# DS740 Midterm: Predicting Groups of Australian Athletes

Adam Bruce

2024-03-02

## Backround and Data Preparation

Collected by the Australian Institute of Sports, the Athletes dataset contains 202 observations for individuals competing in either track, water/gym, or ball sports. The dataset contains 13 variables:

**Sex: 0 = male, 1 = female**

**Ht: Height in centimeters**

**Wt: Weight in Kilograms**

**LBM: Lean Body Mass**

**RCC: Red Blood Cell Count**

**WCC: White Blood Cell Count**

**Hc: Hematocrit Level**

**Hg: Hemoglobin Level**

**Ferr: Plasma Ferritin Level**

**BMI: Body Mass Index weight/(height)^2**

**SSF: Sum of Skin Folds**

**bfat: Body Fat Percentage**

**Sport_group: Track, water/gym, ball**

To make interpretation easier, Height was converted from European measurements in centimeters to American measurements in inches using **cm X 0.394**. Additionally, Weight was transformed from European to American scale using **weight kg X 2.21**. Finally, Sex was left as a numeric indicator variable as preferred in KNN Classification.

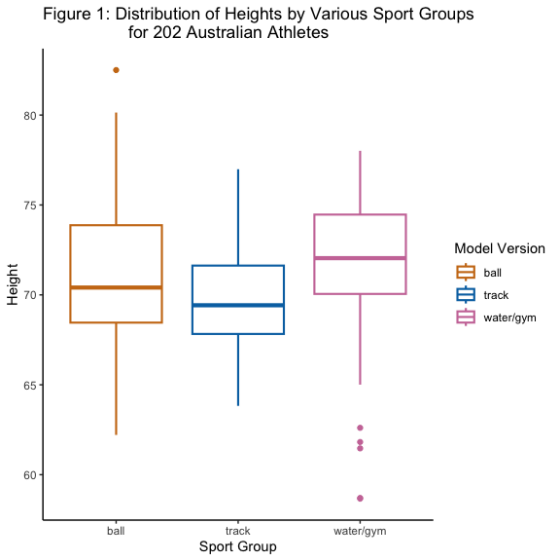
Overall, this analysis aimed to produce a best model for predicting the category of sport an athlete participates in using some or all of the predictor variables in the dataset. To accomplish this, two non-linear classification models in K-Nearest Neighbors (KNN) Classification and Random Forests were fit to the data using both single cross validation and double cross validation for tuning parameter estimation and honest model assessment respectively. For KNN classification an optimal tuning parameter, **K**, was used to decide on how many nearest neighbors to consider for class estimation. Meanwhile, for Random

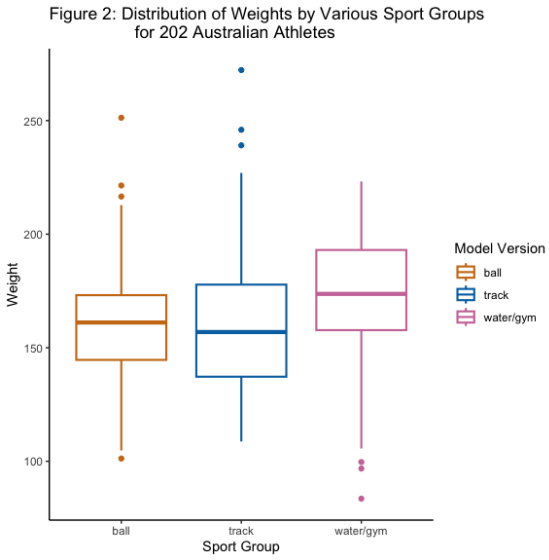
Forests, the optimal tuning parameter **mtry** was used to limit the amount of predictors considered at each node of the tree.

For KNN, K was chosen from a range of whole number integers between 1 and 146 nearest neighbors following the formula for consideration of: **Number of Observations - 1 in steps of 2**. However, the cutoff was set at 146 due to model fitting limitations at higher K's. For Random Forest, the optimal mtry value was considered from a range of whole number integers between 0 and 8 following the formula for consideration of: **Squareroot Number of Predictors = Sqrt(12) = 3.46**. This mtry value contains the expected optimal for classification around 4, but also accounts for a smaller and larger set of predictors to possibly limit or increase variance.

