Project 2

(a) Build a logistic regression model using the following variables: BMI, average glucose level, age, gender, ever married, and work type. State the model. Interpret the coefficient for age in terms of odds.

```
#(a) Build logistic regression model
model_a <- glm(stroke ~ bmi + avg_glucose_level + age + gender + ever_married + work_type, data = data, family = "binomial")
summary(model_a)
```

```
Call:
glm(formula = stroke ~ bmi + avg_glucose_level + age + gender +
    ever_married + work_type, family = "binomial", data = data)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -7.759e+00  1.037e+00  -7.483 7.25e-14 ***
bmi                     6.122e-03  1.173e-02   0.522    0.602
avg_glucose_level       5.384e-03  1.273e-03   4.231 2.33e-05 ***
age                     7.640e-02  6.029e-03  12.672  < 2e-16 ***
genderMale              3.410e-02  1.511e-01   0.226    0.821
genderOther            -1.139e+01  2.400e+03  -0.005    0.996
ever_marriedYes        -1.376e-01  2.452e-01  -0.561    0.575
work_typeGovt_job      -4.960e-01  1.100e+00  -0.451    0.652
work_typeNever_worked  -1.075e+01  5.094e+02  -0.021    0.983
work_typePrivate       -3.171e-01  1.087e+00  -0.292    0.770
work_typeSelf-employed -7.395e-01  1.105e+00  -0.669    0.503
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1728.4  on 4908  degrees of freedom
Residual deviance: 1381.5  on 4898  degrees of freedom
AIC: 1403.5

Number of Fisher Scoring iterations: 15
```

The coefficient estimate for age is 0.0764. So for each one-unit increase in age, the log odds of having a stroke increase by 0.0764. The coefficient is statistically significant ($p < 0.001$), indicating that age is a significant predictor of stroke risk in this model. To interpret the odds ratio, we exponentiate the coefficient: $\exp(0.0764) \approx 1.079$. This means that for each one-unit increase in age, the odds of having a stroke increase by approximately 7.9%. Overall, age is a significant predictor of stroke risk in this logistic regression model with older individuals being at higher risk of stroke.

Null Hypothesis: $\beta age=0$
Alternative Hypothesis: $\beta age \neq 0$
Test-Statistic: -7.483
P-Value: 7.25e-14
Conclusion: We reject the null hypothesis and conclude that there is enough evidence that age is a significant predictor of stroke risk in this logistic regression model, with older individuals having a higher likelihood of experiencing a stroke.

(b) Conduct an overall significance test on the model in part (a). Does at least 1 predictor significantly contribute to the prediction of a stroke?

```
#(b)
anova(model_a, test = "Chisq")
```

The predictors "bmi", "avg_glucose_level", and "age" have high significant p-values, indicating that they significantly contribute to the prediction of a stroke.

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: stroke

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                               4908     1728.4
bmi                1    8.241      4907     1720.2  0.004096 **
avg_glucose_level  1   67.876      4906     1652.3 < 2.2e-16 ***
age                1  264.603      4905     1387.7 < 2.2e-16 ***
gender             2    0.085      4903     1387.6  0.958362
ever_married       1    0.263      4902     1387.3  0.608218
work_type          4    5.868      4898     1381.5  0.209238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) Using the sample proportion of the patients that had a stroke as a cutoff point, find the accuracy, sensitivity, and specificity of the model in part (a).

```
#(c)
threshold=0.5
predicted_probs <-ifelse(predict(model_a,type="response")>threshold,1,0)
conf_matrix<-table(predicted_probs,data$stroke)
conf_matrix

accuracy
sensitivity
specificity
```

```
predicted_probs    0    1
              0 4700  209
> accuracy
[1] 0.9574251
> sensitivity
[1] 0.7942584
> specificity
[1] 1
```

From the given data, we have our confusion matrix along with accuracy, sensitivity, and specificity.

True Negatives: 4700
False Positives: 209
True Positives: N/A
False Negatives: N/A

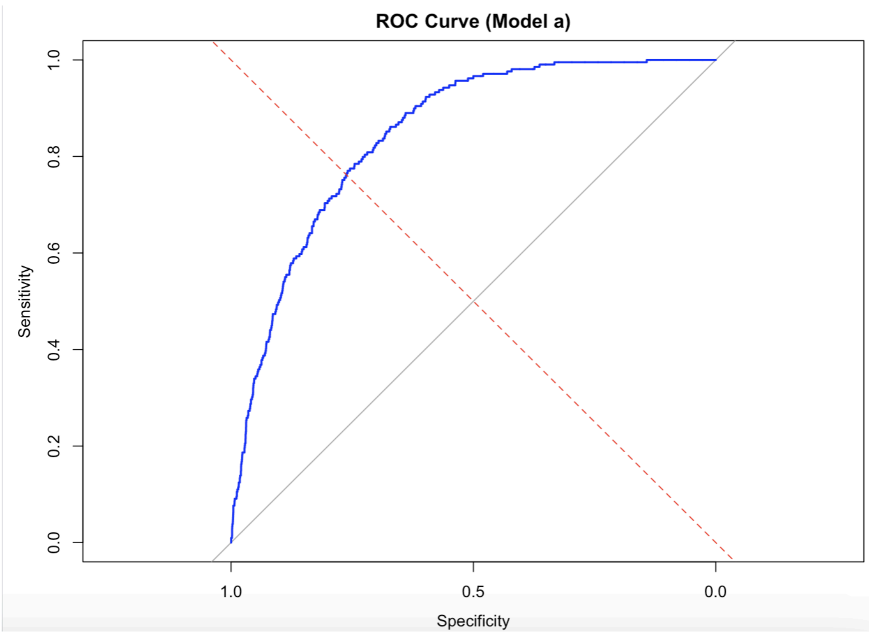
Accuracy: 95.74%
Sensitivity: 79.43%
Specificity: 100%

