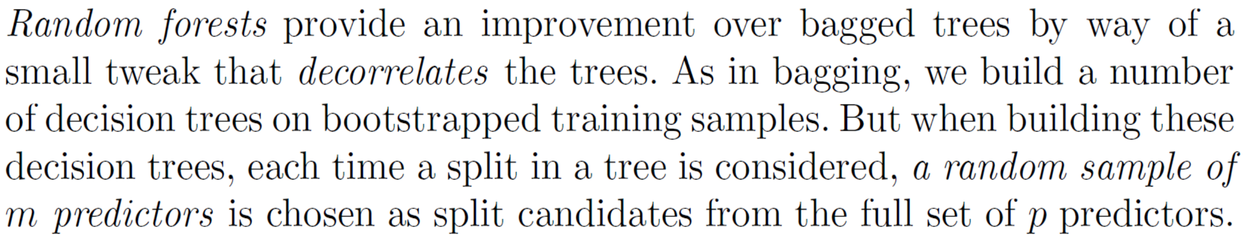
# Response surface methodology motivated from a random forest algorithm

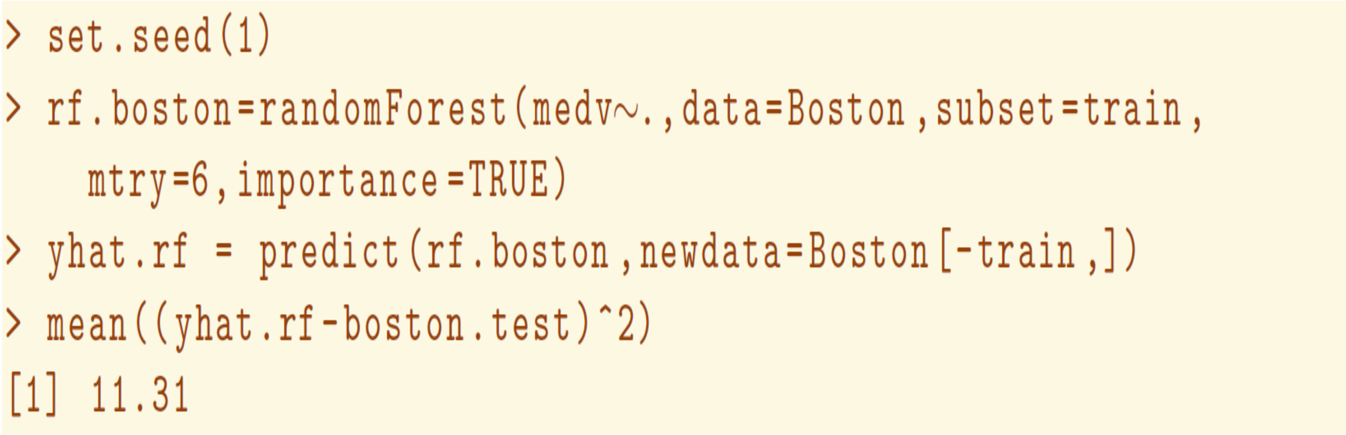
## Random Forest



*Random forests* provide an improvement over bagged trees by way of a small tweak that *decorrelates* the trees. As in bagging, we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, *a random sample of $m$ predictors* is chosen as split candidates from the full set of $p$ predictors.

## Tuning machine learning hyperparameters

### Random Forest



```
> set.seed(1)
> rf.boston=randomForest(medv~.,data=Boston,subset=train,
    mtry=6,importance=TRUE)
> yhat.rf = predict(rf.boston,newdata=Boston[-train,])
> mean((yhat.rf-boston.test)^2)
[1] 11.31
```

```
randomForest(x, y=NULL, xtest=NULL, ytest=NULL, ntree=500,
      mtry=if (!is.null(y) && !is.factor(y)) max(floor(ncol(x)/3), 1) else
      floor(sqrt(ncol(x))), replace=TRUE, classwt=NULL, cutoff,
      strata, sampsize = if (replace) nrow(x) else
      ceiling(.632*nrow(x)), nodesize = if (!is.null(y) && !is.factor(y))
      5 else 1, maxnodes = NULL, importance=FALSE,
      localImp=FALSE, nPerm=1, proximity, oob.prox=proximity,
      norm.votes=TRUE, do.trace=FALSE, keep.forest=!is.null(y) &&
      is.null(xtest), corr.bias=FALSE, keep.inbag=FALSE, ...)
```

