PREDICT BODY PERFORMANCE CLASS, BASED ON PHYSICAL CHARACTERISTICS & PERFORMANCE ON PHYSICAL TESTS

M.Y. Kavinda
(s15055)
IS4007-Project

1.Abstract

Physical performance is a multifaceted aspect of overall health and well-being, influenced by a variety of physical and physiological factors. So, predicting an individual's physical performance class is a very interesting topic. This study aims to investigate the relationship between physical characteristics and performance on physical tests and develop a machine learning model for predicting an individual's performance class based on their physical characteristics and test results and through that try to significantly expands the scope by addressing the broader prediction of overall body performance, going beyond the sole focus on sports performance. Cleaned dataset which used some preprocessing methods and built some new variables by combining some variables, utilized in this study consists of individuals aged between 20 and 64 years, encompassing both males and females. The physical characteristics considered like height & weight and Physical tests were conducted to assess strength, power, endurance, and flexibility. Employing a machine learning approach, this study demonstrates the potential for accurately predicting an individual's performance class based on their physical characteristics and performance test results. The developed model provides valuable insights into the factors that contribute to overall body performance. By using Feature selection methods identified that the variables “sit and bend forward”, “sit -ups”, “age”, etc., excluding “MAP value”, play an important role in prediction of the class of body performance. After comparing accuracy, f1 -score and lower error rate of each model, “XGBoost” model was selected as the best model that provides a good fit for the data with approximately 75% accuracy level for the test data.

2.Table of Content

3.List if Figures and Tables

FIGURES

TABLES

4.Introduction

The prediction of body performance data has become an area of great interest in recent years. Body performance refers to an individual's physical abilities and achievements, including factors such as strength, speed, endurance, and agility. Understanding and accurately predicting body performance class, based on physical characteristics and performance on physical tests can provide valuable insights for personalized training programs, health assessments, and performance monitoring.

The objective of this study is twofold. Firstly, it aims to develop a predictive model that accurately determines an individual's performance class, based on their physical characteristics and performance on physical tests. By analyzing a comprehensive dataset, It strive to create a robust model capable of accurately categorizing individuals into different performance classes. This model will offer personalized guidance and support individuals in achieving their fitness goals. As well as, secondly, It seek to identify the factors that significantly influence body performance and determine their impact on determining performance data. By examining the relationships between physical characteristics, performance on physical tests, and performance class, it aims to uncover the most influential factors that contribute to an individual's performance. This understanding will enhance our knowledge of human performance and help optimize training strategies and performance outcomes. Significance of the study

The significance of this study can be seen in various domains. Firstly, the developed predictive model can be applied in real-world settings, providing personalized training program recommendations, health assessments, and performance monitoring. This can assist individuals in optimizing their training efforts and achieving their fitness goals. In other hand, the study's findings will contribute to the broader research field by expanding our understanding of the interactions between physical characteristics, performance on physical tests, and performance class. By uncovering the key factors that influence body performance, this research can inspire further investigations and studies in related areas. It may also lead to advancements in performance prediction and optimization. And lastly, the comprehensive body performance class predictive project created through this study will serve as a valuable resource for researchers, healthcare professionals, and fitness enthusiasts. It can facilitate collaborations, knowledge sharing, and further studies on performance prediction and related fields. This project will contribute to the collective knowledge in the field and provide a foundation for future research endeavors.

5. Literature Review

Predicting body performance in sports and daily activities has become a popular research area. This literature review explores relevant studies that focus on predictive models and techniques, examining their purpose, factors considered, and the methods used for predicting body performance.

1. **Hindawi (2022) "Prediction of Sports Performance and Analysis of Influencing Factors Based on Machine Learning and Big Data Statistics", Journal of Sensors (6): 1-9.**

The purpose of this study is to formulate short-term, medium-term, and long-term sports development planning and policy services for decision-makers. It focuses on predicting and analyzing sports test scores based on big data. Machine learning algorithms based on big data statistics are explored, including naive Bayesian classification and big data statistical classification algorithms are used as techniques and methods and the study identifies significant differences in the skills of male and female students.

2. **Dr. J. Stebbins (2021) "Machine-learning models for activity class prediction", Journal of Gait & Posture volume 89.**

This study aims to identify an appropriate combination of feature subsets and prediction algorithms for activity class prediction using raw acceleration data collected from the hip. The research suggests that simple time-domain (TD) features can be sufficient for predicting activity classes. Various feature selection methods, including filter-based, wrapper-based, and embedded algorithms, are employed. Classification is performed using artificial neural networks (ANN), support vector machines (SVM), and random forests (RF).

3. **Şimşek, Mehmeta; Kesilmiş, İnci (2022)"Predicting athletic performance from physiological parameters using machine learning", Journal of Sports Analytics, vol 8.**

This study investigates the impact of balance parameters on performance class in bocce athletes. Dynamic balance features are considered, and techniques such as hierarchical agglomerative clustering (HACA) and support vector machine (SVM) classification methods are utilized, and support vector machines-radial basis function (SVM-RBF) kernel correctly predicted all athletes from the high-performance bocce player (HPBP) cluster and 75% of the athletes in the low-performance bocce player (LPBP) cluster. Using machine learning to predict athletic performance from balance data was found to be a time-saving approach for selecting high-potential bocce athletes.

4. **H. ZhaoriGetu (2022) "Prediction of Sports Performance", Hindawi, Volume 2022.**

The objective of this study is to analyze sports achievements from multiple angles and factors. Specifically, the research focuses on solving the gradient disappearance problem in deep learning DNN models. The study compares gradient compression algorithms across three models, and they conclude that ProbComm-LPAC model has higher average accuracy, lower average loss rate, and the best performance.