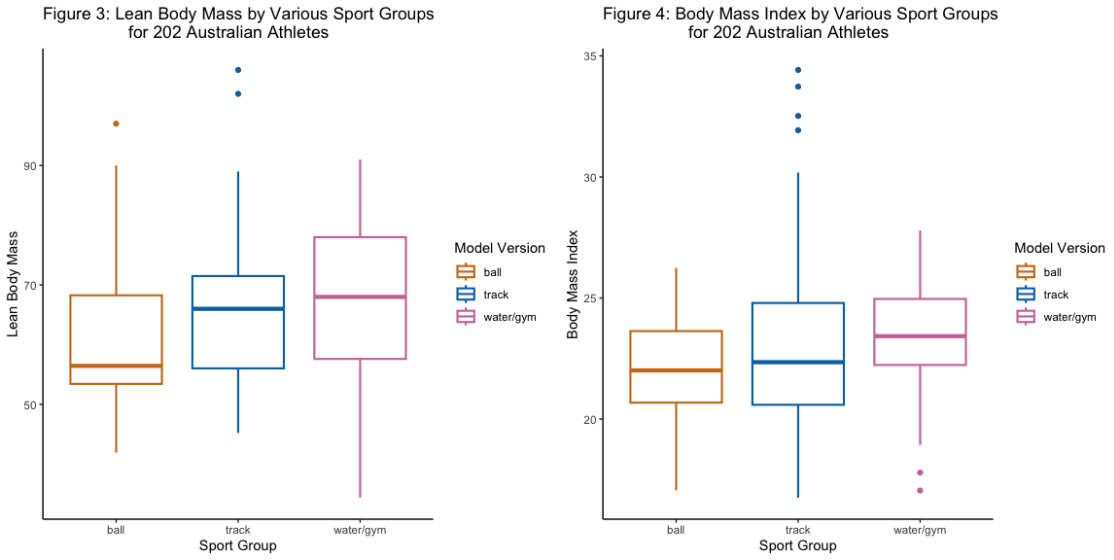
Prior to model fitting, the dataset was checked for missing values, but it contained none. Ultimately, the distributions of the response variables and their correlation was not checked. This was because the non-linear classification models used here handle non-linear predictors, and multicolinearity does not affect their accuracy. However, valuable insights into the relationship between the response variable Sport_group and the numeric predictors were gained by analyzing boxplots. Additionally, a conditional probability barplot for the categorical variable Sex treated as a numeric indicator in our models was produced.

