**Assignment 5**

**Author:** Keerthy Raaj Shanmugam, 8779954

**Course:** PROG8430

**Part – A**

**1 Preliminary Data Preparation**

* 1. **. As demonstrated in class and conducted in previous assignments (MLR), make sure that the data is free from outliers or unnecessary data.**

**1. Missing Value Filter & Outlier,**

Age\_KS Sex\_KS Bone\_KS

Min. :0.0000 Min. :0.0000 Min. :0.00

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00

Median :1.0000 Median :0.0000 Median :1.00

Mean :0.5467 Mean :0.4367 Mean :0.61

3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.00

Max. :1.0000 Max. :1.0000 Max. :1.00

Marrow\_KS Lung\_KS Pleura\_KS

Min. :0.0 Min. :0.0000 Min. :0.0000

1st Qu.:0.0 1st Qu.:0.0000 1st Qu.:0.0000

Median :0.5 Median :0.0000 Median :0.0000

Mean :0.5 Mean :0.4467 Mean :0.1967

3rd Qu.:1.0 3rd Qu.:1.0000 3rd Qu.:0.0000

Max. :1.0 Max. :1.0000 Max. :1.0000

Liver\_KS Brain\_KS Skin\_KS

Min. :0.00 Min. :0.0000 Min. :0.00

1st Qu.:0.00 1st Qu.:0.0000 1st Qu.:0.00

Median :0.00 Median :0.0000 Median :0.00

Mean :0.13 Mean :0.3667 Mean :0.18

3rd Qu.:0.00 3rd Qu.:1.0000 3rd Qu.:0.00

Max. :1.00 Max. :1.0000 Max. :1.00

Neck\_KS Supra\_KS Axil\_KS

Min. :0.0000 Min. :0.0000 Min. :0.0000

1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:1.0000

Median :1.0000 Median :0.0000 Median :1.0000

Mean :0.8233 Mean :0.2167 Mean :0.7633

3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:1.0000

Max. :1.0000 Max. :1.0000 Max. :1.0000

Media\_KS Out\_KS

Min. :0.0000 Min. :0.0

1st Qu.:0.0000 1st Qu.:0.0

Median :0.0000 Median :0.5

Mean :0.3733 Mean :0.5

3rd Qu.:1.0000 3rd Qu.:1.0

Max. :1.0000 Max. :1.0

From the above-given output, it is clear that no values are missing, so no variable is removed using the **missing value filter** and there are no **outliers** because no value is 1.5 times of Inter Quartile Range (IQR).

