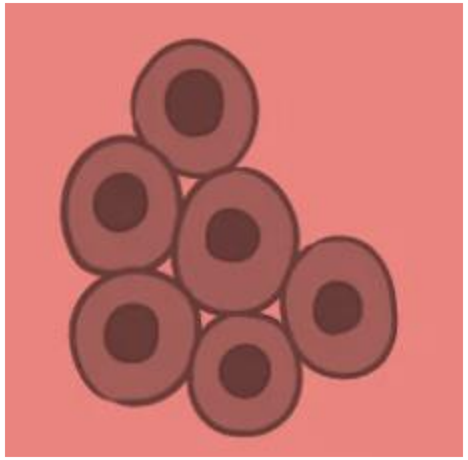
PSY810: Tumor Cell Classification by Nuclear Morphology

Christian D’Andrea

June 3 2020

Research Question:

How can we separate **benign** tumor cells from **malignant** tumor cells based on morphology of the nucleus?

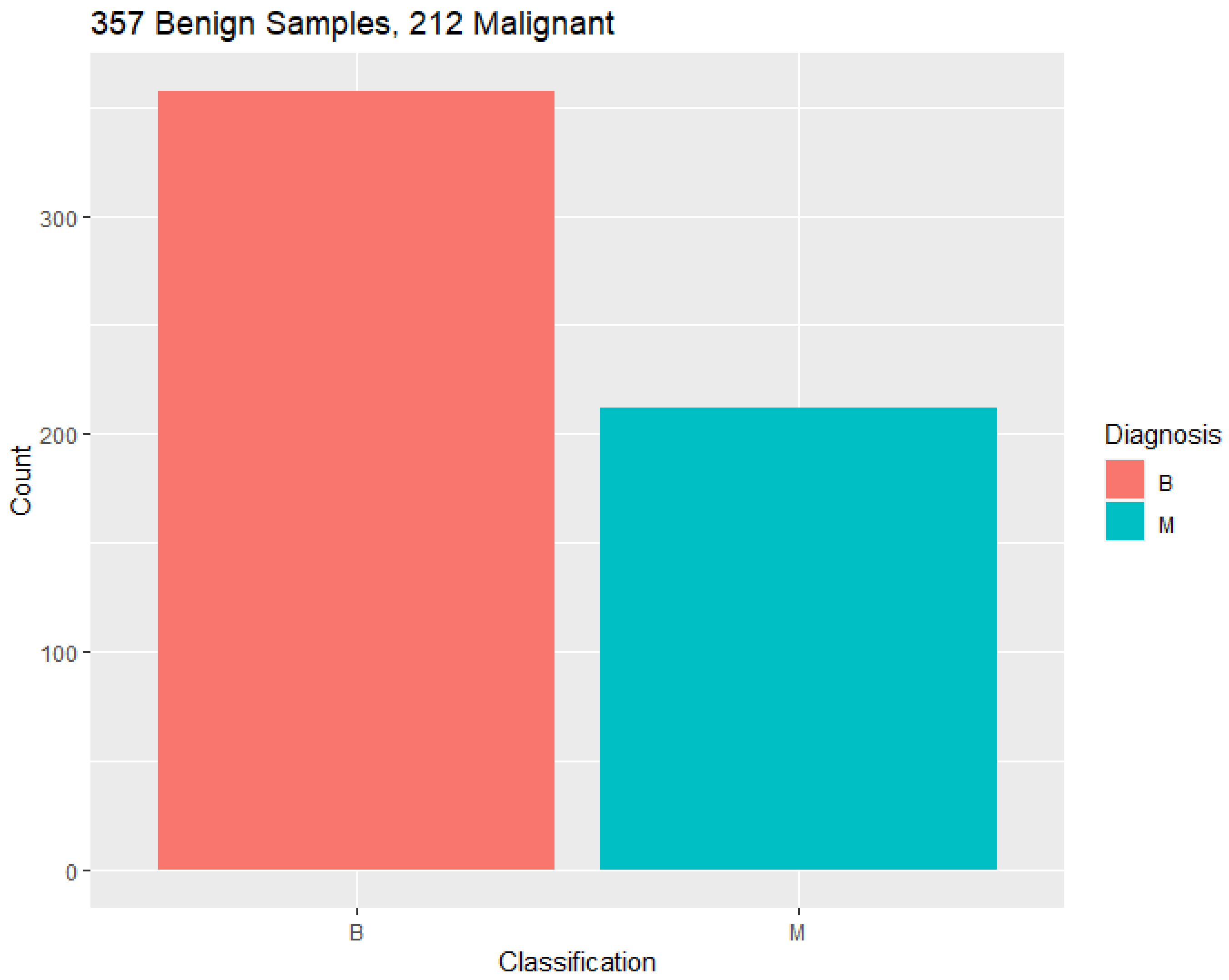


“Histological Samples”

Predictors (mean, max, SE of):

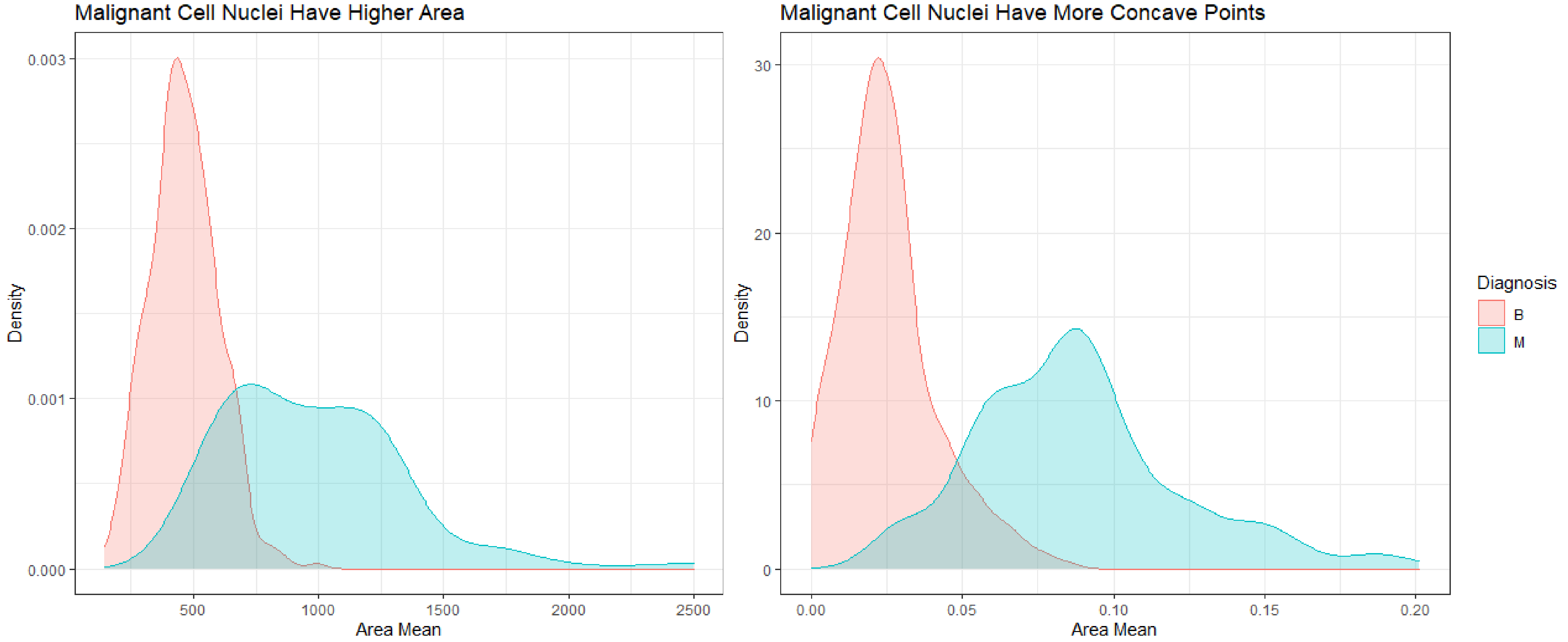
- a) radius (mean of distances from center to points on the perimeter)
- b) texture (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) smoothness (local variation in radius lengths)
- f) compactness (perimeter^2 / area - 1.0)
- g) concavity (severity of concave portions of the contour)
- h) concave points (number of concave portions of the contour)
- i) symmetry (long axis/short axis)
- j) fractal dimension ("coastline approximation" – describes complexity of border pattern with respect to scale of measurement)

