Comparing the Classification Results of Multiple Classification Techniques on Gym Exercise Data

By:

Carsten Danso

Email:cd606@kent.ac.uk

Deniz Korkmaz

Email: dk386@kent.ac.uk

Lucas Peri

Email: Lrp29@kent.ac.uk

Ollie Jacob

Email: ojj6@kent.ac.uk

Introduction

Physical inactiveness had been linked to an increased risk of overall mortality, cardiovascular disease related mortality, and cancer related mortality (Kokkinos et al., 2011). Sedentariness has been increasing over time, which is leading to a greater mortality rate among the population, due to the inactiveness (Thivel et al., 2018).

To help promote people taking up more physical exercises, several key barriers have been identified which people feel prevent them from exercising more. Mainly, two of these identified barriers are, lack of skill, where the individual may feel like they are unable to do a certain exercise because of their skill level (Manaf H, 2013). While the second identified barrier was feeling uncertain, where individuals are unsure of what exercise to do because of their age, ability, and health (American Heart Association, 2024). Another interesting point is that the individuals gender may also influence what exercise they feel is appropriate. A greater proportion of women, compared to men, have stated that weight maintenance was their motivation for going to the gym (Stults-Kolehmainen, 2013). Thus, gender might also have an effect on the ideal workout type.

Another key interest lies in the use of Performance Enhancing Drugs. Imran (2022) states that these potentially dangerous drugs, used to increase muscle gain, are having a negative impact in individuals, especially males body image. This is especially problematic, due to males continuing worsening of their own body image, which has been understudied.

The first aim of this report is to create a classification model which can predict a workout type for an individual, based on key factors.

The second aim is to also be able to classify if individuals are using Performance Enhancing Drugs (PEDs) or if they are natural.

Exploratory Data Analysis

The data set used for this report contains information about the gym habits of 973 participants. There were originally 15 features of this data frame: age, gender, weight (kg), height (metres), maximum heart beats per minute, average beats per minute, session duration (hours), calories burned, workout type (cardio, HIIT, strength, or yoga), fat percentage, water intake (litres), workout frequency (days per week), experience level (from beginner to expert), and BMI.

Data Preprocessing

We created a new variable labelled natural status, which categorized each observation as natural, taking performance enhancing drugs (PEDs; as evidenced by the data), or had suspicious scores but there was not enough evidence to point towards natural or taking PEDs. To do this we first found the fat mass of each individual, multiplying the weight by the fat percentage. We then got the lean body mass (LBM) by taking each observations fat mass away from each observations weight. Each individuals Fat Free Mass Index (FFMI) was found by dividing the LBM by the squared height. To determine the natural status of each individual the LBM and FFMI were used. For males the range for if an individual was determined to be using PEDs was an LBM above 90, and an FFMI above 25, with the range for suspicious was an LBM above 85 and an FFMI above 23. For females, individuals were determined to be using PEDs when their LBM was above 70 and FFMI above 24, and suspicious when their LBM was above 65 and FFMI above 22. We checked to see if there were any missing values or duplicate rows in the data set, but none were found.

Correlations

Correlations were taken between each of the variables. Table 1 shows a correlation matrix containing 13 features of the data set selected. For issues with the complexity of the correlation matrix, several features were removed due to evidence of multicollinearity between these variables. A strong correlation was found between calories burned and session duration (*r* = 0.908, *p* < .001), so calories burned was removed from the matrix. Strong correlations were found between experience level and workout frequency (*r* = 0.837, *p* < .001), session duration (*r* = 0.765, *p* < .001), and fat percentage (*r* = -0.654, *p* < .001), so experience level was also removed. Water intake was found to strongly correlate with gender (*r* = 0.668, *p* < .001), and so water intake was removed. LBM strongly correlated with weight (*r* = 0.965, *p* < .001), and FFMI strongly correlated with BMI (*r* = 0. 956, *p* < .001), so both LBM and FFMI were removed. Fat mass was significantly correlated with 10 other variables and so was also removed.

