BREAST CANCER PREDICTOR

ANTHONY MUNOZ

5/22/2020

ABSTRACT

- Predict if the tumor is benign or malignant
- Using regression modeling in order to predict the variables relationship to predict the tumor cells outcome

KEYWORDS

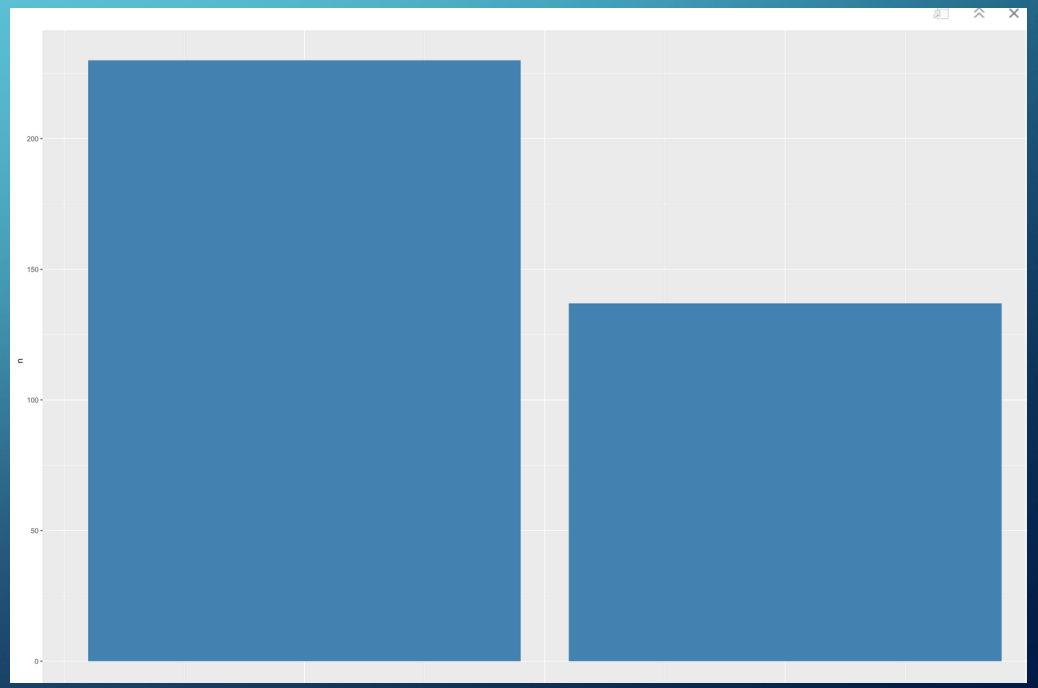
- Breast Cancer, tumor cells, regression.