569 Samples

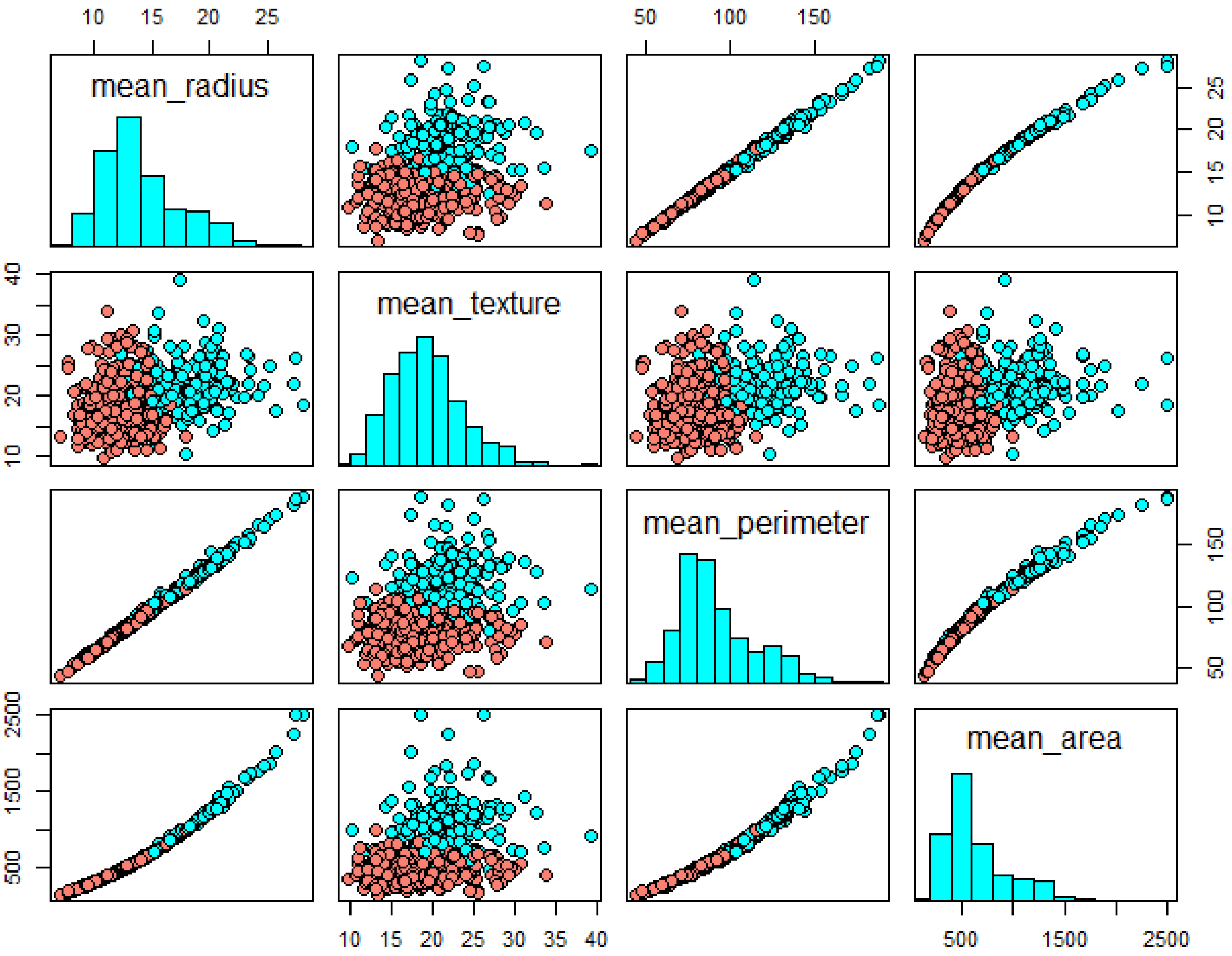


357 Benign Samples, 212 Malignant
Count
300
200
100
0
B
M
Classification
Diagnosis
B
M

Fairly distinct distribution across predictors

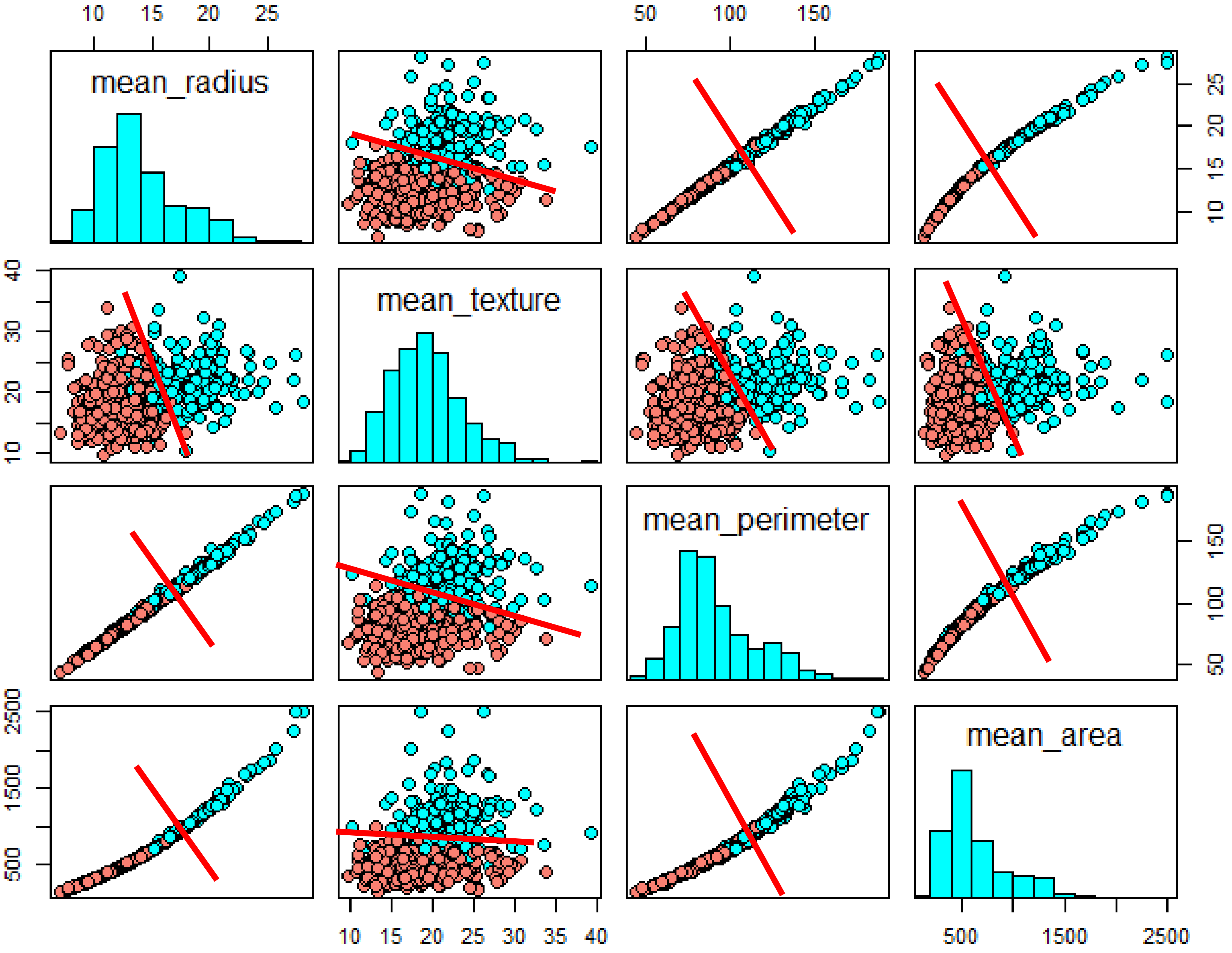


How do the decision boundaries look?