## Exploratory Analysis

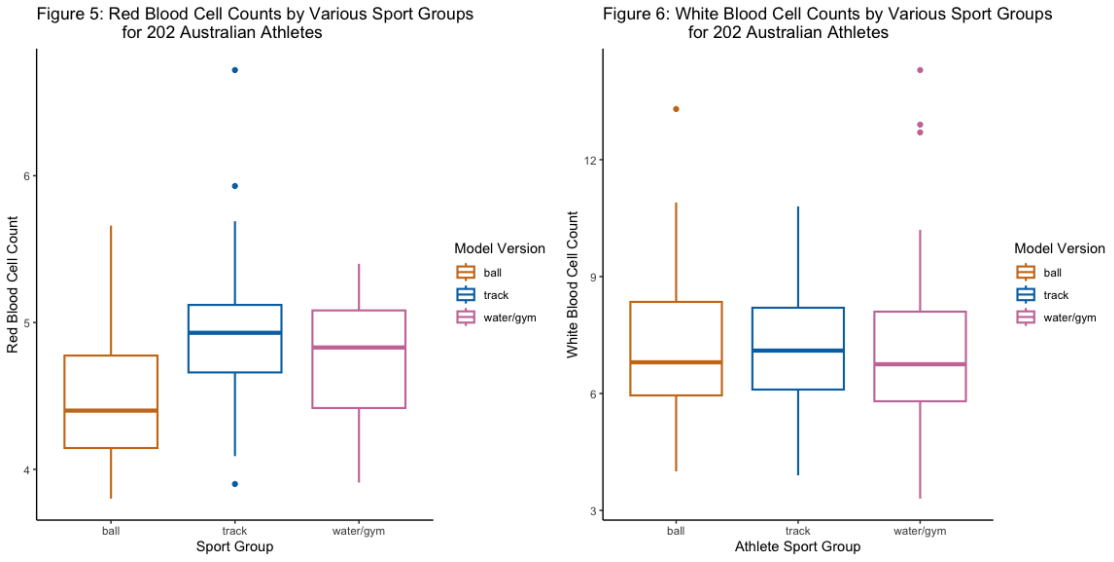



In Figure 1 and 2, we observe the values of height and weight for the three sport groups. For height, there is significant overlap in the range of the three boxplots. Thus, we would not expect height to be a significant predictor in our final model. For weight, the distribution overlap is even tighter than with height. However, water/gym does occur more frequently at higher values. Therefore, we might expect a cutoff of around 175 and above to predict a water/gym athlete in or final model. Overall, the distributions for all three groups are rather similar for both variables, so we would not suspect them to be extremely important in our model.




In Figure 3, we observe that Lean Body Mass has many quite large values in the water/gym group and many low values in the ball group. It appears that LBM greater than 70 is most likely to indicate water/gym while LBM less than about 55 most often indicates ball. For Figure 4, the BMI indices of the athletes are essentially even throughout the three groups. These observations follow what we would expect, as gym athletes are typically more focused on bulking muscle mass than track and ball playing athletes. Overall, we might expect our model to show more importance to Lean Body Mass than BMI, but neither appears to be extremely indicative of our response.




On the left, Figure 5 shows one of the better differences between groups among our predictors and response thus far. Clearly, we see that ball athletes have quite lower Red Blood Cell counts compared to both track and water/gym groups. However, we also see that track athletes have the largest values of RBC's, though there is a good amount of overlap with them and water/gym. Certainly, this is not surprising, as sports like track and swimming prioritize cardio based exercise, which builds significantly better oxygen

transport throughout the body. Such transport requires an abundance of Red Blood Cells, along with hemoglobin and hematocrit, to act as carriers/transporters for the oxygen molecules.

In Figure 6, we observe the tightest grouping of the data thus far. Clearly, White Blood Cell Counts are very similar among all athletes, and this should be expected given they function solely for immune activities. Unless an athlete was sick, most individuals should be within a standard range for WBC regardless of sport. Therefore, we might expect RBC to have a significant role in our model, but not WBC.

Figure 7: Hematocrit Levels by Various Sport Groups for 202 Australian Athletes

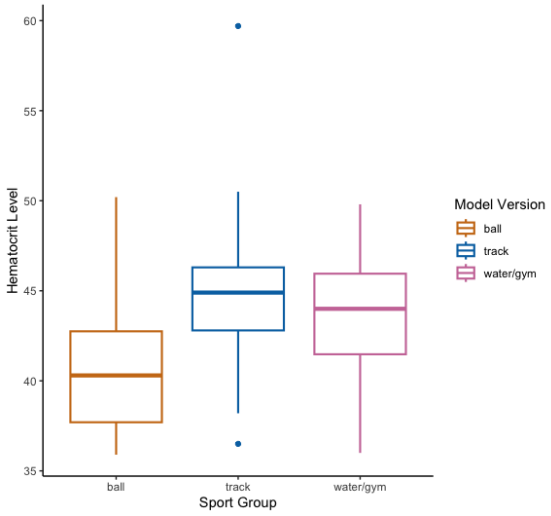
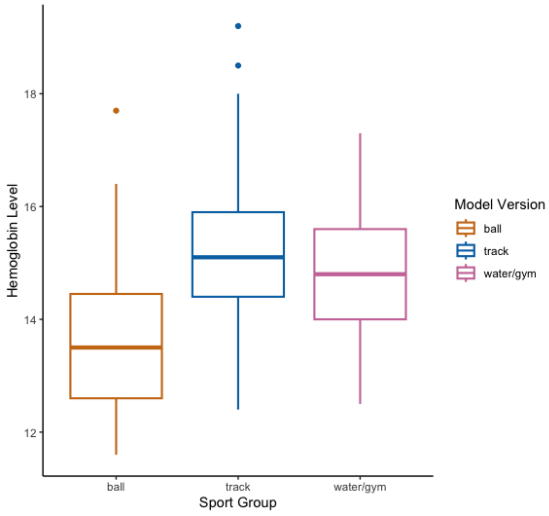
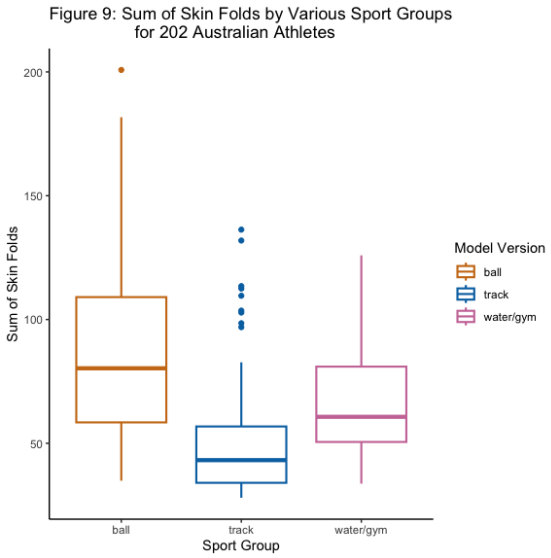