This literature review shows a lack of research papers focused on predicting body performance outside of sports. Although sports performance prediction has been studied, there is a clear need for more exploration in other areas. Difficulties in establishing consistent measurement standards and acquiring diverse datasets may be contributing to this imbalance. However, we can overcome this gap by expanding research to different domains and promoting collaborations between different fields. By doing so, we can unlock the potential of predictive models in various areas, such as healthcare, injury prevention, and overall well-being, leading to improved outcomes for individuals.

6.**Theory and Methodology**

In this study it should mention that python is the only programming language that is used for all preprocessing, EDA and Advance analysis parts. So, in different places used different functions which in built in python and some functions are defined for some situations.

In this study some techniques and methods were used for preprocessing part and those are explained in Data section in this report and cleaned data set were used for further analysis parts.

After preprocessing part, Scatterplots and heatmap are used for exploratory analysis to visualize the relationships between variables. Scatterplots help in identifying any linear or non-linear relationships between two quantitative variables, while heatmaps provide a graphical representation of the relationships between variables using color-coding. For this plotting part "sns.scatterplot()" and "sns.heatmap()" function with their other characteristics in seaborn package used in this study.

Through scatter plots and Heatmap I identified that there were higher relationships between "Height" & "Weight" variables and between "Diastolic' & "systolic" variables which we known in practical world as BMI Body Mass Index) value and MAP (Mean Arterial Pressure). So, I created BMI value variable for represent "Height" & "Weight" variables and Created MAP value for represent "Diastolic' & "systolic" variables by define functions as below,

$$\text{BMI} = \frac{(\text{weight in kilograms})}{\text{height in meters}^2}$$

Figure 1

$$\text{MAP} = \frac{2 * \text{DBP} + \text{SBP}}{3}$$

Figure 2

Then in the latter part of Exploratory analysis, Boxplots are employed by using sns.boxplot() function in seaborn package to explore the relationships between all quantitative variables such as performance and performance class variable. Also, by including gender variability, it helps identify any differences in performance based on gender.

Then for the Advance analysis part, Encoding was applied to the gender variable and the performance class variable. Encoding categorical variables refers to the process of representing categorical data in a format that can be used by machine learning algorithms. Categorical variables are variables that take on a limited and fixed number of values or categories gender and performance class. So, for encoding part use "data.replace()" function.

Then for access the importance of independent variables in explaining the variance in the dependent variable, VIP values calculations were used. So, for that Variable importance in

$$\mathrm{VIP}_j = \sqrt{\frac{p \sum_{a=1}^{A} \left[SS(q_a t_a)(w_{ja} / \parallel w_a \parallel)^2 \right]}{\sum_{a=1}^{A} SS(q_a t_a)}}$$

Figure 3

projection (VIP) score estimates the importance of each variable in the projection used in a PLS model by “PLSRegression()” function and the VIP score for the *j*th variable, in PLS model with *A* principal components, can be calculated as follows,

Then after calculate and plot a graph for VIP values, Used “train_test_split()” function with thir characteristics for split the data set into training and testing parts. In here Training data set also divided into two parts as train_features and train_ labels and testing dataset also divided into two parts as test_features and test_labels.

Then after finishing the Exploratory analysis part, it was started to Advance analysis part.

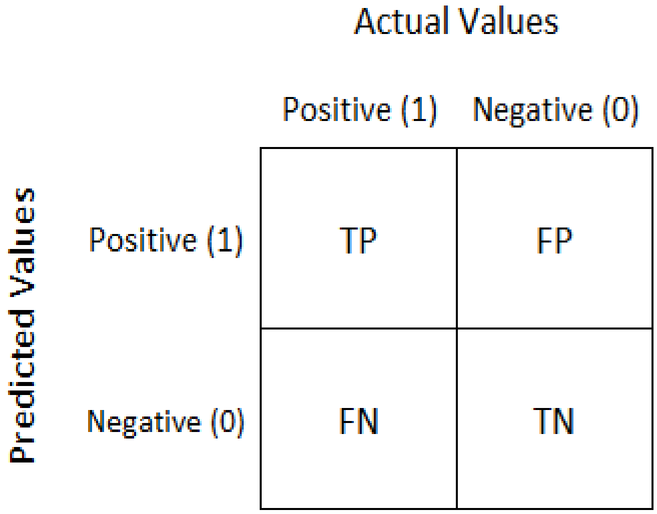
For the find good model with better accuracy, Different machine learning techniques are applied to build appropriate models with higher accuracy rate for training data set and test dataset as well. In here some regularization approaches, namely Ridge and Lasso as well as advanced statistical techniques such as random forest, gradient boosting and few others were utilized for the model building process because it was determined that multicollinearity is present in the data. So, build model below techniques and functions were used,

Random forest technique by “RandomForestClassifier()”
XGBoost technique by “XGBClassifier()”
SVC (Support Vector Classifier) techniques by “SVC (probability=True)”
LDA (Linear Discriminent Analysis) technique by “LDA ()”
Multiple logistic regression technique by “LogisticRegression(solver='liblinear')”
KNN (k- Nearest Neighbors) technique by “KNeighborsClassifier()”
Gaussian Bayes technique by “GaussianNB()”
Gradiant boost technique by “GradientBoostingClassifier(criterion="friedman_mse")”
Ridge regression technique by “LogisticRegression(solver='saga')”
Lasso regression technique by “LogisticRegression(solver='saga')”

But in the results set, non-regularization approaches were giving much better accuracy compared to regularization approaches and these algorithms offer a diverse range of modelling approaches to capture the relationships with the data.