How do the decision boundaries look?

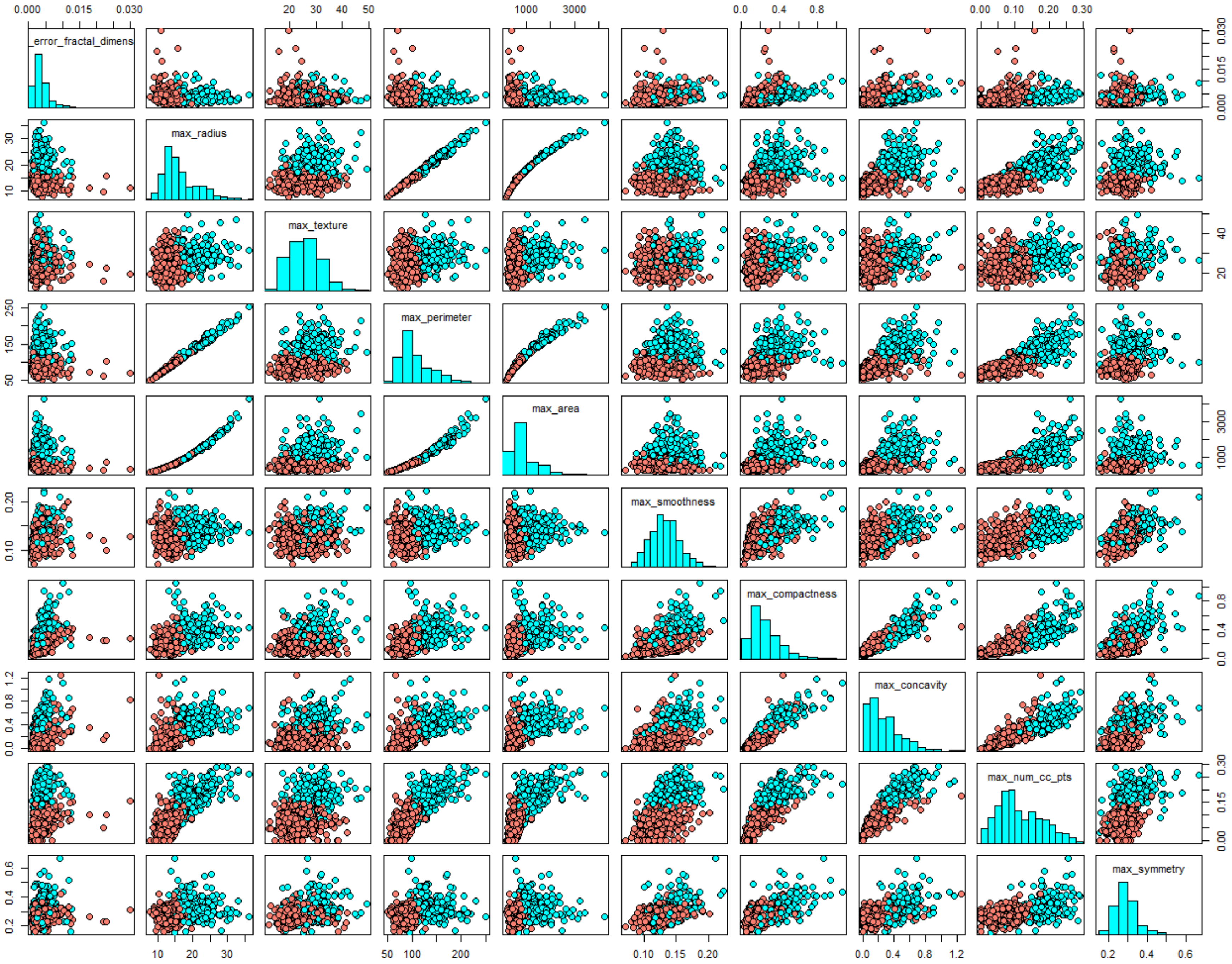
At first glance, all boundaries seem close to linear, nothing resembling radial



How do the decision boundaries look?

At first glance, All boundaries seem linear, nothing resembling radial

Same for all predictors



Some Possible Methods:

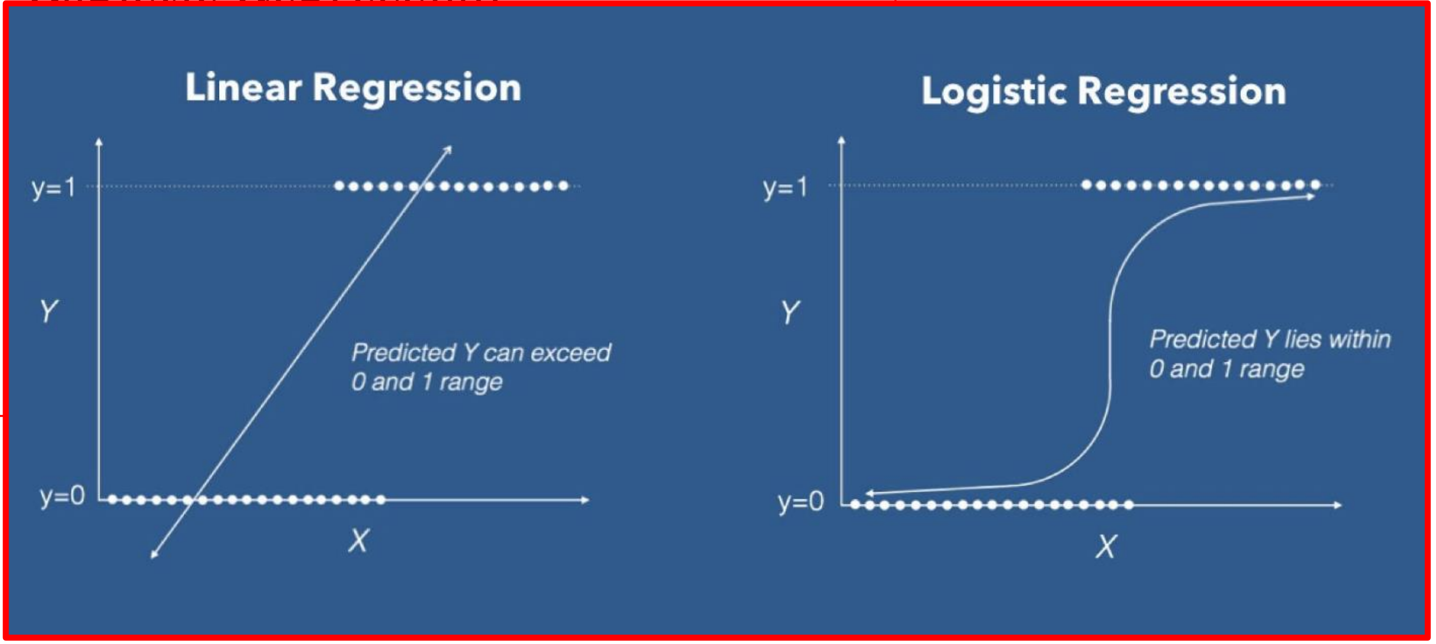
Linear Regression
Logistic Regression
LDA
QDA
Linear SVM
Radial SVM
Random Forest

