with the dependent variable (VO2max). Performance metrics revealed a significant correlation between anthropometric measurements and cardiorespiratory fitness (CRF), with a 24% improvement in training, although test accuracy was –0.8%. Including additional variables, such as sex and age, they have improved the predictive equations. However, the ability of the model to predict VO2max was limited, suggesting the complexity of the relationship between these factors. In a comprehensive evaluation, five linear regression models achieved a correlation accuracy of 22% with the complete data set.

KEYWORDS

predictive model, machine learning, cardiorespiratory capacity, anthropometric data, multiple linear regression

1 INTRODUCTION

Today, physical education and sports development for secondary school students face significant challenges, particularly in a world where technology is rapidly advancing [1]. With the widespread use of teaching acquisition devices such as

Vega, J.S.C., Gomez, G.A.O., Puche, E.S., Ovalle, C. (2024). Predictive Model for Physical Performance in Athletics: Correlation between Anthropometric Data and Cardiorespiratory Capacity in Students from a Private School. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(15), pp. 93–108. https://doi.org/10.3991/ijoe.v20i15.52857

Article submitted 2024-08-12. Revision uploaded 2024-10-05. Final acceptance 2024-10-14.

mobile phones and cameras and the fast-paced evolution of social media, teaching plays an increasingly vital role in daily life. However, the limitations in the performance of teaching processing systems introduce distortions that can diminish the aesthetic quality of teaching and hinder information acquisition [2]. [3] highlights that a significant challenge is extracting relevant features for evaluating multimedia teaching in physical education. Robust methods are needed to identify and assess critical features that reflect teaching quality. This necessitates studying methods for evaluating teaching quality to optimize the performance of acquisition equipment and teaching methods [4]. Technology has positively impacted education, as evidenced by research addressing artificial intelligence (AI), which has affected areas such as education [5, 6]. Modern education must adapt to societal needs, and technological tools have generated significant public interest and continue to evolve [7]. AI offers substantial benefits due to its ability to analyze large data volumes [8, 9].

In the educational sphere, AI has proven to be an invaluable tool for data analysis. This technology has been successfully integrated into various educational courses using specialized algorithms that provide valuable insights into each student's performance. In this context, its application in physical education is especially promising, offering many possibilities to enhance the learning experience and sports performance [10–12]. Childhood obesity is a significant public health challenge worldwide, with over 340 million children and adolescents being obese in 2016 [11]. Consequently, overweight children face higher cardiorespiratory risks during childhood, influenced by lifestyle, health, habits, and physical activity [13]. In this context, physical education teachers are crucial in promoting various strategies to motivate students according to their capabilities [14]. However, many negative determinants affect students' comprehension, such as scheduling conflicts, sleep disorders, low levels of physical activity, and health issues such as being overweight [15]. Moreover, the lack of variety in techniques and methodologies for optimal physical performance assessment contributes to the gaps that students face [16]. Physical inactivity is a leading cause of death, accounting for 9.4% of deaths, and ranks fourth in mortality, with health consequences such as diabetes, obesity, and cancer [17]. Additionally, poor student health can lead to difficulties in performing physical activities, causing cardiorespiratory problems [18]. Another significant factor in the current educational landscape is the digital divide among students [19]. Technological tools are not being implemented in personalized physical conditioning for students, exacerbating the challenges above [20].

Previous research has explored using AI in sports evaluation and physical performance enhancement. For instance, a study [21] developed an individual evaluation methodology incorporating AI in sports instruction and fitness testing. It used a chronic disease questionnaire to create a system that improved exercise prescriptions. With a sample of 100 individuals and using backpropagation neural networks (BPNN), the system achieved 92.5% accuracy in fitness evaluation. However, further research is needed to explore its applicability in real educational contexts.

Furthermore, physical education is fundamental in promoting physical health and overall well-being. Numerous studies and sources emphasize its importance in preventing chronic diseases. For instance, [22] evaluated 60 university students to obtain data on their cardiorespiratory fitness (CRF), endurance, agility, and flexibility. Unsupervised machine learning was used to group students into four physical fitness profiles: high performance, endurance, strength, and flexibility. The results indicated that students in the high-performance profile had better cardiovascular health than others ($p < 0.05$), highlighting the importance of more personalized and effective physical education curricula.

Similarly, [23] examined the relationship between CRF and risk variables in 52 healthy men. Using the Ruffier test, CRF was correlated with anthropometric and

cardiovascular measures. Results indicated that CRF was inversely associated with the risk of cardiovascular diseases and type 2 diabetes. This suggests that CRF assessment could be helpful in the early identification of chronic disease risks and that AI could be used to predict these risks. Additionally, [24] applied machine learning techniques to predict academic success in primary school students by evaluating variables such as strength, speed, flexibility, and aerobic capacity. Machine learning methods, including random forest (RF), support vector machine (SVM), and k-nearest neighbors (KNN), showed accuracy rates of 80.5%, 78.9%, and 77.4%, respectively, indicating that physical health can be a good predictor of academic success. In the study conducted by [25], machine learning algorithms, such as decision trees and SVMs, were used to estimate body mass index (BMI) accurately. With a dataset of 500 participants, the decision tree model predicted BMI with 80% accuracy, and the SVM model achieved 85%. The most critical factors for determining BMI were age, sex, height, and weight, demonstrating the potential of these algorithms to identify risks associated with obesity. Although interventions across various media or systems address this challenge, results are of a moderate level, necessitating a more effective and personalized approach [17–26]. In response, this study proposes an innovative solution to the problem of low physical performance and CRF during exercise [27]. By developing a system that analyzes personalized anthropometric data (weight, sex, BMI, age, and height) collected from a student population, it aims to revolutionize physical activity and improve physical capacity [28]. This system will use AI techniques and data analysis to objectively and accurately assess students through a cardiorespiratory test, providing a predictive model of their maximum oxygen consumption based on anthropometric measurements [29–31].

