BREAST CANCER DIAGNOSIS

Presentation by:

SAMUEL DARTEH
Data Science MS

Group 4
Data Science MS

Introduction | Recap

- The goal of the study is to **create a model to predict the possible diagnosis of breast cancer**—whether **benign** (healthy) or **malignant** (cancerous).
- **Data source:** Breast Cancer Wisconsin (Diagnostic) dataset University of California (Irvine): http://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
  - Features were computed from a digitized image of a fine needle aspirate (FNA) medical procedure of a breast mass, and **relevant features** selected.
- Breast cancer is not only the **commonest cancer** amongst women; it's now the most diagnosed cancer worldwide and the **second deadliest** second to lung cancer. (https://www.cancer.net/)
- **Early detection** is key to the fight against breast cancer.

Data Structure

- **Sample size: 569** observations
- **30 predictors**

  — all variables were continuous.

  — no categorical variable.

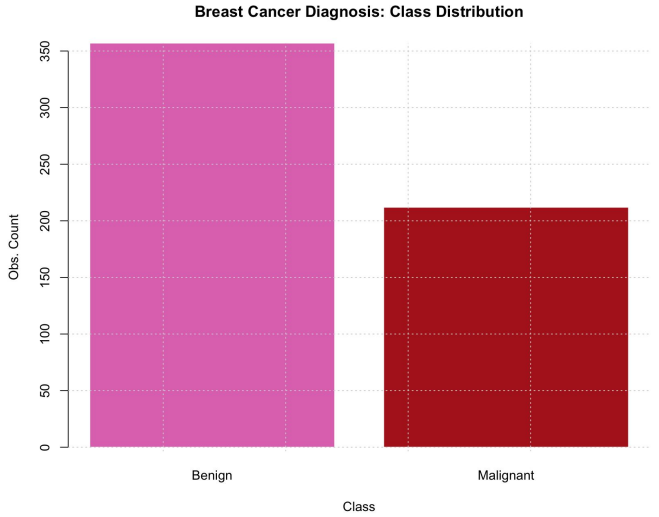
  - **Mean, standard error (se), and worst values** computed for 10 selected features.

- **1 response variable** (Diagnosis)
  - ✔ Categorical — **Classification problem**
  - ✔ Encoded as factor with 2 levels: B, M
  - ✔ Unbalanced distribution

    Benign (B): 357 (63%),

    Malignant (M): 212 (37%)

```
'data.frame':   569 obs. of  32 variables:
 $ id                     : int  842302 842517 84300903 84348301 84358402 843786 844359 84458202 844981 8450
 $ diagnosis              : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
 $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
 $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
 $ area_mean              : num  1001 1326 1203 386 1297 ...
 $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
 $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
 $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
 $ concave_points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
 $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
 $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
 $ radius_error           : num  1.095 0.543 0.746 0.496 0.757 ...
 $ texture_error          : num  0.905 0.734 0.787 1.156 0.781 ...
 $ perimeter_error        : num  8.59 3.4 4.58 3.44 5.44 ...
 $ area_error             : num  153.4 74.1 94 27.2 94.4 ...
 $ smoothness_error       : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
 $ compactness_error      : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
 $ concavity_error        : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
 $ concave_points_error   : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
 $ symmetry_error         : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
 $ fractal_dimension_error: num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
 $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
 $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
 $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
 $ area_worst             : num  2019 1956 1709 568 1575 ...
 $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
 $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
 $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
 $ concave_points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
 $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
 $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
```



Data Structure

10 Features selected – each with "mean, error, worst" values

Predictor	Description
id	Identifier for each sample
diagnosis	Response—Benign (B) or Malignant (M)
radius	Mean distances from center to points on the perimeter
texture	Standard deviation of gray-scale values
perimeter	Perimeter of the tumor cells
area	Area of the tumor cells
smoothness	Smoothness of the tumor cells
compactness	Local variation in radius lengths
concavity	Perimeter^2 / area - 1.0
concave points	Severity of concave portions of the contour
symmetry	Number of concave portions of the contour
fractal dimension	Coastline approximation- 1

"radius_mean"	"texture_mean"	"perimeter_mean"	"area_mean"	"smoothness_mean"
"compactness_mean"	"concavity_mean"	"concave_points_mean"	"symmetry_mean"	"fractal_dimension_mean"
"radius_error"	"texture_error"	"perimeter_error"	"area_error"	"smoothness_error"
"compactness_error"	"concavity_error"	"concave_points_error"	"symmetry_error"	"fractal_dimension_error"
"radius_worst"	"texture_worst"	"perimeter_worst"	"area_worst"	"smoothness_worst"
"compactness_worst"	"concavity_worst"	"concave_points_worst"	"symmetry_worst"	"fractal_dimension_worst"

Data Anomalies

- No **duplicates** found in data

```
> # duplicates
> dup_cols = sum(duplicated(bre_can_dgn)==TRUE)
> dup_cols
[1] 0
```

- **Negative values** in data: **None**

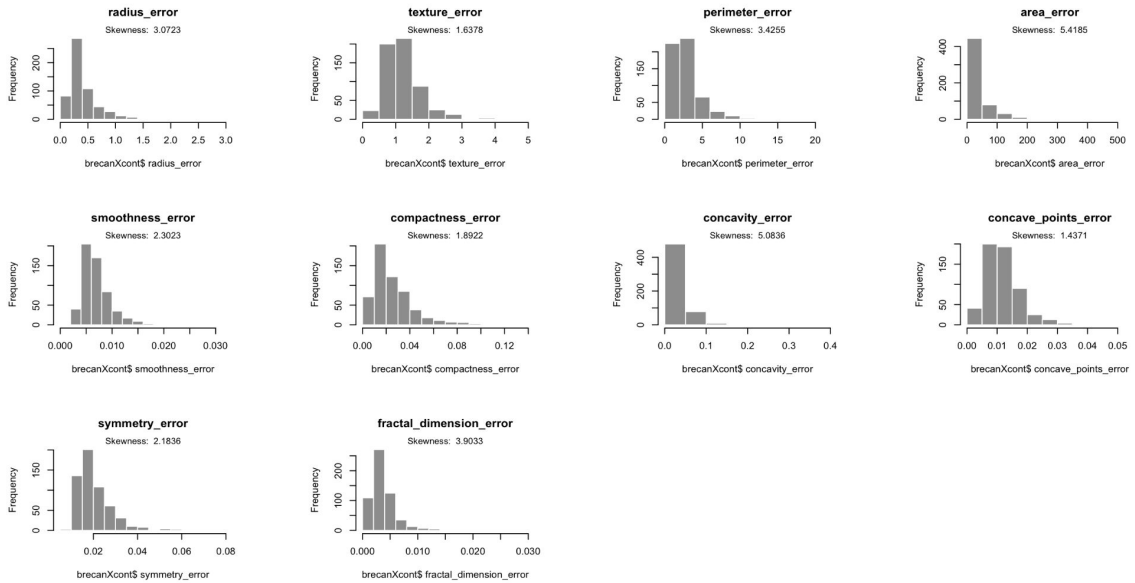
```
> neg_cols_count # none ...
[1] 0
> pos_cols_count # equals sample 569 * 30...
[1] 17070
> nrow(bre_can_dgn) * ncol(bre_can_dgn)
[1] 17070
```

- No categorical variables
  - **No degenerate** (zero/near-zero variance) predictors
  - **No dummy variable** encoding

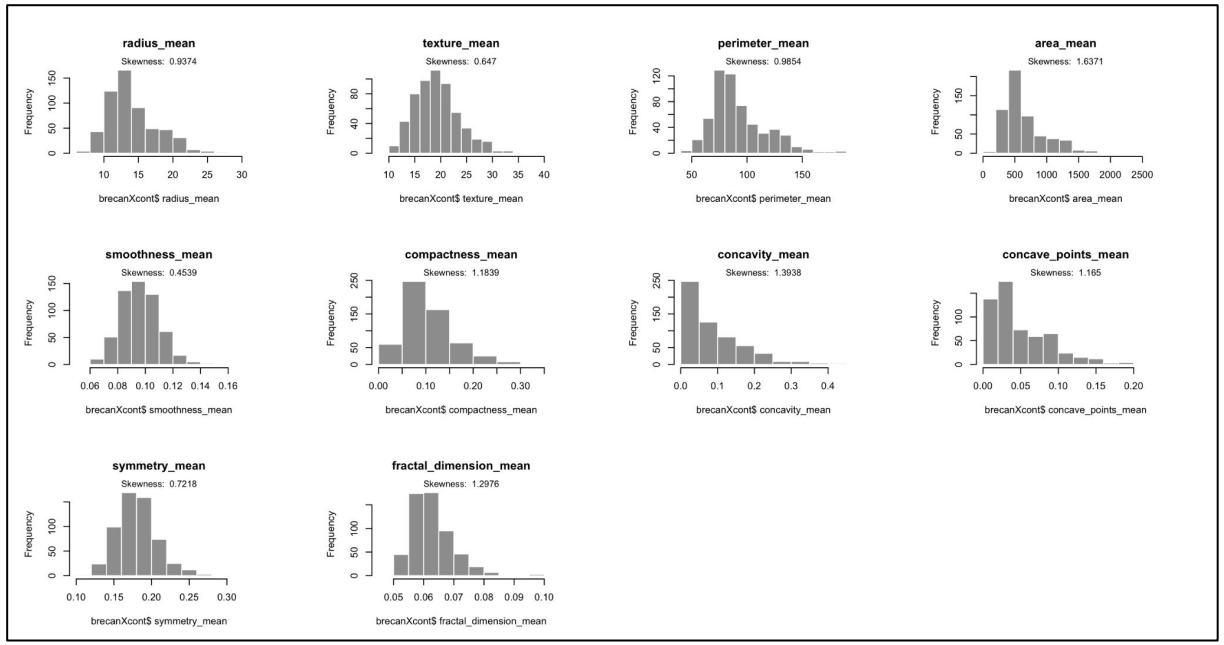
- There were **no missing values. No imputation** was necessary.

