Predicting the Presence Breast Cancer Using Blood Test Data

Members: Edward Wang, Andrew Tran, Aayush Sharma, Raymond Li

1. Introduction:

Cancer is a class of over 100 diseases involving uncontrolled cell growth and the formation of life-threatening cell masses known as tumors (National Institute of Health, 2007). Of all cancers, breast cancer is the most common in women, where 2.3 million women were diagnosed in 2020 alone (Cicinas, 2021). Studies have found certain biomarkers to be predictive of various cancers, including body mass index (BMI), glucose, insulin, the homeostatic model assessment (HOMA), leptin, adiponectin, resistin, and monocyte chemoattractant protein-1 (MCP.1) (Deshmane et al., 2009; Jamaluddin et al., 2012; Lappin & Hantzidiamantis, 2022; Macunso, 2016; Nuttal, 2015; Wallace et al., 2004; Wilcox, 2005). By analysing a dataset from University of California Irvine Machine Learning Repository, the present study aims to explore the relationship between the aforementioned biomarkers and breast cancer.

2. Analysis Methods:

We use the Breast Cancer Coimbra Data Set from the University of California Irvine Machine Learning Repository for our analysis. The dataset contains 116 total records representing 64 patients with breast cancer and 52 control participants. The age, BMI, HOMA, and levels of glucose, insulin, leptin, adiponectin, resistin, and MCP.1 were collected from all participants to be used as potential predictors for breast cancer. Below, we include the explanations of the variables present in our data set:

i. **Breast Cancer (Binary, Response)**: Breast cancer is a disease involving the uncontrolled growth of cells in the breast (Centre for Disease Control, 2022). Participants were classified as a patient (Classification = 1) if they were diagnosed with breast cancer, and as healthy (Classification = 0) if they did not have a diagnosis of breast cancer.

ii. **Age (Discrete, Explanatory)**: Age here is measured in years.

iii. **BMI (Continuous, Explanatory):** Body Mass Index (BMI) is a measure of adult body weight calculated by taking the kilogram body weight divided by height in squared metres. A high BMI indicates excess body fat and has been associated with various health risks such as under-nutrition, obesity, and cancer (Nuttal, 2015).

iv. **Glucose (Continuous, Explanatory):** Glucose is a sugar commonly found in foods, and is measured in milligrams per deciliter. When food is eaten, the carbohydrates in the food are broken down into glucose, raising blood glucose levels, which can either be converted into adenosine triphosphate (ATP) for immediate energy, or converted into glycogen for stored energy (Lappin & Hantzidiamantis, 2022).

v. **Insulin (Continuous, Explanatory):** Insulin is a hormone that regulates blood sugar levels measured in micro-units per milliliter. In response to food, insulin releases into the bloodstream from the pancreas to transport glucose into cells for later use as energy. High levels of insulin are maladaptive because of hypoglycaemia (lack of blood sugar) and insulin resistance (lack of response to insulin, resulting in overly high blood sugar), and imflammation. The latter two risk factors are associated with obesity and cancers respectively. (Wilcox, 2005).

vi. **HOMA (Continuous, Explanatory):** Homeostatic Model Assessment (HOMA) is a method used to estimate insulin resistance and the progression of diabetes. However, methodologists such as Wallace have indicated that insulin itself is a better measure of insulin resistance compared to HOMA, despite its' prevalent use (Wallace et al., 2004).

vii. **Leptin (Continuous, Explanatory):** Leptin is a peptide hormone measured in units of nanograms per millilitre. It is shown to proliferate, migrate, and invade within breast cancer cell lines, possibly promoting the development of cancer (Macunso, 2016).

viii. **Adiponectin (Continuous, Explanatory):** Adiponectin is an amino acid protein measured in micrograms per milliliter. It is a hormone with antiatherogenic properties (able to reduce artery plaque build-up) while affecting insulin sensitivity and inflammation. Abnormally low levels of adiponectin have been associated with cancer (Macunso, 2016).

ix. **Resistin (Continuous, Explanatory):** Resistin - named for its' ability to resist insulin - is a hormone measured in nanograms per millilitre. It is produced by fat cells that contributes to insulin resistance, inflammation, and diabetes, which in turn have been associated with cancer development (Jamaluddin et al., 2012).

x. **MCP.1 (Continuous, Explanatory):** Monocyte chemoattractant protein-1 (MCP.1) is measured in picograms per deciliter and is a chemokine (protein associated with cell trafficking) associated with immune response and inflammatory processes and is involved in the development of obesity. They are hypothesized to play a role in the development of cancer through growth-altering influence on cancer cells (Deshmane et al., 2009).

2.1. Data Analytic Plan:

Our data is split randomly between training (0.8)& test (0.2) such that the training set is used to fit & select models and test set to report our results in terms of best model selected. To explore the patterns in our given data, we will visualize the training set: using boxplot & histogram on the explanatory variables, we wish to detect any outliers & observe distributional patterns of the selected biomarkers. Descriptive statistics will be calculated in terms of values of explanatory variables in the training set. Correlation between explanatory variables will be examined, further decisions on high-correlated variables will be made using the VIF criteria from the baseline model.

