Abstract

The goal of this project was to create a web app that highlights various aspects of the Fitness Registry and Importance of Exercise International Database (FRIEND) to increase interest and data contributions to the registry. I created a data pipeline that uses data from FRIEND to create summaries and regressions that are then deployed through a web app. The resulting web app is for researchers, clinicians, and the general public. It highlights trends in health metrics and also has practical applications in that it can be used to determine an individual's fitness percentile.

Design

The American Heart Association recommends that fitness (i.e., maximum oxygen consumption [VO_{2max}]) be considered a clinical vital sign that is regularly assessed in the same manner as other risk factors (e.g., blood pressure, cholesterol levels, and body weight). Like other risk factors, the accurate interpretation of fitness requires reference standards to stratify patient risk by first determining what constitutes a "normal" value for that individual. FRIEND consists of >100,000 tests and has published reference standards for fitness, yet a larger and more representative dataset could improve upon current reference standards. I was recruited by the FRIEND research committee to create a web app that highlights various aspects of the database in real time to increase interest in the project and therefore increase the contributions of test data to FRIEND.

Data

The data for this project comes from FRIEND. This is an international database consisting of >100,000 fitness test results from an exercise test as well as data from pre-test health screenings. For the regression models, not every test within the dataset contained the necessary features. The number of tests with the needed features was 35,120.

Algorithms

Database creation and storage: The FRIEND dataset is available as an Excel spreadsheet. I converted this .xlsx file (available on my local computer) to a SQL database and saved it on Google Cloud. Considering FRIEND is not updated frequently, the data pipeline involves copying the SQL database from Google Cloud back to my local computer when updates to the web app are needed.

Summary figures: Summaries of different variables within FRIEND were created. These included locations of where the data was collected, distributions of variables (eg, age), and trends in variables (eg, changes to fitness with increasing age). All summaries are interactive to increase interest in FRIEND.

Publications summary: A list of publications from FRIEND was collected via webscraping. This list was not saved with the SQL database since it is more regularly updated.

Regression models: Fitness is not often directly measured, so an OLS linear regression model was created to predict an individual's fitness from their age, sex, height, weight, test exercise mode, and country. From the prediction of fitness, the individual's fitness percentile could be calculated using current fitness reference standards.

Testing and Robustness: Data cleaning during processing was used to eliminate extraneous values prior to the creation of the data summaries and regressions. Various lists were also created from the dataset and then used as options in the user interface of the web app to improve robustness of the data pipeline.

Tools

The database was created with SQLite and then stored on Google Cloud. Data cleaning and robustness testing was done with NumPy and Pandas. Summary figures were created with Plotly, Matplotlib, and Seaborn. The linear regression was created with sklearn. Webscraping was done with BeautifulSoup. The web app was created with Streamlit.

Communication

Slides and visuals were presented and posted on my personal GitHub page. The web app is available at: https://share.streamlit.io/jimpeterman/metis_engineering/main/app.py