# Classification using Logistic Regression

**Author:** Bharat Thakur

## Background

Analyze the tumor dataset, which contains medical information used in the pre-screening diagnosis of tumors. First logistic regression is used to determine the factors that predict the probability of tumor diagnosis. Then naïve Bayes classifier and linear discriminant analysis are used for the classification task. Finally, the performance of the three classifiers is compared.

## Part 1

## 1. Data Transformation and Cleaning (Description)

### Rename

The columns of the tumor dataset were renamed for easier interpretation.

```
names(Tumor_BT)

 [1] "Outcome_BT"          "Age_BT"              "Sex_BT"
 [4] "Bone_Density_BT"     "Bone_Marrow_BT"      "Lung_Spot_BT"
 [7] "Pleura_BT"           "Liver_Spot_BT"       "Brain_Scan_BT"
[10] "Skin_Lesions_BT"     "Stiff_Neck_BT"       "Supraclavicular_BT"
[13] "Axillar_BT"          "Mediastinum_BT"
```

### Normalization

All the columns were normalized to have values in set {0,1}. Additionally, values in Supraclavicular and Axillar columns were swapped so that 0 and 1 consistently represent “No” and “Yes” respectively. The updated Data Dictionary is shown below:

**Data Dictionary (After Transformation)**

| Name | Description |
|---|---|
| Out | Tumor is present=1, Is not preset=0 |
| Age | Older =1, Younger=0 |
| Sex | Male=1, Female=0 |
| Bone | Bone Density Test: Good=0, Bad=1 |
| Marrow | Bone Marrow: Good=0, Bad=1 |
| Lung | Spot on Lung: Yes=1, No=0 |
| Pleura | Pleura: Yes=1, No=0 |
| Liver | Spot on Liver: Yes=1, No=0 |
| Brain | Brain Scan: Yes=1, No=0 |
| Skin | Lesions: Yes=1, No=0 |

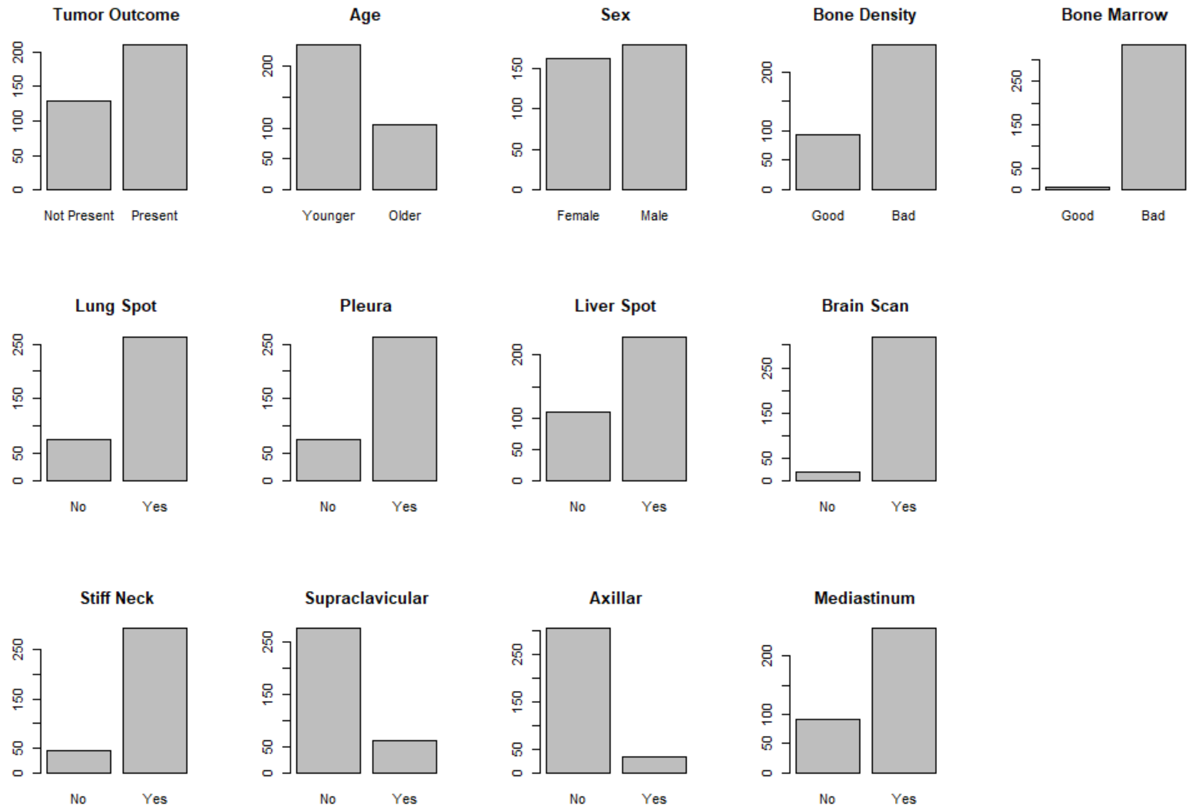
| | |
|---|---|
| Neck | Stiff Neck? Yes=1, No=0 |
| Supra | Supraclavicular: Yes=1, No=0 |
| Axil | Axillar: Yes=1, No=0 |
| Media | Mediastinum: Yes=1, No=0 |