The study also explores the relationship between physical health and academic success, offering insights into how physical performance may influence academic outcomes. This can provide new perspectives for enhancing education and student well-being [32]. By optimizing and tracking personalized progress in each student's physical activity, an active lifestyle is encouraged [33]. In summary, the clear delineation of the topic demonstrates the feasibility of the study, as it focuses on a specific and relevant problem: developing an innovative predictive model that correlates anthropometric data and cardiorespiratory capacity in secondary school students using AI. Integrating AI into physical education allows for personalized and optimized physical performance evaluation, addressing the growing concern of childhood obesity and the lack of effective assessment methods. Consequently, this project innovatively combines technology and physical education to improve and provide the recommended physical conditioning for each student. By offering precise and tailored analysis of students characteristics, this study aims to enhance health and physical performance in an educational context, overcoming current limitations and promoting a more effective and personalized approach to school physical activity.

2 METHODOLOGY

The present research was organized using the cross-industry standard process for data mining (CRISP-DM) methodology, a standard for data mining projects consisting of phases. These phases are executed iteratively, which allows a flexible and adaptable approach to developing effective data mining solutions [34]. The applied methodology consists of five stages meticulously prepared to achieve the most effective application of the predictive model, specially adapted to our objectives. The following sections detail the first phase: data collection and construction. This is an essential step in developing a predictive model that will examine the relationship between the anthropometric indices of high-school students and their CRF.

In this way, a machine learning model will be used, as shown in Figure 1, where the proposed methodology is used, which consists of five phases such as data construction, dataset preparation, data splitting, model training, and algo-rhythm evaluation.

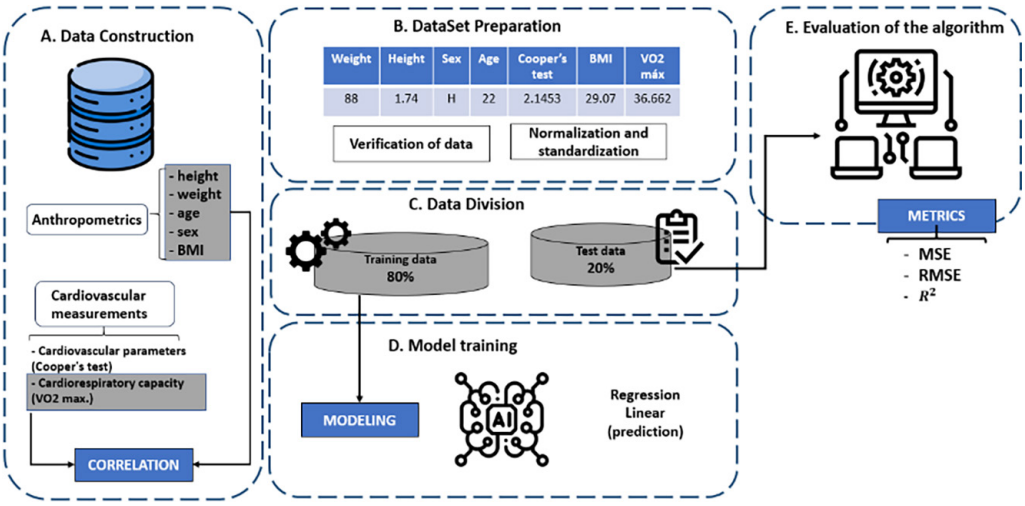

Fig. 1. Visual diagram of the proposed methodology

2.1 Data construction

The data construction phase for developing a predictive model was carried out thoroughly and exhaustively. Over approximately three weeks, accurate and relevant data was collected from high-school students in grades one through five in a controlled educational environment. These data included anthropometric measurements, such as BMI, age, sex, height, and weight, all obtained with precise instruments (see Figure 2a). In addition, cardiovascular measurements were carried out using the Cooper test, which measures the distance traveled by each student in a 12-minute running circuit, as seen in Figure 2b. The evaluation sample comprised 179 students, whose data was meticulously recorded in an Excel spreadsheet. This data calculated BMI and maximum oxygen consumption (VO2 max) using conventional formulas. The primary objective of this data collection was to conduct a comprehensive assessment of the students' cardiorespiratory capacity and physical fitness. Moreover, this data is paramount in developing a predictive model that establishes a relationship between anthropometric measurements and VO2 max, thereby enabling the prediction of an individual's cardiorespiratory capacity.

Fig. 2. Data collection: (a) Taking anthropometric data from the students, (b) Students performing the cardiorespiratory test

2.2 Dataset preparation

Once the anthropometric data, BMI, and VO2 max (maximal oxygen consumption) have been collected, the results must be reliable and meaningful conclusions to be drawn from them. During this data set preparation phase, a series of essential procedures were followed to ensure the validity of the results, as shown in Table 1. In addition, a number of crucial steps were taken in this preparation phase, involving your active participation to access, import, and process the collected data. Key Python libraries were imported, as these are essential for working with data and performing analysis. Similarly, importing data from a .csv file format was also involved.

