**Classification**

**Author:** Harshita Singh, 8786434

**Course:** PROG8430

**Background**

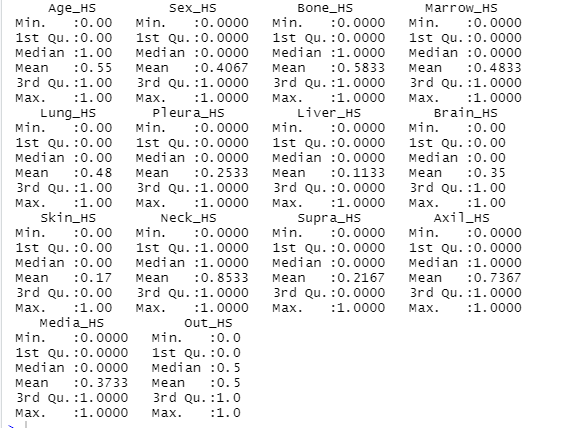
The data has been covered to determine the individuals who went through the process of pre-screening diagnosis of tumor. This analysis will help to discover the predicted probability of tumor diagnosis using logistic regression.

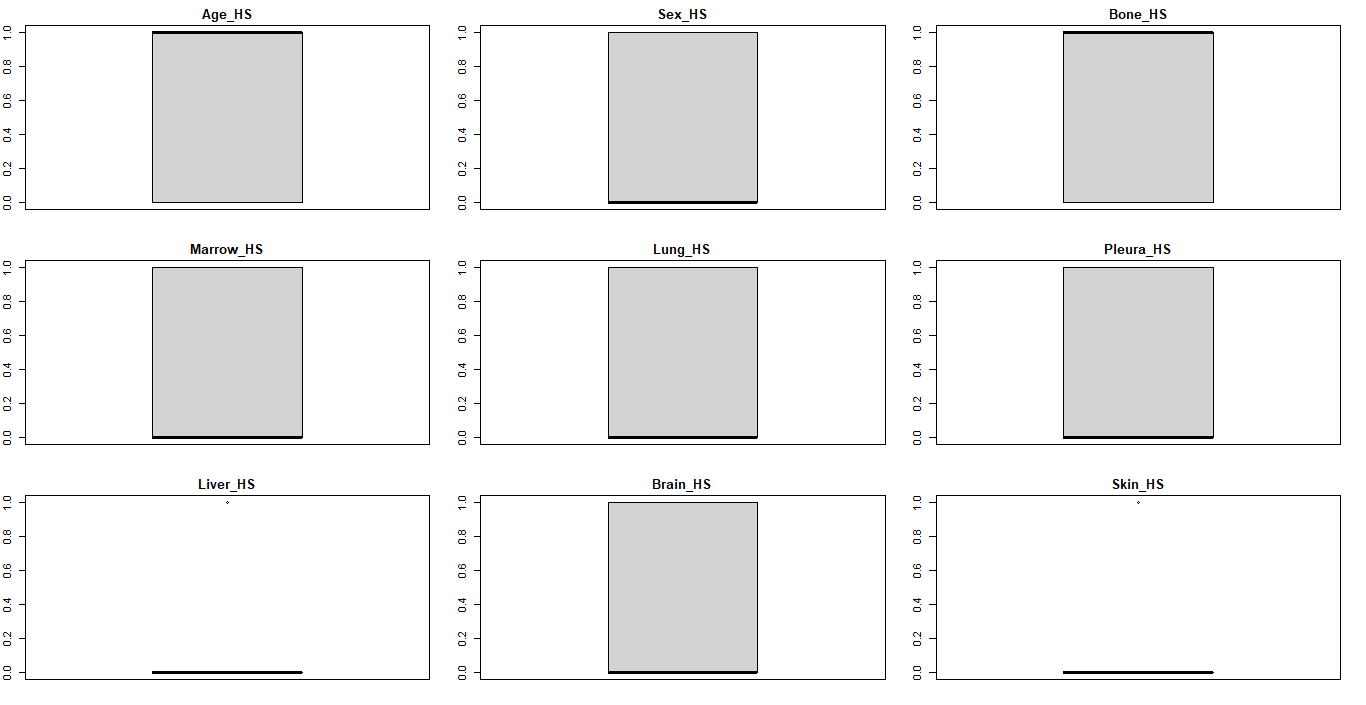
**Data Source**

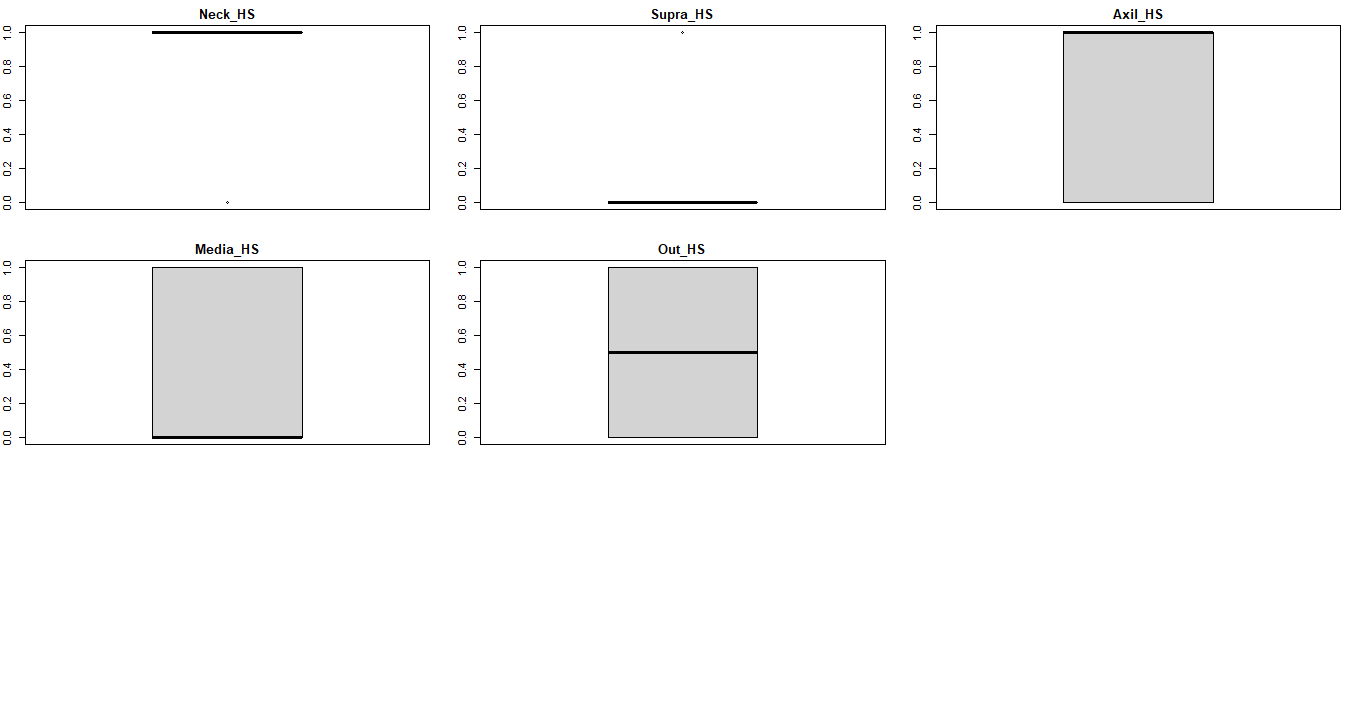
Data contains medical information which includes information such as age, sex, bone density test, bone marrow, and spot on lung, pleura, liver, brain, lesion, stiff neck, supraclavicular, axillar, and mediastinum.

**Part A**

1. **Preliminary Data Preparation:**



****



From the summary statistics we conclude that the prediction of pre-screening diagnosis of tumor worked properly. Also, all of the data look reasonable.

**2. Exploratory Analysis:**

**Part 1:**

1. Co-linear Variables:
   1. Sex\_HS and Bone\_HS are co-linear variables
   2. Liver\_HS and Brain\_HS are co-linear variables



1. Multi-collinear Issues: No multi-collinear variables found.

Threshold p-value = 0.05; According to Shapiro Wilk Test:

1. If p-value >0.05 then it implies that the distribution of the data are not significantly different from normal distribution and are not a useful predicator because of the large P-Value. Example:
2. If p-value <0.05 then we do not consider as normality; these variables are good predicator because of the lower P-Value; strong evidence that the Coefficient estimate is non-zero. Example: Yellow highlighted p-value of the variables

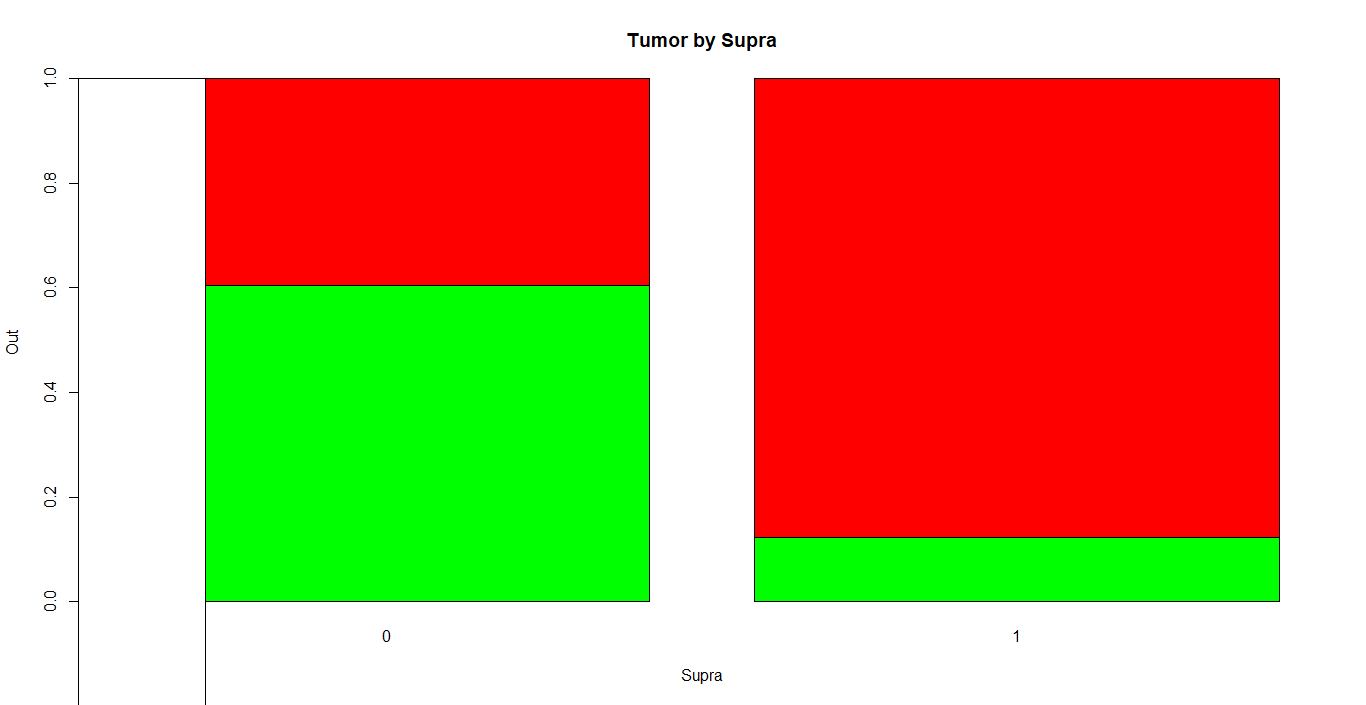
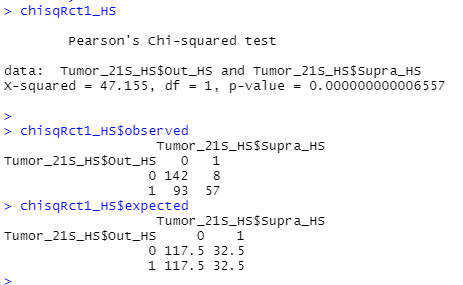
|  |  |  |
| --- | --- | --- |
|  | **statistic** | **p.value** |
| **Age\_HS** | 0.63284 | 7.03E-25 |
| **Sex\_HS** | 0.6236892 | 3.86E-25 |
| **Bone\_HS** | 0.6263052 | 4.58E-25 |
| **Marrow\_HS** | 0.6360829 | 8.71E-25 |
| **Lung\_HS** | 0.6359049 | 8.61E-25 |
| **Pleura\_HS** | 0.5411787 | 2.76E-27 |
| **Liver\_HS** | 0.3669979 | 5.74E-31 |
| **Brain\_HS** | 0.6028969 | 1.03E-25 |
| **Skin\_HS** | 0.4536066 | 2.97E-29 |
| **Neck\_HS** | 0.4212709 | 6.45E-30 |
| **Supra\_HS** | 0.507336 | 4.45E-28 |
| **Axil\_HS** | 0.5493074 | 4.35E-27 |
| **Media\_HS** | 0.6127153 | 1.91E-25 |
| **Out\_HS** | 0.6364871 | 8.95E-25 |