We use the Akaike Information Criterion (AIC) as our criteria for model selection. Given a logistic regression model, AIC computes a score that decreases when the estimated parameters of the model are more likely given the data but increases as you place more parameters into the model, or the parameters are less likely. Therefore, with given data, AIC is a good measurement of how well the model fits as it selects models with most likely parameters & also penalizes on overfitting by simply feeding more terms into the models. AIC can be computed as follows:

$$AIC = -2(l(\hat{\pi}; y) - p)$$

Where $l(\hat{\pi}, y)$ is the log-likelihood of the given data & p is the number of parameters estimated in the data.

After removing highly collinear variables, we will eventually build our models from two methods: 1). we will use "bestglm" package to select the best model (in terms of AIC) without interaction & then add in the interaction terms on our reduced set to achieve the minimum AIC with variables in the best model without interaction, 2). we will use our intuition to compare variables & add interaction terms. The model among two with lower AIC will be our final model, and lastly, we will evaluate it using our test sets & report our results in terms of confusion matrix & Receiver Operating Characteristic plot (ROC).

3. Analysis Results:

3.1 Data Analysis Results:

3.1.1. Outliers Detection & Distribution of the Explanatory Variables:

Based on our visualization of values of explanatory variables in the training set, we observe that Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP.1 are all right skewed with outliers found on the upper tail of each one of mentioned variables. Notice that Glucose also has an outlier on the lower tail. There are no outliers on Age or BMI, but their distribution appears to be bimodal. Since we haven't fitted any model yet, we take notes of the presence of these outliers but choose not to prune them. As nine explanatory variables in our data represent different levels of some human body measurements, we intuitively speculate that these outliers may be helpful in identifying individuals with breast cancer.

From table 1, we see that not all variables are on the same scale: notably, Leptin & Adiponectin have smaller units than other variables & MCP-1 has a significantly bigger scale & variance than others. However, this difference in scale should not influence our linear model more than scaling some coefficients.

3.1.2. Correlation Between the Explanatory Variables:

From the heatmap & the paired plot, we observe that Leptin & BMI (0.56), Glucose & Insulin (0.54), Glucose & HOMA (0.73), Insulin & HOMA (0.93) are highly correlated which might cause unstable estimates of coefficients in the output model (due to collinearity). However, we choose not to prune the variables now due to high correlation but rather use the VIF values from the baseline model as our criteria if any highly collinear variables found.

3.2. Model Selection Results:

3.2.1. Removing Variables with High Collinearity:

Using the full model without interaction as our baseline, we calculate VIF values as shown in table 2. From our correlation plot, Glucose & Insulin (0.54), Glucose & HOMA (0.73), Insulin & HOMA (0.93) are highly correlated. This information combined with high VIF values calculated in table 2 suggests high collinearity (>10) in Insulin (75.03) & HOMA (82.11), which leads us to remove these two variables. In reality, the high correlation & collinearity among HOMA & Insulin can be partially explained by the fact that HOMA is a method used to quantify Insulin resistance and beta-cell function, and as for Insulin & glucose, we know insulin is a hormone that metabolizes glucose, which explains the relatively strong correlation between the two variables.

Baseline Model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \widehat{\beta_0} + \widehat{\beta_{age}} x_{i,age} + \widehat{\beta_{BMI}} x_{i,BMI} + \widehat{\beta_{Glucose}} x_{i,Glucose} + \widehat{\beta_{Insulin}} x_{i,Insulin} + \widehat{\beta_{HOMA}} x_{i,HOMA} + \widehat{\beta_{Leptin}} x_{i,Leptin} + \widehat{\beta_{Adiponectin}} x_{i,Adiponectin} + \widehat{\beta_{Resistin}} x_{i,Resistin} + \widehat{\beta_{MCP.1}} x_{i,MCP.1}$$

3.2.2. Selecting Variables & Interaction Terms Based on Intuition:

After removing two highly collinear variables (HOMA & Insulin), we put our focus on Leptin & BMI, the remaining pair of the highly correlated explanatory variables (0.56). We notice that there is also a possibility that Leptin and BMI representing duplicated information as a lack of Leptin can lead to obesity (Macunso, 2016), in which case the BMI of the subject would be high as well. In order to test whether Leptin and BMI represent a similar criterion in this data set, different regression models were constructed, and their AIC was compared to each other, leading to the removal of Leptin to achieve a lower AIC.

When selecting any possible significant interaction terms, we first looked into age as it can be safely assumed that age can interact will all other variables as a person gets more prone to adverse effects of irregular biological indicators as they get older. Upon further testing, it was clear that including the interaction terms of age with every other variable decreased the AIC of our regression model and lead to these interaction terms being added into the manual model.

The next variable which we looked into was Adiponectin, as it is a protein hormone that is involved in the process of regulating glucose and fatty acid breakdown (Macunso, 2016), which suggests that there might be a significant interaction between Adiponectin and BMI and Glucose. This is backed up by the fact that there is a decrease in AIC when adding these interaction terms in the manual model.