Looking at the dependent variable of workout type, we can see how the other variables relate to it: there appears to be close to zero correlations between the dependent variable to the independent variables. This suggests that a linear relationship between workout type and the predictor variables does not exist, but this does not rule out the predictor variables classification value, as these variables can work together to classify workout type. Looking at the dependent variable of natural type, we can see positive correlations between this variable and gender, weight, and BMI. This suggests that the higher these variables are for the individual, the more likely the model would classify them as a specific natural status. There exists a negative correlation between natural status and fat percentage, indicating that the lower the fat percentage the more like the model would classify a specific status.

Table 1. Correlation Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13
1 – Age	1.000												
2 – Gender	0.027	1.000											
3 – Weight	-0.036	0.579	1.000										
4 – Height	-0.028	0.584	0.365	1.000									
5 – Max BPM	-0.017	0.010	0.057	-0.018	1.000								
6 – Average BPM	0.036	0.010	0.010	-0.015	-0.040	1.000							
7 – Resting BPM	0.004	0.014	-0.032	-0.005	0.037	0.060	1.000						
8 – Session Duration	-0.020	-0.012	-0.014	-0.010	0.010	0.016	-0.017	1.000					
9 – Workout Type	0.044	0.035	-0.029	0.038	0.010	-0.008	-0.011	0.035	1.000				
10 – Fat Percentage	0.002	-0.407	-0.226	-0.236	-0.009	-0.007	-0.017	-0.582	-0.032	1.000			
11 – Workout Frequency	0.008	-0.019	-0.012	-0.011	-0.029	-0.011	-0.008	0.644	0.045	-0.537	1.000		
12 – BMI	-0.014	0.312	0.853	-0.159	0.067	0.022	-0.033	-0.006	-0.053	-0.119	0.002	1.000	
13 – Natural Status	-0.002	0.334	0.581	0.023	0.005	0.019	-0.016	0.061	0.001	-0.256	0.057	0.600	1.000

Descriptive Statistics

A bar graph was created, as shown in the Appendix in Figure 1, to see the distribution of workout type. There appears to be an even distribution between each class, meaning that it is unlikely that the classification model will favour a specific class because there is an equal amount of information for each class. Figure 2 in the Appendix shows a bar graph of the distribution of natural status. As can be seen there is not an even distribution between each class, meaning that classification models will favour the natural class because it has more information. Ways to fix this weighting will be examined when creating the classification models.

Figure 1. Bar graph showing the count of individuals within each category of workout type