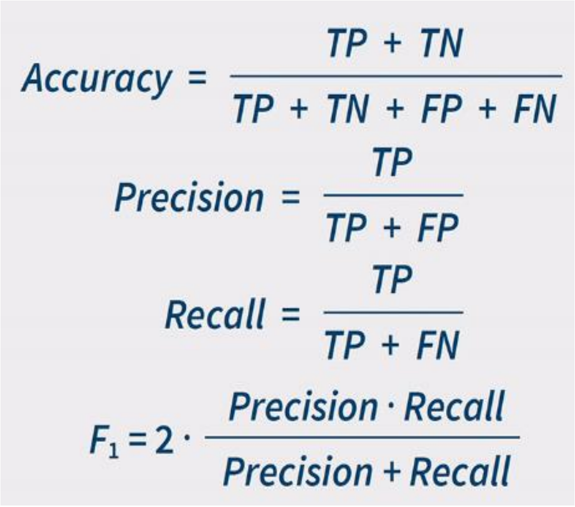
Because of the low accuracy of regularization approaches, non-regularization approaches are used for further analysis parts.

Accuracy, Precision, Recall and F1 values are calculated as Below by hand also,



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 5



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Figure 4

Because of the low accuracy level with low F1 values for testing dataset, among those techniques eliminate some from further comparisons.

After selecting three techniques which had higher accuracy values and F1 values, feature selection techniques such as “SelectKBest()” function used for find the feature scores and “RandmForestClassifier()” function used for calculate the feature importance.

Then according to the Feature scores and Feature importance, remove some features step by step and refit model by techniques which selected previous step with higher accuracy. Then compare all accuracy values among all steps and among three techniques, Find the best techniques with higher accuracy and higher F1 value with selected features.

7.**Data**

7.1 **Description About the Dataset**

The dataset “Body performance Data” is taken from Kaggle and consists of 13393 observations and 12 variables with two qualitative variables which “age” and “Performance class” and 10 quantitative variables. Response “Performance class” variable which is a categorical variable with 4 classes indicating “A” for best performance and “D” for worst.

All variable names, their descriptions and types can summaries as below,

Variable	Description	Type
Age	20 to 64 (years old)	Quantitative
Gender	F (Female) & M (Male)	Qualitative
body fat_%	Body fat in percentage	Quantitative
height_cm	Height in centimeters	Quantitative
weight_kg	Weight in kilograms	Quantitative
diastolic	Blood pressure, the bottom number	Quantitative
systolic	Blood pressure, the top number	Quantitative
gripForce	Person grip strength in kilogram	Quantitative
sit and bend forward_cm	Forward bend measured in centimeters. (flexibility)	Quantitative
sit-ups count	Sit up in 1 repetition	Quantitative
broad jump_cm	High jump measured in centimeter	Quantitative
class	Performance score A, B, C, D (A best) / (Target Variable)	Qualitative

Table 1

7.2 Data Pre-Processing

At the initial stage identified that there were no missing values in this dataset and had one duplicate value and it was removed before further analysis.
Then by using “plt.boxplot()” function in “matplotlib.pyplot” package, plot the box plots for all numerical variables in the dataset for identify the outliers and the distribution of them.

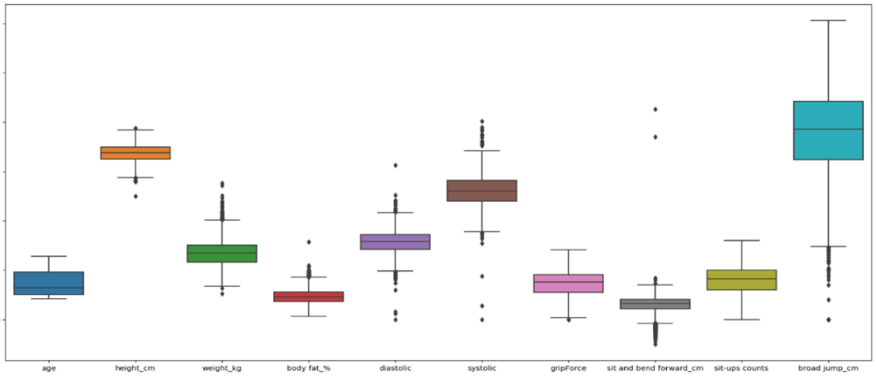


Figure 6

In figure 6 shows that there are many outliers in the dataset. Hence through ground research of all variables was used to identify some unrealistic data of the dataset and as mentioned before, to obtain meaningful results, two new variables created as “BMI val” and “MAP”.

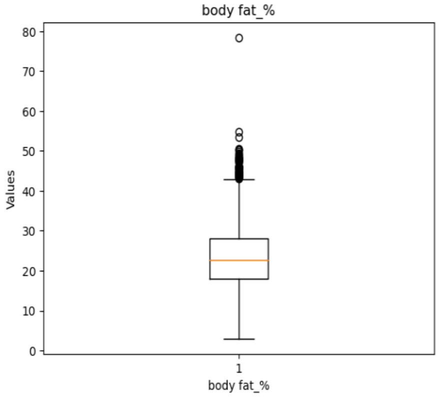


Figure 7

In practical terms, the highest body fat percentage that a person could realistically achieve while still maintaining essential bodily functions is around 60-70%, and even this level of body fat would be considered extremely unhealthy and life-threatening. So, it can be seen an outlier which takes the body fat percentage (78.4%) which is quite unreal. Therefore, this outlier was removed.

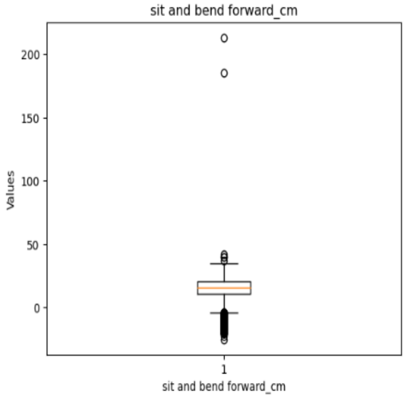


Figure 8

At a glimpse of the “sit and bend forward (cm)” boxplot there are 2 outliers which takes the values 185.0cm and 213.0 cm which are impossible in real world scenarios as per the height and weight values of this persons 164.4cm and 64.4kg and 165.6cm and 60.4kg respectively cannot take such test values. Since the persons’ height cannot be higher than the test values. So those two values were eliminated.