## 2. Descriptive Data Analysis

### Quantitative Summary

```
   Outcome_BT         Age_BT           Sex_BT      Bone_Density_BT
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :1.0000   Median :0.0000   Median :1.0000   Median :1.0000  
 Mean   :0.6195   Mean   :0.3097   Mean   :0.5251   Mean   :0.7227  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
 Bone_Marrow_BT    Lung_Spot_BT      Pleura_BT      Liver_Spot_BT   
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000  
 Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000  
 Mean   :0.9823   Mean   :0.7758   Mean   :0.7758   Mean   :0.6755  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
 Brain_Scan_BT    Skin_Lesions_BT  Stiff_Neck_BT    Supraclavicular_BT
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000    
 1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000    
 Median :1.0000   Median :1.0000   Median :1.0000   Median :0.0000    
 Mean   :0.9381   Mean   :0.9351   Mean   :0.8673   Mean   :0.1829    
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000    
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000    
   Axillar_BT      Mediastinum_BT  
 Min.   :0.00000   Min.   :0.0000  
 1st Qu.:0.00000   1st Qu.:0.0000  
 Median :0.00000   Median :1.0000  
 Mean   :0.09735   Mean   :0.7286  
 3rd Qu.:0.00000   3rd Qu.:1.0000  
 Max.   :1.00000   Max.   :1.0000  
```

### Graphical Summary



**Comment**: The dataset doesn't have an equal representation from older people, as such results can only be generalized to a population with similar sample statistics.