1. Most significant predictors of Tumors:
   1. Variables with the high correlation are the significant predictors of Out\_HS that are Brain\_HS and Supra\_HS.

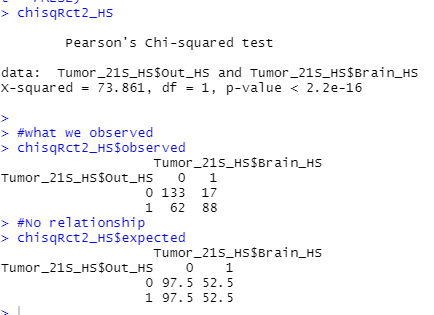
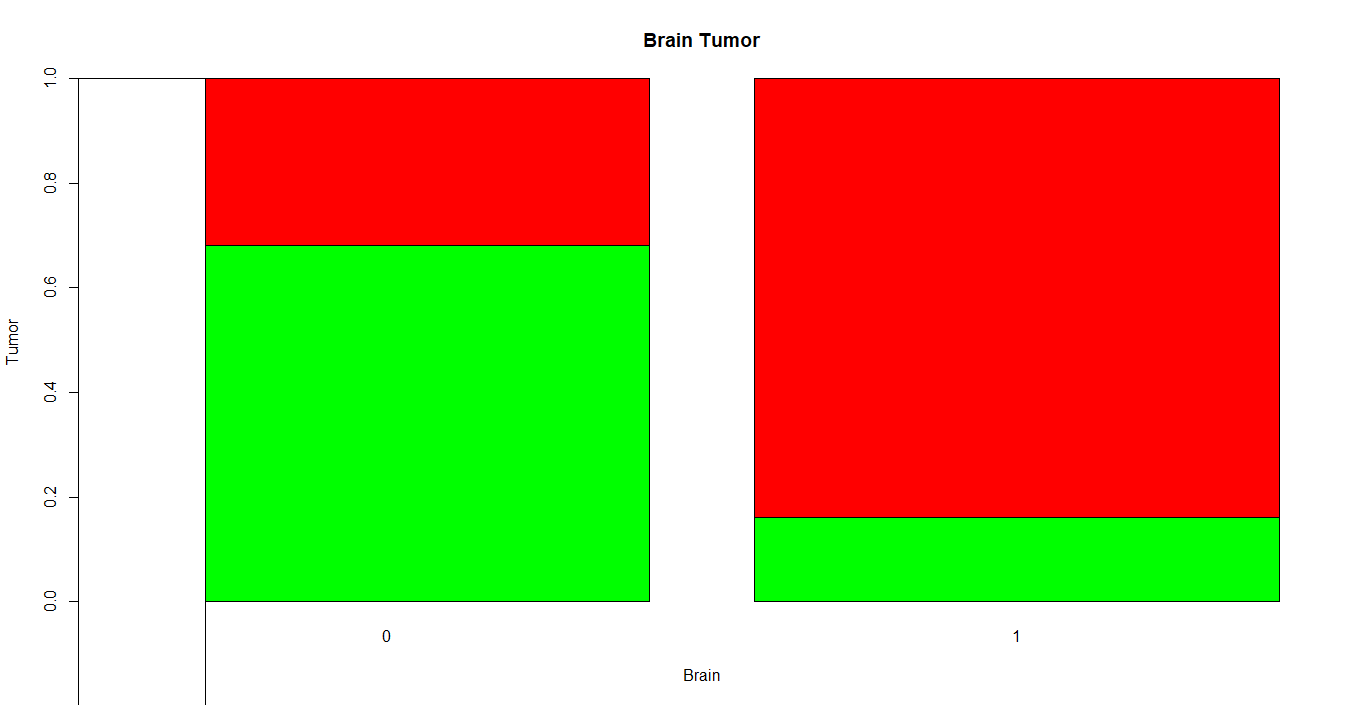


* 1. Using Contingency Table: Chi-Squared test examines where rows and columns of a contingency table are statistically significantly associated. Lower value of Chi-Squared is a high correlation between the two data set.

