# Body Composition Analysis of Athletes

Isuri Gamage

DS 740 – Data Mining and Machine Learning

## Executive Summary

Monitoring body composition in athletes is helpful for many reasons. One of the reasons is to create a dietary invention for strengths athletes. By analyzing body composition data of elite athletes, we can get an idea of optimal fat, blood levels, and total mass values for a sport and position. Master athletes' body composition determinants are vital to identifying athletes' highest performance categories: ball, track, and water/gym. This study helps identify the most critical factors and methods to predict athletes' standards to perform better. We analyzed data on 102 male and 100 female athletes collected at the Australian Institute of Sport. The data set had 13 columns, including athletes' height, weight, BMI, blood levels, percent body fat, and the Sum of skin folds. There are 59 athletes in the ball category, 63 in tracks, and 80 in the water/gym category. The dataset does not have any missing data, and we identified outliers in many variables in the dataset. Therefore we standardize data before building and validating our model, which will put all the variables on the same scale to use in the model. BMI represents the weight and height of the athlete; therefore, we do not use weight and height as predictors. We used two methods to analyze our data,

1. K Nearest Neighbour (KNN)
2. Random Forest

Unlike most algorithms, KNN and Random Forest are non-parametric models, which means they do not make assumptions about the data set. Therefore these two algorithms are more effective since they can handle actual data. We used both algorithms to identify the importance of the variables to predict sports groups, and surprisingly Sex (Male/Female) is the least important variable identified by both methods. Figure 1 shows the male and female BMI vs. Body Fat graphs, and there is no significant difference between the two graphs. The Sum of skin

folds (SSF) is the most crucial factor in categorizing athletes. SSF is one of the most prominent techniques used by sports professionals to assess body composition.