Table 1. Viewing modified collected data

	ID	Weight	Size	Sex	Age	BMI	Cooper	VO2max
0	1	88.40	1.74	0	17	29.198045	2.09421	35.519688
1	2	52.40	1.60	0	14	20.468750	1.95452	32.397477
....	....	....	....	....	....	....	....	....
178	179	62.30	1.69	1	16	21.812962	1.87620	30.64.6946

In Figure 3, the "sep = ' ; ' " parameter is used to specify that the CSV file uses the semicolon (' ; ') as the field delimiter instead of the comma (' ; '), the default delimiter. This setting is critical for correctly interpreting the file, ensuring that the program reads and splits the data according to the actual delimiter used in the file. With this setting, the data processing could be correct, with incorrectly interleaved or mixed fields. Therefore, "sep = ' ; ' " allows the software to correctly recognize and handle CSV files that use the semicolon as a separator.

```
##IMPORT OF LIBRARIES
#Used for data processing and analysis
import pandas as pd
#Used to work with file paths
import os
#Used to work with graphics
import matplotlib.pyplot as plt
#Used to work with matrices, vectors and advanced mathematical functions.
import numpy as np
```

```
##Import of data in .csv format
mainpath = 'D:/ESTUDIOS_UPN/Ciclos/DECIMO_CICLO/CAPSTONE/ML'
filename = 'DataFinal.csv'
fullpath = os.path.join(mainpath,filename)
##Data reading without ';' termination
data = pd.read_csv(fullpath,sep=';')
data
```

Fig. 3. Import of libraries required for processing and reading data

2.3 Data division

In this data set preparation phase, significant actions related to the distribution of the dependent and independent variables were carried out. In Figure 4a, a new data frame called "filtrate" was generated using the Pandas "filter ()" method. This method allows the selection of specific columns from the original dataset that are

interesting for the analysis. In this context, the columns selected include "weight," "height," "age," "BMI," and "gender." In constructing and evaluating linear regression models, Python libraries were used, specifically Scikit-learn's "linear_model" library for linear regression modeling. Additionally, two metrics were imported to evaluate model performance: "mean_squared_error" to calculate the mean squared errors (MSE) between actual values and model predictions and "r2_score" to evaluate the ability of the model to fit the observed data. Moreover, it predicts the variability of the observed data.

When splitting data into training and test sets, the "train_test_split" function was used to achieve a random and controlled partition, as shown in Figure 4b. The proportion of data assigned to the test sets was set to 20%, and a fixed value of "random_state" was set to 42 to ensure the reproducibility of the division. This process was repeated for the "filtered" and "data" data sets, ensuring that both followed the same split structure.

a)

```
filtrados=data.filter(items=['weight','size','age','IMC','sex'])
filtrados
```

	weight	size	age	IMC	sex
0	88.40	1.74	17	29.198045	0
1	52.40	1.60	14	20.468750	0
2	67.55	1.71	16	23.101125	0
3	71.40	1.73	18	23.856460	0
4	65.20	1.70	16	22.560554	1
...	...	...	...	...	...

b)

```
##Linear regression model
from sklearn import linear_model
##Error measurements and approximations
from sklearn.metrics import mean_squared_error, r2_score
```

```
##Data partitioning is performed
from sklearn.model_selection import train_test_split
X_trainf, X_testf = train_test_split(filtrados, test_size=0.20, random_state =42)
X_traind, X_testd = train_test_split(data, test_size=0.20, random_state =42)
```

Fig. 4. Data division: (a) "filtrados" DataFrame, (b) Data division

2.4 Model training

In our study work's crucial phase of model formation, the multiple linear regression model was instantiated using the sci-kit-learn library. This instantiation became the fundamental tool for understanding and predicting the relationship between anthropometric variables such as sex, age, weight, height, BMI, and maximal oxygen consumption (VO2 max).

This process was divided into three essential steps. First, the independent variables were assigned, i.e., the characteristics used to predict the variable of interest, VO2 max. Next, the multiple linear regression model was fitted, allowing the model to find the best regression line that fit the training data. In this way, coefficients representing the contribution of each independent variable in predicting VO2 max were obtained. Finally, in Figure 5, we made predictions using the model previously fitted to the test data. This step was crucial as it allowed us to accurately estimate VO2 max values based on anthropometric characteristics. It represents a significant leap in our analysis, bringing us closer to understanding the intricate relationships and patterns that link these key variables to maximal oxygen consumption.

```
##Linear regression library called
modelo = linear_model.LinearRegression()
x=X_trainf #Independent variables are found
y=X_traind.VO2max

##We created a linear regression model.
modelo.fit(x,y)
```

```
LinearRegression()
```

Fig. 5. Linear regression model training

2.5 Algorithm evaluation

In the evaluation phase of our predictive model, an exhaustive analysis was performed to assess its ability to correlate critical variables, such as sex, age, weight, height, and BMI, with high-school students' maximal oxygen uptake (VO2 max). This process is fundamental to determining the model's effectiveness in its prediction objective.

A code used to calculate the performance metrics of the model is realized, which is crucial to evaluate the effectiveness of the model based on various metrics such as the calculation of the MSE, the root mean squared error (RMSE), and the coefficient of determination (R^2), allowing a comprehensive view of how the model behaves in terms of classification and prediction and allowing a detailed evaluation of its performance. The code exemplified in Figure 6a includes the functions necessary for generating these metrics and how the results are organized to facilitate their analysis. As part of our process, we selected a subset of the first 10 rows related to 20% of the collected data (test data), as shown in Figure 6b. This approach allowed us to visualize the first 10 data records, a practical step that was useful in previewing how the model predictions compare to the actual values in a specific portion of the test data.

