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

*Dataset description “Dataset\_Study4”:*

The publicly available POPANE dataset focuses on psychophysiological responses to positive and negative emotions of 1157 healthy young adults, collected across seven studies. Specifically, the attached dataset originates from “Study 4”, which explores the impact of anxiety on consumer product evaluations among 83 healthy subjects.

The investigation delves into the effects of anxiety on assimilation (group-linked products) and differentiation needs (unique products) among consumers with different self-construal types. Indeed, emotions determine whether individuals prefer to assimilate or differentiate from others via consumer choices. Physiological responses to anxiety, measured by the sympathetic nervous system activation, were analyzed to understand how emotions influence consumer choices.

The procedure was the following: after baseline, a linguistic task was provided to prime independent or interdependent self-construal. In the next step, participants watched a movie to elicit anxiety or a neutral condition. Finally, subjects evaluated six pairs of products.

*Dataset composition:*

* 166 observations (83 subjects x 2 conditions: baseline and elicitation).
* 3 labels indicating the different conditions from which the features were extracted. Specifically:
  + Baseline condition = -1
  + Fear = 210
  + Neutral State = 110.
* Features extracted from the RR signal, which is the time elapsed between two successive R waves of the QRS signal on the electrocardiogram, including time and frequency domain features, as well as non-linear features.

*Features – Time Domain:*

* 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.

*Features – Frequency Domain:*

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:*

Non-linear features include features from the 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 🡪  Ratio of *SD1* to *SD2*. Describes the ratio of short term to long term variations in HRV.
* 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.

*Objective:*

The student has to write a script for data mining and analysis, encompassing import, cleaning, exploration, visualization, and normalization. Employ various techniques such as feature selection, principal component analysis, clustering, and classification, to build a robust model capable of predicting different emotional states. These are just some examples of techniques, so don’t be frightened and feel free to select your best strategy.

*Expected Results:*

Students should submit a **script in Python (or their preferred environment) with annotated code**, showcasing their approach to exploring and understanding the dataset. The focus should be on selecting relevant features for building an effective classifier to discern between the emotional states. Additionally, **a technical report describing the project steps, methods, and achieved results should be submitted at least 3 days before the exam**.

This data science project aims to assess students' ability to navigate the dataset, make informed decisions on feature selection, and construct a model for emotional state prediction. Students are encouraged to leverage various techniques to showcase their understanding of data mining and analytics concepts learned throughout the course.