Midterm Project Executive Summary:

Classifying Athletes into their Respective Sports Group

Tej Rai

University of Wisconsin La-Crosse

DS 740: Data Mining & Machine Learning

Professor Jessica Kraker & Professor Marie-Claire Kossi

Tyreek Hill is considered a top 3 NFL wide receiver currently, but in 2023 he finished first in the USA Track and Field Masters Indoor Championship for the 60-meter dash. Moreover, in high school and college, Tyreek Hill excelled in both football and running. So, what sport was he truly meant to play? Athletes, coaches, and sports organizations agree that it is extremely hard to play different sports full-time and professionally due to the different physical demands, rigorous training, and time each sport requires. That is why predicting the classification of athletes into their respective sport groups based on physiological measurements can provide invaluable insights for coaches, sports scientists, athletic trainers, and even athletes themselves. Accurate predictions can help in tailoring training programs, optimizing performance, and preventing injuries. This analysis aims to leverage machine learning techniques to predict the sport group (ball, track, water/gym) of athletes using various physical measurements. By identifying key predictors and building robust models, we can enhance decision-making in sports science and athlete management.

Data Preparation

The dataset used in this analysis consisted of physiological measurements of athletes. Data cleaning, preparation, and exploration were critical steps to ensure the quality of the analysis. There were a few adjustments made and things of note regarding the dataset.

The dataset had no missing values, so NaN values were not of concern. Categorical variables like `Sex` and `Sport_group` needed to be handled appropriately in the analysis. Both categorical variables were treated as factors. Also, there was 11 numerical variables: `Bfat`, `Ht`, `Wt`, `LBM`, `RCC`, `WCC`, `Hc`, `Hg`, `Ferr`, `BMI`, `SSF`. However, there were some variables that exhibited skewness. Log transformations were used to effectively normalize the distributions of the skewed variables. Therefore, the dataset was now better suited for fitting machine learning models. Figure 1 contains the skewness values before and after log-transformation of the variables (Ferr, SSF, BMI, Bfat, and WCC).

	Skewness	
	Before Log Transformation	After Log Transformation
Ferr	1.2710864	-0.19288100
SSF	1.1659565	0.29288751
BMI	0.9394956	0.42382441
Bfat	0.7539148	0.17208851
WCC	0.8288574	0.08211105

Figure 1: Skewness Value Table

After exploring the dataset and committing those changes, the athlete's data frame was finally comprised of Sex, Ht, Wt, LBM, RCC, Hc, Hg, Sport_group, Log_Ferr, Log_SSF, Log_BMI, Log_Bfat, and Log_WCC. Our next goal was to fit and train the models on this data set.

Fitting the Models

Two machined learning methods were employed in this analysis: K-Nearest Neighbors (KNN) and Random Forest (RF). However, both models required adjustments to their tuning parameters.

The primary tuning parameter for KNN is the number of neighbors ('k'). A range of 'k' values from 1 to 30 was explored using cross-validation to find the optimal value.

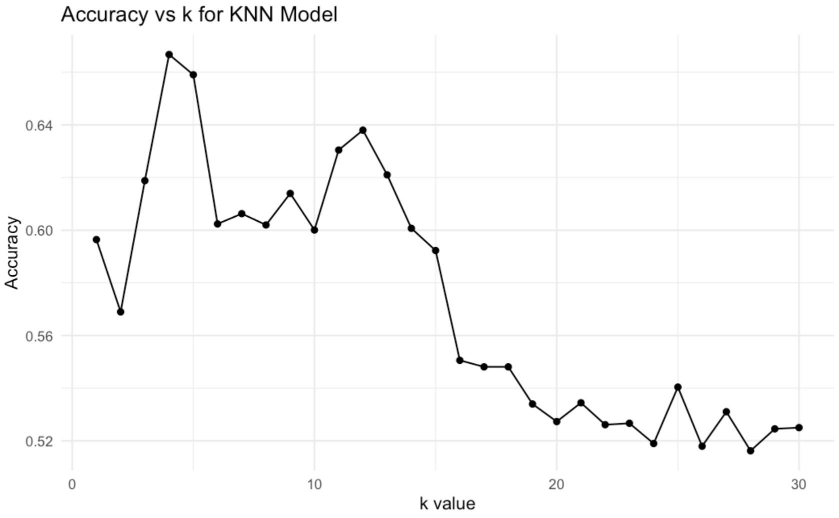

Figure 2: Accuracy vs k for KNN Model graph

As seen from Figue 2, the accuracy of model begins to decrease as ‘k’ increases and the trend begins around ‘k’ = 4/5. The best ‘k’ value found was 4.