Data Anomalies: *Skewness*

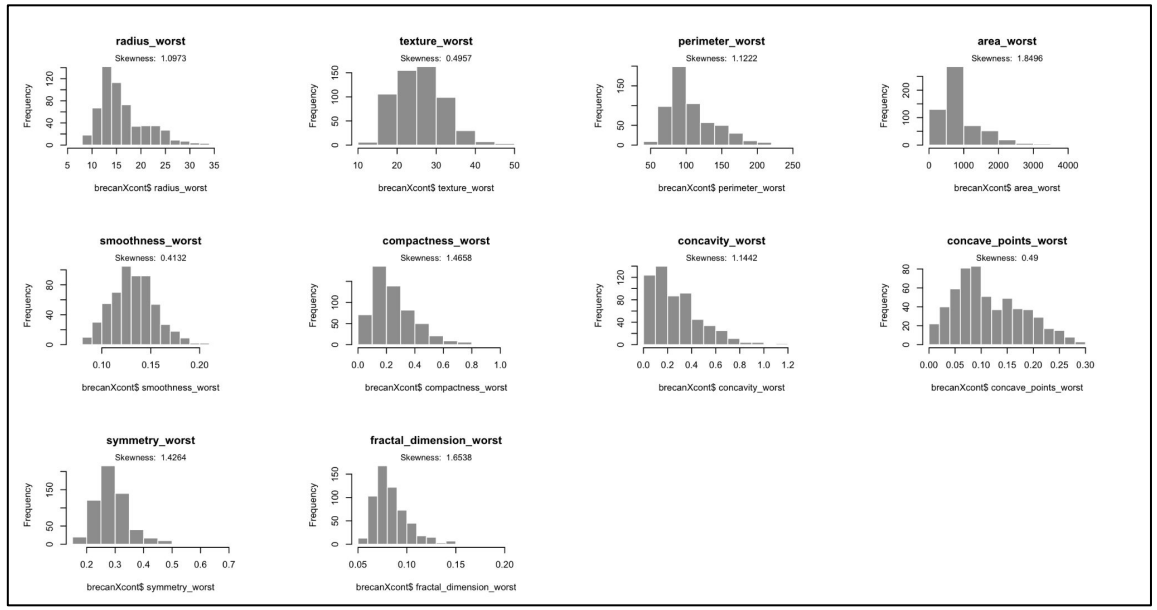
skewness in error values



skewness in mean values



skewness in worst values



Data Anomalies: *Skewness*

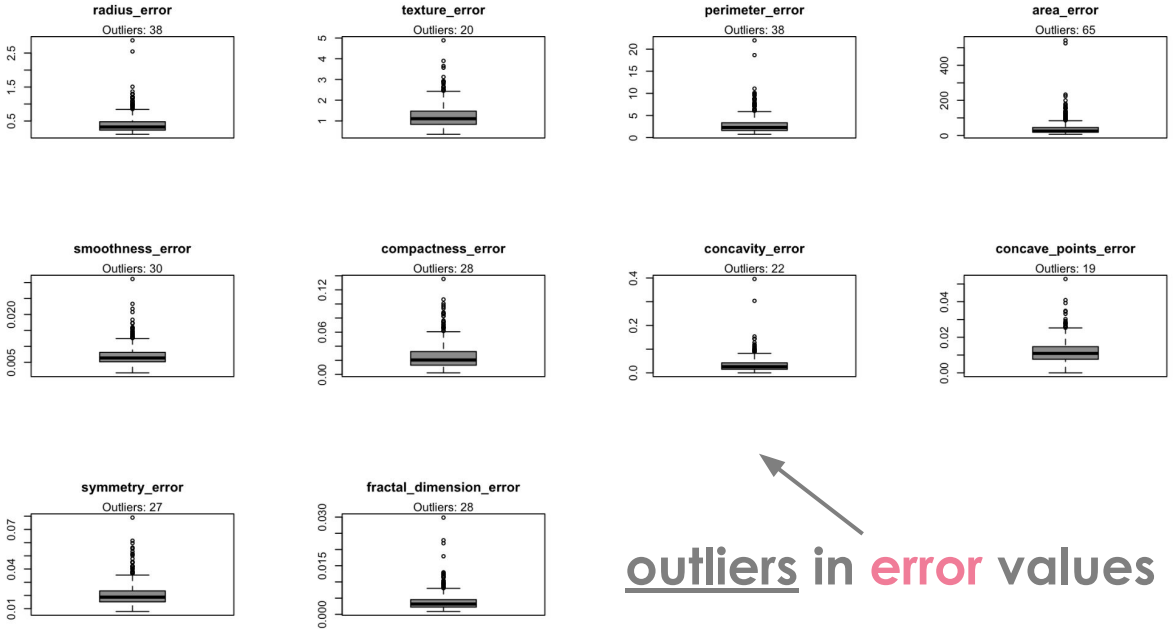
- 22 of the 30 predictors are heavily **skewed to the right**

Predictor	Skewness	Interpretation
radius.mean	0.9374168	Moderately right skewed
texture.mean	0.6470241	Moderately right skewed
perimeter.mean	0.9854334	Moderately right skewed
area.mean	**1.6370654**	**Heavily right skewed**
smoothness.mean	0.4539207	Approximately symmetric
compactness.mean	**1.1838556**	**Heavily right skewed**
concavity.mean	**1.3938008**	**Heavily right skewed**
concave_points.mean	**1.1650124**	**Heavily right skewed**
symmetry.mean	0.7217877	Moderately right skewed
fractal_dimension.mean	**1.2976191**	**Heavily right skewed**
radius.se	**3.0723468**	**Heavily right skewed**
texture.se	**1.6377733**	**Heavily right skewed**
perimeter.se	**3.4254803**	**Heavily right skewed**
area.se	**5.4185001**	**Heavily right skewed**
smoothness.se	**2.3022616**	**Heavily right skewed**

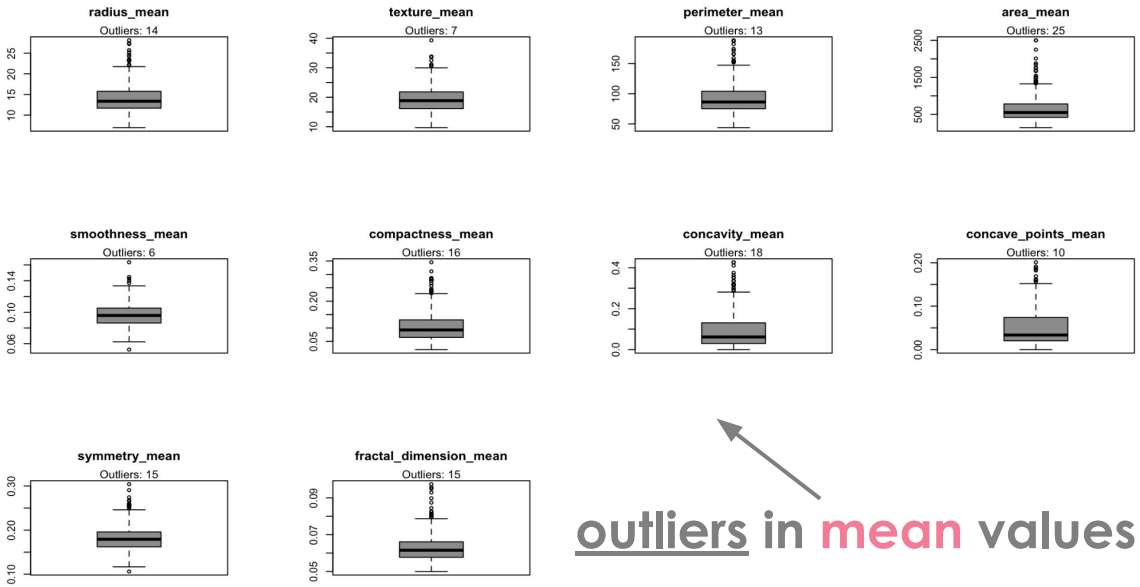
Predictor	Skewness	Interpretation
compactness.se	**1.8922032**	**Heavily right skewed**
concavity.se	**5.0835502**	**Heavily right skewed**
concave_points.se	**1.4370701**	**Heavily right skewed**
symmetry.se	**2.1835728**	**Heavily right skewed**
fractal_dimension.se	**3.9033041**	**Heavily right skewed**
radius.worst	**1.0973059**	**Heavily right skewed**
texture.worst	0.4956970	Approximately symmetric
perimeter.worst	**1.1222227**	**Heavily right skewed**
area.worst	**1.8495814**	**Heavily right skewed**
smoothness.worst	0.4132383	Approximately symmetric
compactness.worst	**1.4657948**	**Heavily right skewed**
concavity.worst	**1.1441794**	**Heavily right skewed**
concave_points.worst	0.4900213	Approximately symmetric
symmetry.worst	**1.4263764**	**Heavily right skewed**
fractal_dimension.worst	**1.6538237**	**Heavily right skewed**