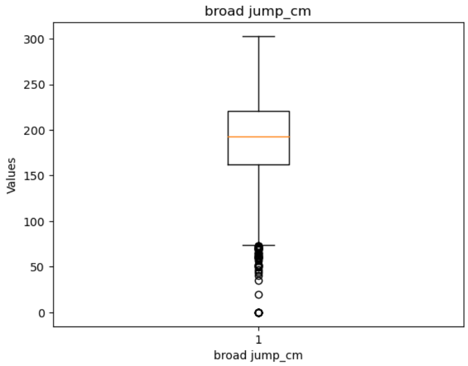


Figure 9

When investigated further it could be noticed that in the “broad jump (cm)” variable, there are 10 values of 0 which means none of distance have jumped which indicates outliers here. So, those 0 valued rows were eliminated. Because it is recorded the longest distance jumped, the best of three attempts where one can jump any of three times. There is one 0 MPA value and it was also removed and after eliminating possible outliers and creating new variables below figure 10 shows the boxplot of all quantitative variables.

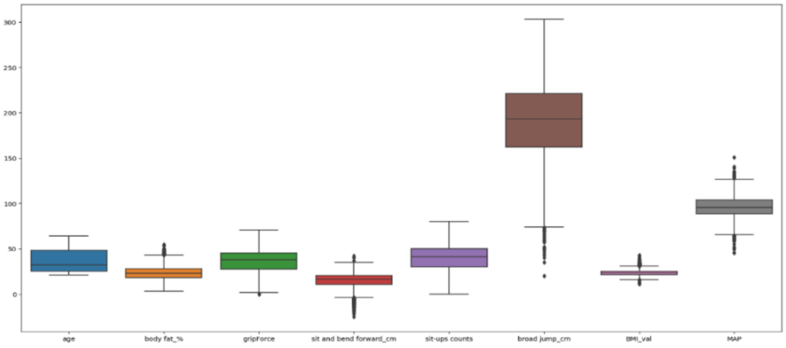


Figure 10

8.Exploratory Data Analysis

A proper descriptive analysis was carried out to identify the important association between body performance class which has four categories and other predictor variables.

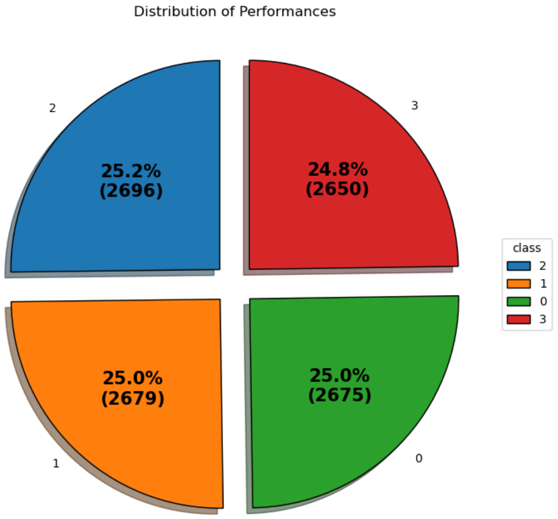


Figure 12

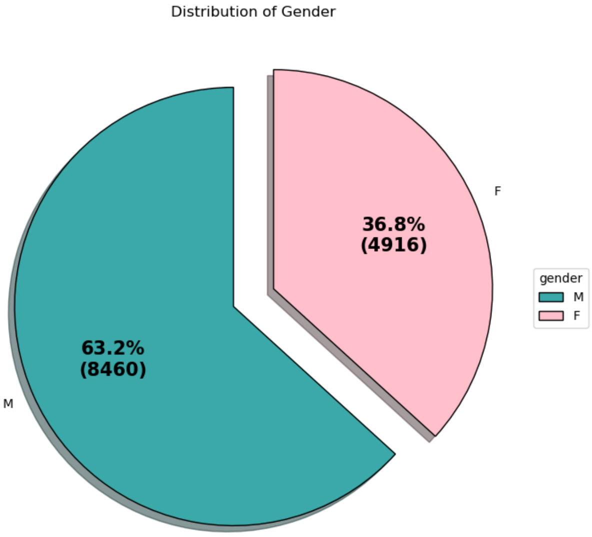


Figure 11

By looking above figure 12 we can see that There is no imbalance in the response variable's categories and figure 11 shows the male population is almost twice the female count in the dataset.