**2. Low Variance Filter**

Age\_KS Sex\_KS Bone\_KS

nbr.val 300.00000000 300.00000000 300.00000000

nbr.null 136.00000000 169.00000000 117.00000000

nbr.na 0.00000000 0.00000000 0.00000000

min 0.00000000 0.00000000 0.00000000

max 1.00000000 1.00000000 1.00000000

range 1.00000000 1.00000000 1.00000000

sum 164.00000000 131.00000000 183.00000000

median 1.00000000 0.00000000 1.00000000

mean 0.54666667 0.43666667 0.61000000

SE.mean 0.02878953 0.02868284 0.02820731

CI.mean.0.95 0.05665576 0.05644581 0.05551000

var 0.24865106 0.24681159 0.23869565

std.dev 0.49864923 0.49680136 0.48856489

coef.var 0.91216323 1.13771304 0.80092605

Marrow\_KS Lung\_KS Pleura\_KS

nbr.val 300.00000000 300.00000000 300.00000000

nbr.null 150.00000000 166.00000000 241.00000000

nbr.na 0.00000000 0.00000000 0.00000000

min 0.00000000 0.00000000 0.00000000

max 1.00000000 1.00000000 1.00000000

range 1.00000000 1.00000000 1.00000000

sum 150.00000000 134.00000000 59.00000000

median 0.50000000 0.00000000 0.00000000

mean 0.50000000 0.44666667 0.19666667

SE.mean 0.02891575 0.02875078 0.02298676

CI.mean.0.95 0.05690416 0.05657951 0.04523632

var 0.25083612 0.24798216 0.15851728

std.dev 0.50083542 0.49797807 0.39814229

coef.var 1.00167084 1.11487629 2.02445230

Liver\_KS Brain\_KS Skin\_KS

nbr.val 300.00000000 300.00000000 300.00000000

nbr.null 261.00000000 190.00000000 246.00000000

nbr.na 0.00000000 0.00000000 0.00000000

min 0.00000000 0.00000000 0.00000000

max 1.00000000 1.00000000 1.00000000

range 1.00000000 1.00000000 1.00000000

sum 39.00000000 110.00000000 54.00000000

median 0.00000000 0.00000000 0.00000000

mean 0.13000000 0.36666667 0.18000000

SE.mean 0.01944893 0.02786867 0.02221813

CI.mean.0.95 0.03827413 0.05484359 0.04372373

var 0.11347826 0.23299889 0.14809365

std.dev 0.33686535 0.48269958 0.38482937

coef.var 2.59127189 1.31645340 2.13794096

Neck\_KS Supra\_KS Axil\_KS

nbr.val 300.00000000 300.00000000 300.00000000

nbr.null 53.00000000 235.00000000 71.00000000

nbr.na 0.00000000 0.00000000 0.00000000

min 0.00000000 0.00000000 0.00000000

max 1.00000000 1.00000000 1.00000000

range 1.00000000 1.00000000 1.00000000

sum 247.00000000 65.00000000 229.00000000

median 1.00000000 0.00000000 1.00000000

mean 0.82333333 0.21666667 0.76333333

SE.mean 0.02205614 0.02382505 0.02458046

CI.mean.0.95 0.04340494 0.04688602 0.04837262

var 0.14594203 0.17028986 0.18125975

std.dev 0.38202360 0.41266191 0.42574612

coef.var 0.46399627 1.90459345 0.55774600

Media\_KS Out\_KS

nbr.val 300.00000000 300.00000000

nbr.null 188.00000000 150.00000000

nbr.na 0.00000000 0.00000000

min 0.00000000 0.00000000

max 1.00000000 1.00000000

range 1.00000000 1.00000000

sum 112.00000000 150.00000000

median 0.00000000 0.50000000

mean 0.37333333 0.50000000

SE.mean 0.02797249 0.02891575

CI.mean.0.95 0.05504789 0.05690416

var 0.23473802 0.25083612

std.dev 0.48449769 0.50083542

coef.var 1.29776168 1.00167084

From the above-given output, it is clear that no variables have almost the same data in almost all rows, so no variables are removed using **Low Variance Filter.**