**Hyperparameters**

| | |
|---|---|
| mtree | Number of trees to grow |
| mtry | Number of variables to use at each split |
| replace | Sampling with or without replacement |
| nodesize | Minimum number of instances in each terminal node |
| classwt | Prior probabilities for each of the classes |
| cutoff | Threshold for binary classification |
| maxnodes | Maximum number of terminal nodes |

| Factor | Low level (-1) | High level (+1) |
|---|---|---|
| A: mtree | 100 | 500 |
| B: mtry | 2 | 4 |
| C: replace | FALSE | TRUE |
| D: nodesize | 1 | 3250 |
| E: classwt | 1 | 10 |
| F: cutoff | 0.2 | 0.8 |
| G: maxnodes | 5 | NULL |

A $2_{IV}^{7-2}$ design. Generators: F = ABCD; G = ABDE

**I=ABCDF=ABDEG=CEFG**

| Row | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | -1 | -1 | -1 | -1 | -1 | 1 | 1 |
| 2 | 1 | -1 | -1 | -1 | -1 | -1 | -1 |
| 3 | -1 | 1 | -1 | -1 | -1 | -1 | -1 |
| 4 | 1 | 1 | -1 | -1 | -1 | 1 | 1 |
| 5 | -1 | -1 | 1 | -1 | -1 | -1 | 1 |
| 6 | 1 | -1 | 1 | -1 | -1 | 1 | -1 |
| 7 | -1 | 1 | 1 | -1 | -1 | 1 | -1 |
| 8 | 1 | 1 | 1 | -1 | -1 | -1 | 1 |
| 9 | -1 | -1 | -1 | 1 | -1 | -1 | -1 |
| 10 | 1 | -1 | -1 | 1 | -1 | 1 | 1 |
| 11 | -1 | 1 | -1 | 1 | -1 | 1 | 1 |
| 12 | 1 | 1 | -1 | 1 | -1 | -1 | -1 |
| 13 | -1 | -1 | 1 | 1 | -1 | 1 | -1 |
| 14 | 1 | -1 | 1 | 1 | -1 | -1 | 1 |
| 15 | -1 | 1 | 1 | 1 | -1 | -1 | 1 |
| 16 | 1 | 1 | 1 | 1 | -1 | 1 | -1 |
| 17 | -1 | -1 | -1 | -1 | 1 | 1 | -1 |
| 18 | 1 | -1 | -1 | -1 | 1 | -1 | 1 |
| 19 | -1 | 1 | -1 | -1 | 1 | -1 | 1 |
| 20 | 1 | 1 | -1 | -1 | 1 | 1 | -1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 21 | -1 | -1 | 1 | -1 | 1 | -1 | -1 |
| 22 | 1 | -1 | 1 | -1 | 1 | 1 | 1 |
| 23 | -1 | 1 | 1 | -1 | 1 | 1 | 1 |
| 24 | 1 | 1 | 1 | -1 | 1 | -1 | -1 |
| 25 | -1 | -1 | -1 | 1 | 1 | -1 | 1 |
| 26 | 1 | -1 | -1 | 1 | 1 | 1 | -1 |
| 27 | -1 | 1 | -1 | 1 | 1 | 1 | -1 |
| 28 | 1 | 1 | -1 | 1 | 1 | -1 | 1 |
| 29 | -1 | -1 | 1 | 1 | 1 | 1 | 1 |
| 30 | 1 | -1 | 1 | 1 | 1 | -1 | -1 |
| 31 | -1 | 1 | 1 | 1 | 1 | -1 | -1 |
| 32 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

The effects (2x coefficients) are obtained from a linear model based estimation procedure.

