Data Science for Engineers

Week 7 assignment

1. Which among the following is not a type of cross-validation technique?

   (a) LOOCV

   (b) k-fold croos validation

   (c) Validation set approach

   (d) Bias variance trade off

   Answer: (d)

2. Which among the following is a classification problem?

   (a) Predicting the average rainfall in a given month.

   (b) Predicting whether a patient is diagnosed with a disease or not.

   (c) Predicting the price of a house.

   (d) Predicting whether it will rain or not tomorrow.

   Answer: (b, d)

3. Consider the following confusion matrix for the classication of Hatchback and SUV:

		True	
		Hatchback	SUV
Prediction	Hatchback	55	5
	SUV	0	40

   (i) Find the accuracy of the model.

   (a) 0.95

   (b) 0.55

   (c) 0.45

   (d) 0.88

   Answer: (a)

   (ii) Find the sensitivity of the model.

   (a) 0.95

(b) 0.55
(c) 1
(d) 0.88

Answer: (c)

4. Under the 'family' parameter of glm() function, which one of the following distributions correspond to logistic regression for a variable with binary output?

   (a) Binomial
   (b) Gaussian
   (c) Gamma
   (d) Poisson

   Answer: (a)

   Use the following information to answer Q6, Q7, Q8, Q9, and Q10:

   Load the dataset iris.csv (add the link sent in the email) as a dataframe irisdata, with the first column as index headers, first row as column headers, dependent variable as factor variable, and answer the following questions.

   The iris dataset contains four Sepal and Petal features (Sepal Length, Sepal Width, Petal Length, Petal Width, all in cm) of 50 equal samples of 3 different species of the iris flower (Setosa, Versicolor, and Virginica).

5. What is the dimension of the dataframe?

   (a) (150, 5)
   (b) (150, 4)
   (c) (50, 5)
   (d) None of the above

   Answer: (a)

6. What can you comment on the distribution of the independent variables in the dataframe?

   (a) The variables Sepal Length and Sepal Width are not normally distributed
   (b) All the variables are normally distributed
   (c) The variable Petal Length alone is normally distributed
   (d) None of the above

   Answer: (b)

7. How many rows in the dataset contain missing values?

   (a) 10

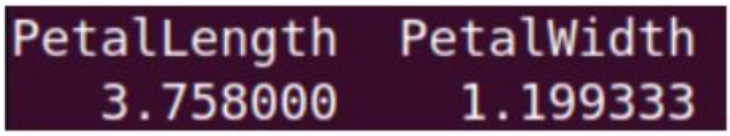
(b) 5

(c) 25

(d) 0

Answer: (d)

8. Which of the following code blocks can be used to summarize the data (finding the mean of the columns PetalLength and PetalWidth), similar to the one given below.



(a) lapply(irisdata[, 3:4], mean)

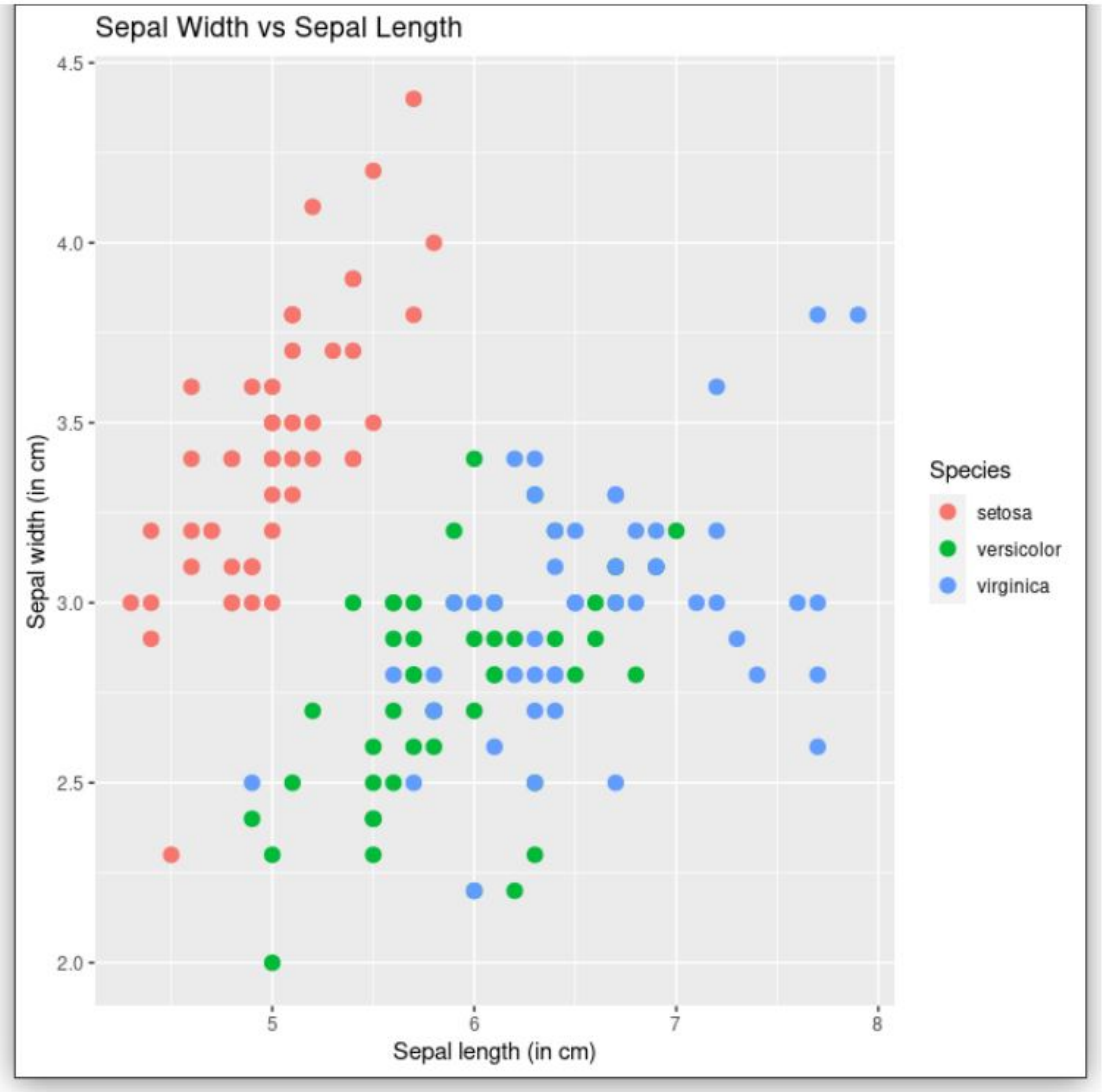
(b) sapply(irisdata[, 3:4], 2, mean)

(c) apply(irisdata[, 3:4], 2, mean)

(d) apply(irisdata[, 3:4], 1, mean)

Answer: (a, c)

9. What can be interpreted from the plot shown below?



(a) Sepal widths of Versicolor flowers are lesser than 3 cm.

(b) Sepal lengths of Setosa flowers are lesser than 6 cm.

(c) Sepal lengths of Virginica flowers are greater than 6 cm.

(d) Sepals of Setosa flowers are relatively more wider than Versicolor flowers.

Answer: (b, d)