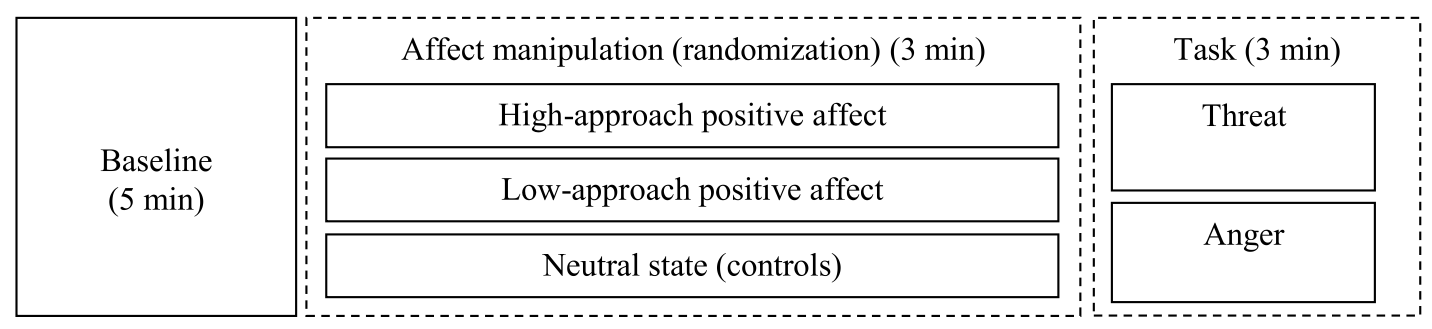
**Exam of: Data Mining and Analytics Course**

*Dataset description “Dataset\_Study2”:*

The dataset in attachment contains some features extracted by processing physiological responses of 186 healthy subjects. The aim of the study was measuring cardiovascular, respiratory, and electrodermal responses to positive stimuli that differed in approach intensity. Specifically, this study examined how high approach and low approach positive emotions buffer stress responses. Also, it was examined whether reactivity to anger (a high approach negative emotion) vs. threat would interact with high approach positive affect. The procedure followed by the study is illustrated in the next figure:



The experiment began with a 5-min baseline, followed by 3 min of watching affective pictures (high-approach positive affect, low-approach positive affect, or neutral depending on randomization), and 3 min of speech preparation (social threat or anger depending on randomization).

A dataset containing 19 features was obtained by processing the provided ECG signals, which is composed as follows:

* 558 observations (186 subjects x 3 phases, that is baseline, affect manipulation and task).
* 6 labels indicating the different conditions from which the features are extracted. Specifically:
  + Baseline condition = -1
  + Threat = 209
  + Anger = 208
  + Low-approach positive emotions = 308
  + High-approach positive emotions = 309
  + Neutral State = 108.

Features are extracted from the RR signal, which is the time elapsed between two successive R waves of the QRS signal on the electrocardiogram. Specifically,

* The features extracted in the time domain are:
  + MeannNN [ms] 🡪 The mean of the RR interval.
  + SDNN [ms] 🡪 The standard deviation of the RR intervals.
  + RMSSD [ms] 🡪 The square root of the mean of the squared successive differences between adjacent RR intervals.
  + Prc20NN [ms] 🡪 The 20th percentile of the RR intervals
  + Prc80NN [ms] 🡪 The 80th percentile of the RR intervals.
  + PNN50 🡪 The proportion of RR intervals greater than 50ms, out of the total number of RR intervals.
  + HTI 🡪 The HRV triangular index, measuring the total number of RR intervals divided by the height of the RR intervals histogram.
* The PSD was estimated using the Welch’s method. The features in the frequency domain are extracted by using the following band limits for the very low frequency (VLF), Low Frequency (LF) and High Frequency (HF) bands: (1) VLF: [0Hz – 0.04Hz]; (2) LF: [0.04Hz-0.15Hz]; (3) HF: [0.15Hz-0.4Hz]. The following features are obtained:
  + VLF [ms2/Hz] 🡪 The spectral power of very low frequencies.
  + LF [ms2/Hz] 🡪 The spectral power of low frequencies.
  + HF [ms2/Hz] 🡪 The spectral power of high frequencies.
  + TP [ms2/Hz] 🡪 The total spectral power.
  + LFHF 🡪 The ratio obtained by dividing the low frequency power by the high frequency power.
* Non-linear features (Poincaré plot and complexity indices corresponding to entropy or fractal dimension):
  + SD1 🡪 standard deviation perpendicular to the line of identity. It is an index of short-term RR interval fluctuations, i.e., beat-to-beat variability.
  + SD2 🡪 Standard deviation along the identity line. Index of long-term HRV changes.
  + SD1SD2
  + DFA\_alpha1 🡪 The monofractal detrended fluctuation analysis of the HR signal, corresponding to short-term correlations.
  + DFA\_alpha2 🡪 The monofractal detrended fluctuation analysis of the HR signal, corresponding to long-term correlations.
  + ApEn 🡪 Approximate entropy, which is a technique used to quantify the amount of regularity and the unpredictability of fluctuations over time-series data.
  + SampEn 🡪 Sample Entropy, which is a modification of ApEn used for assessing complexity of physiological time series signals.

*Expected Results:*

The student has to write a script in Python able to process the dataset following the main steps of data mining and analysis learned during the course such as: Data import, Data cleaning and check of missing data, exploration, visualization and normalization of features, best feature selection approach, principal component analysis, regression, clustering, classification etc.

These are just some examples of techniques, so don’t be frightened and feel free to select your best strategy. The aim of this data science project is to assess the approach of the student to explore the dataset, to understand which features should be used and discarded to build the best model able to predict the different emotional states.

**NOTE:** At least 2 days before the exam we expect that each student/group send us by email:

1. Python, R, Jamovi (you can choose the desired environment) script with annotated code.
2. Technical report with description of your project step by step mainly focusing on Method and achieved Results.