(a)

```
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
# Calculate performance metrics
mse = mean_squared_error(y, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y, y_pred)
```

```
pred_y_data_20 = pd.DataFrame({'Real value':var, 'Predicted value':pred,'Difference':var-pred})
pred_y_data_20[0:10]
```

(b)

	Real value	Predicted value	Difference
78	12.742007	21.536819	-8.794811
16	28.621052	26.129738	2.491313
65	30.379405	31.282568	-0.903163
114	16.331131	29.083682	-12.752551
76	25.684354	25.879491	-0.195137
19	36.190665	32.525576	3.665089
122	33.110026	33.716896	-0.606870
24	32.191847	20.930873	11.260974
66	21.601944	26.364747	-4.762804
152	29.974181	29.930542	0.043639

Fig. 6. Evaluation of the predictive model: (a) Performance metrics (b) Evaluation with 20% testing data

Likewise, the same feature selection process was applied to a dataset comprising 10 rows, representing 80% of the data used for model training. This subset of training data was used to fit the model and evaluate its performance. In Figure 7, the differential between the actual data and the data predicted by the model is visualized. This visualization is fundamental to understanding how the model learns and generalizes from the training data.

```
pred_y_data = pd.DataFrame({'Real value':y, 'Predicted value':y_pred,'Difference':y-y_pred})
pred_y_data[0:10]
```

	Real value	Predicted value	Difference
158	18.304724	19.332136	-1.027412
31	29.889694	29.644330	0.245364
12	32.926972	26.850003	6.076968
51	27.741540	29.734493	-1.992953
41	21.449957	30.298927	-8.848970
85	27.270381	31.129528	-3.859147
93	29.840969	24.943838	4.897131
56	44.233002	25.470268	18.762734
38	26.965960	22.303217	4.662743
144	28.170455	25.579244	2.591211

Fig. 7. Evaluation with 80% training data

3 RESULTS

3.1 Evaluation of the model

As described in the proposed methodology, the present research work focused on applying a predictive model that correlates anthropometric measurements and maximal oxygen consumption of a high-school student through a cardiorespiratory test (Cooper's test). The application procedure was divided into five phases, emphasizing data collection and construction, data set preparation, data partitioning, model training, and algorithm evaluation. In addition, anthropometric data, such as weight, height, age, sex, and BMI, were extracted. Figure 8 present a detailed and comprehensive view of the relationships between these variables, highlighting relevant patterns and trends for further analysis.

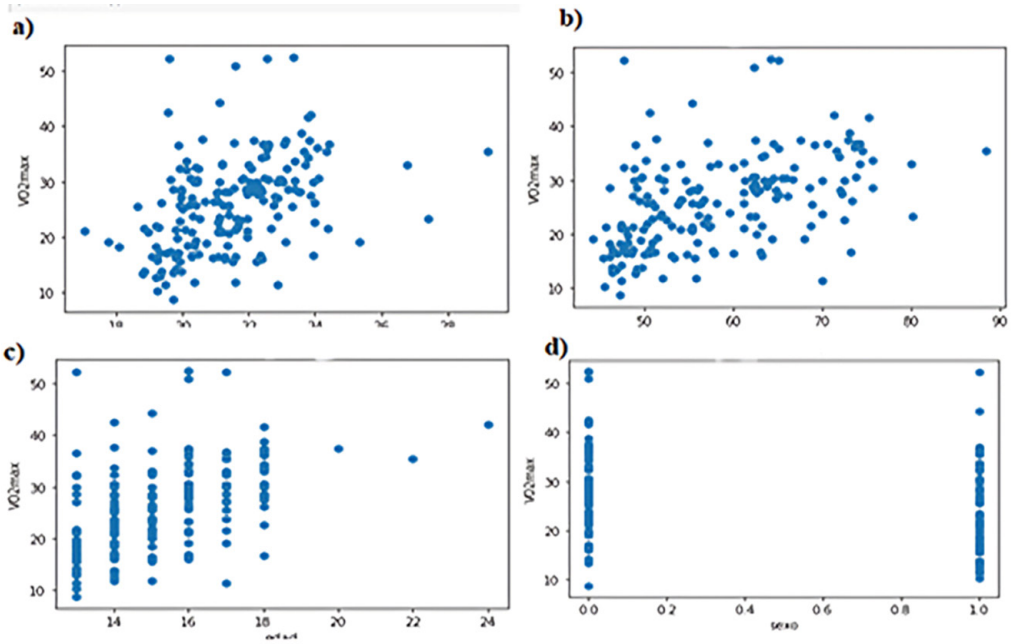

Fig. 8. Scatter diagrams based on the correlation of anthropometric variables:
a) BMI vs. VO2max, b) Weight vs. VO2max, c) Age vs. VO2max, d) Sex vs. VO2max

In addition, a fundamental tool was used: the correlation matrix. Figure 9 provides valuable information on the relationships between the different variables measured in the study. Here are some analytical conclusions based on the correlation coefficients presented in the graph:

a) **Strong positive correlation:** There is a strong positive correlation between weight and height (0.929), which is expected since generally, greater height is associated with greater weight. Age also shows a strong positive correlation with both weight (0.949) and height (0.930), indicating general physical growth as age increases.
b) **Moderate correlation:** The BMI shows moderate correlations with weight (0.933) and height (0.739), which reflects the relationship commonly used between weight and height to evaluate body composition.
c) **Negative correlation:** Sex presents a negative correlation with other variables, suggesting that these variables vary differently between men and women. This pattern is evident in the correlation with weight (–0.423), height (–0.426), age (–0.396), and BMI (–0.371).
d) **Correlation with performance variables:** Physical performance, measured through the Cooper test and VO2max, shows a moderate correlation with each other (0.457), indicating that the results of the Cooper test are related to maximum oxygen capacity.

