Question Paper

Exam Date & Time: 02-Dec-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

Answer all the questions
Data missing if any, may be suitably assumed

ADVANCED DATA SCIENCE - PART III [CRA 4062]

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

Descriptive

Answer all the questions.

1A)	Using R, apply the *caret* package to analyze the "iris.csv" dataset comprehensively. Begin by loading the dataset and addressing any potential missing values, presenting summary statistics and the dataset's structure. Subsequently, split the data into training (70%) and testing (30%) sets. Utilize the k-Nearest Neighbors (kNN) algorithm for classification with 5-fold cross-validation during training. Evaluate the resulting model on the testing set, reporting crucial classification metrics such as accuracy, precision, and recall.	(5)
1B)	Write R code to check for overfitting in a model. Consider a scenario where the in-sample error is significantly lower than the out-of-sample error.	(3)
1C)	Consider a medical diagnostic scenario where a new test is developed to detect a rare disease. The test has a sensitivity of 90% and a specificity of 95%. The prevalence of the disease in the population is 0.1%. Calculate the Positive Predictive Value (PPV) and Negative Predictive Value (NPV) of the test. Assume a population of 1,000 individuals.	(2)
2A)	Demonstrate the use of model-based prediction approach and the models that leverage the same approach with suitable R commands.	(5)
2B)	Developer X intends to retrieve stock prices from the website www.stocks.com. Below are the corresponding use cases; provide the equivalent R code and parameter decription for each: A. Fetch the daily high, low, open, and close prices. B. Subset the time series based on user-specified start and end points. C. Run exponential smoothing model on training data	(3)
2C)	For the given data below in Figure Q6, check if the following statement is true, using the Naïve Bayes rule. "Players will play if the weather is sunny"	(2)

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Figure Q6: Sample data

3A) Write the process to build a classfication tree.Write a R program for constructing a classification tree using the IRIS dataset, comprising four attributes: sepal length, petal length, sepal width, and petal width, along with a class label denoted as Type. (5)

3B) Identify which among the methods (Bagging/Boosting/Ensemble) may be applied for the following scenarios given. Justify your answer. (3)

A. In a telecommunication company, the goal is to predict customer churn (whether a customer will leave the service) based on various features such as call duration, customer satisfaction ratings, and usage patterns. The dataset is large and diverse, with potentially noisy data.

B. In a financial system, there's a need to identify potentially fraudulent transactions among a vast number of transactions. The dataset is imbalanced, with only a small fraction of transactions being fraudulent.

3C) Consider the iris dataset comprising four attributes: sepal length, petal length, sepal width, and petal width, along with a class label denoted as Species. Provided the code for k-means clustering, identify any error/s present and justify for why it/they are erroneous. (2)

```
data(iris)

inTrain < - createDataPartition(y=iris$Species,p=0.7, list=FALSE)

training < - iris[inTrain,]; testing < - iris[-inTrain,]

kMeans1 < - kmeans(subset(training, centers=3)

training$clusters < - as.factor(kMeans1$cluster)

p1 < - qplot(Petal.Width,Petal.Length,colour=clusters,data=training) +

ggtitle("Clusters Classification")

p2 < - qplot(Petal.Width,Petal.Length,colour=Species,data=training) +

ggtitle("Species Classification (Truth)")

grid.arrange(p1, p2, ncol = 2)
```

4A) Develop the following using googleVis package by using the given data frame given in Table (5)

Table Q10: Sample Data

Serial No	Country	population	GDP	Year
1	US	331	21	2020

2	INDIA	1366	2	2020
3	FRANCE	67	2	2020
4	US	327	20	2018
5	INDIA	1345	2	2018
6	FRANCE	66	2	2018

A. A geographical chart that visualizes a dataset containing countries and corresponding population (in million). Use the data frame with fields country and value.

B. Demonstrate the creation of a motion chart in R to show the variation of GDP over the specified years, with suitable R code snippet.

4B) Demonstrate how shinyUI() is utilized for defining the user interface, and shinyServer() is employed to articulate the server logic. Utilize code snippets to illustrate these functions and their functionalities (3)

4C) Construct a map and insert markers into the map using a leaflet. (2)

5A) Create a presentation titled "Data Science Course" with the help of R Markdown code, which includes the following elements, (5)

(a)Title, a top-level subheading, and a second-level subheading

(b) List and ordered list

(c) Italicised text and bold text.

5B) Suggest a suitable method for creating dynamic and visually engaging maps in R. With appropriate code snippets create a basic map centered around the coordinates (37.0902, -95.7129). (3)

5C) With the help of a sample R code snippet, show how an independent user interface component can be created within a Shiny application. (2)

-----End-----