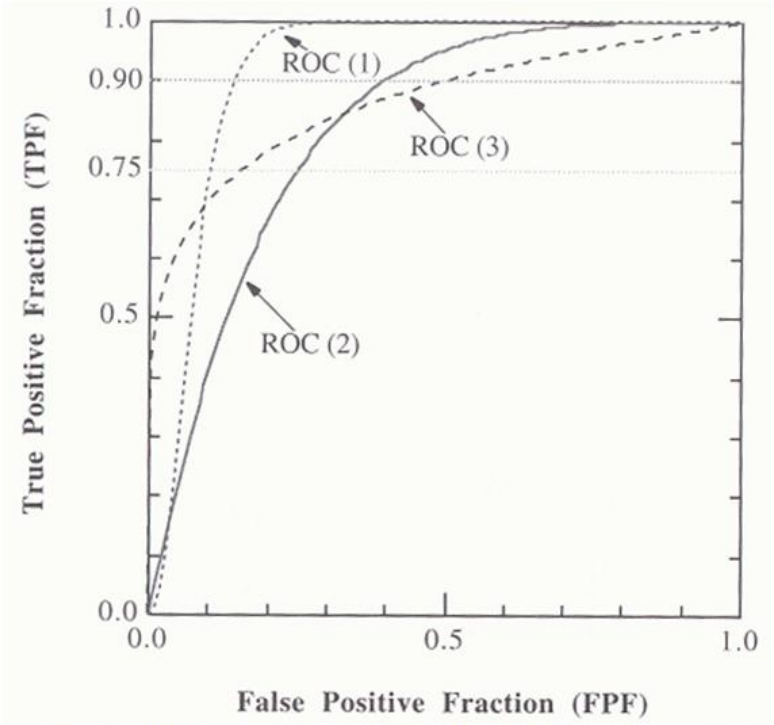
QUESTION 1



The ROC curves of three binary classification models on a balanced data are given above. For question 1-4, Check all the statements which are true.

1. The ROC curve with larger area under it performs generally better than the other curves.

- ○ True
- ○ False

QUESTION 2

2. For specificity less than 10%, model 3 performs better than model 1.

- ○ True
- ○ False

QUESTION 3

3. For lower thresholds, model 2 performs better than model 3.

- ○ True
- ○ False

QUESTION 4

4. For sensitivity higher than 85%, Model 2 performs better than model 3.

- ○ True
- ○ False

QUESTION 5

Which of the following methods dominates the others if decision boundary is non-linear and n is very large compared to p.

- ○ Naive Bayes
- ○ LDA
- ○ QDA
- ○ Logistic Regression
- ○ KNN

QUESTION 6

Which of the following methods is equivalent to QDA , if the covariance matrices are constant for each class?

- ○ Poisson Regression
- ○ Multiple Linear Regression
- ○ LDA
- ○ Naive Bayes
- ○ KNN

QUESTION 7

If we two dimensional normal distribution is used in naive Bayes classifier, the new method will be equivalent to QDA.