Figure 8: Hemoglobin Levels by Various Sport Groups for 202 Australian Athletes




As eluded to previously, track and water athletes are expected to be higher in hemoglobin and hematocrit levels as those are key factors for oxygen transport throughout the body. Therefore, the trends observed in Figure 7 and 8 come as no surprise. Both of these groups are likely to be predicted for values greater than 43 and 14.5 for hematocrit and hemoglobin respectively. Any value below these marks are extremely likely to be predicted in the ball group, and overall one or both variables should be rather important in our final models.



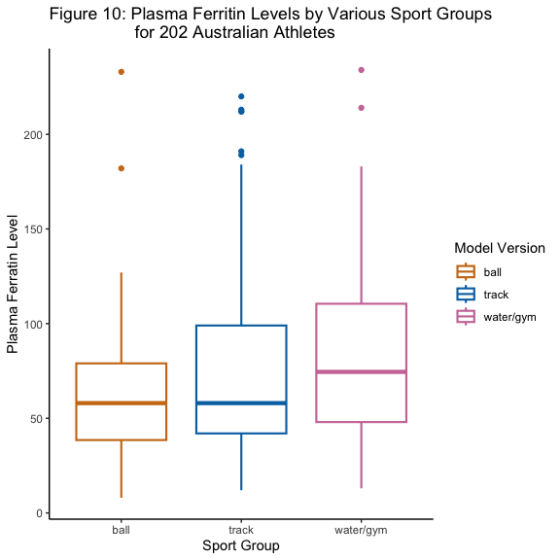
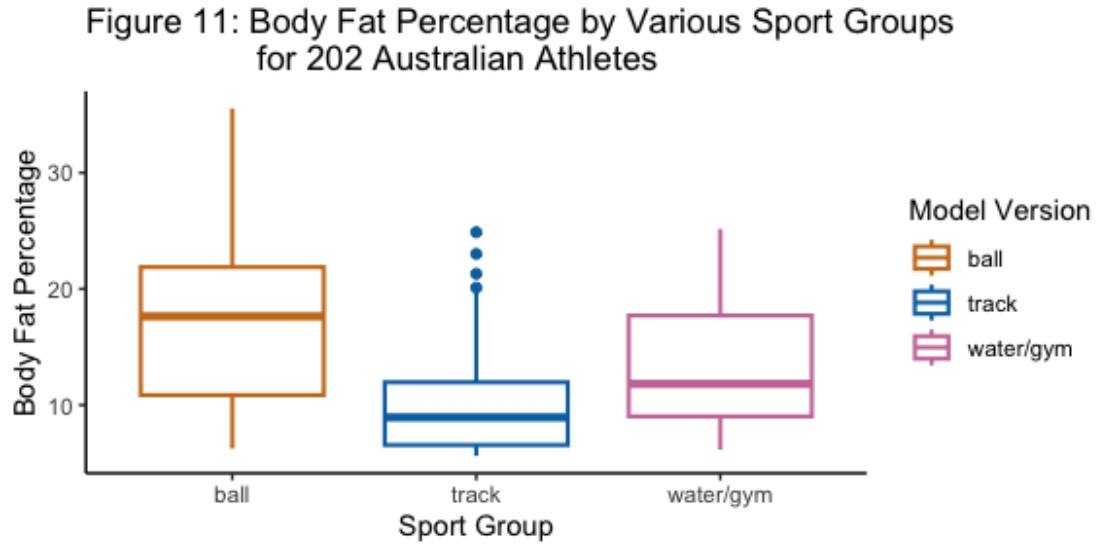



