Capstone – Sheila Lynch – Participating / earning a medal at Olympics over age 30

In consideration of machine learning as it pertains to my main question, I am focusing on Supervised Machine learning as it has input variable(x) and an output variable(y) and there is data to run an algorithm to learn the output. My goal is that I will train and build the code based on the sport Equestrianism which I realize does have a higher mean age than the Olympic Sports overall and evaluate medals earned by those over 30. Then I will test using Rowing which in the article mentioned being a sport that mean age is higher as well as initial data shows and I will see if there is the same chance in proportion to earning a medal over 30. Based on the test set validating the training set, I feel you could use the same analysis to calculate other sport or even by Team the probability of someone over 30 participating and possibly earning a medal.

As this is supervised machine learning, I originally thought I would be using classification plan to use classification method as my data is categorized and separated into groups such as over 30 years old, Sport, Medal and NOC. However as I considered regression, while my data is not numeric, I am looking to predict a future trend so I have begun building my model coding with regression and linear regression modeling.

The main features I plan to use are Age, Sport, Medals and Team/NOC. The first feature is age over 30, then focus on Sport and if a Medal is used. Currently the Team/NOC is planned to be used as a by product.

I plan to focus on linear regression charts to evaluate the data and make conclusions. As such I will create a training and test set. I will create my code for the training set and then I will use the test set to confirm the conclusions are supported. In working with this I have found the data set is still large and my R squared percentage is low. I will continue to work on this to see if I can find more items that could raise this percentage.

I also started to look at Clustering, however on initial review my dataset is too large to run. I am working through some options to reduce the dataset size. While I am not sure Clustering is the best method I am trying to explore if it provides support on this conclusion.