Retreat - Team 4 - Summary

Task

Our task was to predict oral food syndrom and dermatitis/treatment of dermatitis based on IgE levels.

Missing Data

Our first preprocessing step involves handling missing values in the covariates. Some of the clinical variables in our dataset contain missing values. To address this, we use a single imputation approach with the `IterativeImputer` function from `sklearn`. This method estimates the conditional distribution of each feature given the others and imputes missing values accordingly. Specifically, we use the default setting, which employs Bayesian Ridge regression for imputation.

Future work could explore more sophisticated imputation techniques, such as using alternative iterative regressors (e.g., Distributional Random Forests) or adopting multiple imputation methods to better capture uncertainty.

Missing Oral Symptoms: Re-weighting. A substantial proportion of entries in our dataset have missing values for the target variable, `Oral Symptom` (e.g., 562 out of 1139 entries are missing for chip 3). Since we cannot use these instances directly for supervised training, we apply a re-weighting strategy to mitigate the bias introduced by missingness.

To achieve this, we define a binary mask variable that takes the value 1 if `Oral Symptom` is missing and 0 otherwise. We then train a logistic regression classifier to predict this mask variable using all available covariates except `Oral Symptom`. The output of this logistic regression model gives an estimated probability p_i that a given sample x_i has a missing target.

We define a weight for each instance as:

$$w_i = \frac{1}{p_i}$$

To avoid extreme values, we clip the weights to a predefined range (e.g., [0.1, 10]). These weights are then used to train our final classifier on the observed instances of `Oral Symptom`. This adjustment helps correct for the bias introduced by only training on observed values and improves the model's performance on both observed and unobserved instances, as reflected in an improved weighted accuracy.

Pre-processing

We performed experiments with several simple preprocessing techniques for the IgE measurements. The main reason for exploring this direction was that some measurements in the datasets have large outliers. We tested the following pre-processing steps: (1) Ignore very sparse features - almost all measurements were 0 for some IgE's. (2) Clipping - set the maximum value to the 90% or 95% largest measurement for each IgE. All larger measurements are set to this value. (3) Normalize - all measurements are mapped to the interval $[0, 1]$. (4) Bucketing - create a bucket from the 0% quantile to the 10% quantile, a bucket from 10% to 20%

etc. (5) Binary - all non-zero measurements are mapped to 1. This is an extreme version of bucketing. (6) We combined techniques (2) and (3) by clipping before normalizing.

We saw some improvements from pre-processing, but other experiments showed slightly lower utility. As such, we did not reach any strong conclusions about the best approach for pre-processing the dataset. However, clipping generally seemed to improve performance.

We believe that there is much more to explore in this direction. The distribution of measurements vary a lot for different IgE's (See Figure 5). It might be better to use multiple different kinds of pre-processing. The technique is then chosen for each IgE based on the distribution instead of performing the same pre-processing for all of them. For example, clipping large values is important when there are huge outliers, but that is not the case for all IgE's.

Classifiers

In order to train a classifier to predict the target variable ´Oral Syndrom´, many methods exist. Therefore, we implemented a pipeline to systematically test different methods and optimize their hyper-parameters. This pipeline leverages randomized search to look for the best hyper-parameters of each classifier. The classifiers implemented are: Logistic Regression, Decision Tree, Random Forest and Gradient Boosting. Note however that any methods working with `sklearn`'s API can be added.

Without further pre-processing (cf. section above), primary results show that one can attain an accuracy of 80% in the prediction of ´Oral Syndrom´ (Figure 1).