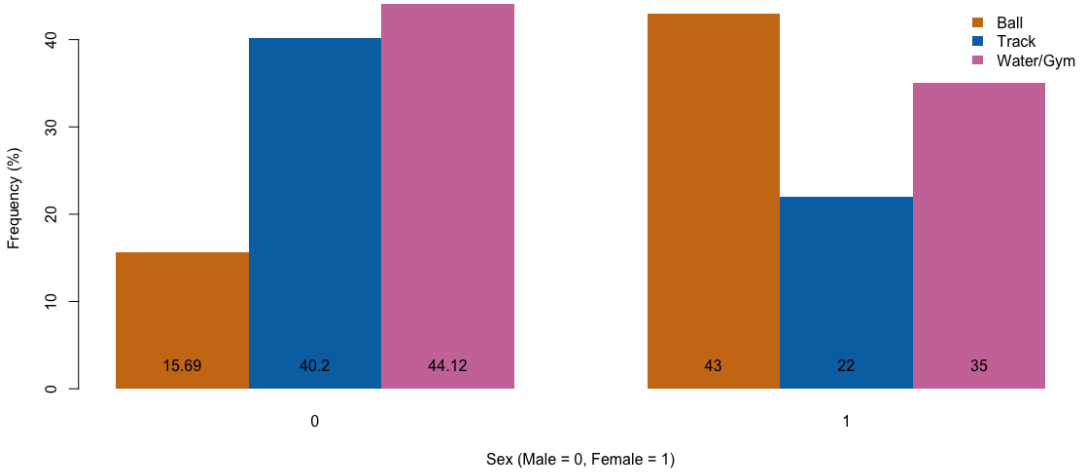

Figure 9 on the left shows track athletes are most likely to have a sum of skin fold below 50. Perhaps a cause for this trend would be that distance athletes are known for being extremely lean from the large amount of miles they run each week. On the other hand, we see that ball athletes are most likely when the sum of skin folds is above 75. This also makes sense as the range of fitness for sports like basketball, rugby, soccer, football, or baseball players can be very wide. In Figure 10, we see another indicator for blood oxygen level in Ferratin. This is an essential blood protein with iron, which is needed to produce hemoglobin for oxygen supply. Thus, we observe values lower than around 60 to most likely be predicted as ball athletes. Overall, Sum of Skin Folds appears as though it will have major importance to our final model. However, Plasma Ferratin Levels appear they might be somewhat important, but perhaps less so than Hematocrit, Hemoglobin, and Lean Body Mass.




In Figure 11, we observe very similar trends to what was found in Figure 9 for Sum of Skin Folds. In this case, track athletes appear to be much less likely to be predicted when Body

Fat Percentage is below around 15. In the middle, we might suspect water/gym athletes to be predicted, and then ball athletes are most likely when Body Fat Percentage is above about 18%. Overall, we should expect both Sum of Skin Folds and Body Fat Percentage to be important to our final prediction model given these findings.

**Figure 12: Frequency of Males and Females Across Three Sport Groups for 202 Australian Athletes**




Above, Figure 12 breaks down the percentage of Male and Female athletes in each of our three sport groups. Of the 102 male observations, most individuals are in the water/gym (44.12%) and track (40.20%) groups. For the 100 females, a majority are in the ball (43%) and water/gym (35%) categories. The large difference between males and females in the ball group will likely influence our model predictions. The same can be said about the track groups. Overall, it is likely our model will predict males in the track group more often than females and females in the ball group much more often than males. The water/gym groups are fairly close for both sexes, so we would not suspect a large influence of the sex variable on this group. Certainly though, it appears sex will be significant in making predictions of sport group.

