Formulating a Prediction Model to Determine Outcome of a Tumor in Breast Cancer Patients

Aditya Pendyala

2023-12-10

Introduction:

This dataset "cancer" has data from a study of breast cancer in Wisconsin. The purpose of the study was to determine whether a new procedure called fine needle aspiration which draws only a small sample of tissue could be effective in determining tumor status.
The data has 10 variables which also consists of a Bernoulli variable "Class" which is the tumor status. The variables are:

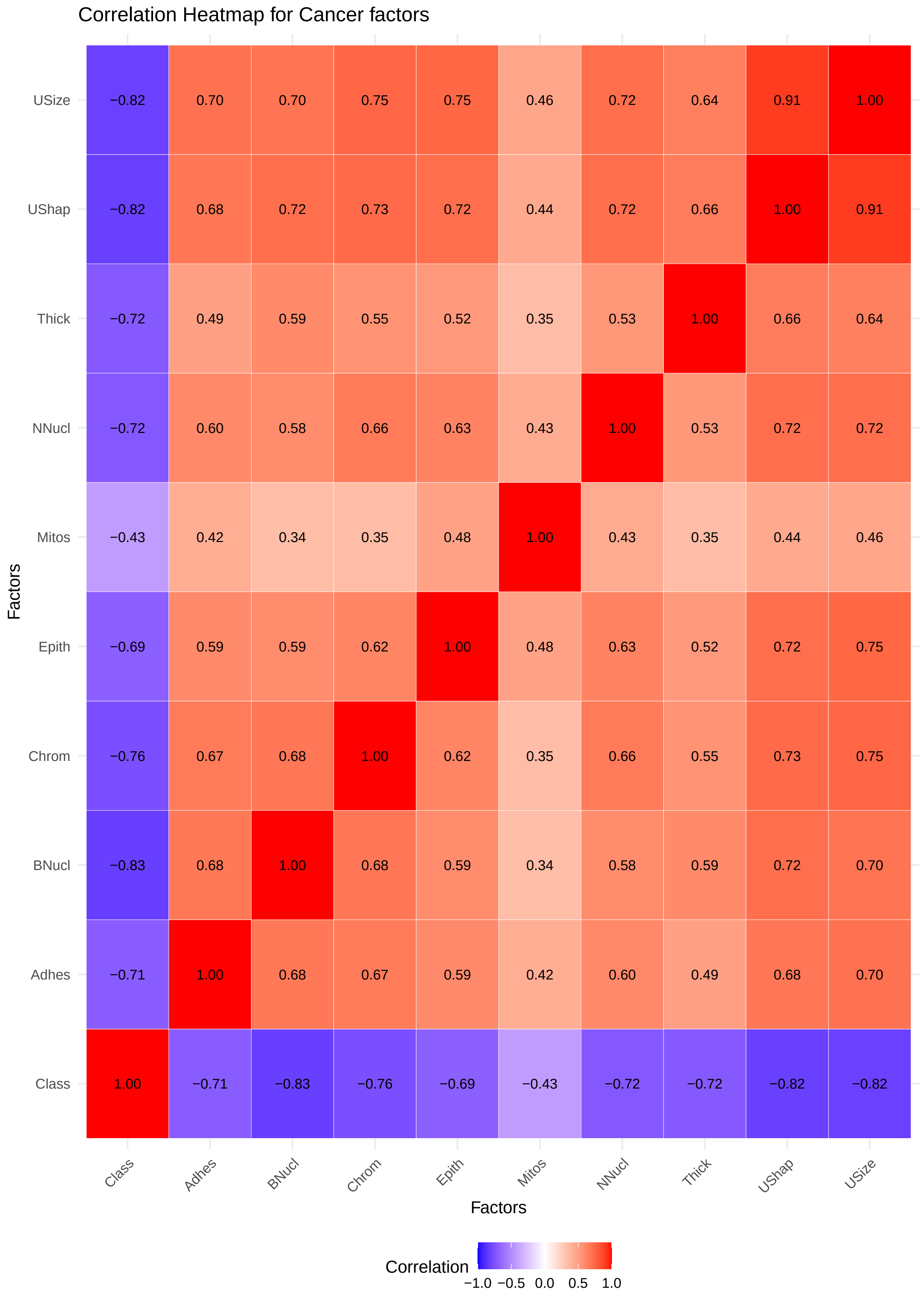
1. Class: 0 if malignant, 1 if benign
2. Adhes: marginal adhesion
3. BNucl: bare nuclei
4. Chrom: bland chromatin
5. Epith: epithelial cell size
6. Mitos: mitoses
7. NNucl: normal nucleoli
8. Thick: clump thickness
9. UShap: cell shape uniformity
10. USize: cell size uniformity