Another variable of interest is Resistin, which is theorized to link obesity to diabetes and similar to the case of Adiponectin, including the interaction terms between Resistin and BMI and Glucose leads to a decrease in the AIC in the manual model.

The best model after manual selection contained the variables Age, BMI, Glucose, Adiponectin, Resistin, MCP.1 and the interaction terms of Age with the other 5 variables (Output 1), Adiponectin with BMI and Glucose, and Resistin with BMI and Glucose. the regression model derived from these variables and interaction terms had AIC of 81.364.

3.2.3. Using Exhaustive Methods to Find Best Models in terms of AIC:

Using the bestglm() function, we exhaustively tested the variables given in the original training set excluding the variables HOMA and Insulin due to their high collinearity and the model obtained contained 5 variables, namely: Age, BMI, Glucose, Resistin, and MCP.1 and had an AIC of 100.06 (Output 2).

The variables present in the model above are used as a basis to fit the best model with interactions using the glmulti() function. We decided to use only the variables found in the model without interaction for the sake of reducing the computational costs, and it does not make sense to exclude those variables for valid reasons while including their interaction terms. The best model obtained through these computations contained the variables Glucose, Resistin, and the interaction terms between BMI : Age, Glucose : Age, and Resistin: BMI and had an AIC of 78.976 (Output 3).

4. Conclusion:

As we use AIC as our model selection criteria, our final model would be the model with interaction found using the exhaustive method since it has lower AIC value than the interaction model fitted based on intuition (78.976 < 81.364). Evaluating the model based on the test set using the default cut-off of 0.5, we obtained a result of 79.2% overall accuracy. According to the confusion matrix, among 5 misdiagnoses (out of 24 test cases in total) observed, the model fails to diagnose 3 patients with cancer & mistakenly identify 2 healthy controls as patients (Table 3).

In general, we would consider this model sufficient to be a good reference when identifying breast cancer and would argue that it might perform better if we use cross-validation to train a better cut-off using the training data. However, at the same time, there are also concerns about suboptimal model with interactions obtained due to computational constraints as we fail to find the best model with interactions using the exhaustive method on the entire set of explanatory variables. Beyond that, there are also questions raised about using AIC or log-likelihood based criteria to select model as maximizing them do not necessarily lead to the optimal solution (this can be seen from graph 12 & 13 as the manually fitted model has lower AIC value but appears better in terms of AUC when evaluating it against the test set).

5. Reference:

1. Centre for Disease Control. (2022, March 9). What Is Breast Cancer? Centres for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm.

2. Deshmane, S. L., Kremlev, S., Amini, S., & Sawaya, B. E. (2009). Monocyte Chemoattractant Protein-1 (MCP-1): An Overview. Journal of Interferon & Cytokine Research, 29(6), 313–326. https://doi.org/10. 1089/jir.2008.0027.

3. Hantzidiamantis, P. J., & Lappin, S. L. (2023). Physiology, Glucose. In StatPearls. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK545201/.

4. Jamaluddin, M. S., Weakley, S. M., Yao, Q., & Chen, C. (2012). Resistin: Functional roles and therapeutic considerations for cardiovascular disease. British Journal of Pharmacology, 165(3), 622–632. https://doi.org/10.1111/j.1476-5381.2011.01369.x.

5. Mancuso, P. (2016). The role of adipokines in chronic inflammation. ImmunoTargets and Therapy, 5, 47–56. https://doi.org/10.2147/ITT.S73223.

6. National Institutes of Health, & Study, B. S. C.(2007). Understanding Cancer. In NIH Curriculum Supplement Series [Internet]. National Institutes of Health (US). https://www.ncbi.nlm.nih.gov/books/NBK20362/.