The primary tuning parameter for RF is the number of variables randomly selected at each split ‘mtry’. Given the number of predictor variables (12), a range of ‘mtry’ values around the $\sqrt{12}$ (approximately 3.46) was considered, specifically 1, 2, 3, 4, 5 and 12.

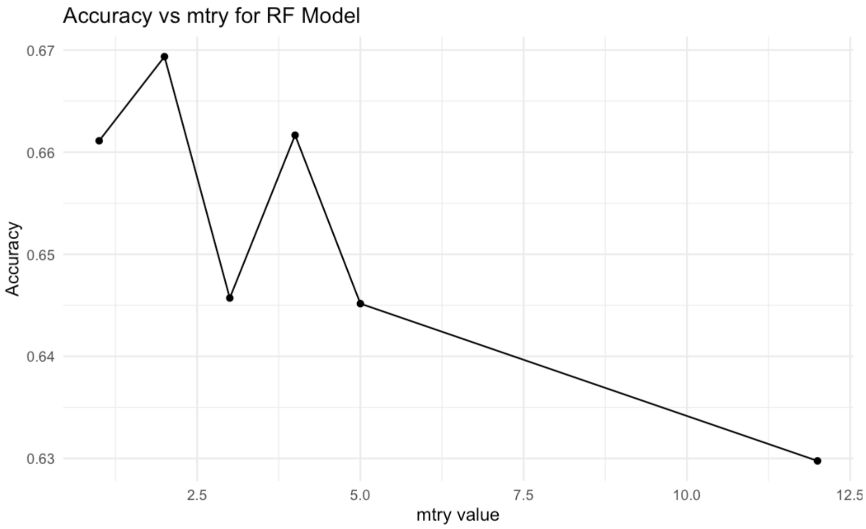

Figure 3: Accuracy vs mtry for RF Model graph

Figure 3 displays that the accuracy of the model begins to decrease as 'mtry' increases around 0.75 – 2.5. The best 'mtry' value found was. 2.

After tuning for best parameter value for each model, we created a for loop to conduct an outer-layer of 5-fold cross-validation containing both model types to decide the best model. The accuracy results for the folds were visualized below:

