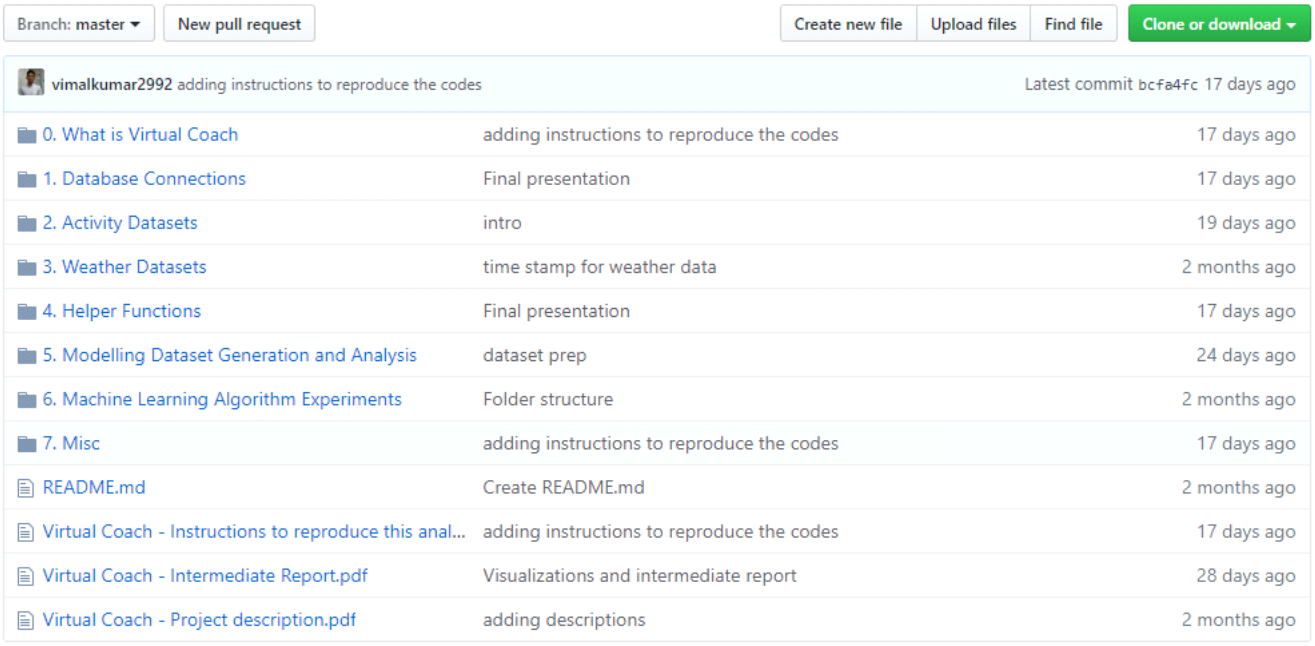
Virtual Coach

Saturday, December 21, 2019 5:35 PM

The link for the complete project is here



Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

vimalkumar2992 adding instructions to reproduce the codes — Latest commit bcfa4fc 17 days ago

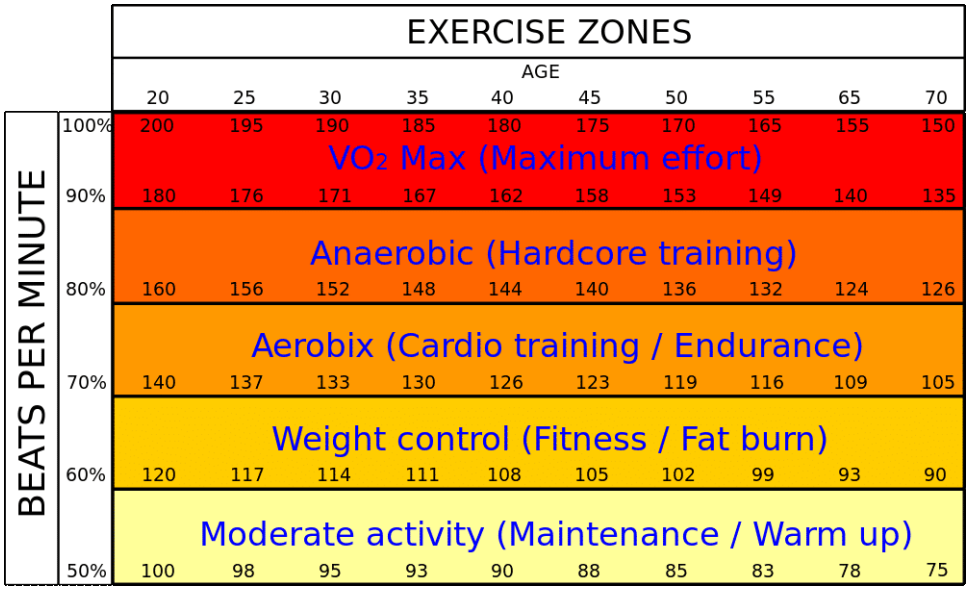
0. What is Virtual Coach	adding instructions to reproduce the codes	17 days ago
1. Database Connections	Final presentation	17 days ago
2. Activity Datasets	intro	19 days ago
3. Weather Datasets	time stamp for weather data	2 months ago
4. Helper Functions	Final presentation	17 days ago
5. Modelling Dataset Generation and Analysis	dataset prep	24 days ago
6. Machine Learning Algorithm Experiments	Folder structure	2 months ago
7. Misc	adding instructions to reproduce the codes	17 days ago
README.md	Create README.md	2 months ago
Virtual Coach - Instructions to reproduce this anal...	adding instructions to reproduce the codes	17 days ago
Virtual Coach - Intermediate Report.pdf	Visualizations and intermediate report	28 days ago
Virtual Coach - Project description.pdf	adding descriptions	2 months ago

