[Knowledge for Development]

KIBABII UNIVERSITY

UNIVERSITY EXAMINATIONS
2017/2018 ACADEMIC YEAR

END OF SEMESTER EXAMINATIONS
YEAR ONE SEMESTER ONE EXAMINATIONS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
(INFORMATION TECHNOLOGY)

COURSE CODE : PIT 913

COURSE TITLE : ADVANCED DATA ANALYSIS IN COMPUTING

DATE: 19/10/2018 TIME: 9.00 AM -12.00 PM

INSTRUCTIONS TO CANDIDATES

ANSWER QUESTIONS ONE AND ANY OTHER TWO.

SECTION A (COMPULSORY QUESTION)

QUESTION ONE (30 MARKS)

a. In detail, discuss the data analysis process.

(10 Marks)

b. The problem definition starts with high-level questions such as how to track differences in behavior between groups of customers, or what's going to be the gold price in the next month.

Discuss the types of data analysis questions that you may encounter in a typical process.

(10 Marks)

c. What is the role of visualization during data analysis ?

(5 Marks)

d. List and Discuss functionalities of five (5) tools that can be used to perform data analysis.

(5 Marks)

SECTION B (ANSWER ANY 2 QUESTIONS FROM THIS SECTION)

QUESTION 2 (20 Marks)

Write short notes on the following data analysis concepts;

a. Predictive analytics
b. Guestimation
c. Scaling Models
d. Probability Models

(5 Marks Each)

QUESTION 3 (20 Marks)

Much of the raw data contained in databases is unpreprocessed, incomplete, and noisy. For example, the databases may contain: Fields that are obsolete or redundant, Missing values, Outliers, Data in a form not suitable for data mining models, Values not consistent with policy or common sense.

a. To be useful for data mining purposes, the databases need to undergo preprocessing. Why is this necessary? Explain.

Data Analysis

(10 Marks)

b. Describe at least Two (2) ways that the preprocessing can be accomplished.

(10 Marks)

QUESTION 4 (20 Marks)

Briefly, explain the following data mining Techniques, For each give examples of where and how they can be used.

a. Decision Trees
b. Neural Networks
c. k-Nearest Algorithm
d. Hierarchical and k-Means Clustering

(4 Marks Each)

QUESTION 5 (20 Marks)

When dealing with univariate data, we are usually mostly concerned with the overall *shape* of the distribution. Discuss some of the initial questions we may ask.

When we are dealing with a data set that consists of *two* variables (that is, a *bivariate* data set), we are mostly interested in seeing whether some kind of relationship exists between the two variables and, if so, what kind of relationship this is. Plotting one variable against another is pretty straightforward.

Discuss some of the tools and transformations that can be applied to characterize the nature of the relationship between the two inputs.

(20 Marks)