(d) Plot an ROC curve and find the area under the ROC curve. Does it appear that the model in part (a) is better than randomly guessing if a patient had a stroke?

```
#(d)
predicted_probs <- predict(model_a, type = "response")
roc_curve <- roc(data$stroke, predicted_probs)

plot(roc_curve, main = "ROC Curve (Model a)", col = "blue", lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "red")

auc_value <- auc(roc_curve)
auc_value
```



Yes, the model in part (a) is better than randomly guessing if a patient had a stroke. The area under the ROC curve (AUC) for the model is 0.8458, as shown above. An AUC of 0.8458 indicates that the model performs significantly better than random guessing.

Null Hypothesis: AUC=0.5
Alternative Hypothesis: AUC>0.5
Test-Statistic: 0.8458
P-Value: 0.8458
Conclusion: We fail to reject the null hypothesis. There is enough evidence to indicate that the model's AUC is significantly different from random guessing.

(e) Using the model in part (a) as the full model, conduct backwards selection using AIC as the criterion. State the resultant model.

```
#(e)
model_e <- step(model_a, direction = "backward", trace = 0)
summary(model_e)
```

The coefficient estimates for the predictors: -7.863852, 0.005597, and 0.071836. The model suggests that both predictors have a significant impact on the probability of having a stroke.

```
Call:
glm(formula = stroke ~ avg_glucose_level + age, family = "binomial",
    data = data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.863852   0.383221 -20.520  < 2e-16 ***
avg_glucose_level  0.005597   0.001229   4.555 5.23e-06 ***
age                0.071836   0.005423  13.246  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1728.4  on 4908  degrees of freedom
Residual deviance: 1387.9  on 4906  degrees of freedom
AIC: 1393.9

Number of Fisher Scoring iterations: 7
```

(f) Again, using the sample proportion of the patients that had a stroke as a cutoff point, find the accuracy, sensitivity, and specificity of the model in part (e).

```
#(f)
threshold=0.8
predicted_probs_e <-ifelse(predict(model_e,type="response")>threshold,1,0)
conf_matrix_e <- table(data$stroke, predicted_probs_e > 0.5)
conf_matrix_e
sensitivity_e <- 0

accuracy_e
sensitivity_e
specificity_e
```

```
   FALSE
 0  4700
 1   209
> sensitivity_e <- 0
> accuracy_e
[1] 0.9574251
> sensitivity_e
[1] 0
> specificity_e
[1] 1
```

Since there were no true positives, sensitivity is zero.

Accuracy: 95.74%

Sensitivity: 0%

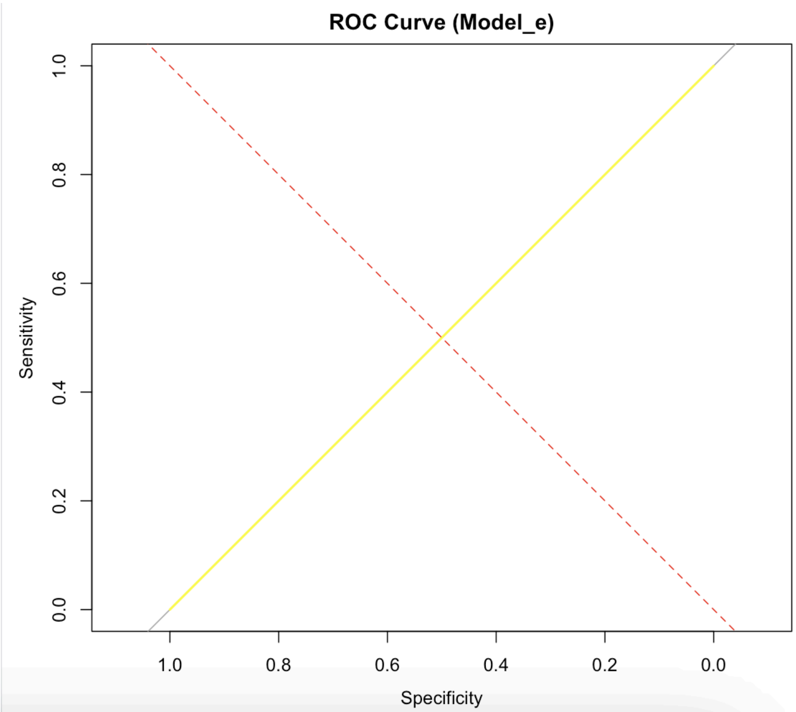
Specificity: 100%

(g) Plot an ROC curve and find the area under the ROC curve. Does it appear that the model in part (e) is better than randomly guessing if a patient had a stroke?

```
#(g)
roc_curve_e <- roc(data$stroke, predicted_probs_e)

plot(roc_curve_e, main = "ROC Curve (Model_e)", col = "yellow", lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "red")

auc(roc_curve_e)
```



