Sarcopenia Dataset

Writting Report: Statistical Analysis

Hanna Gloyna

1 Statistical Analysis

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 study 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 could be selected, whereas 18 serve cases did not match the definition and 7 cases would be serve by definition, but were not marked as such.

Feature	p	f1
TNCM	0.00	14.97
Demencia	0.00	15.25
FuerzaPrension	0.01	8.96
Pantorrilla	0.10	40.60
IMM	0.14	24.10
LawtonAR	0.17	3.15
Dolor	0.17	2.87
Congiciòn	0.17	4.14
Charlson	0.18	2.37
CorreccionVisual	0.20	9.43

Table 1: Features and their corresponding p-value and f1-value to *sarcopenia*

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

The results of a χ_2 and *ANOVA* analysis are displayed in Table 1. The 10 features with the lowest p-value from χ_2 analysis are almost the same as the ones with the highest f1 score. Only Ocupacion, NortonAlto and Edad (engl. age) achieved a

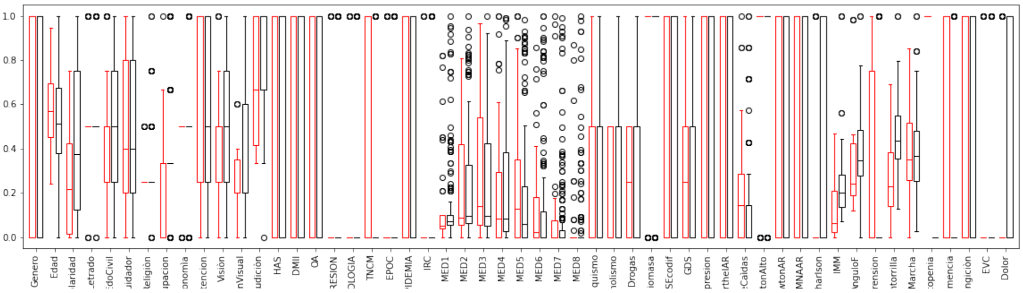

Figure 1: Boxplot of all remaining features, black boxes are created from all mild cases and red from all serve, respectively

Feature 1	Feature 2	r_s
OA	Visiòn	-0.77
MED8	MED7	0.76
Drogas	Visiòn	-0.70

Table 2: Features with high correlation coefficient

higher f1 score, but a had also a higher p-value.

Figure 1 displays the plotted boxes of all remaining 53 features, where black boxes are generated from the subset of mild cases and red boxes from serve cases, respectively. Features with a difference at their mean and inter quartile range are especially interesting and very likely to be important for decision making. As can be seen features such as TNCM, Charlson, IMM, FuerzaPrension, Demencia and Dolor vary a lot for serve and mild cases. This also supports the calculated values and assignment of importance from the calculation of the p- and f1-value.

2 Hyperparameter Tuning

A *Support Vector Machine* (hereinafter: SVM) was used to predict the severity of sarcopenia. A cross validated grid search was applied to get an impression which setting of parameters might be a good fit. Fig. 2 displays the results, where C, γ and 3 different kernels were considered. In every setting the *Radial Basis Function* (rbf) did perform at least as good as the two other kernel. Furthermore it can be clearly seen that $C = 2$ and $\gamma = scale$ are the best setting with the rbf kernel.

Beside the SVM also a *Random Forest* (hereinafter: RF) was trained. Again a cross validated grid search was used. For tuning the depth d of the trees in a set S, amount of trees $|S|$ as well as the splitting criteria were considered. As can be seen in Fig. 3 there is no huge difference between the splitting criteria and a RF performed best with its parameters as follows: $|S| = 75, criteria = gini, d = 20$.

3 Feature Selection

By calculating the p-value and f1 score between each feature and sarcopenia we can get an idea for feature selection. A low p-value indicates that the two features are independent and therefore provide information about each other regarding predictions by statistical models. Table 3 displays the all features for different methods such as χ_2, ANOVA and permuted feature importance for SVM and RF. We calculated the feature importance, f1-value and χ_2-value, respectively. Afterwards the scores were normalised to be comparable and added up to obtain a final ranking which is also listed in Table 3 in the column *Sum*.