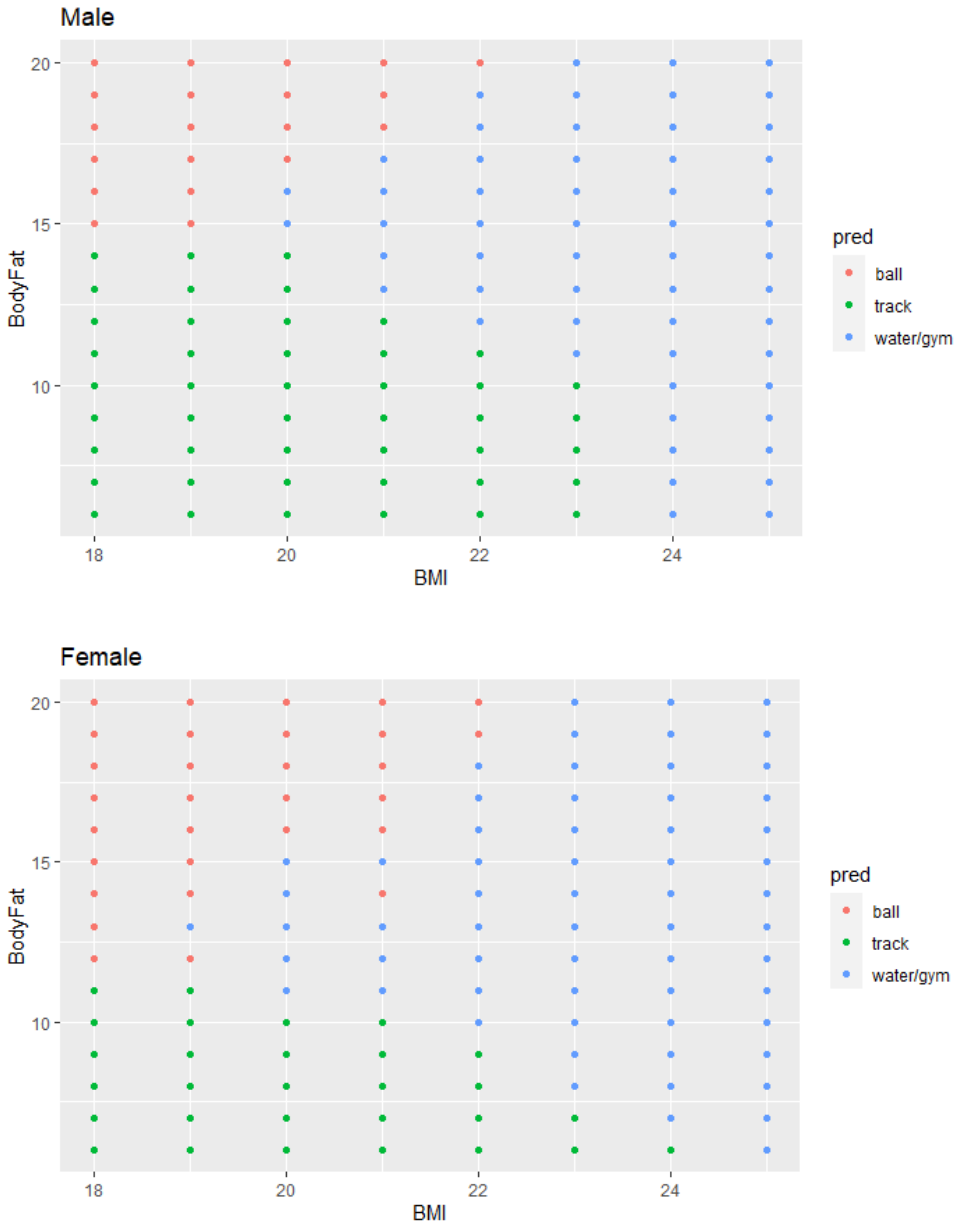


Figure 1

The next step of our study was to identify the best model with tuning parameters to predict new data. Even though the Sex variable is the least important factor identified by algorithms and Sports activity seems to slightly reduce differences in the whole body, we believe the differences

between the sexes in body composition are well established. Therefore we analyzed two models with and without sex variables using both KNN and random forest methods.

We used two model assessment methods to tune hyperparameters and identify the best model to predict new data. The highest accuracy model was the Random Forest model with the sex variable.

Our model,

***randomForest(Sport_group ~ ., data=Athlete, ntree=500, mtry=8, importance=TRUE)***

Even though the Random Forest algorithm did a better job than KNN, the accuracy of the 'ball' and 'water/gym' classes is low. That means our model is poorly classifying the athletes for the ball and water/gym categories. The ROC curves in figure 2 indicate track in green, ball in blue, and water/gym in red. The curve is closer to 1 means it is an excellent model to predict the class (AUC of track -0.9, ball - 0.78, water/gym = 0.75).

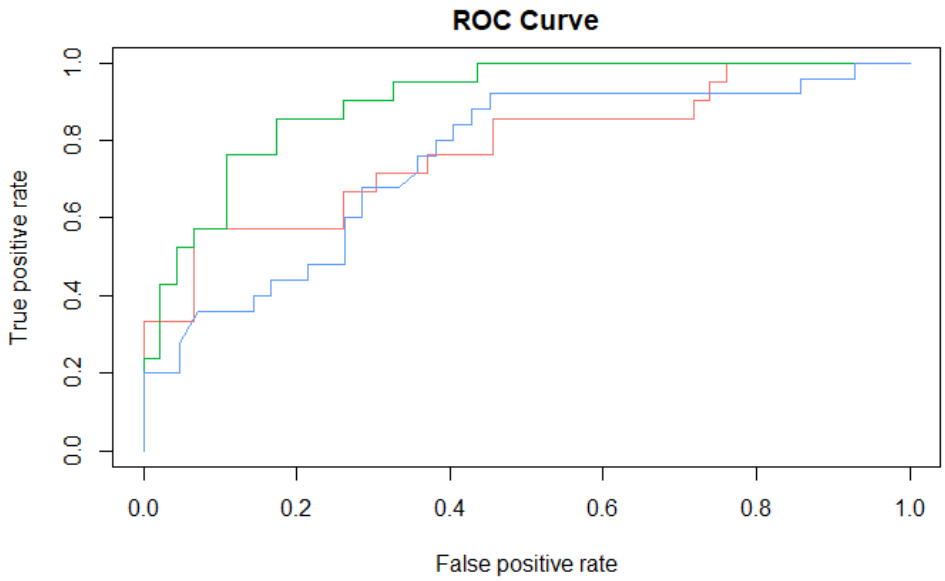


Figure 2

Figure 3 shows the importance of the variables to predict new data in the selected model.

