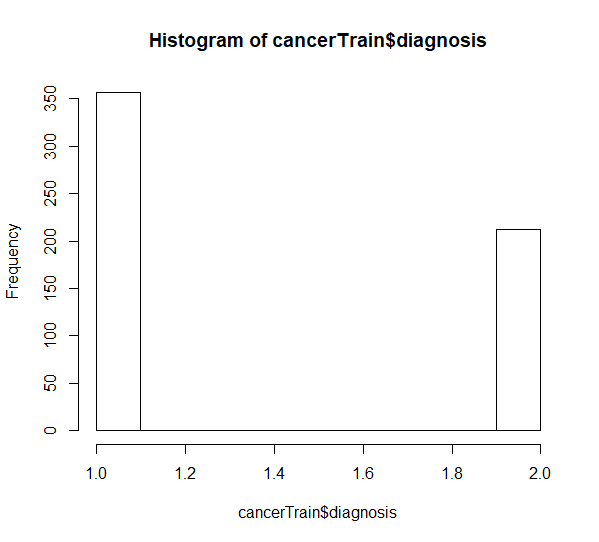
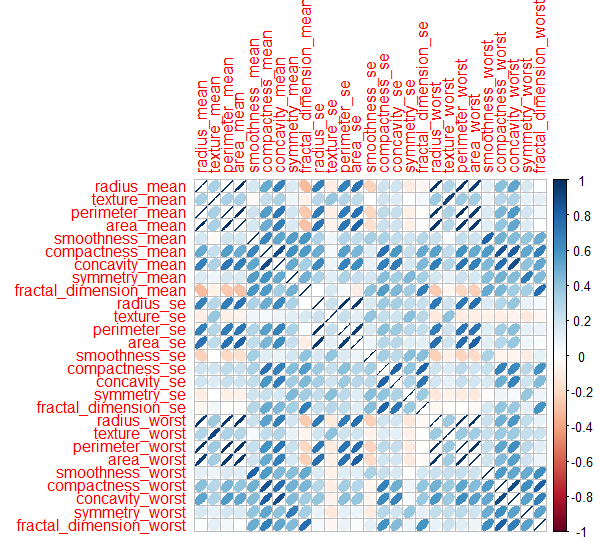
**a. Logistic Regression Approach**

• Exploring features and Data Preparation which includes missing value treatment and

Outlier Detection

• Visualizing relationships among features





• Split the data into train and test data and build sophisticated Machine Learning models

Train data.

Total = 569 | Test data(20%) = 114

TRUE

1 88

2 26

|  |  |  |
| --- | --- | --- |
|  | FALSE | TRUE |
| 1 | 26 | 88 |
| 2 | 88 | 26 |

1=> Benine | 2=> Malaignant | No.of Benine (B) is more

• Evaluating Model performance on test data using Precision, Recall, Accuracy and ROC

curve metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | FALSE | TRUE | total | False Negative(FN) = 86/114 = 0.77 |
| Actually NO | 26 | 88 | 114 | False Positive(FP) = 86/114 = 0.77 |
| Actually YES | 88 | 26 | 114 | True Positive(TP) = 26/114 = 0.23 |
| total | 114 | 114 |  | True Negative(TN) = 26/114 = 0.23 |

**Precision** (P) =TP/(TP+FP) = 0.23/(0.23+0.77) **= 0.23**

**Recall (**R} =TP/(TP+FN) = 0.23/(0.23+0.77) **= 0.23**

**Accuracy (Residual deviance) = 23.9% ROC Curve**

• Determining the factors driving the cancer. **Features highlighted in pink colour have linear correlation with diagnosis**

> summary(logistic)

Call:

glm(formula = diagnosis ~ radius\_mean + texture\_mean + perimeter\_mean +

area\_mean + smoothness\_mean + compactness\_mean + concavity\_mean +

concave.points\_mean + symmetry\_mean + fractal\_dimension\_mean +

radius\_se + texture\_se + perimeter\_se + area\_se + smoothness\_se +

compactness\_se + concavity\_se + concave.points\_se + symmetry\_se +

fractal\_dimension\_se + radius\_worst + texture\_worst + perimeter\_worst +

area\_worst + smoothness\_worst + compactness\_worst + concavity\_worst +

concave.points\_worst + symmetry\_worst + fractal\_dimension\_worst,

data = train\_proj)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.6191 -0.1572 -0.0317 0.1312 0.7720

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.252e+00 4.955e-01 -2.527 0.011869 \*

radius\_mean -1.659e-01 1.935e-01 -0.857 0.391701

texture\_mean 1.726e-02 9.208e-03 1.874 0.061572 .

perimeter\_mean 1.550e-02 2.868e-02 0.540 0.589195

area\_mean 3.466e-04 5.777e-04 0.600 0.548821

smoothness\_mean 8.971e-01 2.357e+00 0.381 0.703635

compactness\_mean -3.950e+00 1.576e+00 -2.507 0.012565 \*

concavity\_mean 1.220e+00 1.191e+00 1.025 0.306108

concave.points\_mean 1.269e+00 2.266e+00 0.560 0.575802

symmetry\_mean -2.431e-01 8.333e-01 -0.292 0.770596

fractal\_dimension\_mean 1.303e+00 6.188e+00 0.211 0.833361

radius\_se 3.682e-01 3.414e-01 1.079 0.281367

texture\_se 2.331e-02 4.309e-02 0.541 0.588880

perimeter\_se -2.901e-02 4.693e-02 -0.618 0.536723

area\_se -6.977e-05 1.731e-03 -0.040 0.967870

smoothness\_se 1.413e+01 7.208e+00 1.961 0.050552 .

compactness\_se -8.304e-01 2.358e+00 -0.352 0.724835

concavity\_se -3.579e+00 1.449e+00 -2.470 0.013905 \*

concave.points\_se 1.200e+01 6.467e+00 1.856 0.064125 .

symmetry\_se 5.403e-01 3.011e+00 0.179 0.857661

fractal\_dimension\_se -4.664e+00 1.249e+01 -0.373 0.709083

radius\_worst 1.877e-01 6.159e-02 3.047 0.002456 \*\*

texture\_worst -9.272e-04 7.903e-03 -0.117 0.906659

perimeter\_worst 1.203e-03 6.612e-03 0.182 0.855754

area\_worst -1.147e-03 3.427e-04 -3.346 0.000893 \*\*\*

smoothness\_worst 8.060e-01 1.592e+00 0.506 0.612860

compactness\_worst 1.793e-01 4.123e-01 0.435 0.663934

concavity\_worst 4.139e-01 3.006e-01 1.377 0.169314

concave.points\_worst 6.232e-01 1.027e+00 0.607 0.544219

symmetry\_worst 5.287e-01 5.551e-01 0.952 0.341397

fractal\_dimension\_worst 3.485e+00 2.562e+00 1.360 0.174573

---

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

(Dispersion parameter for gaussian family taken to be 0.05648101)

Null deviance: 109.965 on 454 degrees of freedom

Residual deviance: 23.948 on 424 degrees of freedom

AIC: 15.525

Number of Fisher Scoring iterations: 2

• Choosing best model based on the accuracy and other measures.

**Logistic Regression (Accuracy) Residual deviance ~23.9%**

b. **Logistic Regression Approach**

**5. Problem Statement**

1. Build Machine Learning Models to predict the type of Breast Cancer (Malignant or

Benign) as well as identify the drivers of cancer.

Apply the concepts of - Logistic Regression and Random Forest.

Note: You need to submit associate R project along with the screenshot.