INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

Mid-Semester Examination, February 2018

Program Code & Semester: B.Tech.(IT) – 6th Semester
Paper Title: IDMW632C

Paper Setter: Dr.Manish Kumar &Dr. Ranjana Vyas

Max Marks: 30 **Duration: 2 Hours**

1. The following table consists of training data from an employee database. The data have been generalized. For example, "31 : : : 35" for *age* represents the age range of 31 to 35.For a given row entry, *count* represents the number of data tuples having the values for *department, status, age*, and *salary* given in that row.

department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

Let *status* be the class label attribute.

a) How would you modify the basic decision tree algorithm to take into consideration the *count* of each generalized data tuple (i.e., of each row entry)?

[05]

b) Use your algorithm to construct a decision tree from the given data by suggesting how the Training and Test Data will be finalized. **[05]**

2. a) Draw diagram showing steps of KDD process.

b) Discuss five major issues in data mining process. **[4+4]**

3. A database has five transactions. Let *min sup* D 60% and *min conf* D 80%.

TID	items_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

a) Find all frequent item-sets using Apriori and FP-growth respectively and compare the efficiency of the two mining processes. **[06]**

b). List all the *strong* association rules (with support *s* and confidence *c*) matching the following metarule, where *X* is a variable representing customers, and *itemi* denotes variables representing items (e.g., "*A,*" "*B,*"): *buys*.(*X*, *item*1) ^ *buys*(*X*, *item*2)=>*buys*(*X*, *item*3) [*s*, *c*]. **[04]**

c) How would the ARM results can help creating appropriate discounting policy for super market business Analyst, make necessary assumptions to address the limitations of the ARM (Ex. Rare Item Problem, Min Sup Problem etc) to propose an Utility based mining rather than only frequency based one. **[02]**