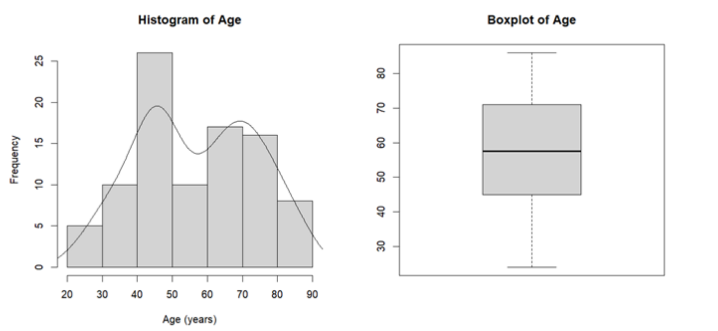
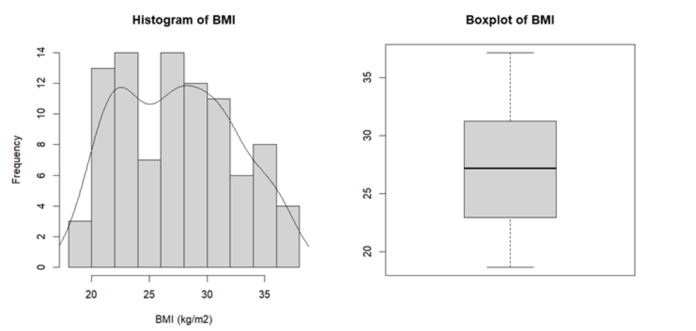
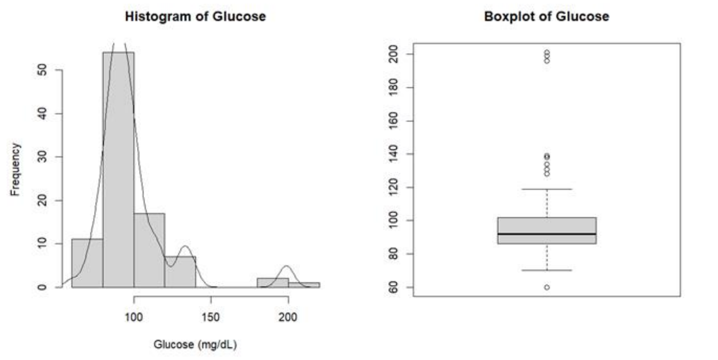
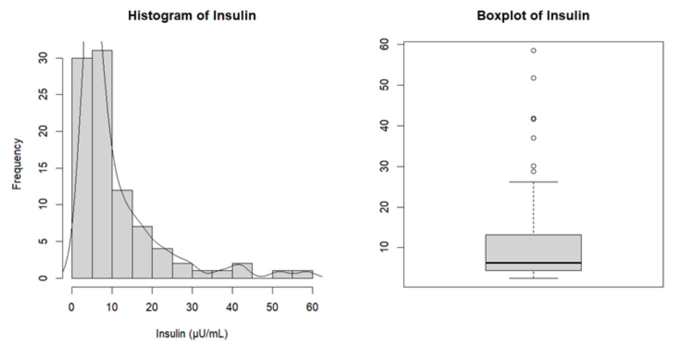
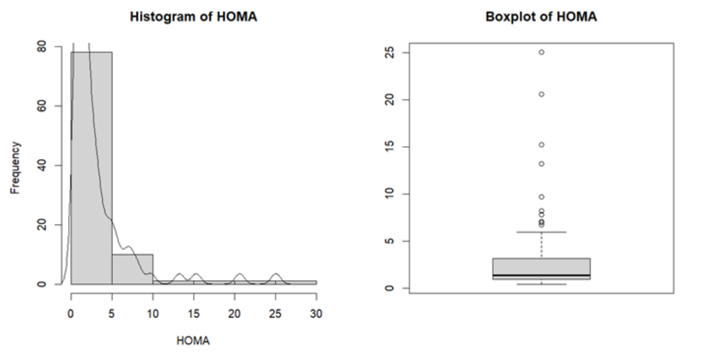
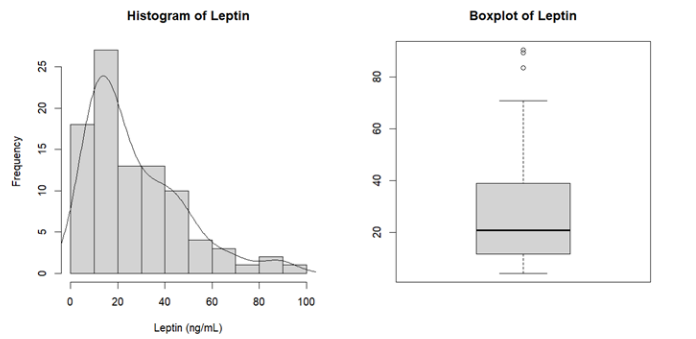
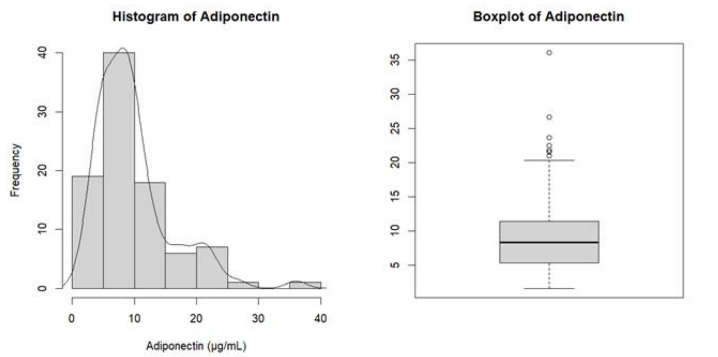
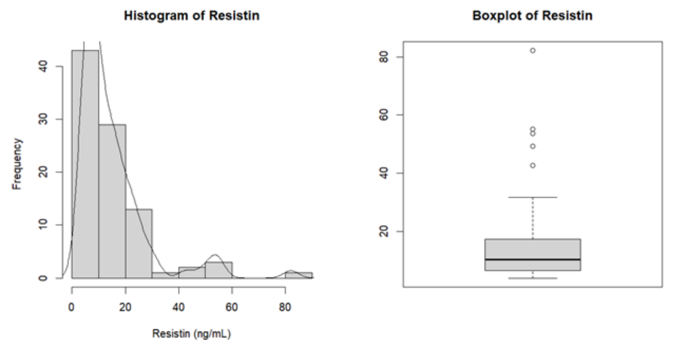
7. Nuttall, F. Q. (2015). Body Mass Index: Obesity, BMI, and Health: A Critical Review. Nutrition Today, 50(3), 117–128. https://doi.org/10.1097/NT.0000000000000092 UCI Machine Learning Repository. (n.d.). Retrieved April 8, 2023, from https://archive.ics.uci.edu/ml/index.php.