Data Anomalies: Outliers

- Total outliers in data: **608**

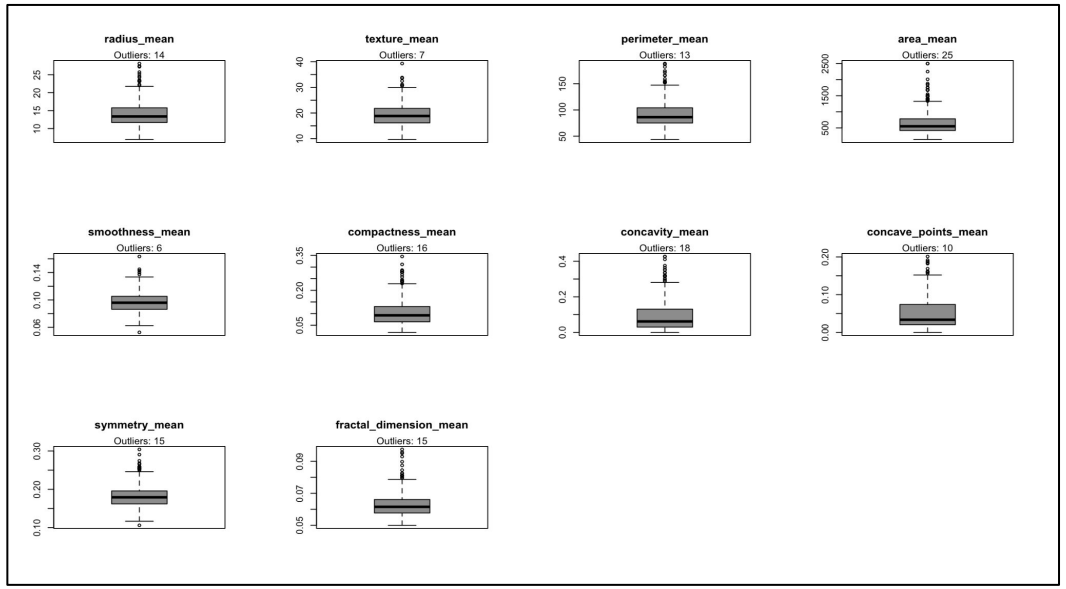


outliers in error values

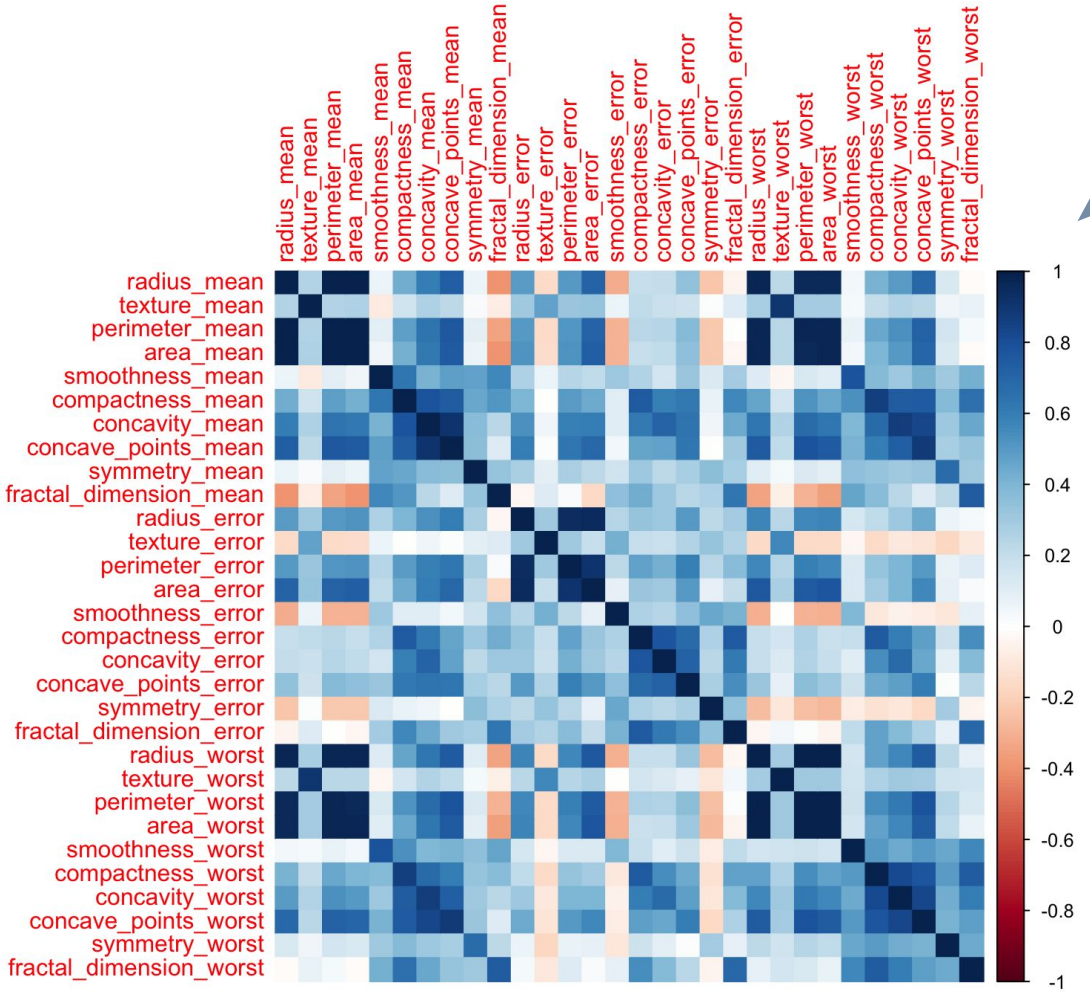


outliers in mean values

outliers in worst values



Data Anomalies: Correlation



Correlation matrix for all 30 predictors

Bivariate relationship between predictor features

```
> corr_bre_can_dgn[1:10,1:10]
                        radius.mean texture.mean perimeter.mean area.mean smoothness.mean
radius.mean               1.0000000    0.32378189      0.9978553 0.9873572      0.17058119
texture.mean              0.3237819    1.00000000      0.3295331 0.3210857     -0.02338852
perimeter.mean            0.9978553    0.32953306      1.0000000 0.9865068      0.20727816
area.mean                 0.9873572    0.32108570      0.9865068 1.0000000      0.17702838
smoothness.mean           0.1705812   -0.02338852      0.2072782 0.1770284      1.00000000
compactness.mean          0.5061236    0.23670222      0.5569362 0.4985017      0.65912322
concavity.mean            0.6767636    0.30241783      0.7161357 0.6859828      0.52198377
concave_points.mean       0.8225285    0.29346405      0.8509770 0.8232689      0.55369517
symmetry.mean             0.1477412    0.07140098      0.1830272 0.1512931      0.55777479
fractal_dimension.mean   -0.3116308   -0.07643718     -0.2614769 -0.2831098     0.58479200
```

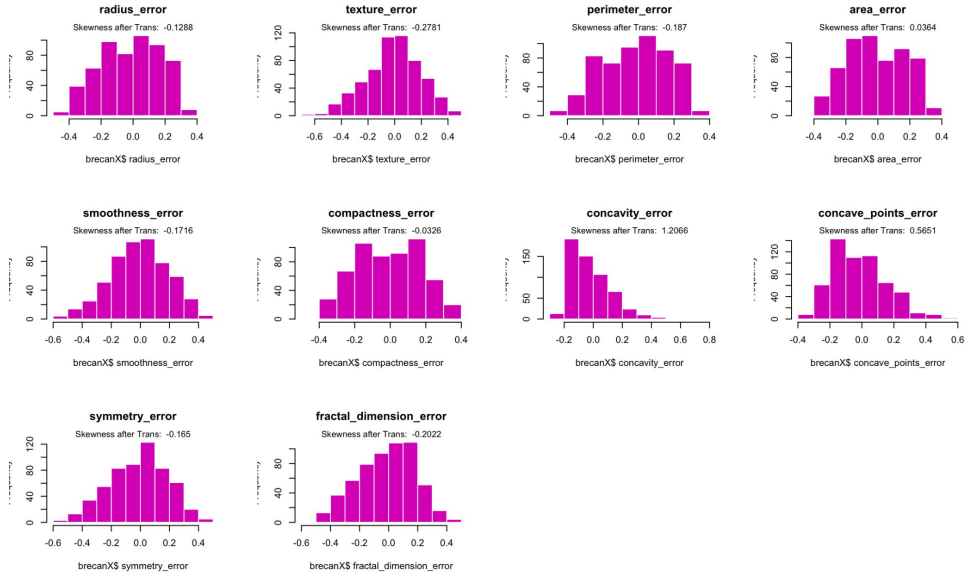
```
compactness.mean concavity.mean concave_points.mean symmetry.mean fractal_dimension.mean
       0.5061236      0.6767636           0.8225285     0.14774124             -0.31163083
       0.2367022      0.3024178           0.2934641     0.07140098             -0.07643718
       0.5569362      0.7161357           0.8509770     0.18302721             -0.26147691
       0.4985017      0.6859828           0.8232689     0.15129308             -0.28310981
       0.6591232      0.5219838           0.5536952     0.55777479              0.58479200
       1.0000000      0.8831207           0.8311350     0.60264105              0.56536866
       0.8831207      1.0000000           0.9213910     0.50066662              0.33678336
       0.8311350      0.9213910           1.0000000     0.46249739              0.16691738
       0.6026410      0.5006666           0.4624974     1.00000000              0.47992133
       0.5653687      0.3367834           0.1669174     0.47992133              1.00000000
```

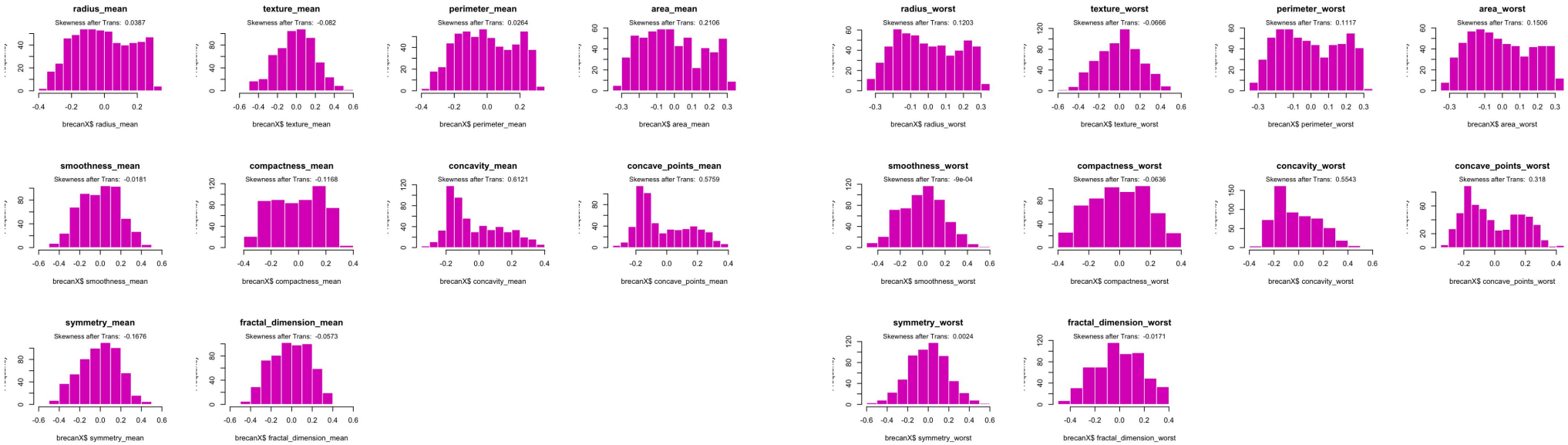
Correlation coefficients of first 10 predictors (**mean values**) showing collinearity in data