8.1. Body Fat VS Body performance

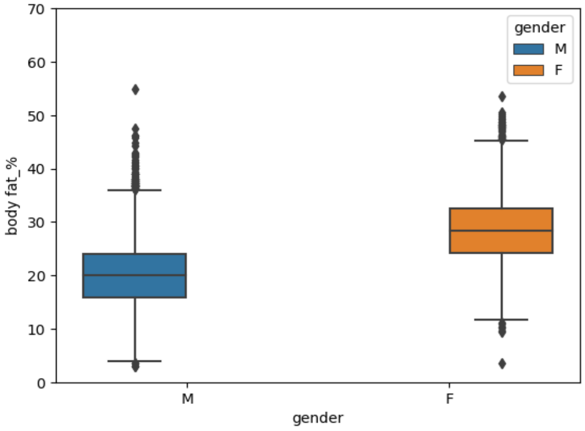


Figure 13

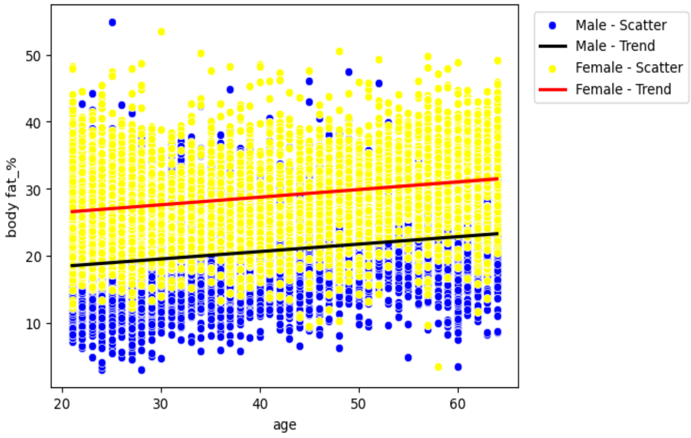


Figure 14

Figure 13 show that the female has higher body fat percentage than that of male and according to figure 14 we can see that when the age variable increasing in both male and female groups, the body fat percentage is also increasing, and the rates are approximately equal for both male and females.

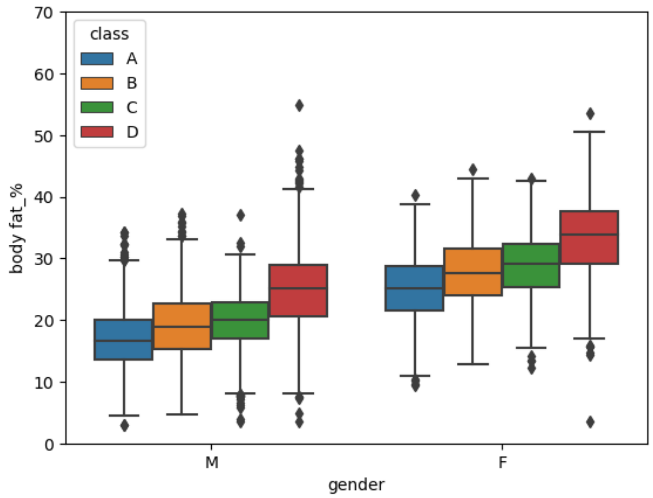


Figure 15

The figure 15 shows that in both male and female categories the body fat percentages gradually higher when the performance comes A to D. It's same as we know in practical world as the people with lesser number of fat percentages are having best performance of their body and when the fat percentage increases the performances are getting decrease. And we know that the fat percentage of females are higher than the male's fat percentages in the real situations also.

8.2. Sit and Bend FORWARD Vs Body Performance

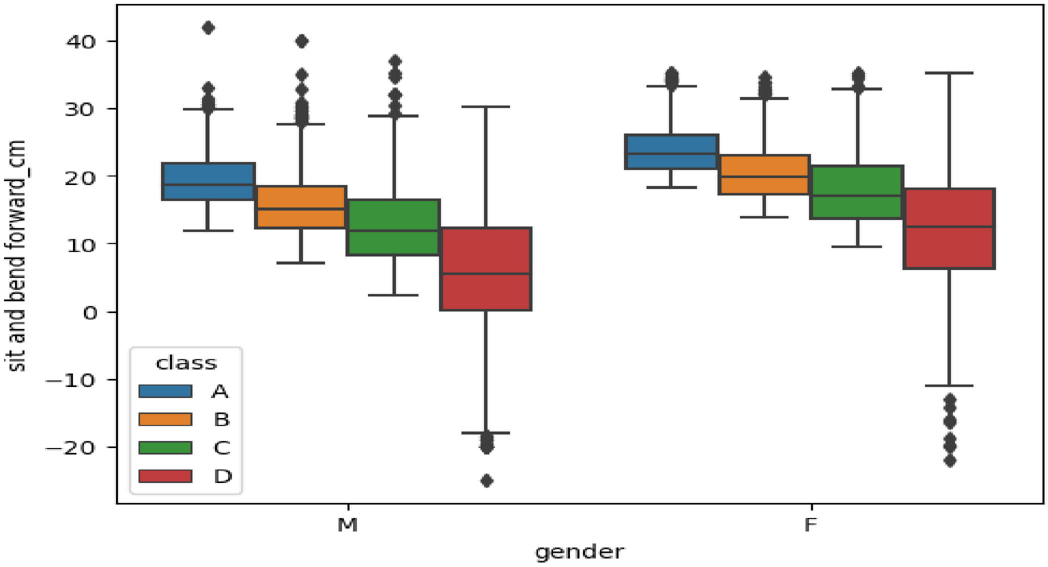


Figure 16

The figure 16 show that, Sit and bend forward values are less for who shows the worst body performance (D) on men and women and it gradually lower when the performance comes A to D. Also, this indicates that females obtain higher sit and bend forward values while giving the evident of flexibility of females. AS well as we can identify that the values are little bit higher in female category than the male category in all classes.

8.3. Broad Jump Vs Body Performance

Figure 17 show that broad jump value is less for persons who shows the worst body performance both men and women and it gradually lower when the performance comes A to D. As well in here males are showing higher distance than the females and when we considering class wise class D males are having approximately higher values than the all-female classes.