**3. High Correlation Filter**

Age\_KS Sex\_KS Bone\_KS Marrow\_KS

Age\_KS 1.000000000 0.032221269 -0.028005386 0.013391790

Sex\_KS 0.032221269 1.000000000 0.083915285 -0.006720800

Bone\_KS -0.028005386 0.083915285 1.000000000 -0.088843340

Marrow\_KS 0.013391790 -0.006720800 -0.088843340 1.000000000

Lung\_KS -0.016880683 0.033616569 0.072307289 0.067049193

Pleura\_KS -0.004267639 -0.080541312 -0.051409034 -0.025158615

Liver\_KS 0.073269822 0.019384805 -0.016053780 0.029735052

Brain\_KS 0.137096760 0.069268236 -0.015599918 0.179845761

Skin\_KS -0.009062933 0.077321468 -0.016721152 0.104115841

Neck\_KS -0.018024903 0.055391902 -0.047844016 0.078660390

Supra\_KS 0.121357335 -0.006253573 -0.043960036 0.121367009

Axil\_KS -0.113216692 -0.015759603 -0.043252194 -0.007842474

Media\_KS 0.038392290 0.042981448 -0.004521316 0.096480623

Out\_KS 0.174093265 0.141136809 -0.006834103 0.333333333

Lung\_KS Pleura\_KS Liver\_KS Brain\_KS

Age\_KS -0.01688068 -0.004267639 0.073269822 0.13709676

Sex\_KS 0.03361657 -0.080541312 0.019384805 0.06926824

Bone\_KS 0.07230729 -0.051409034 -0.016053780 -0.01559992

Marrow\_KS 0.06704919 -0.025158615 0.029735052 0.17984576

Lung\_KS 1.00000000 -0.022828904 0.151123311 0.06771319

Pleura\_KS -0.02282890 1.000000000 0.058101867 -0.13283968

Liver\_KS 0.15112331 0.058101867 1.000000000 -0.15014761

Brain\_KS 0.06771319 -0.132839677 -0.150147607 1.00000000

Skin\_KS -0.01954647 -0.057190459 -0.026315064 0.14763779

Neck\_KS 0.06457862 -0.100635127 -0.028847306 0.13481720

Supra\_KS 0.08083310 0.024766676 -0.010826563 0.18749127

Axil\_KS 0.02702774 0.019007124 0.075322427 -0.06455456

Media\_KS 0.04121644 -0.087152305 -0.072950918 0.12775377

Out\_KS 0.22796726 -0.092248255 -0.009911684 0.47036584

Skin\_KS Neck\_KS Supra\_KS Axil\_KS

Age\_KS -0.009062933 -0.01802490 0.121357335 -0.113216692

Sex\_KS 0.077321468 0.05539190 -0.006253573 -0.015759603

Bone\_KS -0.016721152 -0.04784402 -0.043960036 -0.043252194

Marrow\_KS 0.104115841 0.07866039 0.121367009 -0.007842474

Lung\_KS -0.019546473 0.06457862 0.080833096 0.027027742

Pleura\_KS -0.057190459 -0.10063513 0.024766676 0.019007124

Liver\_KS -0.026315064 -0.02884731 -0.010826563 0.075322427

Brain\_KS 0.147637788 0.13481720 0.187491272 -0.064554557

Skin\_KS 1.000000000 0.01228469 0.153740777 -0.024904036

Neck\_KS 0.012284689 1.00000000 0.010253945 -0.011172601

Supra\_KS 0.153740777 0.01025394 1.000000000 -0.087884492

Axil\_KS -0.024904036 -0.01117260 -0.087884492 1.000000000

Media\_KS -0.002870046 0.08649283 0.045723062 -0.024212697

Out\_KS 0.173526402 0.23598117 0.250825152 0.023527422

Media\_KS Out\_KS

Age\_KS 0.038392290 0.174093265

Sex\_KS 0.042981448 0.141136809

Bone\_KS -0.004521316 -0.006834103

Marrow\_KS 0.096480623 0.333333333

Lung\_KS 0.041216440 0.227967257

Pleura\_KS -0.087152305 -0.092248255

Liver\_KS -0.072950918 -0.009911684

Brain\_KS 0.127753767 0.470365835

Skin\_KS -0.002870046 0.173526402

Neck\_KS 0.086492827 0.235981171

Supra\_KS 0.045723062 0.250825152

Axil\_KS -0.024212697 0.023527422

Media\_KS 1.000000000 0.220527139

Out\_KS 0.220527139 1.000000000

From the above-given output, it is clear that no variables have high a correlation, so no variables are removed using **a High Correlation Filter.**

**2 Exploratory Analysis**

**2.1. Correlations: Create numeric correlations (as demonstrated) and comment on what you see. Are there co-linear variables?**

Age\_KS Sex\_KS Bone\_KS Marrow\_KS Lung\_KS Pleura\_KS

Age\_KS 1.00 0.03 -0.03 0.01 -0.02 0.00

Sex\_KS 0.03 1.00 0.08 -0.01 0.03 -0.08

Bone\_KS -0.03 0.08 1.00 -0.09 0.07 -0.05

Marrow\_KS 0.01 -0.01 -0.09 1.00 0.07 -0.03

Lung\_KS -0.02 0.03 0.07 0.07 1.00 -0.02

Pleura\_KS 0.00 -0.08 -0.05 -0.03 -0.02 1.00

Liver\_KS 0.07 0.02 -0.02 0.03 0.15 0.06

Brain\_KS 0.14 0.07 -0.02 0.18 0.07 -0.13

Skin\_KS -0.01 0.08 -0.02 0.10 -0.02 -0.06

Neck\_KS -0.02 0.06 -0.05 0.08 0.06 -0.10

Supra\_KS 0.12 -0.01 -0.04 0.12 0.08 0.02

Axil\_KS -0.11 -0.02 -0.04 -0.01 0.03 0.02

Media\_KS 0.04 0.04 0.00 0.10 0.04 -0.09

Out\_KS 0.17 0.14 -0.01 0.33 0.23 -0.09

Liver\_KS Brain\_KS Skin\_KS Neck\_KS Supra\_KS Axil\_KS

Age\_KS 0.07 0.14 -0.01 -0.02 0.12 -0.11

Sex\_KS 0.02 0.07 0.08 0.06 -0.01 -0.02

Bone\_KS -0.02 -0.02 -0.02 -0.05 -0.04 -0.04

Marrow\_KS 0.03 0.18 0.10 0.08 0.12 -0.01

Lung\_KS 0.15 0.07 -0.02 0.06 0.08 0.03

Pleura\_KS 0.06 -0.13 -0.06 -0.10 0.02 0.02

Liver\_KS 1.00 -0.15 -0.03 -0.03 -0.01 0.08

Brain\_KS -0.15 1.00 0.15 0.13 0.19 -0.06

Skin\_KS -0.03 0.15 1.00 0.01 0.15 -0.02

Neck\_KS -0.03 0.13 0.01 1.00 0.01 -0.01

Supra\_KS -0.01 0.19 0.15 0.01 1.00 -0.09

Axil\_KS 0.08 -0.06 -0.02 -0.01 -0.09 1.00

Media\_KS -0.07 0.13 0.00 0.09 0.05 -0.02

Out\_KS -0.01 0.47 0.17 0.24 0.25 0.02

Media\_KS Out\_KS

Age\_KS 0.04 0.17

Sex\_KS 0.04 0.14

Bone\_KS 0.00 -0.01

Marrow\_KS 0.10 0.33

Lung\_KS 0.04 0.23

Pleura\_KS -0.09 -0.09

Liver\_KS -0.07 -0.01

Brain\_KS 0.13 0.47

Skin\_KS 0.00 0.17

Neck\_KS 0.09 0.24

Supra\_KS 0.05 0.25

Axil\_KS -0.02 0.02

Media\_KS 1.00 0.22

Out\_KS 0.22 1.00

From the above-given output, it is clear that there are **no** **co-linear variables** because there is no strong correlation between any variables. **Out\_KS (Tumor is present=1, Is not present=0**) is the only variable that has a correlation with some variables.