1. Using Supra\_HS:

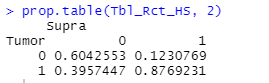
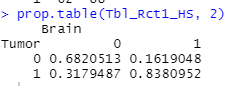


1. Using Brain\_HS:

’ 

**Contingency Table:**

**chisqRct1\_HS chisqRct2\_HS**

**Compare:**

1. **X-squared Value**: chisqRct1\_HS: 47.155 and chisqRct2\_HS: 73.861; It measures the difference between the observed and expected results of the data set or the variable.
2. **P-value**: chisqRct1\_HS and chisqRct2\_HS have p-value that are less than 0.05; indicating p-value <0.05 then we do not consider as normality; strong evidence that the Coefficient estimate is non-zero. This also concludes that the observed proportions are significantly different from the specified proportions.
3. **Contingency Table**: For chisqRct1\_HS: 87% are people who have Tumor and no Supraclavicular; 39% are who have Tumor and Supraclavicular; 12% who do not have tumor and no Supraclavicular; 60% who do not have tumor and they have Supraclavicular. For chisqRct2\_HS: 83% are people who have Tumor and no Brain scan; 39% are who have Tumor and Brain scan; 12% who do not have tumor and no Brain scan; 60% who do not have tumor and they have Brain scan.

**Conclusion**: We reject the null hypothesis and two variable used in Chi-Squared Test are in fact dependent. In chisqRct1\_HS, Out\_HS and Supra\_HS have significant relationship, same as in chisqRct2\_HS, Out\_HS and Brain\_HS have significant relationship. X-squared value for chisqRct2\_HS > chisqRct1\_HS

**3. Model Development:**

1. **Forward Selection Model:**

Eliminated the variables that are less significant in forward selection; currently we have 9 variables whereas in Full baseline selection model, we had 13 variables; AIC works: the lower the number, the better result we get. **Last AIC drop: AIC=-600.81**

1. F-Stat:

* Bigger F score is better for the model; Numerator: 9; Denominator: 290; higher F-static produces conflicting result because F-stat significance assesses all of the coefficient examines them individually.
* Cannot be compared the p value of the model to the significance level to access the null hypothesis

1. P-Value = <2.2e-16; F-Stat= 31.58 which means the p-value is extremely close to 0 but not 0. Alternative hypothesis: true probability of success is not equal to 0.5. This concludes that the forward selection model fits the data better compared to the model with no independent variables.
2. R-Squared value:

* Value=0.3614 which indicates that it takes into the account of how many exploratory variables are in the model
* 36.14% of the exploratory variables are in the model are elaborated.

1. Residuals (Residual Standard Error):

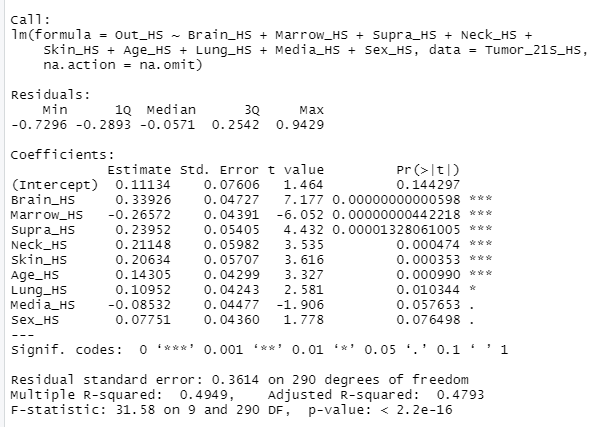
* It is an average error of the model.
* Our all variable model is giving us an average error of 0.3614 when predicting individual model
* Adding variables to the model will increase the value of Residuals

1. Significant variables:

* Almost all of them because we have dropped the insignificant variables in forward selection and have received lowest AIC.

1. Variable Co-Efficient:

* By looking at the co-efficient(Intercept Value/Estimates)Negative correlation of variable :Media\_HS and Marrow\_HS with Out\_HS(Tumor)



1. **Part 1:**

**Annual Selection Model “User Model 1”: (Dropping Supra\_HS due to moderate correlation with Out\_HS)**

* **Number of fisher scoring iteration:** 6 (It does not converge because User Model 2 has 5 iterations)
* **AIC:** 214.6 lower than “User Model 2”; User Model 1 better then User Model 2; the lower AIC, better the model
* **Residual Symmetry:** Significant reduction from 415.89 to 213.60;
* **Deviance Residuals:** data received is between the data and the best fitting line, which means the deviance residuals are square root of the contribution that each point has to overall dataset. They are equally distributed and close to 0, hence they are roughly **symmetric. Residual deviance:** related to model that has independent variable mentioned in the table below. **Null Deviance:** model that has only as a constant coefficient.
* **Z test:** lower the value the better it is for the model; p-value >0.05 (Columns highlighted in red have higher p value) these variables are not a useful predicator because of the large P-Value. P-value < 0.05 (Columns highlighted in green have lower p value) these variables are good predicator because of the lower P-Value; strong evidence that the Coefficient estimate is non-zero. For example: Each unit increment in Supra\_HS increases the Tumor odds of finding out by 2.52845 and p-value indicates that it is somewhat significant in determining the Tumor.
* **Variable Co-efficient:** The coefficient estimate of the variable Lung\_HS is 0.89169, which is positive. This indicates that an increase in spot on Lung is associated with increase in probability of Tumor. Whereas, the coefficient for the Liver\_HS is -0.15455 which is negative. This indicates that an increase in spot on Liver is associated with decrease in probability of Tumor.

