Seat No.: _____                                          Enrolment No._____

# GUJARAT TECHNOLOGICAL UNIVERSITY
## MBA - SEMESTER– III EXAMINATION – WINTER 2019

**Subject Code: 4539251**                                      **Date: 03-12-2019**

**Subject Name: Data Warehousing and Data Mining**

**Time: 10:30 AM TO 1.30 PM**                                  **Total Marks: 70**

**Instructions:**
1. **Attempt all questions.**
2. **Make suitable assumptions wherever necessary.**
3. **Figures to the right indicate full marks.**

**Q.1** **(a)** Give definition of following terms:                                      **14**

    1. Normalization

    2. Data warehouse

    3. Data mining

    4. Metadata

    5. Lazy learners

    6. Outlier

    7. Web mining

**Q.2** **(a)** Explain various features of SQL.                                      **07**

    **(b)** Explain Boyce – Codd Normal form (BCNF) with proper example.     **07**

**OR**

    **(b)** Explain Data manipulation language with proper example.         **07**

**Q.3** **(a)** Explain difference between DWH and OLTP based database management system environments.     **07**

    **(b)** What is dimensional modeling? Explain various steps involved in dimensional modeling.     **07**

**OR**

**Q.3** **(a)** Explain data warehouse design process in detail.                     **07**

    **(b)** Explain various reporting and query tools in detail.            **07**

**Q.4** **(a)** What is cluster analysis? Explain applications of cluster analysis.   **07**

    **(b)** Explain K means clustering method in detail.                   **07**

**OR**

**Q.4** **(a)** Explain decision tree induction in detail with proper examples.      **07**

    **(b)** Explain naïve Bayesian classification in detail.               **07**

**Q.5** Healthcare provider organizations are faced with a rising number of financial pressures. Both administrators and physicians need help analyzing large numbers of clinical and financial data when making decisions. To assist them, Rush-Presbyterian–St. Luke's     **14**

Medical Center and Hitachi America, Ltd. (HAL), Inc., have partnered to build an enterprise data warehouse and perform a series of case study analyses. For the initial level they used Microsoft OLAP Services for the multidimensional database server and Knosys ProClarity to do the reporting, that is, to display grids and graph. Typical problems that data mining addresses are how to classify data, cluster data, find associations between data items, and perform time series analysis. Numerous data mining techniques have been invented for each type of problem.

Each problem requires data mining techniques to analyze large quantities of data. Two techniques for data mining were used: patient rule induction method (PRIM) and weighted item sets (WIS), a type of association rule technique. PRIM and WIS are described next. PRIM is a technique that does not fall exactly into one of the business problem categories listed earlier. PRIM finds the optimal region, that is, a subset of data points with the highest average value, given a set of input attributes and a minimum size of the region specified by the user. Data records contain input variables and an output variable (variables are record attributes or derived attributes, and the output variable must be a measure). A record's location in a dimensional space is based on the value of its attributes, for example, "attending physician," "payer," and "LOS" in a hospital database. PRIM finds regions where the output variable has a high average value compared to the average value for the entire set of records. PRIM could also be used to find regions with minimum average value by maximizing the negative values of the output variable. WIS is an association rule tool that finds relationships between various attributes in a database; some of the attributes can be derived measures. The relationships are defined in terms of if-then rules that show the frequency of records appearing in the database that satisfies the rule. For example, ninety out of one hundred patients in the database with DRG "999" have a length of stay greater than or equal to ten days.

**(a)** Discuss the pros and cons of data warehousing method used by HAL

**(b)** Discuss the usefulness of various data mining outcomes with consideration of cost, quality and betterment of patients.

**OR**

**Q.5** **(a)** Can you suggest better data warehousing methods? Explain with examples. **07**

**(b)** Considering the two methods, Which types of other data mining possibilities are there? **07**

**\*\*\*\*\*\*\*\*\*\*\*\***