PGDM/PGDM(IB), 2018-20
Predictive Business Analytics
Subject Code- DM-341/IB-316
Trimester-III, End-Tem Examination: March 2019

Time: 2 hrs. 30 mins.
Max Marks: 50

Section B (Computational)

You may answer 2 of the 3 questions in this section.

B1. **[17.5 marks]**

(a) Split the dataset into training and test in the ratio 50:50. Ensure that each dataset contains a representative proportion of *Malignant* and *Benign* sample records - i.e. proportion close to that in the full dataset. Tabulate the mean values of *concave points*, *symmetry* and *fractal dimension* for each of the two sets.

(b) How does the *mean* radius (field 3), texture (field 4) and symmetry (field 11) vary between the malignant and *benign* samples? Perform three *t tests*[1] to find out if the radius, the texture and the symmetry are significantly different for malignant and benign samples.

(c) Use the *t tests* for the fields $3, 4 \cdots 7$ to identify 3 the *top predictors* and use Decision Tree classification using the selected variables. What is the accuracy, sensitivity and specificity of the classification (for the test data)?

B2. **[17.5 marks]**

(a) Split the dataset into training and test in the ratio 50:50. Ensure that each dataset contains a representative proportion of *Malignant* and *Benign* sample records - i.e. proportion close to that in the full dataset. Tabulate the mean values of *concave points*, *symmetry* and *fractal dimension* for each of the two sets.

(b) Train a Naive Bayes classifier and use it to classify your test dataset records based on fields $3, 4 \cdots 7$.

(c) Train a Neural Network classifier and use it to classify your test dataset records based on fields $3, 4 \cdots 7$. Use the same training-test data split as in part (a) of this question.

(d) Tabulate and comment on the comparative accuracy, sensitivity and specificity of the two models.

B3. **[17.5 marks]**

(a) Split the dataset into training and test in the ratio 25:75. Ensure that each dataset contains a representative proportion of *Malignant* and *Benign* sample records.

(b) Perform a k-NN classification for the test dataset (fields $3, 4 \cdots 12$ as predictors), for values of k from 3 to 6, and tabulate the accuracy, sensitivity and specificity

(c) Examine the impact on the model accuracy, sensitivity and specificity if we reduce the predictors in the model to fields $3, 4 \cdots 7$ only.

[1] For performing t-test in R, you need two vectors x and y of values for the two categories. The sizes of the two vectors need not be the same. The R command is t.test(x,y)