# Machine Learning - SS 2021
# Exercise 0

Federico Ambrogi, e1449911@student.tuwien.ac.at
Adam Höfler, e11847620@student.tuwien.ac.at
Matteo Panzeri 12039996@student.tuwien.ac.at
TU Wien

March 28, 2021

## Contents



## Introduction

In this report we describe and analyse the main features of the datasets *Drug consumption (quantified) Data Set* [1] and *Cuff-Less Blood Pressure Estimation Data Set* [2]. They differ in size of the data set, number of features, type of features, etc. to give us a variety of possible scenarios to work with.

## 1 Data Set: Cuff-Less Blood Pressure Estimation

The original data comes from the Multi-parameter Intelligent Monitoring in Intensive Care II (MIMIC II) online waveform database [3][4]. The data set was extracted from MIMIC II and reduced onto Photoplethysmograph (PPG), Electrocardiogram (ECG) and arterial blood pressure (ABP) waveform signals [5]. The data was collected between the years 2001 and 2008 from a variety of Intensive Care Units and were sampled at rate of 125 Hz with 8 bit accuracy. The preprocessing on the original included the smoothing of the waveforms and the removal blocks with a) unreasonable blood pressures, b) unreasonable heart rate, c) severe discontinuities and d) big differences in the PPG signal correlation between neighbouring blocks.

**Dataset** The data set contains 144,000 rows with 3 attributes each, i.e. the PPG, ECG and ABP. Due to the preprocessing, there are no missing values or outliers. The features are of numeric type. The signals and the distributions of their values, can be inspected in Fig. 1. While the data measured intrinsically positive values, we found that 9179 entries for the ECG have negative values, which can be seen also from the corresponding distribution. We will therefore take care of this anomaly during the data processing.

**Task** This data set is our choice for the regression part with ABP being the target variable. The most common approach to measure (BP) is by using a cuff, wrapped around the upper arm and inflated. A pressure sensor inserted in the cuff records the arterial pulsations during the cuff deflation, and the amplitudes of these pulsations are used to calculate systolic and diastolic BP. The limitations of the approach include that it relies on a set of empirical coefficients to map the pulse amplitudes to systolic and diastolic BP, specific to the used device. Furthermore, this kind of measurements in patients with atherosclerosis or obese patients, whose pulse amplitudes can be weak, are subject to large errors.

For these reasons, it is very interesting to study the possibility to study a regression model able to predict the BP by using the PPG and/or the ECG signals.

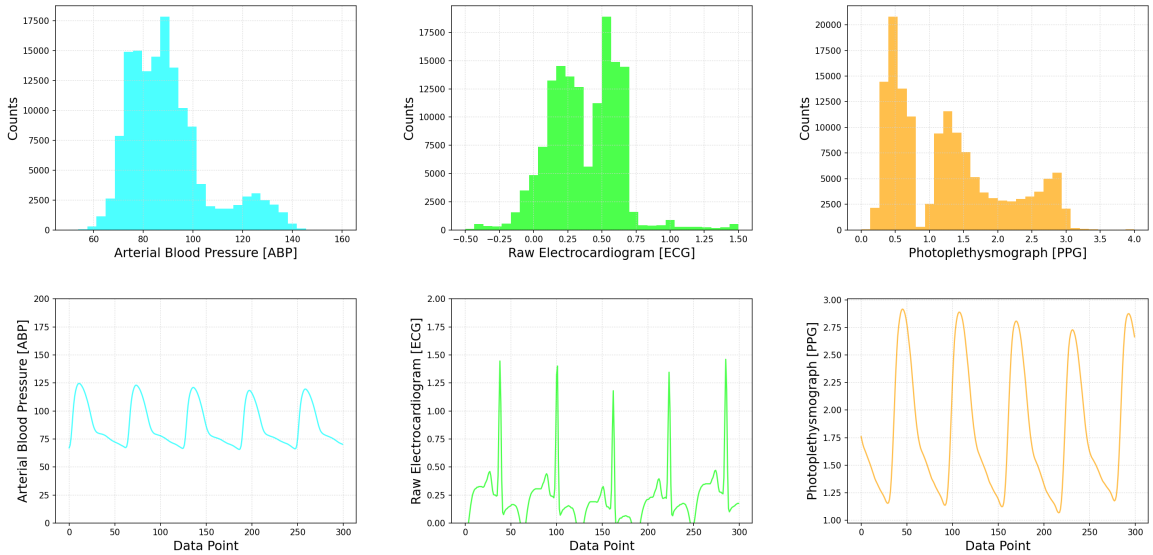


Figure 1: Distribution for the ABP, ECG and PPG signals.