Overall, based on this investigation, we might expect the best prediction model places heavy importance on the variables Sum of Skin Folds, Body Fat Percentage, Hemoglobin Blood Levels, and Hematocrit Blood Levels. Additionally, some importance will likely be placed on Red Blood Cell Count and Sex, though less so than the others.

## Model Building

### Single Cross Validation

*Table 1: Results of Single 10-Fold Cross Validation with K-Nearest Neighbors and Random Forest for the Athletes Dataset*

| Model | Parameter_Range | Optimal_Parameter | Maximum_Accuracy |
|---|---|---|---|
| K-Nearest Neighbord | K = (1:73)*2 | K = 6 | 0.6593 |
| Random Forest | mtry = 1:8 | mtry = 1 | 0.7122 |

Utilizing a single round of 10-Fold Cross Validation on the Athletes dataset, a K-Nearest Neighbors and Random Forest model that optimizes accuracy of predictions was obtained. Overall, Table 1 (Above) shows the range of parameters used in cross validation for each model at K = (1:73)*2 and mtry = (1:8) respectively.

Overall, it was observed that a Random Forest model with 1 parameter (mtry = 1) considered at each split was optimal during this cross validation process. This model obtained a maximum accuracy of 0.7122 (71.22%) when predicting Athletes Sport Groups where the best KNN model obtained a maximum accuracy of only 0.6593 (65.93%). Going forward, we will perform double cross validation with a 5-fold outer split and 10-fold inner split to honestly assess our model's prediction performance on new, unseen data using a train/test split. Then, we will assess our "best" model obtained here.

## Double Cross Validation

*Table 2: Results of Double 5-Fold Cross Validation with K-Nearest Neighbors and Random Forest for the Athletes Dataset*

| Outer_Loop | Optimal_Model | Optimal_Tune | Maximum_Accuracy |
|---|---|---|---|
| 1 | Random Forest | mtry = 5 | 0.6623 |
| 2 | Random Forest | mtry = 4 | 0.6909 |
| 3 | Random Forest | mtry = 6 | 0.7226 |
| 4 | Random Forest | mtry = 3 | 0.7228 |
| 5 | Random Forest | mtry = 1 | 0.6912 |

Figure 13: Athletes Confusion Matrix for Honest Predictions Assessment with 5-Fold Double CV

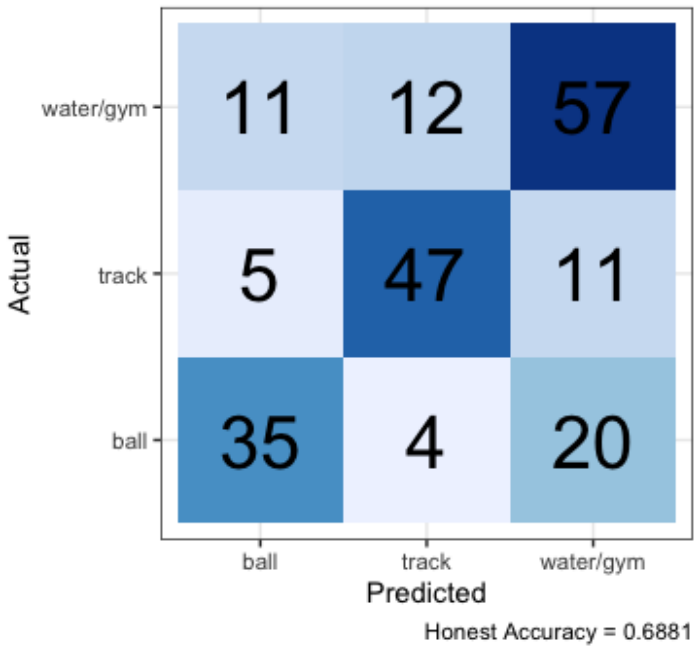


Table 2 (Above Top) provides the output for the optimal models at each of the 5 outer CV folds. Overall, we observed that Random Forests always optimized accuracy better than K-Nearest Neighbors, but the tuning parameter mtry varied between folds. For accuracy, we are most concerned with predictions on the test/validation set at each split. At each fold, we fit the optimal model to the test/validation set to obtain predictions for the observations in that set. These predictions were then saved before moving onto the next fold.