- ○ True
- ○ False

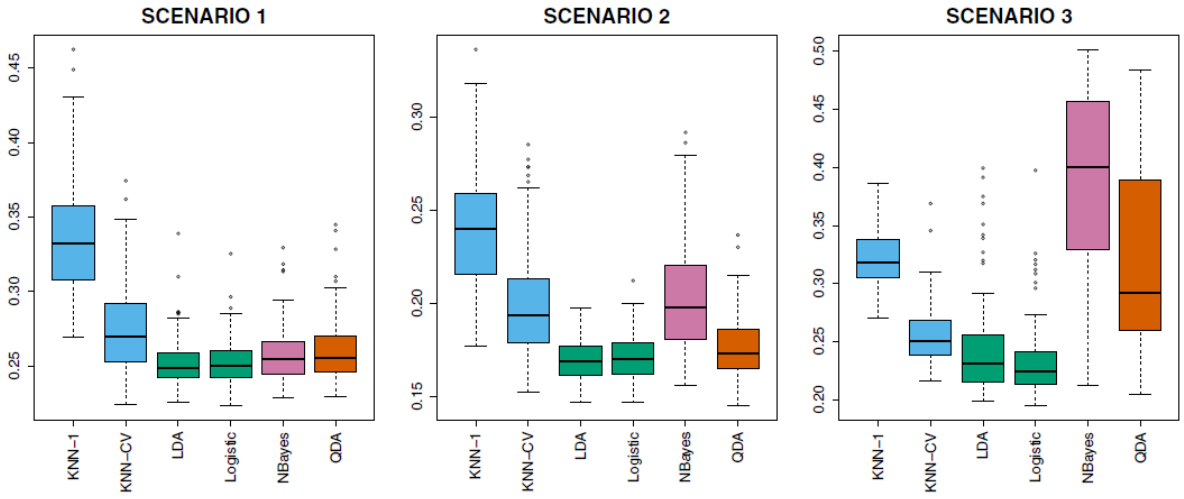
QUESTION 8

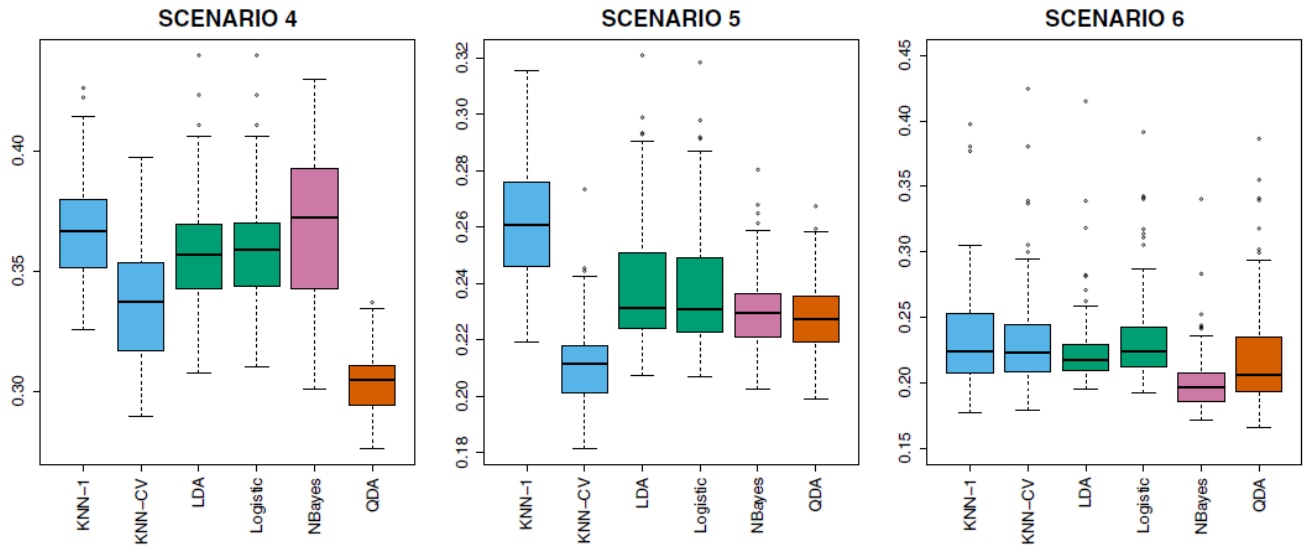
Which of the following method is NOT a good choice if we are interested to determine important predictors in a classification problem?

- ○ Logistic Regression
- ○ Naive Bayes
- ○ QDA
- ○ LDA
- ○ KNN

QUESTION 9

Match the following scenerios with the most appropriate graph.

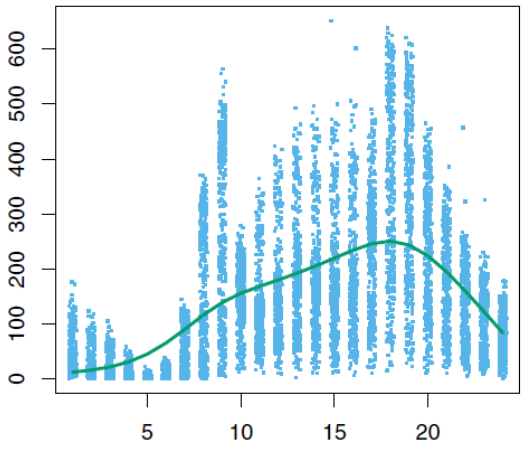




- [ ] The data were generated from a normal distribution, with a correlation of 0.5 between the predictors in the first class, and correlation of −0.5 between the predictors in the second class.
- [ ] There were 20 training observations in each of two classes. The observations within each class were random normal variables with a different mean in each class. The two predictors had a correlation of −0.5.
- [ ] There were 20 training observations in each of two classes. The observations within each class were uncorrelated random normal variables with a different mean in each class.
- [ ] The data were generated from a normal distribution with uncorrelated predictors. Then the responses were sampled from the logistic function applied to a complicated non-linear function of the predictors.
- [ ] There were 50 training observations in each of two classes. The observations within each class were generated from the t-distribution. The two predictors had a correlation of −0.5.
- [ ] The observations were generated from a normal distribution with a different diagonal covariance matrix for each class. The sample size was n = 6 in each class.