## Steepest ascent

In case of of only main effects, new experimentations could be perfomed along the gradient from the center of the design until no improvement.

For example

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = f(x)$$

$$\nabla f(x) = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

So $x_1 = 1,\ x_2 = \dfrac{\hat{\beta}_2}{\hat{\beta}_1}$ and $x_3 = \dfrac{\hat{\beta}_3}{\hat{\beta}_1}$ is a potential direction for improvement or a path of steepest ascent. For minimizing, the path of improvement is along the negative gradient (steepest descent).

**The usefulness of center runs**

Center runs may provide us with a model independent estimate of error (pure error) and a test for lack of fit.

Suppose the design has $i = 1, 2, \ldots m$ distinct points, each replicated $r_i$ times.

At each of these points the residuals are given by

$$y_{ik} - \hat{y}_i = \left(y_{ik} - \overline{y}_i\right) - \left(\hat{y}_i - \overline{y}_i\right),\ i=1,\ 2,\ \ldots, m,\ k = 1, 2, \ldots, r_i .$$

Thereby the sum of squares for residuals is given by:

$$SS_R = \sum_{i=1}^{m}\sum_{k=1}^{r_i}\left(y_{ik} - \hat{y}_i\right)^2 = \sum_{i=1}^{m}\sum_{k=1}^{r_i}\left(y_{ik} - \overline{y}_i\right)^2 + \sum_{i=1}^{m} r_i\left(\hat{y}_i - \overline{y}_i\right)^2$$

or $SS_R = SS_{PE} + SS_{LOF}$.

If $p$ terms are fitted to the data, $SS_R$ has $n-p$ degrees of freedom where $n$ is the total number of runs.

Further $SS_{PE}$ has $\sum_{i=1}^{m}\left(r_i - 1\right) = n - m$ degrees of freedom which leaves $m-p$ degrees of freedom for $SS_{LOF}$.

Thereby a test for lack of fit can be performed using the F-observator: $F_{(m-p),(n-m)} = \dfrac{SS_{LOF} \,/\left(m-p\right)}{SS_{PE} \,/\left(n-m\right)}$.

If the test leads to rejection, we conclude there is a significant lack of fit.

Typically a lack of fit test is performed to check for curvature, or if there should be second order terms in the model. When a two-level screening design is performed and analysed, possibly with several iterations, it may happen that some interactions are active. Thereby a gradient based steepest ascent/descent may not give the best direction to improve the function value. In this situation, and whenever the F-test indicates lack of fit, we normally augment the design to be able to estimate second order terms. The most commonly used designs are the Central Composite Designs (CCD) and the Box-Behnken Designs (BBD).

## Standard procedure

Perform a screening experiment possibly with center runs added.

If only main effects are active, proceed with experimentation along the gradient (steepest ascent) or along the opposite direction (steepest descent).

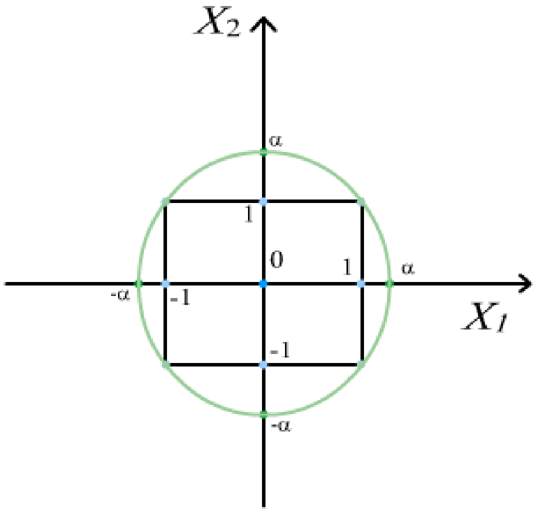
Otherwise, perfom a Central Composite design cube+ center + star runs or a Box-Behnken design to estimate quadratic effects.

