MET regression models

To predict the MET values of the activities, different features are needed for our prediction models. Multiple data sources were combined to prepare the different features for the prediction models. For the creation and calculation of the features we used the activPAL accelerometer data, Vyntus One data and a file containing the respondent characteristics that was supplied by CBS. This file contains different characteristics from all the respondents that participated in the lab research. The following characteristics were used for the creation of our features: length, weight, gender, age category, if the respondent meets the balance guidelines, if the respondent meets the bone and muscle guidelines and if the respondent is sporting.

Since most of these characteristics were non-numerical or string values, all these features were converted to numerical values. These numerical values differed between simple True/False converted to 1 or 0 and numerical values that represent a category, in our case the age category. The age category '15-19' got a numerical value of 0, '20-24' got a numerical value of 1, etcetera. By converting our features to numerical values our Machine Learning models were able to be trained and evaluated (Brownlee, 2020).

A few other features have more complex computations. Since MET is measured in minutes (more in this in chapter... Van Mark), the following features were also resampled to 1 minute. The 'sum of magnitude of acceleration', which means the total acceleration within a certain timespan, is resampled to 1 minute. The formula (Measurement of Physical Activity Using Accelerometers, 2016) can be found in **equation #add number**. The *X*, *Y* and *Z* data from the activPAL accelerometer is used to calculate the acceleration.

$$\sqrt{x^2 + y^2 + z^2}$$

// **Equation #**

The last feature is the speed, which is also resampled to 1 minute for every activity. **Equation #add number** (Calculate speed from accelerometer, 2014) was used to calculate the velocity from the acceleration of the X, Y and Z axis. The velocity is needed to calculate the speed.

$$v(t) = v(0) + \sum a \times \delta t$$

// **Equation #**

The *t* is the *time interval* of the x, y or z velocity. In this case 0.05 seconds (Why activPAL?, z.d.). The *a* is the *acceleration* of the *X, Y* or *Z* axis. After calculating the velocity for the X, Y and Z axis it is possible to calculate the speed. **Equation #add number** (Calculate speed from accelerometer, 2014) has been used to calculate |*v*| which is the total speed. The *X, Y and Z* inputs are taken from **Equation (velocity number)**.

$$|v| = \sqrt{v_x^2 + v_y^2 + v_z^2}$$

// **Equation #**

Once the features were created, 2 different ensembled decision tree regression models were configured. The Random Forest and XGBoost model were chosen and configured as identical as possible to pick the best performing model for each activity. To pick the optimal combination of features, the implementation of Recursive Feature Selection (RFE) was applied on both models. The chosen features were extracted from the prepared data frame with the method described in chapter from Mark (about train/test/split method). Finding the optimal amount of decision trees was decided experimentally. A function was written to find the most optimal number of trees between a certain range of the related model. Hyperparameter tuning was eventually applied on both models to make sure the configuration was implemented in the best possible way.