Statistical Modelling in R

Time: 2 hours

Total Marks: 60 marks

Note:

1) The candidate has option to either attempt question 3A or question 3B. Rest all questions are mandatory.
2) Numbers in the right indicate full marks.
3) The candidate will be provided with the formula sheet and graph papers [if required] for the examination.
4) Use of approved scientific calculator is allowed.

Q1A Install the package "nycflights13" and use "flights" dataset.

I. Write a line of code that gives the first few rows of flights that happened on November 3rd and carrier was 'American Airlines (AA)' [1]

II. Write a line of code that to return the flight data ordered by year [1]

III. Write a line of code to return fields – month, carrier and air time from flights data from the month of September. [1]

IV. Add a new field to the data titled 'new_col' with values as difference between arrival delay and departure delay. [2]

Q1B Use the "Default" dataset from the ISLR package.

I. Load the dataset in the "data" variable and provide its summary. [1]

II. Find total number of rows in the dataset [1]

III. Fit a logistic regression model for default with respect to student, balance, income [1]

IV. Provide a summary for the above model [1]

V. Fit a logistic regression model for default with respect to student, balance [1]
Comment on the changes in AIC values

Q1C In a clinical experiment, 7 patients were treated for cancer. They were treated with therapy A and therapy B. Their response to the therapy was observed. Given below are the results for the same.

DATA: CANCER DATA

PATIENT NO	THERAPY	RESPONSE
1	A	POSITIVE
2	B	NEGATIVE
3	B	POSITIVE
4	A	NEGATIVE
5	B	NEGATIVE
6	A	POSITIVE
7	A	NEGATIVE

I. Create a csv file of CANCER DATA and import into R [1]

II. Create frequency table for THERAPY and RESPONSE [2]

III. Obtain pie-chart for the variable RESPONSE [2]

Q2 A The insurance dataset is given by the Organisation, based on various factors you are asked to perform a Linear Regression.

I. Import Libraries ggplot2 & read the dataset insurance.csv. Display first rows. [1]
II. Provide summary of the dataset [1]
III. Create a linear model of Charges ~ Age, BMI, Smoker [1]
IV. Create a linear model of Charges ~ Age, Sex, BMI, Children, Smoker, Region [1]
V. Compare the R^2 value of both the models and comment on the accuracy of the model [1]

Q2 B Below mentioned are students' dataset of a University They represent the scores of 10 students and the additional marks given to the students belonging to the sports students.

DATA: Students

No	Name	Marks
1	Anmesh	576
2	Suresh	525
3	Akshit	540
4	Mayank	578
5	Kanchan	558
6	Akash	542
7	Naitri	521
8	Keval	560
9	Roshni	577
10	Dishant	525

DATA: Marks_Add

No	Additional_Marks
3	21
5	24
8	12
9	11
1	17

Perform the following in R and display each output:

I. Create CSV files of two datasets and display the outputs. [1]
II. Sort the dataset Marks_Add as per No [1]
III. Merge both the datasets according to No and name this data as Stu_M, Remove the NA values from Stu_M [1]
IV. Calculate TM as Marks + Additional_Marks [1]
V. Summarize TM using Min, Max, Mean and SD [1]

Q2 C Perform the following in R:

I. Load libraries: datasets, caTools, party, dplyr, magrittr & load data “readingSkills” [1]
II. Display the data “readingSkills” & create a train – test split with a ratio 0.8 [2]
III. Run a decision tree model using ctree (nativeSpeaker ~ shoesize+age+score) and plot the model [2]

Q3 A

Import the diamonds dataset from the built-in dataset in R. Perform the following:

I. Read the data and and view the dataset. [1]
II. View the structure of the diamonds dataset. [20]
    i. View top 6 observations of the dataset [2]
    ii. Provide the summary of variables of diamonds dataset [2]
    iii. Provide the dimensions of the diamonds dataset [2]
    iv. Plot a histogram of diamond price [2]
    v. Provide mean of price of the diamonds dataset [2]
    vi. View a scatter plot between carat and price of diamonds [2]
    vii. Find out the price per carat of diamonds across different colours of diamonds using boxplot (Hint: Use in y price/carat in the plot) [4]
    viii. Create a histogram of carat in diamonds, use fill = color [4]
III. Run a linear model over Price as a target variable and other columns as predictor variables. [4]
    i. Mention the R2 Value [2]
    ii. Mention the coefficients [2]

Q3 B

Load iris data. Perform the following:

I. Import library party in the interface. Load the data and store it in the data variable. [1]
II. Find:
    i. Maximum value of Sepal.Length, Sepal.Width, Petal.Length, Petal.Width [2]
    ii. Mean value of Sepal.Length, Sepal.Width, Petal.Length, Petal.Width [2]
    iii. Minimum value Sepal.Length, Sepal.Width, Petal.Length, Petal.Width [2]
III. Display first 6 rows of the data [1]
IV. Check if there are any NA values [1]
V. Plot histograms: [6]
    i. Sepal.Length
    ii. Sepal.Width
    iii. Petal.Length
    iv. Petal.Width
VI. Create a decision tree model (rpart) for Species with respect Sepal.Length, Sepal.Width, Petal.Length, Petal.Width. [7]

Name the model: iris_tree

VII. Create a decision tree model this time, use method = “anova” in the model. Name the mode: iris_tree_anova [8]