## Exploratory Data Analysis

### 2.1 Correlations

```
                   Outcome_BT Age_BT Sex_BT Bone_Density_BT Bone_Marrow_BT
Outcome_BT               1.00   0.08   0.29            0.18           0.03
Age_BT                   0.08   1.00   0.10            0.20           0.04
Sex_BT                   0.29   0.10   1.00            0.08           0.01
Bone_Density_BT          0.18   0.20   0.08            1.00           0.17
Bone_Marrow_BT           0.03   0.04   0.01            0.17           1.00
Lung_Spot_BT             0.02  -0.04   0.00            0.08          -0.02
Pleura_BT               -0.01   0.07  -0.07            0.00          -0.02
Liver_Spot_BT           -0.10  -0.22  -0.04           -0.18           0.00
Brain_Scan_BT            0.15   0.07   0.05            0.06           0.06
Skin_Lesions_BT          0.11   0.15   0.04            0.10          -0.04
Stiff_Neck_BT            0.28   0.07   0.22           -0.01          -0.05
Supraclavicular_BT      -0.16  -0.14  -0.05            0.00           0.06
Axillar_BT               0.07  -0.18   0.13           -0.06          -0.03
Mediastinum_BT           0.34   0.04   0.11            0.08           0.07
```

```
                        Lung_Spot_BT Pleura_BT Liver_Spot_BT Brain_Scan_BT
Outcome_BT                      0.02     -0.01         -0.10          0.15
Age_BT                         -0.04      0.07         -0.22          0.07
Sex_BT                          0.00     -0.07         -0.04          0.05
Bone_Density_BT                 0.08      0.00         -0.18          0.06
Bone_Marrow_BT                 -0.02     -0.02          0.00          0.06
Lung_Spot_BT                    1.00      0.08          0.10          0.01
Pleura_BT                       0.08      1.00         -0.01          0.01
Liver_Spot_BT                   0.10     -0.01          1.00         -0.05
Brain_Scan_BT                   0.01      0.01         -0.05          1.00
Skin_Lesions_BT                 0.00      0.00          0.00          0.08
Stiff_Neck_BT                  -0.04     -0.06         -0.20         -0.06
Supraclavicular_BT              0.00      0.02          0.10         -0.01
Axillar_BT                     -0.01      0.06          0.10         -0.04
Mediastinum_BT                  0.21      0.17          0.13          0.26
                   Skin_Lesions_BT Stiff_Neck_BT Supraclavicular_BT Axillar_BT
Outcome_BT                    0.11          0.28              -0.16       0.07
Age_BT                        0.15          0.07              -0.14      -0.18
Sex_BT                        0.04          0.22              -0.05       0.13
Bone_Density_BT               0.10         -0.01               0.00      -0.06
Bone_Marrow_BT               -0.04         -0.05               0.06      -0.03
Lung_Spot_BT                  0.00         -0.04               0.00      -0.01
Pleura_BT                     0.00         -0.06               0.02       0.06
Liver_Spot_BT                 0.00         -0.20               0.10       0.10
Brain_Scan_BT                 0.08         -0.06              -0.01      -0.04
Skin_Lesions_BT               1.00          0.11              -0.06      -0.24
Stiff_Neck_BT                 0.11          1.00              -0.29      -0.16
Supraclavicular_BT           -0.06         -0.29               1.00       0.31
Axillar_BT                   -0.24         -0.16               0.31       1.00
Mediastinum_BT               -0.03         -0.02              -0.17      -0.05
                   Mediastinum_BT
Outcome_BT                   0.34
Age_BT                       0.04
Sex_BT                       0.11
Bone_Density_BT              0.08
Bone_Marrow_BT               0.07
Lung_Spot_BT                 0.21
Pleura_BT                    0.17
Liver_Spot_BT                0.13
Brain_Scan_BT                0.26
Skin_Lesions_BT             -0.03
Stiff_Neck_BT               -0.02
Supraclavicular_BT          -0.17
Axillar_BT                  -0.05
Mediastinum_BT               1.00
```

Comment: Out of all the variables Mediastinum and Sex have the highest correlation coefficients. Stiff Neck is the variable with the next highest correlation factor.

In terms of collinearities, the correlation matrix suggests the following relationships:

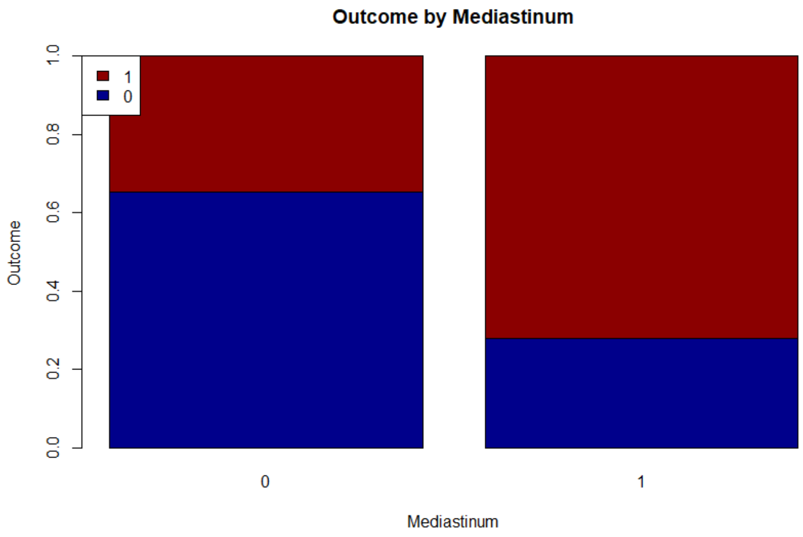
1. Supraclavicular and Stiff Neck, and Supraclavicular and Axillar.
2. Mediastinum and Brain Scan.

