**Unit -1(Feb – 2022, November—2023, )**

1. Elaborate different criteria on the basis of which data mining Techniques are classified
2. By Applying Min Max Normalization transform the 73000 with minimum and maximum values for income attribute 12000 and 98000 respectively to new range [0.0 , 1.0]
3. Elaborate the identification problem faced while integrating the data
4. With an apt example difference between supervised and unsupervised discretization
5. Define Concept hierarchy and its types ? Explain How it can be used in data reduction with example
6. How Sampling of data can improve Quality of data mining ? Difference Between Simple random Sampling without replacement and with replacement
7. How efficient and scalable frequent itemsets can be mined from given data set by apriori algorithm ? Explain the importance of prune steps in improving its efficiency
8. What do you mean by pattern evaluators ? formulate any two evaluators
9. How binning can be used for removing noise from given data
10. Explain the following problems along with one solution faced during data integration : schema integration and redundancy

Unit – 2

1. State the importance of training and testing phase of any classification approach
2. What important role does bias value and weight play in Multilayer feed feed forward neural network model? Explain
3. Different method that can be used evaluating classifiers along with their formula and examples
4. Different phases of classifier model with importance of training and testing phase
5. Formulate the following Classifier Evaluators: accuracy, specificity and sensitivity with apt example
6. How agglomerative clustering approach can be implemented using dendograms
7. Which method is used for clustering high dimensional data
8. How centoid based technique i.e k-means method is used for clustering the given data by applying it recursively
9. How DBSCAN method can be used to create arbitrary shaped clusters by using density reachable and density connected points
10. Difference between active and lazy learners with two examples of each
11. How bagging and boosting methods can be used to improve classifiers performance
12. Elaborate how DBSCAN density-based algorithm can be used for handling noise effectively

Unit – 3

1. Explain different types of digital data along with their data access methods and data management mechanism respectively of each type
2. Five differences between structured and un structured data also explain two access method of each type
3. Check pointing and heart beat signals meant for HDFS
4. Purpose of maintaining edit logs and fsi image by secondary nodes in HDFS cluster
5. With diagram detail the anatomy followed for HDFS data flow for file write operations in HDFS
6. Detail Hadoop architecture for handling user query and roles performed by job tracker and task tracker in handling the query
7. YARN has introduced in Hadoop 2.0 for enhancement the performance of MapReduce comment
8. How data is compressed and copied from local to HDFS and HDFS to local in Hadoop environment
9. For proper Fault tolerance system? Rack awareness is very Important? Support your Answer
10. Elaborate 7V’s responsible for making data handling and challenging task
11. What do you mean by serialization and deserialization and its importance in Hadoop Working environment
12. Elaborate the HDFS architecture and list down different daemons working in HDFS Cluster

Unit – 4

1. Detail Working of three major components Code Driver, Mapper and Reducer of Map Reduce framework
2. Purpose of using Dump Statement and for each operator in PIG
3. Explain Sorting, Shuffling, Spilling of Data carried out in Map reduce Phase
4. Map Reduce Types and formats
5. What makes Hadoop ecosystem? brief working of any five such components
6. Components of Pig Environment? how pig coding converted into MapReduce
7. How data is managed in hive tables top over HDFS in terms of column family, map and time stamps data types
8. Mention any four features provided by R software tool for analysing and visualizing big data
9. List various functions performed by Pig execution engine to successfully run the pig queries
10. Why partitions and buckets are created in Hive? Explain with the help of example. Adding, renaming, and dropping a partition in hive
11. How and why HBase is Gaining popularity Over HDFS data storage cluster for specific applications
12. Elaborate how a request from a client is propagated in HBase architecture that is handled by its different Components