Group 1: Kelly Phalen, Emily Wang, and Xinyu Wu

Performance of Olympic Athletes

Of the project ideas you submitted in the previous deliverable, select the one you want to work on throughout this project. In making your decision, please refer to the feedback you have on the previous deliverable. If all ideas were found plausible, feel free to select the one that you liked best **as a team**. You will be working on this project idea during the remainder of the semester.

Complete this section based on the previous deliverable. After deciding on your topic idea, simply copy and paste the same information from the Topic Proposals document here.

1. Problem Statement

In this project, we will be training a machine learning model to understand the question: what factors are influential in Olympic medal winners? In general, we hope to develop a model that will identify the likelihood of a given individual winning a Gold, Bronze, Silver, or no medal at the Olympics. Inspired by the upcoming 2021 Summer Olympics and the betting that comes with it, we hope that we can learn more about which sorts of athletes are likely to succeed. Do athletes with certain heights and weights perform better in a certain sport? How do athletes from different countries compare? Is body mass index (BMI) a good determinant of an athlete's athletic performance? These are just a few of the many questions that our analysis and machine learning model will answer.

2. Significance of the Problem

At first glance, this project may seem to be for entertainment purposes rather than something that should be taken seriously. However, we believe that this problem is significant because of the history of the Olympics, it's importance in our society, and the insights and applications that our project may provide. Held every four years and two years respectively, the Summer and Winter Olympics are an important tradition where athletes from across the globe are able to come together and compete, representing their country. Particularly in a time period where countries do have political tensions and conflicts, the Olympics provide a platform that puts peace and honor above all. Just like most other problems that exist in our world, wealthier, more developed countries tend to have the upper hand when you look at the distribution of Olympic medals between countries. From our project, we will be able to understand whether that difference is significant, or if other factors such as an athlete's height and weight may be more influential in their performance. In a sense, we hope that our analysis can provide insight on how countries should pick their athletes, and ideally our findings can identify countries who have historically performed worst in the Olympics, and why that may be.

3. Dataset(s)

https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

We obtained our data from a dataset available on Kaggle that covers 120 years of Olympic history.

Dataset File

Download or scrape your data from the source you identified above. Save your dataset as a CSV file. The first row of the file should contain variable names.

Your dataset should have at least 1000 rows, corresponding to samples/records, and 10 columns, corresponding to features and target variables. This is the bare minimum. The more, the better!

Describe your variables below (add more rows if necessary):

Variable name in file	**Description**	**Feature/ Outcome**
ID	Unique identifier of athlete	feature
Name	Athlete’s name	feature
Sex	Male or Female (M or F)	feature
Age	Athlete’s age (integer)	feature
Height	Athlete’s height (cm)	feature
Weight	Athlete’s weight (kg)	feature
Team	Team name (country)	feature
NOC	National Olympic Committee 3-letter code of country	feature
Games	Year and season (i.e. 20212 Summer)	feature
Year	Integer year of games	feature
Season	Summer or winter	feature
City	Host city	feature
Sport	Sport (i.e. basketball)	feature
Event	Specific event (i.e. speed skating women’s 500 metres)	feature
Medal	Gold, silver, bronze, (corresponding to first, second or third place) or N/A	outcome
In the Feature/Outcome column, indicate whether the variable is a feature or outcome variable. You need to have at least one outcome variable, with several feature variables.		

Based on what we discussed regarding machine learning (Week 3), does your dataset include a set of feature variables and one outcome variable that you can use for a supervised machine learning task? Please explain. You need to meet this requirement and show us you understand that you are required to use a predictive model in your project.

Yes, our dataset includes a number of feature variables and a single outcome variable. Using the set of feature variables such as age, height, weight, sport, event, etc, we will be able to run a supervised learning algorithm and predict the medal placement (outcome variable) of olympic athletes. From this, we may find that some variables such as height and weight may be better predictors of medal placement in comparison to the country/team.

Further info on submitting the dataset:
Submit a CSV file, or multiple files, containing your data. If the dataset is too large, you can upload it to GitHub, or any other online repository, and provide a public link.

If you have scraped your data, you should also submit a Jupyter Notebook containing your Python code used to scrape the data. Please be reasonable and comment your code out whenever it makes sense to do so.