120 Years of Olympic History

Anna Frazier, Alex Carson, Nikhil Murty, Arion Morshedian, Justin Kim

Final Project Proposal
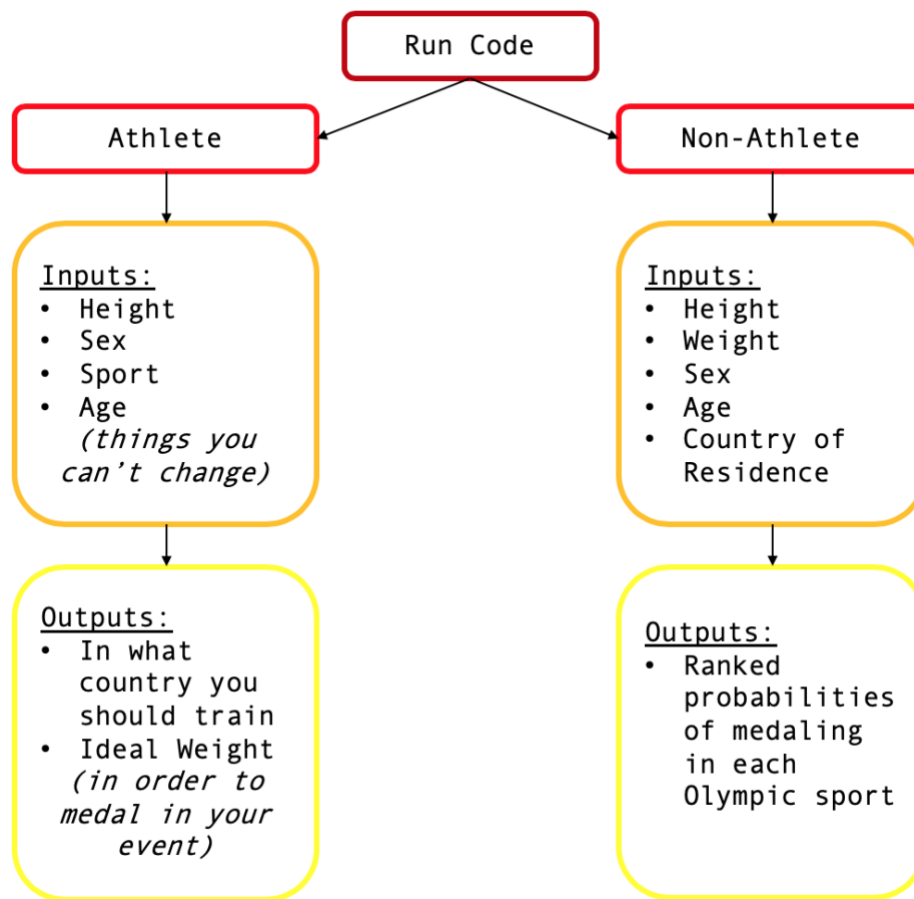
560.348 - Probability & Statistics for Engineers

I.   Why We Are Interested?

Our group was initially interested in how the outcomes of sporting events can be predicted based on the traits of the players involved. When we found the dataset of Olympic athletes and their performance results, we became interested in the patterns we could find in this data set. Specifically, our group wants to explore how a non-athlete's performance, in particular if they will win a medal (bronze, silver, or gold) in any sports can be predicted by their physical traits (sex, height, weight) and the country they live in. For athletes, we want to provide suggestions for their training programs (location, target weight, etc.) based on the sports they are playing to maximize the probability of winning a medal.

II.   What is the Model You Will Build and Code?

```
                          ┌──────────────┐
                          │   Run Code   │
                          └──────────────┘
                    ↙                          ↘
        ┌──────────────┐              ┌──────────────┐
        │   Athlete    │              │ Non-Athlete  │
        └──────────────┘              └──────────────┘
```

Inputs:
- Height
- Sex
- Sport
- Age
  (things you can't change)

Inputs:
- Height
- Weight
- Sex
- Age
- Country of Residence

Outputs:
- In what country you should train
- Ideal Weight *(in order to medal in your event)*

Outputs:
- Ranked probabilities of medaling in each Olympic sport

In the model shown above, we have two main options for our code: athlete and non-athlete. For the athlete option, this is "statistical advising" for an athlete of a particular height competing in a specific sport. They will be given the best country in which they should train and the ideal weight they should attain based on the dataset entries. Additionally a heat map of the world will be output that shows the focus of medals for each sport. For the non-athlete option, this is a fun opportunity for individuals to see what sport they would excel in with their given traits.

III.   **Sources for Data**

The source of our data will be a dataset from Kaggle named "120 Years of Olympic History: Athletes and Results." This dataset includes the following content:

- ID - *Unique number for each athlete*
- Name - *Athlete's name*
- Sex - *M or F*
- Age - *Integer*
- Height - *In centimeters*
- Weight - *In kilograms*
- Team - *Team name*
- NOC - *National Olympic Committee 3-letter code*
- Games - *Year and season*
- Year - *Integer*

- Season - *Summer or Winter*
- City - *Host city*
- Sport - *Sport*
- Event - *Event*
- Medal - *Gold, Silver, Bronze, or NA*

The link to this dataset is: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results#athlete_events.csv

It is important to not that while the dataset comes with pre-set sports and events, our group will be creating our own categories to use for our project, which may involve combining or splitting up various sports based on event.

IV. Expected Outcome

We expect the final data to show what height, weight, and nationality a person should have if they want to obtain a medal in a certain event in the olympics. We hope to create a code that can be run and if the user inputs their information, it will output what events they are most likely to medal in. We also expect there to be clear patterns in which countries athletes train in and the events that they are able to medal in, because various countries are known for their success in specific sports/events.