Sarcopenia Dataset
Writting Report: Statistical Analysis

Hanna Gloyna

The used data set contained initially 250 entries, where each entry represents a patient, with 84 features. After removing ID, 17 entries were deleted as they were duplicates. Furthermore 8 misses in *Marcha* (engl. walking speed) were imputed with the mean. The encoding of the categorical data was changed, so that afterwards 0 did always correspond to either *Not answered*, *Not required* or *Regular value*. To obtain comparable values for variance of each feature a min-max-normalisation was applied, given by Equation 1.

$$\hat{x} = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

Where x corresponds to the original value of a feature, $min(x)$ the minimal values of this feature and $max(x)$ to the maximal value, respectively. Afterwards all features are within $[0, 1]$.

In accordance with specialists every representation of MMSE, Barthel, Norton, Lawton, and MNA, except the binary, was removed to reduce the dimensionality of the datset. Additionally MM was dropped, because of its high correlation to IMM.

We dropped further features with too low variance or entropy, i.e. 90% of the values were the same. Additionally 7 features were deleted, because they contained a lot of misses or were included in another feature. Afterwards again all duplicates were dropped.

The remaining 231 entries included 42 serve and 189 mild sarcopenia cases. The dataset consists of 166 women and 65 men between 60 to 97 years old and a median age of 79. By definition a serve sarcopenia cases is a decrease of IMM, grip strength and walking pace. If only a decrease of IMM and grip strength or walking speed is detected, it is classified as a mild case. With this definition 24 of 42 serve cases can be selected

Feature 1	Feature 2	r_s
OA	Visiòn	-0.769
MED8	MED7	0.759
Drogas	Visiòn	-0.703

Table 1: Features with high correlation coefficient

The features with the highest Spearman correlation coefficient r_s are listed in Table 1. The used threshold for a high correlation was $|r_s| > 0,7$ as suggested by Akoglu in [1]. In each row of Table 1 either feature 1 or feature 2 could be removed.

The χ_2

References

[1] Haldun Akoglu. "User's guide to correlation coefficients". In: *Turkish Journal of Emergency Medicine* 18.3 (Sept. 2018), pp. 91–93. ISSN: 24522473. DOI: 10.1016/j.tjem.2018.08.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S2452247318302164 (visited on 03/23/2022).

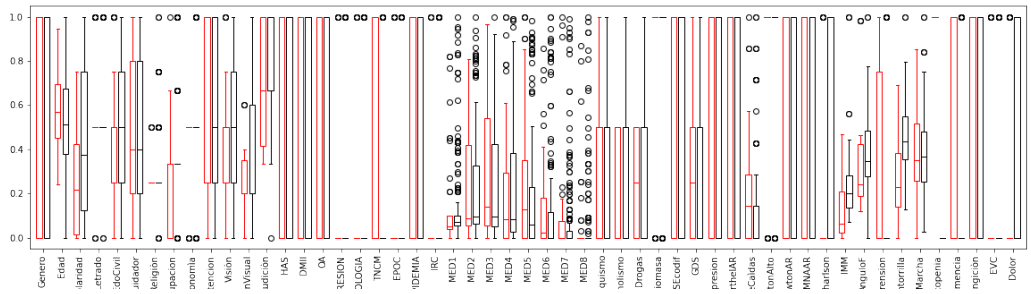

Figure 1: Boxplot of all remaining features, black boxes are from all mild cases and red from all serve, respectively