Data Mining Task

The task is to find players that are close to user-based height, weight, age of players in the dataset. The task will use K-Nearest Neighbor Classifier.

Dataset

The dataset is comprised of 1035 records of Major League Baseball players with their name, team, position, weight, height, and age. The data are SOCR Data which are of current players.[1] Dataset was obtained from different resources.

Methodology

I will use the K-Nearest Neighbor Classifier which will use weight, height and age as training Index. For this purpose, we will most likely round age, since it is given as floating point. The idea would be to find players that are close to the user. The K-nearest Neighbor Classifier will fit the best in this case as does not require training and is easy and simple to implement. Also, in our current dataset the computation time would not be as high as we only have 1035 records. KNN is a non-parametric algorithm. We could potentially create formulate for height factor + weight factor + age factor = factor and use factor as single point for indices.

Final Product

Main outcome of the project is to mine this data and use KNN classifier. It is also interesting application that many users would find interesting to try out. The point of project is to learn how to use KNN classifier and apply it to real world data and create an interesting application at the same time. I think the results will be good from this project. The success of the project would be measures on how close and accurate will the algorithm display players to the user. This algorithm could be used as fun addon to the Major League Basketball website to make fans more interested in players who are close to them.

[1] "SOCR Data MLB HeightsWeights." Socr RSS. Accessed November 5, 2019. http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_MLB_HeightsWeights.