

**JK Lakshmipat University Jaipur**  
**INSTITUTE OF ENGINEERING AND TECHNOLOGY**

End Term Examination (Regular), May 2019

B. Tech. in Computer Science and Engineering Semester VI (IBM-BDA only)

Roll No.....

CSEBD601: PREDICTIVE ANALYTICS MODELER

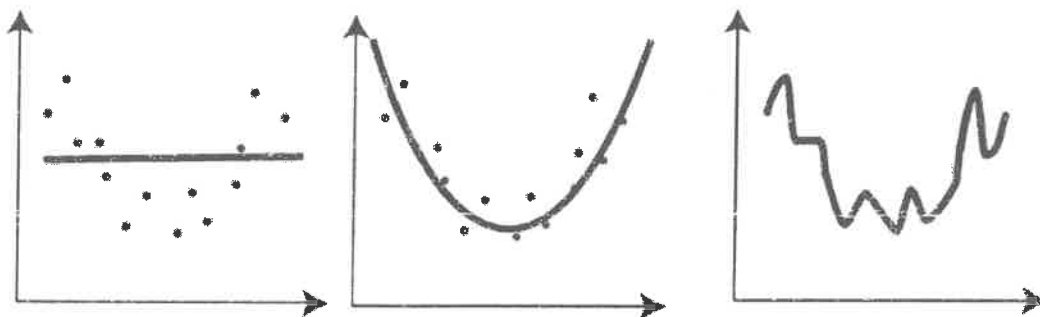
Time: 02 hours

Max. Marks: 50

**Instructions to students:**

1. Do not write anything other than your roll number on question paper.
2. Assume suitable data wherever required and mention it clearly.
3. Writing appropriate units, nomenclature, and drawing neat sketches/schematics/ flow chart wherever, required are an integral part of the answer.

Q.1A If we were evaluating the model performance with the help below visualizations which we have plotted for three different models for same regression problem on same training data. What do you conclude after seeing this visualization? 05



Q.1B Some patient features are expensive to collect (e.g., brain scans) whereas others are not (e.g., temperature). Therefore, we have decided to first ask our classification algorithm to predict whether a patient has a disease, and if the classifier is 80% confident that the patient has a disease, then we will do additional examinations to collect additional patient features. In this case, which classification methods do you recommend: neural networks, decision tree, or naive Bayes? Justify your answer. 05

Q.2A Consider the following 5 scenarios and note which are multivariate studies? Describe the most appropriate analytical method(s) to be used in each case and what statistical relationships/effects the analyses will test. 10

(i) A clinical psychologist interested in panic disorders wanted to determine the effectiveness of two different forms of therapy: relaxation and cognitive. Participants were assigned to four possible treatment groups (relaxation only, cognitive only, both relaxation and cognitive, and control with no therapy) gave ratings on three subscales of a panic disorder questionnaire tapping three different theoretical components of panic disorder.

(ii) A treatment and control group are compared on reading ability (assessed by two moderately correlated indices: a reading speed measure and a reading

comprehension test score).

(iii) An investigator is attempting to predict success in a sales job jointly from high school exam performance and a battery of personality variables given after recruitment.

(iv) Samples of lower and middle class school children (from the same schools) were repeatedly assessed in years 6, 7, and 8 on differing scholastic abilities (tests of reading comprehension, maths and science). The investigator wished to determine whether there are social class differences on these abilities and whether the differences change over time.

(v) In the same study as (iv) the researchers wanted to assess social class influences on "school citizenship" over the same 3 timepoints, using 3 citizenship indices (numbers of detentions received during the year; tutor ratings of participation in extra-curricular activities, merit points for homework completion).

Describe and discuss (with hypothetical examples) the two key advantages of multivariate analysis of variance over univariate analysis of variance.

Q.2B Can you name a possible method of improving the accuracy of a linear regression model? after evaluation of model, it is identified that we have high bias in our model. What could be the possible ways to reduce it? 05

Q.3A How can we use clustering method in supervised machine learning challenge? 05

Q.3B A dealer has a warehouse that stores a variety of fruits and vegetables. When fruit is brought to the warehouse, various types of fruit may be mixed together. The dealer wants a model that will sort the fruit according to type. Justify with reasons how machine learning model is efficient compared to feature based classification technique? Make the model with all the criteria? 05

OR

Q.3B Demonstrate how to interpret SPSS Modeler results (for example, using Evaluation node, Analysis node, and data visualizations). Identify SPSS Modeler reporting methods? 05

Q.4A Why use Predictive analytics or Predictive Modeling? Why is getting more important now as these techniques are available from decades and decades ago? 05

Q.4B A regression model is suffering with multicollinearity. How will you deal with this situation without losing much information? 05

Q.5A Why is it wrong to rank the features with all the training set and then run cross validation on nested subsets defined by the ranking to determine the optimum number of features? 05

OR

Q.5B The goal in predictive analysis is to use training data to learn a model that can make predictions on new data. Suppose we increased the size of the training set. Would this likely improve or deteriorate the performance of the model on new data? Why? 05