

Nearest Major League Baseball player

Mykhailo Bykhovtsev

Table of Contents

I.	Introduction	3
II.	Data Mining Task	3
III.	Technical Approach	4
IV.	Evaluation Methodology	5
V.	Results and Discussion	6
VI.	Lessons Learned	7

I. Introduction

Finding which Major League baseball player user is, based on user input. I think this project is a great example of applying datamining or machine learning for practical uses that are not necessary, but interesting. It is similar in the idea to the applications that will show you which celebrity or dog you are based on your input.

II. 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 uses K-Nearest Neighbor Classifier. This is an interesting project because many people on the internet wonder which Major League baseball player they are close to.

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.

The dataset should be partitioned into training and testing sets. Currently using 20% of data as testing data.

III. Technical Approach

Dataset Partitioning

The dataset was partitioned into balanced training and testing sets with ration 80% training and 20% testing.

Hyperparameter Tuning

The KNN algorithm was used as library in Python. In current implementation we can adjust number of neighbors. We can also specify weight as parameters or which algorithm to use.

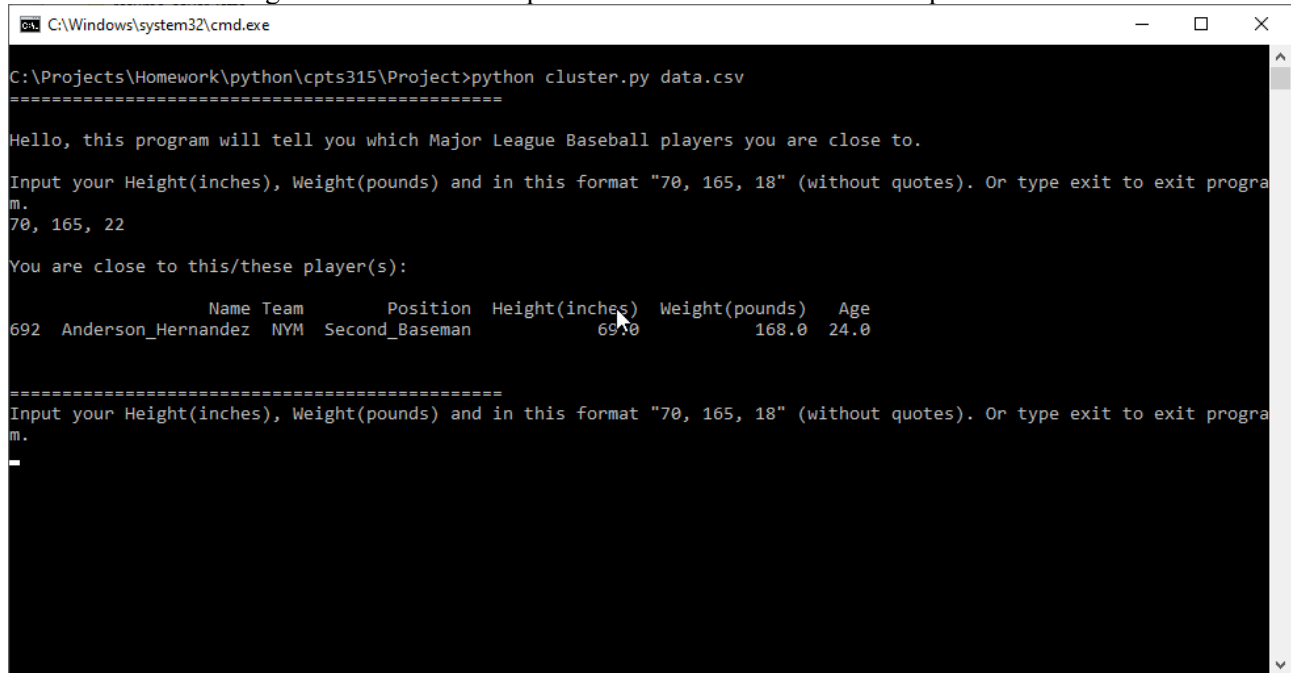
IV. Evaluation Methodology

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.

We can evaluate result by either manually testing input or finding which input will result to which output and automatically test it with unit tests.

V. Results and Discussion

Results are that our algorithm works well. I provide screenshot below as example of execution



```
C:\Windows\system32\cmd.exe

C:\Projects\Homework\python\cpts315\Project>python cluster.py data.csv
=====
Hello, this program will tell you which Major League Baseball players you are close to.
Input your Height(inches), Weight(pounds) and in this format "70, 165, 18" (without quotes). Or type exit to exit program.
70, 165, 22
You are close to this/these player(s):

      Name Team      Position Height(inches) Weight(pounds) Age
692 Anderson_Hernandez NYM Second_Baseman      69.0      168.0 24.0

=====
Input your Height(inches), Weight(pounds) and in this format "70, 165, 18" (without quotes). Or type exit to exit program.
_
```

VI. Lessons Learned

Summary

It is really amazing to learn new algorithms like K Nearest Neighbor. We did implement it before in any of our assignments, so it was great experience for me to learn how to implement and use it in code. I was always fascinated with this algorithm. I am happy that the results of this project are positive, and we are able to practically use it for entertainment.

The ratio of training to testing data is not arbitrary

The dataset was split 80-20 between training and testing sets