**2.2. Identify the two most significant predictors of tumors and provide statistical evidence (in addition to the correlation coefficients) that suggests they are associated with tumors (Think of the contingency tables we did in class).**

**1. Brain**

**1. Correlation Percentage Table:**

Brain

Out 0 1

0 0.6789474 0.1909091

1 0.3210526 0.8090909

**2. Chi-squared test**

Pearson's Chi-squared test

data: Tum\_dat$Out\_KS and Tum\_dat$Brain\_KS

X-squared = 66.373, df = 1, p-value = 3.731e-16

**3. Observed**

Tum\_dat$Brain\_KS

Tum\_dat$Out\_KS 0 1

1. 129 21
2. 61 89

**4. Expected**

Tum\_dat$Brain\_KS

Tum\_dat$Out\_KS 0 1

0 95 55

1 95 55

From the above-given output, it is clear that the Variable **Brain\_KS** isasignificant predictor of tumors because it has correlation coefficients of **0.47**, the good percentage in the Correlation Percentage Table, Chi-squared value is good, degree of freedom is 1, the p-value is almost equal to 1, it has a good percentage in observed value and finally same value in expected value.

**2. Marrow**

**1. Correlation Percentage Table:**

Marrow

Out 0 1

0 0.6666667 0.3333333

1 0.3333333 0.6666667

**2. Chi-squared test**

Pearson's Chi-squared test

data: Tum\_dat$Out\_KS and Tum\_dat$Marrow\_KS

X-squared = 33.333, df = 1, p-value = 0.000000007764

**3. Observed**

Tum\_dat$Marrow\_KS

Tum\_dat$Out\_KS 0 1

0 100 50

1. 50 100

**4. Expected**

Tum\_dat$Marrow\_KS

Tum\_dat$Out\_KS 0 1

0 75 75

1 75 75

From the above-given output, it is clear that the Variable **Marrow\_KS** isasignificant predictor of tumors because it has a correlation coefficient of **0.33**, the good percentage in the Correlation Percentage Table, Chi-squared value is good, degree of freedom is 1, the p-value is almost equal to 1, it has a good percentage in observed value and finally same value in expected value.

**3 Model Development**

**As demonstrated in class, create three logistic regression models.**

**3.1. A forward selection model.**

> summary(Tum\_fsm)

Call:

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

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

Supra\_KS + Axil\_KS + Media\_KS, family = "binomial", data = Tum\_dat,

na.action = na.omit)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.69379 -0.63704 -0.00555 0.67480 2.19194

Coefficients:

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

(Intercept) -5.0238 0.8034 -6.253 0.000000000402 \*\*\*

Age\_KS 0.7227 0.3149 2.295 0.021717 \*

Sex\_KS 0.6911 0.3174 2.177 0.029471 \*

Bone\_KS 0.1188 0.3247 0.366 0.714418

Marrow\_KS 1.4139 0.3182 4.443 0.000008886660 \*\*\*

Lung\_KS 1.0848 0.3190 3.401 0.000672 \*\*\*

Pleura\_KS 0.2130 0.3932 0.542 0.588083

Liver\_KS 0.1863 0.4602 0.405 0.685576

Brain\_KS 2.1182 0.3565 5.941 0.000000002832 \*\*\*

Skin\_KS 0.7236 0.4192 1.726 0.084353 .

Neck\_KS 1.5127 0.4445 3.403 0.000667 \*\*\*

Supra\_KS 1.0352 0.4075 2.540 0.011070 \*

Axil\_KS 0.4706 0.3933 1.197 0.231442

Media\_KS 0.8709 0.3196 2.725 0.006437 \*\*

---

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: 263.80 on 286 degrees of freedom

AIC: 291.8

Number of Fisher Scoring iterations: 5

**3.2. Two additional models using variables that you select based on the above output (recall lecture slides on variable selection). We will refer to these models as “User Model 1” and “User Model 2”. Make sure you mention why you chose the variables you did. For each model, interpret and comment on the main measures we discussed in class:**

**User Model 1,**

Call:

glm(formula = Out\_KS ~ Age\_KS + Sex\_KS + Marrow\_KS + Lung\_KS +

Liver\_KS + Brain\_KS + Skin\_KS + Neck\_KS + Supra\_KS + Axil\_KS +

Media\_KS, family = "binomial", data = Tum\_dat, na.action = na.omit)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.68111 -0.63493 -0.01154 0.67056 2.17445