## 2.2 Significant Predictors

Mediastinum and Sex appear to be the most significant predictors based on correlation coefficients.

**Mediastinum**: Observed data shows that when Mediastinum is “No”, only 34% of time tumor was present, on the other hand when Mediastinum is “Yes” then 72% of the time tumor was present.

```
       Mediastinum
Outcome         0         1
      0 0.6521739 0.2793522
      1 0.3478261 0.7206478
```



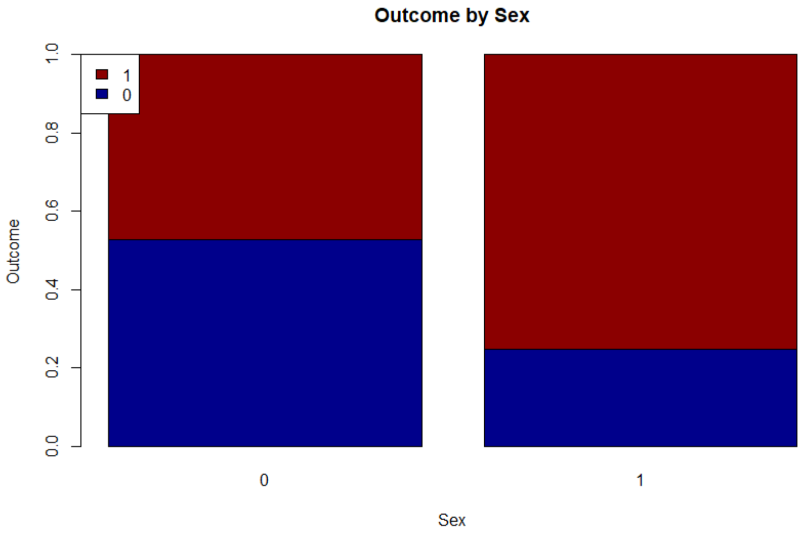
This is further confirmed by the Chi-squared test which results in p-value less than .05, which means the null hypothesis (no relationship) should be rejected. The table also shows how the distribution of outcome should be if there were no relationship which is significantly different from what is observed.

```
Pearson's Chi-squared test

 data:  Tumor_BT$Outcome_BT and Tumor_BT$Mediastinum_BT
 X-squared = 39.526, df = 1, p-value = 0.0000000003238

  chisq_Media_BT$observed   # What we observed

                    Tumor_BT$Mediastinum_BT
 Tumor_BT$Outcome_BT   0   1
                  0   60  69
                  1   32 178

  chisq_Media_BT$expected   # If there were no relationship



                      Tumor_BT$Mediastinum_BT
 Tumor_BT$Outcome_BT        0         1
```

```
                    0 35.00885  93.99115
                    1 56.99115 153.00885
```

**Sex:** Observed data shows that when Sex is "Female", only 47% of the time tumor was present, on the other hand when Sex is "Male" then 75% of the time tumor was present.

```
       Sex
Outcome         0         1
      0 0.5279503 0.2471910
      1 0.4720497 0.7528090
```



Chi-squared test further confirms this. The p-value of less than .05 indicates that the null hypothesis (no relationship) should be rejected. The table also shows how the distribution of outcome should be if there were no relationship which is significantly different from what is observed.

```
Pearson's Chi-squared test

 data:  Tumor_BT$Outcome_BT and Tumor_BT$Sex_BT
 X-squared = 28.269, df = 1, p-value = 0.0000001056

  chisq_Sex_BT$observed   # What we observed

                   Tumor_BT$Sex_BT
 Tumor_BT$Outcome_BT   0   1
                  0  85  44
                  1  76 134

  chisq_Sex_BT$expected   # If there were no relationship



                   Tumor_BT$Sex_BT
 Tumor_BT$Outcome_BT        0         1
```

```
                   0 61.26549  67.73451
                   1 99.73451 110.26549
```