Data Transformation: *BoxCox*

- **Centering and scaling** applied
- **BoxCox applied—Skewness reduced** across predictors

```
> preprocTrans
Created from 569 samples and 30 variables

Pre-processing:
  - Box-Cox transformation (24)
  - centered (30)
  - ignored (0)
  - scaled (30)
  - spatial sign transformation (30)

Lambda estimates for Box-Cox transformation:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.0000 -0.5000 -0.3000 -0.3583  0.0000  0.3000

> brecanXcontTrans <- predict(preprocTrans, brecanXcont)
> dim(brecanXcontTrans)
[1] 569  30
```

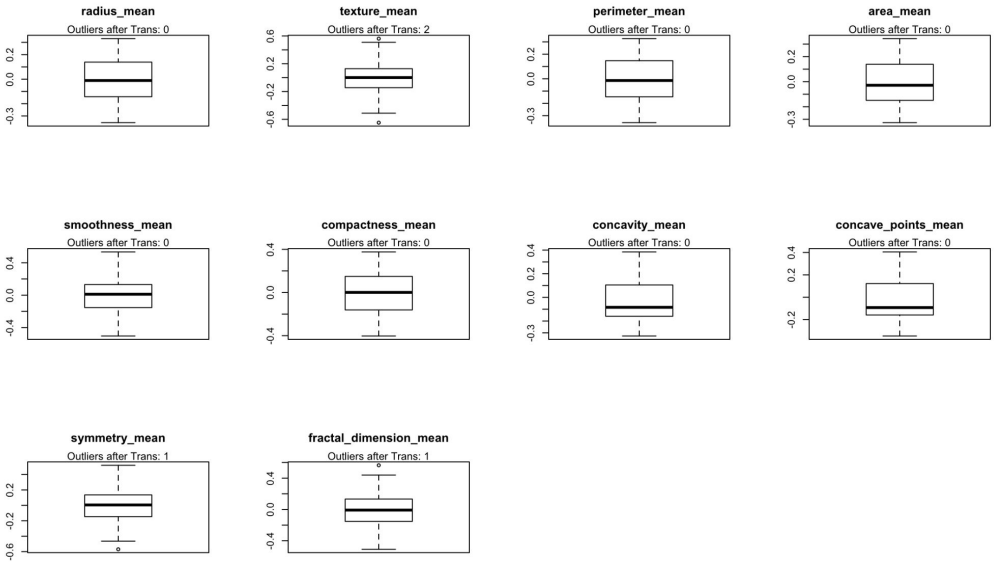


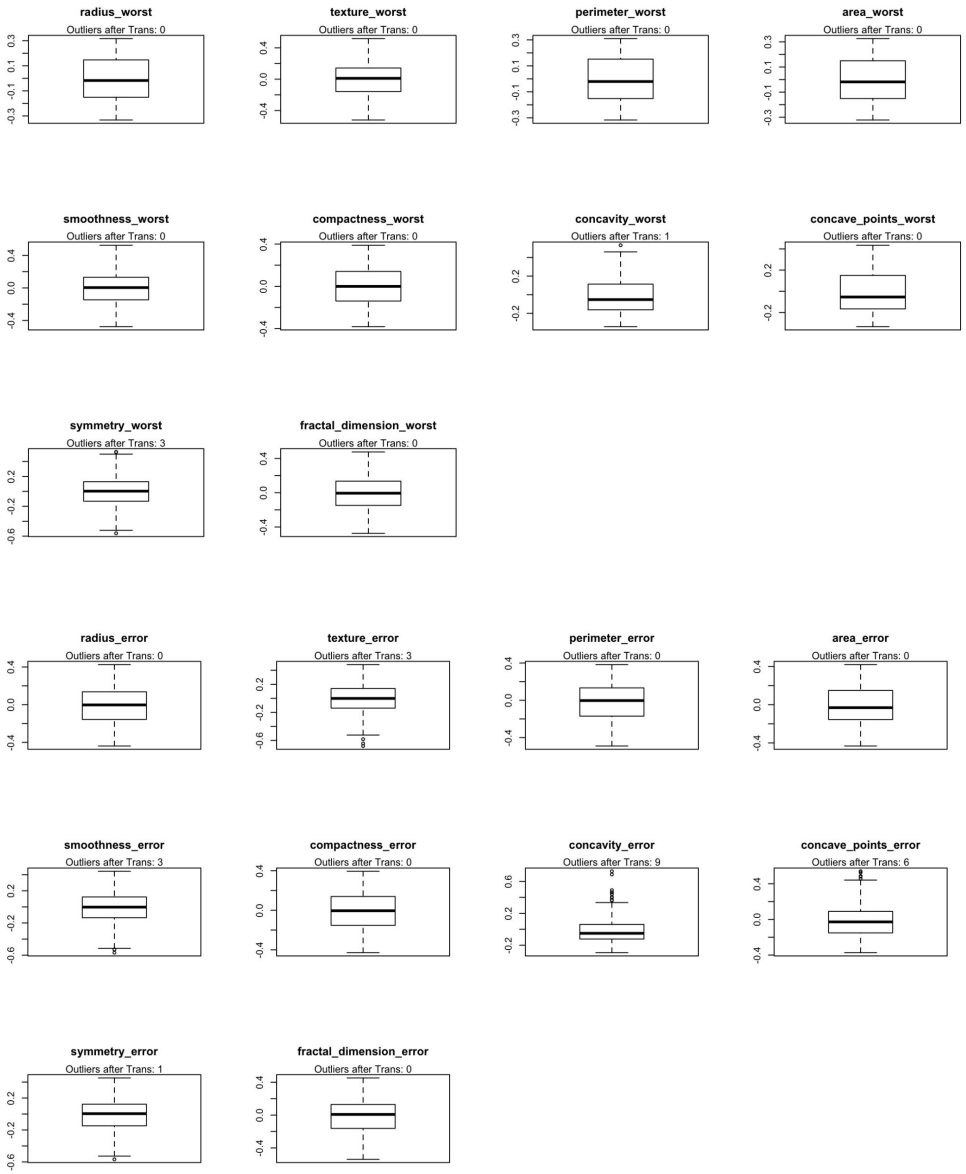


Data Transformation:
spatialSign

608 outliers

→ 30 outliers



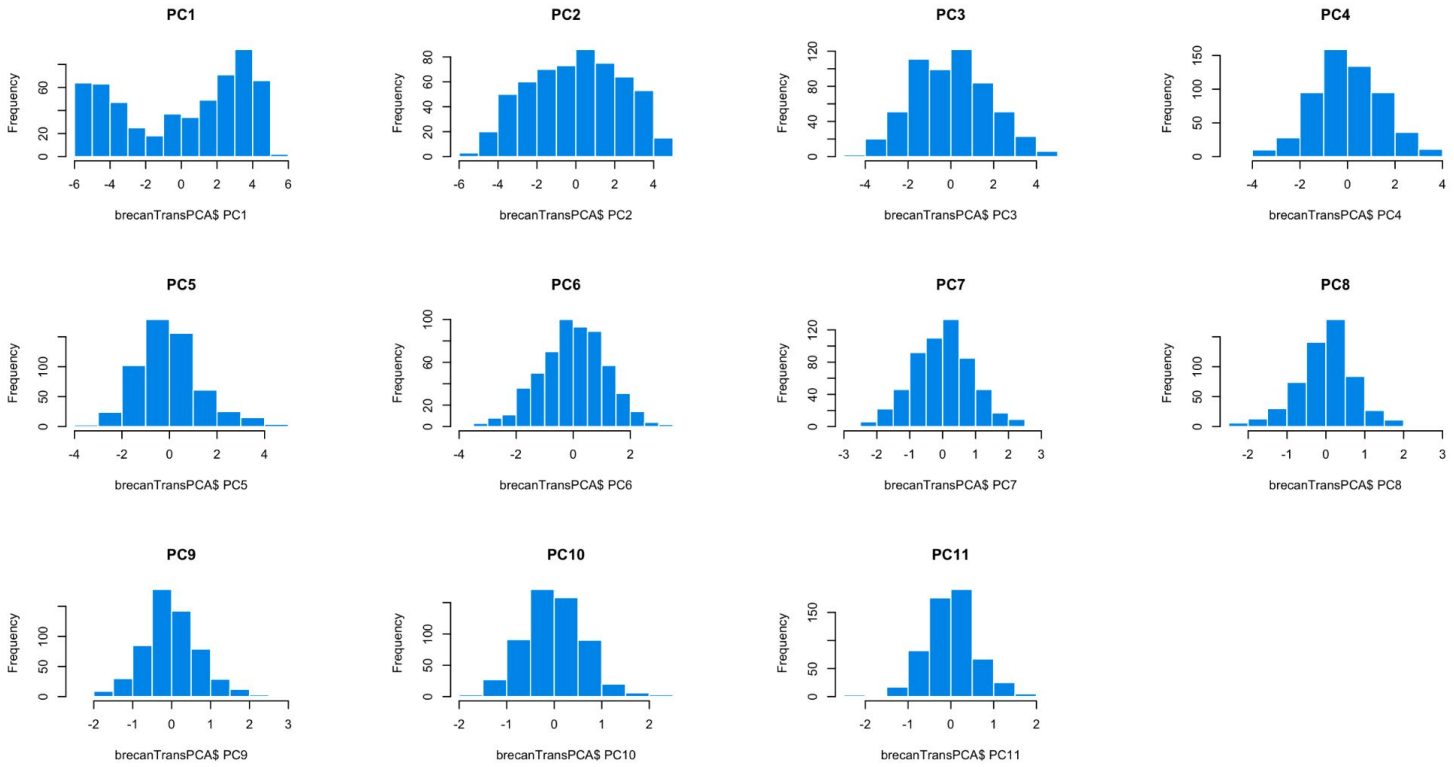


Data Transformation: Highly Correlated Predictors

Cut-off	Total	Predictors to remove
0.75	16	[1] "concavity_mean" "concave_points_worst" "concave_points_mean" [4] "compactness_mean" "perimeter_worst" "concavity_worst" [7] "radius_worst" "area_worst" "compactness_worst" [10] "perimeter_mean" "area_mean" "area_error" [13] "perimeter_error" "compactness_error" "smoothness_mean" [16] "texture_mean"
0.80	**12**	[1] "concavity_mean" "concave_points_worst" "compactness_mean" [4] "perimeter_worst" "concavity_worst" "radius_worst" [7] "area_worst" "perimeter_mean" "area_mean" [10] "area_error" "perimeter_error" "texture_mean" 
0.85	11	[1] "concavity_mean" "concave_points_worst" "compactness_mean" [4] "perimeter_worst" "radius_worst" "area_worst" [7] "perimeter_mean" "area_mean" "area_error" [10] "perimeter_error" "texture_mean"

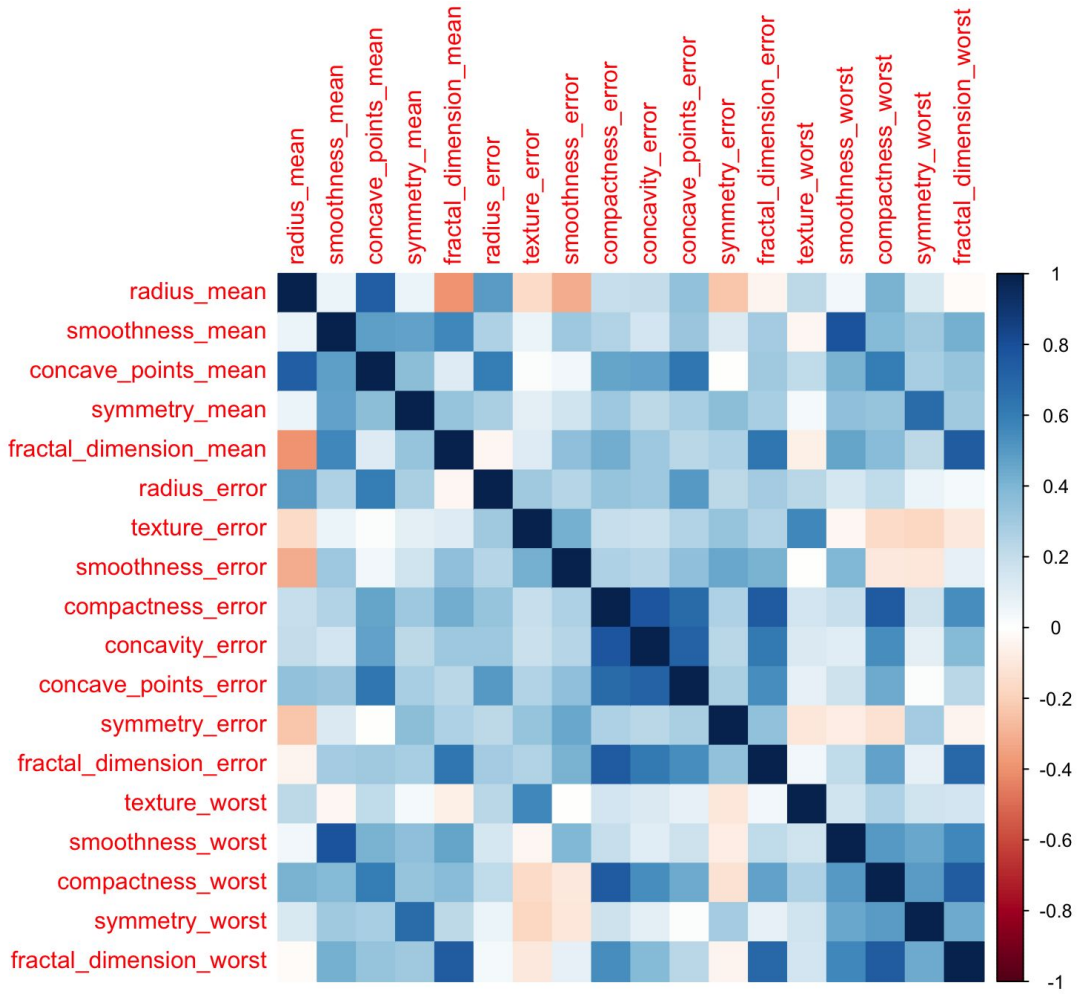
- **Applying cut-off value of 0.8 removed 12** out of 30 predictors.

Data Transformation: "...what if" PCA was used ??



- **PCA needed 11 components** to capture 95% of the variance
- This feature reduction by PCA will **use only 11 out of the 30 predictors** — which may have a high **impact on the sensitivity in the dataset**.
- Handling the collinearity by directly r**emoving high correlated predictors** with 0.8 threshold **was considered the better alternative**.