Some Possible Methods:

~~Linear Regression~~

Logistic Regression

Not suitable for binary classification



Strong 1st option (Zoe's suggestion), commonly used for binary classification [1]

Some Possible Methods:

~~Linear Regression~~
Logistic Regression
LDA
QDA
Linear SVM
Radial SVM
Random Forest

Not as suitable as LR for binary classification

McCormick, T.H., Raftery, A.E., Madigan, D. and Burd, R.S. (2012), Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. Biometrics, 68: 23-30. doi:10.1111/j.1541-0420.2011.01645.x

Strong 1st option (Zoe's suggestion), commonly used for binary classification [ref]

May work, assumptions must be met to beat LR

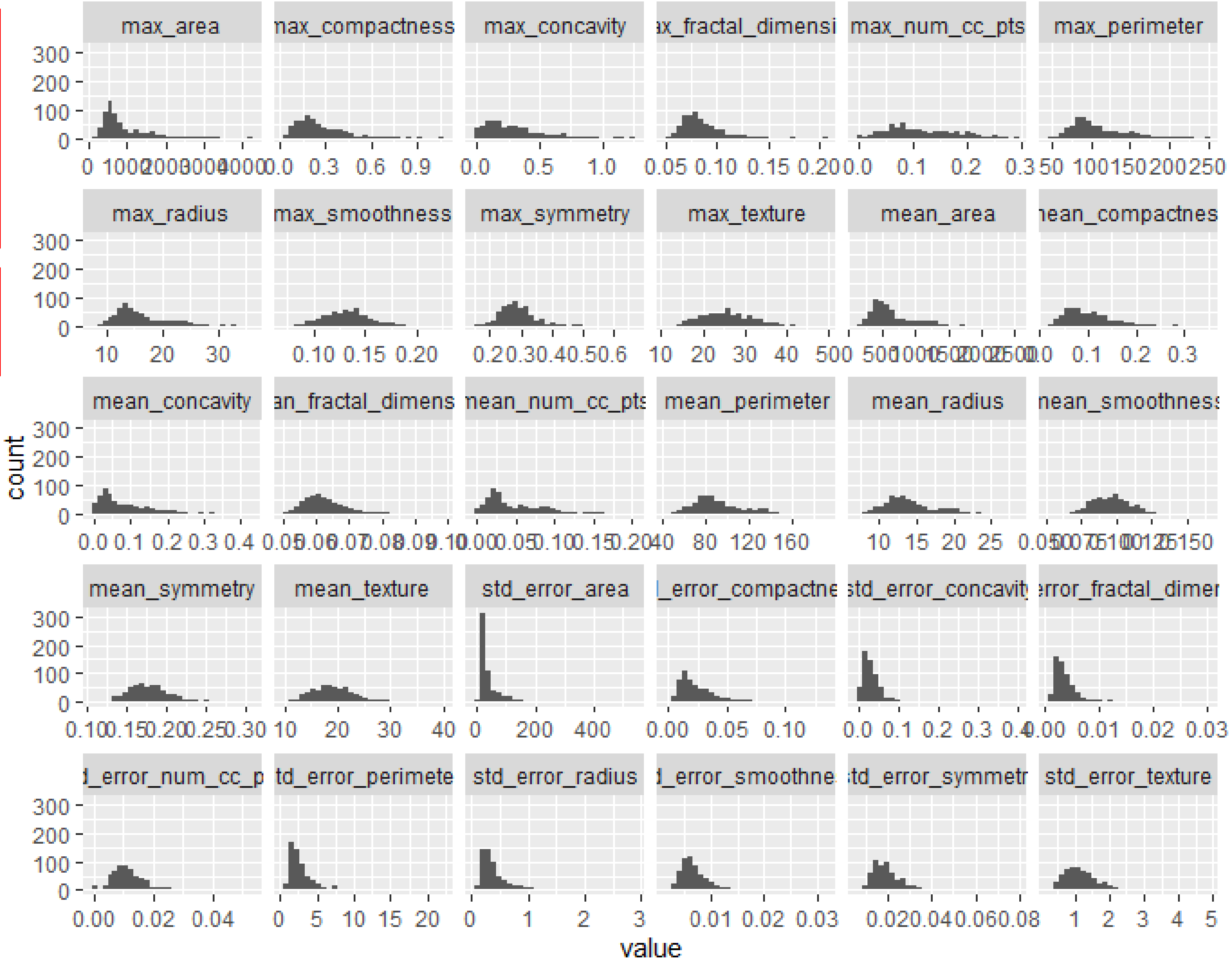
Some Possible Methods:

~~Linear Regression~~
Logistic Regression
LDA
QDA
Linear SVM
Radial SVM
Random Forest

Not as suitable as LR for binary classification

Variable distributions are loosely normal

LDA and QDA may work well



Variable distributions are loosely normal

LDA and QDA may work well

But variance is probably not common, so the LDA assumption of common variance is violated

```
> min(colVariance)
[1] 7.001692e-06
> max(colVariance)
[1] 324167.4
> mean(colVariance)
[1] 15063.22
> sd(colVariance) #....no shot for LDA
[1] 62593.55
```

Strong 1st option (Zoe's suggestion), commonly used for binary classification [ref]

May work, assumption must be met to beat LR

May provide better fit with decreased interpretability

Some Possible Methods:

~~Linear Regression~~
Logistic Regression
~~LDA~~
QDA
Linear SVM
Radial SVM
Random Forest

Not as suitable as LR for binary classification

Variance assumption not met

Logistic Regression

vs

QDA,
Linear SVM,
Radial SVM, and
Random Forest

LR with all 30 predictors did fit the data,
but z=0 indicates perfect separation, can't assess ß ± SE