Purpose:

The goal of this project is to formulate a prediction model, that predicts the outcome of a tumor in breast cancer patients. This project aims to compare different approaches to devising a prediction model.

Exploring the dataset:

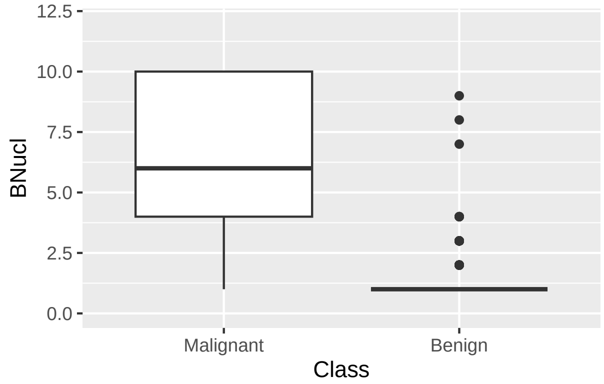
First step in analyzing the dataset is to find the correlation of the factors. By doing this we can get a better understanding of which variables can be the possible factors for determining the Class.


Correlation Heatmap for Cancer factors
USize −0.82 0.70 0.70 0.75 0.75 0.46 0.72 0.64 0.91 1.00
UShap −0.82 0.68 0.72 0.73 0.72 0.44 0.72 0.66 1.00 0.91
Thick −0.72 0.49 0.59 0.55 0.52 0.35 0.53 1.00 0.66 0.64
NNucl −0.72 0.60 0.58 0.66 0.63 0.43 1.00 0.53 0.72 0.72
Mitos −0.43 0.42 0.34 0.35 0.48 1.00 0.43 0.35 0.44 0.46
Epith −0.69 0.59 0.59 0.62 1.00 0.48 0.63 0.52 0.72 0.75
Chrom −0.76 0.67 0.68 1.00 0.62 0.35 0.66 0.55 0.73 0.75
BNucl −0.83 0.68 1.00 0.68 0.59 0.34 0.58 0.59 0.72 0.70
Adhes −0.71 1.00 0.68 0.67 0.59 0.42 0.60 0.49 0.68 0.70
Class 1.00 −0.71 −0.83 −0.76 −0.69 −0.43 −0.72 −0.72 −0.82 −0.82
Factors
Class
Adhes
BNucl
Chrom
Epith
Mitos
NNucl
Thick
UShap
USize
Factors
Correlation
−1.0 −0.5 0.0 0.5 1.0

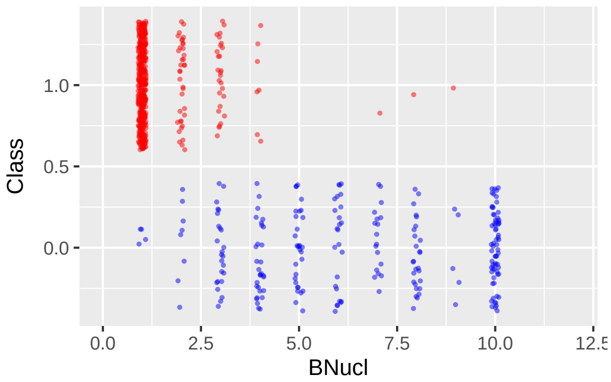
Looking at the heat map, we can see that most of the variables associated with Class have large negative correlations. Thus, we can expect the prediction model to have most of the variables.

Now we are creating box plots and scatter plots to understand each factor's relationship with Class.