Coefficients:

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

(Intercept) -4.8125 0.7093 -6.785 0.0000000000116 \*\*\*

Age\_KS 0.7245 0.3146 2.303 0.021294 \*

Sex\_KS 0.6701 0.3120 2.148 0.031748 \*

Marrow\_KS 1.3882 0.3144 4.415 0.0000100933602 \*\*\*

Lung\_KS 1.0844 0.3174 3.416 0.000635 \*\*\*

Liver\_KS 0.1729 0.4587 0.377 0.706162

Brain\_KS 2.0884 0.3494 5.977 0.0000000022796 \*\*\*

Skin\_KS 0.7089 0.4198 1.689 0.091257 .

Neck\_KS 1.4627 0.4361 3.354 0.000795 \*\*\*

Supra\_KS 1.0387 0.4084 2.544 0.010971 \*

Axil\_KS 0.4559 0.3907 1.167 0.243187

Media\_KS 0.8604 0.3191 2.696 0.007015 \*\*

---

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: 264.19 on 288 degrees of freedom

AIC: 288.19

Number of Fisher Scoring iterations: 5

**User Model 2,**

Call:

glm(formula = Out\_KS ~ Age\_KS + Sex\_KS + Marrow\_KS + Lung\_KS +

Brain\_KS + Neck\_KS + Supra\_KS + Media\_KS, family = "binomial",

data = Tum\_dat, na.action = na.omit)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.72680 -0.65178 -0.02665 0.68643 2.12964

Coefficients:

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

(Intercept) -4.2888 0.6003 -7.144 0.000000000000903 \*\*\*

Age\_KS 0.6732 0.3060 2.200 0.027793 \*

Sex\_KS 0.7057 0.3087 2.286 0.022260 \*

Marrow\_KS 1.4229 0.3114 4.569 0.000004899145360 \*\*\*

Lung\_KS 1.1089 0.3122 3.552 0.000382 \*\*\*

Brain\_KS 2.0935 0.3380 6.193 0.000000000590663 \*\*\*

Neck\_KS 1.4247 0.4287 3.323 0.000890 \*\*\*

Supra\_KS 1.1152 0.4013 2.779 0.005452 \*\*

Media\_KS 0.8167 0.3134 2.606 0.009155 \*\*

---

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: 268.72 on 291 degrees of freedom

AIC: 286.72

Number of Fisher Scoring iterations: 5

**1. AIC**

AIC of **user model 2** **(286.72)** is **Lowe**r and better than **user model 1** and **Forward Model.**

**2. Deviance**

Deviance of **Forward Model (263.80)** is **Lowe**r and better than **user model 1** and **user model 2.**

**3. Residual symmetry**

Residual symmetry was **almost similar** for all the 3 models.

**4. z-values**

Z -values of **user model 2** have no value greater than 0.05 and so it is **superior**.

**5. Variable Co-Efficient**

All the variable co-efficient of **user model 2** is similar to the correlation coefficient (positive or negative) and so it is **superior**.

In **User model 1** the **Bone\_KS** and **Pleura\_KS** variable was removed because the **z-value** of the two variables was **greater than 0.05** and **Variable Co-Efficient** is also **different** from the **correlation coefficient** (positive or negative).

In **User model 2** the **Liver\_KS, Axil\_KS,** and **Skin\_KS** variables were removed because the **z-value** of the two variables was **greater than 0.05**, and **Variable Co-Efficient** is also **different** from the **correlation coefficient** (positive or negative).

**4. Model Evaluation**

**4.1. For User Model 1 and User Model 2, create and evaluate the confusion matrices. Set the default predictive level to 50% for “success”. Based on the confusion matrix, calculate and comment on:**