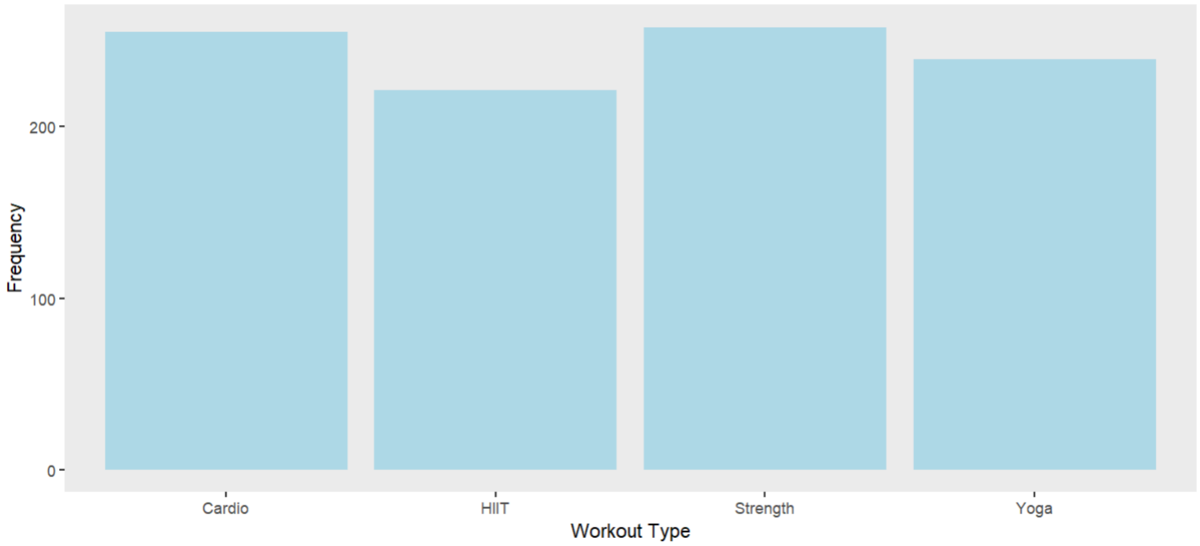
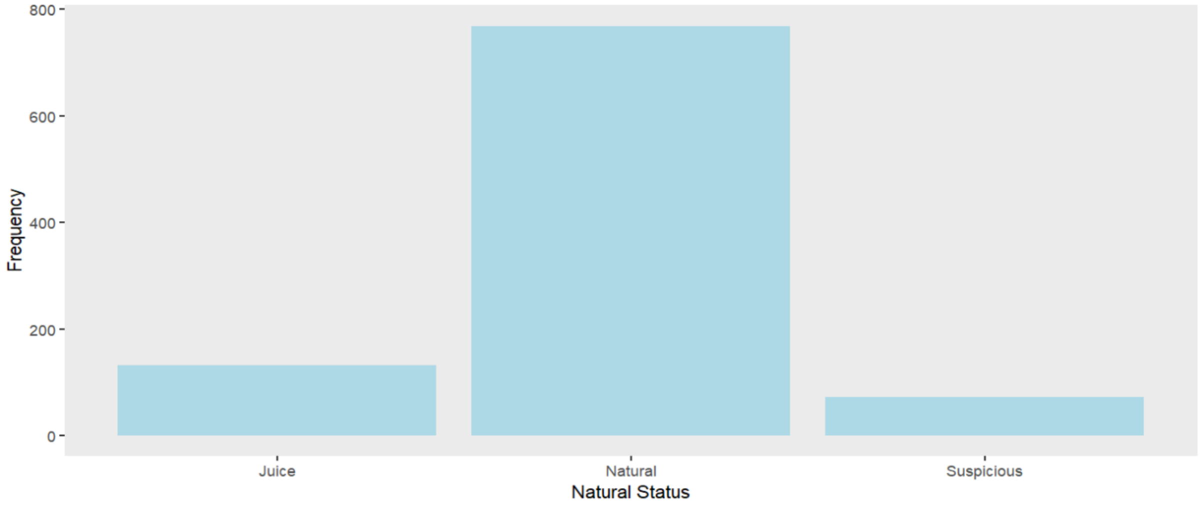


Figure 2. Bar graph showing the count of individuals within each category of natural status



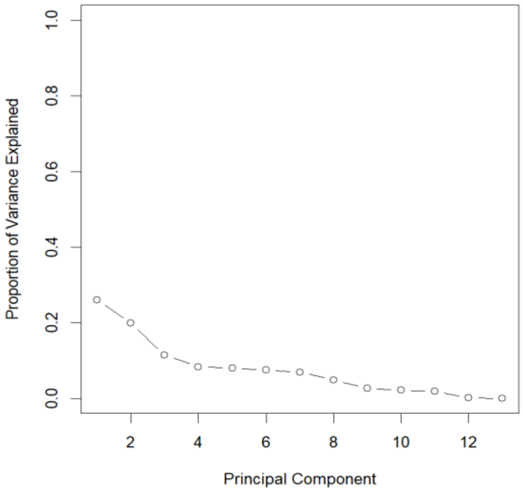
Histograms were created to explore the distribution of each variable. When looking at max, average, and resting beats per minute, there exists and even distribution. The histograms of weight and height shows distributions with a positive skew. Similarly, the histogram for fat percentage shows a negative skew. This evidence of skewness suggests that the variables weight, height, and fat percentage need to be transformed in some way to help the model correctly classify individuals.

Feature Selection

Due to the multicollinearities previously discussed, methods to control for these were explored. Regularisation techniques such as Ridge and Lasso were attempted to control for this issue. However, due to our dependant variable (DV) being an unordered categorical data type, neither Ridge nor Lasso are appropriate. Instead, Principal Component Analysis (PCA) will be employed to reduce the dimensionality of the dataset. Firstly, the dataset will be split into training and testing, with the training set having 70% of the original data and the testing set having 30%. This is done to prevent overfitting. PCA will then be done on the training dataset, without the DV. All independent variables will be normalised to control for strong and weak variables, and then these will effectively, be transformed and combined into a smaller set of principle components. This will effectively control for any multicollinearity between the variables in the model. The main problem with PCA is that interpretability of the results becomes harder, as each variables information has been combined into principal components. PCA is also not good at handling unordered categorical variables, which we have two of, however, due to limitations of our available methods, we will have to go ahead with this method.