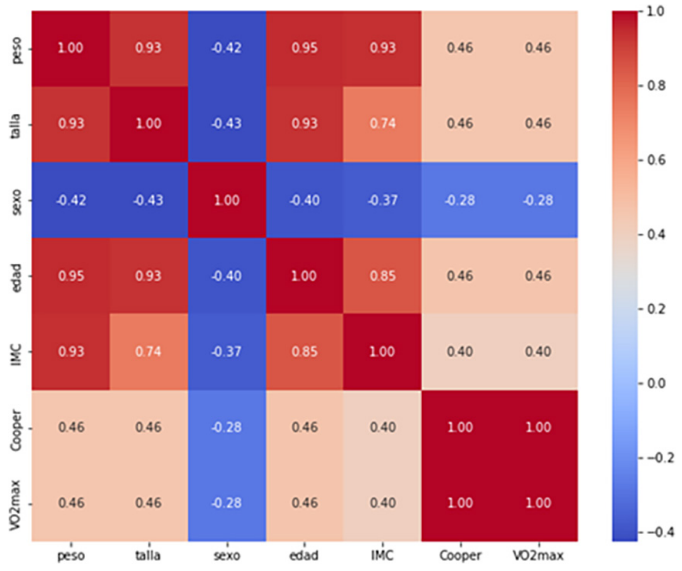

Fig. 9. Correlation matrix

Besides, to complete the predictive model's evaluation, a series of regression models were carried out sequentially to observe the behavior of the independent variables when correlated with maximum oxygen consumption. All models are statistically significant, as indicated by the extremely low p-value (e-value) in each case. This suggests that the null hypothesis that the variables have no impact on VO2max is unlikely. As the construction of the predictive models advanced, a significant improvement was observed in their ability to explain and predict VO2 max. From Model 1, which solely considered weight as a predictor of VO2max, to Model 5, which incorporated all variables, there was a clear and gradual increase in the R^2 and in the adjusted R^2. This increase indicates a substantial enhancement in the models'

capacity to elucidate variability in VO2max. Despite the slight decrease in overall significance due to the inclusion of additional variables such as age, sex, height, and BMI, an overall and optimal improvement in the correlation of the anthropometric variables evaluated was observed. This continuous refinement process underscores the importance of considering multiple factors in predicting cardiorespiratory performance. A summary of the models evaluated is presented in Table 2.

Table 2. Summary of the models

Name	Definition	F-Statistic	p-Value	CSR	R^2	Adjusted R^2	Mistake
Model 1	VO2max ~weight	46.9	1.19e–10	7.52	0.209	0.205	29.12%
Model 2	VO2max ~weight + age	24.24	5.01e–10	7.51	0.216	0.207	29.08%
Model 3	VO2max ~weight + age + sex	16.96	1.03e–09	7.47	0.225	0.212	28.91%
Model 4	VO2max ~weight + age + sex + height	12.79	3.82e–09	7.48	0.227	0.209	28.96%
Model 5	VO2max ~weight + age + sex + height + BMI	10.25	1.28e–08	7.47	0.229	0.206	28.93%

Finally, Table 3 provides additional statistics related to the overall quality of the predictive model. The Omnibus statistic of 34.822 indicates a significant lack of fit of the model to the data, indicating that the variables do not completely explain the variability in the data. The associated probability being close to zero reinforces this observation, suggesting that the model does not fit adequately and leaves patterns of variability unexplained. The value of the Durbin-Watson statistic of 2.207 indicates the possible presence of autocorrelation in the residuals, which could affect the validity of the inferences. The Jarque-Bera test, with a high value of 65.517 and an extremely low associated probability (5.93e–15), indicates that the residuals do not follow a normal distribution. Furthermore, the positive skewness (0.936) and high kurtosis (5.297) suggest asymmetry and heavy tails in the distribution of the residuals. The high condition number (2.33e+04) indicates the possible presence of multicollinearity, which could compromise the stability of the coefficient estimates. These results suggest the need for a thorough review of the model and consideration of possible improvements to increase its robustness and validity.

Table 3. Statistical analysis

Statistical Analysis	
Bus	34,822
Prob (Omnibus)	0.000
Skew	0.936
Kurtosis	5,297
Durbin-Watson	2,207
Jarque-Bera (JB)	65,517
Prob (JB)	5.93e–15
Cond. No.	2.33e+04

3.2 Expert evaluation

To ensure the validity and accuracy of our predictive model of physical performance in athletics, we have relied on the judgment of experts in exercise physiology and anthropometry. The evaluation of the data and methods used in our study, which explores the correlation between anthropometric characteristics and cardiorespiratory capacity in private school students, has been carried out with the participation of professionals with extensive experience in sports research and physical performance analysis. These experts have reviewed both the methodology used and the results obtained, providing critical feedback that has allowed the predictive model to be adjusted and refined. Table 4 shows that each of the 10 aspects evaluated received an individual rating from the experts. The overall average of these ratings was 4.18, with a standard deviation of 0.49. This average exceeds the acceptability threshold of 4, indicating that the experts perceive the predictive model and methods used as acceptable and sound.