**User Model 1,**

Predicted

Act Tumor 0 1

0 119 31

1 33 117

**User Model 2,**

Predicted

Act Tumor 0 1

1. 121 29

1 32 118

1. **Accuracy**

User Model 1 = **0.786,** User Model 2 **=0.796**

1. **Specificity**

User Model 1 **= 0.793,** User Model 2 = **0.806**

1. **Sensitivity**

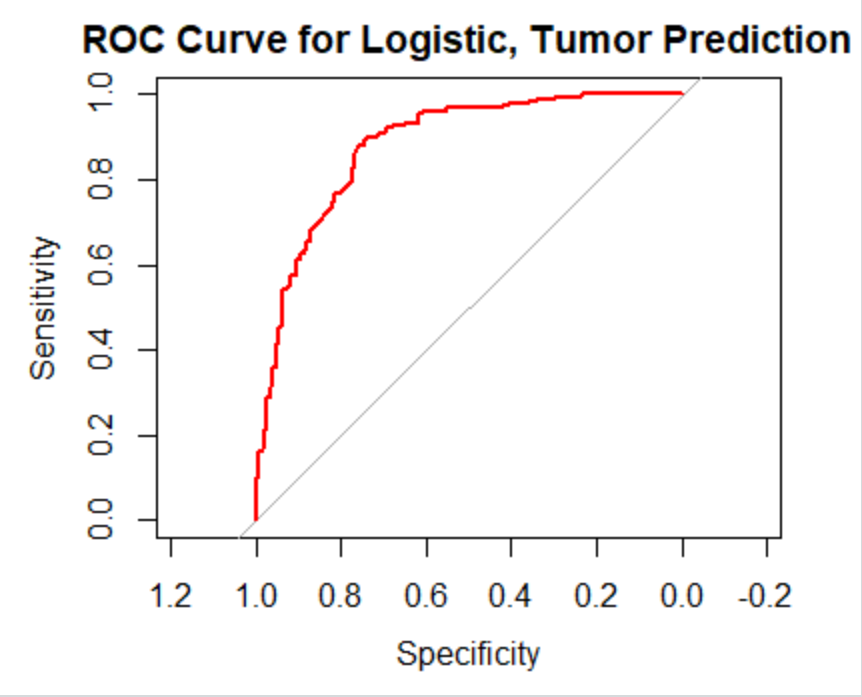
User Model 1 = **0.780,** User Model 2 = **0.786**

1. **Precision**

User Model 1 = **0.790,** User Model 2 = **0.493**

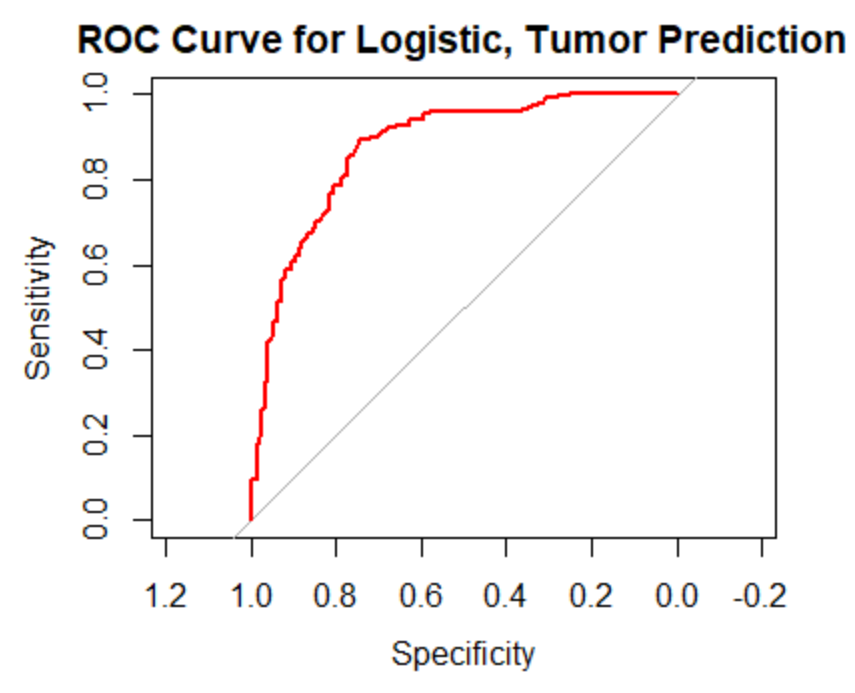
**4.2. For each of the two models, create the ROC curve and calculate the AUC. Comment on how you interpret each of them.**

**User Model 1,**

****

**The area under the curve: 0.8796**

**User Model 2,**

****

**The area under the curve: 0.8762**

From the above output, the Area under the curve of **User Model 1** is **slightly** **greater** than **User Model 2** and so user model 1 will get **better results** (Does not have much impact).

**5 Final Recommendation**

**5.1. Based on your preceding analysis, recommend which model should be selected and explain why.**

From all the above analysis, **User model 2 is the best** model because,

Although,

1. **Deviance** of user model 1 is slightly smaller than user model 2. (Does not have much impact)

2. **Residual Symmetry** is almost similar for both models 1 and 2.

3. **Area under the curve** is slightly higher for model 1. (Does not have much impact)

But,

4. **AIC** of user model 2 is smaller than user model 2 and so it is **superior**.

5. **Z -values** of **user model 2** have no value greater than 0.05 and so it is **superior**.

6. All the **variable co-efficient** of **user model 2** is similar to the **correlation coefficient** (positive or negative) and so it is **superior**.

7. **Type 2 error**,

False Negative value of **user model 1 = 33**

False Negative value of **user model 2 = 32**

False Negative value should be **less** while predicting diseases. So, **User model 2 is the best** overall.