Features	χ_2	f1	RF	SVM	Sum
Demencia	0.93	0.38	0.00	0.77	2.08
IMM	0.18	0.59	1.00	0.00	1.77
Pantorrilla	0.22	1.00	0.10	0.11	1.42
TNCM	1.00	0.37	0.00	0.04	1.41
FuerzaPrension	0.61	0.22	0.00	0.50	1.33
CARDIOOLOGIA	0.02	0.01	0.03	1.00	1.06
HAS	0.11	0.08	0.02	0.75	0.96
EPOC	0.07	0.02	0.00	0.86	0.95
Congiciòn	0.15	0.10	0.00	0.66	0.91
MED5	0.13	0.10	0.00	0.63	0.85
LawtonAR	0.15	0.08	0.02	0.55	0.80
Depresion	0.07	0.05	0.02	0.63	0.77
MNAAR	0.04	0.03	0.01	0.60	0.68
CorreccionVisual	0.13	0.23	0.02	0.29	0.67
Charlson	0.14	0.06	0.00	0.46	0.67
Dolor	0.15	0.07	0.01	0.40	0.63
Genero	0.00	0.00	0.02	0.57	0.61
DISLIPIDEMIA	0.01	0.00	0.00	0.56	0.57
MED7	0.01	0.00	0.04	0.52	0.57
BarthelAR	0.00	0.00	0.00	0.56	0.57
NùmeroDeCaìdas	0.09	0.08	0.03	0.34	0.55
Audiciòn	0.01	0.02	0.03	0.49	0.54
Ocupacion	0.12	0.20	0.03	0.19	0.54
Tabaquismo	9.85×10^{-5}	0.00	0.01	0.49	0.50
MMSEcodif	0.00	0.00	0.03	0.47	0.50
EVC	0.12	0.04	0.00	0.33	0.48
IRC	0.06	0.02	0.00	0.39	0.47
EdoCivil	0.05	0.10	0.00	0.31	0.46
ExpBiomasa	0.00	0.00	0.00	0.45	0.45
AnguloF	0.02	0.09	0.09	0.23	0.43
DEPRESION	0.02	0.01	0.00	0.41	0.43
OA	0.00	0.00	0.00	0.42	0.42
DMII	0.05	0.03	0.02	0.31	0.42
Edad	0.03	0.11	0.05	0.21	0.40
Cuidador	0.00	0.00	0.02	0.36	0.39
NortonAlto	0.06	0.19	0.00	0.13	0.38
MED1	6.00×10^{-6}	0.00	0.05	0.31	0.36
GDS	0.02	0.01	0.04	0.27	0.34
MED8	0.00	0.00	0.00	0.33	0.34
Alcoholismo	0.03	0.03	0.02	0.26	0.33
Religiòn	0.00	0.02	0.00	0.31	0.33
Marcha	0.00	2.62×10^{-5}	0.07	0.26	0.33
Letrado	0.00	0.00	0.00	0.32	0.32
Economìa	0.01	0.03	0.00	0.25	0.30
MED4	0.00	0.00	0.02	0.27	0.30
MED6	0.00	0.00	0.00	0.27	0.27
Escolaridad	0.04	0.06	0.03	0.11	0.24
Manutencion	0.06	0.08	0.00	0.09	0.23
Drogas	0.04	0.03	0.02	0.11	0.20
MED2	0.00	0.01	0.02	0.15	0.19
Visiòn	0.01	0.02	0.00	0.09	0.12
MED3	0.03	0.03	0.04	8.33×10^{-16}	0.10

Table 3: Score of features by different methods

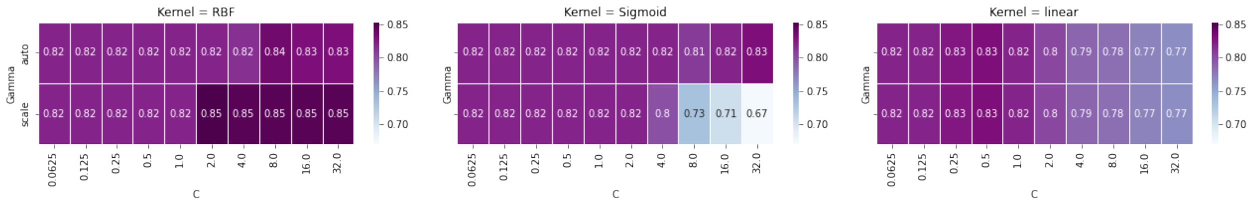

Figure 2: Results of hyperparameter tuning for SVM

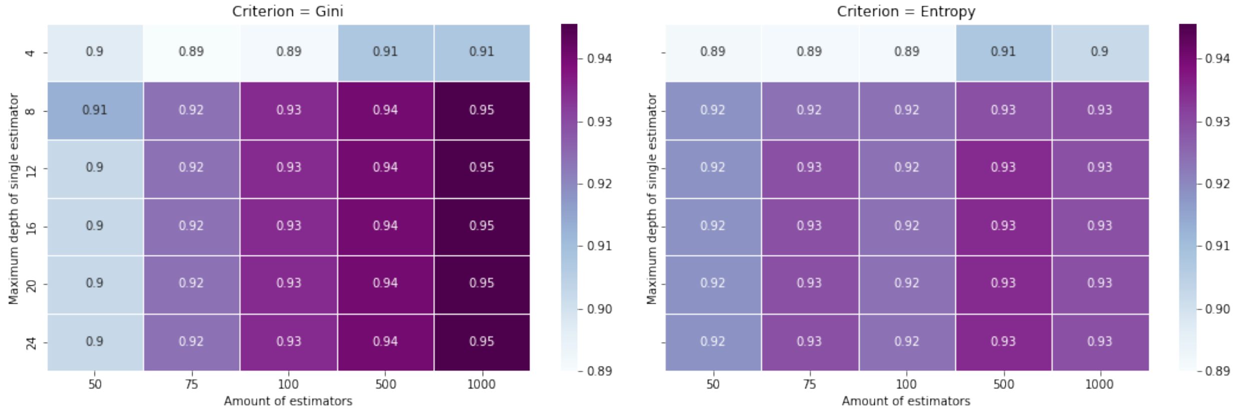

Figure 3: Results of hyperparameter tuning for RF

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).