SRN | | | | | | | | | | | | |

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE17CS322
(NKS/SJ)

END SEMESTER ASSESSMENT (ESA) B.TECH. 5th SEMESTER-December 2019

UE17CS322: DATA ANALYTICS

Time: 3 Hrs | Answer All Questions | Max Marks: 100

1a	Data quality can be assessed in terms of accuracy, completeness, and consistency. Propose other dimensions of data quality.	6
b	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. i. What is the mean of the data? What is the median? ii. What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.). iii. What is the midrange of the data?	6
c	In many applications, new data sets are incrementally added to the existing large data sets. Thus an important consideration for computing descriptive data summary is whether a measure can be computed efficiently in incremental manner. Use count, standard deviation, and median as examples to show that a distributive or algebraic measure facilitates efficient incremental computation, whereas a holistic measure does not.	8
2a	Use a flow chart to summarize the following procedures for attribute subset selection: Stepwise Backward elimination.	6
b	Explain different sources from variables Xj can come in Linear Regression model and least squares.	6
c	Write an algorithm for Multiple regression by Successive Orthogonalization.	8
3a	Write a flowchart for Model identification in ARIMA model	6
b	Based on the plots on Daily demand for omelette at Die Another Day (DAD) hospital for the past 115 days. Conclude that the time series is stationary or not. Develop an appropriate ARIMA model that DAD hospital can use for forecasting demand for Omelette.	6

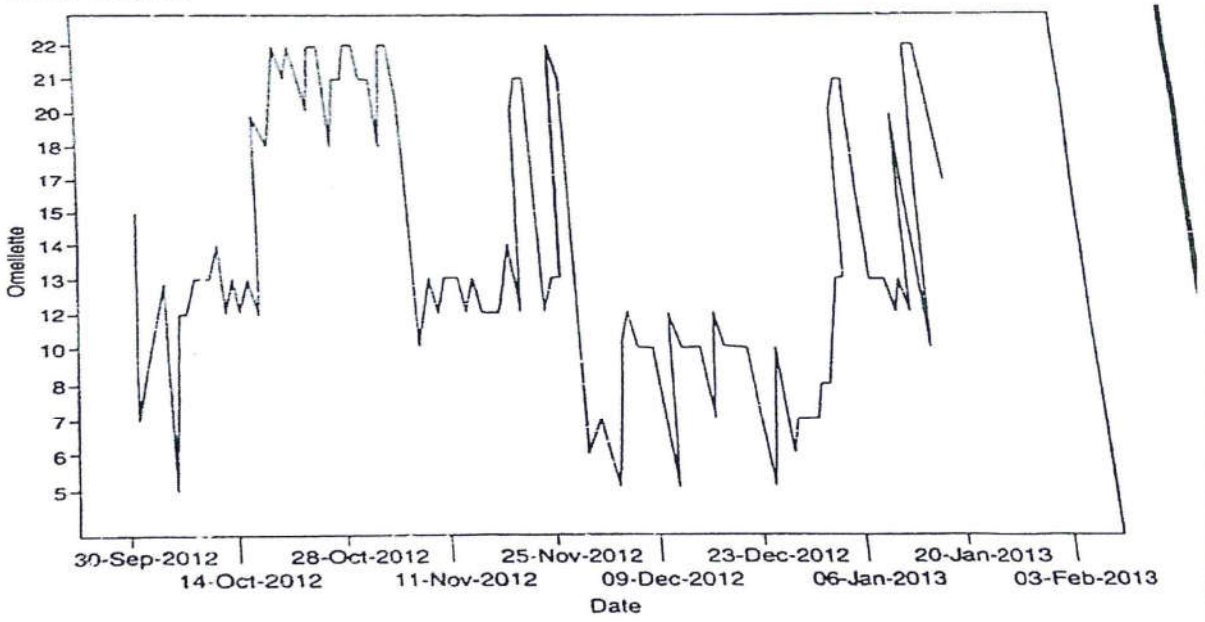

Fig1: Time series plot of demand for Omelette at DAD hospital

SRN

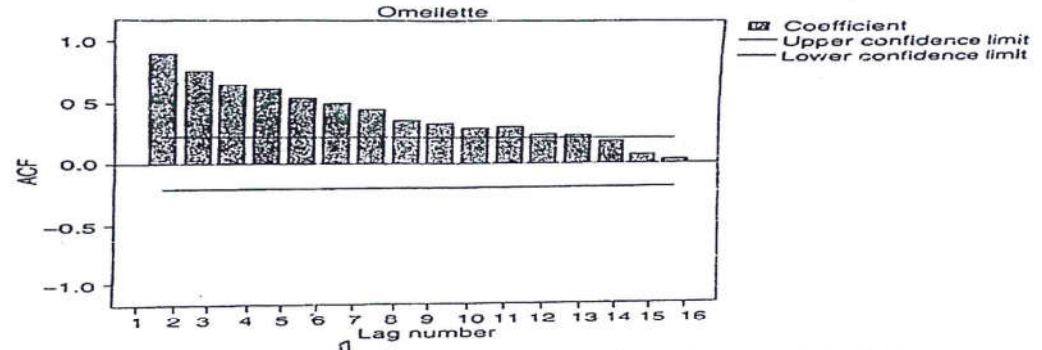

FIGURE 1 ACF plot of demand for Omelette at DAD hospital.