Table 4. Expert evaluation

Items	Evaluation of Technical Aspects	Experts				Average	Standard Deviation
		1	2	3	4		
1	Accuracy of the predictive model	5	5	4	5	4.75	.43
2	Relevance of anthropometric data	3	4	4	4	3.75	.43
3	Quality of cardiorespiratory data	4	5	3	4	4.00	.71
4	Robustness of the model	4	4	3	4	3.75	.43
5	Usability of the model	5	4	4	3	4.00	.71
6	Consistency of results	3	4	5	4	4.00	.71
7	Applicability in real-world contexts	5	4	4	5	4.50	.50
8	Study methodology	5	5	4	5	4.75	.43
9	Interpretation of results	3	4	4	4	3.75	.43
10	Innovation and contribution to the field	4	5	4	4	4.25	.43
	Total					4.16	.49

A significant limitation of the research was the difficulty in generalizing the results obtained in studies with adults or controlled settings to secondary school settings. Although accurate in previous studies, AI models may not be directly applicable to high-school students due to differences in the population and conditions of educational settings. This limitation was addressed by conducting pilot tests in various high schools, which allowed us to adapt the predictive model to the specific characteristics of the students and validate its effectiveness in real educational contexts.

For future research, it is crucial to expand the validation of the predictive model on a broader variety of school settings and student populations to ensure its widespread applicability. In addition, integration of the model with other educational and technological approaches should be explored to create a more comprehensive and personalized solution. Research should also focus on evaluating the long-term impact of the model on student physical performance and health, as well as the effectiveness of implementing the model in different educational settings. Finally, developing clear guidelines for data protection and student privacy is essential, ensuring ethical and safe use of the information collected.

4 DISCUSSION

The study underscores the complexity of understanding CRF in high-school students, emphasizing the need to consider multiple variables, such as age, sex, height, and BMI, for more accurate estimates. The intricate interplay of these factors highlights the importance of adopting comprehensive approaches to address cardiovascular health in adolescents. However, despite the significant improvement of the models by including new variables, the F-statistics, low p-values, and adjusted R^2 and R^2 values suggest that, although an improved fit is achieved, there remains a limitation in the ability of these models to explain the full variability in CRF. This highlights the necessity for further research and refinement of analytical approaches, as the current models have their limitations.

The study [21], with a sample of 100 people aged 20–45 years, used BPNN to analyze data from health and fitness surveys, obtaining health assessments and exercise suggestions. The study [25], with 500 participants, applied decision trees and SVMs, achieving 80% and 85% accuracy in estimating BMI using anthropometric and clinical measures. The study [24], which included 380 elementary school students, employed algorithms such as RF and SVM to predict physical health-based academic success using strength, speed, flexibility, and aerobic capacity measures. Finally, the study [35], using data from the MPED program, applied exercise adherence prediction models with machine learning techniques and the DiPS tool, showing high sensitivity and specificity in prediction. For further details, refer to Table 5, which compares our proposed work with other existing work in the literature, classifying them by their respective models, results, and more.

Table 5. Comparison of the proposed work with existing work in the literature

Ref.	Sample (Population)	Algorithms	AI Model	Tools	Results
[21]	100 cases (20–45 years)	BPNN (Back Propagation Neural Networks)	Deep learning	Chronic Disease Survey Questionnaire	Health assessment, fitness scores, exercise suggestions
[25]	500 participants	Decision Trees, Support Vector Machines	Machine Learning	Anthropometric and clinical measurements	Accuracy of 80% (decision tree model), 85% (support vector machine model) for estimating BMI
[24]	380 primary school students	Random Forest, Support Vector Machine, K-Nearest Neighbor	Machine Learning	Strength, speed, flexibility, aerobic capacity	Prediction of academic success based on physical health
[35]	mPED Program Data	Adhesion Prediction	Machine Learning	DiPS (Disruption Prediction Score)	High sensitivity and specificity in predicting exercise adherence

5 CONCLUSION

In the present study, we utilized multiple linear regression to develop a predictive model that connects anthropometric measures to maximal oxygen uptake (VO2 max) in high-school students. Our goal was to comprehend the link between CRF and anthropometric indices. To ensure the reliability of our variables, we conducted a comprehensive data preparation process, which included meticulous modification, normalization, and verification. The predictive models initially showed the complexity of the relationship between anthropometric factors and CRF, although

their ability to predict VO2max was limited. Incorporating additional variables, such as height, sex, BMI, and age, significantly improved model performance. This underscores the importance of considering multiple factors when assessing cardiovascular health in adolescents. The model that initially only included body weight as a predictor of VO2max had modest predictive ability. However, the model's predictive capacity improved considerably when additional variables were included. The model's accuracy was evidenced by a significant reduction in the MSE, which decreased from 46.9 to 10.25. This indicates a marked improvement in the model's ability to explain variability in cardiorespiratory fitness.

The study suggests a correlation between anthropometric measures and CRF in high-school students but acknowledges that several variables influence this relationship. The researchers recommend conducting further studies with larger samples and including additional variables, such as dietary and exercise patterns, to better understand CRF in this population. They also suggest using advanced machine learning and analytical techniques to improve the model's predictive capacity, emphasizing the collection of high-quality and diverse data. These recommendations strengthen the validity and relevance of future studies and advance knowledge about cardiovascular health among high-school students.

6 AUTHOR'S CONTRIBUTION

To carry out this study, Christian Ovalle was instrumental in developing the predictive model, being in charge of the design and implementation that correlates anthropometric data with cardiorespiratory capacity, and performing an exhaustive literature review to provide a solid basis for the study. Everardo Sánchez designed the methodology, coordinated the data collection, and performed the statistical analysis to validate the model's accuracy. Jairo Cornejo Vega developed and optimized the AI system, conducting pilot tests in school settings and documenting the process for publication. Génesis Andrea Ortiz Gómez contextualized the study by reviewing the literature, supervised the implementation of the model in schools, and participated in writing the final report, integrating the findings and contributions of all authors for the final product.