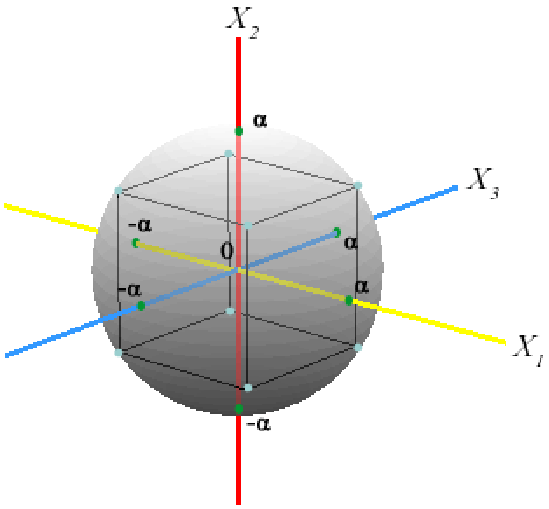
Find the stationary point.

If that is far away from the center, use canonical analysis to find a new direction for improvement.

Iterate.

## Central Composite Designs





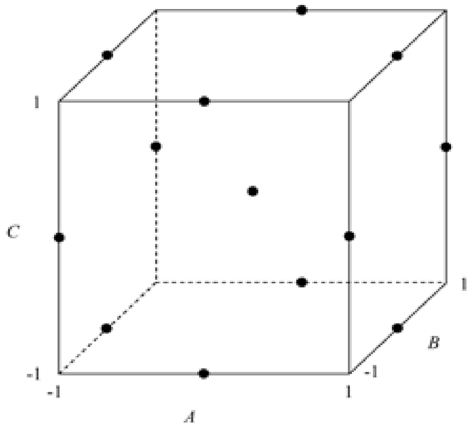
In a CCD we add two extra runs on the axis. Two choices of $\alpha$ are common:

$\alpha = \sqrt{k}$ where $k$ is the number of factors or $\alpha = \left(n_f\right)^{\frac{1}{4}}$. Here $n_f$ is the number of factorial runs.

A **CCD** with three factors, 14+ $n_C$ runs

| Runs | Factors | | |
|---|---|---|---|
| | **A** | **B** | **C** |
| 1 | -1 | -1 | -1 |
| 2 | 1 | -1 | -1 |
| 3 | -1 | 1 | -1 |
| 4 | 1 | 1 | -1 |
| 5 | -1 | -1 | 1 |
| 6 | 1 | -1 | 1 |
| 7 | -1 | 1 | 1 |
| 8 | 1 | 1 | 1 |
| 9 | $\alpha$ | 0 | 0 |
| 10 | $-\alpha$ | 0 | 0 |
| 11 | 0 | $\alpha$ | 0 |
| 12 | 0 | $-\alpha$ | 0 |
| 13 | 0 | 0 | $\alpha$ |
| 14 | 0 | 0 | $-\alpha$ |
| | **0** | **0** | **0** |

**Box-Behnken Designs** have all their runs on the cube



| Runs | Factors | | |
|---|---|---|---|
| | A | B | C |
| 1 | -1 | -1 | 0 |
| 2 | -1 | 1 | 0 |
| 3 | 1 | -1 | 0 |
| 4 | 1 | 1 | 0 |
| 5 | -1 | 0 | -1 |
| 6 | -1 | 0 | 1 |
| 7 | 1 | 0 | -1 |
| 8 | 1 | 0 | 1 |
| 9 | 0 | -1 | -1 |
| 10 | 0 | -1 | 1 |
| 11 | 0 | 1 | -1 |
| 12 | 0 | 1 | 1 |
| 13 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 |

