# Background

We have data from 6 clinical trials on whether subjects met ASAS and ASDAS improvement criteria at 12 weeks on the drug arm of each study. 3 of the studies also provide ASAS and ASDAS outcomes at 24 weeks. We also have data on several baseline metrics including swollen and tender joints, BASDAI and BASFI criteria, CRP levels, disease duration and use of some DMARDs, alongside some baseline demographic and clinical information.

The objective of thiss study is to identify some baseline variables that in combination can be predictive of response.

# Descriptiion

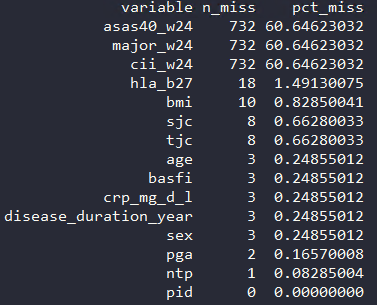
We have data from 1207 patients from 6 clinical studies. Ages ranged from 18 to 75 years (mean of 39 years), with 75% of subjects being males.

[I'll add to this]

# Modeling

To prepare for modeling, we transformed the data in the following ways:

1. Removed features that had very low variance or zero variance across subjects (ank, ibd, psoriasis, tnf)
2. log-transformed the CRP values given the long-tailed nature of its distribution
3. Imputed the small number of missing values using k-nearest neighbors (less than 1% missing on 10 predictors, 1.5% missing HLA B27 data)



1. Standardized all the continuous predictors

We ended up with 22 predictors, namely, the individual BASDAI scores, BASFI, BMI, CRP, disease duration, MTX use, NTP, PGA, SJC, SSZ use, Steroid use, TBP, TJC, HLA B27 status, gender and uveitis.

We investigated several machine learning models in this enterprise, including

1. Decision trees and random forests
2. Nearest neighbors
3. Gradient boosted trees
4. Linear discriminant analysis (LDA)
5. Quadratic discriminant analysis
6. Support vector machines.
7. A one-layer neural network

We evaluated three outcomes, namely ASAS40 and both major and clinically important improvements based on ASDAS criteria. We concentrated on the 12-week outcomes since we had the most subjects who provided outcomes at that time point, and, for each response, coded it as a binary outcome. We evaluated the models based on accuracy and area under the ROC curve after 5-fold cross-validation, to get reasonable estimates of the error rates for new data sets.

From this exercise we found that the ASAS40 was hard to predict, with accuracy getting to 63% at best, and AUROC getting to 67%. For the ASDAS measures, however, we easily reached 75% or greater accuracy and 80% or higher AUROC. Recursive Feature Elimination using 5-fold cross-validation also showed that we could predict ASDAS outcomes well with as few as 5 predictors, while it would take all 22 predictors to achieve the higher levels of accuracy for ASAS40.

# Final model

We chose to use linear discriminant analysis as our final model. We believe this model performed well given that we have a relatively small dataset where more complicated ML models have a hard time determining an optimal model. It also topped our accuracy and AUROC metrics for both ASDAS metrics.

We can see the top predictors based on permutation importance for the major improvement and the CII below, using LDA

|  |  |
| --- | --- |
| Major improvement | CII |
|  |  |

Our final models after recursive feature elimination chose 5 features that were predictive of the ASDAS outcomes. These were

|  |  |
| --- | --- |
| Major | CII |
| CRP | CRP |
| PGA | Age |
| BMI | PGA |
| TJC | BASDAI 2 |
| BASFI | BASDAI 3 |

Each model achieved over 75% accuracy and over 78% AUROC based on 5-fold cross-validation. If you consider AUROC as the metric of choice, the optimal model for CII uses the same predictors as the model for Major improvement.

## A note about LDA

All the predictors do not necessarily meet the requirements of LDA, since some predictors are binary. However, the LDA gives very good predictive performance, and the final models only have continuous predictors. There is also some evidence in the literature of robustness of the LDA models even when the predictors aren’t exactly normally distributed.

Here is what Hastie et al. have to say about it (in context of two-class LDA) in The Elements of Statistical Learning, section 4.3:

Since this derivation of the LDA direction via least squares does not use a Gaussian assumption for the features, its applicability extends beyond the realm of Gaussian data. However the derivation of the particular intercept or cut-point given in (4.11) does require Gaussian data. Thus it makes sense to instead choose the cut-point that empirically minimizes training error for a given dataset. This is something we have found to work well in practice, but have not seen it mentioned in the literature.