Utilizing the confusion matrix in Figure 12 (Above Bottom), we observe how the optimal models in Double CV performed in their predictions. However, most importantly, the honest accuracy can be calculated at 139 correct predictions out of 202 observations. This gives an honest accuracy = 0.6881 (68.81%), which indicates we can expect our "best" model from Single CV to predict the Sport Group of an Australian Athlete correctly on average about 68.81% of the time.

## Model Assessment

### Variable Relationships

**Figure 14: Variable Importance via Gini-Index Decrease for the Best Random Forest Model from 10-Fold Single CV**

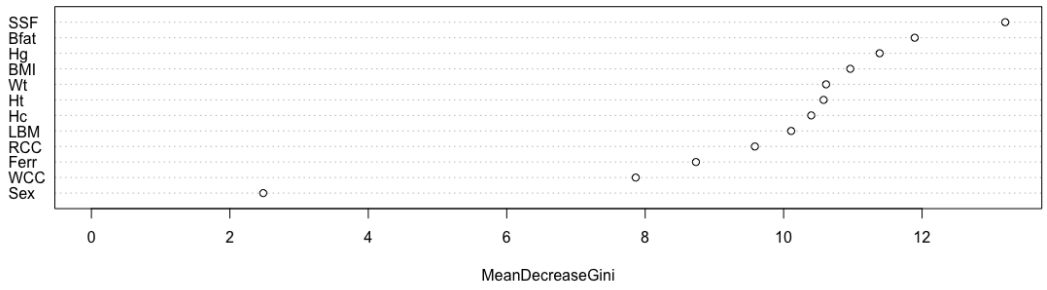
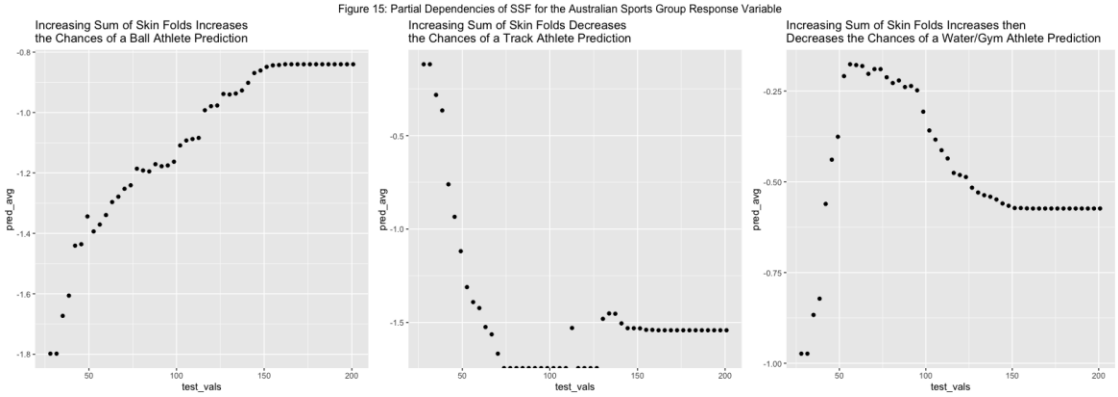



Figure 15: Partial Dependencies of SSF for the Australian Sports Group Response Variable