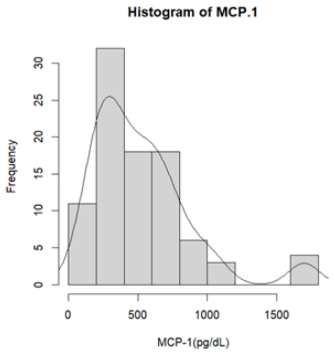
8. Wallace, T. M., Levy, J. C., & Matthews, D. R. (2004). Use and abuse of HOMA modeling. Diabetes Care, 27(6), 1487–1495. https://doi.org/10.2337/diacare.27.6.1487.

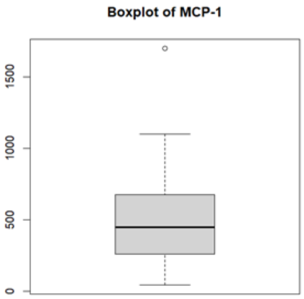
9. Wilcox, G. (2005). Insulin and Insulin Resistance. Clinical Biochemist Reviews, 26(2), 19–39.

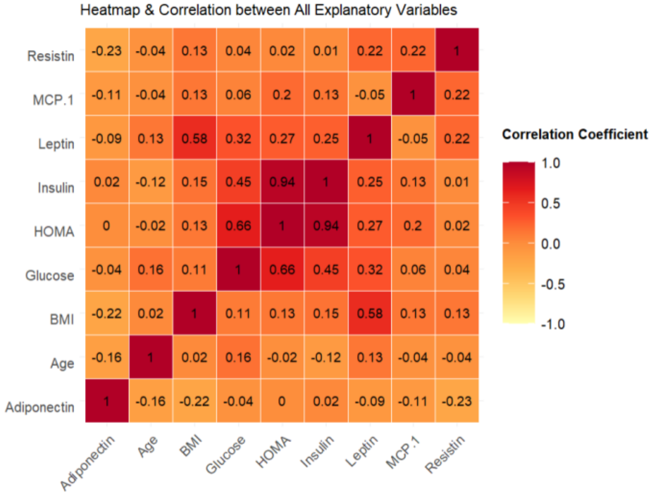
Appendix:

6. Graphs:

Graph 1. Histogram & Boxplot of Age	Graph 2: Histogram & Boxplot of BMI
	
Graph 3. Histogram & Boxplot of Glucose	Graph 4. Histogram & Boxplot of Insulin
	
Graph 5. Histogram & Boxplot of HOMA	Graph 6. Histogram & Boxplot of Leptin
	
Graph 7. Histogram & Boxplot of Adiponectin	Graph 8. Histogram & Boxplot of Resistin
	
Graph 9. Histogram & Boxplot of MCP.1	Graph 10. Heatmap & Correlation between All Explanatory Variables