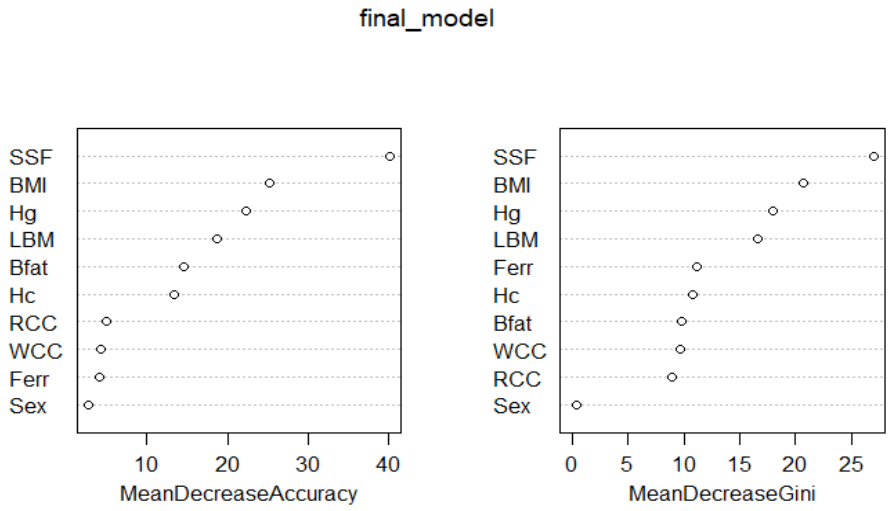


Figure 3

SSF is the most important, and Sex is the least important variable. The model chosen shows that body fat is not highly effective for prediction. Still, in general, there is a positive relationship between SSF and body fat of the general population and athletes. Figure 4 shows the relationship between Sum of skin fold and body fat of athletes dataset.

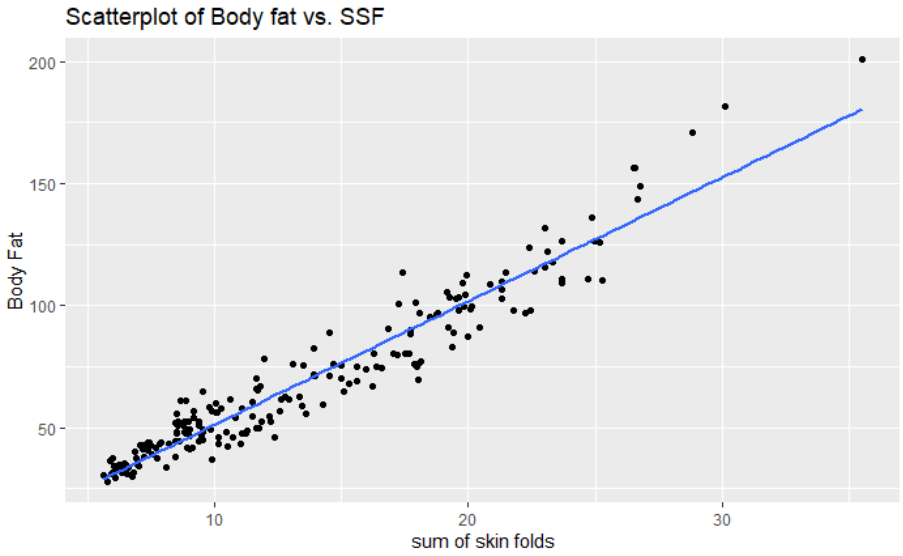


Figure 4

Body fat is vital to body composition in athletics; for example, top runners are light and skinny, and their body fat averages about 7% of males and 12% of females. Top swimmers are big made and usually carry more body fat than other athletes: averages 10-12% for men and 19-21% for women. Therefore we believe body fat is essential to predict sports groups as SSF, but the model says otherwise. One reason for this is insufficient data to train the model (this data set has only 202 data); therefore, we suggest collecting more data to train the model. Also, body composition is different for males and females; thus, adding more data will help identify the difference of sports between Sex.

We recommend categorizing the sports group into specific sports because each sport has its standard level of athletes.

In conclusion, our model is a good prediction for track athletes but not for other categories because of the lack of data to train the model. Also, our model output shows the Body fat and LBM as low importance variables compares to SSF, BMI, and HG. In general, Body fat and LBM are two critical factors in identifying athletes; therefore, this model is ineffective in predicting sports groups.