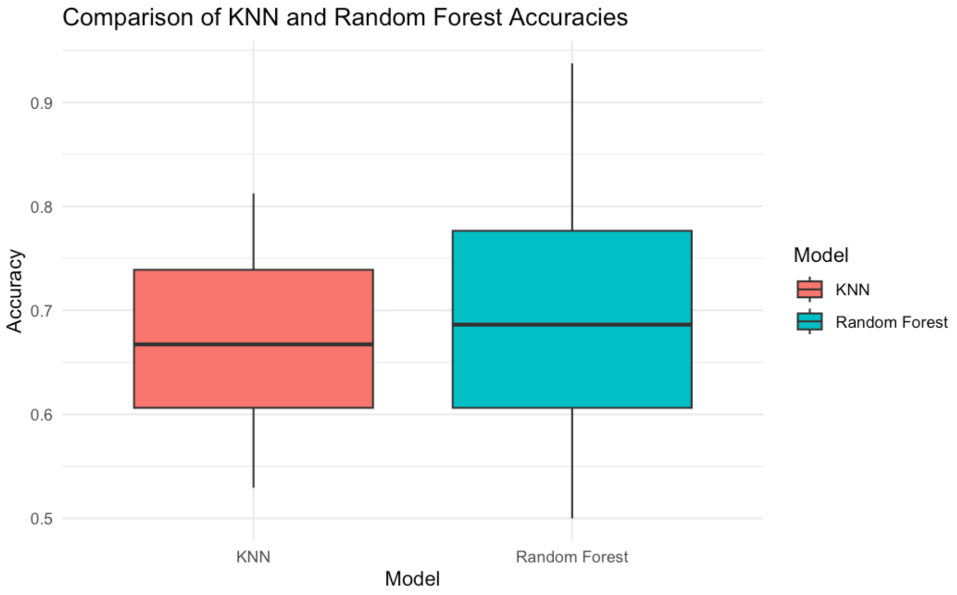

Figure 4: Comparison of KNN and RF Accuracies Graph

As seen from Figure 4, the RF model has a greater median accuracy, and the third quartile has more values and a greater max accuracy. Therefore, the RF model with 'mtry' = 2 was selected as the best model. The Random Forest model was trained using the entire dataset with the best 'mtry' value, assessed using cross-validation, and the final model's accuracy was found to be about 71.35%. Regarding the confusion matrix generated for the RF model, the kappa statistic (= 0.6173) suggests substantial agreement between the predicted and actual classes. Furthermore, the p-value for the accuracy being greater than the No Information Rate (0.4) is 7.463e-06 is very low, indicating that the model's accuracy is significantly better than what would be expected by random chance. Regarding the sensitivity, the model correctly identified 66.76% of athletes who belong to the “ball” group, 83.33% who belong to the “track” group, and 75% who belong to the “water/gym” group. Regarding the specificity, the model correctly identified 89.29% who don’t belong to the “ball” sports group, 96.43% who don’t belong to the “track” group, and 75% who don’t belong to the “water/gym” group. Figure 5 sums up the confusion matrix values for each class:

	Ball	Track	Water/Gym
Sensitivity	66.67%	83.33%	75.00%
Specificity	89.29%	96.43%	75.00%
Positive Predictive Value (Precision)	72.73%	90.91%	66.67%
Negative Predictive Value	86.21%	93.10%	81.82%

Figure 5: Confusion Matrix Results by Sport Group

Overall, the random forest model showed a good level of performance, with a high accuracy and substantial agreement between predicted and actual sport groups. The class-specific metrics indicate that the model performs particularly well for the “track” class, but less so for the “ball” and “water/gym” classes. Methods to improve that performance will be discussed later.

Interpreting the Best Model

The final RF model identified the most important predictors for classifying athletes into their sport groups. The variable importance plot (Figure 6) highlights the top predictors:

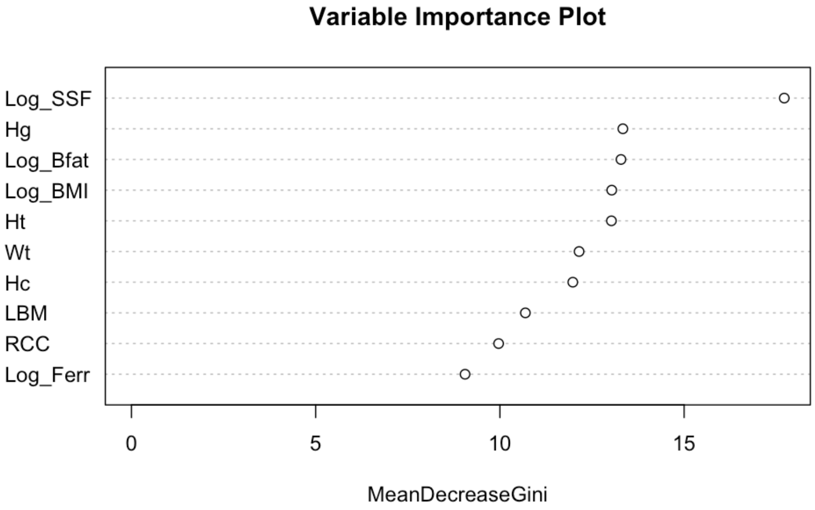

Figure 6: Variable Importance Plot

Log_SSF (sum of skin folds) had the highest importance, indicating that it was a crucial factor in distinguishing between different sport groups. Hg (hemoglobin) and Log_Bfat (percent body fat) were extremely close in importance and came second as crucial indicators.

Sum of skin fold (a proxy for body fat percentage) varies significantly among athletes in different sports. For instance, sports like swimming often require leaner body compositions for optimal performance, while sports like football may have athletes with higher body fat percentages. It does make sense why Log_SSF is an important variable for predicting which sport group an athlete belongs too. This would go for Log_Bfat (Percent Body Fat) as well, since the two variables are directly correlated. Athletes in endurance sports (like track) tend to have lower body fat percentages to enhance their efficiency and endurance. In contrast, athletes in sports requiring strength like football may have higher body fat percentages. Therefore, body is a key factor in classifying athletes into their respective sport groups.

Hemoglobin is crucial for oxygen transport in the blood. High hemoglobin levels are often found in endurance athletes because it enhances their aerobic capacity and endurance performance. Conversely, sports that rely more anaerobic performance might not exhibit such elevated levels of hemoglobin. Therefore, hemoglobin levels help distinguish between athletes involved in endurance sports versus other types of sports.

These three variables—Log_SSF, Log_Bfat, and Hg—make sense as the most important predictors because they capture fundamental physiological differences that are influenced by the specific physical demands and training regimens of various sports.

