Unit – I

1 – Introduction to Data Mining

1) What is data mining? Why there is need of data mining? Who can use data mining?
2) List and explain applications of data mining?
3) Draw and explain Data mining Process?
4) Explain data mining techniques?
5) Differentiate between Data mining and Machine learning?

1. What is the full form of KDD _____

2. Which learning approach is used by Database Segmentation?
(a) Supervised Learning (b) Unsupervised Learning

3. Links between the individual record, or sets of records in a database is called _____

4. What are the two types of predictive modeling?
(a) _____ (b) _____

5. Deviation detection can be performed by using _____ and _____ techniques.

6. Predictive Modeling is developed using a supervised learning approach.
(a) True (b) False

7. Data mining is
_____
(a) The process of automated discovery of previously unknown patterns in large volumes of data.
(b) The stage of selecting the right data for a KDD process
(c) A subject-oriented integrated time variant non-volatile collection of data in support of management
(d) None of these

8. Value prediction uses the traditional statistical techniques of _____ and _____

9. Classification falls under which technique of data mining?
(a) Predictive modeling (b) Database segmentation
(c) Link analysis (d) Deviation detection.

10. Regression falls under which technique of data mining?
(a) Predictive modeling (b) Database segmentation
(c) Link analysis (d) Deviation detection.

11. Clustering falls under which technique of data mining?
(a) Predictive modeling (b) Database segmentation
(c) Link analysis (d) Deviation detection.

12. Visualization is core part for which of following data mining technique?
(a) Predictive modeling (b) Database segmentation
(c) Link analysis (d) Deviation detection.

13. Association mining falls under which technique of data mining?
(a) Predictive modeling (b) Database segmentation
(c) Link analysis (d) Deviation detection.

Answer Keys:
1. Knowledge Discovery in Database; 2. (b); 3. Associations;
4. (a) Classification (b) Value prediction 5. Statistics, Visualization
6. (a); 7. (a); 8. Linear regression, Nonlinear regression;
9. (a); 10. (a); 11. (b); 12. (d);

2 – Data Preprocessing

1. What is Data Preprocessing? Why there is need of Data Preprocessing?
2. Draw and Explain Data Cleaning method of Data Preprocessing?
3. Draw and Explain Data Integration method of Data Preprocessing?
4. Draw and explain Data transform method of Data Preprocessing?

5. What is the difference between equal interval binning and equal frequency binning? Explain with
6. examples.

3 - Classification

1. What is classification? Explain binary and multiclass classification?
2. Explain types of classification (Posteriori classification and Priori classification)
3. Explain the concept of Input and Output in classification with example?
4. Explain the concept of information theory with example?
5. What is the role of information theory in building the decision tree?
6. What is decision tree? Explain the concept of Gini index with example?
7. What are the advantages and disadvantages of decision tree classification?
8. Why is Naïve Bayes called naïve? Explain the working of the NB algorithm by proving the Bayes theorem.
9.

Identify the attribute that will act as the root node of a decision tree to predict golf play for following database with Gini Index. Indicate all the intermediate steps.

Outlook	Wind	PlayGolf
rain	strong	no
sunny	weak	yes
overcast	weak	yes
rain	weak	yes
sunny	strong	yes
rain	strong	no
overcast	strong	no

10.

Make a decision tree for the following database using Gini Index. Indicate all intermediate steps.

Example	*Colour*	*Shape*	*Size*	*Class*
1	Red	Square	Big	+
2	Blue	Square	Big	+
3	Red	Circle	Big	+
4	Red	Circle	Small	–
5	Green	Square	Small	–
6	Green	Square	Big	–

11.Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

A	*B*	*Class Lable*
T	F	+
T	T	+
T	T	+
T	F	–
T	T	+
F	F	–
F	F	–
F	F	–
T	T	–
T	F	–

12.

Predict that X will buy the computer or not, for following training database by using Naïve Bayes.
X= (age = 31...40, income = medium, student = yes, credit-rating = fair)

rec	*Age*	*Income*	*Student*	*Credit_rating*	*Buys_computer*
r1	<=30	High	no	Fair	no
r2	≤30	High	no	Excellent	no
r3	30...40	High	no	Fair	yes
r4	>40	Medium	no	Fair	yes
r5	>40	Low	yes	Fair	yes
r6	>40	Low	yes	Excellent	no
r7	30...40	Low	yes	Excellent	yes

rec	*Age*	*Income*	*Student*	*Credit_rating*	*Buys_computer*
r8	<=30	Medium	no	Fair	no
r9	<=30	Low	yes	Fair	yes
r10	>40	Medium	yes	Fair	yes
r11	<=30	Medium	yes	Excellent	yes
r12	30...40	Medium	no	Excellent	yes
r13	30...40	high	yes	Fair	yes
r14	>40	Medium	no	Excellent	no