Call:

glm(formula = Out\_HS ~ Age\_HS + Sex\_HS + Bone\_HS + Marrow\_HS +

Lung\_HS + Pleura\_HS + Liver\_HS + Brain\_HS + Skin\_HS + Neck\_HS +

Supra\_HS + Axil\_HS + Media\_HS, family = "binomial", data = Tumor\_21S\_HS,

na.action = na.omit)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.30988 -0.56919 -0.02758 0.46886 2.56879

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: | |  |  |  |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | -2.87877 | 0.75937 | -3.791 | 0.000150 \*\*\* |
| Age\_HS | 1.15337 | 0.36173 | 3.188 | 0.001430 \*\* |
| Sex\_HS | 0.60082 | 0.36397 | 1.651 | 0.098790 . |
| Bone\_HS | -0.05461 | 0.36007 | -0.152 | 0.879458 |
| Marrow\_HS | -2.03567 | 0.37634 | -5.409 | 0.00000006332 \*\*\* |
| Lung\_HS | 0.89169 | 0.35745 | 2.495 | 0.012611 \* |
| Pleura\_HS | 0.13443 | 0.42203 | 0.319 | 0.750088 |
| Liver\_HS | -0.15455 | 0.55344 | -0.279 | 0.780052 |
| Brain\_HS | 2.46851 | 0.42503 | 5.808 | 0.00000000633 \*\*\* |
| Skin\_HS | 1.91592 | 0.54951 | 3.487 | 0.000489 \*\*\* |
| Neck\_HS | 1.9352 | 0.53618 | 3.609 | 0.000307 \*\*\* |
| Supra\_HS | 2.52845 | 0.55397 | 4.564 | 0.00000501392 \*\*\* |
| Axil\_HS | -0.36231 | 0.40756 | -0.889 | 0.374018 |
| Media\_HS | -0.61719 | 0.36606 | -1.686 | 0.091785 . |

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 415.89 on 299 degrees of freedom

Residual deviance: 213.60 on 286 degrees of freedom

AIC: 241.6

Number of Fisher Scoring iterations: 6

**Part 2:**

**Annual Selection Model “User Model 2”: (Dropping Brain\_HS due to moderate correlation with Out\_HS)**

* + Fisher’s Scoring Iteration: 5 (It does not converge because User Model 1 has 6 iterations)
  + AIC : 281.26 higher than “User Model 1”; User Model 1 better then User Model 2; the lower AIC, better the model
* Deviance Residual: data received is between the data and the best fitting line, which means the deviance residuals are square root of the contribution that each point has to overall dataset. They are equally distributed and close to 0, hence they are roughly **symmetric.**
* **Residual deviance:** related to model that has independent variable mentioned in the table below.
* **Null Deviance:** model that has only as a constant coefficient.
* **Residual Symmetry:** Significant reduction from 415.89 to 255.26
* **Z test:** lower the value the better it is for the model; p-value >0.05 (Columns highlighted in red have higher p value) these variables are not a useful predicator because of the large P-Value. P-value < 0.05 (Columns highlighted in green have lower p value) these variables are good predicator because of the lower P-Value; strong evidence that the Coefficient estimate is non-zero. For example: Each unit increment in Supra\_HS increases the Tumor odds of finding out by 2.41115 and p-value indicates that it is somewhat significant in determining the Tumor.
* **Variable Co-efficient:** The coefficient estimate of the variable Lung\_HS is 0.79362, which is positive. This indicates that an increase in spot on Lung is associated with increase in probability of Tumor. Whereas, the coefficient for the Bone\_HS is -0.33255 which is negative. This indicates that an increase in bone density test is associated with decrease in probability of Tumor.

Call:

glm(formula = Out\_HS ~ Age\_HS + Sex\_HS + Bone\_HS + Marrow\_HS +

Lung\_HS + Pleura\_HS + Liver\_HS + Skin\_HS + Neck\_HS + Supra\_HS +

Axil\_HS + Media\_HS, family = "binomial", data = Tumor\_21S\_HS,

na.action = na.omit)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.34408 -0.66736 -0.03324 0.59834 3.09407

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: |  |  |  |  |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | -1.70566 | 0.65462 | -2.606 | 0.009173 \*\* |
| Age\_HS | 1.2704 | 0.33413 | 3.802 | 0.000143 \*\*\* |
| Sex\_HS | 0.6417 | 0.32699 | 1.962 | 0.049712 \* |
| Bone\_HS | -0.33255 | 0.31885 | -1.043 | 0.296971 |
| Marrow\_HS | -2.02091 | 0.33835 | -5.973 | 0.00000000233 \*\*\* |
| Lung\_HS | 0.79362 | 0.32556 | 2.438 | 0.014781 \* |
| Pleura\_HS | 0.31285 | 0.37453 | 0.835 | 0.403548 |
| Liver\_HS | 0.01026 | 0.50293 | 0.02 | 0.983719 |
| Skin\_HS | 2.03397 | 0.49043 | 4.147 | 0.00003363674 \*\*\* |
| Neck\_HS | 1.689 | 0.48895 | 3.454 | 0.000552 \*\*\* |
| Supra\_HS | 2.41115 | 0.49026 | 4.918 | 0.00000087379 \*\*\* |
| Axil\_HS | -0.41288 | 0.36411 | -1.134 | 0.256811 |
| Media\_HS | -1.03198 | 0.33342 | -3.095 | 0.001967 \*\* |