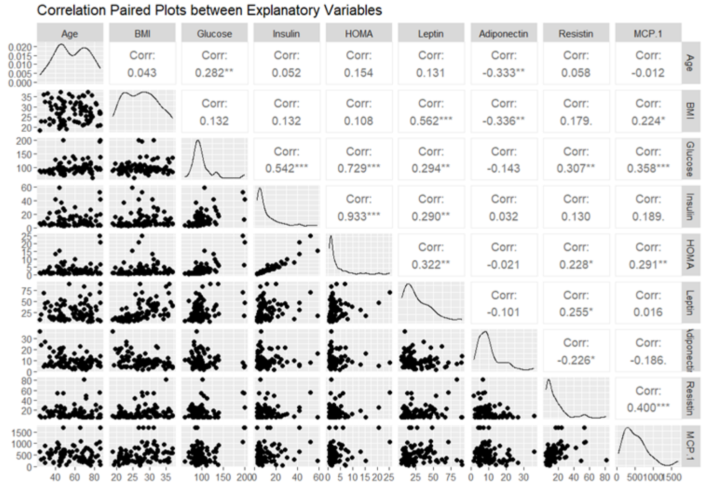


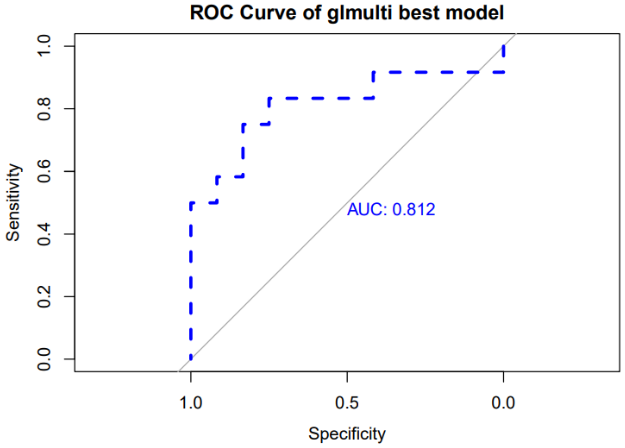




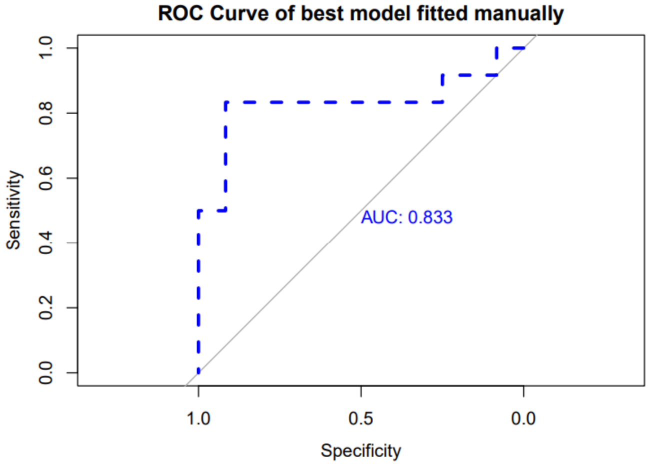
Graph 11. Correlation Paired Plots between Explanatory Variables

Graph 12. ROC Plot of the Model with Best AIC





Graph 13: ROC Plot of Best Model Fitted Manually



7. Tables:

Table 1: Summary Statistics of the Values of Explanatory Variables in the Training Set (using 2 decimal places):

Variables:	Mean:	Median:	Max:	Min:	SD:
Age (years)	56.95	57.50	86.00	24.00	16.59
BMI (kg/m2)	27.40	27.19	37.11	18.67	4.93
Glucose (mg/dL)	98.66	92.00	201.00	60.00	24.02
Insulin (µU/mL)	10.89	6.34	58.46	2.54	10.75
HOMA	2.99	1.42	25.05	0.47	3.99
Leptin (ng/mL)	27.16	20.83	90.28	4.31	19.87
Adiponectin (µg/mL)	9.81	8.35	36.06	1.66	6.09
Resistin (ng/mL)	14.87	10.34	82.10	4.06	13.24
MCP-1(pg/dL)	525.05	446.60	1698.44	45.84	355.66

Table 2: VIF Values of the Full Model without Interaction (2 decimal places) :

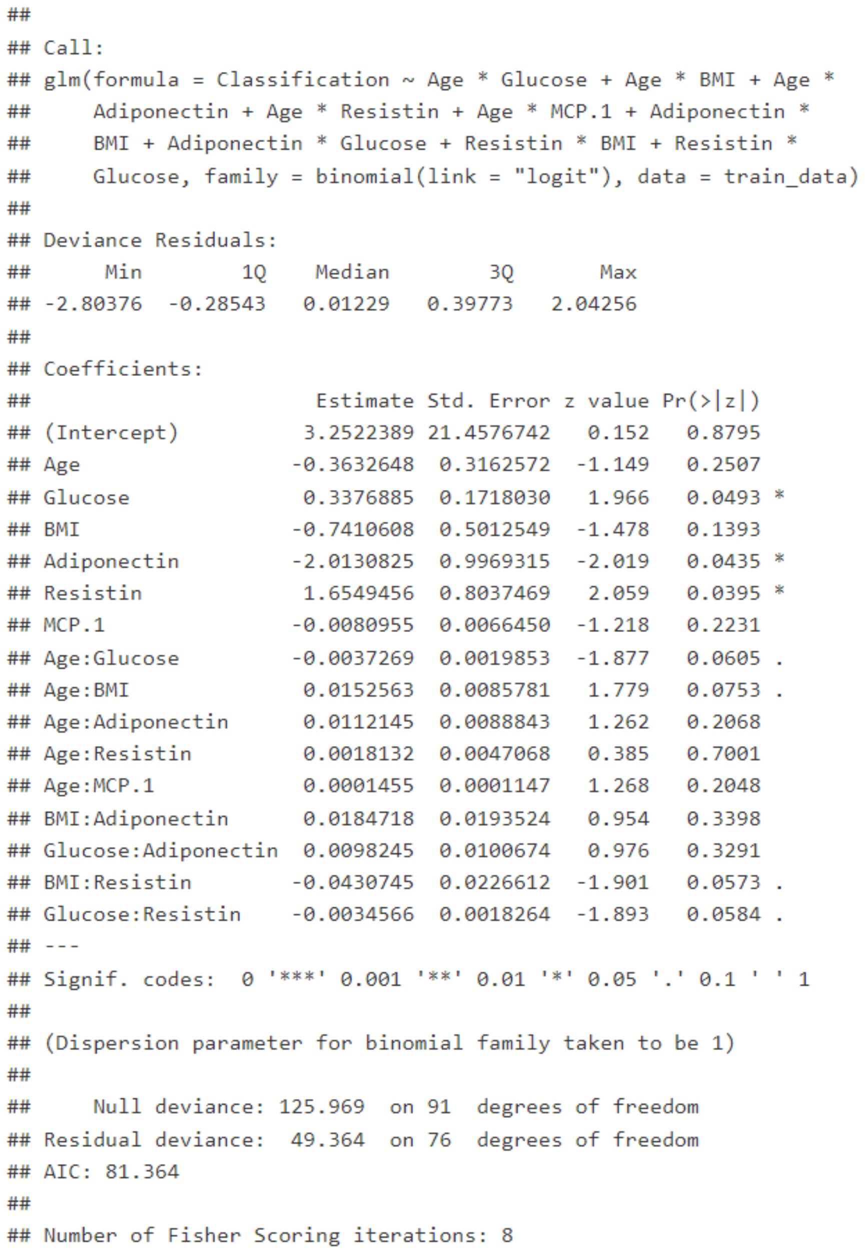
Age	BMI	Glucose	Insulin	HOMA
1.42	2.43	3.40	75.03	82.11
Leptin	Adiponectin	Resistin	MCP-1	
2.14	1.35	1.21	1.32	