Hmm…
Let's get rid of variables causing perfect separation

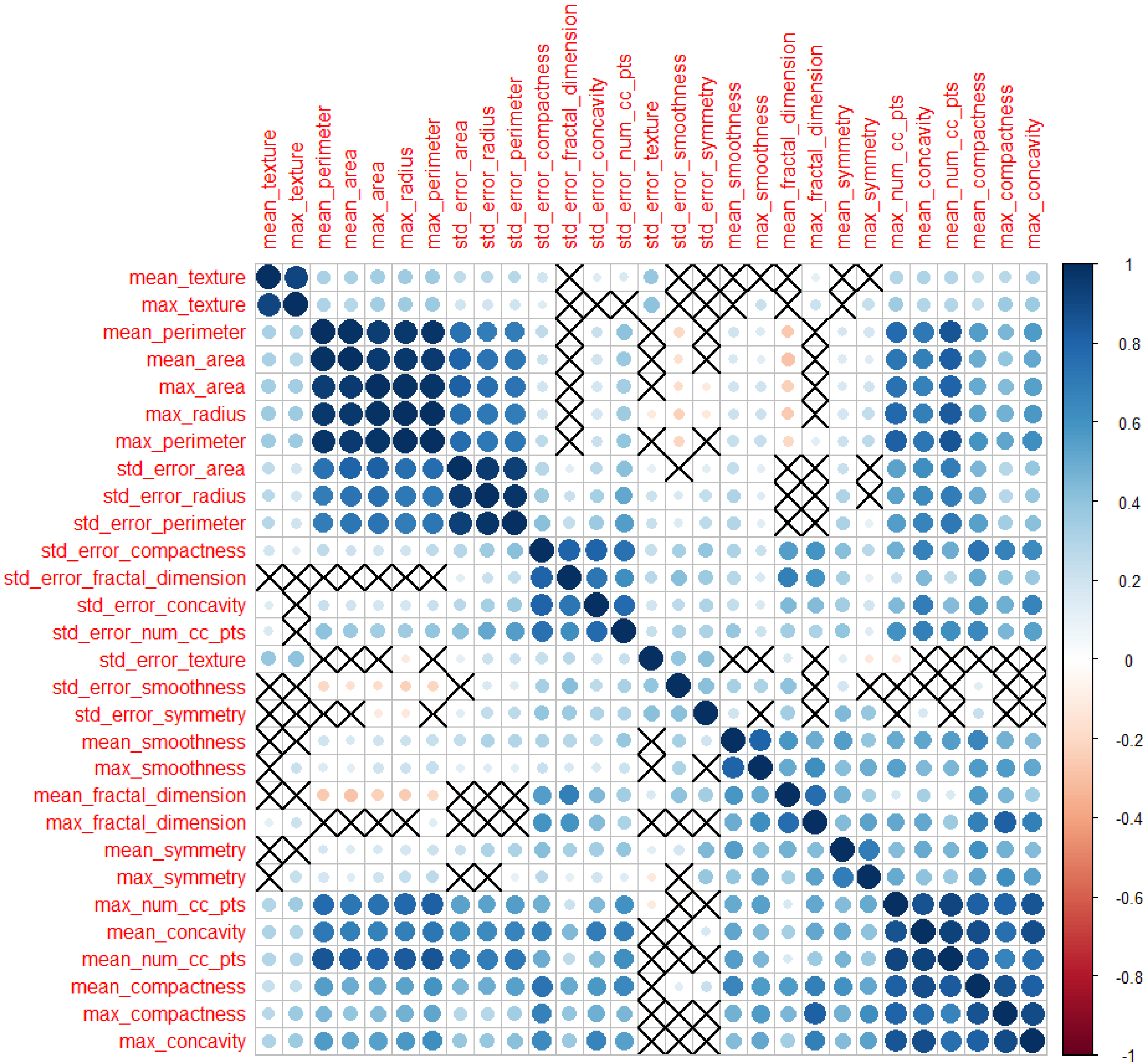
Check for collinear variables (two highly correlated independent variables)

Large SEs, Wald's test failed, normally should reject variables

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.543e+02	6.974e+06	0.000	1
mean_radius	-1.043e+03	7.462e+06	0.000	1
mean_texture	1.890e-01	1.612e+05	0.000	1
mean_perimeter	4.712e+01	9.022e+05	0.000	1
mean_area	6.262e+00	2.636e+04	0.000	1
mean_smoothness	1.318e+04	6.640e+07	0.000	1
mean_compactness	-9.352e+03	3.028e+07	0.000	1
mean_concavity	1.418e+02	2.537e+07	0.000	1
mean_num_cc_pts	7.873e+03	3.732e+07	0.000	1
mean_symmetry	5.506e+02	2.851e+07	0.000	1
mean_fractal_dimension	-5.957e+03	5.436e+07	0.000	1

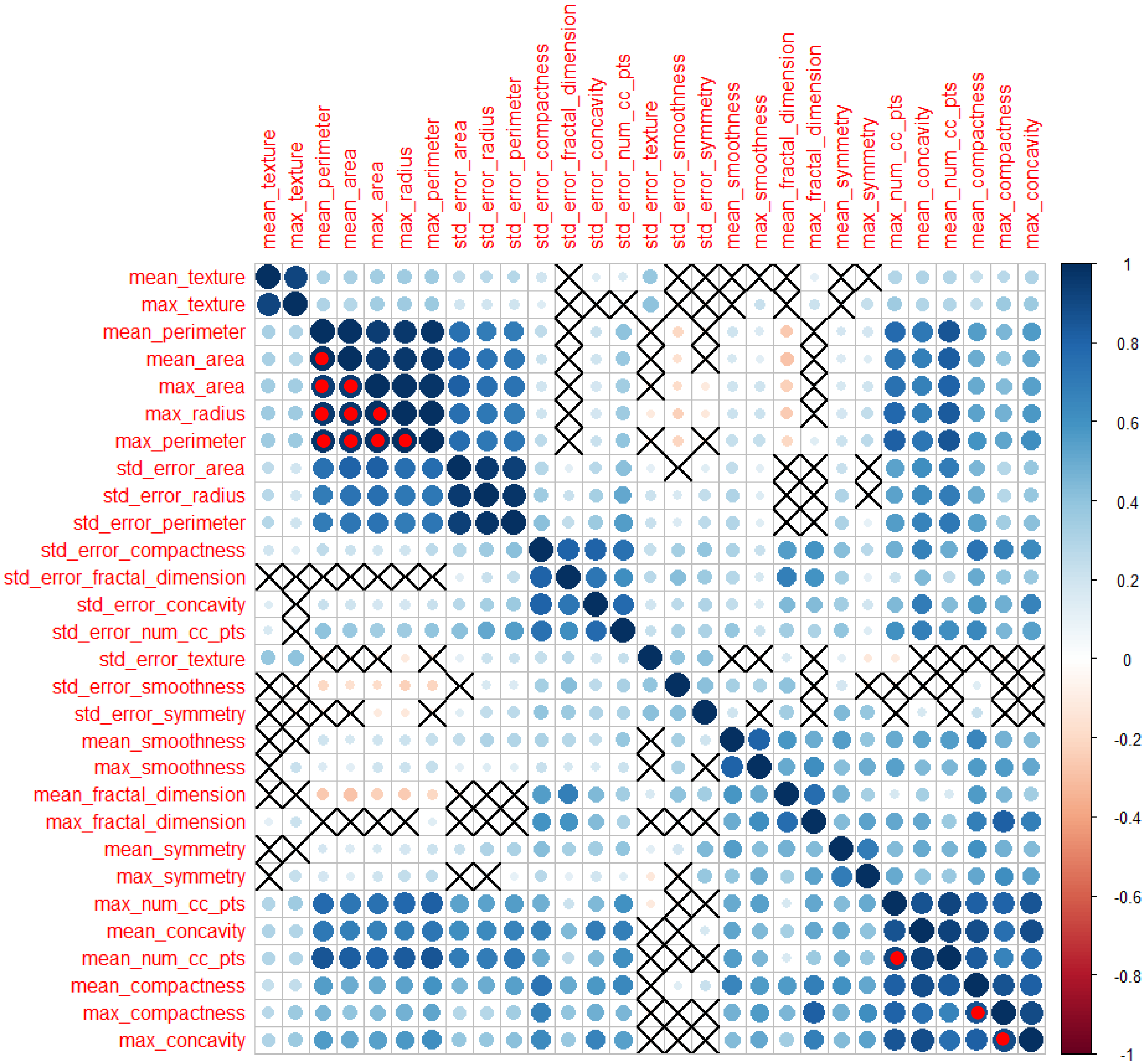
Collinearity?



Collinearity?

Yes

Obscures the attribution of predictive power



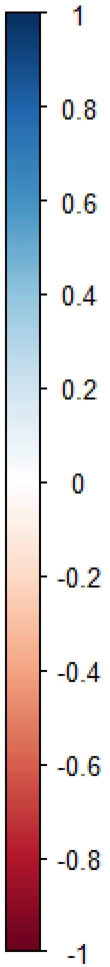
Collinearity?