Motivation

Health and Physical fitness has become the top priority of the current era and with the everyday busy schedule, it has become challenging to be disciplined in our physical activities. Wearable devices solve majority of the challenges, by reminding us to move and sweat everyday. Too much of anything is not good, but how do we know that we are overworking and getting burnt out? Our motivation is to build an intelligence layer on top of the current wearable devices, that leverages the features that are currently collected through wearable devices and provide real time recommendations to the user during the activity.

Context

Every human body is different and Heartrate is a direct measure of the extent to which we are pushing our body. The highest heartrate for a human is approximately (220 bpm - Age of the individual). For someone who is 20 years of age, the maximum heartrate is limited to 200 bpm. The range of 100bpm leading to the maximum limit is the activity zone. For a 20 year old individual, the activity range is between 100 to 200. This range can be split into 5 zones.



- Zone 1: 100 - 120 Easy / Moderate Activity
- Zone 2: 121 - 140 Fat burn / Weight Control
- Zone 3: 141 - 160 Cardio / Aerobic
- Zone 4: 161 - 180 Perform / Anaerobic
- Zone 5: 181 - 200 Peak / Maximum effort / Vo2 Max

Depending on the individual goals, we need to ensure that the user is on the respective zone. For someone who is a beginner, they should reside in Zone 1 / 2, for someone who trains regularly, should reside in Zone 4. For someone who wants to get back to shape should reside in Zone 3.

Objective

To build an intelligence layer on top the wearable devices, that learns the pattern of heartrate across different activities, temperature, altitude, body temperature conditions that guides the user to stay on the aspired zone.

Dataset

An athlete's sports activity data with their consent has been used for this analysis. The dataset captures the information about the athlete every second during an activity. About 36 activity data has been collected, and the features measured were Heartrate, Altitude, Distance covered, Stride count. External environmental variables such as Temperature, Humidity are procured externally for the day of the activity.

- Activity data
  Dataset count: 36
  Features per dataset: 11
  Observations per dataset: 30,000 (Approximately)

- Weather and external variables
  Features per dataset: 3 (Temperature, Humidity, Precipitation)
  Observations: Day level

- Parsing and Importing
  Activity datasets are in csv format, and the datasets are pushed into SQLite3 database with a

indicator for every activity

- Analysis
  The key questions below were answered through writing SQL queries
  Weather dataset was sparse and we were not able to collect data for all the activities
  1. How much time an athlete spends in every zone?
  2. Do they switch between zones or they move in a linear fashion?
  3. What is the time taken by an athlete to warmup before entering into another zone?
  4. Did the athlete get into the risk Zone 5?
  5. How long did the athlete spend in Zone 5?

- Inference
  Based on the above insights, we decided that predicting the risk of athlete entering the risk zone 3 minutes head will be helpful, not too early not too late, as the athlete can still act on it and avoid the risk

- Utility Functions
  There was scope for us to define custom metrics to explain the zone transference better so we consumed the raw features and computed the below mentioned custom features. We have modularized the python codes as utility functions for future reference. Also a set of inputs are accepted from the user, which will help us define the target zone and other assumptions required for zone definition.
  1. Heart rate metrics
  2. Elevation metrics
  3. Speed metrics
  4. Zone metrics
  5. Missing value treatment
     i. This step is contextual, if heartrate metric is missing we replace it with the previous second metric
     ii. If elevation gain is missing, we replace it with 0
  6. Data prep module
     i. Based on the previous analysis, we knew the activities during which the athlete has actually ventured into Zone 5, so based on that we sampled the activities into train and test
     ii. This module computes the heartrate after 180 seconds for every datapoint and labels that as response variable
     iii. Standard scaling is performed on the dataset
     iv. PCA is performed to reduce the number of components for model train
     v. Similarly the above steps were done for test dataset as well
  7. Model build and Testing module
     i. This module builds SVM linear kernel model and predicts the risk of the athlete venturing into Risk zone 180 seconds ahead of time
     ii. The model is tested on the hold out dataset and the performance was observed as 97%
     iii. The performance has lot of scope of improvement as the case of Zone 5 is sparse and most of the cases are still predicted as non-risk cases

- Learnings
  1. Even when the raw data is available in a flat file format, there can be some anomalies in it and it has to be treated
  2. The treatment need not be same for all metrics, few metrics like heartrate can be treated with the previous value, and elevation gain has to be treated differently
  3. The most exciting piece in any prediction is feature engineering, so we need to come up

with features that clearly exhibit the characteristic that we are interested in

4. The algorithm can be chosen based on experiment, we won't be able to assess the required algorithm based on the problem at hand
5. Clear documentation is the key, especially when multiple people are working on the same project