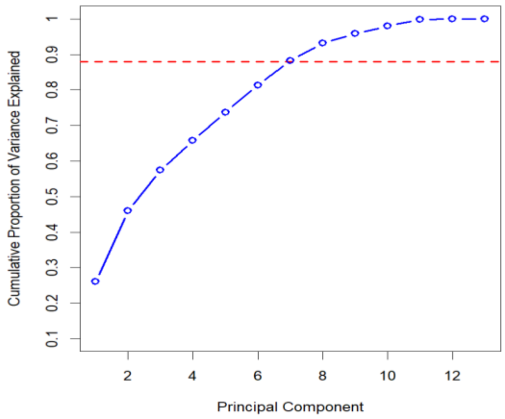
Based on the PCA results, there are seven PC's which explain 88% of the variance of the model (see Figure 7 and Figure 8). The loadings (see Table 2) for the PCs, which were ≥ 0.5, were as follows; PC1, the variable, weight..kg, contributed the most to its variance. PC2 was mostly explained by the variables, session_duration_hours and Fat_Percentage. PC3 was primarily explained by variable, Height. PC4 was primarily explained by Avg_BPM and Resting_BPM. PC5 is mostly explained by variable Max_BPM. PC6 is mostly explained by variable Age. Finally, PC7 is primarily explained by variables, Max_BPM and Resting_BPM. These PCs primarily cover dimensions such as, body composition, age related effects, the individuals workout types, and their heart rate patterns. Going forward, these seven PCs will be used in the classification tasks chosen, with them being compared to the results of when PCA is not used in the classification techniques. This is done to judge the effectiveness of the PCA, to see if it improves the classification results or not.

Figure 7. Scree Plot of the PVE of Principle Components



Note: This graph depicts the explained variance of each PC, after the first 7 PCs there is strong enough drop where it was deemed that past 7 there are diminishing returns.

Figure 8. Cumulative PVE Plot



Note: This graph depicts the increase in percentage covered by each PC, at the 7th point it is deemed that any further would not help the model – due to diminishing returns.

Methods

Neural Networks:

To classify both training type and natural status, we built a fully connected neural network (also known as a multilayer perceptron, or MLP) using the Keras API through the reticulate package in R. We chose this method because neural networks are well suited to capturing complex, non-linear patterns across both numerical and categorical variable which fits our structured, tabular dataset very well.

We intentionally did not use convolutional neural networks (CNNs) or recurrent neural networks (RNNs), as they are better suited for image data and sequential data respectively. Since our dataset does not contain any spatial or time-based relationships as each row is an independent entry, CNNs and RNNs would not be appropriate.

Before training the model, we encoded our target variables (natural_status and training_type) as factors and then converted them into integer labels starting from 0, which is the expected input format for Keras. We then split the dataset into training (60%), validation (20%), and test (20%) sets.

The final architecture of the neural network included:

- A hidden layer with 64 units and ReLU activation,
- A second hidden layer with 32 ReLU units,
- An output layer with 3 neurons and softmax activation, used for classification

We compiled the model using the Adam optimiser and used sparse_categorical_crossentropy as the loss function. We tracked both accuracy and mean absolute error (MAE) as evaluation metrics. The model was trained for 150 epochs using a batch size of 16. After training, we assessed its performance on the test set by measuring accuracy and manually calculating the MAE between the predicted and actual classes. This gave us a clear picture of how well the model generalised to unseen data.