DATA EXPLORATION

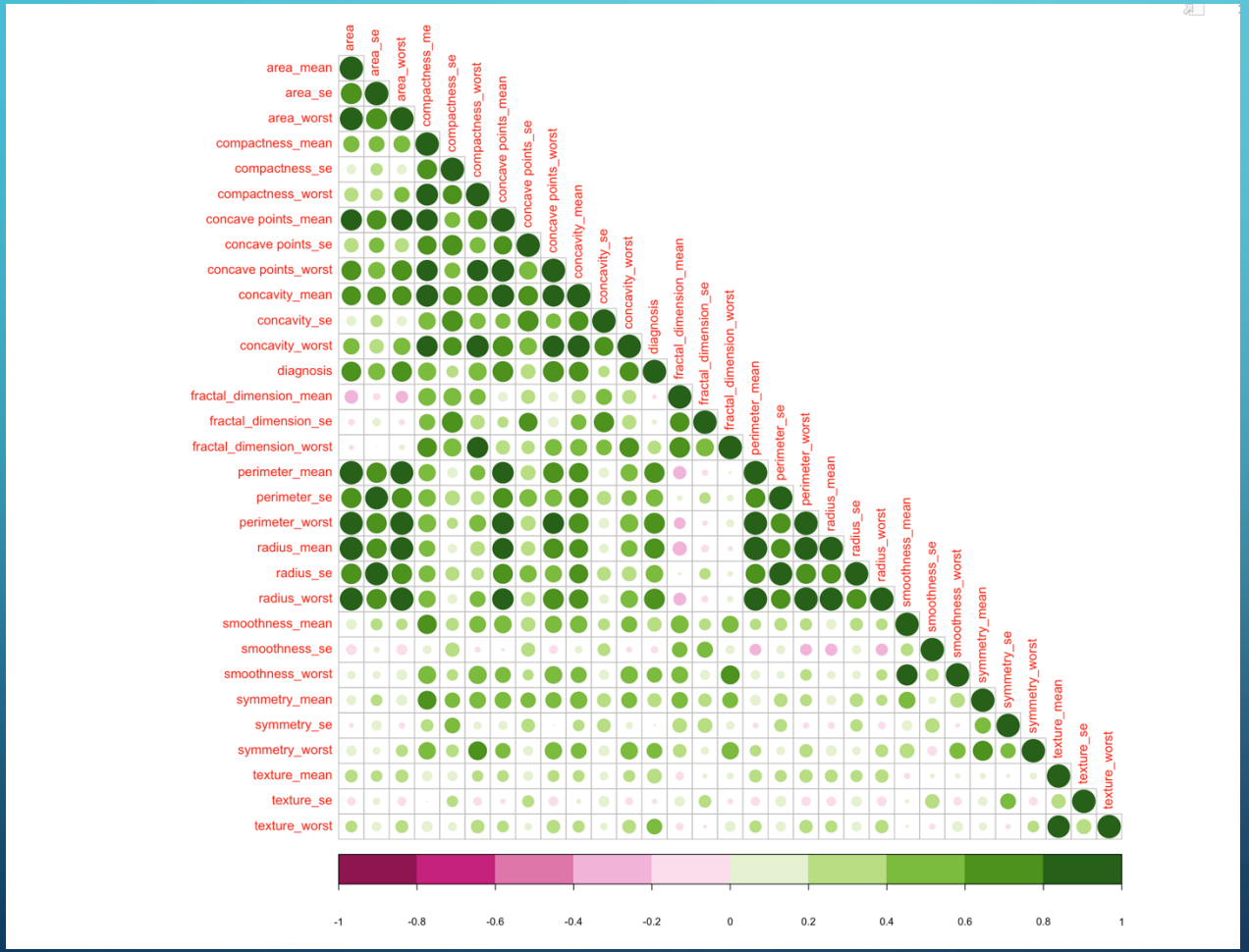
```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       367 obs. of  31 variables:
 $ diagnosis              : num  1 1 1 1 1 1 1 1 1 1 ...
 $ radius_mean            : num  18 19.7 11.4 12.4 13 ...
 $ texture_mean           : num  10.4 21.2 20.4 15.7 21.8 ...
 $ perimeter_mean         : num  122.8 130 77.6 82.6 87.5 ...
 $ area_mean              : num  1001 1203 386 477 520 ...
 $ smoothness_mean        : num  0.118 0.11 0.142 0.128 0.127 ...
 $ compactness_mean       : num  0.278 0.16 0.284 0.17 0.193 ...
 $ concavity_mean         : num  0.3 0.197 0.241 0.158 0.186 ...
 $ concave points_mean    : num  0.1471 0.1279 0.1052 0.0809 0.0935 ...
 $ symmetry_mean          : num  0.242 0.207 0.26 0.209 0.235 ...
 $ fractal_dimension_mean : num  0.0787 0.06 0.0974 0.0761 0.0739 ...
 $ radius_se              : num  1.095 0.746 0.496 0.335 0.306 ...
 $ texture_se             : num  0.905 0.787 1.156 0.89 1.002 ...
 $ perimeter_se           : num  8.59 4.58 3.44 2.22 2.41 ...
 $ area_se                : num  153.4 94 27.2 27.2 24.3 ...
 $ smoothness_se          : num  0.0064 0.00615 0.00911 0.00751 0.00573 ...
 $ compactness_se         : num  0.049 0.0401 0.0746 0.0335 0.035 ...
 $ concavity_se           : num  0.0537 0.0383 0.0566 0.0367 0.0355 ...
 $ concave points_se      : num  0.0159 0.0206 0.0187 0.0114 0.0123 ...
 $ symmetry_se            : num  0.03 0.0225 0.0596 0.0216 0.0214 ...
 $ fractal_dimension_se   : num  0.00619 0.00457 0.00921 0.00508 0.00375 ...
 $ radius_worst           : num  25.4 23.6 14.9 15.5 15.5 ...
 $ texture_worst          : num  17.3 25.5 26.5 23.8 30.7 ...
 $ perimeter_worst        : num  184.6 152.5 98.9 103.4 106.2 ...
 $ area_worst             : num  2019 1709 568 742 739 ...
```

