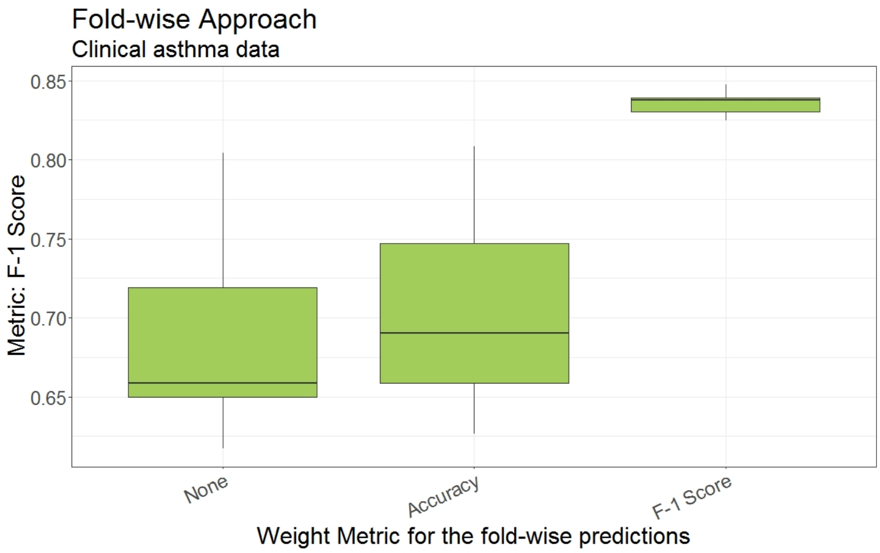

Figure 29: Results of the 'Fold-wise' approach on the clinical asthma data.

4.2.2 Priority-Lasso

This part of the thesis compares the predictive performances of the different priority-Lasso adaptions from Hagenberg [21] on the clinical asthma data. These diverse adaptions can directly deal with block-wise missingness. In the beginning the theoretical principles are explained and then the predictive performance of the different approaches are investigated.

Theoretical principles

The priority-Lasso is an extension of the lasso method. Lasso stands for *least absolute shrinkage and selection operator* and was introduced by Tibshirani in 1996 [48]. The lasso method is a regularised least squares estimator that adds the absolute values of all parameters as an additionally penalty. The loss function is defined as:

$$\|y - \beta_0 \mathbb{1} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \beta_j \tag{13}$$

In the equation 13 the parameter λ is a hyperparameter ≥ 0. The bigger λ, the stronger the added penalty for each $\beta \neq 0$. The additional penalty leads to two major advantages over the regular least squares estimator. Firstly,

the lasso model can be fitted on data with more features than observations. Secondly, the model is far better with variable selection as a regular linear model, as only parameters of the important variables are estimated with a value different from 0.
In 2018 the priority-Lasso was introduced by Klau et al. [15] as an extension of the classical lasso method. The priority-Lasso can be applied to data with different feature-blocks. It is described as a "hierarchical regression method which builds prediction rules for patient outcomes [...] from different blocks of variables including high-throughput molecular data" [[15], p. 2]. Firstly the feature-blocks of the data are ordered in descending order of importance for the target variable. Then a separate lasso model is fitted to the each of these feature-blocks. The models that were fitted on the later blocks (the later, the lower the importance of the block for the target variable) are only used to improve the prediction from the previous blocks. "In order to assure that later blocks only improve the prediction of the previous blocks, the fitted linear predictor of block m is used as the offset when fitting block m + 1" [21]. For more details have a look into the second chapter of Hagenberg's thesis [21] or see Klau et al. [15].
Hagenberg [21] developed different adaptions of the priority-Lasso such that it can directly deal with block-wise missingness. Then - as for the priority-Lasso - a separate model is fitted on each of the feature-blocks of the data. Mainly there are two methods to deal with the missing data.
(1) Ignore: The lasso model for every feature-block is only fitted with the observations that have no missing values for this block. For observations that have missing values in the current feature-block no offset can be calculated for the next block. For observations that miss the first block, the offset for the second feature-block is either set to zero - *PL-ignore, zero* - or to the estimated intercept of the first feature-block - *PL-ignore, intercept*.
(2) Impute: As in the "Ignore" approach, a separate lasso model for each feature-block is only fitted with the observations that have no missing values in this block. For observations with missing values the offset is not carried forward from the previous block, but imputed instead. "This has the advantage that instead of imputing a possibly very high number of covariates, only one value [- the offset -] is imputed" [21]. There are two strategies with this approach. The strategy *max. n* chooses the pattern with the most observations that can be used for the imputation model. The second strategy *max. blocks* chooses the pattern that uses the most high priority feature-blocks for the imputation model.

In total there are four different adaptions of the priority-Lasso - 'ignore, zero', 'ignore, intercept', 'impute, max. n' and 'impute. max. blocks'. All models are fitted in a sequential way and only differ in the way how the offset is carried forward for observations with missing values. Closer information to the theoretical principles of these adaptions can be found in chapter three of [21].

Naive block-order

4.2.3 Comparison of the approaches