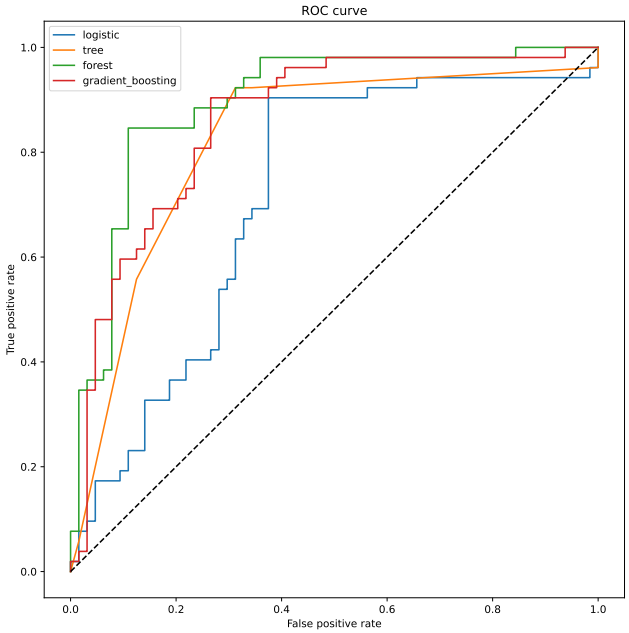

Figure 1: ROC curves for different classifiers.

Feature selection

We ran experiments with different subsets of the features. As a first step, the IgEs were split based on prior knowledge: we selected the ones with the "food" mention in the document. Including clinical data with those relevant IgEs improved the prediction accuracy.

To identify the most relevant variables for predicting oral food syndrom, two data-driven feature importance methods were employed: Random Forest's inherent feature importance and permutation importance. Random Forest importance measures how much each feature reduces impurity in the model trees, while the importance of permutation measures how much model performance degrades when each feature's values are shuffled.

The figure [2] shows the top 100 features, ranked according to their contribution to the model's predictive power, and the figure [3] shows the ranking of features according to the permutation method.

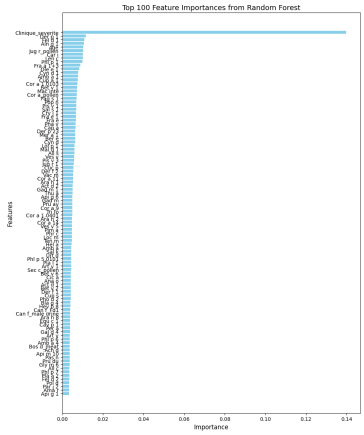

Figure 2: Feature contribution to random forest prediction

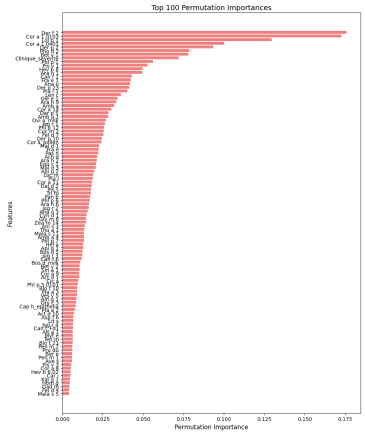

Figure 3: Feature importance using permutation method

Some IgEs identified by both the Random Forest and permutation importance methods are known to be associated with oral allergy symptoms, including: Aln g 1, Bet v 1, Bet v 2, Cor a 1.0103, Cor a 1.0401, Hev b 8, Pho d 2, and Pru du.

Calibration

Beyond classification accuracy, we aim to ensure that our model produces reliable probability estimates. A well-calibrated model provides probabilities that reflect true likelihoods, which is crucial for decision-making.

To assess calibration, we use the **Brier score**, which quantifies both the calibration and sharpness of probability estimates. Additionally, we analyze **calibration curves**, which visually compare predicted probabilities to actual observed frequencies.

To improve calibration, we implement a cross-validation-based calibration technique. Specifically, in a k-fold setting, the model is trained on a subset of the data, and its probability estimates are adjusted using the held-out validation set. The final calibrated model is then an ensemble of these k adjusted estimators. We apply this method using `CalibratedClassifierCV` with both `sigmoid` (Platt scaling) and `isotonic` regression approaches, with the choice of method depending on empirical performance.

Preliminary results suggest that this recalibration might not be enough, with relatively high Brier scores and suboptimal calibration curves. Further investigation is needed to understand the underlying causes and explore alternative calibration techniques.