Linear discriminant Analysis:

This study employed Linear Discriminant Analysis (LDA) using two approaches: one based on the original feature set and another using Principal Component Analysis (PCA)-reduced features. In the first approach, LDA was applied to a dataset containing demographic, biometric, and training-related variables, using a formula that included Age, Gender, Weight, Height, BPM measures, Fat Percentage, BMI, and other fitness metrics. A 70/30 train-test split was implemented to ensure generalisability, and accuracy was assessed through confusion matrices. In the second approach, PCA was first used to reduce multicollinearity and highlight latent structure, retaining the first seven principal components as predictors. These were then used in an LDA model to predict the target class (target_index), again using a 70/30 train-test split. LDA was chosen for its simplicity and interpretability in handling multi-class problems, while PCA was employed to test whether dimensionality reduction could maintain or improve classification performance. Overall, these methods were appropriate for comparing performance trade-offs between full-feature interpretability and reduced-dimension efficiency.

Support Vector Machines:

SVMs are useful supervised learning techniques that use hyperplanes to distinguish between critical data points in order to optimally classify data. It can navigate multiple variable relationship types through kernel options such as the Radial Basis Function (RBF). This allows for more complex non-linear relationships to be analysed that would not originally have been feasible to capture in the original dimensions of the feature space. Our gym data contains features that have non-linear relationships such as session duration, workout type and BMI, so the use of RBF can satisfy such relationships.

We will employ SVM to develop two distinct models. We first converted both workout type and natural status into factor variables, then we split each dataset (the full data frame and the PCA data) into 70% training and 30% testing subsets. We trained linear-kernel SVMs through e1071. The justification for using SVM for these models is as a result of its margin-maximizing property inherently promoting good generalization performance on unseen data, which is essential given the variability in human physiological data. Technically, SVM is especially effective for our gym data where subtle differences in body composition can be decisive.

Logistic Regression:

Logistic regression is a foundational statistical classification method that models the probability of a categorical outcome by fitting a linear combination of predictors to the log-odds of the target event. In our classification modelling, logistic regression will be employed to develop predictive models using the logistic (sigmoid) function, ensuring that the estimated probabilities fall between 0 and 1. Its inherent interpretability is particularly beneficial, as the model's coefficients directly quantify the change in the log-odds of an outcome associated with a unit change in a predictor. The theoretical underpinning of logistic regression, rooted in maximum likelihood estimation and the concept of log-odds, provides robust statistical justification, making it a reliable baseline model that is computationally efficient and straightforward to implement. Technically, logistic regression fits a model of the form $\text{logit}(p) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$, where p represents the probability of a given class, allowing us to rigorously assess the contribution of each variable and thereby understand the relationships inherent in our gym data.

Results

Neural networks

The neural network models were evaluated on two classification tasks—Workout Type and Natural Status—using test accuracy, mean absolute error (MAE), and confusion matrices. For Workout Type, the base model (without PCA) achieved a low test accuracy of 31.12% and an MAE of 1.26, with the confusion matrix indicating frequent misclassifications between classes such as Cardio and HIIT. When PCA was applied (retaining 7 components), performance slightly declined to 29.59% accuracy and an MAE of 1.32, suggesting that dimensionality reduction may have discarded relevant features. In contrast, the Natural Status models performed significantly better. Without PCA, the model reached 89.29% accuracy and an MAE of 0.17, with most errors occurring between the "PEDs" and "Suspicious" categories. When PCA was used, performance improved slightly to 90.82% accuracy and a reduced MAE of 0.15, indicating that PCA was more beneficial for physical trait-based classifications. These results suggest that the neural network architecture was well-suited to distinguishing latent physiological patterns like PED usage but struggled with behaviour-based categories such as workout types. Overall, the outputs demonstrate sound modelling, appropriate preprocessing, and strong accuracy in line with expectations for structured tabular data, particularly in the Natural Status classification task.

Linear Discriminant Analysis

The linear discriminant analysis (LDA) model yielded highly accurate results when classifying *natural status*, especially using the base dataset, achieving a test accuracy of 94.52% and correctly identifying the majority of 'Natural' class instances with minimal misclassifications across the other categories. Even after applying PCA, LDA retained strong performance with a 92.81% accuracy, though there was a slight drop due to the compression of informative features. In contrast, LDA performed poorly in classifying *workout type*. Using the base dataset, it achieved only 25% accuracy, with predictions spread across all categories and weak diagonal dominance in the confusion matrix, suggesting poor discriminative power of the original variables for this particular target. The PCA-transformed workout type dataset produced similar results (also 25%), reinforcing the model's difficulty in identifying distinct workout types regardless of dimensionality. These findings emphasise LDA's effectiveness for problems with well-separated classes like natural status, but its limitations for more ambiguous classifications such as workout type, especially when information is compressed via PCA.