Yes

Obscures the attribution of predictive power

Remove 1/2 variables with $R^2>0.9$

	mean_texture	mean_concavity	mean_num_cc_pts	mean_perimeter	max_area	max_perimeter	max_radius	std_error_perimeter	std_error_area
mean_texture	1	0.3	0.29	0.33	0.34	0.36	0.35	0.28	0.26
mean_concavity	0.3	1	0.92	0.72	0.68	0.73	0.69	0.66	0.62
mean_num_cc_pts	0.29	0.92	1	0.85	0.81	0.86	0.83	0.71	0.69
mean_perimeter	0.33	0.72	0.85	1	0.94	0.97	0.97	0.69	0.74
max_area	0.34	0.68	0.81	0.94	1	0.98	0.98	0.73	0.81
max_perimeter	0.36	0.73	0.86	0.97	0.98	1	0.99	0.72	0.76
max_radius	0.35	0.69	0.83	0.97	0.98	0.99	1	0.7	0.76
std_error_perimeter	0.28	0.66	0.71	0.69	0.73	0.72	0.7	1	0.94
std_error_area	0.26	0.62	0.69	0.74	0.81	0.76	0.76	0.94	1



Collinearity?

Yes

Obscures the attribution of predictive power

Remove 1/2 variables with $R^2>0.9$

Seems to have fixed the issue

```
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -5.988e+01  5.753e+01  -1.041   0.2979
mean_radius             -3.389e+01  2.408e+01  -1.407   0.1593
mean_texture            -5.206e-01  7.217e-01  -0.721   0.4707
mean_perimeter           4.174e+00  3.345e+00   1.248   0.2121
mean_area                9.365e-02  8.365e-02   1.120   0.2629
mean_smoothness          3.800e+02  2.472e+02   1.537   0.1242
mean_compactness        -2.193e+02  1.777e+02  -1.234   0.2171
mean_concavity           1.153e+02  7.464e+01   1.544   0.1225
mean_symmetry           -1.210e+02  8.925e+01  -1.356   0.1752
mean_fractal_dimension   3.500e+00  4.015e+02   0.009   0.9930
std_error_radius         1.132e+02  5.888e+01   1.922   0.0546 .
std_error_texture       -5.939e+00  3.963e+00  -1.499   0.1339
std_error_perimeter     -9.945e+00  6.602e+00  -1.506   0.1319
std_error_smoothness     1.118e+03  8.795e+02   1.271   0.2036
std_error_compactness    9.409e+02  4.750e+02   1.981   0.0476 *
std_error_concavity     -3.565e+02  1.757e+02  -2.029   0.0425 *
std_error_num_cc_pts     7.813e+02  6.566e+02   1.190   0.2341
std_error_symmetry      -6.131e+02  3.439e+02  -1.783   0.0746 .
std_error_fractal_dimension -8.216e+03  4.134e+03  -1.987   0.0469 *
max_texture              1.247e+00  7.610e-01   1.638   0.1013
max_smoothness          -1.902e+02  1.597e+02  -1.191   0.2338
max_compactness         -1.210e+02  6.749e+01  -1.793   0.0730 .
max_concavity            3.620e+01  2.872e+01   1.261   0.2074
max_num_cc_pts           1.287e+02  1.042e+02   1.235   0.2168
max_symmetry             1.172e+02  5.815e+01   2.016   0.0438 *
max_fractal_dimension    8.600e+02  4.060e+02   2.118   0.0341 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Collinearity?

Yes

Obscures the attribution of predictive power

Remove 1/2 variables with $R^2>0.9$

Seems to have fixed the issue

Strong and significant effectors emerge, although predictor data scaling may affect coefficient magnitudes

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.988e+01	5.753e+01	-1.041	0.2979	
mean_radius	-3.389e+01	2.408e+01	-1.407	0.1593	
mean_texture	-5.206e-01	7.217e-01	-0.721	0.4707	
mean_perimeter	4.174e+00	3.345e+00	1.248	0.2121	
mean_area	9.365e-02	8.365e-02	1.120	0.2629	
mean_smoothness	3.800e+02	2.472e+02	1.537	0.1242	
mean_compactness	-2.193e+02	1.777e+02	-1.234	0.2171	
mean_concavity	1.153e+02	7.464e+01	1.544	0.1225	
mean_symmetry	-1.210e+02	8.925e+01	-1.356	0.1752	
mean_fractal_dimension	3.500e+00	4.015e+02	0.009	0.9930	
std_error_radius	1.132e+02	5.888e+01	1.922	0.0546	.
std_error_texture	-5.939e+00	3.963e+00	-1.499	0.1339	
std_error_perimeter	-9.945e+00	6.602e+00	-1.506	0.1319	
std_error_smoothness	1.118e+03	8.795e+02	1.271	0.2036	
std_error_compactness	9.409e+02	4.750e+02	1.981	0.0476	*
std_error_concavity	-3.565e+02	1.757e+02	-2.029	0.0425	*
std_error_num_cc_pts	7.813e+02	6.566e+02	1.190	0.2341	
std_error_symmetry	-6.131e+02	3.439e+02	-1.783	0.0746	.
std_error_fractal_dimension	-8.216e+03	4.134e+03	-1.987	0.0469	*
max_texture	1.247e+00	7.610e-01	1.638	0.1013	
max_smoothness	-1.902e+02	1.597e+02	-1.191	0.2338	
max_compactness	-1.210e+02	6.749e+01	-1.793	0.0730	.
max_concavity	3.620e+01	2.872e+01	1.261	0.2074	
max_num_cc_pts	1.287e+02	1.042e+02	1.235	0.2168	
max_symmetry	1.172e+02	5.815e+01	2.016	0.0438	*
max_fractal_dimension	8.600e+02	4.060e+02	2.118	0.0341	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Collinearity?

Yes

Obscures the attribution of predictive power

Remove 1/2 variables with R^2>0.9

Seems to have fixed the issue

Strong and significant effectors emerge, although predictor data scaling may affect coefficient magnitudes

log odds of M

Intercept

β

Variable

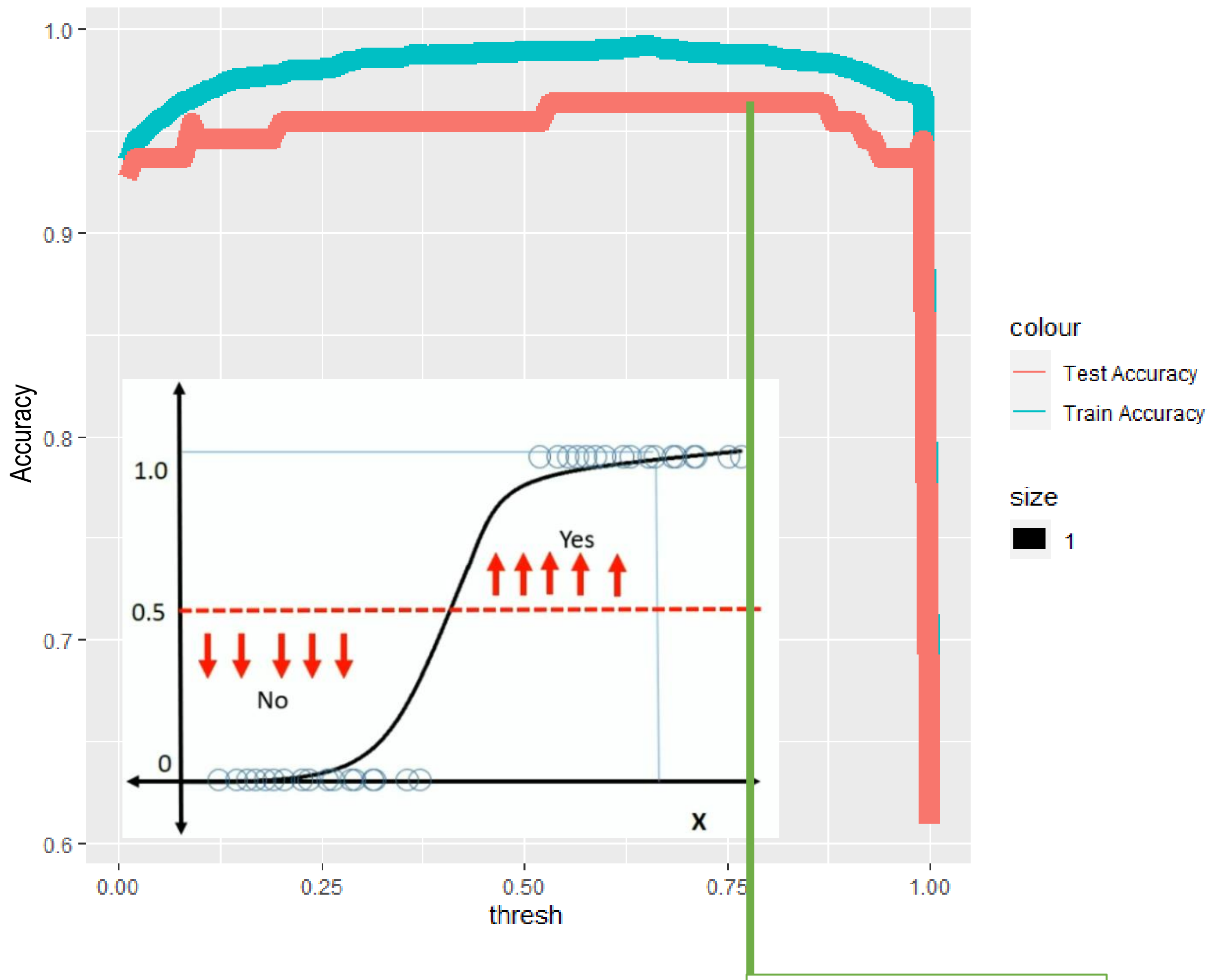
β of 941 means $\ln\left(\frac{p}{1-p}\right) = -599 + (941 * SE_{compactness})$

and let L $= -599 + (941 * SE_{compactness})$

So the effect on probability is $p = \frac{\exp(L)}{\exp(L)+1}$

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -5.988e+01  5.753e+01  -1.041   0.2979
std_error_compactness 9.409e+02  4.750e+02   1.981   0.0476 *
```