Data Anomalies: Correlation



- **Correlation matrix after removing 12** highly correlated predictors
- **18 predictors retained** and will be used to train the model

Spending Data

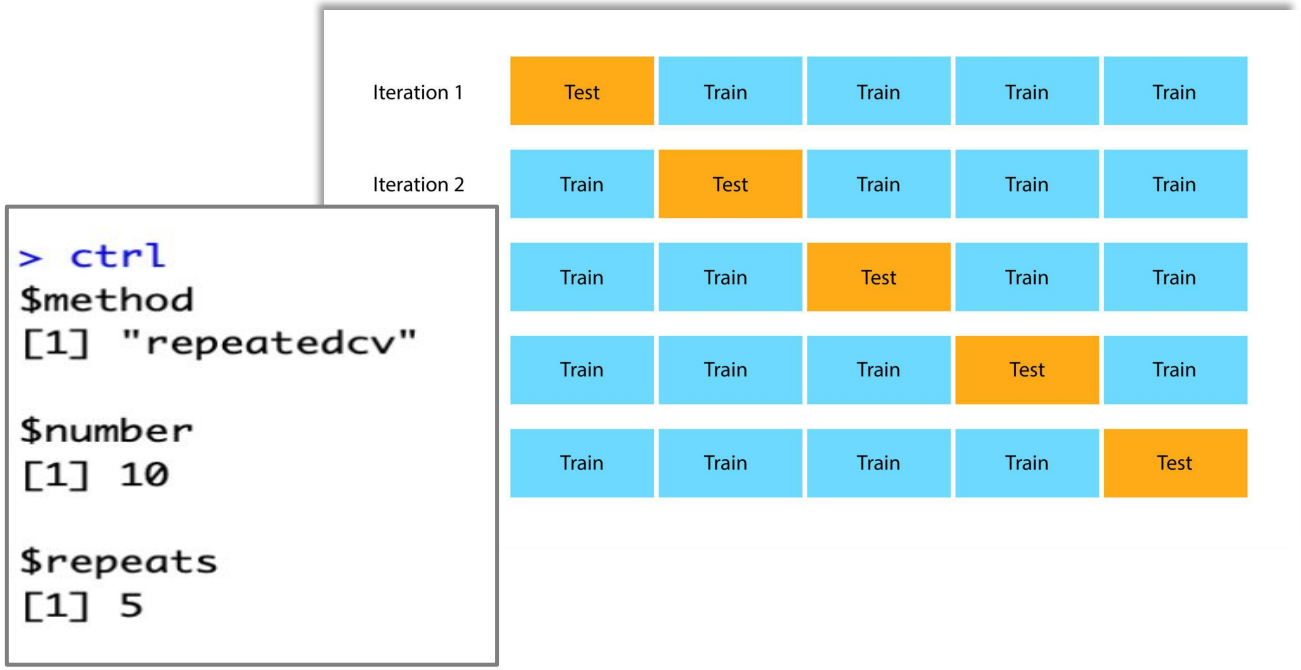
- **Splitting the data**

  Small sample size (569 < 1000). **The initial strategy to use the entire dataset for training was reversed.** Having a new, unseen dataset — as test data — for the model is equally crucial for better model performance **(generalization of the model)**

  80% training. 20% for testing.
  Stratified random sampling used.

- **Resampling technique:**

  Use **k-Fold** (10-Fold) **Cross-Validation;** and **repeat 5 times.**



Spending Data

- **Data partition**

  After data (569 obs) was partitioned…

  - **456 was used to train** the model
  - **113 was used to test** the model
  - 18 predictors were retained after data transformation

```
> dim(brecanX)
[1] 569  18
> # partioned dataset into 80% training and 20% testing...
> dim(trainRows)
[1] 456   1
> dim(brecanX_train)
[1] 456  18
> length(brecanY_train)
[1] 456
> dim(brecanX_test)
[1] 113  18
> length(brecanY_test)
[1] 113
```

BUILDING MODELS

- **5 Linear classification models**
- **8 Non-Linear classification models**

Logistic Regression – LR	Quadratic Discriminant Analysis – QDA
Linear Discriminant Analysis – LDA	Regularized Discriminant Analysis – RDA
PLS Discriminant Analysis – PLSDA	Mixture Discriminant Analysis – MDA
Penalized models	Flexible Discriminant Analysis – FDA
Nearest Shrunken Centroids – NSC	Neural Networks – NNet
	Support Vector Machines – SVM
	K-Nearest Neighbors – KNN
	Naive Bayes model — NB

BUILDING MODELS

- **Train model:** Logistic Regression

```
Generalized Linear Model

456 samples
 18 predictor
  2 classes: 'B', 'M'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 410, 411, 410, 410, 411, 411, ...
Resampling results:

  Accuracy   Kappa
  0.9592754  0.9131085
```

- No Model tuning plot available — **no tuning parameter used**.