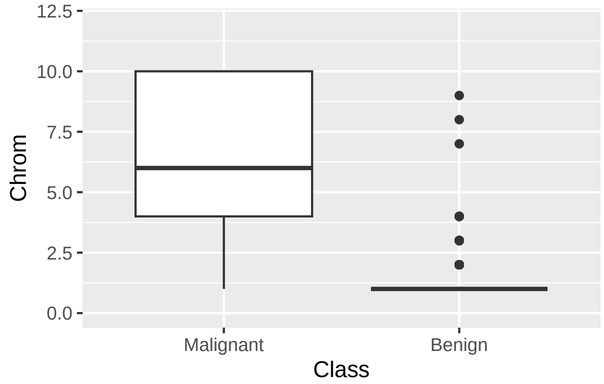
Box plot of BNucl by Class



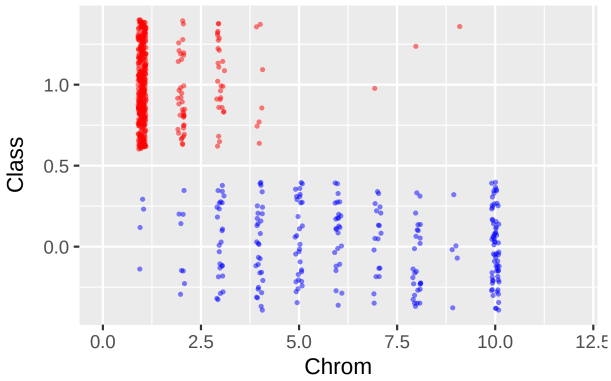
Scatter plot of BNucl by Class



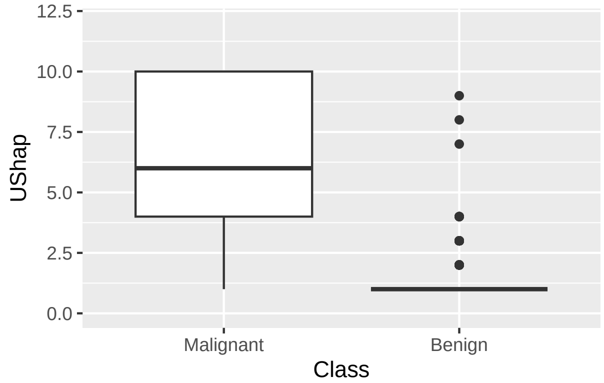
Box plot of Chrom by Class



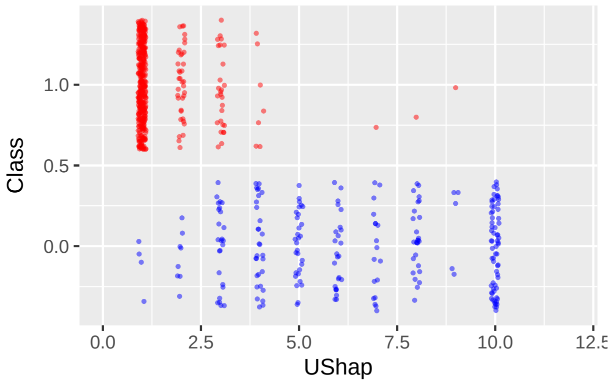
Scatter plot of Chrom by Class



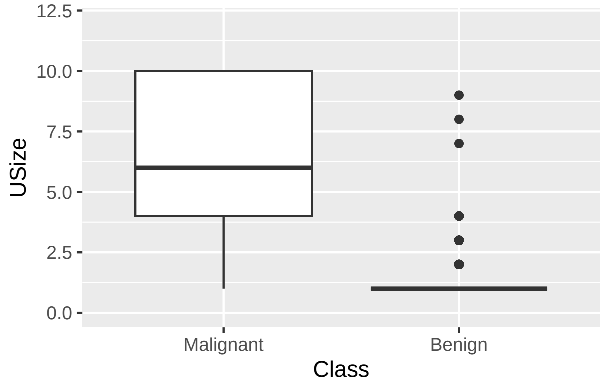
Box plot of UShap by Class



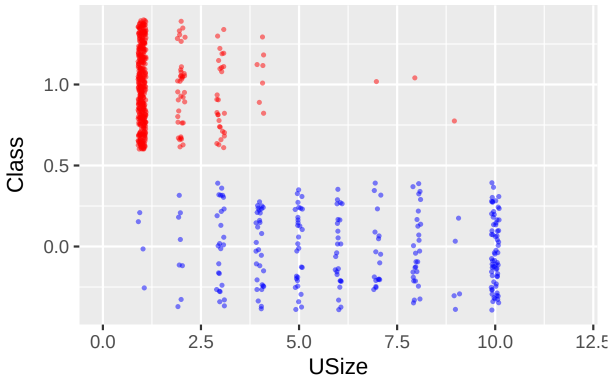
Scatter plot of UShap by Class



Box plot of USize by Class



Scatter plot of USize by Class



Looking at the correlation matrix and the plots, we can see that many variables have good correlation to the "Class" variable. This means that we can expect our ideal prediction model to include many of these terms.

