Question Paper

Exam Date & Time: 08-Jan-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

VII SEMESTER B.TECH MAKEUP EXAMINATIONS, JAN 2024

ADVANCED DATA SCIENCE - PART III [CRA 4062]

Marks: 50 **Duration: 180 mins.**

Descriptive

Answer all the questions. Section Duration: 180 mins

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1) A) Develop a customer churn prediction model for a telecommunications company using the caret package in R. The dataset, named *telecom_churn_data.csv*, contains various features related to customer behavior, services used, and account information. The target variable, Churn, is binary (1 for churn, 0 for no churn). (5)

B) Considering two different models, one exhibiting lower in-sample error and the other displaying lower out-of-sample error, which model would be deemed more robust, and what factors contribute to this determination? (3)

C) Compare two regression models. Model A has an RMSE of 5, and Model B has an RMSE of 8. Which model is performing better, and why? (2)

2) A) Illustrate the use of linear discriminant analysis technique for the output depicted below in Figure Q4 using suitable R commands. Each line in the figure splits the data into two groups 1 vs 2, 2 vs 3, 1 vs 3 and each side of the line represents a region where the probability of one group (1, 2, or 3) is the highest. (5)

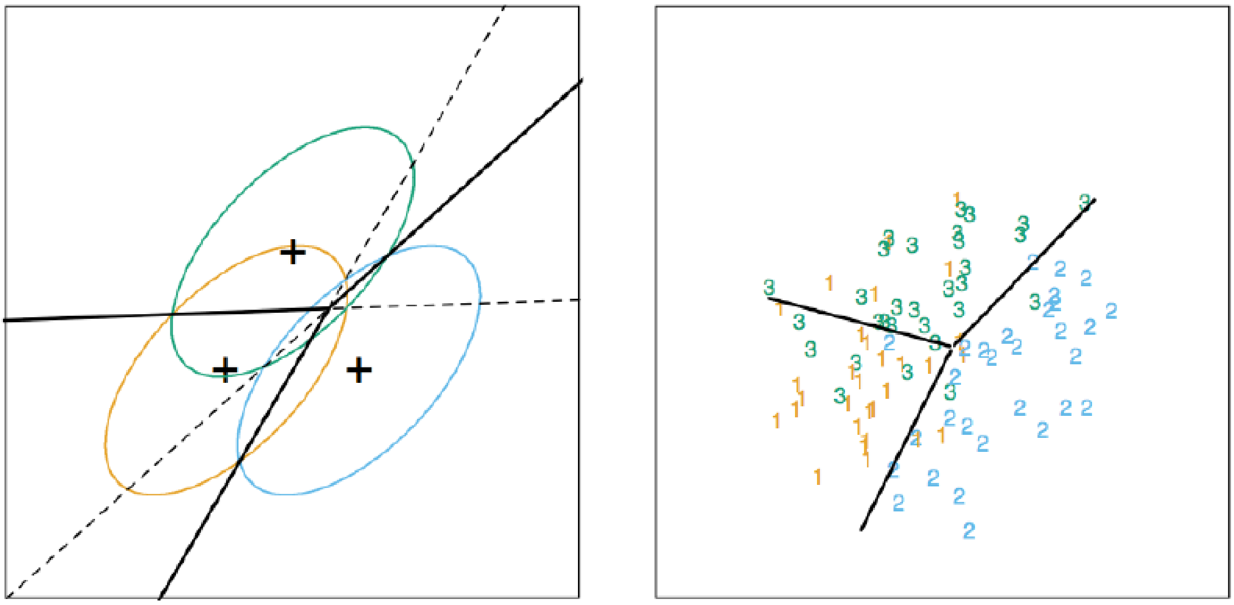

Figure Q4: Sample output

B) Compare and contrast between Linear Discriminant Analysis and Naïve Bayes. (3)

C) Consider the predictions given in the below table. Suppose there are 5 independent classifiers/models and each has 70% accuracy. Find the majority voting accuracy. (2)

Table Q6: Prediction Outcome

	Correct prediction	Wrong prediction
Classifier 1	3	2
Classifier 2	1	4
Classifier 3	4	1
Classifier 4	0	5
Classifier 5	5	0

3) Discuss the process of building a random forest classifier. Examine the given R code for constructing a random forest. Is there an error in the code? Provide a justification for your answer. (5)

A)

```
data(iris)

training < - iris[inTrain,]

testing < - iris[-inTrain,]

modFit < - train(Species~ .,data=training,method="rf",prox=TRUE)

head(getTree(modFit$finalModel,k=2))
```

B) Consider a dataset given in Table Q8, Calculate the gini index for the entire dataset and for the attribute "Color". (3)

Table Q8: Sample Dataset

Example	Color	Target (Class)
1	Red	Good
2	Red	Bad
3	Blue	Good
4	Red	Good

C) Consider the dataset "Ozone" which has four attributes ozone (label), radiation, temperature and wind. Write a R program to build a model using a bagging algorithm. (2)

4) Using the plotly library in R, create and display an interactive 3-D scatter plot to visualize the relationship between temperature, pressure, and time. Provide a suitable code snippet. (5)

A)

B) Discuss the functionality of the "swirl" software package for the R, show its application in the creation of courses and addition of new lessons. Name the output files that are produced while generating the new course. (3)

C) Discuss the key advantages of making R packages public. (2)

5) With the help of R Markdown code, create a presentation titled "My Fancy Presentation." Apply your understanding of the field of data science to highlight the key reasons of high significance of R Markdown. (5)

A)

B) Assume you have a data frame with latitude and longitude, show how all the data points of the data frame can be added to a map at once using a leaflet. (3)

C) Summarize the steps involved in developing an R package. (2)

-----End-----