## 3. Model Development

### 3.1 Stepwise Selection Model

```
summary(stp_Out_glm_BT)


 Call:
 glm(formula = Outcome_BT ~ Sex_BT + Bone_Density_BT + Brain_Scan_BT + 
     Skin_Lesions_BT + Stiff_Neck_BT + Supraclavicular_BT + Axillar_BT + 
     Mediastinum_BT, family = "binomial", data = Tumor_BT, na.action = na.omit)

 Deviance Residuals: 
     Min       1Q   Median       3Q      Max  
 -2.6407  -0.8722   0.4821   0.7060   2.8736  

 Coefficients:
                    Estimate Std. Error z value      Pr(>|z|)    
 (Intercept)         -5.6287     0.9663  -5.825 0.00000000571 ***
 Sex_BT               0.8314     0.2744   3.030       0.00244 ** 
 Bone_Density_BT      0.9148     0.2988   3.062       0.00220 ** 
 Brain_Scan_BT        0.9183     0.5619   1.634       0.10217    
 Skin_Lesions_BT      1.3427     0.5608   2.395       0.01664 *  
 Stiff_Neck_BT        2.0327     0.4302   4.725 0.00000230315 ***
 Supraclavicular_BT  -0.7450     0.3839  -1.941       0.05231 .  
 Axillar_BT           2.1933     0.6790   3.230       0.00124 ** 
 Mediastinum_BT       1.6823     0.3035   5.543 0.00000002968 ***
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 (Dispersion parameter for binomial family taken to be 1)

     Null deviance: 450.41  on 338  degrees of freedom
 Residual deviance: 337.81  on 330  degrees of freedom
 AIC: 355.81

 Number of Fisher Scoring iterations: 5
```

**Model Measures**

1. Fisher's Scoring Iteration: Iteration count of 5 indicates the model had no issues in convergence.
2. AIC: Stepwise selection model results in a final AIC of 355.81. AIC of the model with all the variables included was 363.33, so only marginal drop.
3. Deviance: Residual deviance is smaller than Null deviance, meaning the proposed better does a better job at explaining data than the Null model.
4. Residual symmetry: Deviance residuals are almost symmetric.
5. z-values: p-values corresponding to z-tests indicate Mediastinum and Stiff Neck are the most significant variable. The Sex variable which was found to have a relatively high correlation coefficient is not as significant as the Mediastinum and Stiff Neck variables.

6. Variable Co-Efficients: Mediastinum and Stiff Neck coefficients have a positive sign as we observed in the correlation matrix.

For the two user models, the output of the stepwise selection model (the selected variables) was used as the base model. Then variable selection was performed.

### 4.2 User Model 1

For User Model 1 all the variables as in output of stepwise selection model were selected and the Brain Scan variable was removed. Brain Scan was the least significant variable as per the p-values of 0.1 (>.05) in the z-test. Additionally, removing the Brain Scan from the model would only lead to a marginal increase in AIC as the iterations of the stepwise selection model showed. So we will have a simpler model (fewer variables) with similar performance.

```
 summary(Out_UM_1_BT)


Call:
glm(formula = Outcome_BT ~ Sex_BT + Bone_Density_BT + Skin_Lesions_BT +
    Stiff_Neck_BT + Supraclavicular_BT + Axillar_BT + Mediastinum_BT,
    family = "binomial", data = Tumor_BT, na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5945  -0.8799   0.4908   0.7185   2.8881

Coefficients:
                   Estimate Std. Error z value      Pr(>|z|)
(Intercept)         -4.7961     0.7935  -6.044 0.00000000150 ***
Sex_BT               0.8332     0.2729   3.053       0.00227 **
Bone_Density_BT      0.9090     0.2970   3.060       0.00221 **
Skin_Lesions_BT      1.3584     0.5537   2.454       0.01414 *
Stiff_Neck_BT        1.9720     0.4260   4.629 0.00000366993 ***
Supraclavicular_BT  -0.7173     0.3821  -1.877       0.06046 .
Axillar_BT           2.1079     0.6639   3.175       0.00150 **
Mediastinum_BT       1.7793     0.2978   5.974 0.00000000232 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 450.41  on 338  degrees of freedom
Residual deviance: 340.54  on 331  degrees of freedom
AIC: 356.54

Number of Fisher Scoring iterations: 5
```

**Model Measures**

1. Fisher's Scoring Iteration: Iteration count of 5 indicates the model had no issues in convergence.
2. AIC: User Model 1 results in  AIC of 356.51. So only a marginal increase from the stepwise selection model's AIC.
3. Deviance: Residual deviance is smaller than Null deviance, meaning the proposed better does a better job at explaining data than the Null model.

