# Executive Summary

Competitive sports generate large revenues across the globe. In this environment, sports organizations need to be very discerning about who they choose to compete in a specific sport. When selecting participants, actual trial performance is important. However, it is also beneficial to know about unseen factors that may not appear in trial performance.

If the activity involved with different sports tends to emphasize different physiological factors, then organizations would want to choose athletes who are pre-disposed to those factors at the start. This analysis explains how to predict specific types of performance based on specific types of physiological factors.

## Overview

This analysis breaks multiple sports into two main types of physical effort. Sustained effort is seen in sports like running or basketball where team members need to maintain a sustained amount of effort throughout the competition. Sprint effort is seen in various field and other sports where a burst of energy is required at certain times, and the body is able to rest or recover more frequently.

## Data Used

The physiological factors used to predict these types of physical ability are height, weight, lean body mass, body mass index, percent body fat, and gender. These reflect athletic ability, but one body type may not be best for all types of sports. The study also looks at blood test data which determines red cell count, white cell count, plasma ferritin concentration, Hemoglobin, and Hematocrit levels. These measures are related to how well an athlete is able to carry and maintain oxygen in their system.

## Methodology

Two data mining techniques are used to determine the relationship between these factors and which type of sport an athlete may be predisposed to perform well in. The first is linear regression, which is able to provide a straight forward formula to establish the probability of an athlete’s classification. By setting a threshold, this can categorize players on either side of the threshold. The second technique is tree mapping. Similar to linear regression, this provides a formulaic way to determine an athlete’s classification without complicated math or confusing algorithms.

## Model

With the data set provided, the optimal predictive model is a logistic regression model with the formula:

This formula produces a probability between 0 and 1. A probability above the desired threshold value (.5 by default) will designate an athlete who is ideal for sustained activity. A probability below the threshold will designate an athlete who is ideal of sprint based activity. The threshold can be adjusted to minimize the errors as the model is used with new data. The current data set provided the lowest error (27%) at a threshold of .38. As more data is collected over time, either the ideal model, the optimal threshold, or both may change.