DHARMSINH DESAI UNIVERSITY, NADIAD
FACULTY OF TECHNOLOGY
B.TECH. SEMESTER VII IT

SUBJECT: (IT-704) Data Analysis & Information Extraction

Examination	**: First sessional**	**Seat No.**	**: ________________**
Date	**: 02/08/2018**	**Day**	**: Thursday**
Time	**: 2:30 to 3:45 PM**	**Max. Marks**	**: 36**

INSTRUCTIONS:

1. Figures to the right indicate maximum marks for that question.
2. The symbols used carry their usual meanings.
3. Assume suitable data, if required & mention them clearly.
4. Draw neat sketches wherever necessary.

Q.1 Do as directed. **[12]**

(a) Explain the difference between data and information. **[2]**
(b) Define the term measure in data mining? Also mention various types. **[2]**
(c) Explain what is entity identification problem. Also give an appropriate example. **[2]**
(d) What is the histogram analysis? Also mention the different rules in brief. **[2]**
(e) State which architecture of data mining system is most popular and why? **[2]**
(f) Define the following terms: 1)Utility, 2)Certainty. **[2]**

Q.2 Attempt ***Any Two*** from the following questions. **[12]**

(a) Draw the star schema diagram for the university data warehouse. (Identify at least 5 dimensions and 2 measures.) **[6]**
(b) What is a KDD process? Explain the various steps in brief. **[6]**
(c) Explain the concept hierarchy and where it is used? Give an appropriate example. **[6]**

Q.3 (a) What is a data mining primitive. Explain the various types of data mining primitives. **[6]**
(b) Why do we use Sampling technique? Explain Sampling technique in details. **[6]**

OR

Q.3 (a) Explain what is data transformation and define various methods for data normalization. **[6]**
(b) Use the 3-4-5 rule for the automatic construction of a numeric hierarchy. **[6]**
Suppose that profits at different branches of All Electronics for the year 1999 covers a wide range, from -$351,976.00 to $4,700,896.50. A user wishes to have a concept hierarchy for profit automatically generated. For improved readability, we use the notation (l…r] to represent the interval (l, r]. For example, (-$1,000,000…$0] denotes the range from -$1,000,000(exclusive) to $0(inclusive). Suppose that the data within the 5th percentile and 95th percentile are between -$159,876 and $1,838,761.