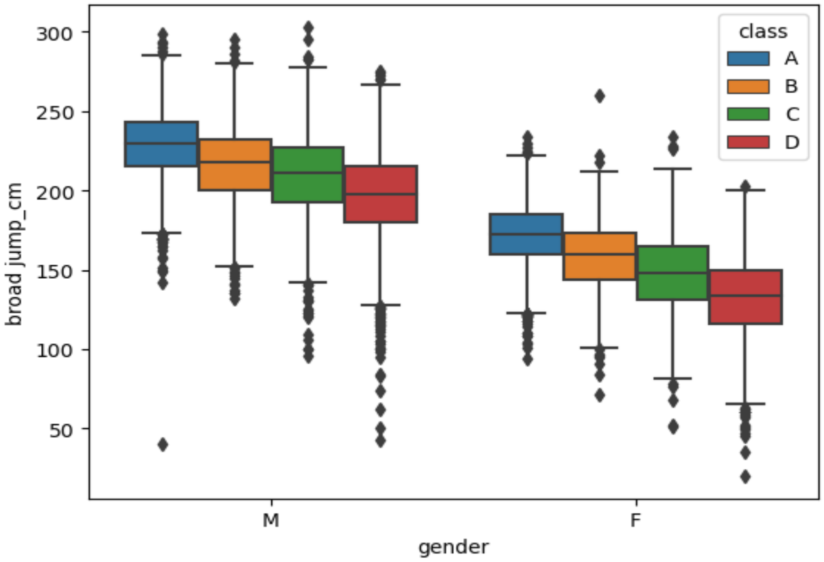


Figure 17

8.4. Grip Force Vs Body Performance

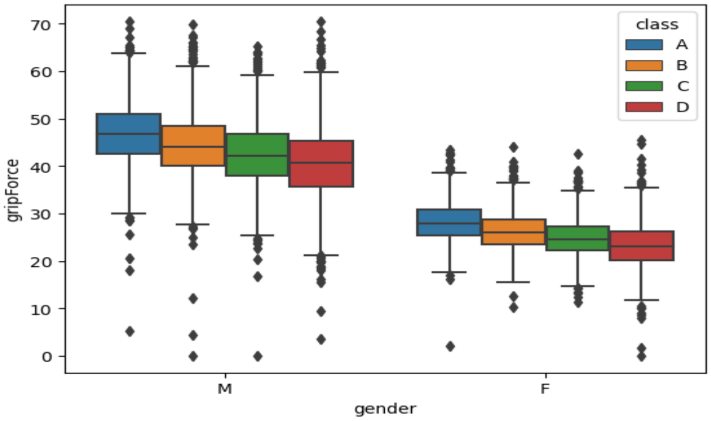


Figure 18

According to figure 18 we can see that the overall grip force is much higher in male category than the female category and it gradually lower when the performance comes A to D in both male and female categories. In here also we can see that the worst performance class males having much higher grip force than all female classes.

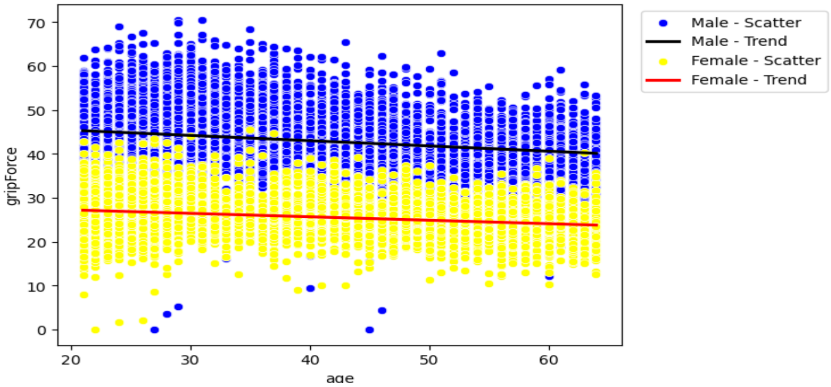


Figure 19

By this scatterplot showing in figure 19 show that in both male and female categories when age variable increasing the force variable is decreasing and the rates are approximately equal.

8.5. Sit Ups Vs Body Performance

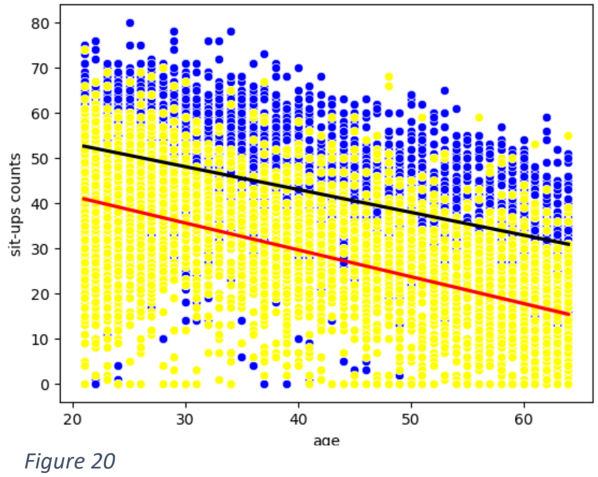


Figure 20

In figure 20 and Figure 21 results are also showing that when comparing gender variable females have lesser counts in all classes than the male variable and it gradually lower when the performance comes A to D in both male and female categories. As well as when the age increasing counts are decreasing in both male and female categories. But in here the decrease rate of male's are little bit lower than the rate of females.