## 2 Dataset: Drug Consumption

This dataset [6] includes records of drug usage of 1885 respondents, together with attributes regarding personal information, collected through an online survey. Data refers to two general categories. The first is related to personal aspects of the respondent and builds the input features: level of education, age, gender, country of residence and ethnicity. Additionally to that, the personality traits of the respondents was classified using the Revised NEO Five-Factor Inventory (NEO-FFI-R) benchmark. Many studies have shown that personality traits such as neuroticism (N), extraversion (E), openness to experience (O), agreeableness (A), and conscientiousness (C) - all quantified by NEO-FFI-R - can be associated to drug consumption[7][8][9]. For example, the personality traits of N,E, and C are highly correlated with dangerous health behaviours. Furthermore, two additional personality traits are considered: the Barratt Impulsiveness Scale version 11 (BIS-11), and the Impulsivity Sensation-Seeking scale (ImpSS).

The second category of data regards the usage of specific drugs acting as central nervous system depressants, stimulants, or hallucinogens: alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers.

For each substance, the respondent had to classify the own usage according to the following frequency categories: "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day".

**Dataset** The dataset contains 1885 rows with each 32 features of which all are of categorical nature - there are no missing values. The first column contains and ID and has now further usage, leaving 31 usable features. The respondents ages were collected in age ranges making them of type nominal instead of ratio. The other non-character related attributes (level of education, gender, country of residence and ethnicity) are also nominal. All of the previously named features were mapped to the real number realm - the mapping was not necessarily monotonic but bijective, keeping the classification intact.

The Revised NEO Five-Factor Inventory (NEO-FFI-R) questionnaire was employed to score the personality traits N, E, O, A and C of the respondents. This ordinal scoring (ranging 20 to 60) was then also mapped to the real number realm. The BIS-11 and ImpSS scores are evaluated in a similar fashion.

The remaining 19 (target) features contain information about the respondents recent consumption of the drugs listed above. The frequency categories of usage were labelled as "CL0"-"CL6".

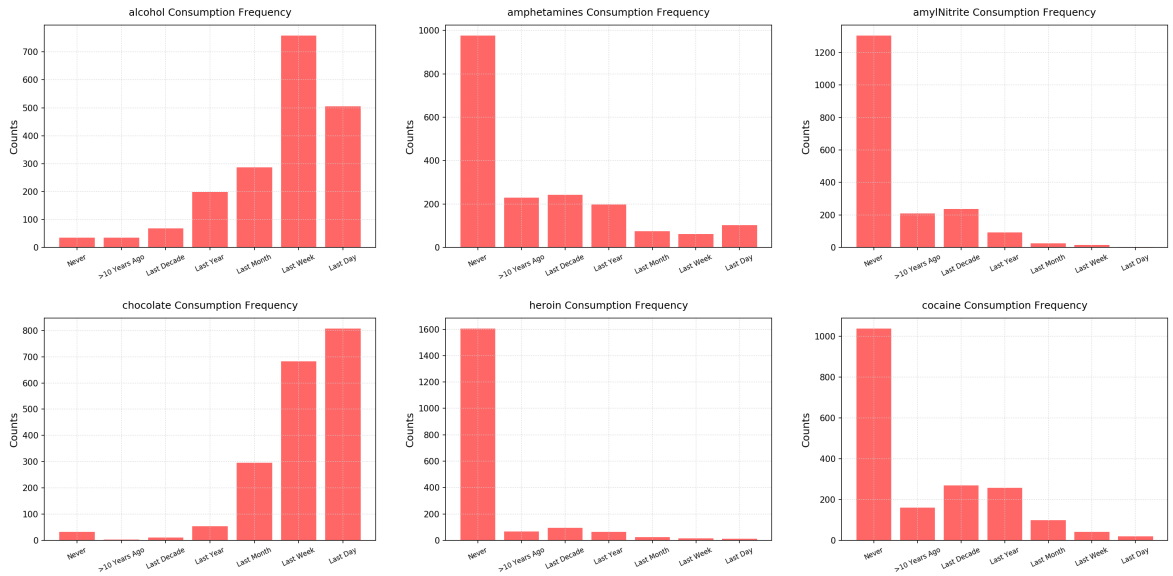


Figure 2: Example of frequency consumption for a selection of drugs.