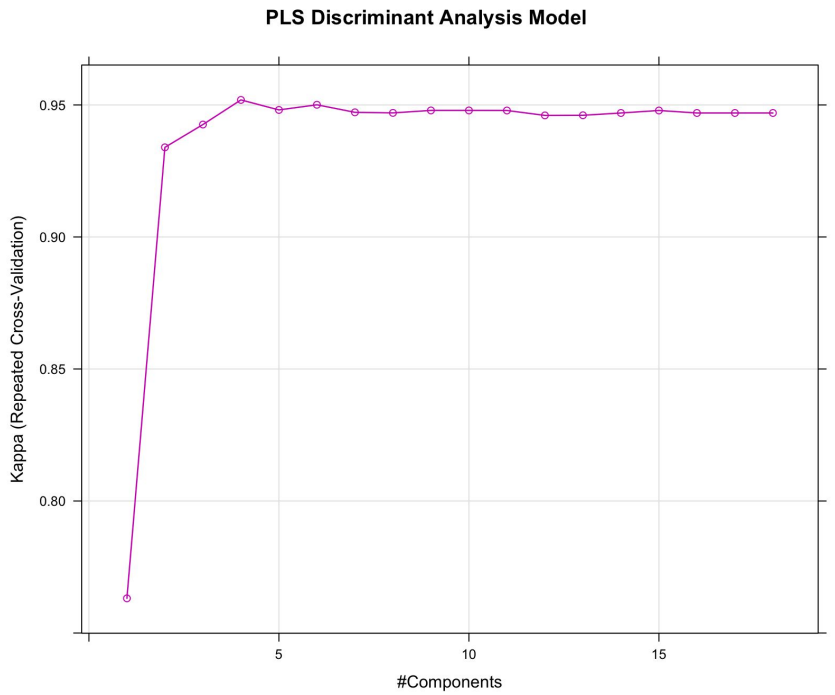
- **Train model:** Linear Discriminant Analysis

```
Linear Discriminant Analysis

456 samples
 18 predictor
  2 classes: 'B', 'M'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 411, 410, 410, 410, 410, 411, ...
Resampling results:

  Accuracy   Kappa
  0.9758841  0.9478061
```

- No Model tuning plot available — **no tuning parameter used**.

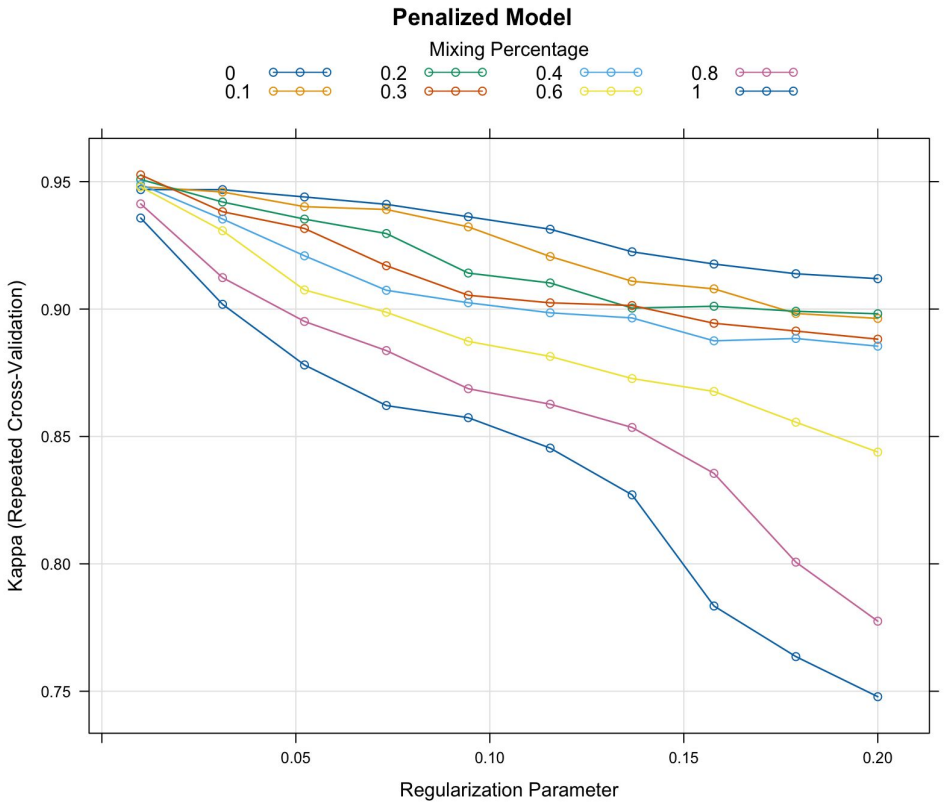
BUILDING MODELS

- **Train model:** PLS Discriminant Analysis



Model tuning plot. Optimal model at "ncomp" = 4

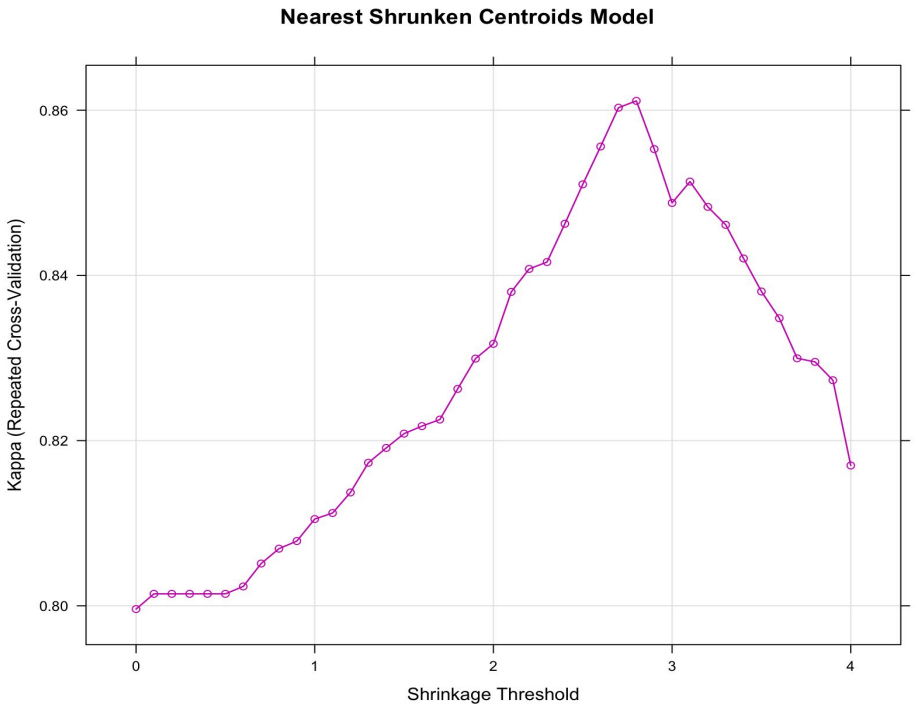
- **Train model:** Penalized models



Optimal model selected at: lambda = 0.3. gamma=0.01

BUILDING MODELS

- **Train** model: Nearest Shrunken Centroids



Model tuning plot. Optimal model at shrinking threshold = 2.8

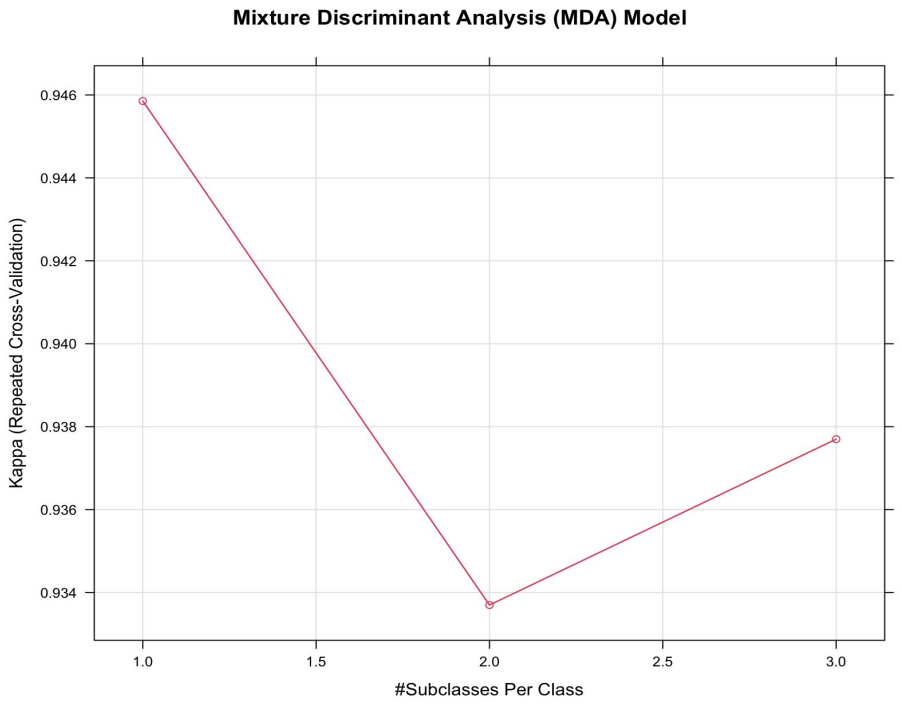
- **Train model:** Quadratic Discriminant Analysis

```
Quadratic Discriminant Analysis

456 samples
 18 predictor
  2 classes: 'B', 'M'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 411, 411, 410, 410, 410, 410, ...
Resampling results:

  Accuracy  Kappa
  0.964     0.9214361
```

- No Model tuning plot available — **no tuning parameter used**.