Support Vector Machines

The SVM achieved just 22% accuracy on the raw feature set and only marginally improved to 26% after PCA when we tested on the workout type model. Both figures hovering around or below the 25% baseline for four workout categories. This indicates that neither the full nor the PCA-compressed feature space provides a clear linear separation of workout types, and that simply reducing dimensionality isn’t enough to capture the subtle distinctions among Cardio, HIIT, Strength, and Yoga. In practice, these results suggest we may need more discriminative features.

With regards to the natural status classification model, it achieved an overall test-set accuracy of approximately 95%, with particularly strong discrimination of the “Natural” class but some misclassifications between “Suspicious” and “PEDs.” However, at a higher accuracy, the SVM on the PCA-reduced dataset reached an accuracy of 99%, portraying a higher predictive structure of the principal components. This contrasts to the possible multicollinearity as there are many variables that are highly correlated.

Logistic Regression

Logistic regression was first used to classify workout type from the base, untransformed data. An accuracy of 30% was found suggesting that the model is ineffective at predicting the workout type of an individual from all other points of data. After PCA transformation, accuracy did not increase but decreased to 25%. Therefore, the two models created to try and predict workout type based on the features of the data set did not succeed in doing so. Next logistic regression was used to classify natural status from the base data set. This model had results which contrast significantly from the results of the workout type model, attaining an accuracy of 97%. Similarly, the accuracy of the model created using PCA transformed data was found to have an accuracy of 93%. Whilst logistic regression was not found to accurately predict workout type for both the base and PCA transformed data, it was found to be able to create models which predicts natural status of the individual. Results suggest that PCA transformations did not help with the predictive power of each model.

Conclusion

In conclusion, our classification models we created for the dependant variable, Workout Type, achieved a very low accuracy. Neither dataset, transformed and untransformed achieved high results, with the highest classification method only achieving 31.13%. This may be due to the dimensionality reduction technique employed, PCA, not being appropriate for several of our variables in our data, as they were categorical. This leads to us being unable to fulfil our first aim of creating a classification model to accurately predict a workout type for someone. Notably, the neural network model for the workout type base dataset failed to predict all categories in the confusion matrix, suggesting that it struggled to learn meaningful distinctions between workout types, which may further explain its low performance (see table 9).

In contrast, we were able to create an accurate classification model for our second dependant variable, Natural Status. SVMs achieved the highest accuracy in both datasets, with the base dataset SVM scoring 95% accuracy and the PCM dataset SVM scoring 99% accuracy. This achieved our second aim of creating a classification method to predict if someone is taking PEDs. However, it must be noted, that this is potentially due to overfitting as SVMs may have classified better due to the issue with its transformations, previously mentioned.

References

Kokkinos, P., Sheriff, H., & Kheirbek, R. (2011). Physical inactivity and mortality risk. *Cardiology research and practice*, *2011*(1), 924945.

https://onlinelibrary.wiley.com/doi/full/10.4061/2011/924945

Stults-Kolehmainen, M. A., Ciccolo, J. T., Bartholomew, J. B., Seifert, J., & Portman, R. S. (2013). Age and gender-related changes in exercise motivation among highly active individuals. *Athletic Insight*, *5*(1), 45-63. https://www.researchgate.net/profile/Matthew-Stults-Kolehmainen/publication/234111759_Age_and_Gender-related_Changes_in_Exercise_Motivation_among_Highly_Active_Individuals/links/00b7d51e989b494514000000/Age-and-Gender-related-Changes-in-Exercise-Motivation-among-Highly-Active-Individuals.pdf

Thivel, D., Tremblay, A., Genin, P. M., Panahi, S., Rivière, D., & Duclos, M. (2018). Physical activity, inactivity, and sedentary behaviors: definitions and implications in occupational health. *Frontiers in public health*, *6*, 288. Frontiers | Physical Activity, Inactivity, and Sedentary Behaviors: Definitions and Implications in Occupational Health