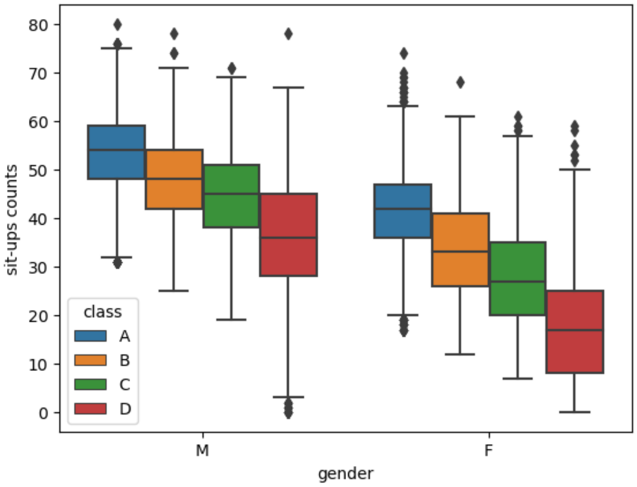


Figure 21

8.6. Correlations

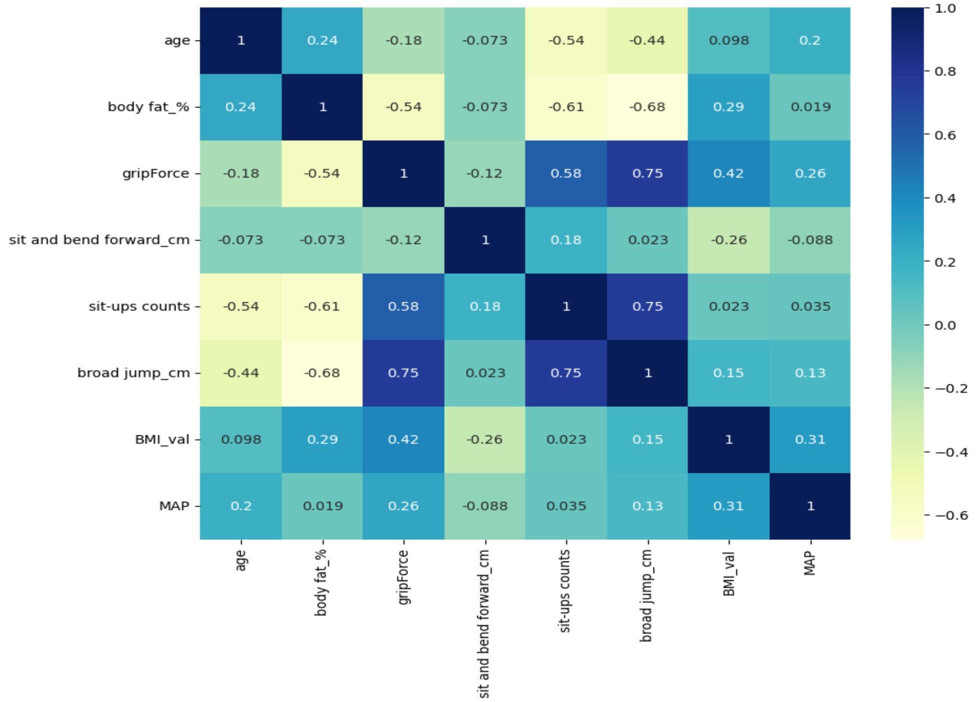


Figure 22

This heatmap in figure 22 shows the correlations between each pair of numerical independent variables. So, we can identify that the grip force and sit-ups to be the most correlated variables with the broad jump variable. Not only that, but there are also many considerably high correlations among predictor variables according to figure 8.12. As a result, it can be concluded that multicollinearity exists between the predictor variables. Due to the presence of multicollinearity regularization techniques will be used to reduce the variance of the coefficient estimates.