Prediction and analysis:

In order to devise the best fitting prediction model, we can find and compare different models. We can choose MSE values as a factor for comparison. We find the prediction model by first using a training dataset, to train the model. Later we fit the prediction model on the testing data to check the accuracy of the model.

1. Finding a prediction model using basic glm.
   We formulate a basic model that considers all the variables as prediction variables.

```
## Coefficients:

## (Intercept)       Adhes       BNucl       Chrom       Epith       Mitos
## 11.05275905 -0.32453598 -0.48879988 -0.62307762  0.06308323 -0.72562562
##       NNucl       Thick       UShap       USize
## -0.34792787 -0.50558267 -0.35975185  0.11699047

## MSE: 0.02476076
```

We can see that the mean standard error of the model is 0.025. Though this is a small value, there's a possibility of over fitting. Over fitting can cause bad prediction results when applied to new data.

2. Finding a prediction model using AIC.
   Since the variable we are trying to predict is Bernoulli, we can perform a logistic regression. One model selection criterion for logistic regression is AIC(Akaike Information Criterion). AIC measures the closeness of the candidate model to the true one.

```
## Morgan-Tatar search since family is non-gaussian.

## Coefficients:

## (Intercept)       Adhes       BNucl       Chrom       Mitos       NNucl
##  11.4457622  -0.3403509  -0.5281079  -0.6655713  -0.7634021  -0.3800429
##       Thick
##  -0.6262485

## MSE: 0.02376475
```

We can see that the mean standard error of the model is 0.024. This is slightly larger than the basic glm. But, unlike the regular model we can see that this model only considers "Adhes", "BNucl", "Chrom", "Mitos", "NNucl", "Thick" as prediction variables. This is fewer than the basic model.

3. Finding a prediction model using BIC glm.
   Another model selection criterion for logistic regression is BIC(Bayesian information criterion). IC (with negative sign) measures how likely the candidate model is the true model from Bayesian perspective.

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## Coefficients:
```

```
## (Intercept)       BNucl       Chrom       Mitos       NNucl       Thick
##  10.7116707  -0.5730144  -0.7115998  -0.7216782  -0.3978999  -0.5922187
```