4. Residual symmetry: Deviance residuals are symmetric for the most part.
5. z-values: p-values corresponding to z-tests indicate that still Mediastinum and Stiff Neck are the most significant variables although the p-values are slightly different now.
6. Variable Co-Efficients: Mediastinum and Stiff Neck coefficients have a positive sign as we observed in the correlation matrix.

### 4.3 User Model 2

Next, we analyze the impact of the variable Supraclavicular. Supraclavicular appeared to be collinear with the Stiff Neck variable as we observed in the correlation matrix. This could be explained by the fact that swollen supraclavicular lymph nodes could lead to a stiff neck (Healthwise Staff, 2020). Furthermore, after the Brain Scan variable, Supraclavicular was the next least significant variable both in terms of AIC and p-value. So, it will be interesting to observe the impact of removing the Supraclavicular variable from the model and keep other variables intact as in the stepwise selection model.

```
 summary(Out_UM_2_BT)


Call:
glm(formula = Outcome_BT ~ Sex_BT + Bone_Density_BT + Brain_Scan_BT + 
    Skin_Lesions_BT + Stiff_Neck_BT + Axillar_BT + Mediastinum_BT, 
    family = "binomial", data = Tumor_BT, na.action = na.omit)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.4299  -0.8022   0.5056   0.7383   2.6850  

Coefficients:
                Estimate Std. Error z value      Pr(>|z|)    
(Intercept)      -5.7110     0.9532  -5.991 0.00000000208 ***
Sex_BT            0.8320     0.2726   3.052       0.00227 ** 
Bone_Density_BT   0.8531     0.2943   2.899       0.00374 ** 
Brain_Scan_BT     0.8736     0.5594   1.562       0.11838    
Skin_Lesions_BT   1.2604     0.5509   2.288       0.02214 *  
Stiff_Neck_BT     2.1293     0.4243   5.019 0.00000052051 ***
Axillar_BT        1.7381     0.6048   2.874       0.00406 ** 
Mediastinum_BT    1.7551     0.2997   5.856 0.00000000474 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 450.41  on 338  degrees of freedom
Residual deviance: 341.58  on 331  degrees of freedom
AIC: 357.58

Number of Fisher Scoring iterations: 4
```

**Model Measures**

1. Fisher's Scoring Iteration: It took 4 indicates for the model to converge faster than the other models.

2. AIC: User Model 2 results in an AIC of 357.58 which is slightly more than both stepwise and user model 1.
3. Deviance: Residual deviance is smaller than Null deviance, meaning the proposed better does a better job at explaining data than the Null model.
4. Residual symmetry: Deviance residuals are also symmetric for the most part.
5. z-values: p-values corresponding to z-tests indicate that still Mediastinum and Stiff Neck are the most significant variable although the p-values are slightly different now.
6. Variable Co-Efficients: Mediastinum and Stiff Neck coefficients have a positive sign as we observed in the correlation matrix.

## 4. Model Evaluation

**User Model 1**

| User Model 1 | | Predicted | | |
|---|---|---|---|---|
| | | 0 | 1 | |
| Actual | 0 | 80 | 49 | 129 |
| | 1 | 26 | 184 | 210 |
| | | 106 | 233 | 339 |

| Parameter | Value |
|---|---|
| Accuracy | 77.9% |
| Specificity | 62.0% |
| Sensitivity | 87.6% |
| Precision | 79.0% |

**User Model 2**

| User Model 2 | | Predicted | | |
|---|---|---|---|---|
| | | 0 | 1 | |
| Actual | 0 | 81 | 48 | 129 |
| | 1 | 25 | 185 | 210 |
| | | 106 | 233 | 339 |

| Parameter | Value |
|---|---|
| Accuracy | 78.5% |
| Specificity | 62.8% |
| Sensitivity | 88.1% |
| Precision | 79.4% |

Comment: User Model 2 slightly outperforms User Model 1 in all the above-calculated aspects.