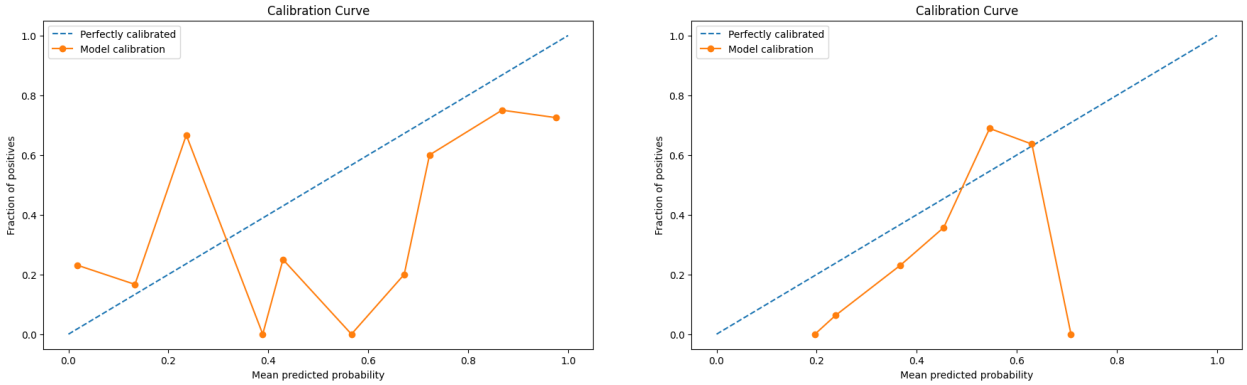

Figure 4: Comparison of calibration curve for before and after applying Cross-validation calibration.

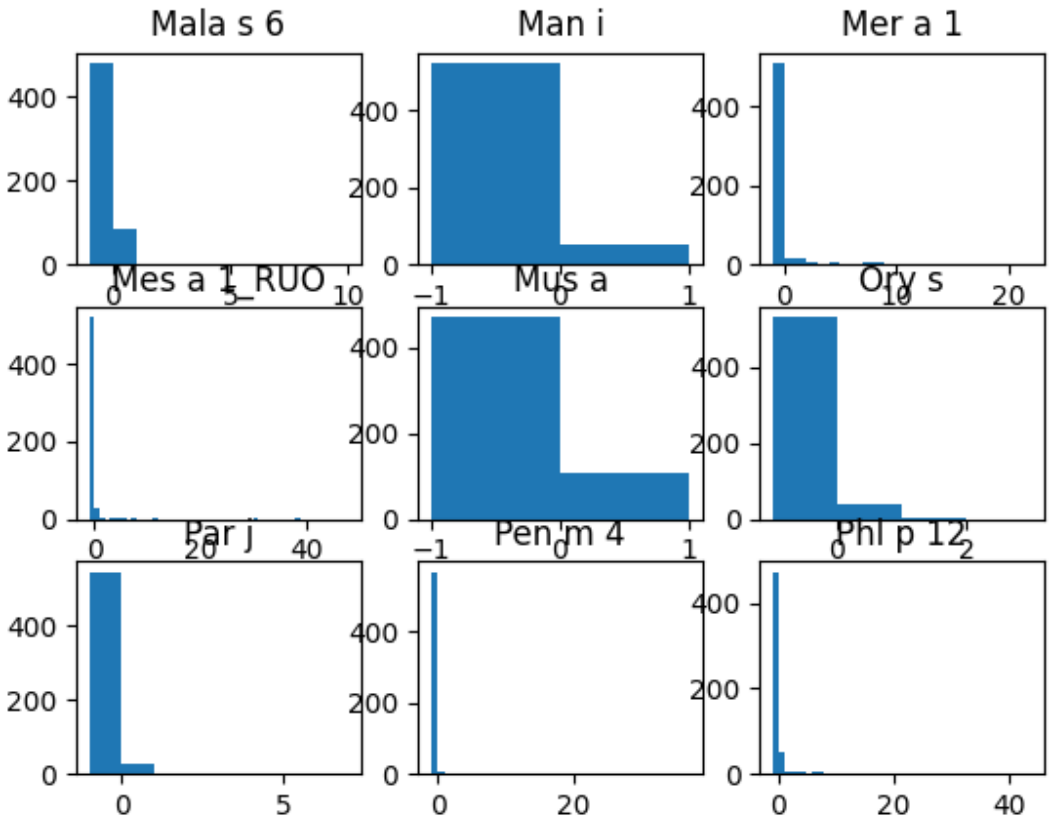

Figure 5: Example of measurements distribution for IgEs