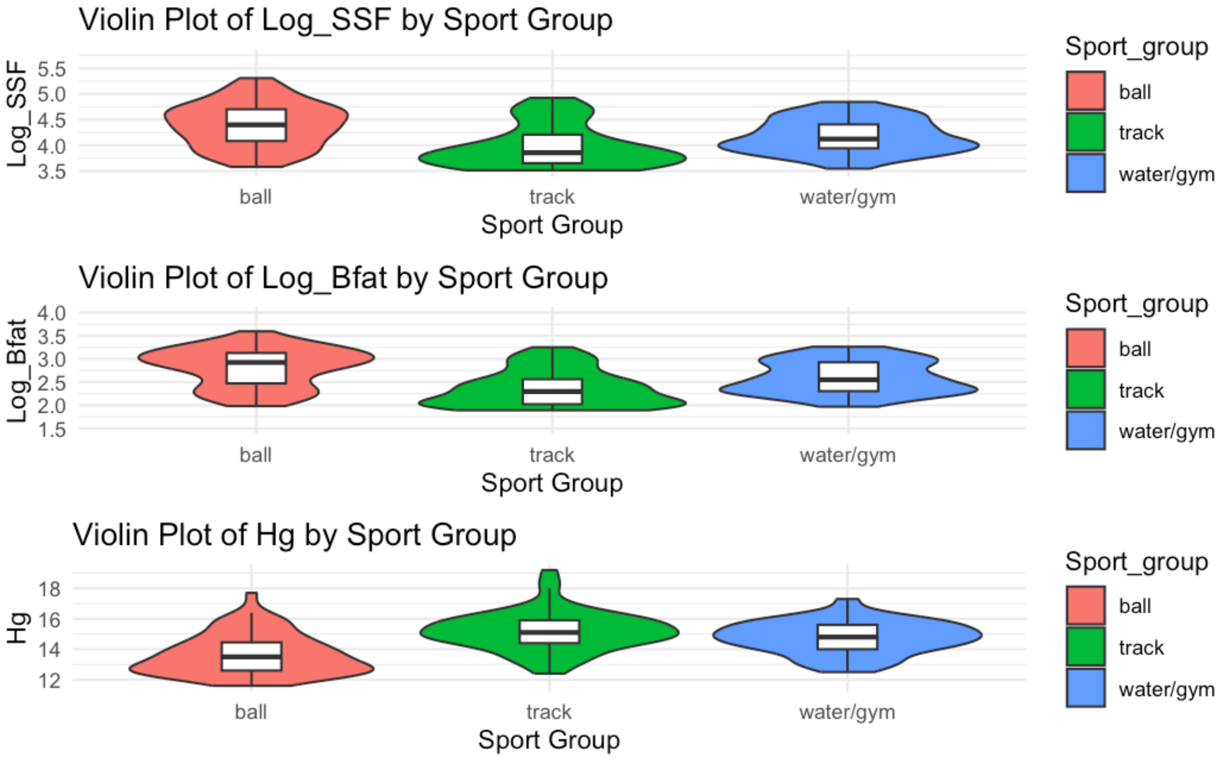

Figure 7: Violin Plots for the 3 Important Predictors

The violin plots in Figure 7 show the relationship between our 3 important predictor variables and the sports group response variable. The relationships do coincide to what we described earlier. For example, the median hemoglobin levels and quartiles for both track and water/gym groups (high aerobic-based sports) are greater than that of ball sports.

Final Thoughts: Model Accuracy and Future Improvements

The accuracy of the random forest model was evaluated using cross-validation, yield an accuracy of approximately 71.35%. While this accuracy is satisfactory, there is room for improvement. Particularly in classifying “ball” and “water/gym” groups, there are ways we could improve. For example, collecting additional predictor variables like VO2 max (maximum amount of oxygen that individual can utilize during intense exercise) or muscle mass could improve the model’s ability. Swimmers likely have higher VO2 max values than ball players, and ball players likely have greater muscle mass than the other two groups. In addition, the model should collect data from different populations and include an age, as people around the world typically have different body types and these values can also change with age.

The goal of this project was to develop a model to classify athletes into their respective sport groups based on physiological and body composition measurements. The Random Forest model was identified as the best performing model with a good accuracy. Enhancing data collection will be key to improving the model’s accuracy and reliability. The model can be refined to provide more precise and actionable insights for coaches, trainers, and athletes.