7 REFERENCES

[1] M. Martínez-Comesaña, X. Rigueira-Díaz, A. Larrañaga-Janeiro, J. Martínez-Torres, I. Ocarranza-Prado, and D. Kreibel, "Impact of artificial intelligence on assessment methods in primary and secondary education: Systematic literature review," *Revista de Psicodidactica* (English ed.), vol. 28, no. 2, pp. 93–103, 2023. https://doi.org/10.1016/j.psicoe.2023.06.002

[2] S.-Y. Kim, "A qualitative exploration of an effective creative/convergent type of class in middle and secondary school physical education," *Int. J. Online Eng. (iJOE)*, vol. 16, no. 15, pp. 23–33, 2020. https://doi.org/10.3991/ijoe.v16i15.18721

[3] M. Baeten, E. Kyndt, K. Struyven, and F. Dochy, "Using student-centered learning environments to stimulate deep approaches to learning: Factors encouraging or discouraging their effectiveness," *Educ. Res. Rev.*, vol. 5, no. 3, pp. 243–260, 2010. https://doi.org/10.1016/j.edurev.2010.06.001

[4] X. Lifu, "Evaluation of multimedia teaching in physical education for data assimilation," *Int. J. Online Eng. (iJOE)*, vol. 14, no. 4, pp. 30–42, 2018. https://doi.org/10.3991/ijoe.v14i04.8366

[5] N. S. Alotaibi and A. H. Alshehri, "Prospers and obstacles in using artificial intelligence in Saudi Arabia higher education institutions—the potential of AI-based learning outcomes," *Sustainability*, vol. 15, no. 13, p. 10723, 2023. https://doi.org/10.3390/su151310723

[6] R. Bucea-Manea-Țoniș *et al.*, "Artificial intelligence potential in higher education institutions enhanced learning environment in Romania and Serbia," *Sustainability*, vol. 14, no. 10, p. 5842, 2022. https://doi.org/10.3390/su14105842

[7] M. Kuznetsova, "Increasing schoolchildren's motivation for physical activity: Theoretical aspects," *Current Res. J. Pedagogics*, vol. 2, no. 10, pp. 247–252, 2021. https://doi.org/10.37547/pedagogics-crjp-02-10-44

[8] B. Feng, "Dynamic analysis of college physical education teaching quality evaluation based on network under the big data," *Computational Intell. Neurosci.*, vol. 2021, no. 1, pp. 1–13, 2021. https://doi.org/10.1155/2021/5949167

[9] C. K. Y. Chan and W. Hu, "Students' voices on generative AI: Perceptions, benefits, and challenges in higher education," *Int. J. Educ. Technol. Higher Educ.*, vol. 20, 2023. https://doi.org/10.1186/s41239-023-00411-8

[10] J. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *Comput. Human Behav.*, vol. 98, pp. 166–173, 2019. https://doi.org/10.1016/j.chb.2019.04.015

[11] M. Ng *et al.*, "Global, regional and national prevalence of overweight and obesity in children and adults during 1980–2013: A systematic analysis for the Global Burden of Disease Study 2013," *Lancet*, vol. 384, no. 9945, pp. 766–781, 2014.

[12] H. Yu and Y. Mi, "Application model for innovative sports practice teaching in colleges using internet of things and artificial intelligence," *Electronics*, vol. 12, no. 4, p. 874, 2023. https://doi.org/10.3390/electronics12040874

[13] R. Ross *et al.*, "Importance of assessing cardiorespiratory fitness in clinical practice: A case for fitness as a clinical vital sign: A scientific statement from the American Heart Association," *Circulation*, vol. 134, no. 24, pp. e653–e699, 2016. https://doi.org/10.1161/CIR.0000000000000461

[14] J. Martins, S. Honório, and J. Cardoso, "Physical fitness levels in students with and without training capacities – A comparative study in physical education classes," *Retos*, vol. 47, pp. 43–50, 2023. https://doi.org/10.47197/retos.v47.94656

[15] M. Henriquez-Beltrán *et al.*, "Association between sleep problems and school performance: Results of the survey of health and school performance in the province of Biobío 2018," *Andes Pediatrica*, vol. 93, no. 2, pp. 235–246, 2022. https://doi.org/10.32641/andespediatr.v93i2.3734

[16] Q. Chen and M. Dong, "Design of assessment judging model for physical education professional skills course based on convolutional neural network and few-shot learning," *Computational Intell. Neurosci.*, vol. 2022, no. 1, pp. 1–11, 2022. https://doi.org/10.1155/2022/7548256

[17] C. Vandelanotte *et al.*, "Increasing physical activity using a just-in-time adaptive digital assistant supported by machine learning: A novel approach for hyper-personalized mHealth interventions," *J. Biomed. Informatics*, vol. 144, p. 104435, 2023. https://doi.org/10.1016/j.jbi.2023.104435

[18] F. Ortega *et al.*, "Physical fitness in childhood and adolescence: A powerful marker of health," *Int. J. Obes.*, vol. 32, pp. 1–11, 2008. https://doi.org/10.1038/sj.ijo.0803774

[19] Y. Ba and Z. Liu, "Design and research of physical education platform based on artificial intelligence," *Scientific Programming*, vol. 2022, no. 1, pp. 1–7, 2022. https://doi.org/10.1155/2022/9327131

[20] M. Zafari, A. Sadeghi-Niaraki, S.-M. Choi, and A. Esmaeily, "A practical model for evaluating the performance of secondary school students based on machine learning," *Appl. Sci.*, vol. 11, no. 23, p. 11534, 2021. https://doi.org/10.3390/app112311534

[21] Y. Li and X. Li, "Artificial intelligence system for generating guidance models in sports education and physical fitness evaluation under deep learning," *Front. Public Health*, vol. 10, 2022. https://doi.org/10.3389/fpubh.2022.917053

