Rough Outline for “What kinds of Medals will the athletes get?”

Purpose

The project aims at employing the Dataset of 120 years of Olympic history to build the model to predict what kind of medals (Gold, Silver, Bronze, or NA) the athletes will get based on Google Cloud Platform.

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

Dataset comes from Kaggle, which is “120 years of Olympic history: athletes and results” including the 15 variables of ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event and Medal.

Preprocessing

- Git rid of rows lack of information
- Some variables are in text format such as “Event”, “Medal”, “City”, etc. We will digitize them.
- We will normalize the numeric columns for building a more accurate and efficient model.

Tentative plan for analysis

As we have learned from class, we plan to load data into Big Query, employ Dataproc to create a Hapood cluster and use Pyspark to build the model. Random forest classification is the machine learning algorithm we plan to use. Also, Datalab will be used for more interactive exploration. Finally, DataStudio will be used to create dashboard to visualize our products to users.