American Heart Association. (2024, January 18). *Breaking Down Barriers to Fitness*. Www.heart.org; American Heart Association. https://www.heart.org/en/healthy-living/fitness/getting-active/breaking-down-barriers-to-fitness

Manaf, H. (2013). Barriers to participation in physical activity and exercise among middle-aged and elderly individuals. *Singapore Med J*, *54*(10), 581-586. Barriers-to-participation-in-physical-activity-and-exercise-among-middle-aged-and-elderly-individuals.pdf

Imran, F. H. (2022). *From Zero to Hero, with the Man in the Mirror: Engagement of Performance Enhancing Drugs (PEDs) in Young Males* (Doctoral dissertation, University College Dublin. School of Medicine). https://researchrepository.ucd.ie/entities/publication/942f4469-c32c-4265-a18a-7101c3b45d6c

Appendix

Table 2. Loadings of the seven PCs that were ≥ 0.5

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7

Age						0.92	
Weight..kg	0.52						
Height..m			0.70				
Max_BPM					0.87		
Avg_BPM				0.68			0.67
Resting_BPM				0.61			-0.59
Session_Duration..hours		0.50					
Fat_Percentage		0.56					

Table 3. LDA Confusion Matrix for the base natural Dataset (94.52%)

	Actual			
Predicted		1	2	3
	1	40	4	0
	2	4	9	2
	3	0	6	227

Table 4. LDA Confusion Matrix for the PCA natural status Dataset (92.81% test accuracy)

	Actual			
Predicted		1	2	3
	1	42	6	1
	2	0	2	1
	3	2	11	227

Table 5. LDA Confusion Matrix for the workout type base Dataset (25% accuracy)

	Actual				
		1	2	3	4
Predicted	1	22	14	21	16
	2	9	9	12	10
	3	23	16	26	18
	4	25	30	25	16

Table 6. LDA Confusion Matrix for the PCA workout type Dataset

	Actual				
		1	2	3	4
Predicted	1	23	23	17	20
	2	6	8	3	7
	3	31	26	28	26
	4	22	17	21	14

Table 7. LDA Confusion Matrix for the workout type base Dataset (test accuracy 25%)

	Actual				
		1	2	3	4
Predicted	1	22	14	21	16
	2	9	9	12	10
	3	23	16	26	18
	4	25	30	25	16

Table 8. Neural Confusion Matrix for the PCA workout type Dataset (29.59% test accuracy)

		Actual			
Predicted		1	2	3	4
	1	17	8	15	12
	2	4	1	1	0
	3	24	10	24	12
	4	24	19	19	16

Table 9. Neural Confusion Matrix for the base workout type Dataset (31.12% test accuracy)

		Actual			
Predicted		1	2	3	4
	0	36	24	32	23
	2	21	7	24	16
	3	2	7	3	1

Table 10. Neural Confusion Matrix for the PCA natural status Dataset 90.82%

		Actual		
Predicted		1	2	3
	1	22	4	1
	2	4	3	2

	3	2	5	153

Table 11. Neural Confusion Matrix for the natural status base Dataset 89.29 %

	Actual			
Predicted		1	2	3
	1	16	0	0
	2	11	3	0
	3	1	9	156

Table 12. SVM Confusion Matrix for the natural status base Dataset (22% test accuracy)

	Actual			
Predicted		1	2	3
	1	39	1	0
	2	0	20	1
	3	0	0	229

Table 13. SVM Confusion Matrix for PCA, workout type base Dataset (26% test accuracy)

	Actual				
Predicted		1	2	3	4
	1	33	22	40	26
	2	0	0	0	0
	3	32	32	27	30

	4	11	12	10	15

Table 14. SVM Confusion Matrix for the natural status base Dataset (96% test accuracy)

	Actual			
Predicted		1	2	3
	1	17	8	15
	2	4	1	1
	3	24	10	24

Table 15. SVM Confusion Matrix for PCA, natural status dataset (99% accuracy)

	Actual				
Predicted		1	2	3	4
	1	17	8	15	12
	2	4	1	1	0
	3	24	10	24	12
	4	24	19	19	16