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 415.89 on 299 degrees of freedom

Residual deviance: 255.26 on 287 degrees of freedom

AIC: 281.26

Number of Fisher Scoring iterations: 5

**4. Model Evaluation:**

**Part 1:**

**I. User Model 1**

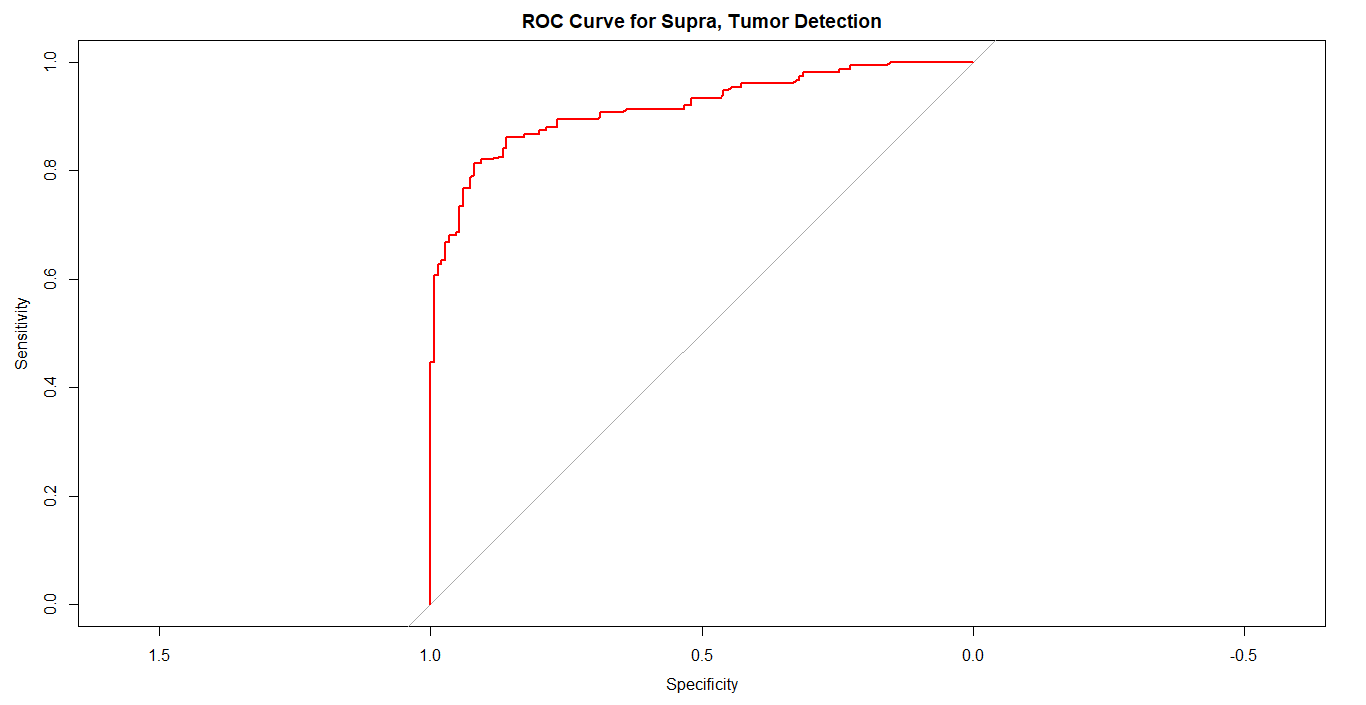
* 1. **Accuracy:** 0.85; 85% means that identification of 2 of every 10 Tumor in Supraclavicular cases is incorrect and 8 is correct.
  2. **Specify:** 0.8666667; 86% means the label of 1 of every 10 Supraclavicular, in actual, are missed by our model as Tumor and 9 are correctly identified as Supraclavicular; identified as true negative fraction
  3. **Sensitivity:** 0.8333333; 83% means the label of 2 of every 10 Supraclavicular in actual, are captured by our model as Tumor and 8 are correctly identified as Supraclavicular; identified as true positive fraction
  4. **Precision:** 0.862069; 86% means the label of 1 of every 10 Tumor cases does not have Supraclavicular and 9 are Tumor.
     1. **User Model 2**

1. **Accuracy:** 0.81; 81% means that identification of 2 of every 10 Tumor in Brain scan cases is incorrect and 8 is correct.
2. **Specify:** 0.84; 84% means the label of 2 of every 10 Brain scan, in actual, are missed by our model as Tumor and 9 are correctly identified as Brain scan; identified as true negative fraction
3. **Sensitivity:** 0.78; 78% means the label of 2 of every 10 Brain scan in actual, are captured by our model as Tumor and 8 are correctly identified as Brain scan; identified as true positive fraction
4. **Precision:** 0.8297872; 82% means the label of 2 of every 10 Tumor cases does not have Brain scan and 8 are Tumor.