Threshold optimization

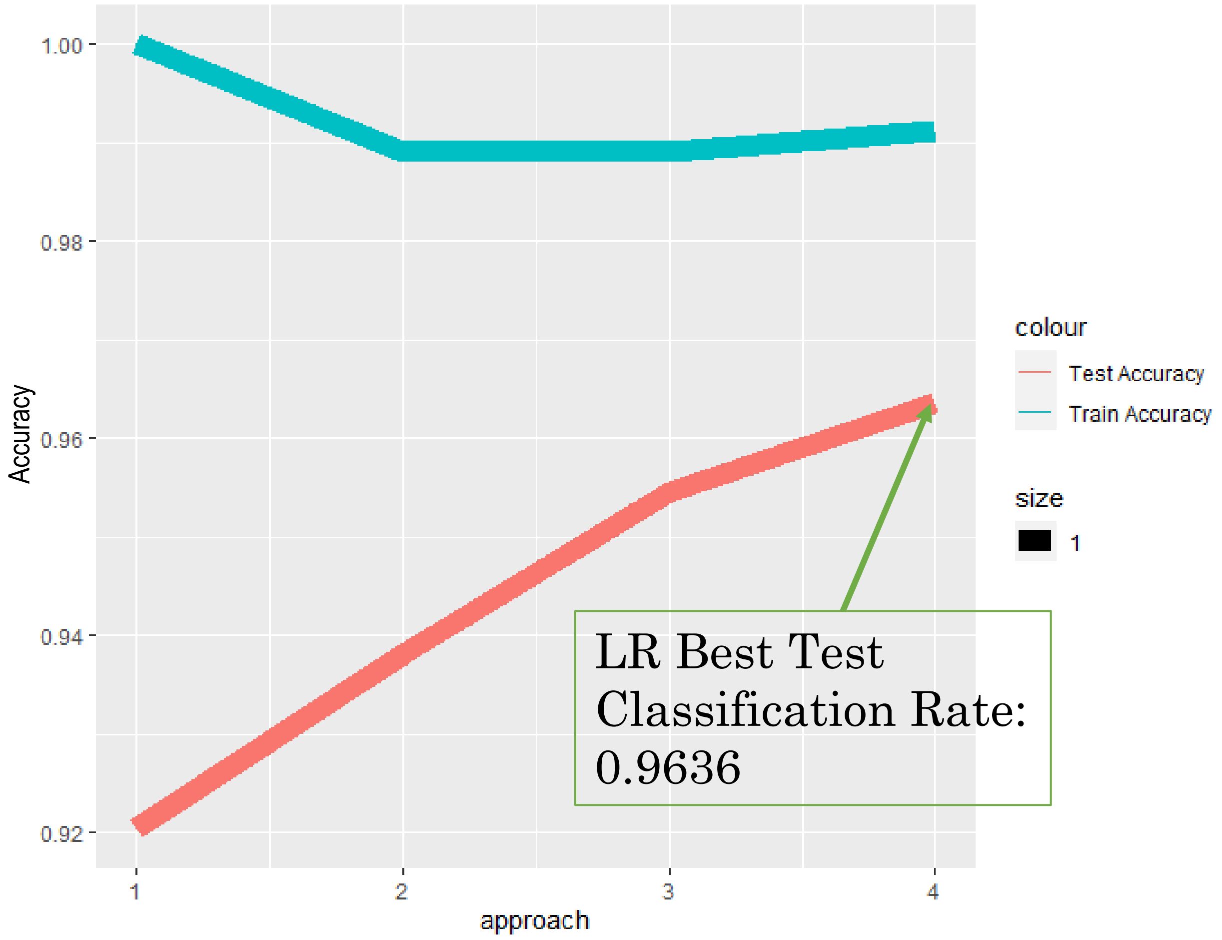


$T_{optimized} = 0.77$

Logistic Regression Results

Approaches:

1. 30 predictors, no CV, T = 0.5
2. 1, sans ½ collinear predictors
3. 2, with 10-fold CV
4. 3, with $T_{optimized} = 0.77$



SVM Results

Accuracy of *tuned* linear and radial SVMs over costs and gammas range with 10-fold CV