## Box-Behnken Designs In three, four and five factors

Three factors, 12+ $n_C$ runs

$$\begin{bmatrix} \pm 1 & \pm 1 & \mathbf{0} \\ \pm 1 & \mathbf{0} & \pm 1 \\ \mathbf{0} & \pm 1 & \pm 1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Four factors, 24+$n_C$ runs

$$\begin{bmatrix} \pm 1 & \pm 1 & \mathbf{0} & \mathbf{0} \\ \pm 1 & \mathbf{0} & \pm 1 & \mathbf{0} \\ \pm 1 & \mathbf{0} & \mathbf{0} & \pm 1 \\ \mathbf{0} & \pm 1 & \pm 1 & \mathbf{0} \\ \mathbf{0} & \pm 1 & \mathbf{0} & \pm 1 \\ \mathbf{0} & \mathbf{0} & \pm 1 & \pm 1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Five factors, 40+$n_C$ runs

$$\begin{bmatrix} \pm1 & \pm1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \pm1 & \mathbf{0} & \pm1 & \mathbf{0} & \mathbf{0} \\ \pm1 & \mathbf{0} & \mathbf{0} & \pm1 & \mathbf{0} \\ \pm1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \pm1 \\ \mathbf{0} & \pm1 & \pm1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \pm1 & \mathbf{0} & \pm1 & \mathbf{0} \\ \mathbf{0} & \pm1 & \mathbf{0} & \mathbf{0} & \pm1 \\ \mathbf{0} & \mathbf{0} & \pm1 & \pm1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \pm1 & \mathbf{0} & \pm1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \pm1 & \pm1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

## Analysis of response surfaces. Canonical analysis.

A fitted second order model can be written on the form:

$$\hat{y} = b_0 + \sum_{j=1}^{k} b_j x_j + \sum_{i \geq j} b_{ij} x_i x_j \, .$$

If we write

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_k \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_k \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} b_{11} & 0.5b_{12} & \cdots & 0.5b_{1k} \\ 0.5b_{12} & b_{22} & \cdots & 0.5b_{2k} \\ \cdots & \cdots & & \cdots \\ 0.5b_{1k} & 0.5b_{2k} & \cdots & b_{kk} \end{bmatrix},$$

we get

$$\hat{y} = b_0 + \mathbf{x}^{\mathrm{t}}\mathbf{b} + \mathbf{x}^{\mathrm{t}}\mathbf{B}\mathbf{x} \, .$$

Let $\mathbf{m}_i , i = 1,2,\cdots,k$ be the orthonormal eigenvectors.

Then $\mathbf{B}\mathbf{m}_i = \mathbf{m}_i \lambda_i , \ i = 1,2,\ldots,k$ and for the k equations we can write:

$\mathbf{B}\mathbf{M} = \mathbf{M}\mathbf{\Lambda}$. Here $\mathbf{\Lambda}$ is a diagonal matrix with the eigenvalues on the diagonal.

Since $\mathbf{M}^{\mathrm{t}}\mathbf{M} = \mathbf{I}$, we have $\mathbf{M}^{\mathrm{t}}\mathbf{B}\mathbf{M} = \mathbf{\Lambda}$ and

$$\hat{y} = b_0 + \mathbf{x}^{\mathrm{t}}\mathbf{M}\mathbf{M}^{\mathrm{t}}\mathbf{b} + \mathbf{x}^{\mathrm{t}}\mathbf{M}\mathbf{M}^{\mathrm{t}}\mathbf{B}\mathbf{M}\mathbf{M}^{\mathrm{t}}\mathbf{x},$$

and writing $\mathbf{X}=\mathbf{M}^{\mathrm{t}}\mathbf{x}$, the vector of principal components, and $\boldsymbol{\theta}=\mathbf{M}^{\mathrm{t}}\mathbf{b}$ we get

$\hat{y}=b_0+\mathbf{X}^{\mathrm{t}}\boldsymbol{\theta}+\mathbf{X}^{\mathrm{t}}\boldsymbol{\Lambda}\mathbf{X}$ or

$$\hat{y}=b_0+\theta_1 X_1+\cdots+\theta_k X_k+\lambda_1 X_1^2+\cdots+\lambda_k X_k^2.$$

Equating the derivatives to zero we get the stationary points in the new coordinates as

$$\mathbf{X}_{\mathrm{s}}=-\frac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\theta}.$$

Introducing $\tilde{\mathbf{X}}=\mathbf{X}-\mathbf{X}_S$ and $\hat{y}_S=b_0+\frac{1}{2}\mathbf{X}_S^{\mathrm{t}}\boldsymbol{\theta}$ we get

$\hat{y}=\hat{y}_S+\tilde{\mathbf{X}}^{\mathrm{t}}\boldsymbol{\Lambda}\tilde{\mathbf{X}}$ or

$$\hat{y}=\hat{y}_S+\sum_{i=1}^{k}\lambda_i\tilde{X}_i^2.$$

If all eigenvalues are negative, $\mathbf{X}_{\mathrm{s}}=-\frac{1}{2}\boldsymbol{\Lambda}^{-1}\boldsymbol{\theta}$ is a maximum.