Table 3: Confusion Matrix of the Best Model with Interaction Found Using Exhaustive Method:

Prediction\ Reference	0	1
0	9	2
1	3	10

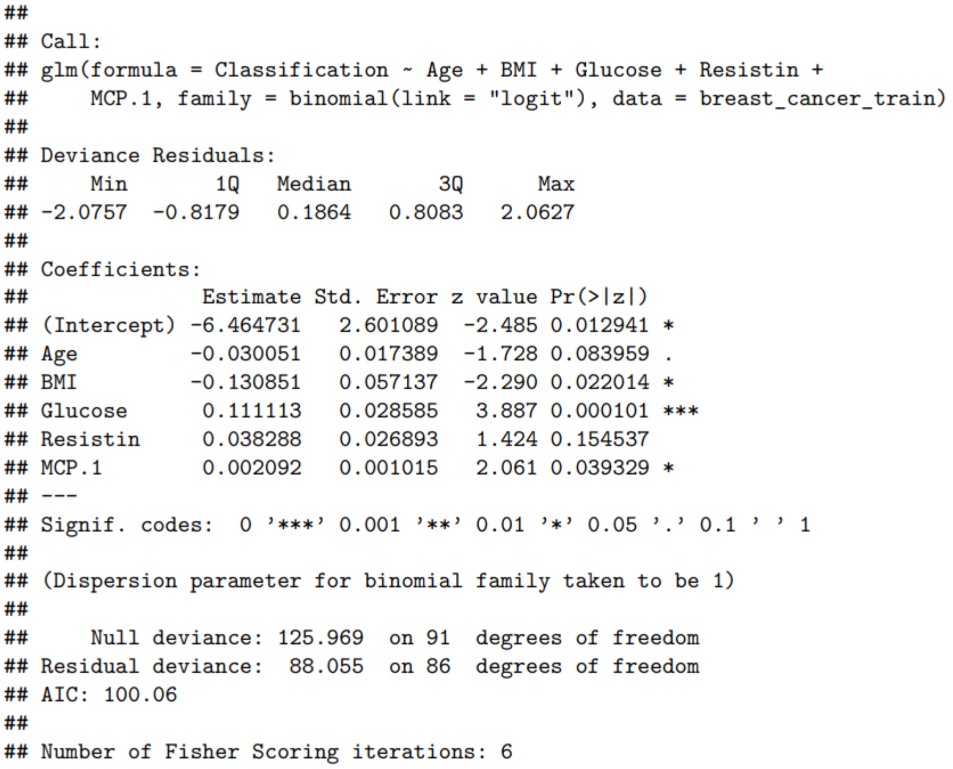
8. R Outputs:

Output 1: R Summary of Manual Selected Model with Interaction Terms:



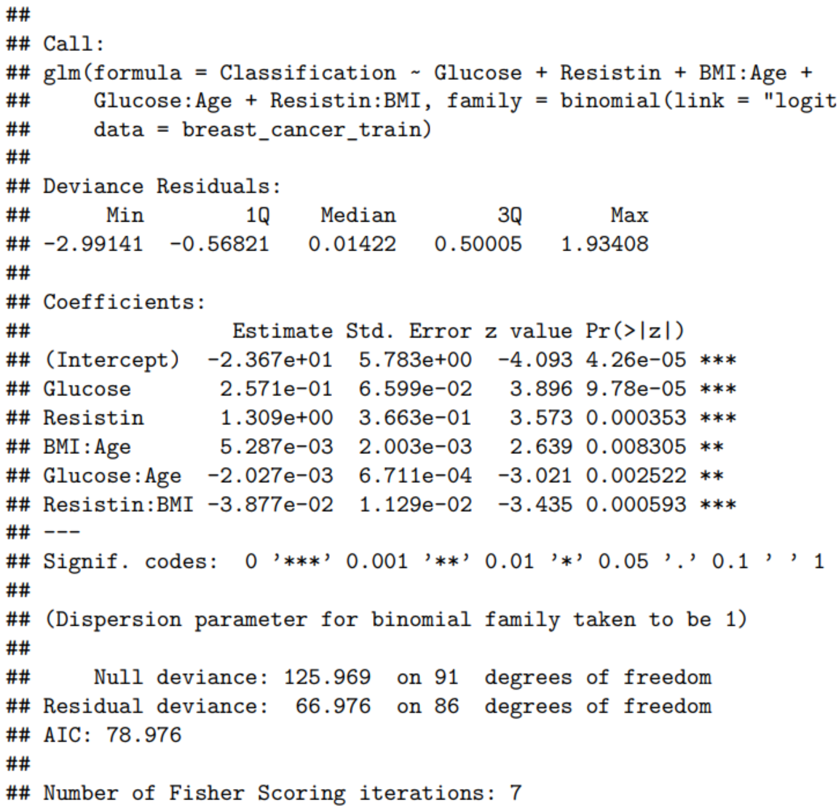
```
##
## Call:
## glm(formula = Classification ~ Age * Glucose + Age * BMI + Age *
##     Adiponectin + Age * Resistin + Age * MCP.1 + Adiponectin *
##     BMI + Adiponectin * Glucose + Resistin * BMI + Resistin *
##     Glucose, family = binomial(link = "logit"), data = train_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.80376  -0.28543   0.01229   0.39773   2.04256
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         3.2522389 21.4576742   0.152   0.8795
## Age                -0.3632648  0.3162572  -1.149   0.2507
## Glucose             0.3376885  0.1718030   1.966   0.0493 *
## BMI                -0.7410608  0.5012549  -1.478   0.1393
## Adiponectin        -2.0130825  0.9969315  -2.019   0.0435 *
## Resistin            1.6549456  0.8037469   2.059   0.0395 *
## MCP.1              -0.0080955  0.0066450  -1.218   0.2231
## Age:Glucose        -0.0037269  0.0019853  -1.877   0.0605 .
## Age:BMI             0.0152563  0.0085781   1.779   0.0753 .
## Age:Adiponectin     0.0112145  0.0088843   1.262   0.2068
## Age:Resistin        0.0018132  0.0047068   0.385   0.7001
## Age:MCP.1           0.0001455  0.0001147   1.268   0.2048
## BMI:Adiponectin     0.0184718  0.0193524   0.954   0.3398
## Glucose:Adiponectin  0.0098245  0.0100674   0.976   0.3291
## BMI:Resistin       -0.0430745  0.0226612  -1.901   0.0573 .
## Glucose:Resistin   -0.0034566  0.0018264  -1.893   0.0584 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 125.969  on 91  degrees of freedom
## Residual deviance:  49.364  on 76  degrees of freedom
## AIC: 81.364
##
## Number of Fisher Scoring iterations: 8
```

Output 2: Best Model without Interaction in terms of AIC



```
##
## Call:
## glm(formula = Classification ~ Age + BMI + Glucose + Resistin +
##     MCP.1, family = binomial(link = "logit"), data = breast_cancer_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0757  -0.8179   0.1864   0.8083   2.0627
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.464731   2.601089  -2.485 0.012941 *
## Age         -0.030051   0.017389  -1.728 0.083959 .
## BMI         -0.130851   0.057137  -2.290 0.022014 *
## Glucose      0.111113   0.028585   3.887 0.000101 ***
## Resistin     0.038288   0.026893   1.424 0.154537
## MCP.1        0.002092   0.001015   2.061 0.039329 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 125.969  on 91  degrees of freedom
## Residual deviance:  88.055  on 86  degrees of freedom
## AIC: 100.06
##
## Number of Fisher Scoring iterations: 6
```

Output 3: Best Model with Interaction using Variables Selected in the Output 2



```
##
## Call:
## glm(formula = Classification ~ Glucose + Resistin + BMI:Age +
##     Glucose:Age + Resistin:BMI, family = binomial(link = "logit"),
##     data = breast_cancer_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.99141  -0.56821   0.01422   0.50005   1.93408
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.367e+01  5.783e+00  -4.093 4.26e-05 ***
## Glucose       2.571e-01  6.599e-02   3.896 9.78e-05 ***
## Resistin      1.309e+00  3.663e-01   3.573 0.000353 ***
## BMI:Age       5.287e-03  2.003e-03   2.639 0.008305 **
## Glucose:Age  -2.027e-03  6.711e-04  -3.021 0.002522 **
## Resistin:BMI -3.877e-02  1.129e-02  -3.435 0.000593 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 125.969  on 91  degrees of freedom
## Residual deviance:  66.976  on 86  degrees of freedom
## AIC: 78.976
##
## Number of Fisher Scoring iterations: 7
```