PGDM (IB), 2020-22

Predictive Business Analytics

IB-341

Trimester-III, End-Term Examination: April 2021

Time allowed: 2 Hrs 30 Min
Max Marks: 50

Instruction: Students are required to write Roll No on every page of the Answer Sheet. All other instructions on the question paper / notifications should be followed meticulously.

```
Submission:
 Section A: you need to submit a Word document containing your answers
       Section B: you need to submit the code file as well as the output file
       Please zip all your output files with the name of the file as 20IB399 (if your
 roll number is 20IB399)

For Section B, download the Happiness dataset for 2019 from the url:
https://raw.githubusercontent.com/amarnathbose/Datafiles/master/happiness_2019.csv

The field **score** denotes the happiness score for each country.
The remaining fields are
   - country
   - gdp:          per capita GDP
   - socsupport:   social support
   - hlifeexp:     healthy life expectancy
   - freechoice:   freedom of choice
   - generosity:   generosity
   - corruption:   corruption
```

Section	Instructions
A	**Total marks for this section is 15** Each question is of 5 marks; you may choose to answer any 3 questions in this section
B	**Total marks for this section is 45;** *however you can score a maximum of 35 only; i.e. any score above 35 will be taken as 35* *B.1 5 marks* *B.2 15 marks* *B.3 15 marks* *B.4 10 marks*

Section A

1. What is the importance of a training set in *supervised learning*? (CILO 1)
2. In kNN classification, why should we normalize the predictor variables? (CILO 1)
3. Explain with an example what is a *false negative* and a *false positive* (CILO 1)
4. How would you find the optimal value of *k* in kNN classification? (CILO 1)
5. Is kNN classification an example of supervised or unsupervised learning? Explain. (CILO 1)

Section B

1. Classify the countries of the dataset into 3 categories based on their happiness score, (CILO 2)
   Low (L): $<$4.5, Medium (M): 4.5-6.5, and High (H): $>$6.5.
   Create a new field called **happinessCateg** with values L, M and H.
2. Create a regression model for happiness score (score) on *gdp*, *socsupport*, *freechoice*, *generosity* and *corruption*. (CILO 2)
   Comment on the regression model. Create a second model retaining the predictors that are significant (p-value < 0.05). Which model is better and why?
3. Create a single linear regression model for the happiness score with the predictors *gdp*, *socsupport*, *hlifeexp* and *freechoice*. (CILO 2)
   Comment on the regression model.
   - Create a second model using the categorical variable **happinessCateg** as an additional predictor. Comment on this model.
   - Using the dummy variables from the categorical predictor, write down the 3 different model equations for the 3 categories, L, M, H of countries. Does the addition of the categorical variable improve the model significantly? [Use **anova** to find the answer]
4. Plot on a single graph, the happiness score (**y-axis**) on the social support (**x-axis**) for the three categories of countries with the colors, (CILO 2)
   L: red, M: gray, and H: green. Superimpose the three lines of best fit for the three happiness categories. Comment.

   [*Hint for Q4*]

- plot(y, x, main="plot caption", xlab="x-axis label", ylab="y-axis label", col=ifelse(r$happinessCateg=='L', 'red', ifelse(r$happinessCateg=='M', 'gray', 'green'))
- Create three separate regression models, e.g.
  lL <- lm(score ~ socsupport, data=r[r$happinessCateg=='L',])
- Superimpose the three lines of best fit by using commands like
  abline(lL, col="red")