If all eigenvalues are positive, it is a minimum.

If some are positive and some negative, we have a saddle point.

**Example**

At some part of an investigation the experimenter ended up with the following model in $x_1, x_2$ and $\mathrm{x}_3$ after a Box Behnken design had been carried out.

$$\hat{y} = 0.75 - 0.0098x_1 - 0.0035x_2 + 0.025x_3 - 0.0015x_1x_2 + 0.0021x_1x_3$$
$$-0.0021x_2x_3 + 0.0007x_1^2 + 0.0015x_2^2 - 0.0007x_3^2$$

This gives the following $\mathbf{B}$ matrix:

$$\mathbf{B} = \begin{bmatrix} 0{,}0007 & -0{,}0008 & 0{,}0011 \\ -0{,}0008 & 0{,}0015 & -0{,}0011 \\ 0{,}0011 & -0{,}0011 & -0{,}0007 \end{bmatrix}$$

with corresponding eigenvectors

$$\begin{bmatrix} -0{,}528460 & -0{,}373265 & 0{,}762498 \\ 0{,}741701 & 0{,}233992 & 0{,}628592 \\ -0{,}413050 & 0{,}897731 & 0{,}153195 \end{bmatrix}$$

and eigenvalues

$$\begin{bmatrix} 0,0026826 \\ -0,0014441 \\ 0,0002615 \end{bmatrix}.$$

With $\boldsymbol{\theta} = \mathbf{M}^{\mathrm{t}}\mathbf{b} = \begin{bmatrix} -0,0077433 \\ 0,0252823 \\ -0,0058427 \end{bmatrix}$,

we get the canonical form

$$\hat{y} = 0.75 - 0.008X_1 + 0.025X_2 - 0.006X_3 + 0.003X_1^2 - 0.001X_2^2 + 0.0003X_3^2$$

.

The stationary point is located in $\mathbf{X}_{\mathrm{s}} = \begin{bmatrix} 1.44 \\ 8.75 \\ 11.17 \end{bmatrix}$ or

$\mathbf{x}_S = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b} = \begin{bmatrix} 4.48 \\ 10.14 \\ 8.97 \end{bmatrix}$ which is pretty far from the center of the design.

Interactions and quadratic effects are involved. To get a direction for improvement, it is often useful to consider the first order terms or the $\boldsymbol{\theta}$ vector. This means we should try to decrease $X_1$, increase $X_2$ and decrease $X_3$. In the rotated coordinate system $\boldsymbol{\theta}$ is roughly pointing in the direction of the vector $\begin{bmatrix} -1 \\ 3 \\ -0.8 \end{bmatrix}$ which gives a direction of improvement using the

$\mathbf{X} = \mathbf{M}^{\mathrm{t}}\mathbf{x}$ vector. Thereby in the original variables we get a possible direction of improvement given by $\mathbf{x} = \mathbf{MX}$ =

$$\begin{bmatrix} -1.2 \\ -0.54 \\ 2.98 \end{bmatrix}.$$

Using balanced accuracy, BACC=$\frac{1}{2}\left[\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\right)+\left(\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right)\right]$, an improvement from 0.64 to 0.81 was obtained.

**Literature**

Box, G. E. P. and Draper, N. R. (2007): Response Surfaces, Mixtures, and Ridge Regression. Wiley.

Myers, R. H., Montgomery, D. C. and Andersen-Cook, C. M. (2016) Response Surface Methodology – Process and Product Optimization Using Designed Experiments. Wiley.