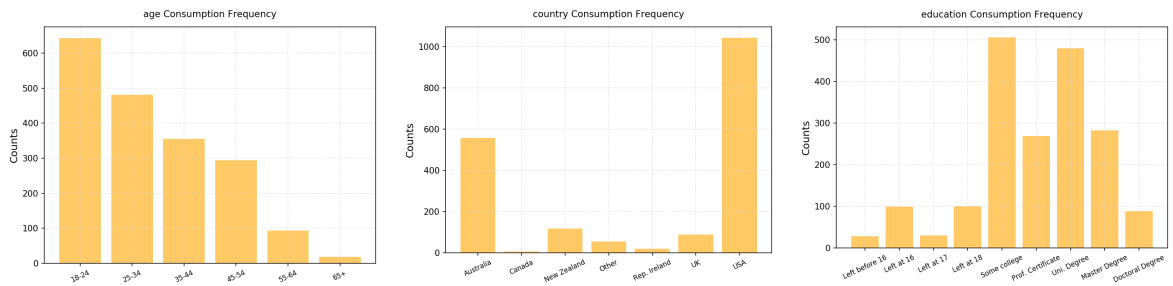


Figure 3: Example distributions of personal characteristics of the respondents: age, country of origin and level of education.

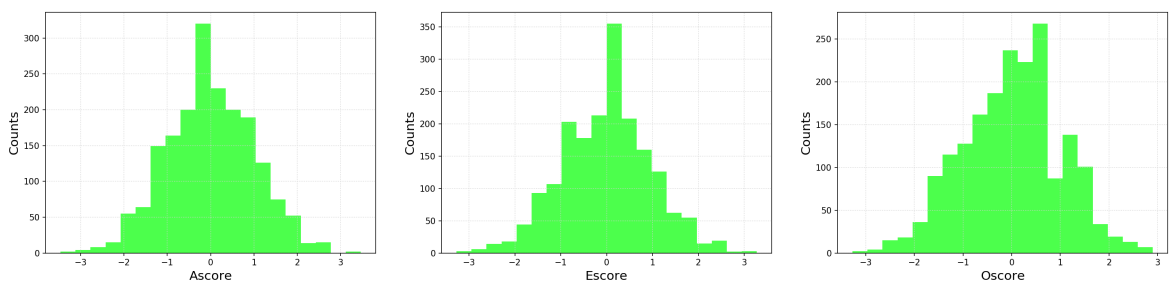


Figure 4: Example of distributions for three personality traits: agreeableness (A), extraversion (E) and openness to experience (O).

**Task** This data set is very appropriate for classification experiments, given the partition of the target features into different classes. Finding a way to predict the consumption behaviour for specific drugs - general correlations between certain character traits and the overall behaviour towards drugs is known[7][8][9] - based on the personality traits and/or ruling out certain traits is desired. The stability related to and effect of unexpected input attributes should be studied to be able to name the scope in which the prediction is reliable.

# References


[1] Evgeny M. Mirkes et al. Uci - drug consumption (quantified) data set. https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29. Accessed: 25.03.2021.

[2] Mohamad Kachuee et al. Uci - cuff-less blood pressure estimation data set. https://archive.ics.uci.edu/ml/datasets/Cuff-Less+Blood+Pressure+Estimation. Accessed: 19.03.2021.

[3] Physionet - multiparameter intelligent monitoring in intensive care (mimic) ii. https://archive.physionet.org/mimic2/. Accessed: 20.03.2021.

[4] A. Goldberger, L. A. Amaral, L. Glass, Jeffrey M. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, and H. Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101 23:E215–20, 2000.

[5] Mohamad Kachuee, Mohammad Mahdi Kiani, Hoda Mohammadzade, and M. Shabany. Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1006–1009, 2015.

[6] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban. The five factor model of personality and evaluation of drug consumption risk. 2017.

[7] C. Roncero, C. Daigre, C. Barral, E. Ros-Cucurull, L. Grau-López, L. Rodríguez-Cintas, N. Tarifa, M. Casas, and S. Valero. Neuroticism associated with cocaine-induced psychosis in cocaine-dependent patients: A cross-sectional observational study. *PLoS ONE*, 9, 2014.

[8] M. Vollrath and S. Torgersen. Who takes health risks? a probe into eight personality types. *Personality and Individual Differences*, 32:1185–1197, 2002.

[9] K. Flory, D. Lynam, R. Milich, C. Leukefeld, and R. Clayton. The relations among personality, symptoms of alcohol and marijuana abuse, and symptoms of comorbid psychopathology: results from a community sample. *Experimental and clinical psychopharmacology*, 10 4:425–34, 2002.