**Part – B**

**1 Logistic Regression – Stepwise**

**1.1. As above, use the forward option in the glm function to fit the model NOTE – These results should match the output from 3.1.**

Call:

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

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

Supra\_KS + Axil\_KS + Media\_KS, family = "binomial", data = Tum\_dat,

na.action = na.omit)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.69379 -0.63704 -0.00555 0.67480 2.19194

Coefficients:

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

(Intercept) -5.0238 0.8034 -6.253 0.000000000402

Age\_KS 0.7227 0.3149 2.295 0.021717

Sex\_KS 0.6911 0.3174 2.177 0.029471

Bone\_KS 0.1188 0.3247 0.366 0.714418

Marrow\_KS 1.4139 0.3182 4.443 0.000008886660

Lung\_KS 1.0848 0.3190 3.401 0.000672

Pleura\_KS 0.2130 0.3932 0.542 0.588083

Liver\_KS 0.1863 0.4602 0.405 0.685576

Brain\_KS 2.1182 0.3565 5.941 0.000000002832

Skin\_KS 0.7236 0.4192 1.726 0.084353

Neck\_KS 1.5127 0.4445 3.403 0.000667

Supra\_KS 1.0352 0.4075 2.540 0.011070

Axil\_KS 0.4706 0.3933 1.197 0.231442

Media\_KS 0.8709 0.3196 2.725 0.006437

(Intercept) \*\*\*

Age\_KS \*

Sex\_KS \*

Bone\_KS

Marrow\_KS \*\*\*

Lung\_KS \*\*\*

Pleura\_KS

Liver\_KS

Brain\_KS \*\*\*

Skin\_KS .

Neck\_KS \*\*\*

Supra\_KS \*

Axil\_KS

Media\_KS \*\*

---

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: 263.80 on 286 degrees of freedom

AIC: 291.8

Number of Fisher Scoring iterations: 5

**This result is as same as the output from 3.1.**

**1. 2. Summarize the results in a Confusion Matrix.**

Predicted

Act Tumor 0 1

0 120 30

1 35 115

**1.3. As demonstrated in class, calculate the time (in seconds) it took to fit the model and include this in your summary.**

Time difference of **0.03300381 secs**

**2 Naïve-Bayes Classification**

**2. 1. As demonstrated in class, transform the variables as necessary for N-B classification.**

'data.frame': 300 obs. of 14 variables:

$ Age\_KS : int 0 1 0 0 1 1 0 1 1 0 ...

$ Sex\_KS : int 1 1 1 1 1 0 0 0 1 0 ...

$ Bone\_KS : int 1 1 0 0 1 0 1 1 0 1 ...

$ Marrow\_KS: int 0 0 0 0 0 0 0 0 0 0 ...

$ Lung\_KS : int 0 1 0 1 0 0 0 0 0 1 ...

$ Pleura\_KS: int 0 0 0 0 0 1 1 0 1 0 ...

$ Liver\_KS : int 0 0 0 0 0 0 1 0 0 0 ...

$ Brain\_KS : int 1 0 1 0 1 0 0 1 0 0 ...

$ Skin\_KS : int 0 0 0 0 0 0 0 0 0 0 ...

$ Neck\_KS : int 1 1 1 1 0 0 1 1 1 1 ...

$ Supra\_KS : int 0 0 0 1 0 1 0 1 0 0 ...

$ Axil\_KS : int 1 1 1 1 0 1 0 1 1 0 ...

$ Media\_KS : int 0 0 1 0 1 1 1 0 1 0 ...

$ Out\_KS : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 1 1 1 2 ...

**2. 2. Use all the variables in the dataset to fit a Naïve-Bayesian classification model.**

> head(pred\_Tum\_bay$class,5)

[1] 1 0 1 1 0

Levels: 0 1

> head(pred\_Tum\_bay$posterior,5)

0 1

[1,] 0.3539715 0.6460285

[2,] 0.7379866 0.2620134

[3,] 0.1495760 0.8504240

[4,] 0.2200528 0.7799472

[5,] 0.8530728 0.1469272

**2.3. Summarize the results in a Confusion Matrix.**

Predicted

Actual 0 1

0 118 32

1 34 116

**2. 4. As demonstrated in class, calculate the time (in seconds) it took to fit the model and include this in your summary.**

Time difference of **0.02600503 secs**

**3 Linear Discriminant Analysis**

**3.1. As demonstrated in class, transform the variables as necessary for LDA classification.**

'data.frame': 300 obs. of 14 variables:

$ Age\_KS : int 0 1 0 0 1 1 0 1 1 0 ...

$ Sex\_KS : int 1 1 1 1 1 0 0 0 1 0 ...

$ Bone\_KS : int 1 1 0 0 1 0 1 1 0 1 ...

$ Marrow\_KS: int 0 0 0 0 0 0 0 0 0 0 ...