[22] D. A. Bonilla *et al.*, "Elaboration of physical fitness profiles of physical education students using unsupervised machine learning," *Int. J. Environ. Res. Public Health*, vol. 20, no. 1, p. 146, 2022. https://doi.org/10.3390/ijerph20010146

[23] K. A. Alahmari *et al.*, "Cardiorespiratory fitness as a correlate of cardiovascular, anthropometric, and physical risk factors: Using the Ruffier test as a model," *Can. Respir. J.*, vol. 2020, pp. 1–10, 2020. https://doi.org/10.1155/2020/3407345

[24] K. Xu and Z. Sun, "Prediction of academic performance associated with physical fitness of primary school students using machine learning methods," *Complementary Therapies Clin. Pract.*, vol. 51, p. 101736, 2023. https://doi.org/10.1016/j.ctcp.2023.101736

[25] I. L. Acosta-Guzman, E. A. Varela-Tapia, C. I. Acosta-Varela, and J. D. Tumbaco-Bravo, "Prediction of body mass index (BMI) using support vector machine and decision tree algorithms of AI," in *Proc. XXI Ibero-American Conf. Systems, Cybernetics Inf. (CISCI 2022)*, N. Callaos, J. Horne, B. Sánchez, and A. Tremante, Eds., 2022, pp. 18–23. https://doi.org/10.54808/CISCI2022.01.18

[26] D. Yang, E.-S. Oh, and Y. Wang, "Hybrid physical education teaching and curriculum design based on an interactive voice artificial intelligence educational robot," *Sustainability*, vol. 12, no. 19, p. 8000, 2020. https://doi.org/10.3390/su12198000

[27] J. Albornoz-Guerrero, R. Zapata-Lamana, D. Reyes-Molina, I. Cigarroa, G. García Pérez de Sevilla, and S. García-Merino, "Overweight/Obesity and low muscle strength in schoolchildren have lower cardiovascular capacity and higher cardiovascular risk: Results from the 2019 School Health Survey of Southern Chile," *Children*, vol. 8, no. 9, p. 734, 2021. https://doi.org/10.3390/children8090734

[28] F. Cao, M. Lei, S. Lin, and M. Xiang, "Application of big data-based artificial intelligence technology in physical education reform," *Mobile Information Systems,* vol. 2022, no. 1, pp. 1–12, 2022. https://doi.org/10.1155/2022/4017151

[29] R. Moore, L. Edmondson, M. Gregory, K. Griffiths, and E. Freeman, "Barriers and facilitators to physical activity and digital intervention in inactive British adolescents in secondary schools: A qualitative study with physical education teachers," *Front. Public Health*, vol. 11, 2023. https://doi.org/10.3389/fpubh.2023.1193669

[30] M. Hu and J. Wang, "Artificial intelligence in dance education: Dance for students with special educational needs," *Technol. Soc.*, vol. 67, p. 101784, 2021. https://doi.org/10.1016/j.techsoc.2021.101784

[31] S. Quinart, F. Mougin, M.-L. Simon-Rigaud, M. Nicolet-Guénat, V. Nègre, and J. Regnard, "Cardiorespiratory fitness assessment through three field tests in obese adolescents: Validity, sensitivity, and prediction of peak VO2," *J. Sci. Medicine Sport*, vol. 17, no. 5, pp. 521–525, 2014. https://doi.org/10.1016/j.jsams.2013.07.010

[32] Z. Teng and S. Cai, "Application of computer-assisted instruction (CAI) in physical education: An analysis of surveys from Chinese universities," *J. Healthcare Eng.*, vol. 2021, no. 1, pp. 1–6, 2021, https://doi.org/10.1155/2021/1328982

[33] A. Ahmed *et al.*, "Portable artificial intelligence for assessing physical activity in secondary school children," *Sustainability*, vol. 15, no. 1, p. 638, 2023. https://doi.org/10.3390/su15010638

[34] P. Haya, "La metodología CRISP-DM en ciencia de datos – IIC," *Instituto de Ingeniería del Conocimiento*, Mar. 15, 2024. [En línea]. Available: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

[35] M. Zhou, Y. Fukuoka, K. Goldberg, E. Vittinghoff, and A. Aswani, "Application of machine learning to predict future adherence to physical activity programs," *BMC Med. Inform. Decis. Making*, vol. 19, 2019. https://doi.org/10.1186/s12911-019-0890-0

8 AUTHORS

Jairo Samir Cornejo Vega is a Systems Engineer graduated from the Universidad Privada del Norte, Lima-Peru. He has a bachelor's degree. His strengths include network infrastructure, data mining, artificial intelligence, technical support and programming (E-mail: jairocornejovega@gmail.com, N00197855@upn.pe).

Genesis Andrea Ortiz Gomez is a Systems Engineer graduated from the Universidad Privada del Norte, Lima-Peru. She has a bachelor's degree. Her strengths include artificial intelligence, computer graphics, and software modeling and analysis (E-mail: andrea_ortizgomez23@hotmail.com, N00211794@upn.pe).

Everardo Sánchez Puche is Scientific Advisor for Performance and Elite Athletes. Research Professor, Bachelor's Program in Physical Education, Latin American University Corporation CUL, Barranquilla-Colombia. He is a candidate for a Doctor of Education (E-mail: esanchezp@ul.edu.co).

Christian Ovalle is a Systems Engineer graduated from the Universidad Peruana Union belonging to the top third, Master in Management and Management of Information Technology, with a Diploma in Process Management and Specialty in operations and Logistics from ESAN University, with extensive experience in business projects, and with a passion for research (E-mail: dovalle@utp.edu.pe).