- The dataset contains 570 rows and 33 columns.
- Most of the variables are numeric except for the diagnosis variables

- Begins vs malignant



CORRELATION PLOT



THE BEST 2 MODELS

MODEL 1

```
##
## Call:
## glm(formula = diagnosis ~ texture_mean + compactness_mean + concavity_mean +
##     radius_se + area_se + smoothness_se + compactness_se + concavity_se +
##     `concave points_se` + radius_worst + area_worst + compactness_worst +
##     symmetry_worst + fractal_dimension_worst, data = df.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6847  -0.1642  -0.0265   0.1183   0.8475
##
 ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 ##
 ## (Dispersion parameter for gaussian family taken to be 0.05904078)
 ##
 ##     Null deviance: 86.109  on 366  degrees of freedom
 ## Residual deviance: 20.782  on 352  degrees of freedom
 ## AIC: 19.749
 ##
 ## Number of Fisher Scoring iterations: 2
```

MODEL 2

```
##
## Call:
## glm(formula = diagnosis ~ radius_mean + perimeter_mean + compactness_mean +
##     fractal_dimension_mean + perimeter_se + area_se + concavity_se +
##     `concave points_se` + fractal_dimension_se + texture_worst +
##     perimeter_worst + smoothness_worst + concavity_worst + symmetry_worst,
##     family = "binomial", data = df.train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.253e-03  -2.000e-08  -2.000e-08   2.000e-08   1.159e-03
 ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 ##
 ## (Dispersion parameter for gaussian family taken to be 0.05904078)
 ##
 ##     Null deviance: 86.109  on 366  degrees of freedom
 ## Residual deviance: 20.782  on 352  degrees of freedom
 ## AIC: 19.749
 ##
 ## Number of Fisher Scoring iterations: 2
```

SELECTED MODEL

```
##
## Call:
## glm(formula = diagnosis ~ radius_mean + perimeter_mean + compactness_mean +
##     fractal_dimension_mean + perimeter_se + area_se + concavity_se +
##     `concave points_se` + fractal_dimension_se + texture_worst +
##     perimeter_worst + smoothness_worst + concavity_worst + symmetry_worst,
##     family = "binomial", data = df.train)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -1.253e-03  -2.000e-08  -2.000e-08   2.000e-08   1.159e-03
##
## (Dispersion parameter for gaussian family taken to be 0.05904078)
##
##     Null deviance: 86.109  on 366  degrees of freedom
## Residual deviance: 20.782  on 352  degrees of freedom
## AIC: 19.749
##
## Number of Fisher Scoring iterations: 2
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 228  13
##          1   1 125
##
##                Accuracy : 0.9619
##                  95% CI : (0.9368, 0.979)
##     No Information Rate : 0.624
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9173
##
##  Mcnemar's Test P-Value : 0.003283
##
##             Sensitivity : 0.9956
##             Specificity : 0.9058
##          Pos Pred Value : 0.9461
##          Neg Pred Value : 0.9921
##              Prevalence : 0.6240
##          Detection Rate : 0.6213
##    Detection Prevalence : 0.6567
##       Balanced Accuracy : 0.9507
##
##        'Positive' Class : 0
##
```

CONCLUSION

- Glm regression model accuracy of 95 %
- 11 variables are the most significant predictors
- The final model just lower a a few points against the evaluation.

Thank You