$ Lung\_KS : int 0 1 0 1 0 0 0 0 0 1 ...

$ Pleura\_KS: int 0 0 0 0 0 1 1 0 1 0 ...

$ Liver\_KS : int 0 0 0 0 0 0 1 0 0 0 ...

$ Brain\_KS : int 1 0 1 0 1 0 0 1 0 0 ...

$ Skin\_KS : int 0 0 0 0 0 0 0 0 0 0 ...

$ Neck\_KS : int 1 1 1 1 0 0 1 1 1 1 ...

$ Supra\_KS : int 0 0 0 1 0 1 0 1 0 0 ...

$ Axil\_KS : int 1 1 1 1 0 1 0 1 1 0 ...

$ Media\_KS : int 0 0 1 0 1 1 1 0 1 0 ...

$ Out\_KS : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 1 1 1 2 ...

**3.2. Use all the variables in the dataset to fit an LDA classification model.**

> head(pred\_Tum\_dis$class,5)

[1] 0 0 1 0 0

Levels: 0 1

> head(pred\_Tum\_dis$posterior,5)

0 1

1 0.5215150 0.4784850

2 0.6286547 0.3713453

3 0.3414817 0.6585183

4 0.6091873 0.3908127

5 0.5805008 0.4194992

**3.3. Summarize the results in a Confusion Matrix.**

Predicted

Actual 0 1

0 120 30

1 35 115

**3.4. As demonstrated in class, calculate the time (in seconds) it took to fit the model and include this in your summary.**

Time difference of **0.02100611 secs**

**4 Compare All Three Classifiers For all questions below please provide evidence.**

**4.1. Which classifier is most accurate? (provide evidence)**

Accuracy of Logistic Regression – Stepwise (forward) model = **0.783**

Accuracy of Naïve-Bayes Classification = **0.780**

Accuracy of Linear Discriminant Analysis = **0.783**

Both **Logistic Regression – Stepwise (forward) model** and **Linear Discriminant Analysis** classifier are most accurate.

**4.2. Which classifier is most suitable when processing speed is most important?**

The processing speed of Logistic Regression – Stepwise (forward) model = **0.03300381 secs**

The processing speed of Naïve-Bayes Classification = **0.02400398 secs**

The processing speed of Linear Discriminant Analysis = **0.02100611 secs**

**Linear Discriminant Analysis** is most suitable when processing speed is most important.

**4.3. Which classifier minimizes Type 1 errors?**

The Type 1 errorof Logistic Regression – Stepwise (forward) model = **30**

The Type 1 errorof Naïve-Bayes Classification = **32**

The Type 1 errorof Linear Discriminant Analysis = **30**

Both **Logistic Regression – Stepwise (forward) model** and **Linear Discriminant Analysis** classifier minimizes Type 1 errors.

**4.4. Which classifier minimizes Type 2 errors?**

The Type 2 errorof Logistic Regression – Stepwise (forward) model = **35**

The Type 2 errorof Naïve-Bayes Classification = **34**

The Type 2 errorof Linear Discriminant Analysis = **35**

**Naïve-Bayes Classification** minimizes Type 2 errors.

**4.5. Which classifier is best overall?**

As we are predicting Tumor (diseases), **Type 2 error** (False Negative) should be **less,** and **accuracy, processing speed** should also be **good.**

So, **Naïve-Bayes classifier** is best overall.

**4.6. How do these classifiers compare to the best model you built in Part 1?**

**Accuracy,**

Accuracy of Logistic Regression – Stepwise (forward) model = **0.783**

Accuracy of Naïve-Bayes Classification = **0.780**

Accuracy of Linear Discriminant Analysis = **0.783**

Accuracy of User Model 2 **(best model in Part 1)** = **0.796**

**User Model 2 (the best model in Part 1)** has the best Accuracy.

**Type 2 errors,**

The Type 2 errorof Logistic Regression – Stepwise (forward) model = **35**

The Type 2 errorof Naïve-Bayes Classification = **34**

The Type 2 errorof Linear Discriminant Analysis = **35**

The Type 2 errorof User Model 2 **(the best model in Part 1)** = **32**

**User Model 2 (the best model in Part 1)** minimizes Type 2 errors.

**Type 1 errors,**

The Type 1 errorof Logistic Regression – Stepwise (forward) model = **30**

The Type 1 errorof Naïve-Bayes Classification = **32**

The Type 1 errorof Linear Discriminant Analysis = **30**

The Type 1 errorof User Model 2 **(the best model in Part 1)** = **29**

**User Model 2 (the best model in Part 1)** minimizes Type 1 errors.

So, **User Model 2 (the best model in Part 1)** is the best overall as it has the best **Accuracy, Type 2 errors,** and **Type 1 errors.**

**5 Professionalism and Clarity**