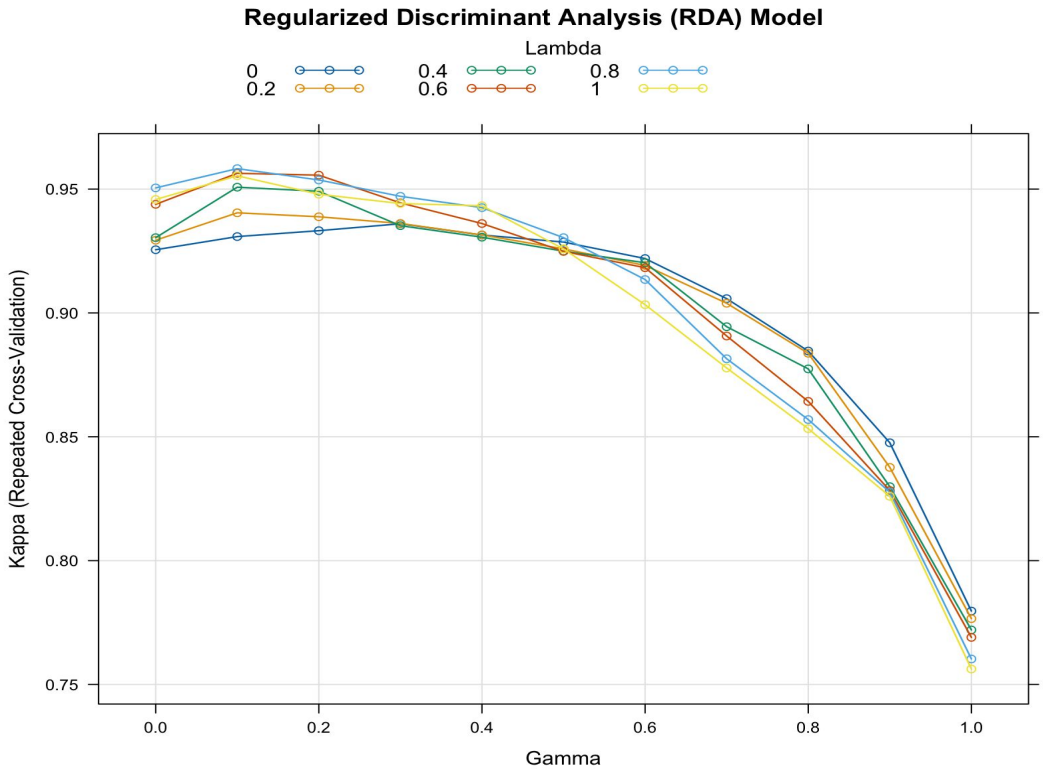
BUILDING MODELS

- **Train model:** Mixture Discriminant Analysis



Model tuning plot. Optimal model at "subclasses" = 1

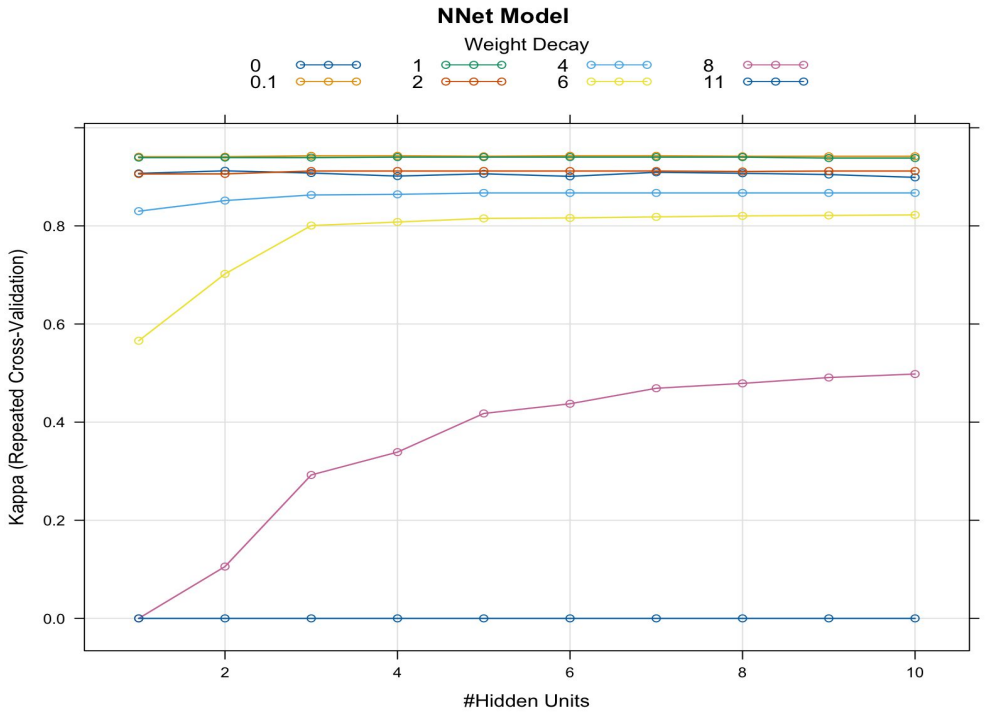
- **Train model:** Regularized Discriminant Analysis model



Optimal model selected at: gamma=0.1. lambda = 0.8.

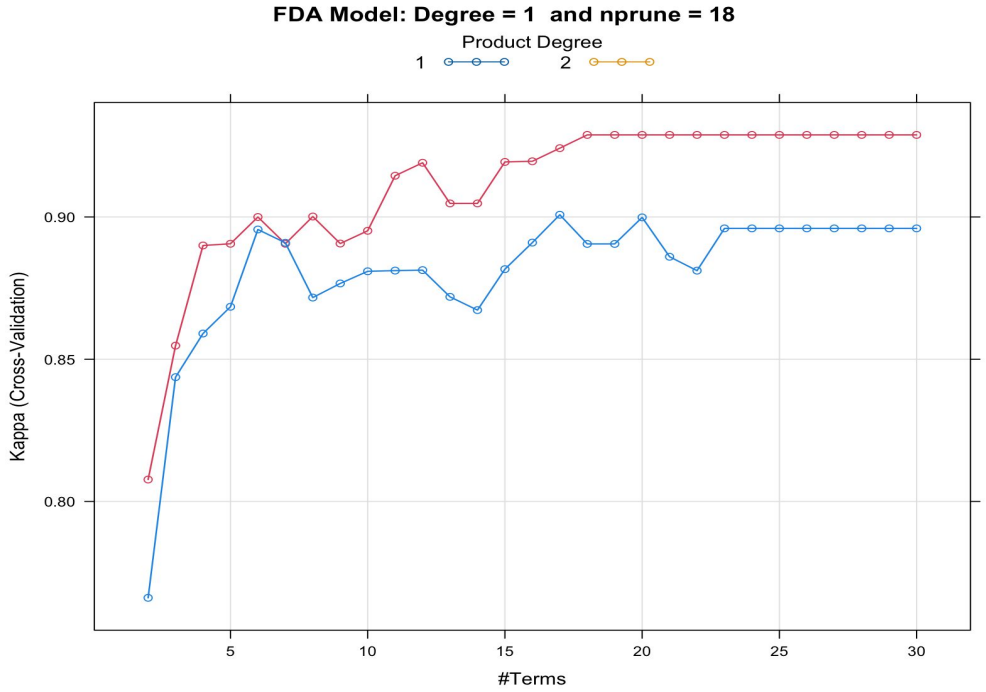
BUILDING MODELS

- **Train** model: Neural Networks



Model tuning plot. Optimal model at size = 3 and decay = 0.1

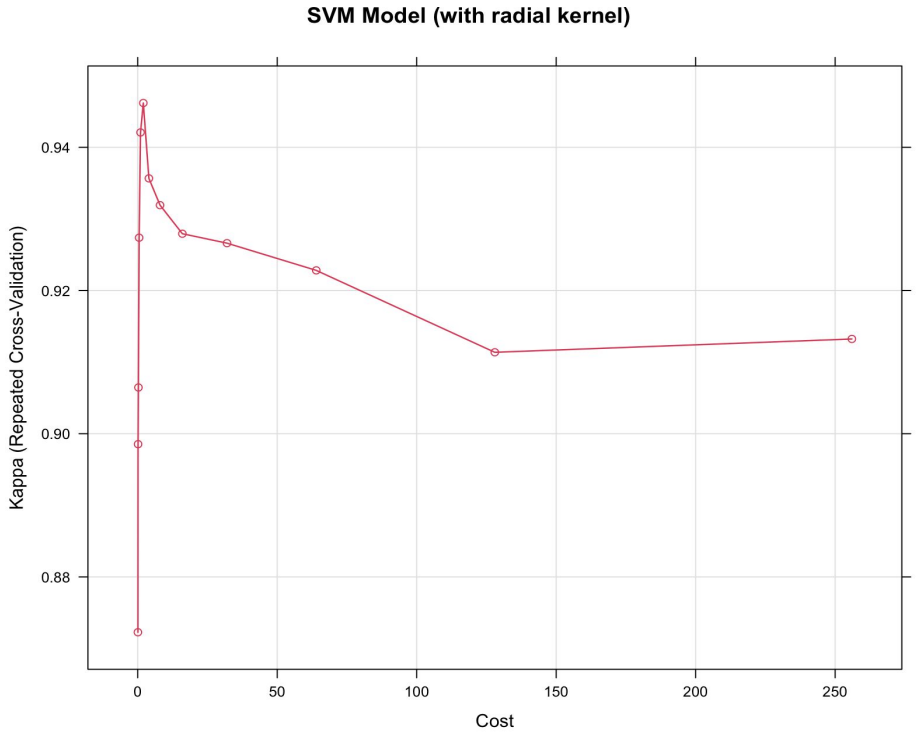
- **Train model:** Flexible Discriminant Analysis