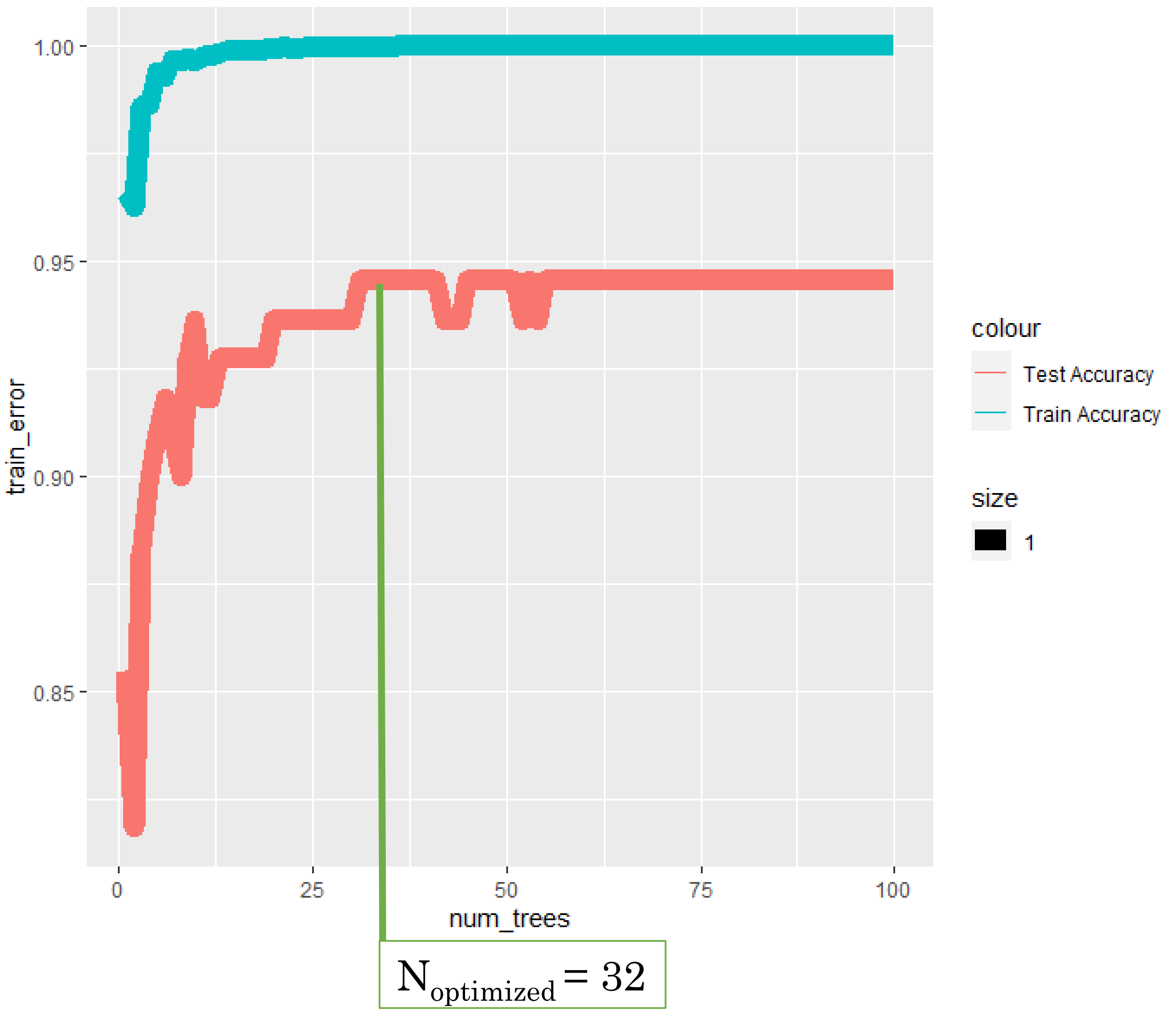
Linear SVM: **0.930** with cost=0.1 and gamma = 1e-5

Radial SVM: **0.960** with cost = 10 and gamma = 0.1

Random Forest Results:
10-fold CV and N_{trees} optimization

Input data is reduced dataset after removal of collinear variables

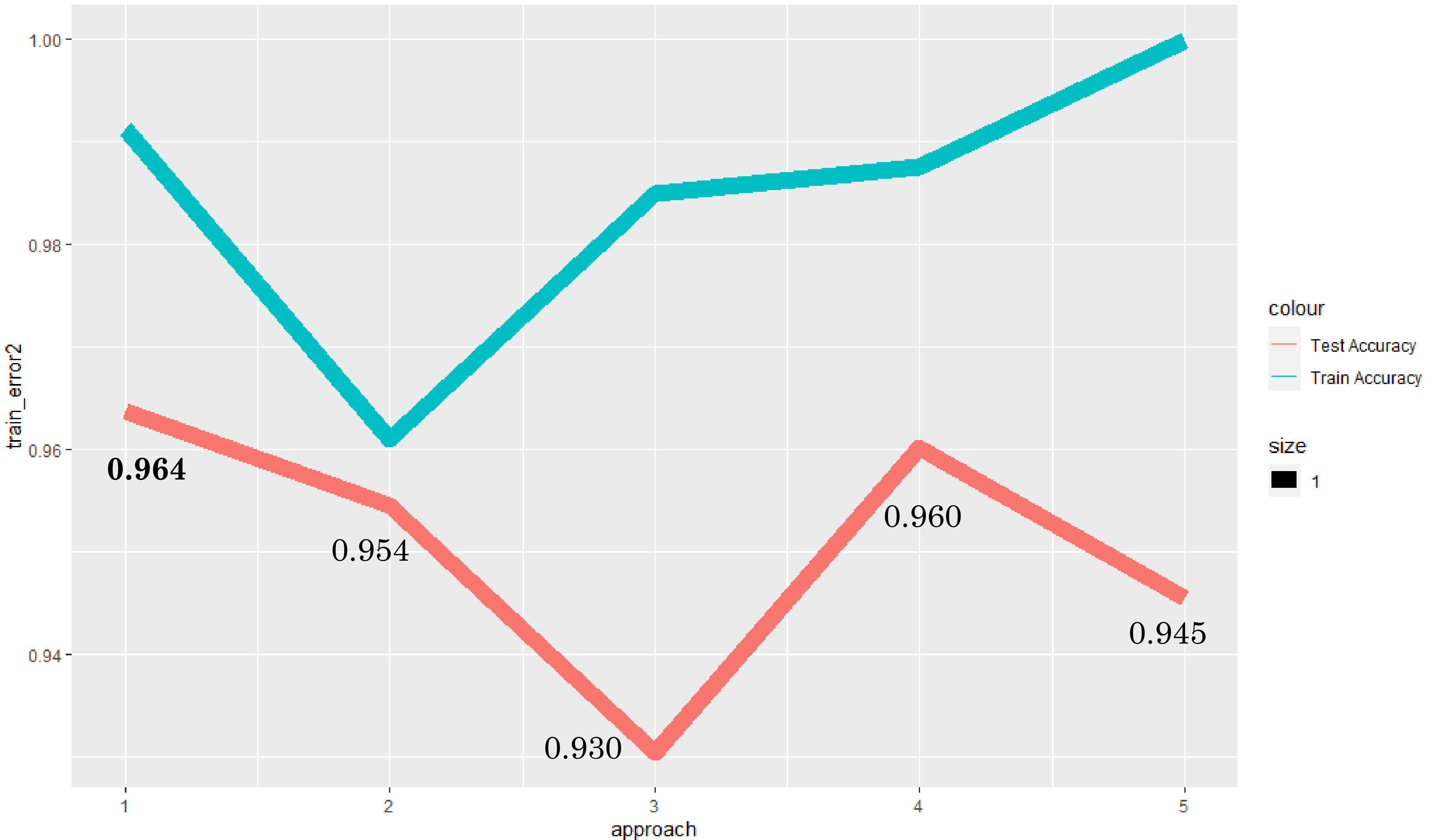
Classification accuracy with optimization was **0.945**



All Approaches Comparison

Approaches:

1. **Logistic Regression**
2. QDA
3. Linear SVM
4. Radial SVM
5. Random Forest



With another month,

- Consider penalties/regularization for logistic regression to further reduce overfitting
- Normalize scales of all independent variables
  - Most are currently within about one order of magnitude, but normalizing all to one scale will give better results of coefficient interpretation
- Correlate to other aspects relating to breast cancer detection such as *ESR1* gene which, when expression is low, indicates poor survival outcomes.