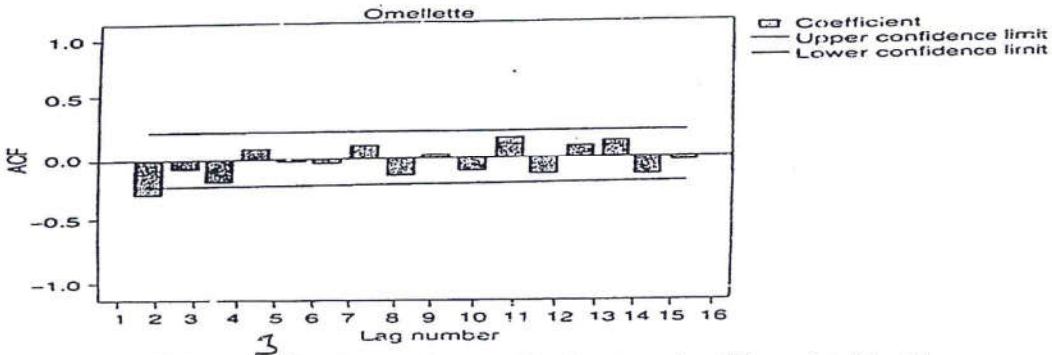

FIGURE 3 ACF plot of demand for Omelette after differencing ($d = 1$).

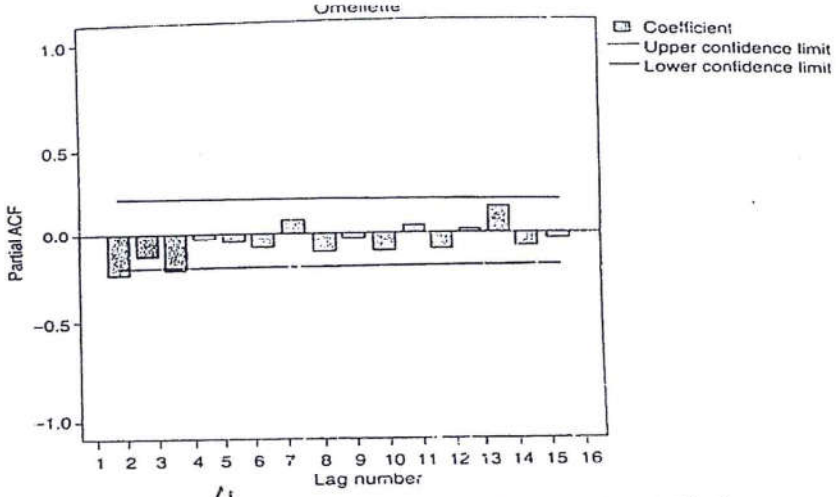

FIGURE 4 PACF plot of demand for Omelette after differencing ($d = 1$).

c	Explain in detail AR Model Identification using ACF and PACF.	8
4a	Briefly explain two basic architecture for a recommendation system.	6
b	Derive the equation for support vectors H_1 and H_2 (Hyperplanes) for SVM.	6

SRN

c

Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Let *D* be the training data shown in the above Table , where there are nine tuples belonging to the class *buys_computer = yes* and the remaining five tuples belong to the class *buys_computer =no*. A (root) node *N* is created for the tuples in *D*.

i. Compute Gini index for impurity of *D*.

ii. Compute Gini index for the attribute income for splitting criterion.

8

5a A meteorologist studying the weather in a region decides to classify each day as simply sunny or cloudy. After analyzing several years of weather records, he finds:

- the day after a sunny day is sunny 80% of the time, and cloudy 20% of the time; and
- the day after a cloudy day is sunny 60% of the time, and cloudy 40% of the time

i. Setup a Markov Chain to modal this process.

ii. Write a transition matrix

6

b What are the research Studies to detect relationships

6

c Compute the stationary distribution for the transition matrix

$$\mathrm{P} = \begin{pmatrix} 0.9 & 0.7 & 0.8 \\ 0.05 & 0 & 0 \\ 0.05 & 0.3 & 0.2 \end{pmatrix}$$

8