8.7. VIP Scores

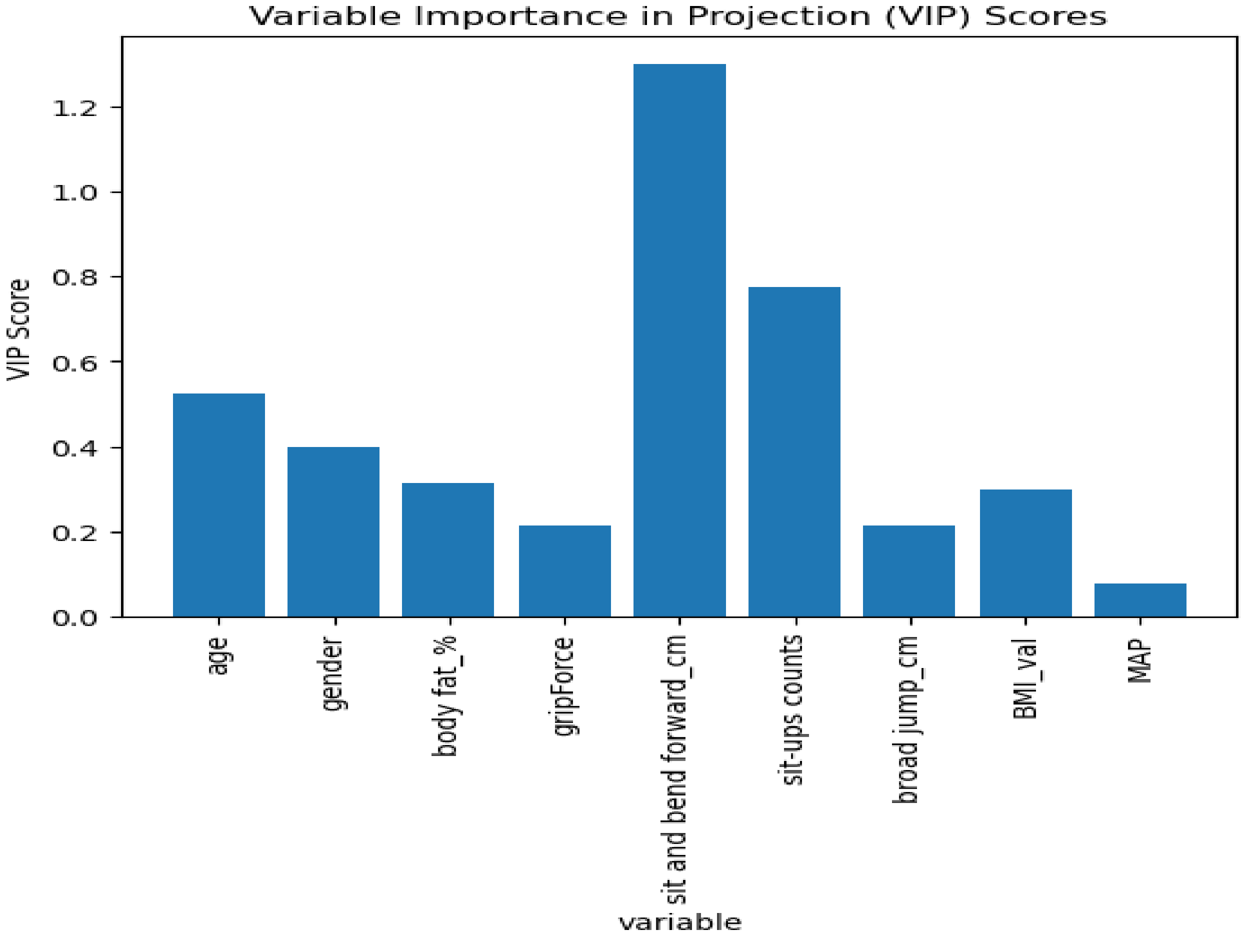


Figure 23

Figure 23 VIP plot is a bar chart that displays the VIP scores for each input variable. It can help identify which variables have the strongest relationship with the response variable and should be retained in the model. VIP plots can also be used to identify redundant or unimportant variables that can be removed from the model to simplify it and reduce overfitting. So, MAP shows the lowest importance according to the plot, a reduced model can be built up excluding that variable.

9.Advanced Analysis

At the initial part of the Advanced analysis, Different classification models were fitted to gain a rough idea on what path the analysis should proceed on, summaries of the different models fitted with default parameters are listed below.

Model	Training Accuracy	Training F1 score	Test Accuracy	Test F1 score
Ridge Logistic	0.578397	0.572799	0.598281	0.596551
Lasso Logistic	0.578303	0.572677	0.598281	0.596551
Multiple Logistic Regression	0.594468	0.572151	0.58296	0.560477
Linear Discriminant Analysis	0.61779	0.61473	0.615845	0.61779
K Nearest Neighbors (KNN)	0.715941	0.716759	0.605007	0.606458
Gaussian Naïve Bayes	0.552607	0.54373	0.573991	0.565867
Random Forest	**0.999907**	**0.999907**	**0.745516**	**0.746555**
Gradient Boosting	**0.773407**	**0.772859**	**0.737294**	**0.737097**
XG Boosting	**0.942628**	**0.942724**	**0.745889**	**0.747344**

Table 2

Hence above bold models in table__ give the higher accuracy and least overfitting, and other models show some overfitting: work better in training data but poorly work on unseen, new data and underfitting: poor performance on both the training and test sets. Those three models are only considering further steps.

Then by using feature selection methods, find the feature scores and feature importance as below.

Feature	Importance value	scores
Age	0.10525024535943626	242.7137362050354
Gender	0.01685769793193364	32.46377659170306
Body fat	0.1011648794191906	3293.8217418497593
Grip Force	0.09125120689069902	599.1452425381503
Sit and bend forward	0.277392387930755	99.17250463858431
Sit-ups counts	0.14749421561032455	11409.224019000758
Broad jump	0.0852492538576471	6023.960322339804
BMI_val	0.10879683329994816	405.35351740844754
MAP	0.06654327970006575	63.254829039436345

Table 3

So, by these values, scores, and results of EDA we can identify that the “MAP” variable has less importance for the dependent variable. So, we can use features excluding “MAP” for further steps and can refit the models according to three techniques which were selected before.

After removing “MAP” variable it gives accuracy values and F1 scores as below,

Model	Training Accuracy	Training F1 score	Test Accuracy	Test F1 score
Random Forest	1	1	0.735426	0.737241
XGBoost	0.930761	0.930902	0.755979	0.758576
Gradient Boost	0.772005	0.771693	0.7358	0.735985

Table 4

By comparing “Random Forest”,” XGBoost” and “Gradient Boost” methods finally, we can identify that the case of least overfitting, least underfitting and higher accuracy with higher F1 score, “XGBoost” is the most suitable model that gives a good fit for the given data and can make accurate predictions on new data.

10. General Discussion and Conclusion

- This Study conduct build a model for predict Body Performance class, base on physical characteristics & performance on physical test and identify the features that affect for the performance class by using Kaggle dataset.
- This study significantly expands the scope by addressing the broader prediction of overall body performance, going beyond the sole focus on sports performance.
- Built new variables such as "BMI value" and "MPA value" by considering relationships among Height & weight and Diastolic & systolic variables by considering their relationships in the data set and considering real world scenarios.
- Through Exploratory Data analysis it identified that there exist special relationships between "Gender" variable and other variables and the existence of multicollinearity of the dataset.
- Some regularization approaches namely Ridge and Lasso models as well as advanced statistical techniques such as random forest, gradient boosting and "XGBoost", were utilized for the model building process because it was determined that multicollinearity is present in the data.
- Analysis of the feature important values obtained for the final gradient boosting model concludes that sit and bend forward, sit -ups, age, etc. excluding MAP value play an important role in prediction of the class of body performance.
- Considering the higher accuracy, f1-score and lower error rate, and the risk of overfitting, "XGBoost" were selected as the best model that a good fit for the data.

11. References

- Hindawi (2022) "Prediction of Sports Performance and Analysis of Influencing Factors Based on Machine Learning and Big Data Statistics", Journal of Sensors (6): 1-9. https://www.researchgate.net/publication/363510822_Prediction_of_Sports_Performance_and_Analysis_of_Influencing_Factors_Based_on_Machine_Learning_and_Big_Data_Statistics

- Dr. J. Stebbins (2021) "Machine-learning models for activity class prediction", Journal of Gait & Posture volume 89. https://www.sciencedirect.com/science/article/pii/S0966636221002332

- Şimşek, Mehmeta; Kesilmiş, İnci (2022)"Predicting athletic performance from physiological parameters using machine learning", Journal of Sports Analytics, vol 8. https://content.iospress.com/articles/journal-of-sports-analytics/jsa200617

- H. ZhaoriGetu (2022) "Prediction of Sports Performance", Hindawi, Volume 2022. https://www.hindawi.com/journals/sp/2022/4082906/#abstract