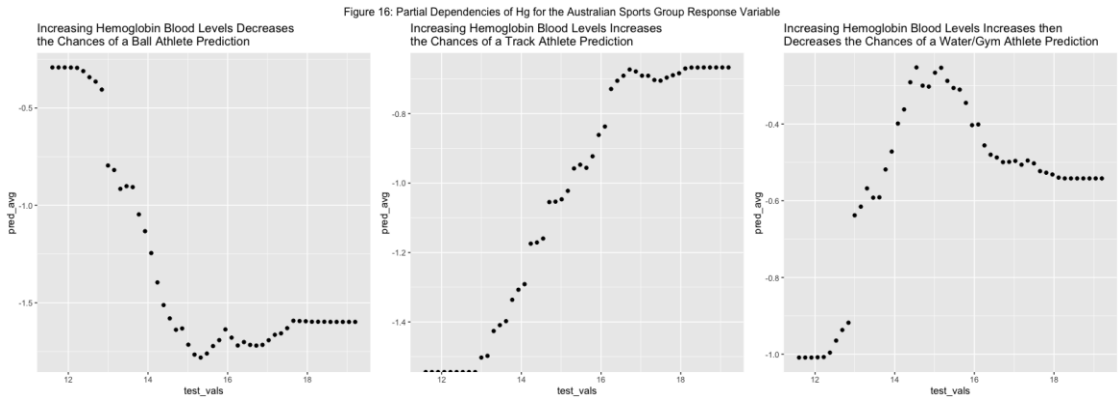



Figure 14 (Above Top) shows the variable importance to the final Random Forest model obtained through 10-Fold Single CV. Ultimately we observe, unsurprisingly, that Sum of Skin Folds (SSF), Body Fat Percentage (Bfat), and Blood Hemoglobin Levels (Hg) are the most important predictors for the model when classifying the Sport Group. This importance is calculated using the Decrease in Gini Index, which is a measure of how dense the observations of a certain category are for a given split. Thus, SSF, Bfat, and Hg splits obtain the most observations of a certain category of our response (either ball, track, or water/gym groups) at one node of the split.

To look at partial dependence between the important predictors and our response, we chose to use SSF and Hg instead of SSF and Bfat. This was because SSF and Bfat follow extremely similar distributions (See Exploratory Analysis) for each response category, and thus the partial dependence plots for these predictors likely follow very similar trends. Ultimately, their was no one single category of the response more important than the other, so to say the variables had a "positive" association with one single level of the response would not give the complete picture for our study. Therefore, a partial dependence plot was made for all three Sport Group categories based on SSF and Hg to get the full picture.

Figure 15 (Above Middle) shows that the probability of an observation being classified as a ball athlete is positively associated with SSF. Additionally, it shows a negative relationship for the probability of an observation being classified as a track athlete with increasing SSF. Finally, it shows that the probability of a water/gym athlete classification first increases with increasing SSF, up until about 60, and then decreases as the values of SSF get larger. These relationships are exactly in line with what we expected to observe based on our exploratory analysis, and they are an initial indication that our model was able to correctly identify important separator variables for our response.

In Figure 16 (Above Bottom), we see first that as Hg Blood Levels increase, the predicted probability of a ball athlete prediction decreases steadily. For track athletes, we see the opposite association as an inceasing Hg Blood Level yields an increase in the probability of a predicted track response. Finally, the water/gym group shows increase Hg Blood Levels yielding a higher probability of a water/gym classification up until about 15, where even higher Hg values begin to decrease the probability of a water/gym classification. Again, these findings are exactly in line with our observations in exploratory analysis, and they

provide more evidence that our model does an adequate job deciding which variables are important for predicting the Sport Group response.

Overall, we are not surprised by any of the associations observed in the partial dependence plots. Every pattern identified follows exactly what we expected to see based on our exploratory analysis. In addition, the model identified SSF, Bfat, and Hg as being very important predictors of Sport Group, which we highlighted it should do previously. We know that track athletes typically have high hemoglobin levels due to oxygen transport needs, and they tend to have low body fat if they are distance focused. Thus, their high probability of prediction at low SSF/Bfat and high Hg levels is reasonable. The same can be said with ball athletes, as we expect them to have a lesser need for oxygen transport and a wider range of fitness levels. Thus, we expected them to be classified most often by high SSF and low Hg while water/gym athletes would be somewhere in between.

## Conclusion

Overall, we observed an honest model accuracy of 68.81% based on assessment with 5-Fold Double CV. Using our confusion matrix, we see that the Sport Group water/gym has the highest frequency of observations in the Athletes dataset at 80. By taking this value divided by the total number of observations, 202, we receive a No-Information Rate of: 0.3960 (39.60%). Thus, if we were to randomly assign any new observation to be the majority class water/gym, we would predict the class correctly only 39.60% of the time. This means that, on average, our model will predict the Sport Group of an Australian Athlete 29.21% more often than randomly guessing. This provides great support for the final Random Forest model being sufficiently accurate to be used as a prediction model going forward. Though our model is sufficient, there may be other useful predictors of Sport Groups not accounted for in this assessment. One such variable would be lung capacity, which likely differs between the three response groups. Gathering additional variables such as this could be useful in producing an even better prediction model in the future.