**Fitbit Case**  A grey and white logo

AI-generated content may be incorrect.

*Unhealthy lifestyles lead to increased premature mortality and are a risk factor for sustaining noncommunicable diseases (NCDs) such as cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes. NCDs caused 63% of all deaths that occurred globally in 2008 . There are four behavioural factors that have a significant influence on the prevention of NDCs: healthy nutrition, not smoking, maintaining a healthy body weight, and sufficient physical activity. Insufficient physical activity is one of the leading risk factors for the major NCDs and not meeting the recommended level of physical activity is associated with approximately 5.3 million deaths that occurred globally in 2008 .*

*A high amount of sedentary time without sufficient daily physical activity leads to a higher rate of all-cause mortality. Besides the increased risk of premature mortality in the long term, the short-term quality of life, being able to work, and social participation is also threatened by insufficient physical activity. Fortunately, these risks are eliminated when this sedentary time is compensated for with sufficient physical activity of moderate intensity.*

*In Western civilization, living a sedentary lifestyle is the rule rather than the exception, as many people work in office environments. In pursuance of preventing the negative effects of insufficient physical activity in the workplace, the Hanze University of Applied Sciences Groningen (HUAS), a large university in the northern part of the Netherlands, started a novel initiative named (in Dutch): ‘Het Nieuwe Gezonde Werken’ (The New Healthy Way of Working; HNGW). With HNGW, the HUAS aims to promote a healthy lifestyle and physical activity during the workday. HNGW consists of providing participants with educational group meetings, food boxes with healthy recipes, and individual coaching sessions supplemented with an activity tracker. Despite the fact that participants are coached every two weeks and measured continuously, it remains difficult for a coach to provide timely personalized feedback. The manual task of creating personalized feedback is time consuming, and as such it is not always possible for the participants to get in-depth and timely daily feedback on their progression. Furthermore, current activity trackers do not provide a prediction for reaching the daily goal.*

*In order to fill this gap, you have to predict whether a person will reach his or her daily average enabling an automated coaching process.*

**Assignment**

The assignment is: find out what the average number of steps per person is at 18.00 on a working day. And is there a model that predicts whether the person will reach his average?  
Taking into account the influence of

1. Seasons
2. Holidays
3. The day of the week

**Deliverables**

For this assignment, we expect you to deliver a .zip container containing the following:

* A report of your process, explanations and insights
* A notebook with code for the relevant elements you’ve described in the report
* Data (if you’ve created any new data, make sure this is mentioned in the report)

In addition to a front page, contents page, and introduction, the report contains at least the following information/chapters:

1. Description of the data. Where does the data come from, what does it reflect?
   1. Draw the Directed Acyclic Graph of the independent variables in the data, and how they lead to the dependent variable you would like to see.
   2. Describe what the column names are, what kind of data each column contains, what does each variable mean, which variables are dependent and independent.
   3. Describe the data, as far as possible, using descriptive statistics.
   4. Describe the correlation matrix and explain the correlation coefficients. Explain which variables are meaningful to use for the predictive modelling.
   5. Visualise the data in a useful way; try to capture relevant patterns
   6. If applicable: visualise the data as a time series (e.g. day/hour/week in combination with the dependent variable).
2. Pipeline of predictive modelling.
   1. Define the features to use for your prediction, and the outcome variable, given the data.
   2. Prepare the data so that it can be used for model fitting.
   3. Make use of additional data sources to explain patterns.
   4. Add the notebook of the algorithms used (e.g. Decision Tree, Random Forest, Linear Regression, K-means Clustering, Logistic Regression, Naive Bayes).
   5. Display the performance of the algorithms with appropriate metrics (e.g. using R2, RMSE, a Confusion Matrix and F1-score, or AUC-ROC) and discuss what the metric tells you.
   6. Argue why certain algorithms were used and what the arguments are for the best model.
3. Conclusion  
   State your conclusions and explain what can be learnt from this data.
4. Justification of tasks

If you worked with someone, make a joint statement on approximately who took up which tasks in the process.

**Available data**

In the accompanying .zip file, you’ll find:

*Data Coaching Fitbit* containing the step data per minute of the various people with numbers (1119,1120,1121,1122,1123,1124,1125).

**Hint**

We’re not necessary looking for perfect accuracy in this task, but rather a logical series of design choices and explanations of the data. If the prediction (using regression) looks something like the following, you’ll have done a fine job at the prediction task itself. Note that, if you can argue why you chose to do a classification, that is perfectly fine too.