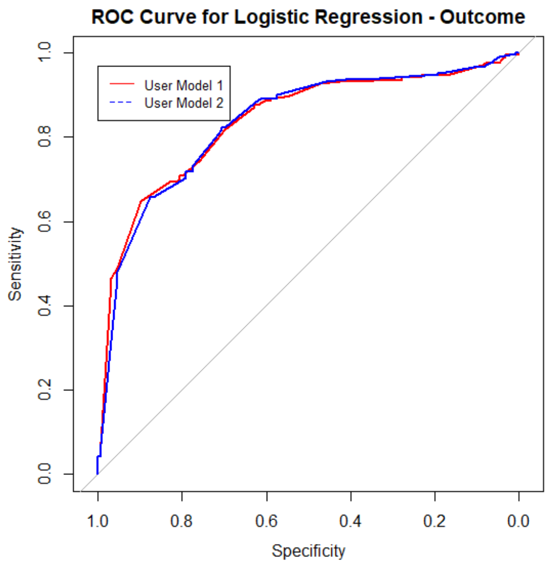
**ROC Curve**

ROC stands for receiving operator characteristics. The ROC curve shows the trade-off between sensitivity and specificity. Ideally, we want a classifier whose ROC curve passes through the top left-corner (~100 % sensitivity and 100 % specificity). When comparing two ROC curves the one closer to the top-left corner indicates a better performance. A random classifier will give points lying along the diagonal. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the classifier is.

In our case, both the models have overlapping ROC, except for higher specificity where User Model 1 seems to be doing better.

**AUC**

AUC means the area under the curve (ROC). In ideal conditions, we want a classifier with an AUC of 1. Generally higher AUC means a better model, but it may change depending upon application whether we want higher sensitivity at low specificity or vice versa. Overall AUC is slightly higher for User Model 1 (.8379) than User Model 2 (.8353).



## 5. Model Evaluation

Overall, both models have similar performance. It is difficult to recommend one over another. User Model 1 appears to be marginally better in terms of ROC curve and AUC whereas User Model 2 has slightly better accuracy, specificity, sensitivity, and precision values.

# Part 2

## 1. Logistic Regression – Stepwise

### 1.1. Model

```
summary(stp_Out_glm_BT)


 Call:
 glm(formula = Outcome_BT ~ Sex_BT + Bone_Density_BT + Brain_Scan_BT +
     Skin_Lesions_BT + Stiff_Neck_BT + Supraclavicular_BT + Axillar_BT +
     Mediastinum_BT, family = "binomial", data = Tumor_BT, na.action = na.omit)

 Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -2.6407  -0.8722   0.4821   0.7060   2.8736

 Coefficients:
                    Estimate Std. Error z value      Pr(>|z|)
 (Intercept)         -5.6287     0.9663  -5.825 0.00000000571 ***
 Sex_BT               0.8314     0.2744   3.030       0.00244 **
```

```
Bone_Density_BT      0.9148     0.2988   3.062       0.00220 **
Brain_Scan_BT        0.9183     0.5619   1.634       0.10217
Skin_Lesions_BT      1.3427     0.5608   2.395       0.01664 *
Stiff_Neck_BT        2.0327     0.4302   4.725 0.00000230315 ***
Supraclavicular_BT  -0.7450     0.3839  -1.941       0.05231 .
Axillar_BT           2.1933     0.6790   3.230       0.00124 **
Mediastinum_BT       1.6823     0.3035   5.543 0.00000002968 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 450.41  on 338  degrees of freedom
Residual deviance: 337.81  on 330  degrees of freedom
AIC: 355.81

Number of Fisher Scoring iterations: 5
```