A. Scenario 2
B. Scenario 3
C. Scenario 4
D. Scenario 1
E. Scenario 6
F. Scenario 5

QUESTION 10



The scatter plot of 24 predictors vs the response (Y) is given above. We would like to fit a linear model using this data. Using the above graph, which of the following statements is true?

- ○ The linear model will suffer from outliers.
- ○ The major violation of the linear model is heteroscedasticity.
- ○ The linear model will suffer from high leverage points.
- ○ Transfroming the response function using Y^2 will give better results in linear model.
- ○ The model suffer from highly correlated variables.

QUESTION 11

	Coefficient	Std. error	z-statistic	p-value
Intercept	4.12	0.01	683.96	0.00
workingday	0.01	0.00	7.5	0.00
temp	0.79	0.01	68.43	0.00
weathersit[cloudy/misty]	-0.08	0.00	-34.53	0.00
weathersit[light rain/snow]	-0.58	0.00	-141.91	0.00
weathersit[heavy rain/snow]	-0.93	0.17	-5.55	0.00

The output of a poisson regression model of bikeshare data is given above. How will the number of bike rentals will be affected if the weather changes from cloudy/misty to light rain/snow.

- ○ 36% as many people will use bikes when it is light rain/snow relative to when it is cloudy/misty.
- ○ 40% as many people will use bikes when it is light rain/snow relative to when it is cloudy/misty.
- ○ 50% as many people will use bikes when it is light rain/snow relative to when it is cloudy/misty.
- ○ 60% as many people will use bikes when it is light rain/snow relative to when it is cloudy/misty.
- ○ 45% as many people will use bikes when it is light rain/snow relative to when it is cloudy/misty.

QUESTION 12

Match the generalized linear model with its link function.

- [-] Linear regression
- [-] Logistic regression
- [-] Poisson regression

A. $\eta(\mu) = \mu$
B. $\eta(\mu) = \log(\mu)$
C. $\eta(\mu) = \log(\mu / 1 - \mu)$

QUESTION 13

The stock market data is trained using logistic regression model. Given the following codes and their output. What is the predicted class of the second observation?

```
glm.probs <- predict(glm.fits, type = "response")
glm.probs[1:3]
```

1: 0.507084133395402 2: 0.481467878454591 3: 0.481138835214201

```
[ ] contrasts(Direction)
```

A matrix: 2 × 1 of type dbl

	Up
Down	0
Up	1

- ○ Up
- ○ Down

QUESTION 14

Given the following code and output. Calculate accuracy, specifity and sensitivity.

```
table(glm.pred, Direction)
```

```
        Direction
glm.pred Down  Up
    Down  145 141
    Up    457 507
```

- ○ Accuracy = 52.16%
  Sensitivity = 21.8%
  Specificity = 76%
- ○ Accuracy = 47.84%
  Sensitivity = 21.8%
  Specificity = 24%
- ○ Accuracy = 47.84%
  Sensitivity = 49.3%
  Specificity = 52.6%
- ○ Accuracy = 52.16%
  Sensitivity = 50.7%
  Specificity = 47.4%
- ○ Accuracy = 52.16%
  Sensitivity = 78.2%
  Specificity = 24%

QUESTION 15

We apply the KNN approach to the Insurance data set. This data set includes 85 predictors that measure demographic characteristics for 5,822 individuals. The response variable is Purchase, which indicates whether or not a given individual purchases a caravan insurance policy. In this data set, only 6,% of people purchased caravan insurance. The company would like to try to sell insurance only to customers who are likely to buy it.

Given the information, which of the following is the best result?

1.

```
         test.Y
knn.pred  No Yes
      No  873  50
      Yes  68   9
```

2.

```
         test.Y
knn.pred  No Yes
      No  920  54
      Yes  21   5
```

3.

```
         test.Y
glm.pred  No Yes
      No  919  48
      Yes  22  11
```

4.

```
         test.Y
knn.pred  No Yes
      No  930  55
      Yes  11   4
```

- ○ 4.
- ○ 1.
- ○ 3.
- ○ 2.
- ○ All of the results are worse than random guessing.