```
## MSE: 0.02656303
```

We can see that the mean standard error of the model is 0.027. This is slightly larger than the AIC model. This model selected “BNucl”, “Chrom”, “Mitos”, “NNucl”, “Thick” as prediction variables. This is fewer than AIC.

4. Performing Bayesian Model Selection.

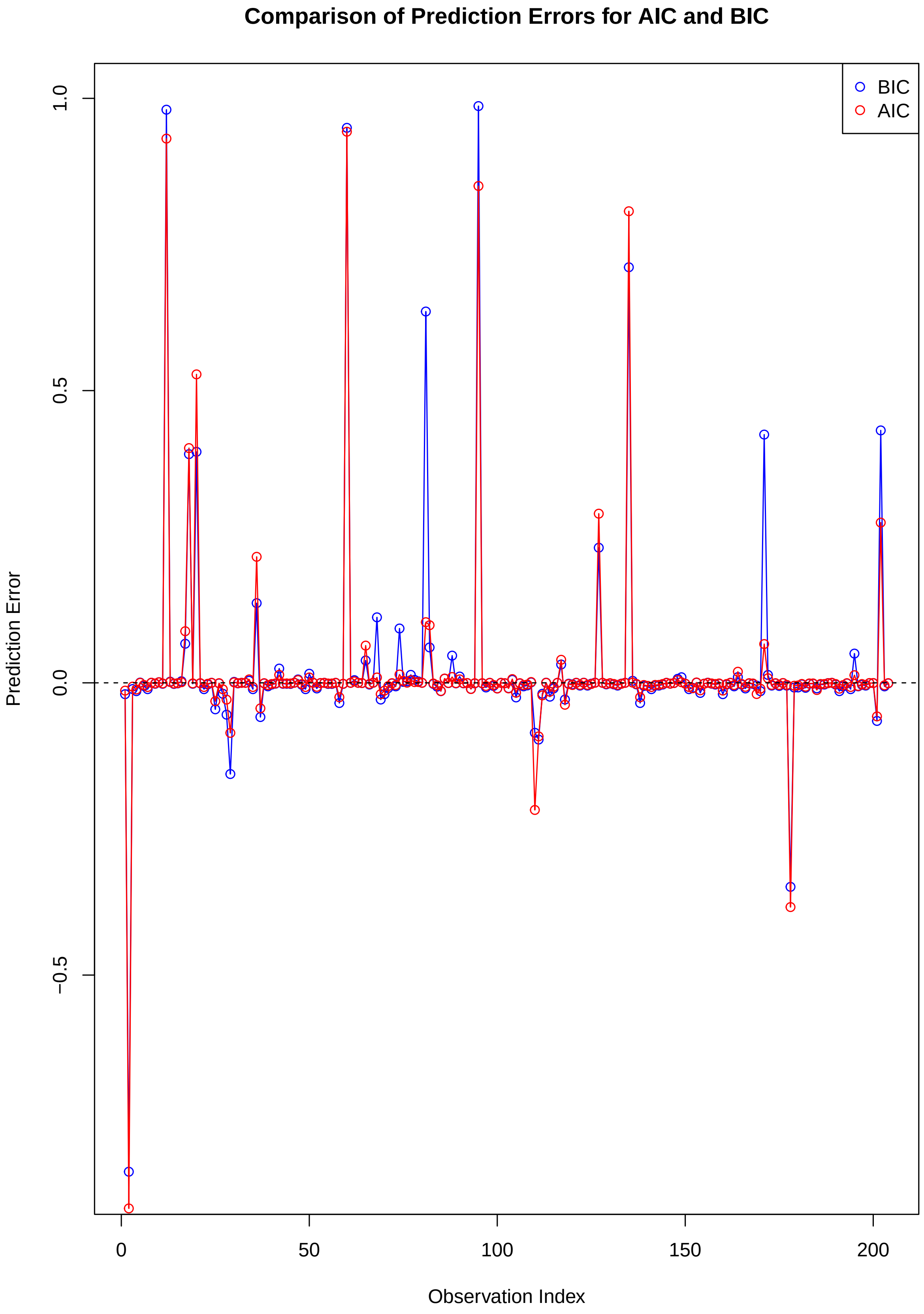
Since this is a logistic regression, another way of finding a prediction model is by performing a Bayesian Model Selection. We use MCMC sampling to find Beta values that are used to find the predicted y values.

We use B and Z from the MCMC algorithm to find Beta, and then use Beta to predict values for Class. We also applied a threshold of 0.5 to select our variables.

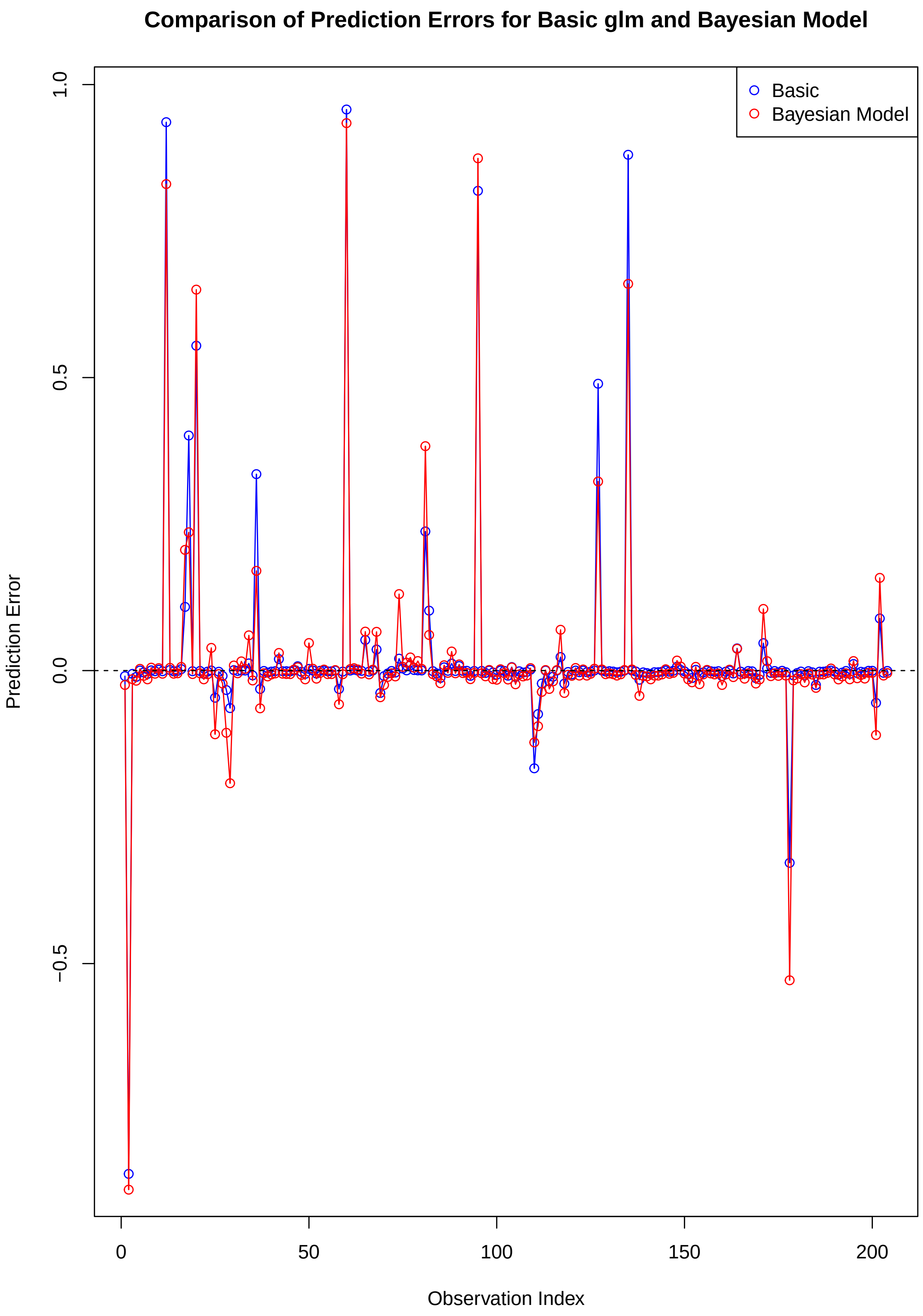
```
## MSE: 0.02370756
```

We can see that the MSE for Bayesian Model Selection is 0.022 which is better than all the other models we found.

Looking at the different MSE values, we can see that for the Bayesian Model Selection we have the lowest Mean Squared Error. This is an indicator of a good prediction model. The comparison of the MSE values with other models, is consistent with statistical findings. Usually, BIC can fit worse than AIC models, and the same can be observed with this data.


Comparison of Prediction Errors for AIC and BIC
○ BIC
○ AIC
Prediction Error
1.0
0.5
0.0
−0.5
0
50
100
150
200
Observation Index

Looking at this plot, we can see that the BIC has a larger prediction error compared to AIC at many points in the graph. This shows that the AIC Model has a lower prediction error value.


Comparison of Prediction Errors for Basic glm and Bayesian Model
○ Basic
○ Bayesian Model
Prediction Error
1.0
0.5
0.0
−0.5
0
50
100
150
200
Observation Index

Looking at this plot, we can see that the Bayesian Model has a smaller prediction error compared to Basic in many points in the graph. This shows that the Bayesian Model has a lower prediction error value.

Conclusion:

On comparison with different models, such as AIC, BIC, and glm it can be inferred that the Bayesian Model has a lower MSE prediction value and high accuracy, making it a suitable choice to predict the tumor status. These findings may change according to the choice of data split, and also the nature of the data. Furthermore, this project compares different models on the basis of MSE, which may not be the universal choice of comparison. Other factors of comparison may include computational power and memory. If the mentioned was the criteria, we can say that the Bayesian Model is memory and power heavy, making it an inefficient choice.