### 1.2. Confusion Matrix

```
           Predicted
Act Outcome   0   1
          0  83  46
          1  26 184
```

### 1.3. Run Time

```
sw_Time_BT

 Time difference of 0.141644 secs
```

## 2. Naïve-Bayes Classification

### 2.1. Data Transformation

The Outcome variable was converted to factor data type to use Naïve Bayes Classifier.

```
 str(Tumor_BT)

'data.frame':    339 obs. of  14 variables:
 $ Outcome_BT        : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 1 ...
 $ Age_BT            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Sex_BT            : num  0 1 1 0 0 0 0 0 0 1 ...
 $ Bone_Density_BT   : num  1 0 0 1 0 1 1 0 1 0 ...
 $ Bone_Marrow_BT    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Lung_Spot_BT      : num  1 0 1 1 1 1 1 1 1 1 ...
 $ Pleura_BT         : num  1 1 1 0 1 1 1 0 1 1 ...
 $ Liver_Spot_BT     : num  0 0 1 1 1 0 1 1 1 1 ...
 $ Brain_Scan_BT     : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Skin_Lesions_BT   : num  1 0 1 1 1 0 1 1 1 1 ...
 $ Stiff_Neck_BT     : num  1 0 1 1 1 1 1 1 1 1 ...
 $ Supraclavicular_BT: num  0 1 0 0 0 0 1 0 1 0 ...
 $ Axillar_BT        : num  0 1 1 0 0 0 0 0 0 0 ...
 $ Mediastinum_BT    : num  1 1 1 1 1 1 1 1 1 0 ...
```

### 2.2. Model

```
Tumor_Naive_BT <- NaiveBayes(Outcome_BT ~ Age_BT + Sex_BT + Bone_Density_BT +
      Bone_Marrow_BT + Lung_Spot_BT + Pleura_BT + Liver_Spot_BT + Brain_Scan_BT +
       Skin_Lesions_BT + Stiff_Neck_BT + Supraclavicular_BT + Axillar_BT +
      Mediastinum_BT, data = Tumor_BT, na.action=na.omit)
```

### 2.3. Confusion Matrix

```
       Predicted
 Actual    0   1
      0   71  58
      1   30 180
```

### 2.4. Run Time

```
  NB_Time_BT

 Time difference of 0.006008148 secs
```

## 3. Linear Discriminant Analysis

### 3.1. Data Transformation

No new transformations are required for LDA.

### 3.2. Model

```
Tumor_Discrim_BT <- lda(Outcome_BT ~ Age_BT + Sex_BT + Bone_Density_BT +
                    Bone_Marrow_BT + Lung_Spot_BT + Pleura_BT + Liver_Spot_BT +
                    Brain_Scan_BT + Skin_Lesions_BT + Stiff_Neck_BT +
                    Supraclavicular_BT + Axillar_BT + Mediastinum_BT,
                         data = Tumor_BT, na.action=na.omit)
```

### 3.3. Confusion Matrix

```
       Predicted
 Actual    0   1
      0   84  45
      1   26 184
```

### 3.4. Run Time

```
  LDA_Time_BT
```

```
Time difference of 0.004987001 secs
```

## 4. Compare All Three Classifiers

| Model | Accuracy | Run Time | Type I Errors | Type II Errors |
|---|---|---|---|---|
| Stepwise Selection | 78.8% | 0.14 secs | 46 | 26 |
| Naïve Bayes | 74.0% | 0.006 secs | 58 | 30 |
| LDA | 79.1% | 0.005 secs | 45 | 26 |
| User Model 2 | 78.5% | 0.02 secs * | 48 | 25 |

* time for stepwise model not included

**4.1.** LDA is the most accurate classifier with an accuracy of 79.1%

**4.2.** LDA has the fastest runtime of .005 secs.

**4.3.** LDA has the lowest Type I errors (45).

**4.4.** Both LDA and stepwise selection models have the lowest Type II errors (26).

**4.5.** In terms of accuracy, Type I and Type II errors the difference between LDA and Stepwise selection model is minimal. But LDA is about 30 times faster than the stepwise selection model. So, in this case, LDA is the best overall classifier.

**4.6.** In case of tumor diagnosis, even one positive case should not go undetected. So, the focus should be on the model with the lowest false negatives, i.e., the one that minimizes Type II errors. Keep that in mind, User Model 2 is recommended, although it does not do as well as LDA on other parameters.

### References

Healthwise Staff. (2020, December 6). *Swollen Lymph Nodes*. Retrieved from www.healthlinkbc.ca: https://www.healthlinkbc.ca/health-topics/aa65796spec