- **Optimal model: 18 terms** (nprune) retained at **degree = 2**

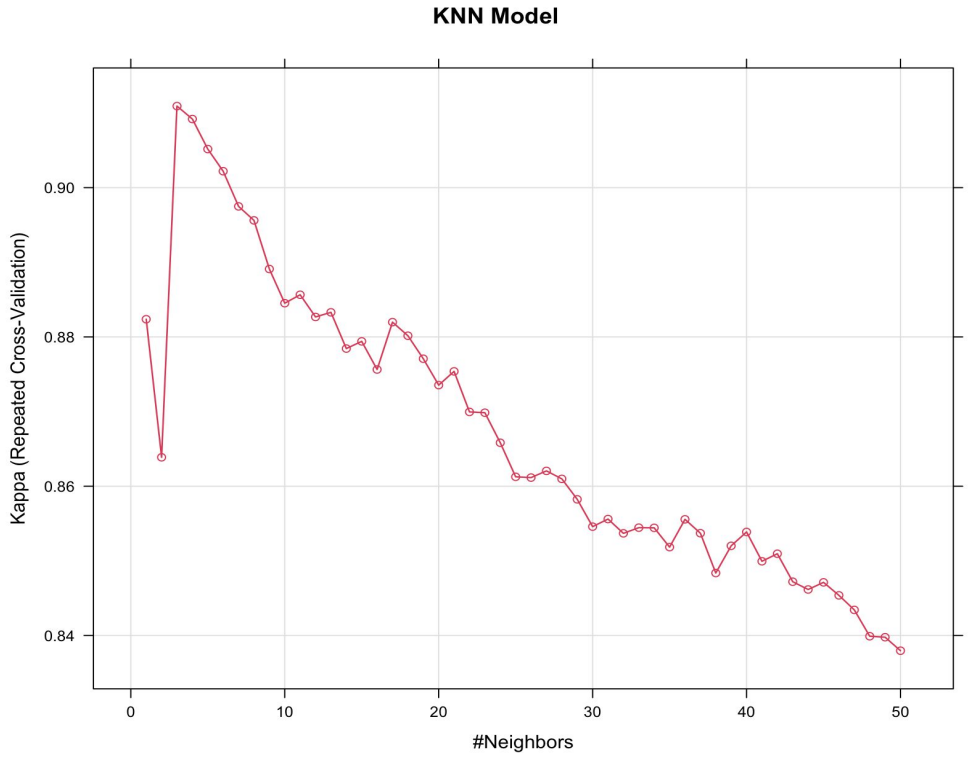
BUILDING MODELS

- **Train model:** Support Vector Machine (SVM)



Model tuning plot. Optimal model at Cost = 2. ***sigma*** was held constant.

- **Train model:** K-Nearest Neighbor (KNN)



Optimal model selected at: k = 3

BUILDING MODELS

- **Train** model: Naïve Bayes

```
Naive Bayes

456 samples
 18 predictor
  2 classes: 'B', 'M'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 410, 410, 410, 411, 410, 410, ...
Resampling results:

  Accuracy   Kappa
  0.9337391  0.8564409

Tuning parameter 'fL' was held constant at a value of 2
Tuning parameter
 'usekernel' was held constant at a value of TRUE
Tuning parameter 'adjust' was
 held constant at a value of TRUE
```

- **Model tuning plot.** No tuning parameter used for the model.
- **fL (Laplace correction) value** was used to "smoothen" (control) the occurrences of zero probability estimates.

BUILDING MODELS

▪ Summary Statistics Table

	Model	Best Tuning Parameter	Training		Testing		
			AR	Kappa	AR	Kappa	ROC AUC
LINEAR MODELS	LR	None	0.9593	0.9131	0.9646 ↑	0.9249 ↑	0.9990**
	LDA	None	0.9759	0.9478	0.9558 ↓	0.9075 ↓	0.9970
	PLSDA	ncomp = 4	0.9776	0.9519	0.9558 ↓	0.9066 ↓	0.9963
	Penalized	**alpha = 0.3, lambda = 0.01**	**0.9780***	**0.9527***	**0.9646 ↓**	**0.9257 ↓**	**0.9977**
	NSC	threshold = 2.8	0.9364	0.8611	0.9381 ↑	0.8680 ↑	0.9805
LINEAR MODELS	QDA	None	0.9640	0.9214	0.9558 ↓	0.9048 ↓	0.9930
	RDA	**Gamma = 0.1, lambda = 0.8**	**0.9808****	**0.9583****	**0.9646 ↓**	**0.9257 ↓**	**0.9966**
	MDA	subclasses = 1	0.9750	0.9459	0.9558 ↓	0.9075 ↓	0.9970
	NNet	size = 3, decay = 0.1	0.9732	0.9429	0.9646 ↓	0.9249 ↓	0.9987*
	FDA	degree = 1, nprune = 18	0.9671	0.9288	0.9381 ↓	0.8693 ↓	0.9903
	SVM	cost, C = 2	0.9749	0.9462	0.9558 ↓	0.9057 ↓	0.9966
	KNN	k = 3	0.9583	0.9109	0.9381 ↓	0.8693 ↓	0.9792
	NB	None	0.9337	0.8564	0.9204 ↓	0.8303 ↓	0.9826

BUILDING MODELS

- Choosing the Best Two (2) Models from Training

1. **Regularized Discriminant Analysis (RDA) model.**

   Kappa value of 95.83%

   Accuracy value of 98.08%

2. **Penalized model**

   Kappa value of 95.27%

   Accuracy value of 97.80%

- **Kappa statistic** was used to train or fit the models. Kappa was considered first followed by **accuracy rate metric**.
- **High Kappa values** indicate there is a strong **concordance** —*agreement* between predictors and the response variable.
- All models except "LR and NSC" **reduced in the evaluation metrics** used on test.

BUILDING MODELS

Best Model (**Penalized Model**) — Confusion Matrix

- Predicting on the test data, the two best models RDA and Penalized coincidentally **had the same values for Kappa and Accuracy rate;** as well as Sensitivity and Specificity.
- The **ROC AUC** value for the penalized model was a little better than the RDA.
- RDA model tend to **underfit** the model a little bit more **compared to** the Penalized model.

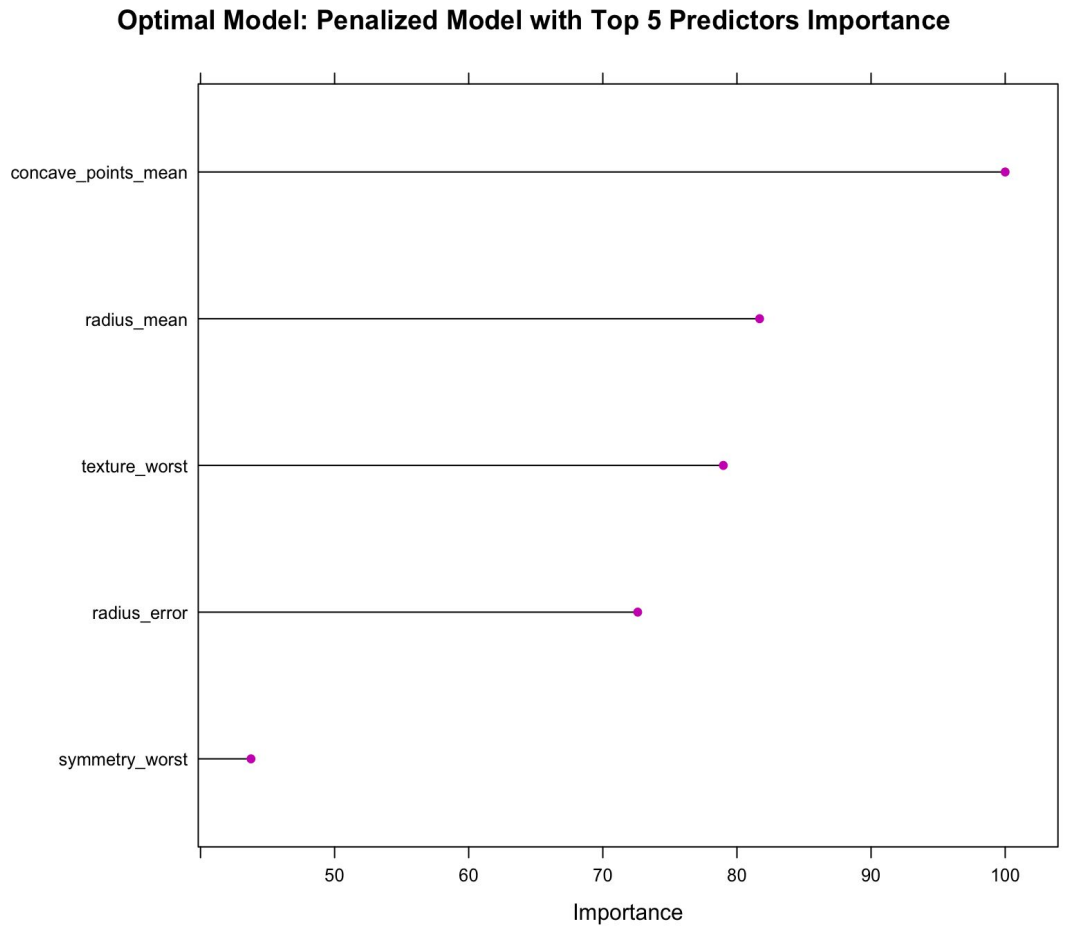
```
> glmnConfMatrix
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 67  0
         M  4 42

               Accuracy : 0.9646
                 95% CI : (0.9118, 0.9903)
    No Information Rate : 0.6283
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9257

 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 0.9437
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9130
             Prevalence : 0.6283
         Detection Rate : 0.5929
   Detection Prevalence : 0.5929
      Balanced Accuracy : 0.9718

       'Positive' Class : B
```

Best Model: Penalized Model

Important Predictors

The **top 5 predictors** using the best model **in order of importance** are as follows:

#	Predictor	Overall
1	concave_points_mean	100.000
2	radius_mean	81.691
3	texture_worst	78.984
4	radius_error	72.602
5	symmetry_worst	43.768



THANK YOU