**Part 2:**

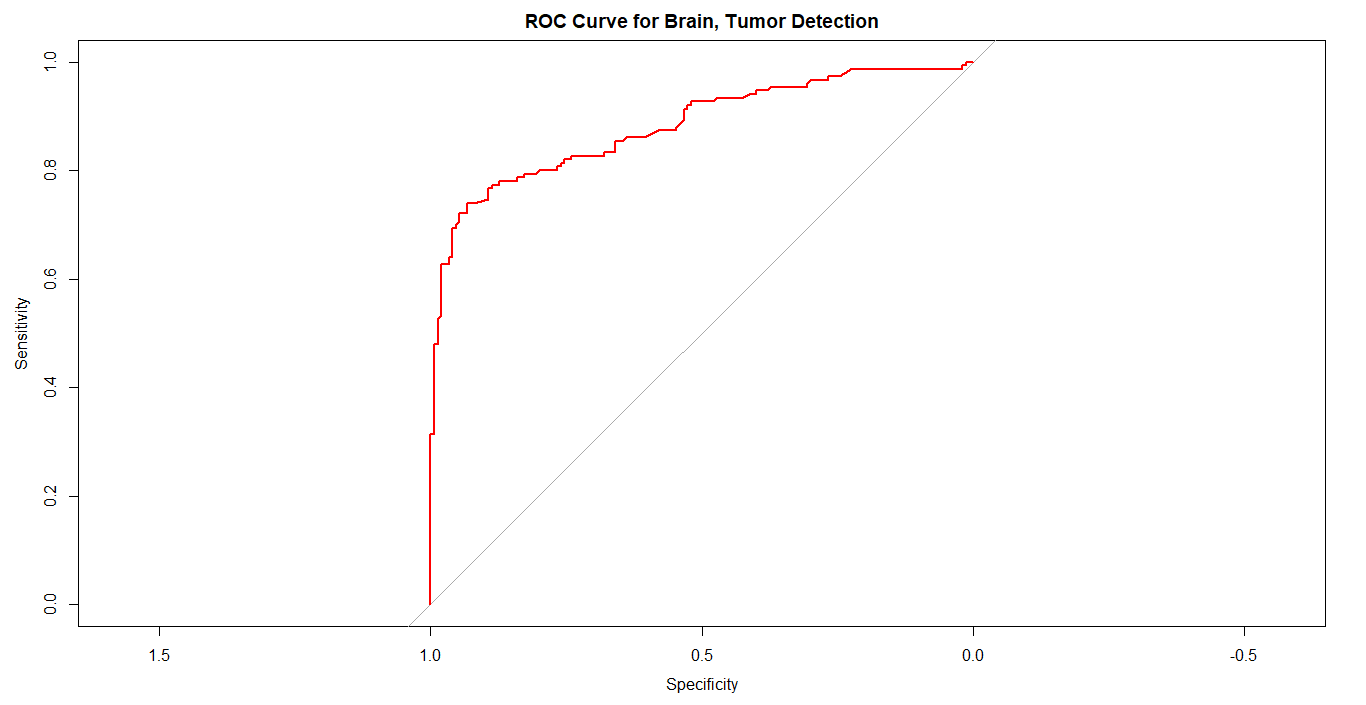
1. **ROC : User Model 1**

Area under the curve: 0.9157; AUC value between 0.5 to 1 where 0.5 indicates bad classifier and 1 indicates excellent classifier. In this case, 0.5<0.9157<1, there is a high chance that the classifier will be able to distinguish between positive and negative points. Classifier is able to identify more positive numbers of True positive and negative compared to false positive and negative. **Graph**: By identifying the (0, 1) top right left coordinates in the Cartesian plan, it indicates that the Sensitivity and Specificity would be higher and classifier would correctly identify the positive and negative class points.



1. **ROC : User Model 2**

Area under the curve: 0.882; AUC value between 0.5 to 1 where 0.5 indicates bad classifier and 1 indicates excellent classifier. In this case, 0.5<0.882<1, there is a high chance that the classifier will be able to distinguish between positive and negative points. Classifier is able to identify more positive numbers of True positive and negative compared to false positive and negative.

****

**4. Final Recommendation:**

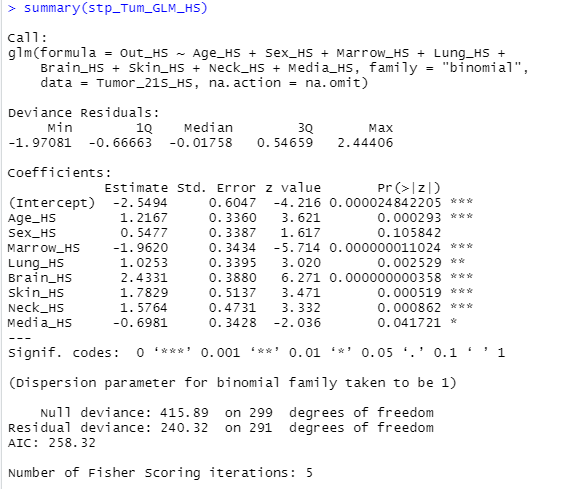
User Model 1 would be the better model in preceding analysis due to:

1. Higher AUC value indicating high classifier
2. More coordinates identified and the (0, 1); Sensitivity and Specificity are higher
3. Higher Accuracy percentage

**PART B**

* 1. **Logistic Regression – Stepwise**

**A**. **Forward option in the glm function**



**B**. **Confusion Matrix**

1. **Accuracy:** 0.8233333; 82% means that identification of 2 of every 10 Tumor in dependent on all the variables (Age\_HS, Sex\_HS, Marrow\_HS, Lung\_HS, Brain\_HS, Skin\_HS, Neck\_HS, Media\_HS) used in this cases is incorrect and 8 is correct.
2. **Specify:** 0.8466667; 84% means the label of 2 of every 10 dependent variables, in actual, are missed by our model as Tumor and 8 are correctly identified as variable; identified as true negative fraction
3. **Sensitivity:** 0.8; 80% means the label of 2 of every 10 variables(Age\_HS, Sex\_HS, Marrow\_HS, Lung\_HS, Brain\_HS, Skin\_HS, Neck\_HS, Media\_HS) in actual, are captured by our model as Tumor and 8 are correctly identified as variables; identified as true positive fraction
4. **Precision:** 0.8391608; 83% means the label of 2 of every 10 Tumor cases does not dependent on dependent variables and 8 are Tumor.