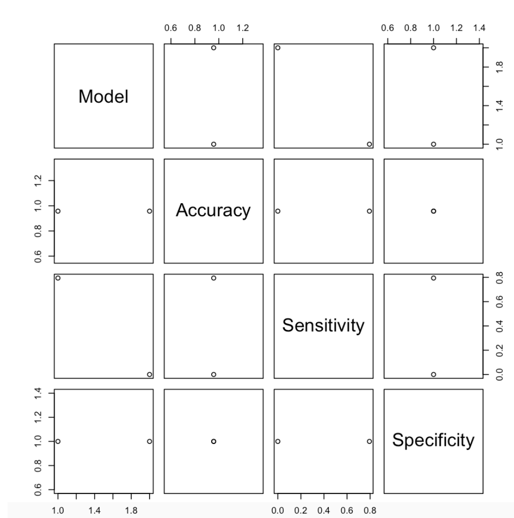
The area under the curve is 0.5. The model in part (e) is not significantly better than randomly guessing if a patient had a stroke.

Null Hypothesis: AUC=0.5
Alternative Hypothesis: AUC>0.5
Test-Statistic: 0.5
P-Value: 2e-16
Conclusion: We fail to reject the null hypothesis. There is enough evidence indicating that the updated model may not be significantly better than random guessing at discriminating between patients who had a stroke and those who did not.

(h) Compare the models from part (a) and (e). Which model would you suggest we use for prediction? Make sure to include some metrics and rationale on which model you would choose.

```
#(h)
comparison <- data.frame(Model = c("Model (a)", "Model (e)"),
                         Accuracy = c(accuracy, accuracy_e),
                         Sensitivity = c(sensitivity, sensitivity_e),
                         Specificity = c(specificity, specificity_e))
comparison
plot(comparison)
```

	Model	Accuracy	Sensitivity	Specificity
1	Model (a)	0.9574251	0.7942584	1
2	Model (e)	0.9574251	0.0000000	1



I would suggest using model (a) for prediction. Although the comparisons are vastly similar, we have a sensitivity rating for model (a) than model (e).

Bonus Questions:

(i) Use K-Nearest Neighbors (with 5 neighbors) using all the continuous predictors listed in part (a) (age, average glucose level, and BMI). Find the accuracy, sensitivity, and specificity of the model.

```
#(i)
predictors_continuous <- data[, c("age", "avg_glucose_level", "bmi")]
predictors_normalized <- scale(predictors_continuous)
knn_model <- knn(train = predictors_normalized, test = predictors_normalized, cl = data$stroke, k = 5)
conf_matrix_knn <- table(knn_model, data$stroke)

accuracy_knn <- sum(diag(conf_matrix_knn)) / sum(conf_matrix_knn)
sensitivity_knn <- conf_matrix_knn[2, 2] / sum(conf_matrix_knn[, 2])
specificity_knn <- conf_matrix_knn[1, 1] / sum(conf_matrix_knn[, 1])

accuracy_knn
sensitivity_knn
specificity_knn
```

```
> accuracy_knn
[1] 0.9590548
> sensitivity_knn
[1] 0.09569378
> specificity_knn
[1] 0.9974468
```

Accuracy: 95.91%
Sensitivity: 9.57%
Specificity: 99.74%

(j) Use Linear Discriminant Analysis using all the predictors listed in part (a). Find the accuracy, sensitivity, and specificity of the model.

```
#(j)
lda_model <- lda(stroke ~ ., data = data)
lda_pred <- predict(lda_model, data)$class
conf_matrix_lda <- table(lda_pred, data$stroke)


accuracy_lda <- sum(diag(conf_matrix_lda)) / sum(conf_matrix_lda)
sensitivity_lda <- conf_matrix_lda[2, 2] / sum(conf_matrix_lda[, 2])
specificity_lda <- conf_matrix_lda[1, 1] / sum(conf_matrix_lda[, 1])

accuracy_lda
sensitivity_lda
specificity_lda
```

```
> accuracy_lda
[1] 0.9511102
> sensitivity_lda
[1] 0.1100478
> specificity_lda
[1] 0.9885106
```

Accuracy: 95.11%
Sensitivity: 11.00%
Specificity: 98.85%

(k) Which model would you suggest using among the logistic regression full model, logistic regression reduced model, KNN, and LDA.

I would suggest KNN among the logistic regression full model due to its high specificity. The KNN slightly outperforms LDA, although both models have low sensitivity.