**C**. **Time**: SWTime\_HS: Time difference of 1.159335 secs

* 1. **Naïve Bayes Classification**

1. **Transform Variables:**

Tumor\_21S\_HS$Out\_HS <- as.factor(Tumor\_21S\_HS$Out\_HS)

Tumor\_21S\_HS$Out\_HS

1. **Naïve Bayes Classification (all variables):**

* TumorNaive\_HS <- NaiveBayes(Out\_HS ~ Age\_HS + Sex\_HS + Bone\_HS + Marrow\_HS + Lung\_HS

+ Pleura\_HS + Liver\_HS +Brain\_HS+ Skin\_HS + Neck\_HS + Supra\_HS

+ Axil\_HS + Media\_HS, data=Tumor\_21S\_HS, na.action = na.omit)

* Below data classified the information based on the all the variables, and they are the direct classifiers

> head(predBay\_HS$class)

1 2 3 4 5 6

1 0 1 1 0 1

Levels: 0 1

1. **Confusion Matrix**
2. **Accuracy:** 0.8033333; 80% means that identification of 2 of every 10 Tumor in dependent on all the variables in this cases is incorrect and 8 is correct.
3. **Specify:** 0.8333333; 83% means the label of 2 of every 10 dependent variables, in actual, are missed by our model as Tumor and 8 are correctly identified variables; identified as true negative fraction
4. **Sensitivity:** 0.7733333; 77% means the label of 2 of every 10 all variables in actual, are captured by our model as Tumor and 8 are correctly identified as variables; identified as true positive fraction
5. **Precision:** 0.822695; 82% means the label of 2 of every 10 Tumor cases does not dependent on dependent variables and 8 are Tumor.
6. **Time**: SWTime\_HS: Time difference of 0.03997087 sec
   1. **Linear Discriminant Analysis:**
      * 1. **Transform Variables:**

Tumor\_21S\_HS$Out\_HS <- as.factor(Tumor\_21S\_HS$Out\_HS)

Tumor\_21S\_HS$Out\_HS

* + - 1. **LDA Classification (all variables):**
* TumorDiscrim\_HS <-lda(Out\_HS ~ Age\_HS + Sex\_HS + Bone\_HS + Marrow\_HS + Lung\_HS

+ Pleura\_HS + Liver\_HS +Brain\_HS+ Skin\_HS + Neck\_HS + Supra\_HS

+ Axil\_HS + Media\_HS, data=Tumor\_21S\_HS, na.action = na.omit)

* > head(predLD\_HS$class)

[1] 0 0 1 0 0 1

Levels: 0 1

* + - 1. **Confusion Matrix**
  1. **Accuracy:** 0.8566667; 80% means that identification of 2 of every 10 Tumor in dependent on all the variables in this cases is incorrect and 8 is correct.
  2. **Specify:** 0.9066667; 90% means the label of 1 of every 10 dependent variables, in actual, are missed by our model as Tumor and 9 are correctly identified variables; identified as true negative fraction
  3. **Sensitivity:** 0.8066667; 80% means the label of 2 of every 10 all variables in actual, are captured by our model as Tumor and 8 are correctly identified as variables; identified as true positive fraction
  4. **Precision:** 0.8962963; 89% means the label of 1 of every 10 Tumor cases does not dependent on dependent variables and 9 are Tumor.
     + 1. **Time**: SWTime\_HS: Time difference of 0.03697896 sec
  5. **Compare Classifiers:**

|  |  |  |
| --- | --- | --- |
| **Question** | **Solution** | **Evidence** |
| Which classifier is most accurate? | Logistic Regression with 82% accuracy | Working with binary outcomes, easier to interpret and lesser sensitive to outliers |
| Which classifier is most suitable when processing speed is most important? | Linear Discriminant Analysis with 0.03697896 sec | Lot more faster compared to any other classifiers |
| Which classifier minimizes Type 1 errors? | (False Positive-FP) Linear Discriminant Analysis: 14 | It has the lowest false positive (Type 1 error). Patient cannot be falsely informed that are tested positive even though they do not have tumor, it can create unnecessary stress. It is practical preferred in the medical field. The lesser the FP, the better it is. |
| Which classifier minimizes Type 2 errors? | (FN-false negative) Linear Discriminant Analysis: 29 | It has the lowest false negative (Type 2 error). Patient cannot be falsely informed, in this case of Tumor patient was infected but the test result was negative. The lesser the FN, the better it is. |
| Which classifier is best overall? | Linear Discriminant Analysis | On the basis of lower FP, FN faster processing speed, and high specify and sensitivity value. |
| How do these classifiers compare to the best model you built in Part 1? | Part 1: Best Model(User Model 1) | 1. Higher AUC value indicating high classifier 2. More coordinates identified and the (0, 1); Sensitivity and Specificity are higher 3. Higher Accuracy percentage 4. Lower AIC value 5. Deviance residual are equally distributed and close to 0 |